Learning When to Stop: Selective Imitation Learning Under Arbitrary Dynamics Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

Behavior cloning provides strong imitation learning guarantees when training and test environments share the same dynamics. However, in many deployment settings the test environment's transitions differ from training, and classical offline IL offers no recourse: the learner must commit to an action at every state, even when its demonstrations are uninformative and could lead to arbitrary degradation of performance. This motivates the study of *selective* imitation, where the learner may choose to *stop* when it cannot act reliably. We introduce a model for selective imitation under arbitrary dynamics shift: given labeled expert demonstrations from a training environment and unlabeled state trajectories from the same expert in a test environment, the learner outputs a *selective policy* that is *complete* (rarely stops in training) and *sound* (incurs low regret before stopping in test). Our algorithm, SeqRejectron, constructs a stopping rule using a small set of *validator policies* whose size is independent of the horizon or policy class. For deterministic policies, this yields horizon-free $\tilde{O}(\log |\Pi|/\epsilon^2)$ sample complexity, assuming sparse costs. For stochastic policies, we obtain analogous horizon-free guarantees using a cumulative Hellinger stopping time. We extend the framework to misspecified experts and different expert policies across train and test and obtain results that gracefully degrade with the amount of misspecification.

1 Introduction

A central challenge in deploying learned policies is *environment shift*: the dynamics governing state transitions at test time may differ from those encountered during training. This arises ubiquitously: in sim-to-real transfer for robotics [Tobin et al., 2017], in autonomous driving where conditions vary between simulation and the road [Codevilla et al., 2019], and in any domain where a policy trained on demonstrations must act in a changed environment.

Classical imitation learning offers strong guarantees when training and test environments coincide, but these guarantees become vacuous under even modest dynamics shift. Even with infinite data, a policy may perform arbitrarily poorly if the test environment's state support differs from the training environment's. The fundamental issue is that it cannot recognize when its training data no longer provides reliable guidance, and has no mechanism to stop before things go wrong. This motivates a basic question: *can a policy learn not only what to do, but when to stop?*

We develop a framework for *learning to imitate with the option to abstain*, drawing on the *PQ learning* paradigm from selective classification [Goldwasser et al., 2020]. We consider a practical setting where the learner is given labeled (state, action) expert demonstrations from a training environment M, but only unlabeled (state-only) expert trajectories from a test environment N. This asymmetry naturally models settings where passive observation is cheap and abundant, but action telemetry is costly or proprietary. For instance, in autonomous driving, a learner might train on rich kinematic data from a

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

sensor-equipped fleet, but must generalize to new cities using massive datasets of public dashcam footage—which provide valid visual states of expert driving, but completely lack the underlying steering commands [Xu et al., 2017, Torabi et al., 2018, Codevilla et al., 2019]. In healthcare, a sepsis treatment policy trained on a fully instrumented ICU system may be deployed to a setting where precise dosage decisions are unavailable [Komorowski et al., 2018, Finlayson et al., 2021]. In algorithmic trading, a model trained on historical market conditions with full access to proprietary executed trades must often navigate novel macroeconomic regimes where only public ticker data is observable, leaving the exact buy and sell actions of top experts completely hidden.

The goal is to learn a *selective policy* that acts when its training data supports confident prediction and stops execution otherwise. We require *completeness* (the policy rarely stops when deployed in the training environment) and *soundness* (whenever the policy chooses to act in the test environment, it matches the expert's performance). A policy that always stops is trivially sound but useless; one that never stops is trivially complete but unsafe. The tension makes the problem nontrivial.

1.1 Our approach and results

We certify trajectory prefixes using *validator policies*: competing explanations of the expert drawn from the training-consistent version space. The learner executes a base policy π_0, continuing only while every validator agrees and abstaining at the first disagreement. Because a small validator set always suffices, the resulting sample complexity depends on the horizon H only through the trajectory cost scale $C_{\max}$ and the complexity of the policy class. We call this approach SeqRejectron. It generalizes the selective classification algorithm of Goldwasser et al. [2020] to sequential decision-making, mirroring recent horizon-independent rates for imitation learning [Foster et al., 2024] while accommodating dynamics shift. We analyze it in three increasingly general settings.

Deterministic policies (Section 3). We first develop our framework for finite classes of deterministic policies. In this setting, each validator triggers at the first state where it disagrees with the base policy. We prove that a surprisingly sparse set of these validators is always sufficient to safely stop execution. Additionally, we demonstrate that this approach is oracle-efficient by providing an exact reduction to multiple-instance learning [Maron and Lozano-Pérez, 1997].

Stochastic policies (Section 4). The binary disagreement test is too coarse for stochastic policies, where two distributions may differ softly at every step without any single step being alarming. We replace it with a cumulative Hellinger stopping time that tracks a trajectory-level divergence budget, preserving the validator template. The main result gives finite-samples bounds on both the stopping rate and the regret. This bound has both a *cost-driven* and *variance-driven* exponent. We prove an $\Omega(1/\epsilon^3)$ labeled-sample lower bound already in the single-step stochastic setting, matching the cost-driven part of the upper bound and leaving the variance-driven exponent open.

Misspecified policy classes (Section 5). Under misspecification, the exact-consistency version space might not contain the expert. We replace it with a symmetrically regularized game that softly penalizes source-data disagreement, ensuring the validator distribution remains well-defined and sparse. We show that the guarantees degrade gracefully under misspecification.

1.2 Related work

Our work builds on three strands of literature.[1] First, it is closely related to imitation learning and recent analyses of behavior cloning. Behavior cloning reduces imitation learning to supervised prediction, but is classically limited by compounding error under distribution shift [Pomerleau, 1988, Ross and Bagnell, 2010, Ross et al., 2011]. Recent theory has further characterized the statistical limits of offline imitation learning and identified settings in which horizon-independent guarantees are possible [Rajaraman et al., 2020, Swamy et al., 2022, Foster et al., 2024]. We study a modified setting, where the learner is trained offline, but the test dynamics may shift arbitrarily.

Second, our framework is inspired by selective classification and learning under arbitrary distribution shift. Closest to our setting is the PQ framework, which studies source-labeled, target-unlabeled learning with abstention under arbitrary shift [Goldwasser et al., 2020, Kalai and Kanade, 2021, Goel et al., 2024]. Our framework can be viewed as a sequential analogue of PQ learning, where the learner abstains along a trajectory through a stopping rule rather than on a single prediction.

[1] We defer a more comprehensive literature review to Section A.

Third, our work is related to robustness and uncertainty in reinforcement learning and robotics, including robust control, sim-to-real transfer, domain randomization, and OOD detection [Iyengar, 2005, Nilim and El Ghaoui, 2005, Wiesemann et al., 2013, Tobin et al., 2017, Peng et al., 2018, Chebotar et al., 2019, Chae et al., 2022, Haider et al., 2023]. These approaches aim to make policies robust or adaptive to changing environments, but are typically heuristic, make strong assumptions about the nature of the shift, or require online access to the test distribution. By contrast, we study distribution-free guarantees under arbitrary dynamics shift using only labeled source demonstrations and unlabeled target trajectories, with abstention[2] as the safety mechanism.

2 Problem Formulation

Markov decision processes. A finite-horizon Markov decision process (MDP) is a tuple $M = (\mathcal{S}, \mathcal{A}, P, c, H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space[3], H is the horizon, $P = \{P_h\}_{h=0}^{H-1}$ specifies the state evolution, and $c = \{c_h\}_{h=1}^{H}$ are per-step cost functions $c_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$. Here $P_0 \in \Delta(\mathcal{S})$ is the initial-state distribution, and for each $h = 1, \dots, H-1$, $P_h(\cdot \mid s, a) \in \Delta(\mathcal{S})$ is the transition kernel. A (possibly nonstationary, possibly randomized) policy is $\pi = \{\pi_h\}_{h=1}^{H}$ with $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$. Together, (M, π) induce a distribution over trajectories $(s_1, a_1), \dots, (s_H, a_H)$ via $s_1 \sim P_0$, $a_h \sim \pi_h(\cdot \mid s_h)$, and $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$ for $h = 1, \dots, H-1$. We denote the expected cumulative cost of π in M by $J_M(\pi; c) \coloneqq \mathbb{E}^{M,\pi}\left[\sum_{h=1}^{H} c_h(s_h, a_h)\right]$.

Stopping times. A key feature of our framework is that the learner may stop execution mid-trajectory. We formalize this decision to abstain via a *stopping time* adapted to the observed history. Denote the canonical sample space of full state-action trajectories by $\Omega = (\mathcal{S} \times \mathcal{A})^H$. Let $\mathbb{G} = \{\mathcal{G}_h\}_{h=1}^{H}$ be the pre-action filtration on Ω, where $\mathcal{G}_h = \sigma(s_1, a_1, \dots, s_{h-1}, a_{h-1}, s_h)$. Thus $\mathcal{G}_h$ contains exactly the information available at step h, just before the learner chooses a_h. For convenience, we also let $\mathcal{G}_{H+1} = \sigma(s_1, a_1, \dots, s_H, a_H)$. We define the point of abstention as a stopping time $\tau : \Omega \to \{1, \dots, H, H+1\}$ adapted to the filtration $\mathbb{G}$. The event $\{\tau = h\}$ signifies that the learner stops execution at step h, having observed the history up to s_h. We adopt the convention that $\tau = H + 1$ if the learner completes the full trajectory without abstaining.

Problem statement. Let $M \coloneqq (\mathcal{S}, \mathcal{A}, P, c, H)$ and $N \coloneqq (\mathcal{S}, \mathcal{A}, Q, c, H)$ be MDPs that share state space, action space, costs, and horizon, but may differ in their dynamics $P = \{P_h\}_{h=0}^{H-1}$ and $Q = \{Q_h\}_{h=0}^{H-1}$. We refer to M as the *training environment* and N as the *test environment*. For any environment $E \in \{M, N\}$ and policy π, we write $\mathcal{P}_E^{\pi}$ for the law of the trajectory under (E, π), and $\mathcal{P}_{E,\text{state}}^{\pi}$ for the law of the state trajectory $s_{1:H}$. Let $C_{\max}$ denote an upper bound on the total trajectory cost $\sum_{h=1}^{H} c_h(s_h, a_h)$.

Let Π be a finite policy class and let $\pi^\star \in \Pi$ denote an expert policy whose behavior we seek to imitate. We do not assume that $\pi^\star$ is optimal for c. The learner observes demonstrations generated by $\pi^\star$, but does not observe realized costs or expert actions in N:

- **Labeled training rollouts:** drawn i.i.d. from $(M, \pi^\star)$, $\mathcal{D}_{\text{train}} = \{(s_{k,1:H}, a_{k,1:H})\}_{k=1}^{m}$.
- **Unlabeled test rollouts:** state-only trajectories drawn i.i.d. from $(N, \pi^\star)$, $\mathcal{D}_{\text{test}} = \{t_{l,1:H}\}_{l=1}^{n}$.

Without the option to abstain, the problem is poorly posed under environment shift: if the expert-induced state supports of M and N are disjoint, even infinite labeled data from M cannot determine the expert's actions on test-time states. To maintain low regret across all valid expert hypotheses, the learner is forced to *abstain* on unseen test states. We therefore allow the learner to stop execution mid-trajectory. We formalize this via the notion of a *selective policy*.

Definition 1 (Selective policy, stopped regret, and stopping rate)**.** *A* selective policy *is a pair (π, τ), where $\pi \in \Pi$ is a policy and τ is a stopping time adapted to the pre-action filtration $\mathcal{G}_h = \sigma(s_1, a_1, \dots, s_{h-1}, a_{h-1}, s_h)$. When deployed, it incurs cost only on the time steps before τ. We define the* stopped cost *and* stopped regret *(in the test environment N) as*

$$J_N(\pi, \tau; c) \coloneqq \mathbb{E}^{N,\pi}\left[\sum_{h=1}^{\tau-1} c_h(s_h, a_h)\right], \qquad \text{Regret}_N(\pi, \tau; c) \coloneqq J_N(\pi, \tau; c) - J_N(\pi^\star, \tau; c).$$

[2]Throughout, will will interchangeably refer to *abstention* as *stopping*.

[3]To simplify presentation, we assume throughout that $\mathcal{S}$ and $\mathcal{A}$ are countable; the definitions and results extend to general spaces with the appropriate measure-theoretic treatment.

We define the stopping rate *(in the training environment M) as* $\alpha_M(\pi, \tau) \coloneqq \Pr_{M,\pi}(\tau \leq H)$.

Intuitively, a selective policy can be thought of as an agent equipped with a fail-safe, executing its learned behavior only as long as it remains confident. For example, consider a self-driving car trained exclusively to navigate dry, sunny roads. If it is suddenly deployed in a blizzard where tire traction and visibility drastically change, a standard imitation policy might blindly attempt a normal turn and crash. A selective policy, by contrast, recognizes that these icy conditions fall outside its training support and safely pulls over before a catastrophic error occurs.

Problem 1 (Selective imitation learning: stopped regret formulation)**.** *Given labeled training rollouts $\mathcal{D}_{\mathrm{train}}$ and unlabeled test rollouts $\mathcal{D}_{\mathrm{test}}$, output a selective policy (π, τ) satisfying:*

- ***Completeness:*** *$\alpha_M(\pi, \tau) \leq \epsilon$. (The policy rarely abstains in M).*
- ***Soundness:*** *$\mathrm{Regret}_N(\pi, \tau; c) \leq \epsilon$. (The policy incurs low stopped regret in N).*

3 Selective Execution for Deterministic Policies

We first present our results for the case when $\pi^\star$ and Π are deterministic policies. We first analyze a per-step baseline that suffers an avoidable horizon factor H by treating rollouts as independent decisions. We then resolve this via a trajectory-level construction based on *validator policies* that certifies entire execution prefixes, yielding an improved sample complexity.

First attempt: Stepwise reduction to PQ-learning. A naive baseline applies Rejectron [Goldwasser et al., 2020] independently at each step. However, converting these per-step guarantees to a trajectory-level bound requires a union bound over H steps, yielding a suboptimal sample complexity of $\tilde{O}(H^2 C_{\max}^2 \log|\Pi|/\epsilon^2)$. This introduces an avoidable H^2 penalty even when costs are sparse; we defer the formal analysis of this stepwise reduction to Section B.2.

Trajectory-level validation via validator policies. The per-step reduction treats a trajectory as H independent decisions, introducing an avoidable horizon factor. However, sequential rollouts have inherent structure: until the learner diverges from the expert, stopping incurs only a prefix of the expert's cost. The danger is failing to stop before this first unrejected deviation. Therefore, our goal is to certify an entire execution prefix at once.

Fix a base policy $\pi_0 \in \Pi_{\mathsf{version}}$. To determine when it is safe to continue executing π_0, we compare it against other policies in the version space. For any $\pi \in \Pi_{\mathsf{version}}$, the first state where π disagrees with π_0 marks a point where π_0 is no longer justified by that competing explanation of the expert. Since no single competitor reliably detects all problematic rollouts across all trajectories, we compare π_0 against a *set* of validators, continuing only while every validator agrees with π_0.

Definition 2 (Validator-induced stopping time)**.** *For any base policy π_0 and set of policies $\Phi \subseteq \Pi_{\mathsf{version}}$, define the stopping time $\tau_{\pi_0,\Phi}$ as the first time along the realized trajectory at which at least one validator in Φ disagrees with π_0, or $H+1$ if no such disagreement occurs:*

$$\tau_{\pi_0,\Phi}(T) \coloneqq \min\left(\left\{h \in [H] : \exists \pi \in \Phi,\ \pi_h(s_h) \neq \pi_{0,h}(s_h)\right\} \cup \{H+1\}\right). \tag{1}$$

The key question is therefore how to choose the validator set Φ. Ideally, we would like Φ to be such that $\Pr_{\pi_0,N}[\tau_{\pi_0,\Phi} \leq \tau_{\pi_0,\{\pi^\star\}}]$ is close to 1. That is, with high probability, Φ tells us to stop before π_0 deviates from $\pi^\star$. This tells us that until the stopping time $\tau_{\pi_0,\Phi}$, the policies π_0 and $\pi^\star$ are approximately coupled, and hence incur similar cost. We would also like Φ to be sparse; this prevents us from rejecting too many trajectories, which is a form of overfitting.

The key difficulty, however, is that Φ must be computed *without knowledge of* $\pi^\star$. We will thus ask: *Can we find a sparse validator set which stops before* any *policy in* Π_{version} *deviates from* π_0*?* There is no obvious reason that one small deterministic validator set should work uniformly against every policy in Π_{version}. Fortunately, however, the following result shows that it is always possible to compute a *distribution* over sparse validator sets which satisfies this guarantee with high probability. We thus sample from this distribution to obtain a sparse validator set for $\pi^\star$.

Definition 3 (ρ-valid validator distribution)**.** *Fix $\pi_0 \in \Pi_{\mathsf{version}}$ and unlabeled test trajectories $\mathcal{D}_{\mathsf{test}} = \{T_j\}_{j=1}^n$. A distribution q over subsets $\Phi \subseteq \Pi_{\mathsf{version}}$ is called ρ-*valid *for π_0 if* $\sup_{\pi \in \Pi_{\mathsf{version}}} \mathbb{E}_{\Phi\sim q}\left[\frac{1}{n}\sum_{j=1}^n \mathbf{1}\left[\tau_{\pi_0,\Phi}(T_j) > \tau_{\pi_0,\{\pi\}}(T_j)\right]\right] \leq \rho$.

In words, a ρ-valid distribution places mass on validator sets that, on almost all empirical test trajectories, stop no later than when *any* fixed competing policy in the version space diverges from π_0. The next lemma shows that this averaged notion of coverage can always be achieved by distributions supported on *small* validator sets. Importantly, the support-size bound depends only on the target violation level ρ; downstream, ρ will only depend on ϵ and $C_{\max}$.

Lemma 1 (Existence of sparse valid validator distributions)**.** *For every $\rho \in (0,1)$, there exists a ρ-valid distribution $q^\star$ over subsets $\Phi \subseteq \Pi_{\mathsf{version}}$ such that every $\Phi \in \mathsf{supp}(q^\star)$ satisfies $|\Phi| \le \left\lceil \frac{1}{\rho} \right\rceil$.*

The proof uses only two ingredients: the minimax theorem and the exchangeability of i.i.d. draws. Importantly, there is no dependence on the state space, action space, horizon, or size of Π. This generality is important, and we will reuse it for stochastic policies (Section 4) and the misspecified setting (Section 5). We state the general version as Lemma 5 in the appendix.

How do we compute $q^\star$? To keep the statistical argument separate from the computation, we package Lemma 1 into a subroutine $\textsc{SparseValidatorDist}(\Pi_{\mathsf{version}}, \pi_0, \mathcal{D}_{\mathsf{test}}, \rho, \xi, \delta)$, which with probability $1-\delta$, returns a $(\rho+\xi)$-valid distribution over subsets of Π_{version} with size $\le \lceil 1/\rho \rceil$. Here ξ is a computational slack term that can be driven arbitrarily close to 0. A concrete, oracle-efficient implementation is given in Section B.6, where we realize this subroutine via no-regret dynamics.

Algorithm and guarantees. We are now ready to present our algorithm for selective imitation with deterministic policies. The algorithm (Algorithm 1) has a clean, simple structure. The labeled data define a version space and a base policy π_0. The unlabeled test trajectories are used to construct a sparse validator distribution, from which we sample a validator set. The returned selector continues while every validator agrees with π_0 and abstains at the first disagreement.

Algorithm 1 SeqRejectron for deterministic policies

Require: deterministic policy class Π, labeled rollouts $\mathcal{D}_{\mathsf{train}} = \{(s_{ih}, a_{ih})\}_{i\in[m],\, h\in[H]}$, unlabeled rollouts $\mathcal{D}_{\mathsf{test}} = \{T_j\}_{j=1}^n$, tolerance η, computational slack ξ, confidence δ

1: Compute the training-consistent version space

$$\Pi_{\mathsf{version}} \leftarrow \{\pi \in \Pi : \pi_h(s_{ih}) = a_{ih} \text{ for all } i \in [m],\ h \in [H]\}\,.$$

2: Choose any base policy $\pi_0 \in \Pi_{\mathsf{version}}$.

3: Let $q^\star \leftarrow \textsc{SparseValidatorDist}(\Pi_{\mathsf{version}}, \pi_0, \mathcal{D}_{\mathsf{test}}, \eta/2, \xi, \delta/5)$.

4: Sample $\Phi_1, \dots, \Phi_k \overset{\text{i.i.d.}}{\sim} q^\star$ with $k \leftarrow \lceil \log_2(5/\delta) \rceil$ and set $\Phi \leftarrow \bigcup_{r=1}^k \Phi_r$.

5: **return** the selective policy $(\pi_0, \tau_{\pi_0,\Phi})$.

Theorem 1 (SeqRejectron guarantee for deterministic classes)**.** *Run Algorithm 1 with hyperparameters $\eta, \delta, \xi > 0$. Define $Z := (\lceil \log_2(5/\delta) \rceil \lceil 2/\eta \rceil + 1) \log|\Pi| + \log(5/\delta)$. Then, with probability at least $1-\delta$, the selective policy $(\pi_0, \tau_{\pi_0,\Phi})$ satisfies $\alpha_M(\pi_0, \tau_{\pi_0,\Phi}) \le \frac{2Z}{m}$ and $\mathrm{Regret}_N(\pi_0, \tau_{\pi_0,\Phi}; c) \le C_{\max}\left(\eta + 2\xi + \sqrt{\frac{2(\eta+2\xi)Z}{n}} + \frac{3Z}{n}\right)$.*

Corollary 1. *Suppose $m = n \ge 6\lceil \log_2(5/\delta) \rceil \log|\Pi|$. Set $\eta := \sqrt{6\lceil \log_2(5/\delta) \rceil \log|\Pi|/n}$. Then, with probability at least $1-\delta$, both the stopping rate and $\mathrm{Regret}_N/C_{\max}$ scale as $\tilde{O}(\sqrt{\log|\Pi|/n})$.*

The role of weight sharing in $\log|\Pi|$. The sample complexity of Theorem 1 scales with $\log|\Pi|$. When Π shares parameters across time steps (e.g., a single neural network mapping states to actions), $\log|\Pi|$ is independent of H and the bound is horizon-free in the strongest sense. In general, H enters only through $C_{\max}$ and $\log|\Pi|$, consistent with Foster et al. [2024].

