REINFORCEMENT LEARNING, ON-POLICY DISTILLATION, OR BOTH? A FINITE-BUDGET THEORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) learns from rewards, while on-policy distillation (OPD) trains a student to imitate a teacher at states the student visits. RL updates can be noisy; OPD can inherit teacher errors. We ask when combining them improves on either method alone under a finite budget. We compare expected final regularized reward regret in a known nonlinear two-stage model, charging for rollouts, rewards, and teacher queries. Under explicit local conditions and bounded learning rates, a fixed mixture can outperform both pure methods even after independent rate tuning and additional updates funded by saved query costs. For the specified estimators, this guarantee covers all permitted RL schedules chosen before training and adaptive OPD rates and stopping. An explicit sufficient window relates the advantage to teacher error, reward noise, learner-dependent visitation, and budget. Retaining sampling variation is essential: average update directions can predict the wrong winner. A separate two-update result quantifies the benefit of adapting RL rates to observed progress and gives sufficient conditions favoring mixing or adaptive RL. Controlled simulations check these criteria. A controlled Qwen study provides complementary evidence: after independent tuning under equal query allowances, mixing has lower mean regret with noisy rewards, while RL has lower mean regret with clean rewards. Together, the results identify a benefit that cannot be explained by rate tuning or cost reallocation alone. Teacher guidance can help when rewards are noisy and teacher mismatch is moderate; reliable rewards, cheaper pure updates, or effective adaptation can remove that benefit.

1 INTRODUCTION

On-policy distillation (OPD) trains a student to imitate a teacher on the student's own trajectories. For a language model, the teacher supplies next-token targets at prefixes the student actually produces, including prefixes containing its mistakes. Unlike supervision from a fixed collection of teacher responses, OPD follows the learner's changing behavior. It also provides detailed guidance when rewards arrive only at the end of a response or are difficult to estimate (Agarwal et al., 2024; Heo et al., 2026; Rang et al., 2026; Zhang et al., 2026a).

Imitation can also pass the teacher's errors to the student. Reinforcement learning (RL) offers a complementary signal: it evaluates actions by their consequences and can improve beyond the teacher, although its updates may be noisy. This motivates combining OPD with RL algorithms such as PPO or GRPO (Schulman et al., 2017; Shao et al., 2024; Zhang et al., 2026b; Wang et al., 2026). Existing approaches use both signals throughout training, reduce teacher guidance over time, or retain it only when it agrees with reward feedback (Schmitt et al., 2018; Tao et al., 2021; Agarwal et al., 2022; Wu et al., 2026). We study the question behind these choices: *under a finite training budget, when does combining RL and OPD outperform either method alone?*

The missing comparison. Co-training methods are available, but when the additional signal justifies its bias and cost remains poorly understood. The intuition that a teacher can help noisy RL does not settle this question. Improvement over RL alone does not rule out OPD as the better choice, and a comparison at a shared learning rate does not account for independent tuning. Teacher queries

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.

also consume resources that pure RL could spend on additional updates. A meaningful comparison must account for both tuning and these savings.

Our contribution. We develop a finite-budget comparative theory that identifies conditions under which a fixed mixture retains an advantage when both pure methods tune independently and use saved query costs for additional updates. All methods start from the same student and are judged by expected final regularized reward regret: the gap to the best policy in the studied class under a reward objective that penalizes deviation from a reference policy. Thus OPD is evaluated by reward, not imitation accuracy. The mixture averages the two update directions; each method pays for its rollouts and the signals it uses.

Our analytical setting. Our main results use a two-stage decision problem: the first action either ends the episode or leads to a second state with two actions. For a language model, this is a response with at most two token-selection steps. The first token ends generation or establishes the prefix for the second, determining which teacher supervision the student receives. This sequential on-policy dependence is absent from a single-decision model. We retain nonlinear policy probabilities, stochastic reward updates, and teacher mismatch while comparing entire classes of learning-rate schedules. Longer responses introduce more coupled states; our guarantees concern the two-stage case, not a full LLM.

The technical challenge. Separate upper bounds on each method's error cannot establish their ordering. We need lower bounds covering every permitted pure rule and an upper bound for one affordable mixture. Obtaining both requires combining stochastic moment analysis with nonlinear error control while jointly tracking teacher bias, reward noise, changing state visits, and different affordable update counts. Retaining sampling variation is essential: an appendix counterexample shows that mean update directions can predict the wrong winner.

What we find. We identify conditions under which mixing outperforms the tuned pure methods and examine how costs and adaptation change this comparison:

1. **A teacher-error, noise, and budget window favors mixing.** Theorem 4.1 establishes the nonlinear comparison; Proposition 4.2 identifies a sufficient teacher-error, noise, and budget window. Teacher mismatch must leave room to improve on OPD without overwhelming the mixture, while reward noise must limit RL despite extra affordable updates. The guarantee covers all permitted RL schedules fixed before training and adaptive OPD rates with stopping. Appendix I sharpens the conditions and quantifies an equal-cost advantage over a continuous nonlinear parameter region.
2. **The extra signal must justify its cost.** A separate scalar comparison in *What Query Costs and Correlation Change* isolates how sharing rollouts and correlation between the signals affect performance under a common budget. Its exact boundary shows when additional pure samples eliminate the mixed advantage.
3. **Adaptive RL can remove an advantage over fixed schedules.** Proposition 4.3 quantifies the extra improvement available when RL chooses its next rate from observed progress. In a separate two-update setting, this gives sufficient conditions for mixing to outperform all permitted adaptive pure baselines, or for an explicit adaptive RL rule to outperform the mixture and OPD. The larger-budget guarantee does not cover adaptive RL.

Controlled simulations check the window and feedback criteria. A complementary Qwen study implements the task through coupled adapter parameters: after independent tuning under equal query allowances, mixing has the lowest observed mean regret with noisy rewards and RL with clean rewards. This supports the qualitative mechanism beyond independent logits, without establishing the theorem's full comparison class for a transformer.

Scope and limitations. The guarantees concern known local models and specified update rules; their extension to learned critics, longer responses, and unseen prompts remains open. Query allowances also differ from measured training time. The results suggest evaluating teacher guidance when rewards are noisy and teacher error and query costs are moderate, then reassessing its value as these conditions change. They do not prescribe an annealing schedule. The relevant baselines are independently tuned RL and OPD under the same total budget, including teacher queries.

2 Related Work

Table 1 summarizes the closest analytical perspectives. Our comparison asks whether mixing improves final regularized reward over both tuned pure methods after accounting for additional updates funded by their saved query costs.

Table 1: Selected related analyses; models, objectives, and access assumptions differ. Our local equal-cost result uses specified estimators and bounded rates, covering prescribed RL schedules and adaptive OPD. Adaptive RL is treated separately at two updates.

Work	Setting	Central result	Comparison focus
Czarnecki et al. (2019)	Policy distillation with reward learning	Non-gradient dynamics and possible cycling	Distillation objective; rollout policy
Zhang et al. (2019)	Noisy quadratic optimization	Finite-training effects of optimizer choices	Batch size; update budget
Tiapkin et al. (2024)	RL with fixed expert demonstrations	Demonstration-dependent sample-complexity bounds	Expert data; RL samples
Sriraman et al. (2026)	Imitation with noisy expert feedback	Online/offline sample-complexity separation	Expert queries; no reward access
Fu et al. (2026); Oh et al. (2026)	Sampled-token OPD	Bias–variance analysis; control-variate baselines	Estimator design and variance
Ours	**Local nonlinear two-stage RL and OPD**	**Conditional mixed advantage over tuned pure methods**	**Final reward regret; equal query allowances**

Teacher-guided learning. Czarnecki et al. (2019) explain why on-student-policy distillation can be nonconservative; Tiapkin et al. (2024) analyze KL-regularized RL with a policy learned from fixed demonstrations. Other analyses address finite-time optimization and stability-based generalization (Ghadimi & Lan, 2013; Bousquet & Elisseeff, 2002; Hardt et al., 2016). Recent work studies noisy-expert imitation without rewards (Sriraman et al., 2026) and proximal teacher updates (Xie et al., 2026). Estimator studies analyze the bias–variance tradeoff of token- versus sequence-level OPD (Fu et al., 2026) and control variates for sampled-token OPD (Oh et al., 2026). We follow changing state visits and update noise through training to compare the resulting policies, rather than individual updates.

Finite-budget analysis. Noisy quadratic models explain finite-training effects (Zhang et al., 2019), and value-aware model-error analysis has precedents in model-based RL (Farahmand et al., 2017; Luo et al., 2018). We use these tools to compare one mixed learner with entire classes of pure learners. The key is to bound what each pure method can achieve, rather than enumerate its possible training histories, and then control the reference model's error in the actual nonlinear process. For adaptive rates, we quantify the additional improvement available from reacting to the current state. This separates an advantage over fixed schedules from one that survives feedback. Policy-gradient and regularized-control foundations include Williams (1992); Sutton et al. (1999); Geist et al. (2019); finite-time stochastic approximation provides a related analytical perspective (Ying, 2019; Chen et al., 2020; Shen & Chen, 2022). These connections do not imply generalization to unseen prompts here.

3 Problem Setup

Empirical Reward and On-Policy Distillation

We define a shared evaluation objective and two update signals on student-generated trajectories. RL seeks actions that improve reward; OPD matches the teacher's action distribution at visited states. Let $S = (x_1, \ldots, x_n) \sim P_{\mathrm{tr}}^n$ be the training prompts and $y = (a_1, \ldots, a_T)$ a response. At $s_t = (x, a_{<t})$, the policy π_θ chooses a_t. For a prompt distribution P, bounded sequence reward R, and reference

policy π_{ref}, define

$$J_P(\theta) = \mathbb{E}_{x\sim P, y\sim\pi_\theta}\left[R(x,y) - \beta\sum_{t=1}^{T}\mathrm{KL}(\pi_\theta(\cdot \mid s_t)\|\pi_{\mathrm{ref}}(\cdot \mid s_t))\right]. \tag{1}$$

The KL term penalizes departure from the reference policy, with strength β. The target is therefore regularized reward rather than raw task reward. J_S replaces P by the empirical distribution on S, and $g_S = \nabla J_S$. A teacher $q(\cdot \mid s)$ and rollout snapshot $\bar{\theta}$ define forward-KL OPD,

$$H_S(\theta; q, \bar{\theta}) = \mathbb{E}_{x\sim S, y\sim\pi_{\bar{\theta}}}\sum_{t=1}^{T} w_t \mathrm{KL}(q(\cdot \mid s_t)\|\pi_\theta(\cdot \mid s_t)). \tag{2}$$

The token weights $w_t \geq 0$ are fixed; all constructions use $w_t = 1$. The snapshot is held fixed while differentiating, giving the teacher direction $u_S = -\nabla_\theta H_S$. OPD thus improves imitation at sampled states without differentiating the dependence of state visitation on the policy. This is a *semi-gradient* when the rollout distribution changes with the student. The distinction matters because later teacher queries depend on earlier student actions.

The concrete comparisons use a normalized sampled direction $\widehat{d}_w = (1-w)\widehat{g} + w\widehat{u}$: $w = 0$ is RL, $w = 1$ is OPD, and $0 < w < 1$ is mixing. With effective step λ, the update is $\theta' = \theta + \lambda\widehat{d}_w$. The common normalization prevents a mixed method from gaining simply by taking a larger effective step. Appendix A.1 records the general bias/noise decomposition and information accounting.

One performance criterion. Fix a target prompt law P and a common comparison class with reward maximizer $\theta^\star_P$. The main comparison uses expected final-iterate regret

$$\mathfrak{R}^a_K = \mathbb{E}\mathcal{L}(\theta^a_K), \qquad \mathcal{L}(\theta) = J_P(\theta^\star_P) - J_P(\theta). \tag{3}$$

This criterion compares the returned policies under the same target distribution and regularized objective, regardless of their training losses. The known-model results do not establish generalization to unseen prompts.

Costs and comparison classes. Each update uses one episode. Charge c_0 for its rollout, c_R for the reward signal, and c_D for a teacher-target bundle. Mixing pays for both signals; a pure method pays for only one and can spend the savings on additional updates. These are fixed allowances, not measured wall-clock times: an unvisited state's reserved teacher cost is not refunded. All methods start from the same policy and use the estimators specified below.

Adaptive rates and stopping. We distinguish a learning-rate schedule fixed before training from a *feedback* rule, which chooses the next rate using observations so far. A decaying schedule chosen in advance is prescribed; reducing the rate in response to the current iterate is feedback. For the two-stage model, let $\mathcal{F}_k$ contain the entire history, current iterate, and internal randomization before episode $k+1$. A feedback RL rule chooses an $\mathcal{F}_k$-measurable step in $[0, 2]$; OPD may choose one in $[0, 4]$. Each may stop at a bounded stopping time and return its *current* iterate. For allowance budget $\mathcal{B}$, the respective update caps are

$$N_R = \left\lfloor \frac{\mathcal{B}}{c_0 + c_R} \right\rfloor, \quad N_D = \left\lfloor \frac{\mathcal{B}}{c_0 + c_D} \right\rfloor, \quad N_M = \left\lfloor \frac{\mathcal{B}}{c_0 + c_R + c_D} \right\rfloor. \tag{4}$$

Each cap is the number of complete updates the method can afford. A rate may use the known model and the observed history, but must be chosen before the next gradient is sampled. Stopping returns the current iterate; retrospective checkpoint selection is excluded. The comparison varies only scalar learning rates and stopping rules; it excludes momentum, preconditioning, alternative estimators, minibatches, and resets. Appendix D.1 gives the full classes.

4 Central Result: A Finite-Budget Three-Way Comparison

We first establish a nonlinear comparison and derive a sufficient condition in terms of teacher error, reward noise, and budget. A scalar model then isolates the effects of query costs and correlation, before we consider adaptive learning rates. The main statements retain free parameters; appendix certificates establish nonempty regions and reversals.

THE MODEL AND A GENERAL COMPARISON

We specialize the setup to the two-stage problem. The student either ends the episode or visits a second state, where it makes another binary choice. The first decision thus controls both reward and access to later supervision. Let $p_\star, s_\star \in (0,1)$ be its optimal probabilities, $h_p = p_\star(1-p_\star)$ and $h_s = s_\star(1-s_\star)$. With logit errors (x, y), write $p = \mathrm{sigmoid}(\mathrm{logit}(p_\star) + x)$ and similarly for s. We reserve h_p, h_s for these fixed optimal variances and $v_p = p(1-p), v_s = s(1-s)$ for current-policy variances. Choosing the leaf rewards as in Appendix B.1 gives the exact regularized reward regret

$$\mathcal{L}(x, y) = \mathrm{KL}(p\|p_\star) + p\mathrm{KL}(s\|s_\star). \tag{5}$$

Here $\mathrm{KL}(p\|p_\star)$ denotes divergence between Bernoulli distributions. The first term measures error at the initial state. The second measures error at the next state, weighted by the student's probability of reaching it. This visitation weight couples the two decisions.

Teacher error and sampling. Start at $\varepsilon(x_0, y_0)$, give the teacher probabilities $q_A = p_\star + \varepsilon h_p h_1$ and $q_B = s_\star + \varepsilon h_s h_2$, and let the fresh reward noise be $\varepsilon\zeta$, where $\mathbb{E}\zeta = 0$, $\mathbb{E}\zeta^2 = \nu$, and $|\zeta| \le T$, independently of actions and history. The positive scale ε places initialization, teacher error, and noise in a common neighborhood of the optimum. It is fixed during training, and all guarantees hold at explicitly bounded positive scales. To specify the estimator, draw $A \sim \mathrm{Ber}(p)$ and, only if $A = 1$, $B \sim \mathrm{Ber}(s)$. Put $a = A - p$, $b = B - s$, $D_s = \mathrm{KL}(s\|s_\star)$, and $\eta = \varepsilon\zeta$. The exact-baseline reward and teacher directions are

$$\begin{aligned} \widehat{g} &= \big(-(x + D_s)a^2 - Ayab + \eta a,\ A[-yb^2 + \eta b]\big), \\ \widehat{u} &= (q_A - p,\ A(q_B - s)). \end{aligned} \tag{6}$$

The factor A turns off the second update when that state is not visited; the unused leaf label can then be set to zero. Even without added reward noise, the RL direction varies with the sampled actions. The baselines subtract known expected returns to reduce this variation; bounded errors are treated in Appendix E.1. Here $0 < h_1 < x_0$: the teacher improves the first decision but remains short of its optimum. All parameters are known in this construction.

Why retain sampling variation? Replacing sampled update coefficients by their means can reverse the predicted winner, even without added reward noise. Proposition A.4 in Appendix A.4 gives a one-update counterexample and a finite nonlinear error bound. We therefore retain sampled coefficients and their second moments; proximity to the optimum alone does not justify dropping this variation.

A simpler model for the comparison. To compare entire schedule classes, we need bounds that avoid enumerating individual training histories. We linearize the updates and freeze action probabilities at the optimum, while retaining random action draws. The resulting *reference model* has closed squared-error recursions, while the actual learner follows the nonlinear updates above. Let $\ell_{1,k}, \ell_{2,k}$ lower-bound the RL second moments, starting at (x_0^2, y_0^2): at each update, minimize each coordinate's next moment over the step interval (65). Allowing separate rates for the coordinates can only help RL, so the resulting lower bound covers every permitted common-rate schedule.

Theorem 4.1 (Parameterized finite-budget nonlinear separation)**.** *Give each regime budget $\mathcal{B}$, rollout cost $c_0 \ge 0$, and channel costs $c_R, c_D > 0$. Put $N_R = \lfloor \mathcal{B}/(c_0 + c_R) \rfloor$, and fix one normalized mixture of $K_M \le \lfloor \mathcal{B}/(c_0 + c_R + c_D) \rfloor$ updates, step $\lambda \le a_{\max}$ and fraction $w \in (0,1)$. Let R_M be its exact reference risk from the four scalar moment recursions* (64)*, and set*

$$\Gamma = \min\{(h_p \ell_{1,N_R} + p_\star h_s \ell_{2,N_R})/2,\ h_p h_1^2/2\} - R_M. \tag{7}$$

Assume the explicit invariant-rectangle conditions in Appendix B.1, with radius cap ε_0. That appendix gives finite constants C_R, C_M, t_D using only the displayed parameters and rectangle. Write $E = \max(C_R, t_D) + C_M$. If $\Gamma > 0$, then, for every $0 < \varepsilon \le \min\{\varepsilon_0, \Gamma/(2E)\}$,

$$\inf_{a \in \mathcal{A}_R \cup \mathcal{A}_D} \mathbb{E}\mathcal{L}(\theta^a) - \mathbb{E}\mathcal{L}(\theta^M) \ge \varepsilon^2 \Gamma/2 > 0. \tag{8}$$

Here $\mathcal{A}_R$ permits every predetermined count through N_R and every prescribed step in $[0, a_{\max}]$; $\mathcal{A}_D$ permits affordable OPD steps in $[0, 4]$, including feedback and stopping. If $E = 0$, only $\varepsilon \le \varepsilon_0$ is required.

Interpretation. Each pure method may make its best permitted choice, while the mixture is fixed. The mixture has lower regret when its reference risk is below both pure-method lower bounds by more than the nonlinear error allowance. The rectangle conditions keep every compared run in the region where this error is controlled. The guarantee holds at every positive scale in the stated range, not only in a vanishing-scale limit.

Proof mechanism: four ingredients.

1. *Schedule-class bound:* the relaxed moment recursions give $R_R^{\mathrm{lb}} = (h_p \ell_{1,N_R} + p_\star h_s \ell_{2,N_R})/2$, an error level below which no permitted prescribed RL schedule can fall.
2. *Teacher floor:* under the allowed steps, OPD cannot pass the teacher's first-state probability, leaving $R_D^{\mathrm{floor}} = h_p h_1^2/2$ in reference units, even with feedback or stopping.
3. *Mixed risk:* four exact moment recursions compute R_M while retaining random actions and visits.
4. *Nonlinear correction:* an explicit error allowance translates these reference quantities back to actual regret.

Together, these give

$$\text{pure-minus-mixed regret} \;\geq \varepsilon^2 \big[\underbrace{\min\{R_R^{\mathrm{lb}}, R_D^{\mathrm{floor}}\} - R_M}_{\text{reference separation}} - \underbrace{\varepsilon E}_{\text{nonlinear correction}} \big]. \tag{9}$$

The reference advantage Γ must therefore exceed the error allowance εE; otherwise the test is inconclusive. Proposition 4.2 identifies a sufficient region of teacher and noise parameters for which this margin is positive.

Robustness and scope. A positive margin also leaves room for small modeling errors. Theorem E.1 quantifies tolerance to predictable value-baseline errors over a continuous family of optimal probabilities. That extension has its own finite neighborhood; the larger region in Appendix C.2 uses a separate numerical certificate. Appendix F extends the comparison to multi-action policies.

A Teacher-Error, Noise, and Budget Condition

We now express the theorem's comparison in terms of teacher error, reward noise, and affordable update counts. We place the mixed risk below a common level and both pure-method lower bounds above it, with a margin for nonlinear error.

Use the scaled model above, with teacher-error vector $h = (h_1, h_2)$. Set $q_1 = h_p x_0^2$, $q_2 = p_\star h_s y_0^2$, and suppose both are positive. Every prescribed RL schedule of at most N updates has sampled-reference risk at least

$$Q_N(\nu) = \frac{1}{2} \sum_{i=1}^{2} \frac{q_i \nu}{\nu + N q_i}. \tag{10}$$

The q_i measure initialization error in reference reward units. Under the stated containment conditions, actual regret is at least $\varepsilon^2 Q_N(\nu) - C_R \varepsilon^3$. More noise raises the reference floor; more pure updates lower it. For the fixed mixture, the exact sampled-reference risk has the form

$$R_M(h, \nu) = F_K(h) + D_K \nu, \qquad F_K(h) = A_K + d_K^\top h + h^\top C_K h, \quad C_K \succeq 0,\ D_K > 0. \tag{11}$$

Its coefficients depend on visitation, initialization, mixture weight, step size, and mixed update count K. Seven evaluations of the finite moment recursion recover them exactly, without statistical fitting. The vector h preserves the direction of teacher error: equally large errors can have different effects at the two decisions.

Proposition 4.2 (A structural equal-cost dominance window)**.** *Under the containment and teacher-floor conditions of Theorem 4.1, take $N = \lfloor \mathcal{B}/(c_0 + c_R) \rfloor \geq 1$ and $1 \leq K \leq \lfloor \mathcal{B}/(c_0 + c_R + c_D) \rfloor$. Certify that each pure lower-bound and mixed upper-bound error, divided by ε^2, is at most $\delta/2$. Choose $\tau, \delta > 0$ with $0 < \tau + \delta < (q_1 + q_2)/2$. If the admissible teacher and noise parameters satisfy*

$$h_1 \geq \sqrt{\frac{2(\tau + \delta)}{h_p}}, \qquad \Phi_N(\tau + \delta) \leq \nu \leq \frac{\tau - \delta - F_K(h)}{D_K}, \qquad 0 < \nu \leq T^2, \tag{12}$$

then the mixture's actual expected regret is at least $\varepsilon^2\delta$ below every affordable prescribed RL schedule in the allowed step range and every affordable OPD rule with steps in $[0, 4]$, including feedback and current-iterate stopping. Here $\Phi_N(t)$ is the explicit inverse of $Q_N(\nu) = t$ in (73)*.*

Reading the condition. Teacher error must be large enough to limit OPD but small enough to keep mixed risk below τ. Reward noise must similarly limit RL without overwhelming the mixture. The buffer δ covers nonlinear error; N, K account for different affordable update counts. At fixed τ and initialization, the lower noise threshold grows linearly with N: more affordable RL updates narrow the certified region. The interval must be nonempty; a failed test is inconclusive. Appendix B.3 proves the result and gives a nonempty finite-scale family.

Sharper quantitative guarantees. Appendix I reduces approximation slack and characterizes the sufficient noise interval (Propositions I.1–I.2). Its separate direct nonlinear certificate (Proposition I.3) covers a continuous parameter region at $\varepsilon = .04$: 20 mixed updates achieve at least 1.4% lower expected final regret than the best permitted pure method with 25 affordable updates under equal costs. The bounds retain prescribed RL schedules and adaptive OPD; they do not follow by substitution into the preceding window. These are theoretical certificates, not new experiments.

When the advantage disappears. Lower reward noise or more affordable pure updates can favor RL; greater teacher error can raise both OPD and mixed risk. Proposition C.2 gives an equal-cost low-noise reversal. These results do not prescribe an optimal switching schedule.

What Query Costs and Correlation Change

The nonlinear window accounts for costs through affordable update counts. A separate scalar model explains how sharing a rollout can offset the cost of a second signal, and why correlation between the signals matters. Assume equal curvature, nonzero initial error, unbiased signals with equal variance and correlation $\rho \in (-1, 1)$, rollout cost $c_{\mathrm{roll}} \geq 0$, and equal signal costs $c_{\mathrm{sig}} > 0$. With no batch-update overhead and positive budgets divisible by both pure and paired observation costs, the exact boundary is

$$\text{strict equal-cost mixed advantage} \quad \Longleftrightarrow \quad \rho < \frac{c_{\mathrm{roll}}}{c_{\mathrm{roll}} + 2c_{\mathrm{sig}}}. \tag{13}$$

Both sides optimize prescribed steps in $[0, 1]$ and positive batch partitions; mixing also optimizes its teacher fraction (Proposition C.1). With independent signals and no rollout overhead, the methods tie: additional pure samples exactly offset the mixed gain. Higher rollout costs make sharing more valuable. Negative correlation allows signal errors to cancel, while strong positive correlation reduces the value of the second signal. This boundary isolates the cost mechanism in an unbiased scalar model; it is not a threshold for the nonlinear problem with a biased teacher. Appendix C treats integer remainders and scalar bias.

When Adaptive Learning Rates Remove the Advantage

The preceding results fix RL rates before training. Adaptive RL can instead reduce its rate or stop in response to observed progress. We quantify this feedback benefit in a separate two-update problem with free model parameters.

On a common invariant rectangle, choose quadratic functions $Q_- \leq L \leq Q_+$, where $L = \mathcal{L}$ is the actual nonlinear regret. The bounds follow directly from the smallest and largest policy variances and visitation probabilities on the rectangle (Appendix D.2). For either quadratic Q, define conditional reward progress and second moment by

$$a_Q(\theta) = -\theta^\top H_Q \mathbb{E}[\widehat{g} \mid \theta], \qquad b_Q(\theta) = \mathbb{E}[\widehat{g}^\top H_Q \widehat{g} \mid \theta], \quad Q(\theta) = \theta^\top H_Q \theta/2.$$

Both quantities use the actual sampled update and current visitation probabilities. With RL step cap $\bar{a}_R \leq 2$, let

$$\Psi(a, b) = \max_{0 \leq t \leq \bar{a}_R} \{ta - t^2 b/2\}. \tag{14}$$

Thus Ψ is the best next quadratic improvement after tuning the rate, including the option to take no update.

For a first RL step α, let Z_α be its random successor. Separate the best prescribed second step from the additional benefit of choosing that step after seeing Z_α:

$$F_Q(\alpha) = \mathbb{E}Q(Z_\alpha) - \Psi\big(\mathbb{E}a_Q(Z_\alpha), \mathbb{E}b_Q(Z_\alpha)\big),$$
$$\mathcal{A}_Q(\alpha) = \mathbb{E}\Psi\big(a_Q(Z_\alpha), b_Q(Z_\alpha)\big) - \Psi\big(\mathbb{E}a_Q(Z_\alpha), \mathbb{E}b_Q(Z_\alpha)\big) \geq 0. \tag{15}$$

F_Q is the best prescribed two-step quadratic risk for this first step. The *feedback credit* $\mathcal{A}_Q$ is the extra improvement from choosing the second rate after observing the outcome. It is large when different outcomes favor different rates. Their difference gives the exact optimal two-step quadratic risk conditional on the first rate, leaving only a scalar first-rate optimization.

Proposition 4.3 (A structural threshold for feedback)**.** *Use exact baselines and a fixed fresh noise law in the two-stage model. Suppose $N_R = N_M = 2$ under the declared costs and budget, and $p_\star < q_A \leq p(\theta_0)$. Let $D = \mathrm{KL}(q_A \| p_\star)$, and let $[L_M, U_M]$ enclose the actual two-update regret of one fixed mixed controller. Then every predictable RL rate in $[0, \bar{a}_R]$, with stopping by two updates, and every affordable adaptive OPD rule with steps in $[0, 4]$, are worse than that mixture by at least*

$$\gamma_{\mathrm{fb}} = \min\left\{\inf_{0 \leq \alpha \leq \bar{a}_R} [F_{Q_-}(\alpha) - \mathcal{A}_{Q_-}(\alpha)], D\right\} - U_M, \tag{16}$$

whenever $\gamma_{\mathrm{fb}} > 0$. Conversely, if some admissible α satisfies

$$F_{Q_+}(\alpha) - \mathcal{A}_{Q_+}(\alpha) < \min\{L_M, D\}, \tag{17}$$

an explicit feedback RL rule beats this mixture and every such OPD rule.

Reading the threshold. The mixed upper bound must lie below OPD's teacher floor and the RL lower bound after subtracting the feedback credit for *every* first RL rate. The converse identifies conditions favoring adaptive RL; either test may be inconclusive. Teacher error and mixing weight affect mixed risk, while noise, visits, and geometry affect the feedback credit. That credit can depend on the full noise law, not just its variance.

The equal-cost condition $2(c_0 + c_R + c_D) \leq \mathcal{B} < 3(c_0 + c_R)$ does not fund a third RL update; OPD keeps all affordable updates. Appendix D.2 gives the proof, clipped rate, and unclipped weighted-variance formula; Appendix D.3 certifies a positive region with bounded baseline errors.

Combining adaptation and additional updates. Appendix J gives sufficient conditions covering adaptive pure rates, stopping, and extra updates funded by saved query costs. Mixing wins if its short-horizon margin over adaptive RL exceeds a bound on its extra progress and mixed regret stays below OPD's teacher floor. State-dependent bounds sharpen this condition; a positive separation over a nonempty unequal-update adaptive region remains unverified.

Reading the comparisons together. Table 2 summarizes the two distinct freedoms given to pure RL: additional updates in the larger-budget results and adaptive rates in the two-update result. The guarantees concern these specified classes, not every budget-feasible algorithm.

Main result	Conclusion	Pure-method freedoms
Theorem 4.1	A positive reference margin exceeding nonlinear error implies dominance.	Affordable prescribed RL schedules; adaptive OPD and stopping.
Proposition 4.2	A sufficient teacher-error and noise window, with costs entering through affordable counts.	Affordable prescribed RL schedules; adaptive OPD and stopping.
Proposition 4.3	Mixed advantage must exceed RL's feedback credit; an upper test can certify the reverse ordering.	Adaptive rates and stopping; two RL/mixed updates, all affordable OPD updates.

Table 2: One reward-regret criterion, three main guarantees. The main statements retain free parameters. Supporting numerical certificates, including the low-noise reversal and baseline-error tolerance, appear in the appendix.

CONTROLLED VALIDATION AND A QWEN MECHANISM STUDY

Research questions. We ask whether *the criteria identify mixed or adaptive-RL advantages* and whether *the noise-dependent ordering persists with a transformer student*.

Experimental setup. The controlled study tests Proposition 4.2 in 54 teacher-error, noise, and cost settings at scale 10^{-8}, and Proposition 4.3 in 27 two-update settings. The complementary study uses Qwen2.5-0.5B-Instruct (Team et al., 2024) with rank-eight LoRA adapters (Hu et al., 2021) on the same two-stage task. Its teacher is an engineered probability policy, not another LLM. Qwen's final regularized reward regret is averaged over 16 fixed prompts. Pure methods receive 25 or 40 updates versus mixing's 20 under equal query allowances, not equal measured compute. Each method selects among 13 schedules using two tuning seeds and is evaluated on 16 fresh paired seeds. Clean and noisy runs share the task, teacher, and mixture weight $1/2$.

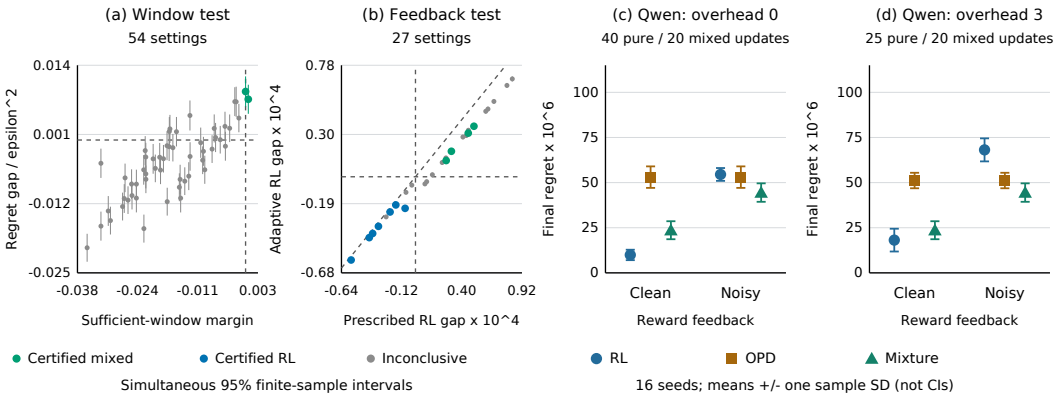


Figure 1: Structural checks and Qwen results. **(a)** Positive window margin certifies mixing; vertical values are the smaller pure-minus-mixed regret gap divided by ε^2. **(b)** Prescribed versus adaptive RL-minus-mixed gaps: feedback helps below the diagonal; lower-right points reverse the ranking. In (a–b), colors denote certificates and bars are simultaneous 95% finite-sample intervals. **(c–d)** Qwen means $\pm$ one sample standard deviation over 16 seeds, *not confidence intervals*. Independently tuned pure methods receive extra updates under equal query allowances. Mixing has the lowest noisy means; the same mixed runs are reused across (c–d), not independent repeats.

Results and interpretation. Controlled simulations agree with all 13 certificates: six mixed wins and seven adaptive-RL wins; 68 settings remain uncertified. These checks support the sufficient conditions, not sharp boundaries. With Qwen, RL has the lowest mean regret with clean rewards, and mixing with noisy rewards. At rollout overhead three, mixing's mean regret under noisy rewards is 34.7% below RL and 13.0% below OPD. Noise reverses the ordering despite independent tuning and extra pure updates, supporting the predicted mechanism but not class-wide dominance for transformers. Aggregate-based checks support all four noisy gaps at the unadjusted two-sided 5% level under paired-t reference assumptions; this is not a distribution-free guarantee. Appendix H gives protocols, complete outcomes, and uncertainty limitations.

5 DISCUSSION

Within the stated classes, noise limits RL and teacher error limits OPD, allowing mixing to beat both tuned pure methods despite extra pure updates. Reliable rewards, larger teacher error, cheaper pure sampling, or effective adaptation can instead favor a pure method.

The larger-budget guarantee covers prescribed RL rates; the two-update result covers feedback. Appendix J gives conditional guarantees combining these freedoms; a positive separation for the unified class remains open. Qwen supports a noise-dependent ordering with coupled parameters, not class-wide dominance or compute savings from an LLM teacher. Practical comparisons should tune RL and OPD independently under one total budget and report means and variability.

AI USE STATEMENT

Generative AI tools assisted with drafting and polishing portions of the manuscript and preparing verification code. The authors remain responsible for all claims and proofs.

REPRODUCIBILITY STATEMENT

Code is available at `https://anonymous.4open.science/r/OPD-RL-028C/`. The appendices provide the complete set of assumptions, proofs, and protocols. Details of the controlled study and the Qwen experiment are provided in the appendices.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Reincarnating reinforcement learning: Reusing prior computation to accelerate progress. *Advances in neural information processing systems*, 35:28955–28971, 2022.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234, 2020.

Wojciech M Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1331–1340. PMLR, 2019.

Amir-massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Artificial Intelligence and Statistics*, pp. 1486–1494. PMLR, 2017.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-policy delta distillation. *arXiv preprint arXiv:2607.15161*, 2026.

Liam Hodgkinson and Michael Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pp. 4262–4274. PMLR, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Fredrik Johansson. Arb: efficient arbitrary-precision midpoint-radius interval arithmetic. *IEEE Transactions on Computers*, 66(8):1281–1292, 2017.

Yuping Luo, Huazhe Xu, Yuanzhi Li, Yuandong Tian, Trevor Darrell, and Tengyu Ma. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. *arXiv preprint arXiv:1807.03858*, 2018.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. Kl for a kl: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026.

Miao Rang, Zhenni Bi, Hang Zhou, Kai Han, Xuechun Wang, An Xiao, Xinghao Chen, Yunhe Wang, and Hanting Chen. Near-policy: Accelerating on-policy distillation via asynchronous generation and selective packing. *arXiv preprint arXiv:2605.05940*, 2026.

Simon Schmitt, Jonathan J Hudson, Augustin Zidek, Simon Osindero, Carl Doersch, Wojciech M Czarnecki, Joel Z Leibo, Heinrich Kuttler, Andrew Zisserman, Karen Simonyan, et al. Kickstarting deep reinforcement learning. *arXiv preprint arXiv:1803.03835*, 2018.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Han Shen and Tianyi Chen. A single-timescale analysis for stochastic approximation with multiple coupled sequences. *Advances in Neural Information Processing Systems*, 35:17415–17429, 2022.

Ved Sriraman, Peihan Liu, Daniel Hsu, and Adam Block. Behavior cloning is not all you need: The optimality of on-policy distillation for noisy expert feedback. *arXiv preprint arXiv:2606.30923*, 2026.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Yunzhe Tao, Sahika Genc, Jonathan Chung, Tao Sun, and Sunil Mallya. Repaint: Knowledge transfer in deep reinforcement learning. In *International conference on machine learning*, pp. 10141–10152. PMLR, 2021.

Qwen Team et al. Qwen2. 5: A party of foundation models, 2024.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Aleksei Naumov, Pierre Perrault, Michal Valko, and Pierre Menard. Regularized rl. In *International Conference on Learning Representations*, volume 2024, pp. 28568–28632, 2024.

Chen Wang, Zhaochun Li, Jionghao Bai, Yining Zhang, Hexuan Deng, Ge Lan, and Yue Wang. Distilled reinforcement learning for llm post-training. *arXiv preprint arXiv:2607.17247*, 2026.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, et al. Seed: Self-evolving on-policy distillation for agentic reinforcement learning. *arXiv preprint arXiv:2607.14777*, 2026.

Zhengpeng Xie, Li Lyna Zhang, Zeke Xie, and Mao Yang. Trust region policy distillation. *arXiv preprint arXiv:2607.04751*, 2026.

Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on learning theory*, pp. 2803–2830. PMLR, 2019.

Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Dahl, Chris Shallue, and Roger B Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. *Advances in neural information processing systems*, 32, 2019.

Ke Zhang, Yunjie Tian, Dongdi Zhao, Yijiang Li, Yuanye Liu, Vishal M Patel, and Di Fu. On-policy distillation with best-of-n teacher rollout selection. *arXiv preprint arXiv:2605.09725*, 2026a.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. In *International Conference on Learning Representations*, volume 2026, pp. 120693–120726, 2026b.

APPENDIX CONTENTS

Appendix Roadmap and Result Dependencies **14**

A Models and Finite-Radius Comparison Tools **15**
A.1 General Updates and Information Accounting . . . 15
A.2 A Two-Stage Example with Learner-Dependent Teacher Queries . . . 15
A.3 One Reference-Value Framework for Both Noise Models . . . 17
A.4 A General Finite-Radius Transfer Statement . . . 19

B Main Nonlinear Separation and Parameter Window **23**
B.1 A Structural Comparison with Free Model and Budget Parameters . . . 23
B.2 An Interpretable Bias–Noise–Budget Window . . . 26
B.3 Proof of the Vector Teacher-Error Window . . . 28

C Query Costs and a Sharper Nonlinear Certificate **29**
C.1 When Saved Query Costs Buy More Samples . . . 29
C.2 Equal-Cost Dominance and Its Reversal . . . 33
C.3 The Teacher Floor and Affordable Comparison Classes . . . 33
C.4 Removing the Constant-Step Restriction on RL . . . 34
C.5 A Noise-Dependent Reversal at the Same Cost . . . 39

D Feedback Learning Rates and Stopping **40**
D.1 Feedback Policies, Design, and Statewise Certificates . . . 40
D.2 Proof of the Structural Feedback Threshold . . . 43
D.3 A Two-Update Feedback Separation . . . 45

E Robustness to Value-Baseline Errors **48**
E.1 Robustness to Predictable Value-Baseline Errors . . . 48
E.2 Predictable Baseline Errors Under Feedback . . . 51

F Multi-Action Comparison Against Pure Schedules **52**
F.1 Explicit Multi-Action Envelopes and a Schedule-Tuned Separation . . . 52
F.2 A Sharper Finite-Radius Certificate for All Pure Schedules . . . 56

G Supporting Enclosures for the Numerical Certificates **58**
G.1 Structural Comparison with Learner-Dependent Visitation . . . 58
G.2 Nonlinear Comparison Against Continuously Tuned Pure Methods . . . 61
G.3 A Computable Comparison Retaining Sampled Update Coefficients . . . 63
G.4 From Interval Enclosures to Tuned-Class Bounds . . . 66

H Validation Protocols and Complete Results **67**
H.1 Validation Aligned with the Structural Propositions . . . 67
H.2 Qwen Implementation of the Two-Stage Mechanism . . . 70
H.3 Prospective Validation Across Model and Cost Parameters . . . 72

I Sharper Structural Windows **75**
I.1 Separate Error Allowances and an Exact Teacher Floor . . . 75
I.2 A Noise Interval and a Certificate-Level Budget Ceiling . . . 76
I.3 Continuous Parameter Windows at a Larger Finite Radius . . . 77
I.4 What Changes in the Existing Parameter Grid? . . . 78

J Adaptive Pure Methods with Additional Affordable Updates **80**
J.1 The Unified Class and an Extra-Update Bound . . . 80
J.2 Bounding Extra Progress with Conditional Moments . . . 81
J.3 State-Dependent Continuation Bounds . . . 82
J.4 Three Adaptive Pure Updates Versus Two Mixed Updates . . . 84

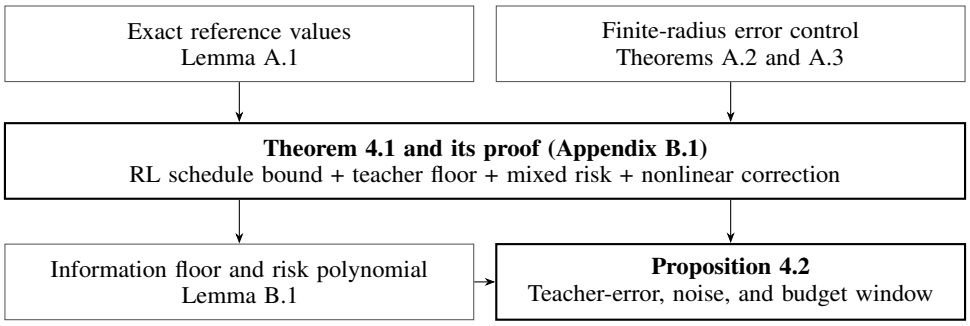
APPENDIX ROADMAP AND RESULT DEPENDENCIES

The appendix separates the main proofs from their general tools, robustness extensions, numerical certificates, and empirical checks. All supporting assumptions and proof dependencies are included. The shortest routes from the main text are:

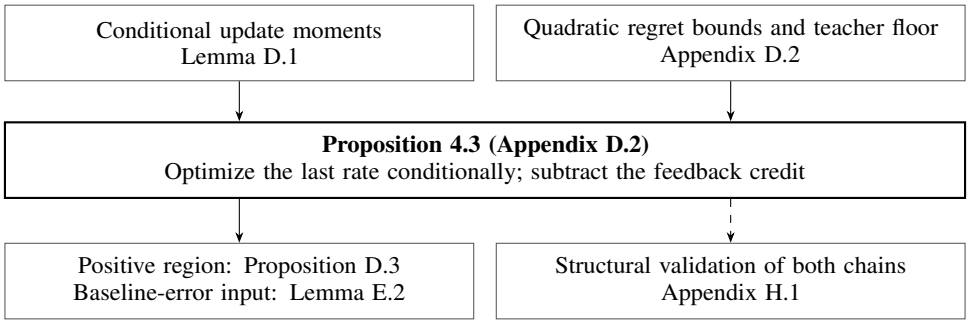
Question or result	Where to start
Theorem 4.1	Appendix B.1 (p. 23): four-part nonlinear comparison proof.
Proposition 4.2	Appendices B.2–B.3: information floor, then the vector teacher-error window.
Scalar cost boundary	Appendix C.1 (p. 29): optimized steps, batching, costs, and correlation.
Proposition 4.3	Appendix D.2 (p. 43): conditional optimization and feedback credit.
Figure 1	Appendices H.1–H.2: structural checks, then the Qwen protocol and uncertainty.

Proof dependencies. Solid arrows identify proof inputs, including constructions established inside a cited proof. Dashed arrows lead to validation, not proof steps. The two chains have different RL comparison classes.

