HYBRID REINFORCEMENT LEARNING UNDER DYNAMICS SHIFT: POLICY OPTIMIZATION, CERTIFICATION, AND ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid reinforcement learning combines offline data with online interaction. When these data come from environments with different transition dynamics, the learner must decide how to use the source and when to rely on target observations. We study policy optimization, target certification, and adaptation under such dynamics shift. For general function classes satisfying realizability and Bellman completeness in both environments, we analyze pessimistic policy optimization with source-consistent critics. Its guarantees depend on target Bellman residuals along the learned policies and a target-optimal comparator, quantified through policy-dependent transfer moduli. A block-based algorithm uses its own target observations to constrain the deployed-policy residual; a separate comparator term and the cost of failed tests remain. An incompatibility test supports a handoff to online learning, while a return-based master adapts between source-informed copies and an online baseline, with coefficient-dependent overheads. Structured upper bounds separate source uncertainty from dynamics shift in tabular, linear, and block MDPs. Lower bounds show that the resulting square-root modulus dependence is necessary on hard families when target observations cannot distinguish the competing models. In finite-class experiments, an optimal source policy can trigger handoff while a better unvisited route remains undetected, and return-based selection partly recovers from a misleading source at a substantial cost relative to target-only controls.

1 INTRODUCTION

Offline data can reduce the interaction needed to learn a good policy, but the data may come from a simulator, an earlier deployment, or a related environment. Even when rewards are unchanged, different transition dynamics can change the value of an action. A learner using such data must account for two limitations. The source may provide little information about some state-action pairs, and the information it provides may describe the wrong dynamics. These limitations affect policies differently. Errors in a region that a policy never visits may be irrelevant to its performance, whereas an error at a frequently visited decision point can change which action is preferable.

Hybrid reinforcement learning uses offline data together with online interaction. Much of its theory considers a common environment for both sources of observations (Song et al., 2023; Wagenmaker & Pacchiano, 2023; Tan & Xu, 2024). With different source and target dynamics, additional source data can improve estimation without resolving the discrepancy between the environments. Target interaction can reveal that discrepancy, but only where the learner collects information. In particular, a source-trained policy can perform poorly even when its predictions are accurate along its own trajectories, because the source undervalues a better route that the policy does not explore.

We study three algorithmic questions in this setting, namely how to optimize a policy using source data, what target observations can certify about the critics used for policy improvement, and how to adapt when a source-informed learner is less effective than learning from scratch. Our approach uses source data to constrain the critics available to a pessimistic policy optimizer. During deployment, target observations provide additional constraints and can trigger a handoff to an online learner. A separate return-based master addresses the broader question of which learning procedure to use, including cases in which on-policy consistency alone does not reveal the source's limitations.

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.

The analysis connects these mechanisms through the target Bellman errors of source-consistent critics. We adapt the transfer-modulus framework of Hanneke & Kpotufe (2024) to these errors, keeping separate the policy in the Bellman backup and the policy determining the evaluation distribution. The source-to-target modulus gives the smallest uniform residual bound implied by source consistency. Its target-conditioned counterpart incorporates constraints supplied by target observations. The algorithms do not take these quantities as inputs, and no single global discrepancy is assumed to describe every policy's transfer difficulty.

Our contributions are as follows. Tables 3 and 4 in Appendix A summarize the settings and guarantees.

1. *Policy optimization from source data.* Under realizability and Bellman completeness in both environments, we analyze a pessimistic policy-optimization procedure with source-consistent critics. The target suboptimality and regret bounds separate an optimization term from transfer terms evaluated along the learned policies and the target-optimal comparator (Theorem 1, and Theorem 21 in Appendix D).
2. *Certification and handoff using target observations.* A block-based variant constrains the critic using observations from the same policy block that it evaluates. In certified blocks, this caps the deployed-policy residual at a target-data radius and replaces the comparator term by its target-conditioned modulus (Theorem 2). With Bernstein calibration, the bound has a $\widetilde{O}(H^2T^{3/4})$ remainder in addition to the comparator transfer term and an explicit charge for fallback blocks (Corollary 3). An empirical incompatibility test supports a handoff bound that combines the certified prefix, the exit-block cost, and an online learner's guarantee (Theorem 4).
3. *Adaptation through observed returns.* A balancing master combines source-informed copies with an online learner having a known anytime square-root regret bound (Theorem 5). Its comparison is with source copies whose candidate bounds are valid at every prefix of their own episodes. The guarantee retains factors depending on the number of copies, the horizon, and the online bound's coefficient; it does not require advance knowledge of a transfer level. We also give a composition with the block-based learner (Proposition 32).
4. *Structured bounds and limits.* For tabular, linear, and block MDPs, upper bounds distinguish source statistical uncertainty, uncovered directions or regions, and dynamics shift. Lower-bound families establish the square-root modulus dependence in target-sample regimes where the competing models remain difficult to distinguish (Section 5). These are worst-case necessities over the stated families, rather than an instancewise converse to every upper bound. Finite-class experiments distinguish critic incompatibility from policy suboptimality, quantify the costs of handoff and return-based selection, and compare with both tabular and candidate-aware target-only controls (Section 6).

Related work. Transfer moduli provide a general framework for relating source and target errors (Hanneke & Kpotufe, 2024). Transfer exponents have also been used in covariate-shift bandits (Suk & Kpotufe, 2021; Cai et al., 2024), and policy transfer coefficients characterize coverage in offline decision-making (Nguyen-Tang & Arora, 2023; 2024). Our analysis applies this approach to policy-dependent Bellman residuals when the source and target operators differ. Same-environment hybrid RL and policy finetuning exploit offline coverage (Song et al., 2023; Tan & Xu, 2024; Wagenmaker & Pacchiano, 2023; Xie et al., 2021b; Li et al., 2023; Huang et al., 2025); dynamics shift adds a bias that source consistency alone cannot remove. In the shifted setting, Qu et al. (2025) use a separability condition, while Wu et al. (2026) give tabular algorithms with per-pair bias bounds and an extension for unknown bias. Correction-based Bellman alignment (Chen et al., 2026b) and adaptation of a pretrained value function (Li & Zhang, 2026) offer other ways to exploit source information. We build on Bellman-consistent pessimism (Jin et al., 2021; Xie et al., 2021a; Zanette et al., 2021) and study its target residuals, the scope of on-policy certification, and adaptation between learning procedures. These guarantees complement structured-shift methods; they do not uniformly improve their rates. Appendix A.2 gives the detailed comparisons.

2 Learning from Source and Target Environments

Data and objective. We study two finite-horizon MDPs $\mathcal{M}_{\mathrm{src}}$ and $\mathcal{M}_{\mathrm{tar}}$ with common state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, initial distribution ρ, and rewards $r_h(s,a) \in [0,1]$. Their transition kernels $P_{\mathrm{src},h}$ and $P_{\mathrm{tar},h}$ may differ. The learner receives n source episodes collected under a behavior policy μ and can collect T target episodes using adaptively chosen policies.

For a policy π, let $d^{\pi}_{\mathcal{M},h}$ denote its state-action distribution at stage h in environment $\mathcal{M}$, and let $Q^{\pi}_{\mathcal{M},h}$ and $V^{\pi}_{\mathcal{M},h}$ denote its action and state values. Write $V^{\pi}_{\mathcal{M},1}(\rho) = \mathbb{E}_{s\sim\rho} V^{\pi}_{\mathcal{M},1}(s)$ and let $\pi^{\star}_{\mathrm{tar}}$ be a target-optimal policy. Our objectives are the target suboptimality of a source-trained policy $\widehat{\pi}$ and the regret of the deployed policies:

$$\mathrm{SubOpt}_{\mathrm{tar}}(\widehat{\pi}) = V^{\pi^{\star}_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar}},1}(\rho) - V^{\widehat{\pi}}_{\mathcal{M}_{\mathrm{tar}},1}(\rho), \qquad \mathrm{Regret}(T) = \sum_{k=1}^{T} \left[V^{\pi^{\star}_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar}},1}(\rho) - V^{\pi^{(k)}}_{\mathcal{M}_{\mathrm{tar}},1}(\rho) \right]. \quad (1)$$

Critics and Bellman equations. A critic $f \in \mathcal{F}$ consists of functions $f_h : \mathcal{S}\times\mathcal{A} \to [0, H-h+1]$, with $f_{H+1} \equiv 0$. For randomized policies, $f_h(s, \pi_h(s))$ abbreviates $\mathbb{E}_{a\sim\pi_h(\cdot|s)} f_h(s,a)$. The backup for policy π is

$$(\mathcal{T}^{\pi}_{\mathcal{M},h} f_{h+1})(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_{\mathcal{M},h}(\cdot|s,a)} f_{h+1}(s', \pi_{h+1}(s')).$$

The analysis must distinguish the policy whose Bellman equation is tested from the policy determining where it is tested. For a backup policy π and an evaluation policy ν, define

$$\mathcal{E}^{\nu,\pi}_{\mathcal{M}}(f)_h := \mathbb{E}_{(s,a)\sim d^{\nu}_{\mathcal{M},h}} \left[\left(f_h(s,a) - \mathcal{T}^{\pi}_{\mathcal{M},h} f_{h+1}(s,a) \right)^2 \right], \qquad \mathcal{E}^{\pi}_{\mathcal{M}}(f)_h := \mathcal{E}^{\pi,\pi}_{\mathcal{M}}(f)_h. \quad (2)$$

Source observations control this error with $\mathcal{M} = \mathcal{M}_{\mathrm{src}}$ and $\nu = \mu$. Policy improvement in the target instead involves target occupancies of both the current policy and a comparator, while retaining the current policy in the backup. Transferring between these errors therefore involves both a change of distribution and a change of Bellman operator.

Function-class assumptions. Realizability and Bellman completeness are imposed on a policy collection $\overline{\Pi}$ fixed before observing data (Munos & Szepesvári, 2008; Chen & Jiang, 2019; Xie et al., 2021a). This collection contains μ, the fixed policies used by the algorithms, and all policies obtainable through at most $L = T + T_{\mathrm{off}}$ exponential-weights updates with critics in $\mathcal{F}$ and at most two uniform mixings. For finite $\mathcal{F}$ and a finite learning-rate set $\mathcal{C}$, its complexity satisfies $\log|\overline{\Pi}| \lesssim (T + T_{\mathrm{off}}) \log(1 + |\mathcal{C}||\mathcal{F}|)$ for fixed base and evaluation-policy lists; Appendix B.1 specifies this closure and its role in uniform concentration.

Assumption 1 (Realizability)**.** *For each $\mathcal{M} \in \{\mathcal{M}_{\mathrm{src}}, \mathcal{M}_{\mathrm{tar}}\}$ and $\pi \in \overline{\Pi}$, the tuple $Q^{\pi}_{\mathcal{M}}$ belongs to $\mathcal{F}$.*

Assumption 2 (Bellman completeness)**.** *For each $\mathcal{M} \in \{\mathcal{M}_{\mathrm{src}}, \mathcal{M}_{\mathrm{tar}}\}$, $\pi \in \overline{\Pi}$, $f \in \mathcal{F}$, and $h \in [H]$, the backup $\mathcal{T}^{\pi}_{\mathcal{M},h} f_{h+1}$ belongs to $\mathcal{F}_h$.*

Here $\mathcal{F}_h$ is the stage-h component class. The comparator $\pi^{\star}_{\mathrm{tar}}$ need not belong to $\overline{\Pi}$: the analysis uses its occupancy, not a realizable critic or a complete backup for that policy. Constants C may change between displays.

2.1 What source consistency controls

Fix the backup policy π. At a population tolerance $\epsilon \geq 0$, the source-consistent critics are

$$\mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon) := \left\{ f \in \mathcal{F} : \mathcal{E}^{\mu,\pi}_{\mathcal{M}_{\mathrm{src}}}(f)_j \leq \epsilon \text{ for every } j \in [H] \right\}.$$

The statistical analysis places the selected critic in this set with high probability, which leaves a deterministic question, namely what target residual bound follows from membership alone. We answer it with the smallest uniform upper bound, adapting the transfer-modulus approach of Hanneke & Kpotufe (2024) to policy-dependent Bellman errors.

Definition 1 (Source-to-target (weak) transfer modulus)**.** *For evaluation and backup policies ν, π, a stage $h \in [H]$, and $\epsilon \geq 0$, let*

$$\delta^{\nu|\pi}_h(\epsilon) := \inf \left\{ b \geq 0 : \mathcal{E}^{\nu,\pi}_{\mathcal{M}_{\mathrm{tar}}}(f)_h \leq b \quad \textit{for every } f \in \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon) \right\}, \qquad \delta^{\pi}_h(\epsilon) := \delta^{\pi|\pi}_h(\epsilon). \quad (3)$$

Equivalently, $\delta^{\nu|\pi}_h(\epsilon)$ is the supremum of the target error over the source-consistent critics, with value zero for an empty set. Under Assumption 1, $Q^{\pi}_{\mathcal{M}_{\mathrm{src}}}$ belongs to that set whenever $\pi \in \overline{\Pi}$.

The modulus is a bound-side object that the algorithms never estimate. It is nondecreasing in ϵ and gives the sharp uniform implication from source consistency to target residual control for the specified class and policies; it need not characterize the regret of an algorithm that acquires additional target information.

Two effects can make this implication weak. Limited source coverage leaves critics unconstrained on target-relevant state-action pairs, and a change of dynamics can leave a nonzero target residual even for the exact source value function. In a common environment with occupancy domination $d^{\nu}_{\mathcal{M},h} \le C d^{\mu}_{\mathcal{M},h}$, by contrast, $\delta_h^{\nu|\pi}(\epsilon) \le C\epsilon$. Appendix B.1 relates the modulus to data-diversity and policy-transfer coefficients (Nguyen-Tang & Arora, 2023; 2024) and illustrates its policy dependence.

3 Pessimistic Policy Optimization from Source Data

The modulus suggests a simple principle, that source data should restrict the set of plausible critics but not be treated as target data. Let $\widehat{\mathcal{E}}^{\mu,\pi}_{\mathrm{src}}(f)_h$ be the empirical analogue of (2) on $\mathcal{D}_{\mathrm{src}}$ (the variance-corrected estimator of Appendix B.4). Given a radius $\beta > 0$, the offline procedure (Algorithm 2 in Appendix C.1) forms the source confidence set, selects the most conservative source-plausible critic for the current policy, and updates the policy by per-state multiplicative weights,

$$\widehat{\mathcal{F}}_{\mathrm{src}}(\beta,\pi) = \big\{ f \in \mathcal{F} : \widehat{\mathcal{E}}^{\mu,\pi}_{\mathrm{src}}(f)_h \le \beta \text{ for all } h \big\}, \qquad f^{(t)} \in \operatorname*{argmin}_{f \in \widehat{\mathcal{F}}_{\mathrm{src}}(\beta,\pi^{(t)})} f_1(\rho, \pi_1^{(t)}), \tag{4}$$

$$\pi_h^{(t+1)}(a|s) \propto \pi_h^{(t)}(a|s) \exp\big(\eta f_h^{(t)}(s,a)\big), \qquad h \in [H], \tag{5}$$

where $f_1(\rho,\pi_1) = \mathbb{E}_{s\sim\rho, a\sim\pi_1(\cdot|s)}[f_1(s,a)]$. The analysis rests on the following uniform validity event, which holds with probability at least $1-\delta$ under standard concentration (Appendix B.4), with $\epsilon_{\mathrm{off}}(n,\delta) = \widetilde{O}(H^2/\sqrt{n})$ up to the closure complexity, improving to the Bernstein scale $\widetilde{O}(H^2/n)$ once n exceeds that complexity,

$$\widehat{\mathcal{F}}_{\mathrm{src}}(\beta,\pi) \subseteq \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon_{\mathrm{off}}(n,\delta)) \quad \text{and} \quad Q^{\pi}_{\mathcal{M}_{\mathrm{src}}} \in \widehat{\mathcal{F}}_{\mathrm{src}}(\beta,\pi) \qquad \text{for all } \pi \in \overline{\Pi}; \tag{6}$$

the membership clause makes every confidence set nonempty, so the pessimistic selection is well defined, and after T iterations the procedure outputs the uniform mixture $\widehat{\pi}$ of its iterates.

Theorem 1 (Hybrid transfer guarantee)**.** *Suppose $\mathcal{F}$ satisfies realizability and Bellman completeness, and choose β so that the event* (6) *holds with probability at least $1-\delta$ at tolerance $\epsilon_{\mathrm{off}} = \epsilon_{\mathrm{off}}(n,\delta)$ (Appendix B.4). Run the offline procedure for T iterations (the offline iteration count, written T_{off} in the closure accounting) with $\eta = \frac{1}{H}\sqrt{8\log|\mathcal{A}|/T}$. Then, on that event and hence with probability at least $1-\delta$,*

$$\mathrm{SubOpt}_{\mathrm{tar}}(\widehat{\pi}) \le \frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\Big[\sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}|\pi^{(t)}}(\epsilon_{\mathrm{off}})} + \sqrt{\delta_h^{\pi^{(t)}}(\epsilon_{\mathrm{off}})}\Big] + H^2\sqrt{\frac{\log|\mathcal{A}|}{2T}}. \tag{7}$$

Reading the bound. The first modulus asks whether the source certifies the iterate's Bellman backups on the pairs that $\pi^\star_{\mathrm{tar}}$ visits, where the target is never observed; the second charges every shifted or uncovered pair the iterates steer into, since pessimism in the source is pessimism in the target only where the two agree. With no shift and good coverage both moduli are of order ϵ_{off} times a concentrability constant; with a shift on the relevant pairs they are of order H^2, which Theorem 7 shows is unavoidable. The proof (Appendix C.2) combines an extended value-difference decomposition (Cai et al., 2020; Shani et al., 2020), per-state Hedge (Freund & Schapire, 1997), and the validity event. It applies to any selection from the confidence set; a max-min variant over a finite comparator class removes the optimization term at the cost of enumerating the class (Proposition 18 in Appendix C.1).

4 Online Adaptation with Target Data

Deploying the pessimistic update online, with pooled target residuals used only to screen the source confidence set (Algorithm 3), gives $\mathrm{Regret}(T) \le T\Gamma_T + H^2\sqrt{T\log|\mathcal{A}|/2}$, where Γ_T averages the two transfer terms of Theorem 1 over the iterates (Theorem 21 in Appendix D.1, which also fixes the source radius β and tolerance ϵ_{off} used below). Its proof uses only source-set membership, so target data do not reduce the transfer terms. The block construction below uses target evidence for that purpose, and Section 4.2 compares learners through observed returns.

Algorithm 1: Lazy Target-Adaptive Hedge

Input: $\mathcal{F}$, $\mathcal{D}_{\mathrm{src}}$, integer block length $1 \le w \le T$, radii β, γ, learning rate η_K, $K = \lceil T/w \rceil$

initialize $\pi^{(1)}$ uniform;
for $k = 1, \dots, K$ **do**
  deploy $\pi^{(k)}$ for $w_k := \min\{w,\, T - (k-1)w\}$ episodes and collect $\mathcal{D}_k$;
  if $k = K$ **then**
    stop; do not test, record a fallback, or update on the final block;
  form $\widehat{\mathcal{G}}_k = \widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi^{(k)}) \cap \{f : \widehat{\mathcal{E}}^{(k)}(f; \pi^{(k)})_h \le \gamma \ \forall h\}$, with $\widehat{\mathcal{E}}^{(k)}$ the corrected residual on $\mathcal{D}_k$ under $\pi^{(k)}$ (same-block certification);
  if $\widehat{\mathcal{G}}_k = \emptyset$ **then**
    record a fallback and set $\pi^{(k+1)} \leftarrow \pi^{(k)}$;
  else
    select $f^{(k)} \in \operatorname{argmin}_{f \in \widehat{\mathcal{G}}_k} f_1(\rho, \pi_1^{(k)})$ and set $\pi_h^{(k+1)}(a|s) \propto \pi_h^{(k)}(a|s) \exp(\eta_K f_h^{(k)}(s,a))$;

4.1 Certification with target observations

The source-only implication does not use the evidence acquired during deployment. Suppose target observations additionally constrain the Bellman residual under a reference policy ν'. The backup policy remains π; the reference policy specifies where target data are informative, while ν specifies where the analysis needs control. Conditioning the source-to-target bound on this evidence gives the following target-localized quantity, following the localization principle of Hanneke & Kpotufe (2024).

Definition 2 (Target-conditioned (strong) transfer modulus)**.** *For policies ν, π, ν', stage $h \in [H]$, and tolerances $\epsilon, \gamma \ge 0$, define*

$$\widetilde{\delta}_h^{\nu|\pi}(\epsilon; \gamma, \nu') := \inf \big\{ b \ge 0 : \mathcal{E}^{\nu,\pi}_{\mathcal{M}_{\mathrm{tar}}}(f)_h \le b \text{ for every } f \in \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon) \text{ such that } \mathcal{E}^{\nu',\pi}_{\mathcal{M}_{\mathrm{tar}}}(f)_j \le \gamma \text{ for every } j \in [H] \big\}. \tag{8}$$

This is equivalently the supremum of the target error over critics satisfying both sets of constraints, with value zero when none do. Removing the target constraint recovers $\widetilde{\delta}_h^{\nu|\pi}(\epsilon; \infty, \nu') = \delta_h^{\nu|\pi}(\epsilon)$.

Additional constraints can only decrease the uniform bound. If $d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h} \le C d^{\nu'}_{\mathcal{M}_{\mathrm{tar}},h}$, then $\widetilde{\delta}_h^{\nu|\pi}(\epsilon; \gamma, \nu') \le \min\{\delta_h^{\nu|\pi}(\epsilon), C\gamma\}$ (Lemma 25). In particular, $\nu = \nu'$ gives a cap of γ. Target evidence along the deployed policy gives no such cap along an arbitrary comparator without a coverage relation, and an empty feasible set, which has value zero by convention, supplies no usable critic.

Algorithm 1 makes this conditioning operational by holding a policy fixed over each block. The critic used for the update after block k must satisfy constraints derived from that block's own observations, with backup and reference policy both equal to $\pi^{(k)}$. This provides the deployed-policy cap while retaining a separate comparator term. A failed empirical feasibility test triggers fallback; Theorem 4 describes a handoff to online learning.

Theorem 2 (Strong-modulus regret bound)**.** *Suppose $\mathcal{F}$ is finite and satisfies realizability and Bellman completeness for both environments. Run Algorithm 1 with block length $1 \le w \le T$, $\eta_K = \frac{1}{H}\sqrt{8 \log|\mathcal{A}|/K}$, target radius $\gamma = \bar{\alpha}(w)/2$ with $\bar{\alpha}(w) := CH^2\sqrt{\log(|\mathcal{F}|HT/\delta)/w}$ for a sufficiently large C, and $\beta, \epsilon_{\mathrm{off}}$ as in Theorem 21. Then, with probability at least $1 - 2\delta$,*

$$\mathrm{Regret}(T) \le \sum_{k \notin \mathcal{B},\, k<K} w \sum_{h=1}^{H} \Big[\sqrt{\widetilde{\delta}_h^{\pi^\star_{\mathrm{tar}}|\pi^{(k)}}\big(\epsilon_{\mathrm{off}}; \bar{\gamma}_k, \pi^{(k)}\big)} + \sqrt{\min\big\{\delta_h^{\pi^{(k)}}(\epsilon_{\mathrm{off}}),\ \bar{\gamma}_k\big\}} \Big] + H^2\sqrt{wT \log|\mathcal{A}|} + wH\,(N_{\mathrm{fb}} + 1), \tag{9}$$

where $\mathcal{B} \subseteq \{1, \dots, K-1\}$ contains the tested blocks whose intersection is empty and $N_{\mathrm{fb}} := |\mathcal{B}|$. The final block K, whether full or partial, is not tested and is not a fallback. Each fallback block and the final block receive no source-driven update and are charged at most wH regret, and $\bar{\gamma}_k = \bar{\alpha}(w)$ for every certified block. In particular, with $w = \lceil T^{2/3} \rceil$ and no fallback, $\mathrm{Regret}(T) \le T\,\widetilde{\Gamma}_T + \widetilde{O}(H^2 T^{5/6})$*, where the strong transfer level* $\widetilde{\Gamma}_T := \frac{1}{K}\sum_{k<K}\sum_h (\widetilde{\delta}_h^{\pi^\star_{\mathrm{tar}}|\pi^{(k)}}(\epsilon_{\mathrm{off}}; \bar{\alpha}(w), \pi^{(k)}))^{1/2}$ *averages the comparator terms over the blocks.*

Reading the bound. The iterate charge, which was the weak modulus and could be of order H^2 on a shifted route, is now capped by the target-data radius $\bar{\gamma}_k = \widetilde{O}(H^2 w^{-1/2})$ whatever the shift, because every source-driven update is checked against the deployed policy's own w episodes; the comparator charge is the strong modulus, which can collapse to the target radius when the iterates visit what $\pi^\star_{\mathrm{tar}}$ visits while the weak modulus stays as large as the shift. The price is the optimization term, $\sqrt{wT}$ instead of $\sqrt{T}$ because the policy moves only T/w times, and one full block per fallback; the Hoeffding calibration balances the two at $w \asymp T^{2/3}$, the Bernstein calibration below at $w \asymp \sqrt{T}$.

Corollary 3 (Fast-rate refinement)**.** *Calibrate the target test at radius $\gamma = 2\gamma_{\mathrm{fast}}(w)$ with $\gamma_{\mathrm{fast}}(w) := C_f H^2 \log(|\mathcal{F}|HK/\delta)/w$. Then Theorem 2 holds with $\bar{\gamma}_k = 6\gamma_{\mathrm{fast}}(w)$ in every certified block, and with $w = \lceil \sqrt{T} \rceil$, unconditionally,* $\mathrm{Regret}(T) \le T\,\widetilde{\Gamma}_T^{\mathrm{fast}} + \widetilde{O}(H^2 T^{3/4}) + 2H\sqrt{T}\,N_{\mathrm{fb}}$*, with $\widetilde{\Gamma}_T^{\mathrm{fast}}$ the strong transfer level at radius $6\gamma_{\mathrm{fast}}(w)$ over the certified blocks, with the sum normalized by the total number K of blocks.*

On the source-validity and block-concentration events, a fallback in a tested block $k < K$ means that the source value witness fails the target test, which certifies $\delta_h^{\pi^{(k)}}(\epsilon_{\mathrm{off}}) > \gamma/2$ for some h under the Hoeffding calibration, or $> \gamma_{\mathrm{fast}}/2$ under the Bernstein calibration (Remark 9). This certifies Bellman incompatibility along the deployed policy, not its suboptimality, and can trigger the following handoff.

Theorem 4 (Certified handoff)**.** *Run Algorithm 1 with the rule that at the first fallback block $k_0 < K$ all remaining episodes are given to a purely online learner $\mathsf{ALG}_{\mathrm{on}}$, started fresh, with a nondecreasing anytime regret bound $R_{\mathrm{on}}(t)$ holding with probability at least $1-\delta$. With probability at least $1-3\delta$, if no fallback occurs the bound of Theorem 2, or Corollary 3 under its calibration, holds with $N_{\mathrm{fb}} = 0$; otherwise*

$$\mathrm{Regret}(T) \;\le\; \sum_{k<k_0} w \sum_{h=1}^{H} \Big[\sqrt{\widetilde{\delta}_h^{\pi^\star_{\mathrm{tar}}|\pi^{(k)}}\big(\epsilon_{\mathrm{off}}; \bar{\gamma}, \pi^{(k)}\big)} + \sqrt{\bar{\gamma}}\Big] + H^2\sqrt{wT\log|\mathcal{A}|} + wH + R_{\mathrm{on}}(T),$$

with $\bar{\gamma} = \bar{\alpha}(w)$ under the Hoeffding calibration and $\bar{\gamma} = 6\gamma_{\mathrm{fast}}(w)$ under the Bernstein one. The exit is conservative, since no fallback occurs at a tested block $k < K$ whose deployed policy satisfies $\max_h \mathcal{E}^{\pi^{(k)},\pi^{(k)}}_{\mathcal{M}_{\mathrm{tar}}}(Q^{\pi^{(k)}}_{\mathcal{M}_{\mathrm{src}}})_h \le \gamma/2$ *(respectively $\le \gamma_{\mathrm{fast}}/2$).*

Reading the bound. Before the exit the learner pays the certified-block bound, after it the online learner's own bound, and the exit costs one block; no factor of M appears because the decision is made by a certificate rather than by comparing returns. The certificate cannot detect a source that is accurate along the deployed route but wrong where $\pi^\star_{\mathrm{tar}}$ goes; the return-based master below addresses that case (proofs in Appendix D.4).

4.2 Adaptation by comparing learner returns

The source-constrained regret bound need not be sublinear when its transfer level is large. We therefore combine source-informed learners with a purely online baseline using regret balancing with elimination (Pacchiano et al., 2020; Cutkosky et al., 2021; Agarwal et al., 2017). Let the online learner have the anytime bound $R_{\mathrm{on}}(t) = \min\{Ht, c_{\mathrm{on}}\sqrt{t\iota_{\mathrm{on}}}\}$, where $c_{\mathrm{on}} \ge H$. Register $M - 1 = \lceil \log_2(HT) \rceil + 1$ source copies with anytime policy updates and candidate bounds $B^{(j)}(t) = \min\{Ht, 2^{-j}Ht + 2H^2\sqrt{t\log|\mathcal{A}|}\}$. Each copy updates only on its own episodes. Write $\mathcal{W}_{\mathrm{ws}}$ for the copies satisfying $\mathrm{Reg}_j(t) \le B^{(j)}(t)$ at every reached prefix of their own episode sequence. This set contains $j = 0$ because its candidate is Ht. Let $\Gamma_{\mathcal{W}}$ be the smallest slope among copies whose slope dominates their prefix-maximal transfer level, a sufficient condition for membership, with $j = 0$ included through the cap (Appendix D.1).

Theorem 5 (Adaptive hybrid learning)**.** *Let $\iota := \log(2MT/\delta)$ and $T \geq MH^2$. Suppose the validity event of Theorem 21 holds with probability at least $1-\delta$ and the envelope R_{on} is a valid anytime bound. Then the balancing-with-elimination master of Proposition 30 satisfies, with probability at least $1-3\delta$, both*

$$\mathrm{Regret}(T) = \widetilde{O}\Big(\big(M + \sqrt{M}\,\tfrac{c_{\mathrm{on}}}{H^2}\big)R_{\mathrm{on}}(T) + H\sqrt{MT\iota} + Mc_{\mathrm{on}}\big(H^2 + \tfrac{c_{\mathrm{on}}}{H}\big)\Big) \qquad \textit{and}$$
$$\mathrm{Regret}(T) = \widetilde{O}\Big(M\Big(\min_{j\in\mathcal{W}_{\mathrm{ws}}} 2^{-j}HT + H^2\sqrt{T\log|\mathcal{A}|}\Big) + H\sqrt{MT\iota} + \sqrt{M}H^3\sqrt{T} + MH^3\Big), \tag{10}$$

hence the minimum of the two, an oracle inequality over $\{\mathsf{ALG}_{\mathrm{on}}\} \cup \mathcal{W}_{\mathrm{ws}}$; the second display also holds with $\Gamma_{\mathcal{W}}$ in place of the minimum (Corollary 28).

Reading the bound. The source comparison is with well-specified copies on their own interleaved episode sequences, not with an uninterrupted source learner. The online branch keeps the multiplier $M+\sqrt{M}\,c_{\mathrm{on}}/H^2$ and the additive term $Mc_{\mathrm{on}}(H^2+c_{\mathrm{on}}/H)$, so its overheads need not be logarithmic in the problem size. Proposition 32 treats block-based copies, whose transfer level includes fallback charges, and Proposition 24 gives a fixed-policy diagnostic.

5 Structured MDPs and Lower Bounds

5.1 Explicit bounds for structured MDPs

In tabular MDPs, let $\Delta_h(s,a) := \|P_{\mathrm{src},h}(\cdot|s,a) - P_{\mathrm{tar},h}(\cdot|s,a)\|_{\mathrm{TV}}$ be the local shift and $\mathrm{Stat}_h(s,a;\epsilon) := \min\{4H^2, \epsilon/d^{\mu}_{\mathcal{M}_{\mathrm{src}},h}(s,a)\}$ the local statistical term.

Theorem 6 (Weak modulus in tabular MDPs)**.** *For every policy π, step h, and $\epsilon \geq 0$, with all critics valued in $[0,H]$,*

$$\delta_h^{\pi}(\epsilon) \leq 2\sum_{s,a} d^{\pi}_{\mathcal{M}_{\mathrm{tar}},h}(s,a)\big[H^2\Delta_h(s,a)^2 + \mathrm{Stat}_h(s,a;\epsilon)\big]. \tag{11}$$

The same bound holds for $\delta_h^{\nu|\pi}(\epsilon)$ with the target visitation of π replaced by that of ν.

Analogous bounds hold in linear MDPs, where covered feature directions are charged through the pseudoinverse of the source feature covariance, unobserved directions through the projection onto its null space, and the dynamics shift through a latent operator discrepancy, and in block MDPs, where the same decomposition applies blockwise (Theorems 36 and 37 in Appendix F.2).

Reading the bound. A pair visited by the evaluation policy contributes $\min\{4H^2, \epsilon/d^{\mu}(s,a)\}$, which vanishes on covered pairs as the source radius shrinks but stays at $4H^2$ on uncovered ones, plus a shift term $H^2\Delta_h(s,a)^2$ that does not decrease with the source sample size. These are upper bounds, so a nonzero shift term alone does not imply a nonzero residual.

5.2 Lower bounds

The weak-modulus dependence is not an artifact of the analysis, since source data may be compatible with several targets that require different optimal actions.

Theorem 7 (General weak-modulus lower bound)**.** *Fix $H \geq 2$, budgets $n, T \in \mathbb{N}$, $K \geq 2$ actions, and levels $\alpha_h \in (0, c_0\sqrt{K/T}\,]$ for the $\lfloor H/2 \rfloor$ decision layers with $\sum_h \alpha_h \leq 1/2$. There is a family of hybrid instances with shared rewards and a common behavior policy such that every instance satisfies $\delta_h^{\pi^\star_{\mathrm{tar}}}(0) = \alpha_h^2$ exactly at every decision layer, and any algorithm using n source and T target episodes outputs a policy with*

$$\mathbb{E}\big[\,\mathrm{SubOpt}_{\mathrm{tar}}(\mathsf{ALG})\big] \geq c\sum_{h=1}^{H}\sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}}(0)} \tag{12}$$

for a universal $c > 0$. The hard class may be taken finite and Bellman complete, and for n large enough the bound persists at radius $\epsilon_{\mathrm{off}}(n)$ up to a constant (Appendix G.2.1).

Reading the bound. At each decision layer every action has the same source transitions, while in the target one hidden action is better by α_h; the source pins down the value function yet says nothing

Table 1: Handoff study: cumulative target pseudo-regret averaged over eight labelings with ten seeds each, with the sample standard deviation of the eight labeling means in parentheses. A2 uses the candidate family but no source samples.

Condition	A: target-only	A2: candidate-aware	C: source blocks	D: handoff
Useful	1676.5 (2.2)	162.2 (16.3)	0.0 (0.0)	1642.7 (9.3)
Misleading	1758.7 (454.6)	105.0 (83.7)	112000.0 (0.0)	24171.0 (456.5)
Poor coverage	2551.5 (581.2)	118.4 (126.6)	37333.3 (0.0)	37333.3 (0.0)
Hidden route	2551.5 (581.2)	118.4 (126.6)	37333.3 (0.0)	37333.3 (0.0)

about which action to take, and every source-consistent critic's target backup is off by exactly α_h at the decision pair. With $\alpha_h \lesssim \sqrt{K/T}$ the target budget cannot separate the actions, so any algorithm loses order $\sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}}(0)}$ per layer. Structured lower bounds matching the shift terms of Theorem 6 appear in Appendix G.1.

6 Experiments: What Source Data Certify, and When to Switch

We test the three algorithmic questions on a finite-class instance in which source constraints, target constraints, and pessimistic optimization are computed exactly. The results separate successful certification from successful adaptation.

6.1 A finite-class implementation with exact evaluation

The four-stage MDP has seven states and two actions. The first decision selects corridor A or B, and a later decision selects a Bernoulli success probability within that corridor; both corridors give reward 0.8, followed by reward 1 on success. Source and target share rewards and differ only in their success probabilities. Starting from $\mathcal{F}_{H+1} = \{0\}$, we close each component class under Bellman backups of a predetermined family of 256 candidate kernels, shared across conditions, for the finite policy set. The product class has component sizes $(49, 49, 256, 1)$, or $614{,}656$ critics; it contains each tested model's policy values, is Bellman complete for the included policies, and is not constructed from the target instance. A layered dynamic program solves feasibility and pessimistic optimization over this class with the variance-corrected excess loss of Appendix B.4. Initial selection implements the max-min variant of Proposition 18, and block updates use greedy improvement within the finite policy set in place of exponential weights, so we study certification and switching without claiming the guarantees of Theorems 2 and 4 (Appendix H.1). Each condition supplies $50{,}000$ source episodes, and cumulative target pseudo-regret $\sum_t (V^\star_{\mathrm{tar}} - V^{\pi_t}_{\mathrm{tar}})$ is computed by dynamic programming. The four conditions are useful transfer, misleading transfer, poor source coverage (low but nonzero visitation) with identical dynamics, and a better target route that the source discourages visiting (Table 6).

We compare an optimistic tabular target-only learner (A), an optimistic learner given the candidate family but no source observations (A2), the source-constrained block learner (C), and C followed by a fresh A once the source–target intersection is empty (D). A and A2 differ in both information and learning rule, and none of these implementations has an established anytime envelope.

6.2 Incompatibility is distinct from policy suboptimality

Table 1 uses $T = 160{,}000$ target episodes and blocks of $w = 32{,}000$, and averages over the eight isomorphic labelings generated by swapping the corridors and the actions at the two later decision states, which controls the label sensitivity of the optimistic baselines.

A correct rejection can induce unnecessary exploration. With the useful source, C plays a target-optimal policy and has zero regret, yet the test rejects the intersection. The rejection is a correct certificate in the sense of Remark 9, since the source value function has a population squared target residual of about 0.0544 at the shifted stage, above $\gamma/2$, while the deployed action remains optimal. Handoff then costs about $1{,}643$, all of it incurred during fresh online learning after the switch (Appendix H.1.3).

Table 2: Return-based selection: cumulative target pseudo-regret, mean $\pm$ sample standard deviation over five held-out seeds at one labeling; each standalone control receives the full budget $T = 168{,}000$.

Condition	Master	C alone	A alone	A2 alone
Useful	140.0 ± 0.0	0.0 ± 0.0	1680.0 ± 46.7	140.0 ± 0.0
Misleading	51671.3 ± 347.0	117600.0 ± 0.0	1400.0 ± 0.0	140.0 ± 0.0
Poor coverage	39340.0 ± 0.0	39200.0 ± 0.0	3084.7 ± 38.3	0.0 ± 0.0
Hidden route	39340.0 ± 0.0	39200.0 ± 0.0	3084.7 ± 38.3	0.0 ± 0.0

Rejection helps with a misleading source but does not establish timing. With the misleading source, D reduces regret from $112{,}000$ to $24{,}171$. A fixed-time control (E), calibrated on separate seeds, selects immediate switching, which coincides with A, and outperforms D on the equally weighted mixture of the four conditions (Appendix H.1.3). The evidence supports recovery after rejecting a misleading source, not certificate-triggered switching times.

Compatibility does not certify optimality. In the hidden-route condition, the source value function has zero target Bellman residual along the deployed route, while the other route is better by 0.2333 per episode, so on-policy data cannot force rejection. Under poor coverage the environments coincide and there is nothing to reject. C and D accumulate $37{,}333$ regret in both conditions, which is the comparator term of the theory. A2 outperforms C and D in three conditions, so these instances admit effective target learning once the candidate family is known.

6.3 Return-based selection: partial recovery and its cost

We run the selection and elimination rule of Proposition 30 with twenty copies of C under different asserted candidate envelopes and one copy of A, at $T = 168{,}000$ and $w = 8{,}000$. Each copy advances only on its own episodes, and the master sees only observed returns. Prefix validity of the envelopes is checked retrospectively along the observed trajectories, so this tests the generic master with asserted envelopes rather than the theorem-specific composition.

With the useful source the master pays 140, as does A2, against C's 0 and A's $1{,}680$. With the misleading source it reduces C's $117{,}600$ to $51{,}671$, closing about 56.7% of the gap to A while remaining far above A ($1{,}400$) and A2 (140); seventeen of the twenty source copies are eliminated, none completes a block, and the online copy receives 57.2% of the episodes. Under poor coverage and the hidden route, no copy is eliminated and the master does not improve on C; there the online copy also exceeds its asserted envelope, so these runs do not instantiate the proposition's comparison with a valid online baseline. No source copy makes a certified policy change in any condition, so the experiment demonstrates return-based selection among unchanged source policies and a learning online copy, not adaptation through certified updates. The trajectory-dependent bound of Proposition 30, minimized over the retrospectively prefix-valid copies, holds in every run; it lies below the reward-based cap VT in two conditions and above it in the other two (Appendix H.1.5), and it is a numerical check rather than an anytime guarantee.

7 Conclusion

We analyze policy optimization, target certification, and adaptation when offline and online data come from different dynamics. Policy-dependent Bellman residuals quantify the error incurred when source consistency is used in the target. Target observations can constrain deployed-policy residuals and support a handoff, and a return-based master compares learning procedures with valid candidate bounds; the guarantees retain comparator terms and adaptation costs. The finite-class experiments show that incompatible critics can accompany an optimal policy, that on-policy compatibility can coexist with a better unvisited route, and that return-based selection reduces harmful transfer at a substantial cost relative to online controls.

Reproducibility statement

The assumptions are stated in Section 2, and complete proofs of all results are given in the appendix. Appendix H describes the experimental protocols, and the supplementary material contains the code and result logs for the finite-class experiments.

Ethics statement

This is a theoretical study of reinforcement learning that raises no specific ethical concerns. Its guarantees concern learning performance and do not certify the safety of a deployed policy.

AI use statement

In this work, we used generative AI tools to help design and provide feedback on the research methodology and experiments, clean and reformat data, and interpret results. We have not used generative AI tools to develop the theoretical framework, formulate mathematical claims, provide critical ingredients for proving them, assist in the writing of proofs, implement methods, generate synthetic data sets or to propose or refine hypotheses, and assisting with translation and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to summarize and analyze existing literature, format references, suggest a structure for the paper, and propose a title and keywords. We have reviewed all AI-assisted work. The authors checked the literature summaries and bibliographic entries against the original papers and their publisher or arXiv records. AI-written scripts converted the raw training logs of the continuous-control study into evaluation curves and tables, and the authors checked the converted values against the original logs. AI tools also drafted interpretations of the experimental results. The authors compared each interpretation with the result logs and removed or qualified any claim the data did not support, such as when a rejection reflected a finite radius rather than a population incompatibility. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 12–38. PMLR, 2017. URL https://proceedings.mlr.press/v65/agarwal17b.html.

Alekh Agarwal, Mikael Henaff, Sham M. Kakade, and Wen Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13399–13412, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/9b3a9fb4db30fc6594ec3990cbc09932-Abstract.html.

Alekh Agarwal, Sham M. Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20095–20107, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/e894d787e2fd6c133af47140aa156f00-Abstract.html.

Philip Amortila, Dylan J. Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=THJEa8adBn.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008. doi: 10.1007/s10994-007-5038-2. URL https://doi.org/10.1007/s10994-007-5038-2.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012. doi: 10.4086/toc.2012.v008a006. URL https://theoryofcomputing.org/articles/v008a006/.

Karl J. Åström and Björn Wittenmark. *Adaptive Control*. Addison-Wesley, 2nd edition, 1995. URL https://lup.lub.lu.se/record/4810314.

Akram S. Awad, Shihab Ahmed, Yue Wang, and George K. Atia. Robust transfer learning with side information. *arXiv preprint arXiv:2603.07921*, 2026. URL https://arxiv.org/abs/2603.07921.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 263–272. PMLR, 2017. URL https://proceedings.mlr.press/v70/azar17a.html.

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1577–1594. PMLR, 2023. URL https://proceedings.mlr.press/v202/ball23a.html.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 122–131. AUAI Press, 2013. URL https://arxiv.org/abs/1309.6821.

Changxiao Cai, T. Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*, 52(1):207–232, 2024. doi: 10.1214/23-AOS2341. URL https://doi.org/10.1214/23-AOS2341.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1283–1294. PMLR, 2020. URL https://proceedings.mlr.press/v119/cai20d.html.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. URL https://doi.org/10.1017/CBO9780511546921.

Jinhang Chai, Elynn Chen, and Lin Yang. Transfer Q-learning with composite MDP structures. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7089–7106. PMLR, 2025. URL https://proceedings.mlr.press/v267/chai25b.html.

Elynn Chen, Xi Chen, and Wenbo Jing. Data-driven knowledge transfer in batch Q^* learning. *Journal of the American Statistical Association*, 121(553):276–288, 2026a. doi: 10.1080/01621459.2025.2603731. URL https://doi.org/10.1080/01621459.2025.2603731.

Elynn Chen, Enpei Zhang, Jinhang Chai, and Yujun Yan. One-step Bellman alignment enables provably efficient transfer in online RL. *arXiv preprint arXiv:2601.21924*, 2026b. URL https://arxiv.org/abs/2601.21924.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051. PMLR, 2019. URL https://proceedings.mlr.press/v97/chen19e.html.

Keru Chen. Information-directed offline-to-online reinforcement learning. *arXiv preprint arXiv:2605.29405*, 2026. URL https://arxiv.org/abs/2605.29405.

Ming-Hong Chen, Kuan-Chen Pan, You-De Huang, Xi Liu, and Ping-Chun Hsieh. Cross-domain policy optimization via Bellman consistency and hybrid critics. In *International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=kTXRFtWHnM.

Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8286–8309. PMLR, 2024. URL https://proceedings.mlr.press/v235/cheung24a.html.

Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, and Manish Purohit. Dynamic balancing for model selection in bandits and RL. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2276–2285. PMLR, 2021. URL https://proceedings.mlr.press/v139/cutkosky21a.html.

Mark Cutler, Thomas J. Walsh, and Jonathan P. How. Real-world reinforcement learning via multifidelity simulators. *IEEE Transactions on Robotics*, 31(3):655–671, 2015. doi: 10.1109/TRO.2015.2419431. URL https://ieeexplore.ieee.org/document/7106543/.

Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1665–1674. PMLR, 2019. URL https://proceedings.mlr.press/v97/du19b.html.

Benjamin Eysenbach, Shreyas Chaudhari, Swapnil Asawa, Sergey Levine, and Ruslan Salakhutdinov. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eqBwg3AcIAK.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504. URL https://doi.org/10.1006/jcss.1997.1504.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020. URL https://arxiv.org/abs/2004.07219.

Yihong Guo, Yu Yang, Pan Xu, and Anqi Liu. MOBODY: Model-based off-dynamics offline reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0bd3512c29ecd987fbd5650891c90562-Abstract-Conference.html.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9867–9877, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/b91f4f4d36fa98a94ac5584af95594a0-Abstract.html.

Steve Hanneke and Samory Kpotufe. Adaptive sample aggregation in transfer learning. *arXiv preprint arXiv:2408.16189*, 2024. URL https://arxiv.org/abs/2408.16189.

Yiting He, Zhishuai Liu, Weixin Wang, and Pan Xu. Sample complexity of distributionally robust off-dynamics reinforcement learning with online interaction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22595–22646. PMLR, 2025. URL https://proceedings.mlr.press/v267/he25n.html.

Ruiquan Huang, Donghao Li, Chengshuai Shi, Cong Shen, and Jing Yang. Augmenting online RL with offline data is all you need: A unified hybrid RL algorithm design and analysis. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1745–1767. PMLR, 2025. URL https://proceedings.mlr.press/v286/huang25a.html.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. doi: 10.1287/moor.1040.0129. URL https://pubsonline.informs.org/doi/abs/10.1287/moor.1040.0129.

Nan Jiang. PAC reinforcement learning with an imperfect model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3334–3341. AAAI Press, 2018. doi: 10.1609/aaai.v32i1.11594. URL https://ojs.aaai.org/index.php/AAAI/article/view/11594.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5084–5096. PMLR, 2021. URL https://proceedings.mlr.press/v139/jin21e.html.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier B. Oliva, Jeff Schneider, and Barnabás Póczos. Gaussian process bandit optimisation with multi-fidelity evaluations. In *Advances in Neural Information Processing Systems*, volume 29, pp. 992–1000, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/605ff764c617d3cd28dbbdd72be8f9a2-Abstract.html.

Lingkai Kong, Haichuan Wang, Tonghan Wang, Guojun Xiong, and Milind Tambe. Composite flow matching for reinforcement learning with shifted-dynamics data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 92500–92541, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/855b928e3f64498da7d98a3996bd4971-Abstract-Conference.html.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1882–1886. PMLR, 2018. URL https://proceedings.mlr.press/v75/kpotufe18a.html.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. URL https://tor-lattimore.com/downloads/book/book.pdf.

Gen Li, Wenhao Zhan, Jason D. Lee, Yuejie Chi, and Yuxin Chen. Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55582–55615, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ade04fd4f26263f86b47ffb535c4cafb-Abstract-Conference.html.