Bounding the test-side abstention. It is also natural to ask how often the selective policy abstains in N. When the environment shift is bounded in total variation, the test-side stopping rate $\alpha_N(\pi_0, \tau)$ inherits the training-side guarantees up to a TV correction term, plus an *additional* sequential penalty that arises because the deployed policy's actions can steer the trajectory out of the safe prefix before the validator stops. See Section B.5 in the appendix for a precise statement.

4 Selective Execution for Stochastic Policies

Throughout this section, Π denotes a finite class of (possibly stochastic, possibly nonstationary) policies, and $\pi^\star \in \Pi$ is the expert. Recall that the deterministic construction rests on a binary primitive: two policies either agree or disagree at each state. For stochastic policies this is too coarse, since two distributions may differ softly at every step without any single discrepancy being alarming. The right question is whether soft discrepancies *accumulate* enough to cause meaningful divergence from the expert. We therefore replace three ingredients from the deterministic construction.

First, we take π_0 to be the MLE and for $\gamma > 0$ define the version space as a log-loss ball,

$$\Pi^{\gamma}_{\mathsf{version}} \coloneqq \big\{\pi \in \Pi : \mathsf{LogLoss}(\pi, \mathcal{D}_{\mathsf{train}}) \le \mathsf{LogLoss}(\pi_0, \mathcal{D}_{\mathsf{train}}) + \gamma\big\},$$

where $\mathsf{LogLoss}(\pi, \mathcal{D}_{\mathsf{train}}) \coloneqq -\frac{1}{m}\sum_{i=1}^{m}\sum_{h=1}^{H} \log \pi_h(a_{ih} \mid s_{ih})$. Second, we replace the first-disagreement stopping time with one based on cumulative squared Hellinger distance[4]: for $p, q \in \Delta(\mathcal{A})$, let $d^2_H(p,q) := 1 - \sum_a \sqrt{p(a)q(a)}$, and for $\theta > 0$ define

$$\tau^{\theta}_{\pi_0,\Phi}(T) \coloneqq \min\left(\left\{h \in [H] : \exists \pi \in \Phi,\ \sum_{k=1}^{h} d^2_H\big(\pi_k(\cdot \mid s_k), \pi_{0,k}(\cdot \mid s_k)\big) > \theta\right\} \cup \{H+1\}\right).$$

Third, we define an analog to the SPARSEVALIDATORDIST subroutine: STOCHASTICSPARSEVALIDATORDIST, which when called with parameters ρ, θ, and δ returns a distribution $q^\star$ supported over subsets $\Phi \subseteq \Pi^{\gamma}_{\mathsf{version}}$ of size at most $\lceil 1/\rho \rceil$, such that[5], with probability $1-\delta$, for every $\pi \in \Pi^{\gamma}_{\mathsf{version}}$, it holds that $\mathbb{E}_{\Phi\sim q^\star}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left[\tau^{\theta}_{\pi_0,\Phi}(T_j) > \tau^{\theta}_{\pi_0,\{\pi\}}(T_j)\right]\right] \le \rho$.

Guarantees. The stochastic SeqRejectron follows the same template as Algorithm 1, save for the changes mentioned above. To formalize our guarantees, let $\mathcal{P}_N^{\pi|_\tau}$ denote the distribution of trajectories induced by π, stopped at time τ in environment N (i.e., $\mathcal{P}_N^{\pi|_\tau}$ is supported over trajectory *prefixes*).[6]

Theorem 2 (SeqRejectron guarantee for stochastic classes)**.** *There exists an algorithm, the stochastic* SeqRejectron*, with hyperparameters $\eta, \theta, \gamma, \delta > 0$. Define $K_{\mathsf{ens}} \coloneqq \lceil \log_2(4/\delta)\rceil \lceil 2/\eta \rceil$ and $Z \coloneqq (K_{\mathsf{ens}}+1)\log|\Pi| + \log(4/\delta)$. If $n \ge 8Z/\eta$ and $m \ge (\log|\Pi| + \log(8/\delta))/\gamma$, then with probability at least $1-\delta$, this algorithm returns a selective policy $(\pi_0, \tau^{\theta}_{\pi_0,\Phi})$ satisfying $\alpha_M(\pi_0, \tau^{\theta}_{\pi_0,\Phi}) \le K_{\mathsf{ens}}\big(\frac{12}{\theta} + 48\big)\gamma$ and $D^2_H\big(\mathcal{P}_N^{\pi_0|_\tau}, \mathcal{P}_N^{\pi^\star|_\tau}\big) \le 3(\theta+\eta)$.*

Corollary 2. *There exists a choice of parameters γ, η, θ for stochastic* SeqRejectron *(depending only on m, n, $\sigma^2_{\pi^\star}$, $\log|\Pi|$, and δ) such that with probability at least $1-\delta$, both $\alpha_M(\pi_0, \tau^{\theta}_{\pi_0,\Phi})$ and $\mathrm{Regret}_N(\pi_0, \tau^{\theta}_{\pi_0,\Phi}; c)/C_{\max}$ are $O\left(\max\left\{\frac{(\log(1/\delta)+\log|\Pi|)^{2/5}}{m^{1/5}}, \frac{(\log(1/\delta)+\log|\Pi|)^{1/2}}{n^{1/4}}\right\}\right)$.*

To prove Corollary 2, we convert the guarantee of Theorem 2 into a total variation bound. A natural question is whether this bound can be sharpened, as Theorem 2 is stated as a guarantee on the *squared Hellinger* distance of the stopped trajectories. Indeed, Foster et al. [2024] use a variance-dependent Hellinger change-of-measure lemma to avoid this apparent looseness in the *full horizon* setting. Their bounds are stated in terms of an expert variance parameter $\sigma^2_{\pi^\star}$; in our setting, however, this variance parameter would depend on the *learned* stopping time. Thus, it is not clear how to make full use of the squared Hellinger guarantee of Theorem 2 in our setting. It is an interesting question for future work whether or not a different algorithm or analysis can sharpen Corollary 2. However, in the following section, we present a natural variation of our objective which is more amenable to the sharper Hellinger change-of-measure, and prove sharper rates accordingly.

Computational implementation and offline oracles. Unlike the deterministic setting, the cumulative Hellinger thresholding required for stochastic policies does not yet admit a known oracle-efficient reduction to standard ERM. We leave the development of computational oracles for this cumulative objective—or rigorous guarantees for practical surrogates—to future work.

[4] For deterministic Π, d^2_H reduces to the disagreement indicator, so $\tau^{\theta}_{\pi_0,\Phi}$ with $\theta < 1$ recovers (1).

[5] The existence of q^* follows the exact same combinatorial argument as Lemma 1, and is stated as Lemma 5 in the appendix.

[6] More formally, $\mathcal{P}_N^{\pi|_\tau}$ is a probability measure over the space of finite histories $\mathcal{H} = \bigcup_{h=1}^{H+1}(\mathcal{S}\times\mathcal{A})^{h-1}\times\mathcal{S}$. A random draw from this measure takes the form $(s_1, a_1, \ldots, s_{\tau-1}, a_{\tau-1}, s_\tau)$ on the event $\{\tau \le H\}$, and is the full trajectory $(s_1, a_1, \ldots, s_H, a_H)$ on the event $\{\tau = H+1\}$.

4.1 Expert handoff interpretation and switched regret guarantees

So far, we have studied policies that can stop execution mid-trajectory. However, in collaborative autonomy—such as a self-driving car handing control back to a human driver upon encountering unfamiliar condition—the episode does not end at abstention. We would like to guarantee that the autonomous system does not drive the system into an unrecoverable state before handing off. Thus, one may consider the *switched policy* that follows the learner until τ and the expert thereafter.

Definition 4 (Switched policy and switched regret)**.** *For a base policy $\pi \in \Pi$, an expert policy $\pi^\star \in \Pi$, and a stopping time τ, the* switched policy $\pi^{\mathrm{sw}}[\pi, \tau, \pi^\star]$ *executes π on steps $1, \dots, \tau - 1$ and $\pi^\star$ from step τ onward. For costs $c = \{c_h\}_{h=1}^H$, we define the* switched regret *(in N) as*

$$\mathrm{Regret}_N^{\mathrm{sw}}(\pi, \tau; c) := J_N(\pi^{\mathrm{sw}}[\pi, \tau, \pi^\star]; c) - J_N(\pi^\star; c). \tag{2}$$

Problem 2 (Selective imitation learning: switched regret formulation)**.** *Given labeled training rollouts $\mathcal{D}_{\mathrm{train}}$ and unlabeled test rollouts $\mathcal{D}_{\mathrm{test}}$, output a selective policy (π, τ) such that*

- ***Completeness:*** *$\alpha_M(\pi, \tau) \le \epsilon$. (The policy rarely abstains in M).*
- ***Soundness:*** *$\mathrm{Regret}_N^{\mathrm{sw}}(\pi, \tau; c) \le \epsilon$. (The policy incurs low switched regret in N).*

The switched regret is directly controlled by our Lemma 2, which generalizes the Hellinger regret decomposition of Foster et al. [2024]. Below, let $\sigma_{\pi^\star}^2 := \sum_{h=1}^H \mathbb{E}_{\pi^\star}\left[(V_h^{\pi^\star}(x_h) - Q_h^{\pi^\star}(x_h, a_h))^2\right] \le C_{\max}^2$ denote the expert-variance parameter of Foster et al. [2024] where (x_h, a_h) is the state-action pair at time h under $\pi^\star$ and $V_h^{\pi^\star}, Q_h^{\pi^\star}$ are the corresponding value and action-value functions.

Lemma 2 (Hellinger-based regret decomposition for stopped trajectories; generalization of Theorem 3.1 of Foster et al. [2024])**.** *Assume $C_{\max} \ge 1$. Let $\pi^\star$ be the expert policy. For any learned policy $\hat{\pi}$ and stopping time τ, let π_{sw} denote the policy that executes $\hat{\pi}$ for $h < \tau$ and executes $\pi^\star$ for all remaining steps $h \ge \tau$. Then, for any $\epsilon \in (0, e^{-1})$, the expected regret of π_{sw} is bounded by:*

$$J(\pi_{\mathsf{sw}}) - J(\pi^\star) \le \sqrt{6\sigma_{\pi^\star}^2 \cdot D_H^2\big(\mathcal{P}_N^{\hat{\pi}_{|\tau}}, \mathcal{P}_N^{\pi^\star_{|\tau}}\big)} + O\left(C_{\max} \log(C_{\max}\epsilon^{-1})\right) \cdot D_H^2\big(\mathcal{P}_N^{\hat{\pi}_{|\tau}}, \mathcal{P}_N^{\pi^\star_{|\tau}}\big) + \epsilon.$$

Corollary 3. *Set $\gamma := (\log|\Pi| + \log(8/\delta))/m$. Let $Z := \log|\Pi| + \log(8/\delta)$. There exists a choice of parameters $\eta = \theta$ (depending only on m, n, $\sigma_{\pi^\star}/C_{\max}$, $|\Pi|$, and δ) such that with probability at least $1 - \delta$: $\alpha_M(\pi_0, \tau^\theta_{\pi_0,\Phi}) = O\left(\max\left\{\frac{(\sigma_{\pi^\star}/C_{\max})^{4/5} Z}{m^{1/5}}, \frac{Z}{m^{1/3}}, \frac{\sigma_{\pi^\star}}{C_{\max}}\frac{Z}{n^{1/4}}, \frac{Z}{n^{1/2}}\right\}\right)$ and $\mathrm{Regret}_N^{\mathrm{sw}}(\pi_0, \tau^\theta_{\pi_0,\Phi}; c) = O\left(\max\left\{\frac{(C_{\max}\sigma_{\pi^\star}^4)^{1/5} Z}{m^{1/5}}, \frac{C_{\max} Z}{m^{1/3}}, \frac{\sigma_{\pi^\star} Z}{n^{1/4}}, \frac{C_{\max} Z}{n^{1/2}}\right\}\right)$.*

Corollary 3 has several parts. Both the stopping rate and regret bounds are a maximum over four terms. The first two terms are variance-driven and cost-driven contributions from the labeled sample, and the last two are their unlabeled-sample counterparts. In Section C.4, we show that the cost-driven labeled-data portion of this upper bound is tight.

Switched regret vs. stopped regret. As previously hinted, the switched regret is driven to zero at a faster rate than the stopped regret in the low $\sigma_{\pi^\star}^2$ regime. This is perhaps unintuitive, as the switched regret bound provides a qualitatively strong guarantee: that the learner does not drive the system into an unrecoverable state before switching; on the other hand, the stopped regret guarantee only guarantees that the learner is competitive with the expert on the prefix for which it acts. Providing a resolution to this (for example, sharpening Corollary 2) is an interesting direction for future work. [7]

5 Extending to Misspecified Policy Classes

Previous sections assumed the expert $\pi^\star$ lies in the learner's policy class Π. In practice, experts often employ richer representations, breaking this realizability assumption. Here, we show our validator framework degrades gracefully under misspecification. Throughout this section, let Π be a finite class of deterministic policies, and let $\pi^\star$ be a deterministic expert. We do not assume $\pi^\star \in \Pi$.

[7] We suggest a partial resolution: define the *asymmetric stopped regret* as $J_N(\pi, \tau; c) - J_N(\pi^\star; c)$—physically, this compares the cost accumulated by the learner before handing over control to the cost the expert would have incurred driving the entire route. It is not hard to see that this is upper bounded by the switched regret, hence our algorithm controls the asymmetric stopped regret at the same rate as the switched regret (i.e., with the same rates as Corollary 3).

Misspecification benchmark. Let τ be as in (1). For any deterministic policy $\pi \in \Pi$, define $d_M(\pi) \coloneqq \Pr_{M,\pi^\star}\left[\tau_{\pi,\{\pi^\star\}} \le H\right]$ and $d_N(\pi) \coloneqq \Pr_{N,\pi^\star}\left[\tau_{\pi,\{\pi^\star\}} \le H\right]$. These are the probabilities that π deviates from the expert at any time during the trajectory (which has horizon H). We define the *policy-specific* and *best-in-class* misspecification, respectively, as $\Delta_\pi \coloneqq \max(d_M(\pi), d_N(\pi))$ and $\Delta \coloneqq \min_{\pi\in\Pi} \Delta_\pi$. Let $\tilde{\pi} = \arg\min_{\pi\in\Pi} \Delta_\pi$ be the policy which achieves the best-in-class misspecification. Finally, we define the empirical analogue of d_M as $\widehat{d}_M(\pi) \coloneqq \frac{1}{m}\sum_{i=1}^m \mathbf{1}\left[\tau_{\pi,\{\pi^\star\}}(S_i) \le H\right]$.

Agnostic version of Algorithm 1. When the expert lies outside the policy class ($\pi^\star \notin \Pi$), the exact-consistency version space used in Algorithm 1 may fail to contain the expert. To ensure the validator set remains well-defined and sparse even under misspecification, we transition from a constrained optimization (forcing zero training disagreement) to a symmetrically regularized game. In this agnostic framework, we allow validators to disagree with the base policy on the source data, but we penalize this disagreement at a rate Λ. This allows the learner to "trade off" a small amount of source-side abstention for a significant reduction in target-side late-stop risk. This shift is captured by the following equilibrium result, which can be thought of as an "agnostic" analog of Lemma 1:

Lemma 3. *Let $\pi_0 \in \arg\min_{\pi\in\Pi} \widehat{d}_M(\pi)$ be an empirical disagreement minimizer. For every penalty parameter $\Lambda > 0$ and integer $K \ge 1$, there exists a distribution $q^\star$ over validator sets $\Phi = (\phi_1, \ldots, \phi_K) \in \Pi^K$ that simultaneously satisfies the following two properties:*

- $\mathbb{E}_{\Phi\sim q^\star}\left[\sum_{k=1}^K \widehat{d}_M(\phi_k)\right] \le K \cdot \widehat{d}_M(\pi_0) + \frac{1}{\Lambda}$,
- $\mathbb{E}_{\Phi\sim q^\star}\left[\frac{1}{n}\sum_{j=1}^n \mathbf{1}\left[\tau_{\pi_0,\Phi}(T_j) > \tau_{\pi_0,\{\pi\}}(T_j)\right]\right] \le \Lambda\Big(\widehat{d}_M(\pi) - \widehat{d}_M(\pi_0)\Big) + \frac{1}{K}$.

Similar to Section 3, Lemma 3 is a non-constructive result, although it may be realized by no-regret play. We package the output Lemma 3 into an abstract subroutine, REGULARIZEDSPARSEVALIDATORDIST and leave the question of its efficient computational implementation to future work.

Theorem 3. *Let $\delta > 0$. Let $\pi_0 \in \arg\min_{\pi\in\Pi} \widehat{d}_M(\pi)$. Fix a penalty $\Lambda > 0$ and committee size $K \ge 1$, and let $q^\star \in \Delta(\Pi^K)$ be the equilibrium distribution from Lemma 3. Define the complexity measure $Z \coloneqq (K+1)\log|\Pi| + \log\frac{4}{\delta}$. Then, with probability at least $1-\delta$, the randomized selective policy $(\pi_0, \tau_{\pi_0,\Phi})$ with validator set drawn $\Phi \sim q^\star$ satisfies*

- $\mathbb{E}_{\Phi\sim q^\star}\left[\alpha_M(\pi_0, \tau_{\pi_0,\Phi})\right] \;\le\; (K+1)\Delta + \frac{1}{\Lambda} + 2(K+1)\sqrt{\frac{Z}{2m}}$,
- $\mathbb{E}_{\Phi\sim q^\star}\left[\mathrm{Regret}_N(\pi_0, \tau_{\pi_0,\Phi}; c)\right] \;\le\; C_{\max}\left(\Delta + \Lambda\Delta + \frac{1}{K} + 2\Lambda\sqrt{\frac{Z}{2m}} + \sqrt{\frac{Z}{2n}}\right)$.

Corollary 4. *There exists a choice of parameters $K = \Lambda$ (depending on $m, n, \Delta, |\Pi|$, and δ) such that, with probability at least $1-\delta$, both $\mathbb{E}_{\Phi\sim q^\star}[\alpha_M(\pi_0, \tau_{\pi_0,\Phi})]$ and[8] $\mathbb{E}_{\Phi\sim q^\star}[\mathrm{Regret}_N(\pi_0, \tau_{\pi_0,\Phi}; c)]/C_{\max}$ are $\widetilde{O}\left(\Delta^{1/2} + (\log|\Pi|/m)^{1/5} + (\log|\Pi|/n)^{1/3}\right)$.*

Extension to off-policy test trajectories. While we have assumed a single expert across both environments, this regularized framework naturally extends to settings where the training demonstrator π^{tr} and test demonstrator π^{te} differ, and neither necessarily lies in Π. By generalizing the disagreement probabilities, we show that our algorithm degrades gracefully with the irreducible off-policy error Δ_{off}. We defer the formal setup and the proof of this guarantee to Section D.4.

6 Proof of Concept

We evaluate SeqRejectron in the LunarLander-v3 environment [Towers et al., 2024], with the goal of empirically characterizing the completeness-soundness tradeoff: the algorithm should stop minimally on train while reliably detecting out-of-distribution dynamics and stopping at test time.

Experimental setup. The step cost c_h is a quadratic penalty on distance to the landing pad, velocity and angular velocity with a cap at 1. If the lander ever crashes, it incurs the maximum step-wise cost for the remainder of the episode. Costs are then normalized to a maximum episode cost of 1.

[8] In contrast to Algorithm 1, the regret guarantee of the misspecified SeqRejectron is stated in expectation over the draw $\Phi \sim q^*$, and thus the algorithm must return the validator distribution q^*. This is because in the agnostic setting, taking the union of k validator sets (which we did to boost the success probability of Algorithm 1) would linearly amplify the irreducible misspecification error (scaling the penalty to $O(k\Delta)$).

For the dynamics shift, the source domain M is windless and uses a reduced initial random impulse, while the target domain N applies a rightward wind force. Thus the environment shift consists of the addition of wind and the stronger initial impulse. The expert $\pi^\star$ is a stochastic state-feedback controller, parameterized by a neural network on the raw LunarLander observation features, tuned to land in the target environment, but not necessarily optimal under the custom cost. The base learner π_0 is fit by MLE on labeled demonstrations $\mathcal{D}_{\mathsf{train}}$ drawn only from M.

To ensure tractability, we replace the game-theoretic construction of Section B.6 with an adversarial posterior-sampling heuristic. First, using the labeled training demonstrations $\mathcal{D}_{\mathsf{train}}$, we form a Bayesian posterior over the policy parameters and sample a pool of candidates approximately consistent with $\pi^\star$ on M. Next, we evaluate this pool on the unlabeled test trajectories $\mathcal{D}_{\mathsf{test}}$ and greedily select the candidate that maximizes the cumulative squared Hellinger disagreement with the base learner π_0 (since our goal is to challenge our base learner on the test data). Repeating this selection $K = 3$ times yields a validator set Φ, which defines a stopping rule as in (2).

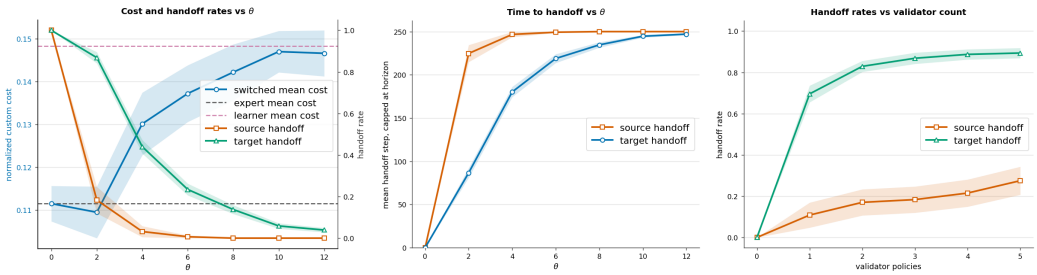

Figure 1: Left: normalized target switched cost and source/target handoff rates as functions of the safety threshold θ. Dashed horizontal lines show the expert and learner mean costs. Middle: mean handoff time as a function of θ, with trajectories that never hand off capped at the episode horizon. Right: source and target handoff rates as a function of the number of validator policies, with $\theta = 2$ fixed. In all plots, shaded bands indicate standard error across trials.

