THE HORIZON OF REWARD GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference feedback can leave parts of the reward function unidentified even when it is sufficient to support a useful policy update. As the policy changes, previously irrelevant ambiguity can become important for deciding how the policy should be updated. Existing pointwise analyses do not account for whether the reward mechanism can adapt fast enough as the policy changes. This makes continued guidance a temporal question: how long can existing preference information and finite reward adaptation continue to certify positive progress? We call this duration the support horizon. In a scalar model, we characterize the support horizon exactly and prove that pointwise feasibility can persist after every admissible reward trajectory has lost the certificate. We further show that this separation can arise from the policy dynamics themselves, including locally under ordinary Euclidean policy gradient. The analysis also shows how additional preference data can extend certified guidance. Rather than requiring full reward identification, it isolates the uncertainty that limits a longer horizon; under Bradley–Terry feedback, resolving this uncertainty requires $\Theta(d_T^{-2}\log(1/\alpha))$ labels, where d_T is the distance to the feasibility boundary.

1 INTRODUCTION

Preference feedback can be sufficient to guide the current policy while leaving some reward directions unresolved. As the policy changes, those unresolved directions can become relevant to its update. The resulting question is not only whether reward predictions remain accurate on new outputs, but whether existing preference information continues to support useful policy changes as training proceeds.

Prior work on partial identifiability asks whether unresolved reward directions still permit reliable downstream decisions (Skalse et al., 2023; Lazzati & Metelli, 2026), while work on reward shift studies how reward estimation changes with the policy distribution (Ackermann et al., 2025). These works do not characterize whether finite reward adaptation can track a moving set of certifying rewards as the policy changes. Even if a suitable reward exists at every policy state, the reward mechanism starts from its current parameters and can adapt only finitely fast. Pointwise feasibility can therefore overstate how long existing preference information remains useful. The set of rewards that certify progress may keep moving while the deployed reward eventually fails to track it.

To address this gap, we formulate continued reward guidance as a constrained reachability problem and introduce the support horizon, the longest interval over which existing preference information and finite reward adaptation can jointly certify progress. At each policy state, we use the source preference information to determine which reward parameters induce policy updates that are guaranteed to make progress for every compatible preference. As the policy changes, this certified set moves, while the reward parameters are constrained by a finite adaptation budget. This view also connects dynamics to data acquisition. Once a target horizon is no longer certified, the analysis identifies the unresolved preference uncertainty that prevents certification and therefore what additional feedback must resolve. Figure 1 previews this mechanism: a source-based trajectory certificate can expire while pointwise guidance remains feasible, and targeted feedback can extend certification for the same reward trajectory.

Our main contributions are summarized below.

- We derive an exact support horizon in a scalar model and prove that pointwise feasibility can persist after every admissible reward trajectory has lost the certificate.

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.

- We show that this separation is not confined to an externally imposed scalar drift. It arises endogenously in a finite policy and reward system and also occurs locally under ordinary Euclidean policy gradient.
- We show how additional preference data can extend certified guidance. Rather than requiring full reward identification, the scalar analysis isolates the uncertainty that limits a longer horizon. Under Bradley–Terry feedback, resolving this uncertainty requires $\Theta(d_T^{-2}\log(1/\alpha))$ labels, where d_T is the distance to the feasibility boundary.

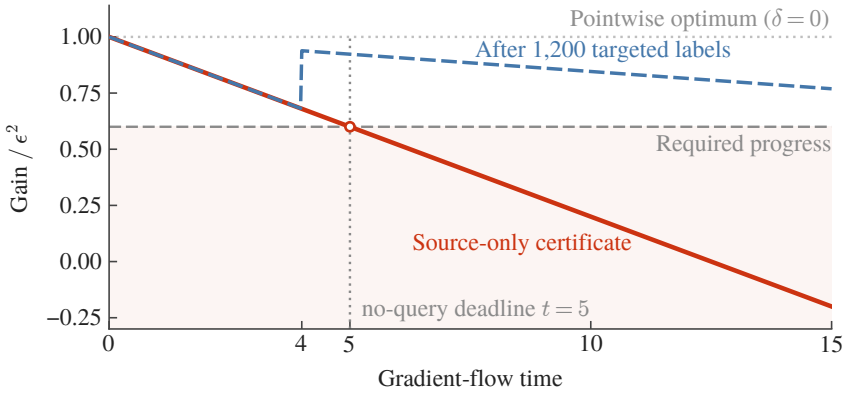

Figure 1: Preview of the central mechanism. The gray dotted curve is the pointwise optimum, which remains above the required progress level. The red curve is the source-only certificate along a fixed reward trajectory and drops below the requirement at the support horizon. After adding targeted preference labels, the blue dashed curve certifies the same trajectory for longer. Thus certification can fail even when pointwise guidance remains feasible, and additional feedback can extend certification without changing the underlying trajectory.

2 Related Work

Work on partial identifiability shows that reward ambiguity need not prevent reliable downstream decisions. Skalse et al. (2023) characterize invariances induced by different forms of reward feedback, while Lazzati & Metelli (2026) study robust decision making when rewards are only partially identified. Our question begins where this static conclusion ends. As the policy changes, an ambiguity that is harmless for the current decision can become relevant later, while the deployed reward has only finite ability to adapt. We therefore ask when pointwise downstream feasibility extends to a guidance trajectory that remains reachable over time.

A changing policy also appears in work on reward shift and deployment generalization. OCRM corrects reward estimation under the distribution induced by the current policy through importance weighting (Ackermann et al., 2025), while Tang et al. (2026) decompose RLHF generalization error into sampling, reward shift, and KL clipping terms. These approaches study prediction or estimation as the policy distribution changes. Our focus is instead on whether fixed preference information can continue to support a realizable guidance trajectory when the reward mechanism itself has limited adaptation.

Once dynamic reachability determines why a target horizon can no longer be certified, a natural next question is what additional preference information is needed to extend it. Existing preference acquisition methods already select feedback according to downstream value or information (Lindner et al., 2021; Feng et al., 2025; Shen et al., 2025), with confidence sets and pessimistic decision rules providing related foundations (Zhu et al., 2023). Our dynamic analysis identifies which unresolved preference uncertainty actually limits continued guidance, turning the control problem into a targeted statistical question rather than requiring complete reward identification. Appendix C treats the related but distinct problem of exact identifiability under a restricted comparison graph.

Our analysis also draws on ideas from safe learning and statistical information theory. Control barrier methods combine certificates with finite control authority (Ames et al., 2017; Taylor et al., 2020;

Capone et al., 2025), while change of measure arguments provide standard tools for lower bounding the information needed to distinguish alternatives (Kaufmann et al., 2016; Garivier & Kaufmann, 2016). In our setting, the control problem determines which reward alternatives threaten continued guidance, and the statistical analysis quantifies how difficult those alternatives are to distinguish from the true preference.

3 Problem Setting for Preference Information, Guidance, and Adaptation

We formalize the picture from the introduction. A guidance reward acts through its induced policy change, source preferences determine which changes can be certified, and finite adaptation determines whether certification can persist as the policy evolves.

3.1 How a reward acts on a policy

Because we study guidance rather than reward prediction, we characterize a guidance reward through the policy change it induces. Let $p_\theta \in \Delta_n^\circ$ denote a policy over a finite output space and $r \in \mathbb{R}^n$ a fixed true reward vector. With fixed guidance features Ψ, write

$$R_\phi = \Psi\phi, \qquad J_r(\theta) = r^\top p_\theta. \tag{1}$$

We consider a declared population policy gradient update

$$\dot\theta = M(\theta) D_\theta^\top \Psi\phi, \qquad D_\theta = \partial_\theta p_\theta, \tag{2}$$

where M is known, smooth, symmetric, and positive definite; ordinary Euclidean policy gradient has $M = I$. Define

$$B_\theta = D_\theta M D_\theta^\top \Psi, \qquad b = B_\theta \phi. \tag{3}$$

Then $b = \dot p$ and

$$\dot J_r = r^\top b. \tag{4}$$

Thus a guidance reward acts through b, whose true gain is $r^\top b$. Because source feedback does not identify r exactly, we next certify this gain over all compatible rewards.

3.2 Preference information and certified gain

Because the true reward is only partially identified, we evaluate each induced policy change against every reward compatible with the source feedback. Fix the additive reward gauge, for example in $\mathcal{H} = \mathbf{1}^\perp$, and let $\mathcal{U}$ be a nonempty compact convex compatibility set, or a confidence set with valid coverage. Define

$$L_{\mathcal{U}}(b) = \inf_{r \in \mathcal{U}} r^\top b. \tag{5}$$

For exact linear feedback $A_0 r = y_0$, let $\mathcal{K}_0 = \ker A_0$. If the unidentified component lies in a radius-η_0 ball, then

$$L_{\mathcal{U}}(b) = \hat r_0^\top b - \eta_0 \left\| \Pi_{\mathcal{K}_0} b \right\|. \tag{6}$$

This identity makes the role of ambiguity explicit: certified gain falls when the update relies on unidentified preference directions. It therefore provides a robust certificate at the current policy state. We next maximize this certificate pointwise before imposing dynamic reachability.

3.3 From pointwise capability to trajectory feasibility

Let Φ denote the admissible guidance parameters. At a fixed policy state, the best certified progress is

$$V(\theta, \mathcal{U}) = \max_{\phi \in \Phi} L_{\mathcal{U}}(B_\theta \phi). \tag{7}$$

This is the pointwise capability, but it ignores whether the deployed reward can reach the maximizing parameters.

