ONLINE EXTRACTION OF PRIVILEGED SIGNALS FOR ASYMMETRIC ACTOR-CRITIC

Anonymous authors
Paper under double-blind review

ABSTRACT

Asymmetric actor-critic methods let the critic, but not the actor, use privileged information during training; which information the critic sees is decided offline, in Informed Asymmetric Actor-Critic (IAAC) by how well it predicts returns on episodes of a fixed policy. We ask what that information does to the actor's update as the policy changes, and whether choosing it online helps. We split what a signal tells the critic into an offset shared by all actions and a contrast between them, and derive the update cost of each. The offset removes noise, but its interaction with the contrast can reverse the benefit: with two actions, an independent signal that shifts the return of the action the policy commits to makes an exact value baseline's update noisier once that action's probability exceeds two thirds, although the signal predicts returns better. A signal's value is thus not stationary, and we propose Online Informed Actor-Critic (OIAC), which re-selects the critic's input during training by a paired confidence bound on the update noise and keeps cross-fitted Monte Carlo updates conditionally unbiased. For bootstrapped advantages, the future offsets add back what the baseline removes, so a one-step actor receives mostly the contrast. In finite informed POMDPs solved exactly, the signal that IAAC's criterion selects lowers the second moment of a Monte Carlo update by up to 64% but raises that of a one-step update by up to 210%. In tabular learning, only OIAC among four selection criteria dropped a signal once it turned harmful; in that example it learned 13% faster than the offline choice under normalized steps but stayed slower than a critic without privileged information, and it was not faster under plain SGD, whereas the exact variance-optimal baseline made the same signal useful.

1 INTRODUCTION

Asymmetric actor-critic methods train an actor on the public history while the critic may also see privileged information that is available only during training (Pinto et al., 2018; Baisero & Amato, 2022; Lambrechts et al., 2024; Ebi et al., 2026). In informed partially observable Markov decision processes (POMDPs), an exact critic conditioned on any state-dependent signal leaves the policy gradient unbiased (Ebi et al., 2026, Thm. 3.1), and a well-chosen signal lowers the variance of the critic's value targets. Informed Asymmetric Actor-Critic (IAAC) therefore selects the critic's signal offline, by how much it improves the critic's return prediction on episodes of a fixed policy (Ebi et al., 2026, §4).

The actor, however, does not use the critic's prediction; it uses the critic's *advantage*, where the critic serves as the baseline subtracted at the current step and, in temporal-difference methods, as the bootstrap that stands in for the future return. In this paper, we split what a signal tells the critic into an *offset*, shared by all actions, and a *contrast*, which separates them, and derive how each changes the noise of the actor's update. As a baseline, the offset removes noise, but its score-weighted interaction with the contrast can reverse the benefit: with two actions, an independent signal that shifts the return of the action the policy commits to makes a value baseline's update noisier once that action's probability exceeds $\frac{2}{3}$, although the signal predicts returns better. The value of a signal is therefore not stationary, and a signal chosen offline can turn harmful as the policy commits. IAAC's authors name criteria that account for direct policy effects and signal selection within the learning loop as future work (Ebi et al., 2026, §6); we make both concrete with Online Informed Actor-Critic (OIAC), which re-evaluates the critic's input during training by the update noise and

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.

keeps cross-fitted Monte Carlo updates conditionally unbiased. As a bootstrap, the future offsets add back what the baseline removed, so a one-step advantage keeps mostly the contrast.

We test the analysis in finite informed POMDPs, where dynamic programming gives the exact return, gradient and update noise of every policy; these exact and tabular experiments isolate the mechanism and test signal selection and update noise, not performance with neural critics. The identities hold to machine precision; the reversal appears along a learning path, and the signal that IAAC's criterion selects helps a Monte Carlo update and, once the policy commits, hurts a one-step one. In learning, OIAC is the only one of four selection criteria that drops a signal once it turns harmful, and whether this shortens training depends on the step rule, consistent with the analysis.

Our contributions are as follows.

- We decompose the actor's update cost into offset and contrast for the baseline and the advantage form, derive from classical baseline weights an explicit sign condition for the value baseline, and show when the offsets cancel under bootstrapped targets (Section 3).
- We propose OIAC, which changes the critic's input only when a paired lower confidence bound on a baseline-form noise statistic clears a switching threshold, and give sufficient conditions under which combining this selection with a K-fold baseline (Kota et al., 2017) keeps Monte Carlo updates unbiased (Section 4).
- We test the analysis exactly and in learning, against offline, return-based selection, a history-only critic, the full privileged critic and an exact oracle schedule (Section 5).

2 Background and Related Work

2.1 Informed POMDPs and Asymmetric Critics

An informed POMDP (Lambrechts et al., 2024; Ebi et al., 2026) is a POMDP whose state s_t also emits an information variable $i_t \sim I(\cdot \mid s_t)$, with the observation drawn from i_t so that $o_t \perp s_t \mid i_t$. The actor $\pi_\theta(a_t \mid h_t)$ sees only the public history $h_t = (o_0, a_0, \ldots, o_t)$; during training, a critic may also see i_t or an *extraction* $z_t = \phi(i_t, h_t)$ of it, such as a subset of its coordinates. With discount factor γ, the return is $G_t = \sum_{k \geq 0} \gamma^k r_{t+k}$ and the objective is $J(\theta) = \mathbb{E}_{\pi_\theta}[G_0]$. For every extraction ϕ, the policy gradient can be written with the exact informed advantage $A_\phi = Q_\phi - V_\phi$ of Section 3 (Ebi et al., 2026, Thm. 3.1; Lemma A.3):

$$\nabla_\theta J = \mathbb{E}\Big[\sum_t \gamma^t A_\phi(h_t, z_t, a_t)\, \nabla_\theta \log \pi_\theta(a_t \mid h_t)\Big]. \tag{1}$$

IAAC estimates the advantage with the one-step temporal-difference error of a learned critic $\hat{V}_\phi$,

$$\delta_t = r_t + \gamma \hat{V}_\phi(h_{t+1}, z_{t+1}) - \hat{V}_\phi(h_t, z_t). \tag{2}$$

It assesses candidate signals on episodes of a fixed policy, independently of the actor's training: a residual-independence test before training, or the gain in held-out return prediction afterwards, retains feature subsets (Ebi et al., 2026, §§4.2–4.3).

2.2 Related Work

Privileged critics support sim-to-real transfer (Pinto et al., 2018), unbiased policy gradients (Baisero & Amato, 2022), asymmetric imitation (Warrington et al., 2021), informed conditioning (Lambrechts et al., 2024; Ebi et al., 2026) and analyses with guarantees (Lambrechts et al., 2025; Cai et al., 2024); the state may also act as an exploration signal (Baisero, 2025). Dual critics, Le Critique and annealing schedules adapt how much the actor trusts a fixed privileged critic (Li et al., 2024; Venkatraman et al., 2026; Haklidir, 2026). OIAC instead adapts which signal the critic sees.

The variance-optimal baseline weights each action by its squared score (Greensmith et al., 2004; Peters & Schaal, 2008), with two actions by swapping their probabilities (Weaver & Tao, 2001), also with information the policy does not see (Mao et al., 2019; Liu et al., 2026), and baselines also change how fast a softmax policy commits (Chung et al., 2021; Mei et al., 2022); centralized critics can even increase the variance of multi-agent policy gradients (Lyu et al., 2021; 2023). Our sign rule

concerns the value baseline that asymmetric methods actually fit, and shows when a more predictive signal increases its update noise.

A critic offset acts on a TD(λ) advantage like a shaping potential (Ng et al., 1999; Schulman et al., 2016), and the difference between the advantages of two critics is known (Liu et al., 2026, App. D); we add when the offset cancels and what a learned informed critic's error does to the actor. TD(λ) targets interpolate between the one-step and the Monte Carlo form (Sutton, 1988).

3 Offset and Contrast in the Actor's Update

In this section, we analyze what a privileged signal does to the actor's update. We fix a policy $\pi = \pi_\theta$ and a step t. A *candidate* is an extraction ϕ that turns the privileged information into the critic's input $z = \phi(i_t, h_t)$, and we write

$$Q_\phi(h, z, a) = \mathbb{E}[G_t \mid h, z, a], \qquad V_\phi(h, z) = \mathbb{E}[G_t \mid h, z], \qquad A_\phi = Q_\phi - V_\phi, \tag{3}$$

with the index 0 for the history-only critic. The actor samples a_t from $\pi(\cdot \mid h_t)$ with its own randomness, so a_t is independent of z given h_t (Lemma A.2); the results on offset and contrast use only this fact, beyond the standing assumptions of Appendix A. $S = \nabla_\theta \log \pi_\theta(a_t \mid h_t)$ is the score, $\omega = \|S\|^2$ its squared norm, and $\bar{s}^2 = \mathbb{E}_{a\sim\pi}\|S(a)\|^2$.

3.1 Offset and Contrast

We measure what a signal adds to the critic by two differences at the same (h, z, a),

$$\Delta_V = V_\phi(h, z) - V_0(h), \qquad \Delta_A = A_\phi(h, z, a) - A_0(h, a). \tag{4}$$

We refer to Δ_V as the *offset* and to Δ_A as the *contrast*. The offset says how much better the situation is than the history alone suggests; the contrast says how much the signal changes the value of the chosen action relative to the others. The return-prediction gain that IAAC's criterion estimates (Ebi et al., 2026, Def. 4.2) is their sum,

$$\Gamma(\phi) = \mathbb{E}[(G_t - Q_0)^2] - \mathbb{E}[(G_t - Q_\phi)^2] = \mathbb{E}[\Delta_V^2] + \mathbb{E}[\Delta_A^2], \tag{5}$$

and it never decreases when the signal is refined (Proposition A.4).

3.2 Update Cost and the Sign Rule

A critic enters the per-sample update in the *advantage form* or in the *baseline form*,

$$g_{\text{adv}} = S\, A_\phi, \qquad g_{\text{base}} = S\,(G_t - V_\phi). \tag{6}$$

With exact critics both have the same mean for every ϕ (Lemma A.3), so a signal changes only the noise. We call the change in the second moment,

$$\Delta M(\phi) = \mathbb{E}\|g_\phi\|^2 - \mathbb{E}\|g_0\|^2, \tag{7}$$

the signal's *update cost*; it is negative when the signal makes the update less noisy.

Theorem 3.1 (Update cost by estimator form)**.** *With exact critics,*

$$\Delta M_{\text{adv}}(\phi) = \mathbb{E}\big[\omega\, \Delta_A^2\big], \qquad \Delta M_{\text{base}}(\phi) = -\,\mathbb{E}\big[\omega\, \Delta_V^2\big] - 2\,\mathbb{E}\big[\omega\, \Delta_V \Delta_A\big]. \tag{8}$$

Proof. See Appendix A. □

In the advantage form the offset is free and the contrast always costs, a counterpart of the Q-form result that a centralized critic cannot lower the variance (Lyu et al., 2021, Thm. 2). In the baseline form the offset always helps, and the contrast enters through its score-weighted correlation with the offset, whose sign depends on which action's return the signal predicts.