(a) Prescribed RL schedules: nonlinear separation and its window



(b) Two-update adaptive RL: feedback credit and its checks



Other routes. Appendix E treats baseline errors; Appendix F treats multiple actions; Appendix G collects supporting certificate constructions. Appendix I separately sharpens the structural window, with retrospective comparisons and a direct-moment parameter certificate. The Qwen study tests the qualitative mechanism, not a proof dependency or a theorem for transformers. Broader stationarity and generalization guarantees are outside the scope of this paper.

Combining adaptation and extra updates. Appendix J extends the feedback framework: Theorem J.1 bounds extra-update progress, Propositions J.2–J.3 give computable and state-dependent refinements, and Corollary J.4 specializes to three pure updates versus two mixed updates. These are sufficient conditions, not a newly certified positive region.

A Models and Finite-Radius Comparison Tools

This section defines the common model and proves the reference-value and finite-radius tools. The main comparison proof begins in Appendix B; the sampling counterexample explains why its reference retains random update coefficients.

A.1 General Updates and Information Accounting

The main text specializes to one-episode score updates and fixed allowance costs. The following notation records the more general algorithm-agnostic update and accounting conventions.

We compare RL-only (R), OPD-only (D), and their mixture (M) through the ideal directions

$$v_k^{\mathsf{R}}(\theta) = g_S(\theta), \quad v_k^{\mathsf{D}}(\theta) = u_S(\theta), \quad v_k^{\mathsf{M}}(\theta) = g_S(\theta) + c_k u_S(\theta). \tag{18}$$

Here $c_k \geq 0$ is the OPD/RL ratio, so R is the $c_k = 0$ special case of M, whereas D is a distinct single-objective baseline. For method $a \in \{\mathsf{R}, \mathsf{D}, \mathsf{M}\}$, let $\widehat{d}_{k+1}^a$ be its implemented stochastic direction and define $b_k^a = \mathbb{E}[\widehat{d}_{k+1}^a \mid \mathcal{F}_k] - v_k^a(\theta_k^a)$ and $\xi_{k+1}^a = \widehat{d}_{k+1}^a - \mathbb{E}[\widehat{d}_{k+1}^a \mid \mathcal{F}_k]$. Then

$$\theta_{k+1}^a = \theta_k^a + \alpha_k^a\{v_k^a(\theta_k^a) + b_k^a + \xi_{k+1}^a\}, \tag{19}$$

with $\mathbb{E}[\xi_{k+1}^a \mid \mathcal{F}_k] = 0$. The three terms separate the intended update, systematic estimator error, and zero-mean sampling noise. The decomposition is exact; an implementation determines b_k^a and the conditional moments of ξ_{k+1}^a. Unless comparing methods explicitly, we analyze M and suppress its superscript, as well as S.

Matched information budget. For method a, let $\mathfrak{B}_a = (K, N_{\mathrm{roll}}, N_{\mathrm{reward}}, N_T, N_{\mathrm{audit}}, N_{\mathrm{val}})_a$ record updates, rollout tokens, reward labels, teacher queries, audit examples, and validation examples. We match this vector or report a common cost conversion. Audits are not free; any extra allowance must be deducted or given to all methods. Held-out-prompt evaluation requires prompts disjoint from training and alignment audits. The Qwen study instead tunes and evaluates on the fixed training-prompt panel, using disjoint tuning and evaluation seeds; it does not test unseen-prompt generalization (Appendix H.2).

Estimator scope. Appendix A.2 derives the sampled updates used in the nonlinear comparison. Applying the analysis to another implementation requires bounds on that implementation's own conditional bias and moments.

A.2 A Two-Stage Example with Learner-Dependent Teacher Queries

This section specifies the environment and estimator used by the main nonlinear comparison. The student changes both its action probabilities and the states on which OPD is applied. The subsequent two-stage certificates use this known-model kernel, with the noise law specified in each result; no fitted model or inexpensive unknown-model audit is assumed.

Environment and exact target regret. The learner first samples $A \sim \mathrm{Ber}(p)$, with $p = \mathrm{sigmoid}(z_\star + x)$. If $A = 0$ the episode terminates with reward zero. If $A = 1$, it reaches a second state and samples $B \sim \mathrm{Ber}(s)$, where $s = \mathrm{sigmoid}(w_\star + y)$, receiving reward $r_0 + w_\star B$. The two-dimensional parameter error is $\theta = (x, y)$; set $p_\star = \mathrm{sigmoid}(z_\star)$ and $s_\star = \mathrm{sigmoid}(w_\star)$. With uniform references at both states and KL coefficient one, choose

$$r_0 = z_\star - \log\{(1 + e^{w_\star})/2\}. \tag{20}$$

The regularized reward and its regret are

$$\begin{aligned} J(x, y) &= p\{r_0 + w_\star s - \mathrm{KL}(s\|1/2)\} - \mathrm{KL}(p\|1/2), \\ \mathcal{L}(x, y) &= J(0, 0) - J(x, y) = \mathrm{KL}(p\|p_\star) + p\mathrm{KL}(s\|s_\star). \end{aligned} \tag{21}$$

Here $\mathrm{KL}(a\|b)$ abbreviates Bernoulli KL. To verify the second equality, use $w_\star s - \mathrm{KL}(s\|1/2) = \log\{(1 + e^{w_\star})/2\} - \mathrm{KL}(s\|s_\star)$, and then apply the same identity at the root. Both KL terms are nonnegative and vanish at $(0, 0)$; since $p > 0$, this is the unique finite-logit optimum.

Writing $v_p = p(1-p)$, $v_s = s(1-s)$, and $D_s = \mathrm{KL}(s\|s_\star)$ gives

$$g(x,y) = \nabla J(x,y) = (-v_p(x+D_s), -pv_s y), \qquad H_R = \nabla^2 \mathcal{L}(0,0) = \mathrm{diag}(v_{p,\star}, p_\star v_{s,\star}). \tag{22}$$

Indeed $D_s' = v_s y$, and the derivative of the root KL is $v_p x$. At the optimum $D_s = D_s' = 0$, so the off-diagonal Hessian entries vanish. These calculations establish the actual objective, not only a chosen update field with an assumed relationship to reward.

On-policy teacher field and its nonconservative part. Fix $h = (h_1, h_2)$ and teacher probabilities $q_A = p_\star + v_{p,\star} h_1$ and $q_B = s_\star + v_{s,\star} h_2$, both strictly between zero and one. The root is always queried; the second teacher is queried only after $A = 1$. The sampled and mean snapshot directions are

$$\widehat{u} = (q_A - p, A(q_B - s)), \qquad u = (q_A - p, p(q_B - s)). \tag{23}$$

For the full on-policy teacher loss $H_{\mathrm{on}} = \mathrm{KL}(q_A\|p) + p\mathrm{KL}(q_B\|s)$, direct differentiation gives

$$\nabla H_{\mathrm{on}} = -u + (v_p \mathrm{KL}(q_B\|s), 0). \tag{24}$$

Moreover $\partial_x u_2 = v_p(q_B - s)$ whereas $\partial_y u_1 = 0$. Thus the implemented teacher field is generally nonconservative even in this tabular model. Changing the root policy changes how often the second teacher is queried; replacing A by a fixed occupancy would remove this mechanism.

An unbiased same-rollout reward estimator. Independently draw $E \in \{-1, 1\}$ with equal probabilities and add reward noise σE. Put $a = A - p$, and $b = B - s$ only on the visited second state. Use

$$\widehat{g}_1 = -(x + D_s)a^2 - Ayab + \sigma Ea, \qquad \widehat{g}_2 = A\{-yb^2 + \sigma Eb\}. \tag{25}$$

Terms multiplied by A are zero when the episode terminates at the root. This is a score-function estimator with exact statewise baselines, available in the known model. To see its origin, let the sampled regularized utility be $U = A\{r_0 + w_\star B - \log(\pi(B)/[1/2])\} - \log(\pi(A)/[1/2])$. The root-centered utility is $U - \mathbb{E}U = -(x + D_s)(A - p) - Ay(B - s)$. Multiplication by the root score gives the first estimator without noise. At the second state the centered continuation utility is $-y(B - s)$, giving the second estimator. The expected derivatives of the sampled log policy penalties vanish by the conditional score identities, so these are unbiased derivatives of the displayed regularized objective.

More directly, $\mathbb{E}a^2 = v_p$, $\mathbb{E}[b \mid A = 1] = 0$, $\mathbb{E}[b^2 \mid A = 1] = v_s$, and the independent sign has zero mean. These four identities give $\mathbb{E}\widehat{g} = g$ in (22). Reward and teacher updates share A, so their errors need not be independent away from the reference optimum.

Reference dynamics. Let (r_a, t_a) equal $(1, 0)$ for RL, $(0, 1)$ for OPD, and $(1, c)$ for the mixture. Method a uses the actual update $\theta' = \theta + \alpha(r_a \widehat{g} + t_a \widehat{u})$. Choose $H_D = H_R =: H$ in the reference model. Its transition, drift, and noise covariance are

$$\begin{aligned} A_a &= I - \alpha(r_a + t_a)H, & d_a &= \alpha t_a H h, \\ Q_a &= \alpha^2 \left[r_a \sigma^2 H + t_a^2 \,\mathrm{diag}\{0, v_{p,\star}(q_B - s_\star)^2\} \right]. \end{aligned} \tag{26}$$

The covariance is the actual update covariance at $(0, 0)$: there $\widehat{g} = \sigma E(A - p_\star, A(B - s_\star))$, whose covariance is $\sigma^2 H$, and the only random teacher component is $A(q_B - s_\star)$. The cross covariance vanishes at this reference point because E is independent and centered. It need not vanish along the actual run. The nonconservative teacher derivative is not silently identified with $-H$; its difference, including the term proportional to $h_2 x$, is charged to e_k^a in Theorem A.2.

All actual conditional moments are explicit finite sums. The reward-using methods have six local outcomes: $(A, E) = (0, \pm 1)$ with probability $(1-p)/2$, and $(A, B, E) = (1, b, \pm 1)$ with probability $ps^b(1 - s)^{1-b}/2$. Let their updates and probabilities be $F_{a,j}(\theta)$ and $w_j(\theta)$. Then

$$\begin{aligned} \mu_a(\theta) &= \sum_j w_j(\theta) F_{a,j}(\theta), \\ C_a(\theta) &= \sum_j w_j(\theta) \{F_{a,j}(\theta) - \mu_a(\theta)\} \{F_{a,j}(\theta) - \mu_a(\theta)\}^\top. \end{aligned} \tag{27}$$

For OPD there are only the two A outcomes, with probabilities $1 - p, p$. Use $e_a = \mu_a - A_a\theta - d_a$ and this covariance in (34). Hence both the model and every residual in the comparison are fully specified without a fitted Jacobian or a frozen trajectory distribution.

A.3 One Reference-Value Framework for Both Noise Models

The reference-value recursion and residual identity below separate the tractable model from actual nonlinear updates. Expectations of actual updates always use their own conditional sampling law.

For a fixed configuration a (RL, OPD, or a fixed mixture) and finite horizon, use the random-affine reference

$$X_{k+1} = M_{k,a}X_k + d_{k,a}. \tag{28}$$

The pair $(M_{k,a}, d_{k,a})$ is fresh and independent of the past, with finite second moments. Its matrix and vector may be correlated, and its law may depend on k and a. Write $\overline{M}_{k,a} = \mathbb{E}M_{k,a}$, $\overline{d}_{k,a} = \mathbb{E}d_{k,a}$, and

$$Q_{k,a}(x) = \mathrm{Cov}(M_{k,a}x + d_{k,a}). \tag{29}$$

This covariance is a quadratic matrix-valued function of x. The additive-noise case $X_{k+1} = A_kX_k + \bar{d}_k + \zeta_k$ is included by taking $M_k = A_k$ deterministic and $d_k = \bar{d}_k + \zeta_k$ with centered fresh noise. Its $Q_k(x) = \mathrm{Cov}(\zeta_k)$ is constant.

Lemma A.1 (Exact random-affine reference values)**.** *For terminal loss $x^\top H_R x/2$, $H_R \succeq 0$, the expected terminal loss starting at state x at time k is*

$$V_{k,a}(x) = \tfrac{1}{2}x^\top P_{k,a}x + p_{k,a}^\top x + \gamma_{k,a}. \tag{30}$$

Suppressing a, initialize $(P_K, p_K, \gamma_K) = (H_R, 0, 0)$ and recurse:

$$\begin{aligned} P_k &= \mathbb{E}[M_k^\top P_{k+1}M_k], \\ p_k &= \mathbb{E}[M_k^\top (P_{k+1}d_k + p_{k+1})], \\ \gamma_k &= \gamma_{k+1} + \tfrac{1}{2}\mathbb{E}[d_k^\top P_{k+1}d_k] + p_{k+1}^\top \mathbb{E}d_k. \end{aligned} \tag{31}$$

Then $P_k \succeq 0$, $V_{0,a}(\theta_0) = \mathcal{R}_K^a$, and

$$V_k(x) = V_{k+1}(\overline{M}_k x + \overline{d}_k) + \tfrac{1}{2}\,\mathrm{tr}\{P_{k+1}Q_k(x)\}. \tag{32}$$

In plain language, the same finite recursion handles additive noise and random coefficients from sampled state visits. Replacing M by its mean generally loses variation because $\mathbb{E}[M^\top PM] \neq (\mathbb{E}M)^\top P(\mathbb{E}M)$. The linear recursion also retains correlation between the sampled matrix and teacher intercept. Random-affine models are established tools (Hodgkinson & Mahoney, 2021); here they support a finite-budget comparison, not a new stochastic-recursion principle.

Proof. Assume (30) at time $k+1$. Expanding $V_{k+1}(M_kx + d_k)$ gives quadratic coefficient $M_k^\top P_{k+1}M_k$, linear coefficient $M_k^\top(P_{k+1}d_k + p_{k+1})$, and constant $\gamma_{k+1} + d_k^\top P_{k+1}d_k/2 + p_{k+1}^\top d_k$. Taking expectations proves every line of (31). Independence from the past makes the coefficients deterministic. Backward induction proves the value formula; positive semidefiniteness is preserved by expectation and congruence. Finite second moments and fresh independent draws give finite iterate second moments at every step of this finite recursion.

For deterministic x, the next reference state's mean is $\overline{M}_kx + \overline{d}_k$ and its covariance is $Q_k(x)$. The expectation of a quadratic at a random vector equals its value at the mean plus one half the Hessian–covariance trace, proving (32). This identity holds for every x, not only states reachable from the original initialization. In the additive case the recursions reduce to $P_k = A_k^\top P_{k+1}A_k$, $p_k = A_k^\top(P_{k+1}\bar{d}_k + p_{k+1})$, and $\gamma_k = \gamma_{k+1} + \bar{d}_k^\top P_{k+1}\bar{d}_k/2 + p_{k+1}^\top \bar{d}_k + \mathrm{tr}(P_{k+1}Q_k)/2$. No contraction or simultaneous diagonalization is needed. □

Changing sampling laws: the exact residual identity. An actual run θ_k^a may have nonlinear, discrete, history-dependent updates. Conditional on its history $\mathcal{F}_k^a$, define

$$m_k^a = \overline{M}_{k,a}\theta_k^a + \overline{d}_{k,a}, \quad e_k^a = \mathbb{E}_k\theta_{k+1}^a - m_k^a, \quad C_k^a = \mathrm{Cov}_k(\theta_{k+1}^a). \tag{33}$$

Here e_k^a measures the error in the predicted next mean, and $C_k^a - Q_{k,a}(\theta_k^a)$ measures the change in update variability. Both may depend on the policy and teacher weight; no shared sampling law is required.

Theorem A.2 (Finite-budget comparison under changing sampling laws)**.** *Suppose the actual run starts at θ_0 and has finite second moments. Set $P = P_{k+1,a}$ and define*

$$Z_k^a = \langle \nabla V_{k+1,a}(m_k^a), e_k^a \rangle + \tfrac{1}{2}(e_k^a)^\top P e_k^a + \tfrac{1}{2}\operatorname{tr}\{P(C_k^a - Q_{k,a}(\theta_k^a))\}. \tag{34}$$

Then the following identity is exact:

$$\tfrac{1}{2}\mathbb{E}[(\theta_K^a)^\top H_R \theta_K^a] - \mathcal{R}_K^a = \sum_{k<K} \mathbb{E} Z_k^a. \tag{35}$$

If $\mathfrak{R}_K^a = \mathbb{E}\mathcal{L}(\theta_K^a)$ and $\mathbb{E}|\mathcal{L}(\theta_K^a) - \frac{1}{2}(\theta_K^a)^\top H_R \theta_K^a| \le \nu_a$, then

$$|\mathfrak{R}_K^a - \mathcal{R}_K^a| \le \rho_a := \nu_a + \sum_{k<K} \mathbb{E}|Z_k^a|. \tag{36}$$

Consequently, $\mathcal{R}_K^b - \mathcal{R}_K^a > \rho_a + \rho_b$ certifies $\mathfrak{R}_K^a < \mathfrak{R}_K^b$.

The identity weights each mean and covariance error by its remaining effect on final reward. It applies without differentiating sampled actions, assuming contraction, or coupling the methods' trajectories. A large sampling-law change can nevertheless overwhelm a reference advantage.

Proof of Theorem A.2. Condition on the actual history $\mathcal{F}_k^a$. Abbreviate $\theta = \theta_k^a$, $m = \overline{M}_{k,a}\theta + \overline{d}_{k,a}$, $e = e_k^a$, $C = C_k^a$, suppress a on the value coefficients, and set $P = P_{k+1}$. By definition, $\mathbb{E}_k \theta_{k+1}^a = m + e$. A quadratic function has the exact conditional expectation identity

$$\mathbb{E}_k V_{k+1}(\theta_{k+1}^a) = V_{k+1}(m + e) + \tfrac{1}{2}\operatorname{tr}(PC). \tag{37}$$

To verify it, write $\theta_{k+1}^a = m + e + \epsilon$; its conditional centered error satisfies $\mathbb{E}_k \epsilon = 0$ and $\mathbb{E}_k \epsilon\epsilon^\top = C$. In the expansion of (30), the terms linear in ϵ have zero conditional mean and the remaining quadratic term is $\mathbb{E}_k[\epsilon^\top P \epsilon]/2 = \operatorname{tr}(PC)/2$.

The quadratic Taylor identity has no remainder:

$$V_{k+1}(m + e) - V_{k+1}(m) = \langle Pm + p_{k+1}, e \rangle + \tfrac{1}{2} e^\top P e. \tag{38}$$

Subtract (32), evaluated at the actual θ, from (37). Since $\nabla V_{k+1}(m) = Pm + p_{k+1}$, the result is exactly

$$\mathbb{E}_k V_{k+1}(\theta_{k+1}^a) - V_k(\theta_k^a) = Z_k^a. \tag{39}$$

No derivative of $Q_k(\theta)$ enters: the two kernels are compared at the same current state, even when they sample different next outcomes. Finite second moments make every term integrable: conditional Jensen bounds the squared conditional means, and the finite deterministic coefficient matrices bound all quadratic terms. Taking total expectations and summing (39) over $k = 0, \dots, K-1$ telescopes to

$$\sum_{k<K} \mathbb{E} Z_k^a = \mathbb{E} V_K(\theta_K^a) - V_0(\theta_0) \tag{40}$$

$$= \tfrac{1}{2}\mathbb{E}[(\theta_K^a)^\top H_R \theta_K^a] - \mathcal{R}_K^a. \tag{41}$$

This proves (35). The triangle inequality and the assumed terminal target approximation then give

$$|\mathfrak{R}_K^a - \mathcal{R}_K^a| \le \mathbb{E}|\mathcal{L}(\theta_K^a) - \tfrac{1}{2}(\theta_K^a)^\top H_R \theta_K^a| + \left| \sum_{k<K} \mathbb{E} Z_k^a \right| \tag{42}$$

$$\le \nu_a + \sum_{k<K} \mathbb{E}|Z_k^a| = \rho_a. \tag{43}$$

Finally, $\mathfrak{R}_K^b - \mathfrak{R}_K^a \ge \mathcal{R}_K^b - \mathcal{R}_K^a - \rho_a - \rho_b$. A strictly positive right side proves the claimed ordering. □

The mean error accounts for nonlinear drift, changing curvature, and estimator bias. The covariance error accounts for policy-dependent sampling, including correlation between the two update channels. There is no requirement that an actual sample field have the reference Hessian. The price of this generality is explicit: the identity alone supplies no small bound on these errors.

A.4 A General Finite-Radius Transfer Statement

Why linearizing the mean is insufficient. Random visitation multiplies the current error, producing variation on the same scale as local reward risk. The counterexample below shows why retaining this variation matters even without added reward noise.

A reference for a class of nonlinear updates. Fix a finite budget K, center the reward optimum at zero, and write $\theta_k = \varepsilon x_k$ for $0 < \varepsilon \le \varepsilon_0$. A recipe a specifies its estimator, steps, and teacher weights. All recipes start at $\theta_0 = \varepsilon x_0$. At update k, its scaled nonlinear map is $f^{\varepsilon}_{k,a,j}(x)$ with probability $\pi^{\varepsilon}_{k,a,j}(x)$, for finitely many outcomes j. Its reference uses affine maps $f^{0}_{k,a,j}(x) = M_{k,a,j}x + d_{k,a,j}$ and state-independent probabilities $\pi^{0}_{k,a,j}$, drawn freshly each step. The laws may differ between recipes; the matrix and drift may be correlated.

Let $B_k = \{x : \|x\| \le R_k\}$, with finite radii, satisfy $x_0 \in B_0$ and $f^{\varepsilon}_{k,a,j}(B_k), f^{0}_{k,a,j}(B_k) \subseteq B_{k+1}$. Assume finite nonnegative constants satisfy, uniformly over the allowed radii, states, and outcomes,

$$\|f^{\varepsilon}_{k,a,j}(x) - f^{0}_{k,a,j}(x)\| \le \varepsilon b_{k,a}, \qquad \sum_j |\pi^{\varepsilon}_{k,a,j}(x) - \pi^{0}_{k,a,j}| \le \varepsilon l_{k,a},$$
$$|\mathcal{L}(\varepsilon x)/\varepsilon^2 - x^\top H_R x/2| \le \varepsilon t_a \quad (x \in B_K), \qquad H_R \succeq 0. \tag{44}$$

The first two bounds apply on B_k; all constants are independent of ε. In words, we control changes in each possible update, its sampling probability, and the terminal reward approximation. Smoothness on verified visited sets supplies such bounds by finite Taylor inequalities; initial error, teacher mismatch, and bounded external noise may scale together. Proximity alone does not establish these assumptions.

Let $V_{k,a}(x) = x^\top P_{k,a} x/2 + p_{k,a}^\top x + \gamma_{k,a}$ be the reference's expected terminal quadratic risk. Its exact backward recursion retains $\mathbb{E}[M^\top P M]$, not $(\mathbb{E}M)^\top P(\mathbb{E}M)$ (Lemma A.1). Define

$$G_{k,a} = \|P_{k+1,a}\|_{\mathrm{op}} R_{k+1} + \|p_{k+1,a}\|,$$
$$O_{k,a} = \tfrac{1}{2}\|P_{k+1,a}\|_{\mathrm{op}} R_{k+1}^2 + 2\|p_{k+1,a}\| R_{k+1}, \quad C_{K,a} = t_a + \sum_{k<K} (G_{k,a} b_{k,a} + O_{k,a} l_{k,a}/2). \tag{45}$$

These weights measure how much a local update or sampling error can affect final reward; $C_{K,a}$ is computable once the stated envelopes are known.

Theorem A.3 (Finite-radius comparison against tuned recipe classes)**.** *Under* (44)*, put* $\mathcal{R}^{\mathrm{ra}}_{K,a} = V_{0,a}(x_0)$ *and* $\mathfrak{R}^{\varepsilon}_{K,a} = \mathbb{E}\mathcal{L}(\theta^a_K)$*. For every* $0 < \varepsilon \le \varepsilon_0$*,*

$$|\mathfrak{R}^{\varepsilon}_{K,a} - \varepsilon^2 \mathcal{R}^{\mathrm{ra}}_{K,a}| \le C_{K,a}\varepsilon^3. \tag{46}$$

Let a fixed mixed recipe have reference margin $\gamma = \inf_{a \in \mathcal{A}_R \cup \mathcal{A}_D} \mathcal{R}^{\mathrm{ra}}_{K,a} - \mathcal{R}^{\mathrm{ra}}_{K,M} > 0$ *over two pure recipe classes. If the envelopes hold uniformly and* $C_P = \sup_{a \in \mathcal{A}_R \cup \mathcal{A}_D} C_{K,a} < \infty$*, then*

$$\inf_{a \in \mathcal{A}_R \cup \mathcal{A}_D} \mathfrak{R}^{\varepsilon}_{K,a} - \mathfrak{R}^{\varepsilon}_{K,M} \ge \varepsilon^2\{\gamma - \varepsilon(C_{K,M} + C_P)\}. \tag{47}$$

Thus a positive bracket certifies strict mixed superiority, including against continuous pure recipe classes.

In plain language, a reference advantage survives nonlinear learning when it exceeds the explicit approximation error. This finite inequality requires verified containment and uniformity, not contraction or a common sampling law. It does not automatically give an advantage or a large useful neighborhood.

The proof below bounds update-map and sampling-law errors separately, then sums their effects on the same future-value function. Appendix G supplies sharper conditional-moment bounds for the numerical certificates. All constructions use known inputs; estimated inputs would require simultaneous coverage and charged queries.

Obtaining the finite-radius envelopes. Theorem A.3 above states the comparison for smooth finite-outcome kernels. Its three envelopes (44) concern update maps, sampling probabilities, and

terminal loss. Outcome labels are held fixed when differentiating a map; their changing probabilities are bounded separately. Initial error, teacher mismatch, and bounded external noise may all scale with the radius. This is a family of local problems, not a claim that arbitrary fixed nonvanishing noise disappears near an optimum.

The envelopes follow from finite Taylor bounds. For example, let $F_{k,a,j}(\theta, \varepsilon)$ be the unscaled sample map with $F_{k,a,j}(0,0) = 0$. If its derivative along $(\theta, \varepsilon) = (tx, t)$ at zero is $M_{k,a,j}x + d_{k,a,j}$, and its second derivative along that segment has norm at most $2b_{k,a}$, the integral Taylor remainder gives the first envelope after division by ε. A bound $l_{k,a}$ on the sum of absolute probability derivatives along the same segments gives the second. If $\mathcal{L}(0) = 0$, $\nabla\mathcal{L}(0) = 0$, and its Hessian is L_H-Lipschitz, the third holds with $t_a = L_H R_K^3/6$. Smooth finite-action policies with bounded scaled noise have these derivative bounds on compact visited sets; containment and uniformity must still be verified. No contraction or equality of sample Jacobians is assumed.

Proof of Theorem A.3. Condition on the actual scaled state $x \in B_k$. For shortness put $V = V_{k+1,a}$, $f_j = f^{\varepsilon}_{k,a,j}(x)$, $f^0_j = f^0_{k,a,j}(x)$, $\pi_j = \pi^{\varepsilon}_{k,a,j}(x)$, and $\pi^0_j = \pi^0_{k,a,j}$. The one-step value discrepancy is

$$\begin{aligned} z_k(x) &= \sum_j \pi_j V(f_j) - \sum_j \pi^0_j V(f^0_j) \\ &= \sum_j \pi_j \{V(f_j) - V(f^0_j)\} + \sum_j (\pi_j - \pi^0_j) V(f^0_j). \end{aligned} \tag{48}$$

Both endpoints of each segment between f_j and f^0_j lie in the convex ball B_{k+1}. On that ball, $\|\nabla V(z)\| = \|P_{k+1,a}z + p_{k+1,a}\| \le G_{k,a}$. The fundamental theorem of calculus therefore bounds the absolute first sum by $\varepsilon G_{k,a} b_{k,a}$.

For the second sum, its coefficients add to zero. Subtract the midpoint of the maximum and minimum of V on B_{k+1} from every value. The absolute result is at most one half its oscillation times $\sum_j |\pi_j - \pi^0_j|$. Since $P_{k+1,a} \succeq 0$, the quadratic part of V ranges between zero and $\|P_{k+1,a}\|_{\mathrm{op}} R^2_{k+1}/2$; the linear part has range width at most $2\|p_{k+1,a}\| R_{k+1}$. Hence its oscillation is at most $O_{k,a}$, and the second sum is bounded by $\varepsilon O_{k,a} l_{k,a}/2$. Combining proves

$$|z_k(x)| \le \varepsilon\{G_{k,a} b_{k,a} + O_{k,a} l_{k,a}/2\} \quad \text{for every } x \in B_k. \tag{49}$$

The second expectation in the first line of (48) is $V_{k,a}(x)$ by the reference recursion. Take total expectations along the actual run and telescope over exactly K steps. This gives

$$|\mathbb{E}[x_K^\top H_R x_K/2] - V_{0,a}(x_0)| \le \varepsilon \sum_{k<K} \{G_{k,a} b_{k,a} + O_{k,a} l_{k,a}/2\}. \tag{50}$$

Add the terminal envelope and multiply by ε^2 to obtain (46).

If all bounds hold uniformly over a recipe class, the same deterministic argument applies to every member; no statistical union bound or finite grid is being used. For each pure recipe a,

$$\mathfrak{R}^{\varepsilon}_{K,a} - \mathfrak{R}^{\varepsilon}_{K,M} \ge \varepsilon^2(\mathcal{R}^{\mathrm{ra}}_{K,a} - \mathcal{R}^{\mathrm{ra}}_{K,M}) - \varepsilon^3(C_{K,a} + C_{K,M}) \tag{51}$$

$$\ge \varepsilon^2\{\gamma - \varepsilon(C_P + C_{K,M})\}. \tag{52}$$

Taking the infimum over each pure class proves the last claim. It requires no attainment of those infima. When $C_P + C_{K,M} = 0$, every allowed radius retains the strict reference gap; otherwise the explicit sufficient restriction is $\varepsilon < \gamma/(C_P + C_{K,M})$. □

Specialization to the two-stage policy. Let p_0, s_0 be its optimal action probabilities and $h_p = p_0(1-p_0)$, $h_s = s_0(1-s_0)$. Scale the initial logit errors, teacher intercepts, and external reward noise by ε. At the reference draw $A \sim \mathrm{Ber}(p_0)$ and, when visited, $B \sim \mathrm{Ber}(s_0)$; bounded centered η is independent. For a normalized recipe with effective step λ and teacher fraction w, put $r = 1 - w$, $t = w$, $a = A - p_0$, $b = B - s_0$. Its samplewise affine map has

$$M = \begin{pmatrix} 1 - \lambda(ra^2 + th_p) & -\lambda r A a b \\ 0 & 1 - \lambda A(rb^2 + th_s) \end{pmatrix}, \qquad d = \lambda \begin{pmatrix} r\eta a + th_p h_1 \\ A(r\eta b + th_s h_2) \end{pmatrix}. \tag{53}$$

These formulas follow by differentiating the displayed sample maps at zero error and zero mismatch, holding the categorical outcome fixed. In particular, the leaf KL in the root estimator has zero first derivative there. Averaging gives the diagonal mean transition and drift, but M itself is random. State-dependent sampling probabilities are accounted for by the separate law-error term in (45). A finite discrete law for η directly fits the theorem; a bounded continuous law fits the same proof by integrating that independent noise instead of summing it.

Why sampling can change the winner. We give an analytic example, distinct from the continuously tuned main proposition. Take $p_0 = s_0 = 3/5$, so both optimal logits are $\log(3/2)$. The nonconstant task rewards are zero, $\log(6/5)$, and $\log(9/5)$; the uniform-reference KL coefficient remains one. Start at logit errors $(0, \varepsilon)$, use teacher probabilities $(3/5, 3/5 + (6/25)\varepsilon h)$, and add no external reward noise. Use one update with effective step one and fractions $w = 0, 1, 1/2$ for RL, OPD, and mixing. The same episode supplies both channels.

The exact random-affine risks are the following polynomials; the terminating decimals are exact rational numbers:

$$\begin{aligned} R_R(h) &= .05693184, \\ R_D(h) &= .05375232 + .01575936h + .00248832h^2, \\ R_M(h) &= .05454720 + .00787968h + .00062208h^2. \end{aligned} \tag{54}$$

They give a complete reference phase diagram on $[0, 1/2]$: put $\tau = (\sqrt{1582} - 38)/18 \approx .098576$. OPD wins for $h < \tau$, mixing for $\tau < h < 3\tau$, and RL for $h > 3\tau \approx .295727$. At either boundary the adjacent methods tie. By contrast, the additive-noise reference, with covariance evaluated at the reward optimum, gives

$$\begin{aligned} \widetilde{R}_R(h) &= .052756992, \\ \widetilde{R}_D(h) &= .052756992 + .017750016h + .00248832h^2, \\ \widetilde{R}_M(h) &= .052756992 + .008875008h + .00062208h^2. \end{aligned} \tag{55}$$

It strictly favors RL for every $h > 0$. This is an unguarded reference prediction, not a failure of the comparison rule with its error bounds.

Proposition A.4 (Visitation noise changes the finite-budget ordering)**.** *For this nonlinear family, simultaneously for all three fixed recipes,* $0 \le h \le 1/2$*, and* $0 < \varepsilon \le 1/10$*,*

$$|\mathfrak{R}_1^a(\varepsilon, h)/\varepsilon^2 - R_a(h)| \le \varepsilon/10. \tag{56}$$

Consequently a polynomial pairwise margin exceeding $\varepsilon/5$ *certifies the actual nonlinear ordering. In particular, for every* $0 < \varepsilon \le 10^{-3}$*, OPD wins throughout* $h \in [.04, .06]$*, mixing wins throughout* $h \in [.14, .16]$*, and RL wins throughout* $h \in [.39, .41]$*. All rankings are strict against both alternatives.*

For every $0 < \varepsilon \le 10^{-3}$, the proposition gives the following actual nonlinear rankings. The mean-coefficient prediction, which also evaluates additive covariance at the optimum, misses the first two winners.

Teacher offset h	Actual winner	Mean-coefficient prediction
$[.04, .06]$	OPD	RL
$[.14, .16]$	Mixing	RL
$[.39, .41]$	RL	RL

This comparison has no added reward noise, yet sampled actions and teacher visitation still create useful differences between the signals. It shows why proximity to the optimum alone cannot justify an additive-noise ordering. It isolates omitted state-dependent sampling variation, not solely the temporal change in visitation probabilities. The three recipes share a normalized effective step, but their rates are not independently optimized here; Theorem 4.1 supplies a separate control. This fixed-rate example is not an equal-cost comparison. Neither result should be substituted for the other.

Proof. Put $p_0 = s_0 = 3/5$, $h_p = h_s = 6/25$, and $H_R = \mathrm{diag}(6/25, 18/125)$. At $x_0 = (0, 1)$, the reference outcomes in (53), without external noise, are

$$x_1^0 = \begin{pmatrix} -rA(A - p_0)(B - s_0) \\ 1 - A\{r(B - s_0)^2 + th_s(1 - h)\} \end{pmatrix}. \tag{57}$$

The three distinct action outcomes $(A, B) = (0, 0), (1, 0), (1, 1)$ have probabilities $2/5, 6/25, 9/25$; the leaf label on termination is a dummy. Square the two coordinates, weight them by $H_R/2$, and sum with these probabilities. Substituting $(r, t) = (1, 0), (0, 1), (1/2, 1/2)$ gives (54) exactly. The differences $R_D - R_M$ and $R_M - R_R$ are positive multiples of $54h^2 + 228h - 23$ and $6h^2 + 76h - 23$, respectively. Each is strictly increasing for $h \geq 0$ and has its unique positive zero at τ or 3τ. Their signs give the stated three-way phase diagram, including its ties. For the additive reference, the mean leaf coordinate is $1 - 18/125 + t(18/125)h$, and its only nonzero variance is $t^2(6/25)(6h/25)^2$. Substitution in the same quadratic risk gives (55).

We now establish a uniform finite error constant, rather than relying on agreement at a few radii. Let $s = \mathrm{sigmoid}(\log(3/2) + \varepsilon)$ and $\delta = s - s_0$. The logistic variance obeys $v \leq 1/4$ and $|v'| \leq 1/10$ globally, since its exact derivative maximum is $1/(6\sqrt{3}) < 1/10$. Thus

$$0 \leq \delta \leq \varepsilon/4, \quad |\delta - h_s\varepsilon| \leq \varepsilon^2/20, \quad D_s = \mathrm{KL}(s\|s_0) \leq \varepsilon^2/8. \tag{58}$$

For $\varepsilon \leq 1/10$, $s \in [3/5, 5/8]$ and $v \geq 1/5$ on this segment, so $\delta \geq \varepsilon/5$. The teacher leaf target is at most $s_0 + (3/25)\varepsilon < s$.

Subtract the reference outcomes from the actual outcomes after division by ε. The root discrepancy is bounded by

$$r\{(3/5)^2 D_s/\varepsilon + A|A - p_0|\delta\} \leq \left(\frac{9}{200} + \frac{1}{10}\right)\varepsilon = \frac{29}{200}\varepsilon. \tag{59}$$

Here $A|A - p_0| \leq 2/5$. At the leaf, expanding $(B - s)^2$ around $(B - s_0)^2$ bounds the reward discrepancy by $2(3/5)\delta + \delta^2 \leq (3/10 + 1/160)\varepsilon$. The teacher discrepancy is at most $\varepsilon/20$. Because $r, t \geq 0$ and $r + t = 1$, their weighted sum is at most $(49/160)\varepsilon$.

Both actual and reference scaled outputs belong to $\mathcal{B} = [-13/50, 13/50] \times [0, 1]$. For the actual root, its magnitude is at most $(9/200)\varepsilon + (2/5)(5/8) \leq .2545 < 13/50$. The actual leaf never increases: both reward and teacher terms are nonpositive. It stays nonnegative because the reward decrement is at most $r\varepsilon$ and the teacher decrement is at most $t\varepsilon/4$. For the reference root the magnitude is at most $(2/5)(3/5) = .24$; its leaf decrement lies between zero and one. This proves the claimed containment, including the segments between the two outputs.

The terminal reference quadratic $Q(x, y) = 3x^2/25 + 9y^2/125$ has coordinate gradient bounds $(6/25)(13/50)$ and $18/125$ on $\mathcal{B}$. Its oscillation is at most $O = \{(6/25)(13/50)^2 + 18/125\}/2 = .080112$. The total absolute difference of the three outcome probabilities is $2p_0\delta \leq (3/10)\varepsilon$. The map/law split (48), using this rectangle in place of a ball, therefore bounds the scaled quadratic-risk error by

$$\varepsilon\left\{\frac{6}{25}\frac{13}{50}\frac{29}{200} + \frac{18}{125}\frac{49}{160} + \frac{O}{2}\frac{3}{10}\right\}. \tag{60}$$

Finally, at an actual output $(\varepsilon x, \varepsilon y)$, the exact loss is the root Bernoulli KL plus its root probability times the leaf KL. The derivative of each KL with respect to its logit error z is $v(z)z$. Integrating $|v(z) - v(0)| \leq |z|/10$ bounds its quadratic remainder by $|z|^3/30$. The actual root probability is at most $3/5 + (1/4)(1/10)(13/50) = 1213/2000$. Adding and subtracting the root-optimum-weighted leaf quadratic, as in (231), gives

$$|\mathcal{L}(\varepsilon x, \varepsilon y)/\varepsilon^2 - Q(x, y)| \leq \varepsilon\left\{\frac{(13/50)^3}{30} + \frac{1213/2000}{30} + \frac{3}{100}\frac{13}{50}\right\}. \tag{61}$$

The sum of the two displayed rational constants is $.0937673 < 1/10$, proving (56).

To verify the three entire teacher intervals, minimize the quadratic differences $R_b(h) - R_a(h)$ on each interval. Their minima occur at an endpoint or an included stationary point. The smallest gaps against both competitors on the three intervals are respectively .000315380736, .000344853504, and .000783053568. Each exceeds $\varepsilon/5$ for every $0 < \varepsilon \leq 10^{-3}$. Subtracting the two error bounds proves all three strict nonlinear rankings. Exact rational arithmetic checks the polynomial margins, and outward-rounded arithmetic independently checks the nonlinear comparison at one point. The uniform claim follows from the error bound above, not from that pointwise check. □

B Main Nonlinear Separation and Parameter Window

The proof of Theorem 4.1 gives the schedule-class bound, teacher floor, mixed risk, and nonlinear correction. An information-floor lemma then supplies the explicit window in Proposition 4.2.

B.1 A Structural Comparison with Free Model and Budget Parameters

This section proves Theorem 4.1. The model parameters, teacher mismatch, noise variance, step cap, and query costs are not fixed to the later numerical construction. All statements hold at explicitly bounded positive radii and finite horizons.

Throughout this section, $h_p = p_\star(1-p_\star)$ and $h_s = s_\star(1-s_\star)$ are fixed variances at the optimum, as in the main text. Current-policy variances are denoted by $v_p = p(1-p)$ and $v_s = s(1-s)$; their lower bounds on a rectangle are $\underline{v}_p, \underline{v}_s$.

Actual objective and estimators. Set $u = \mathrm{logit}(p_\star)$ and $v = \mathrm{logit}(s_\star)$. A root failure pays zero, while a visited leaf action B pays $r_0 + vB$, with $r_0 = u - \log\{(1+e^v)/2\}$. Use uniform reference policies and reference-KL coefficient one. The variational identity for a Bernoulli log-partition gives leaf value $u - \mathrm{KL}(s\|s_\star)$; optimizing the root gives $J^\star - J = \mathrm{KL}(p\|p_\star) + p\mathrm{KL}(s\|s_\star)$. Thus the loss in the main text is an actual regularized reward gap.

For $A \sim \mathrm{Ber}(p)$ and, when $A = 1$, $B \sim \mathrm{Ber}(s)$, put $a = A - p$, $b = B - s$ and $D_s = \mathrm{KL}(s\|s_\star)$. Exact statewise baselines give the sampled RL and teacher directions

$$\widehat{g} = (-(x + D_s)a^2 - Ayab + \varepsilon\zeta a, A[-yb^2 + \varepsilon\zeta b]), \qquad \widehat{u} = (p_\star + \varepsilon h_p h_1 - p, A[s_\star + \varepsilon h_s h_2 - s]). \tag{62}$$

The derivation from raw regularized utilities is the same algebra as Appendix A.2, with arbitrary optimal logits. Fresh noise is independent of the actions and past and has the stated common variance. No symmetry of its distribution is needed.

Exact reference moments. Freeze action probabilities at $p_\star, s_\star$, but retain their draws and sampled coefficients. For scaled state (X, Y), teacher fraction w, reward fraction $r = 1 - w$, and effective step α, the reference is

$$\begin{aligned} X' &= X + \alpha\{r(-Xa^2 - AYab + \zeta a) + wh_p(h_1 - X)\}, \\ Y' &= Y + \alpha A\{r(-Yb^2 + \zeta b) + wh_s(h_2 - Y)\}, \end{aligned} \tag{63}$$

now with $a = A - p_\star$, $b = B - s_\star$. Write $F(z) = z(1-3z)$ for the fourth central moment of a Bernoulli variable with variance z. Let $m = \mathbb{E}X$, $n = \mathbb{E}Y$, $U = \mathbb{E}X^2$, $V = \mathbb{E}Y^2$, and $b_0 = p_\star(1-p_\star)^2 h_s$. Their exact recursions are

$$\begin{aligned} m' &= (1 - \alpha h_p)m + \alpha w h_p h_1, \\ n' &= (1 - \alpha p_\star h_s)n + \alpha w p_\star h_s h_2, \\ U' &= [1 - 2\alpha h_p + \alpha^2\{r^2 F(h_p) + (1-r^2)h_p^2\}]U + \alpha^2 r^2 b_0 V \\ &\quad + 2\alpha w h_p h_1(1 - \alpha h_p)m + \alpha^2 w^2 h_p^2 h_1^2 + \alpha^2 r^2 \nu h_p, \\ V' &= [1 - 2\alpha p_\star h_s + \alpha^2 p_\star\{r^2 F(h_s) + (1-r^2)h_s^2\}]V \\ &\quad + 2\alpha w p_\star h_s h_2(1 - \alpha h_s)n + \alpha^2 w^2 p_\star h_s^2 h_2^2 + \alpha^2 r^2 \nu p_\star h_s. \end{aligned} \tag{64}$$

Initialize $(m, n, U, V) = (x_0, y_0, x_0^2, y_0^2)$, iterate the fixed mixed recipe K_M times, and put $R_M = (h_p U + p_\star h_s V)/2$. These four scalars retain random visitation and reward–teacher dependence; there is no independence assumption between X and Y.

To verify the identities, use $\mathbb{E}a = \mathbb{E}b = 0$, $\mathbb{E}a^2 = h_p$, $\mathbb{E}b^2 = h_s$, $\mathbb{E}a^4 = F(h_p)$, $\mathbb{E}b^4 = F(h_s)$, and $\mathbb{E}[Aa^2] = p_\star(1-p_\star)^2$. In the root square, each term containing one b vanishes after conditioning on A; each term linear in ζ vanishes by centering and independence. The squared root multiplier has expectation $1 - 2\alpha h_p + \alpha^2[r^2 F(h_p) + (2rw + w^2)h_p^2]$. Since $2rw + w^2 = 1 - r^2$, this gives the coefficient of U. The cross term with the deterministic teacher intercept uses the mean multiplier $1 - \alpha h_p$. At the leaf, condition first on $A = 1$; the identical calculation has an additional factor $p_\star$. This proves all four lines, including their shared-action dependence.