Results. We evaluate the framework over 20 independent trials. In each trial, we collect $m = 30$ labeled training trajectories and $n = 30$ unlabeled test trajectories over a range of θ threshold values. Figure 1 summarizes the resulting tradeoffs. In the left panel, the source handoff rate collapses steeply to near 0%, confirming that the validators rarely flag in-distribution trajectories. The target handoff rate decays far more gradually with the trajectory cost mirroring its behavior. As handoff becomes less frequent, the switched policy's cost approaches the learner's cost. The dashed black line represents the expert's average cost on the test environment. The red dashed line represents the expected cost of π_0, and is the cost that would have been obtained by running vanilla behavior cloning, without any abstention mechanism. This shows that SeqRejectron (with a suitable choice of $\theta = 2$) is able to reliably detect and abstain when it encounters environment shift.

The middle panel shows that larger θ delays handoff, as expected, since the validator allows more cumulative disagreement before switching. Finally, the right panel shows that increasing the number of validators raises handoff rates, particularly in the target environment, by making the stopping rule more sensitive to disagreement. This effect quickly saturates, suggesting that SeqRejectron is somewhat robust to the choice of validator set size.

7 Conclusion

We introduced selective imitation learning, a framework where a learner abstains mid-trajectory when training data is uninformative about the test environment. Our algorithm, SeqRejectron, yields horizon-free sample complexity for deterministic classes under sparse costs, with analogous guarantees for stochastic and misspecified classes. We also established an $\Omega(1/\epsilon^3)$ lower bound that matches the labeled-sample cost-driven exponent. Several directions remain open. Computationally, developing practical surrogates for the cumulative Hellinger objective in stochastic settings is needed. Statistically, closing the gap between the variance-driven $\tilde{O}(\epsilon^{-5})$ rate and our lower bound remains open. Finally, incorporating partial structural knowledge about test dynamics (e.g., bounded TV distance, density ratio, or parametric shift) could tighten test-side stopping rates. Ultimately, this work establishes abstention as a principled response to unanticipated dynamics shift.

References

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Aaron D Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *2019 18th European control conference (ECC)*, pages 3420–3431. Ieee, 2019.

OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J Tomlin. Hamilton-jacobi reachability: A brief overview and recent advances. In *2017 IEEE 56th annual conference on decision and control (CDC)*, pages 2242–2253. IEEE, 2017.

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pages 139–153. Springer, 2012.

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):151–175, 2010.

Jongseong Chae, Seungyul Han, Whiyoung Jung, Myungsik Cho, Sungho Choi, and Youngchul Sung. Robust imitation learning against variations in environment dynamics. In *International Conference on Machine Learning*, pages 2828–2852. PMLR, 2022.

Gautam Chandrasekaran, Adam R Klivans, Vasilis Kontonis, Konstantinos Stavropoulos, and Arsen Vasilyan. Efficient discrepancy testing for learning with distribution shift. *NeurIPS*, 2024.

Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 international conference on robotics and automation (ICRA)*, pages 8973–8979. IEEE, 2019.

Xiaoyu Chen, Jiachen Hu, Chi Jin, Lihong Li, and Liwei Wang. Understanding domain randomization for sim-to-real transfer. *arXiv preprint arXiv:2110.03239*, 2021.

C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 2003.

Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9329–9338, 2019.

Siddharth Desai, Ishan Durugkar, Haresh Karnan, Garrett Warnell, Josiah Hanna, and Peter Stone. An imitation from observation approach to transfer learning with dynamics mismatch. *Advances in Neural Information Processing Systems*, 33:3917–3929, 2020.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Anushri Dixit, Lars Lindemann, Skylar X Wei, Matthew Cleaveland, George J Pappas, and Joel W Burdick. Adaptive conformal prediction for motion planning among dynamic agents. In *Learning for Dynamics and Control Conference*, pages 300–314. PMLR, 2023.

Miroslav Dudík, Nika Haghtalab, Haipeng Luo, Robert E Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. Oracle-efficient online learning and auction design. *Journal of the ACM (JACM)*, 67(5):1–57, 2020.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Nicolas Espinosa-Dice, Sanjiban Choudhury, Wen Sun, and Gokul Swamy. Efficient imitation under misspecification. *arXiv preprint arXiv:2503.13162*, 2025.

Benjamin Eysenbach, Swapnil Asawa, Shreyas Chaudhari, Sergey Levine, and Ruslan Salakhutdinov. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. *arXiv preprint arXiv:2006.13916*, 2020.

Arnaud Fickinger, Abderrahim Bendahi, and Stuart Russell. Statistical guarantees for offline domain randomization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

Dylan J Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *Advances in Neural Information Processing Systems*, 37:120602–120666, 2024.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Tesshu Fujinami, Bruce D Lee, Nikolai Matni, and George J Pappas. Domain randomization is sample efficient for linear quadratic control. *arXiv preprint arXiv:2502.12310*, 2025.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Tanmay Gangwani and Jian Peng. State-only imitation with transition dynamics mismatch. *arXiv preprint arXiv:2002.11879*, 2020.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Surbhi Goel, Abhishek Shetty, Konstantinos Stavropoulos, and Arsen Vasilyan. Tolerant algorithms for learning with arbitrary covariate shift. *Advances in Neural Information Processing Systems*, 37: 124979–125018, 2024.

Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, and Omar Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. *Advances in Neural Information Processing Systems*, 33:15859–15870, 2020.

Tom Haider, Karsten Roscher, Felippe Schmoeller da Roza, and Stephan Günnemann. Out-of-distribution detection for reinforcement learning agents with probabilistic dynamics models. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 851–859, 2023.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, and Konstantinos Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12627–12637, 2019.

Gregory Kahn, Adam Villaflor, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. Uncertainty-aware reinforcement learning for collision avoidance. *arXiv preprint arXiv:1702.01182*, 2017.

Adam Tauman Kalai and Varun Kanade. Efficient learning with arbitrary covariate shift. In *Algorithmic Learning Theory*, pages 850–864. PMLR, 2021.

Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. *37th Annual Conference on Learning Theory, COLT 2024 (to appear)*, 2024a.

Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Learning intersections of halfspaces with distribution shift: Improved algorithms and sq lower bounds. *37th Annual Conference on Learning Theory, COLT 2024 (to appear)*, 2024b.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. *arXiv preprint arXiv:1912.05032*, 2019.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. *arXiv preprint arXiv:2107.04034*, 2021.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Niklas Lauffer, Xiang Deng, Srivatsa Kundurthy, Brad Kenstler, and Jeff Da. Imitation learning for multi-turn lm agents via on-policy expert corrections. *arXiv preprint arXiv:2512.14895*, 2025.

Brian Lee and Nikolai Matni. Single trajectory conformal prediction. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 3019–3024. IEEE, 2024.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 8(8):5116–5123, 2023a.

Lars Lindemann, Xin Qin, Jyotirmoy V Deshmukh, and George J Pappas. Conformal prediction for stl runtime verification. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, pages 142–153, 2023b.

Björn Lütjens, Michael Everett, and Jonathan P How. Safe reinforcement learning with model uncertainty estimates. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8662–8668. IEEE, 2019.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

Bhairav Mehta, Manfred Diaz, Florian Golemo, Christopher J Pal, and Liam Paull. Active domain randomization. In *Conference on Robot Learning*, pages 1162–1176. PMLR, 2020.

Natinael Solomon Neggatu, Jeremie Houssineau, and Giovanni Montana. Evaluation-time policy switching for offline reinforcement learning. *arXiv preprint arXiv:2503.12222*, 2025.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR, 2022.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International conference on machine learning*, pages 2817–2826. PMLR, 2017.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

Stephen Prajna and Ali Jadbabaie. Safety verification of hybrid systems using barrier certificates. In *International workshop on hybrid systems: Computation and control*, pages 477–492. Springer, 2004.

Mohit Prashant, Arvind Easwaran, Suman Das, and Michael Yuhas. Guaranteeing out-of-distribution detection in deep rl via transition estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Nived Rajaraman, Lin Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33:2914–2924, 2020.

Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. Epopt: Learning robust neural network policies using model ensembles. *arXiv preprint arXiv:1610.01283*, 2016.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Sahana Rayan and Ambuj Tewari. Learning to partially defer for sequences, 2025. URL `https://arxiv.org/abs/2502.01459`.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

Andrei A Rusu, Matej Večerík, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. In *Conference on robot learning*, pages 262–270. PMLR, 2017.

Fereshteh Sadeghi and Sergey Levine. Cad2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*, 2016.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pages 10022–10032. PMLR, 2021.

Gokul Swamy, Nived Rajaraman, Matt Peng, Sanjiban Choudhury, J Bagnell, Steven Z Wu, Jiantao Jiao, and Kannan Ramchandran. Minimax optimal online imitation learning via replay estimation. *Advances in Neural Information Processing Systems*, 35:7077–7088, 2022.

Jie Tan, Tingnan Zhang, Erwin Coumans, Atil Iscen, Yunfei Bai, Danijar Hafner, Steven Bohez, and Vincent Vanhoucke. Sim-to-real: Learning agile locomotion for quadruped robots. *arXiv preprint arXiv:1804.10332*, 2018.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pages 6215–6224. PMLR, 2019.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

Claire J Tomlin, John Lygeros, and S Shankar Sastry. A game theoretic approach to controller design for hybrid systems. *Proceedings of the IEEE*, 88(7):949–970, 2000.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Adrian Weller, and Volkan Cevher. Robust inverse reinforcement learning under transition dynamics mismatch. *Advances in Neural Information Processing Systems*, 34:25917–25931, 2021.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2174–2182, 2017.

Wenhao Yu, Jie Tan, C Karen Liu, and Greg Turk. Preparing for the unknown: Learning a universal policy with online system identification. *arXiv preprint arXiv:1702.02453*, 2017.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in neural information processing systems*, 33:21024–21037, 2020.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Contents

A Additional Related Work

A.1 Background on imitation learning and reinforcement learning

Reinforcement learning. The standard framework for sequential decision-making is the Markov decision process (MDP) [Sutton et al., 1998], where an agent maximizes cumulative reward by interacting with an environment governed by unknown transition dynamics. In the offline setting, where the agent must learn from a fixed dataset without further environment interaction, pessimism-based approaches [Rashidinejad et al., 2021, Kumar et al., 2020, Levine et al., 2020] provide guarantees when the dataset covers the target policy. Our work operates in a similarly offline regime, but with a key difference: we observe datasets from *two* environments and provide PAC-style guarantees without any online interaction with the test environment.

Imitation learning. The dominant paradigm for learning from demonstrations is *behavior cloning* (BC), which reduces imitation learning to supervised prediction [Pomerleau, 1988, Ross and Bagnell, 2010]. A central limitation of BC is *compounding error*: small per-step mistakes cause the learner to drift to states outside the expert's distribution, where further errors compound. Interactive methods address this by querying the expert on the learner's own rollouts: DAgger [Ross et al., 2011] reduces regret to $O(\epsilon H)$ by collecting on-policy corrections, and subsequent work gave game-theoretic formulations [Swamy et al., 2021] and adversarial approaches via occupancy-measure matching [Ho and Ermon, 2016, Ziebart et al., 2008]. Imitation from observation [Torabi et al., 2018] studies the special case where only state sequences (not actions) are available, which is similar to our assumption of unlabeled test trajectories, but does not provide theoretical guarantees.

Sample complexity of imitation learning. A precise theoretical understanding of BC's sample complexity has emerged over the past several years. Rajaraman et al. [2020] established fundamental limits: offline BC requires $\Omega(H^2/\epsilon)$ trajectories in the worst case, while online methods can achieve $O(H/\epsilon)$, and Swamy et al. [2022] showed that replay estimation achieves minimax-optimal online IL by reducing empirical variance in the expert visitation distribution. This apparent gap between offline and online IL was recently revisited by Foster et al. [2024], who showed that under weight-sharing, log-loss BC achieves *horizon-independent* sample complexity via a trajectory-level Hellinger analysis, whenever cumulative costs are controlled and the policy class shares parameters across time steps; this matches the minimax rate even among interactive algorithms. Our work builds directly on this Hellinger-based analysis, extending it to the environment-shift setting where test dynamics may differ arbitrarily from training, and replacing the standard regret objective with a stopped regret that accounts for the possibility of abstention.

A.2 Background on selective classification and learning with distribution shift

Learning under distribution shift. The classical approach to distribution shift is *covariate shift adaptation*, where the marginal input distribution changes but the labeling rule is preserved. Shimodaira [2000] introduced importance-weighting as a principled correction, and Sugiyama et al. [2007] gave an importance-weighted cross-validation procedure for model selection under this shift. On the theoretical side, Ben-David et al. [2010] and Mansour et al. [2009] bound target error in terms of source error and a divergence between domains. However, Ben-David and Urner [2012] establish that without further assumptions on the relationship between source and target—such as bounded density ratios or overlapping supports—successful domain adaptation is impossible in general. These hardness results make clear that any distribution-free guarantee under arbitrary shift must either make structural assumptions or give up on committing to an action everywhere.

Selective classification and PQ learning. *Selective classification* [Chow, 2003, El-Yaniv et al., 2010, Geifman and El-Yaniv, 2017] offers a way out: by allowing a classifier to abstain on inputs it cannot reliably label, one can trade coverage for accuracy. Recent work on partial deferral for sequence prediction studies one-time handoff to an expert after a chosen point in the sequence, which is closely related in spirit to our expert-handoff view of abstention [Rayan and Tewari, 2025]; however, they do not consider environment shift, which is the central focus of this work. Closest to our formal setup, Goldwasser et al. [2020] introduced the *PQ learning* framework, showing that abstention is precisely the mechanism that circumvents the hardness of arbitrary distribution shift. Given labeled samples from P and unlabeled samples from Q, PQ learning asks for a selective classifier with low error on Q and low rejection rate on P—with no assumptions on the relationship between P and Q. Their algorithm, Rejectron, iteratively builds validator classifiers that agree on P but

disagree on Q, and abstains wherever such disagreement exists. Subsequent work [Kalai and Kanade, 2021] showed that the computational complexity lies in between PAC learning and Agnostic PAC learning. Another closely related model is the Testable Distribution Shift (TDS) Klivans et al. [2024a] which allows for rejecting the entire test distribution instead of selectively predicting. There has been a flurry of recent work designing computationally efficient algorithms for these models under assumptions only on the train distribution [Klivans et al., 2024b, Chandrasekaran et al., 2024, Goel et al., 2024]. Our formulation can be viewed as a sequential analogue of PQ learning (Section 2), and the validator-based construction of SeqRejectron can be thought of as a generalization of Rejectron to trajectory-level certification.

A.3 Robustness, uncertainty quantification, and domain adaptation in RL

Sim-to-real transfer and domain randomization. The *reality gap*—the mismatch between simulated and real-world appearance, physics, and dynamics—is a central obstacle to deploying learned policies on physical systems [Tobin et al., 2017].

The dominant mitigation is *domain randomization*: training over a broad distribution of simulated environments so the real world falls within the training support. This has been applied to visual parameters [Tobin et al., 2017, Sadeghi and Levine, 2016, James et al., 2019], physical parameters such as friction, mass, and actuator delays [Peng et al., 2018, Tan et al., 2018], and scaled to dexterous manipulation and Rubik's-cube solving [Andrychowicz et al., 2020, Akkaya et al., 2019]. Complementing these empirical milestones, recent literature has formalized the theoretical foundations of this approach. Specifically, these works have bounded the sim-to-real gap for uniform domain randomization [Chen et al., 2021], established the method's optimal asymptotic sample efficiency [Fujinami et al., 2025] for linear systems, and proved the statistical consistency of data-driven, offline domain randomization [Fickinger et al., 2026]

A complementary line of work closes the gap *adaptively*: Chebotar et al. [2019] iteratively refine the randomization distribution using real rollouts; Yu et al. [2017] and Kumar et al. [2021] condition policies on online system-identification modules that infer dynamics at deployment; Rusu et al. [2017] use progressive networks for visual transfer; and Mehta et al. [2020] steer randomization toward maximally informative environments. Codevilla et al. [2019] document the resulting failure modes of behavior cloning under sim-to-real shift, motivating principled responses to out-of-distribution conditions.

All of these methods attempt to *bridge* the reality gap. Our work is complementary: rather than correcting for an unknown shift, we learn to *detect* when dynamics have changed too much for reliable execution and abstain, with distribution-free PAC guarantees that require no assumptions on the magnitude or structure of the mismatch.

Robust reinforcement learning. Robust RL seeks policies that perform well under worst-case dynamics drawn from a prespecified uncertainty set. The foundational framework is the robust MDP, in which nature adversarially selects transitions within a rectangular uncertainty set; Iyengar [2005] and Nilim and El Ghaoui [2005] independently showed that such problems admit tractable dynamic-programming solutions, and Wiesemann et al. [2013] extended this to more general uncertainty sets and ambiguity models. In the deep RL setting, Pinto et al. [2017] train a protagonist against an adversary that applies destabilizing forces (RARL), and Rajeswaran et al. [2016] optimize over an ensemble of source models (EPOpt) to guard against model error. Zhang et al. [2020] study robustness to observation and action perturbations in deep RL, and Tessler et al. [2019] isolate the action-perturbation setting. On the theoretical side, Panaganti and Kalathil [2022] and Shi and Chi [2024] establish sample complexity bounds for learning robust policies from offline data under $(\mathbf{s}, \mathbf{a})$-rectangular and distributionally robust uncertainty sets, respectively. All of these methods require the learner to commit ahead of time to an uncertainty set or dynamics prior—implicitly assuming knowledge of the shift's structure or magnitude.

Imitation learning under dynamics mismatch. Classical IL theory assumes that the learner and expert share the same transition dynamics. A growing body of work relaxes this assumption.

Several approaches use adversarial methods to match occupancy measures across environments: Ho and Ermon [2016] introduce GAIL, which frames IL as a distribution-matching problem and provides a foundation for dynamics-robust extensions; Fu et al. [2017] show that adversarial IRL (AIRL) can recover reward functions that transfer under moderate dynamics shift; and Kostrikov et al.

[2019] propose a distribution-correction method (DAC) that compensates for off-policy mismatch via density-ratio reweighting.

Other works directly target the setting where expert and learner operate under different dynamics. Gangwani and Peng [2020] and Desai et al. [2020] use adversarial state-distribution matching and transition-dynamics alignment respectively. Viano et al. [2021] bound performance degradation as a function of the ℓ_1 distance between transition kernels, and Chae et al. [2022] train policies that are robust across a family of perturbed MDPs (RIME). Eysenbach et al. [2020] compensate for dynamics mismatch by modifying the reward with a classifier that distinguishes source from target transitions, though this requires online RL access to the target domain. Lauffer et al. [2025] address the covariate shift that arises when a learned policy diverges into unfamiliar dynamics, using on-policy expert corrections for multi-turn LM agents. Finally, Espinosa-Dice et al. [2025] study the misspecification setting in which the expert falls outside the learner's policy class.

All of these methods attempt to *correct for* or *tolerate* dynamics mismatch via reward modification, distribution matching, or robust optimization, and most require either interactive access to the target environment or assumptions on the magnitude of the shift. Our work is complementary: we make no assumptions on the dynamics and instead learn, from offline data alone, to *detect* when the shift is too large to act reliably and abstain, with distribution-free PAC guarantees on both the stopping rate and the stopped regret.

Control-theoretic safety and reachability. Classical approaches to safe deployment ground safety guarantees in known system dynamics. Hamilton-Jacobi reachability [Tomlin et al., 2000, Bansal et al., 2017] computes the exact set of states from which a safety constraint can be guaranteed, and control barrier functions (CBFs) [Ames et al., 2019] enforce forward invariance of a safe set via Lyapunov-style certificates. Prajna and Jadbabaie [Prajna and Jadbabaie, 2004] give analogous barrier certificates for hybrid systems. These methods offer strong formal guarantees, but require an accurate model of the transition dynamics.

Uncertainty-aware deployment. A natural response to distribution shift is to monitor epistemic uncertainty at deployment and abstain or intervene when it is high; MC dropout [Gal and Ghahramani, 2016] and deep ensembles [Lakshminarayanan et al., 2017] are the dominant practical tools for obtaining such estimates. These estimates have been deployed directly in safety-critical settings: Lütjens et al. [2019] use MC dropout and bootstrapped ensembles to modulate collision-avoidance policies when epistemic uncertainty is high, and Kahn et al. [2017] propagate learned uncertainty estimates through a model-based planner for robot navigation. However, the thresholds used to trigger abstention in these methods are typically heuristic. Neggatu et al. [2025] use uncertainty estimates to switch between an offline RL policy and a more conservative behavior-cloning policy, but do not prove theoretical guarantees, unlike ours. Several works [Haider et al., 2023, Prashant et al., 2025] frame out-of-distribution shift in RL as changes in the transition dynamics, but the proposed methods do not rigorously bound the downstream regret.

Conformal prediction (see Angelopoulos and Bates [2023] for a foundational treatment) offers a distribution-free alternative with rigorous finite-sample coverage guarantees. Tibshirani et al. [2019] extend the framework to covariate shift via importance weighting; and Gibbs and Candes [2021] develop online conformal methods that adapt to distribution shift as it occurs. Recent work has extended these ideas to sequential and dynamical settings. Lindemann et al. [2023a] use conformal prediction to construct probabilistic prediction regions for safe motion planning in dynamic environments; Dixit et al. [2023] apply adaptive conformal prediction to quantify multi-step-ahead uncertainty for MPC among dynamic agents; Lindemann et al. [2023b] give conformal runtime verification guarantees for signal temporal logic specifications; and Lee and Matni [2024] establish conformal coverage guarantees from a single trajectory of temporally correlated data generated by an unknown stochastic dynamical system.

