IN-CONTEXT PURE EXPLORATION IN CONTINUOUS DECISION SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we develop a practical and reusable framework for active sequential testing, also termed *pure exploration*, with continuous recommendations. In this setting, a learner adaptively acquires information to identify an unknown hypothesis with as few queries as possible while satisfying a prescribed prior-averaged accuracy level. Existing methods are predominantly problem-specific, while learned approaches have been limited to finite recommendation spaces. We formulate this objective as Bayesian fixed-confidence pure exploration and characterize the Bellman structure of an ideal learner. Guided by this characterization, we introduce `C-ICPE`, which meta-trains sequential neural architectures over a task prior to jointly learn exploration, stopping, and recommendation. At inference, `C-ICPE` gathers evidence and returns a recommendation without parameter updates. We conclude by analyzing the method with experiments across localization, identification and value estimation problems.

1 INTRODUCTION

Several learning problems are inherently interactive: the learner sequentially performs interventions, observes noisy evidence whose distribution depends on that intervention, and stops once the accumulated evidence supports a reliable conclusion. This type of interactive sequential decision-making problem goes back to Chernoff (1959) and has been formalized through active sequential hypothesis testing (Naghshvar and Javidi, 2013) and pure exploration with fixed confidence (Audibert and Bubeck, 2010; Degenne et al., 2019), where the learner minimizes the number of queries subject to returning an ϵ-accurate recommendation with probability at least $1 - \delta$.

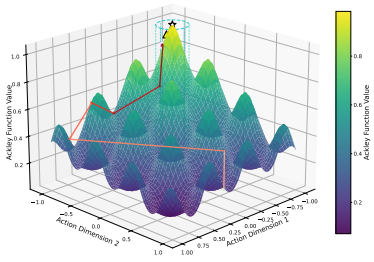


Figure 1: `C-ICPE` sequentially queries a noisy, randomly parametrized inverted Ackley function and localizes the global maximum using an adaptive number of queries.

Fixed-confidence exploration is well understood for finite decision spaces, including best-arm identification (Garivier and Kaufmann, 2016; Degenne et al., 2019; Poiani et al., 2025b) and best-policy identification in finite Markov decision processes (Al Marjani et al., 2021; Taupin et al., 2023). Despite this progress, practical methods remain limited for continuous recommendations. Existing approaches are frequentist and model-specific (Garivier and Kaufmann, 2021; Takemori et al., 2025; Poiani et al., 2025a) or Gaussian-process (GP) based, without optimizing sample-complexity (Wilson, 2024). Even in finite arms, Bayesian fixed-confidence results focus on Gaussian likelihoods and priors (Jang et al., 2024).

Recently, Russo et al. (2026) proposed In-Context Pure Exploration (ICPE), which meta-trains sequential neural policies for finite active-testing problems. We extend this approach to continuous recommendations. The learner must determine which recommendations remain plausible and whether another query is worth its cost. We introduce `C-ICPE`, a practical method that learns to collect data, stop, and recommend directly from trajectories using the same training procedure across task families. This is motivated by applications such as materials discovery (Liu et al., 2017) and dose-finding (O'Quigley et al., 1990), where each trial is costly, observations are noisy, and the objective is to control the prior-averaged probability of returning an ϵ-incorrect recommendation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Contributions. We formulate Bayesian fixed-confidence pure exploration with continuous recommendations and establish its Bellman structure. We prove zero duality gap and recover a correct, sample-optimal policy under conditions that allow non-unique dual minimizers. Then, we introduce C-ICPE, extending the meta-training protocol of (Russo et al., 2026) with Gaussian or mixture inference and continuous exploration rules. We analyze how these design choices affect sample complexity using a multi-modal two-lobe problem, quantifying the cost of Gaussian sampling under approximate learning. We also provide a finite-sample certification procedure for trained learners, and bound the relative sample-suboptimality of an approximate learner in terms of the model error. We also establish a separation between maximum-value estimation and argmax localization under an RBF-GP prior. Finally, we evaluate C-ICPE on localization, continuous-arm identification, value estimation and geochemical surrogate tasks.

2 Problem Setting

We consider a Bayesian family of active sequential decision problems indexed by a latent parameter $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^d$ is compact and $\theta \sim \nu$. Each environment M_θ specifies an initial observation law $\rho_\theta \in \Delta(\mathcal{Y})$ and observation kernels $P_{\theta,t}(\cdot|h_t, a_t)$ over a compact observation space $\mathcal{Y} \subset \mathbb{R}^n$. In sequential testing, the learner interacts with M_θ over time: in round t it observes the history

$$H_t := (Y_1, A_1, Y_2, \dots, A_{t-1}, Y_t),$$

chooses a query $A_t \in \mathcal{A} \subset \mathbb{R}^m$ (compact), and observes $Y_{t+1} \sim P_{\theta,t}(\cdot|H_t, A_t)$. The learner uses this history t to return a recommendation $\hat{x} \in \mathcal{X}$, where $\mathcal{X}$ is compact. We distinguish the query space $\mathcal{A}$, used to collect information, from the recommendation space $\mathcal{X}$. Often $\mathcal{X} = \mathcal{A}$, but in value-identification tasks the learner queries locations and returns a value.

Risk function. Recommendation quality is measured by a task-dependent loss, or risk, function $L_\theta : \mathcal{X} \to [0, \infty)$, satisfying $\inf_{x\in\mathcal{X}} L_\theta(x) = 0$. We say that x is ϵ-optimal for task θ if $L_\theta(x) \le \epsilon$, and define

$$\mathcal{X}_\epsilon(\theta) := \{x \in \mathcal{X} : L_\theta(x) \le \epsilon\}.$$

In the following, we assume that $(\theta, x) \mapsto L_\theta(x)$ is jointly lower semicontinuous (in Sec. C we state the regularity assumptions used in the theoretical results).

For function minimization with $\mathcal{X} = \mathcal{A}$, one may use the value-gap loss $L_\theta(x) := f_\theta(x) - \min_{x'\in\mathcal{X}} f_\theta(x')$, or the localization loss $L_\theta(x) := \|x^\star_\theta - x\|$ around a selected optimizer $x^\star_\theta \in \arg\min_{x'\in\mathcal{X}} f_\theta(x')$. Other examples include best-arm identification, where $f_\theta(x)$ is the mean reward, and noisy binary search, where signed comparisons reveal information about a latent threshold. The recommendation need not be a query: in optimal-value identification, $\mathcal{X} \subset \mathbb{R}$ and $L_\theta(x) := |x - \max_{a\in\mathcal{A}} f_\theta(a)|$. The regularity conditions used in the ideal analysis and their relation to the benchmarks used in the experiments are discussed in Sec. C.1 (remark 1).

Bayesian fixed-confidence objective. A learner is defined by the triplet (π, I, τ): a possibly randomized sampling policy $\pi = (\pi_t)_{t\ge1}$ such that $A_t \sim \pi_t(\cdot \mid H_t)$; a stopping time τ adapted to the observed history and the learner's internal randomization, which is independent of the task and observation noise; an inference rule $I = (I_t)_{t\ge1}$ such that $\hat{x}_\tau = I_\tau(H_\tau)$. For $\epsilon > 0, \delta \in (0, 1/2)$, we minimize the expected number of queries while ensuring correctness

$$\inf_{\tau,\pi,I} \mathbb{E}^\pi_{\theta\sim\nu}[\tau] \quad \text{s.t.} \quad \mathbb{P}^\pi_{\theta\sim\nu}(\hat{x}_\tau \in \mathcal{X}_\epsilon(\theta)) \ge 1 - \delta, \quad \mathbb{E}^\pi_{\theta\sim\nu}[\tau] < \infty. \tag{1}$$

Our formulation is Bayesian: correctness is averaged over tasks drawn from ν. The prior also provides the task distribution for meta-training, allowing the learner to reuse experience across new tasks from the same family, without parameter updates. In applications, ν encodes knowledge about plausible environments. It determines the posterior success probability that we use next to characterize the ideal learner.

3 Theoretical Background

We characterize an ideal Bayesian learner, which motivates the C-ICPE design. Relative to finite ICPE (Russo et al., 2026), in continuous spaces attainment of the Bellman equation over actions is non-trivial, and requires a semicontinuity chain from the observation model to the Q-function. The dual analysis also allows non-unique minimizers through a local closedness condition.

Posterior success and optimal inference. In the fixed-confidence setting with continuous $\mathcal{X}$, the relevant posterior object is the posterior probability that x is ϵ-optimal:

$$q_t(h, x) \coloneqq \mathbb{P}\left(L_\theta(x) \le \epsilon \mid H_t = h\right), \qquad r_t(h) \coloneqq \max_{x \in \mathcal{X}} q_t(h, x).$$

Here $q_t(h, x)$ measures the reliability of a particular recommendation, $r_t(h)$ is the best reliability achievable from history h. We show in Prop. 3 that the maximum is attained and admits a recommendation rule

$$I_t^\star(h) \in \arg\max_{x \in \mathcal{X}} q_t(h, x). \tag{2}$$

Dual formulation and Bellman optimality. We study the fixed-confidence problem in Eq. (1) through its Lagrangian dual, following the ASHT literature (Naghshvar and Javidi, 2013; Russo et al., 2026). Introducing a multiplier $\lambda \ge 0$ for the correctness constraint gives

$$V_\lambda(\pi, I, \tau) \coloneqq -\mathbb{E}^\pi_{\theta\sim\nu}[\tau] + \lambda\left(\mathbb{P}^\pi_{\theta\sim\nu}\left(I_\tau(H_\tau) \in \mathcal{X}_\epsilon(\theta)\right) - 1 + \delta\right), \qquad \inf_{\lambda\ge 0} \sup_{\pi, I, \tau} V_\lambda(\pi, I, \tau). \tag{3}$$

For fixed π and τ, optimizing over I replaces the terminal success probability by $\mathbb{E}^\pi[r_\tau(H_\tau)]$. Hence, for fixed λ, the inner problem is equivalent up to a constant to maximizing $\mathbb{E}^\pi[\lambda r_\tau(H_\tau) - \tau]$. Stopping can also be embedded as an absorbing action a_{stop}: the learner continues with actions in $\mathcal{A}$ until it selects a_{stop}, at which point it stops and outputs $I_t^\star(H_t)$; see Lem. 6. Thus the fixed-λ problem is an optimal-stopping control problem on $\bar{\mathcal{A}} = \mathcal{A} \cup \{a_{\text{stop}}\}$.

For $t \ge 1$ and $h \in \mathcal{H}_t$, define the optimal reward-to-collect value and the Q-values

$$V_t^\star(h; \lambda) \coloneqq \sup_{\bar\pi} \mathbb{E}^{\bar\pi}_{\theta\sim\nu}\left[\lambda r_{\bar\tau}(H_{\bar\tau}) - (\bar\tau - t) \mid H_t = h\right], \tag{4}$$

$$Q^\star_{t,\text{stop}}(h; \lambda) \coloneqq \lambda r_t(h), \qquad Q^\star_{t,\text{cont}}(h, a; \lambda) \coloneqq -1 + \mathbb{E}\left[V^\star_{t+1}(H_{t+1}; \lambda) \mid H_t = h, A_t = a\right].$$

We have the following characterization of the optimal policy for a fixed λ.

Theorem 3.1 (Bellman equation and greedy deterministic optimality)**.** *Assume the regularity conditions of Sec. C.1 (assumptions 2 to 4), then*

$$V_t^\star(h; \lambda) = \max\left\{Q^\star_{t,\text{stop}}(h; \lambda), \sup_{a\in\mathcal{A}} Q^\star_{t,\text{cont}}(h, a; \lambda)\right\}. \tag{5}$$

Moreover, let $a_t^\star(h) \in \arg\max_{a\in\mathcal{A}} Q^\star_{t,\text{cont}}(h, a; \lambda)$. The deterministic policy $\bar\pi^\star$ defined by

$$\bar\pi^\star(h) = a_{\text{stop}} \quad \text{if } Q^\star_{t,\text{stop}}(h; \lambda) \ge Q^\star_{t,\text{cont}}(h, a_t^\star(h); \lambda),$$

and $\bar\pi^\star(h) = a_t^\star(h)$ otherwise, is optimal for the fixed-λ inner problem.

This result gives the desired stop/continue comparison: stop when the recommendation is worth more than the best available query. Attainment of that query, unlike the finite case, requires upper semicontinuity of the reward-to-go formulation on compact $\mathcal{A}$. We prove this theorem in App. C.3.3.

Zero duality gap and Bayesian (ϵ, δ)-correctness. Under two assumptions, time-sharing and strict feasibility, the dual recovers the constrained value of the primal problem. Local closedness additionally yields a primal-optimal policy among the dual minimizers without requiring uniqueness; see Thm. C.2. This permits non-unique multipliers and policies, unlike the uniqueness condition in (Russo et al., 2026). We next use this structure to construct `C-ICPE`, an approximate learner, and analyze its approximate sample complexity.

4 Continuous ICPE: C-ICPE

We now describe `C-ICPE`, a practical method based on the theory of the previous section. It combines an inference network for recommendation, a critic for stopping and an actor rule for exploration. Following (Lee et al., 2023; Russo et al., 2026), we meta-train on simulated tasks from ν, storing partial trajectories and task information in a replay buffer $\mathcal{B}$. Once trained, the model is deployed at test-time on new tasks without parameter updates. We describe the components below; Algorithm 1 summarizes the procedure and Sec. D gives implementation details.

Algorithm 1 C-ICPE

// Training phase
while Training is not over **do**
  Sample environment $M_\theta \sim \nu$ and hypothesis $x^\star_\theta$; observe $Y_1 \sim \rho$ and set $t \leftarrow 1$.
  repeat
    Propose continuation actions A_t using an actor (e.g., TS, TTPS, or TD3).
    If $Q_\psi(H_t, a_{\text{stop}}) \geq Q_\psi(H_t, A_t)$ store the trajectory in $\mathcal{B}$ and break; otherwise execute A_t and observe Y_{t+1}.
    Add partial trajectory $(H_t, A_t, Y_{t+1}, x^\star_\theta)$ to $\mathcal{B}$ and set $t \leftarrow t+1$.
  until $t > T_{\max}$ (max horizon).
  In the fixed confidence, update c according to Eq. (11).
  Sample batch $B \sim \mathcal{B}$ and update models using $\mathcal{L}_{\text{inf}}(B;\phi)$ (Eq. (8)) and $\mathcal{L}_{\text{critic}}(B;\psi)$ (Eq. (10)); for TD3, train π according to $\mathcal{L}_{\text{act}}(B;\psi)$ (Eq. (49)).
end while

// Inference/Deployment phase (models are frozen here)
Sample unknown environment $M \sim \nu$ and collect a trajectory H_τ using π (until $Q_\psi(H_t, A_t) \leq Q_\psi(H_t, a_{\text{stop}})$).
Return $\hat{x}_\tau = \hat{x}_\phi(H_\tau)$ (recommendation).

4.1 Inference Model and Recommendation

The ideal inference rule maximizes $q_t(h, \cdot)$, but learning and maximizing this function over continuous recommendation is costly and non-trivial. Following ICPE and DPT (Russo et al., 2026; Lee et al., 2023), we instead assume access to a fixed measurable simulator-provided supervised target $x^\star_\theta \in \mathcal{X}_\epsilon(\theta)$ for ν-almost every task. Its conditional law $\mathbb{P}(x^\star_\theta \mid H_t = h)$ captures uncertainty about this target given the history. Thus, the model learns the conditional law of a selected target.

Inference model and recommendation. We approximate this conditional law using a mixture of J diagonal Gaussians,

$$I_\phi(\cdot \mid h) = \sum_{j=1}^{J} \omega_{\phi,j}(h)\mathcal{N}\left(\mu_{\phi,j}(h), \text{diag}(\sigma^2_{\phi,j}(h))\right), \qquad \omega_{\phi,j}(h) \geq 0, \quad \sum_j \omega_{\phi,j}(h) = 1. \tag{6}$$

where ϕ denotes the network parameters. At stopping, we return the mean of the highest-weight component,

$$\hat{x}_\phi(h) = \text{Proj}_{\mathcal{X}} \mu_{\phi, j_\phi(h)}(h), \quad \text{where} \quad j_\phi(h) = \arg\max_{1 \leq j \leq J} \omega_{\phi,j}(h). \tag{7}$$

We train ϕ using a log-likelihood loss on a batch of partial trajectories $B = (x^\star_i, H_{t_i})_i$ sampled from $\mathcal{B}$, where $x^\star_i$ is the target for task i:

$$\mathcal{L}_{\text{inf}}(B;\phi) = -\sum_{i=1}^{|B|} \log I_\phi(x^\star_i \mid H_{t_i}), \tag{8}$$

and also denote by $\bar{\phi}$ the target inference parameters, which is later used by the critic and actor models, and it is updated via a soft update $\bar{\phi} \leftarrow (1-\tau_I)\bar{\phi} + \tau_I \phi$ with time constant $\tau_I \in (0,1)$.

Choice of the inference family. Under ordinary population NLL, the optimal single Gaussian matches the conditional mean and coordinate variances of $x^\star_\theta$. Even exact moments can place the mean between plausible targets, requiring further evidence before returning it, whereas a mixture can recommend one component without averaging across alternatives. When the remaining uncertainty is small relative to a neighborhood of correct recommendations, the mean can already be reliable: we analyze this in the appendix, see Eq. (26), where we provide a bound on $\mathbb{P}(L_\theta(\widehat{x}_\phi(H_\tau)) > \epsilon)$. Lastly, the multi-lobe problem in Sec. 4.3 studies this case.

4.2 Critic and Actor Models

Critic. The critic learns whether another query is worth its cost. Following Thm. 3.1, we use a stopping head $Q_\psi(h, a_{\text{stop}})$ and a continuation head $Q_\psi(h, a), a \in \mathcal{A}$, parametrized by ψ. For training, the actor rule supplies a target $a_{\text{tgt}}(h')$ at the next-history h', giving

$$V_{\bar{\psi}}(h') = \max\{Q_{\bar{\psi}}(h', a_{\text{stop}}), Q_{\bar{\psi}}(h', a_{\text{tgt}}(h'))\}.$$

The stopping head evaluates the recommendation that would be returned at stopping. Using the target inference network, we train the stopping head on recommendation success

$$\hat{r}(h, \theta) = \mathbf{1}\left\{L_\theta(\hat{x}_{\bar{\phi}}(h)) \le \epsilon\right\}. \tag{9}$$

Conditionally on h, we have $\mathbb{E}[\hat{r}(H_t, \theta) \mid H_t = h] = q_t(h, \hat{x}_{\bar{\phi}}(h)) \le r_t(h)$. Thus, regression against this reward targets the success of the chosen recommendation, without explicitly maximizing q_t. In Sec. C.8 we discuss a sampled-reward alternative, which averages over inference draws $\hat{x} \sim I_{\bar{\phi}}$, which is generally more stable to variations of the inference model.