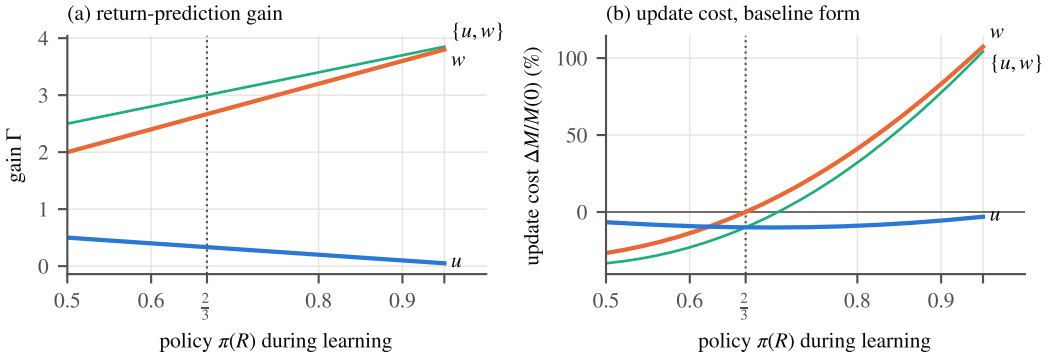


Figure 1: One decision in which independent signals u and w shift the returns of L and R (Example 3.3, $\kappa_L = 1$, $\kappa_R = 2$), as the policy moves from uniform to $\pi(R) = 0.95$. **(a)** Return-prediction gain: the gain of w grows as the policy commits to R. **(b)** Update cost in the baseline form, relative to the history-only second moment; positive values make the update noisier. w turns harmful at $\pi(R) = \frac{2}{3}$ (dotted), while u helps throughout. Exact values for a tabular softmax actor.

Proposition 3.2 (Sign rule). *Fix a history h and let the return be $G = \tilde{G} + \mathbf{1}\{a = \bar{a}\}\, x$ for one action $\bar{a}$, where the signal x seen by the critic is independent of $(a, \tilde{G})$ given h, with mean 0 and variance κ^2. Then $\Gamma(x) = \pi(\bar{a})\kappa^2$ at this history, and*

$$\Delta M_{\text{base}}(x) = -\kappa^2\, \pi(\bar{a})^2 \big(2\|S(\bar{a})\|^2 - \bar{s}^2\big). \tag{9}$$

With two actions, for every policy parameterization,

$$\Delta M_{\text{base}}(x) = -\kappa^2 \pi(\bar{a})^2 \|S(\bar{a})\|^2\, \frac{2 - 3\pi(\bar{a})}{1 - \pi(\bar{a})},$$

so for $\kappa > 0$ and $S(\bar{a}) \neq 0$ the signal reduces the update noise while $\pi(\bar{a}) < \frac{2}{3}$ and increases it once $\pi(\bar{a}) > \frac{2}{3}$.

Proof. See Appendix A. □

We use the following informed POMDP as the running example; Section 5 returns to it.

Example 3.3 (One decision). The actor chooses L or R once, with a constant observation. Hidden bits $u, w \in \{\pm 1\}$ and a nuisance $n \in \{1, \ldots, 6\}$ are uniform and independent, and the return is

$$G = \begin{cases} \kappa_L\, u + \varepsilon, & a = L, \\ 1 + \kappa_R\, w + \varepsilon, & a = R, \end{cases} \qquad \varepsilon \sim \mathcal{N}(0, 1). \tag{10}$$

The critic may see any subset of (u, w, n). The signals u and w are of the kind in Proposition 3.2, for the actions L and R, and n carries no information. The optimal policy plays R, whose return-prediction gain $\pi(R)\kappa_R^2$ grows as the policy commits to it.

Figure 1 shows the reversal along a learning path. The mechanism is the baseline's weights. The value baseline and the variance-optimal baseline (Greensmith et al., 2004; Mao et al., 2019) follow the signal with the weights

$$\pi(\bar{a}) \qquad \text{and} \qquad \nu(\bar{a}) = \frac{\pi(\bar{a})\, \|S(\bar{a})\|^2}{\bar{s}^2},$$

and for two actions $\nu(\bar{a})$ is the other action's probability (Weaver & Tao, 2001), so $\nu(\bar{a}) < \pi(\bar{a})/2$ once $\pi(\bar{a}) > \frac{2}{3}$. Proposition 3.2 is thus an explicit sign condition for one signal model that follows from classical baseline theory, not a new variance principle. Under the variance-optimal baseline more information cannot increase the update noise (Liu et al., 2026, Prop. 4); in the same two-action setting, with an index that shifts the optimal arm's return, Liu et al. (2026, App. A) show that conditioning the value baseline on it tightens an upper bound on its second moment; Proposition 3.2

shows that the second moment itself can increase. The variance-optimal baseline is dominated by the values of rarely taken actions and can make softmax learning slower or less stable (Chung et al., 2021; Liu et al., 2026, App. A).

For plain SGD, update noise enters the expected progress of a step only at second order in the step size, and for a normalized step it can enter at first order (Proposition A.5; cf. Cutkosky & Mehta, 2020, Lemma 2); we expect Adam (Kingma & Ba, 2015) to lie in between but do not prove this.

3.3 Bootstrapped targets

IAAC's actor uses the one-step target $r_t + \gamma V_\phi(h_{t+1}, z_{t+1})$, so the informed critic enters the update twice: as the baseline at t and as the bootstrap at $t+1$. Let $\hat{A}_t^\lambda(V)$ denote the TD(λ) advantage of a critic V, which is the one-step advantage at $\lambda = 0$ and the baseline form at $\lambda = 1$.

Proposition 3.4 (What the informed critic adds to a bootstrapped advantage)**.** *For any critics $V_\phi(h, z)$ and $V_0(h)$, any $\lambda \in [0, 1]$ and every trajectory (the identity is known; Liu et al., 2026, App. D),*

$$\hat{A}_t^\lambda(V_\phi) - \hat{A}_t^\lambda(V_0) = -\Delta_t + \gamma(1-\lambda)\sum_{k\geq 1}(\gamma\lambda)^{k-1}\Delta_{t+k}, \tag{11}$$

where $\Delta_j = V_\phi(h_j, z_j) - V_0(h_j)$.

If, until the step R at which the history learns the outcome that the signal concerns, the offset only discounts forward ($\Delta_{t+k} = \gamma^{-k}\Delta_t$ for $t + k < R$), and the two critics agree from then on ($\Delta_j = 0$ for $j \geq R$), the right-hand side equals $-\lambda^{R-1-t}\Delta_t$.

Proof. See Appendix A.5. □

The baseline removes the current offset and the bootstrap adds the future offsets back. With one-step targets they cancel except in the step before the history learns the signal, whereas $\lambda = 1$ keeps the whole offset; the contrast survives through the next offset, which depends on the action whenever the signal is about the decision. With exact critics, a Markov signal and rewards that the history records, Proposition A.8 gives the one-step form's update cost as

$$\Delta M_{\mathrm{TD}}(\phi) = \Delta M_{\mathrm{base}}(\phi) + \gamma^2\,\mathbb{E}\big[\omega\,\Delta_{t+1}^2\big], \tag{12}$$

and the added term equals the offset's saving $\mathbb{E}[\omega\Delta_t^2]$ when the offset only discounts forward.

Applied to a learned critic $\hat{V}_\phi$ against the exact V_ϕ, Equation (11) also gives the bias that the critic's error $\varepsilon_j = \hat{V}_\phi(h_j, z_j) - V_\phi(h_j, z_j)$ puts into the actor's update when the critic is fitted independently of the rollout it evaluates (Appendix A.5):

$$\mathbb{E}\big[S_t\big(\hat{A}_t^\lambda(\hat{V}_\phi) - \hat{A}_t^\lambda(V_\phi)\big)\big] = \gamma(1-\lambda)\sum_{k\geq 1}(\gamma\lambda)^{k-1}\,\mathbb{E}\big[S_t\,\varepsilon_{t+k}\big]. \tag{13}$$

It enters only through the bootstrap and vanishes at $\lambda = 1$. The informed critic is therefore worth most to a one-step actor as an accurate bootstrap, not as a lower-variance baseline.

4 Online Informed Actor-Critic

In this section, we propose Online Informed Actor-Critic (OIAC), which chooses the critic's input during training. Since a signal's update cost moves with the policy (Proposition 3.2), a choice made offline, on episodes of a fixed policy, can turn wrong as the policy commits. OIAC starts from IAAC's choice and splits each block of B on-policy episodes into two folds. It fits every candidate ψ's critic on one fold, as in the K-fold baseline of Kota et al. (2017), and scores each episode e of the other fold by the baseline-form statistic

$$T_\psi(e) = \sum_t \gamma^{2t}\,\omega_t\big(G_t - \hat{V}_\psi(h_t, z_t)\big)^2, \tag{14}$$

the sum of the squared per-step updates of the episode. Over the last W blocks it compares each challenger ψ with the candidate in force $\hat{\phi}$ on the same episodes, and it switches, from the next

block on, only when the lower confidence bound on the paired reduction $T_{\hat{\phi}} - T_{\psi}$ exceeds c times the mean of $T_{\hat{\phi}}$ (Algorithm 1 in Appendix B). The actor's update in block b,

$$\hat{g}_b = \frac{1}{B} \sum_e \sum_t \gamma^t S_t \big(G_t - \hat{V}_{\hat{\phi}_b}(h_t, z_t)\big), \tag{15}$$

uses the candidate $\hat{\phi}_b$ chosen on earlier blocks, and this keeps it unbiased.

Proposition 4.1 (Online selection keeps cross-fitted Monte Carlo updates unbiased)**.** *Let $\mathcal{F}_{b-1}$ contain the data of the blocks before b and any randomness of the selection rule. Suppose the candidate used in block b and the parameters θ_b are $\mathcal{F}_{b-1}$-measurable, the episodes of block b are drawn independently from π_{θ_b} given $\mathcal{F}_{b-1}$ and assigned to folds without looking at their data, and each episode is scored by a critic that depends only on $\mathcal{F}_{b-1}$ and the other fold. Then $\mathbb{E}[\hat{g}_b \mid \mathcal{F}_{b-1}] = \nabla J(\theta_b)$, whatever the selection rule.*

Proof. See Appendix A.6. □

The switching bound is nominal: our guarantee concerns the unbiasedness of the update, not the coverage of the bound (Appendix B). The cross-fitted baseline is not new; Proposition 4.1 states sufficient conditions under which it stays conditionally unbiased when combined with a selection of the critic's input on earlier blocks, and claims neither that the selection is optimal nor that it improves learning. OIAC's statistic has the baseline form, which is the form of the Monte Carlo update of Equation (15); for an actor that bootstraps, Proposition 3.4 implies that the offsets it rewards largely cancel, so the statistic would credit a signal with savings that such an actor does not receive. Our experiments therefore use Monte Carlo updates.

5 Experiments

In this section, we test the analysis in finite informed POMDPs, where dynamic programming gives the exact return, gradient and update noise of every policy. We compare four criteria for the critic's input, which form a 2×2:

- **IAAC** (offline, return-based): the feature subset with the largest significant held-out return-prediction gain on 512 episodes of the uniform policy, chosen before training (our implementation of IAAC's test, with value critics; the action-value gain used in IAAC's reference code, re-implemented in our learner, gives the same choice on 239 of 240 seed cases, Appendix E.1);
- **offline, update-based**: the same procedure scored by OIAC's statistic;
- **online, return-based**: OIAC's rule with the held-out return-prediction error in place of the update statistic;
- **OIAC** (online, update-based; Section 4).

The references are the history-only critic, the full privileged critic, the variance-optimal baseline of Section 3.2, and an *exact oracle schedule* that uses, at every iteration, the candidate with the smallest exact update second moment.

5.1 Synthetic informed POMDPs

We use two finite informed POMDPs, summarized in Table 1. The one-decision example (Example 3.3) isolates the sign rule; with $\kappa = (1, 2)$ the signal w on the committed action R carries the larger contrast, and with $\kappa = (2, 1)$ the smaller one. The three-step family adds later decisions at which the same signals are offsets, so that the offline choice can be right.