Continued guidance adds this requirement. Fix a horizon $[0, T]$, required progress $\gamma(t) > 0$, an initial guidance reward, and either a speed budget $\|\dot\phi\| \le \rho$ or displacement budgets $\|\phi_{k+1} - \phi_k\| \le s_k$.

Because ϕ_t drives θ_t through Equation equation 2, they form a coupled controlled path. The trajectory certificate is

$$L_{\mathcal{U}}(B_{\theta_t}\phi_t) \geq \gamma(t), \qquad 0 \leq t \leq T. \tag{8}$$

Thus even if

$$V(\theta_t, \mathcal{U}) > \gamma(t)$$

at every time, no trajectory from the declared initialization need maintain the certificate. The following sections study this gap between pointwise availability and trajectory reachability.

The reachability problem therefore depends jointly on the policy dynamics, reward initialization, and adaptation class; changing any of them changes the support horizon. Appendix A.1 summarizes the comparison class.

4 A Finite Support Horizon

The support horizon reflects a mismatch between the motion of the certified reward set and finite reward adaptation. We derive a movement budget from finite updates, solve the resulting scalar reachability problem exactly, then state a general moving-interval criterion and realize the mechanism endogenously in a finite policy system.

4.1 Finite reward updates imply a movement budget

For projected Bradley–Terry–Luce updates (Bradley & Terry, 1952; Luce, 1959) with feature differences $\|d_i\| \leq L$, a step of size η moves the reward parameters by at most ηL. Thus m_k updates give

$$s_k = \sum_{j=1}^{m_k} \eta_{k,j} L_{k,j}. \tag{9}$$

Appendix A gives the derivation. This is a capability envelope, not an optimizer trajectory. Impossibility over the full envelope rules out every algorithm within it, whereas feasibility need not be attained by a particular optimizer. Appendix A.4 gives an example in which startup transients shorten the realized horizon despite the same maximum movement rate.

We next ask whether any path within this budget can track the moving certified set.

4.2 Exact support horizon in a scalar model

We isolate the mechanism with one unresolved preference direction and one adaptive reward coefficient. Suppose the source data identify a common normalized gain $g > 0$ but leave an offset $\delta \in [-\eta_0, \eta_0]$ unidentified. Consider

$$j_\delta(t) = g + \delta[\phi(t) - vt], \qquad \phi(0) = 0, \qquad |\dot{\phi}| \leq \rho < v. \tag{10}$$

Here vt is the moving reward target and $\phi(t)$ can move at speed at most ρ. Since $v > \rho$, every admissible trajectory accumulates lag.

Theorem 1 (Exact support horizon and separation of pointwise and trajectory feasibility)**.** *Require $j_\delta(t) \geq \gamma$ for every compatible δ and every $t \leq T$, where $0 < \gamma < g$, and assume the amplitude constraint does not bind first. Let $\Delta = g - \gamma$. Then the longest support horizon is*

$$\boxed{T_0 = \frac{\Delta}{\eta_0(v - \rho)}}. \tag{11}$$

The path $\phi(t) = \rho t$ attains the bound. For a declared horizon T, the minimum required speed is

$$\rho_{\min}(T) = \max\left\{0,\ v - \frac{\Delta}{\eta_0 T}\right\}. \tag{12}$$

At the same time, if the amplitude set contains $\phi = vt$, the pointwise value remains $V_t = g > \gamma$ at every t.

Proof. The worst compatible gain is $g - \eta_0|\phi - vt|$. Every admissible path satisfies $\phi(t) \le \rho t$, so its lag behind vt is at least $(v - \rho)t$. The certificate fails once that lag reaches Δ/η_0, and $\phi(t) = \rho t$ achieves the smallest possible lag at every time. □

Theorem 1 gives the central separation. The pointwise optimum stays above the required progress level, yet every admissible trajectory loses certification at T_0 because the target moves faster than adaptation permits. Linear drift gives the closed form, but the mechanism only requires a moving certified set and bounded reward speed.

4.3 A GENERAL CRITERION FOR A MOVING FEASIBLE INTERVAL

The reachability logic does not depend on linear drift. A certified set can be nonempty at every time yet admit no bounded-speed path. Suppose the reward values satisfying the current certificate form a known interval $I_t = [\ell(t), u(t)]$, with $I_0 = \{\phi_0\}$. There exists an absolutely continuous path $\phi(t) \in I_t$ with $|\dot{\phi}| \le \rho$ if and only if

$$\ell(s) \le u(t) + \rho|s - t|, \qquad \forall s, t \in [0, T]. \tag{13}$$

This is a cross-time condition rather than pointwise nonemptiness. In discrete time, cumulative displacement replaces $\rho|s - t|$; Appendix A gives the proof and constructive path.

The criterion assumes an externally specified feasible interval. Endogenous intervals require additional analysis, motivating the next construction.

4.4 AN ENDOGENOUS REALIZATION WITH A FINITE POLICY AND REWARD

The moving target need not be imposed externally. Appendix A constructs an explicit five-output policy with full support and a smooth positive definite preconditioner for which

$$\dot{c} = v, \qquad \dot{J}_{r_\delta} = \epsilon^2[1 + \delta(\phi - c)]. \tag{14}$$

Thus the policy dynamics generate the moving target. The same construction leaves the relevant preference direction unresolved by source feedback while providing an informative comparison used later.

The construction still uses a designed preconditioner. The next section asks whether the same separation can arise under ordinary Euclidean policy gradient.

5 A LOCAL SEPARATION UNDER ORDINARY POLICY GRADIENT

The preceding construction establishes realizability but leaves open whether the effect depends on the chosen preconditioner. We therefore ask whether bounded reward adaptation can lose a moving progress certificate under ordinary Euclidean policy gradient.

We use a two-coordinate policy in which the guidance reward controls how policy improvement is allocated between two directions, while the source feedback leaves that tradeoff unresolved. Consider the independent Bernoulli policy on $(q, s) \in \{-1, 1\}^2$,

$$p_\theta(q, s) = \tfrac{1}{4}(1 + xq)(1 + ys), \qquad x = \tanh\theta_q, \quad y = \tanh\theta_s, \tag{15}$$

with ordinary $M = I$ and guidance reward

$$R_\phi = \phi q + (1 - \phi)s, \qquad 0 \le \phi \le 1. \tag{16}$$

The source feedback calibrates only outputs with $q = s$, while the declared true reward class is

$$r_a = aq + (1 - a)s, \qquad a \in [-1, 2]. \tag{17}$$

Let

$$\omega_q = (1 - x^2)^2, \qquad \omega_s = (1 - y^2)^2. \tag{18}$$

Then $\dot{x} = \omega_q\phi$ and $\dot{y} = \omega_s(1 - \phi)$. Requiring a worst case progress of at least γ yields the safe interval

$$\ell(x, y) = \frac{\omega_s + \gamma}{2\omega_q + \omega_s}, \qquad u(x, y) = \frac{2\omega_s - \gamma}{\omega_q + 2\omega_s}, \tag{19}$$

while the pointwise optimum is $V = \omega_q\omega_s/(\omega_q + \omega_s)$. At each policy state, this interval contains the reward coefficients that certify the required progress. The remaining question is whether a coefficient moving at bounded speed can stay inside it as the policy state itself evolves.

Theorem 2 (Ordinary policy gradient can lose a moving certificate)**.** *There exist $\gamma, \rho > 0$ and a full-support interior initialization that is initially certified such that $V > \gamma$ throughout an open neighborhood, yet every admissible reward trajectory with $|\dot{\phi}| \le \rho$ loses certification in finite time.*

Appendix A gives a concrete instance with $\gamma = .06$ and $\rho = .04$. Every admissible control loses the certificate within 0.0100061 time units, while throughout the proof region

$$V \ge .10575039 > .06, \qquad \min_{q,s} p_\theta(q, s) \ge .04655. \tag{20}$$

The lower safe boundary moves faster than any admissible reward coefficient, exhausting the initial margin before the state can leave the proof region. Appendix A proves this for all admissible controls, and Figure 3 visualizes the contact. Thus the separation does not require a designed preconditioner and persists under sufficiently small perturbations because the inequalities are strict.

6 Preference Information Needed Beyond the Horizon

Beyond the support horizon, failure can reflect either true dynamic infeasibility or unresolved alternatives that make the target horizon infeasible. More labels can help only in the second case, when the true task remains feasible. We therefore ask how much additional preference information is needed to rule out those alternatives and certify a longer horizon.

6.1 Dynamic feasibility induces a statistical threshold

For $T > T_0$, Theorem 1 induces a feasibility threshold in preference space. Define

$$a_T = (v - \rho)T, \qquad \delta_c(T) = \frac{\Delta}{a_T}, \qquad d_T = \delta_c(T) - \delta_\star > 0. \tag{21}$$

For $\delta \le \delta_c(T)$, the common control $\phi = \rho t$ completes the task; for $\delta > \delta_c(T)$, no admissible control does. Thus d_T is the true preference's distance to dynamic infeasibility.