We train the critic by TD-Learning. For a batch of transitions $(h_i, a_i, h'_i, d_i, x^\star_i)_i$, let $d_i = 1$ when the query reaches the maximum horizon. We use targets

$$y_i^{\text{stop}} = \hat{r}(h_i, \theta_i), \qquad y_i^{\text{cont}} = -c + d_i \hat{r}(h'_i, \theta_i) + (1 - d_i) V_{\bar{\psi}}(h'_i),$$

where $c = 1/\lambda$ for $\lambda > 0$ is the per-query cost obtained by rescaling the objective. The critic loss is

$$\mathcal{L}_{\text{critic}}(B; \psi) = \frac{1}{2|B|} \sum_i \left[(Q_\psi(h_i, a_i) - y_i^{\text{cont}})^2 + (Q_\psi(h_i, a_{\text{stop}}) - y_i^{\text{stop}})^2\right]. \tag{10}$$

Terminal prefixes additionally supply stopping targets. Every executed query incurs c, including the last one. At rollout time, the policy first proposes A_t, and the learner stops if $Q_\psi(H_t, a_{\text{stop}}) \ge Q_\psi(H_t, A_t)$, and returns $\hat{x}_\phi(H_t)$; otherwise it executes A_t.

Policy and actors. The actor rule selects the next query. We consider three rules, depending on the relation between $\mathcal{X}$ and $\mathcal{A}$. When $\mathcal{X} = \mathcal{A}$ we can use the inference rule itself: the *Thompson Sampling (TS) rule* draws $Z_t \sim I_\phi(\cdot \mid H_t)$ and queries $A_t = \text{Proj}_{\mathcal{A}}(Z_t)$ (Thompson, 1933). This rule is useful when plausible targets are informative queries. Another similar rule that uses the inference model is the *Top Two Posterior Sampling (TTPS) rule* (Russo, 2016). It retains a Thompson sample with probability $1/2$ and otherwise considers a challenger farther from $\hat{x}_\phi(H_t)$, thus TTPS spends some probability mass checking alternative (the challenger condition is further detailed in Sec. D.5). Both rules use $a_{\text{tgt}}(h) = \hat{x}_{\bar{\phi}}(h)$, which provides a stable deterministic target.

While these rules do not directly learn an optimal actor, they are in general efficient, as we see later. Alternatively, TD3 (Fujimoto et al., 2018) learns an actor $\pi_\vartheta(h) \in \mathcal{A}$ by maximizing $Q_\psi(h, \pi_\vartheta(h))$. This includes the case $\mathcal{A} \neq \mathcal{X}$. This remains an approximation to the ideal continuation maximization. Its target actor, twin critics, smoothing, and optional imitation term are detailed in the appendix.

Cost update. Lastly, we update c using a dual update. On a fresh batch of trajectories, we estimate the accuracy rate $\hat{p} = \frac{1}{|B|}\sum_i \mathbf{1}\{L_{\theta_i}(\hat{x}_\phi(H_{\tau_i})) \le \epsilon\}$ and set

$$c \leftarrow \text{Proj}_{[0,1]}\left(c - \eta_c((1 - \delta_{\text{train}}) - \hat{p})\right). \tag{11}$$

where $\delta_{\text{train}} < \delta$ provides a margin to increase the statistical power of the test. If empirical success is below $1 - \delta_{\text{train}}$, the cost decreases and trajectories become longer; otherwise, if above target, the cost increases and stopping becomes more aggressive.

Approximation and sample complexity. We now quantify the sub-optimality of an approximate learner in terms of the errors in recommendation, exploration and stopping. Let $\Pi^\star$ be a constrained-optimal learner with stopping rule $\tau^\star$ and sample complexity $T^\star = \mathbb{E}_\nu^{\Pi^\star}[\tau^\star]$. Similarly, introduce an approximate learner $\hat{\Pi}$ satisfying the correctness constraint, with stopping $\hat{\tau}$, recommendation $\hat{x}_\phi$, critic Q_ψ and complexity $T = \mathbb{E}_\nu^{\hat{\Pi}}[\hat{\tau}]$. Define the recommendation error

$$e_I = \mathbb{E}^{\Pi^\star}[r_{\tau^\star}(H_{\tau^\star}) - q_{\tau^\star}(H_{\tau^\star}, \hat{x}_\phi(H_{\tau^\star}))].$$

This measures recommendation error only where the optimal learner stops. For the recommendation $\hat{x}_\phi(h)$, let $Q_t^\phi(h, a)$ be the optimal continuation value obtained by executing a and subsequently optimizing queries and stopping, while retaining the learned recommendation rule $\hat{x}_\phi(h)$. It satisfies

$$Q_t^\phi(h, a) = -c + \mathbb{E}[V_{t+1}^\phi(H_{t+1}) \mid h, a], \qquad V_t^\phi(h) = \max\{q_t(h, \hat{x}_\phi(h)), \sup_a Q_t^\phi(h, a)\},$$

where the expectation uses the true posterior predictive law. Define the errors

$$\begin{aligned} e_Q &:= \frac{1}{T}\mathbb{E}_\nu^{\hat{\Pi}}\left[\sum_{t=1}^{\hat{\tau}} |Q_\psi(H_t, a_{\text{stop}}) - q_t(H_t, \hat{x}_\phi(H_t))| + |Q_\psi(H_t, A_t) - Q_t^\phi(H_t, A_t)|\right], \\ e_A &:= \frac{1}{T}\mathbb{E}_\nu^{\hat{\Pi}}\left[\sum_{t=1}^{\hat{\tau}} V_t^\phi(H_t) - \max\{q_t(H_t, \hat{x}_\phi(H_t)), Q_t^\phi(H_t, A_t)\}\right]. \end{aligned} \tag{12}$$

The first error quantifies the critic's errors in evaluating stopping and continuing; the second one quantifies the error in stopping with the query $A_t \sim \hat{\Pi}$ rather than the best query. We then have the following result that shows the sub-optimality in sample complexity (the proof is in Sec. C.6).

Proposition 1 (Relative sample complexity)**.** *Denote by $p(\widehat{\Pi}) = 1 - \delta + e_p, e_p \geq 0$, the correctness of a learner $\hat{\Pi}$. Let $\lambda^\star$ be an optimal dual multiplier, $c^\star = 1/\lambda^\star$ and set $e_c := |c - c^\star|$. If $\lambda^\star(e_c + e_Q + e_A) < 1$, then*

$$\frac{\lambda^\star e_p}{T^\star} \leq \frac{T - T^\star}{T^\star} \leq \lambda^\star \frac{(e_Q + e_A) + (e_I + e_p)/T^\star}{1 - \lambda^\star(e_c + e_Q + e_A)}.$$

Thus, relative sample optimality follows when $\lambda^\star(e_c + e_Q + e_A) \to 0$ and $\lambda^\star(e_I + e_p)/T^\star \to 0$. Simultaneously, the lower bound shows that any accuracy surplus e_p has a sample complexity cost.

Correctness certification. Training `C-ICPE` does not automatically provide theoretical correctness guarantees. To certify correctness, we test the model on new independent tasks. We define a *checkpoint* m to be a saved version of the `C-ICPE` models (i.e., inference, actor and critic, all held fixed). We denote the true success probability of m by $p_m := \mathbb{P}^{\pi_m}_{\theta\sim\nu}(L_\theta(\hat{x}_m(H_{\tau_m})) \leq \epsilon)$.

We certify m only when its observed success count is sufficiently high to rule out the hypothesis $H_0 : p_m \leq 1 - \delta$ at some significance level α. To ensure that the power of this test is sufficiently high, we train at an accuracy $\delta_{\text{train}} < \delta$, as outlined in Sec. C.7. We provide an offline and online test: the former is performed at the end of training on some selected checkpoints, while the latter can be performed during training. In the appendix, we discuss both procedures in detail; for both procedures, one can guarantee the following correctness result.

Proposition 2 (Correctness certification)**.** *For either procedure above, let $R = \{\hat{m} \text{ certified}\}$ denote the event that a checkpoint $\hat{m}$ is certified and let $F = \{L_\theta(\hat{x}) > \epsilon\}$ denote the failure event of checkpoint $\hat{m}$ on a fresh task drawn from ν. Then*

$$\mathbb{P}\left(R \cap \{p_{\hat{m}} \leq 1 - \delta\}\right) \leq \alpha, \qquad \mathbb{P}(R \cap F) \leq \delta.$$

In other words, the first bound controls the type-I error rate (i.e., certifying a non-correct model), while the second bound concerns the probability of certifying a model and returning a sub-optimal recommendation in a fresh task. This procedure allows abstention, and does not imply $\mathbb{P}(F \mid R) \leq \delta$.

4.3 Multimodal posterior stress test

In this subsection we investigate how multimodality affects recommendations in a Gaussian and mixture inference using a controlled localization problem, the *multi-lobe environment*.

To fix the ideas, we consider the one-dimensional case, and let $\mathcal{A} = \mathcal{X} = [-1, 1]$. On this environment, we set the target to be $\theta = Sm + U$, where $S \sim \text{Unif}\{-1, +1\}, U \sim \text{Unif}[-w, w]$, thus θ has two modes. We use a risk loss $L_\theta(x) = |x - \theta|$ and assume $0 < w < \epsilon < m - w$ and $m + w < 1$. Each query $A_t = a$ produces an observation $Y_{t+1} \mid (S = s, H_t, A_t = a) \sim \text{Ber}(g_{s,\ell}(a))$, where $g_{s,\ell}(a) = \eta + (1 - 2\eta)e^{-(a-sm)^2/(2\ell^2)}$, $\eta \in (0, 1/2)$. In other words, the scale ℓ controls the width of informative regions around $\pm m$. Observations distinguish the lobes but do not reveal U; since $w < \epsilon$, identifying the lobe is sufficient.

We use this environment to stress-test `C-ICPE` since we can exactly evaluate the posterior distribution. Write $p_t = \mathbb{P}(S = +1 \mid H_t)$ and $I_\pm = [\pm m - w, \pm m + w]$. Then, the target posterior $P_t = \mathbb{P}(\theta \mid H_t)$, its single gaussian projection G_t and mixture M_t with two Gaussians are

$$\begin{aligned} P_t &= p_t \,\text{Unif}(I_+) + (1 - p_t)\,\text{Unif}(I_-), \\ G_t &= \mathcal{N}\big(\mu_t,\, w^2/3 + 4m^2 p_t(1 - p_t)\big), \qquad \mu_t = m(2p_t - 1), \\ M_t &= p_t \mathcal{N}(m, w^2/3) + (1 - p_t)\mathcal{N}(-m, w^2/3). \end{aligned} \tag{13}$$

Then, the posterior success $q_t(H_t, x) = P_t([x - \epsilon, x + \epsilon])$ is the mass within ϵ of recommendation, and $r_t(H_t) = \max\{p_t, 1 - p_t\}$.

Sample Complexity. We compare `C-ICPE` against 3 Thompson Sampling variants, where each uses, respectively, P_t, G_t and M_t. We call these variants Posterior/Moment/Mixture-TS. Moment-TS uses G_t to sample the next query, and we return $\hat{x} = \mu_\tau$ as the recommendation, while the others return the dominant lobe center. Each variant stops as soon as $q_t(h, \hat{x}) \geq 1 - \delta$ and we denote their sample complexity by N_G, N_M, N_P.

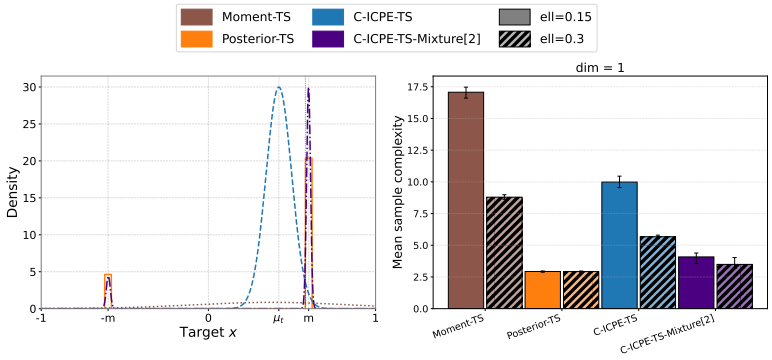


Figure 2: Two-lobe environment with $m = 0.6$, $w = 0.02$, $\epsilon = 0.08$, $\eta = 0.2$. Left ($\ell = 0.10, p_t = 0.75$): an illustrative comparison of the posterior distributions evaluated on the history of C−ICPE with a mixture. Right: sample complexity for $\ell \in \{0.15, 0.30\}$; accuracies are reported in Tab. 19.

Theorem 4.1. *Fix $m, \epsilon, \eta, \delta$, with $0 < \epsilon < m/2$. For $0 < w \leq \min\{\epsilon/2, (1-m)/2\}$ and $0 < \ell \leq \min\{m/2, 1-m\}$, the three TS rules have prior-averaged correctness at least $1-\delta$, and sample complexity*

$$\mathbb{E}N_G \asymp m/\ell, \qquad \mathbb{E}N_M \asymp \mathbb{E}N_P \asymp 1 + w/\ell, \qquad T^\star \asymp 1.$$

Here $T^\star$ is the infimum expected query count among all Bayesian (ϵ, δ)-correct learners. The constants are independent of w, ℓ with the other parameters fixed.

The rates reflect where each rule looks for evidence: the Gaussian tends to query between the lobes, and the Gaussian mean can also require stronger evidence before it is reliable. On the other hand, a mixture can alleviate both issues. In Fig. 2 are depicted the posteriors and the sample complexity for the different methods, showing how a mixture can improve sample complexity vs the single gaussian. Proofs and the 2^d-lobe construction are in Sec. C.9.

5 Numerical Results

We evaluate C−ICPE with $\delta = 0.1$ on noisy binary search, sphere best-arm identification, Ackley minimization, multi-lobe localization, GP max-value estimation, and the geochemical surrogate task (usg, 2004). We ablate over several actor rules: TS, TTPS, TD3, and uniform exploration. For most experiments we use a single Gaussian inference model, and compare to a mixture of Gaussian with TS rule (that we name C−ICPE−TS−Mixture) on the following tasks: Ackley ($d = 5$), binary search ($d = 20$), ϵ-best-arm ($d = 15$), multi-lobe localization and geochem. We compare against Bayesian optimization baselines whenever possible: TPE (Bergstra et al., 2011), GP−UCB and GP−logEI (Srinivas et al., 2010; Ament et al., 2023), and CMA−ES (Hansen, 2016). Each fixed-budget baseline receives a budget equal to the mean stopping time of the most sample-efficient C−ICPE variant; on Ackley and geochem, we also report results with a fixed budget of

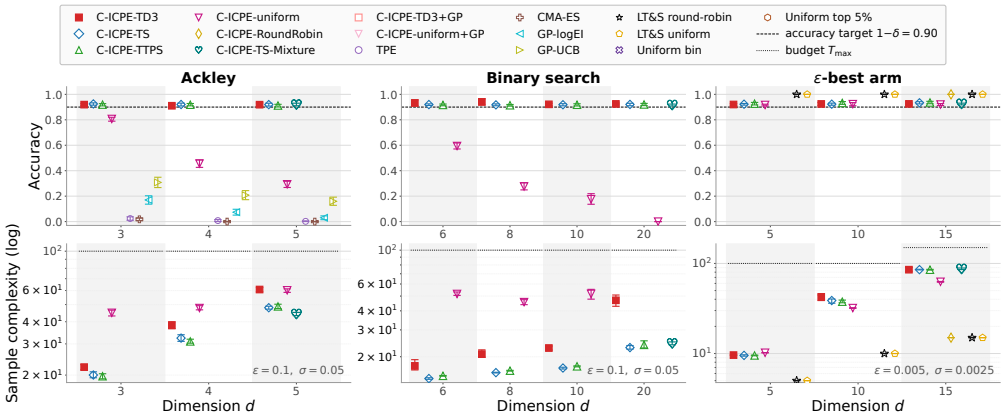


Figure 3: Accuracy (top) and sample complexity (bottom) at the hardest (ε, σ) per benchmark.

100 queries (in the appendix). On the sphere best-arm identification problem we compare against the fixed-confidence Lazy TaS (`LTS`) method (Jedra and Proutiere, 2020), which can use a round-robin strategy over arms, or a uniform strategy. We similarly evaluate `C-ICPE` with a round-robin sampling strategy (`C-ICPE-RoundRobin`), so that we can test the recommendation quality of a single Gaussian while using the same sampling strategy of `LTS`. On value estimation we also test a diagnostic, `C-ICPE-TD3+GP`, that replaces the inference model with a GP at test time. All runs have a maximum horizon, and we report 95% bootstrap intervals. We certify each trained model by rejecting the null $H_0 : p_m \leq 0.9$ at significance $\alpha = 0.01$. We use as training target $\delta_{\text{train}} = 0.0824$. Further details are in Sec. E.

5.1 Benchmarks and Results

Ackley minimization. We sample a shifted Ackley function (Naser et al., 2025) with shape parameters (α, β, γ) and minimizer $\theta^\star \in [-1, 1]^d$. A query a gives $y = \tilde{f}_{\alpha,\beta,\gamma}(a-\theta^\star)+\xi$, where $\tilde{f}$ is normalized and $\xi \sim \mathcal{N}(0, \sigma^2)$. Here $\mathcal{A} = \mathcal{X} = [-1, 1]^d$ and $L_\theta(x) = \|x - \theta^\star\|_2$. The difficulty arises from the function's many local optima and nearly flat outer region, which can trap greedy strategies; the agent must explore globally before converging.

Noisy binary search. We draw $\theta \sim \text{Unif}([-1, 1]^d)$ and use $x^\star_\theta = \theta$, $\mathcal{A} = \mathcal{X} = [-1, 1]^d$, and $L_\theta(x) = \|x - \theta\|_2$. A query a reveals $y_i = \xi_i \operatorname{sign}(\theta_i - a_i)$ in each coordinate, with independent signs satisfying $\mathbb{P}(\xi_i = 1) = 1 - p$. The learner must localize all coordinates from corrupted comparisons, where each coordinate provides one bit of corrupted information per query.

ϵ-best arm on the sphere. We draw θ uniformly on $\mathbb{S}^{d-1}$ and observe $y = \theta^\top a + \xi$, with $\xi \sim \mathcal{N}(0, \sigma^2)$. The effective spaces are $\mathcal{A} = \mathcal{X} = \mathbb{S}^{d-1}$ after normalization, the target is $x^\star_\theta = \theta$, and $L_\theta(x) = 1 - \theta^\top x$. Queries provide scalar information about an unknown direction.

GP max-value estimation. We draw $f \sim \text{GP}(0, k_{\text{RBF}}(\ell, 1))$ on $\mathcal{A} = [0, 1]^d$, with $\ell \sim \text{Unif}[0.05, 0.2]$, and observe $f(a) + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2)$. The target is the scalar $\theta^\star = \max_a f(a)$, with loss $L_\theta(x) = |x - \theta^\star|$. Here the actor queries locations while the inference model estimates the maximum value. We compare TD3 and uniform querying with BO baselines that return the maximum posterior mean, and with uniform-bin and top-5% value estimators (these are two non-parametric baselines). We also provide a characterization showing a separation in sample complexity between value estimation and maximizer localization. We make this distinction precise in Sec. C.11, Thm. C.6, and show under which conditions the value estimation problem becomes asymptotically harder than the argmax localization problem.

Geochemical exploration task. Lastly, we construct a realistic task using data from the USGS Geochemical Survey (usg, 2004), which provides measurements of copper concentration across the United States. The goal is to identify the location of peak copper concentration in an unknown region. We partition the data into geographic regions, fit a sparse variational GP to each, and split regions into training and evaluation. Training regions are used to meta-train `C-ICPE`; we evaluate on held-out regions whose spatial structure was not seen during training. See also Sec. E.5 for more details.