Shangzhe Li and Weitong Zhang. Provably efficient offline-to-online value adaptation with general function approximation. *arXiv preprint arXiv:2604.13966*, 2026. URL https://arxiv.org/abs/2604.13966.

Jinxin Liu, Hongyin Zhang, and Donglin Wang. DARA: Dynamics-aware reward augmentation in offline reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9SDQB3b68K.

Jinxin Liu, Ziqi Zhang, Zhenyu Wei, Zifeng Zhuang, Yachen Kang, Sibo Gai, and Donglin Wang. Beyond OOD state actions: Supported cross-domain offline reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13945–13953, 2024a. doi: 10.1609/aaai.v38i12.29302. URL https://ojs.aaai.org/index.php/AAAI/article/view/29302.

Wei Liu and Ting Long. Target-aligned Bellman backup for cross-domain offline reinforcement learning. *arXiv preprint arXiv:2605.22376*, 2026. URL https://arxiv.org/abs/2605.22376.

Zhishuai Liu and Pan Xu. Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2719–2727. PMLR, 2024. URL https://proceedings.mlr.press/v238/liu24d.html.

Zhishuai Liu, Weixin Wang, and Pan Xu. Upper and lower bounds for distributionally robust off-dynamics reinforcement learning. *arXiv preprint arXiv:2409.20521*, 2024b. URL https://arxiv.org/abs/2409.20521.

Jiafei Lyu, Chenjia Bai, Jing-Wen Yang, Zongqing Lu, and Xiu Li. Cross-domain policy adaptation by capturing representation mismatch. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33638–33663. PMLR, 2024a. URL https://proceedings.mlr.press/v235/lyu24a.html.

Jiafei Lyu, Kang Xu, Jiacheng Xu, Mengbei Yan, Jingwen Yang, Zongzhang Zhang, Chenjia Bai, Zongqing Lu, and Xiu Li. ODRL: A benchmark for off-dynamics reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59859–59911, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6e37ba8a27b53fccd367c648ff365b8f-Abstract-Datasets_and_Benchmarks_Track.html.

Jiafei Lyu, Mengbei Yan, Zhongjian Qiao, Runze Liu, Xiaoteng Ma, Deheng Ye, Jing-Wen Yang, Zongqing Lu, and Xiu Li. Cross-domain offline policy adaptation with optimal transport and dataset constraint. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LRrbD8EZJl.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/003.pdf.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank MDPs. *Journal of Machine Learning Research*, 25(6):1–76, 2024. URL https://jmlr.org/papers/v25/22-0687.html.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008. URL https://jmlr.org/papers/v9/munos08a.html.

Mitsuhiko Nakamoto, Yuexiang Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-QL: Calibrated offline RL pre-training for efficient online fine-tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 62244–62269, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c44a04289beaf0a7d968a94066a1d696-Abstract-Conference.html.

Thanh Nguyen-Tang and Raman Arora. On sample-efficient offline reinforcement learning: Data diversity, posterior sampling and beyond. In *Advances in Neural Information Processing Systems*, volume 36, pp. 61115–61157, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c056d6cf7b7108418f2b8c307dfaab02-Abstract-Conference.html.

Thanh Nguyen-Tang and Raman Arora. On the statistical complexity of offline decision-making. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37900–37928. PMLR, 2024. URL https://proceedings.mlr.press/v235/nguyen-tang24a.html.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216. URL https://pubsonline.informs.org/doi/abs/10.1287/opre.1050.0216.

Haoyi Niu, Shubham Sharma, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, and Xianyuan Zhan. When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36599–36612, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ed3cd2520148b577039adfade82a5566-Abstract-Conference.html.

Haoyi Niu, Tianying Ji, Bingqi Liu, Haocheng Zhao, Xiangyu Zhu, Jianying Zheng, Pengfei Huang, Guyue Zhou, Jianming Hu, and Xianyuan Zhan. H2O+: An improved framework for hybrid offline-and-online RL with dynamics gaps. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1421–1428. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11127503. URL https://ieeexplore.ieee.org/document/11127503.

Aldo Pacchiano, Christoph Dann, Claudio Gentile, and Peter Bartlett. Regret bound balancing and elimination for model selection in bandits and RL. *arXiv preprint arXiv:2012.13045*, 2020. URL https://arxiv.org/abs/2012.13045.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9582–9602. PMLR, 2022. URL https://proceedings.mlr.press/v151/panaganti22a.html.

Zhongjian Qiao, Rui Yang, Jiafei Lyu, Chenjia Bai, Xiu Li, Siyang Gao, and Shuang Qiu. Efficient cross-domain offline reinforcement learning with dynamics- and value-aligned data filtering. *arXiv preprint arXiv:2512.02435*, 2026a. URL https://arxiv.org/abs/2512.02435.

Zhongjian Qiao, Rui Yang, Jiafei Lyu, Xiu Li, Zhongxiang Dai, Zhuoran Yang, Siyang Gao, and Shuang Qiu. Dual-robust cross-domain offline reinforcement learning against dynamics shifts. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/71cc3acb2578bdd1e52bf9f85bbb1af0-Abstract-Conference.html.

Chengrui Qu, Laixi Shi, Kishan Panaganti, Pengcheng You, and Adam Wierman. Hybrid transfer reinforcement learning: Provable sample efficiency from shifted-dynamics data. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1054–1062. PMLR, 2025. URL https://proceedings.mlr.press/v258/qu25a.html.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11702–11716, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/60ce36723c17bbac504f2ef4c8a46995-Abstract.html.

Tyler Sam, Yudong Chen, and Christina Lee Yu. The limits of transfer reinforcement learning with latent low-rank structure. In *Advances in Neural Information Processing Systems*, volume 37, pp. 108262–108330, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c3fa3a7d50b34732c6d08f6f66380d75-Abstract-Conference.html.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8604–8613. PMLR, 2020. URL https://proceedings.mlr.press/v119/shani20a.html.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79903–79917, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fc8ee7c7ab5b5f6b1615045dfb617ed6-Abstract-Conference.html.

Yuda Song, Yifei Zhou, Ayush Sekhari, J. Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: Using both offline and online data can make RL efficient. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yyBis80iUuU.

Joseph Suk and Samory Kpotufe. Self-tuning bandits over unknown covariate-shifts. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 1114–1156. PMLR, 2021. URL https://proceedings.mlr.press/v132/suk21a.html.

Kevin Tan and Ziping Xu. A natural extension to online algorithms for hybrid RL with limited coverage. *Reinforcement Learning Journal*, 3:1252–1264, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper152.html.

Kevin Tan, Wei Fan, and Yuting Wei. Hybrid reinforcement learning breaks sample size barriers in linear MDPs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 120038–120077, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d9251dc20346dffe9b6db86dcc6f8cc9-Abstract-Conference.html.

Matthew E. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(56):1633–1685, 2009. URL https://jmlr.org/papers/v10/taylor09a.html.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=tyrJsbKAe6.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35300–35338. PMLR, 2023. URL https://proceedings.mlr.press/v202/wagenmaker23a.html.

Guojian Wang, Quinson Hon, Xuyang Chen, and Lin Zhao. Target-aligned fusion for decision-sequence learning under dynamics shift. *arXiv preprint arXiv:2511.09173*, 2026. URL https://arxiv.org/abs/2511.09173.

Ruosong Wang, Dean P. Foster, and Sham M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=30EvkP2aQLD.

Xiaoyu Wen, Chenjia Bai, Kang Xu, Xudong Yu, Yang Zhang, Xuelong Li, and Zhen Wang. Contrastive representation for data filtering in cross-domain offline reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52720–52743. PMLR, 2024. URL https://proceedings.mlr.press/v235/wen24a.html.

Zheshun Wu, Renjie Zheng, Jinhang Zuo, Zenglin Xu, and Fang Kong. A unified algorithmic framework for hybrid reinforcement learning in tabular MDPs with shifted transition dynamics. *arXiv preprint arXiv:2607.25207*, 2026. URL https://arxiv.org/abs/2607.25207.

Zhangjie Xia, Yu Yang, and Pan Xu. Localized dynamics-aware domain adaption for off-dynamics offline reinforcement learning. *arXiv preprint arXiv:2602.21072*, 2026. URL https://arxiv.org/abs/2602.21072.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6683–6694, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/34f98c7c5d7063181da890ea8d25265a-Abstract.html.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27395–27407, 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/e61eaa38aed621dd776d0e67cfeee366-Abstract.html.

Tengyang Xie, Dylan J. Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=LQIjzPdDt3q.

Kang Xu, Chenjia Bai, Xiaoteng Ma, Dong Wang, Bin Zhao, Zhen Wang, Xuelong Li, and Wei Li. Cross-domain policy adaptation via value-guided data filtering. In *Advances in Neural Information Processing Systems*, volume 36, pp. 73395–73421, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e8ad87f1076fb0f75d89a45828f186b0-Abstract-Conference.html.

Zhenghai Xue, Lang Feng, Jiacheng Xu, Kang Kang, Xiang Wen, Bo An, and Shuicheng Yan. Policy regularization on globally accessible states in cross-dynamics reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70086–70104. PMLR, 2025. URL https://proceedings.mlr.press/v267/xue25b.html.

Mengbei Yan, Jiafei Lyu, Shengjie Sun, Zhongjian Qiao, Jing-Wen Yang, Zichuan Lin, Deheng Ye, and Xiu Li. Cross-domain offline policy adaptation via selective transition correction. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=TupiNRpgHw.

Andrea Zanette, Martin J. Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13626–13640, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/713fd63d76c8a57b16fc433fb4ae718a-Abstract.html.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D. Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2730–2775. PMLR, 2022. URL https://proceedings.mlr.press/v178/zhan22a.html.

Organization of the Appendix

The appendix is arranged by topic. Each appendix contains the full statements and discussion for one part of the paper, followed by the corresponding proofs. Throughout, "the assumptions" refers to realizability and Bellman completeness (Assumptions 1 and 2).

- Appendix A summarizes the guarantees, compares them with the settings and assumptions of related work, and discusses the related literature in detail.
- Appendix B describes the policy closure on which the assumptions are imposed, relates the weak modulus to the simulation lemma and to shifted rewards, establishes basic properties of the moduli, and collects the tools used in all proofs, namely the Bellman-residual telescoping lemma, the per-state Hedge guarantee, the extended value-difference decomposition, and the source concentration results that connect empirical confidence sets to the moduli.
- Appendix C states the offline algorithm and its max-min variant and proves Theorem 1.
- Appendix D treats online learning with target data. It states the source-constrained regret bound (Theorem 21) and its warm-started version, discusses certification and handoff, gives a plug-in estimator of the modulus for a fixed policy, and proves these results together with Theorems 2 and 4 and Corollary 3.
- Appendix E discusses adaptation through observed returns and proves Theorem 5 together with its slope-adaptive form and its composition with block-based learners.
- Appendix F contains the structured bounds for tabular, linear, and block MDPs and their extension to multiple sources, together with the proofs of these results and of the shifted-reward proposition.
- Appendix G states and proves the lower bounds.
- Appendix H describes the experiments. It gives the protocol and diagnostics for the finite-class study of Section 6, model-based tabular illustrations of the individual mechanisms, and a study of existing deep reinforcement learning methods under friction shift.

A Summary of Results and Related Work

A.1 Summary of results

Table 3 situates the framework among the most closely related theoretical settings.

Table 4 summarizes the guarantees of the paper. For the master rows in Tables 4 and 5, define the complete branch bounds

$$\begin{aligned}
\mathsf{S}_T &:= M\Gamma_{\mathcal{W}}T + MH^2\sqrt{T\log|\mathcal{A}|} + H\sqrt{MT\iota} + \sqrt{M}H^3\sqrt{T} + MH^3,\\
\mathsf{O}_T &:= \left(M + \sqrt{M}\,\frac{c_{\mathrm{on}}}{H^2}\right) R_{\mathrm{on}}(T) + H\sqrt{MT\iota} + Mc_{\mathrm{on}}\left(H^2 + \frac{c_{\mathrm{on}}}{H}\right),\\
\mathsf{L}_T &:= c_{\mathrm{on}}\sqrt{T} + \frac{Mc_{\mathrm{on}}^2}{H} + MH^3\sqrt{T} + H\sqrt{MT\iota}.
\end{aligned}$$

The first two quantities give the weak master's bound $\widetilde{O}(\min\{\mathsf{S}_T, \mathsf{O}_T\})$. The last is the additive remainder in the strong-master row at $w = \lceil\sqrt{T}\rceil$; that row retains Proposition 32's condition $H \le c_{\mathrm{on}} \le c_w$. All transfer levels are those of the specified learner sequences, not universal properties of an environment pair.

A.2 Related work

Table 5 places the paper's guarantees next to the most closely related methods for learning from shifted-dynamics data, at the level of guarantee *shapes*; the settings and parameterizations differ, so the rows are not directly comparable and no row dominates another outside its own assumptions.

Transfer learning and error comparison. Hanneke & Kpotufe (2024) introduce weak and strong transfer moduli to relate source and target errors and to analyze adaptive sample aggregation. Their weak modulus is a continuity measure for these errors; several existing divergences upper-bound it.

Table 3: Comparison with the most closely related theoretical settings for learning with data from shifted or matching dynamics. Policy-localized shift asks whether the difficulty measure is localized to the visitation pattern of individual policies rather than attached to the environment pair globally; shift-agnostic asks whether the guarantees need prior knowledge of the size or structure of the shift. Table 5 in Appendix A.2 compares guarantee shapes quantitatively.

Setting	Shift model	Difficulty measure	Policy-loc. shift	Shift agnostic
Hybrid RL, one environment (Song et al., 2023; Tan & Xu, 2024; Li et al., 2023)	none	coverage, concentrability	n/a	n/a
Policy transfer coeff. (Nguyen-Tang & Arora, 2024)	none (coverage)	transfer exponent, factor	n/a	adaptive hybrid, one env.
HySRL (Qu et al., 2025)	separable margin	global TV with margin	no	detects large shift
MIN-UCB-VI (Wu et al., 2026)	arbitrary, tabular	per-pair bias bounds	no	known-bias base; unknown-bias extension
Bellman alignment (Chen et al., 2026b)	abs. continuous	density ratio, reward class	no	estimated ratios; structural assumptions
Robust MDPs (Iyengar, 2005; Panaganti & Kalathil, 2022)	bounded radius	uncertainty-set radius	no	radius known
This work	arbitrary	weak and strong moduli	yes	yes (Thm. 5)

Table 4: Summary of the guarantees. Upper bounds hold with high probability under their stated assumptions; lower bounds apply to the families and target-sample regimes in the cited theorems. The full master costs S_T, O_T, and L_T are defined in the text. The comparator-only strong average and the master level Γ_{str}, which also includes iterate and fallback charges, are different quantities. Warm-starting uses a known $D \geq \bar{K}$, and $\widetilde{O}$ suppresses logarithmic factors.

Guarantee	Algorithm	Bound (informal)	Where
Offline suboptimality	per-state Hedge	$\frac{1}{T}\sum_{t,h}\big[\sqrt{\delta_h^{\pi^\star\mid\pi^{(t)}}}+\sqrt{\delta_h^{\pi^{(t)}}}\big]+\widetilde{O}\big(\frac{H^2}{\sqrt{T}}\big)$	Thm. 1
Offline suboptimality	max-min	$\sum_h\big[\sqrt{\delta_h^{\pi^\star}}+\sqrt{\delta_h^{\widehat{\pi}}}\big]$	Prop. 18
Online regret	Alg. 3	$T\Gamma_T+\widetilde{O}(H^2\sqrt{T})$	Thm. 21
Online regret, warm start	Alg. 3	$T\Gamma_T+\sqrt{DH^3T/2}$	Thm. 22
Online regret, strong	Alg. 1	$T\widetilde{\Gamma}_T+\widetilde{O}(H^2T^{5/6})$ if $N_{\mathrm{fb}}=0$	Thm. 2
Online regret, fast rate	Alg. 1, $w\asymp\sqrt{T}$	$T\widetilde{\Gamma}_T^{\mathrm{fast}}+\widetilde{O}(H^2T^{3/4})+2H\sqrt{T}N_{\mathrm{fb}}$	Cor. 3
Adaptive regret	balancing master	$\widetilde{O}(\min\{\mathsf{S}_T,\mathsf{O}_T\})$	Thm. 5
Certified handoff	Alg. 1 + exit	certified prefix $+wH+R_{\mathrm{on}}(T)$; no M	Thm. 4
Strong master, native rates	balancing master	$\widetilde{O}\big(\min\{M(T\Gamma_{\mathrm{str}}+H^2T^{3/4}),Mc_{\mathrm{on}}\sqrt{T}\}+\mathsf{L}_T\big)$	Prop. 32
Modulus, structured	tabular, linear, block	source uncertainty + target-weighted shift	Thm. 6
Modulus, multi-source	intersected fitting sets	each (s,a) charged its best source	Prop. 33
Modulus estimation	plug-in over $\widehat{\mathcal{F}}_{\mathrm{src}}$	two-sided sandwich at perturbed radii	Prop. 24
Lower bounds	any algorithm	$\mathbb{E}[\mathrm{SubOpt}]\geq c\sum_h\sqrt{\delta_h^{\pi^\star}(0)}$ and regret analogues	Thm. 7

Their strong modulus captures additional gains from combining source and target evidence. Earlier relatedness measures include domain-classification divergences (Ben-David et al., 2010), discrepancy distances (Mansour et al., 2009), transfer exponents (Kpotufe & Martinet, 2018), and analyses of the value of target samples (Hanneke & Kpotufe, 2019).

Transfer-based coverage measures already appear in sequential decision-making. Transfer exponents govern covariate-shift bandits (Suk & Kpotufe, 2021; Cai et al., 2024). In offline RL, data diversity and policy transfer coefficients relate behavior-policy information to target-policy performance (Nguyen-Tang & Arora, 2023; 2024). The latter work also studies combining offline and online learning under

Table 5: Guarantee shapes for learning with shifted-dynamics source data. Entries summarize the form of each guarantee in its own parameterization; rates across rows are not directly comparable, since the settings, assumptions, and difficulty measures differ. "None" in the prior-knowledge column means the stated guarantee needs no input describing the size or structure of the shift.

Method	Shift knowledge	Guarantee shape
Naive pooling	none	biased by shift; suboptimality $\Theta(H)$ possible at every source size n, linear regret under adversarial shift (Appendix H.2)
Pure online (UCBVI, GOLF)	none (no source use)	$R_{\mathrm{on}}(T) = \widetilde{O}(H^2\sqrt{SAT})$ tabular, GOLF-style under function approximation
HySRL (Qu et al., 2025)	separability margin	tabular sample-complexity gains when the shift is separable at a known margin; general shifted data provably useless without shift information
MIN-UCB-VI (Wu et al., 2026)	per-pair bias bounds	tabular; per-pair minimum of an online bonus and a bias-corrected hybrid bonus; instance-dependent and minimax bounds with matching lower bounds in that parameterization
Bellman alignment (Chen et al., 2026b)	density-ratio class	online regret after one-step alignment under absolute continuity and realizable ratio and reward classes
This work, Thms. 1 and 21	none	$\mathrm{SubOpt} \le A_{\tilde{\iota}} + \widetilde{O}(H^2/\sqrt{T})$ and $\mathrm{Regret} \le T\Gamma_T + \widetilde{O}(H^2\sqrt{T})$ at the modulus level; lower bounds establish necessity on specified hard families and target-sample regimes
This work, Thm. 2, Cor. 3	none	iterate cost capped by the target radius; $T\widetilde{\Gamma}_T^{\mathrm{fast}} + \widetilde{O}(H^2T^{3/4}) + 2H\sqrt{T}\,N_{\mathrm{fb}}$
This work, Thm. 5, Cor. 28	none	$\widetilde{O}(\min\{\mathsf{S}_T, \mathsf{O}_T\})$ with the full baseline-dependent costs defined before Table 4
This work, Thm. 4	none	certified prefix cost $+ wH + R_{\mathrm{on}}(T)$ on certificate firing, no M factor

function approximation. We build on this use of error-comparison quantities; we do not claim the general idea of a transfer modulus as a contribution.

Here the compared quantities are squared Bellman residuals under two different operators. The backup policy and the occupancy on which the residual is evaluated have separate roles. When the environments coincide, the comparison reduces to a coverage question. When their operators differ, even a source-consistent critic can have nonzero target residual. The policy-optimization proof must account for residuals along both the deployed policy and the comparator, while on-policy target data directly constrains only the former. The certification and adaptation results address these algorithmic consequences. Their lower bounds apply to the stated hard families and do not supply a complete joint characterization for every source-target pair.

Hybrid RL with Identical Environments. Song et al. (2023) introduced the hybrid RL framework and showed that combining offline and online data in the same environment yields provable gains over both pure offline and pure online learning for low-bilinear-rank classes, without explicit exploration bonuses. Xie et al. (2021b) studied the closely related policy-finetuning setting, quantifying the value of a reference policy for warm-starting online learning. Tan & Xu (2024) showed that a natural warm start, placing the offline data in the replay buffer of an optimistic online algorithm (DISC-GOLF), achieves provable hybrid gains even without single-policy concentrability. Wagenmaker & Pacchiano (2023) analyzed hybrid RL in linear MDPs without concentrability assumptions, Li et al. (2023) established statistical benefits of reward-agnostic hybrid learning in the tabular case, Tan et al. (2024) obtained the sharpest known hybrid guarantees in linear MDPs, and Huang et al. (2025) gave a unified design achieving state-of-the-art guarantees for both suboptimality and regret. All of these assume the offline and online environments coincide; our work addresses the additional obstruction created by dynamics shift. On the empirical side, offline-to-online methods such as RLPD (Ball et al., 2023) and Cal-QL (Nakamoto et al., 2023) demonstrate the practical value of hybrid data, again under matched dynamics.

Transfer RL with Dynamics Shift. Qu et al. (2025) is the closest prior work in setting, alongside the structured transfer line discussed below. They study hybrid RL where offline data comes from shifted dynamics, under a separability condition in which transitions either match exactly or differ by at least a known margin in total variation, enabling a split of the state-action space into trusted and untrusted regions. Our weak modulus removes this dichotomy. Shift is measured continuously, weighted by target visitation, and attached to individual policies, so partial and subtle shifts are handled by the same guarantee.

In related work, Wu et al. (2026) also drop the separability dichotomy in tabular MDPs, proposing optimism- and pessimism-indexed value iteration (MIN-UCB-VI and MAX-LCB-VI) for regret minimization and best-policy identification under general transition shifts, with instance-dependent and minimax guarantees; their assumptions and tabular guarantees differ from the general-class analysis here. Their main algorithms consume fine-grained per-pair bias bounds $\nu(s,a) \geq \|P_{\mathrm{tar}}(\cdot|s,a) - P_{\mathrm{src}}(\cdot|s,a)\|_1$ as a known input, taking the min of an online and a bias-corrected hybrid confidence bound at each pair, in the lineage of Cheung & Lyu (2024) for bandits with biased offline data, where some such side information is provably necessary by the hardness results of Cheung & Lyu (2024) and Qu et al. (2025). For unknown bias they add a model-selection extension, per-pair EXP3.P experts over a peeled grid of bias estimates analyzed through the black-box reduction of Pacchiano et al. (2020), at a degraded meta-rate. The transfer modulus plays the corresponding role in our framework, but it is defined from the function class rather than assumed, its fixed-policy on-policy version admits a statistical sandwich given a constrained maximization oracle (Proposition 24); this does not provide an efficient estimator of comparator-dependent moduli, and it is defined for arbitrary function classes, with end-to-end guarantees for finite classes and, through the covering extension of Proposition 17, for the linear and block instantiations of Section 5.1, with per-pair bias aggregation recovered as the tabular special case (Theorem 6); our adaptivity operates at learner level via regret balancing with elimination, retaining the base T-rate up to a factor logarithmic in the horizon and budget and polynomial in the candidates' coefficients (Theorem 5). Where they unify regret with best-policy identification, we cover offline suboptimality, online regret, provable worst-case gains from target data (Theorem 2), adaptivity, and estimation of the complexity measure itself; conversely, their tabular instance-dependent rates are sharper than what our general bounds imply in the tabular special case.

Chen et al. (2026b) take a complementary, correction-based route for transfer under task shift. They align source Bellman backups with the target via density-ratio re-weighting and act optimistically with respect to the aligned backups, which requires absolute continuity of the target kernel with respect to the source, estimable density ratios, and a structured (lower-complexity) reward-difference class; in exchange they obtain online regret scaling with the complexity of the task shift.

Offline-to-Online Value Adaptation. Li & Zhang (2026) study the closely related problem of adapting an imperfect offline-pretrained Q-function to a target environment under general function approximation. They prove a minimax lower bound showing that adaptation cannot beat pure online RL on hard instances even when the pretrained function is close to optimal, echoing in the value-adaptation setting the hardness results that motivate side information in ours, and they give O2O-LSVI, whose problem-dependent gains hold under a structural condition on the pretrained values. The adapted object differs, a single pretrained value function there versus a source-constrained version space here, and their structural condition plays the role our modulus plays, a measure of when prior value information transfers; a precise translation between the two conditions is an interesting open question. Chen (2026) formalizes the residual uncertainty that remains after conditioning on offline data and directs exploration by information gain, a natural route toward actively collecting target data that shrink the strong modulus. Our analysis instead assumes realizability and Bellman completeness in both environments and pays policy-dependent residual terms. The lower bounds establish worst-case necessity on the stated hard families and target-sample regimes. Density-ratio realizability has also been exploited for exploration in a single environment (Amortila et al., 2024). Empirically, Eysenbach et al. (2021) train with learned domain classifiers to correct for off-dynamics data (DARC), Niu et al. (2022) adaptively decide when to trust a simulator (H2O), and Niu et al. (2025) extend this to H2O+, validated on physical robots, the closest practical precedent for the deployment setting we study. A rich deep off-dynamics line has followed. Liu et al. (2022) reweight rewards by dynamics ratios (DARA), Lyu et al. (2024a) penalize source transitions by representation mismatch (PAR), Xue et al. (2025) regularize policies toward globally accessible states (ASOR), Kong et al. (2025) estimate the dynamics gap with optimal-transport flow models and actively

collect target data in high-gap regions (CompFlow), an empirical counterpart of the region-level adaptivity we pose as an open problem, Xu et al. (2023) selectively share source transitions whose paired value targets agree across domains (VGDF), and Liu & Long (2026) reweight source data by their alignment with target Bellman backups (TABB), a selection criterion conceptually close to the modulus, which measures exactly the worst-case failure of such alignment over a version space; see also optimal-transport dataset constraints (Lyu et al., 2025), dynamics- and value-aligned filtering (Qiao et al., 2026a), selective transition correction (Yan et al., 2026), fusion of scarce target data with distribution-filtered source fragments (Wang et al., 2026), model-based off-dynamics learning (Guo et al., 2026), and dual-robust formulations (Qiao et al., 2026b). Particularly relevant to our region-level open problem, Xia et al. (2026) exploit *localized* dynamics mismatch by clustering transitions and filtering source data at cluster granularity (LoDADA), an empirical analogue of localized source selection in an offline cross-domain setting; the online, within-policy, region-level regret problem that the block-MDP bound makes precise remains open. Closest in spirit among these, Chen et al. (2026c) combine source and target critics with an adaptive weight through cross-domain Bellman consistency (QAvatar); the delineation from our framework is the same as for TABB, since these methods use Bellman consistency as an empirical reweighting or combination signal for particular critics, whereas the transfer modulus is the *worst-case* source version-space quantity, carries the two-policy separation of backup and occupancy, and comes with finite-horizon upper and lower bounds. These methods, together with RLPD and Cal-QL, are the natural baseline families for the ODRL protocol discussed in Appendix H.2. Jiang (2018) gives PAC guarantees for RL with an imperfect model under a different (model-error) parameterization. Earlier surveys and multi-task analyses include Taylor & Stone (2009) and Brunskill & Li (2013). A conceptually adjacent line is distributionally robust RL (Iyengar, 2005; Nilim & El Ghaoui, 2005; Panaganti & Kalathil, 2022; Shi et al., 2023), which optimizes worst-case performance over an uncertainty set of dynamics centered at the source; robust MDPs deliberately hedge against all shifts of a given radius, whereas the weak modulus is instance-dependent and policy-dependent, penalizing only the shift that the evaluated policy actually encounters. Distributionally robust formulations have also been developed for off-dynamics learning specifically, with provably efficient linear-function-approximation algorithms (Liu & Xu, 2024) and, closest to our concerns, a sample-complexity theory of online interaction with a perturbed training environment whose central quantity, a supremal visitation ratio between training and deployment dynamics, plays a role analogous to our target-weighted shift terms and comes with matching lower bounds (He et al., 2025). Between the two poles, Awad et al. (2026) shrink robust uncertainty sets around constrained kernel estimates using *side information* about the source-target dynamics (moment bounds, distributional distances, density ratios), with finite-sample guarantees on the robust suboptimality gap; this refines the robust pole with the same kind of side information that the bias-bound line takes as input, whereas our certificates are estimated from the data themselves. The robust objective remains worst-case over an uncertainty set, whereas the modulus is anchored at the realized source-target pair. In the structured direction, Sam et al. (2024) give matching upper and lower bounds for representational transfer under latent low-rank structure, organized around a transferability coefficient; that program is the representation-transfer analogue of the modulus-level matching we pursue for dynamics shift. Chai et al. (2025) study online source-to-target transfer when high-dimensional transitions decompose into a shared low-rank component plus a sparse task-specific component, an important structured comparator whose difficulty measure is the sparse discrepancy rather than a version-space quantity, and Chen et al. (2026a) analyze batch transfer of $Q^\star$ under general function approximation with an explicit task-discrepancy measure; both differ from the policy-dependent Bellman modulus, which ties transfer difficulty to the evaluated policy's target occupancy rather than to a global structural gap.

Offline RL with Coverage. The offline RL literature has developed various coverage measures for single-environment settings. Munos & Szepesvári (2008) and Chen & Jiang (2019) formalized all-policy concentrability, measuring how much any policy's distribution can deviate from offline data; Antos et al. (2008) identified the double-sampling issue central to estimating Bellman residuals, which our confidence sets inherit (Appendix B.4). Rashidinejad et al. (2021) proposed single-policy concentrability, a weaker notion sufficient for competing with a fixed policy. Xie et al. (2023) introduced coverability as an intrinsic MDP property. Zhan et al. (2022) achieved guarantees under realizability and single-policy concentrability, and Uehara & Sun (2022) under partial coverage with model-based pessimism. Our weak modulus extends these to cross-environment transfer. When source equals target, $\delta_h^\pi(\epsilon) \le C_{\mathrm{conc}}(\mu, \pi)\,\epsilon$, recovering standard coverage-controlled bounds (Proposition 9).

Two additional offline approaches address related sources of mismatch. Wen et al. (2024) learn contrastive representations to filter source data, while BOSA (Liu et al., 2024a) constrains state-action and transition support. Both concern learning from fixed cross-domain datasets; our online guarantees additionally account for adaptively collected target observations. For distributionally robust off-dynamics RL, Liu et al. (2024b) establish upper and lower bounds under linear structure and use a rarely switching algorithm. Their objective is robust performance over a transition uncertainty set, rather than performance in one realized target MDP characterized through its Bellman residuals.

Pessimistic Policy Optimization. Our algorithms build on pessimism for offline RL: Bellman-consistent pessimism with version spaces (Xie et al., 2021a), pessimistic value iteration (Jin et al., 2021), pessimistic actor-critic methods (Zanette et al., 2021), conservative value estimation (Kumar et al., 2020), and version-space analyses under general data-diversity conditions (Nguyen-Tang & Arora, 2023). The max-min variant of Proposition 18 is precisely Bellman-consistent pessimism built from source data, and our per-state Hedge analysis follows the extended value-difference technique of optimistic policy optimization (Cai et al., 2020; Shani et al., 2020; Agarwal et al., 2020a). The key new difficulty is that pessimism certified in the source need not remain pessimism in the target under dynamics shift; the weak modulus quantifies exactly this failure.

Model Selection over Learners. Our adaptive guarantee (Theorem 5) uses online model selection. Corral (Agarwal et al., 2017) and regret balancing (Pacchiano et al., 2020; Cutkosky et al., 2021) combine base learners with candidate regret bounds and compete with the best well-specified one. We use regret balancing because its conditions, bounded observable episodic returns and anytime base guarantees over own play, are exactly satisfied in our setting.

Multi-Fidelity Optimization and Adaptive Control. A related trade-off between cheap approximate information and expensive accurate information appears in multi-fidelity Bayesian optimization, where low-fidelity evaluations prune the search space and high-fidelity queries refine it (Kandasamy et al., 2016), and in multi-fidelity RL, where simulators of increasing fidelity gate access to real-world interaction (Cutler et al., 2015). These frameworks quantify fidelity through global bias bounds on a fixed objective, whereas our modulus is policy dependent and Bellman-operator based, which is the structure sequential decision-making adds; a fidelity parameter plays the role of a known, uniform shift bound, roughly a robust-MDP radius, while the modulus is instance dependent and certified from data. Adaptive control likewise studies acting under imperfectly known, possibly drifting dynamics (Åström & Wittenmark, 1995), with a focus on stability and asymptotic tracking rather than finite-sample regret against an optimal policy; our fallback certificate can be read as a statistical analogue of the monitoring signals used there.

Structured MDPs. Our explicit bounds leverage structure in MDPs. Jin et al. (2020) introduced linear MDPs with feature-based transitions, Agarwal et al. (2020b) and Modi et al. (2024) studied representation learning in low-rank MDPs, and Du et al. (2019) analyzed block MDPs. We show how the weak modulus simplifies under such structure, particularly for block MDPs where it decomposes over regions with and without shift.

Information-Theoretic Lower Bounds. Our lower bounds use classical techniques from statistical decision theory, namely Le Cam two-point testing, Fano-type multiple hypothesis arguments, and direct sums over independent decision layers; Lattimore & Szepesvári (2020) provide a comprehensive treatment. Azar et al. (2017) derive minimax bounds for tabular online RL, and Wang et al. (2021) establish limits for offline RL with linear function approximation. We extend these to the hybrid transfer setting, showing that the weak-modulus dependence of our upper bounds is information-theoretically necessary.

B Setting, Transfer Moduli, and Analysis Tools

B.1 Assumptions, the simulation lemma, and shifted rewards

The algorithmic policy closure. Assumptions 1 and 2 are quantified over a single *algorithmic policy closure* $\overline{\Pi}$ rather than over all stochastic policies. Fix, before any data are observed, a finite base class $\mathcal{Q}_{\mathrm{base}}$ of full-support policies (the uniform policy and any fixed warm starts under consideration), let $\mathcal{C}$ be the finite collection of learning-rate values used by the algorithms, offline and online, $|\mathcal{C}| = O(T)$, and let Π_{fixed} collect the behavior policy μ and any fixed policies submitted for

evaluation. The closure is defined *operationally*, and this is its sole definition. With the exponential-update operators $U_{c,f}(\pi)_h(a|s) \propto \pi_h(a|s)e^{cf_h(s,a)}$ for $f \in \mathcal{F}$, $c \in \mathcal{C}$, the mixing operator $M_\tau(\pi) = (1-\tau)\pi + \tau\,\mathrm{Unif}(\mathcal{A})$ at $\tau = 1/T$, and $L := T + T_{\mathrm{off}}$, where $T_{\mathrm{off}} \le T$ is the number of offline iterations,

$$\overline{\Pi} := \big\{O_r \circ \cdots \circ O_1(q) :\ q \in \mathcal{Q}_{\mathrm{base}},\ O_j \in \{U_{c,f}\} \cup \{M_{1/T}\},\ \#\{j : O_j \text{ an update}\} \le L,\ \#\{j : O_j = M_{1/T}\} \le 2\big\} \ \cup\ \Pi_{\mathrm{fixed}}.$$

The subfamily with no mixing has the explicit softmax form $\pi_h(a|s) \propto q_h(a|s)\exp\big(\sum_{\ell=1}^m c_\ell f_h^{(\ell)}(s,a)\big)$ with $m \le L$, which is how the algorithms' iterates arise before any mixing; a mixed policy followed by further exponential updates is generally *not* of that form, which is why the operation sequence, including the mixing position, is the definition rather than a description. We write $\overline{\Pi}_{\mathrm{init}} \subseteq \overline{\Pi}$ for the *initialization slice*, the operational members built from at most T_{off} updates and at most one mixing; warm starts are drawn from $\overline{\Pi}_{\mathrm{init}}$, so that their T online updates keep the total within the budget $L = T_{\mathrm{off}} + T$. In particular, *every* data-dependent member of the slice, whether or not its history already contains a mixing, may be mixed again after it is selected, since the closure permits two mixings, and all of its subsequent online updates remain inside the class; a member that has exhausted the full update budget L could not be updated further, which is why initialization is restricted to the slice. A predetermined candidate whose mixed version is wanted as a warm start should have that mixed version inserted into $\mathcal{Q}_{\mathrm{base}}$, making it a length-zero member of the slice; members of Π_{fixed} serve as evaluation comparators only and are not update bases. More generally, a warm start with an a-priori operational description of length m_0 is handled by running the closure at budget $m_0 + T$ in place of L, which replaces $T + T_{\mathrm{off}}$ by $m_0 + T$ in the entropy and hence in the source radii. This one class contains every offline iterate, the data-dependent warm start (the data select a member of the predetermined class, they do not determine the class), and all online descendants under any of the schedules used in the paper, fixed or anytime, with separate offline and online rates. For finite $\mathcal{F}$,

$$|\overline{\Pi}| \ \le\ |\Pi_{\mathrm{fixed}}| + |\mathcal{Q}_{\mathrm{base}}|\,(L+2)^2\big(1 + |\mathcal{C}||\mathcal{F}|\big)^L,$$
$$\log|\overline{\Pi}| \ \lesssim\ (T + T_{\mathrm{off}})\log\big(1 + |\mathcal{C}|\,|\mathcal{F}|\big) + \log|\mathcal{Q}_{\mathrm{base}}| + \log(1 + |\Pi_{\mathrm{fixed}}|),$$

which is the cardinality entering our union bounds; for infinite $\mathcal{F}$ the same expression holds with $|\mathcal{F}|$ replaced by the covering number of Proposition 17, whose projective-metric argument covers mixing at any position at no extra cost. Throughout, $\mathcal{Q}_{\mathrm{base}}$ and Π_{fixed} are short fixed lists (the uniform policy, a bounded number of designated warm starts, the behavior policy, and the policies under evaluation), and we adopt once the convention $\log|\mathcal{Q}_{\mathrm{base}}| + \log(1 + |\Pi_{\mathrm{fixed}}|) \le T\log(T|\mathcal{F}|)$, under which the complexity $T\log(T|\mathcal{F}|H/\delta)$ displayed in the theorems absorbs both; a reader tracking a large base class should substitute the general term $\mathcal{C}_{\mathrm{alg}}$ of Appendix B.4.

Remark 1 (Scope of the assumptions)**.** *Quantifying over all stochastic policies would be incompatible with a finite class, since the value functions $\{Q^\pi\}$ of a nondegenerate MDP form a continuum, so a finite $\mathcal{F}$ cannot contain them all; restricting to $\overline{\Pi}$ is what every proof actually uses. The comparator $\pi^\star_{\mathrm{tar}}$ needs neither assumption in Theorems 1, 21, and 2, since there it enters the analysis only through visitation distributions, never through membership of $Q^{\pi^\star}$ or of its backups in $\mathcal{F}$; the exception is the max-min variant (Proposition 18), whose confidence sets are indexed by the comparator class Π and which therefore requires both assumptions on Π. For structured infinite classes, minima and suprema over $\mathcal{F}$ are interpreted through measurable ε-minimizers, or are attained under the usual compactness of bounded linear and block-constant classes. For infinite classes, Proposition 17 in Appendix B.4 replaces the cardinalities in our radii by sup-norm covering numbers, including the induced cover of the deployed policy class, which gives end-to-end algorithmic guarantees for the linear and block instantiations of Section 5.1. Approximate versions of both assumptions, holding up to a misspecification level ϵ_{app}, enlarge the confidence radii by ϵ_{app} and add $O(\sum_h \sqrt{\epsilon_{\mathrm{app}}})$ to the value guarantees; this extension we state informally and do not develop.*

Throughout, rewards lie in $[0,1]$, functions in $\mathcal{F}_h$ take values in $[0, H-h+1]$ with the convention $f_{H+1} \equiv 0$, so that Bellman backups respect the class ranges and the completeness assumption is internally consistent; the analysis uses only the uniform bound H on all ranges. Statements quoted with the uniform range $[0,H]$, such as the structured bounds, hold verbatim, since they do not invoke completeness of the enlarged class; wherever completeness is invoked, including in the lower-bound constructions, the stepwise convention applies. We use $\|\cdot\|_\infty$ to denote the supremum norm and $\|\cdot\|_{\mathrm{TV}}$ for total variation distance. Constants denoted by C may change from line to line.

Relation to the simulation lemma and shifted rewards.

Remark 2 (Relation to the simulation lemma)**.** *The qualitative principle that model discrepancies matter only where a policy goes is classical; the simulation lemma and the performance-difference lemma bound* $|V^{\pi}_{\mathcal{M}_{\mathrm{src}}} - V^{\pi}_{\mathcal{M}_{\mathrm{tar}}}|$ *by visitation-weighted discrepancies between the* true *kernels. The weak transfer modulus makes a different statement. It measures the ambiguity left by* finite source data within a function class*, coupling statistical coverage with dynamics shift, and it is exactly this data-and-class-dependent quantity, rather than the true-model discrepancy, that the lower bounds of Section 5.2 certify as unavoidable. The two can differ in both directions. With scarce source data the modulus is large even with no shift at all (the coverage regime of Proposition 9), while structure in* $\mathcal{F}$ *can make the modulus far smaller than the simulation-lemma bound, as in block-structured classes where source fitting pins entire regions at once (Theorem 37). The simulation lemma reappears inside our proofs only as the pointwise shift estimate in the structured bounds.*

We assume shared rewards in the main development to isolate dynamics shift, which is also what our lower-bound constructions require. The framework extends directly when the target rewards differ, and the extension is quantitative.

Proposition 8 (Shifted rewards)**.** *Suppose the target rewards differ from the source rewards, and let* $\Delta r_h(s,a) := |r_h^{\mathrm{src}}(s,a) - r_h^{\mathrm{tar}}(s,a)|$*. Definition 1 is unchanged, with the source-fitting set defined through the source Bellman operator (source rewards) and the evaluation through the target Bellman operator (target rewards). In tabular MDPs, for all policies* ν, π*, steps* h*, and* $\epsilon \geq 0$*,*

$$\delta_h^{\nu|\pi}(\epsilon) \leq 3 \sum_{s,a} d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h}(s,a) \big[\, \Delta r_h(s,a)^2 + H^2 \Delta_h(s,a)^2 + \mathrm{Stat}_h(s,a;\epsilon) \,\big],$$

with Δ_h *and* Stat_h *as in Theorem 6. In linear MDPs with per-environment reward parameters and* $\Delta\theta_h := \theta_h^{\mathrm{src}} - \theta_h^{\mathrm{tar}}$*,*

$$\delta_h^{\nu|\pi}(\epsilon) \leq 4\epsilon\, \mathbb{E}_{d^{\nu}} \|\varphi\|^2_{(\Lambda_h^{\mu})^{\dagger}} + 4B_h^2\, \mathbb{E}_{d^{\nu}} \|P_{h,\perp}\varphi\|_2^2 + 4\, \mathbb{E}_{d^{\nu}} \langle \varphi, \Delta\theta_h \rangle^2 + 4H^2 d (\Delta_h^{\mathrm{lin}})^2,$$

and in block MDPs with block-homogeneous rewards in each environment and per-block reward gaps $\Delta r_{h,i}$*,*

$$\delta_h^{\nu|\pi}(\epsilon) \leq 3 \sum_i p^{\nu}_{h,i} \big[\Delta r_{h,i}^2 + H^2 (\Delta_{h,i}^{\mathrm{blk}})^2 + \mathrm{Stat}_{h,i}(\epsilon) \big].$$

Realizability and Bellman completeness are preserved under reward shift, since backups remain linear, respectively block-constant, with each environment's own reward. All algorithmic results are unaffected, because the algorithms only ever fit the rewards they observe.

The proof, a one-line extension of the tabular argument, is in Appendix F.2. A reward-shift experiment in Appendix H.2 illustrates this mechanism; it is not a numerical verification of the bound.

The modulus combines two effects. Even if the source and target dynamics coincide, Bellman residuals under $d^{\mu}_{\mathcal{M}_{\mathrm{src}},h}$ may fail to control residuals under $d^{\pi}_{\mathcal{M}_{\mathrm{src}},h}$ without coverage. When the dynamics differ, there is an additional obstruction because the Bellman operator changes from P_{src} to P_{tar}. The policy dependence of $\delta_h^{\pi}(\epsilon)$ is therefore essential, as illustrated in Figure 1. A policy that avoids shifted regions may have a small modulus even when the two MDPs differ substantially elsewhere, whereas a policy that visits shifted regions may have a large modulus. Basic properties of the modulus (e.g., monotonicity and reduction to standard concentrability when $\mathcal{M}_{\mathrm{src}} = \mathcal{M}_{\mathrm{tar}}$), are collected in Appendix B.2. In that matched case the modulus is a second-moment relative of the data-diversity coefficient of Nguyen-Tang & Arora (2023) and of the policy transfer coefficients of Nguyen-Tang & Arora (2024), which adapted the transfer exponent of Hanneke & Kpotufe (2019) to offline decision-making as a coverage notion; the dynamics shift enters only through the change of Bellman operator between the fitting and the evaluation.

B.2 Properties of the weak transfer modulus

The following elementary properties clarify the role of the modulus.

Proposition 9 (Basic properties of the weak transfer modulus)**.** *For any policy* π*, step* $h \in [H]$*, and tolerances* $\epsilon_1 \leq \epsilon_2$*, the following hold.*

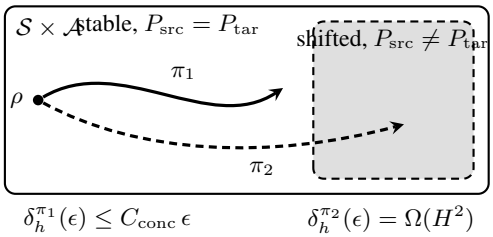


Figure 1: Policy dependence of the weak transfer modulus. Source and target disagree only on the shaded region. The trajectories of π_1 avoid it, so its modulus reduces to a coverage term, while those of π_2 cross it, where source-consistent critics can carry target Bellman error of order H^2 (Proposition 9).

1. ***Monotonicity.***
$$\delta_h^{\pi}(\epsilon_1) \leq \delta_h^{\pi}(\epsilon_2). \tag{13}$$

2. ***Policy dependence.*** *There exist pairs of source and target MDPs for which $\delta_h^{\pi_1}(\epsilon)$ is small while $\delta_h^{\pi_2}(\epsilon)$ is large for two policies π_1 and π_2.*

3. ***Reduction to coverage in the single-environment case.*** *If $\mathcal{M}_{\mathrm{src}} = \mathcal{M}_{\mathrm{tar}} = \mathcal{M}$, then the weak transfer modulus reduces to an off-policy coverage quantity. In particular, if for all measurable nonnegative residual functions g_h induced by $\mathcal{F}$,*
$$\mathbb{E}_{(s,a)\sim d^{\pi}_{\mathcal{M},h}}[g_h(s,a)] \leq C_{\mathrm{conc}}(\mu,\pi)\mathbb{E}_{(s,a)\sim d^{\mu}_{\mathcal{M},h}}[g_h(s,a)], \tag{14}$$
then
$$\delta_h^{\pi}(\epsilon) \leq C_{\mathrm{conc}}(\mu,\pi)\epsilon. \tag{15}$$

Proof. **(1)** is immediate from the nesting $\mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon_1) \subseteq \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon_2)$. **(2)** Consider a two-block instance, $\mathcal{S} \times \mathcal{A} = B_1 \cup B_2$, where $P_{\mathrm{src},h} = P_{\mathrm{tar},h}$ on B_1, while on B_2 the source transition is the uniform mixture of two continuation states g, b whose downstream values differ by $\Theta(H)$, and the target transition places all mass on b; thus $\Delta_h(s,a) = 1/2$ on B_2. Let μ cover both blocks. A policy π_1 that never visits B_2 under either dynamics satisfies $d^{\pi_1}_{\mathcal{M}_{\mathrm{tar}},h} = d^{\pi_1}_{\mathcal{M}_{\mathrm{src}},h}$ (the two kernels agree on every region π_1 reaches), and on this common support the two Bellman operators coincide, so the argument of part (3) applies verbatim and $\delta_h^{\pi_1}(\epsilon) \leq C_{\mathrm{conc}}(\mu,\pi_1)\,\epsilon$. For a policy π_2 that visits B_2 with probability one, take the witness $f = Q^{\pi_2}_{\mathcal{M}_{\mathrm{src}}} \in \mathcal{F}^{\mu,\pi_2}_{\mathrm{src}}(0)$, whose target Bellman residual on B_2 equals the backup gap $\mathbb{E}_{P_{\mathrm{src}}}[f_{h+1}] - \mathbb{E}_{P_{\mathrm{tar}}}[f_{h+1}] = \Theta(H)$, whence $\delta_h^{\pi_2}(\epsilon) \geq \delta_h^{\pi_2}(0) = \Omega(H^2)$ for every $\epsilon \geq 0$. **(3)** is proved as Corollary 35 below; the argument applies verbatim to general measurable residual classes. □

The third property shows that the weak modulus recovers the usual concentrability obstruction from offline RL when there is no source-target shift. Thus, the modulus is not replacing coverage assumptions with a new abstraction; rather, it extends the familiar coverage perspective to the hybrid setting. The second property explains why a worst-case global transfer coefficient can be overly pessimistic. Hybrid RL can be useful because the target policy may only need value information along a restricted subset of trajectories. The weak transfer modulus captures this possibility by allowing transfer difficulty to depend on the policy whose performance is being certified.