If the source information permits parameters on both sides of the threshold, certification becomes a testing problem. We model each new binary comparison as

$$Y_i \mid \mathcal{F}_{i-1} \sim \text{Bernoulli}\{\sigma(c_i + q_i\delta)\}, \qquad |q_i| \le Q. \tag{22}$$

Queries may adapt to past observations but are chosen before their current labels. For $0 < \alpha < 1/2$, a certifying procedure must accept the feasible truth with probability at least $1 - \alpha$, accept every infeasible parameter with probability at most α, and return an admissible control upon acceptance. Source information is included in the KL accounting; Appendix B gives the general form.

Theorem 3 (Preference complexity for a given instance)**.** *For $T > T_0$ and feasible $\delta_\star < \delta_c(T)$, every procedure satisfying the reliability conditions above obeys*

$$\boxed{\mathbb{E}_{\delta_\star} \sum_{i=1}^{\tau} q_i^2 \ge \frac{8}{d_T^2} \operatorname{kl}(1 - \alpha, \alpha)} \tag{23}$$

and consequently

$$\mathbb{E}_{\delta_\star} \tau \ge \frac{8}{Q^2 d_T^2} \operatorname{kl}(1 - \alpha, \alpha). \tag{24}$$

Conversely, if the query pool contains a fixed pure comparison with sensitivity $q > 0$ and $|\delta| \le \eta_0$, let $\kappa = \min_{|z| \le q\eta_0} \sigma'(z) > 0$. A fixed batch succeeds with the stated reliability whenever

$$N \ge \frac{2 \log(1/\alpha)}{(\kappa q d_T)^2}. \tag{25}$$

For fixed q, Q, κ, and $\alpha < 1/4$, the bounds therefore match at

$$\Theta\left(d_T^{-2} \log \frac{1}{\alpha} \right). \tag{26}$$

Infeasible alternatives can approach δ_c arbitrarily closely, while each logistic label contributes only quadratic KL information in their separation. A fixed informative comparison attains the matching upper order. Appendix B gives the proofs and tests.

6.2 Distance to the dynamic boundary is the primitive statistical quantity

Quadratic dependence on the requested horizon arises only in a special case. When $\delta_\star = 0$,

$$d_T = \frac{\Delta}{(v-\rho)T}, \tag{27}$$

so

$$N_T = \Theta\left(\frac{(v-\rho)^2 T^2}{\Delta^2} \log \frac{1}{\alpha}\right). \tag{28}$$

Here $d_T \propto 1/T$, producing the familiar T^2 scaling. In general, however, the primitive quantity is d_T, not T itself. The adaptation budget determines where feasibility ends, while preference information determines how quickly alternatives on the wrong side can be excluded. Figure 2 shows this dependence. Appendix B treats other preference regimes and boundary cases, while Appendix B.4 gives optimizer-specific boundaries and finite label arrival.

7 General Dynamic Alternatives and the Role of Feedback

The scalar analysis is complete because every feasible preference shares one admissible control. In a general reward class, we must separately identify dynamically infeasible preferences and determine whether the remaining preferences share a common control.

Let $\mathfrak{A}_T$ denote the admissible trajectories under the declared initialization, optimizer, reward family, adaptation budget, and progress requirement, and define

$$\mathcal{F}_T(r) = \{u \in \mathfrak{A}_T : \ u \text{ satisfies the task under } r\}. \tag{29}$$

The dynamically infeasible alternatives are

$$\mathrm{Bad}_T = \{r \in \mathcal{U}_0 : \mathcal{F}_T(r) = \varnothing\}. \tag{30}$$

Bad_T generalizes the scalar infeasible side of $\delta_c(T)$. Continued certification does not require full reward identification; it requires distinguishing the true preference from still-compatible alternatives in Bad_T. Under source indistinguishability, let P_r^e denote the label distribution of query $e \in H$. The best worst case information rate against Bad_T is

$$\mathcal{I}_T(r_0) = \sup_{\omega \in \Delta(H)} \inf_{r \in \mathrm{Bad}_T} \sum_e \omega_e \, \mathrm{KL}(P_{r_0}^e \| P_r^e). \tag{31}$$

The same change of measure argument gives the necessary bound

$$\mathbb{E}_{r_0} \tau \geq \frac{\mathrm{kl}(1-\alpha, \alpha)}{\mathcal{I}_T(r_0)}. \tag{32}$$

Thus control determines which alternatives matter, while the query pool and observation model determine how quickly they can be distinguished. Eliminating Bad_T solves only the infeasibility discrimination step. Certification additionally requires one control that works for every preference retained by the updated information set. A sufficient interface is a valid coverage event $r_\star \in \mathcal{U}_N$ together with

$$u_N \in \bigcap_{r \in \mathcal{U}_N} \mathcal{F}_T(r). \tag{33}$$

Individual feasibility is insufficient because retained preferences can require incompatible controls. Thus resolving infeasibility discrimination does not in general resolve control selection, whose matching joint query and control complexity remains open. Appendix B gives source-information and operational details, while Appendix D gives finite-feature interfaces without explicit output enumeration.

8 NUMERICAL VERIFICATION

The numerical experiments test three consequences of the theory in finite synthetic systems: whether certification can fail while pointwise capability remains, whether adaptation resources control the realized horizon, and whether feedback difficulty follows distance to dynamic infeasibility rather than the requested horizon itself.

8.1 CERTIFICATE EXPIRATION AND EXTENSION BY TARGETED FEEDBACK

The first experiment tests whether certification can fail while pointwise capability remains and whether additional information can restore certification for the same trajectory without changing its true gain. In the finite system used for the endogenous realization, Theorem 1 gives $T_0 = 5$ while the pointwise normalized optimum remains one, as shown in Figure 1. Appendix F.1 gives the parameter settings.

For $\delta_\star = 0$, a batch of 1,200 labels collected before the deadline narrows the valid interval enough to certify the same trajectory through $T = 15$. The corresponding one sided test accepts with probability 0.987 at $\delta_\star = 0$ and at most 0.0066 at the feasibility boundary, within $\alpha = .05$. Repeating the source comparison provides no information because its KL divergence is zero across the relevant alternatives. Thus additional information extends the certificate without changing the trajectory or its true gain. We next hold information fixed and vary adaptation resources.

8.2 FINITE REWARD UPDATE RESOURCES EXTEND THE HORIZON

The second experiment varies only reward adaptation resources. In the same finite system, we vary the number m of clipped quadratic calibration updates before each policy step; Appendix F.3 gives the full protocol. More updates should delay certificate loss, while the concrete rule may still fall short of its displacement envelope.

Because the probability map is affine, each finite step satisfies the exact identity

$$\frac{J_{r_\delta}(\theta_{k+1}) - J_{r_\delta}(\theta_k)}{h\epsilon^2} = 1 + \delta(\phi_k - c_k). \tag{34}$$

Table 1 reports the first uncertified step.

Table 1: Support horizon under discrete reward update resources. More updates extend the realized horizon, while the calibration rule can remain below its displacement envelope. A dash indicates no failure for $0 \le t < 15$.

		First uncertified step	
Reward steps m	Budget s	Realized	Ideal envelope
0	0.000	3.35	3.35
1	0.002	4.55	5.05
2	0.004	7.85	10.05
3	0.006	—	—

More reward updates extend the horizon, but the concrete rule does not always realize its full displacement capability. Appendix A.4 gives the continuous analogue with startup loss, and Appendix F.3 gives the full trajectories and verifies full support and the finite step identity. We next return to the information side.

8.3 FEEDBACK COST TRACKS DISTANCE TO THE DYNAMIC BOUNDARY

The final experiment tests whether feedback difficulty is governed by T or by d_T. If T were primitive, instances at the same horizon would have comparable difficulty. Figure 2 shows otherwise: panel (a) gives substantially different label requirements at the same T across true preference parameters, while panel (b) aligns the same instances with the predicted d_T^{-2} dependence.

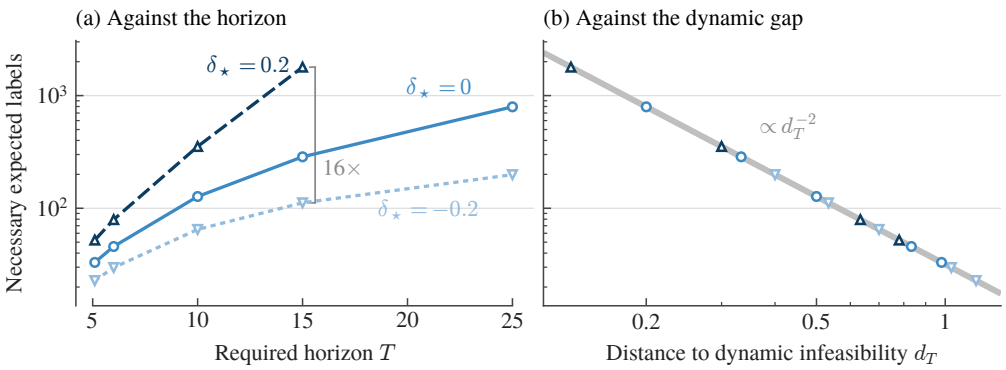

Figure 2: Feedback difficulty is governed by distance to dynamic infeasibility rather than the requested horizon alone. (a) At the same horizon T, different true preference parameters can require substantially different numbers of labels. (b) Replotting the same instances against d_T collapses their lower bounds onto the predicted d_T^{-2} scaling.

Exact binomial calculations verify the fixed batch test across these conditions. Appendix B gives the tests and lower bound calculations, Appendix F.1 records the evaluated grid and verification results, and Appendix F records independent LP and inequality checks for the numerical claims.