Example 5.1 (Three steps)**.** A branch decision $b \in \{L, R\}$ at $t = 0$ is followed by two steps at which the actor must match a public target bit c_t, drawn uniformly and shown in the observation. Hidden bits $u, w, v \in \{\pm 1\}$ are drawn once and a nuisance $n_t \in \{1, 2, 3\}$ at every step, the reward at $t = 0$ is 0, and for $t \geq 1$ the reward in branch b is

$$r_t = \alpha_b + \kappa_b\, x_b + \tfrac{1}{2}\, v + \mathbf{1}\{a_t = c_t\} + \varepsilon_t, \qquad x_L = u,\ x_R = w, \tag{16}$$

Table 1: The two synthetic informed POMDPs. Candidates are the masks of the privileged vector (in brackets: under a budget of at most one feature); the threshold is the return at which learning counts as done.

	one decision (Example 3.3)	three steps (Example 5.1)
decisions per episode	1	3
privileged vector	(u, w, n)	(u, w, v, n_t, b)
role of the signals	contrasts u (on L), w (on R); n uninformative	u, w contrasts at the branch, offsets later; v offset; n_t, b uninformative
candidates	8 (4)	32 (6)
episodes per iteration B	32 (also 128)	64
optimal / uniform return	1 / 0.5	2.6 / 1.3
learning threshold	0.95, i.e. $\pi(R) \geq 0.95$	2.47
IAAC's choice	$\{u, w\}$	$\{u, w, v\}$

Table 2: Update cost of the signal that IAAC's return-prediction criterion selects in the three-step example (exact critics), in percent of the history-only second moment at the same λ, as the policy commits (p: probability of the optimal action at each step). Negative values: the signal makes the actor's update less noisy.

p	$\lambda = 1$ (Monte Carlo)	$\lambda = 0.9$	$\lambda = 0.5$	$\lambda = 0$ (one step)
0.50	-63.5	-60.1	-39.8	-6.2
0.60	-61.1	-56.9	-31.3	$+9.3$
0.70	-54.6	-48.6	-11.8	$+43.2$
0.80	-44.0	-35.1	$+19.5$	$+96.7$
0.95	-20.4	-5.2	$+89.8$	$+210.3$

with $\alpha_L = 0$, $\alpha_R = 0.3$, $\kappa_L = 2$, $\kappa_R = 1$ and $\varepsilon_t \sim \mathcal{N}(0, 1)$. The critic may see any subset of (u, w, v, n_t, b), where the branch b is already public and so adds no information. At the branch step u and w are contrasts; at the later steps they, and v throughout, are offsets.

Learner and protocol. The actor is a tabular softmax over history keys, trained with Monte Carlo baseline-form updates on B fresh episodes per iteration and cross-fitted cell-mean critics, under Adam, a normalized step and plain SGD, with learning rates locked beforehand on the history-only critic (Appendix C). We report the episodes to the learning threshold and the exact update noise $\|\hat{g} - \nabla J\|^2$ over 60 development seeds, and 20 confirmation seeds run once after every setting was locked; all predictions were registered before the runs.

5.2 Exact values

Each identity of Section 3 that we checked holds to within 1.3×10^{-14} (Appendix D). In the one-decision example, the signal w on the action that the policy commits to lowers the value baseline's update noise until $\pi(R) = \frac{2}{3}$ and raises it by 107% at $\pi(R) = 0.95$, while its return-prediction gain keeps growing (Figure 1). In the three-step example, the signal that IAAC's criterion selects lowers the second moment of a Monte Carlo update at every policy, by up to 64%, but makes the one-step update noisier from $p = 0.6$ on, by 210% once the policy commits (Table 2), as Propositions 3.4 and A.8 explain.

5.3 Which criterion drops the harmful signal

With $\kappa = (1, 2)$, IAAC and the offline, update-based criterion select $\{u, w\}$ on every seed, because at the uniform policy both signals reduce the update noise (Figure 1b); the right target alone does not help. The online, return-based criterion moves the wrong way: once u predicts little but still costs critic cells, it drops u and keeps w on all 60 seeds. Only OIAC drops w, on 59 of 60 seeds under Adam, at a median $\pi(R)$ of 0.84 against 0.67 for the exact oracle schedule (Figure 2a), which lowers the late-phase update noise by 15% (Table 3), also on the confirmation seeds (Table 8). Because rare actions carry the harm, few episodes dominate the statistic: OIAC drops w late and re-admits it past

Table 3: Four selection criteria in the one-decision example ($\kappa_L = 1$, $\kappa_R = 2$), 60 seeds, with three reference runs. *Drops* w: median $\pi(R)$ at which the run drops w (Adam). *Late noise*: exact update error once $\pi(R) > \frac{2}{3}$, relative to the history-only critic (Adam). *Episodes*: episodes to the learning threshold ($\times 10^3$) under each step rule. Bold: best of the four criteria in a column.

				episodes ($\times 10^3$)		
run	criterion	drops w	late noise	normalized	Adam	SGD
no budget						
IAAC	offline, return	never	1.43	2.28	4.39	4.07
	offline, update	never	1.43	2.28	4.39	4.07
	online, return	never	1.39	2.24	4.38	4.09
OIAC	online, update	0.84	**1.21**	**1.99**	**4.27**	4.07
history-only	–	–	1.00	1.84	4.23	3.95
full privileged	–	never	1.42	2.14	4.53	4.05
exact oracle schedule	–	0.67	1.05	1.89	4.20	3.95
at most one feature						
IAAC	offline, return	never	1.32	2.12	4.32	4.01
	offline, update	never	1.32	2.12	4.32	4.01
	online, return	never	1.32	2.12	4.32	4.01
OIAC	online, update	0.80	**1.10**	**1.89**	**4.18**	**3.96**
exact oracle schedule	–	0.63	1.04	1.85	4.16	3.95

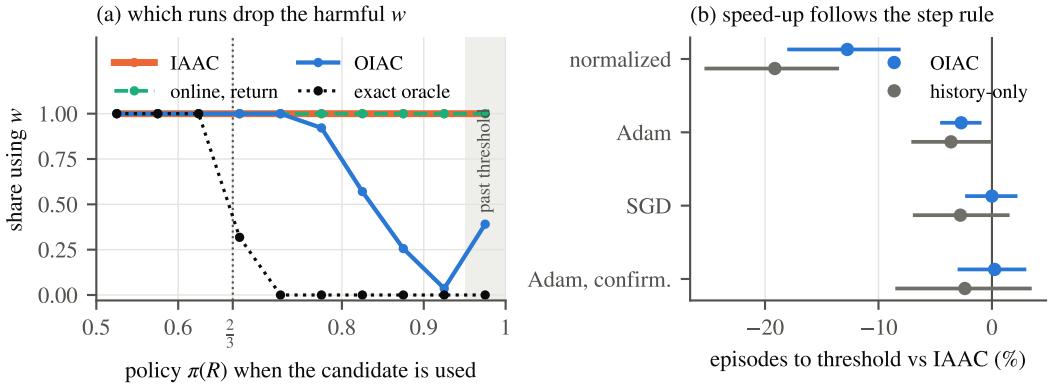


Figure 2: One-decision example with $\kappa = (1, 2)$, 60 seeds. **(a)** Share of iterations whose critic observes w, by the policy at which the candidate is used (Adam). IAAC and the online, return-based criterion keep w throughout; OIAC drops it after the exact oracle does and re-admits it past the threshold (shaded). **(b)** Difference in episodes to threshold against IAAC, in percent of IAAC's mean, with paired 95% intervals; negative values favor the run. The last row repeats Adam on the confirmation seeds.

the learning threshold on most seeds. With $\kappa = (2, 1)$, where the harm of w is four times smaller, OIAC drops w on 33 of 60 seeds under Adam; in the three-step family, whose signals are offsets at the later steps, it mostly keeps IAAC's choice, which there beats the history-only critic (a tie with IAAC under the normalized step, 1% slower under Adam; Appendix E.1).

5.4 Learning speed and the step rule

Whether dropping the signal shortens training depends on the step rule, consistent with Proposition A.5 (Figure 2b). With $\kappa = (1, 2)$, IAAC's signal makes learning slower than no signal at all: the history-only critic needs 19% fewer episodes than IAAC under the normalized step and 4% fewer under Adam. OIAC recovers most of that loss under the normalized step, reaching the threshold 13% sooner than IAAC, but it stays 8% slower than the history-only critic, and even the exact oracle schedule is at most 2% faster than that critic. Under plain SGD, differences in update noise

Table 4: Tabular study, 60 seeds per cell: change in episodes to threshold against IAAC, in percent of IAAC's mean (negative is faster; †: the 95% paired interval contains zero), and the late-phase update error of the pooled score-weighted baseline relative to IAAC. *In-batch* estimates $b^\star$ from the current batch only; *pooled* estimators keep per-action return averages across iterations and reweight them with the current policy, by ν for $b^\star$ and by π for the value baseline V (the matched control); *exact* $b^\star$ is computed by dynamic programming.

environment	step	B	history	in-batch	pooled V	**pooled** $b^\star$	exact $b^\star$	late error
one decision, $\kappa{=}(1,2)$	Adam	32	-4	-4	-6	-12	-13	0.39
one decision, $\kappa{=}(1,2)$	normalized	32	-19	-13	-10	-33	-34	0.40
one decision, $\kappa{=}(2,1)$	Adam	32	$+2^\dagger$	$+4$	-5	-13	-14	0.38
one decision, $\kappa{=}(2,1)$	normalized	32	$+1^\dagger$	$+11$	-9	-33	-34	0.43
one decision, $\kappa{=}(1,2)$	Adam	128	$+1^\dagger$	$-0^\dagger$	$+0^\dagger$	-2	-2	0.45
one decision, $\kappa{=}(1,2)$	normalized	128	-9	-8	-4	-10	-10	0.45
one decision, $\kappa{=}(2,1)$	Adam	128	$+3$	$+2$	$+1^\dagger$	-2	-2	0.46
one decision, $\kappa{=}(2,1)$	normalized	128	$-2^\dagger$	$-1^\dagger$	-4	-9	-9	0.48
three steps	Adam	64	$+8$	$+7$	-16	-19	-24	0.32
three steps	normalized	64	$+4$	$+7$	-21	-32	-35	0.35

did not translate into detectable differences in episodes to threshold, and under Adam OIAC's 3% gain on the development seeds did not replicate on the confirmation seeds (Table 8). Selection can thus avoid a harmful signal, but it cannot make the signal useful.

5.5 The variance-optimal baseline

OIAC addresses harmful input selection for a value baseline; the variance-optimal baseline tests whether changing the baseline's weights can instead make the same information useful, since under it more information cannot increase the update noise. Computed exactly, it reached the threshold 2–35% sooner than IAAC in all ten cells with Adam or normalized steps, and pooled per-action estimates recovered 98% of that reduction in the one-decision cells (Table 4); at fixed policies the signal lowers the second moment under $b^\star$ (0.51 against 0.79 without it at $\pi(R) = 0.9$, relative to the history-only value baseline; Appendix E.2). Estimated inside each batch, it was faster than IAAC in some cells and slower in others, because it is dominated by the values of rarely taken actions; under plain SGD it brought no speed-up, consistent with Proposition A.5 (Appendix E.2).

6 Conclusion

In this paper, we showed that what a privileged signal does to the actor's update is not what it does to the critic's prediction. As a baseline, a signal's offset removes noise and its interaction with the contrast can undo that, so a signal that predicts returns better can make the update noisier as the policy commits; as a bootstrap, the offsets cancel under the conditions of Proposition 3.4. The value of a signal is therefore not stationary. OIAC re-evaluates it online by the update noise, keeps cross-fitted Monte Carlo updates unbiased, and was the only one of four criteria that dropped a signal once it turned harmful; whether that shortens training depends on the step rule, and a variance-optimal baseline, which makes the same signal useful, gained more than any selection rule.

Limitations and future work. The update-cost results assume exact critics, and Proposition 4.1 covers cross-fitted Monte Carlo updates only. OIAC's statistic is dominated by rare-action contributions, so it detects a harmful signal late and can re-admit it; its confidence bounds are nominal, and exhaustive subset selection uses 2^d candidates for d features. Our evidence is exact and tabular by design, with a softmax actor, Monte Carlo targets and small candidate families. The main open question is learned critics and bootstrapped actors on standard informed-POMDP benchmarks such as IAAC's: there, Proposition 3.4 predicts that a baseline-form statistic scores the wrong quantity, and that the critic's error at the bootstrap, rather than its input, decides what the actor receives. A selection statistic in the form of the bootstrapped update, history-dependent extractions $\phi(i, h)$ and estimated variance-optimal baselines are natural next steps.