This perspective also clarifies the best- and worst-case regimes. In the optimistic regime, source and target dynamics may differ globally, but the functions that fit the source Bellman equations remain approximately Bellman-consistent on the target trajectories induced by the learned policy; then $\delta_h^{\pi}(\epsilon)$ is small, and source data can substantially reduce the number of target interactions needed. In the pessimistic regime, source fitting is uninformative for the target, since there exist functions in $\mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon)$ that agree with the source data yet have large target Bellman error. In this case, no hybrid method can safely rely on the source data without additional target correction, and one should expect guarantees to degrade toward those of purely online learning in the target environment.

B.2.1 A CONSEQUENCE OF THE WEAK MODULUS

The weak transfer modulus was defined as

$$\delta_h^\pi(\epsilon) = \sup_{f \in \mathcal{F}_{\mathrm{src}}^{\mu,\pi}(\epsilon)} \mathcal{E}_{\mathcal{M}_{\mathrm{tar}}}^{\pi,\pi}(f)_h,$$

where

$$\mathcal{F}_{\mathrm{src}}^{\mu,\pi}(\epsilon) = \left\{ f \in \mathcal{F} : \mathcal{E}_{\mathcal{M}_{\mathrm{src}}}^{\mu,\pi}(f)_{h'} \le \epsilon \text{ for all } h' \in [H] \right\}.$$

Lemma 10 (From source validity to target residuals)**.** *If $f \in \mathcal{F}_{\mathrm{src}}^{\mu,\pi}(\epsilon)$, then for every $h \in [H]$,*

$$\mathcal{E}_{\mathcal{M}_{\mathrm{tar}}}^{\pi,\pi}(f)_h \le \delta_h^\pi(\epsilon).$$

Consequently,

$$\left| f_1(\rho,\pi_1) - V_{\mathcal{M}_{\mathrm{tar}},1}^\pi(\rho) \right| \le \sum_{h=1}^H \sqrt{\delta_h^\pi(\epsilon)}.$$

Proof. The first inequality is exactly the definition of the supremum in $\delta_h^\pi(\epsilon)$. The second inequality follows by applying Lemma 11 in the target MDP and then using the first inequality at each step. □

B.3 ANALYSIS TOOLS

We begin with several identities used repeatedly in the proofs. Recall that for policies ν, π, MDP $\mathcal{M}$, and function $f \in \mathcal{F}$, the off-policy Bellman error is

$$\mathcal{E}_{\mathcal{M}}^{\nu,\pi}(f)_h = \mathbb{E}_{(s,a)\sim d_{\mathcal{M},h}^\nu} \left[\left(f_h(s,a) - \mathcal{T}_{\mathcal{M},h}^\pi f_{h+1}(s,a) \right)^2 \right],$$

where

$$(\mathcal{T}_{\mathcal{M},h}^\pi f_{h+1})(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_{\mathcal{M},h}(\cdot|s,a)} \left[f_{h+1}(s', \pi_{h+1}) \right].$$

We use the shorthand

$$f_h(s,\pi_h) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}[f_h(s,a)], \qquad f_1(\rho,\pi_1) = \mathbb{E}_{s\sim\rho,a\sim\pi_1(\cdot|s)}[f_1(s,a)].$$

B.3.1 BELLMAN-RESIDUAL TELESCOPING

The first lemma converts Bellman residual control into value error.

Lemma 11 (Bellman residual value bound)**.** *For any MDP $\mathcal{M}$, policy π, and function $f \in \mathcal{F}$ with $f_{H+1} \equiv 0$,*

$$\left| f_1(\rho,\pi_1) - V_{\mathcal{M},1}^\pi(\rho) \right| \le \sum_{h=1}^H \sqrt{\mathcal{E}_{\mathcal{M}}^{\pi,\pi}(f)_h}.$$

Proof. For each step h, define the Bellman residual

$$\Delta_h^f(s,a) = f_h(s,a) - \mathcal{T}_{\mathcal{M},h}^\pi f_{h+1}(s,a).$$

Using the definition of the Bellman operator,

$$f_h(s_h,a_h) = r_h(s_h,a_h) + \mathbb{E}[f_{h+1}(s_{h+1},\pi_{h+1}) \mid s_h,a_h] + \Delta_h^f(s_h,a_h).$$

Taking expectation along a trajectory generated by π in $\mathcal{M}$ and summing over $h = 1,\dots,H$ gives

$$f_1(\rho,\pi_1) = V_{\mathcal{M},1}^\pi(\rho) + \sum_{h=1}^H \mathbb{E}_{(s,a)\sim d_{\mathcal{M},h}^\pi}[\Delta_h^f(s,a)].$$

Therefore,

$$\left| f_1(\rho,\pi_1) - V_{\mathcal{M},1}^\pi(\rho) \right| \le \sum_{h=1}^H \left| \mathbb{E}_{d_{\mathcal{M},h}^\pi}[\Delta_h^f] \right|.$$

By Cauchy-Schwarz,

$$\left| \mathbb{E}_{d_{\mathcal{M},h}^\pi}[\Delta_h^f] \right| \le \sqrt{\mathbb{E}_{d_{\mathcal{M},h}^\pi}[(\Delta_h^f)^2]} = \sqrt{\mathcal{E}_{\mathcal{M}}^{\pi,\pi}(f)_h}.$$

Combining the inequalities proves the claim. □

B.3.2 PER-STATE HEDGE REGRET

The policy update in Algorithms 2 and 3 applies multiplicative weights independently at every state and step, with gains given by the current critic. The following standard guarantee (Freund & Schapire, 1997; Arora et al., 2012) holds deterministically, and hence simultaneously at every state-step pair and for arbitrary, data-dependent gain sequences.

Lemma 12 (Per-state Hedge with arbitrary initialization)**.** *Fix (s,h) and let $g^{(1)},\dots,g^{(T)}:\mathcal{A}\to[0,H]$ be an arbitrary sequence of gain vectors. Let $\pi_h^{(1)}(\cdot|s)=q\in\Delta(\mathcal{A})$ have full support and*

$$\pi_h^{(t+1)}(a|s)\propto\pi_h^{(t)}(a|s)\exp\big(\eta\, g^{(t)}(a)\big).$$

Then, for every $\eta>0$ and every distribution $p^\star\in\Delta(\mathcal{A})$,

$$\sum_{t=1}^{T}\Big\langle g^{(t)},\, p^\star-\pi_h^{(t)}(\cdot|s)\Big\rangle\le\frac{\mathrm{KL}(p^\star\,\|\,q)}{\eta}+\frac{\eta H^2T}{8}.$$

In particular, for $q=\mathrm{Unif}(\mathcal{A})$ and $\eta=\frac{1}{H}\sqrt{\frac{8\log|\mathcal{A}|}{T}}$, the right-hand side is at most $H\sqrt{\frac{T\log|\mathcal{A}|}{2}}$.

Proof. Let $\Phi_t=\mathrm{KL}(p^\star\|\pi_h^{(t)}(\cdot|s))$ and $Z_t=\sum_a\pi_h^{(t)}(a|s)\,e^{\eta g^{(t)}(a)}$. The update rule gives

$$\Phi_{t+1}-\Phi_t=\sum_a p^\star(a)\log\frac{\pi_h^{(t)}(a|s)}{\pi_h^{(t+1)}(a|s)}=-\eta\langle g^{(t)},p^\star\rangle+\log Z_t.$$

Hoeffding's lemma, applied to the random variable $g^{(t)}(a)$ with $a\sim\pi_h^{(t)}(\cdot|s)$ and range H, yields $\log Z_t\le\eta\langle g^{(t)},\pi_h^{(t)}(\cdot|s)\rangle+\eta^2H^2/8$. Rearranging, summing over t, and using $\Phi_{T+1}\ge0$,

$$\sum_{t=1}^{T}\big\langle g^{(t)},\,p^\star-\pi_h^{(t)}(\cdot|s)\big\rangle\le\frac{\Phi_1}{\eta}+\frac{\eta H^2T}{8}=\frac{\mathrm{KL}(p^\star\|q)}{\eta}+\frac{\eta H^2T}{8}.$$

For uniform q we have $\mathrm{KL}(p^\star\|q)\le\log|\mathcal{A}|$, and the stated η balances the two terms, giving the classical Hedge bound (Freund & Schapire, 1997; Arora et al., 2012). The bound is a deterministic statement about the update rule, so it holds simultaneously for every (s,h) and for the data-dependent critics $g^{(t)}=f_h^{(t)}(s,\cdot)$ produced by the algorithm. □

A caution about prefixes is in order. The inequality of Lemma 12 can be *evaluated* at any prefix t, but with a fixed learning rate tuned to a final horizon T its value at prefix t is $\mathrm{KL}/\eta+\eta H^2t/8=\Theta\big(H\sqrt{T\log|\mathcal{A}|}\big)$ for $t\le T$, not $\Theta(H\sqrt{t\log|\mathcal{A}|})$. Whenever a guarantee must hold simultaneously at every prefix, as in the model-selection analysis of Appendix E.2, the following time-varying variant is required.

Lemma 13 (Anytime per-state Hedge)**.** *Fix (s,h) and gains $g^{(1)},g^{(2)},\dots:\mathcal{A}\to[0,H]$. Let the policy at round t be the exponentially weighted distribution over cumulative gains with the nonincreasing schedule $\eta_t=\frac{1}{H}\sqrt{8\log|\mathcal{A}|/t}$, that is,*

$$\pi_h^{(t)}(a|s)\propto\exp\Big(\eta_t\sum_{\tau=1}^{t-1}g^{(\tau)}(a)\Big),$$

which coincides with the multiplicative update of Lemma 12 when the schedule is constant. Then, deterministically, for every prefix $t\ge1$ and every $p^\star\in\Delta(\mathcal{A})$,

$$\sum_{\tau=1}^{t}\big\langle g^{(\tau)},\,p^\star-\pi_h^{(\tau)}(\cdot|s)\big\rangle\le2H\sqrt{t\log|\mathcal{A}|}.$$

Proof. Rescale to $\hat g^{(\tau)}:=g^{(\tau)}/H\in[0,1]^{\mathcal{A}}$ and $\hat\eta_\tau:=H\eta_\tau=\sqrt{8\log|\mathcal{A}|/\tau}$, nonincreasing, let $G_t:=\sum_{\tau\le t}\hat g^{(\tau)}$ with $G_0:=0$, and define the potential $\Phi_t(\eta):=\frac{1}{\eta}\log\big(\frac{1}{|\mathcal{A}|}\sum_a e^{\eta G_t(a)}\big)$. Two

facts. First, $\Phi_t(\eta)$ is nondecreasing in $\eta > 0$, by the power-mean inequality. Second, for the played distribution $\pi^{(t)} \propto e^{\hat{\eta}_t G_{t-1}}$,

$$\Phi_t(\hat{\eta}_t) - \Phi_{t-1}(\hat{\eta}_t) = \frac{1}{\hat{\eta}_t} \log \left\langle \pi^{(t)}, e^{\hat{\eta}_t \hat{g}^{(t)}} \right\rangle \le \left\langle \pi^{(t)}, \hat{g}^{(t)} \right\rangle + \frac{\hat{\eta}_t}{8},$$

by Hoeffding's lemma for $[0,1]$-valued gains. Since $\hat{\eta}_{t+1} \le \hat{\eta}_t$, the first fact gives $\Phi_t(\hat{\eta}_{t+1}) \le \Phi_t(\hat{\eta}_t)$, so summing the second fact telescopes through the intermediate quantities $\Phi_\tau(\hat{\eta}_{\tau+1})$,

$$\Phi_t(\hat{\eta}_{t+1}) \le \Phi_0(\hat{\eta}_1) + \sum_{\tau \le t} \left[\left\langle \pi^{(\tau)}, \hat{g}^{(\tau)} \right\rangle + \frac{\hat{\eta}_\tau}{8} \right] = \sum_{\tau \le t} \left\langle \pi^{(\tau)}, \hat{g}^{(\tau)} \right\rangle + \frac{1}{8} \sum_{\tau \le t} \hat{\eta}_\tau.$$

On the other hand $\Phi_t(\hat{\eta}_{t+1}) \ge \max_a G_t(a) - \log|\mathcal{A}|/\hat{\eta}_{t+1} \ge \langle p^\star, G_t \rangle - \log|\mathcal{A}|/\hat{\eta}_{t+1}$ for every $p^\star \in \Delta(\mathcal{A})$. Rearranging and evaluating the schedule,

$$\sum_{\tau \le t} \left\langle \hat{g}^{(\tau)},\, p^\star - \pi^{(\tau)} \right\rangle \le \frac{\log|\mathcal{A}|}{\hat{\eta}_{t+1}} + \frac{1}{8} \sum_{\tau \le t} \hat{\eta}_\tau \le \sqrt{\frac{(t+1)\log|\mathcal{A}|}{8}} + \frac{\sqrt{8\log|\mathcal{A}|}}{8} \cdot 2\sqrt{t} \le 2\sqrt{t \log|\mathcal{A}|}$$

for $t \ge 1$. Rescaling by H proves the claim. The statement is deterministic, so it holds simultaneously over (s,h) and for data-dependent gains; the telescoping device is the standard analysis of exponentially weighted forecasters with time-varying rates, see Cesa-Bianchi & Lugosi (2006, Section 2.3) and the corrected treatments thereof. □

B.3.3 Extended value-difference decomposition

The next identity is the backbone of both main proofs. It is an instance of the extended value-difference decompositions used in the analysis of optimistic policy optimization (Cai et al., 2020; Shani et al., 2020), specialized to our setting.

Lemma 14 (Decomposition along the comparator)**.** *Let π be any comparator policy, π' any policy, and $f \in \mathcal{F}$ with $f_{H+1} \equiv 0$. Then, in any MDP $\mathcal{M}$,*

$$\begin{aligned} V^{\pi}_{\mathcal{M},1}(\rho) - V^{\pi'}_{\mathcal{M},1}(\rho) = &\sum_{h=1}^{H} \mathbb{E}_{s \sim d^{\pi}_{\mathcal{M},h}} \left[\langle f_h(s,\cdot),\, \pi_h(\cdot|s) - \pi'_h(\cdot|s) \rangle \right] \\ &- \sum_{h=1}^{H} \mathbb{E}_{(s,a) \sim d^{\pi}_{\mathcal{M},h}} \left[\left(f_h - \mathcal{T}^{\pi'}_{\mathcal{M},h} f_{h+1} \right)(s,a) \right] + \left[f_1(\rho, \pi'_1) - V^{\pi'}_{\mathcal{M},1}(\rho) \right], \end{aligned} \tag{16}$$

where in the first sum $d^{\pi}_{\mathcal{M},h}$ denotes the state marginal at step h under π, and in the second sum actions are drawn from π_h.

Proof. Write the three-term split

$$V^{\pi}_1(\rho) - V^{\pi'}_1(\rho) = \underbrace{\left[V^{\pi}_1(\rho) - f_1(\rho, \pi_1) \right]}_{\text{(i)}} + \underbrace{\left[f_1(\rho, \pi_1) - f_1(\rho, \pi'_1) \right]}_{\text{(ii)}} + \underbrace{\left[f_1(\rho, \pi'_1) - V^{\pi'}_1(\rho) \right]}_{\text{(iii)}},$$

all in $\mathcal{M}$. The telescoping identity in the proof of Lemma 11 gives (i) $= -\sum_h \mathbb{E}_{(s,a) \sim d^{\pi}_h}[(f_h - \mathcal{T}^{\pi}_h f_{h+1})(s,a)]$. Now decompose each residual through the backup operator of π', writing

$$f_h - \mathcal{T}^{\pi}_h f_{h+1} = \left(f_h - \mathcal{T}^{\pi'}_h f_{h+1} \right) + \left(\mathcal{T}^{\pi'}_h f_{h+1} - \mathcal{T}^{\pi}_h f_{h+1} \right),$$

and note that $(\mathcal{T}^{\pi'}_h f_{h+1} - \mathcal{T}^{\pi}_h f_{h+1})(s,a) = -\mathbb{E}_{s' \sim P_{\mathcal{M},h}(\cdot|s,a)}[\langle f_{h+1}(s',\cdot), \pi_{h+1}(\cdot|s') - \pi'_{h+1}(\cdot|s') \rangle]$. Since $(s,a) \sim d^{\pi}_h$ followed by $s' \sim P_{\mathcal{M},h}(\cdot|s,a)$ produces $s' \sim d^{\pi}_{h+1}$ (state marginal), summing over h turns these correction terms into $\sum_{h'=2}^{H} \mathbb{E}_{s \sim d^{\pi}_{h'}}[\langle f_{h'}(s,\cdot), \pi_{h'}(\cdot|s) - \pi'_{h'}(\cdot|s) \rangle]$ (the $h' = H+1$ term vanishes because $f_{H+1} \equiv 0$). Finally, (ii) $= \mathbb{E}_{s \sim \rho}[\langle f_1(s,\cdot), \pi_1(\cdot|s) - \pi'_1(\cdot|s) \rangle]$ supplies the $h' = 1$ term. Collecting the pieces yields (16). □

B.4 Source concentration and the validity event

The algorithm constructs empirical source confidence sets. The weak transfer modulus, however, is defined in terms of population Bellman errors. We therefore isolate the statistical event needed to connect the two. Care is required on two counts, the classical double-sampling issue and the fact that the policies encountered by the algorithm are data-dependent.

Corrected empirical residual. For a transition $(s_h^i, a_h^i, r_h^i, s_{h+1}^i)$ from $\mathcal{D}_{\mathrm{src}}$ and functions g, f', write

$$\ell_h^i(g, f') := \big(g(s_h^i, a_h^i) - r_h^i - f'(s_{h+1}^i, \pi_{h+1})\big)^2.$$

The naive empirical squared residual $\frac{1}{n}\sum_i \ell_h^i(f_h, f_{h+1})$ does *not* concentrate around $\mathcal{E}^{\mu,\pi}_{\mathcal{M}_{\mathrm{src}}}(f)_h$. Its expectation equals the Bellman error *plus* the conditional variance of the backup target, the well-known double-sampling phenomenon (Antos et al., 2008). Under Bellman completeness the standard correction subtracts the empirical loss of the best fit to the same targets (Antos et al., 2008; Chen & Jiang, 2019; Xie et al., 2021a), giving

$$\widehat{\mathcal{E}}^{\mu,\pi}_{\mathrm{src}}(f)_h := \frac{1}{n}\sum_{i=1}^{n} \ell_h^i(f_h, f_{h+1}) - \min_{g\in\mathcal{F}_h} \frac{1}{n}\sum_{i=1}^{n} \ell_h^i(g, f_{h+1}). \tag{17}$$

The subtraction removes the variance offset because, by completeness, $\mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{src}},h} f_{h+1} \in \mathcal{F}_h$ realizes the conditional mean of the regression targets. The empirical confidence set is

$$\widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi) = \left\{ f \in \mathcal{F} : \widehat{\mathcal{E}}^{\mu,\pi}_{\mathrm{src}}(f)_h \le \beta \text{ for all } h \in [H] \right\}.$$

Assumption 3 (Uniform empirical-to-population validity)**.** *There exists $\epsilon_{\mathrm{off}} = \epsilon_{\mathrm{off}}(n, \delta)$ such that, with probability at least $1-\delta$, the following event holds simultaneously for all policies used in the analysis.*

$$\widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi) \subseteq \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon_{\mathrm{off}}) \quad \textit{and} \quad Q^{\pi}_{\mathcal{M}_{\mathrm{src}}} \in \widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi) \qquad \textit{for all } \pi \in \Pi.$$

The second requirement guarantees in particular that the confidence sets are nonempty, so the pessimistic critic in (4) is well defined. For finite function and policy classes, the event follows from a standard uniform concentration argument.

Lemma 15 (A finite-class concentration bound)**.** *Suppose $\mathcal{F}$ and Π are finite, rewards lie in $[0,1]$, and all functions in $\mathcal{F}$ take values in $[0, H]$. Then there is a universal constant $C > 0$ such that, with*

$$\beta \ge CH^2\sqrt{\frac{\log(|\mathcal{F}||\Pi|H/\delta)}{n}}, \qquad \epsilon_{\mathrm{off}}(n,\delta) = \beta + CH^2\sqrt{\frac{\log(|\mathcal{F}||\Pi|H/\delta)}{n}},$$

Assumption 3 holds (here n is the number of source episodes, each contributing one transition per step h).

Proof. Fix f, π, h and set $g^\star := \mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{src}},h} f_{h+1}$, which lies in $\mathcal{F}_h$ by completeness. For each sample, the excess loss $\ell_h^i(f_h, f_{h+1}) - \ell_h^i(g^\star, f_{h+1})$ is bounded in magnitude by $O(H^2)$ and has conditional mean $(f_h - g^\star)^2(s_h^i, a_h^i)$, since $g^\star$ is the conditional mean of the target $r_h^i + f_{h+1}(s_{h+1}^i, \pi_{h+1})$. Bernstein's inequality and a union bound over (f, π, h) and over $g \in \mathcal{F}_h$ give, with probability at least $1-\delta$, simultaneously for all (f, π, h),

$$\left|\widehat{\mathcal{E}}^{\mu,\pi}_{\mathrm{src}}(f)_h - \mathcal{E}^{\mu,\pi}_{\mathcal{M}_{\mathrm{src}}}(f)_h\right| \le CH^2\sqrt{\frac{\log(|\mathcal{F}||\Pi|H/\delta)}{n}},$$

where the minimum in (17) is handled in both directions, namely the minimum is at most the empirical loss of $g^\star$, and uniformly over $g \in \mathcal{F}_h$ the empirical excess loss of g against $g^\star$ exceeds its nonnegative conditional mean minus the same deviation term, so the minimum cannot undershoot the empirical loss of $g^\star$ by more than that term; see Xie et al. (2021a, Appendix A) for the identical two-sided argument. The inclusion in Assumption 3 follows from the triangle inequality, and the membership of $Q^{\pi}_{\mathcal{M}_{\mathrm{src}}}$ (whose population residual is zero for every h) follows from the same deviation bound together with the stated choice of β. □

The excess-loss variable in the proof above is self-bounding, its variance is controlled by its mean under Bellman completeness, so the source radius admits a Bernstein refinement at the fast scale.

Lemma 16 (Bernstein refinement of the source radius)**.** *In the setting of Lemma 15, let* $\iota_n := \log(|\mathcal{F}||\Pi|H/\delta)$ *and* $\gamma_n := C_B H^2 \iota_n / n$ *for a sufficiently large universal constant* C_B*. With probability at least* $1-\delta$*, simultaneously for all* $f \in \mathcal{F}$*,* $\pi \in \Pi$*, and* h*,*

$$\tfrac{1}{2}\,\mathcal{E}^{\mu,\pi}_{\mathcal{M}_{\mathrm{src}}}(f)_h - \gamma_n \;\le\; \widehat{\mathcal{E}}^{\mu,\pi}_{\mathrm{src}}(f)_h \;\le\; 2\,\mathcal{E}^{\mu,\pi}_{\mathcal{M}_{\mathrm{src}}}(f)_h + \gamma_n.$$

Consequently Assumption 3 holds with $\beta = \gamma_n$ *and* $\epsilon_{\mathrm{off}} = 4\gamma_n$*, and in general one may take*

$$\epsilon_{\mathrm{off}}(n,\delta) = \min\Big\{CH^2\sqrt{\iota_n/n}\,,\; 4C_B H^2 \iota_n/n\Big\},$$

the smaller of the Hoeffding and Bernstein radii, with the corresponding choice of β*.*

Proof. Fix (f,π,h) and let Z_i be the excess loss of the proof of Lemma 15 on the ith source episode, i.i.d. across episodes, with $\mathbb{E}[Z_i \mid s_i,a_i] = (f_h - g^\star)^2(s_i,a_i)$ and $|Z_i| \le cH^2$. Writing $\mathcal{E}_h(f) := \mathcal{E}^{\mu,\pi}_{\mathcal{M}_{\mathrm{src}}}(f)_h$, the same algebra as in Lemma 27 gives $\mathrm{Var}(Z_i) \le \mathbb{E}[Z_i^2] \le cH^2\,\mathcal{E}_h(f)$, so Bernstein's inequality with a union bound over (f,π,h) yields $|\bar{Z} - \mathcal{E}_h(f)| \le \sqrt{2cH^2\mathcal{E}_h(f)\iota_n/n} + cH^2\iota_n/n \le \frac{1}{2}\mathcal{E}_h(f) + c'H^2\iota_n/n$ by the arithmetic-geometric mean inequality. The corrected estimator's subtracted minimum is handled exactly as in Lemma 27, on a further union event over $g \in \mathcal{F}_h$ whose excess losses against $g^\star$ have nonnegative conditional means, which keeps the empirical minimum within $c'H^2\iota_n/n$ of the empirical loss of $g^\star$ from both sides. Combining the displays proves the two-sided inequality with $C_B = 4c'$. For the consequence, $Q^\pi_{\mathcal{M}_{\mathrm{src}}}$ has zero population residual, so its empirical residual is at most $\gamma_n = \beta$ by the right inequality, and any f with $\widehat{\mathcal{E}}_h(f) \le \beta$ has $\mathcal{E}_h(f) \le 2(\beta+\gamma_n) = 4\gamma_n$ by the left inequality. The minimum form follows by running whichever of the two arguments gives the smaller radius, at the cost of a union of the two events. □

Data-dependent policies. Algorithm 2 evaluates residuals at policies $\pi^{(t)}$ that are themselves functions of $\mathcal{D}_{\mathrm{src}}$, so a union bound over a fixed class does not directly apply. Two standard remedies suffice. First, by construction every iterate, offline or online, lies in the algorithmic policy closure $\overline{\Pi}$ defined in Section 2, an ordered sequence of at most $T + T_{\mathrm{off}}$ critics with *per-critic* rates from the finite collection $\mathcal{C}$ (a single fixed η, the anytime schedule $\{\eta_t\}_{t\le T}$ of Lemma 13, and the offline rate) applied to a base policy in $\mathcal{Q}_{\mathrm{base}}$, with at most two interleaved mixing operators whose positions contribute the absorbed factor $2\log(L+2)$; the class is fixed before any data are observed and has $\log|\overline{\Pi}| \lesssim (T+T_{\mathrm{off}})\log(|\mathcal{C}||\mathcal{F}|) + \log|\mathcal{Q}_{\mathrm{base}}| + \log(1+|\Pi_{\mathrm{fixed}}|)$. Applying Lemma 15 with $\Pi \cup \overline{\Pi}$ inflates the radius to $\epsilon_{\mathrm{off}} = \beta + CH^2\sqrt{\mathcal{C}_{\mathrm{alg}}/n}$, equal to $\beta + CH^2\sqrt{T\log(T|\mathcal{F}|H/\delta)/n}$ for the concrete class, which tends to zero for fixed H when $n/[T\log(T|\mathcal{F}|H/\delta)] \to \infty$. Second, the max-min variant of Proposition 18 only ever evaluates residuals at policies in the prescribed comparator class Π, so no inflation is needed there. For general function classes, cardinalities are replaced by covering numbers or localized complexities in the usual way.

It is convenient to record the union-bound complexity once. Define

$$\mathcal{C}_{\mathrm{alg}} := (T+T_{\mathrm{off}})\log\big(|\mathcal{C}|(|\mathcal{F}|+1)\big) + \log|\mathcal{Q}_{\mathrm{base}}| + \log(1+|\Pi_{\mathrm{fixed}}|) + \log(H/\delta),$$

so that the finite-class source radii read $\beta, \epsilon_{\mathrm{off}} = O\big(H^2\sqrt{\mathcal{C}_{\mathrm{alg}}/n}\big)$ in full generality, or, by Lemma 16 applied to $\Pi\cup\overline{\Pi}$, at the Bernstein scale $\beta, \epsilon_{\mathrm{off}} = O\big(H^2\,\mathcal{C}_{\mathrm{alg}}/n\big)$, which is smaller whenever $\mathcal{C}_{\mathrm{alg}} \le n$; every downstream statement may use the minimum of the two radii. For the concrete class used in this paper, $T_{\mathrm{off}} \le T$, $|\mathcal{C}| = O(T)$, and $\mathcal{Q}_{\mathrm{base}}, \Pi_{\mathrm{fixed}}$ are short fixed lists, so $\mathcal{C}_{\mathrm{alg}} = O(T\log(T|\mathcal{F}|H/\delta))$ and the simplified radii displayed in the theorems follow; the Π_{fixed} term matters only when a large comparator family is inserted, as in Corollary 40, and it remains a centralized complexity term; we do not repeat the general expression in each statement. In the main text we write $\epsilon_{\mathrm{off}}(n)$ or $\epsilon_{\mathrm{off}}(n,\delta)$ to denote the resulting source statistical radius.

A sufficient regime in which the regret bound is sublinear. The source radius depends on how many policy updates the confidence event must cover. To make this cost explicit, suppose the two environments coincide and, for every iterate, both the iterate occupancy and the comparator occupancy

are dominated by $C_{\mathrm{cov}} d^{\mu}_{\mathrm{src},h}$ at every stage. The residual comparison in the proof of Corollary 35 then gives

$$\mathrm{Regret}(T) \le 2HT\sqrt{C_{\mathrm{cov}}\epsilon_{\mathrm{off}}} + H^2\sqrt{T\log|\mathcal{A}|/2}.$$

With the Bernstein source radius this becomes

$$\mathrm{Regret}(T) = O\left(H^2 T\sqrt{\frac{C_{\mathrm{cov}}\mathcal{C}_{\mathrm{alg}}}{n}} + H^2\sqrt{T\log|\mathcal{A}|}\right).$$

Thus, for fixed horizon and action-set size, $n/(C_{\mathrm{cov}}\mathcal{C}_{\mathrm{alg}}) \to \infty$ suffices for sublinear regret as $T \to \infty$. Matching a square-root-in-T bound through this argument requires n of order at least $C_{\mathrm{cov}}\mathcal{C}_{\mathrm{alg}}T$. Under the concrete closure $\mathcal{C}_{\mathrm{alg}} = O(T\log(T|\mathcal{F}|H/\delta))$, this is a source budget of order $C_{\mathrm{cov}}T^2\log(T|\mathcal{F}|H/\delta)$. The Hoeffding source radius instead gives the sufficient scaling n of order $C^2_{\mathrm{cov}}\mathcal{C}_{\mathrm{alg}}T^2$. These are sufficient conditions from the present analysis, not source-sample lower bounds or optimal joint rates. Under dynamics shift, the transfer terms must additionally be controlled; a larger source dataset alone does not guarantee their decay.

B.4.1 Covering-number extension to infinite classes

Proposition 17 (Covering-number version of the concentration)**.** *Let each $\mathcal{F}_h \subseteq [0, H-h+1]^{\mathcal{S}\times\mathcal{A}}$ have finite sup-norm covering numbers, let $N_\alpha := \max_h N_\alpha(\mathcal{F}_h, \|\cdot\|_\infty)$, and suppose Assumptions 1 and 2 hold over $\overline{\Pi}$ defined from $\mathcal{F}$ itself. Let $L := T + T_{\mathrm{off}}$, let $\eta_{\max} := \max\mathcal{C}$, and fix the rate-adapted discretization scale $\bar\alpha := 1/\big(2nH(1+\eta_{\max}HL)\big)$, so that the Lipschitz transfer error below is $CH(1+\eta_{\max}HL)\bar\alpha = O(1/n)$ and $2\eta_{\max}L\bar\alpha \le 1/(nH^2) \le 1$ in all regimes and is absorbed by the statistical radius automatically; $\mathcal{Q}_{\mathrm{base}}$ and $\mathcal{C}$ are as in Section 2, and the data-dependent warm start is covered through the offline iterate sequences inside $\overline{\Pi}$. Then the stagewise conclusions of Lemma 15 and of the data-dependent-policy inflation hold with, at each stage h,*

$$\log(|\mathcal{F}||\Pi|H/\delta) \text{ replaced by } \log N_{\bar\alpha} + \log N_{\Pi,h+1} + \log(H/\delta),$$
$$N_{\Pi,h+1} \le |\Pi_{\mathrm{fixed}}| + |\mathcal{Q}_{\mathrm{base}}|\,(L+2)^2\big(1+|\mathcal{C}|N_{\bar\alpha}\big)^L,$$
$$\log N_{\Pi,h+1} \le L\log\big(1+|\mathcal{C}|N_{\bar\alpha}\big) + \log|\mathcal{Q}_{\mathrm{base}}| + 2\log(L+2) + \log(1+|\Pi_{\mathrm{fixed}}|),$$

up to universal constants, where $N_{\Pi,h+1}$ is the covering number of the algorithm-generated class of stage-$(h+1)$ conditionals *$\pi_{h+1}(\cdot|s)$ (a cover of full H-stage policies is not needed, since the concentration is stagewise; the union over stages contributes the displayed $\log H$), and with an additive discretization error $O(1/n)$ absorbed into the radii by the rate-adapted choice of $\bar\alpha$. Operation sequences containing mixing operators are covered at the* same *scale $\bar\alpha$ and with the same displayed covering numbers, because the mixing operator contracts the projective log-ratio metric used in the proof; the two permitted mixings cost only their position count $2\log(L+2)$, already displayed, so the bounds hold literally as displayed, a detail centralized here and not repeated in the instantiations. (The contraction itself would tolerate any number of mixings; allowing at most L total operations, updates or mixings, would give entropy $L\log(2+|\mathcal{C}|N_{\bar\alpha}) + \log|\mathcal{Q}_{\mathrm{base}}|$, of the same order.) The two-sided Bernstein form of Lemma 16 also extends. With*

$$\gamma_{n,h} := \frac{CH^2\big[\log N_{\bar\alpha} + \log N_{\Pi,h+1} + \log(H/\delta)\big]}{n} + O(1/n),$$

one has $\frac{1}{2}\mathcal{E}_h(f) - \gamma_{n,h} \le \widehat{\mathcal{E}}_h(f) \le 2\mathcal{E}_h(f) + \gamma_{n,h}$ simultaneously over the class, by Lemma 16 on the joint critic and policy net followed by the same Lipschitz transfer, whose additive $O(1/n)$ error matches the Bernstein scale by the choice of $\bar\alpha$; the self-bounding variance survives the projection of the regression center onto the net, since $(f_h - g^\star_{\mathrm{net}})^2 \le 2(f_h - g^\star)^2 + 2\bar\alpha^2$ gives $\mathrm{Var}(Z) \lesssim H^2\big(\mathcal{E}_h(f) + \bar\alpha^2\big)$ with $\bar\alpha^2 = O(1/n^2)$ absorbed; this supplies the Bernstein radii for the linear and block instantiations, in particular the $\widetilde{O}(n^{-1})$ covered-direction term claimed for the linear class. Consequently Theorems 1, 21, 22, and 5 hold for infinite classes (and, through the block-level Remark 8, so do Theorem 2 and Corollary 3), including the concrete warm start produced as Algorithm 2's data-dependent final iterate, with every occurrence of $\log|\mathcal{F}|$ replaced by $\log N_{\bar\alpha}$ in both the Hoeffding (square-root) and Bernstein (linear-in-entropy) radii, and the policy-class terms replaced as displayed; since $T_{\mathrm{off}} \le T$, the stated orders are unchanged. In particular, for the linear class with $\|w\|_2 \le W$ one has the general volumetric bound $\log N_{\bar\alpha} = O\big(d\log(1+$

$W/\bar{\alpha})\big) = O\big(d\log(1+WnH(1+\eta_{\max}HL))\big)$*, which under the standard specialization* $W = 2H\sqrt{d}$ *reads* $\log N_{\bar{\alpha}} = O\big(d\log(dHn(1+\eta_{\max}HL))\big)$*, and analogously* $O\big(\kappa\log(Hn(1+\eta_{\max}HL))\big)$ *for the block-constant class; for the schedules used in this paper the extra dependence sits inside the logarithm, so the displayed linear and block rates are unchanged, and the linear and block instantiations of Section 5.1 carry end-to-end algorithmic guarantees, including Theorem 22 for* every *full-support warm start in the initialization slice* $\overline{\Pi}_{\mathrm{init}}$ *with a known finite divergence bound* D *(and, at budget* $m_0 + T$*, for every warm start of a-priori operational length* m_0*).*

Proof. Fix minimal $\bar{\alpha}$-nets $\mathcal{F}_h^{\mathrm{net}} \subseteq \mathcal{F}_h$. Three Lipschitz facts. First, the squared loss $\ell_h^i(f_h, f_{h+1}) = (f_h(s_h^i, a_h^i) - r_h^i - f_{h+1}(s_{h+1}^i, \pi_{h+1}))^2$ changes by at most $4H\big(\|f_h - f_h'\|_\infty + \|f_{h+1} - f_{h+1}'\|_\infty + H\,\|\pi_{h+1}(\cdot|s_{h+1}^i) - \pi_{h+1}'(\cdot|s_{h+1}^i)\|_1\big)$ under perturbations of the critic pair and of the backup policy, since every quantity involved lies in a range of width at most $2H$. Second, we use a projective-metric contraction argument that handles mixing at any position for free. For full-support conditionals p, p' on $\mathcal{A}$, define the projective log-ratio metric $d_{\mathrm{pr}}(p, p') := \max_a \log \frac{p(a)}{p'(a)} - \min_a \log \frac{p(a)}{p'(a)}$. Two facts hold. (i) A common exponential update obeys $d_{\mathrm{pr}}(U_{c,f}(p), U_{c,f'}(p')) \le d_{\mathrm{pr}}(p, p') + 2c\|f - f'\|_\infty$, since the update adds $c(f - f')(s, a)$ plus an a-independent normalization to the log-ratios and the max-minus-min functional ignores constants. (ii) A common mixing *contracts*, $d_{\mathrm{pr}}(M_\tau p, M_\tau p') \le d_{\mathrm{pr}}(p, p')$: each mixed ratio $\frac{(1-\tau)p(a)+\tau/|\mathcal{A}|}{(1-\tau)p'(a)+\tau/|\mathcal{A}|}$ lies between the original ratio $p(a)/p'(a)$ and 1, and since p, p' are probability vectors the original ratios satisfy $\min_a \le 1 \le \max_a$, so the mixed max-minus-min cannot exceed the original. Every stage-$(h+1)$ conditional the algorithms can deploy is an operation sequence of at most L updates and at most two mixings applied to a member of $\mathcal{Q}_{\mathrm{base}}$; replacing every critic by its net point and keeping the base, the rates, and the mixing positions fixed gives, by (i), (ii), and induction along the sequence, $d_{\mathrm{pr}}(\pi_{h+1}(\cdot|s), \pi_{h+1}'(\cdot|s)) \le 2\eta_{\max}L\bar{\alpha}$ at every state. Normalization brackets the log-ratios around zero ($\min_a \le 0 \le \max_a$), so every log-ratio has magnitude at most $2\eta_{\max}L\bar{\alpha}$ and $\|\pi_{h+1}(\cdot|s) - \pi_{h+1}'(\cdot|s)\|_1 \le e^{2\eta_{\max}L\bar{\alpha}} - 1 \le 4\eta_{\max}L\bar{\alpha}$ whenever $2\eta_{\max}L\bar{\alpha} \le 1$, which the choice of $\bar{\alpha}$ guarantees. The critic net, together with the finite rate collection $\mathcal{C}$, the at most $(L+2)^2$ mixing-position patterns, and the finitely many members of Π_{fixed} covered exactly as themselves, therefore induces a cover of the stage-$(h+1)$ conditionals of $\overline{\Pi}$ of log-cardinality at most $L\log(1+|\mathcal{C}|N_{\bar{\alpha}}) + \log|\mathcal{Q}_{\mathrm{base}}| + 2\log(L+2) + \log(1+|\Pi_{\mathrm{fixed}}|)$ at radius at most $8\eta_{\max}L\bar{\alpha}$ when representatives are chosen from the occupied cells (the $1+$ counts sequences of every length up to L, since $\sum_{m\le L} x^m \le (1+x)^L$), which in particular covers the data-dependent warm start drawn from the initialization slice, its mixed version, and all their descendants, whose total operation count respects the budget L; and we take as cover centers one *actual* deployed-class member from each occupied cell, so that Assumptions 1 and 2 apply to the representatives. Third, population residuals inherit the same moduli of continuity by taking expectations, and the completeness step transfers by projecting $g^\star = \mathcal{T}^\pi_{\mathcal{M},h} f_{h+1}$ onto the net at cost $\bar{\alpha}$. The finite nets need not themselves be Bellman complete: the true regression centers remain in the original class, and their net approximations incur the stated error. Apply the excess-loss concentration argument to the finitely many representative critic–policy triples and regression candidates, and transfer every empirical and population residual to arbitrary (f, π) by the Lipschitz facts, paying an additive $CH(1+\eta_{\max}HL)\bar{\alpha} = O(1/n)$ that the stated radii absorb by the choice of $\bar{\alpha}$. The linear covering bound is the volumetric estimate for a Euclidean ball of radius W, specialized at $W = 2H\sqrt{d}$; the block bound counts $\bar{\alpha}$-grids of the κ block values. □

B.4.2 Proof of the estimation guarantee

Proof of Proposition 24. On the source concentration event, $\mathcal{F}_{\mathrm{src}}^{\mu,\pi}(\beta - c_n) \subseteq \widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi) \subseteq \mathcal{F}_{\mathrm{src}}^{\mu,\pi}(\beta + c_n)$, where the first inclusion holds because a population residual below $\beta - c_n$ forces the empirical residual below β, and the second is Lemma 15. On the target concentration event, $|\widehat{\mathcal{E}}_{\mathrm{tar}}^{\pi,\pi}(f)_h - \mathcal{E}_{\mathcal{M}_{\mathrm{tar}}}^{\pi,\pi}(f)_h| \le c_m$ uniformly over $f \in \mathcal{F}$, by the argument of Lemma 26 with the fixed deployed policy π (the policy is fixed in advance here, so no blocking is needed). Taking suprema over the sandwiched sets and shifting by c_m in each direction yields the two inequalities. □

B.5 Constants and normalizations

The bounds in the main text suppress logarithmic factors and constants depending on the precise concentration inequalities used for empirical Bellman residuals. There are also minor horizon-

Algorithm 2: Pessimistic Hedge for Hybrid Transfer RL

Input: class $\mathcal{F}$, source dataset $\mathcal{D}_{\mathrm{src}}$, iterations T, radius β, learning rate η

initialize $\pi^{(1)}$ uniform at every (s, h);

for $t = 1, \ldots, T$ **do**

  construct $\widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi^{(t)})$ and select the pessimistic critic $f^{(t)}$ as in (4);

  update $\pi^{(t+1)}$ by the per-state Hedge step (5);

Output: mixture policy $\widehat{\pi}$ that samples $t \sim \mathrm{Unif}([T])$ and executes $\pi^{(t)}$

normalization choices depending on whether values are scaled to lie in $[0, H]$ or rewards are rescaled so that values lie in $[0, 1]$. The analysis uses the uniform envelope H for value ranges, which introduces factors of H or H^2 when converting transition discrepancies into Bellman-backup discrepancies; wherever Bellman completeness is invoked, Section 2 adopts the stepwise ranges $f_h \in [0, H - h + 1]$, and the structured population bounds are also stated under the looser $[0, H]$ convention. These factors can be improved in some settings using variance-sensitive or Bernstein-style arguments, but they do not change the central weak-modulus dependence.

C Offline Policy Optimization

C.1 The offline procedure and the max-min variant

The offline procedure and its information-theoretic variant. The three terms in (7) have distinct meanings. The first is the comparator transfer term. Source fitting for the backup operators of the iterates must certify their residuals along the trajectories of the target-optimal policy. The second is the algorithm-dependent transfer term, measuring whether the policies produced by the algorithm stay in regions where source pessimism remains reliable in the target. The final term is the optimization error of the per-state multiplicative-weights update; note that it scales with $\log |\mathcal{A}|$ rather than the cardinality of a policy class, and that the comparator $\pi^{\star}_{\mathrm{tar}}$ need not belong to any prescribed class. When the weak modulus is small for the comparator and the learned policies, source data sharply reduces uncertainty; when it is large, the theorem does not promise transfer.

The proof, given in Appendix C.2, combines three ingredients with the weak modulus. An extended value-difference decomposition, in the style of optimistic policy-optimization analyses (Cai et al., 2020; Shani et al., 2020), splits each performance gap into per-state comparison terms and two Bellman-residual terms, one evaluated along the comparator and one along the iterate; the event (6) and Definition 1 convert the residuals into the weak-modulus terms $\delta_h^{\pi^{\star}_{\mathrm{tar}}|\pi^{(t)}}$ and $\delta_h^{\pi^{(t)}}$; and the per-state Hedge guarantee (Freund & Schapire, 1997; Arora et al., 2012) controls the comparison terms simultaneously at every state, contributing the final optimization term. The comparator residual is an *off-policy* object. The analysis requires exactly the two-policy modulus of Definition 1, which is why this analysis does not yield a guarantee in terms of $\delta_h^{\pi^{\star}_{\mathrm{tar}}}$ alone. Figure 2 summarizes the structure of the argument and where each modulus enters.

Remark 3 (Role of pessimism)**.** *The bounds of Theorem 1 and of the online Theorem 21 hold verbatim for* any *measurable selection from the confidence set, since their proofs use only membership; this is a stronger statement than one about the pessimistic choice in particular. The same is true of the max-min variant (Proposition 18), whose proof uses only membership and the maximization over* Π*. Pessimism does sharpen the iterate term. On the validity event the confidence set* $\widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi^{(t)})$ *contains* $Q^{\pi^{(t)}}_{\mathcal{M}_{\mathrm{src}}}$*, so the pessimistic critic satisfies* $f_1^{(t)}(\rho, \pi_1^{(t)}) \le V^{\pi^{(t)}}_{\mathcal{M}_{\mathrm{src}},1}(\rho)$*, and the iterate term* $(\mathrm{III})_t = f_1^{(t)}(\rho, \pi_1^{(t)}) - V^{\pi^{(t)}}_{\mathcal{M}_{\mathrm{tar}},1}(\rho)$ *in the proof of Theorem 1 is at most* $\min\{\sum_h (\delta_h^{\pi^{(t)}}(\epsilon_{\mathrm{off}}))^{1/2},\ V^{\pi^{(t)}}_{\mathcal{M}_{\mathrm{src}},1}(\rho) - V^{\pi^{(t)}}_{\mathcal{M}_{\mathrm{tar}},1}(\rho)\}$*. The same refinement holds for the max-min variant with* $\widehat{\pi}_{\mathrm{mm}}$ *in place of* $\pi^{(t)}$*. The pessimistic learner therefore pays for the iterates' own route only to the extent that the source overestimates their target value.*

If one is willing to forgo computational considerations and optimize over a finite comparator class Π directly, the on-policy moduli suffice.

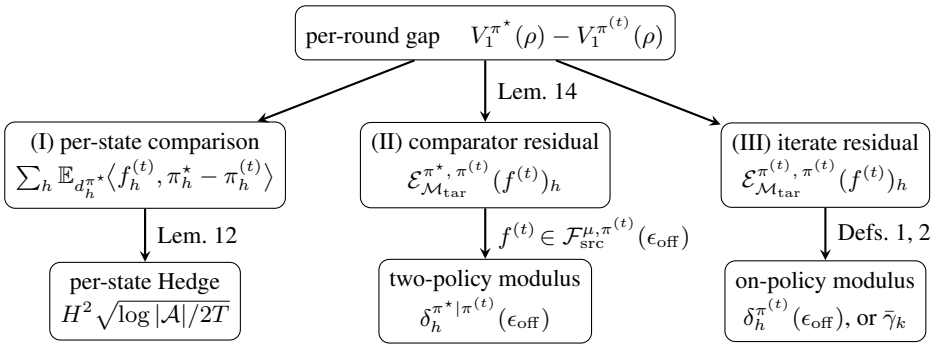


Figure 2: Structure of the proofs of Theorems 1, 21, and 2. The extended value-difference lemma splits each per-round gap into three terms. The per-state Hedge bound controls (I) deterministically at every state. Source membership of the critic converts (II) into the two-policy weak modulus and (III) into the on-policy modulus. With lazy updates and same-block target certification, (III) is instead capped by the target radius $\bar{\gamma}_k$, and (II) by the strong modulus.