9 Discussion and Conclusion

The central implication of our results is that continued reward guidance is not a pointwise property. A suitable reward can remain available at every policy state while finite adaptation prevents the deployed reward from tracking the moving certified set. The support horizon separates this reachability failure from failures of preference information. When the true task remains feasible, additional feedback can remove compatible alternatives that make the target horizon infeasible; if the true preference itself is dynamically infeasible, more data cannot repair the adaptation limit. In general, certification also requires a common control for the retained preferences. Continued guidance therefore depends jointly on what the feedback identifies, how the policy uses the reward, and how quickly the reward mechanism can adapt.

This perspective extends partial identifiability from a static decision question to a dynamic one (Skalse et al., 2023; Lazzati & Metelli, 2026) and distinguishes reward prediction from guidance under finite adaptation (Ackermann et al., 2025; Tang et al., 2026). Reward prediction, a current robust gain certificate, and a trajectory certificate answer different questions. For preference acquisition, dynamic feasibility identifies the uncertainty that matters for sustaining guidance rather than requiring complete reward identification (Lindner et al., 2021; Feng et al., 2025; Shen et al., 2025; Zhu et al., 2023). In the scalar setting, this closes the statistical problem because all feasible preferences share one admissible control; more generally, Bad_T identifies the alternatives that must be distinguished, while common control selection remains separate.

Together, the results establish a temporal view of reward guidance: finite adaptation creates a support horizon, and the resulting feasibility structure determines what additional preference information is needed to extend certified progress. We characterize this mechanism exactly in the scalar setting, realize it within finite policy dynamics and locally under ordinary Euclidean policy gradient, and quantify its statistical consequence under Bradley–Terry feedback. The current evidence establishes the mechanism in finite systems. More broadly, the support horizon connects control reachability to statistical discrimination and makes the duration of certified reward guidance a concrete object of analysis.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript, to check the correctness and internal consistency of mathematical derivations and proofs, and to assist with experimental plotting.

REPRODUCIBILITY STATEMENT

The theoretical assumptions, derivations, and proofs are provided in the main text and appendices. Numerical parameter settings, evaluated conditions, statistical procedures, and verification checks are reported in Appendix F, with additional experimental settings in the corresponding referenced sections and tables.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

Johannes Ackermann, Takashi Ishida, and Masashi Sugiyama. Off-policy corrected reward modeling for reinforcement learning from human feedback. In *Second Conference on Language Modeling*, 2025.

Aaron D. Ames, Xiangru Xu, Jessy W. Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8): 3861–3876, 2017. doi: 10.1109/TAC.2016.2638961.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Alexandre Capone, Ryan Kazuo Cosner, Aaron Ames, and Sandra Hirche. Learning safe control via on-the-fly bandit exploration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6726–6745, 2025.

Yanxi Chen. Ranking from pairwise comparisons in general graphs and graphs with locality. *arXiv preprint arXiv:2304.06821*, 2023.

Yunzhen Feng, Ariel Kwiatkowski, Kunhao Zheng, Julia Kempe, and Yaqi Duan. PILAF: Optimal human preference sampling for reward modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16744–16776, 2025.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, 2016.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Filippo Lazzati and Alberto Maria Metelli. Robustness in the face of partial identifiability in reward learning. In *International Conference on Learning Representations*, 2026.

David Lindner, Matteo Turchetta, Sebastian Tschiatschek, Kamil Ciosek, and Andreas Krause. Information directed reward learning for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3850–3862, 2021.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

Yunyi Shen, Hao Sun, and Jean-François Ton. Active reward modeling: Adaptive preference labeling for large language model alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54410–54430, 2025.

Joar M. V. Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32033–32058, 2023.

Kenton Tang, Yuzhu Chen, and Fengxiang He. Generalisation of RLHF under reward shift and clipped KL regularisation. *arXiv preprint arXiv:2602.21765*, 2026.

Andrew J. Taylor, Andrew Singletary, Yisong Yue, and Aaron D. Ames. Learning for safety-critical control with control barrier functions. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 708–717. PMLR, 2020.

Tom Verhoeff. Settling multiple debts efficiently: An invitation to computing science. *Informatics in Education*, 3(1):105–126, 2004. doi: 10.15388/infedu.2004.08.

Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or K-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43037–43067, 2023.

A Dynamic Results and Proof Details

This appendix supplies the proof details and auxiliary dynamic results used in the main text. We first collect the assumptions that define the comparison class. We then prove the movement budget and corridor criterion, realize the scalar mechanism inside the finite policy system, analyze the explicit calibration transient, and give the local proof for ordinary Euclidean policy gradient.

A.1 Scope and modeling conventions

Table 2 collects assumptions that remain active throughout the main analysis. The purpose is to delimit the comparison class rather than to claim coverage of arbitrary RLHF training dynamics.

Table 2: Scope of the main analysis. Each row lists an assumption, the results that rely on it, and what is outside the present theory.

Assumption	Used for	Not covered
Fixed true preference and declared structural class	All results	Human preferences that vary arbitrarily over time
Fixed features and declared optimizer	Action map and adaptation budget	Full PPO, GRPO, or Adam dynamics
Source indistinguishability or explicit source KL	Label lower bounds	Membership in a confidence set alone
Known scalar corridor or explicit endogenous model	Exact support horizon	General multidimensional viability kernels
Bounded adaptation and positive progress requirement	Dynamic separation	Claims beyond a task with constrained resources
BTL temperature, query sensitivity, and parameter range	Upper bounds for binary feedback	Interchange of exact values, Gaussian ratings, and binary labels

The main text fixes the environment while measuring policy improvement. If the policy also depends on an evolving environment state z, then the total derivative contains the additional term $r^\top \partial_z p\, \dot{z}$. We do not count that environment term as policy improvement.

A.2 Movement budgets and the full trajectory corridor condition

For a BTL observation with feature difference d_i, the loss is

$$\ell_i(\phi) = \log(1 + \exp(\phi^\top d_i)) - Y_i \phi^\top d_i. \tag{35}$$

Its gradient is

$$\nabla \ell_i = (\sigma(\phi^\top d_i) - Y_i) d_i, \tag{36}$$

so $\|\nabla \ell_i\| \le L$; the same bound holds for a batch average. Since $\Pi_\Phi(\phi) = \phi$ and projection is nonexpansive,

$$\|\phi_{\mathrm{new}} - \phi\| \le \eta L. \tag{37}$$

Summing the stepwise displacements gives Equation equation 9. If the objective also contains $\lambda\|\phi\|^2/2$ and $\|\phi\| \le C$, replace L by $L + \lambda C$.

For the continuous corridor, necessity follows from any feasible path.

$$\ell(s) \le \phi(s) \le \phi(t) + \rho|s - t| \le u(t) + \rho|s - t|. \tag{38}$$

For sufficiency, define

$$f(t) = \sup_s [\ell(s) - \rho|t - s|]. \tag{39}$$

Bounded intervals make this upper envelope finite. Every term is ρ-Lipschitz, hence so is f. Setting $s = t$ gives $f(t) \ge \ell(t)$, while Equation equation 13 implies $f(t) \le u(t)$. Replacing the interval at time zero by a singleton fixes the initialization. A Lipschitz function is absolutely continuous with derivative bounded almost everywhere.

In discrete time, let

$$d_s(i,j) = \sum_{k=\min(i,j)}^{\max(i,j)-1} s_k. \tag{40}$$

Necessity follows by accumulating the movement bounds at each step. Sufficiency uses

$$\phi_j = \max_i [\ell_i - d_s(i,j)]. \tag{41}$$

The triangle inequality for the weighted path metric gives both the neighboring displacement bound and the interval constraints. Zero budgets are allowed. This statement concerns a known candidate corridor. It is not a generic planning algorithm for a closed loop system with endogenous intervals.

A special case with a cumulative resource bound is also useful. If $c(0) = 0$ and $c'(t) \geq \rho(t) \geq 0$, every admissible path satisfies

$$\phi(t) \leq \int_0^t \rho(s)\, ds. \tag{42}$$

The minimum lag

$$a(t) = c(t) - \int_0^t \rho(s)\, ds \tag{43}$$

is nondecreasing. Uniform progress is feasible if and only if $\eta_0 a(T) \leq \Delta$, attained by using the maximum allowed speed at every time. If the drift reverses or the reward can catch up, the endpoint alone is insufficient and the full corridor condition is required.

A.3 Finite policy realization of Theorem 1

The explicit five-output construction used in the main text is

$$p_\theta = \tfrac{1}{5}\mathbf{1} + \epsilon V\theta, \qquad r_\delta = V_1 + \delta V_2, \qquad R_\phi = V_3 + \phi V_4, \tag{44}$$

with $D = \epsilon V$ and $V^\top V = I$. With guidance coefficient vector

$$c_\phi = (0, 0, 1, \phi)^\top, \tag{45}$$

we have $\dot\theta = \epsilon M c_\phi$. Block multiplication yields

$$\dot\theta_1 = \epsilon, \quad \dot\theta_2 = \epsilon(\phi - c), \quad \dot\theta_3 = \epsilon d, \quad \dot\theta_4 = \epsilon d\phi. \tag{46}$$

Therefore $\dot c = k\epsilon d$. Since

$$J_{r_\delta} = \epsilon(\theta_1 + \delta\theta_2), \tag{47}$$

Equation equation 14 follows. The minimum eigenvalue of the block matrix M is

$$d - \|K(c)\|_{\mathrm{op}}, \tag{48}$$

which is strictly positive on the declared domain.