Reproducibility statement

The formal results are proved in Appendix A, and the identities are checked against exact dynamic programming in Appendix D. The tabular learner, the selection rules and the seeds are specified in Appendices C and B; code will be released.

AI usage disclosure

In this work, we used generative AI tools (large language models) to help develop the theoretical framework, formulate the mathematical claims, provide ingredients of their proofs and write the proofs, propose and refine hypotheses, design and give feedback on the experiments, implement the methods and the experiment code, and interpret the results. We have not used generative AI tools to generate synthetic data sets: all data come from running the learners in the environments specified in Appendix C. Translation, data-set cleaning and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to identify, search for and summarize related literature, to draft and edit parts of the paper, to create the code for the figures and tables, and to format the references. We have reviewed all AI-assisted work: the identities used in the proofs were checked numerically against an exact dynamic-programming oracle (Appendix D), and the bibliographic entries were checked against the publishers' records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Andrea Baisero. Role of state in partially observable reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Doctoral Consortium*, pp. 2917–2919. IFAAMAS, 2025.

Andrea Baisero and Christopher Amato. Unbiased asymmetric reinforcement learning under partial observability. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 44–52. IFAAMAS, 2022.

Yang Cai, Xiangyu Liu, Argyris Oikonomou, and Kaiqing Zhang. Provable partially observable reinforcement learning with privileged information. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63790–63857. Curran Associates, Inc., 2024.

Wesley Chung, Valentin Thomas, Marlos C. Machado, and Nicolas Le Roux. Beyond variance reduction: Understanding the true impact of baselines on policy optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1999–2009. PMLR, 2021.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2260–2268. PMLR, 2020.

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

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

Mehmet Haklidir. When does adaptive guidance help? Belief-aware privileged distillation for autonomous driving under partial observability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 907–915, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Nithyanand Kota, Abhishek Mishra, Sunil Srinivasa, Xi Chen, and Pieter Abbeel. A K-fold method for baseline estimation in policy gradient algorithms. *arXiv preprint arXiv:1701.00867*, 2017.

Jakub Grudzien Kuba, Muning Wen, Linghui Meng, Shangding Gu, Haifeng Zhang, David Mguni, Jun Wang, and Yaodong Yang. Settling the variance of multi-agent policy gradients. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13458–13470. Curran Associates, Inc., 2021.

Gaspard Lambrechts, Adrien Bolland, and Damien Ernst. Informed POMDP: Leveraging additional information in model-based RL. *Reinforcement Learning Journal*, 2:763–784, 2024.

Gaspard Lambrechts, Damien Ernst, and Aditya Mahajan. A theoretical justification for asymmetric actor-critic algorithms. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32375–32405. PMLR, 2025.

Jinqiu Li, Enmin Zhao, Tong Wei, Junliang Xing, and Shiming Xiang. Dual critic reinforcement learning under partial observability. In *Advances in Neural Information Processing Systems*, volume 37, pp. 116676–116704. Curran Associates, Inc., 2024.

Zhenya Liu, Yang Meng, Zhuokai Zhao, Xuefeng Liu, and Yuxin Chen. Shared actors need not share critics: Effects of value mismatch in parallel reinforcement learning. *arXiv preprint arXiv:2608.26481*, 2026.

Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. Contrasting centralized and decentralized critics in multi-agent reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 844–852. IFAAMAS, 2021.

Xueguang Lyu, Andrea Baisero, Yuchen Xiao, Brett Daley, and Christopher Amato. On centralized critics in multi-agent reinforcement learning. *Journal of Artificial Intelligence Research*, 77:295–354, 2023.

Hongzi Mao, Shaileshh Bojja Venkatakrishnan, Malte Schwarzkopf, and Mohammad Alizadeh. Variance reduction for reinforcement learning in input-driven environments. In *International Conference on Learning Representations*, 2019.

Jincheng Mei, Wesley Chung, Valentin Thomas, Bo Dai, Csaba Szepesvári, and Dale Schuurmans. The role of baselines in policy gradient optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17818–17830. Curran Associates, Inc., 2022.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pp. 278–287. Morgan Kaufmann, 1999.

Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, 2008.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In *Robotics: Science and Systems XIV*, 2018.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pp. 1057–1063. MIT Press, 2000.

Siddarth Venkatraman, Matthieu Dinot, and Laurence Aitchison. Le Critique: Privileged value functions for LLM reinforcement learning. *arXiv preprint arXiv:2608.16739*, 2026.

Andrew Warrington, J. Wilder Lavington, Adam Ścibior, Mark Schmidt, and Frank Wood. Robust asymmetric learning in POMDPs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11013–11023. PMLR, 2021.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 538–545. Morgan Kaufmann, 2001.

A PROOFS

In this appendix, we prove the results of Sections 3 and 4 and state the variance-optimal baseline used in Section 5. Results of the main text are restated before their proofs.

A.1 SETTING AND TWO LEMMAS

Assumption A.1. The informed POMDP of Section 2 has finite action and observation spaces, a finite horizon H and $\gamma \in (0, 1]$; states, information and rewards may be continuous. Let $\mathcal{H}_t$ be generated by $(s_{0:t}, i_{0:t}, o_{0:t}, a_{0:t-1}, r_{0:t-1})$, everything that has happened before the action at step t.

(A1) The actor samples with its own randomness: $a_t = \mathsf{a}_\theta(h_t, \varepsilon_t)$ for i.i.d. $\varepsilon_0, \dots, \varepsilon_{H-1}$ that are independent of the environment.

(A2) Rewards, returns and signals are square integrable.

(A3) $\pi_\theta(a \mid h) > 0$, and $\theta \mapsto \pi_\theta(a \mid h)$ is continuously differentiable.

We fix a step t and write $(h, z, a, G) = (h_t, z_t, a_t, G_t)$ with $z = \phi(i_t, h_t)$. The score is $S = \nabla_\theta \log \pi_\theta(a \mid h)$ and $\omega = \|S\|^2$; $S(a')$ and $\omega(a')$ denote their values at an action a', so that $\bar{s}^2 = \sum_{a'} \pi(a' \mid h)\, \omega(a')$.

Lemma A.2 (The actor's choice carries no information about the signal)**.** $\Pr(a \in \cdot \mid \mathcal{H}_t) = \pi(\cdot \mid h)$. *Hence $a \perp z \mid h$ for every extraction ϕ, and $\mathbb{E}[S \mid \mathcal{H}_t] = 0$.*

Proof. By (A1), for every bounded f,

$$\mathbb{E}[f(a) \mid \mathcal{H}_t] = \mathbb{E}\big[f(\mathsf{a}_\theta(h, \varepsilon_t)) \mid \mathcal{H}_t\big] = \sum_{a'} \pi(a' \mid h)\, f(a'),$$

which is a function of h alone, while z is $\mathcal{H}_t$-measurable. For the score,

$$\mathbb{E}[S \mid \mathcal{H}_t] = \sum_{a'} \pi(a' \mid h)\, \nabla_\theta \log \pi(a' \mid h) = \nabla_\theta \sum_{a'} \pi(a' \mid h) = 0. \qquad \square$$

The tower property and Lemma A.2 give

$$\mathbb{E}[Q_\phi \mid h, a] = Q_0, \qquad \mathbb{E}[V_\phi \mid h, a] = V_0, \qquad \mathbb{E}[A_\phi \mid h, z] = 0, \qquad \mathbb{E}[A_0 \mid h, z] = 0,$$

where the second and the fourth identity use $a \perp z \mid h$. For the offset and the contrast of Equation (4) this means

$$\mathbb{E}[\Delta_V \mid h, a] = 0, \qquad \mathbb{E}[\Delta_A \mid h, a] = 0, \qquad \mathbb{E}[\Delta_A \mid h, z] = 0. \tag{17}$$

Lemma A.3 (Every candidate gives the same mean update)**.** *For every extraction ϕ, $\mathbb{E}[S\, A_\phi] = \mathbb{E}[S\, (G - V_\phi)] = \mathbb{E}[S\, G]$, and $\nabla_\theta J = \sum_{t<H} \gamma^t\, \mathbb{E}[S_t G_t]$.*

Proof. Given h, S is a function of a and V_ϕ a function of z, which are conditionally independent (Lemma A.2); hence $\mathbb{E}[S\, V_\phi \mid h] = \mathbb{E}[S \mid h]\, \mathbb{E}[V_\phi \mid h] = 0$. Since S is a function of (h, a), $\mathbb{E}[S\, Q_\phi] = \mathbb{E}[S\, G]$. The last claim is the likelihood-ratio policy gradient (Sutton et al., 2000); rewards before t drop out because $\mathbb{E}[S_t \mid \mathcal{H}_t] = 0$. $\square$

A.2 OFFSET AND CONTRAST

Proposition A.4 (Return-prediction gain)**.** *The gain of Equation (5) is $\Gamma(\phi) = \mathbb{E}[(Q_\phi - Q_0)^2] = \mathbb{E}[\Delta_V^2] + \mathbb{E}[\Delta_A^2]$, and it never decreases when the signal is refined: if ϕ is a function of ϕ', then $\Gamma(\phi') \geq \Gamma(\phi)$.*

Proof. The residual $G_t - Q_\phi$ is orthogonal to every square-integrable function of (h, z, a), in particular to $Q_\phi - Q_0$, so

$$\mathbb{E}\big[(G_t - Q_0)^2\big] = \mathbb{E}\big[(G_t - Q_\phi)^2\big] + \mathbb{E}\big[(Q_\phi - Q_0)^2\big].$$

Write $Q_\phi - Q_0 = \Delta_V + \Delta_A$. Since Δ_V is a function of (h, z), Equation (17) gives $\mathbb{E}[\Delta_V \Delta_A] = \mathbb{E}\big[\Delta_V \, \mathbb{E}[\Delta_A \mid h, z]\big] = 0$. For the monotonicity, let $z = f(z')$ with $z' = \phi'(i, h)$. Then $Q_\phi = \mathbb{E}[Q_{\phi'} \mid h, z, a]$, so $Q_{\phi'} - Q_\phi$ is orthogonal to $Q_\phi - Q_0$ and

$$\Gamma(\phi') = \mathbb{E}\big[(Q_{\phi'} - Q_\phi)^2\big] + \Gamma(\phi) \geq \Gamma(\phi). \qquad \square$$

Theorem 3.1 (Update cost by estimator form)**.** *With exact critics,*

$$\Delta M_{\text{adv}}(\phi) = \mathbb{E}\big[\omega \, \Delta_A^2\big], \qquad \Delta M_{\text{base}}(\phi) = -\,\mathbb{E}\big[\omega \, \Delta_V^2\big] - 2\,\mathbb{E}\big[\omega \, \Delta_V \Delta_A\big]. \tag{8}$$

Proof. Advantage form. By Equation (17),

$$\mathbb{E}\big[A_\phi^2 \mid h, a\big] = \mathbb{E}\big[(A_0 + \Delta_A)^2 \mid h, a\big] = A_0^2 + \mathbb{E}\big[\Delta_A^2 \mid h, a\big].$$

Multiplying by ω, a function of (h, a), and taking expectations gives $\Delta M_{\text{adv}}(\phi) = \mathbb{E}[\omega \, \Delta_A^2]$.