All of these methods share our high-level goal of detecting when a deployed agent has ventured outside its training support and should stop. Ensemble and conformal methods differ from ours in a fundamental way, however: ensemble approaches rely on heuristic variance thresholds to trigger abstention, while conformal methods provide rigorous marginal coverage for *prediction sets* at each step. Neither directly bounds the regret accumulated along a trajectory before abstention. Our approach avoids both heuristics and prediction sets, instead providing frequentist PAC-style guarantees directly on the stopping time that bound the stopped regret in the test environment.

B Additional Results and Proofs from the Deterministic Section

B.1 Supporting technical lemmas

Lemma 4 (Prefix coupling for deterministic policies)**.** *Let π and $\pi^\star$ be deterministic policies on a horizon-H MDP with pre-action filtration $\mathcal{G}_h = \sigma(s_1, a_1, \dots, s_{h-1}, a_{h-1}, s_h)$. Define the first-deviation time*

$$\tau_{\pi,\{\pi^\star\}} \coloneqq \min\big\{h \in [H] : \pi_h(s_h) \neq \pi^\star_h(s_h)\big\},$$

with the convention $\min \emptyset = H + 1$. Let τ be any stopping time adapted to $\{\mathcal{G}_h\}$, and define $h^\star \coloneqq \min(\tau, \tau_{\pi,\{\pi^\star\}})$, and define the stopped sigma-algebra

$$\mathcal{G}_{h^\star} \coloneqq \big\{A \subseteq \Omega : A \cap \{h^\star = h\} \in \mathcal{G}_h \text{ for every } h \in [H+1]\big\}.$$

Intuitively, $\mathcal{G}_{h^\star}$ represents the information available up until the (random) time $h^\star$. Then, for any environment M, every event $\mathcal{E} \in \mathcal{G}_{h^\star}$ satisfies $\Pr_{M,\pi}[\mathcal{E}] = \Pr_{M,\pi^\star}[\mathcal{E}]$. Consequently, for every bounded $\mathcal{G}_{h^\star}$-measurable random variable Z, we also have $\mathbb{E}_{M,\pi}[Z] = \mathbb{E}_{M,\pi^\star}[Z]$.

Proof. We first check that $h^\star$ is a stopping time. For each $k \in [H]$, the event $\{\tau_{\pi,\{\pi^\star\}} \leq k\} = \bigcup_{j=1}^{k} \{\pi_j(s_j) \neq \pi^\star_j(s_j)\}$ belongs to $\mathcal{G}_k$ because each s_j is $\mathcal{G}_j \subseteq \mathcal{G}_k$-measurable and $\pi_j, \pi^\star_j$ are deterministic; for $k = H + 1$ the claim is trivial. Hence $\tau_{\pi,\{\pi^\star\}}$ is a stopping time, and so is $h^\star = \min(\tau, \tau_{\pi,\{\pi^\star\}})$.

It remains to show that π and $\pi^\star$ induce the same distribution on $\mathcal{G}_{h^\star}$-measurable events. Fix $k \in [H]$ and any pre-action prefix $\mathbf{t}_k = (s_1, a_1, \dots, s_{k-1}, a_{k-1}, s_k)$ with $h^\star(\mathbf{t}_k) = k$. On the event $\{h^\star = k\}$ we have $k \leq \tau_{\pi,\{\pi^\star\}}$, so $\pi_j(s_j) = \pi^\star_j(s_j)$ for all $j < k$, and therefore

$$\begin{aligned}\Pr_{M,\pi}[(s_1, a_1, \dots, s_{k-1}, a_{k-1}, s_k) = \mathbf{t}_k,\, h^\star = k] &= P_0(s_1) \prod_{j=1}^{k-1} P_j(s_{j+1} \mid s_j, \pi_j(s_j)) \\ &= \Pr_{M,\pi^\star}[(s_1, a_1, \dots, s_{k-1}, a_{k-1}, s_k) = \mathbf{t}_k,\, h^\star = k].\end{aligned}$$

When $k = H + 1$, the same argument applied to full trajectories shows that the probabilities also agree on the slice $\{h^\star = H + 1\}$, since $\pi_j(s_j) = \pi^\star_j(s_j)$ for every $j \in [H]$ there. For any $\mathcal{E} \in \mathcal{G}_{h^\star}$, the slice $\mathcal{E} \cap \{h^\star = k\}$ is $\mathcal{G}_k$-measurable, so its probability is a sum of prefix probabilities of the form above; summing over k gives $\Pr_{M,\pi}[\mathcal{E}] = \Pr_{M,\pi^\star}[\mathcal{E}]$. Because the two laws are equal on $\mathcal{G}_{h^\star}$, their expectations agree for every bounded $\mathcal{G}_{h^\star}$-measurable random variable. □

Lemma 5 (General statement of Lemma 1)**.** *Let M be an positive integer and let $\mathcal{T} \subseteq [M]^n$ be any set of vectors. For any target $\epsilon \in (0, 1)$, set $K = \lceil 1/\epsilon \rceil$. There exists a probability distribution $q^\star$ over subsets $S \subseteq \mathcal{T}$ of size at most K with the following two properties.*

1. ***(Coverage)** For every $\tau \in \mathcal{T}$:*

$$\mathbb{E}_{S \sim q^\star}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[\min_{v \in S} v_i > \tau_i\right]\right] < \epsilon.$$

2. ***(Sparsity)** The support of $q^\star$ consists entirely of sets of size at most $\lceil 1/\epsilon \rceil$.*

Proof. Define a finite zero-sum game in which the maximizer chooses $p \in \Delta(\mathcal{T})$ and the minimizer chooses $q \in \Delta(\mathcal{T}^K)$, with payoff

$$\phi(p, q) := \mathbb{E}_{\tau \sim p}\, \mathbb{E}_{S \sim q}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[\min_{v \in S} v_i > \tau_i\right]\right].$$

Since ϕ is bilinear in (p, q), von Neumann's minimax theorem gives

$$\min_q \max_p \ \phi(p, q) = \max_p \min_q \ \phi(p, q).$$

To bound the maximin, fix any $p \in \Delta(\mathcal{T})$ and let p^K denote the strategy of forming $S = \{v_1, \ldots, v_K\}$ by drawing $v_1, \ldots, v_K$ i.i.d. from p. Then

$$\min_q \ \phi(p,q) \le \phi(p,p^K) = \frac{1}{n}\sum_{i=1}^{n} \Pr_{\tau,\, v_1,\ldots,v_K \sim p}\left(\min_{j\in[K]} (v_j)_i > \tau_i\right).$$

For a fixed coordinate i, the event $\{\min_j (v_j)_i > \tau_i\}$ holds if and only if τ_i is strictly the smallest value among the $K+1$ i.i.d. draws $\tau_i, (v_1)_i, \ldots, (v_K)_i$. Since these draws are exchangeable, each is equally likely to be the strict minimum, and since these $K+1$ events are mutually exclusive, each occurs with probability at most $1/(K+1)$. Therefore

$$\phi(p,p^K) \le \frac{1}{n}\sum_{i=1}^{n} \frac{1}{K+1} = \frac{1}{K+1}.$$

Since this holds for every $p \in \Delta(\mathcal{T})$, we conclude $\max_p \min_q \ \phi(p,q) \le 1/(K+1)$.

Combining with the minimax equality and using $K = \lceil 1/\epsilon \rceil$ so that $K+1 > 1/\epsilon$:

$$\min_q \max_p \ \phi(p,q) = \max_p \min_q \ \phi(p,q) \le \frac{1}{K+1} < \epsilon.$$

Hence there exists $q^\star \in \Delta(\mathcal{T}^K)$ with $\max_p \ \phi(p,q^\star) < \epsilon$. Taking $p = \delta_\tau$ for any fixed $\tau \in \mathcal{T}$ yields the coverage condition. The sparsity condition holds because every S in the support of $q^\star$ satisfies $|S| \le K = \lceil 1/\epsilon \rceil$. □

B.2 Analysis of stepwise reduction to PQ-learning

Assume $\mathcal{A} = \{0,1\}$ and $\pi^\star \in \Pi$ is deterministic. For each $h \in [H]$, let μ_h^M and μ_h^N denote the expert's step-h state marginals under M and N. We apply Rejectron [Goldwasser et al., 2020] at each step h with labeled examples $(s, \pi_h^\star(s))$ for $s \sim \mu_h^M$ and unlabeled examples from $\bar{\mu}_h := \frac{1}{2}\mu_h^M + \frac{1}{2}\mu_h^N$, yielding a predictor-selector pair $(\tilde{\pi}_h, g_h)$. Define $\tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_H)$, $g = (g_1, \ldots, g_H)$, and $\tau_g = \min\{h : g_h(s_h) = 0\}$. We run each step-$h$ learner with confidence δ/H and union-bound over h, so all guarantees below hold simultaneously with probability at least $1-\delta$.

For tolerance γ, the PQ guarantee gives for every $h \in [H]$:

$$\Pr_{s\sim\mu_h^M}[g_h(s) = 0] \le \gamma, \tag{3}$$

$$\Pr_{s\sim\bar{\mu}_h}[g_h(s) = 1,\ \tilde{\pi}_h(s) \ne \pi_h^\star(s)] \le \gamma. \tag{4}$$

Since $\bar{\mu}_h$ averages μ_h^M and μ_h^N, both marginals inherit a 2γ accepted-error bound from (4).

Abstention on M. Let E_h be the event that step h is the *first* time the learner either abstains or makes an unrejected mistake and let $A_h := \bigcap_{t=1}^{h-1}\Big\{g_t(s_t) = 1,\ \tilde{\pi}_t(s_t) = \pi_t^\star(s_t)\Big\}$. Thus A_h is the event that, before step h, the learner has neither abstained nor made an unrejected mistake. By definition,

$$E_h = A_h \cap \Big(\{g_h(s_h) = 0\} \cup \{g_h(s_h) = 1,\ \tilde{\pi}_h(s_h) \ne \pi_h^\star(s_h)\}\Big).$$

Since the event inside the intersection is $\mathcal{G}_h$-measurable, by Lemma 4, we may change measure from $\Pr_{M,\tilde{\pi}}$ to $\Pr_{M,\pi^\star}$:

$$\begin{aligned}
\Pr_{M,\tilde{\pi}}(E_h) &= \Pr_{M,\pi^\star}\Big(A_h \cap \Big(\{g_h(s_h) = 0\} \cup \{g_h(s_h) = 1,\ \tilde{\pi}_h(s_h) \ne \pi_h^\star(s_h)\}\Big)\Big) \\
&\le \Pr_{M,\pi^\star}\Big(\{g_h(s_h) = 0\} \cup \{g_h(s_h) = 1,\ \tilde{\pi}_h(s_h) \ne \pi_h^\star(s_h)\}\Big) \\
&\le \Pr_{M,\pi^\star}\big(g_h(s_h) = 0\big) + \Pr_{M,\pi^\star}\big(g_h(s_h) = 1,\ \tilde{\pi}_h(s_h) \ne \pi_h^\star(s_h)\big) \\
&= \Pr_{s\sim\mu_h^M}\big[g_h(s) = 0\big] + \Pr_{s\sim\mu_h^M}\big[g_h(s) = 1, \tilde{\pi}_h(s) \ne \pi_h^\star(s)\big] \le 3\gamma.
\end{aligned}$$

Since abstaining implies E_h occurs for some h, we have

$$\alpha_M(\tilde{\pi}, g) \le \sum_{h=1}^{H} \Pr(E_h) \le 3H\gamma.$$

Regret on N. Let F_h be the event that the first unrejected misstep occurs at step h prior to any abstention. Similar to the argument above, if $A_h := \bigcap_{t=1}^{h-1}\{g_t(s_t)=1,\ \tilde{\pi}_t(s_t)=\pi_t^\star(s_t)\}$, then

$$F_h = A_h \cap \{g_h(s_h)=1,\ \tilde{\pi}_h(s_h)\neq\pi_h^\star(s_h)\}.$$

We may first change measure from $\Pr_{N,\tilde{\pi}}$ to $\Pr_{N,\pi^\star}$ while retaining the factor A_h, and then drop A_h. Thus

$$\Pr_{N,\tilde{\pi}}(F_h) \le \Pr_{N,\pi^\star}\big(g_h(s_h)=1,\ \tilde{\pi}_h(s_h)\neq\pi_h^\star(s_h)\big) = \Pr_{s\sim\mu_h^N}\big[g_h(s)=1,\ \tilde{\pi}_h(s)\neq\pi_h^\star(s)\big] \le 2\gamma.$$

Since the learner incurs zero stopped regret unless an unrejected misstep occurs—which costs at most $C_{\max}$—we obtain

$$\mathrm{Regret}_N(\tilde{\pi}, g; c) \le C_{\max}\sum_{h=1}^{H}\Pr(F_h) \le 2C_{\max}H\gamma.$$

Sample complexity. Setting $\gamma = \Theta(\epsilon/(HC_{\max}))$ to achieve $\mathrm{Regret}_N \le \epsilon$, and using the PQ sample complexity of $\tilde{O}(\log|\Pi|/\gamma^2)$ per step, yields $m = n = \tilde{O}\left(\frac{H^2C_{\max}^2\log|\Pi|}{\epsilon^2}\right)$ trajectories (since one trajectory provides an example for each step). This H^2 penalty is unavoidable in this approach: the reduction certifies H isolated decisions rather than a single trajectory prefix, and the horizon factor enters through the per-step tolerance regardless of the cost structure.

B.3 Proof of Lemma 1

Let $M = H+1$. For each policy $\pi \in \Pi_{\mathsf{version}}$, let $v^\pi \in [M]^n$ be its empirical stopping-time vector on the n test trajectories, where the j-th coordinate is given by $(v^\pi)_j = \tau_{\pi_0,\{\pi\}}(T_j)$. Let $\mathcal{T} = \{v^\pi : \pi \in \Pi_{\mathsf{version}}\}$ be the finite set of all such vectors.

Applying Lemma 5 to the set $\mathcal{T}$ with target error $\epsilon = \rho$ implies the existence of a distribution $q^\star$ over subsets $\Phi \subseteq \Pi_{\mathsf{version}}$ satisfying:

- For all $\pi \in \Pi_{\mathsf{version}}$: $\mathbb{E}_{\Phi\sim q^\star}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}[\tau_{\pi_0,\Phi}(T_j) > \tau_{\pi_0,\{\pi\}}(T_j)]\right] < \rho$.
- For all $\Phi \in \mathsf{supp}(q^\star)$: $|\Phi| \le \lceil 1/\rho\rceil$.

This is exactly the $q^\star$ required by Lemma 1.

B.4 Proof of Theorem 1

We prove the theorem through four steps: constructing the validator distribution, bounding the empirical risk, generalizing to the population, and bounding the regret on N.

Throughout, let $K = \lceil\log_2(5/\delta)\rceil\lceil 2/\eta\rceil$ denote the maximum ensemble size and

$$Z \coloneqq (K+1)\log|\Pi| + \log\frac{5}{\delta}$$

the complexity term appearing in generalization bounds.

Step 1 (Existence of $q^\star$). By Proposition 1, Algorithm 2 yields a $(\eta/2+\xi)$-valid distribution $q^\star$ over subsets of Π_{version} such that every $\Phi' \in \mathsf{supp}(q^\star)$ satisfies $|\Phi'| \le \left\lceil\frac{2}{\eta}\right\rceil$ with probability $1-\delta/5$.

Step 2 (Bound the empirical risk). The output $\Phi = \bigcup_{i=1}^{k}\Phi_i$ is the union of $k = \lceil\log_2(5/\delta)\rceil$ independent draws from $q^\star$, so $|\Phi| \le k\cdot\lceil 2/\eta\rceil = K$ deterministically. By Markov's inequality applied to the $\eta/2+\xi$ coverage guarantee,

$$\Pr_{\Phi_i\sim q^\star}\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}[\tau_{\pi_0,\Phi_i}(T_j) > \tau_{\pi_0,\{\pi^\star\}}(T_j)] > \eta+2\xi\right)$$
$$\le \frac{\mathbb{E}_{\Phi_i\sim q^\star}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}[\tau_{\pi_0,\Phi_i}(T_j) > \tau_{\pi_0,\{\pi^\star\}}(T_j)]\right]}{\eta+2\xi} < \frac{\eta/2+\xi}{\eta+2\xi} = \frac{1}{2}.$$

We refer to the term on the left hand side as the *risk*. Since $\tau_{\pi_0,\Phi}(T_j) = \min_i \tau_{\pi_0,\Phi_i}(T_j)$, then if any Φ_i achieves risk $\leq \eta$, then so does Φ. Equivalently, Φ only has high risk if all k independent draws have high risk. Since these draws are independent,

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}[\tau_{\pi_0,\Phi}(T_j) > \tau_{\pi_0,\{\pi^\star\}}(T_j)] > \eta + 2\xi\right) \leq \prod_{i=1}^{k}\frac{1}{2} \leq \left(\frac{1}{2}\right)^{\lceil \log_2(5/\delta)\rceil} \leq \delta/5.$$

For the training data, all policies in $\Phi \cup \{\pi_0\}$ belong to Π_{version} by construction, meaning they agree with $\pi^\star$ on every observed state-action pair in $\mathcal{D}_{\mathsf{train}}$. Therefore, for all $i \in [m]$:

$$\tau_{\pi_0,\Phi}(S_i) = \tau_{\pi_0,\{\pi^\star\}}(S_i) = H + 1,$$

where the first equality holds because no policy in Φ ever disagrees with π_0 on any training trajectory (since both agree with $\pi^\star$ on all observed states), and the second holds because $\pi^\star$ and π_0 both belong to Π_{version} and hence agree on all observed actions. Since $\tau_{\pi_0,\Phi}(S_i) = H + 1 > H$ for all i, the empirical stopping rate $\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}[\tau_{\pi_0,\Phi}(S_i) \leq H] = 0$, and since $\tau_{\pi_0,\Phi}(S_i) = \tau_{\pi_0,\{\pi^\star\}}(S_i)$ for all i, the empirical risk $\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}[\tau_{\pi_0,\Phi}(S_i) > \tau_{\pi_0,\{\pi^\star\}}(S_i)] = 0$ as well.

Step 3 (Bound generalization error on the expert's distribution). We use uniform convergence over $\mathcal{H}_K = \{(\pi_0, \Phi) : \pi_0 \in \Pi,\ \Phi \subseteq \Pi, |\Phi| \leq K\}$, which satisfies $|\mathcal{H}_K| \leq |\Pi|^{K+1}$ as π_0 and each of the at most K elements of Φ are chosen from Π. We identify three failure events each occurring with probability at most $\delta/5$.

Test risk generalization. Let $Y(\pi_0, \Phi) = \mathbf{1}[\tau_{\pi_0,\Phi}(T) > \tau_{\pi_0,\{\pi^\star\}}(T)]$, so that $\hat{\mathbb{E}}[Y] = \frac{1}{n}\sum_{j=1}^{n} Y(T_j)$ is the empirical test risk over $\mathcal{D}_{\mathsf{test}}$ and $\mathbb{E}_N^{\pi^\star}[Y] = \Pr_{N,\pi^\star}[\tau_{\pi_0,\Phi}(T) > \tau_{\pi_0,\{\pi^\star\}}(T)]$ is the true test risk. Since $Y \in \{0,1\}$, $\widehat{\mathrm{Var}}(Y) \leq \hat{\mathbb{E}}[Y^2] = \hat{\mathbb{E}}[Y]$. By the empirical Bernstein inequality applied to each fixed hypothesis in $\mathcal{H}_K$, then union-bounded over all $|\mathcal{H}_K|$ hypotheses, with probability at least $1 - \delta/4$, every $(\pi_0, \Phi) \in \mathcal{H}_K$ simultaneously satisfies:

$$\mathbb{E}_N^{\pi^\star}[Y] \leq \hat{\mathbb{E}}[Y] + \sqrt{\frac{2\hat{\mathbb{E}}[Y]\log(5|\mathcal{H}_K|/\delta)}{n}} + \frac{3\log(5|\mathcal{H}_K|/\delta)}{n}.$$

Since $|\mathcal{H}_K| \leq |\Pi|^{K+1}$, we have $\log(5|\mathcal{H}_K|/\delta) \leq (K+1)\log|\Pi| + \log(5/\delta) = Z$. Since the empirical step guarantees $\hat{\mathbb{E}}[Y] \leq \eta + 2\xi$, this implies:

$$\mathbb{E}_N^{\pi^\star}[Y] \leq \eta + 2\xi + \sqrt{\frac{2(\eta+2\xi)Z}{n}} + \frac{3Z}{n}.$$

Train abstention generalization. The true stopping rate of (π_0, Φ) on expert trajectories is $\Pr_{M,\pi^\star}[\tau_{\pi_0,\Phi}(S) \leq H]$ and the empirical stopping rate over $\mathcal{D}_{\mathsf{train}}$ is $\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}[\tau_{\pi_0,\Phi}(S_i) \leq H]$. For any fixed $(\pi_0, \Phi) \in \mathcal{H}_K$ with $\Pr_{M,\pi^\star}[\tau_{\pi_0,\Phi}(S) \leq H] > \frac{Z}{m}$, each of the m i.i.d. draws $S_i \sim (M, \pi^\star)$ independently fails to trigger abstention with probability at most $1 - \frac{Z}{m}$, so the probability of observing zero empirical abstention is at most $(1 - \frac{Z}{m})^m \leq e^{-Z}$. A union bound over all $|\mathcal{H}_K| \leq |\Pi|^{K+1}$ hypotheses gives failure probability at most $|\mathcal{H}_K|e^{-Z} \leq \delta/5$.

Train risk generalization. The true train risk of $(\pi_0, \tau_{\pi_0,\Phi})$ is $\Pr_{M,\pi^\star}[\tau_{\pi_0,\Phi}(S) > \tau_{\pi_0,\{\pi^\star\}}(S)]$ and the empirical train risk is $\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}[\tau_{\pi_0,\Phi}(S_i) > \tau_{\pi_0,\{\pi^\star\}}(S_i)]$. Just like above, any hypothesis with $\Pr_{M,\pi^\star}[\tau_{\pi_0,\Phi}(S) > \tau_{\pi_0,\{\pi^\star\}}(S)] > \frac{Z}{m}$ achieves zero empirical train risk with probability at most $(1 - \frac{Z}{m})^m \leq e^{-Z}$, and a union bound over $\mathcal{H}_K$ gives failure probability at most $\delta/5$.