Using a Helmert basis in the zero sum subspace,

$$V_1 = (1, -1, 0, 0, 0)^\top/\sqrt{2}, \qquad V_2 = (1, 1, -2, 0, 0)^\top/\sqrt{6}. \tag{49}$$

The reward difference for the source comparison $(1, 2)$ is $\sqrt{2}$, whereas the informative comparison $(4, 3)$ has difference $\sqrt{2/3}\,\delta$. Repeating the source comparison therefore cannot identify δ, while the second comparison can.

The finite realization requires positive output probabilities, a positive definite preconditioner, and a nonbinding amplitude budget over the stated horizon. The experiments check these conditions along the actual trajectory. Extending the horizon or changing parameters requires checking them again.

A.4 Explicit clipped calibration and startup loss

For a known target $c(t) = vt$, consider the explicit calibration flow

$$\dot{\phi} = \mathrm{clip}(k(c - \phi), -\rho, \rho). \tag{50}$$

This is a structural loss for guidance calibration, not a BTL prediction loss. With lag $e = vt - \phi$ and

$$t_s = k^{-1} \log \frac{v}{v - \rho}, \tag{51}$$

we obtain

$$e(t) = \frac{v}{k}(1 - e^{-kt}), \quad t \le t_s, \qquad e(t) = \frac{\rho}{k} + (v - \rho)(t - t_s), \quad t \ge t_s. \tag{52}$$

If the allowed lag is $a_0 = \Delta/\eta_0 > \rho/k$, the boundary time is

$$T_{\mathrm{clip}} = t_s + \frac{a_0 - \rho/k}{v - \rho}. \tag{53}$$

If $a_0 \le \rho/k$, the boundary occurs during the unsaturated phase at $-k^{-1}\log(1 - ka_0/v)$. Thus two procedures can share the same maximum speed while a concrete optimizer attains a shorter support horizon because it does not immediately realize that speed. Appendix F.2 verifies this transient numerically for the parameters used in the calibration experiment.

A.5 Local proof for all admissible controls in Theorem 2

For the concrete instance used in the proof, set

$$\gamma = .06, \qquad \rho = .04, \qquad x_0 = .8, \qquad y_0 = 0, \tag{54}$$

with

$$\phi_0 = \ell(x_0, y_0) + .001 = .8428043202. \tag{55}$$

The initial worst-case progress is $.0612592 > \gamma$.

Write

$$A = (1 - x^2)^2, \qquad B = (1 - y^2)^2. \tag{56}$$

The true progress for r_a is

$$aA\phi + (1 - a)B(1 - \phi). \tag{57}$$

Minimizing over $a \in [-1, 2]$ and solving the two endpoint inequalities gives the interval in Equation equation 19.

Consider the box

$$x \in [.795, .81], \qquad y \in [-.005, .02], \qquad \phi \in [.835, .86]. \tag{58}$$

Inside it,

$$\dot{A} = -4x(1 - x^2)A\phi, \qquad \dot{B} = -4y(1 - y^2)B(1 - \phi), \tag{59}$$

and

$$\dot{\ell} = \frac{-2(B + \gamma)\dot{A} + (2A - \gamma)\dot{B}}{(2A + B)^2}. \tag{60}$$

Let

$$A_{\min} = (1 - .81^2)^2, \quad A_{\max} = (1 - .795^2)^2, \quad B_{\min} = (1 - .02^2)^2. \tag{61}$$

Because $2A - \gamma > 0$ in the box, bound the positive and negative contributions by

$$L_{\mathrm{pos}} = 8(.795)(1 - .81^2)A_{\min}(.835)(B_{\min} + .06), \tag{62}$$

$$L_{\mathrm{neg}} = (2A_{\max} - .06)4(.02)(1 - .835). \tag{63}$$

Then

$$\dot{\ell} \ge \frac{L_{\mathrm{pos}} - L_{\mathrm{neg}}}{(2A_{\max} + 1)^2} > .1399398. \tag{64}$$

Hence $h = \phi - \ell$ obeys $\dot{h} \le -.0999398$ while $h(0) = .001$, which yields an upper bound of 0.0100061 on the contact time. After contact, the vector field points strictly out of the safe set as long as the trajectory stays in the box.

It remains to rule out earlier box exit. The bounds

$$\dot{x} \le A_{\max}(.86), \qquad 0 \le \dot{y} \le .165, \qquad |\dot{\phi}| \le .04 \tag{65}$$

show that the earliest possible exit through any reachable face is no earlier than $.0858746$. The state variables x and y are monotone and cannot exit through their lower faces. Finally,

$$\frac{A_{\min} B_{\min}}{A_{\min} + B_{\min}} = .105750394\ldots, \tag{66}$$

while the probability lower bound is

$$\frac{(1 - .81)(1 - .02)}{4} = .04655. \tag{67}$$

The strict inequalities make the local mechanism stable to sufficiently small perturbations of the initialization and parameters.

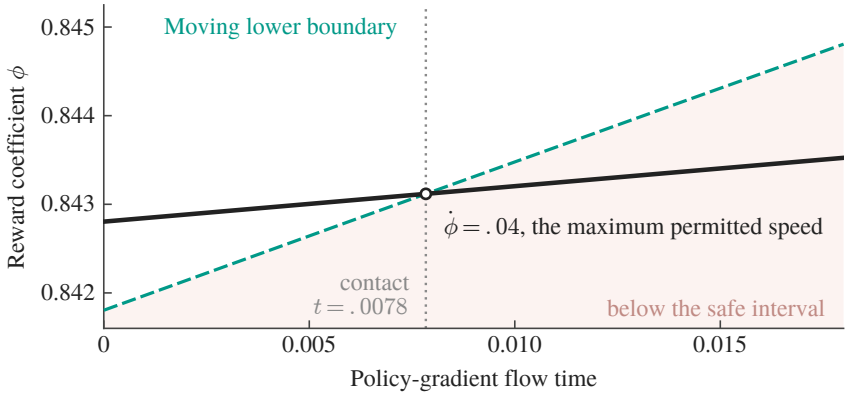

Figure 3: Instance from Theorem 2 under ordinary policy gradient. The solid curve is the reward coefficient under the maximum permitted positive speed, and the dashed curve is the moving lower boundary of the safe interval. Their contact marks loss of the certificate for that path, while the statement covering all admissible controls follows from the analytic box bound.

A.6 A LOCAL CONDITION WITH MULTIPLE WORST CASE REWARDS

For a general state equation $\dot{z} = f(z, \phi)$, define

$$h = L_{\mathcal{U}}(B(z)\phi) - \gamma(t), \qquad \mathcal{U}_\star = \arg\min_{r \in \mathcal{U}} r^\top B\phi. \tag{68}$$

Directional differentiation of the minimum of linear functions on a compact set gives

$$Dh[1, f, u] = \min_{r \in \mathcal{U}_\star} \left\{ r^\top DB(z)[f]\phi + r^\top Bu - \dot{\gamma} \right\}. \tag{69}$$

The admissible velocities are

$$T_\Phi(\phi) \cap \{u : \|u\| \le \rho\}. \tag{70}$$

If the maximum achievable directional derivative is strictly negative on the boundary, forward maintenance of the certificate is impossible from that point. Conversely, if a locally Lipschitz feedback makes $Dh \ge -\kappa_h h$ in a neighborhood, comparison gives

$$h(t) \ge e^{-\kappa_h t} h(0). \tag{71}$$

Passing from the tangency condition to a complete invariance theorem requires the usual assumptions on solution existence and tangent cone regularity from control theory (Ames et al., 2017).

One cannot select an arbitrary active worst-case reward. For

$$h(t, \phi) = \min(-t + \phi, -t - \phi), \tag{72}$$

at the origin with $|u| \leq 2$, the best directional derivative is

$$\max_u(-1 - |u|) = -1. \tag{73}$$

Keeping only the first active constraint would incorrectly give $+1$.

B Statistical Proofs for Preference Certification

This appendix proves the statistical results used once the requested horizon exceeds the certificate supported by the source information. The dynamic boundary determines the alternatives that must be distinguished. We first prove the lower bound by comparing a feasible truth with alternatives just beyond that boundary. We then give a fixed batch test, a batched procedure for an unknown gap, the corresponding boundary for a declared algorithm, and the correction required when the source data already distinguish the alternatives.

B.1 Lower bound for every $T > T_0$

Because $T > T_0$, we have $\delta_c < \eta_0$. Choose any $\varepsilon > 0$ small enough that

$$\delta_1 = \delta_c + \varepsilon \leq \eta_0. \tag{74}$$

No admissible control completes the task under δ_1, while $\delta_\star$ is feasible. For the logistic exponential family, the log partition function is

$$A(z) = \log(1 + e^z), \qquad A''(z) \leq \frac{1}{4}. \tag{75}$$

Hence one query satisfies

$$\mathrm{KL}(P_{\delta_\star,i} \| P_{\delta_1,i}) \leq \frac{q_i^2(\delta_1 - \delta_\star)^2}{8}. \tag{76}$$

The queries may adapt to the history but must be selected before observing their current labels. Conditional KL decomposition over a stopped history and data processing through the final accept event give

$$\frac{(\delta_1 - \delta_\star)^2}{8} \mathbb{E}_{\delta_\star} \sum_{i=1}^{\tau} q_i^2 \geq \mathrm{kl}(1 - \alpha, \alpha). \tag{77}$$

The expectation is under the true instance and does not depend on the proof parameter ε. Letting $\varepsilon \downarrow 0$ yields Equation equation 23. Random stopping requires the usual finite expectation or integrability condition; infinite expected sample size does not violate the lower bound. The change of measure step follows standard best arm identification arguments (Kaufmann et al., 2016; Garivier & Kaufmann, 2016).