Proposition 18 (Information-theoretic variant)**.** *Assume* $\pi^\star_{\mathrm{tar}} \in \Pi \subseteq \overline{\Pi}$. *Let* $\widehat{\pi}_{\mathrm{mm}} \in \mathrm{argmax}_{\pi\in\Pi} \min_{f\in\widehat{\mathcal{F}}_{\mathrm{src}}(\beta,\pi)} f_1(\rho,\pi_1)$. *On the event* (6)*,* $\mathrm{SubOpt}_{\mathrm{tar}}(\widehat{\pi}_{\mathrm{mm}}) \le \sum_{h=1}^H \big[\sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}}(\epsilon_{\mathrm{off}})} + \sqrt{\delta_h^{\widehat{\pi}_{\mathrm{mm}}}(\epsilon_{\mathrm{off}})}\big]$.

The max-min scheme adapts Bellman-consistent pessimism (Xie et al., 2021a) to source-generated data. It achieves the on-policy modulus bound without an optimization term, at the cost of enumerating Π; its comparator must belong to that class. The Hedge scheme instead makes one pessimistic-evaluation oracle call per iteration. The max-min proof appears in Appendix C.2.

Remark 4 (Oracle efficiency, and relation to the policy-finetuning template)**.** *The per-state Hedge scheme is the computationally motivated variant. Each iteration makes a single call to a pessimistic-evaluation oracle, the constrained minimization in* (4)*, which is the same oracle underlying pessimistic offline RL, and the policy update is closed-form; no enumeration of a policy class and no min-max game over* Π *is required, which is why the guarantee scales with* $\log|\mathcal{A}|$ *rather than* $\log|\Pi|$*. In the tabular instantiation of Appendix H.2, the oracle is exact robust policy evaluation over per-pair total-variation balls, computable in time polynomial in* $|\mathcal{S}|$*,* $|\mathcal{A}|$*, and* H*, so the entire algorithm is polynomial there. In the same-environment policy-finetuning setting, Xie et al. (2021b) analyze a simple template that rolls out a reference policy and then runs offline RL on the collected data. One round of that template is exactly one block of Algorithm 1 without the source constraint. The template alone does not transfer across environments, since dropping the source constraint discards the source information and keeping it without a certificate reintroduces the shift bias; Algorithm 1 is the certified multi-round extension, where the intersection decides at each tested block how much of the source certificate survives target evidence.*

C.2 Proof of the main suboptimality bound

We now prove the main hybrid transfer guarantee. We use the notation from Algorithm 2. The algorithm generates policies $\pi^{(1)},\dots,\pi^{(T)}$ and pessimistic critics $f^{(1)},\dots,f^{(T)}$, and outputs the uniform mixture policy $\widehat{\pi}$ that samples $t\sim\mathrm{Unif}([T])$ and executes $\pi^{(t)}$. Throughout, write $\pi^\star := \pi^\star_{\mathrm{tar}}$ and abbreviate all values and distributions in the target MDP.

Theorem 19 (Hybrid transfer guarantee, detailed version)**.** *Suppose Assumption 3 holds with radius* ϵ_{off} *for all policies produced by the algorithm. Run Algorithm 2 with* $\eta = \frac{1}{H}\sqrt{\frac{8\log|\mathcal{A}|}{T}}$*. Then, on the event of Assumption 3,*

$$\mathrm{SubOpt}_{\mathrm{tar}}(\widehat{\pi}) \le \frac{1}{T}\sum_{t=1}^T\sum_{h=1}^H\left[\sqrt{\delta_h^{\pi^\star|\pi^{(t)}}(\epsilon_{\mathrm{off}})} + \sqrt{\delta_h^{\pi^{(t)}}(\epsilon_{\mathrm{off}})}\right] + H^2\sqrt{\frac{\log|\mathcal{A}|}{2T}}.$$

Consequently, defining the whole per-iterate transfer term

$$A_t := \sum_{h=1}^{H} \left[\sqrt{\delta_h^{\pi^\star|\pi^{(t)}}(\epsilon_{\mathrm{off}})} + \sqrt{\delta_h^{\pi^{(t)}}(\epsilon_{\mathrm{off}})} \right],$$

in which both *moduli depend on the iterate, and letting* $\widetilde{\pi} := \pi^{(\widetilde{t})}$ *for any* $\widetilde{t} \in \operatorname{argmax}_{t \le T} A_t$,

$$\mathrm{SubOpt}_{\mathrm{tar}}(\widehat{\pi}) \le A_{\widetilde{t}} + H^2 \sqrt{\frac{\log|\mathcal{A}|}{2T}} = \sum_{h=1}^{H} \left[\sqrt{\delta_h^{\pi^\star|\widetilde{\pi}}(\epsilon_{\mathrm{off}})} + \sqrt{\delta_h^{\widetilde{\pi}}(\epsilon_{\mathrm{off}})} \right] + H^2 \sqrt{\frac{\log|\mathcal{A}|}{2T}}.$$

Proof. Since $\widehat{\pi}$ is the uniform mixture over the iterates,

$$\mathrm{SubOpt}_{\mathrm{tar}}(\widehat{\pi}) = \frac{1}{T} \sum_{t=1}^{T} \left(V_1^{\pi^\star}(\rho) - V_1^{\pi^{(t)}}(\rho) \right).$$

Fix $t \in [T]$ and apply Lemma 14 in $\mathcal{M}_{\mathrm{tar}}$ with comparator $\pi = \pi^\star$, policy $\pi' = \pi^{(t)}$, and $f = f^{(t)}$, obtaining

$$V_1^{\pi^\star}(\rho) - V_1^{\pi^{(t)}}(\rho) = (\mathrm{I})_t - (\mathrm{II})_t + (\mathrm{III})_t, \tag{18}$$

where

$$\begin{aligned}
(\mathrm{I})_t &= \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^\star}} \left[\left\langle f_h^{(t)}(s,\cdot),\, \pi_h^\star(\cdot|s) - \pi_h^{(t)}(\cdot|s) \right\rangle \right], \\
(\mathrm{II})_t &= \sum_{h=1}^{H} \mathbb{E}_{(s,a) \sim d_h^{\pi^\star}} \left[\left(f_h^{(t)} - \mathcal{T}^{\pi^{(t)}}_{\mathcal{M}_{\mathrm{tar}},h} f_{h+1}^{(t)} \right)(s,a) \right], \\
(\mathrm{III})_t &= f_1^{(t)}(\rho, \pi_1^{(t)}) - V_1^{\pi^{(t)}}(\rho).
\end{aligned}$$

Comparator residual term. By Jensen's inequality and the definition of the off-policy Bellman error,

$$|(\mathrm{II})_t| \le \sum_{h=1}^{H} \sqrt{\mathcal{E}^{\pi^\star,\pi^{(t)}}_{\mathcal{M}_{\mathrm{tar}}}(f^{(t)})_h}.$$

On the event of Assumption 3, $f^{(t)} \in \widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi^{(t)}) \subseteq \mathcal{F}^{\mu,\pi^{(t)}}_{\mathrm{src}}(\epsilon_{\mathrm{off}})$, so Definition 1 (with evaluation policy $\nu = \pi^\star$ and backup policy $\pi^{(t)}$) gives

$$|(\mathrm{II})_t| \le \sum_{h=1}^{H} \sqrt{\delta_h^{\pi^\star|\pi^{(t)}}(\epsilon_{\mathrm{off}})}.$$

Iterate residual term. By Lemma 10 applied with policy $\pi^{(t)}$ and the same membership,

$$|(\mathrm{III})_t| \le \sum_{h=1}^{H} \sqrt{\delta_h^{\pi^{(t)}}(\epsilon_{\mathrm{off}})}.$$

Optimization term. For every fixed (s,h), the update (5) is exactly the exponential-weights update of Lemma 12 with gains $g^{(t)} = f_h^{(t)}(s,\cdot) \in [0,H]^{\mathcal{A}}$. Applying the lemma with $p^\star = \pi_h^\star(\cdot|s)$, then taking expectations over $s \sim d_h^{\pi^\star}$ (the bound is deterministic and uniform in s) and summing over $h \in [H]$,

$$\sum_{t=1}^{T} (\mathrm{I})_t \le H \cdot H \sqrt{\frac{T \log|\mathcal{A}|}{2}} = H^2 \sqrt{\frac{T \log|\mathcal{A}|}{2}}.$$

Combining the three bounds in (18), summing over t, and dividing by T proves the first claim; the second follows by bounding the average per-iterate transfer term by its maximum. □

Algorithm 3: Target-Adaptive Pessimistic Hedge

Input: full-support initial policy $\pi^{(1)}$ (uniform, or a final offline iterate), $\mathcal{D}_{\mathrm{src}}$, radii β and $\{\alpha_t\}$, learning rate η, episodes T

$\mathcal{D}_{\mathrm{tar}}^{(0)} \leftarrow \emptyset$;
for $t = 1, \dots, T$ **do**
  execute $\pi^{(t)}$ in $\mathcal{M}_{\mathrm{tar}}$, collect τ_t, and set $\mathcal{D}_{\mathrm{tar}}^{(t)} \leftarrow \mathcal{D}_{\mathrm{tar}}^{(t-1)} \cup \{\tau_t\}$;
  construct $\widehat{\mathcal{F}}_t(\pi^{(t)})$ as in (20) and select $f^{(t)} \in \operatorname{argmin}_{f \in \widehat{\mathcal{F}}_t(\pi^{(t)})} f_1(\rho, \pi_1^{(t)})$;
  update $\pi^{(t+1)}$ by the per-state Hedge step (5);

Remark 5 (Why the off-policy modulus is necessary)**.** *The term* $(\mathrm{II})_t$ *measures the residual of the backup operator* $\mathcal{T}^{\pi^{(t)}}$ *of the* current iterate *along the* comparator's *trajectories. The source data constrain* $f^{(t)}$ *only relative to* $\mathcal{T}^{\pi^{(t)}}$*, since that is how the confidence set* $\widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi^{(t)})$ *is built. Converting* $(\mathrm{II})_t$ *into a modulus therefore requires precisely the two-policy quantity* $\delta^{\pi^\star|\pi^{(t)}}$ *of Definition 1. Stating the guarantee with the on-policy comparator modulus* $\delta^{\pi^\star}$ *instead would implicitly require source Bellman consistency of* $f^{(t)}$ *with respect to* $\mathcal{T}^{\pi^\star}$*, which the algorithm never enforces. The max-min variant below restores the on-policy form by optimizing over a policy class directly.*

C.2.1 The max-min variant

Proposition 20 (Information-theoretic variant, restated from Proposition 18)**.** *Let* Π *be a finite policy class with* $\pi^\star_{\mathrm{tar}} \in \Pi$*, and define*

$$L(\pi) := \min_{f \in \widehat{\mathcal{F}}_{\mathrm{src}}(\beta,\pi)} f_1(\rho, \pi_1), \qquad \widehat{\pi}_{\mathrm{mm}} \in \operatorname*{argmax}_{\pi \in \Pi} L(\pi).$$

On the event of Assumption 3,

$$\mathrm{SubOpt}_{\mathrm{tar}}(\widehat{\pi}_{\mathrm{mm}}) \le \sum_{h=1}^{H} \left[\sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}}(\epsilon_{\mathrm{off}})} + \sqrt{\delta_h^{\widehat{\pi}_{\mathrm{mm}}}(\epsilon_{\mathrm{off}})} \right].$$

Proof. On the event of Assumption 3, every confidence set $\widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi)$ with $\pi \in \Pi$ contains $Q^\pi_{\mathcal{M}_{\mathrm{src}}}$ and is therefore nonempty, so L is well defined. Decompose

$$\mathrm{SubOpt}_{\mathrm{tar}}(\widehat{\pi}_{\mathrm{mm}}) = \big[V_1^{\pi^\star}(\rho) - L(\pi^\star)\big] + \big[L(\pi^\star) - L(\widehat{\pi}_{\mathrm{mm}})\big] + \big[L(\widehat{\pi}_{\mathrm{mm}}) - V_1^{\widehat{\pi}_{\mathrm{mm}}}(\rho)\big].$$

The middle term is nonpositive by the definition of $\widehat{\pi}_{\mathrm{mm}}$. For the first term, let $\widetilde{f}$ attain $L(\pi^\star)$; then $\widetilde{f} \in \mathcal{F}^{\mu,\pi^\star}_{\mathrm{src}}(\epsilon_{\mathrm{off}})$, and Lemma 10 gives $V_1^{\pi^\star}(\rho) - L(\pi^\star) = V_1^{\pi^\star}(\rho) - \widetilde{f}_1(\rho, \pi_1^\star) \le \sum_h \sqrt{\delta_h^{\pi^\star}(\epsilon_{\mathrm{off}})}$. For the third term, let $\widehat{f}$ attain $L(\widehat{\pi}_{\mathrm{mm}})$; the same lemma with $\pi = \widehat{\pi}_{\mathrm{mm}}$ gives $|\widehat{f}_1(\rho, \widehat{\pi}_{\mathrm{mm},1}) - V_1^{\widehat{\pi}_{\mathrm{mm}}}(\rho)| \le \sum_h \sqrt{\delta_h^{\widehat{\pi}_{\mathrm{mm}}}(\epsilon_{\mathrm{off}})}$. Combining the three estimates proves the claim. □

D Online Learning, Certification, and Handoff

D.1 Source-constrained online learning and warm starts

The online procedure and its intersection confidence set. The suboptimality guarantee in Theorem 1 measures the quality of the source-informed mixture policy $\widehat{\pi}$ produced from offline data. We keep two offline outputs distinct. The episode-level mixture $\widehat{\pi}$ is the object the offline theorem evaluates, and it is not, in general, a per-state Markov policy; separately, any full-support *final Hedge iterate* of Algorithm 2 is a Markov policy that may initialize the online learner, which is the warm start Theorem 22 covers. In this section, we analyze the regret of the online phase, in which the learner interacts with $\mathcal{M}_{\mathrm{tar}}$ while constraining its critics by the source data. Theorem 21 treats the source-constrained learner itself; Theorem 22 then quantifies the additional effect of warm-starting the policy from a final offline iterate.

The online phase follows the same high-level principle as the offline procedure, but now target data are incorporated as they arrive. At round t, the learner deploys the current policy $\pi^{(t)}$ in the target MDP, collects a trajectory, updates its target dataset, and then forms a critic class constrained by both the source confidence set and the accumulated target evidence. The source data provide a statistically efficient prior over plausible critics, while the target data progressively correct source-induced bias.

For a policy π, let $\widehat{\mathcal{E}}^{\pi,\pi}_{\mathrm{tar},t}(f)_h$ denote the corrected empirical Bellman error (17), with backup policy π, computed on the pooled target data collected up to episode t. Since the pooled data were generated by the past policies $\pi^{(1)}, \dots, \pi^{(t)}$, this quantity estimates the residual under the empirical mixture occupancy $\bar{d}_t = \frac{1}{t}\sum_{\tau \le t} d^{\pi^{(\tau)}}$, not under the current policy's own occupancy, and the constraint below is accordingly a mixture-weighted constraint. This suffices for the screening role the target set plays in this section, where the regret guarantee uses only source membership; *on-policy* control of target residuals is provided by the block construction of Algorithm 1 in Section 4.1. Given a target confidence radius α_t, define

$$\widehat{\mathcal{F}}_{\mathrm{tar},t}(\alpha_t, \pi) = \left\{ f \in \mathcal{F} : \widehat{\mathcal{E}}^{\pi,\pi}_{\mathrm{tar},t}(f)_h \le \alpha_t \text{ for all } h \in [H] \right\}. \tag{19}$$

The online learner uses the intersection

$$\widehat{\mathcal{F}}_t(\pi) = \begin{cases} \widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi) \cap \widehat{\mathcal{F}}_{\mathrm{tar},t}(\alpha_t, \pi) & \text{if this intersection is nonempty,} \\ \widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi) & \text{otherwise,} \end{cases} \tag{20}$$

matching the order of Algorithm 3, which collects the trajectory τ_t, updates the target dataset, and only then constructs the critic set, so the newly collected episode enters the target constraint without a one-episode delay. Thus, at the first round the target set is computed from a single episode and the critic set is dominated by the source-informed prior, and as target data accumulate, the critic set is refined using target evidence. Theorem 21 is unaffected by the indexing, since its proof uses only source-set membership. The fallback case is not vacuous. When the dynamics shift is large, no source-consistent critic need fit the target data, and the intersection can become empty; an empty intersection certifies empirical incompatibility between the tested source version space and the target residual constraint, at the current policy pair and radius. We note that the adaptive scheme of Section 4.2 operates on observed returns and does not consume this signal; the certificate is a complementary diagnostic.

The regret of Algorithm 3 is

$$\mathrm{Regret}(T) := \sum_{t=1}^{T} \left(V^{\pi^\star_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar}},1}(\rho) - V^{\pi^{(t)}}_{\mathcal{M}_{\mathrm{tar}},1}(\rho) \right). \tag{21}$$

This is the regret of the target-adaptive learner, not the regret of the internal offline optimization iterates.

Theorem 21 (Source-constrained online regret)**.** *Suppose $\mathcal{F}$ is finite and satisfies realizability and Bellman completeness. Let $\beta = CH^2\sqrt{T\log(T|\mathcal{F}|H/\delta)/n}$, $\alpha_t = CH^2\sqrt{\log(|\mathcal{F}|HT/\delta)/t}$, and $\eta = \frac{1}{H}\sqrt{8\log|\mathcal{A}|/T}$, and run the online procedure with uniform initialization. Then, with probability at least $1-\delta$,*

$$\mathrm{Regret}(T) \le \sum_{t=1}^{T}\sum_{h=1}^{H} \left[\sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}|\pi^{(t)}}(\epsilon_{\mathrm{off}})} + \sqrt{\delta_h^{\pi^{(t)}}(\epsilon_{\mathrm{off}})} \right] + H^2\sqrt{\frac{T\log|\mathcal{A}|}{2}}, \tag{22}$$

with $\epsilon_{\mathrm{off}} = \beta + CH^2\sqrt{T\log(T|\mathcal{F}|H/\delta)/n}$. Writing Γ_T for the average over t of the bracketed transfer terms, the bound reads $\mathrm{Regret}(T) \le T\Gamma_T + H^2\sqrt{T\log|\mathcal{A}|/2}$; the closure contributes entropy of order $T\log(T|\mathcal{F}|H/\delta)$, so its iteration dependence enters the radius outside the logarithm as well, and Lemma 16 permits the fast-scale radius $\beta = C_B H^2 T\log(T|\mathcal{F}|H/\delta)/n$.

Reading the bound. If $\Gamma_T \lesssim T^{-1/2}$, Theorem 21 gives a square-root regret bound, with the source sample size and class complexity entering through Γ_T. A constant transfer level makes this upper bound linear; it does not imply that the algorithm's realized regret is linear. The proof uses source-set membership and remains valid when the empirical intersection is empty. It does not exploit the target constraint to reduce its transfer terms. Section 4.1 uses blockwise target evidence for that purpose, and Section 4.2 compares learners through observed returns. Warm-starting changes the explicit optimization penalty and may also change the policies, occupancies, and realized transfer level (Theorem 22).

Discussion of Theorem 21. The choice of β ensures that the source empirical confidence set is contained in the population source-fitting set at radius ϵ_{off}, uniformly over the (data-dependent) policies encountered online; the linear-in-T factor multiplying the logarithm is the policy-closure entropy, the log-cardinality of the class of all possible data-dependent critic sequences, $\log|\overline{\Pi}| = O((T+T_{\mathrm{off}})\log(T|\mathcal{F}|)+\log|\mathcal{Q}_{\mathrm{base}}|)$, not a union over rounds, which would contribute only $\log T$. Two remarks are in order about the role of target data. First, the bound only uses membership of each critic in the *source* confidence set, which holds under both cases of (20); the theorem is therefore valid despite the possible emptiness of the intersection. Second, the target constraint can only shrink the critic set, so it improves the critics selected in practice, but it cannot improve the worst-case bound; relatedly, the displayed α_t is a screening threshold for the mixture-occupancy residual, not a proved on-policy confidence radius, and it plays no role in the bound. The moduli in (22) are suprema over the source-fitting set, and pessimism over on-policy target data provides no exploration with which to reduce them. Obtaining more requires different mechanisms; target evidence provably helps through the block construction of Section 4.1, and misleading sources are handled by the model selection of Section 4.2. For general function classes, the logarithmic factors above are replaced by the corresponding uniform convergence or localized complexity terms. The theorem is stated for uniform per-state initialization. Theorem 22 below quantifies the effect of warm-starting the policy from a source-informed initialization, replacing $\log|\mathcal{A}|$ by a divergence measured along the target-optimal trajectory distribution.

The first term in (22) is the comparator transfer cost. Source data must certify the residuals of the deployed policies' Bellman operators along trajectories of the target-optimal policy for the warm start to be reliable. The second term is the cumulative transfer cost incurred by the policies generated during online adaptation. The final term is the usual online optimization error. Thus, offline data improves the online phase only when source Bellman consistency transfers along both the comparator trajectories and the policies visited by the target-adaptive learner. Equivalently, define

$$\Gamma_T := \frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\left[\sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}|\pi^{(t)}}(\epsilon_{\mathrm{off}})}+\sqrt{\delta_h^{\pi^{(t)}}(\epsilon_{\mathrm{off}})}\right]. \tag{23}$$

Then (22) becomes

$$\mathrm{Regret}(T) \le T\Gamma_T + H^2\sqrt{\frac{T\log|\mathcal{A}|}{2}}. \tag{24}$$

For fixed remaining parameters, $\Gamma_T = O(T^{-1/2})$ gives a square-root regret upper bound. If $\Gamma_T = o(1)$, the bound gives vanishing average regret. If Γ_T remains bounded away from zero, this bound no longer certifies sublinear regret; the actual regret can still be smaller. Appendix D.2 proves the inequality.

Warm starts. The offline phase also produces a policy, and initializing at it is standard practice. With the initialization divergence

$$\bar{K}(\pi^{(1)}) := \sum_{h=1}^{H}\mathbb{E}_{s\sim d^{\pi^\star_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar},h}}}\left[\mathrm{KL}\big(\pi^\star_{\mathrm{tar},h}(\cdot|s)\,\big\|\,\pi_h^{(1)}(\cdot|s)\big)\right], \tag{25}$$

the next result quantifies the effect of a warm start.

Theorem 22 (Warm-started regret)**.** *Under the assumptions and radii of Theorem 21, initialize Algorithm 3 at any full-support $\pi^{(1)}$ in the initialization slice $\overline{\Pi}_{\mathrm{init}}$ of the closure, in particular at any final offline iterate, and run it with $\eta=\sqrt{8D/(H^3T)}$ for a known $D\ge\bar{K}(\pi^{(1)})$ chosen from a predetermined grid. Then, with probability at least $1-\delta$, $\mathrm{Regret}(T)\le T\Gamma_T+\sqrt{DH^3T/2}$. For the uniform initialization $D=H\log|\mathcal{A}|$ recovers Theorem 21; for the final offline iterate of Algorithm 2 the explicit bound $D=H\log|\mathcal{A}|+H\sqrt{8T_{\mathrm{off}}\log|\mathcal{A}|}$ is valid, and mixing any initialization with the uniform policy at weight $1/T$ makes $D=H\log(T|\mathcal{A}|)$ valid in general.*

Reading the bound. The initialization appears explicitly through a known bound $D\ge\bar{K}(\pi^{(1)})$ in the optimization term $\sqrt{DH^3T/2}$. The transfer term has the same formula as before, but its value can change because initialization changes the subsequent policy and critic sequence. A warm start helps this explicit penalty when a smaller valid D is available. Mixing provides the bound $D=H\log(T|\mathcal{A}|)$, which controls a poor initialization at an additional logarithmic cost relative

to the uniform value $H \log |\mathcal{A}|$. Source-trained value suboptimality alone does not imply a small initialization divergence (Appendix D.2).

When $\pi^\star_{\mathrm{tar}}$ is deterministic, $\bar{K}(\pi^{(1)}) = \sum_h \mathbb{E}_{s\sim d_h^{\pi^\star}}[\log(1/\pi_h^{(1)}(a_h^\star(s)|s))]$ is the cumulative log-loss of the initialization on the optimal actions along the optimal trajectory distribution. If the initialization places probability at least $1-\alpha$ on the optimal action at the states $\pi^\star_{\mathrm{tar}}$ visits, with $\alpha \le 1/2$, then $\bar{K} \le 2\alpha H$ and the optimization term improves from $H^2\sqrt{T\log|\mathcal{A}|/2}$ to $H^2\sqrt{\alpha T}$. We make three remarks. First, warm-starting from the final offline iterate is admissible in the two senses required. Every offline Hedge iterate obeys the pointwise lower bound $\pi_h^{(T_{\mathrm{off}}+1)}(a|s) \ge |\mathcal{A}|^{-1}e^{-\eta_{\mathrm{off}}HT_{\mathrm{off}}}$, since the cumulative softmax exponents differ by at most $\eta_{\mathrm{off}}HT_{\mathrm{off}}$ across actions, so $\bar{K}(\pi^{(T_{\mathrm{off}}+1)}) \le H\log|\mathcal{A}| + \eta_{\mathrm{off}}H^2T_{\mathrm{off}}$, which at the rate of Theorem 1 equals $H\log|\mathcal{A}| + H\sqrt{8T_{\mathrm{off}}\log|\mathcal{A}|}$, an explicit finite D known before the data are seen; full support alone would not suffice on a general state space, where pointwise positivity does not imply integrability of $\log(1/\pi)$. Furthermore, the algorithmic policy closure $\overline{\Pi}$, fixed *before* the source data are observed and containing every possible offline iterate and all its online descendants, places the realized warm start inside the uniform validity event; it is this membership, not full support, that licenses the concentration argument for a data-dependent initialization (the data select a member of the predetermined class, they do not determine the class). Moreover, any candidate π_0 whose mixed version has been inserted into $\mathcal{Q}_{\mathrm{base}}$ before the data, and any member of the initialization slice $\overline{\Pi}_{\mathrm{init}}$ selected after the data, mixed or not (a further mixing and the T online updates are still permitted, since the closure allows two mixings), can be mixed as $(1-\frac{1}{T})\pi_0 + \frac{1}{T}\,\mathrm{Unif}(\mathcal{A})$, which guarantees $\bar{K} \le H\log(T|\mathcal{A}|)$ in all cases while increasing a finite $\bar{K}(\pi_0)$ by at most $2H/T$ for $T \ge 2$; mixing fixes full support and the divergence bound, not class membership, which must hold a priori. Second, the guarantee separates the two channels through which offline data helps online. Initialization appears explicitly in the optimization term and can also change the realized transfer terms through the subsequent policies and critics. Theorem 1 controls the value gap of $\widehat{\pi}$, not its action agreement with $\pi^\star_{\mathrm{tar}}$ along $d^{\pi^\star}$, and a near-optimal policy can still disagree with $\pi^\star_{\mathrm{tar}}$ action by action, so the theorem quantifies when warm-starting helps rather than presuming that it does. Third, for the uniform initialization the choice $D = H\log|\mathcal{A}|$ is valid and recovers Theorem 21 exactly, and for the offline warm start the explicit bound above gives $D = H\log|\mathcal{A}| + H\sqrt{8T_{\mathrm{off}}\log|\mathcal{A}|}$; for an arbitrary full-support initialization no finite prior bound need exist, and the practical recipe is the mixing of Remark 7, under which $D = H\log(T|\mathcal{A}|)$ is always valid. The proof is in Appendix D.2, and Appendix H.2 ablates the two channels empirically (Figure 15), with different observed effects of initialization in the aligned and shifted instances; these observations do not establish that initialization leaves the transfer term invariant.

D.2 Proof of the online regret bounds

We prove the regret guarantee for the sequence of policies $\pi^{(1)}, \dots, \pi^{(T)}$ deployed by Algorithm 3.

Theorem 23 (Regret bound, detailed version)**.** *Suppose Assumption 3 holds with radius ϵ_{off} for all policies deployed by Algorithm 3, and run the algorithm with uniform initialization and $\eta = \frac{1}{H}\sqrt{\frac{8\log|\mathcal{A}|}{T}}$. Then, on that event,*

$$\mathrm{Regret}(T) \le \sum_{t=1}^{T}\sum_{h=1}^{H}\left[\sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}|\pi^{(t)}}(\epsilon_{\mathrm{off}})} + \sqrt{\delta_h^{\pi^{(t)}}(\epsilon_{\mathrm{off}})}\right] + H^2\sqrt{\frac{T\log|\mathcal{A}|}{2}}.$$

Proof. By definition,

$$\mathrm{Regret}(T) = \sum_{t=1}^{T}\left(V^{\pi^\star_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar}},1}(\rho) - V^{\pi^{(t)}}_{\mathcal{M}_{\mathrm{tar}},1}(\rho)\right),$$

and we bound each summand by the decomposition (18) of Lemma 14, applied with comparator $\pi^\star_{\mathrm{tar}}$, policy $\pi^{(t)}$, and critic $f^{(t)}$.

The key observation is that the proof only uses membership of $f^{(t)}$ in the *source* confidence set. Under both cases of the selection rule (20) we have $f^{(t)} \in \widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi^{(t)})$, hence $f^{(t)} \in \mathcal{F}^{\mu,\pi^{(t)}}_{\mathrm{src}}(\epsilon_{\mathrm{off}})$ on the validity event; the target-side constraint, when active, only shrinks the set from which $f^{(t)}$

is chosen and affects no step of the argument. (Nonemptiness of the fallback set is guaranteed by Assumption 3, which places $Q^{\pi^{(t)}}_{\mathcal{M}_{\mathrm{src}}}$ in $\widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi^{(t)})$.)

Exactly as in the proof of Theorem 19, the comparator residual term at round t is bounded by $\sum_h \sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}|\pi^{(t)}}(\epsilon_{\mathrm{off}})}$, and the iterate residual term by $\sum_h \sqrt{\delta_h^{\pi^{(t)}}(\epsilon_{\mathrm{off}})}$; summing over t gives the first two terms of the theorem. The per-state Hedge terms are handled directly by the uniform-initialization case of Lemma 12, since Algorithm 3 initializes the per-state distributions uniformly here and applies the same exponential-weights recursion, so that

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{s\sim d^{\pi^\star_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar},h}}} \left[\left\langle f_h^{(t)}(s,\cdot), \pi^\star_{\mathrm{tar},h}(\cdot|s) - \pi_h^{(t)}(\cdot|s)\right\rangle\right] \le H^2\sqrt{\frac{T\log|\mathcal{A}|}{2}}.$$

Combining the three bounds proves the claim. □

Remark 6 (Role of the target data)**.** *Theorem 23 does not improve when the target constraint is active, and this is not an artifact. The moduli are suprema over the source-fitting set, the algorithm performs no exploration, and pessimistic selection from on-policy data cannot reduce a worst-case supremum. The benefit of the target constraint is instance-wise (it removes critics already falsified by target evidence) and diagnostic (an empty intersection certifies empirical incompatibility of the source confidence set with the target test, at the current policy pair and radius). Theorem 5 provides a guarantee via return-based model selection; it does not use the diagnostic, which remains a complementary signal.*

Substituting the definition of Γ_T from (22) into Theorem 23 yields the summary form (22) stated in the main text; in particular, $\Gamma_T = O(H^2\sqrt{\log|\mathcal{A}|/T})$ implies $\mathrm{Regret}(T) = O(H^2\sqrt{T\log|\mathcal{A}|})$.

D.2.1 Warm-started initialization

Proof of Theorem 22. The proof of Theorem 23 uses the initialization only through the per-state Hedge terms; the comparator and iterate residual terms depend on critic membership in the source-fitting sets alone and are unchanged. For each (s,h), Lemma 12 with initialization $q = \pi_h^{(1)}(\cdot|s)$ (full support by assumption) and comparator $p^\star = \pi^\star_{\mathrm{tar},h}(\cdot|s)$ gives, deterministically,

$$\sum_{t=1}^{T} \left\langle f_h^{(t)}(s,\cdot),\, \pi^\star_{\mathrm{tar},h}(\cdot|s) - \pi_h^{(t)}(\cdot|s)\right\rangle \le \frac{\mathrm{KL}\big(\pi^\star_{\mathrm{tar},h}(\cdot|s)\,\big\|\,\pi_h^{(1)}(\cdot|s)\big)}{\eta} + \frac{\eta H^2 T}{8}.$$

Taking expectations over $s \sim d^{\pi^\star_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar},h}}$, summing over $h \in [H]$, and applying the definition (25) bounds the total optimization contribution by

$$\frac{\bar K(\pi^{(1)})}{\eta} + \frac{\eta H^3 T}{8} \le \frac{D}{\eta} + \frac{\eta H^3 T}{8} = \sqrt{\frac{DH^3T}{2}}$$

for $\eta = \sqrt{8D/(H^3T)}$, using $\bar K(\pi^{(1)}) \le D$ in the inequality. Combining with the comparator and iterate residual bounds exactly as in the proof of Theorem 23 proves the claim. For uniform initialization, $\mathrm{KL}(\pi^\star_{\mathrm{tar},h}(\cdot|s)\,\|\,\mathrm{Unif}(\mathcal{A})) \le \log|\mathcal{A}|$ pointwise, so $\bar K \le H\log|\mathcal{A}|$; with $D = H\log|\mathcal{A}|$ the optimization term is $\sqrt{H\log|\mathcal{A}|\cdot H^3T/2} = H^2\sqrt{T\log|\mathcal{A}|/2}$ and the learning rate $\eta = \sqrt{8H\log|\mathcal{A}|/(H^3T)} = \frac{1}{H}\sqrt{8\log|\mathcal{A}|/T}$ coincides with that of Theorem 23, so the two theorems agree. □

Remark 7 (Mixing for admissibility)**.** *If a candidate initialization π_0, fixed before the data (with its mixed version inserted into $\mathcal{Q}_{\mathrm{base}}$, making it a length-zero member of the initialization slice) or any member of $\overline{\Pi}_{\mathrm{init}}$ selected after the data, mixed or not (a further mixing and the online updates are still permitted, since the closure allows two mixings, so the mixed version and all its later updates remain members), lacks full support, replace it by $\pi^{(1)} = (1-\frac{1}{T})\pi_0 + \frac{1}{T}\,\mathrm{Unif}(\mathcal{A})$; mixing repairs support and the divergence bound only, not the a-priori class membership that the concentration requires. Then $\pi_h^{(1)}(a|s) \ge \frac{1}{T|\mathcal{A}|}$ for all (s,a,h), so $\bar K(\pi^{(1)}) \le H\log(T|\mathcal{A}|)$ always, and $\pi_h^{(1)}(a|s) \ge (1-\frac{1}{T})\pi_{0,h}(a|s)$ gives $\bar K(\pi^{(1)}) \le \bar K(\pi_0) + H\log\frac{1}{1-1/T} \le \bar K(\pi_0) + \frac{2H}{T}$ for $T \ge 2$ whenever $\bar K(\pi_0)$ is finite.*

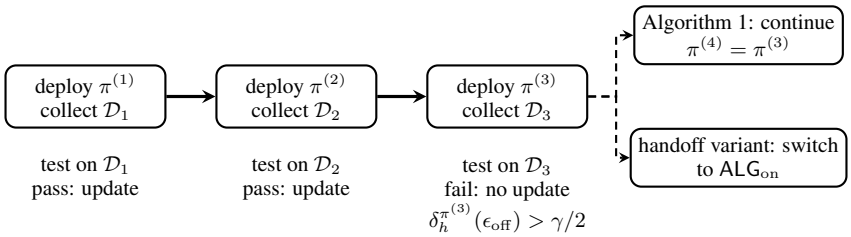


Figure 3: Same-block certification, with block $3 < K$ the first fallback. Each tested block evaluates its critic on its own on-policy data. The two outgoing branches are alternative procedures: Algorithm 1 freezes the policy, while Theorem 4 starts the online learner immediately after block 3, with no additional frozen-policy block. The displayed certificate uses the Hoeffding calibration. The final block of Algorithm 1 is deployed without a test or update and receives its trivial regret charge.

D.3 Certification, handoff, and estimation of the modulus

Discussion of the strong modulus and certification. The strong modulus asks a sharper question than the weak one. Among critics that fit the source data *and* have small target Bellman residual along the trajectories of a reference policy ν', how large can the residual still be along ν? It is monotone in both radii, never exceeds the weak modulus, and satisfies $\widetilde{\delta}_h^{\nu|\pi}(\epsilon;\gamma,\nu') \le C\gamma$ whenever $d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h} \le C\, d^{\nu'}_{\mathcal{M}_{\mathrm{tar}},h}$ pointwise (Lemma 25). The second property is the lever. If the evaluation policy stays close to the reference policy that generated the target data, the target constraint alone controls the residual, and the source-side ambiguity becomes irrelevant.

To exploit this lever we modify the online algorithm in one respect. Policies are held fixed over blocks of w consecutive episodes, so that each block produces on-policy data for a single policy, and the critic that drives the update out of block k is required to fit block k's *own* data, a same-block certification. Freezing the policy makes this possible. The deployed policy $\pi^{(k)}$ is determined before its block's data are collected, so the conditional concentration is valid; the concentration is uniform over critics, so selecting the critic after the data are observed is legitimate; and because the certification is directly on-policy, no argument about the closeness of consecutive policies is needed anywhere in the analysis. Figure 3 depicts the block mechanics, including the update suppression on fallback and the certified exit of Theorem 4.

Theorem 2 is a term-wise sharpening of Theorem 21 in two respects, at the price of a larger optimization term and with the comparison understood along a common policy sequence, since the two algorithms deploy different sequences. First, the iterate transfer cost is *decoupled from the shift*. Whatever the weak modulus of the deployed policies, their residual term is capped by the target-data radius $\bar{\gamma}_k = \widetilde{O}(H^2 w^{-1/2})$, which vanishes as blocks grow; summed over the horizon this contributes only the $\widetilde{O}(H^2 T^{5/6})$ term. The cap is a direct consequence of same-block certification, and it carries no drift factor, since the certified residual is measured on the deployed policy's own data. Second, the comparator cost is measured by the strong modulus rather than the weak one; for the same block-policy sequence, every strong-modulus comparator term is no larger than its corresponding weak-modulus term, by Lemma 25. The gap can be large. If target evidence along the deployed policies pins down the relevant Bellman backups (for instance, in a block MDP where the iterates visit every block that $\pi^\star_{\mathrm{tar}}$ visits), the strong modulus collapses to the target radius while the weak modulus remains as large as the dynamics shift. When the fallback does trigger, the affected blocks certify $\delta_h^{\pi^{(k)}}(\epsilon_{\mathrm{off}}) > \gamma - c(w)$ for some h, an *on-policy* certificate along the deployed policy itself (Remark 9), so the source version space fails to certify the deployed policy at the tested radius; those blocks are charged their worst-case per-block regret, and the certificate additionally *suppresses* the source-driven update, so the policy is carried unchanged into the next block and every source-driven update, including the first, is target-certified. The proof is given in Appendix D.4.

The $T^{5/6}$ rate is not the limit of the mechanism. The block radius above is calibrated at the Hoeffding scale $\widetilde{O}(H^2/\sqrt{w})$, but the corrected squared-residual estimator concentrates at the fast Bernstein scale, since its excess-loss variance is controlled by its mean under Bellman completeness. Certifying

the radius $\widetilde{O}(H^2/w)$ from w on-policy episodes lets shorter blocks do the same work, and rebalancing the block length against the optimization cost improves the rate.

The mechanism experiment in Appendix H.2 (Figure 10) illustrates the value of target constraints with a heuristic critic set; it does not establish the $1/m$ rate. The proof, a fast-rate version of the block concentration lemma followed by rebalancing, is in Appendix D.4.

The certificate can do more than suppress a single update. Because a fallback is a data-driven statement that the deployed policy's own weak modulus is large, it can trigger a permanent exit to purely online learning, and this exit carries a guarantee with no model-selection factor.

Theorem 4 resolves, in its basic form, the certificate-triggered handoff question raised by the ablation of Appendix H.2. It complements rather than replaces the model selection of Theorem 5. The handoff carries no factor of M and no elimination overhead, and it exits when the data certify that the deployed policy's source information has failed. It cannot detect a failure on the comparator side, where a large comparator strong-modulus term coexists with small on-policy residuals; the return-based master addresses that case. The exit is also forced when the source is incompatible along the deployed policy. On the same events, every accepted critic lies in $\mathcal{F}_{\mathrm{src}}^{\mu,\pi^{(k)}}(\epsilon_{\mathrm{off}})$ and has target residual at most $\bar{\gamma}$ at every stage along $\pi^{(k)}$, so a fallback occurs whenever every critic in $\mathcal{F}_{\mathrm{src}}^{\mu,\pi^{(k)}}(\epsilon_{\mathrm{off}})$ has target residual above $\bar{\gamma}$ at some stage. Together with the conservative direction stated in the theorem, the exit is gated between the thresholds $\gamma/2$ and $\bar{\gamma}$. The handoff arm of Figure 18 is a data-reusing blockwise handoff that illustrates, but does not instantiate, this theorem.

Certifying transfer from data. The moduli in the theorems above are analysis quantities, and no algorithm needs them to run. For diagnostics, however, the modulus of a fixed candidate policy can itself be estimated, given an additional *version-space maximization* oracle returning $\sup_{f\in\widehat{\mathcal{F}}_{\mathrm{src}}(\beta,\pi)} \widehat{\mathcal{E}}_{\mathrm{tar}}^{\pi,\pi}(f)_h$. This maximization is a different computation from the pessimistic minimization the algorithms use, and it need not be efficient in general, consistent with the supervised moduli framework, where the modulus need not be efficiently estimable (Hanneke & Kpotufe, 2024). In tabular MDPs, robust pessimistic and optimistic policy evaluation over per-pair TV balls gives a model-based surrogate for this supremum, which is the implementation used in the model-based illustrations of Appendix H.2; the corrected-residual confidence set itself is a count-weighted quadratic constraint at each stage rather than a product of per-pair balls, and block MDPs reduce to the same computation at block granularity.

Proposition 24 (Plug-in estimation of the weak modulus)**.** *Fix a policy π chosen independently of both datasets and inserted into Π_{fixed} before either dataset is observed, so that the globally scoped assumptions apply to it, and suppose $\mathcal{F}$ is finite and Bellman complete for both environments. Given n source episodes from μ and m on-policy target episodes from π, define*

$$\widehat{\delta}_h := \sup_{f\in\widehat{\mathcal{F}}_{\mathrm{src}}(\beta,\pi)} \widehat{\mathcal{E}}_{\mathrm{tar}}^{\pi,\pi}(f)_h,$$

with both empirical residuals given by the corrected estimator (17)*. Let $c_n = CH^2\sqrt{\log(|\mathcal{F}|H/\delta)/n}$ and $c_m = CH^2\sqrt{\log(|\mathcal{F}|H/\delta)/m}$, and take $\beta \geq c_n$. Then, with probability at least $1-\delta$, for every h,*

$$\delta_h^{\pi}(\beta - c_n) - c_m \;\leq\; \widehat{\delta}_h \;\leq\; \delta_h^{\pi}(\beta + c_n) + c_m.$$

A small $\widehat{\delta}_h$ certifies a small on-policy residual over the smaller-radius source set, up to the target estimation error, but this alone does not certify the policy's suboptimality or deployment safety; a large value exhibits a source-plausible critic with large on-policy target residual, so the version space does not certify π (it does not mean every source-consistent critic is falsified). The price of certification is evaluation of the modulus at slightly perturbed radii, which is intrinsic to plug-in functionals. By the self-bounding variance argument (Lemma 16, and Lemma 27 for the on-policy target episodes), the radii c_n and c_m may be taken at the Bernstein scale $\widetilde{O}(H^2/n)$ and $\widetilde{O}(H^2/m)$, at the price of the factor-of-two multiplicative form of those lemmas in place of the additive perturbations displayed. If the candidate policy was itself learned from the same source data, the guarantee requires either splitting the source sample or a union bound over the policy class that produced it, as in Appendix B.4; without one of these the diagnostic is anti-conservative. The proof is in Appendix B.4, and Figure 14 in Appendix H.2 compares the estimate with a zero-radius witness proxy across a shift sweep; those

curves are not bounds on the modulus. Uniform estimation over policy classes without deploying each candidate runs into off-policy evaluation barriers and remains open.

D.4 Proofs for certification and handoff

Throughout this section, all values, visitation distributions, and Bellman errors refer to the target MDP unless stated otherwise, and we write $\pi^\star := \pi^\star_{\mathrm{tar}}$.

D.4.1 Properties of the strong modulus

Lemma 25 (Basic properties)**.** *For all policies ν, π, ν', step h, and radii $\epsilon, \gamma \geq 0$, the following hold.*

1. *$\widetilde{\delta}_h^{\nu|\pi}(\epsilon; \gamma, \nu')$ is nondecreasing in ϵ and in γ, and $\widetilde{\delta}_h^{\nu|\pi}(\epsilon; \gamma, \nu') \leq \delta_h^{\nu|\pi}(\epsilon)$.*
2. *If $d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h}(s,a) \leq C\, d^{\nu'}_{\mathcal{M}_{\mathrm{tar}},h}(s,a)$ for all (s,a), then $\widetilde{\delta}_h^{\nu|\pi}(\epsilon; \gamma, \nu') \leq C\gamma$. Consequently $\widetilde{\delta}_h^{\nu|\pi}(\epsilon; \gamma, \nu') \leq \min\left\{\delta_h^{\nu|\pi}(\epsilon),\, C\gamma\right\}$ under the same domination.*

Proof. (1) Both feasible sets are nested in the corresponding radius, and dropping the target constraint ($\gamma = \infty$) recovers the weak modulus. (2) For any feasible f, the squared residual $g(s,a) = (f_h - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{tar}},h} f_{h+1})^2(s,a)$ is nonnegative, so $\mathcal{E}^{\nu,\pi}_{\mathcal{M}_{\mathrm{tar}}}(f)_h = \mathbb{E}_{d^{\nu}_h}[g] \leq C\, \mathbb{E}_{d^{\nu'}_h}[g] = C\, \mathcal{E}^{\nu',\pi}_{\mathcal{M}_{\mathrm{tar}}}(f)_h \leq C\gamma$. □

D.4.2 Block concentration

Same-block certification removes any need for a slow-drift comparison between consecutive policies; the certified residual is measured on the deployed policy's own data, and the following conditional concentration is all the analysis uses.

Lemma 26 (Block concentration)**.** *Suppose $\mathcal{F}$ is finite and Bellman complete for $\mathcal{M}_{\mathrm{tar}}$. For every tested block $k < K$, the policy $\pi^{(k)}$ is fixed before collecting $\mathcal{D}_k$. Conditional on the preceding history, this dataset consists of w independent episodes from that policy. With probability at least $1-\delta$, simultaneously for all tested blocks $k < K$, stages h, and critics $f \in \mathcal{F}$,*

$$\left|\widehat{\mathcal{E}}^{(k)}(f; \pi^{(k)})_h - \mathcal{E}^{\pi^{(k)},\pi^{(k)}}_{\mathcal{M}_{\mathrm{tar}}}(f)_h\right| \leq c(w) := CH^2\sqrt{\frac{\log(|\mathcal{F}|HK/\delta)}{w}}.$$

In particular, under realizability and on this event, for every tested block $k < K$, $Q^{\pi^{(k)}}_{\mathcal{M}_{\mathrm{tar}}}$ satisfies the target constraint at radius γ whenever $\gamma \geq c(w)$, and every f passing the constraint satisfies the on-policy *bound $\mathcal{E}^{\pi^{(k)},\pi^{(k)}}_{\mathcal{M}_{\mathrm{tar}}}(f)_h \leq \gamma + c(w)$.*

Proof. The policy $\pi^{(k)}$ is measurable with respect to $\sigma(\mathcal{D}_{\mathrm{src}}, \mathcal{D}_1, \ldots, \mathcal{D}_{k-1})$, depending recursively on the entire history through the sequence of certified updates and freezes; the data $\mathcal{D}_k$ is collected afterwards. Conditioned on the past, $\mathcal{D}_k$ is therefore an i.i.d. sample of w_k episodes from the fixed policy $\pi^{(k)}$, and the residual class is indexed by $f \in \mathcal{F}$ with both the evaluation distribution and the backup policy fixed and equal to $\pi^{(k)}$; the concentration is uniform over f, so selecting the critic after observing $\mathcal{D}_k$ is legitimate. The claim now follows exactly as in Lemma 15, using the corrected estimator (17) on $\mathcal{D}_k$, Bellman completeness for $\mathcal{M}_{\mathrm{tar}}$, and a union bound over $f \in \mathcal{F}$, $h \in [H]$, and the tested indices $k < K$, all of which have $w_k = w$. No concentration claim at radius $c(w)$ is used for the final block. The membership of $Q^{\pi^{(k)}}_{\mathcal{M}_{\mathrm{tar}}}$ follows because its population residual with backup $\pi^{(k)}$ vanishes identically, at every evaluation distribution. □

D.4.3 Proof of Theorem 2

Proof. Work on the intersection of the source validity event (Assumption 3 with radius ϵ_{off}, probability $1-\delta$) and the block concentration event of Lemma 26 (probability $1-\delta$). Within block k, of actual

length $w_k \le w$ (only the last block may be shorter), the deployed policy is $\pi^{(k)}$, so $\mathrm{Regret}(T) = \sum_k w_k(V_1^{\pi^\star}(\rho) - V_1^{\pi^{(k)}}(\rho))$, and since every summand is nonnegative,

$$\mathrm{Regret}(T) \le \sum_{k=1}^{K} w \left(V_1^{\pi^\star}(\rho) - V_1^{\pi^{(k)}}(\rho) \right),$$

Let $\mathcal{C} := \{k < K : k \notin \mathcal{B}\}$ be the certified blocks. Each fallback block contributes at most wH regret and performs no update. The final block is not tested, supplies no critic or update, and contributes at most $w_K H \le wH$. These disjoint charges sum to at most $wH(N_{\mathrm{fb}} + 1)$. In every block $k \in \mathcal{C}$, apply Lemma 14 with comparator $\pi^\star$, deployed policy $\pi^{(k)}$, and the critic selected after observing $\mathcal{D}_k$. This deterministic identity remains valid for that data-dependent critic.