Bounding all prescribed RL schedules. We approximate the updates locally while retaining the random action draws. The resulting reference model has closed recursions for the two RL second moments. Retain the coupling coefficient b_0 above and start $(\ell_{1,0}, \ell_{2,0}) = (x_0^2, y_0^2)$. For a pure step cap $a_{\max}$, define

$$\ell_{1,k+1} = \min_{0 \le a \le a_{\max}} \{(1 - 2ah_p + a^2 F(h_p))\ell_{1,k} + a^2 b_0 \ell_{2,k} + a^2 \nu h_p\},$$
$$\ell_{2,k+1} = \min_{0 \le a \le a_{\max}} \{(1 - 2ap_\star h_s + a^2 p_\star F(h_s))\ell_{2,k} + a^2 \nu p_\star h_s\}. \tag{65}$$

Each minimum is a scalar quadratic evaluated at its clipped vertex. Allowing the two coordinates to choose different steps gives a lower bound for *every* common prescribed schedule. The minimizing choices need not form an implementable schedule, but the bound is valid for all of them and requires work linear in the budget.

Why the two-coordinate minimization covers every RL schedule. Set $w = 0$ in (64). Both next second moments are affine functions of (U, V) with nonnegative coefficients: the diagonal coefficients are expectations of squared sample multipliers, and the cross-coordinate coefficient is $\alpha^2 b_0 \ge 0$. Suppose $U_k \ge \ell_{1,k}$ and $V_k \ge \ell_{2,k}$. Substitute these lower endpoints at any prescribed next step and then minimize each coordinate over its allowed step interval. This proves $U_{k+1} \ge \ell_{1,k+1}$ and $V_{k+1} \ge \ell_{2,k+1}$ by induction. Separate coordinate minimizers need not correspond to one implementable schedule; allowing them can only reduce the lower bound.

More explicitly, put $q_1 = F(h_p)\ell_1 + b_0\ell_2 + \nu h_p$ and $q_2 = p_\star F(h_s)\ell_2 + \nu p_\star h_s$. The minimizing steps are $\mathrm{clip}(h_p\ell_1/q_1; [0, a_{\max}])$ and $\mathrm{clip}(p_\star h_s \ell_2/q_2; [0, a_{\max}])$. If a denominator is zero, its moment and linear term are zero and step zero is valid. Since zero is allowed, both lower sequences are nonincreasing. Their N_R-update value therefore also bounds every shorter prescribed count. Alternatively, pad a shorter schedule with zero steps. This proof does not apply to state-dependent RL steps: their correlation with (X, Y) prevents taking the step outside the moment expectation.

Explicit uniform nonlinear constants. Choose a rectangle $\mathcal{B} = [-R_1, R_1] \times [-R_2, R_2]$ and $\varepsilon_0 > 0$. Let p_-, p_+ be the logistic endpoints at $u \pm \varepsilon_0 R_1$, and similarly s_-, s_+ at $v \pm \varepsilon_0 R_2$. Define

$$\begin{aligned} a_- &= \min(p_-, 1 - p_+), & a_+ &= \max(p_+, 1 - p_-), \\ b_- &= \min(s_-, 1 - s_+), & b_+ &= \max(s_+, 1 - s_-), \\ \underline{v}_p &= \min_{z \in \{p_-, p_+\}} z(1 - z), & \underline{v}_s &= \min_{z \in \{s_-, s_+\}} z(1 - z). \end{aligned} \tag{66}$$

The following directly checkable conditions suffice:

$$\begin{aligned} &|x_0| \le R_1, \quad |y_0| \le R_2, \quad 0 < \lambda \le a_{\max} \le 4, \quad a_{\max} \max(a_+^2, b_+^2) \le 1, \\ &\varepsilon_0 R_2^2/8 + R_2 b_+/a_- + T/a_- \le R_1, \qquad T/b_- \le R_2, \\ &h_p|h_1| \le R_1 \underline{v}_p, \quad h_s|h_2| \le R_2 \underline{v}_s, \quad 0 < h_p h_1 \le x_0 \underline{v}_p. \end{aligned} \tag{67}$$

For every $0 < \varepsilon \le \varepsilon_0$, these conditions contain every actual and reference state in the scaled rectangle, for all allowed pure schedules and the fixed mixture. They also ensure valid teacher probabilities. Their failure means this sufficient test abstains.

Here is the containment proof. At a visited leaf the actual RL update, divided by ε, is $(1 - \alpha b^2)Y + \alpha b^2 \zeta/b$; at an unvisited leaf it is unchanged. The root is the convex combination of X and $-D_s/\varepsilon - AYb/a + \zeta/a$. Because $D_s \le \varepsilon^2 R_2^2/8$, the first two lines of (67) bound these target points by R_2, R_1. For OPD, each visited coordinate map has derivative $1 - \alpha\,\mathrm{sigmoid}' \ge 0$ at steps through four and fixes the teacher logit. The teacher inequalities place those logits inside the rectangle, using the derivative lower bound along each segment. Thus OPD maps the rectangle to itself, and a normalized mixture is a convex combination of two contained maps. The reference has the same argument at the optimal probabilities, with $D_s = 0$ and teacher targets (h_1, h_2). Induction proves containment without any statement about contraction in expectation.

Put $R = \max(R_1, R_2)$, $H_0 = h_p + p_\star h_s$, and define

$$\begin{aligned} B_R &= \max\{R_1^2/2 + 3R_2^2/8 + R_1R_2/4 + TR_1/4, R_2^2/2 + TR_2/4\}, \\ B_D &= \max(R_1^2, R_2^2)/20, \qquad L_\pi = (R_1 + p_\star R_2)/2, \\ Q_{\max} &= (h_p R_1^2 + p_\star h_s R_2^2)/2, \\ L &= 1 + a_{\max}(1 - p_\star)[\max(s_\star, 1 - s_\star) - (1 - p_\star)]_+, \\ t &= R_1^3/30 + R_2^3/30 + h_s R_1 R_2^2/8, \\ C_R &= t + H_0 R a_{\max} B_R \sum_{j=0}^{N_R - 1} L^j + N_R Q_{\max} L_\pi/2, \\ C_M &= t + H_0 R \lambda\{(1 - w)B_R + wB_D\} \sum_{j=0}^{K_M - 1} L^j + K_M Q_{\max} L_\pi/2. \end{aligned} \tag{68}$$

These are finite sums, including when $L = 1$; an empty sum is zero. They favor transparent dependence over tight numerical radii.

For detail, $|p - p_\star| \le \varepsilon R_1/4$ and $|s - s_\star| \le \varepsilon R_2/4$. Comparing the two raw RL maps term by term gives scaled coordinate errors at most $\alpha\varepsilon B_R$. The root contributions are respectively $R_1^2/2$, $R_2^2/8$, $R_2(R_1 + R_2)/4$, and $TR_1/4$. The leaf contributions are $R_2^2/2$ and $TR_2/4$. The global logistic second derivative has magnitude at most $1/(6\sqrt{3}) < 1/10$; Taylor's integral remainder gives the teacher map error $\alpha\varepsilon B_D$. The three branch probabilities $(1 - p, p(1 - s), ps)$ have total absolute discrepancy at most $2|p - p_\star| + 2p_\star|s - s_\star| \le \varepsilon L_\pi$. Integrate over the unchanged noise law; no finite noise support is required.

Every reference sample matrix has infinity norm at most L. Only a visited root row can exceed one: its row sum is $1 + \alpha(1 - p_\star)(|B - s_\star| - (1 - p_\star))$. Nonnegative diagonal multipliers follow from the box tests. Teacher matrices have norm at most one, and mixture matrices are their convex combinations with RL matrices. A reference value with j steps remaining consequently has gradient one-norm at most $H_0 R L^j$: differentiate each affine history, use its matrix norm at most L^j, and bound the terminal quadratic gradient by $H_0 R$. Its value lies in $[0, Q_{\max}]$ on the entire rectangle. The map/law telescoping proof of Theorem A.3, now on this convex rectangle, therefore gives map charge $H_0 R L^j \alpha\varepsilon B$ and law charge $\varepsilon Q_{\max} L_\pi/2$ per step. Finally $D'_p(x) = xp'(x)$ gives root terminal error $\varepsilon^3 R_1^3/30$; the leaf KL and changed root probability add $\varepsilon^3(R_2^3/30 + h_s R_1 R_2^2/8)$. Multiplying the telescoped scaled bound by ε^2 proves the actual/reference error bounds $C_R\varepsilon^3$ and $C_M\varepsilon^3$, uniformly over the full prescribed RL class.

An OPD floor valid under feedback and stopping. The final inequality in (67) ensures that the initial root probability is at least $q = p_\star + \varepsilon h_p h_1 > p_\star$. Every root OPD step through four preserves the interval between its current logit and the teacher logit. Hence every OPD stopping point has root probability at least q and regret at least $\mathrm{KL}(q\|p_\star)$. For $q_- = p_\star - \varepsilon_0 h_p|h_1| > 0$ and $q_+ = p_\star + \varepsilon_0 h_p|h_1| < 1$, set

$$t_D = \frac{h_p^3|h_1|^3}{6}\{q_-^{-2} + (1 - q_+)^{-2}\}. \tag{69}$$

These additional segment inequalities are required for this displayed constant. In probability coordinates the KL has first derivative zero and second derivative $1/h_p$ at $p_\star$; its third derivative is $-q^{-2} + (1 - q)^{-2}$. Taylor's finite remainder therefore proves $\mathrm{KL}(q\|p_\star) \ge \varepsilon^2 h_p h_1^2/2 - \varepsilon^3 t_D$. This bound is pathwise, so adaptive steps and stopping do not invalidate it.

Completion of Theorem 4.1. Every affordable prescribed RL schedule has reference risk at least $(h_p\ell_{1,N_R} + p_\star h_s \ell_{2,N_R})/2$ and actual risk at least that quantity times ε^2, minus $C_R\varepsilon^3$. Every OPD rule has the lower bound just proved. The fixed mixture has actual risk at most $\varepsilon^2 R_M + C_M\varepsilon^3$. Subtract its upper bound from the smaller pure lower bound to obtain $\varepsilon^2(\Gamma - \varepsilon E)$. The stated positive radius makes this at least $\varepsilon^2\Gamma/2$. Counts are determined by the same declared query costs, and no optimizing pure schedule was excluded by the coordinate relaxation. This proves the result. □

A concrete positive radius, and its limitation. For $p_\star = s_\star = 1/2$, $(x_0, y_0) = (1, 1)$, $(h_1, h_2) = (1/2, -1/2)$, $\nu = 16/25$, $T = 4/5$, $a_{\max} = 2$, $K_M = 20$, $\lambda = 3/5$, and $w = 1/2$, costs $(c_0, c_R, c_D) = (3, 1, 1)$ and $B = 100$ give $N_R = 25$. The recursions yield $\Gamma > .00369$. Taking $(R_1, R_2) = (10, 3)$ and $\varepsilon_0 = 1/25$ passes the box tests and gives $E < 15081$; thus every $0 < \varepsilon \le 10^{-7}$ has a strict certified nonlinear gap. This elementary radius is intentionally conservative, not evidence of a large useful neighborhood. The direct nonlinear moment certificates in Appendix C.4 and the new validation below avoid this coarse remainder and test the larger scale $\varepsilon = .04$. The parameterized statement is useful for its explicit structural dependence and schedule elimination; the sharper certificates supply its quantitative complement.

B.2 An Interpretable Bias–Noise–Budget Window

This section supplies the information floor and polynomial construction used in Proposition 4.2. A closed-form RL floor avoids optimizing its learning-rate schedule, while an exact polynomial exposes how teacher error and reward noise enter the mixed risk. Nonlinear discrepancy bounds are still required; none of the reference inequalities is silently transferred to actual training.

Setting and notation. Use the scaled two-stage model of Theorem 4.1, with $x_0 > 0$, $y_0 \neq 0$, and $h_1 = b$, $h_2 = -b$, where $0 < b < x_0$. The fixed optimal variances are $h_p = p_\star(1 - p_\star)$ and $h_s = s_\star(1 - s_\star)$. Thus actual initial logit errors are $\varepsilon(x_0, y_0)$, teacher probability offsets are $\varepsilon(h_p b, -h_s b)$, and the independent centered reward noise is $\varepsilon\zeta$, with variance $\varepsilon^2\nu$ and support $[-\varepsilon T, \varepsilon T]$. This is a one-parameter teacher family, not a claim about every teacher with a given error norm. Let $K \ge 1$ be the fixed mixture's update count, with step $\lambda > 0$ and fraction $w \in (0, 1)$. Let $N \ge 1$ bound the allowed pure-RL counts. Define

$$c_1 = h_p, \qquad c_2 = p_\star h_s, \qquad q_1 = c_1 x_0^2, \quad q_2 = c_2 y_0^2. \tag{70}$$

Both q_i are positive. These quantities describe the reference, in which action and visitation coefficients remain random.

Lemma B.1 (A finite-budget information floor and an exact mixed polynomial)**.** *For every $\nu > 0$, every predetermined pure-RL count $k \le N$, and every prescribed nonnegative step sequence, the sampled-coefficient reference obeys*

$$\mathcal{R}_k^{\mathsf{R}} \ge Q_N(\nu) = \frac{1}{2}\left\{\frac{q_1\nu}{\nu + Nq_1} + \frac{q_2\nu}{\nu + Nq_2}\right\}. \tag{71}$$

For the fixed mixture there are coefficients determined by $(p_\star, s_\star, x_0, y_0, K, \lambda, w)$ such that

$$\mathcal{R}_K^{\mathsf{M}}(b, \nu) = A_K + B_K b + C_K b^2 + D_K \nu, \qquad C_K \ge 0, \quad D_K > 0. \tag{72}$$

Four evaluations of the finite moment recursion compute these coefficients exactly in $O(K)$ scalar operations. For $0 < t < (q_1 + q_2)/2$, $Q_N(\nu) \ge t$ is equivalent to $\nu \ge \Phi_N(t)$, where

$$\Phi_N(t) = N\frac{t(q_1 + q_2) - q_1 q_2 + \sqrt{(q_1 q_2)^2 + t^2(q_1 - q_2)^2}}{q_1 + q_2 - 2t}. \tag{73}$$

The floor is weaker than the coupled moment relaxation in Theorem 4.1, but has a direct interpretation: the reward information available in N updates limits how far any prescribed schedule can reduce the reference loss. It is a bound for the specified estimator, not a minimax lower bound over all RL algorithms. The polynomial is an identity, not a fitted response surface.

Proof. We derive the pure floor, the mixed polynomial, and the inverse noise threshold in turn.

Step 1: retain the noise, then relax the other second moments. Let $U_k = \mathbb{E}X_k^2$, $V_k = \mathbb{E}Y_k^2$ for pure RL. Set $w = 0$ in (64). Since $F(h_p) - h_p^2 = h_p(1 - 4h_p) \ge 0$ and

$$p_\star F(h_s) - (p_\star h_s)^2 = p_\star h_s\{1 - (3 + p_\star)h_s\} \ge 0, \tag{74}$$

the diagonal multiplier squares are at least the squares of their means. The cross-coordinate second-moment coefficient is nonnegative and can be dropped for a lower bound. Thus, at any prescribed next step $a \ge 0$,

$$U_{k+1} \ge (1 - ac_1)^2 U_k + a^2\nu c_1, \qquad V_{k+1} \ge (1 - ac_2)^2 V_k + a^2\nu c_2. \tag{75}$$

Dropping these positive terms is a relaxation, not the incorrect assertion that sampled coefficients equal their means. In particular, the resulting formula is not used as an exact risk prediction.

Step 2: complete the square and telescope reciprocal moments. For $z \geq 0$, $c > 0$, and $\nu > 0$,

$$(1 - ac)^2 z + a^2 \nu c = c(cz + \nu) \left(a - \frac{z}{cz + \nu} \right)^2 + \frac{z\nu}{\nu + cz}. \tag{76}$$

The last term is nondecreasing in z; its derivative is $\nu^2/(\nu + cz)^2$. Starting from $z_0 = x_0^2$ or y_0^2, induction in (75) therefore gives

$$U_k \geq \frac{x_0^2 \nu}{\nu + k c_1 x_0^2}, \qquad V_k \geq \frac{y_0^2 \nu}{\nu + k c_2 y_0^2}. \tag{77}$$

Equivalently, the relaxed scalar recursion adds c/ν to the inverse moment at each step. Taking separate minimizing steps only enlarges the allowed behavior, so the bound covers every common step sequence, with or without an upper cap. Weighting by $c_1/2, c_2/2$ and using $k \leq N$ proves (71). The case $k = 0$ is immediate. The proof requires prescribed steps: state-dependent steps cannot be taken outside the expectations in (75).

Step 3: isolate the two inputs to the mixed risk. For a fixed action history, the reference state is an affine function of the teacher amplitude b and the independent reward noises. The terminal quadratic loss therefore has an initialization term, a term linear in b, a term quadratic in b, and a sum of squared noise responses. Centering and independence eliminate noise cross terms, including those with the teacher response. Averaging over action histories proves (72). In particular, C_K is the expected nonnegative quadratic loss of the unit teacher response, so $C_K \geq 0$. The final update alone contributes

$$D_K \geq \frac{\lambda^2 (1 - w)^2}{2} \{ h_p^2 + (p_\star h_s)^2 \} > 0. \tag{78}$$

All remaining independent-noise contributions are nonnegative.

For a constructive calculation write $r(b, \nu) = \mathcal{R}_K^{\mathsf{M}}(b, \nu)$. The following are exact polynomial identities:

$$\begin{aligned} A_K &= r(0,0), & B_K &= \{r(1,0) - r(-1,0)\}/2, \\ C_K &= \{r(1,0) + r(-1,0)\}/2 - r(0,0), & D_K &= r(0,1) - r(0,0). \end{aligned} \tag{79}$$

Each evaluation uses (64). The evaluations at $b = \pm 1$ are algebraic evaluations of the reference polynomial; they do not assert that those teacher values satisfy the actual model's local containment assumptions. This is exact coefficient extraction, not statistical fitting or interpolation of empirical risks.

Step 4: solve the noise inequality. Put $S_q = q_1 + q_2$ and $P_q = q_1 q_2$. Multiplying $Q_N(\nu) \geq t$ by the positive denominator $2(\nu + N q_1)(\nu + N q_2)$ gives

$$(S_q - 2t)\nu^2 + 2N(P_q - t S_q)\nu - 2tN^2 P_q \geq 0. \tag{80}$$

The leading coefficient is positive and the constant term negative, so there is exactly one positive root. Its discriminant simplifies using $S_q^2 - 4P_q = (q_1 - q_2)^2$, yielding (73). This also proves that $\Phi_N(t)$ grows linearly in N for fixed t and geometry. Direct differentiation shows Q_N increases in ν and decreases in N. □

Reference inputs for a nonempty vector window. For the construction in the next subsection, take $p_\star = s_\star = 11/20$, $(x_0, y_0) = (1, 1)$, $\lambda = 3/5$, $w = 1/2$, $h = (b, -b)$, $b \in [43/100, 1/2]$, $\nu \in [13/20, 33/50]$, $N = 25$, and $K \in \{20, \ldots, 25\}$. Exact rational evaluation of (64) gives

$$Q_{25}(13/20) > .02267, \qquad h_p(43/100)^2/2 > .02288, \qquad \max_{K,b} \mathcal{R}_K^{\mathsf{M}}(b, 33/50) < .02045. \tag{81}$$

The mixed maximum uses the two bias endpoints at each of the six horizons: the risk is convex in b because $C_K \geq 0$, and increasing in ν because $D_K > 0$. The pure minimum uses the lowest noise and largest count. These are reference inequalities only. The following proof supplies the nonlinear correction and the resulting strictly positive scale.

B.3 Proof of the Vector Teacher-Error Window

Proposition 4.2 makes the main comparison a condition on model and budget parameters. The scalar teacher family in Appendix B.2 is a special case, not a restriction of the proposition.

Proof of Proposition 4.2. We first derive the vector risk identity, then apply the separate pure bounds and the finite nonlinear correction.

1. The mixed risk is a quadratic function of the teacher vector. Condition on the entire action history in the reference recursion (63). Its coefficients depend on these actions, the fixed step and weight, but not on h or the reward noises. Thus its terminal state can be written

$$Z_K = z_{\mathrm{init}} + T_K h + \sum_{j=0}^{K-1} W_j \zeta_j. \tag{82}$$

Here $z_{\mathrm{init}}, T_K, W_j$ are functions of the action history, and $H = \mathrm{diag}(h_p, p_\star h_s)$ is positive definite. The reference action law is independent of all reward noises. Conditional centering removes every term linear in a noise and independence removes cross terms between noises at different updates. Expanding $\mathbb{E}[Z_K^\top H Z_K]/2$ therefore gives

$$\begin{aligned} A_K &= \tfrac{1}{2}\mathbb{E}[z_{\mathrm{init}}^\top H z_{\mathrm{init}}], & d_K &= \mathbb{E}[T_K^\top H z_{\mathrm{init}}], \\ C_K &= \tfrac{1}{2}\mathbb{E}[T_K^\top H T_K], & D_K &= \tfrac{1}{2}\sum_{j=0}^{K-1} \mathbb{E}[W_j^\top H W_j]. \end{aligned} \tag{83}$$

This proves (11) and $C_K \succeq 0$. The last update's noise contribution alone gives $D_K \geq \lambda^2(1-w)^2\{h_p^2 + (p_\star h_s)^2\}/2 > 0$. These are raw second moments; no independence between the reward and teacher directions on the same rollout has been imposed.

2. Coefficients are recoverable by finite recursion. Write $r(h,\nu) = R_M(h,\nu)$ and let e_1, e_2 be the coordinate vectors. The seven evaluations $r(0,0)$, $r(\pm e_i, 0)$, $r(e_1+e_2,0)$, and $r(0,1)$ give

$$\begin{aligned} A_K &= r(0,0), & (d_K)_i &= [r(e_i,0) - r(-e_i,0)]/2, \\ (C_K)_{ii} &= [r(e_i,0) + r(-e_i,0)]/2 - A_K, & D_K &= r(0,1) - A_K, \end{aligned} \tag{84}$$

The remaining cross coefficient is

$$2(C_K)_{12} = r(e_1+e_2,0) - A_K - (d_K)_1 - (d_K)_2 - (C_K)_{11} - (C_K)_{22}. \tag{85}$$

Each evaluation uses the four scalar recursions (64), so the total work is linear in K. The basis evaluations are algebraic, not assertions that those teacher vectors satisfy the nonlinear containment tests.

3. Compare against the entire prescribed RL class. Lemma B.1 derives $Q_N(\nu)$ by discarding nonnegative multiplier variation and completing a square in each coordinate. Its RL proof contains no teacher parameter. It therefore applies unchanged to an arbitrary h, every prescribed step sequence in the permitted interval, and every count through N. The inverse calculation in (73) shows $Q_N(\nu) \geq \tau + \delta$ exactly when $\nu \geq \Phi_N(\tau+\delta)$, under the displayed range of $\tau + \delta$. This step is not valid for feedback RL, whose rate is correlated with the current random state.

4. Include the teacher floor and nonlinear errors. The first inequality in (12) gives the reference OPD floor $h_p h_1^2/2 \geq \tau + \delta$. The actual floor follows from the pathwise noncrossing argument in Appendix B.1, not from assuming OPD minimizes the reward objective. It covers every affordable OPD history and stopping time. Since $D_K > 0$, the upper noise inequality is equivalent to $R_M(h,\nu) \leq \tau - \delta$. With normalized errors at most $\delta/2$, every compared pure regret is at least $\varepsilon^2(\tau + \delta/2)$, while the mixed regret is at most $\varepsilon^2(\tau - \delta/2)$. Subtraction proves the result. The different counts are exactly the affordable counts in the statement; no teacher query is treated as free. □

A nonempty family with all prescribed RL schedules. The reference inequalities in (81) already verify a continuous family $h = (b, -b)$ and a continuous noise interval with $N = 25$, including $K = 20$. For the full prescribed-schedule class, use the explicit constants of Theorem 4.1. With the same reference parameters, set $(R_1, R_2) = (10, 3)$, $\varepsilon_0 = 10^{-3}$, $a_{\max} = 2$, and bound those constants uniformly over the teacher interval. The rectangle tests (67) hold with strict slack. Let $C_* \geq \max(C_R, C_M, t_D) > 0$ be their finite uniform upper bound. Every explicitly bounded scale

$$0 < \varepsilon \leq \min\{10^{-3}, \delta/(2C_*)\}, \qquad \tau = 43/2000, \quad \delta = 1/1000, \tag{86}$$

satisfies the required nonlinear error allowance. This yields a nonempty finite-scale family against all prescribed schedules and adaptive OPD, including equal spending with $K = 20$ and $N = 25$. The radius cap 10^{-3} alone does not certify this family: the error allowance may require a much smaller scale.

Parameter interpretation and limits. The teacher floor depends on root mismatch, but mixed risk depends on the whole teacher vector through d_K and C_K. Therefore a bound on teacher error norm alone would lose relevant directional information. The noise lower threshold is linear in the affordable pure count N; the upper threshold depends on the mixed count and weight through F_K, D_K. Changing costs can change both counts discontinuously. The support constraint and rectangle tests must also continue to hold. These facts explain why the window is sufficient rather than a universal phase diagram, and why its failure does not establish a reversal.

C Query Costs and a Sharper Nonlinear Certificate

We first solve the separate scalar allocation problem. The nonlinear comparison then permits extra pure updates and all prescribed RL rates. Its direct moment bounds give a sharper finite-radius certificate than the general window. The final proof changes only reward noise to reverse the ordering; it does not claim an adaptive-RL guarantee.

C.1 When Saved Query Costs Buy More Samples

The scalar comparison lets every method optimize deterministic step schedules and batch sizes under one observation allowance. It isolates the value of a paired signal after accounting for the extra samples that a pure method can afford. The following nonlinear certificate addresses a separate model and retains its own comparison class.

A declared cost model. Let $c_0 \geq 0$ be a rollout cost and $c_R, c_D > 0$ the incremental costs of the reward and teacher channels on that rollout. The per-observation costs are

$$a_R = c_0 + c_R, \qquad a_D = c_0 + c_D, \qquad a_M = c_0 + c_R + c_D. \tag{87}$$

With a common finite budget B, each pure method can buy additional observations instead of an unused channel:

$$N_R = \lfloor B/a_R \rfloor, \qquad N_D = \lfloor B/a_D \rfloor, \qquad N_M = \lfloor B/a_M \rfloor. \tag{88}$$

These are stipulated costs, not measured GPU times. The scalar calculation charges each observation but no separate batch-update overhead. All model parameters are known when the oracle selects a recipe. It concerns the three stated acquisition modes, not policies that alternate arbitrarily between paired and unpaired purchases or adapt purchases to observations.

Scalar recipes with arbitrary deterministic batching. Use equal-curvature scalar fields, initial error $x_0 \neq 0$, and reward regret $H\mathbb{E}x_K^2/2$, with $H > 0$. Each fresh pair (z, e) is centered, independent of earlier pairs, and has positive-definite covariance

$$\Sigma = \begin{pmatrix} v_R & C \\ C & v_D \end{pmatrix} \succ 0. \tag{89}$$

The teacher additionally has a fixed bias h. For any positive integer batch partition $n_0 + \cdots + n_{K-1} = N$, the paired update is

$$x_{k+1} = (1 - \lambda_k)x_k + \lambda_k\{(1 - w_k)\bar{z}_k + w_k(h + \bar{e}_k)\}, \qquad \lambda_k, w_k \in [0, 1]. \tag{90}$$

Steps, fractions, batch sizes, and the finite update count are chosen before sampling. Pure RL and OPD set every fraction to zero and one, respectively. Every method may use any positive batch partition compatible with a common update cap; a single batch is allowed. At zero observations it can only return the initial point. The proposition below assumes all three counts in (88) are positive.

Proposition C.1 (Exact comparison after cost reallocation and schedule tuning)**.** *For the scalar model above, the optimal paired risk at N observations is*

$$R_M^\star(N,h) = \frac{H}{2} \min_{\substack{r,d\geq 0\\ r+d\leq 1}} \left[\{(1-r-d)x_0 + dh\}^2 + \frac{v_R r^2 + 2Crd + v_D d^2}{N} \right]. \tag{91}$$

The pure optima $R_R^\star(N,h)$ and $R_D^\star(N,h)$ restrict this minimum to $d=0$ and $r=0$, respectively. These optima are attainable for every *positive batch partition of N, with suitable deterministic steps. Thus the exact equal-cost comparison evaluates the three optima at N_R, N_D, N_M, not at a shared observation count.*

For $h=0$, define

$$\begin{aligned} w_\star &= \mathrm{clip}\left(\frac{v_R - C}{v_R + v_D - 2C}; [0,1]\right), \\ v_\star &= (1-w_\star)^2 v_R + 2w_\star(1-w_\star)C + w_\star^2 v_D, \\ I_R &= N_R/v_R, \qquad I_D = N_D/v_D, \qquad I_M = N_M/v_\star. \end{aligned} \tag{92}$$

The optimized risks and exact strict-synergy criterion are

$$R_a^\star = \frac{H}{2(x_0^{-2} + I_a)}, \qquad R_M^\star < \min\{R_R^\star, R_D^\star\} \iff I_M > \max\{I_R, I_D\}. \tag{93}$$

If its unbiased risk gap is $\delta_B > 0$, the unbiased-optimal mixed recipe, kept fixed, still beats both pure methods reoptimized for teacher bias h whenever

$$2E(h) < \delta_B, \qquad E(h) = H\left(\frac{|x_0 h|}{4} + \frac{h^2}{2}\right). \tag{94}$$

For arbitrary h, (91) *gives the exact comparison;* (94) *is only a sufficient robustness condition.*

In plain language, each regime can slow down, change its step at every update, and buy as many observations as its own query costs permit. With an unbiased teacher, the deciding quantity is total noise precision purchased, not variance per paired rollout alone. For a biased teacher, the same comparison is a two-variable convex quadratic problem. The averaging and convex-optimization tools are standard; the purpose is to make the resource qualification in the three-way comparison exact.

Proof. **Step 1: unroll every scheduled recipe.** Set

$$a_0 = \prod_{j=0}^{K-1}(1-\lambda_j), \qquad b_k = \lambda_k \prod_{j=k+1}^{K-1}(1-\lambda_j), \qquad v_k = b_k(1-w_k, w_k)^\top. \tag{95}$$

All these weights are nonnegative, and telescoping the products gives $a_0 + \sum_k b_k = 1$. Write $(r,d)^\top = \sum_k v_k$. Finite unrolling of (90) yields

$$x_K = a_0 x_0 + dh + \sum_k b_k\{(1-w_k)\bar{z}_k + w_k \bar{e}_k\}. \tag{96}$$

Thus $a_0 = 1-r-d$, and independence across batches gives the exact mean and variance

$$\mathbb{E}x_K = (1-r-d)x_0 + dh, \qquad \mathrm{Var}(x_K) = \sum_k v_k^\top \Sigma v_k / n_k. \tag{97}$$

The fractions may differ at every update; no constant-mixture assumption has been used.

Step 2: aggregate the variance without losing a possible optimum. Let $v = (r, d)^\top$ and $N = \sum_k n_k$. Expanding squares gives

$$\sum_k \frac{v_k^\top \Sigma v_k}{n_k} - \frac{v^\top \Sigma v}{N} = \sum_k n_k \left(\frac{v_k}{n_k} - \frac{v}{N}\right)^\top \Sigma \left(\frac{v_k}{n_k} - \frac{v}{N}\right) \geq 0. \tag{98}$$

Indeed, the cross term is $-2v^\top \Sigma \sum_k v_k/N$, and the final term is $v^\top \Sigma v/N$, so the equality is exact. Equality holds when $v_k = n_k v/N$. Consequently every scheduled recipe has at least the risk in (91) at its aggregate weights.

Step 3: construct steps attaining every aggregate pair. Fix any $r, d \geq 0$ with $r + d \leq 1$ and any positive batch partition. If $r + d = 0$, choose all steps zero. Otherwise put

$$w_k = \frac{d}{r+d}, \qquad \widetilde{b}_k = \frac{n_k}{N}(r+d), \qquad S_k = 1 - r - d + \sum_{j=0}^{k} \widetilde{b}_j, \qquad \lambda_k = \widetilde{b}_k / S_k. \tag{99}$$

Each denominator is positive, every step lies in $[0, 1]$, and $S_{K-1} = 1$. For $k \geq 1$, $1 - \lambda_k = S_{k-1}/S_k$; the same identity at $k = 0$ uses $S_{-1} = 1 - r - d$, possibly zero. Therefore $\prod_{j>k}(1 - \lambda_j) = S_k$ and the unrolled weight is $b_k = \widetilde{b}_k$. Hence $v_k = n_k(r, d)^\top/N$, attaining equality in (98). This proves both directions of the optimization claim, including the boundary $r + d = 1$. It also proves attainability under any positive finite update cap because one may place all observations in a single batch.

Using additional observations cannot increase the optimum: hold the aggregate weights fixed and increase N in (91). Thus buying all affordable observations gives the pure and paired optima under (88); a method may still choose zero steps.

Step 4: solve the unbiased case. For $h = 0$, put $t = r + d$ and, when $t > 0$, $w = d/t$. The expression in brackets is

$$(1 - t)^2 x_0^2 + \frac{t^2}{N}\{(1 - w)^2 v_R + 2w(1 - w)C + w^2 v_D\}. \tag{100}$$

The variance quadratic is strictly convex since $v_R + v_D - 2C > 0$, so completing the square gives $w_\star$ in (92). With its minimum $v_\star > 0$, differentiation in t gives

$$t_\star = \frac{x_0^2}{x_0^2 + v_\star/N} \in (0, 1), \qquad \min_t\{(1 - t)^2 x_0^2 + t^2 v_\star/N\} = \frac{1}{x_0^{-2} + N/v_\star}. \tag{101}$$

The pure calculation replaces $v_\star$ by v_R or v_D and uses its own count. The risk is strictly decreasing in precision, proving (93). Because $N_M \leq N_R, N_D$, a strict equal-cost gain cannot arise from a boundary fraction that discards a purchased channel.

Step 5: control bias uniformly before retuning. For any scheduled recipe, its aggregate teacher mass is $d \in [0, 1 - a_0]$. Changing h from zero changes its risk by

$$\frac{H}{2}\{2a_0 x_0 dh + d^2 h^2\}. \tag{102}$$

Since $a_0 d \leq 1/4$ and $d^2 \leq 1$, its absolute value is at most $E(h)$, uniformly over steps, fractions, batches, and observation counts. Taking an infimum preserves this uniform error bound. The fixed unbiased-optimal mixture therefore has biased risk at most $R_M^\star(0) + E(h)$, whereas either reoptimized pure method has risk at least $R_a^\star(0) - E(h)$. Subtraction proves a gap of at least $\delta_B - 2E(h) > 0$, as claimed. □

Explicit solution with an arbitrary biased teacher. The two-dimensional minimization need not be delegated to a numerical optimizer. Put $s = (x_0, x_0 - h)^\top$ and $Q = ss^\top + \Sigma/N \succ 0$. The unscaled risk is $x_0^2 - 2x_0 s^\top (r, d)^\top + (r, d)Q(r, d)^\top$. Its unconstrained minimizer is $Q^{-1}x_0 s$. If this point lies in the triangle it is the minimizer; otherwise a minimum lies on one of the three edges.

Their respective minimizing candidates are

$$\begin{aligned} d = 0: \quad & r = \frac{x_0^2}{x_0^2 + v_R/N}, \\ r = 0: \quad & d = \mathrm{clip}\left(\frac{x_0(x_0 - h)}{(x_0 - h)^2 + v_D/N}; [0, 1]\right), \\ r + d = 1: \quad & d = \mathrm{clip}\left(\frac{v_R - C}{Nh^2 + v_R + v_D - 2C}; [0, 1]\right), \quad r = 1 - d. \end{aligned} \tag{103}$$

Each follows by differentiating a strictly convex one-variable quadratic and projecting onto its edge. Evaluate the objective at these candidates and any feasible interior point. Compactness ensures a minimum exists; an interior minimum has zero gradient, and a boundary minimum must minimize its edge, so this finite list is complete. If a vertex is globally minimizing, it is also an edge minimum and is included. These candidates can be evaluated with exact rational arithmetic. Checking their first-order inequalities against all triangle vertices suffices to certify global optimality of this convex quadratic.

What shared rollout cost and correlation change. If $B/a_R, B/a_D, B/a_M$ are integers, the exact unbiased criterion reduces to

$$a_M v_\star < \min\{a_R v_R, a_D v_D\}. \tag{104}$$

Without rollout overhead ($c_0 = 0$) and with $C \geq 0$,

$$v_\star \geq \frac{v_R v_D}{v_R + v_D}, \qquad (c_R + c_D)\frac{v_R v_D}{v_R + v_D} \geq \min\{c_R v_R, c_D v_D\}. \tag{105}$$

The first inequality follows by dropping the nonnegative covariance term before minimizing; the second is a convex combination of $c_R v_R$ and $c_D v_D$, with weights $v_D/(v_R + v_D)$ and $v_R/(v_R + v_D)$. Thus there is no strict equal-cost mixed gain in this case, even though mixing independent signals strictly improves variance at a shared paired sample count.

For balanced channels $c_R = c_D = c$ and $v_R = v_D = v$, write $C = \rho v$, where $-1 < \rho < 1$. Then $w_\star = 1/2$ and $v_\star = v(1 + \rho)/2$. Under the same integer-divisibility condition, (104) becomes the exact boundary

$$\boxed{\text{strict equal-cost mixed gain} \quad \Longleftrightarrow \quad \rho < \frac{c_0}{c_0 + 2c}.} \tag{106}$$

At equality the optimized risks tie; above it either identical pure channel is better. Negative correlation can pay for the second channel without rollout savings. Shared rollout costs can make mixing useful even with positive correlation. These are exact finite-budget statements in this scalar model, not claims that all OPD pipelines have balanced channels.

c_0	ρ	$N_R = N_D$	N_M	Pure risk	Mixed risk	Comparison
0	0	12	6	$1/26$	$1/26$	Tie
0	$1/2$	12	6	$1/26$	$1/18$	Pure
0	$-1/2$	12	6	$1/26$	$1/50$	Mixed
1	0	6	4	$1/14$	$1/18$	Mixed

Table 3: Exact cost-aware optima with $B = 12$, $c_R = c_D = v_R = v_D = H = x_0 = 1$, $h = 0$, and arbitrary deterministic step schedules and positive batch partitions. The first row's mixed gain at a matched count disappears when pure methods spend their savings on twice as many observations.

Integer costs matter at small budgets: without divisibility one must use (93), not (104). For example, $c_0 = 0$, $(c_R, c_D) = (2, 3)$, $(v_R, v_D) = (3/2, 1)$, $C = h = 0$, and $B = 5$ give counts $(2, 1, 1)$. The paired precision $5/3$ exceeds both pure precisions $4/3$ and 1, solely because the mixed purchase uses budget that either pure method leaves unspent. No large-budget approximation is needed to account for this effect.

Why the optimized scalar risk does not depend on batch partition. The affine oracle has exogenous independent observations and unrestricted deterministic steps in $[0, 1]$. Equation (99) can realize the same aggregate estimate through one batch or many updates. This collapse is a property of the idealized model, not a claim that update count is irrelevant for nonlinear learning. Bias can also remove the advantage; the explicit biased problem, rather than a variance-only rule, must then decide. For instance, $B = 100$, $c_0 = 3$, $c_R = c_D = v_R = v_D = H = x_0 = 1$ and $C = 0$ give unbiased gap $1/52 - 1/82$. Both biases $h = \pm.01$ satisfy (94), so a strict cost-aware advantage survives even after the pure methods retune for the biased teacher.

C.2 Equal-Cost Dominance and Its Reversal

Does such an advantage actually survive the nonlinear correction? The next proposition answers yes in an explicit region. It then changes only reward noise and obtains the opposite winner. The proof follows the same comparison strategy but uses sharper direct nonlinear moment bounds, not an automatic substitution into the theorem's radius estimate.

A certified numerical instance. Use (6) with actual teacher offsets $q_A = p_\star + h_p\widetilde{h}_1$ and $q_B = s_\star + h_s\widetilde{h}_2$. Fix

$$\begin{aligned} p_\star = s_\star = \mathrm{sigmoid}(.1), \quad \theta_0 = (.04, .04), \quad S = .032, \\ \widetilde{h}_1 \in [.019, .021], \qquad \widetilde{h}_2 \in [-.021, -.019]. \end{aligned} \tag{107}$$

At every update, draw fresh centered reward noise η from any fixed law supported on $[-S, S]$, independently of actions and history; write $\nu = \mathbb{E}\eta^2$. Use exact statewise baselines and exact teacher targets. The fixed mixture M takes 20 updates with $\lambda = .6, w = .5$.

Proposition C.2 (Equal-cost dominance and a noise-dependent reversal)**.** *Under* (107)*, give every method allowance $\mathcal{B} = 100$ and costs $(c_0, c_R, c_D) = (3, 1, 1)$. Let $\mathcal{A}_R$ contain all prescribed RL schedules of any predetermined length from zero to 25 with steps in $[0, 2]$. Let $\mathcal{A}_D$ contain all OPD rules with history-dependent steps in $[0, 4]$ and current-iterate stopping by update 25. Write $\mathfrak{R}(a)$ for a rule's expected returned-policy regret. Uniformly over the teacher rectangle and the indicated noise laws:*

1. *If $9S^2/10 \le \nu \le S^2$, then*
$$\inf_{a \in \mathcal{A}_R \cup \mathcal{A}_D} \mathfrak{R}(a) - \mathfrak{R}^{\mathsf{M}}_{20} > 5 \cdot 10^{-8}. \tag{108}$$
2. *If $0 \le \nu \le 2S^2/5$, the prescribed RL schedule $\alpha_k = 6/(k + 3)$, $k = 0, \dots, 24$, denoted R_0, satisfies*
$$\min\{\mathfrak{R}^{\mathsf{M}}_{20}, \inf_{a \in \mathcal{A}_D} \mathfrak{R}(a)\} - \mathfrak{R}^{\mathsf{R}_0}_{25} > 10^{-6}. \tag{109}$$

What the proposition establishes. **Equal cost:** a pure update costs four units and a mixed update costs five. The pure methods can therefore buy 25 updates while the mixture buys 20. Its high-noise advantage exceeds $5 \cdot 10^{-8}$ even after that extra training and independent pure-method tuning. This small certified gap establishes strict dominance, not a large practical gain. With less noisy rewards, RL instead beats the same mixture and OPD. Nothing else changes: teacher, initialization, costs, and mixed recipe are held fixed. The result does not rule out a better retuned mixture or locate the transition between the two noise regions. Appendix C gives the nonlinear bounds proving both claims.

C.3 The Teacher Floor and Affordable Comparison Classes

Use the numerical inputs of (107). Let $a_R = c_0 + c_R$, $a_D = c_0 + c_D$, and $a_M = c_0 + c_R + c_D$ be per-update allowances. The fixed 20-update mixture and every affordable prescribed RL count through 25 are covered whenever

$$20a_M \le B < 26a_R. \tag{110}$$

The first inequality makes the mixture affordable; the second gives $\lfloor B/a_R \rfloor \le 25$. OPD needs no count restriction beyond affordability: for steps in $[0, 4]$, its root error has the pathwise floor

$$\mathcal{L}(\theta^{\mathsf{D}}_k) \ge \mathrm{KL}(q_- \| p_\star) > 45.027 \cdot 10^{-6}, \qquad q_- = p_\star + h_p \cdot .019. \tag{111}$$

In particular, allowing more teacher updates, history-dependent steps, or stopping cannot cross this floor.

Proof of the pathwise teacher floor. Write $z_k = .1 + x_k$ for the root logit and $z_q = \log(q_A/(1-q_A))$ for its teacher target. The root is always visited, so its implemented OPD update is exactly

$$z_{k+1} = T_{\alpha_k}(z_k), \qquad T_\alpha(z) = z + \alpha\{q_A - \mathrm{sigmoid}(z)\}. \tag{112}$$

The teacher rectangle ensures $p_\star < q_- \le q_A \le q_+ = p_\star + p_\star(1-p_\star)\cdot .021 < \mathrm{sigmoid}(.14)$, hence $.1 < z_q < z_0 = .14$. For $0 \le \alpha \le 4$, the global logistic derivative bound gives

$$T'_\alpha(z) = 1 - \alpha\,\mathrm{sigmoid}(z)\{1 - \mathrm{sigmoid}(z)\} \ge 1 - \alpha/4 \ge 0. \tag{113}$$

Also $T_\alpha(z_q) = z_q$ and $T_\alpha(z) \le z$ for $z \ge z_q$. Therefore $z_q \le T_\alpha(z) \le z$ whenever $z \ge z_q$. Induction proves $z_q \le z_k \le z_0$ for every finite k, for every realized sequence of allowed steps. The argument is pathwise: steps may depend on all earlier leaf visits or external randomization, without an independence assumption. It also includes zero steps and zero updates.