If $T \leq T_0$, no new feedback is required because $\phi = \rho t$ certifies the full prior class. The dependence on the requested horizon varies with the true preference. If $\delta_\star < 0$, the gap d_T need not vanish over long feasible horizons. If $\delta_\star > 0$, d_T shrinks as the requested horizon approaches the true feasibility boundary, so certification becomes arbitrarily difficult before feasibility is lost. At $\delta_\star = \delta_c$, infeasible alternatives approach arbitrarily closely, so finite expected sample size is incompatible with both the stated power and false acceptance guarantees.

B.2 One sided upper bound using a fixed batch

Let

$$p_\delta = \sigma(q\delta), \qquad s_N = \sqrt{\frac{\log(1/\alpha)}{2N}}. \tag{78}$$

The accept event is $\hat{p}_N + s_N \leq p_c$. For any infeasible parameter, $p_\delta > p_c$, so false acceptance implies

$$\hat{p}_N - p_\delta < -s_N. \tag{79}$$

A one sided Hoeffding bound controls this probability by α.

Under the true feasible parameter, with probability at least $1 - \alpha$,

$$\hat{p}_N \le p_\star + s_N. \tag{80}$$

If $2s_N \le p_c - p_\star$, this event guarantees acceptance, equivalently

$$N \ge \frac{2\log(1/\alpha)}{(p_c - p_\star)^2}. \tag{81}$$

On the bounded parameter interval, $\sigma' \ge \kappa$, so

$$p_c - p_\star \ge \kappa q d_T, \tag{82}$$

which gives Equation equation 25.

B.3 A BATCHED PROCEDURE FOR AN UNKNOWN GAP

Set

$$N_j = 2^j, \qquad \alpha_j = \frac{6\alpha}{\pi^2(j+1)^2}, \qquad s_j = \sqrt{\frac{\log(1/\alpha_j)}{2N_j}}. \tag{83}$$

At checkpoint j, accept if $\hat{p}_{N_j} + s_j \le p_c$. A union bound keeps the total false acceptance probability below α. Under a strictly feasible instance, on the event that all upper deviations are controlled, the procedure accepts by the first checkpoint satisfying

$$N_j \ge \frac{2\log(1/\alpha_j)}{(p_c - p_\star)^2}. \tag{84}$$

This is a stopping guarantee that holds with high probability. We do not claim constant optimal expected sample size for the procedure when the gap is unknown.

The dependence of the lower bound on the dynamic gap d_T is illustrated in Figure 2 in the main text for three representative values of the true preference parameter.

B.4 BOUNDARIES FOR A DECLARED ALGORITHM AND FEEDBACK ARRIVAL TIME

If the target of certification is a declared algorithm $\mathcal{A}$, fixed independently of the new labels, define its maximum nonnegative lag

$$E_{\mathcal{A}}(T) = \max_{t \le T}\{c(t) - \phi_{\mathcal{A}}(t)\}. \tag{85}$$

Then $\delta_c^{\mathcal{A}} = \Delta / E_{\mathcal{A}}(T)$ replaces δ_c. The lower bound now excludes parameters under which that algorithm cannot continue; those parameters need not coincide with the infeasible set of the entire control class. The transient of the explicit clipped gradient rule in Appendix A.4 can be substituted directly.

Labels also take time to obtain. If at most Λ new labels arrive per unit policy time, the requested batch must be collected within the task schedule. A conservative plan under uninterrupted training completes certification with the fixed batch for the target horizon before the old certificate expires. We do not prove an optimal general online query schedule.

B.5 GENERAL DYNAMIC ALTERNATIVES AND SOURCE KL

For any $r \in \mathrm{Bad}_T$, source indistinguishability and the accept event give

$$\sum_e \mathbb{E}_{r_0} N_e \,\mathrm{KL}(P_{r_0}^e \| P_r^e) \ge \mathrm{kl}(1-\alpha, \alpha). \tag{86}$$

For finite positive $\mathbb{E}\tau$, define

$$\omega_e = \frac{\mathbb{E}N_e}{\mathbb{E}\tau}. \tag{87}$$

Taking the infimum over alternatives and comparing with the optimal allocation yields Equation equation 32. If $\mathcal{I}_T = 0$, no procedure with finite expected sample size satisfies the task. If $\mathrm{Bad}_T = \varnothing$, this lower bound is zero, but information may still be necessary to select a common control.

If the source distributions differ, the correct information inequality is

$$\mathrm{KL}(P^{\mathrm{src}}_{r_0} \| P^{\mathrm{src}}_{r}) + \sum_e \mathbb{E}_{r_0} N_e \, \mathrm{KL}(P^e_{r_0} \| P^e_r) \geq \mathrm{kl}(1-\alpha, \alpha). \tag{88}$$

Membership in the same realized confidence set does not justify setting the first term to zero. Conditioning on one confidence set also does not preserve the original unconditional error rates for free.

If a valid coverage event gives $r_\star \in \mathcal{U}_N$ and the procedure outputs a common control as in Equation equation 33, then violation of the progress requirement is bounded by the failure probability of the coverage event. This is a sufficient interface; it does not imply that the multidimensional intersection is computationally easy.

Operationally, one first asks whether the target horizon is dynamically feasible under the declared policy dynamics and adaptation class. Preference data are useful for excluding unresolved alternatives that obstruct certification, while execution additionally requires a common feasible trajectory for the retained preferences. If no such trajectory exists, either the target task or the adaptation capability must change.

C Comparison Structure and Adaptive Preference Confidence Sets

This appendix separates two supporting questions from the dynamic results in the main text. The first asks what exact comparison information identifies. The second asks how to maintain a valid confidence set under adaptively collected Bradley–Terry data. Both describe what the feedback reveals, but neither replaces the trajectory feasibility problem.

C.1 Exact comparison graphs answer identifiability, not dynamic necessity

Assume first that each observed source comparison edge reveals an exact reward difference, or equivalently the population BTL probability. Let the source graph have c connected components C_k. Differences are identified inside each component, while each component retains an unknown offset. For an induced direction b with zero sum, define

$$z_k(b) = \sum_{i \in C_k} b_i. \tag{89}$$

Then $r^\top b$ is identifiable if and only if every $z_k(b) = 0$, because any two compatible rewards differ by a componentwise constant.

Let $W \subseteq \mathbf{1}_c^\perp$ be a declared space of future component demands and let H restrict the available new comparisons. The minimum number of new edge types needed to identify every vector in W exactly is

$$m_H(W) = c - \max_{\mathcal{P}} |\mathcal{P}|, \tag{90}$$

where every block $P \in \mathcal{P}$ satisfies

$$\mathbf{1}_P^\top w = 0 \qquad \text{for every } w \in W, \tag{91}$$

and $H[P]$ is connected. The final comparison graph merges the original components into balanced connected blocks. A block containing s original components requires at least $s-1$ edges, and a spanning tree attains that count. If no valid partition exists, exact identification is impossible.

For example, a one dimensional demand $W = \mathrm{span}(1, 2, 4, -7)$ requires three new comparison types under a complete query pool, while $(1, -1, 2, -2)$ requires two. The partition problem in the zero sum subspace is the same as a minimum transfer problem studied by Verhoeff (2004). Related computational hardness follows classical partitioning results (Garey & Johnson, 1979). Exact identification of all demands is stronger than finding one feasible control with margin, so Equation equation 90 is not a lower bound on the number of labels required to continue training.

C.2 A BTL CONFIDENCE EVENT THAT HOLDS UNIFORMLY OVER TIME

In the fixed gauge space $\mathcal{H}$, let

$$r_\star \in \Theta_S = \{r : \|r\| \le S\}, \qquad d = n - 1. \tag{92}$$

A query direction $d_t = e_i - e_j$ is chosen from the history before observing Y_t, with

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{Bernoulli}(\sigma(d_t^\top r_\star)). \tag{93}$$

Given the parameter range and BTL temperature, define

$$\kappa = \min_{|z| \le \sqrt{2}S} \sigma'(z) > 0, \qquad V_t = \lambda I + \sum_{i \le t} d_i d_i^\top. \tag{94}$$

Use the constrained regularized logistic MLE

$$\hat{r}_t = \arg\min_{r \in \Theta_S} \left\{ \sum_{i \le t} [\log(1 + e^{d_i^\top r}) - Y_i d_i^\top r] + \frac{\kappa\lambda}{2}\|r\|^2 \right\}. \tag{95}$$

A self normalized martingale bound (Abbasi-Yadkori et al., 2011) gives the following statement with probability at least $1 - \alpha$, simultaneously for all t,

$$\|\hat{r}_t - r_\star\|_{V_t} \le \beta_t, \tag{96}$$

where

$$\beta_t = \frac{1}{2\kappa}\sqrt{2\log\frac{\det(V_t)^{1/2}}{\lambda^{d/2}\alpha}} + \sqrt{\lambda}S. \tag{97}$$

On the same event, every adaptive demand b_t satisfies

$$r_\star^\top b_t \ge \hat{r}_t^\top b_t - \beta_t\sqrt{b_t^\top V_t^{-1} b_t}. \tag{98}$$

The score noise is conditionally sub-Gaussian with parameter $1/2$. The self-normalized inequality controls its norm at every time, while the Hessian of the regularized loss is at least κV_t on the convex parameter set. The variational inequality for the constrained optimum and strong monotonicity then give Equation equation 98 through dual norms.