Results. Figs. 3-4 report accuracy and sample complexity at the hardest (ε, σ) per benchmark; full sweeps over (ε, σ, d), experimental details, and robustness to prior misspecification are in Sec. E. Every critic-based `C-ICPE` variant is certified in all seeds and configurations (largest p-value 7.5×10^{-4}); failures occur only for the uniform-query and critic-free ablations. On binary search and Ackley, the learned active `C-ICPE` variants remain near the accuracy target, while the uniform variant loses accuracy in the harder settings. On the hardest Ackley setting $d = 5$ we see that using a mixture inference further reduces the sample complexity. On the ϵ-best arm problem, by rotational symmetry, round-robin over an orthonormal basis is efficient (Jedra and Proutiere, 2020). With round-robin queries, `C-ICPE` matches LTS, whereas the other actor variants typically require more samples, and the gap lies in a sub-optimal exploration rule. In Fig. 4 we show the other results. On the multi-lobe setting, explained in Sec. 4.3, we observe how a mixture can achieve similar sample complexity to Posterior-TS, and as dimensionality increases we see how a single Gaussian can become severely sub-optimal. On the GP value estimation problem, whose prior is itself a GP, GP-logEI and TPE are slightly more accurate than C-ICPE-TD3 at matched budget. We also test a variant, `C-ICPE-TD3+GP`, which replaces the inference model with a GP at test time, and improves the overall accuracy. On geochem, we test `C-ICPE-TD3`, `C-ICPE-TS` and `C-ICPE-TS-Mixture` against the BO baselines. The `C-ICPE` variants are empirically near the target, and the mixture

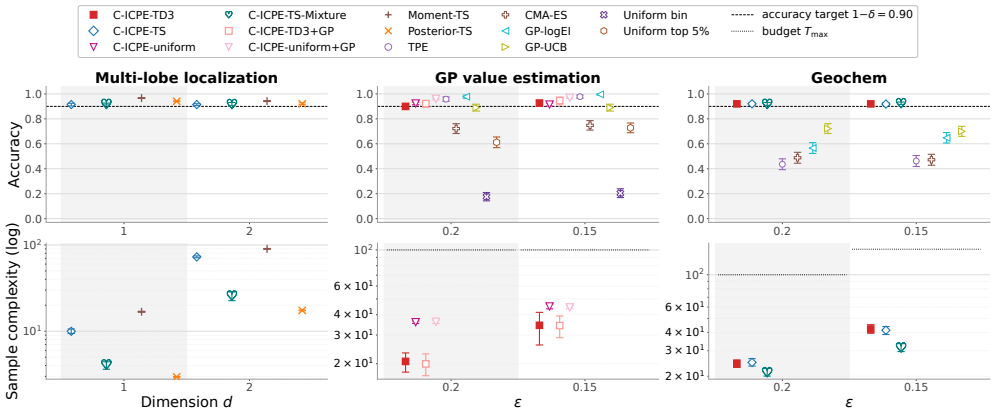


Figure 4: Accuracy (top) and mean queries (bottom): multi-lobe localization at $\ell = 0.15$, GP value estimation at the larger noise for each ϵ, and geochemistry at the reported $\sigma = 0$.

variant displays smaller sample complexity. The BO baselines fall below the target at the matched budgets, whereas GP-logEI attains higher accuracy when given 100 queries. In Sec. E.5 we report the experimental details, and results for different values of ϵ. Lastly, stopping on the inference model's own belief (+CF, see Sec. E in the appendix) is generally miscalibrated: it is overconfident on four benchmarks, where certification fails, and underconfident on geochem (2.2-2.5$\times$ more queries).

We note these results also test C−ICPE outside the assumptions of the theoretical analysis. Multi-lobe localization satisfies them; binary search has a discontinuous likelihood, while Gaussian observation models have unbounded support; the GP benchmark additionally has unbounded latent values. The observed performance provides evidence of empirical robustness beyond these conditions.

6 Discussion and Conclusions

Active sequential hypothesis testing studies how experiments and stopping should be chosen together to reach a reliable decision with few observations (Chernoff, 1959; Naghshvar and Javidi, 2013). This is also the objective of C−ICPE, but the decisions are learned across simulated tasks rather than computed from explicit likelihoods at deployment. Bayesian experimental design provides the broader expected-utility perspective (Rainforth et al., 2024); here the utility concerns the success of the returned recommendation and the cost of further queries. For structured observation models with finite arms, fixed-confidence bandit algorithms turn this tradeoff into explicit exploration and stopping strategies (Garivier and Kaufmann, 2016; Jedra and Proutiere, 2020). This approach extends to continuous-arm linear bandits (Takemori et al., 2025) and infinite-answer problems (Poiani et al., 2025a). Bayesian results include prior-averaged identification in finite Gaussian bandits (Jang et al., 2024) and GP-based stopping with probabilistic value-regret bounds (Wilson, 2024). These analyses exploit the structure to obtain sharp guarantees. C−ICPE instead targets task families for which a simulator is available but deriving a dedicated allocation and stopping rule is difficult. A related approach is Deep Adaptive Design: its likelihood-free extension learns history-dependent policies before deployment by optimizing information objectives (Foster et al., 2021; Ivanova et al., 2021). DPT instead learns in-context decision-making through supervised prediction of optimal actions (Lee et al., 2023). The stop-versus-continue critic follows ICPE (Russo et al., 2026); Cheng and Huan (2026) later use a similar comparison for continuous design, but trade information gain against a fixed cost, whereas C−ICPE enforces a prescribed accuracy through the dual cost, returns an explicit continuous recommendation, and certifies the trained model. See Sec. B for more related work.

In conclusion, we introduced C−ICPE for Bayesian pure exploration with continuous recommendations. The theory identifies the ideal structure, while the approximation analysis explains possible additional query costs. The experiments test C−ICPE across localization, identification, and value-estimation tasks, including settings outside the ideal assumptions. The main resulting benefit is a policy that adapts to new tasks within its family without parameter updates. Extending this reliability to changed task priors and reducing dependence on supervised targets remain important directions; limitations are discussed in Appendix A.

AI USE STATEMENT

We used generative AI to assist with manuscript editing and LaTeX, literature searches, code debugging, and the formulation and checking of mathematical statements and proofs. It also provided feedback on experimental design and interpretation. These outputs served as research and writing assistance, not independent verification. The authors take responsibility for the final text, mathematical claims, citations, code, and reported results.

ETHICS STATEMENT

The reported study uses simulated tasks and a surrogate benchmark built from public USGS geochemical measurements; it does not involve human subjects, animal experiments, or a new field-sampling campaign. A prior-averaged accuracy guarantee is not a per-instance safety guarantee. Deployment in scientific experimentation would require domain-specific validation and safeguards, including environmental constraints where relevant. Further limitations are discussed in Sec. A.

REPRODUCIBILITY STATEMENT

The assumptions and proofs are given in Sec. C. The history encoder, training objectives, actor rules, and update procedures are described in Sec. D. Task generation, evaluation protocols, computational resources, and geochemical preprocessing are described in Secs. E and E.5. The certification test and its independence requirements are specified in Sec. C.7; the reported test summaries are listed in the appendix. For further details, we refer the reader to the code and the `README.md` file.

REFERENCES

The National Geochemical Survey: Database and documentation. Report 2004-1001, 2004.

T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

A. Al Marjani, A. Garivier, and A. Proutiere. Navigating to the best policy in markov decision processes. In *Advances in neural information processing systems*, volume 34, pages 25852–25864, 2021.

S. Ament, S. Daulton, D. Eriksson, M. Balandat, and E. Bakshy. Unexpected improvements to expected improvement for bayesian optimization. *Advances in neural information processing systems*, 36:20577–20612, 2023.

J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p, 2010.

M. Balandat, B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538, 2020.

J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

D. Bertsekas and S. E. Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization: Theoryand Examples*. Springer, 2006.

C. Cheng and X. Huan. Optimal Stopping for Sequential Bayesian Experimental Design, June 2026. URL http://arxiv.org/abs/2509.21734. arXiv:2509.21734 [stat.ME].

H. Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755 – 770, 1959. doi: 10.1214/aoms/1177706205. URL https://doi.org/10.1214/aoms/1177706205. Publisher: Institute of Mathematical Statistics.

R. Degenne, W. M. Koolen, and P. Ménard. Non-asymptotic pure exploration by solving games. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché Buc, E. Fox, and R. Garnett, editors, *Advances in neural information processing systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8d1de7457fa769ece8d93a13a59c8552-Paper.pdf.

B. Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

I. Ekeland and R. Temam. *Convex analysis and variational problems*. SIAM, 1999.

E. A. Feinberg and P. O. Kasyanov. MDPs with setwise continuous transition probabilities. *Operations Research Letters*, 49(5):734–740, 2021.

A. Foster, D. R. Ivanova, I. Malik, and T. Rainforth. Deep Adaptive Design: Amortizing Sequential Bayesian Experimental Design. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3384–3395. PMLR, July 2021.

S. Fujimoto, H. Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. *Proceedings of the 29th Conference on Learning Theory*, 49:998–1027, 2016.

A. Garivier and E. Kaufmann. Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96, 2021. Publisher: Taylor & Francis.

R. Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.

N. Hansen. The cma evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772*, 2016.

A. Hantoute and MA. López. Characterizations of the subdifferential of the supremum of convex functions. *Journal of Convex Analysis*, 15:831–858, 2008.

P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *Journal of Machine Learning Research*, 13(57):1809–1837, 2012. ISSN 1533-7928.

O. Hernández-Lerma and J. B. Lasserre. *Discrete-Time Markov Control Processes*. Springer, New York, NY, 1996. ISBN 978-1-4612-6884-0 978-1-4612-0729-0. doi: 10.1007/978-1-4612-0729-0. URL http://link.springer.com/10.1007/978-1-4612-0729-0.

J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive Entropy Search for Efficient Global Optimization of Black-box Functions. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/6488484c982e9af5c35689523ba1abfe-Abstract.html.

D. Huang, Y. Guo, L. Acerbi, and S. Kaski. Amortized Bayesian Experimental Design for Decision-Making, Jan. 2025a. URL http://arxiv.org/abs/2411.02064. arXiv:2411.02064 [stat.ML].

D. Huang, X. Wen, A. Bharti, S. Kaski, and L. Acerbi. ALINE: Joint Amortization for Bayesian Inference and Active Data Acquisition, Oct. 2025b. URL http://arxiv.org/abs/2506.07259. arXiv:2506.07259 [stat.ML].

D. R. Ivanova, A. Foster, S. Kleinegesse, M. U. Gutmann, and T. Rainforth. Implicit Deep Adaptive Design: Policy-Based Experimental Design without Likelihoods, Nov. 2021.

K. Jang, J. Komiyama, and K. Yamazaki. Fixed Confidence Best Arm Identification in the Bayesian Setting. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 17789–17829. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1fb0a4de9c14f5557eeea886e22569cd-Paper-Conference.pdf.

Y. Jedra and A. Proutiere. Optimal best-arm identification in linear bandits. In *Advances in neural information processing systems*, volume 33, pages 10007–10017, 2020.

M. Jourdan, R. Degenne, D. Baudry, R. de Heide, and E. Kaufmann. Top Two Algorithms Revisited. In *Advances in Neural Information Processing Systems*, volume 35, pages 26791–26803, Dec. 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ab5f5f22e3e09f4424592ffb06840ab0-Abstract-Conference.html.

E. Kaufmann and W. M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

J. Lee, A. Xie, A. Pacchiano, Y. Chandak, C. Finn, O. Nachum, and E. Brunskill. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36:43057–43083, 2023.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Y. Liu, T. Zhao, W. Ju, and S. Shi. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177, 2017.

M. Naghshvar and T. Javidi. Active Sequential Hypothesis Testing. *The Annals of Statistics*, 41(6): 2703–2738, 2013. ISSN 0090-5364.

NASA/GSFC/METI/ERSDAC/JAROS, and U.S./Japan ASTER Science Team. Mountain pass mine, california, 2010. URL https://www.jpl.nasa.gov/images/pia13979-mountain-pass-mine-california. Image PIA13979, acquired March 28, 2010.

M. Naser, M. K. Al-Bashiti, A. T. G. Tapeh, A. Naser, V. Kodur, R. Hawileh, J. Abdalla, N. Khodadadi, A. H. Gandomi, and A. D. Eslamlou. A review of benchmark and test functions for global optimization algorithms and metaheuristics. *Wiley Interdisciplinary Reviews: Computational Statistics*, 17(2):e70028, 2025.

J. O'Quigley, M. Pepe, and L. Fisher. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, 46(1):33–48, Mar. 1990. ISSN 0006-341X.

R. Poiani, M. Bernasconi, and A. Celli. Pure Exploration with Infinite Answers, May 2025a.

R. Poiani, M. Jourdan, E. Kaufmann, and R. Degenne. Best-Arm Identification in Unimodal Bandits. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pages 2233–2241. PMLR, Apr. 2025b. URL https://proceedings.mlr.press/v258/poiani25a.html. ISSN: 2640-3498.

T. Rainforth, A. Foster, D. R. Ivanova, and F. Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024. Publisher: Institute of Mathematical Statistics.

R. T. Rockafellar. *Conjugate duality and optimization*. Society for Industrial and Applied Mathematics, 1974.

R. T. Rockafellar and R. J. Wets. *Variational analysis*. Springer, 1998.

A. Russo, R. Welch, and A. Pacchiano. In-context learning for pure exploration. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 14493–14509, 2026.

D. Russo. Simple Bayesian Algorithms for Best Arm Identification. In *Conference on Learning Theory*, pages 1417–1418. PMLR, June 2016. URL https://proceedings.mlr.press/v49/russo16.html.

N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022, 2010.

S. Takemori, Y. Umeda, and A. Gopalan. Instance-optimal pure exploration for linear bandits on continuous arms. In *Forty-second International Conference on Machine Learning*, 2025.

J. Taupin, Y. Jedra, and A. Proutiere. Best policy identification in discounted linear MDPs. In *Sixteenth european workshop on reinforcement learning*, 2023.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326 – 339, 1948. doi: 10.1214/aoms/1177730197. URL https://doi.org/10.1214/aoms/1177730197.

J. T. Wilson. Stopping bayesian optimization with probabilistic regret bounds. *Advances in Neural Information Processing Systems*, 37:98264–98296, 2024.

CONTENTS

1 Introduction **1**

2 Problem Setting **2**

3 Theoretical Background **2**

4 Continuous ICPE: C-ICPE **3**

4.1 Inference Model and Recommendation 4

4.2 Critic and Actor Models 4

4.3 Multimodal posterior stress test 6

5 Numerical Results **7**

5.1 Benchmarks and Results 8

6 Discussion and Conclusions **9**

Appendix **16**

A Limitations and Broader Impact **17**

B Extended Related Work **18**

C Appendix: Theoretical Results **20**

C.1 Problem Modeling 21

C.2 Posterior distribution over the true hypothesis and inference rule optimality 23

C.3 Fixed-confidence setting: formulation and optimal rules 26

C.3.1 Optimal Inference Rule 27

C.3.2 Stopping as an action (equivalence) 27

C.3.3 Optimal Policy 27

C.3.4 Reward Shaping and Removal of the Stop Action 32

C.4 Fixed-confidence setting: Bayesian (ϵ, δ)-correctness and zero-duality gap 34

C.4.1 Main Result 34

C.5 Bayesian Correctness of Dual-Optimal Points 35

C.5.1 Zero Duality Gap via Perturbation Values 38

C.6 Approximation and relative sample complexity 42

C.7 Correctness certification 43

C.8 Choice of inference model and reward modeling 46

C.9 Multimodal posterior stress test 51

C.9.1 Posterior and recommendation 52

C.9.2 Information collected by the query rules 53

C.9.3 From information to sample complexity . . . 54
C.9.4 The 2^d-lobe construction . . . 55
C.10 Robustness to prior misspecification . . . 56
C.11 Sample Complexity: Value Estimation vs Argmax Localization . . . 59
C.11.1 Max-value estimation complexity . . . 60
C.11.2 Argmax localization complexity . . . 63

D Appendix: Algorithms **69**
D.1 History Encoder and Time Pooling Layer . . . 69
D.2 Replay Buffer and Prefix Sampling . . . 69
D.3 Inference Update and Its Regularization . . . 70
D.4 Reward, Critic Update, and Critic Regularization . . . 70
D.5 Actor Rules: TS, TTPS, and TD3 . . . 71
D.6 Cost Update for Fixed Confidence . . . 73

E Appendix: Numerical Results **74**
E.1 Synthetic Benchmarks: description . . . 75
E.1.1 Noisy binary search. . . . 76
E.1.2 ϵ-best-arm identification on the sphere. . . . 76
E.1.3 Ackley minimizer identification. . . . 76
E.1.4 GP max-value estimation . . . 77
E.1.5 Multi-Lobe Localization . . . 78
E.2 Synthetic Benchmarks: baselines . . . 79
E.3 Synthetic Benchmarks: numerical results . . . 82
E.3.1 Noisy Binary Search . . . 82
E.3.2 ϵ-best arm problem . . . 87
E.3.3 Ackley minimization . . . 92
E.3.4 GP max-value estimation . . . 97
E.3.5 Multi-Lobe Localization . . . 100
E.4 Synthetic Benchmarks: robustness . . . 104
E.4.1 Noisy Binary Search . . . 104
E.4.2 ε-best-arm problem . . . 104
E.4.3 Ackley minimization . . . 104
E.4.4 GP max-value estimation . . . 107
E.4.5 Multi-Lobe Localization . . . 108
E.5 Geochemical Exploration: Experimental Details and Numerical Results . . . 109
E.5.1 Dataset and Motivation . . . 109
E.5.2 Region Partitioning and GP Fitting . . . 109
E.5.3 Ground Truth Construction . . . 109

E.5.4 Task Prior and Train/Test Split 110
E.5.5 Numerical Results 114

A LIMITATIONS AND BROADER IMPACT

Parametric inference model. C−ICPE models the posterior law of the target $x^\star_\theta$ with a diagonal Gaussian or a finite mixture. For a single Gaussian with nondegenerate conditional moments, ordinary population NLL gives a moment projection: the mean matches $\mathbb{E}[x^\star_\theta | H_t]$ and the diagonal covariance matches the posterior coordinate variances. This is only a surrogate for the Bayes ϵ-rule $\arg\max_{x\in\mathcal{X}} q_t(H_t, x)$. When the posterior over $x^\star_\theta$ is multimodal, or poorly summarized by first and second moments, the mean recommendation may lie between plausible targets and the covariance may misrepresent uncertainty. This can affect both exploration and stopping, since inference supplies recommendations and, for TS, query proposals. Richer posterior families, such as normalizing flows, or a direct model of $q_t(h, x)$, could reduce this mismatch at the cost of additional optimization and training complexity.

Note that this limitation is most benign in localization benchmarks, where $\mathcal{X}_\epsilon(\theta)$ is a ball or interval around $x^\star_\theta$. It is more pronounced in value-gap tasks such as geochemical optimization, where the ϵ-optimal set can be anisotropic, nonconvex, or multimodal. These tasks are therefore useful stress tests for the Gaussian inference model.

Selector availability. Training the inference model requires access to the selector $x^\star(\theta)$ for each sampled task $\theta \sim \nu$. In Ackley and binary search it is available in closed form; in the implemented GP benchmark the target is obtained from the grid representation. In general, obtaining a target with a known success margin may itself require numerical optimization. Suboptimal targets are allowed, but a smaller margin can increase the evidence required for a reliable recommendation (Example 2). Noisy or uncontrolled target errors require additional assumptions.