Taking a union bound over all failure events, with probability at least $1 - \delta$, uniform convergence holds simultaneously for all three failure events across all hypotheses in $\mathcal{H}_K$.

Step 4 (Bound stopping rate and regret on the learned policy's distribution). We first address the stopping rate on M under $(\pi_0, \tau_{\pi_0,\Phi})$. The generalization steps for the train abstention and train risk, respectively, give $\Pr_{M,\pi^\star}[\tau_{\pi_0,\Phi} \leq H] \leq \frac{Z}{m}$ and $\Pr_{M,\pi^\star}[\tau_{\pi_0,\Phi} > \tau_{\pi_0,\{\pi^\star\}}] \leq \frac{Z}{m}$. Define the safe event $B = \{\tau_{\pi_0,\Phi} \leq \tau_{\pi_0,\{\pi^\star\}}\}$, so $\Pr_{M,\pi^\star}[B^c] \leq \frac{Z}{m}$. Then,

$$\Pr_{M,\pi_0}[\tau_{\pi_0,\Phi} \leq H] \leq \Pr_{M,\pi_0}[\{\tau_{\pi_0,\Phi} \leq H\} \cap B] + \Pr_{M,\pi_0}[B^c].$$

To apply Lemma 4, let $h^\star = \min(\tau_{\pi_0,\Phi}, \tau_{\pi_0,\{\pi^\star\}})$. Both $\{\tau_{\pi_0,\Phi} \le H\} \cap B$ and B^c lie in $\mathcal{G}_{h^\star}$: for each $h \in [H]$, the slice $\{\tau_{\pi_0,\Phi} \le H\} \cap B \cap \{h^\star = h\}$ equals $\{\tau_{\pi_0,\Phi} = h\} \cap \{\tau_{\pi_0,\{\pi^\star\}} \ge h\}$, which is $\mathcal{G}_h$-measurable since $\tau_{\pi_0,\Phi}$ is a stopping time and $\{\tau_{\pi_0,\{\pi^\star\}} \ge h\}$ depends only on whether π_0 and $\pi^\star$ agree on $s_1, \dots, s_{h-1}$ (the case $h = H+1$ is trivial as the slice is the empty set); the same argument applies to B^c. Lemma 4 therefore gives

$$\Pr_{M,\pi_0}[\{\tau_{\pi_0,\Phi} \le H\} \cap B] = \Pr_{M,\pi^\star}[\{\tau_{\pi_0,\Phi} \le H\} \cap B] \le \Pr_{M,\pi^\star}[\tau_{\pi_0,\Phi} \le H] \le \frac{Z}{m},$$

and $\Pr_{M,\pi_0}[B^c] = \Pr_{M,\pi^\star}[B^c] \le \frac{Z}{m}$. Combining yields $\Pr_{M,\pi_0}[\tau_{\pi_0,\Phi} \le H] \le \frac{2Z}{m}$.

To bound the stopped regret on N, note that $\Pr_{N,\pi_0}[B^c] = \Pr_{N,\pi^\star}[B^c] \le \eta + 2\xi + \sqrt{\frac{2(\eta+2\xi)Z}{n}} + \frac{3Z}{n}$ by the test generalization guarantee and Lemma 4. Decomposing the expected stopped cost of $(\pi_0, \tau_{\pi_0,\Phi})$ over B and B^c,

$$\begin{aligned}
\mathrm{Regret}_N(\pi_0, \tau_{\pi_0,\Phi}; c) &= \mathbb{E}_N^{\pi_0}\left[\Big(\sum_{h=1}^{\tau_{\pi_0,\Phi}-1} c_h\Big)\mathbf{1}(B)\right] + \mathbb{E}_N^{\pi_0}\left[\Big(\sum_{h=1}^{\tau_{\pi_0,\Phi}-1} c_h\Big)\mathbf{1}(B^c)\right] - \mathbb{E}_N^{\pi^\star}\left[\sum_{h=1}^{\tau_{\pi_0,\Phi}-1} c_h\right] \\
&\le \mathbb{E}_N^{\pi^\star}\left[\Big(\sum_{h=1}^{\tau_{\pi_0,\Phi}-1} c_h\Big)\mathbf{1}(B)\right] + C_{\max} \cdot \Pr_{N,\pi_0}[B^c] - \mathbb{E}_N^{\pi^\star}\left[\sum_{h=1}^{\tau_{\pi_0,\Phi}-1} c_h\right] \\
&= -\mathbb{E}_N^{\pi^\star}\left[\Big(\sum_{h=1}^{\tau_{\pi_0,\Phi}-1} c_h\Big)\mathbf{1}(B^c)\right] + C_{\max} \cdot \Pr_{N,\pi_0}[B^c] \\
&\le C_{\max} \cdot \Pr_{N,\pi_0}[B^c] \qquad \text{(costs are nonnegative)} \\
&\le C_{\max}\left(\eta + 2\xi + \sqrt{\frac{2(\eta+2\xi)Z}{n}} + \frac{3Z}{n}\right),
\end{aligned}$$

where the first inequality applies Lemma 4 on B (because $\tau_{\pi_0,\Phi} \le \tau_{\pi_0,\{\pi^\star\}}$ on B, the selective policy's expected prefix cost $\sum_{h=1}^{\tau_{\pi_0,\Phi}-1} c_h$ matches the expert's expected cost over that prefix) and bounds the total cost on B^c by $C_{\max}$.

B.5 Transferring the completeness guarantee to the test environment

While our objective does not require bounding the test-side stopping rate $\alpha_N(\pi_0, \tau) \coloneqq \Pr_{N,\pi_0}[\tau \le H]$ distribution-free, it is natural to ask how this rate scales if the environment shift is bounded. Let $d^{\mathrm{state}}_{\pi^\star}(M, N)$ denote the total variation distance between the state-trajectory distributions induced by the expert in M and N.

Define the safe event $B = \{\tau \le \tau_{\pi_0,\{\pi^\star\}}\}$ and let $h^\star = \min(\tau, \tau_{\pi_0,\{\pi^\star\}})$. Both $\{\tau \le H\} \cap B$ and B^c lie in $\mathcal{G}_{h^\star}$ by the same measurability argument as in the proof of Theorem 1. By a union bound and two applications of Lemma 4,

$$\begin{aligned}
\alpha_N(\pi_0, \tau_{\pi_0,\Phi}) &\le \Pr_{N,\pi_0}[\{\tau_{\pi_0,\Phi} \le H\} \cap B] + \Pr_{N,\pi_0}[B^c] \\
&= \Pr_{N,\pi^\star}[\{\tau_{\pi_0,\Phi} \le H\} \cap B] + \Pr_{N,\pi^\star}[B^c] \\
&\le \Pr_{N,\pi^\star}[\tau_{\pi_0,\Phi} \le H] + \Pr_{N,\pi^\star}[\tau_{\pi_0,\Phi} > \tau_{\pi_0,\{\pi^\star\}}].
\end{aligned}$$

For a fixed validator set, the abstention event $\{\tau \le H\}$ is strictly a measurable function of the state trajectory. Thus, its probability under the expert can differ between the test environment (N) and training environment (M) by at most the TV distance:

$$\Pr_{N,\pi^\star}[\tau_{\pi_0,\Phi} \le H] \le \Pr_{M,\pi^\star}[\tau_{\pi_0,\Phi} \le H] + d^{\mathrm{state}}_{\pi^\star}(M, N).$$

Finally, substituting the high-probability bounds from the proof of Theorem 1 (where the train-side abstention is bounded by Z/m and the late-stop risk by $\eta + \sqrt{2\eta Z/n} + 3Z/n$) yields

$$\alpha_N(\pi_0, \tau) \le \underbrace{\frac{Z}{m} + d^{\mathrm{state}}_{\pi^\star}(M, N)}_{\text{PQ-style TV transfer}} + \underbrace{\eta + \sqrt{\frac{2\eta Z}{n}} + \frac{3Z}{n}}_{\text{Sequential penalty}}.$$

The first two terms are the sequential analogue of the rejection-rate transfer in batch selective classification [Goldwasser et al., 2020]: the abstention event is a fixed measurable set, so its probability shifts by at most the TV distance. The final term has no batch analogue; it arises because the deployed policy's actions can steer the trajectory out of the safe prefix before the validators stop it.

B.6 Constructing a validator distribution using no-regret dynamics

The nonconstructive minimax proof of Lemma 1 can be made algorithmic via the well-known connection between approximate minimax equilibria and no-regret play. In Section B.6, we formulate the problem as a repeated game where a maximization player maintains a distribution over the version space using a no-regret algorithm, and the validation player plays an (approximate) best-response by sampling a sparse validator set. Averaging the validation player's choices over T rounds yields a distribution that converges to the minimax optimum. This gives an explicit, oracle-efficient implementation of the SPARSEVALIDATORDIST subroutine.

Fix the version space Π_{version}, a base policy $\pi_0 \in \Pi_{\mathsf{version}}$, and unlabeled test trajectories $\mathcal{D}_{\mathsf{test}} = \{T_j\}_{j=1}^n$. Write each test trajectory as $T_j = (s_{j1}, \dots, s_{jH})$.

Algorithm 2 SPARSEVALIDATORDIST$(\Pi_{\mathsf{version}}, \pi_0, \mathcal{D}_{\mathsf{test}}, \rho, \xi, \delta)$

Require: version space Π_{version}, base policy π_0, unlabeled test trajectories $\mathcal{D}_{\mathsf{test}} = \{T_j\}_{j=1}^n$, tolerance $\rho \in (0,1)$, computational slack ξ, confidence δ, no-regret algorithm $\mathcal{A}$ with regret bound Reg_T.

1: Set $K \leftarrow \lceil 1/\rho \rceil$.
2: Set T large enough that $\frac{\mathrm{Reg}_T}{T} + \sqrt{\frac{\log(1/\delta)}{2T}} \le \xi$.
3: Initialize any distribution p^1 over Π_{version}.
4: **for** $t = 1, \dots, T$ **do**
5: Draw $\pi_1^t, \dots, \pi_K^t \overset{\text{i.i.d.}}{\sim} p^t$ and set $\Phi^t \leftarrow \{\pi_1^t, \dots, \pi_K^t\}$.
6: For each $\pi \in \Pi_{\mathsf{version}}$, define the payoff $u^t(\pi) \coloneqq \frac{1}{n}\sum_{j=1}^n \mathbf{1}\left[\tau_{\pi_0,\Phi^t}(T_j) > \tau_{\pi_0,\{\pi\}}(T_j)\right]$.
7: Update p^{t+1} using $\mathcal{A}$ for maximizing the payoffs u^t.
8: Return the empirical distribution $\bar{q} \coloneqq \frac{1}{T}\sum_{t=1}^T \delta_{\Phi^t}$.

The interpretation of the payoff is straightforward: $u^t(\pi)$ is the fraction of empirical test trajectories on which the validator set Φ^t stops strictly later than the single competing policy π. Thus the maximization player tries to expose competitors against which the current validator set is weak, while the validation player responds by sampling a fresh set from the current maximizer distribution.

Proposition 1 (Generic no-regret construction of $q^\star$)**.** *Let $\bar{q}$ be the output of Algorithm 2. With probability at least $1-\delta$, we have*

$$\sup_{\pi \in \Pi_{\mathsf{version}}} \mathbb{E}_{\Phi\sim\bar{q}}\left[\frac{1}{n}\sum_{j=1}^n \mathbf{1}\left[\tau_{\pi_0,\Phi}(T_j) > \tau_{\pi_0,\{\pi\}}(T_j)\right]\right] \le \rho + \xi.$$

In particular, using Hedge yields $\mathrm{Reg}_T \le \sqrt{2T\log|\Pi|}$, meaning $T = O\left(\frac{\log|\Pi| + \log(1/\delta)}{\rho^2}\right)$ rounds suffice to return an $O(\rho)$-valid distribution supported on sets of size at most $\lceil 1/\rho \rceil$.

The no-regret construction above is explicit, but inefficient (when using Hedge) when the version space is large. A standard choice in online learning with large action spaces is to use Follow-the-Perturbed-Leader (FTPL) instead of Hedge. However, a naive FTPL implementation would perturb one coordinate per policy in Π_{version}, which is also intractable. Instead, we show that by perturbing the *dataset*, and then solving a multiple-instance learning problem [Dietterich et al., 1997, Maron and Lozano-Pérez, 1997] over this perturbed dataset, we can implement an *oracle-efficient* no-regret algorithm for the maximizer. This places our implementation in the setting of generalized FTPL [Dudík et al., 2020]. The full formulation and proofs of this reduction are deferred to Section B.8.

B.7 Proof of Proposition 1

By the definition of $\bar{q}$, the expected coverage for any fixed $\pi \in \Pi_{\mathsf{version}}$ is exactly the average payoff over all rounds:

$$\mathbb{E}_{\Phi\sim\bar{q}}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}\left[\tau_{\pi_0,\Phi}(T_j) > \tau_{\pi_0,\{\pi\}}(T_j)\right]\right] = \frac{1}{T}\sum_{t=1}^{T}u^t(\pi).$$

By the definition of external regret, this average payoff is bounded by the algorithm's average expected payoff plus the regret term:

$$\frac{1}{T}\sum_{t=1}^{T}u^t(\pi) \le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi'\sim p^t}[u^t(\pi')] + \frac{\mathrm{Reg}_T}{T}.$$

It therefore remains to control the average expected payoff of the mixed strategy p^t against the sampled validator set Φ^t.

Conditioned on the history $\mathcal{H}_{t-1}$ up to round $t-1$, the distribution p^t is fixed, and Φ^t is obtained by drawing $K = \lceil 1/\rho \rceil$ policies i.i.d. from p^t. The exchangeability argument from Lemma 1 bounds the conditional expected payoff:

$$\mathbb{E}\Big[\mathbb{E}_{\pi'\sim p^t}[u^t(\pi')] \;\Big|\; \mathcal{H}_{t-1}\Big] \le \frac{1}{K+1} < \rho.$$

Since the expected payoffs $\mathbb{E}_{\pi'\sim p^t}[u^t(\pi')]$ are bounded in $[0,1]$, we can apply the Azuma–Hoeffding inequality. With probability at least $1-\delta$, the empirical average concentrates around its conditional expectation upper bound:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi'\sim p^t}[u^t(\pi')] \le \rho + \sqrt{\frac{\log(1/\delta)}{2T}}.$$

Combining these bounds completes the proof.

B.8 Oracle-efficient constructions of a validator distribution

As mentioned in Section B.6, the generic no-regret construction of a sparse validator distribution is inefficient. In this section, we instantiate the generalized FTPL framework of Dudík et al. [2020] to obtain an oracle-efficient implementation of the no-regret strategy required by Algorithm 2.

The generalized FTPL framework guarantees efficient no-regret learning, provided an online game can be reduced to a linearly perturbed offline optimization problem. To achieve this, we must map our abstract policy class into a structured vector space via a translation matrix. Crucially, this matrix must satisfy two structural conditions to ensure the random perturbations effectively smooth the leader's choices: admissibility (the matrix must have bounded entries and separate distinct policies to control the geometric complexity of the action space) and implementability (any payoff generated by the adversary must be expressible as a linear combination of the matrix's columns). If these conditions hold, FTPL can simulate the online game using synthetic datasets passed to a standard offline optimization oracle.

We now formulate the online learning problem by defining the strategy space X, the adversary action space Y, and the translation matrix Γ. This is very similar to the online learning problem constructed in Section B.6, except slightly more pedantic to ensure that our game satisfies the technical conditions required by generalized FTPL.

- **Strategy space X:** Quotient Π_{version} by empirical stopping-time equivalence: $\pi \sim \pi' \iff \tau_{\pi_0,\{\pi\}}(T_j) = \tau_{\pi_0,\{\pi'\}}(T_j)$ for all $j \in [n]$. Let X denote the set of equivalence classes, and for $\pi \in X$ write $x_\pi \coloneqq (\tau_{\pi_0,\{\pi\}}(T_1), \ldots, \tau_{\pi_0,\{\pi\}}(T_n)) \in \{1, \ldots, H+1\}^n$.
- **Adversary action space Y:** Let $Y \coloneqq \{1, \ldots, H+1\}^n$. For any validator set Φ, we define its empirical cutoff vector $h^\Phi \coloneqq (\tau_{\pi_0,\Phi}(T_1), \ldots, \tau_{\pi_0,\Phi}(T_n))$, which is an element of Y.

We define the payoff function $f\colon X \times Y \to [0,1]$ as $f(\pi, h) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}[x_{\pi,j} < h_j]$. Observe that for any validator set Φ, the payoff evaluates to $f(\pi, h^{\Phi}) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}[\tau_{\pi_0,\Phi}(T_j) > \tau_{\pi_0,\{\pi\}}(T_j)]$. Thus, the validator-distribution problem is exactly equivalent to online learning over X against adversary actions in Y with payoff function f.

Finally, we define the generalized FTPL translation matrix Γ. Its rows are indexed by the strategy space X, its columns by the environment time steps $(j, c) \in [n] \times \{1, \ldots, H\}$, and its entries map the policies into structured cumulative features:

$$\Gamma_{\pi,(j,c)} \coloneqq \mathbf{1}[\tau_{\pi_0,\{\pi\}}(T_j) \le c].$$

We now prove that Γ satisfies the technical conditions necessary for generalized FTPL.

Lemma 6 (Admissibility and implementability of the cutoff matrix)**.** *The matrix Γ is 1-admissible and implementable with complexity 1, in the sense of Dudík et al. [2020].*

Proof. We first prove admissibility. Let $\pi, \pi' \in X$ be distinct rows. Since they represent distinct stopping-time vectors, there exists some trajectory j such that $\tau_{\pi_0,\{\pi\}}(T_j) \neq \tau_{\pi_0,\{\pi'\}}(T_j)$. Assuming without loss of generality that $\tau_{\pi_0,\{\pi\}}(T_j) < \tau_{\pi_0,\{\pi'\}}(T_j)$, we set $c^\star \coloneqq \tau_{\pi_0,\{\pi'\}}(T_j) - 1$. Then $\Gamma_{\pi,(j,c^\star)} = 1$ and $\Gamma_{\pi',(j,c^\star)} = 0$. Thus any two distinct rows are separated by some column. Because all matrix entries lie in $\{0, 1\}$, distinct values in a column differ by exactly 1, making Γ 1-admissible.

To prove implementability, we must show that each column can be simulated by a synthetic adversary action. Fix a column (j, c) and define the adversary action $y^{(j,c)} \in Y$ by $y_j^{(j,c)} = c+1$ and $y_{j'}^{(j,c)} = 1$ for all $j' \neq j$. Then for any $\pi \in X$:

$$\begin{aligned} f(\pi, y^{(j,c)}) &= \frac{1}{n}\left(\mathbf{1}\left[\tau_{\pi_0,\{\pi\}}(T_j) < c+1\right] + \sum_{j' \neq j} \mathbf{1}\left[\tau_{\pi_0,\{\pi\}}(T_{j'}) < 1\right]\right) \\ &= \frac{1}{n}\,\mathbf{1}\left[\tau_{\pi_0,\{\pi\}}(T_j) \le c\right] = \frac{1}{n}\,\Gamma_{\pi,(j,c)}, \end{aligned}$$

since $\tau_{\pi_0,\{\pi\}}(T_{j'}) \ge 1$ always. Therefore, the weighted singleton dataset $S_{j,c} \coloneqq \{(n, y^{(j,c)})\}$ implements column (j, c), because for all $\pi, \pi' \in X$, $\sum_{(w,y)\in S_{j,c}} w\big(f(\pi, y) - f(\pi', y)\big) = \Gamma_{\pi,(j,c)} - \Gamma_{\pi',(j,c)}$. Thus, each column is implementable by a single synthetic adversary action. □

By Corollary 2.11 of Dudík et al. [2020], the admissibility and implementability of Γ guarantee the existence of an oracle-efficient, randomized no-regret maximization player. Specifically, because our translation matrix Γ is 1-admissible (i.e., $\delta = 1$) and contains $N = nH$ columns, the generalized FTPL framework yields an expected regret bound of $O(N\sqrt{T}/\delta) = O(nH\sqrt{T})$. It remains only to identify the specific offline optimization problem required by the oracle on each round.

It remains only to identify the specific offline optimization problem required by the oracle on each round. According to the generalized FTPL algorithm Dudík et al. [2020], on each round t, the maximization player draws independent, non-negative perturbations $\alpha_{j,c}^{t}$ for every column of the translation matrix. The player must then compute the perturbed best response against the empirical history of the adversary's actions—which, in our game, are the cutoff vectors h^{Φ^s} of the validator sets Φ^s sampled on previous rounds $s < t$. Formally, the oracle must solve:

$$\arg\max_{\pi \in \Pi_{\text{version}}} \left\{ \sum_{s=1}^{t-1} f(\pi, h^{\Phi^s}) + \sum_{j=1}^{n}\sum_{c=1}^{H} \alpha_{j,c}^{t}\,\Gamma_{\pi,(j,c)} \right\}, \tag{5}$$

where the first term represents the cumulative historical payoff and the second term injects the shared linear perturbations. We now connect this oracle optimization problem to the multiple instance learning (MIL) framework of Maron and Lozano-Pérez [1997].

Problem (5) **reduces to weighted MIL over the disagreement class.** To implement the oracle efficiently, we show that the perturbed leader problem (5) can be cast as a (weighted) MIL problem. The intuition behind this reduction is straightforward: a policy stops by step c on trajectory j if and only if it disagrees with the base policy π_0 on *at least one* state in the trajectory prefix up to step c. This "at least one" condition exactly mirrors the definition of a positive bag in MIL. By treating each

trajectory prefix as a "bag" of states, and any disagreement with π_0 as a positive instance label, we can formally map the FTPL objective to a realizable MIL objective.