Intersecting the resulting ellipsoids with Θ_S over time produces nested information sets. An empty intersection signals a modeling or numerical problem; it is not an unconditional safety certificate. The curvature constant can be conservative, and we do not claim confidence widths that are optimal in dimension or constants.

C.3 TWO LIMITS OF PSEUDOINVERSES AND FISHER APPROXIMATIONS

The unregularized count Laplacian is

$$G_t = \sum d_i d_i^\top. \tag{99}$$

If the comparison graph is disconnected, the pseudoinverse quantity $b^\top G_t^\dagger b$ removes null space components. When the unknown component offsets have no finite range and $b \notin \text{Range}(G_t)$, the appropriate uncertainty is infinite rather than zero. A finite prior must retain that component explicitly or use a regularized confidence set.

The true Fisher matrix is

$$I_t(r) = \sum \sigma'(d_i^\top r) d_i d_i^\top, \tag{100}$$

with

$$\kappa G_t \preceq I_t(r) \preceq G_t/4. \tag{101}$$

Effective resistance and pairwise estimation are related in general comparison graphs (Chen, 2023); our use of this geometry is only to connect it to the current policy demand.

For fixed positive definite V, demand b, comparison direction d, and weight w, Sherman–Morrison gives

$$b^\top(V + wdd^\top)^{-1}b = b^\top V^{-1}b - \frac{w(d^\top V^{-1}b)^2}{1 + wd^\top V^{-1}d}. \tag{102}$$

Repeated measurements of one comparison can therefore leave a positive uncertainty floor for the declared task, at which point a new query direction is required. Equation equation 102 is a design matrix identity; a single random label need not improve the full certificate because the estimate center, confidence radius, and Fisher weight all change.

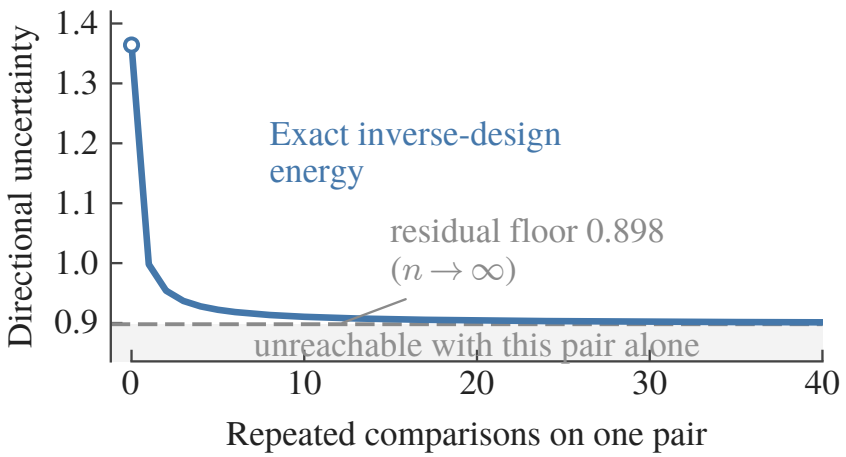

Figure 4: Uncertainty relevant to the task under repeated observations of one comparison direction in the rank one design check. Repetition reduces uncertainty only along the queried direction and can leave a positive residual floor for the declared demand.

D Finite Step Interfaces without Explicit Jacobians

The finite state derivations in the main text do not require explicit enumeration of every output when the relevant feature changes can be estimated directly. This appendix gives two interfaces. The first certifies a finite policy change using fixed reward features and an independent rollout estimate. The second expresses the first order gain through feature expectations and score vector products, avoiding an explicit policy Jacobian.

D.1 Finite policy certificate with fixed features

Suppose the true utility belongs to a fixed feature family

$$r_w(y) = w^\top \psi(y), \tag{103}$$

with w in a valid preference set. For a fixed candidate policy π' chosen independently of the validation batch, define

$$a = \mathbb{E}_{\pi'}\psi - \mathbb{E}_{\pi}\psi. \tag{104}$$

Then the finite policy gain is exactly

$$J_w(\pi') - J_w(\pi) = w^\top a. \tag{105}$$

If an independent unlabeled rollout estimates a by $\hat{a}$ with $\|a - \hat{a}\| \le \epsilon_B$ and $\|w\| \le S$, then

$$J_w(\pi') - J_w(\pi) \ge \inf_{w \in \mathcal{U}} w^\top \hat{a} - S\epsilon_B. \tag{106}$$

If every feature coordinate is bounded by H in absolute value and the feature dimension is d, draw N independent outcomes from each policy. Hoeffding concentration and a union bound give, with probability at least $1 - \alpha_B$,

$$\|a - \hat{a}\|_2 \le 2H\sqrt{\frac{2d\log(4d/\alpha_B)}{N}} = \epsilon_B. \tag{107}$$

Combining Equation equation 107 with $w \in \mathcal{U}$ and $\|w\| \le S$ yields Equation equation 106. Evaluating multiple candidates or repeatedly reusing the same validation batch requires independent validation or simultaneous error control.

D.2 First order interface without an explicit Jacobian

Let the guidance reward be $\phi^\top \chi(y)$ and let $s_\theta = \nabla_\theta \log \pi_\theta$. Define

$$C_\psi = \mathbb{E}[s_\theta \psi^\top], \qquad C_\chi = \mathbb{E}[s_\theta \chi^\top]. \tag{108}$$

Under $\dot{\theta} = MC_\chi \phi$,

$$\dot{J}_w = w^\top A_\theta \phi, \qquad A_\theta = C_\psi^\top M C_\chi. \tag{109}$$

The demand induced by a candidate direction can be estimated by JVPs of feature expectations, score vector products, or independent sampling without enumerating all outputs. Multiplying two estimated matrices from the same batch introduces correlated error and should not be called unbiased without further analysis.

If the true reward includes a residual e with $\|e\|_\infty \le \epsilon_r$, subtract

$$\epsilon_r \|p' - p\|_1 \tag{110}$$

from the certificate, or the coarser bound $2\epsilon_r$. This residual range is an additional assumption, not a property obtained automatically from a frozen reward head.

For a parameter update $\theta' = \theta + \eta v$, if J_r is L smooth along the line segment and $c \le r^\top B\phi$, the standard smoothness bound gives

$$J_r(\theta') - J_r(\theta) \ge \eta c - \frac{L}{2}\eta^2 \|v\|^2. \tag{111}$$

A suitable step size therefore certifies true gain, but continuous policy gradient, a concrete discrete optimizer, and optimizers with history state must be modeled separately.

E Prediction, Guidance, and Weak Identification

This appendix makes the distinction between prediction and guidance explicit under the same source information. The first example shows that source data can identify a common improvement direction without supporting uniformly accurate prediction of the underlying rewards. The second keeps full support while making one direction weakly identified, which shows that prediction risk and uncertainty in an identified gain can scale differently.

E.1 Prediction and guidance separate under the same source data

Let the source ratings contain only $(q, s) = (1, 1)$ and $(-1, -1)$ with values 1 and -1. The declared true reward class is

$$r_a = aq + (1 - a)s. \tag{112}$$

The source fit with minimum norm is

$$\hat{R}_0 = (q + s)/2. \tag{113}$$

At a target checkpoint with mean $(.8, 0)$, the four outcome probabilities are $(.45, .45, .05, .05)$. The source data cannot distinguish

$$r_A = 2q - s, \qquad r_B = -q + 2s. \tag{114}$$

Any predictor using only the source data has worst case MSE over these two targets of at least

$$\frac{1}{4}\mathbb{E}[(r_A - r_B)^2] = 4.5, \tag{115}$$

and the midpoint predictor attains the bound. Under ordinary policy gradient, the static source fit gives true progress rates $(-.3704, .9352)$ for (r_A, r_B).

Choose instead the dynamic guidance coefficients

$$\phi_q = 625/706, \qquad \phi_s = 81/706, \tag{116}$$

which increase the two mean coordinates at the same rate. Table 3 reports the resulting prediction errors and true progress.

Dynamic guidance is a worse predictor for r_B, so it has no uniform prediction advantage. What the source data identify is a common improvement direction, not the two independent factor preferences. A query on disagreeing outputs, whose difference is $4a - 2$, supplies the missing distinction. Exact real valued differences and binary preference bits carry different amounts of information and should not be conflated.

Table 3: Prediction and guidance in the example with four outputs. Static denotes the source fit with minimum norm $\hat{R}_0$, and dynamic denotes the guidance coefficients $(\phi_q, \phi_s) = (625/706, 81/706)$. Mean squared error is evaluated at the target checkpoint, while progress is the instantaneous rate under ordinary policy gradient.