Consequently the root probability obeys $p_k \ge q_A \ge q_- > p_\star$. The derivative of $\mathrm{KL}(p\|p_\star)$ with respect to p is $\log(p/(1-p)) - \log(p_\star/(1-p_\star))$, positive for $p > p_\star$. The leaf term of the exact loss (5) is nonnegative. Dropping it gives

$$\mathcal{L}(\theta_k^{\mathsf{D}}) \ge \mathrm{KL}(p_k\|p_\star) \ge \mathrm{KL}(q_-\|p_\star). \tag{114}$$

This holds simultaneously at all finite checkpoints, so it survives any finite stopping rule or checkpoint selection, and taking expectations. No optional-stopping theorem or final-iterate convergence claim is needed.

Outward-rounded evaluation at 256 bits gives $45.02730023 \cdot 10^{-6} < \mathrm{KL}(q_-\|p_\star) < 45.02730024 \cdot 10^{-6}$ and $\mathrm{sigmoid}(.14) - q_+ > .00472686$. These checks establish the stated numerical floor and initial ordering. □

For equal channel costs $c_R = c_D = c$ and $B = 20(c_0 + 2c)$, (110) is equivalent to $c_0 > 7c/3$. This is a sufficient cost region, not a sharp boundary: equality allows 26 RL updates, which the certificate below does not cover. The following moment argument bounds every prescribed RL schedule in the affordable class.

C.4 Removing the Constant-Step Restriction on RL

We cover every prescribed sequence of steps in $[0, 2]$, without enumerating a 25-dimensional schedule grid. The construction bounds actual nonlinear moments directly. Its lower bound is a relaxation of the learning process, not a claim to have found its optimal schedule.

Proposition C.3 (Nonlinear comparison against all prescribed RL steps)**.** *Throughout the numerical input region* (107)*, with $9S^2/10 \le \nu \le S^2$, the fixed 20-update mixture obeys*

$$\mathfrak{R}_{20}^{\mathsf{M}} + 5 \cdot 10^{-8} < \inf_{\substack{0 \le k \le 25 \\ (\alpha_0,\ldots,\alpha_{k-1}) \in [0,2]^k}} \mathfrak{R}_k^{\mathsf{R}}(\alpha_0, \ldots, \alpha_{k-1}). \tag{115}$$

The empty schedule returns the initial point. Counts and steps are chosen before training; they may depend on the known model and noise law. Together with (111)*, the same gap holds against both pure methods in the cost region* (110)*, with arbitrary prescribed RL steps and adaptive OPD steps/stopping as stated there.*

In plain language, a decreasing or otherwise varying RL learning rate cannot remove this particular equal-cost advantage. The bound is conservative, and its positive margin establishes strict dominance rather than a large quantitative gain. It is not a sharp estimate of the advantage. Adaptive RL steps, adaptive RL stopping, different estimators, and minibatch allocation remain outside the claim.

Proof. We give the moment induction, its continuous-step verification, the mixed upper bound, and the complete uncertainty partition separately. Every inequality below is finite; no limit in horizon, radius, or noise is used.

Step 1: exact nonlinear conditional moments. Write $p = \mathrm{sigmoid}(.1 + x)$, $s = \mathrm{sigmoid}(.1 + y)$, $p_0 = s_0 = \mathrm{sigmoid}(.1)$, $h_p = p_0(1 - p_0)$, $h_s = s_0(1 - s_0)$, and $D_s = \mathrm{KL}(s\|s_0)$. Put $a = A - p$, $b = B - s$, where $A \sim \mathrm{Ber}(p)$ and the leaf is sampled conditional on $A = 1$. The independent centered noise has variance ν and support $[-S, S]$. The exact RL update obtained from the regularized utilities and their statewise baselines is

$$\begin{aligned} x' &= x - \alpha\{(x + D_s)a^2 + Ayab\} + \alpha\eta a, \\ y' &= (1 - \alpha Ab^2)y + \alpha A\eta b. \end{aligned} \tag{116}$$

Let $v_p = p(1 - p)$, $v_s = s(1 - s)$, and $F(v) = v(1 - 3v)$. For a centered Bernoulli variable of variance v, its fourth, sixth, and eighth moments are respectively

$$F(v), \qquad G(v) = v(1 - 5v + 5v^2), \qquad H(v) = v(1 - 7v + 14v^2 - 7v^3). \tag{117}$$

Each identity follows by summing the two values $(1 - p)(-p)^j + p(1 - p)^j$ and substituting $v = p(1 - p)$. Expansion of (116) therefore gives

$$\begin{aligned} \mathbb{E}[x'^2 \mid x, y] &= (1 - 2\alpha v_p + \alpha^2 F(v_p))x^2 + 2\alpha(\alpha F(v_p) - v_p)xD_s \\ &\quad + \alpha^2 F(v_p)D_s^2 + \alpha^2 p(1 - p)^2 v_s y^2 + \alpha^2 \nu v_p, \\ \mathbb{E}[y'^2 \mid x, y] &= \{1 + p[-2\alpha v_s + \alpha^2 F(v_s)]\}y^2 + \alpha^2 \nu p v_s. \end{aligned} \tag{118}$$

For the first identity, terms containing one factor b vanish after conditioning on A; terms linear in η vanish by independence and centering. Also $\mathbb{E}[Aa^2] = p(1 - p)^2$. For the second, use $A^2 = A$. Thus these are identities for the actual changing action probabilities, not identities for a frozen reference law.

For later use define

$$\begin{aligned} A_4(\alpha, v) &= 1 - 4\alpha v + 6\alpha^2 F(v) - 4\alpha^3 G(v) + \alpha^4 H(v), \\ B_2(\alpha, v) &= v - 2\alpha F(v) + \alpha^2 G(v). \end{aligned} \tag{119}$$

They equal $\mathbb{E}_b(1 - \alpha b^2)^4$ and $\mathbb{E}_b[(1 - \alpha b^2)^2 b^2]$. Writing $\mu_3 = \mathbb{E}\eta^3$, $\mu_4 = \mathbb{E}\eta^4$, the fourth-moment identity is

$$\begin{aligned} \mathbb{E}[y'^4 \mid x, y] &= \{1 - p + pA_4(\alpha, v_s)\}y^4 + 6\alpha^2 \nu p B_2(\alpha, v_s)y^2 \\ &\quad + 4\alpha^3 \mu_3 p v_s(1 - 2s)\{1 - \alpha(1 - 2v_s)\}y + \alpha^4 \mu_4 p F(v_s). \end{aligned} \tag{120}$$

Indeed, $\mathbb{E}b^3 = v_s(1 - 2s)$ and $\mathbb{E}b^5 = v_s(1 - 2s)(1 - 2v_s)$. The coefficient six and coefficient four are the binomial coefficients in the fourth power of the leaf update. The first-order noise term vanishes; the third-order term must be retained.

Step 2: moment bounds and noise injection. At each update maintain deterministic bounds, uniform over all preceding prescribed steps,

$$\ell_x \le M_x := \mathbb{E}x^2 \le u_x, \quad \ell_y \le M_y := \mathbb{E}y^2 \le u_y, \quad \mathbb{E}x \ge m_-, \quad \mathbb{E}y^4 \le \kappa M_y^2. \tag{121}$$

Initially the moments and mean are exact and $\kappa = 1$. Reachable boxes come from the already proved sample-map containment construction, with the step interval $[0, 2]$ used independently at every update. Its monotonicity tests apply at each step and do not require steps to be equal. Include zero in each coordinate interval when bounding derivatives. Let R_x, R_y bound absolute coordinates, let V_p, L_p bound $|p'|, |p''|$ on the root segment, and let V_s, L_s be the corresponding leaf bounds. Denote actual probability endpoints by p_-, p_+, s_-, s_+ and actual variance extrema by $v_{p,-}, v_{p,+}, v_{s,-}, v_{s,+}$. In the present calculation these variances and h_p, h_s exceed $1/6$. Hence F is positive and decreasing on every variance interval used below.

Taylor's finite inequality and Cauchy–Schwarz give

$$\begin{aligned} \mathbb{E}v_p &\ge n_1 := \max\{v_{p,-}, h_p - L_p\sqrt{u_x}\}, \\ \mathbb{E}p &\ge p_0 + h_p m_- - L_p u_x/2 =: e_p, \\ \mathbb{E}[pv_s] &\ge n_2 := \max\{p_- v_{s,-}, h_s e_p - p_+ L_s\sqrt{u_y}\}. \end{aligned} \tag{122}$$

For example, $|v_s - h_s| \le L_s|y|$ gives $pv_s \ge h_s p - p_+ L_s|y|$; averaging proves the last line. The calculation verifies $n_1, n_2 > 0$ before division by either quantity. The fourth-moment ratio provides the useful cross-moment bound

$$\mathbb{E}[|x|y^2] \le \sqrt{\mathbb{E}x^2\,\mathbb{E}y^4} \le \sqrt{\kappa M_x}\, M_y. \tag{123}$$

No independence of the two state coordinates is assumed.

Step 3: lower bounds for both second moments. Put $c = v_{p,+}$, $f_p = F(v_{p,+})$, and $B_- = \min p(1-p)^2$ on the root probability interval. For $0 \le \alpha \le 2$, $v_p - \alpha F(v_p) \ge 0$. Also $0 \le D_s \le V_s y^2/2$. Drop the nonnegative D_s^2 term in (118), and bound its cross term in absolute value using (123). This gives

$$\begin{aligned} M_x' \ge (1 - 2c\alpha + f_p\alpha^2)M_x + \alpha^2 B_- v_{s,-} M_y \\ - \alpha(c - \alpha f_p)V_s\sqrt{\kappa M_x}\, M_y + \alpha^2 \nu_- n_1, \end{aligned} \tag{124}$$

where $[\nu_-, \nu_+]$ is the current noise-variance slice. The right side is nondecreasing in M_x throughout its rectangle if

$$1 - 4c + 4f_p - \frac{V_s\sqrt{\kappa}\, u_y}{2\sqrt{\ell_x}} \max_{0\le\alpha\le 2} \alpha(c - \alpha f_p) \ge 0. \tag{125}$$

To see sufficiency, its derivative in M_x subtracts $\alpha(c - \alpha f_p)V_s\sqrt{\kappa}M_y/(2\sqrt{M_x})$. The quadratic coefficient of M_x decreases up to step two because $c - 2f_p = c(6c - 1) > 0$; its minimum is $1 - 4c + 4f_p$. The other factors are bounded by u_y and ℓ_x as displayed. This condition is checked at every induction step. Substitute $M_x = \ell_x$. The remaining expression is affine in M_y, so its minimum occurs at $z \in \{\ell_y, u_y\}$. For each endpoint it is the quadratic

$$\ell_x - \alpha\{2c\ell_x + cV_s\sqrt{\kappa\ell_x}\, z\} + \alpha^2\{f_p\ell_x + (B_- v_{s,-} + f_p V_s\sqrt{\kappa\ell_x})z + \nu_- n_1\}. \tag{126}$$

Its positive quadratic coefficient is verified. Evaluate its vertex clipped to $[0, 2]$, then take the smaller value across the two endpoints. This produces the new ℓ_x for every possible prescribed next step.

For the leaf, set

$$\begin{aligned} h &= p_0 h_s + h_s V_p\sqrt{\kappa u_x} + p_+ L_s R_y, \\ f &= p_0 F(h_s) + V_p F(h_s)\sqrt{\kappa u_x} - p_+ L_s R_y(6V_s - 1). \end{aligned} \tag{127}$$

Then

$$M_y' \ge (1 - 2h\alpha + f\alpha^2)M_y + \alpha^2\nu_- n_2. \tag{128}$$

For detail, write $q_\alpha(v) = -2\alpha v + \alpha^2 F(v)$ and split $pq_\alpha(v_s)$ into $p_0 q_\alpha(h_s)$, $(p - p_0)q_\alpha(h_s)$, and $p\{q_\alpha(v_s) - q_\alpha(h_s)\}$. The middle term, after multiplication by y^2 and averaging, is at least $-V_p\{2\alpha h_s - \alpha^2 F(h_s)\}\sqrt{\kappa u_x}M_y$. The last term is at least $-p_+ L_s\{2\alpha + \alpha^2(6V_s - 1)\}R_y M_y$: on the segment, $|q_\alpha'(v)| \le 2\alpha + \alpha^2(6V_s - 1)$. Adding yields exactly (127). Verify that $f > 0$ and $1 - 2h\alpha + f\alpha^2 \ge 0$ on $[0, 2]$. Substitute $M_y = \ell_y$ and minimize the resulting positive quadratic at its clipped vertex. This gives the new ℓ_y. The two coordinate minima need not choose the same step. That enlarges the set of possible moment pairs, so it can only lower the bound; it does not exclude any actual common-step schedule.

Step 4: upper moments and the root mean. Write $F_p^+ = F(v_{p,-})$, $F_p^- = F(v_{p,+})$, $F_s^+ = F(v_{s,-})$, and $B_+ = \max p(1-p)^2$. For a fixed current state, each next second moment is a convex quadratic in α. Its maximum on $[0, 2]$ is therefore at an endpoint. With $a = 2$, valid bounds at the upper endpoint are

$$\begin{aligned} U_x &= (1 - 2av_{p,-} + a^2 F_p^+)u_x + a^2 B_+ v_{s,+} u_y \\ &\quad + a(v_{p,+} - aF_p^-)V_s R_y\sqrt{u_x u_y} + a^2 F_p^+ V_s^2 R_y^2 u_y/4 + a^2\nu_+ v_{p,+}, \\ U_y &= \{1 + p_-(-2av_{s,-} + a^2 F_s^+)\}u_y + a^2\nu_+ p_+ v_{s,+}. \end{aligned} \tag{129}$$

The cross term uses $\mathbb{E}|x|y^2 \le R_y\sqrt{u_x u_y}$; $D_s^2 \le V_s^2 R_y^2 y^2/4$ bounds its square term. For the leaf, $-2av_{s,-} + a^2 F_s^+ = -12v_{s,-}^2 \le 0$, which justifies multiplying by p_- for an upper bound. The coefficients of u_x, u_y in the diagonal terms are nonnegative upper bounds on squared sample multipliers. Set $u_x' = \max(u_x, U_x)$ and $u_y' = \max(u_y, U_y)$, thereby also covering the zero-step endpoints.

The exact root mean update is $\mathbb{E}x' = \mathbb{E}x - \alpha\mathbb{E}[v_p(x + D_s)]$. Since $|(v_p - h_p)x| \le L_p x^2$ and $v_p D_s \le v_{p,+}V_s y^2/2$, put $\beta = L_p u_x + v_{p,+}V_s u_y/2$. Because $1 - \alpha h_p \ge 0$,

$$\mathbb{E}x' \ge (1 - \alpha h_p)m_- - \alpha\beta \ge \min\{m_-, (1 - 2h_p)m_- - 2\beta\} =: m_-'. \tag{130}$$

This is one place where prescribed, rather than state-dependent, steps are essential: the same deterministic α is taken outside expectations.

Step 5: propagating a finite fourth-moment ratio. The elementary bounds on centered bounded noise are

$$\mu_4 \le S^2\nu, \qquad |\mu_3| \le S\nu - \nu^2/S \le T_3 := \max_{v\in[\nu_-,\nu_+]}(Sv - v^2/S). \tag{131}$$

The first follows pointwise from $\eta^4 \le S^2\eta^2$. For the second, expand the nonnegative random variable $(S-\eta)(\eta+\nu/S)^2$ and take expectations, obtaining $-\mu_3 + S\nu - \nu^2/S \ge 0$; reflect η for the other side. The maximum occurs at $\nu = S^2/2$ clipped to the variance slice. Thus no symmetry of the noise law is assumed.

In addition to (128), the actual leaf identity gives $M_y' \ge a_o(\alpha)M_y + \alpha^2\nu n_2$, where

$$a_o(\alpha) = 1 - 2\alpha p_+ v_{s,+} + \alpha^2 p_+ F(v_{s,+}), \qquad a_2(\alpha) = \max\{a_o(\alpha), 1 - 2h\alpha + f\alpha^2\}. \tag{132}$$

Indeed $q_\alpha(v)$ is nonpositive and decreasing in v on this region. Consequently $M_y' \ge a_2 M_y + \alpha^2\nu n_2$. On every step interval let $a_2^- \le a_2 \le a_2^+$ with $a_2^- > 0$. Obtain these bounds by checking the endpoints and clipped vertices of both convex quadratics. Bounds for the maximum follow by taking the maximum of their lower bounds and the maximum of their upper bounds.

Let $a_4^+, b_2^+, c_3^+, d_4^+$ bound, respectively,

$$1 - p + pA_4(\alpha, v_s), \quad pB_2(\alpha, v_s), \quad |pv_s(1-2s)\{1-\alpha(1-2v_s)\}|, \quad pF(v_s). \tag{133}$$

The verified condition $2\max(s_+, 1-s_-)^2 \le 1$ makes $1 - \alpha b^2$ nonnegative. Therefore A_4 and B_2 decrease with α; their step-interval upper bounds use its left endpoint. Their variance derivatives are polynomial and can be bounded on the whole variance interval. When the required sign is certified, evaluate the appropriate variance endpoint; otherwise use interval evaluation of the polynomial itself. Also $A_4 \le 1$, $B_2 \le v_{s,+}$. Write A_4^+, B_2^+ for the resulting polynomial upper bounds. Thus choosing $a_4^+ = 1 - p_- + p_- A_4^+$ and $b_2^+ = p_+ B_2^+$ gives valid upper envelopes. For the other coefficients use

$$\begin{aligned} c_3^+ &= p_+ v_{s,+} \max|1-2s| \max_{\substack{\alpha \text{ at step endpoints} \\ v\in\{v_{s,-}, v_{s,+}\}}} |1-\alpha(1-2v)|, \\ d_4^+ &= p_+ F(v_{s,-}). \end{aligned} \tag{134}$$

The four-corner maximum is valid because the expression inside its absolute value is affine in either variable when the other is fixed.

Average (120), apply $\mathbb{E}|y| \le \sqrt{M_y}$ and (131), and divide by the square of the lower bound on M_y'. For $m = M_y$, put $w = a_2 m/(a_2 m + \alpha^2\nu n_2)$. The resulting ratio is at most

$$\begin{aligned} &\frac{\kappa a_4^+}{(a_2^-)^2} w^2 + \frac{6b_2^+}{a_2^- n_2} w(1-w) + \frac{S^2 d_4^+}{\nu_- n_2^2}(1-w)^2 \\ &\qquad + \frac{4T_3 c_3^+}{\sqrt{a_2^-}(\nu_- n_2)^{3/2}} \sqrt{w(1-w)^3}. \end{aligned} \tag{135}$$

For a step interval $[a_-, a_+]$, monotonicity of the fraction yields

$$\frac{a_2^- \ell_y}{a_2^- \ell_y + a_+^2 \nu_+ n_2} \le w \le \frac{a_2^+ u_y}{a_2^+ u_y + a_-^2 \nu_- n_2}. \tag{136}$$

All denominators are positive. The first line of (135) is a quadratic in w: its maximum is at an endpoint or, when concave, its clipped vertex. If interval rounding does not resolve the curvature sign, its three degree-two Bernstein coefficients provide a valid upper bound instead. The last factor has its maximum at $w = 1/4$ clipped to the displayed interval, as follows by differentiating $w(1-w)^3$. Partition $[0, 2]$ into all 1,024 consecutive rational intervals of width $1/512$. Take the maximum of these interval bounds and one to obtain κ'. This partitions a one-dimensional inequality at each update; it does not sample or discard any multidimensional schedule. At step zero $w = 1$, so the formula also covers zero steps without division by α.

Steps 2–5 complete the induction in (121). Every lower coefficient, upper coefficient, positivity test, and monotonicity test is evaluated with outward rounding; an unresolved test would invalidate the certificate rather than be treated as a successful step.

Step 6: converting moments into actual regret. For $D_p(x) = \mathrm{KL}(p\|p_0)$, the exact identity $D_p'(x) = xp'(x)$ implies $|D_p(x) - h_p x^2/2| \le L_p|x|^3/3$; the analogous leaf bound holds. Expand pD_s around $p_0 h_s y^2/2$ and use (123). The expected true loss satisfies

$$\begin{aligned}\mathbb{E}\mathcal{L}(x,y) &\ge aM_x + bM_y - d\sqrt{M_x}M_y,\\ a = h_p/2 - L_pR_x/3, &\quad b = p_0h_s/2 - p_+L_sR_y/3, \quad d = V_ph_s\sqrt{\kappa}/2.\end{aligned} \tag{137}$$

The derivative in M_x is at least $a - du_y/(2\sqrt{\ell_x})$ and the derivative in M_y is at least $b - d\sqrt{u_x}$. Both are verified nonnegative at every prefix. Substitution of (ℓ_x, ℓ_y) therefore gives a regret lower bound. Taking the minimum of all 26 prefix bounds covers every prescribed count from zero through 25. The initialization and every intermediate second-moment lower bound are verified positive.

Step 7: a sharper mixed upper bound with a finite remainder. To sharpen the mixed upper bound, propagate signed raw moments $m_{ij} = \mathbb{E}x^i y^j$ for $i+j \le 3$ and bound fourth-order remainders. Let $K_{ij}(x,y) = \mathbb{E}[(x')^i(y')^j \mid x, y]$ for its exact nonlinear update. The mixed update adds $.3(q_A - p)$ and $.3A(q_B - s)$ to the two coordinates of (116) with $\alpha = .3$. Use the mixed reachable boxes and second-moment upper bounds; enlarge each box to include zero before bounding derivatives. For a multi-index β, finite Taylor expansion along the segment from zero to (x, y) gives

$$\begin{aligned}K_{ij}(x,y) &= \sum_{|\beta|\le 3} \frac{D^\beta K_{ij}(0)}{\beta!}(x,y)^\beta + r_{ij}(x,y),\\ |r_{ij}(x,y)| &\le \sum_{|\beta|=4} \sup_{z \text{ in enlarged box}} \left|\frac{D^\beta K_{ij}(z)}{\beta!}\right| |(x,y)^\beta|.\end{aligned} \tag{138}$$

For the constant, the fourth directional derivative has coefficient $4!/\beta!$; the integral remainder divides by 3! and integrates $(1-t)^3$ to $1/4$, leaving exactly $1/\beta!$. Every absolute fourth-order moment is bounded by the minimum of the applicable inequalities

$$\begin{aligned}&\mathbb{E}|x|^i|y|^j \le R_x^i R_y^j,\\ &\mathbb{E}|x|^i|y|^j \le R_x^{i-2} R_y^j u_x && (i \ge 2),\\ &\mathbb{E}|x|^i|y|^j \le R_x^i R_y^{j-2} u_y && (j \ge 2),\\ &\mathbb{E}|x|^i|y|^j \le R_x^{i-1} R_y^{j-1}\sqrt{u_x u_y} && (i,j \ge 1).\end{aligned} \tag{139}$$

Factor out the supported powers and apply Cauchy–Schwarz for the last line. These bounds do not require coordinate independence. Average (138). Interval arithmetic on its polynomial part plus the signed remainder interval encloses every next raw moment. Intersect with the independently valid support and absolute-moment bounds. Starting from exact initial monomials, induction proves coverage for all 20 mixed updates. Apply the same expansion to the exact terminal KL loss, then intersect its risk interval with the structural interval. The upper endpoint remains valid.

For completeness, the derivatives are computed by finite bivariate polynomial arithmetic through degree four, not by fitting sampled values. The logistic derivatives through order four are p, v_p, $v_p(1-2p)$, $v_p(1-6v_p)$, and $v_p(1-2p)(1-12v_p)$. For $D_p' = xv_p$ they imply $D_p'' = v_p + xv_p'$, $D_p''' = 2v_p' + xv_p''$, and $D_p'''' = 3v_p'' + xv_p'''$. Store derivatives divided by their multi-index factorials; ordinary truncated multiplication then implements the product rule exactly. Compute each K_{ij} by summing the three action branches with probabilities $1-p, p(1-s), ps$ and expanding the affine noise terms by the binomial theorem. Only noise moments through degree three enter; use (131) for the third. Crucially the branch probabilities are differentiated as well as the sample maps. Interval evaluation at a box center with formal coordinate increments encloses the derivatives at every point of that box. Lost dependence may widen an interval but cannot remove a possible derivative or moment.

Step 8: complete coverage and the certified gap. Split $[.0009216, .001024]$ into 32 consecutive closed equal-width rational intervals, and split each teacher-intercept interval into four equal parts. On each variance slice, Steps 1–6 produce a single RL lower bound for every prescribed schedule and count. Step 7 produces mixed upper bounds on all 4×4 teacher rectangles; take their maximum. Compare these bounds *within the same variance slice*. Taking a global RL minimum and an unrelated global mixed maximum would lose the shared-noise information and would not prove the displayed separation.

The complete 192-bit real-ball calculation yields a smallest slice-wise gap greater than $5.322 \cdot 10^{-8} > 5 \cdot 10^{-8}$. Every saved mixed upper endpoint is also below $36.210 \cdot 10^{-6}$, so the OPD floor in (111) leaves a gap greater than $8.817 \cdot 10^{-6}$ throughout the full input region. The calculation covers all 32 variance slices, all 512 mixed teacher rectangles, and every RL prefix. Endpoints are rounded outwards to multiples of 10^{-12} with one further outward unit; the strict comparison uses these widened bounds, not the unrounded values. A 256-bit calculation checks the complete partition and every prefix inequality. Thus the guarantee covers the complete teacher/noise region and continuous RL schedule class, not selected schedules. Independent tests compare the conditional identities and full short-run trees against unexpanded reward utilities, including both extremal skewed noise laws. These tests supplement the induction and complete interval calculation; they are not the reason untested schedules are covered.

Every admissible noise variance belongs to a verified slice, and every teacher belongs to a verified rectangle. The strict slice-wise inequalities therefore prove (115). The OPD pathwise floor exceeds the displayed uniform mixed upper bound by more than $8.817 \cdot 10^{-6}$. If $20a_M \le B < 26a_R$, the mixture is affordable and RL can use at most 25 updates. Restricting to affordable recipes cannot reduce the proved pure lower bound. Combining with the OPD floor proves the equal-cost conclusion, including its stronger OPD stopping class. □

C.5 A Noise-Dependent Reversal at the Same Cost

The equal-cost advantage above is not solely a property of the teacher, initialization, or number of updates. We keep all of these fixed and vary only the variance of the independent centered reward noise. Its support bound remains $S = .032$; no action-sampling variation is removed.

Proposition C.4 (Two separated noise regions in the same nonlinear problem)**.** *Use the two-stage model, initial point, teacher rectangle, and 100-unit cost model of Proposition C.2. Keep the same fixed 20-update mixture. For every independent centered reward-noise law with support $[-S, S]$ and variance $0 \le \nu \le 2S^2/5$, let R_0 be the 25-update RL recipe*

$$\alpha_k = \frac{6}{k+3}, \qquad k = 0, \dots, 24. \tag{140}$$

Then, uniformly over the teacher rectangle,

$$\mathfrak{R}^{\mathsf{R}_0}_{25} + 10^{-6} < \min\left\{ \mathfrak{R}^{\mathsf{M}}_{20}, \inf_{a \in \mathcal{A}_D} \mathfrak{R}^{a}_{k(a)} \right\}, \tag{141}$$

where $\mathcal{A}_D$ permits every affordable OPD trajectory with history-dependent steps in $[0, 4]$ and stopping. In contrast, for $9S^2/10 \le \nu \le S^2$, Proposition C.3 certifies that the same fixed mixture beats every affordable prescribed RL schedule in $[0, 2]$ and every member of $\mathcal{A}_D$.

In plain language, the reward channel's reliability can reverse the equal-cost decision while the imperfect teacher remains unchanged. The low-noise conclusion does not rule out a different mixture or teacher schedule: it defeats the fixed mixture under study. The intervals are sufficient winning regions, not a sharp transition, and we make no ordering claim for the intervening variance band. The witness schedule is explicit, not asserted to be optimal.

Proof. **Step 1: check the recipes and information costs.** The first step of (140) is two, and its last is $2/9$. All steps are positive and decreasing, so this recipe belongs to the prescribed-step class already considered. It costs $25(3 + 1) = 100$ units, exactly the mixture's $20(3 + 1 + 1) = 100$ units. Neither the costs nor the teacher queries depend on the noise variance.

Step 2: enclose both actual nonlinear risks. Apply the signed-moment induction from Step 7 of Proposition C.3 separately to the explicit RL schedule and the unchanged mixed recipe. For RL, its conditional kernels use (116) at the prescribed step for that update; the teacher terms are zero. The reachable-set construction uses the corresponding singleton step interval, and its containment tests are checked at every update. For mixing the kernels, boxes, and teacher terms are exactly those in the preceding proof. Derivatives of the changing action probabilities are retained for both methods.

The moment induction propagates a two-sided enclosure. This time take the RL *upper* endpoint and mixed *lower* endpoint, reversing their roles in the preceding high-noise comparison. The Taylor

remainder is bounded in absolute value, so the same induction proves both directions. For every variance slice use $|\mathbb{E}\eta^3| \le S\nu - \nu^2/S$, maximized on that slice, rather than assuming symmetric noise. At $\nu = 0$ the third moment is zero and the external noise vanishes almost surely; the action draws remain random. There is no division by the variance in this signed-moment calculation, so the zero endpoint is included without a limiting argument.

Step 3: cover the complete low-noise region. Partition $[0, 2S^2/5]$ into sixteen consecutive closed rational intervals of width $S^2/40$. Split each teacher coordinate interval into four equal parts. For each variance slice i, let $U_{R,i}$ be the RL upper endpoint and let $L_{M,i}$ be the minimum mixed lower endpoint over its sixteen teacher rectangles. Both apply to the same unknown noise law in that slice. The OPD floor in (111) is pathwise and noise-independent, so it applies throughout these new noise intervals. Let L_D be an outward-rounded lower bound on that same floor.

Interval evaluation covers all sixteen variance slices and 256 teacher rectangles at 192-bit precision. Each endpoint is rounded outwards to a multiple of 10^{-12} and the calculation adds one further outward unit. On these widened endpoints the smallest gap is

$$\min_i \{\min(L_{M,i}, L_D) - U_{R,i}\} = 1.452078 \cdot 10^{-6} > 10^{-6}. \tag{142}$$

An independent 256-bit calculation reconstructs the full partition and rechecks every inequality for the stated model and fixed schedule. No interval or teacher corner is excluded. The upper/lower ordering therefore holds for every teacher and noise law in the stated region, proving (141). The high-noise claim is exactly Proposition C.3, with identical costs and recipes. □

Proof of Proposition C.2. With allowance 100 and costs $(3, 1, 1)$, each pure update costs four and each mixed update costs five. Thus the affordable counts are at most 25 and 20, respectively, including the zero-update pure option. At high noise, (107) and the variance restriction are exactly the region of Proposition C.3. Proposition C.3 therefore places the infimum over the entire stated RL class more than $5 \cdot 10^{-8}$ above the unchanged mixture. For OPD, (111) holds pathwise at every permitted checkpoint, hence also at its history-dependent stopping time. Its lower bound exceeds $45.027 \cdot 10^{-6}$, whereas the mixed upper bound is at most $36.210 \cdot 10^{-6}$. Their difference exceeds $8.817 \cdot 10^{-6} > 5 \cdot 10^{-8}$. Taking the smaller of the RL and OPD gaps proves (108).

At low noise, Proposition C.4 applies with the same teacher rectangle, initialization, and billing convention. Its explicit RL schedule has 25 updates and steps in $[0, 2]$, so it is affordable. Equation (141) is precisely (109) after rearranging. Both component certificates hold uniformly over their teacher and noise-law regions; no averaging over that uncertainty or change of comparator is required. □

D Feedback Learning Rates and Stopping

Conditional moments yield Proposition 4.3 and a two-update certificate. The statewise comparison tool also supports the baseline-error extension below. The longer-budget adaptive comparison remains unresolved.

D.1 Feedback Policies, Design, and Statewise Certificates

We allow learning rates to depend on the complete observed history and derive statewise bounds for comparing adaptive learners. A positive comparison margin is still required; Appendix D.3 establishes one for a two-update problem. The corresponding 25-versus-20 adaptive comparison remains unresolved.

Model, notation, and information. Use the raw utility, objective, and score estimator of Appendix A.2. Write $p = \mathrm{sigmoid}(\mathrm{logit}(p^\star) + x)$, $s = \mathrm{sigmoid}(\mathrm{logit}(s^\star) + y)$, $v_p = p(1-p)$, $v_s = s(1-s)$, and $D_s = \mathrm{KL}(s\|s^\star)$. The terminal loss is always $L(x, y) = \mathrm{KL}(p\|p^\star) + pD_s$. The following notation avoids confusing actual and scaled offsets:

Symbol	Meaning
h_p, h_s	Optimal variances $p^\star(1-p^\star), s^\star(1-s^\star)$
v_p, v_s	Current-policy variances $p(1-p), s(1-s)$
H	$\mathrm{diag}(h_p, p^\star h_s)$, not current curvature
$\widetilde{h}_i$	Actual teacher offset: $q_A = p^\star + h_p\widetilde{h}_1$
h_i in a scaled family	$\widetilde{h}_i = \varepsilon h_i$, not $\varepsilon^2 h_i$
η, ν	Actual reward noise and variance; locally $\eta = \varepsilon\zeta$
$B, \mathcal{B}, \tau$	Leaf action, spending allowance, and stopping time

For example, the offsets $.02, -.02$ specify $\widetilde{h}_1, \widetilde{h}_2$. The leaf teacher is $q_B = s^\star + h_s\widetilde{h}_2$. The distinction is the same as in the prescribed-schedule analysis: h_p, h_s remain fixed, while v_p, v_s change with the current policy.

Fix model parameters ϑ, including a declared fresh noise law, independent of current actions and the past. The state is a sufficient statistic for the next transition under this fixed model. The policy may nevertheless use all of $\mathcal{F}_k$, including independent randomization before its next control. Let Π_R^{fb} consist of predictable steps $\alpha_k \in [0,2]$ and stopping times $\tau \le N_R$; Π_D^{fb} uses steps in $[0,4]$ and $\tau \le N_D$. Both output θ_τ, not a retrospectively chosen previous iterate. The exclusions and accounting in (4) apply. Zero steps can pad any allowed stopping rule without changing its output.

The mixed witness is defined separately: $w_0 = 1/2$ and Borel functions $\mu_n : \mathcal{D} \to [0,2]$, using exactly N_M episodes and the direction $(1-w_0)\widehat{g}_R + w_0\widehat{u}_D$. Its coefficients are fixed in advance. A region claim requires one common μ for every $\vartheta \in \Theta$; the pure policies may know ϑ. We do not interchange these quantifiers.

Raw utilities and their conditional means. Put $a = A - p$, $b = B - s$ on a visited leaf. The root utility and continuation utility in Appendix A.2 satisfy

$$U - \mathbb{E}U = -(x + D_s)a - Ayb, \qquad U_B - \mathbb{E}[U_B \mid A = 1] = -yb. \tag{143}$$

To check this from rewards, the continuation utility is an affine function of B with slope $-y$, because its reward slope is the optimal logit while its log-policy slope is the current logit. Its mean equals the optimal root logit minus D_s. Subtracting the root log-policy penalty, whose slope is the current root logit, and then centering gives the first identity. Multiplication by the appropriate scores gives

$$\widehat{g}_R = \big(-(x+D_s)a^2 - Ayab + \eta a, A(-yb^2 + \eta b)\big), \quad \widehat{u}_D = (t_A, At_B), \quad (t_A, t_B) = (q_A - p, q_B - s). \tag{144}$$

Conditional score centering also removes the expected derivative of the sampled log-policy penalties. Hence $\mathbb{E}\widehat{g}_R = (-v_p(x + D_s), -pv_s y) = \nabla J$ and $\mathbb{E}\widehat{u}_D = (t_A, pt_B)$. The latter is the implemented semi-gradient. The full on-policy teacher loss has the additional root occupancy derivative $v_p \mathrm{KL}(q_B \| s)$; it is a different update.

Lemma D.1 (Normalized feedback design from raw second moments)**.** *Let $Q(\theta) = \theta^\top H\theta/2$ and $F(v) = v(1-3v)$. For the nominal exact-baseline estimator, define*

$$\begin{aligned}
a_R &= H_{11}v_p x(x + D_s) + H_{22}pv_s y^2, \\
a_D &= -H_{11}xt_A - H_{22}pyt_B, \\
b_{RR} &= H_{11}\{F(v_p)(x + D_s)^2 + p(1-p)^2 v_s y^2 + \nu v_p\} \\
&\quad + H_{22}\{pF(v_s)y^2 + \nu p v_s\}, \\
b_{DD} &= H_{11}t_A^2 + H_{22}pt_B^2, \\
b_{RD} &= -H_{11}v_p(x + D_s)t_A - H_{22}pv_s yt_B.
\end{aligned} \tag{145}$$

For fixed $w \in [0,1]$, set $a_w = (1-w)a_R + wa_D$ and $b_w = (1-w)^2 b_{RR} + 2w(1-w)b_{RD} + w^2 b_{DD}$. Then

$$\mathbb{E}Q(\theta + \alpha\widehat{d}_w) = Q(\theta) - \alpha a_w + \alpha^2 b_w/2. \tag{146}$$

If $b_w > 0$, its minimizing step in $[0,2]$ is $\alpha_w^Q(\theta) = \mathrm{clip}(a_w/b_w; [0,2])$. If $b_w = 0$, then $a_w = 0$; define $\alpha_w^Q(\theta) = 0$ in this case.

The rule optimizes only the next quadratic loss. The cross moment is a raw second moment, not a covariance; keeping it preserves same-rollout dependence. Neither finite-horizon optimality nor optimal nonlinear regret follows from this design lemma.

Proof. Expand $Q(\theta+\alpha d) = Q(\theta) + \alpha\theta^\top Hd + \alpha^2 d^\top Hd/2$ and condition on the current state. For the root reward square, the squared-score term is $(x+D_s)^2\mathbb{E}a^4 = F(v_p)(x+D_s)^2$; the visited cross term contributes $y^2\mathbb{E}[Aa^2b^2] = p(1-p)^2v_sy^2$; the noise square contributes νv_p. Every other cross term contains one centered leaf score or one independent centered noise. The leaf square similarly gives $pF(v_s)y^2+\nu pv_s$. Squaring the teacher direction gives t_A^2, pt_B^2. Multiplying reward and teacher coordinates before taking expectations gives $-v_p(x+D_s)t_A, -pv_syt_B$. These computations and the conditional means prove all formulas. In particular $b_w = \mathbb{E}\|H^{1/2}\widehat{d}_w\|^2 \geq 0$. If it is zero, positive definiteness of H implies $\widehat{d}_w = 0$ almost surely, and thus $a_w = 0$. For positive b_w, the derivative is $-a_w + \alpha b_w$. It is minimized at zero when $a_w \leq 0$, at two when $a_w \geq 2b_w$, and at a_w/b_w otherwise. This proves all clipping cases. □

Why the prescribed moment proof cannot be recycled. With feedback, $\mathbb{E}[\alpha_k x_k^2]$ is not $\alpha_k\mathbb{E}x_k^2$, and fourth-moment and mean inductions fail at the same step. Conditioning and minimizing first does not give the old lower bound. For example $f(z) = \nu z/(\nu+cz)$ has $f''(z) = -2c\nu^2/(\nu+cz)^3 < 0$ when $c, \nu > 0$. Hence $\mathbb{E}f(Z) \leq f(\mathbb{E}Z)$, the opposite Jensen direction needed to lower-bound a feedback risk by that unconditional expression. The remedy is to keep inequalities statewise until after the control is chosen.

Controlled kernels and Bellman values. For bounded measurable f, let $P_\alpha^R f(\theta) = \mathbb{E}[f(\theta+\alpha\widehat{g}_R) \mid \theta]$ and define $P_{\alpha,w}^M$ analogously. Assume a compact invariant $\mathcal{D}$, continuous sample maps and branch probabilities, and a fixed independent noise law with compact support. Dominated convergence makes the kernel preserve continuity jointly in state and control. With compact action intervals and continuous bounded L, backward induction yields continuous values

$$V_0^R = L, \qquad V_n^R = \min\{L, \inf_{\alpha\in[0,2]} P_\alpha^R V_{n-1}^R\}. \tag{147}$$

To avoid relying on an attained optimizer, take a countable dense rational action set: continuity makes its infimum identical. The first action in this enumeration within a chosen positive tolerance is a Borel selector, since its defining comparisons involve continuous functions. Choosing per-stage tolerances whose sum is arbitrarily small proves that these values equal the infimum over the full history-dependent, randomized, stopping class. Extra history cannot improve the conditional transition at a fixed state/control, while randomized actions average its values.

Theorem D.2 (Statewise comparison against feedback and stopping)**.** *On a common invariant domain, let bounded Borel functions satisfy*

$$\begin{aligned} &\ell_0 \leq L, \qquad \ell_n \leq L, \qquad \ell_n(\theta) \leq P_\alpha^R \ell_{n-1}(\theta) \qquad\qquad (\textit{every } \alpha\in[0,2]),\\ &U_0 \geq L, \qquad U_n(\theta) \geq P_{\mu_n(\theta),w_0}^M U_{n-1}(\theta). \end{aligned} \tag{148}$$

Suppose $p^\star \leq q_A \leq p(\theta_0)$, and let $D_{\min}$ be the teacher root's KL floor (or a uniform lower floor on a declared parameter cell). Then

$$\inf_{\pi\in\Pi_R^{\mathrm{fb}}\cup\Pi_D^{\mathrm{fb}}} \mathbb{E}L(\theta_\tau^\pi) - \mathbb{E}L(\theta_{N_M}^\mu) \geq \min\{\ell_{N_R}(\theta_0), D_{\min}\} - U_{N_M}(\theta_0) =: \Delta_{\mathrm{fb}}. \tag{149}$$

Uniform inequalities for one fixed μ prove the common-controller parameter-region claim. Strict dominance requires $\Delta_{\mathrm{fb}} > 0$; the theorem alone does not assert that positivity.

This replaces an unconditional moment comparison by a local inequality valid for whichever action the history selects. The stopping inequality is essential: an algorithm is allowed to keep its current loss instead of taking another update.

Proof. Fix a pure policy, set $N = N_R$, and define a bounded process $Z_k = \ell_{N-k}(\theta_k)$ before stopping, and $Z_k = L(\theta_\tau)$ at and after stopping. Initially $Z_0 \geq \ell_N(\theta_0)$ even if the rule stops immediately. Before stopping at k, the predictable chosen α_k may be inserted directly into the pointwise kernel inequality. At the next time, whether stopping occurs or not, $Z_{k+1} \geq \ell_{N-k-1}(\theta_{k+1})$, because

$\ell_j \le L$. Consequently $\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \ge Z_k$ on continuing histories; equality holds after stopping. Conditioning also on any pre-draw randomization proves the same statement for randomized rules. Iterating finitely to N gives $\mathbb{E}L(\theta_\tau) = \mathbb{E}Z_N \ge \ell_N(\theta_0)$. No random step has been moved outside an expectation.

For the mixed rule, insert the frozen measurable $\mu_{N_M-k}(\theta_k)$ into its upper inequality. The resulting process $U_{N_M-k}(\theta_k)$ has nonincreasing conditional expectation. Iteration gives $\mathbb{E}L(\theta_{N_M}) \le U_{N_M}(\theta_0)$.

For OPD, its root map in absolute logit z is $z \mapsto z + \alpha(q_A - \mathrm{sigmoid}(z))$. Its derivative is $1 - \alpha p(1-p) \ge 0$ for $\alpha \le 4$; it fixes the teacher logit and moves a point above that logit toward it without crossing. Induction therefore gives $p_k \ge q_A \ge p^\star$ at every stopped iterate. Bernoulli KL increases above $p^\star$, and the leaf KL is nonnegative, so $L(\theta_k) \ge \mathrm{KL}(q_A \| p^\star)$ pathwise. Taking a parameter-cell lower floor, then the smaller pure bound, and subtracting the mixed upper bound proves the result. □

D.2 Proof of the Structural Feedback Threshold

This result keeps the nonlinear transition law. Only terminal regret is sandwiched between quadratics; rates are optimized conditionally before averaging. That order is what accounts for the benefit of adaptation.

Parameters and quadratic bounds. Fix optimal probabilities, initial errors, teacher probabilities, step caps, costs, and a compactly supported fresh noise law independent of the actions and history. The teacher and reward directions are (6), in actual rather than scaled coordinates. Assume a compact rectangle containing the origin and initialization is invariant under all permitted pure and mixed controls. Write p_-, p_+ for its root-probability extrema and put

$$\underline{v}_p = \min_{\mathcal{D}} p(1-p), \quad \overline{v}_p = \max_{\mathcal{D}} p(1-p), \qquad H_- = \mathrm{diag}(\underline{v}_p, p_- \underline{v}_s), \quad H_+ = \mathrm{diag}(\overline{v}_p, p_+ \overline{v}_s), \tag{150}$$

with analogous leaf variances. Set $Q_\pm(\theta) = \theta^\top H_\pm \theta/2$. The identity $D_p'(x) = xp(1-p)$ gives $\underline{v}_p x^2/2 \le D_p(x) \le \overline{v}_p x^2/2$ on both sides of zero, by integration along the segment from zero to x. The same applies at the leaf, and multiplication by the nonnegative root probability gives $Q_- \le L \le Q_+$. No limit or unspecified Taylor remainder is used.