Baseline form. Decompose the two residuals as

$$G - V_\phi = (G - Q_\phi) + A_\phi, \qquad G - V_0 = (G - Q_\phi) + (Q_\phi - Q_0) + A_0.$$

Since $\mathbb{E}[G - Q_\phi \mid h, z, a] = 0$ and ω is a function of (h, a), the term $G - Q_\phi$, weighted by ω, is orthogonal to every function of (h, z, a). The last two terms are orthogonal as well, because ωA_0 is a function of (h, a) and $\mathbb{E}[Q_\phi - Q_0 \mid h, a] = 0$. Hence

$$\mathbb{E}\big[\omega \, (G - V_\phi)^2\big] = \mathbb{E}\big[\omega \, (G - Q_\phi)^2\big] + \mathbb{E}\big[\omega \, A_\phi^2\big],$$
$$\mathbb{E}\big[\omega \, (G - V_0)^2\big] = \mathbb{E}\big[\omega \, (G - Q_\phi)^2\big] + \mathbb{E}\big[\omega \, (Q_\phi - Q_0)^2\big] + \mathbb{E}\big[\omega \, A_0^2\big].$$

Subtracting, and using the advantage form for $\mathbb{E}[\omega A_\phi^2] - \mathbb{E}[\omega A_0^2]$,

$$\Delta M_{\text{base}}(\phi) = \mathbb{E}\big[\omega \, \Delta_A^2\big] - \mathbb{E}\big[\omega \, (\Delta_V + \Delta_A)^2\big] = -\mathbb{E}\big[\omega \, \Delta_V^2\big] - 2\,\mathbb{E}\big[\omega \, \Delta_V \Delta_A\big]. \qquad \square$$

A.3 Proof of Proposition 3.2

Proposition 3.2 (Sign rule)**.** *Fix a history h and let the return be $G = \tilde{G} + \mathbf{1}\{a = \bar{a}\}\, x$ for one action $\bar{a}$, where the signal x seen by the critic is independent of $(a, \tilde{G})$ given h, with mean 0 and variance κ^2. Then $\Gamma(x) = \pi(\bar{a})\kappa^2$ at this history, and*

$$\Delta M_{\text{base}}(x) = -\kappa^2 \, \pi(\bar{a})^2 \big(2\|S(\bar{a})\|^2 - \bar{s}^2\big). \tag{9}$$

With two actions, for every policy parameterization,

$$\Delta M_{\text{base}}(x) = -\kappa^2 \pi(\bar{a})^2 \|S(\bar{a})\|^2 \, \frac{2 - 3\pi(\bar{a})}{1 - \pi(\bar{a})},$$

so for $\kappa > 0$ and $S(\bar{a}) \neq 0$ the signal reduces the update noise while $\pi(\bar{a}) < \frac{2}{3}$ and increases it once $\pi(\bar{a}) > \frac{2}{3}$.

Proof. Condition on the history h throughout. Since x is independent of $(a, \tilde{G})$ and has mean zero,

$$Q_x = Q_0 + \mathbf{1}\{a = \bar{a}\}\, x, \qquad V_x = V_0 + \pi(\bar{a})\, x.$$

For the action a that is taken, the offset and the contrast are therefore

$$\Delta_V = \pi(\bar{a})\, x, \qquad \Delta_A = \big(\mathbf{1}\{a = \bar{a}\} - \pi(\bar{a})\big)\, x,$$

so that $\Gamma(x) = \mathbb{E}[(Q_x - Q_0)^2] = \pi(\bar{a})\kappa^2$ and

$$\Delta_V^2 + 2\Delta_V \Delta_A = \pi(\bar{a})\, x^2 \, \big(2 \cdot \mathbf{1}\{a = \bar{a}\} - \pi(\bar{a})\big).$$

Since x is independent of a, Theorem 3.1 gives

$$\Delta M_{\text{base}}(x) = -\kappa^2 \pi(\bar{a}) \sum_a \pi(a) \, \|S(a)\|^2 \big(2 \cdot \mathbf{1}\{a = \bar{a}\} - \pi(\bar{a})\big) = -\kappa^2 \pi(\bar{a})^2 \big(2\|S(\bar{a})\|^2 - \bar{s}^2\big).$$

Two actions. Let b be the other action. Since $\mathbb{E}[S \mid h] = 0$, $\pi(\bar{a})S(\bar{a}) = -\pi(b)S(b)$, and hence

$$\bar{s}^2 = \pi(\bar{a})\|S(\bar{a})\|^2 + \pi(b)\|S(b)\|^2 = \frac{\pi(\bar{a})}{\pi(b)} \, \|S(\bar{a})\|^2,$$

$$2\|S(\bar{a})\|^2 - \bar{s}^2 = \|S(\bar{a})\|^2 \, \frac{2 - 3\pi(\bar{a})}{1 - \pi(\bar{a})},$$

which gives the two-action formula for every parameterization. $\square$

For a tabular softmax with two actions, $\|S(\bar{a})\|^2 = 2(1-\pi(\bar{a}))^2$, so

$$\Delta M_{\text{base}}(x) = -2\kappa^2\,\pi(\bar{a})^2\big(1-\pi(\bar{a})\big)\big(2-3\pi(\bar{a})\big).$$

For independent signals attached to different actions, Δ_V and Δ_A are sums over the signals and every cross term has mean zero, so the update costs add, as for $\{u, w\}$ in Figure 1.

A.4 The step rule

This is an auxiliary result built on the standard smoothness (ascent) inequality. Its normalized-step bound is of the same type as Cutkosky & Mehta (2020, Lemma 2), whose constants and setting differ; we do not claim it as a new optimization result.

Proposition A.5 (Step rule)**.** *Let $\hat{g} = \nabla J(\theta) + \xi$ with $\mathbb{E}\xi = 0$ and $\mathbb{E}\|\xi\|^2 = v$, and let ∇J be β-Lipschitz. A plain SGD step $\theta' = \theta + \eta\hat{g}$ satisfies*

$$\big|\,\mathbb{E}J(\theta') - J(\theta) - \eta\|\nabla J\|^2\,\big| \le \tfrac{\beta\eta^2}{2}\big(\|\nabla J\|^2 + v\big),$$

whereas a normalized step $\theta' = \theta + \eta\,\hat{g}/\|\hat{g}\|$ (no step if $\hat{g} = 0$) satisfies

$$\mathbb{E}J(\theta') - J(\theta) \ge \eta\big(\|\nabla J\| - 2\sqrt{v}\big) - \tfrac{\beta\eta^2}{2}.$$

Proof. Write $g = \nabla J(\theta)$. Because ∇J is β-Lipschitz, every step Δ satisfies

$$\big|J(\theta+\Delta) - J(\theta) - \langle g, \Delta\rangle\big| \le \tfrac{\beta}{2}\|\Delta\|^2.$$

SGD. Take $\Delta = \eta\hat{g}$ and use $\mathbb{E}\langle g, \hat{g}\rangle = \|g\|^2$ and $\mathbb{E}\|\hat{g}\|^2 = \|g\|^2 + v$.

Normalized step. Let $u = \hat{g}/\|\hat{g}\|$ if $\hat{g} \neq 0$ and $u = 0$ otherwise, and take $\Delta = \eta u$. If $\hat{g} \neq 0$, then, since $g = \hat{g} - \xi$,

$$J(\theta') - J(\theta) \ge \eta\langle u, g\rangle - \tfrac{\beta\eta^2}{2}, \qquad \langle u, g\rangle = \|\hat{g}\| - \langle u, \xi\rangle \ge \|\hat{g}\| - \|\xi\| \ge \|g\| - 2\|\xi\|.$$

If $\hat{g} = 0$, then $\xi = -g$ and $J(\theta') - J(\theta) = 0 \ge \eta(\|g\| - 2\|\xi\|)$. In both cases

$$J(\theta') - J(\theta) \ge \eta\big(\|g\| - 2\|\xi\|\big) - \tfrac{\beta\eta^2}{2},$$

and taking expectations with $\mathbb{E}\|\xi\| \le \sqrt{v}$ gives the claim. □

Under SGD, update noise leaves the first-order progress $\eta\|\nabla J\|^2$ unchanged and enters at order η^2; for a normalized step it enters the bound at first order. For the tabular softmax actors of Appendix C, J is a finite sum of products of softmax probabilities, so ∇J is globally Lipschitz.

Remark A.6 (The noise does enter at first order). In one dimension, with $g > 0$ and $\xi \sim \mathcal{N}(0, \sigma^2)$, the normalized step moves by $u = \operatorname{sign}(g + \xi)$, and

$$\mathbb{E}\langle u, g\rangle = g\,\big(1 - 2\Phi(-g/\sigma)\big),$$

which decreases strictly in σ. In this example, the first-order progress $\eta\,\mathbb{E}\langle u, g\rangle$ of a normalized step thus falls as the update noise grows.

A.5 Bootstrapped advantages

For a critic V of (h_j, z_j) or of h_j alone, let

$$\delta_j(V) = r_j + \gamma V_{j+1} - V_j, \qquad \hat{A}_t^{\lambda}(V) = \sum_{k\ge 0}(\gamma\lambda)^k\,\delta_{t+k}(V),$$

with $V = 0$ after the end of the episode; $\lambda = 0$ gives the one-step advantage $\delta_t(V)$ and $\lambda = 1$ the Monte Carlo advantage $G_t - V_t$.

Proposition 3.4 (What the informed critic adds to a bootstrapped advantage)**.** *For any critics $V_\phi(h, z)$ and $V_0(h)$, any $\lambda \in [0, 1]$ and every trajectory (the identity is known; Liu et al., 2026, App. D),*

$$\hat{A}_t^\lambda(V_\phi) - \hat{A}_t^\lambda(V_0) = -\Delta_t + \gamma(1-\lambda)\sum_{k\geq 1}(\gamma\lambda)^{k-1}\Delta_{t+k}, \tag{11}$$

where $\Delta_j = V_\phi(h_j, z_j) - V_0(h_j)$.

If, until the step R at which the history learns the outcome that the signal concerns, the offset only discounts forward ($\Delta_{t+k} = \gamma^{-k}\Delta_t$ for $t + k < R$), and the two critics agree from then on ($\Delta_j = 0$ for $j \geq R$), the right-hand side equals $-\lambda^{R-1-t}\Delta_t$.

Proof. Since $\delta_j(V_\phi) - \delta_j(V_0) = \gamma\Delta_{j+1} - \Delta_j$,

$$\hat{A}_t^\lambda(V_\phi) - \hat{A}_t^\lambda(V_0) = \sum_{k\geq 0}(\gamma\lambda)^k\big(\gamma\Delta_{t+k+1} - \Delta_{t+k}\big).$$

The coefficient of Δ_t is -1, and that of Δ_{t+k} for $k \geq 1$ is

$$(\gamma\lambda)^{k-1}\gamma - (\gamma\lambda)^k = \gamma(1-\lambda)(\gamma\lambda)^{k-1},$$

which gives Equation (11). Under the second hypothesis, the term of Δ_{t+k} equals $(1-\lambda)\lambda^{k-1}\Delta_t$ for $t + k < R$ and vanishes for $t + k \geq R$, so the right-hand side is

$$-\Delta_t\Big(1 - (1-\lambda)\sum_{k=1}^{R-1-t}\lambda^{k-1}\Big) = -\lambda^{R-1-t}\,\Delta_t. \qquad \square$$

With exact critics the difference in Equation (11) has mean zero given (h_t, a_t), so the two critics give the same mean update. Indeed, $\mathbb{E}[\Delta_j \mid h_j] = 0$ for every j, and h_j contains (h_t, a_t) for $j > t$, so

$$\mathbb{E}[\Delta_t \mid h_t, a_t] = \mathbb{E}[\Delta_t \mid h_t] = 0, \qquad \mathbb{E}[\Delta_j \mid h_t, a_t] = \mathbb{E}\big[\mathbb{E}[\Delta_j \mid h_j] \mid h_t, a_t\big] = 0 \quad (j > t),$$

where the first identity uses Lemma A.2. When n steps remain, the future offsets enter with total weight

$$\gamma(1-\lambda)\sum_{k=1}^{n-1}(\gamma\lambda)^{k-1} \;\longrightarrow\; \frac{\gamma(1-\lambda)}{1-\gamma\lambda} \qquad (n \to \infty),$$

which tends to 1 as $\gamma \to 1$ for every $\lambda < 1$: the bootstrap adds back about as much offset as the baseline removes.