Cost calibration. The per-query cost $c = 1/\lambda$, which controls the exploration–stopping tradeoff, is updated online during training via primal feedback on the empirical success rate. In practice, the convergence of c and the sensitivity of stopping behavior to its value require careful tuning of learning rates and update schedules.

Decoupled action and recommendation spaces. When $\mathcal{X} \neq \mathcal{A}$, C−ICPE requires a separate TD3 actor to learn the exploration policy. This adds architectural complexity and an additional source of approximation error. Our GP max-value experiment exercises this setting.

Bayesian guarantee and prior dependence. The (ϵ, δ)-correctness guarantee is average-case under the task prior ν. When the deployment distribution differs substantially from ν, correctness may degrade. Our robustness experiments and the prior-shift bounds in Sec. C.10 provide some quantitative control, but worst-case guarantees for individual task instances are not provided. This limitation is shared by all Bayesian methods and is analogous to the dependence of frequentist methods on their parametric assumptions.

Scalability. The method's behavior in higher dimensions (larger than $d \geq 50$), where posterior concentration is slower and exploration is harder, remains to be investigated. The LSTM (or transformer) architecture scales with the maximum horizon $t_{\max}$, which may need to grow with dimension, increasing both training and inference cost.

Broader impact. C−ICPE is a general-purpose tool for adaptive experimentation with correctness guarantees. Potential applications include materials discovery, dose-finding in clinical trials, environmental monitoring, and any setting where sequential experiments are costly and the practitioner requires a principled stopping criterion. In such settings, reducing sample complexity directly translates to reduced cost, time, and resource consumption.

We do not foresee direct negative societal impacts from the method itself. However, as with any system that automates experimental decisions, users should be aware that the (ϵ, δ) guarantee is conditional on the modeling assumptions (the task prior ν and the observation model). Deploying C−ICPE in safety-critical domains, such as clinical dose-finding, would require careful validation of these assumptions and, where appropriate, additional safeguards beyond the Bayesian guarantee.

B Extended Related Work

We expand the discussion in Sec. 6, focusing on how sequential testing, model-specific pure exploration, and amortized experimental design address the choice of queries and the decision to stop.

Sequential testing and experimental design. Sequential analysis studies how to accumulate evidence until it supports a reliable decision (Wald and Wolfowitz, 1948). Active sequential hypothesis testing adds the choice of which experiment to perform, coupling this choice with stopping and the final declaration (Chernoff, 1959; Naghshvar and Javidi, 2013). With a specified observation model, likelihoods provide the link between each experiment and the evidence it supplies about competing hypotheses. Bayesian experimental design expresses the value of an experiment through expected utility, often measured by information gain (Lindley, 1956; Rainforth et al., 2024). Our formulation focuses this utility on the success of the returned recommendation: further observations are useful insofar as they help reach the prescribed accuracy with fewer queries.

Pure exploration in bandits and MDPs. Best-arm identification makes this tradeoff explicit in bandit models. Instance-dependent lower bounds characterize the information needed to distinguish the best arm from its alternatives, and allocation rules can be designed to approach these bounds while using valid stopping tests (Garivier and Kaufmann, 2016; Degenne et al., 2019). Top-two methods adapt posterior-driven exploration to identification by allocating samples between a leading candidate and a challenger (Russo, 2016; Jourdan et al., 2022). A related question arises in Markov decision processes, where the learner must collect evidence along a trajectory to identify an optimal policy (Al Marjani et al., 2021). These approaches provide sharp guidance on how observations should be allocated under their respective statistical and interaction structures. They motivate learning the same coupled decisions when deriving a dedicated allocation and stopping rule is difficult.

Continuous recommendations and Bayesian correctness. Continuous recommendation spaces introduce the possibility of infinitely many answers, including multiple acceptable recommendations for the same task. Takemori et al. (2025) develop a tractable, instance-dependent pure-exploration method for linear bandits with continuous arm sets. Poiani et al. (2025a) study infinite-answer problems and derive lower bounds and an asymptotically optimal Track-and-Stop framework. These results concern instance-wise correctness. A different objective averages correctness and query cost over a task prior. In finite Gaussian bandits, Jang et al. (2024) analyze this Bayesian fixed-confidence objective and show that methods designed for the instance-wise criterion can be inefficient for it. `C-ICPE` adopts the prior-averaged objective and allows the recommendation to be continuous.

Bayesian optimization and stopping. Bayesian optimization uses a surrogate posterior to choose evaluations of an unknown objective (Garnett, 2023). Information-based acquisition rules can target uncertainty about its optimizer (Hennig and Schuler, 2012; Hernández-Lobato et al., 2014). Wilson (2024) provide a GP-based stopping criterion that controls the posterior probability of returning a point with excessive value regret. This supplies a principled alternative to a fixed budget. The recommendation and its loss remain important: returning a point with small value regret differs from estimating the numerical maximum value. Our formulation accommodates both by specifying the recommendation space and success loss separately from the query space.

Amortized design and learned stopping. Amortized methods move the cost of constructing an experimental policy before deployment. Deep Adaptive Design learns a history-dependent design network from simulated experiments (Foster et al., 2021), and its implicit extension supports simulators without tractable likelihoods (Ivanova et al., 2021). The learned objective can also be tied directly to a downstream task: TNDP jointly amortizes experimental design and decision-making (Huang et al., 2025a), while ALINE couples inference and acquisition using a self-estimated information-gain reward (Huang et al., 2025b). Related work by Cheng and Huan (2026) learns continuous design policies with value-based stopping, balancing information gain against experimental cost. Their terminal reward is the information gain of an explicitly computed posterior at a fixed per-experiment cost, so the method neither returns a recommendation nor controls its accuracy; applying it to our setting would require a success-based reward and a cost tuned to reach $1 - \delta$, which is what `C-ICPE` provides.

Our method follows the in-context pure-exploration approach of ICPE (Russo et al., 2026). DPT learns in-context decision-making by supervised prediction of optimal actions (Lee et al., 2023); ICPE combines supervised inference and reinforcement learning to learn querying and stopping for finite active-testing problems. `C-ICPE` extends this construction to continuous recommendations. Its objective is to minimize expected deployment queries subject to a prescribed prior-averaged success probability, rather than to maximize information gain. The inference model, actor, and stopping critic are trained together, and independent tests assess the accuracy of the complete frozen policy.

C Appendix: Theoretical Results

The theoretical analysis is organized as follows. We summarize what is standard and what is new relative to the finite ICPE framework of Russo et al. (2026).

- **§C.1: Problem modeling.** We define the environment, learner, and induced probability measures, state the regularity assumptions, and map each benchmark to them (Remark 1).
- **§C.2: Posterior success probability.** We define $q_t(h, x) = \mathbb{P}(L_\theta(x) \leq \epsilon \mid H_t = h)$ and establish its regularity properties. The proofs use standard tools (Radon–Nikodym, reverse Fatou, Portmanteau); the object $q_t(h, x)$ itself, the natural continuous analogue of posterior mass.
- **§C.3: Bellman optimality.** We prove that the optimal value satisfies a stop/continue Bellman equation and that the supremum over continuation actions is attained by a measurable selector. The proof adapts the value-iteration framework of Bertsekas and Shreve (1996) to our setting, with the main technical content being a semicontinuity induction that threads likelihood continuity through posterior weak continuity, predictive weak continuity, Q-function lower semicontinuity, and a sup–inf interchange. None of these steps appear in Russo et al. (2026), where attainment is automatic for finite action spaces. The resulting dependency chain and its coupling with the posterior success payoff are specific to this problem.
- **§C.4–C.5: Correctness and zero duality gap.** §C.4 states the main result (Thm. C.2): under time-sharing and strict feasibility the duality gap is zero, and if the near-optimal cost–success set $\mathcal{K}_{\epsilon_0}(\lambda^\star)$ is closed, some dual-optimal policy is primal optimal and, with the posterior-optimal inference rule, Bayesian (ϵ, δ)-correct. §C.5 proves it. Correctness holds under this local closedness condition, which allows non-unique dual minimizers; the proof uses a subdifferential characterization from Hantoute and López (2008) to derive a contradiction without the monotonicity argument of Russo et al. (2026), and this is the strongest theorem-level novelty in the appendix. Zero duality gap follows from a perturbation argument (§C.5.1, Rockafellar, 1974).
- **§C.6: Approximation and relative sample complexity.** Assuming strong duality and a constrained-optimal learner, we prove Prop. 1, which bounds the expected number of queries of an approximate learner relative to $T^\star$. The lower bound follows from strong duality: an accuracy surplus e_p costs at least $\lambda^\star e_p$ queries. For the upper bound, we accumulate the critic and actor errors along the learner's trajectory, and compare with a learner that keeps the queries and stopping rule of the optimal learner but returns the learned recommendation, which isolates the recommendation error e_I.
- **§C.7: Model certification.** We introduce a checkpointwise certification protocol based on mixture supermartingales (Kaufmann and Koolen, 2021). Unlike the pooled approach in Russo et al. (2026), it tests each frozen checkpoint independently and requires no monotonicity assumption on the training trajectory.
- **§C.8: Inference model and sampled reward.** We relate the implemented Gaussian inference model to the ideal posterior quantities used in the Bellman characterization. We show that the sampled reward is an unbiased estimate of the success probability of the stochastic selector and is conservative relative to the ideal reward $r_t(h)$; the gap is controlled by second moments of the inference distribution (Prop. 8). We also characterize the Gaussian NLL as a moment projection of the posterior law of $x^\star_\theta$ and give conditions under which the NLL mean is near-optimal (Props. 9 and 10).
- **§C.9: Multimodal posterior stress test.** We prove Thm. 4.1. We compute the exact posterior of the two-lobe problem and the success of each recommendation (exact posterior, Gaussian projection, and two-component mixture), derive the information that each Thompson-sampling rule collects per query, and turn it into expected query counts with matching lower bounds. This shows why a single Gaussian needs $\asymp m/\ell$ queries, while the mixture and the exact posterior need $\asymp 1 + w/\ell$ and $T^\star \asymp 1$. We conclude with the 2^d-lobe construction used in the experiments.
- **§C.10: Robustness to prior misspecification.** We analyse the robustness to prior misspecification, and what is the predicted impact on sample complexity and accuracy.

- **§C.11: Sample Complexity of Value Estimation vs Argmax Localization in Gaussian Processes.** In this section we prove that, under a hierarchical RBF-GP prior with a high-probability interior regularity condition, max-value estimation is asymptotically harder than argmax localization. Specifically, under this regularity assumption, we establish that the value estimation problem cannot be circumvented by the geometric structure of the function.
  - The analysis is based on showing a lower bound on the sample complexity of estimating the max-value, and an upper bound on estimating the argmax.
  - To this aim, we introduce an algorithm, Two-Stage Bayesian Argmax Localization (T-BAL) Algorithm 2, for locating the argmax of a Gaussian Process.
  - T-BAL works in two phases: first, searches the domain for a region where $X^\star$ may be located, and then perform gradient ascent using noisy finite differences to approximate the gradients.
  - We provide a sample complexity upper bound of T-BAL and provide Bayesian (ϵ, δ)-guarantees.

C.1 Problem Modeling

We specialize to the Bayesian fixed-confidence (Bayesian (ϵ, δ)-PAC) setting introduced in Section 2, and provide a self-contained definition of the induced probability measures.

We now provide a formal definition of the underlying probability measures of the problem we consider. To that aim, it is important to formally define what a model M is, as well as the definition of policy π and inference rule I (infernece rules are also known as recommendation rules).

Remark 1 (Scope of the regularity assumptions.)**.** *The assumptions state below are sufficient for the ideal Bellman analysis, not prerequisites for executing a learned policy or testing a frozen checkpoint. The benchmark correspondence is:*

Benchmark	*Deviations from the stated ideal assumptions*
Noisy binary search	*Compact parameter and decision spaces and continuous loss, but the sign likelihood is discontinuous at query-target equality.*
Sphere and Ackley	*Continuous positive Gaussian likelihoods at positive noise, but unbounded observation support. The sphere uses normalized feasible actions and recommendations.*
Multi-lobe	*None (with a θ-independent initial observation). Θ is a union of 2^d disjoint cubes (since $w < m$), $\mathcal{A} = \mathcal{X} = [-1, 1]^d$, and observations are binary. The Bernoulli likelihood $g_{s,\ell}(a) \in [\eta, 1-\eta]$ is strictly positive and jointly continuous, since the lobe sign s is constant on each cube; the L_∞ loss is continuous.*
GP max-value	*The grid implementation is finite-dimensional, but Gaussian latent values, scalar target values, and observation noise are unbounded.*
Geochem task	*Noise-free observations ($\sigma = 0$) of random surfaces sampled from region-specific fitted GP priors: observation laws have no density, and the task space is not a compact subset of $\mathbb{R}^d$.*

Hence, we see performance outside these conditions as an evidence of empirical robustness.

Spaces and histories. Let $\Theta \subset \mathbb{R}^d$ be compact. Let $\mathcal{A} \subset \mathbb{R}^m$ be a compact action (query) space and $\mathcal{Y} \subset \mathbb{R}^n$ a compact observation space, each endowed with the Borel σ-algebra. Let $\mathcal{X}$ be a compact hypothesis/decision space (not necessarily equal to $\mathcal{A}$). For $t \in \mathbb{N}$, define the history space

$$\mathcal{H}_t := (\mathcal{Y} \times \mathcal{A})^{t-1} \times \mathcal{Y}, \qquad h_t = (y_1, a_1, \ldots, a_{t-1}, y_t),$$

with its product Borel σ-algebra. We also write $\mathcal{H}_\infty := \mathcal{Y} \times (\mathcal{A} \times \mathcal{Y})^{\mathbb{N}}$ for infinite histories. Since $\mathcal{A}, \mathcal{Y}$ are compact metric spaces, $\mathcal{H}_t$ and $\mathcal{H}_\infty$ are standard Borel spaces.

Environment (observation model). An environment is indexed by $\theta \in \Theta$ and specified by an initial observation law $\rho_\theta \in \Delta(\mathcal{Y})$ and a sequence of (possibly history-dependent) observation kernels

$$P_{\theta,t}(\cdot | h_t, a_t) \in \Delta(\mathcal{Y}), \qquad t \geq 1,$$

such that for every Borel $C \subset \mathcal{Y}$ the map $(\theta, h_t, a) \mapsto P_{\theta,t}(C|h_t, a)$ is measurable. Optionally, one may assume weak continuity in θ. However, we do assume weak continuity in a, as this is later used to prove optimality.

Assumption 1 (Weak continuity of the transition)**.** *For all $\theta \in \Theta$ we assume $a \mapsto P_{\theta,t}(\cdot|h_t, a)$ to be weakly continuous.*[1]

Learner: policy, stopping time, inference rule. A (possibly randomized) sampling policy is a sequence of probability kernels

$$\pi_t(\cdot|h_t) \in \Delta(\mathcal{A}), \qquad t \geq 1,$$

measurable as maps $\mathcal{H}_t \to \Delta(\mathcal{A})$. Let $H_t = (Y_1, A_1, \ldots, A_{t-1}, Y_t)$ be the observed history. Let U collect the learner's random seeds, independent of the task and observation noise before interaction. A stopping time τ is adapted to the available history and seeds (equivalently, to $\mathcal{F}_t = \sigma(H_t, U)$ for this analysis). An inference rule is a sequence of measurable maps $I_t : \mathcal{H}_t \to \mathcal{X}$, and the learner outputs

$$\hat{x}_\tau := I_\tau(H_\tau).$$

Loss and ϵ-optimal set. For each $\theta \in \Theta$, the environment induces a loss function $L_\theta : \mathcal{X} \to [0, \infty)$ with $\inf_{x\in\mathcal{X}} L_\theta(x) = 0$. Define the ϵ-optimal set

$$\mathcal{X}_\epsilon(\theta) := \{x \in \mathcal{X} : \ L_\theta(x) \leq \epsilon\}.$$

In the following we the following conditions on $L_\theta(x)$.

Assumption 2. *We assume joint lower-semicontinuity of $(x, \theta) \mapsto L_\theta(x)$ and joint Borel measurability.*

Remark 2 (Regularity of selector-based localization losses)**.** *Some of our localization examples define the loss through a selected target $x^\star_{\mathrm{sel}}(\theta) \in \mathcal{X}$, for instance $L_\theta(x) = \|x - x^\star_{\mathrm{sel}}(\theta)\|$. In this case, the standing lower-semicontinuity assumption on $L(\theta, x) = L_\theta(x)$ is satisfied whenever the selector $\theta \mapsto x^\star_{\mathrm{sel}}(\theta)$ is continuous. There are two standard ways to obtain such a selector.*

1. *First, if $F(\theta) := \arg\max_{x\in\mathcal{X}} f_\theta(x) = \{x^\star(\theta)\}$ is singleton for every θ, $f(\theta, x)$ is jointly continuous, and $\mathcal{X}$ is compact, then the maximum theorem implies that the unique optimizer is continuous, and thus $x^\star_{\mathrm{sel}}(\theta) = x^\star(\theta)$ is continuous. This covers the shifted Ackley benchmark when the shift determines a unique optimizer continuously, the linear bandit benchmark on the unit sphere where $x^\star(\theta) = \theta$, and noisy binary search when the target map is continuous.*

2. *Second, if $F(\theta)$ is not singleton but has nonempty compact convex values and is Hausdorff-continuous in θ, then the minimum-norm selector $x^\star_{\mathrm{sel}}(\theta) := \arg\min_{x\in F(\theta)} \|x\|^2$ is well-defined and continuous: closed convex values give uniqueness of the minimum-norm point, while Hausdorff continuity gives stability of this point as θ varies.*

Thus the selector-based distance loss is jointly continuous in these cases. This selector also gives a canonical target for the Gaussian inference model: the NLL objective learns a moment projection of the posterior law of $x^\star_{\mathrm{sel}}(\theta)$, while correctness remains defined through the loss L_θ and the set $\mathcal{X}_\epsilon(\theta)$.

Path measures (Ionescu–Tulcea). Fix $\theta \in \Theta$ and a policy π. By the Ionescu–Tulcea theorem, there exists a unique probability measure $\mathbb{P}^\pi_{\theta,t}$ on $(\mathcal{H}_t, \mathcal{B}(\mathcal{H}_t))$ such that for all cylinder sets $C = C_1 \times B_1 \times \cdots \times B_{t-1} \times C_t$ (with $C_i \in \mathcal{B}(\mathcal{Y})$ and $B_i \in \mathcal{B}(\mathcal{A})$),

$$\mathbb{P}^\pi_{\theta,t}(C) = \int_{C_1} \rho_\theta(dy_1) \prod_{s=1}^{t-1} \left[\int_{B_s} \pi_s(\mathrm{d}a_s|h_s) \int_{C_{s+1}} P_{\theta,s}(\mathrm{d}y_{s+1}|h_s, a_s) \right].$$

Analogously, one obtains a unique path measure $\mathbb{P}^\pi_\theta$ on $(\mathcal{H}_\infty, \mathcal{B}(\mathcal{H}_\infty))$.