Conditional quantities are explicit functions of the model. For $H_Q = \mathrm{diag}(H_1, H_2)$, write $v_p = p(1-p)$, $v_s = s(1-s)$, $D_s = \mathrm{KL}(s\|s_\star)$, $\nu = \mathbb{E}\eta^2$, and $F(v) = v(1-3v)$. Lemma D.1 gives

$$\begin{aligned} a_Q &= H_1 v_p x(x + D_s) + H_2 p v_s y^2, \\ b_Q &= H_1\{F(v_p)(x+D_s)^2 + p(1-p)^2 v_s y^2 + \nu v_p\} + H_2\{pF(v_s)y^2 + \nu p v_s\}. \end{aligned} \tag{151}$$

For a mixture, put $t_A = q_A - p$, $t_B = q_B - s$ and $a_D = -H_1 x t_A - H_2 p y t_B$, $b_{DD} = H_1 t_A^2 + H_2 p t_B^2$, $b_{RD} = -H_1 v_p(x + D_s) t_A - H_2 p v_s y t_B$. Its corresponding moments are $a_w = (1-w)a_Q + w a_D$ and $b_w = (1-w)^2 b_Q + 2w(1-w)b_{RD} + w^2 b_{DD}$. These expressions expose teacher mismatch, noise, visitation, geometry, and same-sample dependence without replacing random coefficients by their means. Expected quadratic change at a selected step t is $-t a_w + t^2 b_w/2$. Actual terminal mixed regret can instead be integrated directly over its two successive kernels.

Proof of Proposition 4.3. We optimize the last action at each state, identify the exact feedback credit, and use the two terminal bounds for the two possible orderings.

1. Optimize the last rate before averaging. At any state, quadratic expansion gives

$$\mathbb{E}[Q(\theta + t\widehat{g}) \mid \theta] = Q(\theta) - t a_Q(\theta) + t^2 b_Q(\theta)/2. \tag{152}$$

If $b_Q > 0$, the optimal admissible rate is $t_Q(\theta) = \mathrm{clip}(a_Q/b_Q; [0, \bar{a}_R])$. If $b_Q = 0$, positive definiteness of H_Q implies the sampled update is zero almost surely, hence $a_Q = 0$; choose rate zero. More explicitly,

$$\Psi(a,b) = \begin{cases} 0, & a \le 0, \\ a^2/(2b), & 0 < a < \bar{a}_R b, \\ \bar{a}_R a - \bar{a}_R^2 b/2, & a \ge \bar{a}_R b, \end{cases} \tag{153}$$

with $\Psi(0,0) = 0$. The formula is used only for realizable moment pairs, so $b = 0, a \neq 0$ cannot occur. The maximum includes zero, which accounts for stopping rather than forcing an additional update.

2. Separate a prescribed continuation from its feedback credit. Fix the first rate α and let Z_α be its successor under the actual nonlinear kernel. A deterministic second rate has best quadratic risk $\mathbb{E}Q(Z_\alpha) - \Psi(\mathbb{E}a_Q(Z_\alpha), \mathbb{E}b_Q(Z_\alpha)) = F_Q(\alpha)$. A rate chosen after observing that state has best risk

$$\mathbb{E}\{Q(Z_\alpha) - \Psi(a_Q(Z_\alpha), b_Q(Z_\alpha))\} = F_Q(\alpha) - \mathcal{A}_Q(\alpha). \tag{154}$$

For every fixed second rate, its expected improvement is at most the expectation of the pointwise maximum. Taking the maximum over that fixed rate proves $\mathcal{A}_Q(\alpha) \geq 0$. This argument remains valid when clipping is active; it does not interchange a supremum and an expectation as an equality.

3. Cover every feedback history and stopping rule. The fixed known model makes the present state sufficient for the next transition, although a policy may use the full past and pre-draw randomization. Conditional on any such history, its last action cannot improve on the pointwise optimum in step 1. Randomized first controls average the corresponding values and cannot beat their infimum over α. Conversely, the deterministic first rate followed by t_Q realizes (154). The clipped control is Borel measurable; equivalently, continuous moments and compact controls allow arbitrarily accurate measurable minimizing selectors. Stopping after the first step is represented by a zero second step. Stopping immediately is also covered: taking both rates zero returns the initial state. It follows that the infimum quadratic risk over the full two-update feedback/stopping class is exactly $\inf_\alpha[F_Q(\alpha) - \mathcal{A}_Q(\alpha)]$.

Since $Q_- \leq L$, its value is a lower bound on every actual RL risk. For OPD, the monotone root map at steps through four cannot cross its teacher logit. The assumed ordering therefore gives the actual, pathwise floor $D = \mathrm{KL}(q_A \| p_\star)$ at every affordable stopping time. Taking the smaller lower bound and subtracting U_M proves (16).

4. Certify the opposite ordering with an implementable RL rule. Use the first rate in (17), then use t_{Q_+} after observing its successor. Its actual expected regret is at most its expected Q_+, exactly $F_{Q_+}(\alpha) - \mathcal{A}_{Q_+}(\alpha)$. That upper bound lies below the mixed lower bound L_M and the OPD floor D. Thus this one admissible adaptive RL rule beats the specified mixture and all permitted OPD rules. This is not a claim that every adaptive RL rule wins or that no retuned mixture could improve.

5. Account for the budget and parameter quantifiers. Let $C_R = c_0 + c_R$ and $C_M = c_0 + c_R + c_D > C_R$. The conditions $\lfloor \mathcal{B}/C_R \rfloor = \lfloor \mathcal{B}/C_M \rfloor = 2$ are equivalent to $2C_M \leq \mathcal{B} < 3C_R$. The interval is nonempty only if $2c_D < C_R$. OPD's floor holds for any number of its affordable steps, even if its cost is low enough to buy more than two. For a parameter-region claim, all inequalities and the mixed bounds must hold uniformly for one fixed controller; pure RL may know each model's parameters. No controller is silently retuned within a claimed common-controller region. □

A more explicit interpretation when clipping is inactive. Fix α and suppose $b_Q(Z_\alpha) > 0$ almost surely and all pointwise optimal rates $t_* = a_Q(Z_\alpha)/b_Q(Z_\alpha)$ lie in $[0, \bar{a}_R]$. Their b_Q-weighted mean $\bar{t} = \mathbb{E}a_Q(Z_\alpha)/\mathbb{E}b_Q(Z_\alpha)$ lies in the same interval. Substituting the quadratic branch of (153) and completing a square gives the identity

$$\mathcal{A}_Q(\alpha) = \frac{1}{2}\left\{\mathbb{E}\frac{a_Q(Z_\alpha)^2}{b_Q(Z_\alpha)} - \frac{(\mathbb{E}a_Q(Z_\alpha))^2}{\mathbb{E}b_Q(Z_\alpha)}\right\} = \frac{1}{2}\mathbb{E}[b_Q(Z_\alpha)(t_* - \bar{t})^2]. \tag{155}$$

The credit vanishes when all successors favor the same rate and grows with their weighted disagreement. Reward noise can therefore both impede RL and make adaptation more valuable. With clipping, the definition in (15) remains exact but this weighted-variance formula need not apply.

Why variance alone cannot determine the threshold. Although (151) at a fixed state uses only the noise variance, Z_α is random and Ψ is generally a nonlinear clipped function. Its expectation can depend on the complete first-noise law. A two-point or three-point noise calculation must not be advertised as uniform over all bounded noise laws with that variance. Finite support permits explicit branch enumeration; other laws require valid integration or moment-envelope bounds.

A positive region with baseline errors. At zero baseline error, the function m in (163) is precisely $Q_- - \Psi(a_{Q_-}, b_{Q_-})$. Thus its complete first-action enclosure lower-bounds the infimum in (16). The mixed terminal calculation supplies U_M. The positive parameter region in Appendix D.3 therefore verifies this structural criterion, rather than just the formal possibility that a gap could be positive. Its error-box lower bound is conservative also at zero error. For nonzero predictable baseline errors, its direct argument permits the last-step second moment to improve as well as worsen; the guarantee is not obtained by assuming critic errors harm RL. The general Lipschitz charges in Theorem E.3 provide an alternative sufficient extension whenever a nominal margin exceeds those charges.

D.3 A Two-Update Feedback Separation

We establish a nonempty parameter region in which one mixed controller outperforms all permitted adaptive pure rules, even with predictable baseline errors. This applies the feedback-credit criterion of Proposition 4.3 to a two-update problem. Its budget and teacher region differ from those of Proposition C.3; the result does not extend that 25-versus-20 comparison to adaptive RL.

A certified parameter region. For this result, write $L = \mathcal{L}$, let $H = \mathrm{diag}(h_p, p_\star h_s)$, and set $Q(\theta) = \theta^\top H\theta/2$. Define a common mixed controller by taking step two on the first update and, on the second, choosing

$$\alpha^Q_{1/2}(\theta) = \operatorname*{arg\,min}_{0\le\alpha\le 2} \mathbb{E}_{\mathrm{nom}} Q\big(\theta + \alpha(\widehat{g} + \widehat{u})/2\big). \tag{156}$$

The nominal expectation uses optimal logits $(.1, .1)$, exact baselines, teacher offsets $(.039, -.02)$, and noise probabilities $\Pr(\eta = \pm.032) = .45$, $\Pr(\eta = 0) = .1$. Choose zero if the direction has zero second moment. In words, the mixture takes a fixed first step, then chooses the second step that minimizes predicted local quadratic loss at its new state. The scalar minimization has an explicit clipped solution (Lemma D.1). The same nominal model is used throughout; the controller does not retune to the actual teacher or noise law.

Proposition D.3 (A common feedback witness with baseline-error tolerance)**.** *Use optimal logits* $(.1, .1)$*, initial error* $(.04, .04)$*, noise support* $S = .032$*, and independent noise with probabilities* $\Pr(\eta = \pm S) = \zeta/2$*,* $\Pr(\eta = 0) = 1 - \zeta$*. Let the* actual *teacher offsets and activation range over*

$$\widetilde{h}_1 \in [.0385, .0395], \quad \widetilde{h}_2 \in [-.021, -.019], \quad \zeta \in [.89, .91]. \tag{157}$$

Take costs $(c_0, c_R, c_D) = (3, 1, 1)$ *and allowance* $\mathcal{B} = 10$*, so each pure method and the mixture can afford two updates. Let* $\widetilde{\Pi}^{\mathrm{fb}}_R$ *and* Π^{fb}_D *respectively permit every predictable RL step in* $[0, 2]$ *and OPD step in* $[0, 4]$*, with current-iterate stopping at any stopping time* $\tau \le 2$*. In both RL and the mixture, replace* $\widehat{g}$ *by* $\widetilde{g} = \widehat{g} - (e_A a, Ae_B b)$*, where* $|e_A|, |e_B| \le 10^{-4}$ *are chosen before the corresponding action and reward noise; OPD targets remain exact.*

Fix one mixed controller for the whole region: $w_0 = 1/2$*,* $\mu_2 = 2$*, and* $\mu_1(\theta) = \alpha^Q_{1/2}(\theta)$ *defined above. Then, uniformly over the region and these baseline-error policies,*

$$\inf_{\pi \in \widetilde{\Pi}^{\mathrm{fb}}_R \cup \Pi^{\mathrm{fb}}_D} \mathbb{E}L(\theta^\pi_\tau) - \mathbb{E}L(\widetilde{\theta}^\mu_2) > 10^{-5}. \tag{158}$$

Here $\widetilde{\theta}^\mu_2$ *is the mixed iterate with these baseline errors. The certified relative reduction against the better pure optimum exceeds* 8%*. Both are conservative guarantees, not estimates of the actual gap.*

What the proposition adds. The pure methods can observe their progress, adjust the next rate, or stop, yet the mixture still wins. The guarantee concerns actual nonlinear regret, not merely the quadratic used to design its controller, and survives the stated errors in the reward baselines. The budget buys two updates for every method, so pure-method savings cannot purchase a third. The gain above 8% belongs to this separate short-horizon problem; extending it to the 25-versus-20 budget with adaptive RL remains unestablished.

The final rate includes a zero step on part of the state space. The comparison does not cover algorithms that switch between pure channels.

Continuation bounds. For $L = \mathcal{L}$, the required inequalities are

$$\ell_0 \le L, \quad \ell_n \le \min\{L, P_\alpha^R \ell_{n-1}\}\ (\forall \alpha \in [0,2]), \qquad U_0 \ge L, \quad U_n \ge P^M_{\mu_n,1/2} U_{n-1}. \tag{159}$$

Here P averages the indicated function over the nonlinear next-state distribution. The function ℓ bounds every pure RL continuation from below, while U bounds one fixed mixed controller from above. Because the inequalities hold for every allowed pure step at each state, they also hold for the step selected from the observed history. Iteration and the OPD floor $D_{\min}$ give the gap $\min\{\ell_{N_R}(\theta_0), D_{\min}\} - U_{N_M}(\theta_0)$.

Proof of Proposition D.3. We give the analytic reduction, the complete continuous-control enclosure, and the resulting numerical bounds. No fitted risk surface or Monte Carlo mean is used in this certificate.

1. A common invariant domain, including errors. Take $\mathcal{D} = [-.2, .2] \times [-.08, .08]$. Let p_-, p_+, s_-, s_+ be its exact logistic endpoints, and set $a_- = \min(p_-, 1-p_+)$, $a_+ = \max(p_+, 1-p_-)$ and analogous b_-, b_+. The verifier checks

$$2a_+^2 < 1, \quad 2b_+^2 < 1, \quad Y^2/8 + Y b_+/a_- + (S+\delta)/a_- < X, \quad (S+\delta)/b_- < Y \tag{160}$$

at $X = .2$, $Y = .08$, $\delta = 10^{-4}$. It also checks that both teacher probabilities lie strictly between their coordinate endpoint probabilities, and that $p^\star < q_A < p(\theta_0)$. These inequalities use the exact logistic endpoints, not rounded endpoints treated as equalities. The convex-combination argument of Appendix E.2 proves containment for every permitted feedback rule, including errors. For the mixture this follows from the contained component maps at the same step. The OPD floor is therefore valid for every allowed stopping time.

2. Eliminate the last pure feedback action analytically. Let c_p and c_s be lower bounds on $p(1-p)$ and $s(1-s)$ on the whole domain. Integrating the identity $D'(z) = z\,\mathrm{sigmoid}'(\mathrm{logit}(p^\star) + z)$ on either side of zero gives

$$L(x,y) \ge Q_-(x,y) := (c_p x^2 + p_- c_s y^2)/2. \tag{161}$$

This quadratic is only a lower certificate; the compared performance remains the actual nonlinear loss L. Use $H_- = \mathrm{diag}(c_p, p_- c_s)$ in the moment formulas of Lemma D.1, and call their reward entries a_Q^-, b_Q^-. By Lemma E.2, the smallest attainable reward second moment under a baseline error box is

$$\begin{aligned} b_Q^{-,\delta} = b_Q^- &+ (H_-)_{11} v_p \min_{|e|\le\delta}\{e^2 + 2e(x + D_s)(1-2p)\} \\ &+ (H_-)_{22} p v_s \min_{|e|\le\delta}\{e^2 + 2ey(1-2s)\}. \end{aligned} \tag{162}$$

Each minimum is at the projection of the negative linear coefficient onto $[-\delta, \delta]$. The conditional mean progress a_Q^- is unchanged. Thus the function

$$m(\theta) = Q_-(\theta) - \max_{0\le\alpha\le 2}\{\alpha a_Q^- - \alpha^2 b_Q^{-,\delta}/2\} \tag{163}$$

is an exact one-step optimal value for the terminal quadratic under this error class. It satisfies $0 \le m \le Q_- \le L$ and lower-bounds both stopping now and every permitted one-step continuation to L. The maximum is the clipped quadratic optimum, with the zero-denominator convention already proved; the numerical enclosure verifies a strictly positive denominator on every image it uses.

This analytic inequality holds at every state of $\mathcal{D}$ and for every last-step control and predictable offset. Consequently, conditioning after the first episode, every pure two-update policy has risk at least

$$\inf_{\alpha\in[0,2],\, e\in[-\delta,\delta]^2} \widetilde{P}^R_{\alpha,e} m(\theta_0). \tag{164}$$

Immediate stopping is also covered: taking $\alpha = 0$ gives $m(\theta_0) \le L(\theta_0)$. First-step randomized controls average these lower-bounded values. Histories with the same state but different last-step decisions are all covered by the pointwise definition of m.

3. Verify every first action and parameter cell. Partition the two teacher intervals into two equal subintervals each and the activation interval into four. Within each of the resulting 16 cells, partition

$[0, 2]$ into 1,024 closed action intervals with shared endpoints. For each action interval, enumerate all nine first-episode branches with their actual probabilities, enclose the whole error box at the first transition, evaluate an outward lower bound on m at every resulting state box, and sum. Take the smallest lower endpoint over the complete action partition. This verifies (164); it is not optimization on a learning-rate grid. No numerical full-state grid is required here because step 2 proves the last-state inequality analytically.

For clarity, on a state box the verifier bounds the quadratic progress using its upper a_Q^- endpoint and lower positive $b_Q^{-,\delta}$ endpoint. The maximum over nonnegative steps is monotone in those directions. Subtract this upper progress from a lower Q_- endpoint, and intersect with the valid lower bound zero. This is conservative even when the state, step, baseline, and parameters are dependent. Each offset-box minimum in (162) is enclosed at its clipped vertex. Failed positivity or containment aborts the certificate.

4. Bound the fixed mixture and compare within cells. Enumerate its nine first branches at $\mu_2 = 2$, enclose the fixed μ_1 on each whole successor box, then enumerate the nine second branches and bound the actual terminal KL. Baseline errors are enclosed at both transitions. Ignoring dependencies between interval quantities enlarges these enclosures; it does not omit any admissible history. For stable KL evaluation, its minimum is zero when the logit-error interval crosses zero and otherwise occurs at its nearer endpoint; its maximum occurs at an endpoint. Those endpoint values are evaluated outwards with the exact optimal logit. The OPD floor is enclosed directly as $\mathrm{KL}(q_A \| p^\star)$ at the cell's lower root-teacher endpoint.

Interval calculations at 192-bit precision, checked at 256 bits, give across all cells

$$L_R > 1.64985 \cdot 10^{-4}, \qquad D_{\min} > 1.84948 \cdot 10^{-4}, \qquad U_M < 1.51919 \cdot 10^{-4}. \tag{165}$$

Even these deliberately coarsened global endpoints imply a gap greater than $1.3 \cdot 10^{-5}$ and a relative improvement greater than 7.9%. Comparing the bounds *within each shared parameter cell* gives gap greater than $1.3557 \cdot 10^{-5}$ and relative reduction greater than 8.21%, establishing the more conservative advertised claims. Indeed if $R_* \geq L_* > 0$ and $R_M \leq U$, then $(R_* - R_M)/R_* \geq 1 - U/L_*$; the stored rational endpoints verify this inequality cell by cell. This completes the actual-regret comparison. □

What the short certificate verifies. The same mixed controller works throughout the teacher/noise region; it is not refitted inside each cell. The pure learner may know that cell's exact parameters and choose its feedback and baseline policy accordingly. The error tolerance is an actual baseline error of 10^{-4}, equal to 0.3125% of the noise support amplitude, not a relative error in an unknown critic. The estimator, teacher queries, state probabilities, terminal regret, and per-episode costs are unchanged. The teacher root is slightly better than the initial root, but has a strict reward mismatch. The proof does not invoke a common average update or assume the last feedback value is quadratic: m is a clipped rational function of state-dependent moments.

Computational verification and scope. The interval calculation covers all 16 parameter cells and all 1,024 first-action intervals per cell, with outward-rounded endpoints. Every bound is checked at 256-bit precision and compared with independent raw-utility and short-history calculations. After analytic elimination of the final action, construction costs $O(CAJ + CJ^2)$ scalar enclosure evaluations, with $C = 16$, $A = 1024$, and $J = 9$. This is not a general linear-horizon feedback algorithm. A positive uniform separation for the 25-versus-20 adaptive comparison remains unproved; the two-update result does not control the benefit of additional RL updates.

E Robustness to Value-Baseline Errors

The first extension concerns prescribed RL schedules; the second concerns feedback. Both allow baseline errors to improve as well as worsen RL variance, and account for their effects on both pure and mixed learners.

E.1 Robustness to Predictable Value-Baseline Errors

This extension changes the score estimator, not just the teacher or noise parameters. A *value baseline* is the action-independent control variate in a policy-gradient estimate; it is distinct from a pure-method comparison baseline. Pure RL may choose these control variates from its history and current state. Its learning-rate sequence remains prescribed. We first give this stronger comparator the benefit of variance-minimizing control variates, then charge the mixture for bounded baseline errors.

Admissible perturbations and information. At an actual state (x, y), replace (62) by

$$\widetilde{g}_1 = \widehat{g}_1 - e_1(A - p), \qquad \widetilde{g}_2 = \widehat{g}_2 - Ae_2(B - s), \qquad |e_i| \le \varepsilon\tau. \tag{166}$$

The root offset is chosen before the current root action and reward noise; the leaf offset is chosen before its action and that noise, conditional on visitation. Both may depend arbitrarily on past data, current parameters, and independent internal randomness. Condition on these choices in each calculation below. The score identities imply zero conditional mean perturbation. No smoothness of the baseline function is required. Action-dependent offsets or offsets fitted to the current reward noise need not preserve this identity and are not covered. OPD is unchanged. The cost comparison assumes these offsets use already available history: extra data for baseline fitting or auditing must be charged to the budget.

A stronger pure-RL reference. Define $(\ell^{\mathrm{vb}}_{1,k}, \ell^{\mathrm{vb}}_{2,k})$ by (65), replacing $F(h_p)$ by h_p^2 and $F(h_s)$ by h_s^2. Put

$$R^{\mathrm{vb}}_R = (h_p\ell^{\mathrm{vb}}_{1,N_R} + p_\star h_s \ell^{\mathrm{vb}}_{2,N_R})/2, \qquad \Gamma_{\mathrm{vb}} = \min\{R^{\mathrm{vb}}_R, h_p h_1^2/2\} - R_M. \tag{167}$$

These smaller multiplier variances strengthen the pure comparator. The new lower bound does not assume that adding baseline error only hurts RL; an appropriate offset can improve on the original state-value baseline.

Explicit robustness charges. Retain the notation of Appendix B.1. Require its box tests with T replaced by $T + \tau$, and its teacher-segment conditions. In the following constants, however, T is the original noise bound:

$$\begin{aligned} B^{\mathrm{vb}}_R &= \max\{R_1^2/10 + R_2^2/32 + R_2(R_1 + R_2)/4 + TR_1/4, R_2^2/10 + TR_2/4\}, \\ L_{\mathrm{vb}} &= 1 + a_{\max}(1 - p_\star)[\max(s_\star, 1 - s_\star) - p_\star]_+, \\ C^{\mathrm{vb}}_R &= t + H_0 R a_{\max} B^{\mathrm{vb}}_R \sum_{j=0}^{N_R - 1} L^j_{\mathrm{vb}} + N_R Q_{\max} L_\pi/2. \end{aligned} \tag{168}$$

Let $P_{1,k}, P_{2,k}$ be the diagonal Hessian of the mixed reference quadratic value at time k, with terminal $(P_{1,K_M}, P_{2,K_M}) = (h_p, p_\star h_s)$. If a_{11}, a_{22}, c_{12} are the coefficients of U, V in (64) at (λ, w), then

$$P_{1,k} = a_{11}P_{1,k+1}, \qquad P_{2,k} = a_{22}P_{2,k+1} + c_{12}P_{1,k+1}. \tag{169}$$

All these entries are nonnegative. Set $r = 1 - w$, $\chi_p = |1 - 2p_\star|$, $\chi_s = |1 - 2s_\star|$, and

$$\begin{aligned} W_1 &= \sum_{k<K_M} \lambda^2 r^2 P_{1,k+1}/4, & W_2 &= \sum_{k<K_M} \lambda^2 r^2 P_{2,k+1} p_+/4, \\ D_{\mathrm{vb}} &= (W_1 + W_2)/2, & A_{\mathrm{vb}} &= W_1 R_1 \chi_p + W_2 R_2 \chi_s, \\ B_{\mathrm{vb}} &= W_1(R_1^2/2 + R_2^2\chi_p/8 + \varepsilon_0 R_1 R_2^2/16) + W_2 R_2^2/2. \end{aligned} \tag{170}$$

Here p_+ is the actual-probability endpoint defined before (67). The charges are finite backward sums, not fitted sensitivity coefficients.

Theorem E.1 (Finite tolerance to predictable value-baseline errors)**.** *Under the model, costs, and step restrictions of Theorem 4.1, impose the enlarged box tests above. Allow pure RL and the fixed mixture to use any offsets* (166)*. Define*

$$\Delta_\tau = \Gamma_{\mathrm{vb}} - A_{\mathrm{vb}}\tau - D_{\mathrm{vb}}\tau^2, \qquad E_\tau = \max(C_R^{\mathrm{vb}}, t_D) + C_M + B_{\mathrm{vb}}\tau, \tag{171}$$

where C_M, t_D use the original noise bound T. For every $0 < \varepsilon \le \varepsilon_0$, the pure-minus-mixed regret gap is at least

$$\varepsilon^2(\Delta_\tau - \varepsilon E_\tau). \tag{172}$$

The infimum is over all affordable prescribed RL counts and schedules and *their admissible predictable baseline policies, and all allowed adaptive OPD rules. The bound holds for every admissible mixed baseline policy. If $\Delta_\tau > 0$, choosing $\varepsilon \le \min(\varepsilon_0, \Delta_\tau/(2E_\tau))$ gives a gap at least $\varepsilon^2\Delta_\tau/2$; if $E_\tau = 0$, omit the second radius restriction.*

The leading error allowance is explicitly quadratic plus linear in the baseline tolerance. At symmetric optimal probabilities, the linear term vanishes. Away from symmetry its coefficient measures that departure. For $K_M \ge 1$, $D_{\mathrm{vb}} > 0$, and a positive reference margin permits

$$0 \le \tau < \frac{\sqrt{A_{\mathrm{vb}}^2 + 4D_{\mathrm{vb}}\Gamma_{\mathrm{vb}}} - A_{\mathrm{vb}}}{2D_{\mathrm{vb}}}, \tag{173}$$

subject also to containment. This is a tolerance certificate, not a guarantee that a learned critic achieves the required error.

Proof. **1. Conditional variance minimization.** Write $v_p = p(1-p)$ and $v_s = s(1-s)$. The Bernoulli identity $(A-p)^2 = v_p + (1-2p)(A-p)$ gives the variance-minimizing root offset $e_1^\circ = -(x + D_s)(1-2p)$. Likewise $e_2^\circ = -y(1-2s)$. The corresponding directions are

$$g_1^\circ = -v_p(x + D_s) - Ay(A-p)(B-s) + \varepsilon\zeta(A-p), \quad g_2^\circ = A\{-v_s y + \varepsilon\zeta(B-s)\}. \tag{174}$$

For any other predictable offsets, the difference in each direction is $-(e_1 - e_1^\circ)(A-p)$ or $-A(e_2 - e_2^\circ)(B-s)$. Each has conditional mean zero. Its cross product with the corresponding direction in (174) has expectation zero: the constant drift multiplies a centered score; the root cross term has one centered leaf score; and noise terms vanish by independence and centering of ζ. Thus, at any fixed state and step α,

$$\begin{aligned}\mathbb{E}[(x + \alpha\widetilde{g}_1)^2 - (x + \alpha g_1^\circ)^2] &= \alpha^2 v_p(e_1 - e_1^\circ)^2,\\ \mathbb{E}[(y + \alpha\widetilde{g}_2)^2 - (y + \alpha g_2^\circ)^2] &= \alpha^2 p v_s(e_2 - e_2^\circ)^2.\end{aligned} \tag{175}$$

The leaf expression conditions on its offset when visited. Both expressions are nonnegative. We will apply this identity to diagonal quadratic future values, not infer terminal-risk ordering from one-step variance alone.

2. The pure schedule-class bound. Freeze probabilities in (174) at their optima, set $D_s = 0$, and divide the state by ε:

$$X' = (1 - \alpha h_p)X - \alpha AYab + \alpha\zeta a, \qquad Y' = (1 - \alpha A h_s)Y + \alpha A\zeta b. \tag{176}$$

Its second moments obey $U' = (1-\alpha h_p)^2 U + \alpha^2 b_0 V + \alpha^2 \nu h_p$ and $V' = (1 - 2\alpha p_\star h_s + \alpha^2 p_\star h_s^2)V + \alpha^2 \nu p_\star h_s$. The same conditioning identities used for (64) remove all cross terms. All moment coefficients are nonnegative. Minimizing the next two moments separately therefore proves the relaxation defining R_R^{vb} by induction. Zero-step padding includes shorter counts. For any prescribed remaining schedule, the reference future value has a diagonal positive-semidefinite Hessian: the only possible cross coefficient contains $\mathbb{E}[Aab] = 0$. Its linear coefficient is zero because the noise is centered. Consequently (175) also bounds its one-step expectation for every predictable baseline policy.

3. Containment and pure nonlinear correction. For an arbitrary bounded baseline offset, the coordinate-wise effective noise is $\zeta - e_i/\varepsilon$, of magnitude at most $T + \tau$. The convex-combination proof of (67) hence contains all perturbed pure and mixed trajectories. No independence is used for this pathwise containment. The auxiliary optimal directions are also contained, even when their optimal offsets exceed the allowed tolerance: their root multiplier is $1 - \alpha v_p$ and their target is $-D_s/\varepsilon - AYab/v_p + \zeta a/v_p$. When $A = 1$, $a/v_p = 1/p$; for either root action $|a|/v_p \le 1/a_-$. At a

visited leaf the multiplier is $1 - \alpha v_s$ and the target has magnitude at most T/b_-. Thus the same box tests contain these targets and the frozen reference targets. OPD containment and its teacher floor are unchanged.

Comparing the scaled optimal directions with (176) gives map error at most $\alpha\varepsilon B_R^{\mathrm{vb}}$. Indeed, $|v_p - h_p| \le |x|/10$, the root KL contribution is at most $\varepsilon R_2^2/32$, the cross-score contribution is at most $\varepsilon R_2(R_1 + R_2)/4$, and the noise-score contribution is at most $\varepsilon T R_1/4$; the leaf has the last two terms displayed in (168). These use the global logistic derivative bounds, so hold at every state in the rectangle. The root row sum of a visited reference matrix is $1 + \alpha(1 - p_\star)(|B - s_\star| - p_\star)$; other rows have norm at most one. This proves the bound L_{vb}. Reference future values lie in $[0, Q_{\max}]$ and, with j remaining steps, have gradient one-norm at most $H_0 R L_{\mathrm{vb}}^j$. The branch-law discrepancy is still at most εL_π.

Apply the variance inequality in step 1 to the common reference future value at the *actual perturbed state*; then compare the optimal actual map with its reference map and sampling law. This gives a one-sided Bellman error bounded by the map/law charges above, uniformly over that state's history-dependent offsets. Telescoping along the perturbed histories and adding terminal error $t\varepsilon^3$ proves

$$\mathbb{E}\mathcal{L}(\theta^R) \ge \varepsilon^2 R_R^{\mathrm{vb}} - \varepsilon^3 C_R^{\mathrm{vb}}. \tag{177}$$

This argument does not replace the perturbed histories by nominal ones, nor couple discontinuous baseline functions at nearby states.

4. Mixed-risk tolerance. Use the unscaled future value $\varepsilon^2 V_k((x, y)/\varepsilon)$, whose Hessian is still P_k. The mixed reference Hessian is diagonal by the same centered-leaf-score identity, giving (169). Its linear term need not vanish, but baseline corrections have zero mean. At a fixed actual state, expanding this quadratic future value gives the exact unscaled one-step change due to the offsets:

$$\begin{aligned}\frac{\lambda^2 r^2}{2}\big[&P_{1,k+1} v_p\{e_1^2 + 2e_1(x + D_s)(1 - 2p)\} \\ &+ P_{2,k+1} p v_s\{e_2^2 + 2e_2 y(1 - 2s)\}\big].\end{aligned} \tag{178}$$

The teacher cross terms vanish: the root teacher update is constant given the state, and the visited leaf teacher update multiplies a centered leaf score. The reward cross terms follow from $\mathbb{E}(A - p)^3 = v_p(1 - 2p)$ and its leaf analogue. Bound $v_p \le 1/4$, $pv_s \le p_+/4$, and use

$$\begin{aligned}\frac{|x + D_s|}{\varepsilon}|1 - 2p| &\le R_1\chi_p + \varepsilon(R_1^2/2 + R_2^2\chi_p/8 + \varepsilon_0 R_1 R_2^2/16), \\ \frac{|y|}{\varepsilon}|1 - 2s| &\le R_2\chi_s + \varepsilon R_2^2/2.\end{aligned} \tag{179}$$

These follow by multiplying $|1 - 2p| \le \chi_p + \varepsilon R_1/2$ and $|x + D_s|/\varepsilon \le R_1 + \varepsilon R_2^2/8$. Summing (178) yields at most $\varepsilon^2(A_{\mathrm{vb}}\tau + D_{\mathrm{vb}}\tau^2) + \varepsilon^3 B_{\mathrm{vb}}\tau$. At the same actual states, the nominal mixed map/law error is bounded by C_M, using the original noise bound; its reference future values and all compared maps are contained in the common rectangle. Bellman telescoping and terminal Taylor error therefore give

$$\mathbb{E}\mathcal{L}(\theta^M) \le \varepsilon^2(R_M + A_{\mathrm{vb}}\tau + D_{\mathrm{vb}}\tau^2) + \varepsilon^3(C_M + B_{\mathrm{vb}}\tau). \tag{180}$$

Subtract this from the smaller of (177) and the unchanged OPD floor $\varepsilon^2 h_p h_1^2/2 - \varepsilon^3 t_D$. This proves (172), the radius statement, and the positive-root tolerance in (173). □

A continuous, nonsymmetric certified family. Take $p_\star, s_\star \in [.505, .515]$, $(x_0, y_0) = (1, 1)$, $(h_1, h_2) = (1/2, -1/2)$, $\nu = 16/25$, $T = 4/5$, and $\tau = 1/20$. Use $a_{\max} = 2$, $\lambda = 3/5$, $w = 1/2$, $K_M = 20$, and costs $(c_0, c_R, c_D) = (3, 1, 1)$ at budget 100, so $N_R = 25$. With $(R_1, R_2) = (21/5, 19/10)$ and $\varepsilon_0 = 1/100$, the box tests and the bound

$$\Delta_\tau - 10^{-6} E_\tau > 2 \cdot 10^{-3} \tag{181}$$

hold throughout this rectangle. Hence every $0 < \varepsilon \le 10^{-6}$ has pure-minus-mixed actual regret gap greater than $.002\varepsilon^2$, uniformly over all admissible baseline policies. All these leaf rewards are nonconstant. The allowed value baseline error is $\varepsilon/20$, or 6.25% of the noise support amplitude, not 6.25% of an unknown optimal critic's value.

The numerical calculation covers every cell of a complete 32×32 interval partition using Arb arithmetic, with exact rational input endpoints; it does not infer a continuous claim from grid samples. It propagates lower moment endpoints using monotonicity, and each cell's bounds are checked at 256-bit precision. Exact rational branch enumeration independently tests the variance and quadratic-charge identities. The numerical certificate supplements, rather than replaces, the proof above. This establishes an explicit estimator and geometry tolerance, but at a small nonlinear radius. It does not extend the separate .04-scale certificate to perturbed baselines.

E.2 Predictable Baseline Errors Under Feedback

This section bounds the effect of value-baseline errors under adaptive learning rates. Teacher targets remain exact, so the OPD floor still applies. All error bounds below are conditional on the complete history before the current episode.

Lemma E.2 (Mean and second-moment changes)**.** *Let e_A, e_B be measurable before the current actions and reward noise, and replace the reward estimator by $\widetilde{g}_R = \widehat{g}_R - (e_A a, A e_B b)$. Its conditional mean is unchanged. The exact raw-moment changes are*

$$\begin{aligned} \mathbb{E}(\widetilde{g}_1^2 - \widehat{g}_1^2) &= v_p\{e_A^2 + 2e_A(x + D_s)(1-2p)\}, \\ \mathbb{E}(\widetilde{g}_2^2 - \widehat{g}_2^2) &= p v_s\{e_B^2 + 2e_B y(1-2s)\}, \\ \mathbb{E}(\widetilde{g}_1\widetilde{g}_2 - \widehat{g}_1\widehat{g}_2) &= v_p v_s y(e_A + e_B). \end{aligned} \tag{182}$$

For $|e_A|, |e_B| \le \delta$,

$$\mathbb{E}\|\widetilde{g}_R - \widehat{g}_R\|^2 = e_A^2 v_p + e_B^2 p v_s \le \delta^2/2. \tag{183}$$

In plain language, prediction errors in an action-independent baseline preserve the reward direction but can improve or worsen its variance. They also change the cross-coordinate moment, which matters for nonquadratic continuation values. They are not independent added noise.

Proof. The mean change vanishes because $\mathbb{E}a = 0$ and $\mathbb{E}[Ab] = 0$. Expanding the root square gives $e_A^2\mathbb{E}a^2 - 2e_A\mathbb{E}[a\widehat{g}_1]$. The second expectation equals $-(x + D_s)\mathbb{E}a^3 = -(x + D_s)v_p(1-2p)$: the term with one b and the term with one η vanish. This proves the first identity. The leaf calculation uses $A^2 = A$ and $\mathbb{E}[b^3 \mid A = 1] = v_s(1-2s)$, giving the second. For the cross moment expand $-e_A\mathbb{E}[a\widehat{g}_2] - e_B\mathbb{E}[Ab\widehat{g}_1] + e_A e_B\mathbb{E}[Aab]$. The last term is zero; both preceding expectations equal $-y v_p v_s$, since $\mathbb{E}[Aa] = v_p$ and the remaining centered scores or noise terms vanish. This proves the third identity. Finally the squared Euclidean norm has no cross-coordinate term and equals $e_A^2 a^2 + A e_B^2 b^2$; average and use $v_p, v_s \le 1/4$, $p \le 1$. If the offsets depend on the current action or current reward noise, these cancellations need not hold. Conditioning on pre-draw internal randomization is permitted and preserves every identity. □

Theorem E.3 (Feedback-gap robustness for a frozen controller)**.** *Suppose the nominal functions in Theorem D.2 are Lipschitz on a common domain invariant also for all errors in $[-\delta, \delta]^2$. Keep the mixed controller fixed as a function of state and remaining budget. Define*

$$E_R(\delta) = \sqrt{2}\delta \sum_{j=0}^{N_R - 1} \mathrm{Lip}(\ell_j), \quad E_M(\delta) = \sqrt{2}(1 - w_0)\delta \sum_{j=0}^{N_M - 1} \mathrm{Lip}(U_j). \tag{184}$$

Allow arbitrary predictable baseline-error policies for pure RL and the mixture. Then

$$\begin{aligned} \inf_{\pi \in \widetilde{\Pi}_R^{\mathrm{fb}} \cup \Pi_D^{\mathrm{fb}}} \mathbb{E}L(\theta_\tau^\pi) - \mathbb{E}L(\widetilde{\theta}_{N_M}^{\mu}) &\ge \min\{\ell_{N_R}(\theta_0) - E_R(\delta), D_{\min}\} - U_{N_M}(\theta_0) - E_M(\delta) \\ &\ge \Delta_{\mathrm{fb}} - E_R(\delta) - E_M(\delta). \end{aligned} \tag{185}$$

This charges errors in both the comparator and the witness. It does not assume a perturbed critic must harm RL, nor that its history is summarized by the policy parameters. A nonzero certified tolerance requires a positive nominal gap and computed charges plus perturbed containment. The two-update example instead bounds the error box directly, avoiding a potentially loose global Lipschitz charge.

Proof. Couple nominal and perturbed transitions at the same current state, control, actions, and reward noise. Their Euclidean distance is $\alpha\|\widetilde{g}_R - \widehat{g}_R\|$ for RL, and $\alpha(1-w_0)\|\widetilde{g}_R - \widehat{g}_R\|$ for the mixture. By Cauchy–Schwarz and (183), for a G_f-Lipschitz function,

$$|(\widetilde{P}^R_\alpha - P^R_\alpha)f| \le G_f\alpha\delta/\sqrt{2}, \qquad |(\widetilde{P}^M_{\alpha,w_0} - P^M_{\alpha,w_0})f| \le G_f\alpha(1-w_0)\delta/\sqrt{2}. \tag{186}$$

These inequalities are uniform over all allowed offsets at the present history. Thus they hold even though the perturbed kernel may depend on critic history; no false Markov assumption is needed.

Subtract the cumulative pure charge through n from ℓ_n, and add the mixed charge through n to U_n. At $n = 0$ the charges vanish. The step cap two and (186) show, exactly by induction on the remaining horizon, that the modified functions obey the perturbed lower/upper transition inequalities. The pure stopping inequality remains valid because a nonnegative number was subtracted. Apply the stopped-process proof to every history-dependent offset policy. The teacher floor did not change. Taking the smaller pure bound proves the first line; $\min(a-e,b) \ge \min(a,b) - e$ for $e \ge 0$ proves the second. □

Perturbed containment and information costs. Here is the pathwise containment argument used by the short certificate. On $|x| \le X$, $|y| \le Y$, let a_-, a_+ bound the nonzero absolute root scores and b_-, b_+ the leaf scores. For $0 \le \alpha \le 2$,

$$\begin{aligned} x' &= (1-\alpha a^2)x + \alpha a^2\left(-D_s - \frac{Ayb}{a} + \frac{\eta - e_A}{a}\right), \\ y' &= (1-\alpha b^2)y + \alpha b^2\frac{\eta - e_B}{b} \quad \text{at a visited leaf.} \end{aligned} \tag{187}$$

If $2a_+^2, 2b_+^2 \le 1$, each line is a convex combination. Since $D_s \le Y^2/8$, the target points remain inside the rectangle when

$$Y^2/8 + Yb_+/a_- + (S+\delta)/a_- \le X, \qquad (S+\delta)/b_- \le Y. \tag{188}$$

An unvisited leaf is unchanged. OPD maps each coordinate interval into itself for steps through four when the teacher logit is contained: its map is increasing, fixes that logit, and moves toward it without crossing. A normalized mixed step through two is a convex combination of these contained RL and OPD maps at the same step. Induction proves containment under every history-dependent choice and stopping rule; no independence of baseline errors is required for this argument.

These tests establish containment, not separation. Learning the geometry or critic would require additional fitting queries charged to the allowance. Exact teacher targets are essential to the unchanged OPD floor.

F Multi-Action Comparison Against Pure Schedules

Explicit softmax envelopes and a noncommuting example extend the analysis beyond binary actions. Their comparison class is distinct from the two-stage feedback result.

F.1 Explicit Multi-Action Envelopes and a Schedule-Tuned Separation

The two-stage construction retains changing teacher visitation. A different setting tests whether the comparison depends on that construction's diagonal value recursion. Here a categorical softmax policy has any finite number of actions; its sampled coefficient matrices need not commute. We derive all envelopes in Theorem A.3 from global softmax inequalities, rather than assuming that a local audit has supplied them. The resulting constants are conservative. A three-action witness nevertheless certifies a mixed advantage against *all deterministic three-step pure schedules* in a continuous cube, not only constant rates.

Model, estimators, and information. Fix $m \ge 2$ and an interior target distribution $p^\star \in \mathbb{R}^m$. Let $p(\theta) = \mathrm{softmax}(\log p^\star + \theta)$ and use the uniform reference, KL coefficient one, and task rewards $r_j = \log(mp^\star_j)$. Then

$$J(\theta) = \sum_j p_j(\theta)r_j - \mathrm{KL}(p(\theta)\|\mathbf{1}/m) = -\mathrm{KL}(p(\theta)\|p^\star), \qquad \mathcal{L}(\theta) = \mathrm{KL}(p(\theta)\|p^\star). \tag{189}$$

The optimum is zero in centered logit coordinates. Constants in all logits are immaterial; one may restrict to $\mathbf{1}^\top\theta = 0$. Put $H = \mathrm{diag}(p^\star) - p^\star(p^\star)^\top$, initialize at $\theta_0 = \varepsilon x_0$, and use teacher $q = p^\star + \varepsilon Hh$, assumed strictly positive for the allowed positive radii. Write $s_j(\theta) = e_j - p(\theta)$. At each update draw $A \sim p(\theta)$ and, independently of A and the past, centered noise η from a fixed law with $|\eta| \le S$. The directions are

$$\widehat{g} = -s_A(\theta)s_A(\theta)^\top\theta + \varepsilon\eta s_A(\theta), \qquad \widehat{u} = q - p(\theta). \tag{190}$$

The teacher supplies its full categorical target on the student's prompt; there is no state-visitation mechanism in this one-state model.

For completeness, the sampled regularized utility is $U_j = r_j - \log(mp_j(\theta))$ and its exact baseline is $J(\theta)$. Writing $Z(\theta) = \log\sum_j p_j^\star e^{\theta_j}$ gives $U_j = -\theta_j + Z(\theta)$ and $J = -p(\theta)^\top\theta + Z(\theta)$. Thus $(U_A - J + \varepsilon\eta)s_A$ is exactly (190), with expectation ∇J. The teacher direction is the negative forward-KL gradient. Rewards and baselines are known in this construction, not estimated without cost. Every recipe receives the same K fresh episodes and paired reward/teacher allowance; unused channels are discarded, not converted to extra episodes.