Corollary A.7 (Bias from a learned critic)**.** *Let $\hat{V}_\phi$ be fitted independently of the rollout it evaluates, so that $\varepsilon_j = \hat{V}_\phi(h_j, z_j) - V_\phi(h_j, z_j)$ is a function of (h_j, z_j). Then Equation (13) holds.*

Proof. The proof of Equation (11) uses no property of the two critics. Applied to $\hat{V}_\phi$ and V_ϕ, it holds with ε_j in place of Δ_j. Multiply by S_t and take expectations; the term $-\mathbb{E}[S_t\varepsilon_t]$ vanishes by Lemma A.2. $\square$

For $\lambda < 1$ the bias is carried by the correlation between the current action and the critic's later errors, and it can be large when the signal omits a part of the state that the action changes.

Proposition A.8 (Update cost of the one-step form)**.** *Let the critics be exact and the signal Markov relative to the history, $\mathbb{E}[G_{t+1} \mid h_{t+1}, z_{t+1}, z_t] = \mathbb{E}[G_{t+1} \mid h_{t+1}, z_{t+1}]$, as when z is the state or a fixed hidden variable. Then the one-step advantage has update cost*

$$\begin{aligned}\Delta M_{\mathrm{TD}}(\phi) &= \mathbb{E}\big[\omega\,\delta_t(V_\phi)^2\big] - \mathbb{E}\big[\omega\,\delta_t(V_0)^2\big] \\ &= \Delta M_{\mathrm{base}}(\phi) + \gamma^2\,\mathbb{E}\big[\omega\,\Delta_{t+1}^2\big] + 2\gamma\,\mathbb{E}\big[\omega\,\big(r_t - \mathbb{E}[r_t \mid h_{t+1}]\big)\,\Delta_{t+1}\big],\end{aligned}$$

where the last term vanishes when the reward is part of the history.

Proof. Pathwise,

$$\delta_t(V_\phi) = \delta_t(V_0) + \gamma\Delta_{t+1} - \Delta_t, \qquad G_t - V_\phi = (G_t - V_0) - \Delta_t.$$

Expanding both second moments, subtracting, and using

$$G_t - V_0(h_t) - \delta_t(V_0) = \gamma\big(G_{t+1} - V_0(h_{t+1})\big),$$

we obtain

$$\begin{aligned}\Delta M_{\mathrm{TD}} - \Delta M_{\mathrm{base}} = \gamma^2\,\mathbb{E}\big[\omega\,\Delta_{t+1}^2\big] + 2\gamma\,\mathbb{E}\big[\omega\,\delta_t(V_0)\,\Delta_{t+1}\big] \\ - 2\gamma\,\mathbb{E}\big[\omega\,\Delta_t\Delta_{t+1}\big] + 2\gamma\,\mathbb{E}\big[\omega\,\big(G_{t+1} - V_0(h_{t+1})\big)\,\Delta_t\big].\end{aligned}$$

In the second term, ω, $V_0(h_t)$ and $V_0(h_{t+1})$ are functions of h_{t+1} and $\mathbb{E}[\Delta_{t+1} \mid h_{t+1}] = 0$, so of $\delta_t(V_0)$ only $r_t - \mathbb{E}[r_t \mid h_{t+1}]$ remains. In the last term, ω and Δ_t are functions of (h_{t+1}, z_t), and the Markov assumption gives $\mathbb{E}[G_{t+1} - V_0(h_{t+1}) \mid h_{t+1}, z_{t+1}, z_t] = \Delta_{t+1}$; the last term therefore cancels the third. □

If the offset only discounts forward, $\gamma\Delta_{t+1} = \Delta_t$, as in the second part of Proposition 3.4 for $t < R - 1$, then $\delta_t(V_\phi) = \delta_t(V_0)$ pathwise and $\Delta M_{\mathrm{TD}} = 0$: the next offset pays back exactly what the current one saves.

A.6 Proof of Proposition 4.1

Proposition 4.1 (Online selection keeps cross-fitted Monte Carlo updates unbiased)**.** *Let $\mathcal{F}_{b-1}$ contain the data of the blocks before b and any randomness of the selection rule. Suppose the candidate used in block b and the parameters θ_b are $\mathcal{F}_{b-1}$-measurable, the episodes of block b are drawn independently from π_{θ_b} given $\mathcal{F}_{b-1}$ and assigned to folds without looking at their data, and each episode is scored by a critic that depends only on $\mathcal{F}_{b-1}$ and the other fold. Then $\mathbb{E}[\hat{g}_b \mid \mathcal{F}_{b-1}] = \nabla J(\theta_b)$, whatever the selection rule.*

Lemma A.9 (Cross-fitting)**.** *Under the hypotheses of Proposition 4.1, let $\mathcal{G}$ be generated by $\mathcal{F}_{b-1}$ and the data of the fold that does not contain episode e. If $\hat{V}$ is $\mathcal{G}$-measurable, then $\mathbb{E}[S_t\,\hat{V}(h_t, z_t) \mid \mathcal{G}] = 0$ for every step t of episode e.*

Proof. Given $\mathcal{F}_{b-1}$, episode e is independent of the other fold, so it has the same law given $\mathcal{G}$ as given $\mathcal{F}_{b-1}$, and $\hat{V} = v$ is a fixed function given $\mathcal{G}$. With $\mathcal{H}_t$ the past of episode e before its action at step t, Lemma A.2 holds conditionally on $\mathcal{G}$, so

$$\mathbb{E}\big[S_t\,v(h_t, z_t) \mid \mathcal{G}, \mathcal{H}_t\big] = v(h_t, z_t)\,\mathbb{E}\big[S_t \mid \mathcal{G}, \mathcal{H}_t\big] = 0.$$

Taking expectations given $\mathcal{G}$ proves the claim. □

Proof of Proposition 4.1. Conditioning on $\mathcal{F}_{b-1}$ fixes θ_b and the candidate $\hat{\phi}_b$. For an episode in either fold, Lemmas A.3 and A.9 give

$$\mathbb{E}\Big[\sum_t \gamma^t S_t\big(G_t - \hat{V}(h_t, z_t)\big)\,\Big|\,\mathcal{G}\Big] = \nabla J(\theta_b),$$

and the tower property removes $\mathcal{G}$. Averaging over the episodes gives $\mathbb{E}[\hat{g}_b \mid \mathcal{F}_{b-1}] = \nabla J(\theta_b)$. The proof uses no property of the selection rule or of the quality of $\hat{V}$. □

Remark A.10 (What breaks it). If the candidate is selected on the scored block, $\hat{\phi}_b$ depends on the block's own actions, and $a_t \perp z_t \mid h_t$ can fail. If a tabular critic is fitted on the episodes it scores, it returns $\hat{V}(h_t, z_t) = G_t$ in every cell visited once, which sets those terms of $\hat{g}_b$ to zero, the overfitting that motivates K-fold baselines (Kota et al., 2017, §3). Episodes must also be independent across folds, which fails in an environment that does not reset between episodes. Appendix B detects the first two on sampled data.

A.7 THE VARIANCE-OPTIMAL BASELINE

Proposition A.11 (Variance-optimal baseline)**.** *Let y be the actor's target, a Monte Carlo return or a target computed with a critic that does not depend on the baseline, and let $b(h, z)$ be a baseline. Then, at histories with $\bar{s}^2(h) > 0$ (elsewhere every baseline gives the same update):*

(i) $\mathbb{E}[S\, b(h, z)] = 0$*;*

(ii) the second moment $\mathbb{E}[\omega\,(y - b)^2]$ *is minimized by*

$$b^\star_\phi(h, z) = \frac{\mathbb{E}[\omega\, y \mid h, z]}{\mathbb{E}[\omega \mid h, z]} = \sum_a \nu(a \mid h)\, Q^y_\phi(h, z, a), \qquad \nu(a \mid h) = \frac{\pi(a \mid h)\,\omega(a)}{\bar{s}^2(h)},$$

where $Q^y_\phi(h, z, a) = \mathbb{E}[y \mid h, z, a]$*, and any other baseline pays* $\mathbb{E}[\bar{s}^2(h)\,(b - b^\star_\phi)^2]$ *more;*

(iii) the minimum never increases when the signal is refined.

Proof. (i) Given h, $a \perp z$ (Lemma A.2), so $\mathbb{E}[S\, b \mid h] = \mathbb{E}[S \mid h]\,\mathbb{E}[b \mid h] = 0$. (ii) By the same independence, $\mathbb{E}[\omega \mid h, z] = \bar{s}^2(h)$ and $\mathbb{E}[\omega\, y \mid h, z] = \sum_a \pi(a \mid h)\,\omega(a)\, Q^y_\phi(h, z, a)$, which gives the second form of $b^\star_\phi$. Since $\mathbb{E}[\omega\,(y - b^\star_\phi) \mid h, z] = 0$,

$$\mathbb{E}\big[\omega\,(y - b)^2\big] = \mathbb{E}\big[\omega\,(y - b^\star_\phi)^2\big] + \mathbb{E}\big[\bar{s}^2(h)\,(b - b^\star_\phi)^2\big].$$

(iii) Every function of a coarser signal is a function of the finer one, so the minimum over the larger class is no larger. □

Part (ii) is due to Greensmith et al. (2004, Thms. 13 and 16) for the policy's input and to Mao et al. (2019, Thm. 2) for information that the policy does not see, and Liu et al. (2026, Prop. 4) show (iii) for a privileged index; Kuba et al. (2021, Cor. 1) use the tabular-softmax weight $\omega(a) = \|e_a - \pi\|^2$.

Two actions. With actions a and b, $\pi(a)S(a) = -\pi(b)S(b)$ gives $\nu(a) = \pi(b)$ and $\nu(b) = \pi(a)$, a formula that Weaver & Tao (2001) attribute to Dayan. For the target y, the optimal and the value baseline are therefore

$$b^\star_\phi = \pi(b)\, Q^y_\phi(a) + \pi(a)\, Q^y_\phi(b), \qquad V^y_\phi = \pi(a)\, Q^y_\phi(a) + \pi(b)\, Q^y_\phi(b),$$

and by (ii) the value baseline pays

$$\mathbb{E}\big[\bar{s}^2\,(2\pi(a) - 1)^2\,\big(Q^y_\phi(a) - Q^y_\phi(b)\big)^2\big]$$

more than $b^\star_\phi$. The excess vanishes at the uniform policy and grows with the contrast that the signal adds to $Q^y_\phi(a) - Q^y_\phi(b)$. For the signal of Proposition 3.2, $b^\star$ follows x with weight $\nu(\bar{a}) = 1 - \pi(\bar{a})$ and the value baseline with weight $\pi(\bar{a})$. Adding the signal lowers the minimum of (ii) at every policy, by $\bar{s}^2\nu(\bar{a})^2\kappa^2$, which is $2\kappa^2\pi(\bar{a})(1 - \pi(\bar{a}))^3$ for a tabular softmax.

B ONLINE INFORMED ACTOR-CRITIC

In this appendix, we give OIAC's pseudo-code, its statistic and decision rule, and a check of Proposition 4.1 on sampled data.

Algorithm 1 Online Informed Actor-Critic (OIAC), baseline form. The two-fold critic fit is a K-fold baseline (Kota et al., 2017); the selection on earlier blocks is ours.