Mixture law over tasks. Given a prior ν on Θ, define the joint law on $\Theta \times \mathcal{H}_t$ by

$$\mathbf{P}^\pi_t(\mathrm{d}\theta, \mathrm{d}h_t) := \nu(\mathrm{d}\theta)\, \mathbb{P}^\pi_{\theta,t}(\mathrm{d}h_t),$$

and the trajectory marginal $\mathbb{P}^\pi_t(\cdot) = \int \mathbb{P}^\pi_{\theta,t}(\cdot)\, \nu(\mathrm{d}\theta)$. We use $\mathbb{E}^\pi_{\theta\sim\nu}[\cdot]$ and $\mathbb{P}^\pi_{\theta\sim\nu}(\cdot)$ for expectations/probabilities under this mixture.

[1] That is, for all continuous bounded functions f we have that $a \mapsto \int_{\mathcal{Y}} f(y) P_{\theta,t}(\mathrm{d}y|h_t, a)$ is continuous.

Fixed-confidence objective. The learner is Bayesian (ϵ, δ)-correct (under ν) if

$$\mathbb{P}^{\pi}_{\theta\sim\nu}(\hat{x}_\tau \in \mathcal{X}_\epsilon(\theta)) \geq 1-\delta.$$

In the fixed-confidence regime, we seek to minimize the expected number of queries subject to Bayesian (ϵ, δ)-correctness:

$$\inf_{\pi, I, \tau} \mathbb{E}^{\pi}_{\theta\sim\nu}[\tau] \quad \text{s.t.} \quad \mathbb{P}^{\pi}_{\theta\sim\nu}(\hat{x}_\tau \in \mathcal{X}_\epsilon(\theta)) \geq 1-\delta.$$

C.2 Posterior distribution over the true hypothesis and inference rule optimality

We first record a domination assumption that allows us to express likelihoods w.r.t. fixed reference measures.

Assumption 3 (Domination)**.** *There exist probability measures λ_0, λ on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ such that, for all $\theta \in \Theta$, all $t \geq 1$, and all $(h_t, a) \in \mathcal{H}_t \times \mathcal{A}$,*

$$\rho_\theta(\cdot) \ll \lambda_0(\cdot) \quad \textit{and} \quad P_{\theta,t}(\cdot|h_t, a) \ll \lambda(\cdot).$$

Let $p_{\theta,0}(y) := \frac{\mathrm{d}\rho_\theta}{\mathrm{d}\lambda_0}(y)$ and $p_{\theta,t}(y'|h_t, a) := \frac{\mathrm{d}P_{\theta,t}(\cdot|h_t,a)}{\mathrm{d}\lambda}(y')$ be versions of the corresponding densities, chosen jointly measurable in their arguments.

Remark. The assumption holds, for instance, when all ρ_θ and $P_{\theta,t}(\cdot|h_t, a)$ admit densities w.r.t. a common reference measure (e.g., Lebesgue on $\mathcal{Y} \subset \mathbb{R}^n$ or counting measure when $\mathcal{Y}$ is finite).

Under Assumption 3, define the (policy-independent) likelihood of a realized history $h_t = (y_1, a_1, \ldots, a_{t-1}, y_t) \in \mathcal{H}_t$ under parameter θ:

$$\ell_t(\theta, h_t) := p_{\theta,0}(y_1) \prod_{s=1}^{t-1} p_{\theta,s}(y_{s+1}|h_s, a_s), \qquad h_s = (y_1, a_1, \ldots, a_{s-1}, y_s).$$

We now give a posterior kernel representation that is independent of π.

Lemma 1 (Posterior kernel over Θ)**.** *Consider Assumption 3. For each $t \in \mathbb{N}$ there exists a probability kernel $R_t : \mathcal{H}_t \times \mathcal{B}(\Theta) \to [0, 1]$, independent of π, such that for every policy π, all $A \in \mathcal{B}(\Theta)$ and $Z \in \mathcal{B}(\mathcal{H}_t)$,*

$$\mathbf{P}^{\pi}_t(\theta \in A, H_t \in Z) = \int_Z R_t(A|h)\mathbb{P}^{\pi}_t(\mathrm{d}h),$$

where $\mathbf{P}^{\pi}_t(\mathrm{d}\theta, \mathrm{d}h) = \nu(\mathrm{d}\theta)\, \mathbb{P}^{\pi}_{\theta,t}(\mathrm{d}h)$ and $\mathbb{P}^{\pi}_t$ is its $\mathcal{H}_t$-marginal. Moreover, for $\mathbb{P}^{\pi}_t$-a.e. $h \in \mathcal{H}_t$,

$$R_t(A|h) = \frac{\int_A \ell_t(\theta, h)\nu(\mathrm{d}\theta)}{\int_\Theta \ell_t(\theta, h)\nu(\mathrm{d}\theta)}.$$

Consequently, for any measurable map $g : \Theta \to \mathcal{S}$ into a standard Borel space $\mathcal{S}$ and any $B \in \mathcal{B}(\mathcal{S})$,

$$\mathbb{P}(g(\theta) \in B|H_t = h) = R_t(\{\theta :\ g(\theta) \in B\}|h) \quad \textit{for } \mathbb{P}^{\pi}_t\textit{-a.e. } h.$$

Proof. Fix π and t. Define the reference measure on $\mathcal{H}_t$ (depending on π)

$$\nu^{\pi}_t(\mathrm{d}h_t) := \lambda_0(\mathrm{d}y_1) \prod_{s=1}^{t-1} \left[\pi_s(\mathrm{d}a_s|h_s)\, \lambda(\mathrm{d}y_{s+1})\right].$$

By construction and Assumption 3, $\mathbb{P}^{\pi}_{\theta,t} \ll \nu^{\pi}_t$ for every θ, with Radon–Nikodym density

$$\frac{\mathrm{d}\mathbb{P}^{\pi}_{\theta,t}}{\mathrm{d}\nu^{\pi}_t}(h_t) = \ell_t(\theta, h_t),$$

which does not depend on π. Therefore, for $A \in \mathcal{B}(\Theta)$ and $Z \in \mathcal{B}(\mathcal{H}_t)$,

$$\mathbf{P}^{\pi}_t(\theta \in A, H_t \in Z) = \int_A \int_Z \ell_t(\theta, h)\, \nu^{\pi}_t(\mathrm{d}h)\nu(\mathrm{d}\theta),$$

and

$$\mathbb{P}_t^\pi(Z) = \int_Z \int_\Theta \ell_t(\theta, h)\, \nu(\mathrm{d}\theta)\, \nu_t^\pi(\mathrm{d}h).$$

Hence $\mathbf{P}_t^\pi(\theta \in A, \cdot) \ll \mathbb{P}_t^\pi(\cdot)$ and the Radon-Nikodym derivative is the displayed Bayes ratio, which defines the kernel $R_t(A|h)$. Measurability and the fact that $R_t(\cdot|h)$ is a probability measure follow from standard properties of Radon–Nikodym derivatives. Independence of π is immediate from the explicit formula. □

In the following we also need to consider in what cases the mapping $h \mapsto R_t(\cdot|h)$ is weakly continuous. To that aim, we require a further assumption.

Assumption 4 (Likelihood continuity)**.** *The initial density $p_{\theta,0}(y)$ is jointly continuous and strictly positive in (θ, y). For every $s \geq 1$ the density $p_{\theta,s}(y|h_s, a_s)$ is jointly continuous in (θ, y, h_s, a_s) with strictly positive density. Hence, $(\theta, h_t) \mapsto \ell_t(\theta, h_t)$ is jointly continuous.*

Under this assumption, we have the following.

Lemma 2 (Weak continuity of the posterior)**.** *Consider Assumption 3 and Assumption 4. For each t we have that the mapping $h \mapsto R_t(\cdot|h),\ h \in \mathcal{H}_t$, is weakly continuous.*

Proof. Consider any sequence $(h_n)_n \in \mathcal{H}_t$ such that $h_n \to h,\ h \in \mathcal{H}_t$. Fix $f \in C_b(\Theta)$ (continuous and bounded). Since for $\mathbb{P}_t^\pi$-a.e. $h' \in \mathcal{H}_t$ we have $R_t(\mathrm{d}\theta|h') = \frac{\ell_t(\theta,h')\nu(\mathrm{d}\theta)}{\int_\Theta \ell_t(\theta,h')\nu(\mathrm{d}\theta)}$, we have that

$$\int_\Theta f(\theta) R_t(\mathrm{d}\theta|h_n) = \frac{\int_\Theta f(\theta)\ell_t(\theta, h_n)\nu(\mathrm{d}\theta)}{\int_\Theta \ell_t(\theta, h_n)\nu(\mathrm{d}\theta)}.$$

Now, since $\Theta \times (\{h\} \cup \{h_n : n \geq 1\})$ is compact, joint continuity bounds $f(\theta)\ell_t(\theta, h_n)$ uniformly in n, θ. Dominated convergence therefore gives

$$\int_\Theta f(\theta)\ell_t(\theta, h_n)\nu(\mathrm{d}\theta) \to \int_\Theta f(\theta)\ell_t(\theta, h)\nu(\mathrm{d}\theta) \quad \text{and} \quad \int_\Theta \ell_t(\theta, h_n)\nu(\mathrm{d}\theta) \to \int_\Theta \ell_t(\theta, h)\nu(\mathrm{d}\theta).$$

Since $\int_\Theta \ell_t(\theta, h)\nu(\mathrm{d}\theta) > 0$ by strict positivity of the density p_s, we have that

$$\int_\Theta f(\theta) R_t(\mathrm{d}\theta \mid h_n) \to \int_\Theta f(\theta) R_t(\mathrm{d}\theta \mid h),$$

so $h \mapsto R_t(\cdot|h)$ is weakly continuous. □

Optimal inference rule. Fix $\epsilon > 0$ and $t \in \mathbb{N}$. Recall that, for each $\theta \in \Theta$, the ϵ-optimal set is $\mathcal{X}_\epsilon(\theta) = \{x \in \mathcal{X} : L_\theta(x) \leq \epsilon\}$. Given a realized history $h \in \mathcal{H}_t$, define the *posterior success probability* of recommending $x \in \mathcal{X}$ as

$$q_t(h, x) \coloneqq \mathbb{P}_{\theta\sim\nu}^\pi\big(x \in \mathcal{X}_\epsilon(\theta) | H_t = h\big) = R_t\big(\{\theta \in \Theta : L_\theta(x) \leq \epsilon\}|h\big), \tag{14}$$

where $R_t(\cdot|h)$ is the posterior kernel from Lem. 1. We also define

$$r_t(h) \coloneqq \sup_{x\in\mathcal{X}} q_t(h, x). \tag{15}$$

Lemma 3. *Under Assumption 2, we have that $q_t(h, x)$ is jointly Borel measurable in (h, x) and upper semicontinuous in x for each fixed h. If in addition to Assumption 2 we also assume Assumption 4, then $q_t(h, x)$ is jointly upper semicontinuous.*

Proof. We prove the 3 properties separately.

- Regarding measurability, note that $(x, \theta) \mapsto \mathbf{1}\{L_\theta(x) \leq \epsilon\}$ is Boreal measurable by Assumption 2. Using that for every $A \in \mathcal{B}(\Theta)$ we have that $h \mapsto R_t(A|h)$ is Borel-measurable, then $q_t(h, x) = \int_\Theta \mathbf{1}\{L_\theta(x) \leq \epsilon\} R_t(\mathrm{d}\theta|h)$ is jointly measurable since integration preserves measurability.

- Consider now the u.s.c. property of $x \mapsto q_t(h, x)$ for each h. If, for every θ, the map $x \mapsto L_\theta(x)$ is lower semicontinuous on the compact set $\mathcal{X}$, then $\mathcal{X}_\epsilon(\theta)$ is closed and $x \mapsto \mathbf{1}\{x \in \mathcal{X}_\epsilon(\theta)\}$ is upper semicontinuous. Consequently, $x \mapsto q_t(h, x)$ is upper semicontinuous $\mathbb{P}_t^\pi$-a.s.. To see this, let $(x_n)_n$ be a sequence in $\mathcal{X}$ such that $x_n \to x$. Define $y_n = \mathbf{1}\{x_n \in \mathcal{X}_\epsilon(\theta)\}$. By Fatou's reverse lemma we have

$$\limsup_n \mathbb{E}_t[y_n | H_t = h] \leq \mathbb{E}_t[\limsup_n y_n | H_t = h] \leq \mathbb{P}_t(x \in \mathcal{X}_\epsilon(\theta) | H_t = h).$$

  where the last inequality follows from the fact that $\limsup_n y_n \leq \mathbf{1}\{x \in \mathcal{X}_\epsilon(\theta)\}$ from the upper semicontinuity. Thus the posterior is upper semicontinuous on $\mathbb{P}_t^\pi$-a.s.

- Define $C_\epsilon = \{(\theta, x) \in \Theta \times \mathcal{X} : L_\theta(x) \leq \epsilon\}$.By joint lower semicontinuity of L, we have that C_ϵ is closed. Consider any sequence $(h_n, x_n) \to (h, x)$ and define the distribution $\mu_n(\cdot) := R_t(\cdot|h_n) \otimes \delta_{x_n}$ and $\mu(\cdot) := R_t(\cdot|h) \otimes \delta_x$. Since $h \mapsto R_t(\cdot|h)$ is weakly continuous (Lem. 2, follows from Assumption 4), by Portmanteau's lemma we have

$$\limsup_{n \to \infty} \mu_n(C_\epsilon) \leq \mu(C_\epsilon).$$

  But $\mu_n(C_\epsilon) = \int_\Theta \mathbf{1}\{L_\theta(x_n) \leq \epsilon\} R_t(\mathrm{d}\theta|h_n) = q_t(h_n, x_n)$ and $\mu(C_\epsilon) = \int_\Theta \mathbf{1}\{L_\theta(x) \leq \epsilon\} R_t(\mathrm{d}\theta|h) = q_t(h, x)$, hence $(h, x) \mapsto q_t(h, x)$ is jointly upper semicontinuous.

□

Since $\mathcal{X}$ is compact, by the Extreme Value theorem we have that the supremum in (15) is attained (so one may replace sup by max). Under the joint upper-semicontinuity assumption, r_t is also upper semicontinuous: along convergent histories, take maximizing recommendations, extract a convergent subsequence by compactness of $\mathcal{X}$, and apply the joint upper-semicontinuity of q_t.

Proposition 3 (Optimal inference)**.** *Consider a fixed policy π and a fixed time $t \in \mathbb{N}$. Assume Assumption 3 and Assumption 2. Among measurable inference rules $I_t : \mathcal{H}_t \to \mathcal{X}$, the maximal value of $\mathbb{P}_{\theta\sim\nu}^\pi\big(I_t(H_t) \in \mathcal{X}_\epsilon(\theta)\big)$ is achieved by any rule satisfying, for $\mathbb{P}_t^\pi$-a.e. $h \in \mathcal{H}_t$,*

$$I_t(h) \in \arg\max_{x \in \mathcal{X}} q_t(h, x).$$

Furthermore, a measurable arg max *selector exists and $I_t(h)$ is measurable.*

Proof. Fix π and t, and let $\hat{x}_t := I_t(H_t)$. Using the posterior kernel,

$$\begin{aligned}
\mathbb{P}_{\theta\sim\nu}^\pi\big(\hat{x}_t \in \mathcal{X}_\epsilon(\theta)\big) &= \int \mathbf{1}\{\hat{x}_t \in \mathcal{X}_\epsilon(\theta)\}\, \mathbf{P}_t^\pi(\mathrm{d}\theta, \mathrm{d}h) \\
&= \int_{\mathcal{H}_t} \left[\int_\Theta \mathbf{1}\{I_t(h) \in \mathcal{X}_\epsilon(\theta)\}\, R_t(\mathrm{d}\theta|h)\right] \mathbb{P}_t^\pi(\mathrm{d}h) \\
&= \int_{\mathcal{H}_t} q_t\big(h, I_t(h)\big)\, \mathbb{P}_t^\pi(\mathrm{d}h) \\
&\leq \int_{\mathcal{H}_t} \sup_{x \in \mathcal{X}} q_t(h, x)\, \mathbb{P}_t^\pi(\mathrm{d}h) = \int_{\mathcal{H}_t} r_t(h)\, \mathbb{P}_t^\pi(\mathrm{d}h).
\end{aligned}$$

If $I_t(h) \in \arg\max_{x \in \mathcal{X}} q_t(h, x)$ for $\mathbb{P}_t^\pi$-a.e. h, then the inequality holds with equality, yielding the optimal value.

The existence of an arg max rule, and measurability of I_t, follows from an application of (Hernández-Lerma and Lasserre, 1996, Proposition D.5). □

We also note the following lower bound on the posterior q_t.

Lemma 4 (Markov lower bound on posterior success)**.** *Fix $\epsilon > 0$, $t \in \mathbb{N}$, and a realized history $h \in \mathcal{H}_t$. For any decision $x \in \mathcal{X}$, define the posterior success probability*

$$q_t(h, x) := \mathbb{P}_{\theta\sim\nu}^\pi\big(L_\theta(x) \leq \epsilon | H_t = h\big),$$

and the posterior mean loss

$$\bar{L}_t(h,x) \coloneqq \mathbb{E}^{\pi}_{\theta\sim\nu}\big[L_\theta(x)|H_t=h\big].$$

Then

$$q_t(h,x) \geq 1 - \frac{\bar{L}_t(h,x)}{\epsilon}.$$

Moreover, with the (clipped) shaped reward $r_\epsilon(\theta,x) \coloneqq \big[1-L_\theta(x)/\epsilon\big]_+$,

$$q_t(h,x) \geq \mathbb{E}^{\pi}_{\theta\sim\nu}\big[r_\epsilon(\theta,x)|H_t=h\big].$$

Proof. Since $L_\theta(x)\geq 0$, Markov's inequality yields

$$\mathbb{P}^{\pi}_{\theta\sim\nu}\big(L_\theta(x)>\epsilon|H_t=h\big) \leq \frac{\bar{L}_t(h,x)}{\epsilon}.$$

Taking complements gives the first claim. For the second, note that $\mathbf{1}\{L_\theta(x)\leq\epsilon\}\geq[1-L_\theta(x)/\epsilon]_+$ pointwise, and take conditional expectations. □

C.3 Fixed-confidence setting: formulation and optimal rules

We consider the fixed-confidence problem from Section 2 in its Bayesian (task-averaged) form. A learner is a triplet (π, I, τ) with sampling policy π, inference rule $I=(I_t)_{t\geq1}$, and stopping time τ. The objective is

$$\inf_{\pi,I,\tau}\ \mathbb{E}^{\pi}_{\theta\sim\nu}[\tau] \qquad \text{s.t.} \qquad \mathbb{P}^{\pi}_{\theta\sim\nu}\big(I_\tau(H_\tau)\in\mathcal{X}_\epsilon(\theta)\big)\ \geq\ 1-\delta, \quad \mathbb{E}^{\pi}_{\theta\sim\nu}[\tau]<\infty. \tag{16}$$

Throughout this section, $H_t=(Y_1,A_1,\ldots,A_{t-1},Y_t)$ is the history, $\hat{x}_\tau = I_\tau(H_\tau)$ and $\mathcal{F}_t=\sigma(H_t,U)$, as above.

Posterior success. For each t and realized history $h\in\mathcal{H}_t$, define the posterior success probability of recommending $x\in\mathcal{X}$ as

$$q_t(h,x)\ \coloneqq\ \mathbb{P}^{\pi}_{\theta\sim\nu}\big(x\in\mathcal{X}_\epsilon(\theta)|H_t=h\big),$$

and recall $r_t(h)\coloneqq\sup_{x\in\mathcal{X}}q_t(h,x)$. We have the following lemma that relates the success probability to the expected posterior success.