Comparator residual. In every certified block, $f^{(k)} \in \widehat{\mathcal{F}}_{\mathrm{src}}(\beta, \pi^{(k)}) \subseteq \mathcal{F}_{\mathrm{src}}^{\mu,\pi^{(k)}}(\epsilon_{\mathrm{off}})$, and $f^{(k)}$ passes the same-block target constraint, so by Lemma 26 and the choice of γ, $\mathcal{E}_{\mathcal{M}_{\mathrm{tar}}}^{\pi^{(k)},\pi^{(k)}}(f^{(k)})_{h'} \le \gamma + c(w) \le 2\gamma = \bar{\alpha}(w) = \bar{\gamma}_k$ for all h'. Hence in every certified block, including block 1, the critic $f^{(k)}$ is feasible for the program defining $\widetilde{\delta}_h^{\pi^\star|\pi^{(k)}}(\epsilon_{\mathrm{off}}; \bar{\gamma}_k, \pi^{(k)})$, whose conditioning policy is the deployed policy itself, and the comparator residual term is bounded by $\sum_h (\widetilde{\delta}_h^{\pi^\star|\pi^{(k)}}(\epsilon_{\mathrm{off}}; \bar{\gamma}_k, \pi^{(k)}))^{1/2}$.

Iterate residual. In every certified block, $\mathcal{E}_{\mathcal{M}_{\mathrm{tar}}}^{\pi^{(k)},\pi^{(k)}}(f^{(k)})_h \le \delta_h^{\pi^{(k)}}(\epsilon_{\mathrm{off}})$ by source membership, and the same-block certification bound above gives $\mathcal{E}_{\mathcal{M}_{\mathrm{tar}}}^{\pi^{(k)},\pi^{(k)}}(f^{(k)})_h \le \bar{\gamma}_k$ *directly*, the certified residual being measured on the deployed policy's own data; no slow-drift argument and no condition on K are needed. This yields the $\min$ in (9).

Optimization term. Only certified blocks contribute critic gains and Hedge updates. Fallback blocks freeze the policy, and the terminal block supplies no gain. Thus the policies on $\mathcal{C}$ form exactly the Hedge sequence driven by $(f^{(k)})_{k\in\mathcal{C}}$, with $m := |\mathcal{C}| = K - 1 - N_{\mathrm{fb}} \le K$. At the fixed rate η_K, Lemma 12 gives, for every state and stage,

$$\sum_{k\in\mathcal{C}} \langle f_h^{(k)}(s,\cdot), \pi_h^\star(\cdot \mid s) - \pi_h^{(k)}(\cdot \mid s) \rangle \le \frac{\log|\mathcal{A}|}{\eta_K} + \frac{\eta_K H^2 m}{8} \le H\sqrt{K \log|\mathcal{A}|/2}.$$

No term of unknown sign is removed from this Hedge sum: the sum is defined on precisely the certified subsequence. Integrating against the comparator's stagewise occupancies, summing over stages, and multiplying by w bounds the optimization contribution by

$$wH^2\sqrt{K\log|\mathcal{A}|/2} \le H^2\sqrt{wT\log|\mathcal{A}|},$$

because $w^2 K \le w(T+w) \le 2wT$ for $1 \le w \le T$. The convention also covers $K = 1$: there are no tested blocks and the terminal charge alone bounds regret.

Combining the certified-block bounds with the fallback and final-block charges proves (9). For the simplified bound of Theorem 2, take $w = \lceil T^{2/3} \rceil$. If no fallback triggers, the iterate terms over certified blocks are bounded by $\sum_{k<K} w \sum_h \sqrt{\bar{\alpha}(w)} \le C'TH \cdot H\big(\log(|\mathcal{F}|HT/\delta)/w\big)^{1/4} = \widetilde{O}(H^2T^{5/6})$, and the final block contributes $wH \le 2HT^{2/3}$. The optimization term is $H^2\sqrt{wT\log|\mathcal{A}|} = \widetilde{O}(H^2T^{5/6})$. Finally, the comparator terms sum to at most $wK\widetilde{\Gamma}_T \le (T+w)\widetilde{\Gamma}_T \le T\widetilde{\Gamma}_T + 2H^2 w$, using $\widetilde{\Gamma}_T \le 2H^2$; all the extra $O(H^2T^{2/3})$ terms are absorbed into the $\widetilde{O}(H^2T^{5/6})$ slack. This proves the claim. □

D.4.4 Fast-rate refinement, fallback certificates, and handoff

Lemma 27 (Fast-rate block concentration)**.** *In the setting of Lemma 26, let $\iota := \log(|\mathcal{F}|HK/\delta)$ and $\gamma_{\mathrm{fast}}(w) := C_f H^2 \iota / w$. With probability at least $1-\delta$, simultaneously for all tested blocks $k < K$, steps h, and $f \in \mathcal{F}$,*

$$\tfrac{1}{2}\,\mathcal{E}_{\mathcal{M}_{\mathrm{tar}}}^{\pi^{(k)},\pi^{(k)}}(f)_h - \gamma_{\mathrm{fast}}(w) \;\le\; \widehat{\mathcal{E}}^{(k)}(f;\pi^{(k)})_h \;\le\; 2\,\mathcal{E}_{\mathcal{M}_{\mathrm{tar}}}^{\pi^{(k)},\pi^{(k)}}(f)_h + \gamma_{\mathrm{fast}}(w).$$

Proof. Fix (f,h,k), write $g^\star := \mathcal{T}_{\mathcal{M}_{\mathrm{tar}},h}^{\pi^{(k)}} f_{h+1} \in \mathcal{F}_h$ (completeness), and let $Z_i := \ell_h^i(f_h, f_{h+1}) - \ell_h^i(g^\star, f_{h+1})$ be the excess losses on $\mathcal{D}_k$, which are i.i.d. conditioned on the past exactly as in

Lemma 26. Conditionally on (s_i, a_i),

$$Z_i = (f_h - g^\star)(s_i,a_i)\big[(f_h - g^\star)(s_i,a_i) + 2\big(g^\star(s_i,a_i) - y_i\big)\big], \qquad \mathbb{E}[y_i \mid s_i, a_i] = g^\star(s_i,a_i),$$

so $\mathbb{E}[Z_i \mid s_i,a_i] = (f_h - g^\star)^2(s_i,a_i)$, $|Z_i| \le cH^2$, and $\mathrm{Var}(Z_i) \le \mathbb{E}[Z_i^2] \le cH^2\,\mathbb{E}\big[(f_h - g^\star)^2(s_i,a_i)\big] = cH^2\,\mathcal{E}_h(f)$, where $\mathcal{E}_h(f)$ abbreviates $\mathcal{E}^{\pi^{(k)},\pi^{(k)}}_{\mathcal{M}_{\mathrm{tar}}}(f)_h$. Bernstein's inequality with a union bound over (f,h,k) gives, with probability at least $1-\delta/2$,

$$\big|\bar{Z} - \mathcal{E}_h(f)\big| \le \sqrt{\frac{2cH^2\,\mathcal{E}_h(f)\,\iota}{w}} + \frac{cH^2\iota}{w} \le \tfrac{1}{2}\,\mathcal{E}_h(f) + \frac{c'H^2\iota}{w}$$

by the arithmetic-geometric mean inequality. The same Bernstein bound applied to the excess losses of every $g \in \mathcal{F}_h$ against $g^\star$, whose conditional means $(g-g^\star)^2$ are nonnegative, shows (on a further $1-\delta/2$ event) that the empirical minimum in (17) lies within $c'H^2\iota/w$ of the empirical loss of $g^\star$, from below as well as from above. Since $\widehat{\mathcal{E}}^{(k)}(f;\pi^{(k)})_h = \bar{Z} + \big[\widehat{L}(g^\star) - \min_g \widehat{L}(g)\big]$, combining the two displays yields the claim with $C_f = 4c'$. □

Remark 8 (Infinite classes at block granularity)**.** *Lemmas 26 and 27 extend to infinite critic classes without any induced policy-cover term. As recorded in Lemma 26, the deployed policy $\pi^{(k)}$ is determined before its block data $\mathcal{D}_k$ are collected, so only the critic class requires discretization. A sup-norm net at scale $\alpha = 1/(wH)$ perturbs every empirical and population squared residual, including the subtracted minimum of the corrected estimator, by $O(H\alpha) = O(1/w)$, which is absorbed by $\gamma_{\mathrm{fast}}(w)$ and a fortiori by the Hoeffding radius $c(w)$, so both lemmas hold with $\log|\mathcal{F}|$ replaced by $\log N_{1/(wH)}$. Combined with Proposition 17 for the source-side radii, Theorem 2 and Corollary 3 hold for the linear class with $\log N_{1/(wH)} = O(d\log(dHw))$ and for the block-constant class with $O(\kappa\log(Hw))$, giving end-to-end strong-modulus guarantees in both structured settings.*

Proof of Corollary 3. Repeat the proof of Theorem 2 with Lemma 27 in place of Lemma 26 and test radius $\gamma = 2\gamma_{\mathrm{fast}}(w)$. On the fast-rate event, $Q^{\pi^{(k)}}_{\mathcal{M}_{\mathrm{tar}}}$ passes the target constraint, since its population residual vanishes so its empirical residual is at most $\gamma_{\mathrm{fast}}(w) \le \gamma$, and every critic passing the constraint satisfies $\mathcal{E}_h(f) \le 2\big(\widehat{\mathcal{E}}_h(f) + \gamma_{\mathrm{fast}}(w)\big) \le 2\big(2\gamma_{\mathrm{fast}}(w) + \gamma_{\mathrm{fast}}(w)\big) = 6\gamma_{\mathrm{fast}}(w) = \bar{\gamma}_k$ by the left inequality of the lemma, applied to the same-block data $\mathcal{D}_k$; no drift argument is involved. The optimization accounting is unchanged under the fixed learning rate of Theorem 2. With $w = \lceil\sqrt{T}\rceil$, the iterate cost over certified blocks is at most

$$\sum_{k<K} w \sum_h \sqrt{\bar{\gamma}_k} \le 2T\cdot H \cdot 3H\sqrt{\frac{C_f\iota}{w}} = \widetilde{O}\big(H^2T^{3/4}\big),$$

the optimization cost is $2H^2\sqrt{wT\log|\mathcal{A}|} = \widetilde{O}(H^2T^{3/4})$, and the final block contributes $wH = \widetilde{O}(H\sqrt{T})$. Writing $\widetilde{\Gamma}^{\mathrm{fast}}_T = K^{-1}\sum_{k\in\mathcal{C}}\sum_h \sqrt{\widetilde{\delta}^{\pi^\star|\pi^{(k)}}_h(\epsilon_{\mathrm{off}}; 6\gamma_{\mathrm{fast}}(w), \pi^{(k)})}$, the comparator contribution is $wK\widetilde{\Gamma}^{\mathrm{fast}}_T \le T\widetilde{\Gamma}^{\mathrm{fast}}_T + 2H^2w$. The last term is absorbed by the stated remainder. For the unconditional display, each fallback block k is charged its trivial regret $\sum_{t\in\mathrm{block}\,k}(V_1^\star(\rho) - V_1^{\pi^{(k)}}(\rho)) \le wH \le 2H\sqrt{T}$, since values lie in $[0,H]$ and the per-block bound holds pathwise, and the value-difference decomposition is applied only to the certified blocks. Because Algorithm 1 performs no update on fallback blocks, the deployed policies on certified blocks form the per-state Hedge trajectory driven by the certified-block critics alone, so Lemma 12 applies verbatim to that subsequence and the decomposition drops no Hedge comparison term of indeterminate sign (this is where the suppressed update is used). Summing yields the additional term $2H\sqrt{T}\,N_{\mathrm{fb}}$, and $N_{\mathrm{fb}} = 0$ recovers the conditional display. □

Remark 9 (What the fallback certifies)**.** *Work on source validity and the appropriate block-concentration event, and fix a tested block $k < K$. The witness $q = Q^{\pi^{(k)}}_{\mathcal{M}_{\mathrm{src}}}$ belongs to both the empirical source confidence set and the population source-fitting set. If the empirical intersection is empty, this witness fails at least one target constraint. Under Hoeffding calibration, choosing the constant so that $\gamma \ge 2c(w)$ gives*

$$\mathcal{E}^{\pi^{(k)}}_{\mathcal{M}_{\mathrm{tar}}}(q)_h \ge \widehat{\mathcal{E}}^{(k)}(q;\pi^{(k)})_h - c(w) > \gamma - c(w) \ge \gamma/2.$$

Under Bernstein calibration, with test radius $2\gamma_{\mathrm{fast}}(w)$, the right inequality of Lemma 27 gives

$$2\gamma_{\mathrm{fast}}(w) < \widehat{\mathcal{E}}^{(k)}(q;\pi^{(k)})_h \le 2\mathcal{E}^{\pi^{(k)}}_{\mathcal{M}_{\mathrm{tar}}}(q)_h + \gamma_{\mathrm{fast}}(w),$$

so the population residual exceeds $\gamma_{\mathrm{fast}}(w)/2$. Source membership then gives the same respective lower bounds on $\delta_h^{\pi^{(k)}}(\epsilon_{\mathrm{off}})$. Neither conclusion asserts that the deployed policy is suboptimal or that source information is unhelpful everywhere. The terminal block is not tested and produces no certificate. The balancing master uses returns rather than this statistic.

Remark 10 (Compatibility when pooling frozen-policy data)**.** *For a fixed policy π, compatibility must be measured relative to the source set actually used by the algorithm. In the finite-class setting, conditional on the source data, define*

$$\widehat{r}_{\mathrm{src}}(\pi) := \min_{f\in\widehat{\mathcal{F}}_{\mathrm{src}}(\beta,\pi)} \max_h \mathcal{E}^{\pi,\pi}_{\mathcal{M}_{\mathrm{tar}}}(f)_h.$$

The minimum exists on source validity because the empirical source set is finite and contains $Q^{\pi}_{\mathcal{M}_{\mathrm{src}}}$. Minimizing over the larger population source-fitting set can give a smaller value, even zero, without supplying an empirically admissible critic.

Suppose target data from the same frozen policy are pooled, with a concentration event valid at every pool size at which a test is performed. Write a_m for its Bernstein scale after m full blocks, including the logarithmic cost of simultaneous testing. At test radius $2a_m$, the inequalities of Lemma 27 imply

$$\widehat{r}_{\mathrm{src}}(\pi) \le a_m/2 \implies \textit{the intersection is nonempty},$$
$$\widehat{r}_{\mathrm{src}}(\pi) > 6a_m \implies \textit{the intersection is empty}.$$

For the first implication, a minimizing critic has empirical target residual at most $2(a_m/2)+a_m = 2a_m$ at every stage. For the second, any accepted critic would have population residual at most $2(2a_m+a_m) = 6a_m$ at every stage, a contradiction. At Hoeffding deviation scale c_m and test radius γ_m, the corresponding sufficient conditions are $\widehat{r}_{\mathrm{src}}(\pi) \le \gamma_m - c_m$ for acceptance and $\widehat{r}_{\mathrm{src}}(\pi) > \gamma_m + c_m$ for rejection.

Thus zero empirical-set incompatibility prevents fallback on the simultaneous validity event, whereas positive incompatibility is eventually detected if the frozen-policy sample grows and the radii vanish. Positive dynamics shift alone does not imply positive $\widehat{r}_{\mathrm{src}}(\pi)$. These are implications for the pooled test; an adaptively restarted pooling procedure additionally requires a concentration event covering all its tested starts and pool sizes. Theorem 4 analyzes the single-fallback rule and does not require such a pooling procedure or assert a pooled recovery-time guarantee.

Proof of Theorem 4. Work on the intersection of the source validity event, the block concentration event of Lemma 26 (or Lemma 27 under the Bernstein calibration), and $\mathsf{ALG}_{\mathrm{on}}$'s anytime event, of probability at least $1-3\delta$. If no fallback occurs, the run coincides with Algorithm 1, including its untested terminal block, so Theorem 2 or its Bernstein refinement applies with $N_{\mathrm{fb}} = 0$. Otherwise the first fallback has $k_0 < K$ and is a full tested block, so $t_0 = k_0 w$ episodes have been played at the handoff. Decompose $\mathrm{Regret}(T)$ into the blocks before k_0, block k_0 itself, and the post-handoff episodes.

Every block $k < k_0$ is certified, so the certified-block accounting of Theorem 2 applies verbatim to those blocks. The comparator term of block k is bounded by $w\sum_h (\widetilde{\delta}_h^{\pi^\star|\pi^{(k)}}(\epsilon_{\mathrm{off}};\bar{\gamma},\pi^{(k)}))^{1/2}$ and the iterate term by $w\sum_h\sqrt{\bar{\gamma}}$, by the same-block certification bound at the stated calibration. The Hedge comparison terms of blocks $k < k_0$ form the per-state Hedge trajectory driven by exactly those blocks' critics, and Lemma 12 at the fixed rate η_K over $k_0 - 1 \le K$ rounds gives, per state and step, at most $\log|\mathcal{A}|/\eta_K + \eta_K H^2 K/8 = H\sqrt{K\log|\mathcal{A}|/2}$; multiplying by the block length, taking expectations over $d_h^{\pi^\star}$, and summing over h bounds the optimization contribution by $wH^2\sqrt{K\log|\mathcal{A}|/2} \le H^2\sqrt{wT\log|\mathcal{A}|}$ exactly as in Theorem 2. Block k_0 is charged its trivial regret wH. Its empty intersection supplies no critic, and no source-driven update is performed.

Conditional on the history at the handoff, the fresh online learner uses new target episodes and randomness; its anytime guarantee therefore fails with conditional probability at most δ. No union bound over possible handoff times is needed. After the handoff, $\mathsf{ALG}_{\mathrm{on}}$ owns every remaining episode, so its realized regret over those $T - t_0$ episodes is at most $R_{\mathrm{on}}(T-t_0) \le R_{\mathrm{on}}(T)$ on its event, R_{on} being nondecreasing. Summing the three parts proves the display.

For the gating claims, the fallback direction is Remark 9. For the no-false-exit direction, fix a tested block $k < K$ with $E := \max_h \mathcal{E}^{\pi^{(k)},\pi^{(k)}}_{\mathcal{M}_{\rm tar}}(Q^{\pi^{(k)}}_{\mathcal{M}_{\rm src}})_h \le \gamma/2$. The witness $Q^{\pi^{(k)}}_{\mathcal{M}_{\rm src}}$ passes the empirical source test by the membership clause of Assumption 3, and it passes the empirical target test since $\widehat{\mathcal{E}}^{(k)}(Q^{\pi^{(k)}}_{\mathcal{M}_{\rm src}}; \pi^{(k)})_h \le E + c(w) \le \gamma/2 + \gamma/2 \le \gamma$ on the Hoeffding block event with $\gamma \ge 2c(w)$. Under the Bernstein calibration, $E \le \gamma_{\rm fast}/2$ gives $\widehat{\mathcal{E}}^{(k)} \le 2E + \gamma_{\rm fast} \le 2\gamma_{\rm fast}$, the test radius, by the right inequality of Lemma 27. The intersection is then nonempty and no fallback fires at block k. □

E Adaptation by Comparing Learner Returns

E.1 Discussion and the slope-adaptive form

Discussion of adaptivity. Theorem 21 can give a linear upper bound when its transfer level is large, whereas the purely online baseline has its own sublinear guarantee. Neither bound dominates in every setting. The master compares these guarantees through observed returns, with the coefficients and additive terms stated below. The lower bounds concern particular hard families and target-sample regimes; they do not require every learner on every instance to incur a linear modulus penalty. The construction follows regret balancing with elimination (Pacchiano et al., 2020; Cutkosky et al., 2021; Agarwal et al., 2017).

Concretely, let $\mathsf{ALG}_{\rm src}$ denote Algorithm 3 and let $\mathsf{ALG}_{\rm on}$ be any purely online algorithm with a known anytime regret bound $R_{\rm on}(t)$ holding with probability at least $1-\delta$ (for instance, $R_{\rm on}(t) = \widetilde{O}(H^2\sqrt{SAt})$ in the tabular case (Azar et al., 2017), or a GOLF-style bound under general function approximation (Xie et al., 2023)). Because the transfer level of $\mathsf{ALG}_{\rm src}$ is unknown, we register it with a geometric grid of capped candidate regret bounds $B^{(j)}(t) := \min\{Ht,\, 2^{-j}Ht + 2H^2\sqrt{t\log|\mathcal{A}|}\}$ for $j = 0, 1, \ldots, \lceil \log_2(HT)\rceil$, so that every candidate vanishes at $t = 0$; a candidate is well specified as soon as $2^{-j}H$ upper bounds the realized average transfer level of the source-informed learner at *every prefix* of its own episodes, since elimination tests the candidate bound at all times, not only at the final horizon. For the same reason, the registered copies are run with a time-varying learning-rate schedule whose optimization regret is $O(H^2\sqrt{t\log|\mathcal{A}|})$ uniformly over prefixes (Corollary 29 in Appendix E.2); with a rate tuned to the final horizon, the optimization cost is of order $H^2\sqrt{T\log|\mathcal{A}|}$ at *every* prefix, and the candidate would be violated at early prefixes even under benign transfer. The master plays base learners to keep their balanced regret estimates comparable and eliminates any learner that violates its candidate bound.

Slope-adaptive form.

Corollary 28 (No $T^{3/4}$ price of adaptivity)**.** *With Γ either $\min_{j\in\mathcal{W}_{\rm ws}} 2^{-j}H$ or $\Gamma_{\mathcal{W}}$, with probability at least $1-3\delta$,* $\mathrm{Regret}(T) = \widetilde{O}\big(M\Gamma T + MH^2\sqrt{T\log|\mathcal{A}|} + H\sqrt{MT\iota} + \sqrt{M}H^3\sqrt{T} + MH^3\big)$.

The two displays of Theorem 5 combine into the single oracle inequality over $\{\mathsf{ALG}_{\rm on}\} \cup \mathcal{W}_{\rm ws}$,

$$\mathrm{Regret}(T) = \widetilde{O}\Bigg(\min\Bigg\{ \Big(M + \sqrt{M}\,\frac{c_{\rm on}}{H^2}\Big) R_{\rm on}(T) + Mc_{\rm on}\Big(H^2 + \frac{c_{\rm on}}{H}\Big), \\ \min_{j\in\mathcal{W}_{\rm ws}} M\big(2^{-j}HT + H^2\sqrt{T\log|\mathcal{A}|}\big) + \sqrt{M}H^3\sqrt{T} + MH^3 \Bigg\} + H\sqrt{MT\iota}\Bigg).$$

where $\mathcal{W}_{\rm ws} := \{j :\ \mathrm{Reg}_j(r) \le B^{(j)}(r)$ at every own-play prefix $r\}$ is the set of pathwise well-specified source copies, the exact set over which Proposition 30 holds simultaneously, so the minimum is the master's oracle inequality itself. It always contains $j = 0$, whose capped candidate is $B^{(0)}(t) = \min\{Ht,\, Ht + 2H^2\sqrt{t\log|\mathcal{A}|}\} = Ht$, satisfied pathwise by every learner since returns lie in $[0, H]$. A transfer-interpretable certificate for membership is supplied by the realized levels. With $\bar{\Gamma}_j$ copy j's own prefix-maximal realized transfer level (22) over its own episodes, the subset $\mathcal{W}_\Gamma := \{0\} \cup \{j \ge 1 :\ 2^{-j}H \ge \bar{\Gamma}_j\}$ satisfies $\mathcal{W}_\Gamma \subseteq \mathcal{W}_{\rm ws}$ by Corollary 29, and we abbreviate $\Gamma_{\mathcal{W}} := \min_{j\in\mathcal{W}_\Gamma} 2^{-j}H \ge \min_{j\in\mathcal{W}_{\rm ws}} 2^{-j}H$, so the display holds a fortiori with $\Gamma_{\mathcal{W}}$; the levels enter through the certificate only, because the decomposition (22) bounds a realized level by $O(H^2)$

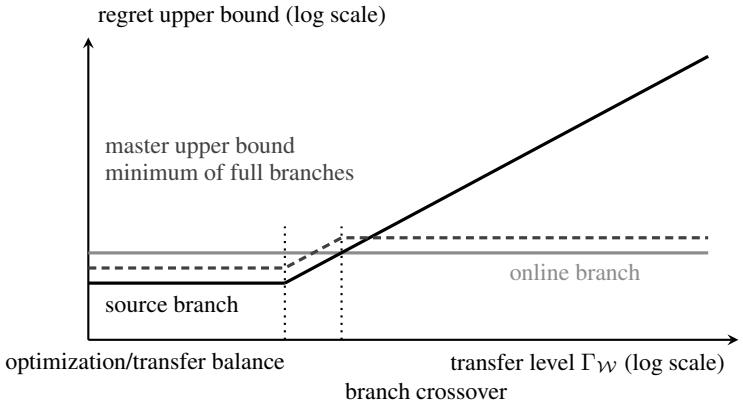


Figure 4: Schematic comparison of the two upper-bound branches in Theorem 5, holding other parameters fixed. The source branch grows with its transfer level, while the online branch does not. The master satisfies their minimum up to the stated constants and logarithmic factors. Both branches include their complete coefficient-dependent and additive costs; the curves are not observed regrets or exact crossover thresholds. Explicit expressions appear before Table 4.

rather than H, which is why $j = 0$ enters through the cap. A single badly misled copy cannot spoil the source guarantee. The coarser form with $\bar{\Gamma} := \max_j \bar{\Gamma}_j$ in place of the minimum follows from the three-case grid selection in the proof, up to a factor of two and an additive one. When the envelope's cap branch is active at the horizon, $R_{\mathrm{on}}(T) = HT$ and the first display holds trivially through the maximal-regret bound; when the square-root branch is active, $R_{\mathrm{on}}(T) = c_{\mathrm{on}}\sqrt{T\iota_{\mathrm{on}}}$ and the display is the substitution of Lemma 31(b). In particular, the adaptive learner is never worse than the online baseline by more than a multiplicative factor logarithmic in HT and polynomial in the baseline's coefficient, with no degradation in the T-rate, and it inherits the source-informed guarantee up to an additive $\widetilde{O}\big((MH^2 + \sqrt{M}H^3)\sqrt{T} + MH^3\big)$ model-selection overhead, with *no* $T^{3/4}$ term and no prior knowledge of the weak transfer modulus, since the slope-weighted $T^{3/4}$ expression of Lemma 31(a) is absorbed by the arithmetic-geometric mean inequality into the transfer term and the $\sqrt{T}$ overhead (Corollary 28). Both extra costs originate in the play-count accounting of the elimination analysis; the overhead is dominated by the transfer term whenever $\Gamma_{\mathcal{W}} \gtrsim (H^2 + H^3/\sqrt{M})/\sqrt{T}$.

In the source branch there is no separate $T^{3/4}$ overhead, since the additive model-selection overhead is always $\widetilde{O}\big((MH^2 + \sqrt{M}H^3)\sqrt{T} + MH^3\big)$, and the only super-$\sqrt{T}$ term in this upper bound is $O(\Gamma T)$; necessity statements remain restricted to the lower-bound families and regimes; the explicit master pays only the logarithmic factor $M = O(\log(HT))$ on it. The interpolation is visible in the slope-dependent form of Lemma 31(a), whose middle term collapses by the arithmetic-geometric mean inequality,

$$M^{1/4}\sqrt{\Gamma}\, H^{3/2}T^{3/4} = \sqrt{\Gamma T}\cdot\sqrt{\sqrt{M}H^3\sqrt{T}} \le \tfrac{1}{2}\Gamma T + \tfrac{1}{2}\sqrt{M}H^3\sqrt{T},$$

both absorbing terms being already present. Consequently the source branch is $\widetilde{O}(\sqrt{T})$ already when $\Gamma \lesssim (H^2 + H^3/\sqrt{M})/\sqrt{T}$, a $T^{-1/2}$ threshold rather than $1/T$; in particular, when a copy at the bottom grid slope $2^{-\lceil \log_2(HT)\rceil}H \le 1/T$ is pathwise well specified, its linear contribution over T episodes is at most 1 and

$$\mathrm{Regret}(T) = \widetilde{O}\Big(\big(MH^2 + \sqrt{M}H^3\big)\sqrt{T} + MH^3\Big),$$

with the same event and no change to the algorithm. The proof is the proof of Theorem 5 with the slope-dependent display of Lemma 31(a) carried through and the middle term absorbed as above.

Figure 4 summarizes sufficient regimes from the bounds. Transfer-based certification of a candidate must hold at every reached prefix, not only at the final count. For per-episode costs $(1, 0, \ldots, 0)$, the maximum prefix average is 1 while the final average is $1/T$, so nonincreasing costs do not make

the two coincide. Nondecreasing costs are a sufficient condition for the final average to be maximal. Even failure of this transfer-based condition need not imply elimination: the regret bound may be loose. Theorem 5 compares all pathwise well-specified copies, and the transfer-certified subset gives the bound with $\Gamma_{\mathcal{W}}$. Each condition concerns the copy's own observed episode sequence.

The proof, given in Appendix E.2, does not invoke an off-the-shelf model-selection theorem; the composition of a geometric source grid with an arbitrary $\sqrt{t}$-shaped online candidate falls outside the stated conditions of the results of Pacchiano et al. (2020) and Cutkosky et al. (2021), so we state and prove the master guarantee we need (Proposition 30), following their template. Episodic returns lie in $[0, H]$ and are observed after each episode; base learners update only on their own episodes, so Corollary 29 governs each copy under interleaved play; and with probability $1 - 2\delta$ at least one registered candidate is pathwise well specified (the online learner always, and the source copy whose slope dominates its own prefix-maximal realized transfer level). This result partially settles the adaptivity question raised in Section 7. Reliability of source guidance need not be known, or even tested through Bellman residuals, in order to be exploited, because observed returns suffice. The construction also composes with the strong-modulus algorithm. Proposition 32 in Appendix E.2 registers lazy copies at episode granularity, their capped prefix candidates being fully valid under same-block certification, together with the online learner registered through its native envelope, and the master inherits the strong transfer level of the best well-specified copy at the strong learner's own $T^{3/4}$ rate and the online $\sqrt{T}$ envelope simultaneously, with no prior knowledge of either. The one subtlety is an accounting choice: the online learner carries a much smaller square-root coefficient than the lazy copies and is therefore played far more often, and charging its plays through the elimination test would cost a linear term, whereas charging it by its own envelope, which is valid whatever the comparator, costs only its own $\sqrt{T}$ price. What remains open is adaptivity at the finer, region-level granularity, namely exploiting source guidance in the parts of the state space where it is sound while learning online elsewhere *within* a single policy.

E.2 Proof of the adaptive guarantee

We prove Theorem 5 in three steps. First we record a prefix-uniform regret guarantee for the source-informed learner, which the fixed-learning-rate Theorem 23 does not provide (see the caution after Lemma 12). Then we state and prove, self-contained, a balancing-with-elimination master guarantee for anytime candidates; the argument follows the regret-balancing-with-elimination template of Pacchiano et al. (2020) and Cutkosky et al. (2021), but we do not invoke their theorems, whose stated conditions (in particular the slope-ordering condition on misspecified candidates in the simplified form of the former) do not cover the composition of a geometric source grid with an arbitrary $\sqrt{t}$-shaped online candidate. Finally we apply the master to our candidate family.

Corollary 29 (Anytime source-constrained regret)**.** *Run Algorithm 3 with uniform initialization, radii β and $\{\alpha_t\}$ as in Theorem 21, and the cumulative-gain policy update of Lemma 13 with schedule $\eta_t = \frac{1}{H}\sqrt{8\log|\mathcal{A}|/t}$. On the validity event of Theorem 23, simultaneously for every prefix $t \geq 1$,*

$$\mathrm{Regret}(t) \leq t\,\Gamma_t + 2H^2\sqrt{t\log|\mathcal{A}|},$$

where Γ_t is the cumulative transfer level (22) *over the first t episodes.*

Proof. The proof of Theorem 23 applies verbatim at each prefix t. The comparator and iterate residual terms are sums over the first t rounds by construction, giving $t\Gamma_t$, and the optimization term is bounded at every prefix by Lemma 13, contributing $2H^2\sqrt{t\log|\mathcal{A}|}$ after taking expectations over $d_h^{\pi^\star}$ and summing over h. The validity event is a single event uniform over rounds, so no further union bound is needed. □

The registered learners. The master maintains M base learners, namely $\mathsf{ALG}_{\mathrm{on}}$ with candidate $\min\{Ht, R_{\mathrm{on}}(t)\}$ and, for $j = 0, \ldots, \lceil \log_2(HT) \rceil$, an independent copy of $\mathsf{ALG}_{\mathrm{src}}$ run with the anytime schedule of Corollary 29 and registered with the capped candidate

$$B^{(j)}(t) = \min\left\{Ht,\ 2^{-j}Ht + 2H^2\sqrt{t\log|\mathcal{A}|}\right\},$$

so every candidate vanishes at $t = 0$, increases by at most H per episode, and has nonincreasing average $B^{(j)}(t)/t$. In each episode the master selects one active learner and deploys the policy

that learner proposes; the full observed trajectory is fed back to that learner only, which updates its internal datasets and its counter t_i, while the master records the episodic return, a value in $[0, H]$, for its elimination statistics. Each learner is played once in the first M episodes. The probability ledger has exactly three events, the shared source-validity event behind Corollary 29, the online learner's anytime event, and the return-concentration event of Proposition 30, giving $1 - 3\delta$; the registered source copies are Algorithm 3 instances whose guarantee uses source membership only, so no per-copy target events are needed.

Proposition 30 (Balancing with elimination over anytime candidates)**.** *Let M base learners be given, learner i carrying a nondecreasing candidate R_i with $R_i(0) = 0$, $R_i(t) \le Ht$, per-episode increments at most H, and nonincreasing average $R_i(t)/t$; set $\iota := \log(2MT/\delta)$. The capped families of Theorem 5 satisfy all of these. Monotonicity, the cap, and the vanishing origin are immediate. For the increments, write $g_i(t) := \epsilon_i t + c_i\sqrt{t\iota_i}$ and note the discrete bound $g_i(t) - g_i(t-1) = \epsilon_i + c_i\sqrt{\iota_i}/(\sqrt{t} + \sqrt{t-1}) \le \epsilon_i + c_i\sqrt{\iota_i/t}$. If the non-cap branch is active at t, then $g_i(t) \le Ht$, which is exactly $\epsilon_i + c_i\sqrt{\iota_i/t} \le H$, so the increment of R_i is at most H. If the minimum switches from the cap at $t-1$ to g_i at t, then $R_i(t) - R_i(t-1) \le Ht - H(t-1) = H$ directly. If the cap is active at both, the increment is exactly H. Returns lie in $[0, H]$. After an initial round-robin in which each learner is played once, the master plays an active learner minimizing $R_i(t_i)$ and, after each episode, eliminates any active learner i for which some active j satisfies*

$$\widehat{U}_j(t_j) - \widehat{U}_i(t_i) > \frac{R_i(t_i)}{t_i} + \kappa(t_i) + \kappa(t_j), \qquad \kappa(t) := H\sqrt{2\iota/t},$$

where $\widehat{U}_j$ denotes learner j's average realized return. Call learner $i^\star$ pathwise well specified *if its realized regret obeys $\mathrm{Reg}_{i^\star}(t) \le R_{i^\star}(t)$ at every value of its own counter. Then, with probability at least $1 - \delta$, simultaneously for every pathwise well-specified $i^\star$,*

$$\mathrm{Regret}(T) \le M\big(R_{i^\star}(T) + 3H\big) + 2H\sqrt{2MT\iota} + \sum_{s=1}^{T} \rho_{i^\star}(s),$$

where $\rho_{i^\star}(s) := \min\big\{H,\ R_{i^\star}(t_{i^\star}(s))/t_{i^\star}(s) + 2\kappa(t_{i^\star}(s))\big\}$ for $t_{i^\star}(s) \ge 1$, with $t_{i^\star}(s)$ the count of $i^\star$ after s episodes and $\rho_{i^\star}(s) := H$ while $t_{i^\star}(s) = 0$; the clip at H is what makes the sequence nonincreasing from count zero onward, and it is valid because per-episode regret never exceeds H.

Proof. The event. The master selects a copy before observing the next trajectory, and that copy updates only on its own episodes. Conditional expected return therefore equals the deployed policy's value under the global history. Applying the martingale bound to the sequence of plays of each copy (or equivalently its independent, sequentially revealed episode stream) is valid under adaptive scheduling. Each learner's returns form a martingale difference sequence around the values of the deployed policies, so Azuma-Hoeffding with a union bound over $i \in [M]$ and $t \le T$ gives, with probability at least $1 - \delta$, $|\widehat{U}_i(t) - \frac{1}{t}\sum_{s \le t} V^{\pi_{i,s}}(\rho)| \le \kappa(t)$ for all i and t. Work on this event.

Well-specified learners survive. For any active j and pathwise well-specified $i^\star$ at counts (t_j, t), the event gives $\widehat{U}_j(t_j) \le V^\star + \kappa(t_j)$ and $\widehat{U}_{i^\star}(t) \ge V^\star - R_{i^\star}(t)/t - \kappa(t)$, so the elimination inequality never fires at $i = i^\star$.

Per-learner damage. Fix a learner i with final count $t_i \ge 2$ and let σ_i be the episode of its last play. At that moment i was active, so the test with $j = i^\star$ had not fired at counts $(t_i - 1,\ t^\star)$, $t^\star := t_{i^\star}(\sigma_i)$. Using the event for $\widehat{U}_i$ and for $\widehat{U}_{i^\star}$, and the survival bound for $V^\star - \widehat{U}_{i^\star}(\sigma_i)$,

$$\mathrm{Reg}_i(t_i) \le R_i(t_i) + 2t_i\,\kappa(t_i) + t_i\,\rho_{i^\star}(\sigma_i) + 2H.$$

This holds with the *clipped* $\rho_{i^\star}$. When the minimum selects the unclipped value, the display is exactly the chain above; when it selects H, the display holds trivially, since $\mathrm{Reg}_i(t_i) \le t_i H$ pathwise (returns lie in $[0, H]$) while the remaining terms on the right are nonnegative.

Balancing. When i was last played it minimized the candidates among active learners, and $i^\star$ was active, so $R_i(t_i - 1) \le R_{i^\star}(t^\star) \le R_{i^\star}(T)$ and hence $R_i(t_i) \le R_{i^\star}(T) + H$. Summing, $\sum_i R_i(t_i) \le M(R_{i^\star}(T) + H)$.

Deviation and cross terms. $\sum_i 2t_i\kappa(t_i) = 2H\sqrt{2\iota}\sum_i \sqrt{t_i} \le 2H\sqrt{2MT\iota}$ by Cauchy-Schwarz. The clipped $\rho_{i^\star}$ is nonincreasing in s from count zero onward. It equals H while $t_{i^\star}(s) = 0$, is

at most H from the first play on, and for counts $t \geq 1$ both summands inside the minimum are nonincreasing, $R_{i^\star}(t)/t$ by hypothesis and $\kappa(t)$ by definition. (Without the clip the sequence can *increase* at the first play, since $R_{i^\star}(1) + 2H\sqrt{2\iota}$ generally exceeds H, and the next step would fail for plays occurring before $i^\star$'s first play.) Hence $t_i\, \rho_{i^\star}(\sigma_i) \leq \sum_{s \in \mathrm{plays}(i)} \rho_{i^\star}(s)$, and summing over i gives $\sum_{s \leq T} \rho_{i^\star}(s)$. Collecting terms and absorbing the round-robin and overshoot additives ($\leq 2MH$) proves the claim. □

Lemma 31 (Cross term for the registered family)**.** *In the setting of Theorem 5, the registered candidates are $R_i(t) = \min\{Ht,\ \epsilon_i t + c_i\sqrt{t\iota_i}\}$, where the source copies have $\epsilon_j = 2^{-j}H$, $c_j = 2H^2$, and $\iota_j = \log|\mathcal{A}|$ (so that $\epsilon_j t + c_j\sqrt{t\iota_j}$ is exactly the registered $B^{(j)}$), and the online learner has $\epsilon_{\mathrm{on}} = 0$ and coefficient $c_{\mathrm{on}} \geq H$ with $\iota_{\mathrm{on}} \geq 1$. Write ι' for the largest logarithmic factor in play. Then, on the event of Proposition 30, the following hold up to universal constants and polylogarithmic factors. For a source comparator, with slope $\epsilon^\star \in [0, H]$, coefficient $c^\star = 2H^2$, and $\iota^\star = \log|\mathcal{A}|$, and $T \geq MH^2$,*

$$\textit{(a)} \qquad \sum_{s \leq T} \rho_{i^\star}(s) \ \leq\ \epsilon^\star T + \widetilde{O}\Big(\sqrt{M}H^3\sqrt{T} + M^{1/4}\sqrt{\epsilon^\star}\,H^{3/2}\,T^{3/4} + MH^3\Big) \leq\ \epsilon^\star T + \widetilde{O}\big(M^{1/4}H^{5/2}T^{3/4}\big),$$

where the first, slope-dependent form is the one the proof establishes and the second follows from $\epsilon^\star \leq H$ and $T \geq MH^2$, with no dependence on c_{on} in either; the middle term of the first form is absorbed into its neighbors by the arithmetic-geometric mean inequality once multiplied into the master bound, which is how Corollary 28 removes the $T^{3/4}$ term entirely; and for the online comparator, with $\epsilon^\star = 0$ and coefficient c_{on},

$$\textit{(b)} \qquad \sum_{s \leq T} \rho_{i^\star}(s) \ \leq\ \widetilde{O}\Big(\sqrt{M}\Big(1 + \frac{c_{\mathrm{on}}}{H^2}\Big)c_{\mathrm{on}}\sqrt{T} + Mc_{\mathrm{on}}\Big(H^2 + \frac{c_{\mathrm{on}}}{H}\Big)\Big).$$

Proof. Throughout, $\rho_{i^\star}$ denotes the clipped quantity of Proposition 30; since the clip only decreases it, it suffices to bound the sums using the unclipped majorant $R_{i^\star}(t)/t + 2\kappa(t)$, which is what the computations below do.

Capacity. For any level $B \geq 0$ and any learner i, if $R_i(t) \leq B$ then at least one branch of the minimum is at most B, so $t \leq B/H$ or $c_i\sqrt{t\iota_i} \leq \epsilon_i t + c_i\sqrt{t\iota_i} \leq B$, whence

$$n_i(B) := \max\{t :\ R_i(t) \leq B\} \ \leq\ \frac{B}{H} + \frac{B^2}{c_i^2\iota_i}.$$

No identification of cap-active rounds is needed. If a learner $i \neq i^\star$ is selected at pre-play count r, the balancing rule gives $R_i(r) \leq R_{i^\star}(t_{i^\star})$, and the increment condition then gives $R_i(r+1) \leq R_{i^\star}(t_{i^\star}) + H \leq R_{i^\star}(k) + H = B_k$ while $t_{i^\star} \leq k$; since R_i is nondecreasing, $n_i(B_k)$ therefore bounds learner i's *cumulative* play count over that period, the $+H$ overshoot being exactly what the definition of B_k carries. Hence the number of episodes elapsed before $t_{i^\star}$ reaches k satisfies

$$s(k) \ \leq\ k + \sum_{i \neq i^\star} n_i\big(B_k\big), \qquad B_k := R_{i^\star}(k) + H.$$

Split $\rho_{i^\star}(s) \leq \epsilon^\star + \gamma/\sqrt{t_{i^\star}(s)}$ for $t_{i^\star}(s) \geq 1$, with $\gamma \leq c^\star\sqrt{\iota'} + 2\sqrt{2}H\sqrt{\iota}$; the linear part contributes at most $\epsilon^\star T$ in total, and the round-robin plus the $t_{i^\star} = 0$ convention contribute at most $2MH$.

Case (a), source comparator. Here $B_k \leq \epsilon^\star k + c^\star\sqrt{k\iota^\star} + H$ and $\gamma \leq c^\star\sqrt{\iota^\star} + 2\sqrt{2}H\sqrt{\iota} \leq CH^2\sqrt{\iota'}$. Every competitor satisfies $c_i^2\iota_i \geq H^2$. For the online learner this follows from $c_{\mathrm{on}} \geq H$ and $\iota_{\mathrm{on}} \geq 1$; for a source copy, $c_j^2\iota_j = 4H^4\log|\mathcal{A}| \geq 4(\log 2)H^4 \geq H^2$ for $H \geq 1$ and $|\mathcal{A}| \geq 2$ (we assume $|\mathcal{A}| \geq 2$ throughout without loss of generality, since the one-action case has zero regret). Hence $n_i(B_k) \leq B_k/H + B_k^2/H^2$, with *no* reference to c_{on}. Therefore

$$s(k) \ \leq\ k + M\Big(\frac{B_k}{H} + \frac{B_k^2}{H^2}\Big) \ \leq\ k\Big(1 + CM\Big(1 + \frac{(c^\star)^2\iota'}{H^2}\Big)\Big) + \frac{CM(\epsilon^\star)^2k^2}{H^2} + CM\frac{c^\star\sqrt{k\iota'}}{H} + CM.$$

Inverting the three regimes, $t_{i^\star}(s) \geq c\min\{s/a_1,\ \sqrt{sH^2/(M(\epsilon^\star)^2)},\ (sH/(Mc^\star\sqrt{\iota'}))^2\}$ with $a_1 := CM(1+(c^\star)^2\iota'/H^2) \leq CMH^2\iota'$, so

$$\sum_{s\leq T}\frac{\gamma}{\sqrt{t_{i^\star}(s)}} \;\leq\; C\gamma\Big(\sqrt{a_1 T}+\Big(\frac{M(\epsilon^\star)^2}{H^2}\Big)^{1/4}T^{3/4}+\frac{Mc^\star\sqrt{\iota'}}{H}\log T\Big).$$

The three contributions are, up to polylogs, as follows. The first is $\sqrt{M}H^3\sqrt{T}$. The second is $C\gamma\,(M(\epsilon^\star)^2/H^2)^{1/4}T^{3/4} \;=\; C\,M^{1/4}\sqrt{\epsilon^\star}\,\gamma H^{-1/2}\,T^{3/4} \;\leq\; \widetilde{O}\big(M^{1/4}\sqrt{\epsilon^\star}H^{3/2}T^{3/4}\big)$, using $\gamma \leq CH^2\sqrt{\iota'}$ (note the square root on $\epsilon^\star$, from the quarter power of $(\epsilon^\star)^2$). The third is $\widetilde{O}(MH^3)$. Collecting terms proves the slope-dependent display of (a) exactly as written; the coarse form follows since $\sqrt{M}H^3\sqrt{T} \leq M^{1/4}H^{5/2}T^{3/4}$ for $T \geq MH^2$, $\sqrt{\epsilon^\star} \leq \sqrt{H}$, and $MH^3 \leq M^{1/4}H^{5/2}T^{3/4}$ in the same regime.