	OOD MSE		True progress	
True reward	Static	Dynamic	Static	Dynamic
r_A	4.5000	2.4853	-0.3704	0.1147
r_B	4.5000	7.1085	0.9352	0.1147

E.2 The separation persists with full support and weak identification

Now let a fixed balanced source design place fraction $1-\varepsilon$ of its mass on agreeing features $q = s$ and fraction $\varepsilon > 0$ on disagreeing features, with Gaussian rating noise variance σ^2. Its Gram matrix is

$$G = N \begin{pmatrix} 1 & 1-2\varepsilon \\ 1-2\varepsilon & 1 \end{pmatrix}. \tag{117}$$

The common and difference directions have eigenvalues $2N(1-\varepsilon)$ and $2N\varepsilon$. If the target feature second moment is I, then the OLS prediction risk and the variance of a common gain estimate for $c = (\gamma_0, \gamma_0)$ are

$$\frac{\sigma^2}{2N}\left(\frac{1}{1-\varepsilon} + \frac{1}{\varepsilon}\right), \qquad \frac{\sigma^2\gamma_0^2}{N(1-\varepsilon)}. \tag{118}$$

The two quantities therefore depend differently on a weakly identified direction even though the design has full support. This is an exact calculation for the fixed design, not a general nonparametric minimax theorem.

With $N = 2000$, $\sigma = 1$, and $\gamma_0 = .1$, we evaluate eight values of ε and 50,000 independent Gaussian error samples at each value. The maximum relative errors between simulation and the two analytic quantities are approximately 1.33% and 1.61%.

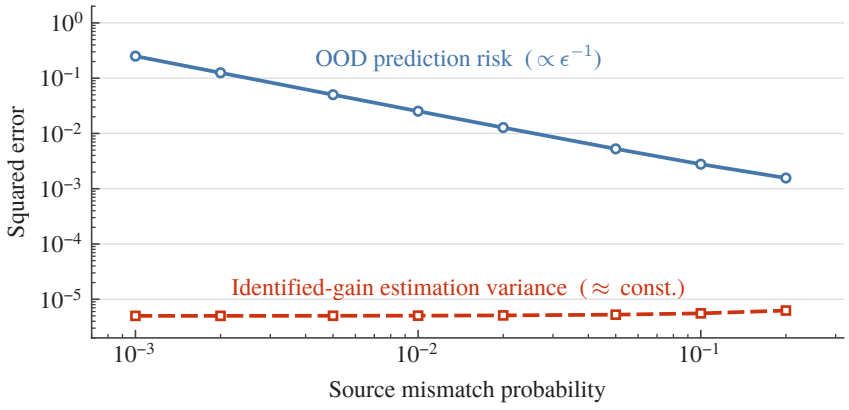

Figure 5: Weak identification with full support in the fixed Gaussian design. The upper curve is OOD prediction risk and the lower curve is the variance of the identified common gain estimate. The two curves are different squared error quantities and should not be interpreted as two methods on a common performance scale.

This example keeps the reward modeling learning layer visible. Complete prediction can be difficult because a weak direction is poorly estimated, while an action of the reward already supported by the source information can remain easy to estimate. The main theorems study the additional restriction created by finite adaptation once such information is used during policy change.

F Numerical Settings and Verification

This appendix records the numerical parameter settings and verification checks used for the reported numerical claims.

F.1 Parameters for the main numerical experiments

For the certificate expiration and targeted feedback experiment, we use $\epsilon = .04$, $v = .12$, $g = 1$, $\gamma = .6$, $\eta_0 = 1$, $\rho = .04$, and $|\phi| \le 2$, with the displayed trajectory $\phi = .04t$. Theorem 1 then gives $T_0 = 5$. For the feedback extension, $\delta_\star = 0$ and the representative batch contains 1,200 labels with 600 positive outcomes.

For the instance complexity experiment, we evaluate $T \in \{5.1, 6, 10, 15, 25\}$ and $\delta_\star \in \{0, -.2, .2\}$ over the strictly feasible conditions. Exact binomial calculations at the sufficient fixed batch sizes give minimum power 0.991 and maximum boundary acceptance probability 0.0077, both within the target $\alpha = .05$.

F.2 Clipped calibration startup loss

Set $v = .12$, $\rho = .04$, $k = 1$, and allowed lag $a_0 = .4$. The path that uses the ideal speed envelope reaches its deadline at 5. Equation equation 53 gives $T_{\text{clip}} = 4.905465108$ for the clipped calibration flow, and an independent ODE solve returns the same value.

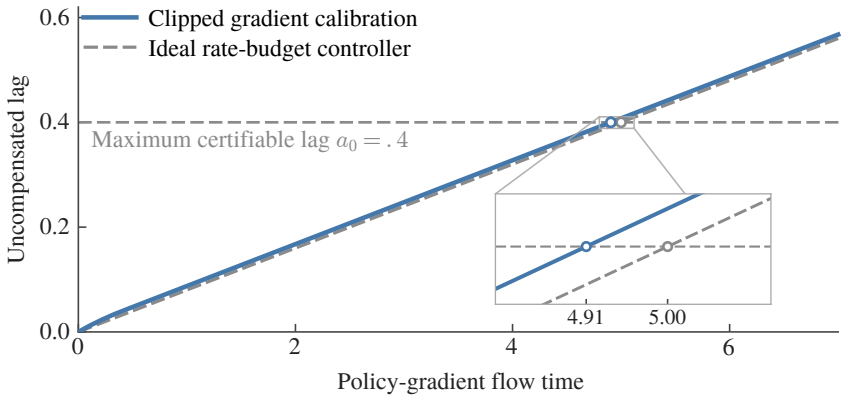

Figure 6: Lag behind the moving target for the path that uses the ideal speed envelope and the explicit clipped gradient calibration flow. Both share the same maximum speed $\rho = .04$, but the clipped flow spends an initial unsaturated phase before reaching that speed. The horizontal line marks the maximum certifiable lag $a_0 = .4$. The inset resolves the two crossing times, 4.91 for the clipped flow and 5.00 for the ideal path.

Figure 6 shows the source of the difference. The explicit algorithm does not instantly attain the maximum allowed speed, so its lag accumulates faster during startup. This experiment verifies the declared guidance calibration algorithm; it does not assert that a BTL likelihood update automatically implements the same control.

F.3 Complete trajectories for the discrete resource experiment

The discrete resource experiment uses policy Euler step $h = .05$. Before each policy step, we perform m reward updates at the current c_k,

$$\phi \leftarrow \Pi_{[-2,2]}\left[\phi - \eta\,\mathrm{clip}(\phi - c_k, -G, G)\right], \qquad \eta = .01, \quad G = .2. \tag{119}$$

The reward is then frozen for one policy step. The round displacement envelope is $s = m\eta G$, and we evaluate $m \in \{0, 1, 2, 3\}$ for 300 policy steps. This is a quadratic guidance calibration rule rather than training a Bradley–Terry reward model.

Figure 7 plots the complete progress trajectories underlying Table 1. The diagnostic simulations continue to the endpoint after certification is lost only to expose the mechanism.

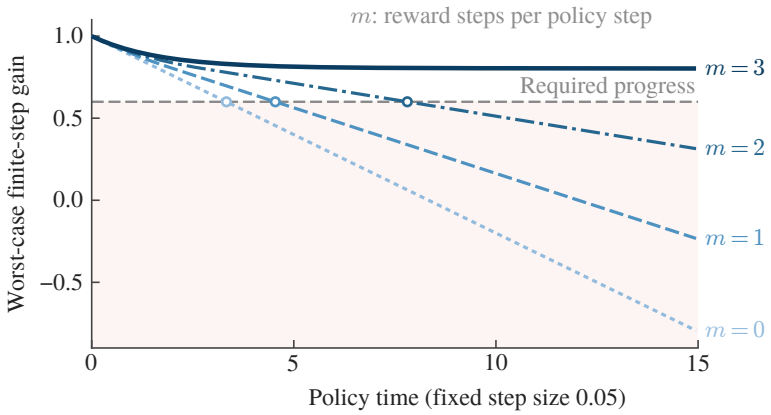

Figure 7: Worst case progress in the finite step reward resource experiment. The policy step, reward learning rate, and clipping threshold are fixed, while only the number m of reward updates before each policy step changes. Each curve is generated by the discrete iteration in Equation equation 119. The dashed horizontal line is the required normalized progress, the shaded band is the uncertified region, and open markers mark the first uncertified time for each m.

Across all discrete reward resource settings, the output probabilities remain at least 0.11686156, the preconditioner eigenvalues remain at least 1.75964264, and the maximum numerical error in Equation equation 34 is 7.06×10^{-13}. These diagnostics support the finite step construction but do not add a separate empirical claim beyond the horizon comparison in Table 1.

F.4 Auxiliary numerical verification

Table 4 collects additional numerical and consistency checks. These checks validate numerical constructions and elementary inequalities; they do not replace the analytic proofs or add a separate empirical claim.

Table 4: Auxiliary numerical verification. Each row lists the construction checked, the evaluation scope, and the property the check establishes.

Object	Evaluation	What it checks
Nonuniform discrete corridors	180 instances, 120 feasible and 60 infeasible; all agree with an independent LP	Corridor criterion and constructive path
Piecewise continuous corridor checks	150 instances; all agree with an LP	Agreement with the corridor criterion
Update displacement and logistic KL	10,000 random checks, no violations	Elementary movement and KL bounds
Restricted comparison pools	36 pools agree with exhaustive edge subset search	Comparison graph calculation
Anytime BTL	100 trajectories, 40 labels each, 8 checkpoints; simultaneous coverage 1.00	Finite simulation of a conservative bound
Feature and JVP interface	Identity error 1.73×10^{-18}; finite difference error 5.32×10^{-11}	Interface consistency without an explicit Jacobian