Proposition 2. *For the base policy π_0 and any policy $\pi \in \Pi_{\mathsf{version}}$, define the disagreement hypothesis $h_\pi(h, s) \coloneqq \mathbf{1}[\pi_h(s) \neq \pi_{0,h}(s)]$, and let $\mathcal{H}_{\pi_0} \coloneqq \{h_\pi : \pi \in \Pi_{\mathsf{version}}\}$ be the corresponding disagreement class. On round t, the perturbed leader problem in* (5) *is exactly equivalent to the following weighted realizable multiple-instance disagreement problem over $\mathcal{H}_{\pi_0}$:*

$$\arg \max_{\pi \in \Pi_{\mathsf{version}}} \sum_{j=1}^{n} \sum_{c=1}^{H} w_{j,c}^{t-1} \, \mathbf{1}\left[\exists (h, s_{jh}) \in B_{j,c} : h_\pi(h, s_{jh}) = 1\right], \tag{6}$$

where the aggregated weights and prefix bags are defined as

$$w_{j,c}^{t-1} \coloneqq |\{s < t : \tau_{\pi_0, \Phi^s}(T_j) - 1 = c\}| + n\alpha_{j,c}^{t}, \qquad B_{j,c} \coloneqq \{(1, s_{j1}), \ldots, (c, s_{jc})\}.$$

Proof. Multiplying the objective in (5) by n (which preserves the argmax) and expanding the definitions of f and Γ yields:

$$\arg \max_{\pi \in \Pi_{\mathsf{version}}} \left\{ \sum_{s=1}^{t-1} \sum_{j=1}^{n} \mathbf{1}\left[\tau_{\pi_0,\{\pi\}}(T_j) \leq \tau_{\pi_0,\Phi^s}(T_j) - 1\right] + n \sum_{j=1}^{n} \sum_{c=1}^{H} \alpha_{j,c}^{t} \, \mathbf{1}\left[\tau_{\pi_0,\{\pi\}}(T_j) \leq c\right] \right\}.$$

Regrouping the first term by the cutoff value c and combining it with the perturbation term gives the simplified objective $\sum_{j=1}^{n} \sum_{c=1}^{H} w_{j,c}^{t-1} \mathbf{1}[\tau_{\pi_0,\{\pi\}}(T_j) \leq c]$. We map this to the MIL framework using the disagreement hypothesis $h_\pi(h, s) = \mathbf{1}[\pi_h(s) \neq \pi_{0,h}(s)]$. By definition, the event $\tau_{\pi_0,\{\pi\}}(T_j) \leq c$ occurs if and only if there is some step $h \leq c$ where $\pi_h(s_{jh}) \neq \pi_{0,h}(s_{jh})$. This is precisely the condition that the prefix bag $B_{j,c}$ is labeled positive by h_π, establishing (6). Finally, since any $\pi \in \Pi_{\mathsf{version}}$ matches the expert on the labeled training data, h_π is strictly zero on those states, satisfying the negative singleton constraints of realizable MIL. □

Problem (5) **reduces to weighted MIL over the original class.** Alternatively, for finite action spaces $\mathcal{A}$, we can avoid the binary disagreement wrapper and reduce directly over Π_{version} via data augmentation. By duplicating each state in a prefix for every alternative action $a \neq \pi_0(s)$, we create an augmented bag that a standard multi-class MIL solver evaluates as positive if π outputs *any* of those alternative actions. This inflates bag sizes by $|\mathcal{A}| - 1$.

Proposition 3 (Exact perturbed best response as multi-class MIL over Π_{version})**.** *Assume a finite action space $\mathcal{A}$. On round t, the perturbed leader problem in* (5) *is exactly equivalent to the following weighted realizable multi-class multiple-instance problem over Π_{version}:*

$$\arg \max_{\pi \in \Pi_{\mathsf{version}}} \sum_{j=1}^{n} \sum_{c=1}^{H} w_{j,c}^{t-1} \, \mathbf{1}\left[\exists ((h, s_{jh}), a) \in \tilde{B}_{j,c} : \pi_h(s_{jh}) = a\right], \tag{7}$$

with weights $w_{j,c}^{t-1}$ as in Proposition 2 and augmented prefix bags:

$$\tilde{B}_{j,c} \coloneqq \left\{ \big((h, s_{jh}), a\big) \;\middle|\; h \in \{1, \ldots, c\},\ a \in \mathcal{A} \setminus \{\pi_{0,h}(s_{jh})\} \right\}.$$

Proof. The simplified FTPL objective matches Proposition 2; we only need to verify the positive bag condition. A policy π disagrees with π_0 on the prefix up to step c if and only if $\pi_h(s_{jh}) = a$ for some step $h \leq c$ and alternative action $a \in \mathcal{A} \setminus \{\pi_{0,h}(s_{jh})\}$. This exactly matches the condition that π predicts at least one target label a in the augmented bag $\tilde{B}_{j,c}$. Finally, since all $\pi \in \Pi_{\mathsf{version}}$ perfectly match the expert on the labeled training data, they never predict alternative actions on those states, naturally satisfying the negative singleton constraints of realizable MIL. □

Agnostic oracle and computational-statistical tradeoff. So far, we have established the existence of an oracle-efficient subroutine to compute a validator distribution. However, the specific optimization problem, (weighted) MIL, is somewhat of an exotic class of problem and it is not clear such oracles may actually be implemented or approximated in practice. It turns out, however, that we can relax the requirement of a MIL oracle for Π to a standard *agnostic learning* oracle for Π (in fact, we can

relax this to a weaker oracle known as a *reliable learning* oracle [Kalai and Kanade, 2021]). This is simply due to the stepwise reduction to PQ-learning of Section B.2.

Of course, the cost is a degradation in our statistical guarantees. In particular, by moving from our trajectory-wise problem to the stepwise reduction, we solve an easier computational problem but pick up a dependence on H in our sample complexity. We leave the question of whether this tradeoff is fundamental, or just a by-product of our algorithms, to future work.

C Proofs from the Stochastic Section

Algorithm 3 SeqRejectron for stochastic policies

Require: policy class Π, labeled rollouts $\mathcal{D}_{\mathsf{train}} = \{(s_{ih}, a_{ih})\}_{i\in[m],h\in[H]}$, unlabeled rollouts $\mathcal{D}_{\mathsf{test}} = \{T_j\}_{j=1}^{n}$, tolerance η, confidence δ, threshold θ, version-space radius γ

1: Compute the MLE base policy $\pi_0 \leftarrow \arg\min_{\pi\in\Pi} \mathsf{LogLoss}(\pi, \mathcal{D}_{\mathsf{train}})$.
2: Form the version space $\Pi^{\gamma}_{\mathsf{version}}$.
3: Let $q^\star \leftarrow \textsc{StochasticSparseValidatorDist}(\Pi^{\gamma}_{\mathsf{version}}, \pi_0, \mathcal{D}_{\mathsf{test}}, \eta/2, \theta, \delta/4)$.
4: Sample $\Phi_1, \ldots, \Phi_k \overset{\text{i.i.d.}}{\sim} q^\star$ with $k \leftarrow \lceil \log_2(4/\delta) \rceil$ and set $\Phi \leftarrow \bigcup_{r=1}^{k} \Phi_r$.
5: **return** the selective policy $(\pi_0, \tau^{\theta}_{\pi_0,\Phi})$.

C.1 Proof of Theorem 2

Let $(\pi_0, \tau^{\theta}_{\pi_0,\Phi})$ be the output of Algorithm 3. We first control the abstention probability on M, and then the stopped trajectory Hellinger distance on N.

Step 1 (Control abstention on M). The quantity we want to bound is the probability that the cumulative Hellinger distance exceeds θ. Define $\gamma^\star := \frac{\log|\Pi|+\log(8/\delta)}{m}$, and let γ be the version space radius which is used by Algorithm 3. By assumption on m, we have that $\gamma^\star \le \gamma$. Thus, by Lemma 7, with probability at least $1-\delta/4$, we have $\pi^\star \in \Pi^{\gamma}_{\mathsf{version}}$ and every $\pi \in \Pi^{\gamma}_{\mathsf{version}}$ satisfies

$$D^2_H(\mathcal{P}^{\pi}_M, \mathcal{P}^{\pi^\star}_M) \le \frac{\gamma}{2} + \gamma^\star \le 1.5\gamma.$$

Combining this with the Hellinger triangle inequality, we have that for every $\pi \in \Pi^{\gamma}_{\mathsf{version}}$,

$$D^2_H(\mathcal{P}^{\pi}_M, \mathcal{P}^{\pi_0}_M) \le 2D^2_H(\mathcal{P}^{\pi}_M, \mathcal{P}^{\pi^\star}_M) + 2D^2_H(\mathcal{P}^{\pi_0}_M, \mathcal{P}^{\pi^\star}_M) \le 6\gamma.$$

Now consider any validator $\pi \in \Phi \subseteq \Pi^{\gamma}_{\mathsf{version}}$. Let E be the event that the cumulative Hellinger distance between π and π_0 exceeds θ. To bound the probability of E under π_0, we introduce the geometric mixture policy $\bar\pi \propto \sqrt{\pi\pi_0}$, using the convention in Lemma 8. By Lemma 8, the trajectory affinity satisfies

$$1 - D^2_H(\mathcal{P}^{\pi}_M, \mathcal{P}^{\pi_0}_M) \le \mathbb{E}_{T\sim\mathcal{P}^{\bar\pi}_M}\left[e^{-\sum_{h=1}^{H} d^2_H(\pi_h,\pi_{0,h})}\right] \le \Pr_{\mathcal{P}^{\bar\pi}_M}[E]e^{-\theta} + 1 - \Pr_{\mathcal{P}^{\bar\pi}_M}[E].$$

Rearranging this inequality yields $\Pr_{\mathcal{P}^{\bar\pi}_M}[E] \le \frac{1}{1-e^{-\theta}} D^2_H(\mathcal{P}^{\pi}_M, \mathcal{P}^{\pi_0}_M) \le \frac{1+\theta}{\theta} D^2_H(\mathcal{P}^{\pi}_M, \mathcal{P}^{\pi_0}_M)$. To shift this probability from the mixture policy to the deployed policy π_0, we apply a Hellinger change-of-measure inequality (e.g., Lemma 3.1 of Foster et al. [2024]) alongside Lemma 9:

$$\begin{aligned}
\Pr_{\mathcal{P}^{\pi_0}_M}[E] &\le 2\Pr_{\mathcal{P}^{\bar\pi}_M}[E] + 4D^2_H(\mathcal{P}^{\bar\pi}_M, \mathcal{P}^{\pi_0}_M) \\
&\le \frac{2(1+\theta)}{\theta} D^2_H(\mathcal{P}^{\pi}_M, \mathcal{P}^{\pi_0}_M) + 6D^2_H(\mathcal{P}^{\pi}_M, \mathcal{P}^{\pi_0}_M) \\
&= \left(\frac{2}{\theta} + 8\right) D^2_H(\mathcal{P}^{\pi}_M, \mathcal{P}^{\pi_0}_M).
\end{aligned}$$

Substituting our version space bound $D^2_H(\mathcal{P}^{\pi}_M, \mathcal{P}^{\pi_0}_M) \le 6\gamma$, the probability of E under π_0 is at most $\left(\frac{12}{\theta} + 48\right)\gamma$. Note that the algorithm abstains ($\tau^{\theta}_{\pi_0,\Phi} \le H$) if and only if at least one of these cumulative distances exceeds θ. By a union bound over the $|\Phi| \le K_{\mathsf{ens}}$ validators,

$$\alpha_M(\pi_0, g) \le K_{\mathsf{ens}}\left(\frac{12}{\theta} + 48\right)\gamma.$$

Step 2 (Bound Hellinger Distance of stopped trajectories on N). Let $M = H + 1$. For each policy $\pi \in \Pi^{\gamma}_{\mathsf{version}}$, let $v^{\pi} \in [M]^n$ be its empirical stopping-time vector on the n test trajectories, where the j-th coordinate is given by $(v^{\pi})_j = \tau^{\theta}_{\pi_0,\{\pi\}}(T_j)$. Let $\mathcal{T} = \{v^{\pi} : \pi \in \Pi^{\gamma}_{\mathsf{version}}\}$ be the finite set of all such vectors.

By Lemma 5 applied to $\mathcal{T}$ with target error $\eta/2$, there exists a distribution $q^\star$ over subsets $\Phi' \subseteq \Pi^{\gamma}_{\mathsf{version}}$ satisfying $|\Phi'| \le \lceil 2/\eta \rceil$ such that, because $\pi^* \in \Pi^{\gamma}_{\mathsf{version}}$:

$$\mathbb{E}_{\Phi' \sim q^\star}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left[\tau^{\theta}_{\pi_0,\Phi'}(T_j) > \tau^{\theta}_{\pi_0,\{\pi^\star\}}(T_j)\right]\right] \le \frac{\eta}{2}.$$

For a validator set Φ, define the empirical risk $\widehat{p}(\Phi) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}[\tau^{\theta}_{\pi_0,\{\pi^\star\}}(T_j) < \tau^{\theta}_{\pi_0,\Phi}(T_j)]$. By an identical argument to Theorem 1 (Markov's inequality and multiple draws) we have $\widehat{p}(\Phi) \le \eta$ with probability at least $1 - \delta/4$.

To generalize to expert trajectories in N, note the final validator set satisfies $|\Phi| \le K_{\mathsf{ens}}$. Consider the indicator class

$$\mathcal{Y}_{K_{\mathsf{ens}}} \coloneqq \{T \mapsto \mathbf{1}[\tau^{\theta}_{\pi_0,\{\pi^\star\}}(T) < \tau^{\theta}_{\pi_0,\Phi}(T)] : \pi_0 \in \Pi, \Phi \subseteq \Pi, |\Phi| \le K_{\mathsf{ens}}\},$$

which has cardinality at most $|\Pi|^{K_{\mathsf{ens}}+1}$. Applying the empirical Bernstein inequality with a union bound, with probability at least $1 - \delta/4$, simultaneously for all such (π_0, Φ),

$$p(\Phi) \le \widehat{p}(\Phi) + \sqrt{\frac{2\widehat{p}(\Phi)Z}{n}} + \frac{3Z}{n} \le \eta + \sqrt{\frac{2\eta Z}{n}} + \frac{3Z}{n},$$

where $p(\Phi) \coloneqq \Pr_{T\sim(\pi^\star,N)}\left(\tau^{\theta}_{\pi_0,\{\pi^\star\}}(T) < \tau^{\theta}_{\pi_0,\Phi}(T)\right)$. Under the assumption $n \ge 8Z/\eta$, each correction term is at most $\eta/2$, yielding $p(\Phi) \le 2\eta$.

Finally, to transfer this to the deployed trajectories, let $\mathcal{E} \coloneqq \{\tau^{\theta}_{\pi_0,\{\pi^\star\}}(T) < \tau^{\theta}_{\pi_0,\Phi}(T)\}$ and define $\tau'(T) \coloneqq \tau^{\theta}_{\pi_0,\Phi\cup\{\pi^\star\}}(T)$. Applying Lemma 10 with validator set $\Phi \cup \{\pi^\star\}$ and comparison policy $\pi^\star$, we have deterministically $D^2_H(P^{\pi_0}_{|\tau'}, P^{\pi^\star}_{|\tau'}) \le \theta$, because the event $\tau^{\theta}_{\pi_0,\{\pi^\star\}}(T) < \tau^{\theta}_{\pi_0,\Phi\cup\{\pi^\star\}}(T)$ has probability zero under (π_0, N). Since $\mathcal{E}$ is measurable with respect to the trajectory stopped at τ', the binary Hellinger change-of-measure inequality yields

$$\begin{aligned}
\Pr_{(\pi_0,N)}(\mathcal{E}) &\le \left(\sqrt{\Pr_{(\pi^\star,N)}(\mathcal{E})} + \sqrt{2D^2_H\left(\mathcal{P}_N^{\pi_{0|\tau'}}, \mathcal{P}_N^{\pi^\star_{|\tau'}}\right)}\right)^2 \\
&\le 2\Pr_{(\pi^\star,N)}(\mathcal{E}) + 4D^2_H\left(\mathcal{P}_N^{\pi_{0|\tau'}}, \mathcal{P}_N^{\pi^\star_{|\tau'}}\right) \\
&\le 4\eta + 4\theta,
\end{aligned}$$

where the last step uses $p(\Phi) \le 2\eta$. A final application of Lemma 10, now with validator set Φ, comparison policy $\pi^\star$, and failure probabilities $4\eta + 4\theta$ and 2η, gives

$$\begin{aligned}
D^2_H(P_N^{\pi_{0|\tau}}, P_N^{\pi^\star_{|\tau}}) &\le \theta + \frac{(4\eta + 4\theta) + 2\eta}{2} \\
&= 3\theta + 3\eta.
\end{aligned}$$

The four failure events above each occur with probability at most $\delta/4$, so the claimed bounds hold simultaneously with probability at least $1 - \delta$.

C.2 Supporting Hellinger lemmas

We collect the proofs of the four Hellinger lemmas used in the proof of Theorem 2.

Lemma 7 (Approximate minimizers are Hellinger-close)**.** *Let π_0 be the MLE on a dataset of m trajectories, and define $\gamma^\star \coloneqq (\log|\Pi| + \log(2/\delta))/m$. With probability at least $1 - \delta$:*

1. *$\pi^\star \in \Pi^{\gamma^\star}_{\mathsf{version}}$, i.e., $\mathsf{LogLoss}(\pi^\star) \le \mathsf{LogLoss}(\pi_0) + \gamma^\star$.*

2. *Every* $\pi \in \Pi$ *with* $\mathsf{LogLoss}(\pi) \le \mathsf{LogLoss}(\pi_0) + \gamma$ *satisfies* $D_H^2(\mathcal{P}_M^{\pi}, \mathcal{P}_M^{\pi^\star}) \le \gamma/2 + \gamma^\star$.

Proof. Let $L(\pi) = \prod_{k=1}^{m} \prod_{h=1}^{H} \frac{\pi_h(a_h^{(k)}|s_h^{(k)})}{\pi_h^\star(a_h^{(k)}|s_h^{(k)})}$ denote the likelihood ratio of the dataset under π versus $\pi^\star$. For a single trajectory, the likelihood ratio under $\mathcal{P}_M^{\pi^\star}$ sums only the $\mathcal{P}_M^{\pi}$-mass on trajectories in the support of $\mathcal{P}_M^{\pi^\star}$, so its expectation is at most 1 (and exactly 1 whenever $\mathcal{P}_M^{\pi} \ll \mathcal{P}_M^{\pi^\star}$). Since $L(\pi)$ is a product of m independent trajectory likelihood ratios, we have $\mathbb{E}[L(\pi)] \le 1$ for all π. Thus, by Markov's inequality and a union bound,

$$\Pr\left[\sup_{\pi \in \Pi} L(\pi) > \frac{2|\Pi|}{\delta}\right] \le \frac{\delta}{2}.$$

Taking the log and dividing by m, with probability at least $1 - \delta/2$ the log-loss gap satisfies $\mathsf{LogLoss}(\pi^\star) \le \mathsf{LogLoss}(\pi_0) + \gamma^\star$, proving Part 1.

For Part 2, if π is a γ-approximate minimizer then $\mathsf{LogLoss}(\pi) \le \mathsf{LogLoss}(\pi^\star) + \gamma$ (since π_0 is the MLE), which rearranges to

$$-\frac{1}{m}\sum_{k=1}^{m} \log\left(\prod_{h=1}^{H} \frac{\pi_h(a_h^{(k)} \mid s_h^{(k)})}{\pi_h^\star(a_h^{(k)} \mid s_h^{(k)})}\right) \le \gamma \implies \sum_{k=1}^{m} \log \sqrt{\prod_{h=1}^{H} \frac{\pi_h(a_h^{(k)} \mid s_h^{(k)})}{\pi_h^\star(a_h^{(k)} \mid s_h^{(k)})}} \ge -\frac{m\gamma}{2} \tag{8}$$

$$\implies \prod_{k=1}^{m} \sqrt{\prod_{h=1}^{H} \frac{\pi_h(a_h^{(k)} \mid s_h^{(k)})}{\pi_h^\star(a_h^{(k)} \mid s_h^{(k)})}} \ge e^{-m\gamma/2}. \tag{9}$$

By definition, the expectation of each factor in the outer product equals $1 - D_H^2(\mathcal{P}_M^{\pi}, \mathcal{P}_M^{\pi^\star})$, so the LHS has expectation $(1 - D_H^2(\mathcal{P}_M^{\pi}, \mathcal{P}_M^{\pi^\star}))^m \le e^{-mD_H^2(\mathcal{P}_M^{\pi}, \mathcal{P}_M^{\pi^\star})}$. For any π with $D_H^2(\mathcal{P}_M^{\pi}, \mathcal{P}_M^{\pi^\star}) > \gamma/2 + \gamma^\star$, Markov's inequality applied to (9) gives failure probability at most $\delta/(2|\Pi|)$. A union bound over Π completes the proof with overall probability at least $1 - \delta$. □

Lemma 8 (Hellinger tensorization under geometric mixture)**.** *Let* $\mathcal{P}^{\pi}$ *and* $\mathcal{P}^{\pi'}$ *denote the trajectory distributions under policies* π *and* π' *in an MDP* M*. Define the geometric mixture policy by* $\bar{\pi}_h(a \mid s) \coloneqq \sqrt{\pi_h(a \mid s)\,\pi'_h(a \mid s)}/Z_h(s)$*, where* $Z_h(s) \coloneqq \sum_a \sqrt{\pi_h(a \mid s)\,\pi'_h(a \mid s)}$*.*[9] *Then*

$$1 - D_H^2(\mathcal{P}^{\pi}, \mathcal{P}^{\pi'}) = \mathbb{E}_{T \sim \mathcal{P}^{\bar{\pi}}}\left[\prod_{h=1}^{H} \left(1 - d_H^2(\pi_h(\cdot \mid s_h), \pi'_h(\cdot \mid s_h))\right)\right].$$

Proof. By definition, the trajectory-level squared Hellinger affinity is

$$1 - D_H^2(\mathcal{P}^{\pi}, \mathcal{P}^{\pi'}) = \sum_{T} \sqrt{\mathcal{P}^{\pi}(T)\,\mathcal{P}^{\pi'}(T)}.$$

Expanding each trajectory probability,

$$\sqrt{\mathcal{P}^{\pi}(T)\,\mathcal{P}^{\pi'}(T)} = P_0(s_1)\left(\prod_{h=1}^{H-1} P_h(s_{h+1} \mid s_h, a_h)\right)\left(\prod_{h=1}^{H} \sqrt{\pi_h(a_h \mid s_h)\,\pi'_h(a_h \mid s_h)}\right).$$

Let $Z_h(s_h) \coloneqq \sum_a \sqrt{\pi_h(a \mid s_h)\,\pi'_h(a \mid s_h)} = 1 - d_H^2(\pi_h(\cdot \mid s_h), \pi'_h(\cdot \mid s_h))$. With the convention in the lemma statement, the identity $\sqrt{\pi_h(a_h \mid s_h)\,\pi'_h(a_h \mid s_h)} = \bar{\pi}_h(a_h \mid s_h)\,Z_h(s_h)$ holds for every s_h, a_h: it is the definition when $Z_h(s_h) > 0$, and both sides are zero when $Z_h(s_h) = 0$. Substituting,

$$\sqrt{\mathcal{P}^{\pi}(T)\,\mathcal{P}^{\pi'}(T)} = \mathcal{P}^{\bar{\pi}}(T)\prod_{h=1}^{H} Z_h(s_h).$$

Summing over all trajectories T gives the result. □

[9] If $Z_h(s) = 0$, define $\bar{\pi}_h(\cdot \mid s)$ arbitrarily. In this case $\pi_h(\cdot \mid s)$ and $\pi'_h(\cdot \mid s)$ have disjoint support, so $\sqrt{\pi_h(a \mid s)\pi'_h(a \mid s)} = 0$ for every a, and the identities below still hold.