Lemma 5 (Stopped success as expected posterior success)**.** *For any policy π, almost-surely finite stopping time τ, and measurable inference rule I,*

$$\mathbb{P}^{\pi}_{\theta\sim\nu}\left(\hat{x}_\tau\in\mathcal{X}_\epsilon(\theta)\right)=\mathbb{E}^{\pi}\left[q_\tau\left(H_\tau,\hat{x}_\tau\right)\right].$$

Proof. For each t, $\{\tau=t\}$ and $I_t(H_t)$ are $\mathcal{F}_t$-measurable, and the posterior given $\mathcal{F}_t$ is $R_t(\cdot\mid H_t)$. Therefore,

$$\mathbb{E}^{\pi}[\mathbf{1}\{\tau=t\}\mathbf{1}\{L_\theta(I_t(H_t))\leq\epsilon\}]=\mathbb{E}^{\pi}[\mathbf{1}\{\tau=t\}q_t(H_t,I_t(H_t))].$$

Sum over t and use Tonelli's theorem and $\tau<\infty$ almost surely. □

Therefore, we have that $\mathbb{P}^{\pi}_{\theta\sim\nu}\big(I_\tau(H_\tau)\in\mathcal{X}_\epsilon(\theta)\big)=\mathbb{E}^{\pi}\left[q_\tau\big(H_\tau,I_\tau(H_\tau)\big)\right]$.

Lagrangian dual. Define, for $\lambda\geq0$, the Lagrangian value

$$V_\lambda(\pi,I,\tau)\coloneqq\mathbb{E}^{\pi}_{\theta\sim\nu}[\tau]\ +\lambda\Big((1-\delta)-\mathbb{P}^{\pi}_{\theta\sim\nu}\big(I_\tau(H_\tau)\in\mathcal{X}_\epsilon(\theta)\big)\Big).$$

Using Lem. 5, this can be written as

$$V_\lambda(\pi,I,\tau)=\lambda(1-\delta)+\mathbb{E}^{\pi}\left[\tau-\lambda q_\tau\big(H_\tau,I_\tau(H_\tau)\big)\right]. \tag{17}$$

The Lagrangian dual of (16) is then

$$\sup_{\lambda\geq0}\ \inf_{\pi,I,\tau}\ V_\lambda(\pi,I,\tau). \tag{18}$$

C.3.1 Optimal Inference Rule

Fix π, λ and t. Among all measurable inference rules $I_t : \mathcal{H}_t \to \mathcal{X}$, the maximal probability of ϵ-success at time t is achieved by any

$$I_t(h) \in \arg\max_{x \in \mathcal{X}} q_t(h, x),$$

equivalently, $q_t(h, I_t(h)) = r_t(h)$ for $\mathbb{P}_t^\pi$-a.e. h. (see Prop. 3.)

Since τ is adapted, plugging the optimal inference rule into Eq. (17) yields the simplified dual objective

$$\sup_{\lambda \geq 0} \inf_{\pi, \tau} \lambda(1 - \delta) + \mathbb{E}^\pi \left[\tau - \lambda r_\tau(H_\tau) \right]. \tag{19}$$

C.3.2 Stopping as an action (equivalence)

The additional optimization over stopping rules can be avoided by introducing an additional stopping aciton a_{stop}. Introduce an augmented action space $\bar{\mathcal{A}} := \mathcal{A} \cup \{a_{\text{stop}}\}$, where choosing a_{stop} terminates interaction (no new observation is collected). Let $\bar{\tau} := \inf\{t \geq 1 : A_t = a_{\text{stop}}\}$.

Lemma 6 (Embedding stopping times as a stop action)**.** *For every triplet (π, I, τ) with $\tau < \infty$ a.s., there exists a policy $\bar{\pi}$ on $\bar{\mathcal{A}}$ such that, under $\bar{\pi}$, $(\bar{\tau}, H_{\bar{\tau}})$ has the same distribution as (τ, H_τ) under π. In particular, for every $\lambda \geq 0$, $V_\lambda(\pi, I, \tau) = V_\lambda(\bar{\pi}, I, \bar{\tau})$.*

Proof. On $\{\tau \geq t\}$, let $B_t = a_{\text{stop}}$ if $\tau = t$, and $B_t = A_t$ otherwise. Define $\bar{\pi}_t(\cdot \mid h)$ as the regular conditional law of B_t given $H_t = h$ and $\tau \geq t$, choosing it arbitrarily at null histories. Such kernels exist on the standard Borel spaces. The posterior given the history, survival, and learner seeds is $R_t(\cdot \mid h)$, so the continuation observation kernel is still $\bar{P}_t$. Induction over t now shows equality of the surviving-history laws and of the stopping mass at each history. Thus $(\bar{\tau}, H_{\bar{\tau}})$ and (τ, H_τ) have the same law. Since the task posterior at each stopped history is also the same, the expected time and success agree, proving the Lagrangian identity. □

C.3.3 Optimal Policy

Lem. 6 shows that (for fixed λ) the inner problem in Eq. (19) can be viewed as an optimal-stopping control problem on the augmented action space: each continuation step incurs unit cost, while stopping at history $h \in \mathcal{H}_t$ incurs terminal cost $-\lambda\, r_t(h)$.

This subsection uses cost-to-go values, which are the negatives of the reward-to-go values in the main text; the minimization Bellman equation below is therefore equivalent to Eq. (5). Define the optimal cost-to-go (for fixed λ) from a history $h \in \mathcal{H}_t$ as

$$V_t^\star(h; \lambda) := \inf_{\bar{\pi} = (\bar{\pi}_i)_{i \geq t}} \mathbb{E}_{\theta \sim \nu}^{\bar{\pi}} \left[\sum_{s=t}^{\bar{\tau}-1} 1 - \lambda r_{\bar{\tau}}(H_{\bar{\tau}}) \,\middle|\, H_t = h \right], \tag{20}$$

where the infimum is over policies on $\bar{\mathcal{A}}$ and $\bar{\tau}$ is the first time a_{stop} is chosen. Similarly to (Russo et al., 2026), we can define the following optimal Q-functions

$$Q_{t,\text{stop}}^\star(h; \lambda) := -\lambda r_t(h), \qquad Q_{t,\text{cont}}^\star(h, a; \lambda) := 1 + \mathbb{E}\left[V_{t+1}^\star(H_{t+1}; \lambda) | H_t = h, A_t = a \right].$$

where the latter expectation is over the posterior mixture, defined as

$$\bar{P}_t(y' \in Y | H_t = h, A_t = a) = \int P_{\theta,t}(y' \in Y | H_t, A_t = a) R_t(\mathrm{d}\theta | H_t = h), \quad \forall Y \in \mathcal{B}(\mathcal{Y}).$$

Furthermore, similarly to (Russo et al., 2026), a standard decomposition yields the Bellman optimality relation

$$V_t^\star(h; \lambda) = \min \left\{ Q_{t,\text{stop}}^\star(h; \lambda), \inf_{a \in \mathcal{A}} Q_{t,\text{cont}}^\star(h, a; \lambda) \right\}. \tag{21}$$

However, in order to guarantee that the infimum , $\inf_{a \in \mathcal{A}} Q_{t,\text{cont}}^\star(h, a; \lambda)$ is attained, since $\mathcal{A}$ is compact, we need to guarantee that the Q-value is lower semicontinuous. We begin by showing that $V_t^\star$ is lower semicontinuous. To that aim, we need some results first. We begin by showing that the mixture posterior is weakly continuous.

Lemma 7 (Weak continuity of the mixture posterior)**.** *Fix t and $h \in \mathcal{H}_t$. Let $R_t(\cdot|h)$ be the posterior on Θ and define the posterior predictive kernel*

$$\bar{P}_t(\cdot|h,a) \coloneqq \int_\Theta P_{\theta,t}(\cdot|h,a)\, R_t(\mathrm{d}\theta|h).$$

Under Assumption 1, $a \mapsto \bar{P}_t(\cdot|h,a)$ is weakly continuous. If in addition we assume Assumption 4, then $(h,a) \mapsto \bar{P}_t(\cdot|h,a)$ is jointly weakly continuous.

Proof. We prove weak continuity in the action first. Fix $f \in C_b(\mathcal{Y})$ (continuous and bounded) and a sequence $(a_n)_n$ such that $a_n \to a$. For each θ, define

$$g_n(\theta) \coloneqq \int f(y)\, P_{\theta,t}(\mathrm{d}y|h,a_n), \qquad g(\theta) \coloneqq \int f(y)\, P_{\theta,t}(\mathrm{d}y|h,a).$$

By Assumption 1,

$$g_n(\theta) \to g(\theta).$$

Moreover, $\left|\int f(y)\, P_{\theta,t}(\mathrm{d}y|h,a_n)\right| \le \|f\|_\infty < \infty$ for all θ, n, and thus is also bounded. By dominated convergence,

$$\begin{aligned}
\int f(y)\, \bar{P}_t(\mathrm{d}y|h,a_n) &= \int_\Theta g_n(\theta)\, R_t(\mathrm{d}\theta \mid h), \\
&\to \int_\Theta \underbrace{\left(\int f(y)\, P_{\theta,t}(\mathrm{d}y|h,a)\right)}_{=g(\theta)} R_t(\mathrm{d}\theta \mid h), \\
&= \int f(y)\bar{P}_t(\mathrm{d}y|h,a). \qquad \text{(Fubini-Tonelli)}
\end{aligned}$$

Regarding the second part, it is less straightforward to prove. We assume Assumption 4. Recall that $R_t(\mathrm{d}\theta \mid h) = \frac{\ell_t(\theta,h)\nu(\mathrm{d}\theta)}{\int_\Theta \ell_t(\theta,h)\nu(\mathrm{d}\theta)}$. We omit the normalization constant for simplicity, and simply include it in ℓ_t.

Define some sequence $(h_n, a_n) \to (h,a)$, and consider

$$\int f(y)\bar{P}_t(\mathrm{d}y \mid h_n,a_n) = \int_\Theta \underbrace{\int f(y) P_{\theta,t}(\mathrm{d}y \mid h_n,a_n)}_{=:G_f(h_n,a_n,\theta)} \ell_t(\theta,h_n)\nu(\mathrm{d}\theta) = \int_\Theta G_f(h_n,a_n,\theta)\ell_t(\theta,h_n)\nu(\mathrm{d}\theta).$$

Clearly G_f is bounded and ℓ_t is jointly continuous by assumption. Since Θ is compact, we have that along any convergence sequence $h_n \to h$, the likelihood ℓ_t is uniformly bounded. If G_f is jointly continuous in (h,a), then, by dominated convergence, we obtain

$$\int_\Theta G_f(h_n,a_n,\theta)\ell_t(\theta,h_n)\nu(\mathrm{d}\theta) \to \int_\Theta G_f(h,a,\theta)\ell_t(\theta,h)\nu(\mathrm{d}\theta) = \int f(y)\bar{P}_t(\mathrm{d}y \mid h,a),$$

which shows the claim. Hence, we need to show that $(h,a) \mapsto P_{\theta,t}(\cdot \mid h,a)$ is weakly continuous for each θ.

Using that $\mathrm{d}P_{\theta,t}(\cdot \mid h,a) = p_{\theta,t}(y \mid h,a)\mathrm{d}\lambda$, and recalling that by Assumption 4 $p_{\theta,t}$ is jointly continuous. We have

$$\int f(y)P_{\theta,t}(\mathrm{d}y \mid h_n,a_n) = \int f(y)p_{\theta,t}(y \mid h_n,a_n)\mathrm{d}\lambda$$

using again compactness (of $\mathcal{Y}$) and continuity of the arguments, we derive weak continuity of $(h,a) \mapsto P_{\theta,t}(\cdot \mid h,a)$, which concludes the proof.

□

The main technical challenge in the continuous case is showing that the infimum $\inf_{a\in\mathcal{A}} Q^\star_{t,\mathrm{cont}}(h,a)$ in the Bellman equation is attained. In finite $\mathcal{A}$ this is trivial; in compact continuous $\mathcal{A}$ it requires

lower semicontinuity of $a \mapsto Q^\star_{t,\mathrm{cont}}(h,a)$. We establish this through a value-iteration construction that propagates lower semicontinuity from the observation model through the posterior predictive to the Q-function. The proof adapts the general template of negative dynamic programming (Bertsekas and Shreve, 1996; Hernández-Lerma and Lasserre, 1996) to our stop/continue structure, where the stopping payoff involves $r_t(h) = \sup_x q_t(h,x)$, itself a supremum over a continuous set whose upper semicontinuity must be established first (Lem. 3). We build a sequence of values $W_t^{(n)}(h)$ and show that these are l.s.c. in h.

Starting from $V_t^\star$, we construct $W_t^\star$ and build an increasing sequence $W_t^{(n)}$ from below. We show that each $W_t^{(n)}$ is l.s.c. in h and that for each fixed h, the map $a \mapsto Q_t^{(n)}(h,a)$ is l.s.c. on $\mathcal{A}$. We then show that $W_t^{(n)}$ approaches $W_t^\star$. We conclude by showing that $W_t^\star = V^\star + \lambda$, proving that $V_t^\star$ is $l.s.c.$

Define $W_t^{(0)}(h) \coloneqq 0$ for all t, h. Recursively for $n \geq 0$, define

$$
\begin{aligned}
Q_t^{(n)}(h,a) &\coloneqq 1 + \int_{\mathcal{Y}} W_{t+1}^{(n)}(h,a,y)\bar{P}_t(\mathrm{d}y \mid h,a),\\
C_t^{(n)}(h) &\coloneqq \inf_{a\in\mathcal{A}} Q_t^{(n)}(h,a),\\
W_t^{(n+1)}(h) &\coloneqq \min\{\lambda - \lambda r_t(h); C_t^{(n)}(h)\}.
\end{aligned}
$$

Also define $W_t^\star(h) = V_t^\star(h;\lambda) + \lambda$. We also define the operator $\mathcal{T}_t$ as follows:

$$
(\mathcal{T}_t u)(h) \coloneqq \min\left\{\lambda(1 - r_t(h)), 1 + \inf_{a\in\mathcal{A}} \int_{\mathcal{Y}} u(h,a,y)\bar{P}_t(\mathrm{d}y \mid h,a)\right\},
$$

therefore $W_t^{(n+1)} = \mathcal{T}_t W_{t+1}^{(n)}$. One can clearly show that the operator is monotone, as for $u(h) \leq v(h)$ we have $(\mathcal{T}_t u)(h) \leq (\mathcal{T}_t v)(h)$. Therefore $U_t(h) \coloneqq \sup_{n\geq 0} W_t^{(n)}(h)$ exists pointwise, and satisfies $0 \leq U_t(h) \leq \lambda$.

Lemma 8. *Assume Assumption 2 and Assumption 4. For every $t \in \mathbb{N}$, the iterates $W_t^{(n)}$ and $C_t^{(n)}$ are lower semicontinuous, and for each fixed h, $a \mapsto Q_t^{(n)}(h,a)$ is lower semicontinuous on $\mathcal{A}$ and attains its minimum.*

Proof. Let $g_t(h) = \lambda(1 - r_t(h))$ and fix t. We prove the argument by induction on n: for a fixed n, $W_s^{(n)}$ is l.s.c. for every s.

Base Step. Clearly, for all t, $W_t^{(0)}$ is l.s.c. , and since $Q_t^{(0)}(h,a) = 1$ and $C_t^{(0)}(h) = 1$, we have $W_t^{(1)}(h) = \min(g_t(h), 1)$ which is l.s.c. if $-r_t$ is l.s.c. (which is, by Lem. 3).

Then, assume the induction hypothesis. We use that $W_{t+1}^{(n)}$ is l.s.c. and prove that $W_t^{(n+1)}$ is l.s.c. We do so in 3 steps: first we prove that $a \mapsto Q_t^{(n)}(h,a)$ is l.s.c. Then, we prove that $h \mapsto C_t^{(n)}(h)$ is l.s.c. Lastly we show that $W_t^{(n+1)}$ is l.s.c.

Step 1. Fix h and consider a sequence $a_m \to a$. Define $\mu_m \coloneqq \delta_{a_m} \otimes \bar{P}_t(\cdot \mid h, a_m)$ and $\mu \coloneqq \delta_a \otimes \bar{P}_t(\cdot \mid h, a)$. By Lem. 7 we have that $a \to \bar{P}_t(\cdot \mid h, a)$ is weakly continuous, and thus $\mu_m \Rightarrow \mu$. Since $W_{t+1}^{(n)}$ is l.s.c. and bounded, the Portmanteau theorem yields

$$
\liminf_{m\to\infty} \int_{\mathcal{Y}\times\mathcal{A}} W_{t+1}^{(n)}(h,a',y)\mathrm{d}\mu_m \geq \int_{\mathcal{Y}\times\mathcal{A}} W_{t+1}^{(n)}(h,a',y)\mathrm{d}\mu.
$$

therefore $a \mapsto Q_t^{(n)}(h,a)$ is l.s.c. on compact $\mathcal{A}$, and the infimum is attained.

Step 2. We now show that $h \mapsto C_t^{(n)}(h)$ is l.s.c. Define a sequence $(h_m)_m$ such that $h_m \to h$. Choose a subsequence $(h_{m_k})_k$ such that $C_t^{(n)}(h_{m_k}) \to \liminf_{m\to\infty} C_t^{(n)}(h_m)$. For each k choose $a_{m_k} \in \arg\min_{a\in\mathcal{A}} Q_t^{(n)}(h_{m_k}, a)$. Since $\mathcal{A}$ is compact, by passing to a further subsequence if necessary, we may assume $a_{m_k} \to a$. Define $\mu_{m_k} \coloneqq \delta_{(h_{m_k}, a_{m_k})} \otimes \bar{P}_t(\cdot \mid h_{m_k}, a_{m_k})$ and $\mu \coloneqq$

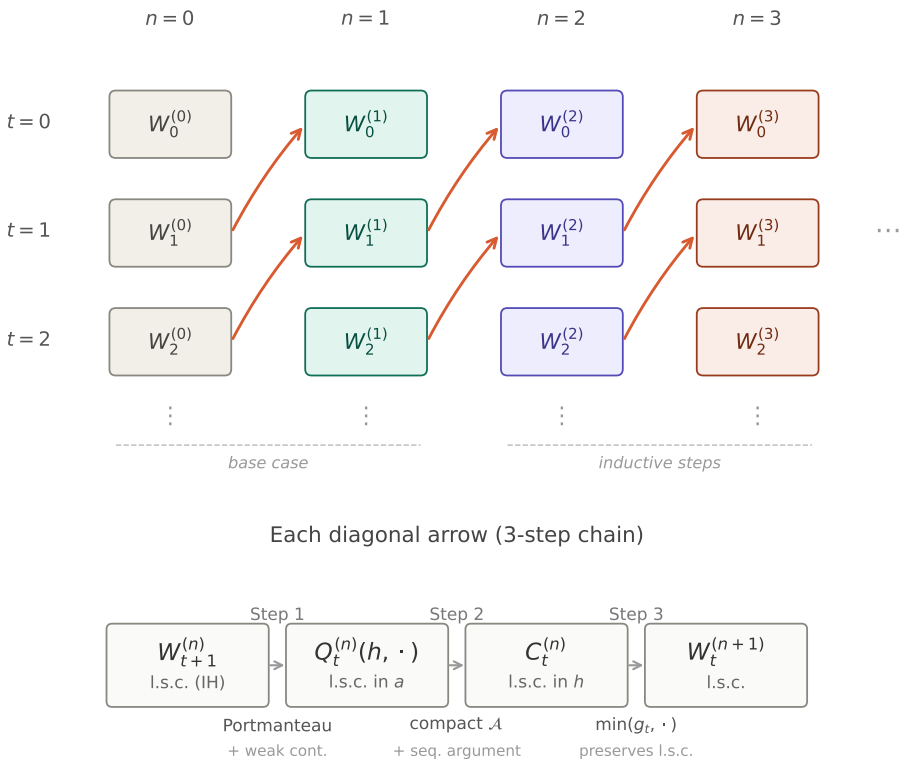


Figure 5: Induction diagram used in the proof of Lem. 8.