Case (b), online comparator. Here $\epsilon^\star = 0$, $B_k \leq c_{\mathrm{on}}\sqrt{k\iota_{\mathrm{on}}} + H$, and $\gamma \leq Cc_{\mathrm{on}}\sqrt{\iota'}$. Every competitor is a source copy with the explicit product $c_j^2\iota_j = 4H^4\log|\mathcal{A}| \geq 4(\log 2)H^4$ for $|\mathcal{A}| \geq 2$, so

$$s(k) \;\leq\; k + M\Big(\frac{B_k}{H}+\frac{B_k^2}{4H^4\log|\mathcal{A}|}\Big) \;\leq\; k\Big(1+\frac{CMc_{\mathrm{on}}^2\iota_{\mathrm{on}}}{H^4}\Big)+\frac{CMc_{\mathrm{on}}\sqrt{k\iota_{\mathrm{on}}}}{H}+CM.$$

Inverting the two regimes, $t_{i^\star}(s) \geq c\min\{s/a_1',\ (sH/(Mc_{\mathrm{on}}\sqrt{\iota_{\mathrm{on}}}))^2\}$ with $a_1' := C(1 + Mc_{\mathrm{on}}^2\iota_{\mathrm{on}}/H^4)$, so

$$\sum_{s\leq T}\frac{\gamma}{\sqrt{t_{i^\star}(s)}} \;\leq\; C\gamma\Big(\sqrt{a_1' T}+\frac{Mc_{\mathrm{on}}\sqrt{\iota_{\mathrm{on}}}}{H}\log T\Big) \;\leq\; \widetilde{O}\Big(\sqrt{M}\Big(1+\frac{c_{\mathrm{on}}}{H^2}\Big)c_{\mathrm{on}}\sqrt{T}+\frac{Mc_{\mathrm{on}}^2}{H}\Big),$$

and adding the $2MH$ additive and the trivial bound $Mc_{\mathrm{on}}H^2$ for the first rounds after the round-robin gives (b). □

Proof of Theorem 5. Work on the intersection of the shared source validity event, the online learner's event, and the event of Proposition 30, of probability at least $1-3\delta$. The registered candidates satisfy the proposition's hypotheses, as checked in its statement. The online learner is pathwise well specified through its envelope by assumption. By Corollary 29, a source copy is pathwise well specified whenever its slope dominates its own prefix-maximal realized transfer level. Hence every copy $j \geq 1$ with $2^{-j}H \geq \bar{\Gamma}_j$ is pathwise well specified, and the copy $j = 0$ is pathwise well specified unconditionally, since its capped candidate equals Ht and every learner's realized regret is at most Ht (returns lie in $[0, H]$); note the realized levels themselves are bounded only by $O(H^2)$ under the decomposition (22), so the unconditional membership of $j = 0$ comes from the cap, not from the levels. Instantiating Proposition 30 with case (b) of Lemma 31 at $i^\star = \mathsf{ALG}_{\mathrm{on}}$ gives the first display of (10). Since the proposition holds simultaneously for every pathwise well-specified learner, instantiating case (a) at each $j \in \mathcal{W}_{\mathrm{ws}}$, whose candidate at T is at most $2^{-j}HT + 2H^2\sqrt{T\log|\mathcal{A}|}$, absorbing the slope-weighted middle term of the sharp display of Lemma 31(a) by the arithmetic-geometric mean inequality, $M^{1/4}\sqrt{\epsilon^\star}H^{3/2}T^{3/4} = \sqrt{\epsilon^\star T}\cdot\sqrt{\sqrt{M}H^3\sqrt{T}} \leq \frac{1}{2}\epsilon^\star T + \frac{1}{2}\sqrt{M}H^3\sqrt{T}$, into terms already present, and taking the minimum over $\mathcal{W}_{\mathrm{ws}}$ gives the second display; the certificate subset $\mathcal{W}_\Gamma \subseteq \mathcal{W}_{\mathrm{ws}}$ then gives the transfer-interpretable bound with $\Gamma_{\mathcal{W}}$ a fortiori. The coarser $\bar{\Gamma}$ form follows by a three-case grid selection with $\bar{\Gamma} := \max_j \bar{\Gamma}_j$. If $\bar{\Gamma} \geq H$, the copy $j = 0$ gives $\Gamma_{\mathcal{W}} \leq H \leq \bar{\Gamma}$. If $\bar{\Gamma} \in [2^{-\lceil\log_2(HT)\rceil}H,\ H)$, some grid point satisfies $2^{-j}H \in [\bar{\Gamma}, 2\bar{\Gamma}]$; its slope dominates every copy's level, in particular its own, so $j \in \mathcal{W}_\Gamma$ and $\Gamma_{\mathcal{W}} \leq 2\bar{\Gamma}$. If $\bar{\Gamma}$ is below the bottom grid slope, the bottom copy lies in $\mathcal{W}_\Gamma$ and its linear contribution over T episodes is at most 1. □

Composition with the strong-modulus learner. A lazy copy holds its policy fixed for each block of w of its own episodes. Its update uses the cumulative-gain version of Lemma 13, with learning rate indexed by the number of certified updates; a fallback adds no gain and advances no update index. The global budget T is an upper bound on every copy's own budget, and the terminal-block convention of Algorithm 1 uses $K = \lceil T/w\rceil$. A prefix at which the master inspects a copy is not a new terminal horizon for that copy.

Let $\mathcal{J}_j(t)$ be the fully observed, tested blocks of copy j by its own episode count t, and let $\mathcal{B}_j$ be its fallback blocks. Thus $k \in \mathcal{J}_j(t)$ requires $kw \le t$ and $k < K$. At a reached boundary $u = w\ell \le t$, define

$$\Gamma^{\mathrm{str}}_{j,u} := \frac{1}{u}\left[\sum_{k\in\mathcal{J}_j(u)\setminus\mathcal{B}_j} w\sum_h \left(\sqrt{\widetilde{\delta}_h^{\pi^\star|\pi_j^{(k)}}(\epsilon_{\mathrm{off}};\bar{\gamma},\pi_j^{(k)})} + \sqrt{\bar{\gamma}}\right) + wH|\mathcal{J}_j(u)\cap\mathcal{B}_j|\right].$$

Write $\bar{\Gamma}^{\mathrm{str}}_j$ for the maximum over reached block boundaries, with maximum zero if none has been reached. This level includes the iterate and fallback costs as well as the comparator residual. It is distinct from the comparator-only average in Corollary 3.

The tested certified blocks form the actual Hedge update subsequence. Their optimization cost through own count t is at most $c_w\sqrt{t}$, with $c_w = 2H^2\sqrt{w\log|\mathcal{A}|}$. The only unaccounted episodes belong to one unfinished or untested terminal block, whose regret is at most wH. Consequently, simultaneously at every reached own prefix,

$$\mathrm{Reg}_j(t) \le t\bar{\Gamma}^{\mathrm{str}}_j + c_w\sqrt{t} + wH, \qquad B^{(j)}_{\mathrm{lazy}}(t) := \min\{Ht, 2^{-j}Ht + c_w\sqrt{t} + wH\}.$$

The candidate is valid when its slope dominates $\bar{\Gamma}^{\mathrm{str}}_j$; copy $j = 0$ is always valid through the cap. This argument also covers a completed full terminal block, which is not tested and is charged by the same wH offset.

Validity under interleaved play. Let $\mathcal{H}_s$ contain all observations through global episode s. The master chooses the next copy using $\mathcal{H}_s$, before the next trajectory is generated. For a frozen block of a selected copy, condition on the global history when that block starts. One may generate in advance w independent target trajectories from its fixed policy and reveal the next unused trajectory whenever the master selects that copy. Other copies' outcomes and all later scheduling decisions depend only on revealed observations and fresh independent randomness. Thus Lemmas 26 and 27 apply to each potential full block conditional on its start history. Union-bound these events over copies and block indices before restricting to the blocks actually completed; do not condition on completion, which can depend on the observed returns. This yields uniform block validity even if a copy is paused or eliminated mid-block. Source validity is shared and uniform over the predetermined policy closure, and episodic return deviations are martingale differences under the global filtration.

The master proof charges the online learner by its own envelope when comparing against a lazy copy. This avoids charging its potentially much larger play count through the lazy comparator's average candidate. The remaining competitors share the coefficient c_w, permitting the capacity argument below.

Proposition 32 (Composition with the strong-modulus learner at both native rates)**.** *Fix an integer block length $1 \le w \le T$ and let $c_w := 2H^2\sqrt{w\log|\mathcal{A}|}$. Register with the master of Proposition 30 the $M = O(\log(HT))$ learners consisting of, for $j = 0,\dots,\lceil\log_2(HT)\rceil$, an independent copy of Algorithm 1 run at block length w with the anytime block schedule, same-block certification at the Bernstein calibration of Corollary 3, and the capped candidate $B^{(j)}_{\mathrm{lazy}}(t) = \min\{Ht,\ 2^{-j}Ht + c_w\sqrt{t} + wH\}$ over its own episode counter, together with $\mathsf{ALG}_{\mathrm{on}}$ registered through its* native *envelope $R_{\mathrm{on}}(t) = \min\{Ht, c_{\mathrm{on}}\sqrt{t\iota_{\mathrm{on}}}\}$, $H \le c_{\mathrm{on}} \le c_w$. Assume $T \ge M$. Let $\mathcal{W}_{\mathrm{str}} := \{0\}\cup\{j \ge 1 : 2^{-j}H \ge \bar{\Gamma}^{\mathrm{str}}_j\}$ with $\bar{\Gamma}^{\mathrm{str}}_j$ copy j's prefix-maximal strong transfer level defined above. Allocate failure probability $\delta/(M+2)$ to the shared source event, each of the $M-1$ copy block events, the online event, and the return event. Then, with probability at least $1-3\delta$, both*

$$\mathrm{Regret}(T) = \widetilde{O}\Big(Mc_{\mathrm{on}}\sqrt{T} + \frac{Mc_{\mathrm{on}}^2}{H} + H\sqrt{MT\iota}\Big)$$

and, for every $j \in \mathcal{W}_{\mathrm{str}}$,

$$\mathrm{Regret}(T) = \widetilde{O}\Big(M\big(2^{-j}HT + c_w\sqrt{T} + wH\big) + \sqrt{M}\,c_w\sqrt{T} + MwH^3 + c_{\mathrm{on}}\sqrt{T}\Big).$$

With $w = \lceil\sqrt{T}\rceil$, so that $c_w = \widetilde{O}(H^2T^{1/4})$, and $T\Gamma_{\mathrm{str}} := \min_{j\in\mathcal{W}_{\mathrm{str}}} 2^{-j}HT$,

$$\mathrm{Regret}(T) = \widetilde{O}\Big(\min\Big\{M\big(T\Gamma_{\mathrm{str}} + H^2T^{3/4}\big),\ Mc_{\mathrm{on}}\sqrt{T}\Big\} + c_{\mathrm{on}}\sqrt{T} + \frac{Mc_{\mathrm{on}}^2}{H} + MH^3\sqrt{T} + H\sqrt{MT\iota}\Big),$$

so the master inherits the strong transfer level of the best well-specified lazy copy at the strong learner's own $T^{3/4}$ rate and *the online learner's native $\sqrt{T}$ envelope, simultaneously and with no prior knowledge of either; the additive $c_{\mathrm{on}}\sqrt{T}$ is the online learner's own price and never exceeds the online branch.*

The point of the proof is a single accounting choice. The elimination analysis of Proposition 30 bounds the regret of every competitor of the comparator through the elimination test, which charges each of the competitor's plays with the comparator's current average candidate. For competitors carrying the same square-root coefficient as the comparator this is harmless, but the online learner carries the much smaller coefficient c_{on}, so the balancing rule plays it up to $(c_w/c_{\mathrm{on}})^2$ times more often than a lazy comparator, and charging those plays through the test yields a linear-in-T term that would prevent the intended sublinear comparison. The online learner, however, is pathwise well specified on its own event whatever the comparator, so its regret can be charged directly by its envelope; its play count then never enters the analysis, and the lazy competitors, which share the comparator's coefficient, are charged through the test exactly as before.

Proof. Hypotheses and ledger. Each $B^{(j)}_{\mathrm{lazy}}$ is capped, zero at the origin through the cap, has nonincreasing average since $2^{-j}H + c_w/\sqrt{t} + wH/t$ is nonincreasing, and has increments at most H by the active-branch argument of Proposition 30, the constant offset wH cancelling from increments; copy j is pathwise well specified on its own episodes whenever $2^{-j}H \geq \bar{\Gamma}^{\mathrm{str}}_j$, by the candidate construction above, and $j = 0$ unconditionally through the cap. The native envelope is valid on the online event. The ledger has $M + 2$ events: shared source validity, $M - 1$ copy block events, the online event, and return concentration. The allocation stated above bounds total failure by δ, hence also by 3δ, without requiring independence between events. The interleaved-play argument above supplies the block events and changes only their logarithmic factors. Capacities are $n_j(B) \leq B/H + B^2/c_w^2$ for every lazy copy, since its noncap branch forces $c_w\sqrt{t} \leq B$, and $n_{\mathrm{on}}(B) \leq B/H + B^2/(c_{\mathrm{on}}^2\iota_{\mathrm{on}})$ for the online learner.

Online comparator. Take $i^\star = \mathsf{ALG}_{\mathrm{on}}$, so $B_k \leq c_{\mathrm{on}}\sqrt{k\iota_{\mathrm{on}}} + H$ and every competitor is a lazy copy. The number of episodes elapsed before $t_{i^\star}$ reaches k is $s(k) \leq k + M(B_k/H + B_k^2/c_w^2) \leq CM\iota_{\mathrm{on}}\, k + CMc_{\mathrm{on}}\sqrt{k\iota_{\mathrm{on}}}/H + CM$, using $c_{\mathrm{on}} \leq c_w$. Inverting the two regimes as in case (b) of Lemma 31, with $\rho_{i^\star}(s) \leq Cc_{\mathrm{on}}\sqrt{\iota'}/\sqrt{t_{i^\star}(s)}$, gives $\sum_s \rho_{i^\star}(s) \leq \widetilde{O}(\sqrt{M}c_{\mathrm{on}}\sqrt{T} + Mc_{\mathrm{on}}^2/H)$; adding $M(R_{\mathrm{on}}(T) + 3H)$ and $2H\sqrt{2MT\iota}$ from Proposition 30 gives the first display.

Lazy comparator. Fix $j^\star \in \mathcal{W}_{\mathrm{str}}$ with slope $\epsilon^\star = 2^{-j^\star}H$ and write $i^\star$ for that copy. Decompose $\mathrm{Regret}(T) = \sum_i \mathrm{Reg}_i(t_i)$ into the comparator, the other lazy copies, and the online learner. The comparator contributes $\mathrm{Reg}_{i^\star}(t_{i^\star}) \leq R_{i^\star}(T)$ by well-specification. The online learner is pathwise well specified on its event, so $\mathrm{Reg}_{\mathrm{on}}(t_{\mathrm{on}}) \leq R_{\mathrm{on}}(t_{\mathrm{on}}) \leq c_{\mathrm{on}}\sqrt{T\iota_{\mathrm{on}}}$; this is the only place the online learner enters. For a lazy competitor $j \neq j^\star$ with final count $t_j \geq 2$, the per-learner damage bound in the proof of Proposition 30, which uses only the return-concentration event and the fact that $i^\star$ is never eliminated, gives $\mathrm{Reg}_j(t_j) \leq R_j(t_j) + 2t_j\kappa(t_j) + t_j\rho_{i^\star}(\sigma_j) + 2H$ with σ_j the episode of j's last play, and balancing gives $R_j(t_j) \leq R_{i^\star}(T) + H$, so $\sum_{j\neq j^\star} R_j(t_j) \leq (M-2)(R_{i^\star}(T) + H)$ and $\sum_j 2t_j\kappa(t_j) \leq 2H\sqrt{2MT\iota}$. By the monotonicity of the clipped $\rho_{i^\star}$, $t_j\rho_{i^\star}(\sigma_j) \leq \sum_{s\in\mathrm{plays}(j)} \rho_{i^\star}(s)$, so the cross term is at most $\sum_{s\in\mathcal{P}} \rho_{i^\star}(s)$ with $\mathcal{P}$ the set of plays of lazy competitors, of cardinality at most T.

Capacity over lazy plays only. A lazy competitor j selected at pre-play count r while $t_{i^\star} \leq k$ satisfies $R_j(r) \leq R_{i^\star}(t_{i^\star}) \leq R_{i^\star}(k)$, because $i^\star$ is active, so its cumulative plays while $t_{i^\star} \leq k$ number at most $n_j(B_k)$ with $B_k := R_{i^\star}(k) + H \leq \epsilon^\star k + c_w\sqrt{k} + (w+1)H$. Hence the number $N(k)$ of plays in $\mathcal{P}$ that occur while $t_{i^\star} \leq k$ satisfies, by the capacity bound, $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, $\epsilon^\star \leq H$, the arithmetic-geometric mean bound $c_w\sqrt{k}/H \leq H^2 w\log|\mathcal{A}| + k$, and $(w+1)^2H^2/c_w^2 \leq Cw$,

$$N(k) \;\leq\; (M-2)\Big(\frac{B_k}{H} + \frac{B_k^2}{c_w^2}\Big) \;\leq\; CMk + \frac{3M(\epsilon^\star)^2k^2}{c_w^2} + s_0, \qquad s_0 := CMH^2w\iota',$$

with no reference to the online learner. Enumerate the plays in $\mathcal{P}$ in time order, $m = 1, \ldots, |\mathcal{P}|$, and let k_m be the comparator's count at the m-th such play; all m plays occurred while $t_{i^\star} \leq k_m$, so $N(k_m) \geq m$, and inverting the bound on N gives $k_m \geq c\min\{m/M,\ (c_w/\epsilon^\star)\sqrt{m/(3M)}\}$

for $m \geq 2s_0$. Charging the first $2s_0$ plays by $\rho \leq H$ and the remaining ones by $\rho_{i^\star}(k_m) \leq \epsilon^\star + Cc_w\sqrt{\iota'}/\sqrt{k_m} + wH/k_m$,

$$\sum_{s\in\mathcal{P}} \rho_{i^\star}(s) \;\leq\; \epsilon^\star T + \widetilde{O}\big(\sqrt{M}c_w\sqrt{T} + M^{1/4}\sqrt{\epsilon^\star c_w}\,T^{3/4} + MwH^3 + \sqrt{MwT}\big),$$

where the linear part is summed over at most T plays, the wH/k part contributes $\widetilde{O}(MwH)$ in the linear regime and at most $C\sqrt{MwT}$ in the square-root regime, and the middle term is absorbed by the arithmetic-geometric mean inequality, $M^{1/4}\sqrt{\epsilon^\star c_w}\,T^{3/4} \leq \frac{1}{2}\epsilon^\star T + \frac{1}{2}\sqrt{M}c_w\sqrt{T}$. Collecting the comparator's own regret, the balancing sum, the deviation term, the cross term, the online learner's regret, and the round-robin additive $2MH$ gives the second display. The instantiation substitutes $w = \lceil\sqrt{T}\rceil$ and $c_w = \widetilde{O}(H^2T^{1/4})$, and $\min\{a + c,\, b\} \leq \min\{a, b\} + c$ for $c \geq 0$ moves the online learner's own price outside the minimum. □

Remark 11 (Leveling and the source-branch overhead of Theorem 5)**.** *Registering the online learner through an envelope leveled to c_w also composes, at the cost of a weaker online branch; the accounting above avoids this. The same accounting sharpens the source branch of Theorem 5, where the online learner's plays were charged through the elimination test under the worst-case coefficient bound $c_{\mathrm{on}} \geq H$: charging it by its envelope instead replaces the overhead $\sqrt{M}H^3\sqrt{T} + MH^3$ by $c_{\mathrm{on}}\sqrt{T\iota_{\mathrm{on}}}$, and since both accountings are valid on the same event the smaller of the two applies.*

Remark 12 (Granularity of adaptation)**.** *The master adapts at the level of whole learners. It discovers whether source-informed learning is globally competitive. It does not adapt within a policy, e.g., trusting source guidance in some regions of the state space and exploring online in others. Achieving such region-level adaptivity with matching instance-dependent guarantees is an open problem that the block-MDP structure of Theorem 37 makes precise. One would like regret scaling with $\sum_i p_i^{\pi^\star} \min\{\mathit{shift}_i, \mathit{online}_i\}$ rather than $\min$ of the aggregates.*

F Structured MDPs

F.1 Discussion and multiple sources

The weak transfer modulus is an abstract measure of when source Bellman fitting is informative for target performance. We now instantiate it in three standard structured settings. Across these examples, the same principle emerges. Transfer is easy when the target policy visits regions where the source data provide reliable Bellman constraints and where the source and target dynamics are aligned; transfer is hard when either source coverage or dynamics alignment fails.

Discussion of the tabular, linear, and block bounds. The bound is a purely population-level statement. No concentration event is needed, and randomness enters only through the source radius $\epsilon = \epsilon_{\mathrm{off}}(n, \delta)$ supplied by Appendix B.4. If, as the tabular setting permits, the source-fitting set is instead defined through local residual constraints at each (s, a, h) with visit counts $n_h(s, a)$, standard concentration localizes the statistical term to $\mathrm{Stat}_h(s, a) = \widetilde{O}\big(H^2(S + \log(HSA/\delta))/\max\{1, n_h(s, a)\}\big)$, capped at $4H^2$, for unrestricted tabular continuations; see Appendix F.2. The tabular bound shows that a state-action pair contributes to transfer difficulty only if it is visited by the target policy and either poorly covered by source data or significantly shifted between source and target dynamics. When $\mathcal{M}_{\mathrm{src}} = \mathcal{M}_{\mathrm{tar}}$, the dynamics-shift term vanishes and the bound reduces to the usual coverage obstruction from offline RL.

The $\sqrt{d}$ normalization matches the standard linear-MDP convention $\|\int g\, d\mu_{\mathcal{M},h}\|_2 \leq \sqrt{d}\,\|g\|_\infty$ (Jin et al., 2020), so that $\Delta_h^{\mathrm{lin}} \in [0, 2]$.

The linear bound separates three effects. The coverage term is small when the source behavior policy excites the feature directions the target policy uses, and it uses the pseudoinverse, so it charges only directions the source actually covers. The projector term charges the directions the behavior policy never excites at all, at the worst-case coefficient ambiguity B_h; the lower bounds of Section 5.2 show that it is this null-space term, not weak coverage, that admits a floor persisting in the source sample size. The shift term is controlled by the discrepancy between the source and target latent transition operators. Since the tolerance ϵ enters the coverage term linearly, instantiating $\epsilon = \epsilon_{\mathrm{off}}$

at the Bernstein scale of Lemma 16, extended to the infinite linear class by the Bernstein clause of Proposition 17, improves the statistical part of the covered directions from $\widetilde{O}(n^{-1/2})$ to $\widetilde{O}(n^{-1})$ in the regime $\mathcal{C}_{\mathrm{alg}} \le n$; the null-space and shift terms are unaffected, as the lower bounds require. The shift term is moreover target-weighted. Theorem 36 replaces Δ_h^{lin} by the smaller quantity $\Delta_h^{\mathrm{lin},\nu}$, in which the integrated shift is measured in the norm of the target feature covariance $\Sigma_h^\nu = \mathbb{E}_{d^\nu_{\mathrm{tar}},h}[\varphi\varphi^\top]$ of the evaluation policy, so that shift along feature directions the target policy never excites is not charged; the global form with Δ_h^{lin} follows from $\|(\Sigma_h^\nu)^{1/2}\|_{\mathrm{op}} \le 1$, and the decomposition into source uncertainty plus target-weighted shift holds literally in all three structured settings.

The block bound makes the policy dependence of the modulus especially transparent. A global source-target discrepancy may be large, but transfer can still be easy for policies that remain in stable blocks. Conversely, policies that enter shifted blocks pay a transfer penalty proportional to how often they visit those regions. No block structure is imposed on policies, and none is needed on the continuation values. In the Bellman backup, the continuation value enters only through its integral against the transition kernel, which is common to all state-action pairs of a block by homogeneity, so the backup is block-constant whatever the continuation function is; block-measurability of the critic class is used only to make f_h itself block-constant. The proof in Appendix F.2 makes this explicit.

Common structure. The three bounds show that the weak transfer modulus recovers familiar coverage quantities while adding the source-target shift terms specific to hybrid learning. In tabular MDPs, transfer difficulty is local in state-action space; in linear MDPs, it is governed by feature coverage and latent operator shift; in block MDPs, it decomposes over regions and captures partial transfer. Across these settings,

$$\text{weak transfer difficulty} \;\approx\; \text{source statistical uncertainty} + \text{target-weighted dynamics shift}. \tag{26}$$

Detailed proofs are deferred to Appendix F.2.

Multiple sources. The framework extends to several source environments, and the extension exhibits a benefit that no single source provides. Suppose sources $\mathcal{M}^{(1)}_{\mathrm{src}}, \dots, \mathcal{M}^{(J)}_{\mathrm{src}}$ come with behavior policies μ_j and radii ϵ_j, write $\mathcal{F}^{\mu_j,\pi}_{\mathrm{src},j}(\epsilon_j)$ for the corresponding source-fitting sets, and define the multi-source weak modulus by intersecting them,

$$\delta^{\nu|\pi}_{h,\mathrm{multi}}(\vec{\epsilon}) := \sup\Big\{\mathcal{E}^{\nu,\pi}_{\mathcal{M}_{\mathrm{tar}}}(f)_h \;:\; f \in \textstyle\bigcap_{j=1}^J \mathcal{F}^{\mu_j,\pi}_{\mathrm{src},j}(\epsilon_j)\Big\}, \qquad \sup\emptyset := 0. \tag{27}$$

Proposition 33 (Multi-source weak modulus)**.** *For all policies ν, π, steps h, and radii $\vec{\epsilon}$, the following hold. (i) $\delta^{\nu|\pi}_{h,\mathrm{multi}}(\vec{\epsilon}) \le \min_{j\in[J]} \delta^{\nu|\pi}_{h,(j)}(\epsilon_j)$, where $\delta_{h,(j)}$ denotes the single-source modulus of source j. (ii) In tabular MDPs, with $\Delta_h^{(j)}(s,a)$ the total-variation shift between source j and the target at (s,a,h) and $\mathrm{Stat}_h^{(j)}(s,a;\epsilon_j) := \min\{4H^2,\, \epsilon_j/q_h^{\mu_j}(s,a)\}$,*

$$\delta^{\nu|\pi}_{h,\mathrm{multi}}(\vec{\epsilon}) \le 2\sum_{s,a} d^\nu_{\mathcal{M}_{\mathrm{tar}},h}(s,a)\, \min_{j\in[J]} \Big[H^2\Delta_h^{(j)}(s,a)^2 + \mathrm{Stat}_h^{(j)}(s,a;\epsilon_j)\Big], \tag{28}$$

and the same bound holds blockwise under the assumptions of Theorem 37.

The point of (28) is that the minimum sits *inside* the sum. Each state-action pair is charged only its best source, so complementary sources jointly certify what no single source can. If source 1 matches the target on one region and source 2 on the complementary region, every single-source modulus is of order H^2 for a policy visiting both regions, while the multi-source modulus is purely statistical, so the gap between (28) and the minimum of the aggregate single-source bounds is unbounded. Two caveats apply. First, the intersection can be empty when the sources are mutually inconsistent, in which case the modulus is vacuous by the convention above; since $\sup\emptyset = 0$ would otherwise conflate perfect transfer with incompatible sources, the meaningful report is the pair (feasibility indicator, modulus). An empty *empirical* intersection is itself a certificate of cross-source disagreement, and the algorithmic counterpart is a fallback to the largest subset of sources whose empirical sets still intersect, in the spirit of (20), with the minimum in (28) restricted to that subset; this selection rule carries a formal guarantee, since on the simultaneous per-source validity event the inclusions hold for every source individually and hence for every data-dependent subset automatically, with no union over the 2^J subsets (Corollary 38 in Appendix F.2). Second, Algorithms 2 and 3 extend by

intersecting the empirical source confidence sets, and their guarantees hold with $\delta_{h,\mathrm{multi}}$ in place of δ_h on the intersection of the J concentration events, whenever the intersected set is nonempty at the visited rounds (otherwise the subset fallback applies). The proof is in Appendix F.2. What remains open is complementarity beyond the pointwise best source, for instance sources whose *combination* constrains a backup that no individual source constrains at all.

F.2 Proofs for structured MDPs

We now prove the explicit weak-modulus bounds for tabular, linear, and block-structured MDPs.

F.2.1 Tabular MDPs

For tabular MDPs, define the local transition discrepancy

$$\Delta_h(s,a) = \|P_{\mathrm{src},h}(\cdot|s,a) - P_{\mathrm{tar},h}(\cdot|s,a)\|_{\mathrm{TV}}\,.$$

Theorem 34 (Tabular weak-modulus bound)**.** *Suppose $\mathcal{S}$ and $\mathcal{A}$ are finite, rewards lie in $[0,1]$, and $f_h(s,a) \in [0,H]$ for all $f \in \mathcal{F}$. Then, for any policies ν, π, step h, and tolerance $\epsilon \geq 0$,*

$$\delta_h^{\nu|\pi}(\epsilon) \leq 2\sum_{s,a} d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h}(s,a)\left[H^2\Delta_h(s,a)^2 + \mathrm{Stat}_h(s,a;\epsilon)\right],$$

where $\mathrm{Stat}_h(s,a;\epsilon) := \min\left\{4H^2,\ \epsilon/d^{\mu}_{\mathcal{M}_{\mathrm{src}},h}(s,a)\right\}$ with the convention $\epsilon/0 = \infty$. Taking $\nu = \pi$ gives Theorem 6.

Proof. Fix $f \in \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon)$ and write the source residual $\varrho_h(s,a) := f_h(s,a) - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{src}},h} f_{h+1}(s,a)$. For each (s,a,h), decompose the target residual as

$$f_h(s,a) - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{tar}},h} f_{h+1}(s,a) = \varrho_h(s,a) + \left[\mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{src}},h} f_{h+1}(s,a) - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{tar}},h} f_{h+1}(s,a)\right]. \tag{29}$$

Shift term. Since rewards are shared, the Bellman operators differ only through the transition kernel. Because $f_{h+1}(\cdot,\pi_{h+1})$ takes values in $[0,H]$, centering at $H/2$ gives

$$\begin{aligned}
&\left|\mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{src}},h} f_{h+1}(s,a) - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{tar}},h} f_{h+1}(s,a)\right| \\
&= \left|\int \left(f_{h+1}(s',\pi_{h+1}) - \tfrac{H}{2}\right) d\big(P_{\mathrm{src},h} - P_{\mathrm{tar},h}\big)(s'|s,a)\right| \\
&\leq H\,\Delta_h(s,a),
\end{aligned}$$

using $\|P-Q\|_1 = 2\|P-Q\|_{\mathrm{TV}}$ and $\|f_{h+1} - H/2\|_\infty \leq H/2$.

Statistical term. The source-fitting constraint is a population statement, namely

$$\sum_{s,a} d^{\mu}_{\mathcal{M}_{\mathrm{src}},h}(s,a)\,\varrho_h(s,a)^2 = \mathcal{E}^{\mu,\pi}_{\mathcal{M}_{\mathrm{src}}}(f)_h \leq \epsilon.$$

Hence, pointwise, $\varrho_h(s,a)^2 \leq \epsilon/d^{\mu}_{\mathcal{M}_{\mathrm{src}},h}(s,a)$ wherever $d^{\mu} > 0$. Moreover $|\varrho_h(s,a)| \leq 2H$ always, since $f_h \in [0,H]$ and $\mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{src}},h} f_{h+1} \in [0, 1+H]$. Therefore $\varrho_h(s,a)^2 \leq \mathrm{Stat}_h(s,a;\epsilon)$ at every (s,a).

Combining via $(x+y)^2 \leq 2x^2 + 2y^2$ and averaging over $(s,a) \sim d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h}$,

$$\mathcal{E}^{\nu,\pi}_{\mathcal{M}_{\mathrm{tar}}}(f)_h \leq 2\sum_{s,a} d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h}(s,a)\left[\mathrm{Stat}_h(s,a;\epsilon) + H^2\Delta_h(s,a)^2\right].$$

Taking the supremum over $f \in \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon)$ proves the theorem. □

Remark 13 (Localized, count-based statistical term)**.** *A local confidence set can retain the dependence on individual visit counts, but its radius must also control the continuation class. Here is an explicit construction for known rewards and the unrestricted tabular class. Write $S = |\mathcal{S}|$, $A = |\mathcal{A}|$, $N = n_h(s,a)$, and let $\widehat{P}_h(\cdot|s,a)$ be the empirical source kernel. For $N \geq 1$, set*

$$b_h(s,a) = H\min\left\{1, \sqrt{\frac{S\log 2 + \log(2HSA/\delta)}{2N}}\right\}.$$

Conditional on the number of visits to a fixed pair at a fixed stage, the observed successors are independent samples from its kernel. Hoeffding's inequality applied to every subset of successor states, followed by a union bound over pairs and stages, gives, with probability at least $1-\delta$,

$$\sup_{g:\mathcal{S}\to[0,H]} |(\widehat{P}_h - P_{\mathrm{src},h})g|(s,a) = H\|\widehat{P}_h - P_{\mathrm{src},h}\|_{\mathrm{TV}} \le b_h(s,a)$$

simultaneously. This event covers data-dependent continuations as well.

Require the absolute empirical Bellman residual at each observed pair to be at most $b_h(s,a)$. The source value function is feasible on this event, and every feasible critic has absolute population source residual at most $2b_h(s,a)$. Leave unobserved pairs unconstrained and set

$$\mathrm{Stat}_h^{\mathrm{loc}}(s,a) = \begin{cases} 4b_h(s,a)^2, & N \ge 1, \\ 4H^2, & N = 0. \end{cases}$$

The proof of Theorem 34 then bounds every critic in this local *set by the same target-weighted sum with $\mathrm{Stat}^{\mathrm{loc}}$ replacing $\mathrm{Stat}(\epsilon)$. This is a separate confidence-set construction; it does not change the population modulus or the algorithms analyzed above.*

For a fixed finite class $\mathcal{G}_h \subseteq [0,H]^{\mathcal{S}}$ containing every continuation $f_{h+1}(\cdot,\pi_{h+1})$ under consideration, a direct union bound replaces $S\log 2$ by $\log\max_h|\mathcal{G}_h|$. A covering argument requires the corresponding approximation term. In particular, unrestricted tabular continuations give a squared radius of order $H^2(S+\log(HSA/\delta))/N$, capped at $4H^2$, rather than a dimension-free $\widetilde{O}(H^2/N)$. The population structured bounds themselves need no such concentration argument.

Corollary 35 (Single-environment reduction)**.** *If $\mathcal{M}_{\mathrm{src}} = \mathcal{M}_{\mathrm{tar}} = \mathcal{M}$ and*

$$d^{\pi}_{\mathcal{M},h}(s,a) \le C_{\mathrm{conc}}(\mu,\pi) d^{\mu}_{\mathcal{M},h}(s,a)$$

for all (s,a,h), then

$$\delta_h^{\pi}(\epsilon) \le C_{\mathrm{conc}}(\mu,\pi)\epsilon.$$

Proof. When the source and target MDPs coincide, $\Delta_h(s,a) = 0$. For any $f \in \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon)$,

$$\mathcal{E}^{\pi,\pi}_{\mathcal{M}}(f)_h = \mathbb{E}_{d^{\pi}_{\mathcal{M},h}}[(f_h - \mathcal{T}^{\pi}_{\mathcal{M},h} f_{h+1})^2].$$

By concentrability,

$$\mathcal{E}^{\pi,\pi}_{\mathcal{M}}(f)_h \le C_{\mathrm{conc}}(\mu,\pi)\mathbb{E}_{d^{\mu}_{\mathcal{M},h}}[(f_h - \mathcal{T}^{\pi}_{\mathcal{M},h} f_{h+1})^2] \le C_{\mathrm{conc}}(\mu,\pi)\epsilon.$$

Taking the supremum over f gives the result. □

F.2.2 Linear MDPs

Assume a common feature map $\varphi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ with $\|\varphi(s,a)\|_2 \le 1$, shared linear rewards $r_h(s,a) = \langle\varphi(s,a),\theta_h\rangle$, and linear transition representations

$$P_{\mathcal{M},h}(\cdot|s,a) = \langle\varphi(s,a),\mu_{\mathcal{M},h}(\cdot)\rangle.$$

Let, as in Section 5.1,

$$\Lambda^{\mu}_h = \mathbb{E}_{(s,a)\sim d^{\mu}_{\mathcal{M}_{\mathrm{src}},h}}[\varphi(s,a)\varphi(s,a)^{\top}],$$

$$\Delta^{\mathrm{lin}}_h = \frac{1}{\sqrt{d}} \sup_{\|g\|_\infty\le 1} \left\| \int g(s')\, d\big(\mu_{\mathrm{src},h} - \mu_{\mathrm{tar},h}\big)(s') \right\|_2 .$$

Theorem 36 (Linear MDP weak-modulus bound)**.** *Under the linear MDP assumptions above, suppose every $f \in \mathcal{F}$ is linear, $f_h = \langle\varphi(\cdot,\cdot), w_{f,h}\rangle$, with values in $[0,H]$ and $\|w_{f,h}\|_2 \le W$, and adopt the standard normalizations $\|\theta_h\|_2 \le \sqrt{d}$ and $\|\int g\, d\mu_{\mathcal{M},h}\|_2 \le \sqrt{d}\,\|g\|_\infty$ (Jin et al., 2020). Let $P_h := (\Lambda^{\mu}_h)^{\dagger}\Lambda^{\mu}_h$ be the orthogonal projector onto the range of Λ^{μ}_h, let $P_{h,\perp} := I - P_h$, and set $B_h := W + (1+H)\sqrt{d}$. Define the target feature covariance and the target-weighted shift*

$$\Sigma^{\nu}_h := \mathbb{E}_{(s,a)\sim d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h}}\big[\varphi(s,a)\varphi(s,a)^{\top}\big],$$

$$\Delta^{\mathrm{lin},\nu}_h := \frac{1}{\sqrt{d}} \sup_{\|g\|_\infty\le 1} \left\| (\Sigma^{\nu}_h)^{1/2} \int g(s')\, d\big(\mu_{\mathrm{src},h} - \mu_{\mathrm{tar},h}\big)(s') \right\|_2 \le \Delta^{\mathrm{lin}}_h,$$

where the inequality holds because $\|(\Sigma_h^\nu)^{1/2}\|_{\mathrm{op}} \le 1$ (features have norm at most one). Then for any policies ν, π, step h, and $\epsilon \ge 0$,

$$\delta_h^{\nu|\pi}(\epsilon) \le 3\,\epsilon\,\mathbb{E}_{(s,a)\sim d^\nu_{\mathcal{M}_{\mathrm{tar},h}}}\|\varphi(s,a)\|^2_{(\Lambda_h^\mu)^\dagger} + 3B_h^2\,\mathbb{E}_{(s,a)\sim d^\nu_{\mathcal{M}_{\mathrm{tar},h}}}\|P_{h,\perp}\varphi(s,a)\|_2^2 + 3H^2 d(\Delta_h^{\mathrm{lin},\nu})^2.$$

Since $\Delta_h^{\mathrm{lin},\nu} \le \Delta_h^{\mathrm{lin}}$, the same display holds with the global shift Δ_h^{lin}, which is the form quoted in the main text. If $\lambda_{\min}(\Lambda_h^\mu) > 0$, then $P_{h,\perp} = 0$, the constant improves to 2, and the bound is at most $2\epsilon/\lambda_{\min}(\Lambda_h^\mu) + 2H^2 d(\Delta_h^{\mathrm{lin}})^2$. Taking $\nu = \pi$ gives the on-policy bound $\delta_h^\pi(\epsilon) \le 3\epsilon\,\mathbb{E}_{d^\pi_{\mathcal{M}_{\mathrm{tar},h}}}\|\varphi\|^2_{(\Lambda_h^\mu)^\dagger} + 3B_h^2\,\mathbb{E}_{d^\pi_{\mathcal{M}_{\mathrm{tar},h}}}\|P_{h,\perp}\varphi\|_2^2 + 3H^2 d(\Delta_h^{\mathrm{lin}})^2$.

Proof. For any continuation function f_{h+1} and $\mathcal{M} \in \{\mathcal{M}_{\mathrm{src}}, \mathcal{M}_{\mathrm{tar}}\}$, the Bellman backup in a linear MDP is linear in the features, taking the form

$$\mathcal{T}^\pi_{\mathcal{M},h} f_{h+1}(s,a) = \langle \varphi(s,a), w^{\pi,f}_{\mathcal{M},h}\rangle, \qquad w^{\pi,f}_{\mathcal{M},h} = \theta_h + \int f_{h+1}(s', \pi_{h+1})\, d\mu_{\mathcal{M},h}(s').$$

Fix $f \in \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon)$ with coefficient $w_{f,h}$ at step h.

Statistical term. Since both f_h and the source backup are linear, the population source constraint is exactly a covariance-weighted quadratic, namely

$$\mathcal{E}^{\mu,\pi}_{\mathcal{M}_{\mathrm{src}}}(f)_h = \mathbb{E}_{(s,a)\sim d^\mu_{\mathcal{M}_{\mathrm{src},h}}}\big\langle \varphi(s,a),\, w_{f,h} - w^{\pi,f}_{\mathrm{src},h}\big\rangle^2 = \big\|w_{f,h} - w^{\pi,f}_{\mathrm{src},h}\big\|^2_{\Lambda_h^\mu} \le \epsilon.$$

No regression or concentration argument is needed, since the constraint itself is the population statement. Write $e := w_{f,h} - w^{\pi,f}_{\mathrm{src},h}$ and $\Lambda := \Lambda_h^\mu$, and decompose $\langle \varphi, e\rangle = \langle \varphi, P_h e\rangle + \langle \varphi, P_{h,\perp} e\rangle$. For the range component, $P_h = \Lambda^{\dagger/2}\Lambda^{1/2}$ (functional calculus on the common eigenbasis), so

$$\big\langle \varphi(s,a), P_h e\big\rangle^2 = \big\langle \Lambda^{\dagger/2}\varphi(s,a),\, \Lambda^{1/2} e\big\rangle^2 \le \|\varphi(s,a)\|^2_{\Lambda^\dagger}\,\|e\|^2_\Lambda \le \epsilon\,\|\varphi(s,a)\|^2_{(\Lambda_h^\mu)^\dagger}.$$

For the null component, $\langle \varphi, P_{h,\perp} e\rangle = \langle P_{h,\perp}\varphi, e\rangle$, and the normalizations give $\|e\|_2 \le \|w_{f,h}\|_2 + \|\theta_h\|_2 + \|\int f_{h+1}(\cdot,\pi_{h+1})\, d\mu_{\mathrm{src},h}\|_2 \le W + (1+H)\sqrt{d} = B_h$, so

$$\big\langle \varphi(s,a), P_{h,\perp} e\big\rangle^2 \le B_h^2\,\|P_{h,\perp}\varphi(s,a)\|_2^2.$$

Note that $\|e\|^2_\Lambda \le \epsilon$ constrains only the range component, which is why the null component must be charged at the worst-case ambiguity B_h; the lower bound in Appendix G.2.3 shows this is not an artifact.

Shift term. Since rewards are shared,

$$w^{\pi,f}_{\mathrm{src},h} - w^{\pi,f}_{\mathrm{tar},h} = \int f_{h+1}(s', \pi_{h+1})\, d(\mu_{\mathrm{src},h} - \mu_{\mathrm{tar},h})(s'),$$

and since $\|f_{h+1}(\cdot,\pi_{h+1})\|_\infty \le H$, the definition of Δ_h^{lin} gives $\|w^{\pi,f}_{\mathrm{src},h} - w^{\pi,f}_{\mathrm{tar},h}\|_2 \le H\sqrt{d}\,\Delta_h^{\mathrm{lin}}$. Averaging the squared shift component over $(s,a) \sim d^\nu_{\mathcal{M}_{\mathrm{tar},h}}$ gives an exact covariance-weighted norm rather than a pointwise bound,

$$\mathbb{E}_{(s,a)\sim d^\nu_{\mathcal{M}_{\mathrm{tar},h}}}\big\langle \varphi(s,a), w^{\pi,f}_{\mathrm{src},h} - w^{\pi,f}_{\mathrm{tar},h}\big\rangle^2 = \big\|w^{\pi,f}_{\mathrm{src},h} - w^{\pi,f}_{\mathrm{tar},h}\big\|^2_{\Sigma_h^\nu} \le H^2 d(\Delta_h^{\mathrm{lin},\nu})^2,$$

where the inequality applies the definition of $\Delta_h^{\mathrm{lin},\nu}$ with test function $g = f_{h+1}(\cdot,\pi_{h+1})/H$, since $w^{\pi,f}_{\mathrm{src},h} - w^{\pi,f}_{\mathrm{tar},h} = H\int g(s')\, d(\mu_{\mathrm{src},h} - \mu_{\mathrm{tar},h})(s')$; the pointwise bound $\|\varphi\|_2 \le 1$ recovers the global form with Δ_h^{lin}.

The target residual at each (s,a) is the sum of the two statistical components and the shift component, so $(x+y+z)^2 \le 3(x^2+y^2+z^2)$; averaging over $(s,a) \sim d^\nu_{\mathcal{M}_{\mathrm{tar},h}}$, the two statistical components give the displayed expectations and the shift component gives the exact covariance norm above, which yields the claim. When Λ_h^μ has full rank the null component is zero, the split needs only two terms with constant 2, and $\|\varphi(s,a)\|^2_{(\Lambda_h^\mu)^\dagger} \le 1/\lambda_{\min}(\Lambda_h^\mu)$ gives the simplified form. Taking the supremum over f completes the proof. □

F.2.3 BLOCK MDPS

Suppose

$$\mathcal{S} \times \mathcal{A} = \bigcup_{i=1}^{\kappa} B_i$$

is a disjoint partition. Define

$$p^{\pi}_{h,i} = \Pr^{\pi}_{\mathcal{M}_{\mathrm{tar}}}[(s_h, a_h) \in B_i],$$

and

$$\Delta^{\mathrm{blk}}_{h,i} = \sup_{(s,a)\in B_i} \|P_{\mathrm{src},h}(\cdot|s,a) - P_{\mathrm{tar},h}(\cdot|s,a)\|_{\mathrm{TV}}.$$

Theorem 37 (Block MDP weak-modulus bound)**.** *Suppose rewards and transitions are homogeneous within each block in both environments, and every $f \in \mathcal{F}$ is block-measurable (constant on each B_i at every step) with values in $[0, H]$. Let $q^{\mu}_{h,i} := \Pr^{\mu}_{\mathcal{M}_{\mathrm{src}}}[(s_h, a_h) \in B_i]$. Then for any policies ν, π, step h, and $\epsilon \geq 0$,*

$$\delta^{\nu|\pi}_h(\epsilon) \leq 2\sum_{i=1}^{\kappa} p^{\nu}_{h,i}\left[H^2(\Delta^{\mathrm{blk}}_{h,i})^2 + \mathrm{Stat}_{h,i}(\epsilon)\right], \qquad \mathrm{Stat}_{h,i}(\epsilon) := \min\left\{4H^2, \frac{\epsilon}{q^{\mu}_{h,i}}\right\},$$

and, when each block is either unchanged or arbitrarily shifted,

$$\delta^{\nu|\pi}_h(\epsilon) \leq 2\sum_{i=1}^{\kappa} p^{\nu}_{h,i}\left[4H^2\,\mathbf{1}\{B_i \text{ is shifted at step } h\} + \mathrm{Stat}_{h,i}(\epsilon)\right].$$

Taking $\nu = \pi$ gives the on-policy bound $\delta^{\pi}_h(\epsilon) \leq 2\sum_i p^{\pi}_{h,i}[H^2(\Delta^{\mathrm{blk}}_{h,i})^2 + \mathrm{Stat}_{h,i}(\epsilon)]$.

Proof. The proof repeats the tabular argument at the block level. Fix $f \in \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon)$ and write $g(s') := f_{h+1}(s', \pi_{h+1}) = \sum_{a'} \pi_{h+1}(a'|s') f_{h+1}(s', a')$ for the continuation value. The source residual $\varrho_h = f_h - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{src}},h} f_{h+1}$ is constant on each block, for two separate reasons. First, f_h is block-constant by assumption. Second, on block B_i the backup is $\mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{src}},h} f_{h+1}(s,a) = r_{h,i} + \int g\, dP^{\mathrm{src}}_{h,i}$, and by within-block homogeneity both the reward $r_{h,i}$ and the kernel $P^{\mathrm{src}}_{h,i}$ are the same for every $(s,a) \in B_i$, so the backup takes a single value on B_i *whatever* the function g is. In particular, no block structure is required of the continuation g or of the policy π; the continuation enters only integrated against the block-common kernel. Write $\varrho_{h,i}$ for the residual's value on B_i. The source constraint reads

$$\sum_{i=1}^{\kappa} q^{\mu}_{h,i}\, \varrho^2_{h,i} \leq \epsilon \qquad \Longrightarrow \qquad \varrho^2_{h,i} \leq \min\left\{4H^2,\ \epsilon/q^{\mu}_{h,i}\right\} = \mathrm{Stat}_{h,i}(\epsilon),$$

where the range bound $|\varrho_{h,i}| \leq 2H$ holds as in the tabular proof. The Bellman-operator difference is uniformly bounded over B_i by $H\Delta^{\mathrm{blk}}_{h,i}$, by the same centering argument as in Theorem 34. Combining via $(x+y)^2 \leq 2x^2 + 2y^2$ and averaging over $(s,a) \sim d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h}$, which weights block B_i by $p^{\nu}_{h,i}$, yields the first claim; the indicator form follows from $\Delta^{\mathrm{blk}}_{h,i} = 0$ on unchanged blocks and $(\Delta^{\mathrm{blk}}_{h,i})^2 \leq 4$ on shifted blocks. Taking the supremum over f completes the proof. □

F.2.4 SHIFTED REWARDS

Proof of Proposition 8. Fix $f \in \mathcal{F}^{\mu,\pi}_{\mathrm{src}}(\epsilon)$. As in the proof of Theorem 34, the source residual satisfies $\varrho^{\mathrm{src}}_h(s,a)^2 \leq \mathrm{Stat}_h(s,a;\epsilon)$ pointwise. The operator difference splits into a reward part and a transition part,

$$\left[\mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{src}},h} - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{tar}},h}\right] f_{h+1}(s,a) = \left(r^{\mathrm{src}}_h - r^{\mathrm{tar}}_h\right)(s,a) + \int f_{h+1}(s', \pi_{h+1})\, d\left(P_{\mathrm{src},h} - P_{\mathrm{tar},h}\right)(s'|s,a),$$

whose terms are bounded in absolute value by $\Delta r_h(s,a)$ and $H\Delta_h(s,a)$ respectively, the latter by the centering argument of the tabular proof. The target residual is the sum of the source residual and the operator difference, so

$$\varrho^{\mathrm{tar}}_h(s,a)^2 \leq 3\left[\mathrm{Stat}_h(s,a;\epsilon) + \Delta r_h(s,a)^2 + H^2\Delta_h(s,a)^2\right]$$

pointwise, by $(x+y+z)^2 \le 3(x^2+y^2+z^2)$. Averaging over $(s,a) \sim d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h}$ and taking the supremum over feasible f proves the tabular claim. The linear and block cases repeat their respective arguments with the reward-discrepancy term carried through the operator difference in the same way. □

F.2.5 Multiple sources

Proof of Proposition 33. (i) The feasible set of (28) is contained in each single-source feasible set, and a supremum over a subset is no larger. If the intersection is empty the left side is zero and the claim is trivial.