Lemma 9 (Geometric mixture Hellinger bound)**.** *Let $\pi, \pi', \bar{\pi}$ be as in Lemma 8 and let $\gamma = D_H^2(\mathcal{P}^\pi, \mathcal{P}^{\pi'})$. Then*

$$D_H^2(\mathcal{P}^{\bar{\pi}}, \mathcal{P}^{\pi'}) \le \frac{3}{2} D_H^2(\mathcal{P}^\pi, \mathcal{P}^{\pi'}).$$

Proof. Define

$$P := \mathcal{P}^\pi, \qquad Q := \mathcal{P}^{\pi'}, \qquad \bar{P} := \mathcal{P}^{\bar{\pi}}.$$

From the proof of Lemma 8, we have that pointwise $\bar{P}(T)Z(T) = \sqrt{P(T)Q(T)}$ where $Z(T) = \prod_{h=1}^H Z_h(s_h) \le 1$. Therefore $\bar{P}(T) \ge \sqrt{P(T)Q(T)}$, and

$$1 - D_H^2(\bar{P}, Q) = \sum_T \sqrt{\bar{P}(T)Q(T)} \ge \sum_T \sqrt{\sqrt{P(T)Q(T)} \cdot Q(T)} = \sum_T P(T)^{1/4} Q(T)^{3/4}.$$

Applying Hölder's inequality with conjugate exponents $3/2$ and 3, we obtain

$$\begin{aligned}
\sum_T P(T)^{1/2} Q(T)^{1/2} &= \sum_T \big(P(T)^{1/4} Q(T)^{3/4}\big)^{2/3} P(T)^{1/3} \\
&\le \left(\sum_T \Big[\big(P(T)^{1/4} Q(T)^{3/4}\big)^{2/3} \Big]^{3/2} \right)^{2/3} \left(\sum_T \big(P(T)^{1/3}\big)^3 \right)^{1/3} \\
&= \left(\sum_T P(T)^{1/4} Q(T)^{3/4} \right)^{2/3} \left(\sum_T P(T) \right)^{1/3} \\
&= \left(\sum_T P(T)^{1/4} Q(T)^{3/4} \right)^{2/3}.
\end{aligned}$$

Hence $1 - D_H^2(\mathcal{P}^\pi, \mathcal{P}^{\pi'}) \le \big(\sum_T P(T)^{1/4} Q(T)^{3/4}\big)^{2/3}$. Raising to the $3/2$ power, $\sum_T P(T)^{1/4} Q(T)^{3/4} \ge (1 - D_H^2(\mathcal{P}^\pi, \mathcal{P}^{\pi'}))^{3/2}$, so by Bernoulli's inequality:

$$D_H^2(\bar{P}, Q) \le 1 - (1 - D_H^2(\mathcal{P}^\pi, \mathcal{P}^{\pi'}))^{3/2} \le \frac{3}{2} D_H^2(\mathcal{P}^\pi, \mathcal{P}^{\pi'}).$$

□

Lemma 10 (Stopped trajectory Hellinger bound)**.** *Let $\pi_0, \pi' \in \Pi$ and $\Phi \subseteq \Pi$. Let $\tau \coloneqq \tau^\theta_{\pi_0, \Phi}(T)$. If Φ satisfies*

$$\Pr_{T \sim (\pi_0, N)} \big(\tau^\theta_{\pi_0, \{\pi'\}}(T) < \tau \big) < \eta \quad \textit{and} \quad \Pr_{T \sim (\pi', N)} \big(\tau^\theta_{\pi_0, \{\pi'\}}(T) < \tau \big) < \eta'$$

then, with probability at least $1 - \eta$ over $T \sim (\pi_0, N)$, the cumulative per-step Hellinger distance up to τ is bounded by θ:

$$\sum_{h=1}^{\tau - 1} d_H^2(\pi'_h(\cdot \mid s_h), \pi_{0,h}(\cdot \mid s_h)) \le \theta.$$

Furthermore, the stopped trajectory distributions satisfy

$$D_H^2\big(\mathcal{P}_N^{\pi_{0|\tau}}, \mathcal{P}_N^{\pi'_{|\tau}}\big) \le \theta + \frac{\eta + \eta'}{2}.$$

Proof. Let

$$\tau' \coloneqq \tau^\theta_{\pi_0, \{\pi'\}}(T), \qquad \mathcal{E} \coloneqq \{\tau \le \tau'\}.$$

By hypothesis, $\Pr_{(\pi_0, N)}(\mathcal{E}^c) < \eta$, so $\Pr(\mathcal{E}) \ge 1 - \eta$. Conditioned on $\mathcal{E}$, the sum

$$\sum_{h=1}^{\tau - 1} d_H^2(\pi'_h(\cdot | s_h), \pi_{0,h}(\cdot | s_h)) \le \theta$$

by the definition of τ' and monotonicity of the sum.

For the second claim, define $\tau^\star(T) := \min(\tau, \tau')$. Since $\tau^\star \le \tau'$ on every trajectory, we have $\sum_{h<\tau^\star} d_H^2(\pi'_h(\cdot \mid s_h), \pi_{0,h}(\cdot \mid s_h)) \le \theta$ pathwise. The stopped affinity therefore satisfies

$$\sum_T \sqrt{\mathcal{P}_N^{\pi_{0|\tau^\star}}(T)\mathcal{P}_N^{\pi'_{|\tau^\star}}(T)} = \mathbb{E}_{T\sim\mathcal{P}^{\bar\pi}}\left[\prod_{h<\tau^\star}\left(1 - d_H^2(\pi'_h(\cdot|s_h), \pi_{0,h}(\cdot|s_h))\right)\right]$$
$$\ge \mathbb{E}_{T\sim\mathcal{P}^{\bar\pi}}\left[1 - \sum_{h<\tau^\star} d_H^2(\pi'_h(\cdot|s_h), \pi_{0,h}(\cdot|s_h))\right] \ge 1-\theta,$$

so $D_H^2(\mathcal{P}_N^{\pi_{0|\tau^\star}}, \mathcal{P}_N^{\pi'_{|\tau^\star}}) \le \theta$.

Write $A(P,Q) = \sum_T \sqrt{P(T)Q(T)}$ for the Hellinger affinity. On $\mathcal{E}$, the stopped distributions at τ and $\tau^\star$ agree, so

$$A(\mathcal{P}_N^{\pi_{0|\tau}}, \mathcal{P}_N^{\pi'_{|\tau}}) \ge \sum_{T\in\mathcal{E}} \sqrt{\mathcal{P}_N^{\pi_{0|\tau^\star}}(T)\mathcal{P}_N^{\pi'_{|\tau^\star}}(T)}.$$

On the other hand,

$$A(\mathcal{P}_N^{\pi_{0|\tau^\star}}, \mathcal{P}_N^{\pi'_{|\tau^\star}}) = \sum_{T\in\mathcal{E}} \sqrt{\mathcal{P}_N^{\pi_{0|\tau^\star}}(T)\mathcal{P}_N^{\pi'_{|\tau^\star}}(T)} + \sum_{T\in\mathcal{E}^c} \sqrt{\mathcal{P}_N^{\pi_{0|\tau^\star}}(T)\mathcal{P}_N^{\pi'_{|\tau^\star}}(T)}.$$

By AM-GM, the second sum satisfies

$$\sum_{T\in\mathcal{E}^c} \sqrt{\mathcal{P}_N^{\pi_{0|\tau^\star}}(T)\mathcal{P}_N^{\pi'_{|\tau^\star}}(T)} \le \sum_{T\in\mathcal{E}^c} \frac{\mathcal{P}_N^{\pi_{0|\tau^\star}}(T) + \mathcal{P}_N^{\pi'_{|\tau^\star}}(T)}{2} \le \frac{\eta+\eta'}{2}$$

Therefore $A(\mathcal{P}_N^{\pi_{0|\tau}}, \mathcal{P}_N^{\pi'_{|\tau}}) \ge A(\mathcal{P}_N^{\pi_{0|\tau^\star}}, \mathcal{P}_N^{\pi'_{|\tau^\star}}) - \frac{\eta+\eta'}{2}$, which gives

$$D_H^2(\mathcal{P}_N^{\pi_{0|\tau}}, \mathcal{P}_N^{\pi'_{|\tau}}) \le D_H^2(\mathcal{P}_N^{\pi_{0|\tau^\star}}, \mathcal{P}_N^{\pi'_{|\tau^\star}}) + (\eta+\eta')/2 \le \theta + \frac{\eta+\eta'}{2}$$

□

C.3 Proof of Lemma 2

Because π_{sw} and $\pi^\star$ are valid policies evaluated over the full-horizon MDP N, we can directly invoke Theorem 3.1 of Foster et al. [2024] to bound their regret difference:

$$J(\pi_{\mathsf{sw}}) - J(\pi^\star) \le \sqrt{6\sigma_{\pi^\star}^2 \cdot D_H^2(\mathcal{P}^{\pi_{\mathsf{sw}}}, \mathcal{P}^{\pi^\star})} + O\left(C_{\max}\log(C_{\max}\epsilon^{-1})\right) \cdot D_H^2(\mathcal{P}^{\pi_{\mathsf{sw}}}, \mathcal{P}^{\pi^\star}) + \epsilon,$$

where $\mathcal{P}^{\pi_{\mathsf{sw}}}$ and $\mathcal{P}^{\pi^\star}$ are the full-horizon trajectory distributions of π_{sw} and $\pi^\star$, respectively.

To evaluate the full-horizon Hellinger distance, we expand the Hellinger affinity between $\mathcal{P}^{\pi_{\mathsf{sw}}}$ and $\mathcal{P}^{\pi^\star}$. Recall that the squared Hellinger distance satisfies $1 - D_H^2(P,Q) = \sum_T \sqrt{P(T)Q(T)}$. Let $T_{|\tau}$ denote the prefix of the trajectory up to the stopping time τ, and let $T_{>\tau}$ denote the suffix. Therefore, the full-horizon trajectory distribution for any policy π factors as $P^\pi(T) = P^\pi(T_{|\tau})P^\pi(T_{>\tau} \mid T_{|\tau})$.

By construction, the mixed policy π_{sw} executes the learner $\hat\pi$ up to τ and the expert $\pi^\star$ for all steps $h > \tau$. Thus, its prefix distribution is exactly the stopped learner distribution, $\mathcal{P}^{\pi_{\mathsf{sw}}}(T_{|\tau}) = \mathcal{P}^{\hat\pi_{|\tau}}(T_{|\tau})$. Crucially, because τ is a stopping time, the event $\{\tau = t\}$ depends exclusively on the trajectory prefix $T_{1:t}$. Consequently, for any prefix where the switch occurs at step t, the conditional distribution of the remaining trajectory is governed entirely by the expert policy, yielding $\mathcal{P}^{\pi_{\mathsf{sw}}}(T_{t+1:H} \mid T_{1:t}) = \mathcal{P}^{\pi^\star}(T_{t+1:H} \mid T_{1:t})$. Thus,

$$1 - D_H^2(\mathcal{P}^{\pi_{\mathsf{sw}}}, \mathcal{P}^{\pi^\star}) = \sum_{T_{|\tau}}\sum_{T_{>\tau}} \sqrt{\mathcal{P}^{\pi_{\mathsf{sw}}}(T_{|\tau})\mathcal{P}^{\pi_{\mathsf{sw}}}(T_{>\tau} \mid T_{|\tau}) \cdot \mathcal{P}^{\pi^\star}_{|\tau}(T_{|\tau})\mathcal{P}^{\pi^\star}(T_{>\tau} \mid T_{|\tau})}$$
$$= \sum_{T_{|\tau}} \sqrt{\mathcal{P}^{\hat\pi_{|\tau}}(T_{|\tau})\mathcal{P}^{\pi^\star_{|\tau}}(T_{|\tau})} \sum_{T_{>\tau}} \mathcal{P}^{\pi^\star}(T_{>\tau} \mid T_{|\tau}).$$

Because $P^{\pi^\star}(T_{>\tau} \mid T_{|\tau})$ is a valid probability distribution over the suffix trajectories, the inner sum evaluates exactly to 1 for every prefix $T_{|\tau}$. The remaining sum is exactly the Hellinger affinity of the stopped trajectories. Therefore, the full-horizon trajectory distance perfectly collapses to the distance accumulated up to the stopping time:

$$D_H^2(\mathcal{P}^{\pi_{\mathsf{sw}}}, \mathcal{P}^{\pi^\star}) = D_H^2\big(\mathcal{P}^{\hat{\pi}_{|\tau}}, \mathcal{P}^{\pi^\star_{|\tau}}\big). \tag{10}$$

Substituting (10) into the bound from Theorem 3.1 of Foster et al. [2024] concludes the proof.

C.4 Single-step lower bound for stochastic policies

How tight are these rates? The upper bound of Corollary 3 has two regimes: a variance-driven $\widetilde{O}(\epsilon^{-5})$ term and a cost-driven $\widetilde{O}(\epsilon^{-3})$ term. The next result shows that the cubic dependence is already unavoidable in the single-step setting, matching the cost-driven part.

Theorem 4 (Stochastic PQ lower bound)**.** *There exists a constant $c_0 > 0$ such that:*

For any dimension $d \geq 1$ and tolerance $\epsilon \in (0, 1/16]$, there exist a state space $\mathcal{X}$, a stochastic policy class Π of log-cardinality d, and fixed known distributions P and Q over $\mathcal{X}$ for which the following holds: For any proper[10] *selective learner, if the labeled sample size is $m \leq c_0 \frac{d}{\epsilon^3}$, then there exists an expert $\pi^\star \in \Pi$ and a bounded cost function $c : \mathcal{X} \times \{0,1\} \to [0,1]$ such that*

$$\mathbb{E}\left[\alpha_P(\pi, \tau)\right] > \epsilon \quad \text{or} \quad \mathbb{E}\left[\mathrm{Regret}_Q(\pi, \tau; c)\right] > \epsilon$$

where the outer expectation is over the labeled sample and the learner's internal randomization, which in particular determine π and τ.

Theorem 4 shows that the ϵ^{-3} dependence in labeled sample complexity is already unavoidable in the single-step stochastic setting. Crucially, because this lower bound is established in a single-step environment ($H = 1$), there is no future horizon over which errors can accumulate after an abstention. Consequently, the stopped regret and the switched regret are equal in this construction, meaning this lower bound applies to both formulations. It therefore matches the $C_{\max}$-driven $\widetilde{O}(\epsilon^{-3})$ portion of Corollary 3, while leaving a gap to the current variance-driven $\widetilde{O}(\epsilon^{-5})$ labeled-sample rate. For deterministic classes, the lower bound of Goldwasser et al. [2020] for PQ learning implies that the $\tilde{O}(1/\epsilon^2)$ labeled complexity is tight. Whether the sequential stochastic upper bound can be improved, or whether stronger lower bounds are possible, remains open.

C.5 Proof of Theorem 4

Let the state space be $\mathcal{X} = \{1, 2, \ldots, N\}$. Let P be uniform over $\mathcal{X}$, Q be uniform over a known subset $S_Q = \{1, \ldots, d\}$, and let $\Delta \in [0, 1/2]$ be a parameter to be chosen later. Let Π be the class of stochastic policies parameterized by $\sigma \in \{-1, 1\}^d$: for $x \in S_Q$, the expert $\pi^\star$ takes action 1 with probability $1/2 + \sigma_x \Delta$; for $x \notin S_Q$, the probability is $1/2$. Similarly, define the cost function $c_\sigma : \mathcal{X} \times \{0,1\} \to [0,1]$ by $c_\sigma(x, 1) = 1\{\sigma_x = -1\}$ and $c_\sigma(x, 0) = 1\{\sigma_x = 1\}$ for $x \in S_Q$ and 0 otherwise.

We now apply Yao's minimax principle. It is enough to lower bound the expected performance of a deterministic proper learner under the uniform prior $\sigma \sim \mathrm{Unif}\big(\{-1, 1\}^d\big)$. Let D be the m labeled samples from P the learner observes. Since the learner is deterministic, after observing D, it will, for each state x, choose an acceptance probability $\alpha_x(D) \in [0, 1]$ and a bernoulli parameter $q_x(D) = \hat{\pi}(1|x)$. Crucially, because the learner is proper, it must output a policy in Π, meaning $q_x(D) \in \{1/2 - \Delta, 1/2 + \Delta\}$.

For $x \in S_Q$, define the one-step selective regret at state x as

$$\begin{aligned} R_x(D, \sigma) &\coloneqq \mathbb{E}_{a \sim \hat{\pi}_D(\cdot|x)}[c_\sigma(x, a)] - \mathbb{E}_{a \sim \pi_\sigma(\cdot|x)}[c_\sigma(x, a)] \\ &= \Delta + \sigma_x \left(\frac{1}{2} - q_x(D)\right). \qquad \text{(direct computation)} \end{aligned}$$

[10] By proper selective learner, we mean that the base policy π_0 always lies in Π. This holds for SeqRejectron. In the batch (i.e., single step) setting this holds, for example, for Rejectron [Goldwasser et al., 2020] and the Slice-and-Dice algorithm of Kalai and Kanade [2021].

Now, define $\eta_x = \mathbb{E}[\sigma_x|D]$. By conditioning on D, we get

$$\mathbb{E}[R_x(D,\sigma)|D] = \Delta + \eta_x\left(\frac{1}{2} - q_x(D)\right) \geq \Delta - \Delta|\eta_x|$$

where we used the proper learner constraint $\left|\frac{1}{2} - q_x(D)\right| = \Delta$.

Multiplying by α_x and then taking the expectation over D gives

$$\mathbb{E}[\alpha_x(D)R_x(D,\sigma)] \geq \Delta\mathbb{E}[\alpha_x(D)] - \Delta\mathbb{E}[|\eta_x|]$$

where the full expectation is now over σ and D.

Let $n_x(D) \sim \text{Binomial}(m, 1/N)$ be the number of times state x is observed. The posterior bias satisfies

$$\begin{aligned}\mathbb{E}[|\eta_x(D)| \mid n_x(D) = k] &= \text{TV}\left(\text{Binomial}\left(k, \frac{1}{2} + \Delta\right), \text{Binomial}\left(k, \frac{1}{2} - \Delta\right)\right) \\ &\leq 2\sqrt{2}\Delta\sqrt{k}\end{aligned}$$

by Pinsker's inequality. Taking expectations yields $\mathbb{E}[|\eta_x(D)|] \leq 2\sqrt{2}\Delta\mathbb{E}\left[\sqrt{n_x(D)}\right]$. Then, averaging over the possibilities of x gives

$$\mathbb{E}[\text{Regret}_Q(\hat{\pi}, \tau; c)] \geq \Delta\mathbb{E}[\overline{\alpha}(D)] - \frac{2\sqrt{2}\Delta^2}{d}\sum_{x=1}^{d}\mathbb{E}\left[\sqrt{n_x(D)}\right]$$

where $\overline{\alpha}(D) = \frac{1}{d}\sum_{x=1}^{d}\alpha_x(D)$. Jensen's inequality on the second expectation then yields

$$\mathbb{E}[\text{Regret}_Q(\hat{\pi}, \tau; c)] \geq \Delta\mathbb{E}[\overline{\alpha}(D)] - 2\sqrt{2}\Delta^2\sqrt{\frac{m}{N}}.$$

We now bound the other expectation. Since P is uniform, $\mathbb{E}_{x\sim P}[1 - \alpha_x(D)] = 1 - \frac{1}{N}\sum_{x=1}^{N}\alpha_x(D)$. Thus, if the learner satisfies $\mathbb{E}[\mathbb{E}_{x\sim P}[1 - \alpha_x(D)]] \leq \epsilon$, then $\frac{1}{N}\sum_{x=1}^{N}\mathbb{E}[\alpha_x(D)] \geq 1 - \epsilon$. Even if the learner accepts with probability 1 on all states outside S_Q, this still forces $\mathbb{E}[\bar{\alpha}(D)] \geq 1 - \frac{N\epsilon}{d}$.

Now, setting $N = \frac{d}{2\epsilon}$ and $\Delta = 4\epsilon$ gives us the bounds

$$E[\text{Regret}_Q(\hat{\pi}, \tau; c)] \geq 2\epsilon - 32\sqrt{2}\epsilon^2\sqrt{\frac{2\epsilon m}{d}}.$$

Requiring this expected regret to be at most ϵ yields:

$$1 \leq 32\sqrt{2}\epsilon\sqrt{\frac{2\epsilon m}{d}} \implies m \geq \frac{d}{4096\epsilon^3}.$$

Setting $c_0 = 1/4096$ completes the proof.