Require: candidates $\Phi \ni \phi_0$ (history-only); current choice $\hat{\phi} \leftarrow$ IAAC's offline choice; block size B; window W; level δ; cost c

for block $b = 1, 2, \dots$ **do**
  sample B episodes with π_θ and split them into two folds by episode index
  for each candidate $\psi \in \Phi$ **do**
    fit a critic $\hat{V}_\psi$ on each fold and evaluate it on the other fold
    $T_\psi(e) \leftarrow \sum_t \gamma^{2t} \|S_t\|^2 \big(G_t - \hat{V}_\psi(h_t, \psi(i_t, h_t))\big)^2$ for every episode e ▷ update noise
  update θ with $\hat{g} = \frac{1}{B} \sum_e \sum_t \gamma^t S_t \big(G_t - \hat{V}_{\hat{\phi}}(h_t, \hat{\phi}(i_t, h_t))\big)$ ▷ $\hat{\phi}$ chosen before block b
  over the episodes of the last W blocks, for each challenger $\psi \neq \hat{\phi}$:
  $D_\psi \leftarrow \mathrm{mean}_e \big[T_{\hat{\phi}}(e) - T_\psi(e)\big]$, $\mathrm{LCB}_\psi \leftarrow D_\psi - q_{1-\delta/m}\, \mathrm{SE}(D_\psi)$, $m = |\Phi| - 1$
  if $\max_\psi \mathrm{LCB}_\psi > c \cdot \mathrm{mean}_e\, T_{\hat{\phi}}(e)$ **then**
    $\hat{\phi} \leftarrow \arg\max_\psi \mathrm{LCB}_\psi$ ▷ used from block $b + 1$ on

Statistic. Let $g_e = \sum_t \gamma^t S_t \big(G_t - \hat{V}_\psi(h_t, z_t)\big)$ be the update of episode e. The statistic $T_\psi(e)$ keeps the diagonal terms of $\|g_e\|^2$; the cross-step terms vanish when different steps use disjoint actor parameters, as for our tabular actors, whose keys include the step, and then $T_\psi(e) = \|g_e\|^2$. At one step, a critic $\hat{V}$ fitted without the scored episode gives

$$\mathbb{E}\big[\omega\,(G - \hat{V})^2\big] = \mathbb{E}\big[\omega\,(G - V_\psi)^2\big] + \mathbb{E}\big[\omega\,(V_\psi - \hat{V})^2\big] + 2\,\mathbb{E}\big[(V_\psi - \hat{V})\,\mathbb{E}[\omega\, A_\psi \mid h, z]\big].$$

The first term is the second moment with the exact critic, to which Theorem 3.1 applies. The second is the score-weighted estimation error, through which a candidate that fragments the critic's data pays for it. The third has either sign and vanishes when the fit is unbiased at every (h, z). Cross-fitting couples the episodes of a fold, so T_ψ does not directly give the variance of the batch update $\hat{g}$.

Decision rule. Candidates are compared through paired differences on the same episodes, which removes the between-episode variation that they share. In Algorithm 1, SE is the standard deviation of the paired differences divided by the square root of their number, q_p is the standard normal p-quantile, and the level is Bonferroni-adjusted over the m challengers. Cross-fitting dependence, policy changes and repeated, overlapping windows leave the coverage of these bounds nominal, and we claim no anytime-valid guarantee. The switching threshold c keeps the incumbent unless a challenger promises a reduction of at least a fraction c of the statistic, so when the statistic cannot separate the candidates, OIAC keeps the current one, initially IAAC's choice. OIAC uses no episodes beyond the training batch, so its sample budget equals IAAC's.

Checks on sampled data. We checked Proposition 4.1 with the learner's own estimators in a four-step informed POMDP with 32 candidate masks, averaging 20,000 fresh blocks of 64 episodes at five policies and comparing the mean update with the exact gradient. No estimator that satisfies the hypotheses, including online selection on the previous block, showed a detectable bias. Both violations of Remark A.10, selection on the scored block and an in-sample critic, were detected; the squared norm of every detected bias, corrected for sampling noise, was below 1% of the update's mean squared error at $B = 64$.

C Experimental setup

In this appendix, we give the details of the tabular study of Section 5. All runs are single-threaded NumPy on one workstation CPU. Exact returns, gradients, update errors and variance-optimal baselines come from dynamic programming on the finite models.

Environments. Table 1 summarizes the two environments. In the three-step family the nuisance n_t is redrawn at every step, the discount is 1, and the privileged vector also contains the branch b,

which the history already shows. At the branch step Proposition 3.2 holds exactly with κ replaced by 2κ. On the development seeds, IAAC's offline criterion selects $\{u, w\}$ on every seed of the one-decision example with $\kappa = (1, 2)$ and on 59 of 60 seeds with $\kappa = (2, 1)$, where one seed adds n; in the three-step family it selects $\{u, w, v\}$ on every seed.

Learner. The actor is a tabular softmax over full-history keys, initialized at $\theta = 0$. Each iteration samples B fresh on-policy episodes and applies the Monte Carlo baseline-form update

$$\hat{g} = \frac{1}{B} \sum_{e} \sum_{t} S_t \big(G_t - \hat{b}(h_t, z_t)\big)$$

with Adam, the normalized step $\theta \leftarrow \theta + \eta\, \hat{g}/\|\hat{g}\|$ or plain SGD. Each learning rate was locked before any comparison, on seed 0 with the history-only critic, as the grid value whose threshold crossing is closest to 150 iterations (Table 5).

Baselines. No baseline $\hat{b}$ uses the episodes whose update it evaluates, so every update is unbiased (Lemma A.2).

- *Value critics* (history-only, full privileged, fixed masks, and the candidates of every selection rule): cell means of the returns over the history key and the candidate's features, cross-fitted on two folds of the batch split by the parity of the episode index. An empty cell falls back to the key's mean and then to the global mean.
- *In-batch* estimates of $b^\star$ (Proposition A.11) with the full privileged vector and the same cross-fitting: the score-weighted cell mean $\sum_t \omega_t G_t / \sum_t \omega_t$, and $\sum_a \nu(a \mid h)\, \hat{Q}(h, z, a)$ from per-action cell means $\hat{Q}$.
- *Pooled* estimates: $\hat{b} = \sum_a \nu(a \mid h)\, \hat{Q}(h, z, a)$, with $\hat{Q}(h, z, a)$ the running mean of all past returns in the cell and ν from the current policy. The statistics are updated after the actor step, so a batch's baseline contains none of its returns; in the three-step family the counts and sums are multiplied by 0.99 once per iteration, because the later-step policy changes Q. The *pooled value* control uses the same $\hat{Q}$ with π in place of ν.
- *Exact* $b^\star$: computed by dynamic programming at the current policy, a reference that no learner has.

Selection criteria. IAAC's choice is made before training, from 512 episodes of the uniform policy that are not charged to its budget, by the held-out return-prediction gain with a 1,000-resample bootstrap at level 0.05 (Ebi et al., 2026, §§4.2–4.3). IAAC's authors apply this test post hoc to select among feature subsets; it needs only episodes of a fixed policy, and we apply it before training. We implement the held-out gain test and not the residual-independence test; their population nulls coincide here, because the only features without gain (n, and b in the three-step family) are independent of the return residual by construction. The offline, update-based criterion replaces the gain by the statistic T_ψ of Equation (14) under the uniform policy. The online, return-based criterion runs Algorithm 1 with the held-out squared prediction error $\sum_t \big(G_t - \hat{V}_\psi(h_t, z_t)\big)^2$ in place of T_ψ. Both online rules start from IAAC's choice, with W, δ and c fixed in advance (Table 5). The candidates are all masks of the privileged vector or, under a budget of at most one feature, the empty mask and the single features. The exact oracle schedule uses, at every iteration, the candidate with the smallest exact update second moment at the current policy.

Metrics and protocol. Episodes to threshold is B times the first iteration at which the exact return reaches the threshold; runs that never reach it within the iteration budget are censored and counted. The update error of an iteration is the exact $\|\hat{g} - \nabla J(\theta)\|^2$. It is compared across runs at matched policies, with iterations binned by the exact $\pi(R)$ at the first step, and its late phase is $\pi(R) > \frac{2}{3}$. Intervals are 95% percentile bootstraps over seeds, and comparisons between runs are paired by seed. The predictions of each study were written down before its runs, those of the pooled estimators before any pooled run. Where a cell existed in an earlier study of ours, its history-only and IAAC runs were re-run with the modified code and reproduced the stored runs exactly.

Table 5: Locked settings. Learning rates were locked on seed 0 with the history-only critic; the one-decision value was locked on $\kappa = (2, 1)$ with $B = 32$ and reused for $(1, 2)$ and for $B = 128$.

	one decision	three steps
episodes per iteration B / iterations	32 and 128 / 500	64 / 400
learning rate: Adam ($\beta_1, \beta_2 = 0.9, 0.999$)	0.02	0.02
learning rate: normalized step	0.05	0.1
learning rate: plain SGD	0.1	–
online rules: window W / level δ / cost c	10 / 0.05 / 0.05	10 / 0.05 / 0.05
development seeds / confirmation seeds (Adam)	0–59 / 100–119	0–59 / –

D EXACT CHECKS

In this appendix, we check the identities by exact dynamic programming. The critics are exact, $V_\phi(h, z) = \mathbb{E}[G \mid h, z]$ and $V_0(h) = \mathbb{E}[G \mid h]$, with the full history as memory. Every trajectory is enumerated, and the Gaussian reward noise is integrated in closed form. The squared score is that of the tabular softmax actor, $\omega_t = \|e_{a_t} - \pi(\cdot \mid h_t)\|^2$.

Environments and policies.

- One decision, both settings of κ: $\pi(R) \in \{0.5, 0.6, \frac{2}{3}, 0.7, 0.8, 0.9, 0.95\}$.
- Three steps, interpolated toward the optimal policy: $\pi(R) = p$ at the branch and $\pi(a = c_t) = p$ at the later steps, for $p \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$.
- A delayed-outcome chain ($H = 6$, $\gamma = 0.9$): a hidden bit x, the signal, fixes a reward of ± 1 at the last step, and the observation reveals x from step $R = 4$ on. Both actions have the same effect, so the signal has no contrast.

The signals are IAAC's choice ($\{u, w\}$ and $\{u, w, v\}$) and the full privileged vector, which give identical exact values because exact critics ignore the independent nuisance and the public branch.

Results. Table 6 lists the largest absolute error of each identity; for Equation (13), the exact informed critic of the three-step family was perturbed by a random table of scale 0.3 over (t, h, z), at $p = 0.5$ and 0.8. In the three-step family the rewards depend on u, w and v and are not part of the history, so Proposition A.8 needs its reward term, which reaches 2.75 at the first later step under the uniform policy and 0.27 at $p = 0.95$; in the other two environments the reward term vanishes. Table 7 extends Table 2 to all policies. Its payback column is the share of the offset's saving that the next offset returns to a one-step update (Proposition A.8). The payback exceeds 100% from $p = 0.7$ on, because the next offset depends on the branch taken, which is the signal's contrast at the branch step. With a single decision all targets coincide.

Table 6: Largest absolute error of each identity over all environments, policies, steps and signals.

identity	max. abs. error
Theorem 3.1, baseline form	1.8×10^{-15}
Proposition 3.4, Equation (11)	1.8×10^{-15}
Proposition 3.4, delayed outcome (chain)	1.1×10^{-16}
Equation (11), mean zero given (h_t, a_t)	1.2×10^{-14}
Proposition A.8, with its reward term	1.3×10^{-15}
Equation (13), perturbed critic	2.3×10^{-15}

Table 7: Update cost of IAAC's signal $\{u, w, v\}$ in the three-step family (exact critics), in percent of the history-only second moment at the same λ, summed over steps; negative values: less noise. p: probability of the optimal action at every step. Payback: $\gamma^2\mathbb{E}[\omega\Delta_{t+1}^2]/\mathbb{E}[\omega\Delta_t^2]$, in percent.

p	$\lambda = 1$ (Monte Carlo)	$\lambda = 0.9$	$\lambda = 0.5$	$\lambda = 0$ (one step)	payback
0.50	-63.5	-60.1	-39.8	-6.2	70
0.60	-61.1	-56.9	-31.3	$+9.3$	85
0.70	-54.6	-48.6	-11.8	$+43.2$	103
0.80	-44.0	-35.1	$+19.5$	$+96.7$	122
0.90	-29.3	-16.4	$+63.3$	$+168.6$	143
0.95	-20.4	-5.2	$+89.8$	$+210.3$	153

E Additional results

In this appendix, we report the full results of the selection criteria and of the variance-optimal baseline.