A deterministic recipe prescribes $\lambda_k, w_k \in [0,1]$ and uses

$$\theta_{k+1} = \theta_k + \lambda_k\{(1-w_k)\widehat{g}_k + w_k\widehat{u}_k\}. \tag{191}$$

RL and OPD have $w_k = 0$ and $w_k = 1$, respectively. Their step schedules may be chosen with full knowledge of the problem. Data-dependent choices of these parameters are outside the reference class below.

Proposition F.1 (Explicit envelopes for any finite action set)**.** *For the preceding model, let $R_0 \ge \|x_0\|$ and $D \ge \max\{3S/2, \|Hh\|\}$. Set $R_k = R_0 + kD$ and*

$$b_k = 3R_k^2/2 + SR_k/2, \qquad l_k = 3R_k/4,$$
$$C_K^{\mathrm{cat}} = \frac{R_K^3}{4} + \sum_{k<K}\left[\frac{R_K b_k}{2} + \frac{l_k}{2}\left\{\frac{R_{k+1}^2}{4} + (K-k-1)DR_{k+1}\right\}\right]. \tag{192}$$

Let $s_j^\star = e_j - p^\star$. The exact random-affine reference draws $A \sim p^\star$ and the same independent noise law, with

$$M_{k,A} = I - \lambda_k\{(1-w_k)s_A^\star(s_A^\star)^\top + w_k H\},$$
$$d_{k,A,\eta} = \lambda_k\{(1-w_k)\eta s_A^\star + w_k Hh\}. \tag{193}$$

Its terminal loss is $x^\top Hx/2$ and its exact risk is $\mathcal{R}_{K,a}^{\mathrm{cat}}$. For every admissible $\varepsilon > 0$, simultaneously over all deterministic recipes in (191)*,*

$$|\mathfrak{R}_{K,a}^{\varepsilon} - \varepsilon^2\mathcal{R}_{K,a}^{\mathrm{cat}}| \le C_K^{\mathrm{cat}}\varepsilon^3. \tag{194}$$

Consequently any reference margin against a prescribed pure schedule class greater than $2C_K^{\mathrm{cat}}\varepsilon$ certifies actual mixed superiority.

In words, this is a class-level way to verify the local theorem's hypotheses: action count, target probabilities, and prescribed schedules may vary, and containment is proved directly. It does not ensure a positive reference gap or a large usable radius. Initial error, teacher error, and external noise scale together; fixed nonvanishing noise is not being discarded.

Proof. We give the derivative, containment, map, law, and value bounds separately. For any probability vector p, put $H(p) = \mathrm{diag}(p) - pp^\top$. It is positive semidefinite since $v^\top H(p)v$ is the variance of v_A. Its maximum absolute row sum is $2p_i(1-p_i) \le 1/2$, hence $\|H(p)\|_{\mathrm{op}} \le 1/2$. Also $\|e_j - p\| \le \sqrt{2} \le 3/2$.

For vectors u, v, differentiation of an expectation under softmax gives

$$u^\top DH(p(\theta))[v]u = \mathbb{E}[(u_A - \mathbb{E}u_A)^2(v_A - \mathbb{E}v_A)]. \tag{195}$$

The absolute last factor is at most $\max_i v_i - \min_i v_i \le \sqrt{2}\|v\|$; the expectation of the squared factor is at most $\|u\|^2/2$. As $DH[v]$ is symmetric, its operator norm is at most $\|v\|/\sqrt{2} \le 3\|v\|/4$.

Integration on any logit segment therefore gives

$$\begin{aligned} \|p(\theta) - p^\star\| &\le \|\theta\|/2, \\ \|p(\theta) - p^\star - H\theta\| &\le 3\|\theta\|^2/8, \\ \sum_j |p_j(\theta) - p_j^\star| &\le 3\|\theta\|/4. \end{aligned} \tag{196}$$

For the last line, along the segment the sum of absolute probability derivatives is $\mathbb{E}|\theta_A - \mathbb{E}\theta_A|$, at most $\sqrt{\theta^\top H(p)\theta} \le \|\theta\|/\sqrt{2}$. Thus the bound has no factor depending on the number of actions.

Next establish containment without a local monotonicity test. Write $\theta = \varepsilon x$ and $\overline{H}_x = \int_0^1 H(p(t\varepsilon x))\,dt$, so $(p(\varepsilon x) - p^\star)/\varepsilon = \overline{H}_x x$. For each sampled outcome the scaled actual update is

$$f^\varepsilon(x) = \big[I - \lambda\{(1-w)s_A(\varepsilon x)s_A(\varepsilon x)^\top + w\overline{H}_x\}\big]x + \lambda\{(1-w)\eta s_A(\varepsilon x) + wHh\}. \tag{197}$$

The matrix subtracted from the identity is positive semidefinite, with operator norm at most $\lambda\{2(1-w) + w/2\} \le 2$. The bracket therefore has operator norm at most one. The final vector has norm at most $\lambda\{(1-w)3S/2 + w\|Hh\|\} \le D$. Hence $\|f^\varepsilon(x)\| \le \|x\| + D$. Exactly the same argument holds for the reference map. Induction proves that both maps send the entire ball of radius R_k into the ball of radius R_{k+1}, uniformly over all recipes. This is a bound on the image norm, not a claim that the nonlinear map is globally Lipschitz with constant one.

On the current ball, abbreviate $s = s_A(\varepsilon x)$ and $s_0 = s_A^\star$. By (196),

$$\|ss^\top - s_0 s_0^\top\| \le (\|s\| + \|s_0\|)\|s - s_0\| \le 3\varepsilon R_k/2. \tag{198}$$

The reward part of the map discrepancy is therefore bounded by $\varepsilon(3R_k^2/2 + SR_k/2)$, and the teacher part by $3\varepsilon R_k^2/8$. Multiplying by λ and the nonnegative fractions, whose sum is one, bounds the total by εb_k. The action-law discrepancy is at most εl_k by the last line of (196). Appending the common independent noise law does not change that bound. Finite noise laws fit the displayed theorem directly; for a general bounded noise law the same sum over actions is integrated against that common probability measure.

For the terminal loss, differentiation of $\mathcal{L}(\theta) = p(\theta)^\top\theta - Z(\theta)$ gives $\nabla\mathcal{L}(\theta) = H(p(\theta))\theta$. Integration along the segment, followed by the derivative bound above, gives

$$\mathcal{L}(\theta) = \int_0^1 t\theta^\top H(p(t\theta))\theta\,dt, \tag{199}$$

$$|\mathcal{L}(\theta) - \theta^\top H\theta/2| \le \int_0^1 t(3t/4)\|\theta\|^3\,dt = \|\theta\|^3/4. \tag{200}$$

Thus the terminal scaled error is at most $\varepsilon R_K^3/4$.

Finally let (P_k, p_k, γ_k) denote the reference value coefficients. The reference matrices satisfy $\|M_k\| \le 1$ and drifts satisfy $\|d_k\| \le D$. Lemma A.1 implies

$$0 \preceq P_k, \quad \|P_k\| \le 1/2, \qquad \|p_k\| \le (K-k)D/2. \tag{201}$$

Indeed, the quadratic bound follows by backward induction, and $\|p_k\| \le D/2 + \|p_{k+1}\|$ proves the linear bound from $p_K = 0$. The weights in (45) are consequently

$$G_k \le R_K/2, \qquad O_k \le R_{k+1}^2/4 + (K-k-1)DR_{k+1}. \tag{202}$$

Substitution into Theorem A.3 proves (194). All bounds are uniform over the full deterministic recipe class. Subtracting the two error bounds proves the margin condition, including for continuous schedule classes. □

A noncommuting example with continuously tuned schedules. Take

$$p^\star = (1/5, 3/10, 1/2), \quad x_0 = (1, -1, 0), \quad h = (2/5, 2/5, -4/5), \quad \eta \in \{-1/2, 1/2\} \tag{203}$$

with equal noise probabilities, and $K = 3$. Task rewards are nonconstant. The score outer products for different actions do not commute; neither a scalar reduction nor the two-stage diagonal backward recursion is used. The fixed mixture has $w_k = 1/2$ and $\lambda_k = 9/10$ at every update.

Proposition F.2 (Mixed advantage over every three-step pure schedule)**.** *For* (203)*, every* $0 < \varepsilon \le 10^{-4}$*, and each* $a \in \{\mathsf{R}, \mathsf{D}\}$*,*

$$\inf_{(\lambda_0,\lambda_1,\lambda_2)\in[0,1]^3} \mathfrak{R}^{\varepsilon}_{3,a}(\lambda_0,\lambda_1,\lambda_2) - \mathfrak{R}^{\varepsilon}_{3,\mathsf{M}} > .0221\varepsilon^2. \tag{204}$$

For the same model and recipe classes, the computer-assisted refinement starting on page 56 gives

$$\inf_{[0,1]^3} \mathfrak{R}^{\varepsilon}_{3,a} - \mathfrak{R}^{\varepsilon}_{3,\mathsf{M}} > .01\varepsilon^2, \qquad 0 < \varepsilon \le 1/20. \tag{205}$$

This separation extends the analysis to multi-action policies with noncommuting sampled matrices. It covers prescribed three-step schedules and gives a sharper finite-radius guarantee than the analytic bound above. It is not a claim about learned baselines, adaptive schedules, large training errors, or economical unknown-model selection. The numerical witness was chosen after exploratory calculations; its proof covers the complete displayed classes and does not infer coverage from the exploratory grid. The teacher itself has lower regularized regret than the initial student throughout the certified region, so the separation does not require an initially worse teacher.

Proof. We separate the exact reference inequality from its nonlinear transfer. Every reference map is affine in its own step. Its final quadratic loss therefore has degree at most two in each of the three step variables, even after averaging the independently sampled actions and noise. All coefficients are rational for (203).

For each pure method, evaluate the exact reference second-moment recursion at the 27 points $\{0, 1/2, 1\}^3$. These are interpolation nodes for the known degree, not test points used to extrapolate an inequality. For a one-variable quadratic with values $(v_0, v_{1/2}, v_1)$, its degree-two Bernstein coefficients on $[0, 1]$ are

$$(v_0,\ 2v_{1/2} - (v_0 + v_1)/2,\ v_1). \tag{206}$$

Applying this linear transform along each coordinate gives the exact 27-coefficient tensor. The tensor-product Bernstein basis is nonnegative and sums to one on the cube; the minimum coefficient is therefore a lower bound on the polynomial throughout the cube.

To sharpen the bound, split each coordinate three times at its midpoint, producing $8^3 = 512$ closed boxes covering $[0, 1]^3$. For coefficients (a, b, c), midpoint subdivision produces

$$\left(a, \frac{a+b}{2}, \frac{a+2b+c}{4}\right), \qquad \left(\frac{a+2b+c}{4}, \frac{b+c}{2}, c\right). \tag{207}$$

Apply this identity to every one-dimensional slice of each tensor. All operations are exact rational additions and divisions; taking the minimum of every coefficient in every box proves the following bounds:

$$\begin{aligned} \inf_{[0,1]^3} \mathcal{R}^{\mathrm{cat}}_{3,R} &> .13230711, \\ \inf_{[0,1]^3} \mathcal{R}^{\mathrm{cat}}_{3,D} &> .12910169, \\ \mathcal{R}^{\mathrm{cat}}_{3,M} &< .09244978. \end{aligned} \tag{208}$$

Exact rational subdivision gives bounds on every box; the displayed decimals are rounded outwards. In particular, no continuous pure schedule is omitted, and no optimizer's local minimum is assumed to be global.

Now $\|x_0\| = \sqrt{2} < 3/2$ and $Hh = (3/25, 9/50, -3/10)$ has norm below $1/2$. We may take $R_0 = 3/2$ and $D = 3/4$, giving $R_k = 3/2 + 3k/4$. Direct rational substitution in (192) yields

$$C_3^{\mathrm{cat}} = 72.5888671875. \tag{209}$$

The teacher $p^\star + \varepsilon Hh$ is strictly positive on the stated radius interval. Proposition F.1 applies simultaneously to the mixture and every pure schedule. For OPD, the scaled nonlinear gap is strictly greater than

$$.12910169 - .09244978 - 2(10^{-4})(72.5888671875) = .0221341365625 > .0221. \tag{210}$$

The RL lower bound is larger, so the same conclusion follows for it. Multiplying by ε^2 proves (204) uniformly on its stated radius interval. The larger-radius bound is proved in the next subsection.

To check teacher quality, put $u = Hh$. Here $|u_i|/p_i^\star = 3/5$ and $\sum_i u_i^2/p_i^\star = 9/25$. Integrating the Hessian of KL along the probability segment $p^\star + tu$ yields

$$\mathrm{KL}(q\|p^\star) \le \frac{.18\varepsilon^2}{1 - .6\varepsilon} < .181\varepsilon^2. \tag{211}$$

The initial quadratic risk is $x_0^\top H x_0/2 = .245$. The terminal inequality proved above, applied at initialization, gives $\mathcal{L}(\varepsilon x_0) \ge \varepsilon^2\{.245 - \varepsilon(3/2)^3/4\} > .244\varepsilon^2$. These strict inequalities prove the initial teacher advantage on the smaller interval. At every radius up to $1/20$, the same two estimates give a teacher risk below $.186\varepsilon^2$ and an initial student risk above $.202\varepsilon^2$, so the advantage also holds on the enlarged interval. □

F.2 A Sharper Finite-Radius Certificate for All Pure Schedules

The preceding dimension-free constant bounds every intermediate state by a common growing ball. It proves the separation analytically, but gives a small usable radius. For the same three-action model and three-update budget, we now bound the third radius derivative of the actual expected regret directly. This retains the exact reference comparison while reducing the nonlinear error. The resulting certificate covers every radius in $(0, 1/20]$ and every pure schedule in $[0, 1]^3$, not only selected radii or schedule points. It uses full short-horizon trees and rigorous interval arithmetic; unlike the two-stage structural construction, its cost is exponential in the number of updates.

An exact finite Taylor inequality. Fix a deterministic schedule a and let $F_a(e) = \mathfrak{R}^e_{3,a}$ for the model in (203). This is a smooth function of the real radius e on the closed interval $[0, 1/20]$. Indeed, it is a finite sum over action/noise histories of smooth probabilities times a smooth terminal KL loss. Softmax denominators are strictly positive, and all teacher probabilities remain positive on this interval.

At zero radius every iterate and loss is zero. Differentiating the update at zero, while holding the action/noise labels fixed, gives exactly the random-affine recursion (193). The terminal loss has zero value and gradient there, and Hessian H. Thus, writing $R_a = \mathcal{R}^{\mathrm{cat}}_{3,a}$,

$$F_a(0) = F_a'(0) = 0, \qquad F_a''(0)/2 = R_a. \tag{212}$$

To justify the last equality including the changing sampling law, write $F_a(e) = \sum_\omega P_a(e,\omega) L_a(e,\omega)$ over the finitely many histories. Both $L_a(0,\omega)$ and $L_a'(0,\omega)$ vanish. In the second product derivative, the terms multiplying P_a' and P_a'' therefore vanish; the remaining term is the quadratic loss of the first iterate derivative, averaged under the reference history probabilities. This is $2R_a$, including all sampled coefficient matrices.

Taylor's integral identity, with no limiting step, now reads

$$F_a(e) = e^2 R_a + \frac{1}{2}\int_0^e (e-t)^2 F_a'''(t)\, dt. \tag{213}$$

If $d_{a,-} \le F_a'''(t)/6 \le d_{a,+}$ on the entire interval $0 \le t \le e_0$, then for every $0 < e \le e_0$,

$$\begin{aligned} R_a + e d_{a,-} \le F_a(e)/e^2 \le R_a + e d_{a,+}, \\ R_a + e_0 \min(0, d_{a,-}) \le F_a(e)/e^2 \le R_a + e_0 \max(0, d_{a,+}). \end{aligned} \tag{214}$$

The factor in the first line follows because $3\int_0^e (e-t)^2\, dt = e^3$. The second line is uniform down to every positive radius; no division of an interval containing zero is performed.

In plain language, the exact reference risk is corrected by a verified finite remainder. Different schedule boxes receive different corrections. Small approximation error is demonstrated by the calculation, not assumed from proximity alone.

Enclosing derivatives, not fitting a polynomial. The verifier evaluates the full nonlinear expected-risk recursion with four-coefficient real-ball power series (Johansson, 2017). For a radius interval

I, its input coefficients are $(I, 1, 0, 0)$; the formal variable represents an increment about any radius in I. Step intervals are constant coefficients. Addition, multiplication, division, and exponentiation propagate the first four Taylor coefficients with outward rounding. The coefficient of degree three therefore encloses $F_a'''(t)/6$ for every $t \in I$ and every step in the box. It is not a third-degree fit to finitely many evaluations. Dependence lost by interval arithmetic can widen the enclosure but cannot exclude a possible value. Every division requires a verified positive softmax denominator; a nonfinite enclosure is not a certificate.

For additional detail, at a node with logit series $\Theta(z)$ let $p(z) = \mathrm{softmax}(\log p^\star + \Theta(z))$. A child with action j and noise sign $b \in \{-1, 1\}$ has logit series

$$\begin{aligned}\mathcal{A}_j^b(z) &= -\Theta_j(z) + p(z)^\top \Theta(z) + \tfrac{1}{2} b(e + z),\\ \Theta_j^b(z) &= \Theta(z) + \lambda(1 - w)\mathcal{A}_j^b(z)\{e_j - p(z)\} + \lambda w\{q(e + z) - p(z)\}.\end{aligned} \tag{215}$$

The node value is the sum of its child values weighted by $p_j(z)/2$. This differentiates the sampling probabilities as well as the update maps. For pure OPD, all children have the same state and can be combined into one branch using $\sum_j p_j(z) = 1$; for a zero step, the unchanged state can likewise be propagated directly. These exact simplifications do not omit any probability derivative.

At a terminal node the loss is $\Theta(z)^\top p(z) - \log Z(z)$, where $Z(z) = \sum_j p_j^\star \exp \Theta_j(z)$. To avoid an unnecessarily wide subtraction, denote the coefficient vectors by Θ_i, p_i. The exact identity $(\log Z)' = \Theta'^\top p$ gives the loss coefficients, for $n = 1, 2, 3$,

$$[z^n]\mathcal{L}(\Theta(z)) = \Theta_0^\top p_n + \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) \Theta_i^\top p_{n-i}. \tag{216}$$

The degree-zero coefficient is evaluated from the original loss. To check (216), the product coefficient is $\sum_{i=0}^{n} \Theta_i^\top p_{n-i}$, whereas the logarithm coefficient is $\sum_{i=1}^{n} (i/n)\Theta_i^\top p_{n-i}$. Subtracting cancels the highest logit coefficient exactly before interval evaluation. This is an algebraic tightening, not a different loss or an omitted remainder.

Complete schedule and radius coverage. Set $e_0 = 1/20$ and split $[0, e_0]$ into eight closed intervals of width $1/160$. On each schedule box, take the smallest lower and largest upper third-coefficient endpoints across all eight intervals. These are valid $d_{a,-}, d_{a,+}$ in (214). The mixed schedule is the fixed point $(.9, .9, .9)$ with fraction $1/2$.

For each pure method begin with 4^3 boxes covering $[0, 1]^3$ and the exact reference Bernstein tensors from the preceding proof. On a box B let r_B be its minimum reference Bernstein coefficient. Then

$$\ell_B = r_B + e_0 \min(0, d_{B,-}), \qquad u_M = R_M + e_0 \max(0, d_{M,+}) \tag{217}$$

give a uniform pure lower bound and mixed upper bound on scaled actual regret. If $\ell_B - u_M$ does not exceed .01, split the box at the midpoint of its longest side, breaking ties by coordinate order, and apply exact Bernstein subdivision to its reference polynomial. Both children are retained. Refinement stops after at most nine additional bisections along a branch; a remaining unresolved leaf would prevent the claim.

Every retained leaf records its base box, binary subdivision path, exact endpoints, reference bound, and all eight derivative enclosures. The verification checks that the paths in each base box are distinct and prefix-free and that $\sum_{\text{leaves}} 2^{-\text{path length}} = 1$. These conditions prove complete coverage: each split partitions its parent into two equal-volume boxes, prefix-freeness makes interiors disjoint, and the finite union of closed leaves with full volume cannot omit an open portion of the base box. The verifier also reconstructs every box and Bernstein tensor from its path and checks the complete radius partition. Consequently no schedule, including boundary schedules, is discarded.

Verified finite-radius inequality. At 192-bit precision the complete calculation terminates with 1,084 RL leaves and 81 OPD leaves. Taking the smallest lower bound over each full partition and rounding conservatively gives

$$\inf_{[0,1]^3} \frac{F_R(e)}{e^2} > .10647256, \qquad \inf_{[0,1]^3} \frac{F_D(e)}{e^2} > .10664086, \qquad \frac{F_M(e)}{e^2} < .09643646 \tag{218}$$

simultaneously for every $0 < e \le 1/20$. The displayed pure lower and mixed upper endpoints leave gaps of .01003610 and .01020440, respectively. Both exceed .01, which proves (205).

The computer-assisted proof encloses every schedule box and radius interval, with all bounds checked at 256-bit precision. Independent raw-utility trees and high-precision derivatives check the nonlinear risk and derivative implementation, including the sampling-law terms and the terminal cancellation identity. Those point checks are diagnostics; the complete interval coverage and the finite Taylor proof establish the uniform inequality. The radius and refinement rule were chosen during development of this constructive witness, not in a prospective empirical study.

This enlarges the certified radius by a factor of 500 without changing the teacher, the training algorithm, or the pure comparator class. It remains a local, known-model, three-update result with jointly scaled initialization, teacher error, and reward noise. It is not a guarantee for adaptive schedules, fixed nonvanishing noise at arbitrarily small radii, or a computationally economical long-horizon selector.

G Supporting Enclosures for the Numerical Certificates

The direct certificate uses reachable boxes, moment bounds, and valid intersections of reference enclosures. The additive and sampled-coefficient constructions provide complementary bounds whose intersection sharpens the final comparison.

G.1 Structural Comparison with Learner-Dependent Visitation

We bound the two-stage model without enumerating trajectories. In the parameterized family of Appendix A.2, the initialization, two optimal logits, two teacher intercepts, noise level, step size, and horizon may vary subject to the explicit tests below. The model and its statewise baselines remain known. Its changing second-state visitation and nonconservative teacher field are retained; no frozen-occupancy approximation is imposed on the actual run.

Conditional moments that retain shared-rollout dependence. Suppress the method index and write $(r,t) = (1,0),(0,1),(1,c)$. Let $p_0 = p_\star$, $s_0 = s_\star$, $h_p = p_0(1-p_0)$, $h_s = s_0(1-s_0)$, and $H = \mathrm{diag}(h_p, p_0 h_s)$. The affine reference has diagonal transition entries $a_i = 1 - \alpha(r+t)H_{ii}$ and drift $d_i = \alpha t H_{ii} h_i$. Its covariance Q is (26). The actual mean is

$$\mu(x,y) = \begin{pmatrix} x + \alpha\{-r v_p (x + D_s) + t(q_A - p)\} \\ y + \alpha p\{-r v_s y + t(q_B - s)\} \end{pmatrix}. \tag{219}$$

For its conditional covariance C, the diagonal entries are exactly

$$\begin{aligned} C_{11}/\alpha^2 &= r\{(x+D_s)^2 v_p (1-2p)^2 + y^2 p(1-p)^2 v_s + \sigma^2 v_p\}, \\ C_{22}/\alpha^2 &= pr\{y^2 v_s (1-2s)^2 + \sigma^2 v_s\} + v_p\{-r v_s y + t(q_B - s)\}^2. \end{aligned} \tag{220}$$

The second square includes reward–teacher covariance; it is not the sum of two independent channel variances. Off-diagonal covariance can also be nonzero. Because every reference value Hessian is diagonal, only these two entries enter its exact trace residual.

To verify (220), condition first on the root action. For a Bernoulli score $a = A - p$, $\mathrm{Var}(a^2) = v_p(1-2p)^2$. In the first reward coordinate, the centered a^2 term is uncorrelated with $Ayab$ because the leaf score $b = B - s$ has conditional mean zero. Also $\mathbb{E}(Aa^2b^2) = p(1-p)^2 v_s$. The independent centered sign removes the remaining cross terms and contributes $\sigma^2 v_p$. The root teacher update is deterministic. At the second state, the conditional mean direction given $A = 1$ is $-r v_s y + t(q_B - s)$, and its conditional variance is $r\{y^2 v_s (1-2s)^2 + \sigma^2 v_s\}$. The law of total variance gives the second line, including its visitation variance. These identities hold for $r \in \{0,1\}$, as used here.

Analytic reachable boxes. Start with $l_0 = u_0 = \theta_0$ and componentwise second moments $M_{i,0} = \theta_{i,0}^2$. At step k put $R_i = \max\{|l_{i,k}|, |u_{i,k}|\}$ and $S_i = \sqrt{M_{i,k}}$; indices $1,2$ denote the root and leaf. On the symmetric intervals $|x| \le R_1$, $|y| \le R_2$, calculate

$$V_p = \max v_p, \quad L_p = \max |v_p'|, \quad T_p = \max |1-2p|, \qquad (V_s, L_s, T_s) \text{ analogously}, \tag{221}$$

and $\overline{p} = \max p$, $B_p = \max p(1-p)^2$. For V_p, maximize $p(1-p)$ at the probability endpoints and at $1/2$ when included. For L_p, maximize $p(1-p)|1-2p|$ at the endpoints and the included points $(3 \pm \sqrt{3})/6$; T_p uses the endpoints. The leaf constants use the same rules. For B_p, use the endpoints and $p = 1/3$ when included. Thus every constant requires only a fixed number of scalar calculations.

Let $F_{A,B,E}$ be the sample update defined by (25) and (23). Require $R_2 \le 1$ when $r = 1$, and the two sufficient tests

$$\begin{aligned} 1 - \alpha\left[r\{1 + V_p(2(R_1 + V_s R_2^2/2) + R_2 + \sigma)\} + tV_p\right] &\ge 0, \\ 1 - \alpha\{r(1 + 2R_2 V_s + \sigma V_s) + tV_s\} &\ge 0. \end{aligned} \tag{222}$$

Form the at most six parameter points

$$\mathcal{E}_k = \{l_{1,k}, u_{1,k}\} \times \{l_{2,k}, u_{2,k}, \mathrm{clip}(0; [l_{2,k}, u_{2,k}])\}. \tag{223}$$

Set $l_{i,k+1}$ and $u_{i,k+1}$ to the minimum and maximum of $F_{A,B,E,i}$ over these points and all local action/sign outcomes. These are exact extrema over the current box under (222), not a sampled state net.

Here is the derivative argument. Put $a = A - p$ and $b = B - s$. Direct differentiation gives

$$\begin{aligned} \partial_x F_1 &= 1 + \alpha\{r[-a^2 + v_p\{2(x + D_s)a + Ayb - \sigma E\}] - tv_p\}, \\ \partial_y F_2 &= 1 + \alpha A\{r[-b^2 + 2ybv_s - \sigma E v_s] - tv_s\}, \\ \partial_y F_1 &= \alpha r a\{-Ab + pv_s y\}, \qquad \partial_x F_2 = 0. \end{aligned} \tag{224}$$

Since $D_s \le V_s y^2/2$ and both scores have magnitude at most one, the tests make each map nondecreasing in its own coordinate. For $A = 0$, $\partial_y F_1 = -\alpha r p^2 v_s y$, so its only possible interior extremum is at zero. For $A = 1, B = 1$ this derivative is $-\alpha r(1-p)(1-s)(1-psy) \le 0$ when $|y| \le 1$; for $A = 1, B = 0$ it is $\alpha r(1-p)s\{1 + p(1-s)y\} \ge 0$. Thus the displayed points contain every needed extremum. If $r = 0$, the root map is independent of y and the restriction $R_2 \le 1$ is unnecessary. Induction proves almost-sure containment in the propagated boxes, irrespective of the policy-dependent probabilities of the maps.

Moment and residual bounds. All quantities in the remainder of this construction use the current box. Define nonnegative coefficients

$$\begin{aligned} A_1 &= (r + t/2)L_p, & B_1 &= rV_pV_s/2, & & \\ A_2 &= th_s|h_2|V_p, & B_2 &= (r+t)h_sV_p, & D_2 &= \overline{p}(r + t/2)L_s. \end{aligned} \tag{225}$$

The following vectors bound first absolute moments and root-mean-square errors of the actual next mean relative to its reference:

$$\begin{aligned} \eta_1 &= \alpha(A_1M_1 + B_1M_2), & \eta_2 &= \alpha(A_2S_1 + B_2S_1S_2 + D_2M_2), \\ \beta_1 &= \alpha(A_1R_1S_1 + B_1R_2S_2), & \beta_2 &= \alpha\{A_2S_1 + B_2\min(R_1S_2, R_2S_1) + D_2R_2S_2\}. \end{aligned} \tag{226}$$

Specifically $\mathbb{E}|e_i| \le \eta_i$ and $\mathbb{E}e_i^2 \le \beta_i^2$. The A_2 term charges the nonconservative teacher's linear dependence on root visitation; it vanishes for an exactly matched leaf teacher, not by assuming a fixed visitation distribution.

For covariance envelopes, put $m_0 = th_sh_2$, $L_m = (r+t)V_s$, and

$$\begin{aligned} N_1 &= r\{V_pT_p^2(S_1 + V_sR_2S_2/2)^2 + B_pV_sM_2\}, \\ N_2 &= r\overline{p}V_sT_s^2M_2. \end{aligned} \tag{227}$$

Valid bounds $\mathbb{E}|C_{ii} - Q_{ii}| \le \omega_i$ and $\mathbb{E}C_{ii} \le \upsilon_i$ are

$$\begin{aligned} \omega_1 &= \alpha^2(N_1 + r\sigma^2L_pS_1), \\ \omega_2 &= \alpha^2\left[N_2 + r\sigma^2(h_sV_pS_1 + \overline{p}L_sS_2) + L_pm_0^2S_1 + V_p(2|m_0|L_mS_2 + L_m^2M_2)\right], \\ \upsilon_1 &= \alpha^2(N_1 + r\sigma^2V_p), \\ \upsilon_2 &= \alpha^2\{N_2 + r\sigma^2\overline{p}V_s + V_p(|m_0| + L_mS_2)^2\}. \end{aligned} \tag{228}$$

Propagate second moments by

$$M_{i,k+1} = \min\{R_{i,k+1}^2, (|a_i|S_i + |d_i| + \beta_i)^2 + \upsilon_i\}. \tag{229}$$

Let $P_{i,k+1}$ and $p_{i,k+1}$ denote the diagonal quadratic coefficient and linear coefficient of the reference value. Define

$$\begin{aligned} G_{i,k} &= P_{i,k+1}(|a_i|R_i + |d_i|) + |p_{i,k+1}|, \\ \delta_k &= \sum_{i=1}^{2}\{G_{i,k}\eta_i + \tfrac{1}{2}P_{i,k+1}(\beta_i^2 + \omega_i)\}, \qquad \Delta_K = \sum_{k<K}\delta_k. \end{aligned} \tag{230}$$

At the terminal box, use $\overline{M}_i = \min\{M_{i,K}, 2(\mathcal{R}_K + \Delta_K)/H_{ii}\}$ and

$$\nu_K = \frac{L_p R_1 \overline{M}_1}{3} + \frac{\overline{p} L_s R_2 \overline{M}_2}{3} + \frac{V_p h_s}{2}\min\{R_1\overline{M}_2, R_2\sqrt{\overline{M}_1\overline{M}_2}\}. \tag{231}$$

Proposition G.1 (Linear-horizon comparison with changing teacher visitation)**.** *For the two-stage family, assume finite initial logits, $\alpha > 0$, $c, \sigma \geq 0$, valid teacher probabilities, and a finite integer $K \geq 0$. If the stated box tests hold at every step, then the propagated boxes contain every actual iterate and $\mathbb{E}\theta_{i,k}^2 \leq M_{i,k}$. Moreover,*

$$|\mathfrak{R}_K - \mathcal{R}_K| \leq \rho_K^{\mathrm{branch}} := \Delta_K + \nu_K. \tag{232}$$

The construction uses a number of scalar arithmetic operations linear in K, including the backward reference values. A reference gap exceeding the two corresponding radii certifies actual expected reward-risk ordering.

In plain language, the calculation retains both ways that the root policy changes teacher learning: it changes the average leaf update and its sampling variance. A box controls all possible states; moment bounds avoid treating every extreme history as likely. This provides an informative comparison without tracing the exponentially many possible training paths. It does not learn the model, guarantee that every parameter choice is certifiable, or optimize the pure methods' learning rates.

Proof. Containment was established using (224). We prove the remaining bounds inductively, beginning with $M_{i,0} = \theta_{i,0}^2$. Subtract the reference mean from (219). With $\delta p = p - p_0$, $\delta s = s - s_0$, the two errors are

$$\begin{aligned} e_1/\alpha &= -r(v_p - h_p)x - r v_p D_s - t(\delta p - h_p x), \\ e_2/\alpha &= t h_s h_2 \delta p - (r+t)h_s \delta p\, y - p\{r(v_s - h_s)y + t(\delta s - h_s y)\}. \end{aligned} \tag{233}$$

The derivative envelopes imply $|\delta p| \leq V_p|x|$, $|v_p - h_p| \leq L_p|x|$, $|\delta p - h_p x| \leq L_p x^2/2$, with analogous leaf bounds; also $D_s \leq V_s y^2/2$. Therefore

$$|e_1| \leq \alpha(A_1 x^2 + B_1 y^2), \qquad |e_2| \leq \alpha(A_2|x| + B_2|xy| + D_2 y^2). \tag{234}$$

Cauchy–Schwarz gives $\mathbb{E}|xy| \leq S_1 S_2$. For root-mean-square bounds, containment gives $\|x^2\|_{L^2} \leq R_1 S_1$, $\|y^2\|_{L^2} \leq R_2 S_2$, and $\|xy\|_{L^2} \leq \min(R_1 S_2, R_2 S_1)$. Apply Minkowski to the displayed pointwise errors to obtain (226).

For the first covariance entry, $\|x + D_s\|_{L^2} \leq S_1 + V_s R_2 S_2/2$. The two nonnoise terms of (220) are nonnegative and have expectation at most N_1. The remaining deviation from reference is bounded using $\mathbb{E}|v_p - h_p| \leq L_p S_1$, proving ω_1. Using $v_p \leq V_p$ instead gives υ_1.

For the second entry, write its conditional-on-visitation mean as $m(y) = m_0 - r v_s y - t\delta s$. Then $|m(y) - m_0| \leq L_m|y|$. The action-induced leaf variance is bounded by N_2 and

$$\begin{aligned} \mathbb{E}|p v_s - p_0 h_s| &\leq h_s V_p S_1 + \overline{p} L_s S_2, \\ \mathbb{E}|v_p m(y)^2 - h_p m_0^2| &\leq L_p m_0^2 S_1 + V_p(2|m_0|L_m S_2 + L_m^2 M_2). \end{aligned} \tag{235}$$

The second inequality follows by adding and subtracting $v_p m_0^2$, expanding $m(y)^2 - m_0^2$, and taking absolute values. These bounds give ω_2; bounding p, v_s, v_p by their box maxima and applying Minkowski to $m(y)$ gives υ_2. This proves (228) without discarding any same-rollout covariance before bounding it.

Since $\mu_i = a_i\theta_i + d_i + e_i$, Minkowski gives $\|\mu_i\|_{L^2} \le |a_i|S_i + |d_i| + \beta_i$. The law of total second moments and the covariance upper bound yield the second argument of the minimum in (229); containment supplies the first. This completes the moment induction.

The reference Hessians are diagonal with nonnegative entries. The absolute value of the ith reference gradient at $A\theta + d$ is at most $G_{i,k}$. Applying the first absolute moment, squared mean-error, and diagonal covariance bounds to (34) gives $\mathbb{E}|Z_k| \le \delta_k$. Thus

$$\tfrac{1}{2}\sum_i H_{ii}\mathbb{E}\theta_{i,K}^2 \le \mathcal{R}_K + \Delta_K. \tag{236}$$

Each summand is nonnegative, so this additionally bounds each terminal moment by $2(\mathcal{R}_K + \Delta_K)/H_{ii}$, justifying $\overline{M}_i$ without using the target-risk conclusion.

Finally decompose the terminal target error as

$$\begin{aligned}\mathcal{L}(x,y) - \tfrac{1}{2}h_p x^2 - \tfrac{1}{2}p_0 h_s y^2 = \mathrm{KL}(p\|p_0) - \tfrac{1}{2}h_p x^2 \\ + p\{D_s - \tfrac{1}{2}h_s y^2\} + \tfrac{1}{2}h_s(p-p_0)y^2.\end{aligned} \tag{237}$$

Integration of the respective derivatives $v_p x$ and $v_s y$ bounds the first two absolute remainders by $L_p|x|^3/3$ and $\overline{p}L_s|y|^3/3$. The last is at most $V_p h_s |x|y^2/2$. Containment and Cauchy–Schwarz give

$$\mathbb{E}|x|y^2 \le \min\{R_1\mathbb{E}y^2, R_2\sqrt{\mathbb{E}x^2\mathbb{E}y^2}\}. \tag{238}$$

Substitute the terminal moment bounds to obtain (231). Theorem A.2 now proves the target-risk radius. Each stage evaluates a fixed number of box points and at most six outcomes, plus scalar moment operations. The backward reference also takes linear work. This proves the computational claim. □

G.2 Nonlinear Comparison Against Continuously Tuned Pure Methods

A prescribed mixture can win merely because its effective step is larger. We now give each pure method its own constant step in a continuous interval, while fixing one *normalized* mixed recipe. The result uses the same two-stage nonlinear training kernel, including learner-dependent teacher visitation, with explicit teacher and noise uncertainty. No claim about arbitrary step schedules or unbounded step sizes is made.

Recipe classes and uncertainty. Use the known reward geometry of Proposition G.1, finite horizon K, deterministic initialization, teacher rectangle $\mathcal{H}$, and independent centered reward noise supported on $[-S, S]$ with a common variance $v \in [v_-, v_+]$. Sample maps are affine in the external noise, so support extrema still occur at $\pm S$. Centering and independence remove its linear cross terms; conditional second moments use v, not S^2. This justifies applying the preceding bounds to arbitrary noise laws with this support and variance. Both pure methods may choose any constant step $\alpha \in [0, L]$, separately. The mixed update is

$$\theta_{k+1} = \theta_k + \lambda\{(1-w)\widehat{g}_k + w\widehat{u}_k\}, \qquad 0 < \lambda \le L, \quad 0 < w < 1. \tag{239}$$

Thus its total effective scale lies in the same interval as the pure steps. The existing formulas apply with $\alpha = \lambda(1-w)$ and $c = w/(1-w)$; their indicator r remains in $\{0, 1\}$. Substituting a fractional r into formulas that used $r^2 = r$ would not be valid. All methods retain the same allocated training information.

For every candidate write $(r, t) = (1, 0), (0, 1), (1, c)$ as before. Partition a pure step interval into finitely many closed intervals $I = [a, b]$ whose union is $[0, L]$. The mixed candidate needs only the singleton interval $\{\lambda(1-w)\}$. We give bounds valid at *every* step in each interval, not only at the partition endpoints.

Reference risks over a whole step interval. For scalar intervals use the usual inclusion-preserving arithmetic: the product takes the smallest and largest of the four endpoint products; a square has lower endpoint zero when the interval contains zero and otherwise the smaller endpoint square. Addition is endpointwise. Apply these operations separately to each coordinate. Define

$$\begin{aligned}\mathcal{A}_i &= 1 - I(r+t)H_i, & \mathcal{D}_i &= ItH_i[h_i^-, h_i^+], \\ \mathcal{Q}_1 &= I^2 r[v_-, v_+]H_1, & \mathcal{Q}_2 &= I^2\{r[v_-, v_+]H_2 + t^2 H_1 h_s^2[h_2^-, h_2^+]^2\}.\end{aligned} \tag{240}$$

Start $\mathcal{M}_{i,0} = \{\theta_{i,0}\}$ and $\mathcal{V}_{i,0} = \{0\}$, and propagate for exactly K updates

$$\mathcal{M}_{i,k+1} = \mathcal{A}_i \mathcal{M}_{i,k} + \mathcal{D}_i, \qquad \mathcal{V}_{i,k+1} = \mathcal{A}_i^2 \mathcal{V}_{i,k} + \mathcal{Q}_i. \tag{241}$$

Let $[\underline{R}_I, \overline{R}_I]$ be the interval $\frac{1}{2}\sum_i H_i(\mathcal{M}_{i,K}^2 + \mathcal{V}_{i,K})$. These endpoints enclose every reference risk at constant $\alpha \in I$ and every admissible teacher and variance. Repeated occurrences of the same input are not treated as statistically independent. Interval arithmetic simply discards some algebraic dependence, which may enlarge the enclosure but cannot exclude a valid constant-input trajectory.

Uniform nonlinear error bounds. Use common boxes for all inputs and all steps in I. Local score bounds sharpen the sufficient monotonicity test. On the current symmetric box define $Q_p = \max\{\max p, 1 - \min p\}$ and $Q_s = \max\{\max s, 1 - \min s\}$, and require, with $R_2 \le 1$ if $r = 1$,

$$\begin{aligned} 1 - b\left[r\{Q_p^2 + 2(R_1 + V_s R_2^2/2)Q_p V_p + R_2 V_p Q_s + S V_p\} + t V_p\right] \ge 0, \\ 1 - b\{r(Q_s^2 + 2R_2 Q_s V_s + S V_s) + t V_s\} \ge 0. \end{aligned} \tag{242}$$

Evaluate each sample map at the state points $\mathcal{E}_k$, both step endpoints, the four teacher corners, and action/noise-support endpoints. Their coordinate extrema form the next box. These are uniform extrema, not evaluations of training risk on a rate grid.

Let $A_i^+ = \max_{z \in \mathcal{A}_i} |z|$, $D_i^+ = \max_{z \in \mathcal{D}_i} |z|$, and $\bar{h}_i = \max(|h_i^-|, |h_i^+|)$. In the nonnegative mean and covariance envelopes (226)–(228), use $\alpha = b$, $|h_2| = \bar{h}_2$, and $\sigma^2 = v_+$. The support bound in the box tests remains S, not $\sqrt{v_+}$. Call these uniform envelopes $\eta_i, \beta_i, \omega_i, \upsilon_i$. Starting with $M_{i,0} = \theta_{i,0}^2$, propagate

$$M_{i,k+1} = \min\{R_{i,k+1}^2, (A_i^+ \sqrt{M_{i,k}} + D_i^+ + \beta_i)^2 + \upsilon_i\}. \tag{243}$$

To bound the backward reference coefficients, set $\overline{P}_{i,K} = H_i$, $\overline{p}_{i,K} = 0$ and use

$$\overline{P}_{i,k} = (A_i^+)^2 \overline{P}_{i,k+1}, \qquad \overline{p}_{i,k} = A_i^+(\overline{P}_{i,k+1} D_i^+ + \overline{p}_{i,k+1}). \tag{244}$$

In the residual sum use

$$\begin{aligned} \overline{G}_{i,k} &= \overline{P}_{i,k+1}(A_i^+ R_{i,k} + D_i^+) + \overline{p}_{i,k+1}, \\ \Delta_I &= \sum_{k<K} \sum_i \left\{\overline{G}_{i,k}\eta_i + \tfrac{1}{2}\overline{P}_{i,k+1}(\beta_i^2 + \omega_i)\right\}. \end{aligned} \tag{245}$$

Apply the existing terminal Taylor bound (231) on the terminal common box with $\overline{M}_i = \min\{M_{i,K}, 2(\overline{R}_I + \Delta_I)/H_i\}$, obtaining ν_I. Define $\rho_I = \Delta_I + \nu_I$ and

$$\ell_I = \max(0, \underline{R}_I - \rho_I), \qquad u_I = \overline{R}_I + \rho_I. \tag{246}$$

Lemma G.2 (Uniform additive-reference risk enclosure)**.** *Assume valid teacher probabilities throughout $\mathcal{H}$, a finite $L > 0$, and the uniform box tests above at every update of every partition interval and the fixed mixed candidate. Then* (246) *contains the actual risk for every step and every admissible teacher and noise law in that interval. For fixed arithmetic precision, constructing one enclosure takes $O(K)$ scalar operations.*

The enclosure covers an entire rate interval, not just evaluated grid points. It is an ingredient of the numerical certificate, not a separate dominance claim. Uniform comparisons combine the enclosures only after every interval in the pure class is covered.

Proof. Fix one method, one interval I, and any admissible constant step, teacher, and noise law. The exact reference transitions, drifts, and diagonal noise variances belong to (240). Inductively, every product and sum in its mean and variance recursions belongs to the corresponding interval in (241). Squaring the final means by the stated interval-square rule and summing the nonnegative Hessian-weighted terms gives $\underline{R}_I \le \mathcal{R}_K \le \overline{R}_I$. Off-diagonal covariance is immaterial for this diagonal reference value, not assumed absent from the actual training process.