(ii) Fix $f \in \bigcap_j \mathcal{F}^{\mu_j,\pi}_{\mathrm{src},j}(\epsilon_j)$ and a pair (s,a). For each source j, the tabular argument of Theorem 34 applies verbatim with source j in place of the single source. The source-j residual obeys $\varrho^{(j)}_h(s,a)^2 \le \mathrm{Stat}^{(j)}_h(s,a;\epsilon_j)$ pointwise, from the constraint $\sum_{s,a} q^{\mu_j}_h(s,a)\,\varrho^{(j)}_h(s,a)^2 \le \epsilon_j$ together with the range bound, and the operator discrepancy obeys $|\mathcal{T}^{\pi}_{\mathcal{M}^{(j)}_{\mathrm{src}},h} f_{h+1} - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{tar}},h} f_{h+1}|(s,a) \le H\Delta^{(j)}_h(s,a)$ by centering. Writing the target residual through source j and using $(x+y)^2 \le 2x^2+2y^2$,

$$\begin{aligned}\varrho^{\mathrm{tar}}_h(s,a)^2 &= \Big(\varrho^{(j)}_h(s,a) + \big[\mathcal{T}^{\pi}_{\mathcal{M}^{(j)}_{\mathrm{src}},h} - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathrm{tar}},h}\big] f_{h+1}(s,a)\Big)^2 \\ &\le 2\,\mathrm{Stat}^{(j)}_h(s,a;\epsilon_j) + 2H^2\Delta^{(j)}_h(s,a)^2.\end{aligned}$$

The left side does not depend on j, and the inequality holds for every j simultaneously because f lies in all J fitting sets, so the right side may be replaced by twice the minimum over $j \in [J]$, pair by pair. Averaging over $(s,a) \sim d^{\nu}_{\mathcal{M}_{\mathrm{tar}},h}$ and taking the supremum over feasible f yields (28). The block version repeats the argument at the block level exactly as in Theorem 37, with the minimum taken per block. □

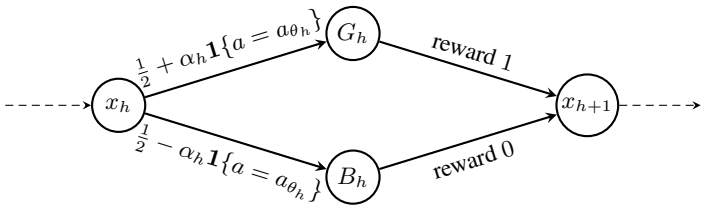
Corollary 38 (Selected-subset guarantee)**.** *Work on the intersection of the J per-source validity events (each an instance of Assumption 3 for its own source, total failure probability at most $J\delta$), on which, for* every *source j simultaneously, $\widehat{\mathcal{F}}_{\mathrm{src},j}(\beta_j,\pi) \subseteq \mathcal{F}^{\mu_j,\pi}_{\mathrm{src},j}(\epsilon_j)$ and $Q^{\pi}_{\mathcal{M}^{(j)}_{\mathrm{src}}} \in \widehat{\mathcal{F}}_{\mathrm{src},j}(\beta_j,\pi)$ for all $\pi \in \overline{\Pi}$. Let $S_t \subseteq [J]$ be* any *data-dependent subsets selected by the learner, for instance the largest subsets whose empirical sets still intersect, subject only to the selected empirical intersections being nonempty at the visited rounds. Then Algorithms 2 and 3, run with the intersected sets over S_t, satisfy Theorems 1 and 21 with $\delta^{S_t}_{h,\mathrm{multi}}$, the multi-source modulus of* (28) *restricted to the sources in S_t, in place of δ_h at each round. No union bound over the 2^J subsets is needed. The per-source inclusions hold simultaneously for every j on the single event above, so for any realized S_t the selected critic lies in $\bigcap_{j\in S_t} \mathcal{F}^{\mu_j,\pi}_{\mathrm{src},j}(\epsilon_j)$, which is exactly the feasible set of the restricted modulus; the data select the subset, they do not alter the event.*

G Lower Bounds

G.1 Statements and discussion

The upper bounds above show that hybrid transfer is controlled by the weak transfer modulus. We now show that this dependence is necessary. Source data may be compatible with multiple target environments that require different optimal actions. In such cases, any algorithm must either collect enough target information to resolve the ambiguity or suffer suboptimality. The weak modulus therefore reflects an information-theoretic difficulty of the problem, beyond its role as a complexity term in the analysis.

The bound (12) of Theorem 7 holds for a universal constant $c > 0$. Moreover, for all n sufficiently large (depending on the instance), $\delta^{\pi^\star_{\mathrm{tar}}}_h(\epsilon_{\mathrm{off}}(n)) \le 2\,\delta^{\pi^\star_{\mathrm{tar}}}_h(0)$ at every decision step. At the branch steps $\delta^{\pi^\star_{\mathrm{tar}}}_j(0) = 0$ and $\delta^{\pi^\star_{\mathrm{tar}}}_j(\epsilon) \le 2K\epsilon$, because branch transitions coincide in the two environments and each branch pair has behavior probability $1/(2K)$. Hence, once also $\epsilon_{\mathrm{off}}(n) \le c\min_h \alpha_h^2/K$, the sum over all steps of $(\delta^{\pi^\star_{\mathrm{tar}}}_j(\epsilon_{\mathrm{off}}(n)))^{1/2}$ is at most a constant times $\sum_h \alpha_h$, and the bound holds at radius $\epsilon_{\mathrm{off}}(n)$ up to a constant. In particular, the comparator weak-modulus term cannot



in the source MDP every action reaches G_h and B_h with probability $\frac{1}{2}$ each

Figure 5: One layer of the general lower-bound construction (target instance θ). All K actions at x_h behave identically in the source, while in the target the hidden action a_{θ_h} reaches the rewarding branch with probability elevated by α_h. Rewards are shared, transitions out of G_h and B_h are deterministic, and layers compose in series, so per-layer losses add and $\delta_h^{\pi^\star_{\mathrm{tar}}}(0) = \alpha_h^2$ exactly.

be removed, nor replaced by any uniformly smaller function of the modulus, in the following sense. The information-theoretic argument constrains every algorithm regardless of any class, and on these instances every critic in the source-fitting set at radius zero agrees with $Q^\pi_{\mathcal{M}_{\mathrm{src}}}$ on the layered support, for every subclass of the stepwise class containing that source value function. Zero source residual forces $f_h = \mathcal{T}^\pi_{\mathcal{M}_{\mathrm{src}},h} f_{h+1}$ at every stage-h pair the behavior policy reaches, and backward induction along the chain pins f there uniquely; coordinates at states belonging to other layers carry zero occupancy at that stage in *both* environments, so they are unconstrained but enter no source or target residual, and with stage-specific state spaces $\mathcal{S}_h$ and $\mathcal{F}_h = [0, H-h+1]^{\mathcal{S}_h \times \mathcal{A}}$ the fitting set is literally the singleton $\{Q^\pi_{\mathcal{M}_{\mathrm{src}}}\}$. Hence $\delta_h^{\pi^\star_{\mathrm{tar}}}(0) = \alpha_h^2$ *exactly* for every such subclass. The necessity claim therefore covers the population framework and every such class. The stepwise hard class is moreover an *algorithmic* instance as it stands. The construction's state-action space is finite, so each $\mathcal{F}_h = [0, H-h+1]^{\mathcal{S}\times\mathcal{A}}$ has finite sup-norm covering numbers, $\log N_\alpha \le |\mathcal{S}||\mathcal{A}| \log(1 + H/\alpha)$, and the class is realizable and Bellman complete by construction; the covering-number version of the upper bounds (Proposition 17) therefore applies to this class verbatim, and the lower bound is an end-to-end statement against the same algorithmic guarantees, not only against the population framework. Moreover, for any fixed interaction budget (T, T_{off}) and the algorithmic closure of Section 2, the hard class may be chosen *finite* and Bellman complete while containing all witnesses (Corollary 40 in Appendix G.2.1), so the necessity statement matches the finite-class algorithmic guarantees verbatim; the finite class is enormous, but its cardinality is immaterial to the information-theoretic argument.

The construction behind Theorem 7 uses shared rewards and differs only in transition dynamics. Source data is arbitrarily abundant yet structurally uninformative. At each decision layer, all actions have identical source transitions, while in the target family one hidden action is slightly better, by margin α_h in value. Under the full-support behavior policy, every zero-residual source-consistent critic agrees with the source value function $Q^\pi_{\mathcal{M}_{\mathrm{src}}}$ on the layered support relevant to both residuals (literal uniqueness holds in the stage-specific-space formulation), and its target Bellman backup differs from the source backup by exactly α_h at the decision pair, which is the statement $\delta_h^{\pi^\star_{\mathrm{tar}}}(0) = \alpha_h^2$, and with $\alpha_h \lesssim \sqrt{K/T}$ the target budget is information-theoretically insufficient to identify the good action at every layer. The value lost per misidentified layer is $\Theta(\alpha_h) = \Theta(\sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}}(0)})$, matching the square-root dependence in the upper bounds. In the constructed instances $\delta_h^{\pi^\star_{\mathrm{tar}}|\pi}(0) = \delta_h^{\pi^\star_{\mathrm{tar}}}(0)$ for *every* backup policy π, so the lower bound matches the comparator term of Theorem 1 in both its two-policy and on-policy forms. Figure 5 depicts one layer of the construction; the full proof is given in Appendix G.2.1.

We also prove structured lower bounds matching the shift-dependent terms of the explicit upper bounds in Section 5.1; we do not claim matching in all parameters simultaneously. In tabular MDPs the sharpest statement is a *direct sum* (Theorem 41). For any decision layers with visitation weights p_k, downstream value scales L_k obeying the per-episode budget $\sum_k (L_k + 2) \le H$, and local shifts

$\Delta_k \lesssim 1/\sqrt{T p_k}$, there are instances on which any algorithm suffers

$$\mathbb{E}[\mathrm{Regret}(T)] \;\geq\; c\,\min\Big\{HT,\; \sum_k T\,p_k\,L_k\,\Delta_k\Big\},$$

each shifted layer charging its visitation times its downstream scale times its shift, added across layers up to the maximal regret. The single-layer specialization with $L_1 = \Theta(H)$, stated next in the relaxed square-root form for comparison with the modulus bounds, is what the following display records. For every target budget T there are instances, constructed given T, whose shifted pairs lie at a single decision layer controlling a downstream branch of value scale $\Theta(H)$ and whose local shifts satisfy the testing constraint $\Delta_h(s,a) \leq c'/\sqrt{T\, d^{\pi^\star_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar}},h}(s,a)}$, on which any algorithm interacting with the target for T episodes satisfies

$$\mathbb{E}[\mathrm{Regret}(T)] \geq c\sqrt{T\sum_{h=1}^{H} H^2 \sum_{s,a} d^{\pi^\star_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar}},h}(s,a)\Delta_h(s,a)^2}, \tag{30}$$

up to logarithmic and problem-normalization factors. The constraint on Δ_h delimits the regime in which the shift cannot be fully identified within T episodes; outside this regime the shift is learnable and cumulative regret saturates for this instance class, so no bound of the form (30) can hold for all T on a fixed instance. Within the regime, the target-weighted dynamics-shift term in Theorem 6 cannot be removed. The value budget in the direct-sum form above is essential, since $\Theta(H)$ downstream gaps at many layers simultaneously are impossible within episodes of value at most H, and an unbudgeted multi-layer form would exceed the maximal regret. For linear MDPs, we construct instances where any hybrid algorithm with a bounded target budget suffers target suboptimality at least

$$\mathrm{SubOpt}_{\mathrm{tar}}(\mathsf{ALG}) \geq cH\left(\sqrt{\epsilon_{\mathrm{stat}}} + \sqrt{d}\,\Delta^{\mathrm{lin}}\right), \tag{31}$$

with constant probability, and the two terms realize the two source-side terms of Theorem 36. The ϵ_{stat} term is *structural*. It is realized by instances whose behavior policy leaves a feature direction entirely unexcited, so that the source data distribution is identical across the family for every sample size n and the ambiguous value functions lie in a common source-fitting set at radius zero; this is the null-space term $B^2\,\mathbb{E}\|P_\perp\varphi\|_2^2$ of the upper bound, and the floor persists for every n. Directions that are weakly covered, rather than uncovered, behave differently on both sides. A companion lower bound (part 2 of the detailed statement) quantifies the finite-sample price when the ambiguous direction carries covariance eigenvalue $\lambda > 0$, exhibiting instances with unavoidable suboptimality $\Omega\big(H\min\{1,(n\lambda+T)^{-1/2}\}\big)$, and the absence of any persistent floor under full coverage follows from the upper bounds, since Theorems 1 and 36 give suboptimality $O(\sum_h \sqrt{\epsilon_{\mathrm{off}}(n)/\lambda_{\min}})$ plus an optimization remainder, both vanishing under matched dynamics when the offline iteration count grows subcritically relative to the source sample size, $T_{\mathrm{off}}\log(T_{\mathrm{off}}|\mathcal{F}|H/\delta) = o(n)$ in the finite-class case, since the statistical radius carries the iteration-dependent class term. Persistence is therefore a structural noncoverage phenomenon, and both structural noncoverage and latent transition-operator shift are unavoidable. These lower bounds are proved in Appendices G.2.2 and G.2.3.

The lower bounds show that source data can leave target-relevant decisions unresolved even when source Bellman fitting is exact. On the constructed families and within the stated target-sample regimes, the square-root modulus dependence cannot be uniformly improved. This is a worst-case statement over those families, not a condition that every useful source or successful learner must satisfy. The constructions also illustrate why the location of a dynamics discrepancy and its effect on the relevant values matter.

G.2 Proofs of the lower bounds

We now prove the lower bounds.

G.2.1 General weak-modulus lower bound

Theorem 39 (General lower bound, detailed version)**.** *Fix $H \geq 2$, $n, T \in \mathbb{N}$, $K \geq 2$, and levels $\alpha_h \in (0, c_0\sqrt{K/T}\,]$ for each decision layer h, with $\sum_h \alpha_h \leq 1/2$ and c_0 a sufficiently small universal constant. There is a family of hybrid instances with shared rewards, a common behavior*

policy μ, and function class $\mathcal{F}$ given stepwise by all functions $\mathcal{F}_h = \{f_h : \mathcal{S} \times \mathcal{A} \to [0, H-h+1]\}$, which satisfies Assumptions 1 and 2 at every policy since backups of $[0, H-h]$-valued functions with rewards in $[0,1]$ land in $[0, H-h+1]$, such that each instance satisfies $\delta_h^{\pi^\star_{\mathrm{tar}}}(0) = \alpha_h^2$ at every decision layer h, and for any algorithm ALG *using n source episodes and T target episodes,*

$$\mathbb{E}\big[\,\mathrm{SubOpt}_{\mathrm{tar}}(\mathsf{ALG})\big] \geq c \sum_h \sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}}(0)},$$

where the expectation is over a uniform prior on the instance family, the data, and the algorithm's randomization, and $c > 0$ is universal. Moreover, on these instances, $\delta_h^{\pi^\star_{\mathrm{tar}}}(\epsilon) \leq (\alpha_h + 2\sqrt{\epsilon / d_{\min}^{\mu}})^2$ where $d_{\min}^{\mu}$ is the minimum source visitation probability of the layer-h decision pairs, so $\delta_h^{\pi^\star_{\mathrm{tar}}}(\epsilon_{\mathrm{off}}(n)) \leq 2\,\delta_h^{\pi^\star_{\mathrm{tar}}}(0)$ once n is large enough that $\epsilon_{\mathrm{off}}(n) \leq c_1 \alpha_h^2 d_{\min}^{\mu}$.

Proof. Construction. The state space is a chain $x_1, G_1, B_1, x_2, G_2, B_2, \ldots, x_L, G_L, B_L, x_{L+1}$ with $L = \lfloor H/2 \rfloor$ decision layers; layer h occupies steps $2h-1$ (decision state x_h) and $2h$ (branch states G_h, B_h), and x_{L+1} is absorbing with zero reward. At x_h there are K actions; from G_h and B_h every action moves deterministically to x_{h+1}. Rewards are shared between source and target, and every reward is zero except $r(G_h, \cdot) = 1$.

In the source MDP, every action at x_h moves to G_h with probability $1/2$ and to B_h with probability $1/2$. The target family is indexed by $\theta = (\theta_1, \ldots, \theta_L) \in [K]^L$: in instance θ, action a_{θ_h} at x_h moves to G_h with probability $1/2 + \alpha_h$, and every other action with probability $1/2$. The behavior policy μ is uniform. Note $\pi^\star_{\mathrm{tar}}$ plays a_{θ_h} at x_h, and choosing any other action there loses exactly α_h in value; since layer transitions always reach x_{h+1}, the losses across layers add, so a policy π suffers

$$\mathrm{SubOpt}_{\mathrm{tar}}(\pi) = \sum_{h=1}^{L} \alpha_h \left(1 - \pi(a_{\theta_h} \,|\, x_h)\right). \tag{32}$$

Modulus computation. With $\mathcal{F}$ the full tabular class, source Bellman consistency at tolerance 0 pins f on the layered support. The deterministic transitions from G_h, B_h force $f(G_h, \cdot) = 1 + f(x_{h+1}, \pi)$ and $f(B_h, \cdot) = f(x_{h+1}, \pi)$, and the constraint at x_h forces $f(x_h, a) = \frac{1}{2} f(G_h, \pi) + \frac{1}{2} f(B_h, \pi)$ for every a. Hence every member of $\mathcal{F}_{\mathrm{src}}^{\mu,\pi}(0)$ agrees with $Q^\pi_{\mathcal{M}_{\mathrm{src}}}$ at every pair the chain occupies at each stage; coordinates at states belonging to other layers have zero occupancy at that stage in both environments and enter no residual, so they are unconstrained and irrelevant, and with stage-specific state spaces $\mathcal{S}_h$ the set is literally the singleton. Writing f for any such critic, its target residual at (x_h, a_{θ_h}), at the decision step $2h-1$, depends only on the pinned coordinates and equals

$$f(x_h, a_{\theta_h}) - \mathcal{T}^\pi_{\mathcal{M}_{\mathrm{tar}}, 2h-1} f(x_h, a_{\theta_h}) = -\alpha_h \big(f(G_h, \pi) - f(B_h, \pi)\big) = -\alpha_h,$$

and $\pi^\star_{\mathrm{tar}}$ visits (x_h, a_{θ_h}) with probability one, so $\delta_h^{\pi^\star_{\mathrm{tar}}}(0) = \alpha_h^2$ at each decision layer (and 0 elsewhere). For positive tolerance, value errors can accumulate along the chain; we instead bound the local residuals directly. Fix a feasible critic and its backup policy π. Let $e_x(a)$ be its source residual at (x_h, a), and $e_G(a), e_B(a)$ its source residuals at the next-stage branch pairs. Since both branches move to the same state,

$$f(G_h, \pi) - f(B_h, \pi) = 1 + \bar{e}_G - \bar{e}_B, \qquad \bar{e}_G = \sum_a \pi(a|G_h) e_G(a), \quad \bar{e}_B = \sum_a \pi(a|B_h) e_B(a).$$

Uniform behavior gives $d_{\min}^{\mu} = 1/K$ at the decision stage and probability $1/(2K)$ at each branch pair. Consequently,

$$|e_x(a)| \leq \sqrt{K\epsilon}, \qquad \frac{1}{2K} \sum_a (e_G(a)^2 + e_B(a)^2) \leq \epsilon, \qquad |\bar{e}_G - \bar{e}_B| \leq 2\sqrt{K\epsilon},$$

where the last inequality is Cauchy–Schwarz and uses $\sum_a \pi(a|G_h)^2 + \sum_a \pi(a|B_h)^2 \leq 2$. The target residual at the distinguished action is therefore

$$r_{\mathrm{tar}}(a_{\theta_h}) = e_x(a_{\theta_h}) - \alpha_h(1 + \bar{e}_G - \bar{e}_B), \qquad |r_{\mathrm{tar}}(a_{\theta_h})| \leq \alpha_h + (1 + 2\alpha_h)\sqrt{K\epsilon} \leq \alpha_h + 2\sqrt{\epsilon / d_{\min}^{\mu}},$$

since $\alpha_h \leq 1/2$. Squaring and taking the supremum proves the positive-radius claim. For example, $c_1 = (\sqrt{2} - 1)^2/4$ suffices for the factor-two conclusion in the statement.

Information argument. Fix a layer h and the signs (action indices) at all other layers. Write $\mathbb{P}_a$ for the full interaction law when $\theta_h = a$ and $\mathbb{P}_0$ for the law with every action unbiased at this layer. Include the source data, the algorithm's randomization, and its output in these laws. Source observations have the same distribution under both laws. The chain rule for the adaptive transcript cancels every transition except observations of action a at layer h. For $\alpha_h \le 1/4$ it gives

$$\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_a) = \mathbb{E}_0[N_a]\,\mathrm{kl}\big(\tfrac{1}{2} \,\big\|\, \tfrac{1}{2} + \alpha_h\big) \le 3\alpha_h^2 \mathbb{E}_0[N_a],$$

where N_a counts this action's target visits and $\sum_a N_a = T$. Let q_a be the output policy's probability of playing a at x_h. Pinsker's inequality and concavity of the square root yield

$$\frac{1}{K}\sum_{a=1}^{K} \mathbb{E}_a[q_a] \le \frac{1}{K}\sum_a \mathbb{E}_0[q_a] + \frac{1}{K}\sum_a \sqrt{\tfrac{3}{2}\alpha_h^2 \mathbb{E}_0[N_a]} \le \frac{1}{K} + \sqrt{\frac{3\alpha_h^2 T}{2K}}.$$

This is the required uniform-prior average, rather than a bound for just one under-sampled action.

The statement also permits $1/4 < \alpha_h \le 1/2$, including a deterministic endpoint where the displayed KL can be infinite. Couple $\mathbb{P}_a$ and $\mathbb{P}_0$ until the first visit to action a at this layer. Their transcripts agree unless that visit occurs, so

$$\frac{1}{K}\sum_a \mathbb{E}_a[q_a] \le \frac{1}{K} + \frac{1}{K}\sum_a \mathbb{P}_0(N_a > 0) \le \frac{1}{K} + \frac{T}{K} \le \frac{1}{K} + 16c_0^2.$$

For a sufficiently small universal c_0 both success bounds are bounded away from one, because $K \ge 2$ and $\alpha_h^2 T/K \le c_0^2$. These arguments hold for every fixed choice of the other layers' indices even when the algorithm adapts across layers. Averaging those indices under their independent uniform prior and using linearity of expectation in (32) gives

$$\mathbb{E}[\mathrm{SubOpt}_{\mathrm{tar}}(\mathsf{ALG})] \ge c\sum_{h=1}^{L} \alpha_h = c\sum_h \sqrt{\delta_h^{\pi^\star_{\mathrm{tar}}}(0)}.$$

No independence of the adaptively collected layer transcripts is needed. □

Corollary 40 (A finite Bellman-complete hard class)**.** *The stepwise hard class is infinite, but for any fixed interaction budget the finite-class caveat can be closed outright by a backward construction. Fix the finite rate collection $\mathcal{C}$, the finite base class $\mathcal{Q}_{\mathrm{base}}$, and the operation budget $L = T + T_{\mathrm{off}}$ of Section 2; let the environment family consist of the source MDP and the finitely many target instances $\theta \in [K]^{\lfloor H/2 \rfloor}$ above; and include the finite comparator family $\Pi^\star := \{\pi^\star_\theta : \theta \in [K]^{\lfloor H/2 \rfloor}\}$ among the fixed evaluation policies Π_{fixed}, so that the assumptions quantify over the comparators and $\mathcal{C}_{\mathrm{alg}}$ gains only the centralized term $\log(1 + |\Pi_{\mathrm{fixed}}|) = O(H \log K)$. Construct, backwards in h, $\mathcal{F}_{H+1} := \{0\}$ and*

$$\mathcal{F}_h := \big\{\mathcal{T}^\pi_{\mathcal{M},h} f :\ f \in \mathcal{F}_{h+1},\ \mathcal{M} \in \textit{the family},\ \pi_{h+1} \textit{ a stage-}(h{+}1) \textit{ conditional of the closure generated by } \mathcal{F}_{h+1} \textit{ or of } \Pi^\star\big\}.$$

The generating set of policies is finite because a stage-$(h+1)$ conditional of the operational closure depends only on a base policy in $\mathcal{Q}_{\mathrm{base}}$, at most L stage-$(h+1)$ critics from $\mathcal{F}_{h+1}$ with rates in $\mathcal{C}$, and the up to two mixing positions, so each $\mathcal{F}_h$ is finite; its cardinality grows doubly exponentially in the remaining horizon, which is immaterial for an information-theoretic statement. By construction the class is Bellman complete for every environment in the family with respect to its own closure, since completeness only requires backups under stage-$(h+1)$ conditionals generated by stage-$(h+1)$ critics, which is exactly the generating set; it is realizable, since $Q^\pi_{\mathcal{M},H+1} = 0 \in \mathcal{F}_{H+1}$ and $Q^\pi_{\mathcal{M},h} = \mathcal{T}^\pi_{\mathcal{M},h} Q^\pi_{\mathcal{M},h+1} \in \mathcal{F}_h$ by downward induction; it respects the stepwise ranges, since rewards lie in $[0,1]$ and any backup of a $[0, H-h]$-valued function lies in $[0, H-h+1]$; and it contains the witness $Q^\pi_{\mathcal{M}_{\mathrm{src}}}$ for every policy π in the closure and *for every comparator $\pi^\star_\theta$, the latter directly by the downward induction through the $\Pi^\star$ conditionals; the lower bound is evaluated at the deterministic target-optimal policies, which are not generated from full-support bases by finitely many exponential updates, and including $\Pi^\star$ in the generating set is what places their witnesses in the class without any appeal to special structure (for this particular hard source MDP one could alternatively use that source transitions and rewards are action-independent at the decision states, making $Q^\pi_{\mathcal{M}_{\mathrm{src}}}$ policy-independent there, but the construction does not rely on this). Every subclass containing the relevant witness satisfies $\delta_h^{\pi^\star_{\mathrm{tar}}}(0) = \alpha_h^2$ exactly by the support argument in the proof, so the lower bound holds verbatim against the finite-class algorithmic guarantees at budget (T, T_{off}).*

G.2.2 TABULAR LOWER BOUND

Theorem 41 (Tabular lower bound, detailed version)**.** *Fix a target budget T, a number m of decision layers with nonincreasing visitation weights $p_1 \geq \cdots \geq p_m$, $p_k \in (1/T, 1]$, shift levels $\Delta_k \in \big(0,\ c_0/\sqrt{Tp_k}\,\big]$, and downstream scales $L_k \geq 1$ obeying the value budget $\sum_{k\leq m}(L_k+2) \leq H$. There exists a family of tabular hybrid MDP instances with shared rewards in $[0,1]$, in which the kth perturbed pair is visited by $\pi^\star_{\mathrm{tar}}$ with probability p_k, carries local shift Δ_k, and controls a downstream branch of value scale L_k, such that any algorithm interacting with the target MDP for T episodes satisfies*

$$\mathbb{E}[\mathrm{Regret}(T)] \ \geq\ c\,\min\Big\{HT,\ \sum_{k=1}^{m} T\,p_k\,L_k\,\Delta_k\Big\}.$$

In particular, with a single shifted layer carrying the full downstream budget ($m=1$, $L_1 = \lfloor H/2 \rfloor$, $H \geq 6$), writing h_0 for its step and using $Tp_1 \geq 1$,

$$\mathbb{E}[\mathrm{Regret}(T)] \ \geq\ \frac{c}{4}\,Tp_1\,H\Delta_1 \ \geq\ \frac{c}{4}\sqrt{T\sum_h H^2 \sum_{s,a} d^{\pi^\star_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar}},h}(s,a)\Delta_h(s,a)^2},$$

matching the target-weighted shift term of Theorem 6 on this instance class. The constraint $\Delta_k \lesssim 1/\sqrt{Tp_k}$ delimits the testing regime as before, and the value budget is necessary. Per-episode value is at most H under rewards in $[0,1]$, so downstream gaps of scale $\Theta(H)$ at many layers simultaneously would drive the display beyond the maximal regret HT; no bound of that unbudgeted form can hold.

Proof. Construction. A spine of states runs through the horizon, and the m decision points are placed sequentially along it with the spacing that the budget $\sum_k (L_k+2) \leq H$ permits. Ahead of the kth decision point the spine passes a gate that, independently of the action, continues with probability p_k/p_{k-1} (with $p_0 := 1$) and otherwise diverts to a neutral zero-reward corridor for the remainder of the episode; since the p_k are nonincreasing these are valid probabilities, and the probability that an episode reaches decision point k is exactly p_k, for *every* policy. The decision state y_k offers two actions $\{a_+, a_-\}$; the chosen action leads either to a good branch, a corridor of L_k consecutive states paying reward 1 per step, or to a bad branch, a zero-reward corridor of the same length, and both branches rejoin the spine afterwards, so later decision points remain reachable whichever branch is taken. In the source MDP both actions reach the good branch with probability $\frac{1}{2}$. The target instance is indexed by $\sigma \in \{\pm 1\}^m$, and action a_{σ_k} reaches the good branch with probability $\frac{1}{2}+\Delta_k$, the other with probability $\frac{1}{2}$. Rewards are shared and lie in $[0,1]$, total episode value is at most $\sum_k L_k \leq H$, the local shift at the perturbed pair is Δ_k, and $d^{\pi^\star_{\mathrm{tar}}}(y_k, a_{\sigma_k}) = p_k$, so at the single-layer instance

$$\sum_{s,a} d^{\pi^\star_{\mathrm{tar}}}_{\mathcal{M}_{\mathrm{tar}},h_0}(s,a)\Delta_{h_0}(s,a)^2 = p_1\Delta_1^2. \tag{33}$$

Per-layer regret. Fix k and all other signs. Let $\mathbb{P}_+$ and $\mathbb{P}_-$ denote the laws of the adaptive data under the two values of σ_k. Source data are identical under these laws. Each informative target observation is a Bernoulli whose two parameters are $1/2$ and $1/2+\Delta_k$ in one order or the other. Choose $c_0 \leq 1/4$; since $Tp_k > 1$, both directed Bernoulli divergences are at most $C\Delta_k^2$. The chain rule therefore bounds the KL of the history before episode t by

$$\mathrm{KL}(\mathbb{P}_+^{<t} \| \mathbb{P}_-^{<t}) \leq C(t-1)p_k\Delta_k^2 \leq Cc_0^2.$$

Here the expected number of informative observations is at most $(t-1)p_k$, regardless of the actions selected.

In the current episode, reaching y_k depends on action-independent gates and has probability p_k under either sign, independently of previous episodes. Before the action at y_k, all current-episode transition kernels are the same under the two signs. Conditional on reaching y_k, they thus apply the same probability kernel to the previous history and cannot increase its total variation distance. If $e_{+,t}$ and $e_{-,t}$ are the conditional probabilities of a wrong action given this visit, the two-point testing inequality and Pinsker give

$$e_{+,t} + e_{-,t} \geq 1 - \sqrt{Cc_0^2/2} \geq \tfrac{1}{2}$$

for sufficiently small c_0. Each wrong action costs $L_k\Delta_k$ in expected value. Averaging the signs and summing the episodes gives regret at least $Tp_kL_k\Delta_k/4$ at this layer.

Summing layers. The argument holds uniformly for each fixed configuration of other signs, so averaging over their independent prior and adding the expected layer losses proves $c\sum_k Tp_kL_k\Delta_k$. This sum is already at most HT under the value budget and shift constraints, and hence implies the displayed minimum. The single-layer consequence follows from $Tp_1 \geq 1$ and $Tp_1H\Delta_1 \geq H\Delta_1\sqrt{Tp_1}$. The theorem is restricted to this testing regime; it makes no claim of a matching regret characterization outside that regime. □

G.2.3 LINEAR MDP LOWER BOUND

Theorem 42 (Linear MDP lower bound, detailed version)**.** *There are universal constants $c, c_0, c_1 > 0$ such that the following hold.*

1. (Structural noncoverage.) *For every $H \geq 4$, $d \geq d_0$ (a universal constant; $d_0 = 10$ suffices, with four coordinates carrying the two testing pairs, the rest carrying deterministic bookkeeping transitions, and unused coordinates padded), $\Delta^{\mathrm{lin}} \in (0, c_0/\sqrt{d}]$, $\epsilon_{\mathrm{stat}} \in (0, 1/2]$, and* every *source sample size n, there is a family of d-dimensional linear MDPs of horizon H with a shared feature map, shared rewards, and the critic class $\mathcal{F}_h = \{\langle\varphi, w\rangle : \|w\|_2 \leq 2H\sqrt{d}\} \cap \{f : 0 \leq f(s,a) \leq H - h + 1 \ \forall(s,a)\}$, which contains every instance's value functions (they are linear with the stated norm and respect the ranges) and is moreover Bellman complete in both environments of every instance, since any backup of $f \in \mathcal{F}_{h+1}$ is linear with coefficient norm at most $\|\theta_h\|_2 + \sqrt{d}\,\|f\|_\infty \leq (1+H)\sqrt{d} \leq 2H\sqrt{d}$ and takes values in $[0, H-h+1]$, so the hard family satisfies exactly the algorithmic assumptions, with latent-operator shift Δ^{lin} in the normalization of Section 5.1 and a behavior policy whose covariance has a null direction, such that the source data distribution is identical across the family, the ambiguous pair of value functions lies in a common source-fitting set at radius* zero *with $\delta_h^\star(0) \geq c\,H^2\epsilon_{\mathrm{stat}}$ at the decision step, and any hybrid algorithm whose target budget satisfies $T \leq c_1 \min\{1/(d(\Delta^{\mathrm{lin}})^2),\ 1/\epsilon_{\mathrm{stat}}\}$ suffers*
$$\mathrm{SubOpt}_{\mathrm{tar}}(\mathsf{ALG}) \geq cH\left(\sqrt{\epsilon_{\mathrm{stat}}} + \sqrt{d}\,\Delta^{\mathrm{lin}}\right)$$
with constant probability. Each term separately requires only its own testing pair plus bookkeeping; the sum places the two gadgets on disjoint coordinate pairs and mixes them through the initial distribution, as detailed in the proof.

2. (Covered directions admit no persistent floor.) *For every $\lambda \in (0, 1/2]$, every $d \geq d_0$, and all $n, T \geq 1$, there is a family of d-dimensional linear MDPs with* matched *dynamics ($\Delta^{\mathrm{lin}} = 0$) whose behavior covariance satisfies $u^\top\Lambda_h^\mu u = \lambda$ along the ambiguous direction u, on which any hybrid algorithm suffers*
$$\mathbb{E}\big[\,\mathrm{SubOpt}_{\mathrm{tar}}(\mathsf{ALG})\big] \geq cH\min\Big\{1,\ \frac{1}{\sqrt{n\lambda + T}}\Big\}.$$

Part 1 realizes the null-space mechanism of Theorem 36 qualitatively, together with the shift term; the floor in part 1 persists for every n precisely because the behavior policy carries no information about the hidden direction. Part 2 is a finite-sample lower bound for the covered-direction gadget, quantifying the price of ambiguity along a direction with eigenvalue $\lambda > 0$ and locating the crossover between source-limited and target-limited learning at $n\lambda \asymp T$. Being a vanishing lower bound, part 2 cannot by itself rule out floors elsewhere in the class; the complementary no-floor statement under full coverage follows instead from the upper *bounds, since by Theorems 1 and 36 the suboptimality under $\lambda_{\min}(\Lambda_h^\mu) > 0$ and matched dynamics is $O\big(\sum_h \sqrt{\epsilon_{\mathrm{off}}(n)/\lambda_{\min}}\big)$ plus Theorem 1's optimization remainder. Both vanish provided the offline iteration count grows subcritically relative to the source sample size, $T_{\mathrm{off}} \to \infty$ with $\log|\Pi_{T_{\mathrm{off}}}|/n \to 0$, since the radius $\epsilon_{\mathrm{off}}(n)$ carries the T_{off}-dependent class term; for finite $\mathcal{F}$ the condition $T_{\mathrm{off}}\log(T_{\mathrm{off}}|\mathcal{F}|H/\delta) = o(n)$ suffices. Growing the two together proportionally would not. Persistence is therefore a structural noncoverage phenomenon, with part 1 exhibiting the floor at $\lambda = 0$ and the upper bounds excluding it under full coverage.*

Proof. We prove the two terms of part 1 separately and then combine them. The combined instance places the shift gadget on coordinates (e_0, u) and the structural gadget on the disjoint coordinates

(e_0', u'), and the initial distribution ρ puts mass $\frac{1}{2}$ on each gadget's entry state, so no branching state is needed. Fix the other gadget's hidden sign and compare the two full transcript laws obtained by changing the sign in the selected gadget. Only transitions in that gadget contribute to the chain-rule divergence. Its initial mass is $1/2$, so the expected number of informative target observations is at most $T/2$ under either hypothesis. Each gadget's testing bound therefore holds under adaptive allocation, uniformly in the other sign. Averaging over the independent signs gives the two component bounds; the transcript itself need not factor across gadgets. For the constant-probability statement we argue on the branch carrying the larger of the two terms, $A := H\sqrt{d}\,\Delta^{\mathrm{lin}}$ and $B := H\sqrt{\epsilon_{\mathrm{stat}}}$, since $\max\{A, B\} \geq (A + B)/2$; a naive conversion through a bound of the form $L \leq A + B$ would be false here, because refusing to enter a gadget can cost $\Theta(H)$, far above $A + B$. Write q for the output policy's probability of entering that branch's decision state and w for its conditional probability of the suboptimal decision action, so that the total suboptimality obeys

$$L \;\geq\; \tfrac{1}{2}\big[(1-q)\, c_3 H + q\, w\, c_4 \max\{A, B\}\big],$$

the $\frac{1}{2}$ from the initial mixture, the first term from forgoing the entering branch's $\Theta(H)$ continuation, and the second from the decision gap. Since the output policy depends on the data, the hidden sign, and the algorithm's randomization, q and w are random, so the case analysis must be at the level of events, not deterministic branches. Let $E_0 := \{q \leq \frac{1}{2}\}$. If $\Pr[E_0] \geq c_5/4$, then on E_0 the non-entry term alone gives $L \geq c_3 H/4 \geq c'(A + B)$, since $A + B \leq CH$ under the stated constraints on Δ^{lin} and ϵ_{stat}, so the loss event holds with probability at least $c_5/4$. Otherwise $\Pr[E_0] < c_5/4$, and since the branch's two-point argument gives the *unconditional* bound $\mathbb{E}[w] \geq c_5$ under the uniform prior on the hidden sign, while $0 \leq w \leq 1$,

$$\mathbb{E}\big[w\, \mathbf{1}_{E_0^c}\big] \;\geq\; \mathbb{E}[w] - \Pr[E_0] \;\geq\; \tfrac{3c_5}{4}, \qquad \text{hence} \qquad \Pr\big[E_0^c \cap \{w \geq c_5/4\}\big] \;\geq\; \tfrac{3c_5}{4} - \tfrac{c_5}{4} \;=\; \tfrac{c_5}{2},$$

and on that event $q > \frac{1}{2}$ and the decision-gap term gives $L \geq (c_4 c_5/16)\max\{A, B\} \geq (c_4 c_5/32)(A + B)$. In both cases $L \geq c(A + B)$ with probability at least $c_5/4$, which is the claimed constant-probability bound.

Construction details. We record the full linear-MDP specification once; both gadgets follow the same pattern. The state space consists of the two entry states, the decision states x, x', the continuation states g, b, g', b', an absorbing zero-reward sink, and the rewarding corridor states. Beyond the four testing coordinates e_0, u, e_0', u', deterministic transitions are realized through one bookkeeping coordinate per transition *type*, reusing the coordinate across the stages with the stage-dependent latent measures μ_h: a pair (s, a) of a given type has $\varphi(s, a) = e_{\mathrm{type}}$, and μ_h places on e_{type} the point mass at that stage's intended next state, which is well defined because at each stage every reachable pair of a given type has a single intended successor. The bookkeeping roles, none shared across gadgets except the last, are as follows. There is one corridor-advance coordinate per gadget, whose stage-h measure is the point mass at the next corridor state and, at the corridor's end, at the terminal goal state (the terminal self-loop is realized by the same coordinate through the stage dependence); one entry coordinate per gadget; and one globally shared zero-reward sink coordinate, since all zero-value absorptions can lead to a single sink state. That is five bookkeeping coordinates, so with the four testing coordinates $d_0 = 10$ suffices with one coordinate to spare, and unused coordinates carry zero measure and zero reward. All induced kernels are nonnegative and normalized by construction. The decision pairs carry the features and latent measures displayed in the two testing paragraphs, which are probability measures under the stated constraints on κ and Δ^{lin}. Rewards are $r_h(s, a) = \langle \varphi(s, a), \theta_h \rangle$ with θ_h supported on the corridor-advance bookkeeping coordinates with unit coefficients, so $r_h \in [0, 1]$ pointwise and $\|\theta_h\|_2 \leq \sqrt{d}$. Every instance's value functions are linear with coefficient norm at most $\|\theta_h\|_2 + H\sqrt{d} \leq 2H\sqrt{d}$ and respect the stepwise ranges, hence lie in the stated class $\mathcal{F}$, and the continuation gaps are $\Theta(H)$ for $H \geq 4$ through the corridors, which pay reward 1 per step for the remaining steps. Combining the two gadgets changes nothing in this representation, since their coordinate supports are disjoint and the mixing is performed by ρ, not by a transition. Part 2 is proved last.

Shift term. Consider a two-step decision embedded in a d-dimensional linear MDP. As with the structural gadget below, the decision state x is reached from the shift gadget's entry state by a dedicated entering action carried on the gadget's entry bookkeeping coordinate, against a zero-value alternative, so $\pi^\star$ enters and the entry probability q of an arbitrary output policy is defined in the same way for both gadgets. There are two continuation states g (downstream value $\Theta(H)$, realized

by a rewarding absorbing corridor with shared rewards) and b (value 0). Fix orthonormal vectors $e_0, u \in \mathbb{R}^d$ and let the decision state x offer two actions with features

$$\varphi(x, a_\pm) = \frac{e_0 \pm u}{\sqrt{2}}, \qquad \|\varphi(x, a_\pm)\|_2 = 1.$$

The source latent measure is supported on the e_0 coordinate, $\mu_{\mathrm{src}} = \sqrt{2}\left(\frac{1}{2}\delta_g + \frac{1}{2}\delta_b\right) e_0$, so that $P_{\mathrm{src}}(\cdot|x, a_\pm) = \langle \varphi(x, a_\pm), \mu_{\mathrm{src}} \rangle = \frac{1}{2}\delta_g + \frac{1}{2}\delta_b$ for both actions, and the source is uninformative about which action is better. The target family, indexed by $\varsigma \in \{\pm 1\}$, perturbs the latent operator in direction u, setting

$$\mu_{\mathrm{tar}} = \mu_{\mathrm{src}} + \varsigma\, \kappa\, u\, (\delta_g - \delta_b), \qquad \kappa := \frac{\sqrt{d}\, \Delta^{\mathrm{lin}}}{2}.$$

Total mass is preserved because $(\delta_g - \delta_b)(\mathcal{S}) = 0$, and each action's next-state measure equals $\frac{1}{2}(\delta_g + \delta_b) \pm \varsigma \frac{\kappa}{\sqrt{2}}(\delta_g - \delta_b)$, a valid distribution whenever $\kappa \le \frac{1}{\sqrt{2}}$, that is for $\Delta^{\mathrm{lin}} \le c_0/\sqrt{d}$ with $c_0 = \sqrt{2}$. The construction attains the normalized shift of Section 5.1 exactly, since $\sup_{\|g\|_\infty \le 1} \| \int g\, d(\mu_{\mathrm{src}} - \mu_{\mathrm{tar}})\|_2 = 2\kappa = \sqrt{d}\, \Delta^{\mathrm{lin}}$. Under instance ς, the good-branch probability of a_+ minus that of a_- equals $\varsigma\sqrt{2}\,\kappa = \Theta(\varsigma\sqrt{d}\, \Delta^{\mathrm{lin}})$, so the value gap between the two actions is $\Theta(H\sqrt{d}\, \Delta^{\mathrm{lin}})$ and the optimal action is determined by the hidden sign. Source data are identically distributed under $\varsigma = \pm 1$, each target episode yields one Bernoulli observation whose parameter differs by $\Theta(\sqrt{d}\, \Delta^{\mathrm{lin}})$ across the two hypotheses, and with $T \le c_1/(d(\Delta^{\mathrm{lin}})^2)$ the total squared testing distance is $O(c_1)$, so by Le Cam's method the learner errs on the sign with constant probability and loses $\Omega(H\sqrt{d}\, \Delta^{\mathrm{lin}})$.

Structural term (part 1). On a disjoint pair of orthonormal coordinates e_0', u', embed the same two-action gadget with *matched* dynamics, so $\Delta^{\mathrm{lin}} = 0$ on these coordinates. A decision state x' is reached at a fixed step, with probability one, by any policy that elects to enter it from the preceding state, and the entering branch has continuation value $\Theta(H)$ against 0 for the alternative (the good continuation g' opens a corridor paying reward 1 for the remaining $H - 2$ steps), so $\pi^\star$ enters and a sign error costs $(H-2) \cdot \sqrt{2}\,\kappa = \Theta(H\kappa)$ for $H \ge 4$. The decision actions have features $\varphi(x', a'_\pm) = (e_0' \pm u')/\sqrt{2}$, continuation states g' (value $\Theta(H)$) and b' (value 0), and the instance family, indexed by $\varsigma' \in \{\pm 1\}$, has latent measure

$$\mu^{(\varsigma')} = \sqrt{2}\left(\tfrac{1}{2}\delta_{g'} + \tfrac{1}{2}\delta_{b'}\right) e_0' + \varsigma' \kappa\, u'\, (\delta_{g'} - \delta_{b'}), \qquad \kappa := \sqrt{\epsilon_{\mathrm{stat}}},$$

in both environments, a valid MDP family for $\kappa \le 1/\sqrt{2}$. The behavior policy takes the non-entering action at the preceding state, so x' has source visitation zero, u' lies in the null space of Λ_h^μ, and the source data distribution is *identical* under $\varsigma' = \pm 1$, for every sample size n. No amount of source data therefore carries any information about ς'. Moreover, the wrong instance's value function has zero population source residual, since the two instances' Bellman backups agree on every pair the behavior policy visits, while its target Bellman residual at the decision step is $\left(\sqrt{2}\,\kappa \cdot \Theta(H)\right)^2 = \Theta(H^2 \epsilon_{\mathrm{stat}})$ under $d^{\pi^\star}$; this proves the claim $\delta_h^{\pi^\star}(0) \ge cH^2\epsilon_{\mathrm{stat}}$, and the ambiguity is exactly the null-space term of Theorem 36 with $\|P_\perp \varphi(x', a'_\pm)\|_2^2 = \frac{1}{2}$. On the target side, each episode yields at most one observation at x', a Bernoulli with parameter $\frac{1}{2} + \varsigma'\kappa/\sqrt{2}$, so the squared testing distance after T episodes is $O(T\kappa^2) = O(c_1)$ under the stated budget, and Le Cam's two-point method forces a sign error with constant probability, at value cost $\Omega(H\kappa) = \Omega(H\sqrt{\epsilon_{\mathrm{stat}}})$.