E.1 Selection criteria

Tables 9–11 give the runs with paired intervals. The offline, update-based criterion chose the same candidate as IAAC on every seed, so its runs coincide with IAAC's and are not shown.

Other settings.

- *One-feature budget,* $\kappa = (1, 2)$. IAAC and the other two criteria hold $\{w\}$ throughout, while OIAC drops it at a median $\pi(R)$ of 0.80 and reaches the threshold 11% sooner than IAAC under the normalized step (Table 9).
- $\kappa = (2, 1)$. The harm of w is four times smaller; under Adam, OIAC drops w on 33 of 60 seeds and lowers the late update noise by 6%. Under the budget IAAC's choice $\{u\}$ is right, and OIAC keeps it through the threshold on 58 of 60 seeds, whereas the online, return-based criterion switches to the harmful w and is the slowest of the four criteria under Adam and the normalized step (Table 10).
- *Three steps.* The signals are pure offsets at the later steps, which outweighs the contrast cost of w at the branch step. IAAC's choice learns faster than the history-only critic, and OIAC keeps it, with a tie under the normalized step and a 1% loss under Adam; the online, return-based criterion finds the later-step offsets sooner and is the fastest criterion under the normalized step (Table 11).

Failed predictions. Of the 88 pre-registered checks, 62 passed. The failed selection predictions have two causes. First, the population theory ignores estimation cost at 32 episodes per block. The extra critic cells cancel u's small benefit, so in the late phase OIAC uses $\emptyset$ about as often as $\{u\}$, and the online, return-based criterion drops u instead of keeping $\{u, w\}$. Second, the statistic is heavy-tailed. Past the threshold, windows contain few of the rare-action episodes that carry the harm, and OIAC re-admits w on most seeds, although just before the threshold it uses w in only 4% of iterations under Adam. With $\kappa = (1, 2)$, the prediction that OIAC would be the fastest run other than the oracle failed in three of the four Adam and normalized-step cells (with and without the budget), where the history-only critic was faster; OIAC was the fastest of the four criteria in all four, and no criterion was significantly faster than the history-only critic.

Confirmation seeds. Table 8 repeats the Adam runs on seeds 100–119, run once after every setting had been locked. The reduction of the late update noise replicates, whereas no paired interval for the episodes excludes zero, so the small Adam speed-up is not confirmed.

Action-value critics. IAAC's reference code uses action-conditioned critics. With them, in runs added after the pre-registered ones, the offline gain test makes the same choice as the value-based test on 239 of 240 seed cases, and the episodes to threshold of IAAC and of the online, return-based criterion change by at most 0.04×10^3.

Table 8: Confirmation seeds 100–119 (Adam), OIAC against IAAC: late update noise as a ratio, and paired difference in episodes to threshold ($\times 10^3$), with 95% intervals.

	late noise, OIAC / IAAC	episodes, OIAC − IAAC
$\kappa = (1, 2)$	$0.85\ [0.80, 0.89]$	$+0.01\ [-0.12, +0.12]$
$\kappa = (1, 2)$, at most one feature	$0.82\ [0.79, 0.85]$	$-0.04\ [-0.14, +0.06]$
$\kappa = (2, 1)$	$0.94\ [0.91, 0.98]$	$-0.02\ [-0.10, +0.07]$

Table 9: One-decision example with $\kappa = (1, 2)$, development seeds 0–59. Episodes to threshold ($\times 10^3$): IAAC's mean (bold) and, for every other run, the paired difference to IAAC with 95% interval (negative is faster); the history-only and full privileged critics are listed once. Late noise: exact update error once $\pi(R) > \frac{2}{3}$, relative to the history-only critic at the same $\pi(R)$ (Adam).

run	normalized step	Adam	plain SGD	late noise
no budget				
IAAC, episodes	**2.28**	**4.39**	**4.07**	1.43
history-only	$-0.44\ [-0.57, -0.31]$	$-0.16\ [-0.30, -0.01]$	$-0.11\ [-0.28, +0.06]$	1.00
full privileged	$-0.14\ [-0.29, +0.00]$	$+0.15\ [-0.02, +0.31]$	$-0.01\ [-0.20, +0.18]$	1.42
online, return-based	$-0.04\ [-0.13, +0.03]$	$-0.01\ [-0.04, +0.03]$	$+0.02\ [-0.02, +0.07]$	1.39
OIAC	$-0.29\ [-0.41, -0.19]$	$-0.12\ [-0.19, -0.05]$	$+0.00\ [-0.09, +0.08]$	1.21
fixed $\{u\}$	$-0.40\ [-0.54, -0.26]$	$-0.14\ [-0.28, +0.01]$	$-0.12\ [-0.29, +0.05]$	1.02
fixed $\{u, w\}$	$-0.05\ [-0.21, +0.10]$	$-0.08\ [-0.24, +0.09]$	$-0.09\ [-0.27, +0.10]$	1.38
exact oracle schedule	$-0.39\ [-0.54, -0.25]$	$-0.19\ [-0.33, -0.04]$	$-0.11\ [-0.27, +0.05]$	1.05
at most one feature				
IAAC, episodes	**2.12**	**4.32**	**4.01**	1.32
online, return-based	0	0	0	1.32
OIAC	$-0.23\ [-0.30, -0.16]$	$-0.14\ [-0.23, -0.06]$	$-0.05\ [-0.17, +0.06]$	1.10
fixed $\{u\}$	$-0.23\ [-0.33, -0.14]$	$-0.07\ [-0.21, +0.07]$	$-0.06\ [-0.24, +0.11]$	1.02
fixed $\{w\}$	$+0.07\ [-0.06, +0.20]$	$-0.08\ [-0.22, +0.06]$	$-0.04\ [-0.22, +0.14]$	1.33
exact oracle schedule	$-0.27\ [-0.36, -0.18]$	$-0.16\ [-0.30, -0.02]$	$-0.06\ [-0.25, +0.12]$	1.04

Table 10: As Table 9, with $\kappa = (2, 1)$.

run	normalized step	Adam	plain SGD	late noise
no budget				
IAAC, episodes	**2.28**	**4.43**	**4.05**	0.99
history-only	$+0.03\ [-0.12, +0.18]$	$+0.07\ [-0.10, +0.25]$	$-0.05\ [-0.23, +0.14]$	1.00
full privileged	$+0.09\ [-0.05, +0.25]$	$+0.28\ [+0.11, +0.45]$	$-0.08\ [-0.26, +0.12]$	1.21
online, return-based	$+0.01\ [-0.09, +0.12]$	$+0.00\ [-0.03, +0.04]$	$+0.04\ [-0.01, +0.08]$	0.98
OIAC	$-0.07\ [-0.13, -0.01]$	$-0.06\ [-0.11, -0.01]$	$+0.00\ [-0.05, +0.06]$	0.93
fixed $\{u\}$	$-0.07\ [-0.22, +0.09]$	$-0.11\ [-0.27, +0.05]$	$-0.07\ [-0.25, +0.12]$	0.84
fixed $\{u, w\}$	$+0.02\ [-0.11, +0.15]$	$-0.01\ [-0.18, +0.16]$	$-0.05\ [-0.23, +0.15]$	0.99
exact oracle schedule	$-0.04\ [-0.19, +0.10]$	$-0.07\ [-0.23, +0.10]$	$-0.07\ [-0.24, +0.12]$	0.84
at most one feature				
IAAC, episodes	**2.10**	**4.22**	**3.98**	0.79
online, return-based	$+0.07\ [-0.01, +0.15]$	$+0.09\ [+0.04, +0.14]$	$-0.01\ [-0.08, +0.06]$	0.88
OIAC	0	$-0.00\ [-0.00, +0.00]$	$-0.01\ [-0.01, +0.00]$	0.79
fixed $\{u\}$	$+0.11\ [-0.04, +0.25]$	$+0.10\ [-0.06, +0.25]$	$-0.00\ [-0.18, +0.18]$	0.84
fixed $\{w\}$	$+0.28\ [+0.15, +0.44]$	$+0.33\ [+0.17, +0.50]$	$+0.02\ [-0.16, +0.20]$	1.08
exact oracle schedule	$+0.11\ [-0.04, +0.25]$	$+0.10\ [-0.06, +0.25]$	$-0.00\ [-0.18, +0.18]$	0.84

E.2 The variance-optimal baseline

Table 4 gives the learning results, with the estimators of Appendix C. A baseline fitted on other samples than those it is applied to is, given them, a function of (h, z), so by Proposition A.11 it pays $\mathbb{E}[\bar{s}^2(\hat{b} - b^\star_\phi)^2]$ over the optimum. As the policy commits, ν puts most of its mass on the rarely taken actions, so $b^\star_\phi$ is essentially their value and is estimated from few samples.

Table 11: As Table 9, for the three-step family ($B = 64$); the late phase is defined by the branch policy $\pi(R)$.

run	normalized step	Adam	late noise
no budget			
IAAC, episodes	**8.23**	**10.57**	0.85
history-only	$+0.33$ $[+0.11, +0.54]$	$+0.82$ $[+0.62, +1.02]$	1.00
full privileged	$+0.20$ $[-0.01, +0.41]$	$+0.66$ $[+0.47, +0.85]$	0.99
online, return-based	-0.12 $[-0.19, -0.06]$	$+0.03$ $[-0.03, +0.09]$	0.84
OIAC	$+0.02$ $[-0.01, +0.05]$	$+0.09$ $[+0.03, +0.16]$	0.84
at most one feature			
IAAC, episodes	**8.57**	**11.33**	0.83
online, return-based	-0.47 $[-0.56, -0.38]$	-0.22 $[-0.30, -0.15]$	0.84
OIAC	-0.19 $[-0.25, -0.13]$	$+0.00$ $[-0.00, +0.00]$	0.83

- *Exact.* Over 34 policies and all 1,224 nested pairs of signals, the minimum never increased under refinement (largest relative violation 1.8×10^{-15}), and $b^\star$ was never worse than the value baseline with the same signal. The two can differ by more than a factor of three: with $\kappa = (1, 2)$ at $\pi(R) = 0.9$, IAAC's signal $\{u, w\}$ raises the value baseline's second moment to 1.78 times the history-only one, whereas $b^\star$ is at 0.51 with the same signal and at 0.79 without any signal.
- *No signal.* The history-only value baseline reached the threshold sooner than IAAC in three of the four cells with $\kappa = (1, 2)$, by up to 19%, and later in the three-step family, where IAAC's signal includes the pure offset v.
- *In-batch estimates.* The score-weighted cell mean was not significantly faster than IAAC in any of the ten cells, against the prediction we had registered; the per-action estimate was faster in some cells and slower in others. At fixed policies, both approach the exact $b^\star$ only with 512 to 2,048 episodes per batch.
- *Pooled estimates.* Pooling the per-action returns across iterations recovered 98% of the exact $b^\star$'s reduction relative to IAAC over the eight one-decision cells, at a late-phase update error of 0.32–0.48 times IAAC's in all ten cells. Against the pooled value control, which uses the same data with the weights π, the weights ν cut the late update error to 0.47–0.63 of the control's in all fourteen cells and reached the threshold sooner in all ten cells with Adam or normalized steps.
- *Plain SGD.* In the four cells with plain SGD, the weights ν brought no speed-up over the control, and in one they were slower, in line with Proposition A.5.