D Proofs from the Misspecification Section

Notation. For a base policy $\pi_0 \in \Pi$, comparator $\pi \in \Pi$, and validator sequence Φ, define the expert-side late-stop risk

$$r_N(\pi_0, \pi, \Phi) \coloneqq \Pr_{T\sim P_N^\star}\left[\tau_{\pi_0,\Phi}(T) > \tau_{\pi_0,\{\pi\}}(T)\right],$$

with empirical analogue

$$\widehat{r}_N(\pi_0, \pi, \Phi) \coloneqq \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}\left[\tau_{\pi_0,\Phi}(T_j) > \tau_{\pi_0,\{\pi\}}(T_j)\right],$$

where $T_1, \ldots, T_n$ are the unlabeled expert test trajectories. This is the expert-side late-stop quantity that appears in the equilibrium lemma and in the generalization argument.

D.1 Proof of Lemma 3

Define a zero-sum game in which the maximizer chooses $p \in \Delta(\Pi)$ and the minimizer chooses $q \in \Delta(\Pi^K)$, with payoff

$$U(p,q) \coloneqq \mathbb{E}_{\pi\sim p}\,\mathbb{E}_{\Phi\sim q}\left[\widehat{r}_N(\pi_0,\pi,\Phi) - \Lambda\cdot\widehat{d}_M(\pi) + \frac{\Lambda}{K}\sum_{k=1}^{K}\widehat{d}_M(\phi_k)\right].$$

Since U is bilinear, von Neumann's minimax theorem implies $\min_q \max_p U(p,q) = \max_p \min_q U(p,q)$. To bound the maximin, fix any $p \in \Delta(\Pi)$ and let p^K denote the strategy of forming $\Phi = (\phi_1,\dots,\phi_K)$ by drawing $\phi_1,\dots,\phi_K$ i.i.d. from p. Then,

$$\begin{aligned}
\min_q\ U(p,q) \le U(p,p^K) &= \mathbb{E}_{\pi\sim p,\,\Phi\sim p^K}\left[\widehat{r}_N(\pi_0,\pi,\Phi)\right] - \Lambda\cdot\mathbb{E}_{\pi\sim p}\left[\widehat{d}_M(\pi)\right] + \frac{\Lambda}{K}\sum_{k=1}^{K}\mathbb{E}_{\phi_k\sim p}\left[\widehat{d}_M(\phi_k)\right] \\
&= \mathbb{E}_{\pi\sim p,\,\Phi\sim p^K}\left[\widehat{r}_N(\pi_0,\pi,\Phi)\right] && \text{(the other two terms cancel)} \\
&\le \frac{1}{K+1}. && \text{(same exchangeability argument as Lemma 5)}
\end{aligned}$$

Since this holds for every $p \in \Delta(\Pi)$, we conclude $\min_q \max_p U(p,q) = \max_p \min_q U(p,q) \le \frac{1}{K+1}$. Hence, there exists a minimizer strategy $q^\star \in \Delta(\Pi^K)$ such that $\max_p U(p,q^\star) \le \frac{1}{K+1}$. Because this bound holds against any p, it holds in particular for the pure strategies δ_π for any $\pi \in \Pi$.

To prove the first part of Lemma 3, take $p = \delta_{\pi_0}$ to be the pure strategy playing π_0. Since π_0 never deviates from itself, the late-stop risk against any committee is exactly zero ($\widehat{r}_N = 0$). Thus,

$$\begin{aligned}
U(\delta_{\pi_0},q^\star) = 0 - \Lambda\cdot\widehat{d}_M(\pi_0) + \frac{\Lambda}{K}\mathbb{E}_{\Phi\sim q^\star}\left[\sum_{k=1}^{K}\widehat{d}_M(\phi_k)\right] &\le \frac{1}{K+1} \\
\implies \mathbb{E}_{\Phi\sim q^\star}\left[\sum_{k=1}^{K}\widehat{d}_M(\phi_k)\right] &\le K\cdot\widehat{d}_M(\pi_0) + \frac{1}{\Lambda}.
\end{aligned}$$

To prove the second part of Lemma 3, take $p = \delta_\pi$ for an arbitrary target policy $\pi \in \Pi$. We have

$$\begin{aligned}
U(\delta_\pi,q^\star) = \mathbb{E}_{\Phi\sim q^\star}\left[\widehat{r}_N(\pi_0,\pi,\Phi)\right] - \Lambda\cdot\widehat{d}_M(\pi) + \frac{\Lambda}{K}\mathbb{E}_{\Phi\sim q^\star}\left[\sum_{k=1}^{K}\widehat{d}_M(\phi_k)\right] &\le \frac{1}{K+1} \\
\implies \mathbb{E}_{\Phi\sim q^\star}\left[\widehat{r}_N(\pi_0,\pi,\Phi)\right] &\le \Lambda\left(\widehat{d}_M(\pi) - \frac{1}{K}\mathbb{E}_{\Phi\sim q^\star}\left[\sum_{k=1}^{K}\widehat{d}_M(\phi_k)\right]\right) + \frac{1}{K+1} \\
\implies \mathbb{E}_{\Phi\sim q^\star}\left[\widehat{r}_N(\pi_0,\pi,\Phi)\right] &\le \Lambda\left(\widehat{d}_M(\pi) - \widehat{d}_M(\pi_0)\right) + \frac{1}{K}. && (\pi_0 \text{ minimizes } \widehat{d}_M)
\end{aligned}$$

D.2 Proof of Theorem 3

The proof closely follows the proof of Theorem 1 (Section B.4); we highlight the key differences and omit steps that are identical.

Existence of a sparse regularized validator distribution. For parameters Λ, K, let $q^\star$ be the distribution guaranteed by Lemma 3. Throughout, we let Λ, K be free and optimize them at the end.

Bounding the empirical risk. Unlike the deterministic non-misspecified setting—which draws k independent committees to suppress test risk and relies on the version space to prevent source abstention—here we bound the expected empirical metrics over a single regularized sequence $\Phi \sim q^\star$ (recall that this is done so as to not amplify the irreducible error).

For the source data, the policy abstains if any validator in Φ disagrees with π_0. Combining the union bound with the the first part of Lemma 3 bounds the expected empirical abstention:

$$\mathbb{E}_{\Phi\sim q^\star}\left[\widehat{\alpha}_M(\pi_0,\tau_{\pi_0,\Phi})\right] \le \mathbb{E}_{\Phi\sim q^\star}\left[\sum_{k=1}^{K}\widehat{d}_M(\phi_k)\right] \le K\cdot\widehat{d}_M(\pi_0) + \frac{1}{\Lambda}. \tag{11}$$

For the test data, evaluating against any comparator $\pi \in \Pi$, the second part of Lemma 3 directly bounds the expected empirical late-stop risk by the empirical suboptimality of π. In particular, taking our comparator to be $\tilde{\pi} = \arg\min_{\pi \in \Pi} \Delta_\pi$,

$$\mathbb{E}_{\Phi \sim q^\star}\left[\widehat{r}_N(\pi_0, \tilde{\pi}, \Phi)\right] \le \Lambda\Big(\widehat{d}_M(\tilde{\pi}) - \widehat{d}_M(\pi_0)\Big) + \frac{1}{K}. \tag{12}$$

Bounding the generalization error on the expert's distribution. We establish uniform convergence over all $|\Pi|^{K+1}$ possible combinations of base policy π_0 and validator set Φ, exactly as in the non-misspecified proof. Unlike that proof, the empirical source disagreements and target late-stop risks are no longer small due to misspecification, so we use standard Hoeffding bounds rather than fast-rate bounds. Let $Z \coloneqq (K+1)\log|\Pi| + \log(4/\delta)$. By Hoeffding's inequality and a union bound, with probability at least $1-\delta$, simultaneously for all (π_0, Φ):

$$\left|d_M(\pi_0) - \widehat{d}_M(\pi_0)\right| \le \sqrt{\frac{Z}{2m}}, \qquad \left|r_N(\pi_0, \tilde{\pi}, \Phi) - \widehat{r}_N(\pi_0, \tilde{\pi}, \Phi)\right| \le \sqrt{\frac{Z}{2n}}. \tag{13}$$

For the source data, the true stopping rate satisfies $\alpha_M(\pi_0, \tau_{\pi_0,\Phi}) \le \sum_{k=1}^K d_M(\phi_k)$ by a union bound. Taking the expectation over $\Phi \sim q^\star$ and applying (13) for each $\phi_k \in \Phi$,

$$\begin{aligned}
\mathbb{E}_{\Phi \sim q^\star}\left[\alpha_M(\pi_0, \tau_{\pi_0,\Phi})\right] &\le \mathbb{E}_{\Phi \sim q^\star}\left[\sum_{k=1}^K \widehat{d}_M(\phi_k)\right] + K\sqrt{\frac{Z}{2m}} && \text{(by (13))} \\
&\le K \cdot \widehat{d}_M(\pi_0) + \frac{1}{\Lambda} + K\sqrt{\frac{Z}{2m}} && \text{(by (11))} \\
&\le K \cdot \widehat{d}_M(\tilde{\pi}) + \frac{1}{\Lambda} + K\sqrt{\frac{Z}{2m}} && (\pi_0 \text{ minimizes } \widehat{d}_M) \\
&\le K\Delta + \frac{1}{\Lambda} + 2K\sqrt{\frac{Z}{2m}}. && \text{(by (13) and definition of } \Delta)
\end{aligned}$$

For the test data,

$$\begin{aligned}
\mathbb{E}_{\Phi \sim q^\star}\left[r_N(\pi_0, \tilde{\pi}, \Phi)\right] &\le \mathbb{E}_{\Phi \sim q^\star}\left[\widehat{r}_N(\pi_0, \tilde{\pi}, \Phi)\right] + \sqrt{\frac{Z}{2n}} && \text{(by (13))} \\
&\le \Lambda\Big(\widehat{d}_M(\tilde{\pi}) - \widehat{d}_M(\pi_0)\Big) + \frac{1}{K} + \sqrt{\frac{Z}{2n}} && \text{(by (12))} \\
&\le \Lambda\Big(d_M(\tilde{\pi}) - d_M(\pi_0)\Big) + \frac{1}{K} + 2\Lambda\sqrt{\frac{Z}{2m}} + \sqrt{\frac{Z}{2n}} && \text{(by (13))} \\
&\le \Lambda\Delta + \frac{1}{K} + 2\Lambda\sqrt{\frac{Z}{2m}} + \sqrt{\frac{Z}{2n}}, && \text{(definition of } \Delta, \text{ and } -d_M(\pi_0) \le 0)
\end{aligned}$$

where the last step uses $d_M(\tilde{\pi}) - d_M(\pi_0) \le \Delta_{\tilde{\pi}} \le \Delta$.

Bounding the stopping rate and regret on the learned policy's distribution. For the source data, we apply the same union bound and prefix coupling argument as in the proof of Theorem 1. The one difference is that π_0 may now deviate from $\pi^\star$, contributing an additional $d_M(\pi_0)$ term. Since $d_M(\pi_0) \le d_M(\tilde{\pi}) + 2\sqrt{Z/2m} \le \Delta + 2\sqrt{Z/2m}$ by (13) and the empirical minimality of π_0, taking the expectation over $q^\star$ yields:

$$\begin{aligned}
\mathbb{E}_{\Phi \sim q^\star}\left[\alpha_M(\pi_0, \tau_{\pi_0,\Phi})\right] &\le \mathbb{E}_{\Phi \sim q^\star}\left[\alpha_M(\pi^\star, \tau_{\pi_0,\Phi})\right] + d_M(\pi_0) \\
&\le \left(K\Delta + \frac{1}{\Lambda} + 2K\sqrt{\frac{Z}{2m}}\right) + \Delta + 2\sqrt{\frac{Z}{2m}} \\
&= (K+1)\Delta + \frac{1}{\Lambda} + 2(K+1)\sqrt{\frac{Z}{2m}}.
\end{aligned}$$

For the target data, let $B \coloneqq \{\tau_{\pi_0,\{\pi^\star\}} < \tau_{\pi_0,\Phi}\}$ be the event that the deployed policy π_0 deviates from the true expert $\pi^\star$ before the validator set stops execution. As in the realizable case, the event B is $\mathcal{G}_{h^\star}$-measurable with $h^\star = \min(\tau_{\pi_0,\Phi}, \tau_{\pi_0,\{\pi^\star\}})$. Lemma 4 therefore gives $\Pr_{N,\pi_0}[B] = \Pr_{N,\pi^\star}[B]$.

To bound this probability on the expert's distribution, we compare $\pi^\star$ to the best-in-class policy $\tilde{\pi}$. On any expert trajectory, if π_0 deviates from $\pi^\star$ before Φ triggers (event B), then either (1) $\tilde{\pi}$ deviates from $\pi^\star$ at some point during the trajectory, or (2) $\tilde{\pi}$ perfectly matches $\pi^\star$ everywhere, meaning the deviation between π_0 and $\pi^\star$ is simultaneously a deviation between π_0 and $\tilde{\pi}$, and this deviation occurs before Φ triggers. This implies the set containment $B \subseteq \{\tau_{\tilde{\pi},\{\pi^\star\}} \leq H\} \cup \{\tau_{\pi_0,\Phi} > \tau_{\pi_0,\{\tilde{\pi}\}}\}$ on expert trajectories. Applying a union bound, we have $\Pr_{N,\pi^\star}[B] \leq d_N(\tilde{\pi}) + r_N(\pi_0, \tilde{\pi}, \Phi)$.

As in the proof of Theorem 1, the stopped regret is bounded by $C_{\max}$ times the probability of an unrejected deviation from the expert, so taking the expectation over $q^\star$ yields:

$$\begin{aligned}
\mathbb{E}_{\Phi\sim q^\star}\left[\mathrm{Regret}_N(\pi_0, \tau_{\pi_0,\Phi}; c)\right] &\leq C_{\max} \cdot \mathbb{E}_{\Phi\sim q^\star}\left[\Pr_{N,\pi_0}[B]\right] \\
&= C_{\max} \cdot \mathbb{E}_{\Phi\sim q^\star}\left[\Pr_{N,\pi^\star}[B]\right] \\
&\leq C_{\max}\left(d_N(\tilde{\pi}) + \mathbb{E}_{\Phi\sim q^\star}\left[r_N(\pi_0, \tilde{\pi}, \Phi)\right]\right) \\
&\leq C_{\max}\left(\Delta + \Lambda\Delta + \frac{1}{K} + 2\Lambda\sqrt{\frac{Z}{2m}} + \sqrt{\frac{Z}{2n}}\right).
\end{aligned}$$

Setting $\Lambda = K = \Theta(\Delta^{-1/2})$ balances the Δ-dependent and Δ-independent terms in both bounds, yielding the rates in Theorem 3.

D.3 Proof of Corollary 4

Let $c_0 \coloneqq \log|\Pi| + \log(4/\delta)$. We define the $\epsilon^\star \coloneqq \max\left\{(c_0/m)^{1/5}, (c_0/n)^{1/3}\right\}$ and set $K = \Lambda = \lceil(\Delta^{1/2} + \epsilon^\star)^{-1}\rceil$. Substituting this into Theorem 3,

$$K\Delta + \frac{1}{K} \leq \frac{\Delta}{\Delta^{1/2}} + (\Delta^{1/2} + \epsilon^\star) = 2\Delta^{1/2} + \epsilon^\star.$$

For the sample estimation penalties, noting $C \leq (K+1)c_0$ and $K \leq (\epsilon^\star)^{-1}$, the labeled penalty is bounded by $(\epsilon^\star)^{-3/2}\sqrt{c_0/m}$. By definition, $\epsilon^\star \geq (c_0/m)^{1/5}$, which implies $\sqrt{c_0/m} \leq (\epsilon^\star)^{5/2}$. Thus, the labeled penalty is at most $(\epsilon^\star)^{-3/2}(\epsilon^\star)^{5/2} = \epsilon^\star$. An identical balancing argument bounds the unlabeled penalty: $K^{1/2}\sqrt{c_0/n} \leq (\epsilon^\star)^{-1/2}(\epsilon^\star)^{3/2} = \epsilon^\star$.

Because all terms in the abstention and regret bounds of Theorem 3 are upper-bounded by $O(\Delta^{1/2} + \epsilon^\star)$, expanding $\epsilon^\star$ yields the claimed result.

D.4 Off-policy test trajectories under misspecification

So far, we have assumed that the expert π^* is the same across train and test environments. In practice, however, train and test demonstrations often come from different sources—for example, using one expert to pilot a simulated car, while using dashcam aggregated from multiple drivers. This creates a challenging off-policy gap where the training behavior π^{tr} and test behavior π^{te} differ, and neither necessarily lies in the policy class Π. Fortunately, the same idea we used for misspecified experts—finding a validator distribution via a *regularized* game, which induces a stopping time—extends readily to policy shift between environments.

Concretely, suppose $\mathcal{D}_{\mathsf{train}}$ is generated by a deterministic policy π^{tr}, while $\mathcal{D}_{\mathsf{test}}$ is generated by a different deterministic policy π^{te}, with neither policy necessarily in Π. Let d_M^{tr} and d_N^{te} denote the corresponding disagreement probabilities in (M, π^{tr}) and (N, π^{te}), respectively, and let $\mathrm{Regret}_N^{\mathsf{te}}$ denote regret against π^{te} in the test environment. Then, the agnostic SeqRejectron satisfies:

Proposition 4. *There exists an algorithm, the agnostic* SeqRejectron *(which we described in Section 5), with hyperparameters $K, \Lambda, \delta > 0$, with the following guarantee. Let $\Delta_{\mathsf{off}} \coloneqq \min_{\bar{\pi}\in\Pi}\max\{d_M^{\mathsf{tr}}(\bar{\pi}), d_N^{\mathsf{te}}(\bar{\pi})\}$.*[11] *When run with a specific choice of parameters K, Λ (depending on $m, n, \Delta_{\mathsf{off}}, |\Pi|$, and δ), then with probability at least $1-\delta$, the agnostic* SeqRejectron *will return a pair $(\pi_0, q^\star)$ such that both $\mathbb{E}_{\Phi\sim q^\star}[\alpha_M(\pi_0, \tau_{\pi_0,\Phi})]$ and $\mathbb{E}_{\Phi\sim q^\star}[\mathrm{Regret}_N^{\mathsf{te}}(\pi_0, \tau_{\pi_0,\Phi}; c)]/C_{\max}$ are $\widetilde{O}\left(\Delta_{\mathsf{off}}^{1/2} + (\log|\Pi|/m)^{1/5} + (\log|\Pi|/n)^{1/3}\right)$.*

[11] Note that as long as there exists some policy $\pi \in \Pi$ which is able to behave similarly to π^{tr} on M and to π^{te} on N, then the irreducible error Δ_{off} will be small.

D.5 Proof of Proposition 4

The proof is a direct adaptation of the proof of Theorem 3. We first redefine the suboptimality gaps and their empirical estimators with respect to the off-policy behavior policies π^{tr} and π^{te} instead of the expert $\pi^\star$:

$$d_M^{\mathsf{tr}}(\bar{\pi}) \coloneqq \Pr_{M,\pi^{\mathsf{tr}}}\left[\tau_{\bar{\pi},\{\pi^{\mathsf{tr}}\}} \le H\right], \qquad d_N^{\mathsf{te}}(\bar{\pi}) \coloneqq \Pr_{N,\pi^{\mathsf{te}}}\left[\tau_{\bar{\pi},\{\pi^{\mathsf{te}}\}} \le H\right]$$

$$\widehat{d}_M^{\mathsf{tr}}(\pi) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left[\tau_{\pi,\{\pi^{\mathsf{tr}}\}}(S_i) \le H\right], \qquad \widehat{r}_N^{\mathsf{te}}(\pi_0,\pi,\Phi) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left[\tau_{\pi_0,\Phi}(T_j) > \tau_{\pi_0,\{\pi\}}(T_j)\right].$$

Let $\pi_0 \in \arg\min_{\pi\in\Pi} \widehat{d}_M^{\mathsf{tr}}(\pi)$. Applying the symmetrically regularized equilibrium (Lemma 3) to these modified losses guarantees the existence of a distribution $q^\star$ such that:

$$\mathbb{E}_{\Phi\sim q^\star}\left[\sum_{k=1}^{K} \widehat{d}_M^{\mathsf{tr}}(\phi_k)\right] \le K\widehat{d}_M^{\mathsf{tr}}(\pi_0) + \frac{1}{\Lambda},$$

$$\mathbb{E}_{\Phi\sim q^\star}\left[\widehat{r}_N^{\mathsf{te}}(\pi_0,\pi,\Phi)\right] \le \Lambda\big(\widehat{d}_M^{\mathsf{tr}}(\pi) - \widehat{d}_M^{\mathsf{tr}}(\pi_0)\big) + \frac{1}{K} \qquad \text{for every } \pi\in\Pi.$$

From here, the mechanics follow exactly as before. Applying Hoeffding bounds introduces the $O(\sqrt{C/m})$ and $O(\sqrt{C/n})$ concentration penalties to these empirical terms. The prefix-coupling argument bounds the true completeness α_M by the sum of training disagreements, and the test regret by $C_{\max}$ times the late-stop risk against π^{te}.

Finally, to bound the regret against the training demonstrator π^{tr} in the test environment, coupling π^{tr} to $\bar{\pi}$ and then $\bar{\pi}$ to π^{te} naturally introduces the discrepancy terms $d_N^{\mathsf{tr}}(\bar{\pi})$ and $d_N^{\mathsf{te}}(\bar{\pi})$ via the union bound, yielding the secondary result.

Finally, to obtain the optimal rates claimed in the proposition, we evaluate these bounds at the comparator $\bar{\pi}_{\mathrm{off}} \in \arg\min_{\bar{\pi}\in\Pi} \max\{d_M^{\mathsf{tr}}(\bar{\pi}), d_N^{\mathsf{te}}(\bar{\pi})\}$. Setting $K = \Lambda = \left\lceil (\Delta_{\mathrm{off}}^{1/2} + \epsilon^\star)^{-1} \right\rceil$ with $\epsilon^\star \coloneqq \max\{(\log|\Pi|/m)^{1/5}, (\log|\Pi|/n)^{1/3}\}$ and applying the identical balancing argument from the proof of Corollary 4 yields the final rate.