Covered directions (part 2). Modify the gadget in one respect. The behavior policy now enters x' with probability $p = 2\lambda \in (0, 1]$, so that $u'^\top \Lambda_h^\mu u' = p \cdot \frac{1}{2} = \lambda$ (each visit contributes $(u'^\top \varphi)^2 = \frac{1}{2}$ regardless of which action is taken). Keep matched dynamics and tilt $\theta \in (0, 1/\sqrt{2}]$ in place of κ. Every observation at x', whether in the source or the target data, is a Bernoulli with parameter gap $\sqrt{2}\,\theta$ across the two hypotheses, and these are the only observations whose law depends on ς'. A source episode observes x' with probability p, and a target episode observes it at most once whatever policy the algorithm deploys, so by the chain rule the squared Hellinger distance between the joint laws of (source dataset, adaptively collected target data) under $\varsigma' = \pm 1$ is at most $c'(np + T)\theta^2 = c'(2n\lambda + T)\theta^2$. Choosing $\theta = c_0' \min\{1, (n\lambda + T)^{-1/2}\}$ keeps this below a small constant, so the sign is unidentifiable with constant probability, and a sign error costs $\Omega(H\theta)$ because $\pi^\star$ enters x'. This proves part 2. Its scope is limited. It exhibits, for each (n, T, λ), a hard pair of instances whose difficulty vanishes at the $(n\lambda + T)^{-1/2}$ scale; the statement that *no* persistent floor exists under $\lambda_{\min} > 0$ is not a consequence of this lower bound but of the upper bounds, as recorded after the theorem statement. □

H EXPERIMENTS

H.1 FINITE-CLASS EXPERIMENTS

H.1.1 ENVIRONMENT, SHARED CLASS, AND EXACT COMPUTATION

The environment has four layers, namely the start state, the corridor states A2 and B2, the decision states A3 and B3, and the terminal states G4 and W4. The first action chooses a corridor deterministically, the second transition advances within it, and an action at layer 3 reaches G4 with probability $q_A(a)$ or $q_B(a)$ and otherwise reaches W4. The reward is 0.8 at either layer-2 state, 1 at G4, and zero elsewhere, so a policy whose final action has success probability q has value $0.8 + q$. The return bound derived from the known rewards is $V = 1.8$, and the largest value in the candidate family is $V_{\mathcal{F}} = 1.65$. The tabular online baseline uses the reward-based bound.

Let $g = (0.15, 0.383333\ldots, 0.616666\ldots, 0.85)$ be the fixed probability grid. The shared family contains all $4^4 = 256$ assignments of grid values to the four corridor–action pairs, and Table 6 specifies the instances by zero-based grid indices. The eight deterministic policies choose an action at the start, at A3, and at B3. For constructing the critic classes, the policy set also contains three behavior policies, which select corridor A with probability $1/2$, 0.95, and 0.05 respectively and mix the two later actions uniformly. The last policy mirrors the second, so that swapping the corridor labels maps the behavior policy to another member of the class.

Table 6: Four experimental conditions, specified by grid indices. Each pair lists the two actions' indices in that corridor. Uniform behavior selects either corridor with probability $1/2$; skew-A selects A with probability 0.95.

Condition	Source A	Source B	Target A	Target B	Behavior
Useful	$(3, 0)$	$(0, 0)$	$(2, 0)$	$(0, 0)$	Uniform
Misleading	$(3, 0)$	$(0, 0)$	$(0, 3)$	$(0, 0)$	Uniform
Poor coverage	$(2, 0)$	$(3, 3)$	$(2, 0)$	$(3, 3)$	Skew-A
Hidden route	$(2, 0)$	$(0, 0)$	$(2, 0)$	$(3, 3)$	Uniform

With this fixed model family $\mathcal{M}_{\mathrm{cand}}$ and policy set Π_{class}, construct

$$\mathcal{F}_{H+1} = \{0\}, \qquad \mathcal{F}_h = \left\{\mathcal{T}^{\pi}_{M,h} f_{h+1} : M \in \mathcal{M}_{\mathrm{cand}},\ \pi \in \Pi_{\mathrm{class}},\ f_{h+1} \in \mathcal{F}_{h+1}\right\}.$$

Deduplicating component arrays to ten decimal places gives sizes $(49, 49, 256, 1)$, and the learner uses the full Cartesian product of these components. The recursion gives stagewise completeness, and induction gives realizability for every candidate model and every policy in Π_{class}. Both properties refer to this finite policy set; greedy improvement maps back into the eight deterministic policies.

Every residual constraint depends only on consecutive components (f_h, f_{h+1}). Admissible pairs therefore form the edges of a layered directed acyclic graph whose complete paths are the feasible critics. Reachability decides nonemptiness, and dynamic programming minimizes the initial value, maximizes evaluation residuals, and computes bottleneck objectives. These computations are exact up to floating-point tolerance, and no test requires enumerating the product class.

H.1.2 EMPIRICAL CONSTRAINTS AND POPULATION DIAGNOSTICS

For a fixed policy π and candidate continuation f_{h+1}, write the regression target for transition i as $y_i = r_h(s_i, a_i) + \mathbb{E}_{a' \sim \pi_{h+1}(\cdot|s'_i)} f_{h+1}(s'_i, a')$. With N transitions at stage h, the constraint uses

$$\widehat{\mathcal{E}}_h(f; \pi) = \frac{1}{N}\sum_{i=1}^{N} (f_h(s_i, a_i) - y_i)^2 - \min_{g \in \mathcal{F}_h} \frac{1}{N}\sum_{i=1}^{N} (g(s_i, a_i) - y_i)^2.$$

The continuation and policy are the same in both terms. Under completeness the population regression minimizer is the Bellman backup, so the conditional variance of the regression target cancels and the population excess loss equals the squared Bellman residual. Because the empirical minimization biases the excess loss upward, a target rejection uses a concentration allowance. Population residuals and exact occupancies enter only the diagnostics and the evaluation.

The radii used in the experiments are

$$\ell(K) = \log\left(\frac{2H|\Pi_{\text{class}}|m^3 \max\{1, K\}}{0.05}\right), \qquad m = \max_h |\mathcal{F}_h| = 256,$$
$$\beta = \frac{26V_{\mathcal{F}}^2 \ell(1)}{3n_{\text{src}}}, \qquad \gamma = \frac{26V_{\mathcal{F}}^2 \ell(K)}{3w}.$$

Here n_{src} episodes yield n_{src} transitions per stage, and $K = T/w$ bounds the number of blocks. Population diagnostics for the source use $\epsilon_{\text{off}} = (42/13)\beta$. At $n_{\text{src}} = 50{,}000$ these give $\beta \approx 0.0114$ and $\epsilon_{\text{off}} \approx 0.0368$, and the target radius is $\gamma \approx 0.019$ at $w = 32{,}000$ and $\gamma \approx 0.080$ at $w = 8{,}000$. The finite-sample validity of each test is conditional on the corresponding uniform concentration event. The master analysis below uses its own return statistics and retrospectively checked envelopes.

For a fixed deployed policy, the population diagnostic

$$\rho_\pi(\epsilon) = \min_{f:\, \max_h \mathcal{E}_{\text{src}}(f;\pi)_h \le \epsilon} \max_h \mathcal{E}_{\text{tar}}(f;\pi)_h$$

measures how far the source-fitting set lies from target compatibility along that policy. The source residuals are taken under the behavior distribution and the target residuals under the deployed policy's target occupancy. The diagnostic vanishes whenever the fitting set contains a target-compatible critic, for example Q^π_{tar}. Whether a test fires in finite samples depends on both radii and on sampling fluctuations as well as on the size of the shift.

In the useful condition, the source value function has squared target residual $(g_3 - g_2)^2 = 0.054444\ldots$ at layer 3 along the deployed policy A0, while the policy remains optimal. This exceeds $\gamma/2$ at $w = 32{,}000$, so a rejection there is a correct certificate in the sense of Remark 9. At the larger radius ϵ_{off} the fitting set also contains Q^π_{tar}, whose source residual is $0.25 \times 0.054444\ldots \approx 0.0136$, so $\rho_\pi(\epsilon_{\text{off}}) = 0$. In the hidden-route condition, Q^π_{src} has zero source and zero target residual along A0, while target route B is better by $g_3 - g_2 = 0.233333\ldots$. It is therefore a compatibility witness at every radius, and on the containment event it keeps the intersection nonempty while that policy is deployed. In the poor-coverage condition the environments coincide, so low visitation creates no inconsistency between source and target.

H.1.3 Learners, handoff, and the fixed-time control

The initial policy maximizes, over the eight deterministic policies, the minimum initial value among empirically source-feasible critics. This is the max-min selection of Proposition 18. Policy and critic ties use the lowest index. A greedy update keeps the current action when it is a maximizer and otherwise takes the lowest maximizing action. The label permutations described below measure how much the outcomes depend on these conventions.

Learner C deploys its current policy for one block, forms the source constraints and the current-block target constraints for that policy, and checks whether their intersection is empty. If it is nonempty, C selects a pessimistic critic from it and applies the greedy update; otherwise C keeps its current policy. Tests and updates that would not be followed by any deployment are omitted. Learner D agrees with C until the first rejection and then starts A afresh, so the comparison of D with C measures the effect of adding the handoff, including the cost of fresh exploration.

Learner A uses empirical tabular transitions, known rewards, and Hoeffding-style optimism, and replans every 100 episodes. It does not receive the candidate-model family, and we treat it as a batched optimistic baseline without a proved anytime regret bound. Learner A2 eliminates candidate models using target observations and acts optimistically within the remaining family; it receives no source samples. Since A2 differs from A both in its information and in its algorithm, its comparison with A does not isolate the value of the candidate family.

Learner E follows C up to a fixed switch time and then starts A in the same way as D. Switch times of $0, 1, 2, 3, 4$ blocks and "never" are compared on seeds 0–4, weighting the four conditions equally, and the selected time is then fixed for seeds 5–9. In the configuration with $n_{\text{src}} = 50{,}000$, $w = 32{,}000$, and $T = 160{,}000$, calibration selects time zero, so E coincides with A. Table 7 reports the held-out comparison at the original labeling; these runs are separate from the label-averaged results of the main text.

Table 7: Original-label handoff comparison, five held-out seeds, mean ± sample standard deviation. The fixed-time control E selects immediate switching on calibration seeds. The paired D–E column is computed from per-seed differences.

Condition	C	D	E (= A)	D–E
Useful	0.0 ± 0.0	1642.7 ± 39.0	1680.0 ± 73.8	-37.3 ± 51.1
Misleading	112000.0 ± 0.0	23744.0 ± 31.3	1302.0 ± 38.3	22442.0 ± 38.3
Poor coverage	37333.3 ± 0.0	37333.3 ± 0.0	3112.7 ± 81.8	34220.7 ± 81.8
Hidden route	37333.3 ± 0.0	37333.3 ± 0.0	3112.7 ± 81.8	34220.7 ± 81.8

Averaged equally over the four conditions, the held-out regret is $25{,}013.3$ for D and $2{,}301.9$ for E; this comparison depends on the chosen mixture of conditions. In the useful condition, D switches after its first block, which incurs no regret, so all $1{,}642.7$ of its regret occurs during online learning after the switch. In the misleading condition, D's total of $23{,}744$ consists of $22{,}400$ during the first block and $1{,}344$ after the switch. The regret of any policy splits as

$$V^\star - V^\pi = \left[V^\star - \max_{\pi' \text{: same corridor as } \pi} V^{\pi'}\right] + \left[\max_{\pi' \text{: same corridor as } \pi} V^{\pi'} - V^\pi\right],$$

which separates the choice of corridor from the choice of action within it.

Four additional conditions change the shift strengths without reselecting E. In grid-index notation, H1 changes the target A of the useful condition to $(1, 0)$, and H2 changes the target A of the misleading condition to $(1, 2)$. H3 uses source A $(1, 0)$, source B $(0, 0)$, target A $(1, 0)$, and target B $(2, 2)$ with uniform behavior, and H4 has identical source and target with A $(1, 0)$, B $(2, 2)$, and skew-A behavior. Table 8 reports the held-out results at the original labeling. These conditions played no role in selecting the switch rule, although pilot experiments informed the design of the testbed.

Table 8: Additional shift conditions with the fixed switching rule unchanged. Mean ± sample standard deviation over seeds 5–9; original labeling, $T = 160{,}000$.

Condition	C	D	E (= A)	A2
H1 useful	0.0 ± 0.0	2352.0 ± 60.4	2520.0 ± 71.9	280.0 ± 194.5
H2 misleading	37333.3 ± 0.0	9179.3 ± 35.4	1787.3 ± 60.8	177.3 ± 51.1
H3 hidden	37333.3 ± 0.0	37333.3 ± 0.0	1936.7 ± 87.3	93.3 ± 0.0
H4 coverage	37333.3 ± 0.0	37333.3 ± 0.0	1936.7 ± 87.3	93.3 ± 0.0

H.1.4 Label permutations and reporting conventions

The eight-element group swaps the two corridors and, independently, the actions at each layer-3 decision state; kernels, policies, and source behavior are transformed together. For each method and condition, let $r_{g,k}$ denote the regret at labeling g and seed k. Table 1 reports the average of $\bar{r}_g = 10^{-1} \sum_{k=0}^{9} r_{g,k}$ over the eight labelings together with the sample standard deviation of the eight values $\bar{r}_g$. Since the eight labelings exhaust the group, this spread describes label sensitivity and is not a confidence interval.

The regret of C is the same at all eight labelings. D has a small spread across labelings in the useful condition and a much larger one in the misleading condition, where A is also label-sensitive. A2 is label-sensitive as well; its zero regret at the original labeling in two conditions becomes 118.4 after averaging over the group. The comparisons in which A2 outperforms C and D hold in three conditions after averaging.

H.1.5 Master protocol, allocation, and bound check

The master has $M = 21$ registered learners, namely twenty source copies indexed by $j = 0, \ldots, 19$ and one online copy. With reward bound $V = 1.8$, $T = 168{,}000$, and $\delta_{\text{ret}} = 0.05$, define

$$\iota = \log(2MT/\delta_{\text{ret}}), \qquad c = 2V\sqrt{\iota} = 15.5947\ldots,$$
$$R_j(t) = \min\{Vt,\, 2^{-j}Vt + c\sqrt{t}\}, \qquad R_{\text{on}}(t) = \min\{Vt,\, c\sqrt{t}\}.$$

These envelopes, including the online one, are asserted rather than proved for the implemented learners, and the logarithmic factor is included in c. The envelope of copy $j = 0$ is the trivial bound Vt, which keeps the prefix-valid set nonempty. The source copies share the same $50{,}000$ source episodes and initial policy, and each uses $w = 8{,}000$ of its own target episodes per block. A2 is a standalone control.

The master first plays each copy once. Thereafter it plays an active copy minimizing $R_i(t_i)$ at that copy's own count, breaking ties at random, and only the selected copy receives the trajectory and advances its counter. A source copy accumulates its own episodes across interruptions until it completes a block. After every episode the master removes copy i when

$$\max_{j \text{ active}} \{\widehat{U}_j(t_j) - \kappa(t_j)\} > \widehat{U}_i(t_i) + R_i(t_i)/t_i + \kappa(t_i), \qquad \kappa(t) = V\sqrt{2\iota/t},$$

with the benchmark on the left computed from the active set before any simultaneous removals. The master uses only observed returns.

For evaluation, the set W contains the copies whose cumulative pseudo-regret stays within their asserted envelope at every value of their own count that is reached, including the prefix before a copy is eliminated. On the return-concentration event, Proposition 30 compares the master with every copy in W. We evaluate the explicit right-hand side of that proposition, with the return bound V in place of H,

$$\min_{i \in W} \left\{ M(R_i(T) + 3V) + 2V\sqrt{2MT\iota} + \sum_{s=1}^{T} \rho_i(s) \right\},$$
$$\rho_i(s) = \min\{V,\, R_i(t_i(s))/t_i(s) + 2\kappa(t_i(s))\},$$

where $\rho_i(s) = V$ while $t_i(s) = 0$, counts are taken after global episode s, and the sum includes the initial round of plays. The candidate of each comparator is evaluated at T in the first term even when that copy is played less often. The standalone controls are run for the full budget.

Table 9: Master allocation and bound diagnostics, means over five held-out seeds. Completed and accepted blocks refer to source copies; no source policy changed. RHS is the trajectory-dependent expression above, minimized over W. The reward-based cap is $VT = 302{,}400$.

Condition	Online share	Blocks/accepted	Eliminated	$\lvert W\rvert$	Online in W	RHS
Useful	6.2%	13/13	0	21	Yes	289691
Misleading	57.2%	0/0	17	3	Yes	261965
Poor coverage	6.2%	13/13	0	6	No	515028
Hidden route	6.2%	13/13	0	6	No	515028

The inequality of Proposition 30 holds in every run. Its right-hand side lies below the cap VT in the useful and misleading conditions and above it in the other two. No source copy changes its policy. In the misleading condition, seventeen of the twenty source copies are eliminated and none completes a block. For seed 5, the three copies with the loosest envelopes are never eliminated, the online copy receives $96{,}547$ episodes, and copy $j = 18$ first exceeds its envelope at its own episode 497 and is eliminated at global episode $54{,}029$. The first-violation time is computed by exact evaluation, which the master does not have. In the poor-coverage and hidden-route conditions the online copy exceeds its asserted envelope, so those runs do not instantiate the proposition's comparison with a valid online baseline. A single budget does not reveal how these costs change with T.

H.2 Model-based tabular illustrations

The experiments in this part illustrate individual mechanisms of the theory with model-based tabular analogues of the algorithms. They use confidence balls around transition models and heuristic combinations of source and target models in place of the finite-class Bellman constraints of Section 6. Values, occupancies, Bellman residuals, and modulus proxies are computed exactly by dynamic programming, so the value and regret curves carry no evaluation noise; the one exception is the Monte Carlo simulation of Figure 17.

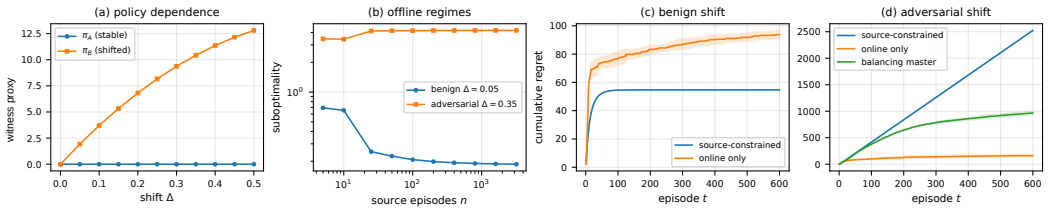


Figure 6: The two-corridor testbed. (a) The source-value witness residual is zero on the stable route and increases with the shift on the crossing route. (b) The offline error decreases with source data under benign shift and approaches the target gap of the source-optimal policy under adversarial shift. (c, d) The source-informed analogue helps under benign shift and has nearly linear regret under adversarial shift, and the experimental master reduces the latter cost.

H.2.1 Testbed and procedures

Environment. The testbed is a two-corridor instance of the block-MDP structure of Section 5.1, with $|\mathcal{S}| = 12$ states, $|\mathcal{A}| = 2$ actions, and horizon $H = 15$. From the start state, action 1 enters corridor A and action 2 enters corridor B; each corridor is a chain of 4 states in which one action advances and the other stays. Corridor A ends in an absorbing goal paying reward 0.6 per step, corridor B in an absorbing goal paying 1.0 per step, and all other rewards are zero. Rewards and every transition outside corridor B are identical in the source and the target. In the target only, each advance in corridor B fails into an absorbing zero-reward sink with probability Δ, so the local shift of Theorem 6 equals Δ at four state-action pairs and zero elsewhere. In the source, corridor B is optimal, with $V_{\mathrm{src}}^{\pi_B} = 10.0$ against $V^{\pi_A} = 6.0$ from the start state, while in the target corridor A becomes optimal once $(1-\Delta)^4 < 0.6$, that is, for $\Delta > 1 - 0.6^{1/4} \approx 0.12$. The single parameter Δ therefore moves the instance from benign transfer through mild shift to adversarial shift, and the two deterministic corridor policies π_A and π_B realize the two cases of Figure 1.

Procedures. The learners are model-based analogues of the offline procedure, the source-informed online procedure, the block procedure, and the balancing master. A time-homogeneous empirical transition model pools counts across stages, with per-pair total-variation radius $\min\{1, c\sqrt{(|\mathcal{S}| + \log(1/\delta))/N(s,a)}\}$ and radius one on unvisited pairs. Robust dynamic programming over these balls supplies pessimistic critics for policy evaluation and optimistic policies for planning. For source-informed updates, the implementation tests whether the source and target balls overlap. At overlapping pairs with at least five target visits it averages the two centers with weights proportional to the inverse squared radii and uses the smaller radius; at separated pairs it keeps the source model and records the separation. The per-episode and previous-block variants always apply their Hedge update, the previous-block variant using only the preceding block's target counts, while the same-block variant skips its update when any sufficiently visited pair separates. The online baseline plans optimistically from target data alone, and the splitting baseline pools counts at overlapping pairs and uses target counts alone at separated pairs. The experimental master runs four source-informed copies and one online copy with experimentally scaled candidate bounds and elimination thresholds. In the figures, "certificate" refers to the transition-ball separation diagnostic, and the labels "Source online" and "Lazy blocks" refer to the analogues of Algorithms 3 and 1; the offline analogue corresponds to Algorithm 2.

Protocol. Source data are episodes of the uniform behavior policy, with $n = 1600$ in the online experiments. The offline sweep uses $n \in \{5, 10, 25, 50, 100, 200, 400, 800, 1600, 3200\}$ in both regimes and averages the values of 150 Hedge iterates. The radius scale is $c = 0.1$ and the learning-rate multiplier is 8. The master uses confidence scale 2, online-bound scale 0.15, optimization coefficient 2, checks every ten episodes, and requires at least eight episodes per copy. Online runs use $T = 600$. The main online, splitting, warm-start, diagnostic, and same-block experiments use eight seeds; the master, boundary, and reward-shift experiments use six; and the offline and fixed-policy residual experiments use ten. Bands show $\pm 2s/\sqrt{m}$, where s is the sample standard deviation over the m seeds. The supplementary material contains the scripts and run configurations.

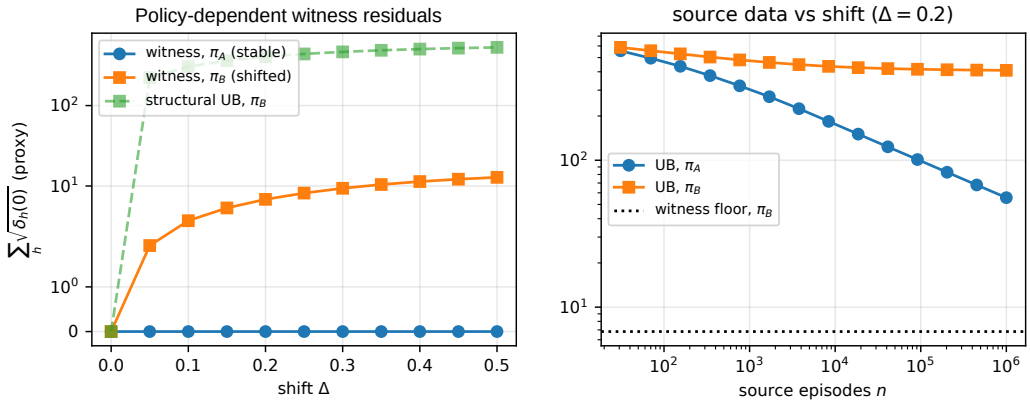


Figure 7: Policy dependence and scaling of the weak modulus. Left, the witness residual $\sum_h (\mathcal{E}^{\pi,\pi}_{\mathrm{tar}}(Q^{\pi}_{\mathcal{M}_{\mathrm{src}}})_h)^{1/2}$ is zero for π_A and increases with Δ for π_B, and the structural upper bound of Theorem 6 follows the same occupancy-weighted trend with a looser constant. Right, as the source size n grows at $\Delta = 0.2$, the bound for π_A is purely statistical and decays, while the bound for π_B levels off at its shift term.

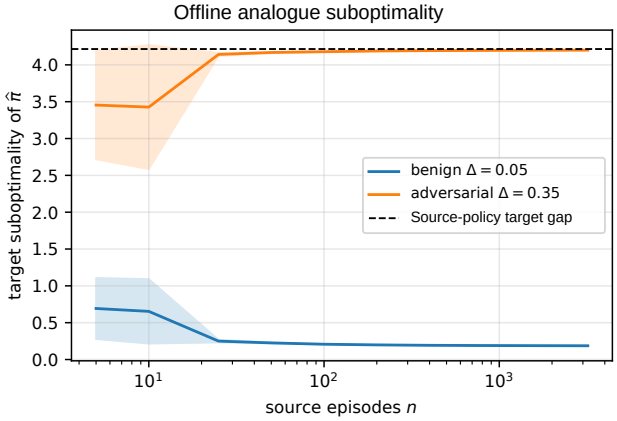


Figure 8: Suboptimality of the offline analogue against the source sample size, ten seeds. The error decreases under benign shift and approaches the target gap of the source-optimal policy (dashed) under adversarial shift.

H.2.2 Results

Witness residuals and structural bounds. Figure 7 plots the source-value witness residual $\sum_h (\mathcal{E}^{\pi,\pi}_{\mathrm{tar}}(Q^{\pi}_{\mathcal{M}_{\mathrm{src}}})_h)^{1/2}$, which bounds the corresponding modulus terms from below, together with the structural upper bound of Theorem 6. The witness residual is zero on the stable route, and on the crossing route it equals

$$10\Delta \sum_{j=0}^{3} (1-\Delta)^{j/2},$$

since each local backup discrepancy equals 10Δ and the probability of reaching the jth shifted pair is $(1-\Delta)^j$. The curve is linear to first order near zero. The structural bound separates a statistical contribution, which decreases with the source sample size, from a shift contribution, which does not.

Offline learning. In Figure 8, the suboptimality of the offline analogue decreases from about 0.69 at $n = 5$ to about 0.21 near $n = 100$ under benign shift. Under adversarial shift it approaches 4.2, the target value gap of the source-optimal policy, as the source sample size grows.

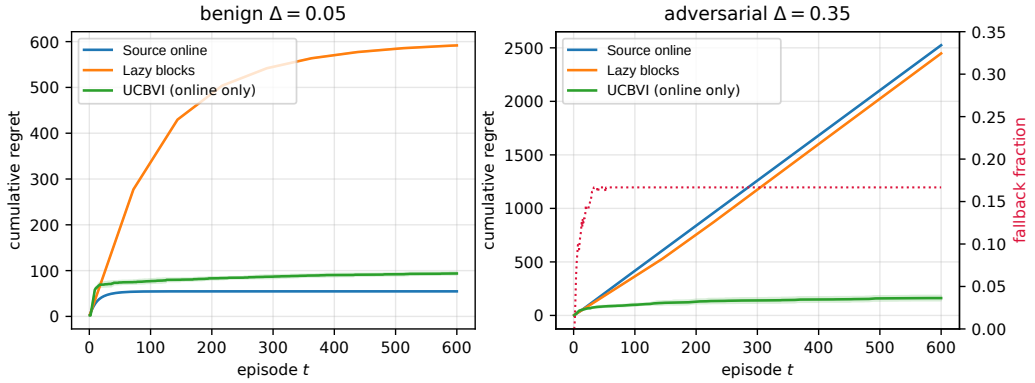


Figure 9: Regret of the online analogues, eight seeds. Under benign shift the source-informed learner outperforms the online baseline, and block updates incur larger regret. Under adversarial shift the source-constrained learners have nearly linear regret over the displayed horizon. The dotted line records the separation of the empirical transition balls at the four shifted pairs.

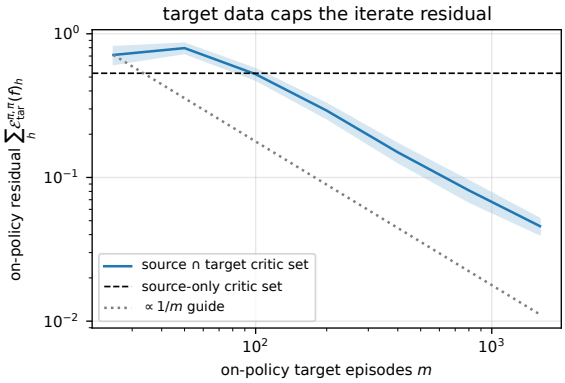


Figure 10: Target constraints at a fixed policy, with π_B fixed, $\Delta = 0.03$, and $n = 100$. The on-policy residual of the combined-model critic decreases as the target sample size grows, while the source-only residual stays constant.

Online regret. In Figure 9, the regret under benign shift is about 55 for the source-informed analogue and about 94 for the online baseline. Under adversarial shift, the regret of the source-informed analogue grows at about 4.2 per episode, the target gap of the source-optimal policy. The separation diagnostic identifies the four shifted pairs. Updating once per block raises the regret under benign shift, which is consistent with the less frequent updates.

Target constraints at a fixed policy. Figure 10 fixes the policy π_B and compares a source-only critic with the combined-model critic. The on-policy residual of the combined critic decreases as the target sample size grows, while the source-only residual stays constant. The critic sets and the test differ from those of Lemma 27, so the figure illustrates the effect of target constraints rather than the rate of the lemma.

Adaptivity. In Figure 11, the master with $M = 5$ candidates incurs about 5.3 times the regret of the better base learner under benign shift and 6.0 times under adversarial shift. Under benign shift it is worse than both base learners at this horizon, and under adversarial shift it stops using the misled source-informed copies and follows the slope of the online learner.

Baselines. Figure 12 compares the source-informed analogue with pooling, online learning, and a splitting baseline inspired by Qu et al. (2025). Pooling performs well under benign shift and keeps

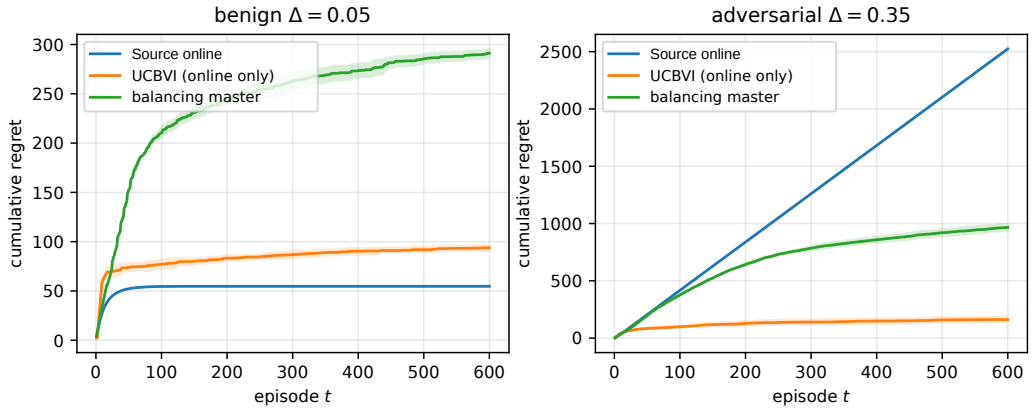


Figure 11: Adaptivity through regret balancing, six seeds. Under adversarial shift the master eliminates the misled source-informed copies and follows the slope of the online learner. Its exploration cost is about 5.3 times the regret of the better base learner under benign shift and 6.0 times under adversarial shift.

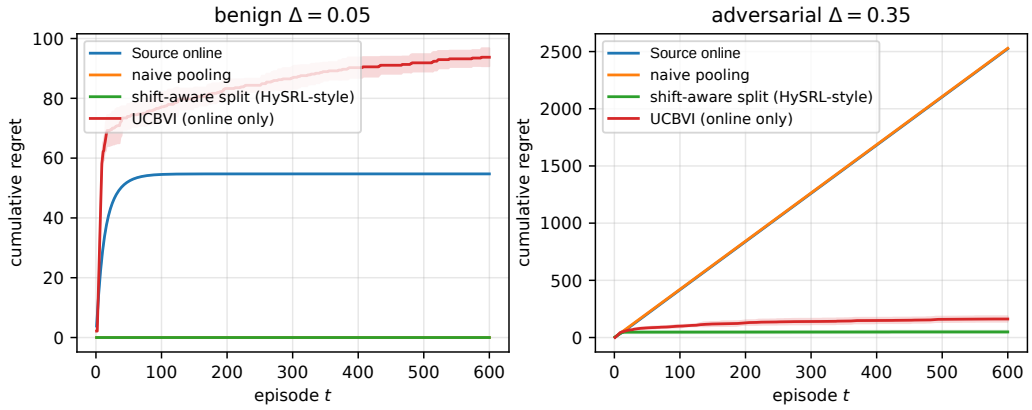


Figure 12: Comparison with model-based baselines, eight seeds. Pooling performs well under benign shift and keeps a persistent error over the adversarial horizon, and splitting is effective on this local and detectable shift.

a substantial error under adversarial shift, while splitting is effective because the shift is local and quickly detected. Our general-class guarantees cover arbitrary shift through their residual terms and need not be sharper than guarantees specialized to tabular models.

Reward shift. The reward-shift variant keeps the dynamics fixed and lowers the target reward of the corridor-B goal by x. Comparing the two corridor values shows that the optimal policy changes at $x = 0.4$ and that the source-optimal corridor has target gap $(10x - 4)_+$. Figure 13 compares the error of the offline analogue with this gap and with the source-value witness residual, illustrating the reward-discrepancy term of Proposition 8.

Estimating the modulus. Figure 14 reports a diagnostic built from three critics, namely the empirical source-model critic and the pessimistic and optimistic robust-evaluation critics, whose residuals are evaluated under a plug-in target model estimated from $m = 200$ on-policy target episodes. The maximum over these critics approximates a model-based supremum from below when the critics are feasible for that set. Because the model-based confidence set differs from the empirical Bellman set of Proposition 24, the diagnostic is compared with the witness residual and is not used as a bound; the offset at small shift reflects finite-sample and model-set effects.

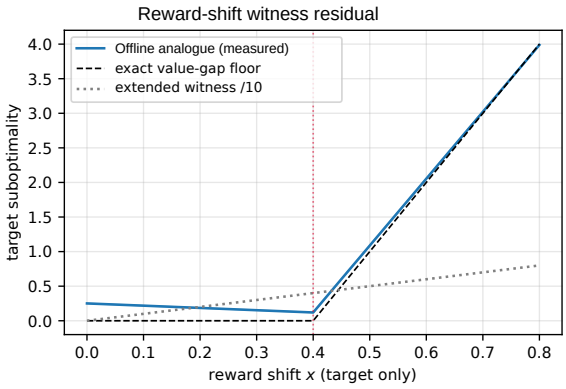


Figure 13: Reward-shift variant, six seeds. The offline analogue is compared with the target gap of the source-optimal policy and with a rescaled source-value witness residual. The boundary $x = 0.4$ follows from comparing the two corridor values.

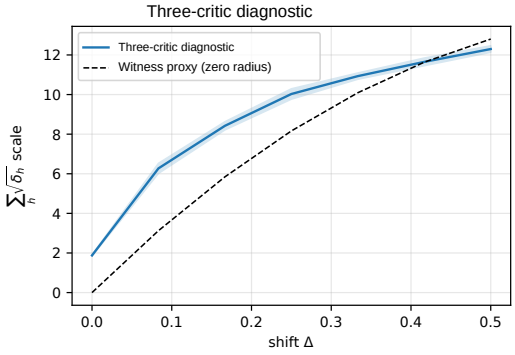


Figure 14: Three-critic model-based diagnostic, eight seeds, with $n = 400$ source and $m = 200$ on-policy target episodes. The diagnostic (band) is compared with the source-value witness residual (dashed).

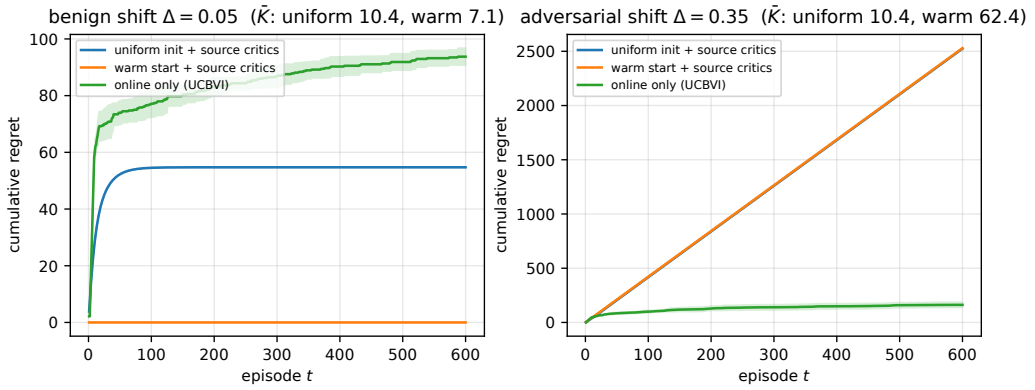


Figure 15: Warm-start ablation, eight seeds. Both source-critic runs share data, radii, and learning rate and differ only in the initial policy. Left, under benign shift, the aligned warm start ($\bar{K} = 7.1$, against $H \log 2 \approx 10.4$ for the uniform start) has nearly zero regret. Right, under adversarial shift, the misled warm start ($\bar{K} \approx 62.4$) neither helps nor hurts.

Warm starts. Theorem 22 separates two channels through which offline data help online learning, the optimization penalty, which depends on a bound for the initialization divergence (25), and the realized transfer term, which can also change with the initialization. We run Algorithm 3 twice with the same source data, radii, and learning rate, starting once from the uniform policy and once from the final offline iterate of Algorithm 2, and compare both runs with the online baseline; $\bar{K}$ is computed exactly by dynamic programming. Both runs use the same learning rate so that only the initialization differs, whereas the theorem tunes the rate to a divergence bound $D \geq \bar{K}$ and gives an optimization term proportional to $\sqrt{D}$. Under benign shift the offline policy agrees with $\pi^{\star}_{\mathrm{tar}}$ along its route, $\bar{K}$ drops from $H \log 2 \approx 10.4$ to 7.1, and the warm-started run has nearly zero regret because its initialization already selects the optimal corridor, while the uniform run incurs about 55 before converging (Figure 15). Under adversarial shift the offline policy is confidently wrong, $\bar{K}$ rises to about 62.4, and the two runs follow the same misled trajectory. The divergence is evaluated without flooring small initialization probabilities, since flooring would understate the adversarial divergence.

Near-boundary sweep. Figure 16 sweeps $\Delta \in \{0.08, 0.10, 0.12, 0.14, 0.16\}$ around the decision boundary $1 - 0.6^{1/4} \approx 0.12$. Below the boundary, pooling and splitting have zero regret because their policies select the target-optimal corridor. Above it, pooling keeps the corridor preferred by the source over this horizon, splitting can switch after detection, and the source-informed and block variants deteriorate. The master reduces regret relative to the source-informed variants at the two largest shifts but can remain worse than the online and splitting baselines.

Covered directions. Figure 17 simulates the covered-direction testing gadget of the linear lower bound at tilt $\theta = 0.15$, source visitation $p = 2\lambda$, and $T = 100$ target observations. The error of the empirical-mean test does not depend on the source sample size when $\lambda = 0$ and decreases with source information when $\lambda > 0$, which matches the two parts of Theorem 42.

Same-block certification. Figure 18 compares same-block updating with skipping on fallback, previous-block updating, and a handoff to target-only learning. The test declares a separation when a pair with at least five visits has an empirical total-variation distance larger than the sum of its source and target radii. The mean numbers of fallbacks are 0, 5.75, and 8 out of eight tests in the matched, benign, and adversarial settings; the terminal block is not tested, since no update follows it. The skip rule can slow improvement under small shift and limits the damage under larger shift. The plotted handoff reuses all previously collected target counts and replans at block boundaries, whereas Theorem 4 restarts a fresh online learner.

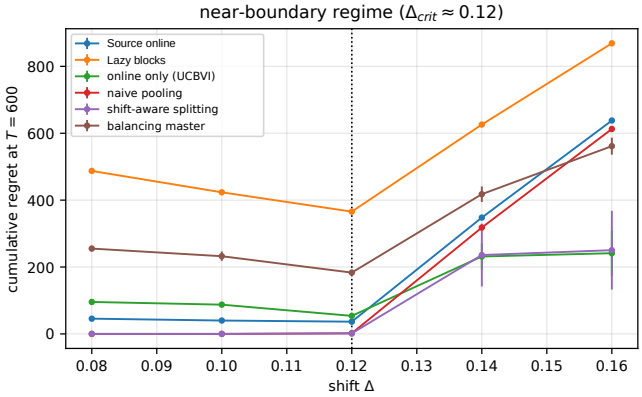


Figure 16: Near-boundary sweep, six seeds per shift, final cumulative regret at $T = 600$. The dotted line marks the boundary $\Delta_{\mathrm{crit}} \approx 0.12$ at which the target-optimal corridor changes.

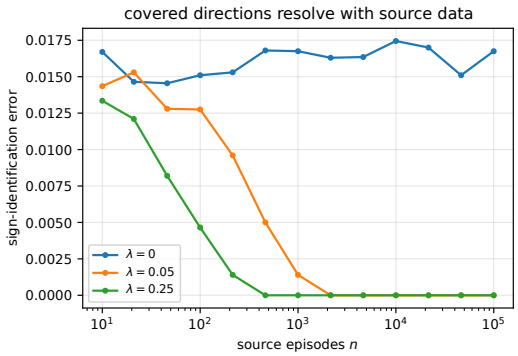


Figure 17: Monte Carlo simulation of the covered-direction gadget, 20,000 replications per point; two-standard-error binomial intervals are below 0.004 and are not drawn. The error at $\lambda = 0$ is constant in n, which corresponds to the persistent floor of part 1 of the linear lower bound, while for $\lambda > 0$ the error vanishes on the scale of $n\lambda$, as in part 2.

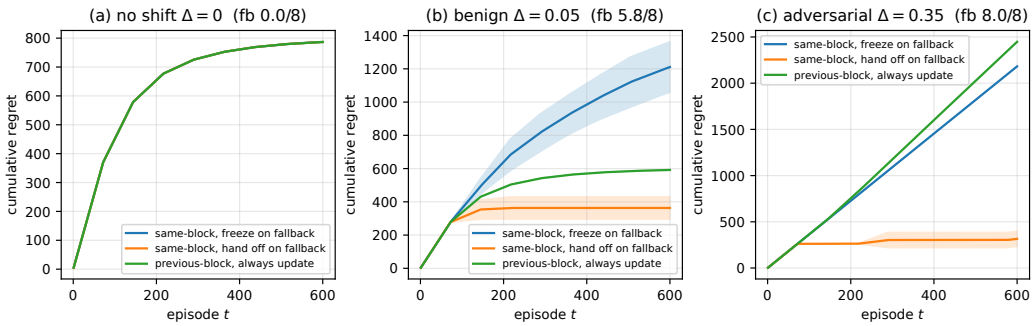


Figure 18: Same-block model-based ablation, eight seeds, with bands of ± 2 sample standard errors. The annotations count the eight nonterminal tests at $T = 600$ and $w = 72$; the terminal block contains 24 episodes.

H.3 Existing deep reinforcement learning methods under friction shift

This study asks how existing deep reinforcement learning methods use source data under dynamics shift, in the offline-source, online-target setting of ODRL (Lyu et al., 2024b). None of the compared methods implements the source-consistent optimizer, the Bellman-residual test, or the master analyzed in this paper; the finite-class study of Section 6 examines those mechanisms directly.

Environment and protocol. We use `halfcheetah-friction`, with the target ground friction scaled by 0.5 or 5.0 relative to the source; ODRL contains further friction levels, which we did not run. The source dataset is D4RL `halfcheetah-medium-v2` (Fu et al., 2020), with 10^6 transitions. Each learner receives this dataset and collects up to 10^6 target-environment transitions. Evaluation takes place every 10^4 target steps over ten episodes. Three methods, two friction levels, and three seeds per condition give 18 runs.

Compared methods. RLPD (Ball et al., 2023) trains on minibatches drawn half from the source buffer and half from the target buffer in the implementation used here, and it does not screen source transitions for dynamics mismatch. BC+VGDF (Xu et al., 2023) uses value-guided filtering to select the source transitions it learns from. CQL+SAC is the ODRL combination of conservative offline learning (Kumar et al., 2020) with online actor-critic learning. The three methods differ in architecture, losses, and update schedules as well as in their use of source data, so their comparison is not a controlled ablation. All three continue to learn from target observations throughout training.

Evaluation measures. For each target condition, we report the normalized score

$$S(t) = 100\,\frac{J(\pi_t) - J_{\mathrm{random}}}{J_{\mathrm{expert}} - J_{\mathrm{random}}},$$

where $J(\pi_t)$ is the mean evaluation return at training step t and the reference scores are those of the benchmark for that condition. A score of 100 corresponds to the expert reference, and larger scores are possible. We report checkpoint scores, the final score, and the area under the evaluation curve normalized by its step range. For recorded checkpoints $t_1 < \cdots < t_m$, this is

$$\mathrm{AUC} = \frac{1}{t_m - t_1} \sum_{j=1}^{m-1} (t_{j+1} - t_j)\,\frac{S(t_j) + S(t_{j+1})}{2}.$$

The weights use environment steps, although Figure 19 uses a logarithmic horizontal axis. Within a fixed condition, a larger AUC corresponds to a smaller integrated shortfall from any fixed reference score. The AUC is computed from evaluation episodes and therefore differs from the cumulative training regret of our theorems.

Results over the interaction budget. Table 10 reports final scores and AUC, and Table 11 reports intermediate checkpoints. At friction 0.5, RLPD reaches a mean final score of 128.1, compared with

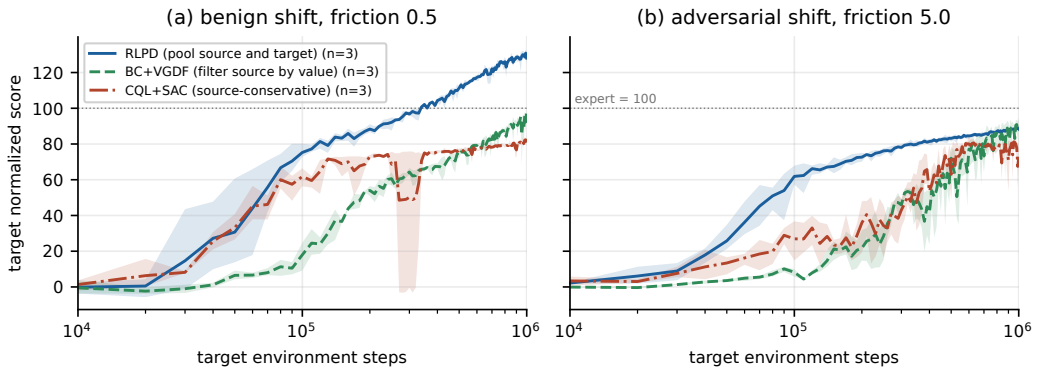


Figure 19: ODRL HalfCheetah with friction multipliers 0.5 and 5.0. Normalized target evaluation score against target-environment steps on a logarithmic axis. Curves are means over three seeds, and shading shows the range across seeds. The panel labels benign and adversarial refer to the two friction levels. RLPD has the highest mean AUC in both conditions; its final-score advantage over BC+VGDF is much smaller at friction 5.0, and CQL+SAC has the lowest final score in both conditions.

Table 10: Final normalized score and step-weighted AUC on ODRL HalfCheetah, mean ± standard deviation over three seeds. Bold marks the largest mean within a condition.

	Friction 0.5		Friction 5.0	
Method	Final score	AUC	Final score	AUC
RLPD	**128.1** ± 3.5	**104.1** ± 1.9	**88.4** ± 1.0	**77.3** ± 0.8
BC+VGDF	98.4 ± 1.3	66.8 ± 0.5	87.5 ± 9.4	56.0 ± 4.4
CQL+SAC	80.9 ± 0.6	70.6 ± 3.0	72.4 ± 6.7	58.7 ± 4.2

98.4 for BC+VGDF and 80.9 for CQL+SAC, and it also has the largest AUC, 104.1, compared with 66.8 and 70.6. At 50k steps, however, CQL+SAC has the higher mean score in this condition, with substantial variability across seeds, so the ranking depends on the interaction budget.

At friction 5.0, RLPD again has the highest mean AUC, 77.3, compared with 56.0 for BC+VGDF and 58.7 for CQL+SAC. The final means of RLPD and BC+VGDF are close, 88.4 and 87.5, while CQL+SAC ends at 72.4; with three seeds, the small difference between the first two is not a reliable ranking. Their different AUCs show that similar endpoints can follow quite different learning curves. The two conditions use different reference scores, so the normalized values are best compared within a condition.

Relation to the theory. Persistent mixing of source and target data remains competitive on this task at both friction levels, so the size of the physical shift alone does not determine which method learns fastest. This agrees with the policy-dependent view of the analysis, in which transfer difficulty depends on residuals along the policies being evaluated and on coverage; we did not estimate these quantities here. The lower bounds concern constructed hard families and specified target budgets and make no prediction about the ranking of these methods. The study includes no target-only baseline, so it does not measure the benefit of the source data itself, and a common deep learner implementing the tests and handoff of Theorems 2 and 4 remains to be evaluated.

Table 11: Intermediate normalized evaluation scores, mean $\pm$ standard deviation over three seeds. Budgets count target-environment steps.

Method	Friction	50k	100k	200k	500k
RLPD	0.5	30.7 ± 26.1	75.3 ± 4.3	88.4 ± 2.4	111.4 ± 1.4
BC+VGDF	0.5	6.5 ± 2.5	18.0 ± 8.0	54.0 ± 2.9	73.5 ± 2.1
CQL+SAC	0.5	33.2 ± 7.2	61.8 ± 5.1	73.1 ± 0.9	76.8 ± 1.2
RLPD	5.0	25.8 ± 7.8	61.9 ± 5.4	72.6 ± 2.4	83.2 ± 1.6
BC+VGDF	5.0	3.6 ± 0.3	8.2 ± 2.2	21.7 ± 9.0	72.1 ± 7.6
CQL+SAC	5.0	13.4 ± 5.5	26.9 ± 11.1	35.9 ± 15.0	63.1 ± 13.3