$\delta_{(h,a)} \otimes \bar{P}_t(\cdot \mid h, a)$. Since by assumption we have that $(h, a) \mapsto \bar{P}_t(\cdot \mid h, a)$ is jointly weakly continuous, then $\mu_{m_k} \Rightarrow \mu$. Similarly to before, we obtain

$$\liminf_{k\to\infty} Q_t^{(n)}(h_{m_k}, a_{m_k}) \geq Q_t^{(n)}(h, a) \geq C_t^{(n)}(h).$$

But $Q_t^{(n)}(h_{m_k}, a_{m_k}) = C_t^{(n)}(h_{m_k})$, therefore $\liminf_{m\to\infty} C_t^{(n)}(h_m) = \lim_{k\to\infty} C_t^{(n)}(h_{m_k}) \geq C_t^{(n)}(h)$. Therefore $C_t^{(n)}$ is l.s.c. in h.

Step 3. Lastly, consider $W_t^{(n+1)} = \min(g_t(h), C_t^{(n)}(h))$. Since both arguments are l.s.c., then $W_t^{(n+1)}$ is l.s.c. Since t was arbitrary, the statements holds for all t.

□

Then, since each $W_t^{(n)}$ is l.s.c., we get that $U_t(h)$ is l.s.c. (arbitrary suprema of l.s.c. functions are l.s.c.).

Lemma 9. *Assume Assumption 2 and Assumption 4. For every $t \in \mathbb{N}$, define*

$$Q_t^U(h, a) \coloneqq 1 + \int_{\mathcal{Y}} U_{t+1}(h, a, y)\bar{P}_t(\mathrm{d}y \mid h, a), \qquad C_t^U(h) \coloneqq \inf_{a\in\mathcal{A}} Q_t^U(h, a),$$

then $U_t(h) = \min\{g_t(h), C_t^U(h)\}$. Furthermore, we have that $a \mapsto Q_t^U(h, \cdot)$ is lower semicontinuous and $Q_t^U(h, a)$ is jointly Borel measurable.

Proof. Let $g_t(h) = \lambda(1 - r_t(h))$ and fix t. Since $W_{t+1}^{(n)} \uparrow U_{t+1}$ and $W_{t+1}^{(n)} \geq 0$, by monotone convergence for each (h, a) we have $Q_t^{(n)} \uparrow Q_t^U$. By Lem. 8, each $Q_t^{(n)}(h, \cdot)$ is l.s.c., thus Q_t^U is

also l.s.c. being the suprema of l.s.c. functions. Furthermore, since each $U_{t+1}(h, a, y)$ is l.s.c., and $(h, a) \mapsto \bar{P}_t(\cdot \mid h, a)$ is a Borel kernel, then $Q_t^U(h, a)$ is jointly Borel measurable.

Next, we show that $\sup_n \inf_a Q_t^{(n)}(h, a) = \inf_a \sup_n Q_t^{(n)}(h, a) = C_t^U(h)$. To show this, let $m_n = \inf_a Q_t^{(n)}(h, a)$ and $m = \sup_n m_n$. First, note that by monotonicity for all N we have $m_n = \inf_a Q_t^{(n)}(h, a) \leq \inf_a Q_t^U(h, a) = C_t^U(h)$, and thus $m \leq C_t^U(h)$. Now, choose some minimizers $a_n \in \arg\min_a Q_t^{(n)}(h, a)$. By compactness, there is some subsequence satisfying $a_{n_k} \to a$. Choose some integers i_0, then by monotonicity for all large k we have

$$Q_t^{(i_0)}(h, a_{n_k}) \leq Q_t^{(n_k)}(h, a_{n_k}) = m_{n_k}.$$

Therefore $\liminf_k m_{n_k} = m \geq Q_t^{(i_0)}(h, a)$. Since this holds for any i_0, we have $m \geq \sup_{i_0} Q_t^{(i_0)}(h, a) = Q_t^U(h, a) \geq C_t^U(h)$, which shows that $\sup_n \inf_a Q_t^{(n)}(h, a) = \inf_a \sup_n Q_t^{(n)}(h, a) = C_t^U(h)$.

Hence, we have that $U_t(h) = \sup_n W_t^{(n+1)}(h) = \sup_n \min\{g_t(h), C_t^{(n)}(h)\}$. Since $C_t^{(n)}(h) \uparrow C_t^U(h)$, we have $U_t(h) = \min\{g_t(h), C_t^U(h)\}$. □

Finally, we show that U_t is actually the true optimal value $W_t^\star$.

Theorem C.1. *Assume Assumption 2 and Assumption 4. For every $t \in \mathbb{N}$, we have that $U_t(h) = V_t^\star(h) + \lambda$, and $V_t^\star$ is l.s.c.*

Proof. Let $g_t(h) = \lambda(1 - r_t(h))$. We first show that U_t is actually a lower bound on the cost of any policy. Then we construct a policy that achieves equality.

Step 1 (lower bound). Consider any admissible policy π. At any history h at time t,

1. if π stops, then $U_t(h) \leq g_t(h)$ since $U_t(h) = \min\{g_t(h), C_t^U(h)\}$.
2. if π continues with some action a, then

$$U_t(h) \leq 1 + \int_{\mathcal{Y}} U_{t+1}(h, a, y) \bar{P}_t(\mathrm{d}y \mid h, a),$$

since $C_t^U(h) \leq Q_t^U(h, a)$.

Let τ be the first timestep the policy π decides to stop. Then, iterating up to $\min(\tau, t + N)$

$$U_t(h) \leq \mathbb{E}^\pi \left[\sum_{s=t}^{\min(\tau, t+N)-1} 1 + \mathbf{1}_{\{\tau \leq t+N\}} g_\tau(H_\tau) + \mathbf{1}_{\{\tau > t+N\}} U_{t+N}(H_{t+N}) \middle| H_t = h \right] =: J_N^\pi(h).$$

If the policy achieves infinite cost, then the inequality is true for all N since $U_t(h)$ is bounded. Then, consider the case where the cost is finite as $N \to \infty$. Since $g_\tau \geq 0$, this implies $\mathbb{E}^\pi[\tau - t \mid H_t = h] < \infty$, therefore $\tau < \infty$ almost surely. Consequently, by dominated convergence the remainder term $+\mathbf{1}_{\{\tau > t+N\}} U_{t+N}(H_{t+N})$ vanishes as $N \to \infty$. Therefore, letting $N \to \infty$ we obtain $U_t(h) \leq \lim_{N \to \infty} J_N^\pi(h) \coloneqq J^\pi(h)$, and taking infimum over π we get $U_t(h) \leq W_t^\star(h)$.

Step 2 (equality). From Lem. 9, $Q_t^U(h, a)$ is l.s.c. on $\mathcal{A}$ for each h, and $\mathcal{A}$ is compact. Furthermore $Q_t^U(h, a)$ is jointly Borel-measurable. Then, there exists a measurable selector $a_t^\star(h) \in \arg\min_a Q_t^U(h, a)$ (Feinberg and Kasyanov, 2021) . Then both $C_t^U(h) = Q_t^U(h, a_t^\star(h))$ is Borel measurable. Consider then a policy $\pi^\star$ that stops at h if $g_t(h) \leq C_t^U(h)$, and otherwise continue with $a_t^\star(h)$. Denote by τ^* this stopping rule. Along this policy, the Bellman minimum is attained with equality at every step, therefore

$$U_t(h) = \mathbb{E}^{\pi^*} \left[\sum_{s=t}^{\min(\tau^*, t+N)-1} 1 + \mathbf{1}_{\{\tau^* \leq t+N\}} g_{\tau^*}(H_{\tau^*}) + \mathbf{1}_{\{\tau^* > t+N\}} U_{t+N}(H_{t+N}) \middle| H_t = h \right].$$

Since $U_t(h)$ is bounded, we have

$$\lambda \geq U_t(h) \geq \mathbb{E}^{\pi^*}[\min(\tau^*, t+N) - t \mid H_t = h].$$

Letting $N \to \infty$, we obtain $\mathbb{E}^{\pi^*}[\tau - t \mid H_t = h] \leq \lambda$. Therefore the remainder term $\mathbf{1}_{\{\tau^* > t+N\}} U_{t+N}$ vanishes as $N \to \infty$ by dominated convergence since $U_{t+N} \leq \lambda$ and $\tau^* < \infty$ almost surely. Therefore we obtain $U_t(h) = J^{\pi^*}(h)$. Since $W^\star(h)$ is the minimal cost we obtain $W_t^\star(h) \leq U_t(h)$, but from the previous step we also have $U_t(h) \leq W_t^\star(h)$. Therefore $U_t(h) = W_t^\star(h) = V_t^\star(h;\lambda) + \lambda$. Finally, $V_t^\star$ is l.s.c. since U_t is l.s.c. □

Hence, we conclude with the following result

Proposition 4 (Actor over $\mathcal{A}$ and stopping action)**.** *Assume Assumption 2 and Assumption 4. and let $a^\star(h) \in \arg\min_{a \in \mathcal{A}} Q^\star_{t,\mathrm{cont}}(h,a)$. Then an optimal policy can be implemented by: (i) selecting $a^\star(h)$ as the continuation action, and (ii) stopping iff $Q^\star_{t,\mathrm{stop}}(h) \leq Q^\star_{t,\mathrm{cont}}(h, a^\star(h))$.*

Proof. By (21), at each history h the optimal action is whichever attains the minimum between $Q^\star_{t,\mathrm{stop}}(h)$ and $\inf_{a \in \mathcal{A}} Q^\star_{t,\mathrm{cont}}(h,a)$. If $a^\star(h)$ attains the infimum over $\mathcal{A}$, then the comparison $Q^\star_{t,\mathrm{stop}}(h) \leq Q^\star_{t,\mathrm{cont}}(h, a^\star(h))$ is equivalent to $Q^\star_{t,\mathrm{stop}}(h) \leq \inf_a Q^\star_{t,\mathrm{cont}}(h,a)$, i.e., stopping is optimal. Otherwise continuing with $a^\star(h)$ is optimal.

□

This last result shows that an algorithm can act in the augmented space $\bar{\mathcal{A}}$ while never explicitly parameterizing a "stop action" in the actor.

Hence an optimal policy can be implemented by:

1. choose a continuation action $a^\star(h) \in \arg\min_{a \in \mathcal{A}} Q^\star_{t,\mathrm{cont}}(h,a)$ (this is the actor over $\mathcal{A}$ only);
2. stop if $Q^\star_{t,\mathrm{stop}}(h) \leq Q^\star_{t,\mathrm{cont}}(h, a^\star(h))$, otherwise continue with $a^\star(h)$.

This is mathematically equivalent to having a single policy over $\bar{\mathcal{A}}$ that selects $\arg\min\{Q^\star_{t,\mathrm{stop}}(h), \min_a Q^\star_{t,\mathrm{cont}}(h,a)\}$, but it decomposes the decision into (i) a continuous control over $\mathcal{A}$ and (ii) an optimal-stopping comparison against a scalar stop value. Therefore, one can learn a Q-value for the stop decision and comparing it to the Q-value of the actor-chosen action: this is consistent with the optimal control structure of (21), provided the actor approximates the minimizer of $Q^\star_{t,\mathrm{cont}}(h,\cdot)$ and the critic estimates are calibrated.

C.3.4 Reward Shaping and Removal of the Stop Action

The stop action need not be explicitly parameterized by the actor. Indeed, for fixed $\lambda > 0$, the Lagrangian objective

$$\mathbb{E}^{\pi}\left[\sum_{s=t}^{\tau-1} 1 - \lambda r_\tau(H_\tau) \mid H_t = h\right]$$

is equivalent, up to an additive constant independent of the policy, to maximizing the shaped return obtained from the one-step reward

$$R_s^{\mathrm{sh}} := -1 + \lambda\left(r_{s+1}(H_{s+1}) - r_s(H_s)\right).$$

Thus the actor may be restricted to continuation actions $a \in \mathcal{A}$, while the stopping decision is implemented by comparing the best continuation advantage against the stop value, which is zero in the shaped formulation. Equivalently, using the rescaled reward

$$-\frac{1}{\lambda} + r_{s+1}(H_{s+1}) - r_s(H_s)$$

gives the same optimal policies.

Lemma 10 (Reward shaping)**.** *Fix $\lambda > 0$. For a continuation policy π and stopping time τ, define the original Lagrangian cost*

$$J_t^\lambda(h;\pi,\tau) \coloneqq \mathbb{E}^\pi\left[\sum_{s=t}^{\tau-1} 1 - \lambda r_\tau(H_\tau)\middle| H_t = h\right],$$

and the shaped return

$$G_t^\lambda(h;\pi,\tau) \coloneqq \mathbb{E}^\pi\left[\sum_{s=t}^{\tau-1} R_s^{\mathrm{sh}}(H_s,A_s,H_{s+1})\middle| H_t = h\right].$$

Then

$$\arg\min_{\pi,\tau} J_t^\lambda(h;\pi,\tau) = \arg\max_{\pi,\tau} G_t^\lambda(h;\pi,\tau).$$

Proof. First, the shaped reward telescopes:

$$\sum_{s=t}^{\tau-1} R_s^{\mathrm{sh}}(H_s,A_s,H_{s+1}) = \sum_{s=t}^{\tau-1}\left[\lambda\left(r_{s+1}(H_{s+1}) - r_s(H_s)\right) - 1\right] = \lambda r_\tau(H_\tau) - \lambda r_t(H_t) - \sum_{s=t}^{\tau-1} 1.$$

Conditioning on $H_t = h$ gives

$$G_t^\lambda(h;\pi,\tau) = \mathbb{E}^\pi\left[\lambda r_\tau(H_\tau) - \sum_{s=t}^{\tau-1} 1 \;\middle|\; H_t = h\right] - \lambda r_t(h).$$

Since

$$J_t^\lambda(h;\pi,\tau) = \mathbb{E}^\pi\left[\sum_{s=t}^{\tau-1} 1 - \lambda r_\tau(H_\tau) \;\middle|\; H_t = h\right],$$

we obtain

$$G_t^\lambda(h;\pi,\tau) = -J_t^\lambda(h;\pi,\tau) - \lambda r_t(h).$$

The term $-\lambda r_t(h)$ does not depend on (π,τ), so maximizing G_t^λ is equivalent to minimizing J_t^λ.

□

Using this reward shaping, we do not need to learn the stopping Q-value, as shown in the next proposition. The actor only needs to output a continuation action $a \in \mathcal{A}$. The stopping rule is a scalar gate:

$$\text{continue iff } Q_{t,\mathrm{cont}}^{\mathrm{sh}}(h, a_t(h)) > 0.$$

Equivalently, the stop value in the shaped formulation is identically zero, and the state-value target is

$$S_t(h) = \max\{0, Q_{t,\mathrm{cont}}^{\mathrm{sh}}(h, a_t(h))\}.$$

Proposition 5 (Reward shaping does not require learning a stopping Q-value)**.** *Let*

$$S_t^\star(h) \coloneqq \sup_{\pi,\tau} G_t^\lambda(h;\pi,\tau)$$

be the optimal shaped value. Then

$$S_t^\star(h) = \max\left\{0, \sup_{a\in\mathcal{A}} \mathbb{E}\left[-1 + \lambda\left(r_{t+1}(H_{t+1}) - r_t(h)\right) + S_{t+1}^\star(H_{t+1})\middle| H_t = h, A_t = a\right]\right\}.$$

Define the shaped continuation Q-value

$$Q_{t,\mathrm{cont}}^{\mathrm{sh}}(h,a) \coloneqq -1 + \mathbb{E}\left[\lambda\left(r_{t+1}(H_{t+1}) - r_t(h)\right) + S_{t+1}^\star(H_{t+1})\middle| H_t = h, A_t = a\right].$$

If the supremum over $a \in \mathcal{A}$ is attained by a measurable selector $a_t^\star(h) \in \arg\max_{a\in\mathcal{A}} Q_{t,\mathrm{cont}}^{\mathrm{sh}}(h,a)$, then an optimal policy is implemented by

$$\textit{stop at } h \quad\Longleftrightarrow\quad Q_{t,\mathrm{cont}}^{\mathrm{sh}}(h, a_t^\star(h)) \le 0,$$

and otherwise continuing with $a_t^\star(h)$.

Proof. Next, in the shaped formulation, stopping immediately produces the empty sum and hence shaped return 0. If instead the learner continues with action a, it receives the immediate shaped reward

$$\lambda\big(r_{t+1}(H_{t+1}) - r_t(h)\big) - 1$$

and then obtains the optimal future shaped value $S^\star_{t+1}(H_{t+1})$. Therefore

$$S^\star_t(h) = \max\left\{0,\ \sup_{a\in\mathcal{A}} Q^{\mathrm{sh}}_{t,\mathrm{cont}}(h,a)\right\}.$$

Thus stopping is optimal exactly when

$$\sup_{a\in\mathcal{A}} Q^{\mathrm{sh}}_{t,\mathrm{cont}}(h,a) \le 0.$$

If $a^\star_t(h)$ attains the supremum, this is equivalent to

$$Q^{\mathrm{sh}}_{t,\mathrm{cont}}(h, a^\star_t(h)) \le 0.$$

Otherwise, if

$$Q^{\mathrm{sh}}_{t,\mathrm{cont}}(h, a^\star_t(h)) > 0,$$

continuing with $a^\star_t(h)$ attains the continuation branch and is optimal. □

C.4 Fixed-confidence setting: Bayesian (ϵ,δ)-correctness and zero-duality gap

In this part of the appendix we study the correctness property of dual solutions, as well as investigate the zero-duality gap. We begin by stating the main result, and then proceed to discuss correctness and analyse the zero-duality gap in the subsequent two subsections.

C.4.1 Main Result

We now state the main result concerning correctness and the zero-duality gap. The perturbation argument provez zero gap and primal recovery; Thm. C.3 gives a separate subgradient argument for feasibility.

Let $c(\pi) := \mathbb{E}^\pi[\tau_\pi]$, $\rho(\pi) := \mathbb{E}^\pi[r_{\tau_\pi}(H_{\tau_\pi})]$, and $\mathcal{K} := \{(c(\pi),\rho(\pi)) : \pi \in \mathcal{T}\}$, where τ_π is the stopping time of a policy whose action space embeds the stopping decision. The time-sharing assumption, stated formally in Assumption 7, says that ex-ante randomization between two admissible policies remains admissible and therefore convexifies $\mathcal{K}$.