For containment, the first two derivatives in (224) are bounded below by (242): replace the score magnitudes by Q_p, Q_s, the leaf KL by $V_s R_2^2/2$, and the noise magnitude by S. All subtracted coefficients are nonnegative, so the largest admissible step b yields a sufficient test for all $\alpha \in I$. The cross-coordinate sign argument holds when $R_2 \le 1$; it does not depend on the teacher intercept

or the step magnitude. Thus, at each fixed input, the state extrema occur in $\mathcal{E}_k$. At each such state and outcome the map is separately affine in step, teacher coordinates, and noise. A separately affine function on a rectangle attains its extrema at vertices: maximize or minimize one coordinate at a time, preserving an endpoint extremizer after each choice. Hence the displayed endpoint calculation contains every next state for every input. Induction proves almost-sure containment, regardless of the changing on-policy outcome probabilities.

The mean and covariance envelopes are nondecreasing in $\alpha \geq 0$, $|h_2|$, and variance, since all coefficients in those bounds are nonnegative. Their errors remain relative to each fixed input's own reference, not to the interval midpoint. The stated endpoint substitutions therefore bound $\mathbb{E}|e_i|$, $\mathbb{E}e_i^2$, $\mathbb{E}|C_{ii} - Q_{ii}|$, and $\mathbb{E}C_{ii}$ uniformly. Conditional second moments and Minkowski give

$$\mathbb{E}\theta_{i,k+1}^2 \leq (|a_i|\sqrt{\mathbb{E}\theta_{i,k}^2} + |d_i| + \sqrt{\mathbb{E}e_i^2})^2 + \mathbb{E}C_{ii}, \tag{247}$$

which is bounded by the second term in (243). Containment gives its first term. This proves the moment induction.

For the exact backward reference coefficients, $P_{i,k} = a_i^2 P_{i,k+1}$ and $p_{i,k} = a_i(P_{i,k+1}d_i + p_{i,k+1})$. Backward induction gives $0 \leq P_{i,k} \leq \overline{P}_{i,k}$ and $|p_{i,k}| \leq \overline{p}_{i,k}$, including zero or negative a_i. Consequently the coordinate magnitude of $\nabla V_{k+1}(A\theta_k + d)$ is at most $\overline{G}_{i,k}$. Substitution in Theorem A.2 bounds the accumulated expected absolute residuals by Δ_I. Moreover,

$$\tfrac{1}{2}\sum_i H_i \mathbb{E}\theta_{i,K}^2 \leq \mathcal{R}_K + \Delta_I \leq \overline{R}_I + \Delta_I. \tag{248}$$

Each term on the left is nonnegative, justifying the terminal moment refinement. The terminal Taylor envelope gives ν_I and hence $|\mathfrak{R}_K - \mathcal{R}_K| \leq \rho_I$. The actual KL reward regret is nonnegative, proving the full enclosure $\ell_I \leq \mathfrak{R}_K \leq u_I$.

Each interval uses one forward reference pass, one backward coefficient pass, and one forward box/moment pass, each with a constant amount of work per update, proving the operation count. Zero step and zero horizon are covered by the same finite recursions. □

Role in the final certificate. The numerical comparison intersects this additive-reference enclosure with the sampled-coefficient enclosure proved below. Both bounds are valid on the full parameter domain, so their intersection preserves coverage. Outward-rounded Arb arithmetic evaluates the exact rational inputs on all 128 subintervals for each pure method and for the fixed mixed candidate.

G.3 A Computable Comparison Retaining Sampled Update Coefficients

The finite-radius result explains why action and visitation noise should be retained, but does not by itself supply tight multi-update constants. We now construct such constants for the same nonlinear two-stage family. Nothing changes in training, the information allocation, or the continuous pure-step classes. Only the reference and the error charged to it change. This distinction permits a controlled test of which approximation limits the comparison.

An exact reference with the same mean and richer covariance. Use the notation of Appendix G.1, including $(r,t) = (1,0),(0,1),(1,c)$ and $r \in \{0,1\}$. Independently at each update draw $A \sim \mathrm{Ber}(p_0)$, $B \sim \mathrm{Ber}(s_0)$, and centered reward noise η with variance σ^2. Put $a = A - p_0$, $b = B - s_0$. The reference is $X' = MX + d$, with

$$M = \begin{pmatrix} 1 - \alpha(ra^2 + th_p) & -\alpha r A a b \\ 0 & 1 - \alpha A(rb^2 + th_s) \end{pmatrix}, \qquad d = \alpha \begin{pmatrix} r\eta a + th_p h_1 \\ A(r\eta b + th_s h_2) \end{pmatrix}. \tag{249}$$

This is the sample-coefficient construction of (53), expressed in step/ratio notation. Its mean map is diagonal: $\mathbb{E}M = \mathrm{diag}(a_1, a_2)$ and $\mathbb{E}d = (d_1, d_2)$, where $a_i = 1 - \alpha(r+t)H_{ii}$ and $d_i = \alpha t H_{ii} h_i$. Here d_i denotes mean drift; sampled drift coordinates are denoted by d_i^{sample} when the distinction is needed. Write $Q^{\mathrm{ra}}(x,y) = \mathrm{Cov}(M(x,y)^\top + d)$. Its diagonal entries are

$$\begin{aligned} Q_{11}^{\mathrm{ra}}/\alpha^2 &= r\{h_p(1-2p_0)^2x^2 + p_0(1-p_0)^2 h_s y^2 + \sigma^2 h_p\}, \\ Q_{22}^{\mathrm{ra}}/\alpha^2 &= p_0 r\{h_s(1-2s_0)^2 y^2 + \sigma^2 h_s\} + h_p\{th_s h_2 - (r+t)h_s y\}^2. \end{aligned} \tag{250}$$

The additive reference retains only their values at zero. The new reference also retains quadratic variation generated by actions and visits, including the dependence between the teacher intercept and its sampled leaf update. There is no claim that the actual sampling probabilities remain fixed. At zero step, formulas displaying division by α^2 are read in their multiplied form: both actual and reference update covariances are zero.

Diagonal backward values remain sufficient. For $P_{k+1} = \mathrm{diag}(P_{1,k+1}, P_{2,k+1})$, define

$$\begin{aligned}
\lambda_{11} &= \mathbb{E}\big[\{1-\alpha(r(A-p_0)^2+th_p)\}^2\big],\\
\lambda_{12} &= \alpha^2 r p_0(1-p_0)^2 h_s,\\
\lambda_{22} &= 1-p_0+p_0\mathbb{E}\big[\{1-\alpha(r(B-s_0)^2+th_s)\}^2\big],\\
\tau_1 &= a_1 d_1, & \tau_2 &= \{1-\alpha(r+t)h_s\}d_2,\\
\kappa_1 &= \alpha^2\{r\sigma^2 h_p + (th_p h_1)^2\}, & \kappa_2 &= \alpha^2 p_0\{r\sigma^2 h_s + (th_s h_2)^2\}.
\end{aligned} \tag{251}$$

The two expectations contain only two Bernoulli outcomes. The exact value coefficients satisfy

$$\begin{aligned}
P_{1,k} &= \lambda_{11}P_{1,k+1}, & P_{2,k} &= \lambda_{12}P_{1,k+1} + \lambda_{22}P_{2,k+1},\\
p_{i,k} &= \tau_i P_{i,k+1} + a_i p_{i,k+1}, & & i=1,2,\\
\gamma_k &= \gamma_{k+1} + \tfrac{1}{2}\sum_i P_{i,k+1}\kappa_i + \sum_i p_{i,k+1}d_i.
\end{aligned} \tag{252}$$

The terminal values remain $(P_K, p_K, \gamma_K) = (H, 0, 0)$. Thus the reference risk is still a scalar quadratic value at initialization, computed in linear work in the finite horizon.

Covariance discrepancies of higher local order. Use the reachable boxes, actual second-moment upper bounds, mean-error envelopes, and variance upper bounds from Appendix G.1. On the current box write

$$\begin{aligned}
X_{12} &= \min\{R_1 M_2, R_2\sqrt{M_1 M_2}\},\\
J_p &= L_p T_p^2 + 4h_p T_p V_p, & J_s &= L_s T_s^2 + 4h_s T_s V_s,\\
L_B &= B_p \max_{p \text{ in box}} |1-3p|, & m &= th_s|h_2|,\\
b_0 &= (r+t)h_s, & j_0 &= (r+t/2)L_s.
\end{aligned} \tag{253}$$

Here the probability maximum in L_B uses its two endpoints. Define

$$\begin{aligned}
\widetilde{\omega}_1/\alpha^2 &= r\left[J_p R_1 M_1 + V_p T_p^2\left(V_s X_{12} + \frac{V_s^2 R_2^2 M_2}{4}\right) + h_s L_B X_{12} + B_p L_s R_2 M_2 + \sigma^2 L_p S_1\right],\\
\widetilde{\omega}_2/\alpha^2 &= r\{h_s(1-2s_0)^2 V_p X_{12} + \overline{p} J_s R_2 M_2\} + r\sigma^2(h_s V_p S_1 + \overline{p} L_s S_2)\\
&\quad + L_p(m^2 S_1 + 2mb_0 S_1 S_2 + b_0^2 X_{12})\\
&\quad + V_p\{2j_0(mM_2 + b_0 R_2 M_2) + j_0^2 R_2^2 M_2\}.
\end{aligned} \tag{254}$$

These quantities bound $\mathbb{E}|C_{ii} - Q_{ii}^{\mathrm{ra}}(\theta_k)|$. The dependence on external noise and teacher mismatch remains explicit; it is not removed by assuming that the mean is locally linear.

Proposition G.3 (Structural certificate with sampled update coefficients)**.** *Under the two-stage model and reachable-box hypotheses of Proposition G.1, use the exact backward values* (252)*. In* (230)*, replace ω_i by $\widetilde{\omega}_i$ from* (254)*. Let the resulting residual sum be $\widetilde{\Delta}_K$ and the reference risk be $\mathcal{R}_K^{\mathrm{ra}}$. Apply the same terminal bound using*

$$\overline{M}_i = \min\{M_{i,K}, 2(\mathcal{R}_K^{\mathrm{ra}} + \widetilde{\Delta}_K)/H_{ii}\}. \tag{255}$$

Then

$$|\mathfrak{R}_K - \mathcal{R}_K^{\mathrm{ra}}| \le \widetilde{\rho}_K := \widetilde{\Delta}_K + \widetilde{\nu}_K. \tag{256}$$

Interval arithmetic over every prescribed step/teacher/variance interval gives uniform risk enclosures, and hence comparisons against continuously tuned pure methods, in $O((M+1)K)$ scalar operations for M pure intervals.

In plain language, the new reference predicts some variability that the old one charged entirely as approximation error. The computation still avoids enumerating training histories. A smaller final radius is not automatic: the value weights and interval widths also change. Both certificates remain valid, so intersecting their intervals is safe and never weakens either available enclosure.

Proof. We first verify the reference moments, then its nonlinear covariance discrepancies, and finally the continuous-input construction.

Conditioning on A and using $\mathbb{E}(B - s_0) = 0$ gives $\mathbb{E}(M_{11}M_{12}) = 0$. Since $M_{21} = 0$, a diagonal P_{k+1} therefore gives diagonal $\mathbb{E}(M^\top P_{k+1} M)$. Its two entries are the first line of (252). Directly, $\mathbb{E}M_{12}^2 = \alpha^2 r \mathbb{E}[A(A-p_0)^2]\mathbb{E}(B-s_0)^2 = \alpha^2 r p_0(1-p_0)^2 h_s$, using $r^2 = r$. The other two squares give λ_{11} and λ_{22}.

The centered independent noise removes all terms linear in η. Consequently $\mathbb{E}(M_{11} d_1^{\text{sample}}) = a_1 d_1$ and $\mathbb{E}(M_{12} d_1^{\text{sample}}) = 0$. For the leaf,

$$\mathbb{E}(M_{22} d_2^{\text{sample}}) = \alpha t h_s h_2 \mathbb{E}[A\{1 - \alpha A(r(B-s_0)^2 + t h_s)\}] \tag{257}$$

$$= \alpha t h_s h_2 p_0 \{1 - \alpha(r+t) h_s\} = \tau_2. \tag{258}$$

This differs from $a_2 d_2$: factoring the matrix and drift would incorrectly remove their shared visitation. Squaring the two sampled drift coordinates and taking expectations gives κ_1, κ_2. Lemma A.1 now proves all remaining backward coefficients. The law of total variance, first conditional on the root action, gives (250); the derivation is the same as (220) with coefficients at p_0, s_0 and leaf conditional mean $t h_s h_2 - (r+t) h_s y$.

For the first covariance entry, put $D_s = \mathrm{KL}(s \| s_0)$ and $F(p) = p(1-p)(1-2p)^2$. Adding and subtracting $h_p(1-2p)^2$ gives

$$|F(p) - F(p_0)| \le L_p T_p^2 |x| + h_p \, 4 T_p |p - p_0| \le J_p |x|. \tag{259}$$

Also, since $0 \le D_s \le V_s y^2/2$,

$$|(x + D_s)^2 - x^2| \le V_s |x| y^2 + V_s^2 y^4 / 4. \tag{260}$$

Thus the discrepancy of the first quadratic term in (220) has expectation at most $J_p R_1 M_1 + V_p T_p^2 (V_s X_{12} + V_s^2 R_2^2 M_2/4)$. For $B(p) = p(1-p)^2$, differentiation with respect to the root logit gives $\frac{d}{dx} B(p(x)) = B(p(x))(1 - 3p(x))$. Therefore $|B(p) - B(p_0)| \le L_B |x|$, and

$$|B(p) v_s - B(p_0) h_s| \le h_s L_B |x| + B_p L_s |y|. \tag{261}$$

Multiplying by y^2, taking expectations, and adding the external-noise bound $\sigma^2 \mathbb{E}|v_p - h_p| \le \sigma^2 L_p S_1$ proves $\widetilde{\omega}_1$. Here we used $\mathbb{E}|x| y^2 \le X_{12}$, $\mathbb{E}|x|^3 \le R_1 M_1$, $\mathbb{E}|y|^3 \le R_2 M_2$, and $\mathbb{E} y^4 \le R_2^2 M_2$.

For the leaf action term, applying the preceding F bound to s gives

$$|p v_s (1-2s)^2 - p_0 h_s (1-2s_0)^2| \le h_s (1-2s_0)^2 V_p |x| + \overline{p} J_s |y|. \tag{262}$$

After multiplication by y^2 this yields the first braces in $\widetilde{\omega}_2$. The external-noise term uses the unchanged bound $\mathbb{E}|p v_s - p_0 h_s| \le h_s V_p S_1 + \overline{p} L_s S_2$.

For visitation variance let $m_a(y) = -r v_s y + t(q_B - s)$ and $m_r(y) = t h_s h_2 - b_0 y$. Taylor bounds give

$$|m_a(y) - m_r(y)| \le (r + t/2) L_s y^2 = j_0 y^2, \qquad |m_r(y)| \le m + b_0 |y|. \tag{263}$$

Add and subtract $v_p m_r(y)^2$ to obtain

$$|v_p m_a(y)^2 - h_p m_r(y)^2| \le L_p |x| (m + b_0 |y|)^2 + V_p \{2 j_0 y^2 (m + b_0 |y|) + j_0^2 y^4\}. \tag{264}$$

Taking expectations, and using $\mathbb{E}|xy| \le S_1 S_2$ in addition to the preceding moment bounds, gives the last two lines of (254). This proves every new covariance envelope without assuming independence of the two coordinates.

The additive and sampled-coefficient references have the same conditional mean, so the first absolute and second mean-error bounds apply to both. The actual variance bounds and moment recursion

also apply to both references. The reference value Hessians are diagonal and nonnegative. Insert the covariance envelopes and exact value coefficients into the conditional-moment residual from Theorem A.2. This proves $\sum_k \mathbb{E}|Z_k| \le \widetilde{\Delta}_K$. Nonnegativity of each terminal quadratic summand then justifies the refined $\overline{M}_i$, before invoking the target-loss approximation. Applying the terminal Taylor bound proves (256).

For a continuous input rectangle, enclose every quantity in (251) by inclusion-preserving interval arithmetic. Propagate the signed linear coefficients and nonnegative quadratic coefficients in (252). Upper endpoints of $P_{i,k}$ and upper absolute endpoints of $p_{i,k}$ give valid residual weights. In (254), use the upper step, variance, and absolute teacher endpoints, as all remaining bound coefficients are nonnegative. The old uniform box tests and moment recursion therefore remain valid on the same whole rectangle.

For completeness, a second reference-risk enclosure is obtained by forward raw moments. If $\mu_i = \mathbb{E}X_i$ and $s_i = \mathbb{E}X_i^2$, then

$$\begin{aligned} \mu_i' &= a_i\mu_i + d_i, \\ s_1' &= \lambda_{11}s_1 + \lambda_{12}s_2 + 2\tau_1\mu_1 + \kappa_1, \\ s_2' &= \lambda_{22}s_2 + 2\tau_2\mu_2 + \kappa_2. \end{aligned} \tag{265}$$

The cross-coordinate terms vanish by the same zero expectations already proved. The forward and backward intervals enclose the same exact risk; their intersection, clipped below at zero, is therefore valid. Similarly, intersecting the old and new *actual-risk* intervals preserves coverage. Taking minima over the complete pure partition then gives the tuned-family enclosures of (269). These are deterministic uniform operations, not data-dependent hypothesis tests. Each reference and bound step uses a fixed number of scalar operations, proving the linear-horizon computational claim. □

A finite inequality explaining the improvement. Suppose on a particular step that $R_i \le \delta$, $M_i \le \delta^2$, $|h_2| \le H_0\delta$, $\sigma^2 \le V_0\delta^2$, and $0 < \delta \le \delta_0$. Use common upper bounds for the box constants over radius δ_0 and put $m_H = th_sH_0$. Direct substitution gives

$$\begin{aligned} \widetilde{\omega}_1 &\le \alpha^2\delta^3 r\{J_p + V_pT_p^2(V_s + V_s^2\delta_0/4) + h_sL_B + B_pL_s + V_0L_p\}, \\ \widetilde{\omega}_2 &\le \alpha^2\delta^3\big[r\{h_s(1-2s_0)^2V_p + \overline{p}J_s + V_0(h_sV_p + \overline{p}L_s)\} \\ &\qquad + L_p(m_H + b_0)^2 + V_p\{2j_0(m_H + b_0) + j_0^2\delta_0\}\big]. \end{aligned} \tag{266}$$

These are inequalities for every positive radius in the stated interval, not asymptotic assertions. They explain why retaining sampled coefficients can matter: the additive reference's envelopes include entire state-quadratic variance terms, whereas these bounds charge their nonlinear discrepancies. Fixed nonvanishing noise or teacher error must instead retain the explicit terms of (254).

G.4 From Interval Enclosures to Tuned-Class Bounds

This aggregation argument turns uniform per-interval bounds into a comparison against an entire constant-rate class. It justifies combining the supporting enclosures without omitting any permitted rate.

Lemma G.4 (Complete-partition enclosure)**.** *Fix a finite step cap $L > 0$ and let ζ denote a common admissible teacher and noise law. Define the pure-family risks*

$$F_a(K;\zeta) = \inf_{\alpha\in[0,L]} \mathfrak{R}_K^a(\alpha;\zeta), \qquad a \in \{\mathsf{R},\mathsf{D}\}. \tag{267}$$

For mixing, F_{M} is the risk of the fixed normalized recipe (λ, w). Suppose each pure method has a finite partition of $[0, L]$ into nonempty intervals, each with a risk enclosure $[\ell_{a,I}, u_{a,I}]$ valid for every step in I and every admissible ζ. This may be an individual enclosure or an intersection of valid enclosures. Put

$$L_a = \min_I \ell_{a,I}, \qquad U_a = \min_I u_{a,I}. \tag{268}$$

Then, simultaneously for every ζ in the stated input region,

$$L_a \le F_a(K;\zeta) \le U_a. \tag{269}$$

For the fixed mixture use its single-recipe enclosure. A strict inequality $U_a < \min_{b\neq a} L_b$ certifies that family a has lower expected regret than both alternatives throughout the entire input region.

In plain language, the pure methods may choose the best constant step with full knowledge of the inputs. Their columns are not risks at a common step, grid estimates of an optimum, or comparisons of upper bounds alone. The mixture is fixed: a pure winner is not claimed to outperform a mixed family that can set its weight to the same pure endpoint.

Proof of the tuned-family enclosure. Fix any admissible ζ. Every $\alpha \in [0, L]$ belongs to a partition interval I, so $\mathfrak{R}_K^a(\alpha;\zeta) \geq \ell_{a,I} \geq L_a$. Taking the infimum over the entire step interval proves the lower bound. For every nonempty I, choose any $\alpha_I \in I$. The uniform upper enclosure gives

$$F_a(K;\zeta) \leq \mathfrak{R}_K^a(\alpha_I;\zeta) \leq u_{a,I}. \tag{270}$$

This holds for every I, hence the minimum of their upper endpoints is also an upper bound. No attainment of the pure risk infimum is required. The minimum of the finitely many upper endpoints is attained; any step in that minimizing interval is an explicit fixed recipe with risk at most U_a throughout the input region. In particular, the upper bound need not exploit input-dependent tuning. Finally, for every competing family,

$$F_b(K;\zeta) - F_a(K;\zeta) \geq L_b - U_a. \tag{271}$$

Taking the smaller of these two gaps proves the ordering criterion. The argument is pointwise for every admissible ζ, so it proves a uniform statement without a statistical union bound. If any interval cannot be enclosed, the complete-family lower bound is not supplied by this construction; that interval cannot be omitted. Overlapping enclosures likewise do not establish equality or reversed performance. □

H Validation Protocols and Complete Results

The structural checks and Qwen study correspond to Figure 1. The final subsection examines reference forecasts and nonlinear certificates at a larger radius. Each study reports its own protocol and complete outcomes, including inconclusive comparisons.

H.1 Validation Aligned with the Structural Propositions

These studies directly test the sufficient conditions in Propositions 4.2 and 4.3 using the known two-stage model. The designs were informed by the theoretical constructions, with all conditions, controllers, predictions, and sample sizes specified before evaluation. They test the predicted mechanisms in a known environment, rather than provide blinded discovery or unknown-model validation.

Window design and actual finite scale. Set $p_\star = s_\star = 11/20$, $(x_0, y_0) = (1, 1)$, $\varepsilon = 10^{-8}$, mixed count $K = 20$, mixed step $3/5$, and teacher fraction $1/2$. Independently cross

$$h_1 \in \{.3, .45, .6\}, \quad h_2 \in \{-.5, 0, .5\}, \quad \nu \in \{.2, .65, 1\}, \quad c_0 \in \{0, 3\}.$$

There are 54 settings, covering nine teacher-error directions without imposing $h_1 = -h_2$. The noise is symmetric $\zeta = \pm\sqrt{\nu}$, with common support bound one. Costs $c_R = c_D = 1$ and budget $20(c_0+2)$ allow 40 pure updates when $c_0 = 0$ and 25 when $c_0 = 3$, against 20 mixed updates. This is equal spending, not matched updates.

Let $P = \min\{Q_N(\nu), h_p h_1^2/2\}$, $M = F_K(h) + D_K\nu$, and $g = P - M$. For $g > 0$, the prespecified rule sets $\tau = (P + M)/2$ and $\delta = g/4$; these depend only on known parameters. The exact rational recursion determines P and M. The constants of Theorem 4.1, with rectangle $(10, 3)$ and radius cap 10^{-3}, give $C_* \geq \max(C_R, C_M, t_D)$. The reported window-test margin is $g - 8\varepsilon C_*$. A positive margin ensures that every normalized nonlinear error is below $\delta/2$. We also explicitly check the two noise thresholds, teacher ordering, and containment with 192-bit outward-rounded arithmetic. Passing certifies an actual regret gap of at least $\varepsilon^2\delta$ under Proposition 4.2; otherwise the output is inconclusive. The scale is deliberately small because these uniform bounds are conservative. It should not be interpreted as an empirically useful training radius.

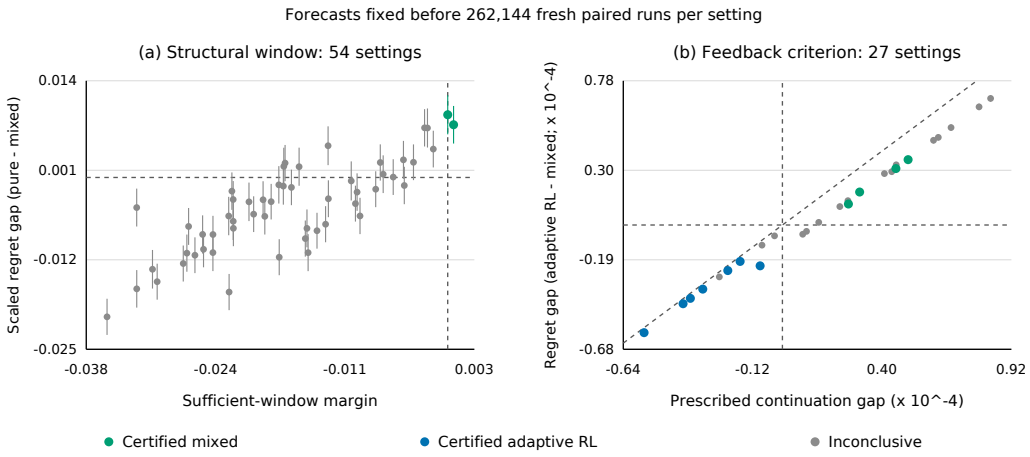


Figure 2: Enlarged view of Figure 1a–b; all structural-validation settings are shown. Left: a positive window-test margin certifies mixing; the vertical axis compares the evaluated pure recipes with mixing in units of ε^2. Right: RL-minus-mixed regret with a prescribed or adaptive second rate, holding the first rate fixed. Below the diagonal, feedback helps RL; the lower-right quadrant shows a ranking reversal. Colors denote analytical certificates, not empirical winners. Vertical bars use simultaneous 95% finite-sample intervals.

Which pure recipes are simulated? The certificate covers all prescribed RL schedules in $[0, 2]$ and all permitted adaptive OPD rules and stopping. Finite simulations cannot enumerate those classes. We select an RL recipe by its known reference risk over 33 equally spaced constant rates in $[0, 2]$, all their prefixes, and the five schedules $2/(1 + ak)$ with $a \in \{.05, .1, .2, .5, 1\}$. OPD uses 65 constant rates in $[0, 4]$ and all prefixes. Ties retain the first enumerated candidate. These model-aware calculations are fixed before simulation; they are not trained selectors or claims of globally optimal baseline tuning. As in the theoretical comparison, the cost budget counts training queries, not the analyst's computation.

The simulation evolves the actual nonlinear policy, not the reference model. To avoid cancellation at the small scale, it stores logit errors divided by ε and computes policy increments using `expm1`. Normalized Bernoulli regret is evaluated through

$$\frac{\mathrm{KL}(\sigma(z + \varepsilon x)\|\sigma(z))}{\varepsilon^2} = x^2 \int_0^1 t\sigma'(z + t\varepsilon x)\, dt,$$

using four-point Gauss–Legendre quadrature. Checks against 80-digit arithmetic test this numerical implementation. The analytic certificates do not rely on that quadrature or the simulated outcomes.

Feedback design and continuous-rate guarantees. Use optimal logits $(.1, .1)$, actual initial errors $(.04, .04)$, and actual teacher offsets independently crossed over

$$h_1^{\mathrm{actual}} \in \{.02, .03, .039\}, \qquad h_2^{\mathrm{actual}} \in \{-.04, -.02, .02\}.$$

The noise law has masses $.45, .1, .45$ at $-S, 0, S$, with $S \in \{.008, .02, .032\}$, giving 27 settings. Here offsets mean teacher probabilities $p_\star + h_p h_1^{\mathrm{actual}}$ and $s_\star + h_s h_2^{\mathrm{actual}}$, not teacher logit errors. Costs $(c_0, c_R, c_D) = (3, 1, 1)$ and budget ten allow two updates per method. No baseline errors are added. The common invariant rectangle is $|x| \le .2$, $|y| \le .08$; its containment and OPD-floor conditions are checked for every setting.

The mixture is the same controller throughout the grid: its first step is two and weight is $1/2$; its second step is the clipped quadratic mixed rule computed at the fixed nominal teacher offsets $(.039, -.02)$, support .032, and activation .9. It may depend on the current state but is not retuned to each setting. We enclose its actual terminal regret by interval enumeration of both updates.

For each setting, integrate the final-step quadratic optimum over every first-step branch and enclose every first RL rate in $[0, 2]$ using 1,024 covering intervals. This directly lower-bounds $\inf_\alpha(F_{Q_-} -$

$\mathcal{A}_{Q_-}$), including stopping. Together with the OPD floor and mixed upper bound, it evaluates (16). The reverse test (17) uses an explicit first RL rate selected on a 1,025-point grid by $F_{Q_+} - \mathcal{A}_{Q_+}$, followed by the state-dependent optimal Q_+ rate. Its expected quadratic risk is reevaluated with interval arithmetic. Selecting a witness from a grid does not claim global optimality; an upper-bound witness suffices for the reverse inequality. The mixed certificate, in contrast, covers the entire first-rate interval, not just that grid.

Measuring feedback rather than retuning the first step. Five controllers are simulated: the adaptive RL witness; a prescribed RL continuation with the *same* first rate; a separately selected prescribed two-rate RL recipe; OPD; and the fixed mixed controller. Prescribed RL minimizes F_{Q_+} over the same first-rate grid, with the optimal constant second rate. OPD uses a first-rate grid in $[0, 4]$ and a clipped Q_+-optimal second rate. Every controller is chosen before the next sampled update. The paired same-first-rate comparison isolates feedback credit: its difference in expected terminal Q_+ is exactly $\mathcal{A}_{Q_+}(\alpha)$. Actual regret differences need not equal that quadratic credit. An independent raw-utility enumeration checks both terminal expectations without using the moment formula to generate outcomes.

Fresh evaluation and simultaneous uncertainty. Every setting has 262,144 independent paired replications in chunks of 32,768. Methods share random uniforms within a replication, while each samples actions from its own policy. The seed prefix is $(20260924, 427043)$, followed by study, setting, and chunk indices. All paired terminal losses are retained. Window losses are reported in units of ε^2. Their pathwise bound is $109/8$ from the certified rectangle. Feedback losses are bounded by $(.2^2 + .08^2)/8$; terminal Q_+ is bounded by $(.2^2 + \sigma(.3).08^2)/8$.

There are $3 \cdot 54 + 11 \cdot 27 = 459$ contrasts: all three window pairs, all ten feedback-controller pairs, and the quadratic feedback credit in each feedback setting. Applying (276) with $m = 459$ and each variable's range gives simultaneous coverage at least 95% over the complete pairwise family under the specified sampling model. This controls sampling uncertainty, not floating-point error; numerical checks and outward-rounded certificates are reported separately. For the left panel of Figure 1, taking the minimum of the two pure-minus-mixed interval endpoints bounds the smaller expected improvement. The right panel holds the first RL rate fixed.

Complete outcomes and limits. The window certifies two mixed wins (root offsets .45 and .6, leaf offset $-.5$, variance .65 and one, respectively, both with overhead three); both are supported by the simultaneous intervals. The remaining 52 settings are inconclusive, not certified pure wins. The feedback test certifies four mixed wins and seven adaptive-RL wins, all supported, while 16 settings remain inconclusive. In all 27 settings the predicted quadratic feedback credit lies in its simulation interval; independent terminal-tree enumeration agrees within $3 \cdot 10^{-20}$ in double precision.

Two of the feedback test's inconclusive settings exhibit a statistically supported reversal of the evaluated prescribed-RL ordering. Both have $S = .02$, with teacher offsets $(.03, .02)$ and $(.039, -.02)$. Relative to mixing, the separately selected prescribed RL mean regrets are higher by $8.18 \cdot 10^{-6}$ and $9.71 \cdot 10^{-6}$; their simultaneous intervals are $[6.75, 9.60] \cdot 10^{-6}$ and $[8.31, 11.11] \cdot 10^{-6}$, rounded outwards. The adaptive RL differences are $-5.03 \cdot 10^{-6}$ and $-3.45 \cdot 10^{-6}$, with intervals $[-6.48, -3.58] \cdot 10^{-6}$ and $[-4.88, -2.02] \cdot 10^{-6}$. These reversals demonstrate the value of feedback for the evaluated controllers; they are not successful applications of the conservative reverse certificate, nor proofs against all prescribed controllers.

These tests support both propositions where their sufficient conditions hold, while also documenting inconclusive cases. They do not establish sharp boundaries, practical-scale window guarantees, or adaptive dominance at longer horizons. The Qwen and larger-radius studies provide complementary evidence; their samples are not pooled with these tests.

Numerical checks. All predictions are checked at 256-bit precision, and the reported intervals are reconstructed from the paired terminal losses. Independent calculations check the nonlinear loss evaluation, cost counts, interval formulas, and raw-tree feedback identity. These checks supplement the proofs; they do not replace the class-wide analytical bounds.

H.2 QWEN IMPLEMENTATION OF THE TWO-STAGE MECHANISM

This experiment replaces the student's two independent logits with a pretrained transformer while retaining the known task, engineered teacher, and exact reward evaluation. It asks whether the noise-dependent ordering persists when the parameters couple the decisions and prompts. The goal is to test this mechanism, not language-generation quality or distillation from another LLM. We first describe the protocol and tuning choices, then report the outcomes and their uncertainty.

The analysis uses aggregate measurements from the accompanying experiment report. The underlying run-level records have not been independently reconstructed. Uncertainty is assessed from the reported marginal moments, without imputing the missing covariance.

Model, task, and endpoint. The student is `Qwen/Qwen2.5-0.5B-Instruct`, with 494,573,440 total parameters and 540,672 trainable LoRA parameters. Rank-eight, scaling-16 adapters modify attention query and value projections. The pretrained base and original vocabulary head are frozen, but the transformer participates in each forward and backward pass. Computation uses FP32 with TF32 disabled. The available output tokens are the original vocabulary tokens `0` and `1`; the leaf prefix preserves the sampled root token.

Sixteen fixed text prompts describe the two-stage task. Root action zero terminates with reward zero; action one visits a leaf with rewards approximately .04875052 and .14875052. With a uniform reference policy and KL coefficient one, both optimal probabilities are $\sigma(.1)$ and the exact regret is (5). Final regret averages this quantity over the uniform 16-prompt training law, without sampling terminal evaluation actions. Thus the experimental replicates are training seeds, not prompts or sampled evaluation completions. This endpoint does not test generalization to unseen prompts.

Frozen prefix-specific decoding offsets center the initial effective logits at .14, giving initial errors $(.04, .04)$. In actual probability coordinates, the engineered teacher has $q_A = p_\star + .02h_p$ and $q_B = s_\star - .02h_s$. Its exact regret is $7.6316498 \cdot 10^{-5}$, below the initial student's $3.0575379 \cdot 10^{-4}$. This establishes useful teacher performance in the known task, not necessarily alignment of its parameter gradient with the reward gradient.

Updates and declared costs. The sampled directions in (6) are differentiated through the actual transformer logits. Only visited leaf states contribute. The mixture uses $(\widehat{g} + \widehat{u})/2$, and all methods use plain SGD without momentum, adaptive preconditioning, channel normalization, or gradient clipping. State-value baselines are exact. Clean runs have zero added reward noise; noisy runs add independent signs of magnitude .032. The task and teacher are unchanged across noise conditions.

Reward and teacher-bundle allowances each cost one. With rollout overhead zero and budget 40, the pure methods receive 40 updates and mixing receives 20. With overhead three and budget 100, the corresponding counts are 25 and 20. These are equal declared query allowances within each comparison, not equal measured compute. Changing both overhead and budget keeps the mixed horizon fixed; it is not a cost intervention at fixed budget or a test of the scalar correlation boundary. The analytic teacher's inexpensive evaluation does not measure the cost of querying another LLM.

Independent tuning and fresh evaluation. One shared fixed divisor, 2157.241291, converts effective learning rates to parameter rates. It is the largest initial two-output parameter-Jacobian Gram eigenvalue across four calibration prompts; it is not adapted during training. Each method receives 13 candidates: a zero-step option, six constant rates, and the same six initial rates followed by $3/(k+3)$ decay. RL and mixing use initial effective rates $\{.05, .1, .3, .6, 1.2, 2\}$; OPD uses $\{.1, .2, .6, 1.2, 2.4, 4\}$. Two paired tuning seeds select each recipe by final regret before 16 fresh paired evaluation seeds. The mixture weight remains $1/2$, but its learning rate is retuned for each noise condition. Thus this is not the paper's low-noise comparison against one fixed mixed recipe. Query budgets apply to the selected training run; tuning is separate and its run counts are reported below, not silently included in those allowances.

Table 4 lists all eight distinct selections. Six reach an upper endpoint, so the results establish a comparison within the tested grid rather than optimality over all rates or schedules. The grid was not expanded after observing the final outcomes.

Method	Noise	Schedule	Updates	First rate	Tuning regret
Mixing	Clean	Constant	20	$2^{\dagger}$	$2.3201444 \cdot 10^{-5}$
Mixing	Noisy	Constant	20	$2^{\dagger}$	$2.9698763 \cdot 10^{-5}$
OPD	Either	Constant	25	1.2	$4.6125343 \cdot 10^{-5}$
OPD	Either	Decay	40	$4^{\dagger}$	$4.5792351 \cdot 10^{-5}$
RL	Clean	Constant	25	$2^{\dagger}$	$1.1764554 \cdot 10^{-5}$
RL	Clean	Constant	40	$2^{\dagger}$	$7.7032596 \cdot 10^{-6}$
RL	Noisy	Constant	25	1.2	$2.4677319 \cdot 10^{-5}$
RL	Noisy	Decay	40	$2^{\dagger}$	$7.0128330 \cdot 10^{-5}$

Table 4: Qwen schedules selected using two tuning seeds. Rates are effective rates; divide by 2157.241291 for parameter rates. A dagger marks an upper grid endpoint. Tuning values are not evaluation results.

Prompt indices, action uniforms, and reward-noise uniforms are paired across methods, while each policy samples its own evolving action distribution. All comparisons use the final iterate. Identical mixed runs are reused across overheads, and OPD runs across reward-noise conditions. Their reported means and variances agree across the reused conditions. This reuse changes neither training allowances nor the number of independent seeds; the two noisy comparisons do not provide 32 independent mixed evaluations.

All reported means and variances. Let $Y = 10^6\mathcal{L}$ for the final policy, averaged over the full prompt panel before aggregation across seeds. The reported mean is $\bar{Y} = \sum_{i=1}^{16} Y_i/16$ and sample variance is $s_Y^2 = \sum_{i=1}^{16}(Y_i - \bar{Y})^2/15$. Table 5 gives both; Figure 1 plots the means with one sample standard deviation s_Y in panels (c–d), not confidence intervals.

Reward	Overhead	RL	OPD	Mixing
Clean	0	9.928 [8.355]	53.030 [35.498]	23.616 [24.850]
Noisy	0	54.518 [12.634]	53.030 [35.498]	44.477 [26.138]
Clean	3	18.142 [39.836]	51.149 [18.447]	23.616 [24.850]
Noisy	3	68.132 [40.684]	51.149 [18.447]	44.477 [26.138]

Table 5: Qwen final regret: mean Y [sample variance of Y] over 16 training seeds. Divide means by 10^6 and variances by 10^{12} for original units. Brackets are sample variances, not confidence intervals.

RL has the lowest observed mean with clean rewards and mixing with noisy rewards. At overhead zero, the noisy mixed mean is 18.4% below RL and 16.1% below OPD; at overhead three, the reductions are 34.7% and 13.0%. These are descriptive reductions in sample means, not lower confidence bounds. The noisy variances put mixing between the pure methods: at overhead zero, RL has the lowest variance and OPD the highest; at overhead three, OPD has the lowest and RL the highest. Final-regret variance measures run-to-run variability, not gradient-estimator variance or injected reward-noise variance.

Uncertainty from the reported summaries. The report provides marginal means and variances but not the paired differences or their covariance. This prevents reconstruction of paired confidence intervals, yet still permits a conservative check of the noisy mean gaps. For a baseline A and mixture M, Cauchy–Schwarz bounds the conventional paired standard error:

$$s_{A-M}^2 = s_A^2 + s_M^2 - 2s_{AM}, \qquad \frac{|s_A - s_M|}{4} \le \frac{s_{A-M}}{\sqrt{16}} \le \frac{s_A + s_M}{4}. \tag{272}$$

Here s_{AM} is the sample covariance. Dividing a positive observed gap $\bar{Y}_A - \bar{Y}_M$ by the upper standard-error bound gives a lower bound on the conventional paired t statistic, whatever the missing covariance. At overhead zero, the noisy summaries give approximate lower bounds 4.634 against RL and 3.090 against OPD; at overhead three, they give 8.234 and 2.837. For example, the latter OPD comparison has gap 6.672 and maximum standard error $(\sqrt{18.447} + \sqrt{26.138})/4 \simeq 2.352$. All four lower bounds exceed the unadjusted two-sided critical value $t_{15,.975} \simeq 2.131$, even under the largest

covariance-compatible standard error. Under the conventional paired-t reference assumptions, each noisy mixed-versus-pure comparison therefore passes the unadjusted 5% threshold.

This calculation is an aggregate-based diagnostic using rounded moments, not an observed paired interval or a distribution-free inference procedure. It does not establish a multiplicity-adjusted result or check the shape of the paired-difference distribution. With only 16 seeds, the run-level outcomes remain important for checking distributional assumptions and outliers, and for applying alternative uncertainty procedures. The diagnostic concerns only the four noisy comparisons; the clean-reward ordering is reported descriptively.

Relation to the theoretical statements. Three features connect the study to the theory: changing reward noise preserves the task and teacher, the pure methods receive extra affordable updates, and each method is tuned independently. The observed reversal therefore supports the qualitative claim that noise can change the preferred regime after tuning and cost accounting.

The formal guarantee does not transfer automatically. Adapter updates couple both logits and multiple prompts, so the independent-logit containment, teacher-floor, and moment arguments need not hold. The study does not certify Theorem 4.1 for a transformer, verify a structural margin, or test the two-update feedback criterion. Its fixed prompt panel, engineered teacher, exact baselines, calibrated initial logits, and short binary responses make it a controlled mechanism test rather than a test of ordinary LLM post-training.

Computation and reproducibility. The experiment report records 208 unique tuning runs and 128 unique evaluation runs on two RTX 5090 D v2 workers, with 9,870 training updates and 19.55 summed worker-training minutes; pipeline elapsed time was 10.87 minutes. Summed training time excludes model loading, final evaluation, checkpoint writing, and reporting. Initial calibration peaked at 2.235 GiB allocated on one GPU; this is not a campaign-wide memory bound.

The accompanying materials contain the report, aggregate measurements, and figure code, but not the server-side training sources, prompt texts, trajectories, adapters, or run-level verification records. The run-level checks described in the report have not been independently repeated. These materials therefore do not provide full empirical reproducibility or the paired outcomes needed to reconstruct confidence intervals. The aggregate-based diagnostic above does not replace run-level validation.

H.3 Prospective Validation Across Model and Cost Parameters

This study compares a reference forecast for three fixed recipes with the nonlinear class certificate in Appendix C.4. Simulations test the former; they cannot establish superiority over every schedule covered by the latter.

Prespecified design. We use the Cartesian design

$$\begin{gathered}\operatorname{logit}(p_\star) \in \{-.2, 0, .2\}, \quad s_\star = 1/2, \quad h_1 = -h_2 \in \{.25, .5, .75\}, \\ T \in \{.6, .8, 1\}, \quad c_0 \in \{0, 3\}. \end{gathered} \tag{273}$$

There are 54 conditions. Set $\varepsilon = .04$, $(x_0, y_0) = (1, 1)$, $\Pr(\zeta = T) = \Pr(\zeta = -T) = 1/2$, and $c_R = c_D = 1$. Each learner receives budget $B = 20(c_0 + 2)$: the mixture takes 20 updates, while each pure method takes 40 updates when $c_0 = 0$ and 25 when $c_0 = 3$. The recipes are fixed in advance: mixed step .6 and teacher fraction $1/2$, pure RL steps $6/(k + 3)$, and pure OPD steps two. The forecasts compare their exact sampled-coefficient reference risks using (64); there is no tuning or fitted boundary. Separately, the direct nonlinear moment verifier computes a mixed upper bound and lower bounds for every prescribed RL schedule through the pure count, steps in $[0, 2]$, and every affordable adaptive OPD rule with steps in $[0, 4]$ and stopping. Overlapping enclosures return "abstain."

The design was informed by analytical and exploratory results, so this is prospective evaluation rather than blinded discovery. Seeds, conditions, sample counts, uncertainty rules, and predictions were fixed before evaluation and were not adjusted to its outcomes.

Fresh nonlinear evaluation. Each condition uses $N = 262{,}144$ independent paired training replications. Common uniforms couple the three methods within a replication, but each samples

from its own evolving policy. The simulation evaluates raw regularized utilities and exact terminal KL regret, independently of the moment recursion. Repeated runs measure uncertainty and do not increase an individual learner's training allowance. The master seed is $(20260919, 270057)$, with condition and chunk indices appended; chunks contain 32,768 replications. All paired terminal losses are retained.

A proved range for every evaluated trajectory. Use the actual-logit rectangle $|x| \le X = .3$, $|y| \le Y = .1$. Throughout the design, the root probability is between $\sigma(-.5)$ and $\sigma(.5)$ and the leaf probability between $\sigma(-.1)$ and $\sigma(.1)$. Write $a_- = \sigma(-.5)$, $a_+ = \sigma(.5)$, $b_- = \sigma(-.1)$, and $b_+ = \sigma(.1)$. For RL steps through two, both sample multipliers are nonnegative since $2\max(a_+^2, b_+^2) < 1$. The convex-combination proof of (67), now at actual scale, applies because

$$Y^2/8 + Yb_+/a_- + .04/a_- < X, \qquad .04/b_- < Y. \tag{274}$$

Every teacher logit lies in this rectangle; its endpoint inequalities are checked with interval arithmetic for all 54 settings. OPD steps through four preserve it, and the normalized mixture averages two contained maps. The common initialization lies inside it. Induction therefore contains every trajectory, not merely the observed ones. Finally,

$$0 \le \mathcal{L}(x, y) \le (x^2 + y^2)/8 \le B_L = 1/80. \tag{275}$$

The bound is deliberately loose: it controls sampling uncertainty without assuming a favorable ordering.

Simultaneous finite-sample intervals. For each of the three pairs in each condition, form the paired loss difference D_i, with mean $\bar{D}$ and unbiased variance $\widehat{v}$. There are $m = 3 \cdot 54 = 162$ contrasts. Use

$$\bar{D} \pm \left\{ \sqrt{2\widehat{v}\log(4m/.05)/N} + \frac{7(2B_L)\log(4m/.05)}{3(N-1)} \right\}. \tag{276}$$

To prove coverage, rescale $D_i \in [-B_L, B_L]$ to $[0, 1]$ and apply Maurer & Pontil (2009, Theorem 4) to it and its complement, each with failure probability $.05/(2m)$. Rescaling restores the width $2B_L$; a union bound over the $2m$ tails gives simultaneous coverage at least .95. No normal approximation or independence across methods or conditions is needed. Independence across training replications is needed. An empirically supported winner has both corresponding intervals strictly in its favor. This statement concerns sampling error under the specified kernels; double-precision simulation is not an interval-arithmetic proof of the underlying expectations.

h_1	c_0	Forecasts (R/D/M)	Mean agreement	Supported forecast	Class certificates
.25	0	1/6/2	9/9	6/9	0/9
.25	3	0/6/3	9/9	8/9	0/9
.50	0	7/0/2	8/9	6/9	0/9
.50	3	2/0/7	9/9	6/9	6/9
.75	0	9/0/0	9/9	9/9	0/9
.75	3	6/0/3	9/9	9/9	0/9
All		25/12/17	53/54	44/54	6/54

Table 6: Complete prospective design, grouped only for readability. Each row contains all three root logits and noise amplitudes. R/D/M denote RL, OPD, and RL+OPD. Mean agreement is descriptive; supported forecasts use the simultaneous intervals (276). Class certificates instead prove nonlinear mixed superiority over the specified entire pure classes.

Outcomes and interpretation. The reference forecasts agree with 53 of 54 empirical mean winners; 44 forecasts are supported against both alternatives by the simultaneous intervals. All remaining ten are unresolved, not confirmed mistakes. The sole mean disagreement occurs at root logit .2, $h_1 = .5$, $T = 1$, and zero overhead. The reference favors RL, whereas the empirical mixed mean is lower by only $.0192 \cdot 10^{-6}$; the interval for mixed minus RL is $[-2.566, 2.528] \cdot 10^{-6}$, rounded outwards. Thus the study does not resolve that near-boundary ranking.

The stronger nonlinear certificate succeeds in six conditions: all three root logits with $h_1 = .5$, $T \in \{.8, 1\}$, and overhead three. Certified gaps exceed $2.86 \cdot 10^{-6}$ throughout these six point

conditions. All six also have empirical intervals favoring mixing over the two evaluated pure recipes. The other 48 conditions have overlapping analytic enclosures; none fails containment. No claimed pure winner is inferred from an abstention. Reference forecasts and direct certificates are complementary: the latter cover more baselines but decide fewer cases. The coarse radius in Theorem 4.1 alone does not certify this evaluation scale; the direct nonlinear bounds are essential here.

Numerical checks. The predictions are recomputed at 256-bit precision, and uncertainty intervals are reconstructed from raw paired losses for every condition, including inconclusive cases. Exact short-history and small-schedule enumerations independently check the reference moments and relaxation.

I SHARPER STRUCTURAL WINDOWS

We refine the sufficient conditions in Proposition 4.2 without weakening its baselines. We separate the methods' error allowances, characterize the noise interval without an auxiliary separating level, and use direct nonlinear moments when the general remainder is too large. RL still uses prescribed rates; OPD may adapt and stop. Nothing here extends the longer-budget result to adaptive RL.

I.1 SEPARATE ERROR ALLOWANCES AND AN EXACT TEACHER FLOOR

Use the parameters, containment conditions, and finite constants of Theorem 4.1. Fix the mixture before comparing methods. Write $L_N^{\mathrm{rec}} = (h_p \ell_{1,N} + p_\star h_s \ell_{2,N})/2$ for its coupled RL relaxation, $M = F_K(h) + D_K \nu$ for the mixed reference risk, and

$$D_\varepsilon = \frac{\mathrm{KL}(p_\star + \varepsilon h_p h_1 \| p_\star)}{\varepsilon^2}. \tag{277}$$

The normalization places all three quantities in the same reference units. Unlike $h_p h_1^2/2$, D_ε is the actual teacher floor divided by ε^2, not a quadratic approximation. The root noncrossing condition is still needed; an arbitrary teacher does not supply this floor.

Proposition I.1 (Direct-margin refinement)**.** *Under the stated conditions, suppose $N \geq 1$, $\nu > 0$, and both initial coordinates are nonzero. Choose either $L_N = Q_N(\nu)$ or $L_N = L_N^{\mathrm{rec}}$. For every $0 < \varepsilon \leq \varepsilon_0$,*

$$\inf_{a \in \mathcal{A}_R \cup \mathcal{A}_D} \mathbb{E}\mathcal{L}(\theta^a) - \mathbb{E}\mathcal{L}(\theta^M) \geq \varepsilon^2 G_\varepsilon,$$
$$G_\varepsilon = \min\{L_N - M - \varepsilon(C_R + C_M),\, D_\varepsilon - M - \varepsilon C_M\}. \tag{278}$$

Thus $G_\varepsilon > 0$ certifies dominance without choosing a common separating level or halving a reference margin. Moreover, $L_N^{\mathrm{rec}} \geq Q_N(\nu)$, so replacing the closed-form floor by the coupled relaxation cannot weaken this test.

RL and the mixture have separate approximation allowances, while OPD uses an exact floor. The RL error constant therefore does not enter the OPD comparison. This sharpens Proposition 4.2 without changing the learning rules or comparison classes.

Proof. The uniform reference comparison in Appendix B.1 gives, for each admissible prescribed RL schedule, $\mathbb{E}\mathcal{L}(\theta^R) \geq \varepsilon^2 L_N - \varepsilon^3 C_R$. The mixture satisfies $\mathbb{E}\mathcal{L}(\theta^M) \leq \varepsilon^2 M + \varepsilon^3 C_M$. The pathwise noncrossing argument gives $\mathcal{L}(\theta^D) \geq \varepsilon^2 D_\varepsilon$ at every permitted OPD stopping point. Subtract the same mixed upper bound from each pure lower bound, take their minimum, and then take the infimum over the pure rules. This proves (278). The bounds are uniform before taking the infimum; no selected or simulated schedule replaces the full comparison class.

For completeness, the coupled relaxation dominates the closed-form floor coordinate by coordinate. Its diagonal multiplier square is at least the square of the mean multiplier, and its root-to-leaf moment coupling is nonnegative. Dropping those terms, and minimizing over all nonnegative rates instead of the bounded interval, can only lower each one-step minimum. For curvature $c > 0$, the resulting map is $f_c(z) = z\nu/(\nu + cz)$, which is nondecreasing in $z \geq 0$. Induction from the common initialization therefore bounds each coupled moment below by the scalar recursion in Lemma B.1. Weighting and summing the two coordinates proves $L_N^{\mathrm{rec}} \geq Q_N(\nu)$.

Finally, if Proposition 4.2 holds with its given τ, δ, its reference pure floors are at least $\tau + \delta$, its mixed risk is at most $\tau - \delta$, and each normalized error is at most $\delta/2$. The RL branch of (278) is therefore at least δ. The exact OPD floor is at least its quadratic floor minus εt_D, so its branch is also at least δ. The direct test retains every certificate of that proposition. □

An explicit finite radius. One can retain a formula valid for every smaller positive scale without solving an exact-KL inequality. Set $D_0 = h_p h_1^2/2$ and

$$\varepsilon_{\mathrm{sep}} = \min\left\{\varepsilon_0, \frac{L_N - M}{C_R + C_M}, \frac{D_0 - M}{t_D + C_M}\right\}. \tag{279}$$

When both numerator margins are positive, every $0 < \varepsilon < \varepsilon_{\text{sep}}$ has a positive guaranteed gap. A zero denominator imposes no restriction when its numerator is positive. Indeed $D_\varepsilon \geq D_0 - \varepsilon t_D$, and substitution in (278) proves the claim. The strict endpoint matters: the guarantee may become zero at the displayed cap. This is a sufficient radius, not the largest radius at which mixing actually wins.

I.2 A Noise Interval and a Certificate-Level Budget Ceiling

The sharper recursive floor is useful for certification, whereas Q_N reveals the structure of the noise window. Fix teacher vector h, geometry, initialization, and the mixed recipe. Use one support bound T and uniform error allowances throughout $0 \leq \nu \leq T^2$; do not change these allowances as a function of ν in the following argument. Put $F = F_K(h) \geq 0$, $D = D_K > 0$, $e_R = \varepsilon(C_R + C_M)$, and $e_M = \varepsilon C_M$. The RL-side margin is

$$g_N(\nu) = Q_N(\nu) - F - D\nu - e_R. \tag{280}$$

It balances the increasing noise floor of pure RL against the mixture's linear noise cost. OPD imposes a separate upper noise threshold. Here $q_1 = h_p x_0^2$, $q_2 = p_\star h_s y_0^2$, and $Q_N(\nu) = \frac{1}{2}\sum_{i=1}^{2} q_i \nu/(\nu + Nq_i)$.

Proposition I.2 (Geometry of the closed-form sufficient window)**.** *Suppose $q_1, q_2 > 0$ and $N \geq 1$. Define $\nu_D = (D_\varepsilon - F - e_M)/D$. The direct test with $L_N = Q_N$ certifies precisely the noise values satisfying*

$$\nu \in (0, T^2], \qquad g_N(\nu) > 0, \qquad \nu < \nu_D, \tag{281}$$

subject to the common containment assumptions. This set is an interval, possibly empty. A necessary condition for it to be nonempty is

$$ND < 1. \tag{282}$$

If $ND < 1$, there is a unique $\widehat{\nu} > 0$ solving $Q'_N(\widehat{\nu}) = D$. Before intersecting with the teacher and support constraints, the positive interval exists if and only if $g_N(\widehat{\nu}) > 0$; its endpoints are the two roots of $g_N = 0$, allowing a left endpoint of zero when $F + e_R = 0$.

The test avoids choosing an auxiliary level τ: find the concave maximum, locate the two roots if it is positive, and apply the teacher and support caps. The budget ceiling describes this particular sufficient test. It does not prove that mixing is impossible when $ND \geq 1$, and it need not apply to the stronger coupled test.

Proof. The two branches of (278) are positive exactly when (281) holds. Differentiating the explicit floor, without any approximation, gives

$$Q'_N(\nu) = \frac{1}{2}\sum_{i=1}^{2} \frac{Nq_i^2}{(\nu + Nq_i)^2}, \qquad Q''_N(\nu) = -\sum_{i=1}^{2} \frac{Nq_i^2}{(\nu + Nq_i)^3} < 0. \tag{283}$$

Consequently g_N is strictly concave, $g_N(0) = -F - e_R \leq 0$, $g'_N(0) = 1/N - D$, and $g_N(\nu) \to -\infty$ as $\nu \to \infty$. The last fact follows from $Q_N(\nu) \leq (q_1 + q_2)/2$ and $D > 0$; it is only a property of the scalar function, not an asymptotic training result. If $ND \geq 1$, the derivative is nonpositive at zero and strictly decreasing thereafter. The margin is negative for every positive noise value. If $ND < 1$, the derivative starts positive and decreases continuously to $-D$, so it has one zero. Strict concavity makes this point the unique maximum. A positive maximum produces precisely one positive interval, with the stated endpoint convention; a nonpositive maximum produces none. Intersecting with two upper intervals preserves the interval property. The support and teacher constraints may still make this intersection empty even when the unconstrained maximum is positive. □

A closed-form illustration without fixed numerical constants. If $q_1 = q_2 = q > 0$, put $A = F + e_R$ and $B = q(1 - ND) - A$. Multiplication by the positive denominator $\nu + Nq$ gives

$$g_N(\nu) > 0 \quad \Longleftrightarrow \quad D\nu^2 - B\nu + NqA < 0. \tag{284}$$

Thus the unconstrained positive window is nonempty exactly when $B > 0$ and $B^2 > 4DNqA$, and its endpoints are

$$\nu_\pm = \frac{B \pm \sqrt{B^2 - 4DNqA}}{2D}. \tag{285}$$

The actual sufficient interval additionally requires $0 < \nu \le T^2$ and $\nu < \nu_D$. To verify the equivalence, expand $q\nu - (A + D\nu)(\nu + Nq)$; its coefficients are exactly the negative of the quadratic in (284). Since $A \ge 0$, a positive interval on the positive axis requires a positive sum of roots and a strictly positive discriminant, giving the two displayed tests. If $A = 0$, the smaller root is zero and the same strict tests apply.

The interpretation is explicit: larger mixed bias F or error allowance e_R decreases B; more affordable pure updates decrease it through ND and increase the constant term NqA. These statements hold with the mixture fixed. Changing its horizon or retuning its weight also changes F, D, so no universal monotonicity under arbitrary retuning is implied.

I.3 Continuous Parameter Windows at a Larger Finite Radius

Separate error allowances cannot repair an intrinsically loose nonlinear remainder. We therefore reuse the direct moment construction in Appendix C.4, now with variable model parameters. It follows the actual changing action probabilities and bounds nonlinear regret directly. This gives a guarantee uniform on a continuous parameter region, rather than a collection of pointwise checks.

What must hold on a parameter cell. On each closed cell, construct reachable state boxes for every pure step in $[0, 2]$ and the fixed mixture. Require the sample-map monotonicity tests used to obtain those boxes, action variances above $1/6$, positive noise-injection and moment denominators, and the root, leaf, and terminal moment monotonicity tests of Appendix C.4. Require valid teacher probabilities and $p_\star < q_A < p(\theta_0)$. Every test must hold uniformly on the entire cell. An unresolved sign is an abstention, not permission to use an endpoint approximation.

Proposition I.3 (Uniform direct-moment parameter certificate)**.** *Let a finite collection of closed parameter cells cover a region $\mathcal{P}$. On every cell C, suppose the preceding tests hold and the conditional-moment construction gives an RL lower bound $L_R(C)$ for every prescribed count through N, a mixed upper bound $U_M(C)$ for one common K-update recipe, and a pathwise OPD lower bound $L_D(C)$. Set*

$$g_{\mathcal{P}} = \min_C \{\min(L_R(C), L_D(C)) - U_M(C)\}. \tag{286}$$

If $g_{\mathcal{P}} > 0$, this common mixture has at least that much smaller expected nonlinear regret than every permitted pure rule at every parameter in $\mathcal{P}$. The comparison is equal-cost whenever

$$K(c_0 + c_R + c_D) \le \mathcal{B} < (N + 1)(c_0 + c_R), \qquad c_0 \ge 0, \quad c_R, c_D > 0. \tag{287}$$

OPD may use all its affordable updates with rates in $[0, 4]$ and current-iterate stopping. The pure RL rule may know the exact parameters within a cell; its rates and count must still be prescribed.

The cells partition parameter uncertainty, not learning-rate schedules. Each RL bound still covers a continuum of schedules. The cost condition lets pure RL buy additional updates; it does not charge the analyst's verification computation to the learner or equate allowances with GPU time.

Proof. Fix a parameter vector in $\mathcal{P}$ and choose a containing cell. The conditional identities (118) and (120) follow from Bernoulli moments at the current p, s. They remain identities for arbitrary optimal probabilities; the particular logit .1 is not required in their derivation. Replace each fixed optimal probability in the subsequent bounds by its interval enclosure on this cell.

Initialize the moment bounds from the common deterministic student. At each update, the verified root monotonicity permits substitution of the lower root moment in (124); the remaining leaf-moment dependence is affine, so both endpoints suffice. Minimize the resulting quadratic (126) at its vertex clipped to the full rate interval. For the leaf use (128) and its verified nonnegative coefficient. The two minimizing rates may differ; this relaxes the learner rather than excluding any shared-rate schedule. Upper second moments, a lower root mean, and a fourth-moment ratio propagate by (129)–(135). Finite interval partitions bound the latter inequality for all rates, including zero, not only for sampled rates. Induction thus preserves the moment bounds for every prescribed schedule and every parameter in the cell. Terminal monotonicity justifies (137); the minimum over every prefix, including zero updates, gives $L_R(C)$.

For the common mixture, expand the exact conditional raw-moment functions through degree three and enclose the degree-four remainder as in (138). Include zero in the derivative boxes, and bound

the absolute fourth moments by (139). These derivative bounds include the sampling probabilities as well as the update maps. Interval induction from the initial monomials therefore covers its actual raw moments; the identical finite expansion of terminal KL supplies a risk upper bound. Its intersection with the two independently valid structural enclosures remains valid. This is $U_M(C)$.

The higher-noise moments are bounded using support and variance through (131); no symmetry or finite support of the noise law is required. Parameter dependence discarded by interval arithmetic widens the bound but does not assert probabilistic independence. For OPD the monotone root update cannot cross its teacher, regardless of the number of permitted updates or the stopping rule. Hence its root KL is a pathwise floor and any uniform lower enclosure is $L_D(C)$. Subtracting $U_M(C)$ from the smaller pure lower bound proves the claim on this cell. The finite cover proves it on $\mathcal{P}$.

Finally, (287) makes the fixed mixture affordable and ensures that pure RL can buy at most N updates. The floor already covers every affordable OPD count. Thus the proof retains cost reallocation and independent pure-method tuning throughout the entire region. □

A verified continuous region. Let $u_\star = \mathrm{logit}(p_\star)$ and $v_\star = \mathrm{logit}(s_\star)$. Consider

$$\begin{gathered} u_\star, v_\star \in [.199, .201], \quad h_1 \in [.44, .46], \quad h_2 \in [-.51, -.49], \quad \nu \in [.63, .67], \\ \varepsilon = .04, \quad \theta_0 = \varepsilon(1, 1), \quad q_A = p_\star + \varepsilon h_p h_1, \quad q_B = s_\star + \varepsilon h_s h_2. \end{gathered} \tag{288}$$

Use fresh independent centered noise $\varepsilon\zeta$ with $|\zeta| \le 1$ and $\mathbb{E}\zeta^2 = \nu$. The same mixture takes $K = 20$ updates with $\lambda = .6, w = .5$ throughout this five-dimensional region. The direct certificate below covers every prescribed RL schedule through $N = 25$ with rates in $[0, 2]$ and adaptive OPD with rates in $[0, 4]$. Its complete cover of 612 verified parameter cells establishes

$$\inf_{a \in \mathcal{A}_R \cup \mathcal{A}_D} \mathbb{E}\mathcal{L}(\theta^a) - \mathbb{E}\mathcal{L}(\theta^M) > 5 \cdot 10^{-7}. \tag{289}$$

Costs $(c_0, c_R, c_D) = (3, 1, 1)$ and $\mathcal{B} = 100$ satisfy (287), giving 25 pure versus 20 mixed updates. More generally, the same certificate applies throughout that cost region.

For the cellwise teacher bound we use $\mathrm{KL}(q_A \| p_\star) \ge 2(q_A - p_\star)^2$. To see this directly, the divergence as a function of its first probability has second derivative $1/[q(1-q)] \ge 4$, with value and first derivative zero at $q = p_\star$; integrate twice. This avoids losing the shared $p_\star$ dependence in an interval evaluation of the exact KL formula. Also $\mathcal{L}(\theta_0) \le \varepsilon^2(1 + \sigma(.241))/8 < 3.125 \cdot 10^{-4}$. Thus the displayed gap exceeds 0.16% of initial regret. This is a modest certified improvement, not an estimate of the empirical gain or of the optimal policy's achievable risk.

The same bounds give a comparison in final-regret units: on every cell, $U_M(C) < .986 \min\{L_R(C), L_D(C)\}$. Since actual mixed risk is at most $U_M(C)$ and the best permitted pure risk is at least this positive minimum, the mixture has at least 1.4% lower expected final regret than the best permitted pure rule throughout the region. This percentage is an equal-cost guarantee after pure-method tuning, not a matched-update simulation gain.

I.4 What Changes in the Existing Parameter Grid?

We retrospectively apply the sharper certificates to the 54 settings in Appendix H.1. This analysis uses the existing experimental results and introduces no additional training runs. The simultaneous simulation intervals compare evaluated recipes, whereas the certificates cover the stated full classes. Those are different claims.

The first gain comes from removing slack, not from discovering a new mechanism. Using the exact teacher floor changes the numerical margin but adds no classifications on this grid. In the representative setting $h = (.45, -.5)$, $\nu = .65$, and $N = 25$, the coarse constants remain approximately $C_R = 42168$ and $C_M = 4819$. Even the separated radius (279) is only about $1.14 \cdot 10^{-7}$. Improved bookkeeping alone therefore does not solve the finite-radius problem.

At scale .04, the four certified settings all have overhead three. They are $(h_1, h_2, \nu) = (.45, -.5, .65)$, $(.45, -.5, 1)$, $(.45, 0, .65)$, and $(.6, -.5, 1)$, with respective gaps exceeding $4.34 \cdot 10^{-6}$, $8.77 \cdot 10^{-7}$, $1.12 \cdot 10^{-6}$, and $1.64 \cdot 10^{-6}$. These discrete settings have $p_\star = s_\star = .55$; they are distinct from the continuous logit region (288). All 162 radius checks, including overlapping enclosures and failed sufficient tests, are reported.

Criterion at $\varepsilon = 10^{-8}$	Positive certificates / 54	Simulation support
Prespecified common-buffer rule	2	2/2
Separate allowances, closed-form RL floor	5	5/5
Separate allowances, coupled RL floor	6	5/6
Coupled RL floor and exact OPD floor	6	5/6

Table 7: Retrospective comparison on the same 54 settings. Simulation support means that both existing pure-minus-mixed intervals are positive for the evaluated recipes. The sixth certificate's interval is inconclusive, not contradictory. The first row uses the prespecified buffer rule, not the best possible choice of its auxiliary parameters.

Scale	Positive direct-moment certificates / 54	Uncertified settings
$\varepsilon = .01$	6	48
$\varepsilon = .02$	6	48
$\varepsilon = .04$	4	50

Table 8: Theoretical radius sweep using the same normalized parameters, counts, and mixed recipe. The entries count theoretical certificates, not simulated wins, and do not certify all radii between the rows.

Verification and limitations. The interval calculation starts with two intervals per parameter axis and bisects unresolved cells, including both children in the resulting cover. Refinement uses only mathematical bounds, not simulated outcomes. Each RL step interval is covered by 256 subintervals for the fourth-moment inequality. Bounds are computed at 192-bit precision and checked at 256 bits for every grid point and final cell. Independent calculations check the complete cover, moment identities, floor ordering, and displayed counts.

Separate error allowances improve the general window; direct nonlinear bounds establish a continuous positive region at a larger scale, but not a sharp phase boundary, a uniform guarantee over unknown environments, or an adaptive-RL result. The region is a constructive theoretical example selected through analysis of the known model, not prospective empirical evidence. These certificate calculations do not constitute additional Qwen or controlled training experiments.

J Adaptive Pure Methods with Additional Affordable Updates

The main comparisons give pure RL two distinct freedoms: additional updates in Theorem 4.1, and adaptive rates and stopping in Proposition 4.3. Here we allow both at once. Theorem D.2 already supplies a general statewise framework for this class. We derive a complementary sufficient condition that quantifies how extra updates can erode an existing margin, then give a state-dependent refinement and a three-versus-two specialization. These are conditional guarantees: a positive three-versus-two parameter region has not been established, and no empirical claim follows from them.

J.1 The Unified Class and an Extra-Update Bound

Model and budget. Use the known nonlinear two-stage model and exact-baseline estimators of Appendix D.1. Fix the model parameters and a fresh bounded, centered noise law, independent of actions and history. On a compact domain $\mathcal{D}$, the loss $L = \mathcal{L}$ is continuous, bounded, and nonnegative. The domain is invariant under every allowed update. Write

$$P_\alpha f(\theta) = \mathbb{E}[f(\theta + \alpha \widehat{g}_R) \mid \theta], \qquad 0 \le \alpha \le \bar{a}_R \le 2. \tag{290}$$

The transition uses current visitation probabilities, not their values at the optimum. Compact controls, continuous sample maps and probabilities, and the fixed bounded noise law give the continuity properties used in (147). In particular, $P_0 f = f$.

Under one allowance budget $\mathcal{B}$, set

$$N = \left\lfloor \frac{\mathcal{B}}{c_0 + c_R} \right\rfloor, \quad N_D = \left\lfloor \frac{\mathcal{B}}{c_0 + c_D} \right\rfloor, \quad 1 \le K \le \left\lfloor \frac{\mathcal{B}}{c_0 + c_R + c_D} \right\rfloor, \qquad c_0 \ge 0,\ c_R, c_D > 0. \tag{291}$$

Let $\Pi_R(N)$ contain all predictable scalar rates in $[0, \bar{a}_R]$ and stopping times at most N; let $\Pi_D(N_D)$ use rates in $[0, 4]$ and stopping times at most N_D. Rates may use the full observed history, known model, and pre-draw randomization. Both return the current iterate, not a retrospectively selected checkpoint. Other choices, including estimators, momentum, and batch sizes, remain fixed as in the main text. Let

$$R_R^\star(n) = \inf_{\pi \in \Pi_R(n)} \mathbb{E} L(\theta_\tau^\pi), \qquad R_D^\star(N_D) = \inf_{\pi \in \Pi_D(N_D)} \mathbb{E} L(\theta_\tau^\pi). \tag{292}$$

The auxiliary class $\Pi_R(m)$ for $m \le N$ has the same information and controls, but a smaller update cap. It is an analytical restriction, not a different estimator or cost model.

Fix one affordable mixed controller, and let $R_M = \mathbb{E} L(\theta_K^M) \le U_M$. The controller can be a constant-rate recipe or a specified feedback rule; these should not be conflated. For a parameter-region claim, the same controller must work throughout the region, while each pure learner may know the exact parameters.

Reachable states and extra progress. Let deterministic compact sets $\mathcal{D}_j \subseteq \mathcal{D}$ contain all possible RL states after j updates. Require $\theta_0 \in \mathcal{D}_0$ and that every permitted transition from $\mathcal{D}_j$ stays in $\mathcal{D}_{j+1}$ almost surely. Since zero is an allowed rate, $\mathcal{D}_j \subseteq \mathcal{D}_{j+1}$. Taking every set equal to $\mathcal{D}$ is valid, but often conservative. For $0 \le m \le N$, choose nonnegative bounds

$$d_j \ge \sup_{\theta \in \mathcal{D}_j,\, 0 \le \alpha \le \bar{a}_R} \{L(\theta) - P_\alpha L(\theta)\}, \qquad C_{m,N} = \sum_{j=m}^{N-1} d_j. \tag{293}$$

An empty sum is zero. Each d_j bounds the largest conditional expected improvement from one additional update, including a rate chosen after observing the current state. It is not a bound on the realized improvement of every sampled update.

Theorem J.1 (Extending an adaptive comparison to extra updates)**.** *Under the preceding assumptions, let $\underline{r}_m \le R_R^\star(m)$. Suppose $p_\star \le q_A \le p(\theta_0)$, and put $D = \mathrm{KL}(q_A \| p_\star)$. Then*

$$\begin{aligned} R_R^\star(N) &\ge [\underline{r}_m - C_{m,N}]_+, \\ \min\{R_R^\star(N), R_D^\star(N_D)\} - R_M &\ge \min\{[\underline{r}_m - C_{m,N}]_+, D\} - U_M =: \Delta_{m,N}, \end{aligned} \tag{294}$$

where $[z]_+ = \max\{z, 0\}$. *In particular, strict mixed dominance follows if*

$$D > U_M, \qquad C_{m,N} < \underline{r}_m - U_M. \tag{295}$$

The conclusion covers all history-dependent rates and current-iterate stopping in both pure classes, including their additional affordable updates. Uniform assumptions and a uniformly positive $\Delta_{m,N}$ *give the same claim throughout a parameter region for one common mixed controller.*

The margin over RL must exceed the progress its extra updates could make. The comparison with OPD is unchanged because its teacher floor holds at every permitted stopping point. Crucially, $\underline{r}_m$ must already cover adaptive RL; a lower bound for prescribed schedules is insufficient.

Proof. Fix an arbitrary rule $\pi \in \Pi_R(N)$ with stopping time τ. After it stops, set every remaining rate to zero and keep its state fixed. This padding is only a proof device: it neither changes the output nor requires additional queries. Denote the padded states by θ_j. They satisfy $\theta_N = \theta_\tau$ and lie in $\mathcal{D}_j$ by induction.

Truncate the original rule at $\min\{\tau, m\}$. This is an admissible member of $\Pi_R(m)$, even if its rates were designed with the larger horizon in mind: the shorter class allows every predictable choice with the same information. Its output is θ_m. Consequently

$$\mathbb{E}L(\theta_m) \geq R^\star_R(m) \geq \underline{r}_m. \tag{296}$$

Condition on the complete history and any randomization used to choose α_j, before sampling the next episode. Before stopping, the fixed model gives the conditional transition P_{α_j}. After stopping, $\alpha_j = 0$ and the loss is unchanged. In both cases,

$$\mathbb{E}[L(\theta_j) - L(\theta_{j+1}) \mid \mathcal{F}_j] = L(\theta_j) - P_{\alpha_j} L(\theta_j) \leq d_j. \tag{297}$$

Here $\mathcal{F}_j$ includes the pre-draw rate randomization. We have not moved a random learning rate outside an expectation. Taking expectations, summing from m to $N-1$, and using boundedness to justify these finite expectations yields

$$\mathbb{E}L(\theta_\tau) = \mathbb{E}L(\theta_N) \geq \mathbb{E}L(\theta_m) - C_{m,N} \geq \underline{r}_m - C_{m,N}.$$

Nonnegativity of L supplies the positive part. Since π was arbitrary, the same lower bound holds after taking its infimum.

For OPD, the root-logit map is $z \mapsto z + \alpha(q_A - \mathrm{sigmoid}(z))$. Its derivative is $1 - \alpha p(1-p) \geq 0$ for $\alpha \leq 4$. It fixes the teacher logit and maps every logit above it into the interval between that logit and the teacher. Starting with $p(\theta_0) \geq q_A$, induction gives $p_j \geq q_A \geq p_\star$, regardless of feedback or stopping. Bernoulli KL is increasing above $p_\star$, and the leaf loss is nonnegative. Thus $L(\theta^D_\tau) \geq D$ pathwise for every affordable OPD rule. Taking the smaller pure lower bound and subtracting $R_M \leq U_M$ proves (294).

Because $U_M \geq R_M \geq 0$, the strict inequalities in (295) make both pure lower bounds greater than U_M. For a region, apply this argument at each parameter value with the same mixed controller and take a uniform positive lower bound on the gap. □

J.2 Bounding Extra Progress with Conditional Moments

The supremum in (293) can be bounded without solving a full adaptive-control problem. Use the quadratic envelopes $Q_- \leq L \leq Q_+$ from (150) on the invariant rectangle. Write $Q_-(\theta) = \theta^\top H_- \theta/2$ and

$$a_-(\theta) = -\theta^\top H_- \mathbb{E}[\widehat{g}_R \mid \theta], \qquad b_-(\theta) = \mathbb{E}[\widehat{g}_R^\top H_- \widehat{g}_R \mid \theta]. \tag{298}$$

These are explicit nonlinear functions in (151). For the same rate cap, set

$$\Psi_-(\theta) = \max_{0 \leq \alpha \leq \bar{a}_R} \{\alpha a_-(\theta) - \alpha^2 b_-(\theta)/2\}. \tag{299}$$

This is the largest expected reduction of the lower quadratic in one step. It uses the actual conditional second moment, including sampled actions and learner-dependent visitation.

Proposition J.2 (A computable one-step progress envelope)**.** *For every state and rate whose successors remain in $\mathcal{D}$,*

$$L(\theta) - P_\alpha L(\theta) \le c(\theta),$$
$$c(\theta) = \min\{L(\theta),\, L(\theta) - Q_-(\theta) + \Psi_-(\theta)\} \ge 0. \tag{300}$$

Hence $d_j = \sup_{\theta\in\mathcal{D}_j} c(\theta)$, or any certified upper bound on that supremum, is valid in Theorem J.1. A coarser alternative replaces $L(\theta) - Q_-(\theta)$ by $Q_+(\theta) - Q_-(\theta)$.

The allowance separates possible quadratic progress from the error in approximating current regret. Using the known loss directly avoids paying for an upper quadratic when it is unnecessary. At a fixed state and geometry, larger reward-noise variance increases b_- without changing a_-, so it cannot increase Ψ_-. This observation does not imply monotonicity of the full comparison: noise also changes reachable states and mixed risk.

Proof. For any selected rate, expand the quadratic before taking the conditional expectation:

$$P_\alpha Q_-(\theta) = Q_-(\theta) + \alpha\theta^\top H_-\mathbb{E}[\widehat{g}_R \mid \theta] + \frac{\alpha^2}{2}\mathbb{E}[\widehat{g}_R^\top H_-\widehat{g}_R \mid \theta] = Q_-(\theta) - \alpha a_-(\theta) + \alpha^2 b_-(\theta)/2.$$

The lower envelope is valid at every successor, so $P_\alpha L \ge P_\alpha Q_-$. Subtracting from the exact current loss gives

$$L - P_\alpha L \le L - Q_- + \alpha a_- - \alpha^2 b_-/2 \le L - Q_- + \Psi_-.$$

Also $P_\alpha L \ge 0$, hence $L - P_\alpha L \le L$. Taking the smaller upper bound proves (300). Both entries of the minimum are nonnegative because $L \ge Q_-$ and rate zero is allowed. The claimed coarser bound follows from $L \le Q_+$.

For completeness, H_- is positive definite on the compact logistic rectangle. If $b_- = 0$, the sampled update vanishes almost surely, so $a_- = 0$ and $\Psi_- = 0$. Otherwise the maximizing rate is

$$\alpha_-^\star(\theta) = \mathrm{clip}(a_-(\theta)/b_-(\theta); [0, \bar{a}_R]).$$

For $a_- \le 0$ the maximum is zero. For $0 < a_- < \bar{a}_R b_-$ it is $a_-^2/(2b_-)$; for $a_- \ge \bar{a}_R b_-$ it is $\bar{a}_R a_- - \bar{a}_R^2 b_-/2$. Thus the envelope requires only the displayed moments and scalar clipping. Any interval implementation must enclose whole state cells, not just evaluate a grid of states or rates. □

Why the uniform bound may be loose. Zero rates keep the initial state reachable at every time. A supremum over reachable states can therefore charge a large improvement available near initialization to a learner that has already made substantial progress. Summing such suprema can also combine improvements whose maximizing states cannot occur along one trajectory. The next result keeps this state dependence instead of charging every prefix the same sum.

J.3 State-Dependent Continuation Bounds

Write $r(\theta, \alpha) = L(\theta) - P_\alpha L(\theta)$, which may be negative. For the remaining updates $j = m, \dots, N-1$, choose bounded continuous functions $W_j \ge 0$ on $\mathcal{D}$, with $W_N = 0$, satisfying

$$W_j(\theta) \ge r(\theta, \alpha) + P_\alpha W_{j+1}(\theta) \quad \text{for every allowed } \alpha,\ \theta \in \mathcal{D}_j. \tag{301}$$

The function W_j bounds the improvement available over the entire remaining horizon from a given state. Negative immediate progress is retained because a temporarily harmful step can improve later opportunities.

Define the Bellman operator $Tf = \inf_{0\le\alpha\le\bar{a}_R} P_\alpha f$. For a terminal function f, $T^m f$ means m successive applications, with $T^0 f = f$. Set

$$B_m(\theta) = [L(\theta) - W_m(\theta)]_+, \qquad B_j = TB_{j+1} \quad (j = m-1, \dots, 0). \tag{302}$$

The modified terminal loss credits each prefix state only with the continuation improvement available there. The prefix is then optimized for this modified loss; it is not assumed to be optimal for the shorter training horizon.

Proposition J.3 (A state-dependent refinement)**.** *Under the unified-class assumptions and* (301)*,*

$$R_R^\star(N) \ge B_0(\theta_0), \qquad \min\{R_R^\star(N), R_D^\star(N_D)\} - R_M \ge \min\{B_0(\theta_0), D\} - U_M. \tag{303}$$

If also $W_m(\theta) \le C_{m,N}$ *on* $\mathcal{D}_m$*, then*

$$B_0(\theta_0) \ge [R_R^\star(m) - C_{m,N}]_+ \ge [\underline{r}_m - C_{m,N}]_+. \tag{304}$$

Thus the refinement is no weaker than the uniform bound whenever its continuation envelope satisfies the stated cap.

This result can exploit the fact that the states with the largest remaining improvement need not be the states with the smallest prefix regret. Computing the Bellman functions is not automatic: the proposition specifies what must be bounded, not a claim that a finite state grid is exact.

Proof. Pad an arbitrary stopped RL rule with zeros as before. Insert each predictable chosen rate into (301). The conditional transition law gives

$$L(\theta_j) - W_j(\theta_j) \le \mathbb{E}[L(\theta_{j+1}) - W_{j+1}(\theta_{j+1}) \mid \mathcal{F}_j].$$

Successive conditioning from $N-1$ down to m, using $W_N = 0$, yields

$$\mathbb{E}[L(\theta_N) \mid \mathcal{F}_m] \ge L(\theta_m) - W_m(\theta_m).$$

The left side is also nonnegative, so it is at least $B_m(\theta_m)$. For each prefix step, the infimum defining B_j implies $B_j(\theta_j) \le \mathbb{E}[B_{j+1}(\theta_{j+1}) \mid \mathcal{F}_j]$. Iterating through the prefix proves $\mathbb{E}L(\theta_N) \ge \mathbb{E}B_m(\theta_m) \ge B_0(\theta_0)$. This argument includes stopped paths and pre-draw randomization. The OPD floor and mixed upper bound give (303).

To prove the ordering, the cap gives $B_m \ge L - C_{m,N}$ on every reachable prefix state. The operator T preserves inequalities and commutes with addition of a constant. Propagating through the m reachable prefix transitions gives $B_0(\theta_0) \ge (T^m L)(\theta_0) - C_{m,N}$. Under the compactness and continuity assumptions, the selector argument for (147) gives $(T^m L)(\theta_0) = R_R^\star(m)$. To see why stopping is included, $TL \le L$ by rate zero, and induction makes $T^n L$ nonincreasing in n. Hence the additional stopping branch in that Bellman equation does not change its value. Equivalently, every stopped policy can be padded with zeros. Finally, $B_m \ge 0$ implies $B_0 \ge 0$, proving the positive parts in (304). □

Exact continuation credit and the limits of the refinement. Let $V_n = T^n L$. The smallest continuation envelope is

$$W_j^\star = L - V_{N-j}, \qquad W_N^\star = 0, \qquad W_j^\star = \sup_{0 \le \alpha \le \bar{a}_R} \{r(\cdot, \alpha) + P_\alpha W_{j+1}^\star\}. \tag{305}$$

Indeed substitution cancels $P_\alpha L$ and leaves $L - \inf_\alpha P_\alpha V_{N-j-1} = L - V_{N-j}$. Nonnegativity follows from $V_n \le L$. Backward induction shows that any envelope satisfying (301) dominates $W_j^\star$ on the corresponding reachable sets. For $W_m = W_m^\star$, the prefix terminal function is V_{N-m} and $B_0 = V_N$ exactly. This identity explains how to recover the full adaptive comparison; it does not provide a closed-form value or a positive mixed gap.

Certified approximation errors. A numerical value approximation needs a uniform residual bound. Suppose bounded Borel functions $v_0, \dots, v_N$ satisfy $v_0 \le L$, $v_n \le L$, and

$$v_n(\theta) \le P_\alpha v_{n-1}(\theta) + e_n \quad \text{for every state and rate,} \qquad e_n \ge 0.$$

Then $R_R^\star(N) \ge [v_N(\theta_0) - \sum_{n=1}^N e_n]_+$. To verify this, set $\widetilde{v}_n = v_n - \sum_{i=1}^n e_i$. These corrected functions satisfy the stopping and transition inequalities of Theorem D.2; nonnegative regret supplies the positive part. Residuals must cover interpolation, integration, parameter uncertainty, and optimization over all rates. A calculation on selected states or rates alone does not establish these inequalities.

J.4 Three Adaptive Pure Updates Versus Two Mixed Updates

The first unequal-count case retains a two-update mixed controller but allows pure RL three adaptive updates. Its exact cost condition is

$$\max\{2(c_0 + c_R + c_D),\, 3(c_0 + c_R)\} \le \mathcal{B} < \min\{3(c_0 + c_R + c_D),\, 4(c_0 + c_R)\}. \tag{306}$$

This gives $N = 3$ and a mixed cap of two; OPD retains its full affordable cap. For example, $(c_0, c_R, c_D) = (1, 1, 1)$ and $\mathcal{B} = 6$ give three pure updates versus two mixed updates. This is a budget example, not a certified positive learning instance.

The two-update adaptive-RL lower bound from Proposition 4.3 is

$$\underline{r}_2 = \inf_{0 \le \alpha \le \bar{a}_R} \{F_{Q_-}(\alpha) - \mathcal{A}_{Q_-}(\alpha)\}. \tag{307}$$

Its derivation concerns the capped learning process and therefore remains a valid auxiliary two-update bound when the actual budget permits three. It does not become a three-update bound without a further argument.

Corollary J.4 (A three-versus-two sufficient condition)**.** *Assume* (306)*, the preceding model and teacher-floor conditions, and a fixed two-update mixture with upper bound U_M. Let d_2 be a valid bound from Proposition J.2. If*

$$D > U_M, \qquad d_2 < \underline{r}_2 - U_M, \tag{308}$$

then the mixture has smaller expected final regret than every adaptive RL rule with at most three updates and every affordable adaptive OPD rule. The guaranteed gap is at least $\min\{[\underline{r}_2 - d_2]_+, D\} - U_M > 0$.

The additional requirement is explicit: even the most useful third RL update must not close its two-update gap to the mixture. This accounts for adaptation and the extra update simultaneously, rather than assuming that one of these freedoms is harmless.

Proof. The lower and upper budget inequalities in (306) give $3 \le \mathcal{B}/(c_0 + c_R) < 4$ and $2 \le \mathcal{B}/(c_0 + c_R + c_D) < 3$, respectively. Taking floors proves the stated counts. The proof of Proposition 4.3 bounds all two-update feedback rules independently of how the cap was financed, so (307) is admissible in Theorem J.1. Substitute $m = 2$, $N = 3$, and $C_{2,3} = d_2$. The OPD floor already covers every affordable count. The strict inequalities give the positive gap. □

A direct state-dependent alternative. The same conditional moments give a one-update lower-value function

$$z_1(\theta) = [Q_-(\theta) - \Psi_-(\theta)]_+.$$

Since $L - c = [Q_- - \Psi_-]_+$ in (300), take $W_2 = c$, $W_3 = 0$ in the state-dependent construction. It follows that

$$R^\star_R(3) \ge (T^2 z_1)(\theta_0). \tag{309}$$

Each application of T optimizes the next rate conditionally before averaging. Valid lower enclosures for the two nested optimizations can therefore replace $[\underline{r}_2 - d_2]_+$ in the comparison. Moreover, if $d_2 \ge \sup_{\mathcal{D}_2} c$, this construction is no weaker than that uniform bound by Proposition J.3. It avoids replacing the last-step opportunity by its largest value at all reachable states, but still requires certified continuum bounds.

What would establish a reversal? An inconclusive lower bound does not show that adaptive RL wins. To certify the opposite ordering, fix an affordable adaptive RL rule and prove its risk is at most U_R. If the specified mixture has a lower bound L_M and $U_R < \min\{L_M, D\}$, that rule outperforms the mixture and every permitted OPD rule. This follows directly by comparing actual risks. It does not exclude a different mixed controller from doing better.

Scope of the extension. The inequalities above unify the comparison classes at the level of finite-budget sufficient conditions. They do not establish positivity of an unequal-count adaptive certificate. In particular, the positive prescribed-schedule regions in Appendix I cannot be imported into this class, and the two-update certificate in Appendix D.3 does not control the benefit of a third RL update. A positive regional claim requires a verified continuation bound, common mixed controller, containment, and cost inequalities on every parameter cell.