Theorem C.2 (Zero duality gap and Bayesian (ϵ,δ)-correctness)**.** *Assume time-sharing (Assumption 7), and assume strict feasibility (Assumption 8), i.e., there exists a feasible $\pi_{\mathrm{sf}} \in \mathcal{T}$ such that $\rho(\pi_{\mathrm{sf}}) > 1-\delta$. Then, the duality gap is zero. Furthermore, let $g(\lambda) := \inf_{\pi\in\mathcal{T}}\{c(\pi) + \lambda(1-\delta-\rho(\pi))\}$, $\lambda^\star \in \arg\max_{\lambda\ge 0} g(\lambda)$, and let $\mathcal{S}(\lambda^\star)$ be the set of dual minimizers at $\lambda^\star$. If there exists $\epsilon_0 > 0$ such that*

$$\mathcal{K}_{\epsilon_0}(\lambda^\star) := \{(c,\rho)\in\mathcal{K} : c + \lambda^\star(1-\delta-\rho) \le g(\lambda^\star) + \epsilon_0\}$$

is closed in $\mathbb{R}^2$, then there exists a dual-optimal policy $\pi^\star \in \mathcal{S}(\lambda^\star)$ that is primal optimal. Consequently, with the posterior-optimal inference rule $I^\star_t(h) \in \arg\max_{x\in\mathcal{X}} q_t(h,x)$,

$$\mathbb{P}^{\pi^\star}_{\theta\sim\nu}\left(L_\theta\left(I^\star_{\tau_{\pi^\star}}(H_{\tau_{\pi^\star}})\right) \le \epsilon\right) \ge 1-\delta.$$

This theorem improves on the corresponding result in (Russo et al., 2026) in two ways. First, we replace the assumption that the dual-optimal policy is unique with a local closedness condition on the near-optimal set $\mathcal{K}_{\epsilon_0}(\lambda^\star)$: this allows multiple dual-optimal policies, which is more natural. Second, while (Russo et al., 2026) derives correctness via a monotonicity argument on the optimal cost, our proof (Sec. C.5) uses a direct subdifferential characterization to show that if all near-optimal policies have $\rho < 1-\delta$, then every subgradient of the negative dual value is strictly negative, contradicting optimality. Zero duality gap follows from a standard perturbation argument (Rockafellar and Wets, 1998; Borwein and Lewis, 2006); see Sec. C.5.1.

C.5 Bayesian Correctness of Dual-Optimal Points

In this section we provide the Bayesian guarantees for the continuous ICPE setting. Compared to the correctness argument in (Russo et al., 2026), our proof differs in three aspects, each addressing a limitation of the finite-space analysis.

1. **Posterior object.** The finite ICPE framework uses $\mathbb{P}(x^\star = x \mid H_t)$, which is ill-defined in continuous $\mathcal{X}$. We replace it with the posterior success probability $r_t(h) = \sup_{x\in\mathcal{X}} \mathbb{P}(L_\theta(x) \le \epsilon \mid H_t = h)$. This is the natural continuous analogue and the only change forced by the setting.
2. **Non-singleton dual optima (main technical improvement).** Russo et al. (2026) assume the dual-optimal policy set $\mathcal{S}(\lambda)$ is a singleton for each λ. We relax this to local closedness of the near-optimal cost-reward set $\mathcal{K}_{\epsilon_0}(\lambda^\star)$ (Assumption 6). This is meaningful in continuous spaces where the policy class is richer and uniqueness is harder to verify. The key proof step is showing that if all policies in $\mathcal{K}_{\epsilon_0}(\lambda^\star)$ have $\rho < 1-\delta$, then a uniform gap propagates to all near-optimal policies (Step 2 below), which requires the closedness assumption in an essential way.
3. **Subdifferential argument (alternative proof technique).** Rather than deriving a contradiction from monotonicity of the optimal cost in λ (as in (Russo et al., 2026)), we use the subdifferential characterization of Hantoute and López (2008) directly. This avoids the monotonicity argument entirely: we show that every subgradient of $f(\lambda) = -g(\lambda)$ at $\lambda^\star$ is strictly negative, contradicting $0 \in \partial f(\lambda^\star)$. This argument works with multiple dual optima without modification.

We begin by stating our assumptions. To do so, let Π denote the class of admissible policies on the augmented action space $\bar{\mathcal{A}}$, and for each π let τ_π be the corresponding stopping time, i.e., the first tiem the policy chooses the stop action. Define

$$\mathcal{T} := \{\pi \in \Pi : \mathbb{E}^\pi[\tau_\pi] < \infty\},$$

be the set of admissible policies with finite expected sample complexity.

Assumption 5 (Strict feasibility)**.** *We assume there exists a policy $\pi \in \mathcal{T}$ such that*

$$\mathbb{E}^\pi[r_{\tau_\pi}(H_{\tau_\pi})] > 1-\delta.$$

We also make the following assumption of closedness on the attainable cost-reward set. For any $\pi \in \mathcal{T}$ define $c(\pi) := \mathbb{E}^\pi[\tau_\pi]$ and $\rho(\pi) := \mathbb{E}^\pi[r_{\tau_\pi}(H_{\tau_\pi})]$. Define also

$$g(\lambda) = \inf_{\pi\in\Pi: c(\pi)<\infty} c(\pi) + \lambda(1-\delta-\rho(\pi)), \qquad \mathcal{S}(\lambda) = \operatorname*{arg\,min}_{\pi\in\Pi: c(\pi)<\infty} c(\pi) + \lambda(1-\delta-\rho(\pi)),$$

and the set of attainable cost-rewards:

$$\mathcal{K} := \{(c(\pi), \rho(\pi)) : \pi \in \mathcal{T}\} \subset [0,\infty) \times [0,1].$$

We have the following properties on $g(\lambda)$.

Lemma 11. *Under Assumption 5, $g(\lambda)$ is finite, concave on $[0,\infty)$ and attains a maximum on $[0,\infty)$.*

Proof. For simplicity, we write the dual value directly on the attainable cost-reward set:

$$g(\lambda) = \inf_{(c,\rho)\in\mathcal{K}} c + \lambda(1-\delta-\rho),$$

Then, $g(\lambda) \ge -\delta\lambda$. Let (c,ρ) be a point satisfying Assumption 5. Then $\rho > 1-\delta$, and thus $g(\lambda) \le c - \lambda\eta$ for $\eta = -(1-\delta-\rho)$. Since $\eta > 0$, and $-\lambda\delta \le g(\lambda) \le c - \lambda\eta$, then $g(\lambda) \to -\infty$ as $\lambda \to \infty$.

Finally, since $\lambda \mapsto c + \lambda(1-\delta-\rho)$ is affine, the infimum of any family of concave functions is concave, therefore $g(\lambda)$ is concave. Then, since g is finite and concave, it's continuous on $(0,\infty)$ and u.s.c. in 0. Therefore g is u.s.c. on $[0,\infty)$, and since $\lim_{\lambda\to\infty} g(\lambda) \to -\infty$, there exists $\lambda^\star \in [0,\infty)$ such that $g(\lambda^\star) = \max_{\lambda\ge 0} g(\lambda)$. □

We impose then the following assumption.

Assumption 6 (Closed optimal cost-reward set)**.** *Let $\mathcal{G} = \arg\max_\lambda g(\lambda)$. We assume that for every maximizer $\lambda^\star \in \mathcal{G}$ there exists $\epsilon_0 > 0$ such that*

$$\mathcal{K}_\epsilon(\lambda^\star) \coloneqq \{(c,\rho) \in \mathcal{K} : c + \lambda^\star(1-\delta-\rho) \leq g(\lambda^\star) + \epsilon\}$$

is closed for all $\epsilon \in (0, \epsilon_0]$

Essentially, this geometric assumption is used to relax the assumption used in (Russo et al., 2026) that $\mathcal{S}(\lambda)$ is a singleton for each λ, and allows us to generalize the argument to multiple optimal dual policies.

To verify correctness, we use the fact that the sub-gradient of the optimal value of the dual problem is non-decreasing. To show this result, we employ the following proposition from (Hantoute and López, 2008) (see Prop. 3.1 therein), which characterizes the subdifferential of the supremum of a family of affine functions.

Proposition 6 (Subdifferential of the supremum of affine functions (Hantoute and López, 2008))**.** *Given a non-empty set $\{(a_t, b_t) : t \in \mathcal{T}\} \subset \mathbb{R}^2$, and the supremum function $f(x) : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$*

$$f(x) = \sup\{a_t x - b_t : t \in \mathcal{T}\},$$

for every $x \in \operatorname{dom} f$ we have

$$\partial f(x) = \bigcap_{\epsilon > 0} \operatorname{cl}\left(\operatorname{conv}\{a_t : t \in \mathcal{T}_\epsilon(x)\} + B(x)\right)$$

with

$$\mathcal{T}_\epsilon(x) \coloneqq \{t \in \mathcal{T} : a_t x - b_t \geq f(x) - \epsilon\},$$

and

$$B(x) \coloneqq \{y \in \mathbb{R} : (y, yx) \in (\overline{\operatorname{conv}}\{(a_t, b_t) : t \in \mathcal{T}\})_\infty\},$$

where C_∞ is the recession cone of a set C and $\overline{\operatorname{conv}(\cdot)}$ denotes the closed convex hull of a set. In particular, if $x \in \operatorname{int}(\operatorname{dom} f)$ we have

$$\partial f(x) = \bigcap_{\epsilon > 0} \overline{\operatorname{conv}}\{a_t : t \in \mathcal{T}_\epsilon(x)\}.$$

This last proposition permits us to define the subdifferential of the supremum of affine functions, and, as we see now, we can also find an upper bound on every subgradient $d \in \partial f(x)$.

We are now ready to prove Bayesian (ϵ, δ)-PAC guarantees for ICPE in the continuous setting.

Theorem C.3 (Bayesian (ϵ, δ)-correctness in the continuous ICPE case)**.** *Under Assumption 5 and Assumption 6 for all $\lambda^\star \in \mathcal{G} \coloneqq \arg\max_\lambda g(\lambda)$ there exists a dual-optimal policy $\pi^\star \in \mathcal{S}(\lambda^\star)$ such that*

$$\mathbb{E}^{\pi^\star}[r_{\tau_{\pi^\star}}(H_{\tau_{\pi^\star}})] \geq 1-\delta, \quad \text{and} \quad \mathbb{P}^{\pi^\star}\left(L_\theta\left(I^\star_{\tau^\star}(H_{\tau^\star})\right) \leq \epsilon\right) \geq 1-\delta.$$

where $I^\star_t(h) \in \arg\max_x q_t(h,x)$ is the optimal inference selector.

Proof. For simplicity, we write the dual value directly on the attainable cost-reward set:

$$g(\lambda) = \inf_{(c,\rho)\in\mathcal{K}} c + \lambda(1-\delta-\rho),$$

and the active set $M(\lambda) \coloneqq \arg\min_{(c,\rho)\in\mathcal{K}} c + \lambda(1-\delta-\rho)$.

Consider Lem. 11, then there exists $\lambda^\star \in \arg\max_\lambda g(\lambda)$ such that $\lambda \in [0, \infty)$. We now prove that $M(\lambda^\star)$ contains at-least a point (c,ρ) with reward at-least $1-\delta$.

By contradiction, (**HYP**) assume that all points $(c,\rho) \in M(\lambda^\star)$ satisfy $\rho < 1-\delta$.

First step: there exists $\alpha > 0$ such that $\rho \leq 1-\delta-\alpha$ for optimal policies. Under (**HYP**), we claim that there exists $\alpha > 0$ such that $\rho \leq 1-\delta-\alpha$ for all $(c,\rho) \in M(\lambda^\star)$. First, note that by Assumption 6 there exists ϵ_0 such that $M(\lambda^\star) = \bigcap_{\epsilon\in(0,\epsilon_0]} \mathcal{K}_\epsilon(\lambda^\star)$. Note that each $\mathcal{K}_\epsilon(\lambda^\star)$ is closed,

non-empty (since g is finite), and since $\rho \in [0,1]$ and $c \leq g(\lambda^\star) + \epsilon + \lambda^\star\delta$, it's bounded. Thus each $\mathcal{K}_\epsilon(\lambda^\star)$ is compact: therefore, also the intersection $M(\lambda^\star)$ is compact.

Since $M(\lambda^\star)$ is compact, we have that the maximum is achieved, and since $(c,\rho) \mapsto \rho$ is continuous, we set $\rho_{\max} = \max\{\rho : (c,\rho) \in M(\lambda^\star)\}$ and $\alpha = 1 - \delta - \rho_{\max} > 0$.

Second step: near optimal policies satisfy $\rho \leq 1 - \delta - \alpha/2$. Always under (**HYP**), we now claim that for near-optimal points we actually have $\rho \leq 1 - \delta - \alpha/2$. Suppose that this is not true: then, by Assumption 6 there exists ϵ_0 such that for every $\epsilon \in (0, \epsilon_0]$ there exists a point (c,ρ) in $\mathcal{K}_\epsilon(\lambda^\star)$ such that $\rho > 1 - \delta - \alpha/2$. Let $\epsilon_n = 1/n$ and consider such a sequence $(c_n, \rho_n) \in \mathcal{K}_{\epsilon_n}(\lambda^\star)$ satisfying $\rho_n > 1 - \delta - \alpha/2$ for all n Therefore, for each n we have that

$$c_n + \lambda^\star(1 - \delta - \rho_n) \leq g(\lambda^\star) + \frac{1}{n}.$$

Since $\rho_n \in [0,1]$ and $c_n \leq g(\lambda^\star) + 1/n + \lambda^\star\delta$, thus bounded, we have that the sequence (c_n, ρ_n) is bounded. Since for each ϵ_n the set $\mathcal{K}_{\epsilon_n}(\lambda^\star)$ is closed, we can take a convergent subsequence that satisfies $(c_n, \rho_n) \to (c, \rho)$ for some $(c,\rho) \in \mathcal{K}_{\epsilon_0}$. Therefore, we have that

$$\lim_{n\to\infty} c_n + \lambda^\star(1 - \delta - \rho_n) \leq g(\lambda^\star) + 1/n \implies c + \lambda^\star(1 - \delta - \rho) \leq g(\lambda^\star).$$

However, for any (c,ρ) we also have $g(\lambda^\star) \leq c + \lambda^\star(1 - \delta - \rho)$, thus $c + \lambda^\star(1 - \delta - \rho) = g(\lambda^\star)$, while also having $\rho \geq 1 - \delta - \alpha/2$, which contradicts $\rho \leq 1 - \delta - \alpha$. Therefore, for all ϵ_0 close points we have $\rho \leq 1 - \delta - \alpha/2$.

We now distinguish the cases $\lambda^\star = 0$ and $\lambda^\star > 0$.

Third step: case $\lambda^\star = 0$ is not optimal. Under (**HYP**), assume $\lambda^\star = 0$ is optimal. We show that this is not the case, and there exists $\lambda : g(\lambda) > g(0)$. We proceed by showing there exists $\lambda > 0$ small such that for any $(c,\rho) \in \mathcal{K}$ we have $c + \lambda(1 - \delta - \rho) > g(0)$, and thus there exists $\lambda > 0$ such that $g(\lambda) > g(0)$.

Let $\lambda > 0$ to be chosen. By Assumption 6 and the previous step there exists ϵ_0 such that for all $(c,\rho) \in \mathcal{K}_\epsilon(0)$ with $\epsilon \in (0, \epsilon_0]$ we have $1 - \delta - \alpha/2 \geq \rho$ and $c \leq g(0) + \epsilon$. Consider then the following two cases:

1. Case where (c,ρ) satisfies $c \geq g(0) + \epsilon_0$ (far away point). Since $\rho \in [0,1]$, we have $1 - \delta - \rho \in [-1, 1]$, and thus $\lambda(1 - \delta - \rho) \geq -\lambda$. Hence
$$g(0) + \epsilon_0 - \lambda \leq c + \lambda(1 - \delta - \rho).$$
2. Case where (c,ρ) satisfies $c \leq g(0) + \epsilon_0$ (near optimal point). Then, in this case we have that $(1 - \delta - \rho) \geq \alpha/2$, thus, using that $g(0) \leq c$, combining the two inequalities we have
$$g(0) + \lambda\alpha/2 \leq c + \lambda(1 - \delta - \rho).$$

Then, choose $\lambda \in (0, \epsilon_0/2)$: we obtain $-\lambda > -\epsilon_0/2$, and $\lambda > 0$, thus

$$g(0) < g(0) + \frac{\epsilon_0}{2} < c + \lambda(1 - \delta - \rho), \qquad g(0) < g(0) + \lambda\frac{\alpha}{2} \leq c + \lambda(1 - \delta - \rho).$$

Therefore $g(0) < c + \lambda(1 - \delta - \rho)$ for all $(c,\rho) \in \mathcal{K}$. This shows that $g(\lambda) > g(0)$, and contradicts the optimality of $\lambda^\star = 0$.

Fourth step: case $\lambda^\star > 0$ is not optimal. Define the function $f(\lambda) = -g(\lambda)$. Writing $f(\lambda) = \sup_{(c,\rho)\in\mathcal{K}} \lambda(\rho - 1 + \delta) - c$ extends it to a finite convex function on the reals, with bounded slopes, $c \geq 0$, and $\mathcal{K}$ contains a finite-cost point. Thus $\mathrm{dom} f = \mathbb{R}$ and $\lambda^\star > 0$ makes the constraint $\lambda \geq 0$ locally inactive. Then, by Prop. 6 , we have

$$\partial f(\lambda^\star) = \bigcap_{\epsilon\in(0,\epsilon_0]} \overline{\mathrm{conv}}\{\rho - (1 - \delta) : (c,\rho) \in \mathcal{K}_\epsilon(\lambda^\star)\}$$

By step 2 we have $1-\delta-\alpha/2 \geq \rho$, therefore $\rho-(1-\delta) \leq -\alpha/2$, implying that

$$d \leq -\frac{\alpha}{2} \qquad \forall d \in \partial f(\lambda^\star).$$

But $\lambda^\star$ minimizes the convex function f and is an interior point of the domain, so necessarily $0 \in \partial f(\lambda^\star)$, which is a contradiction.

Last step. Since both cases are impossible, we conclude there exists $(c,\rho) \in M(\lambda^\star)$ with $\rho \geq 1-\delta$. Hence, there exists a dual optimal policy $\pi \in \mathcal{S}(\lambda^\star)$ such that $c(\pi) = c$ and $\rho(\pi) \geq 1-\delta$. Lastly, since $I_t^\star$ attains the supremum in the definition of $r_t(h)$, we have $r_t(h) = q_t(h, I_t^\star(h)) = \mathbb{P}(L_\theta(I_t^\star(h)) \leq \epsilon \mid H_t = h)$. Thus, we get

$$1-\delta \leq \rho(\pi) = \mathbb{E}^\pi[r_{\tau_\pi}(H_{\tau_\pi})] = \mathbb{E}^\pi_{H_{\tau_\pi}}\left[\mathbb{P}\left(L_\theta\left(I^\star_{\tau_\pi}(H_{\tau_\pi})\right) \leq \epsilon \middle| H_{\tau_\pi}\right)\right] = \mathbb{P}\left(L_\theta\left(I^\star_{\tau_\pi}(H_{\tau_\pi})\right) \leq \epsilon\right).$$

□

Remark 3. *The above result yields a* Bayesian fixed-confidence *guarantee under the prior ν. It is therefore the natural continuous counterpart of the fixed-confidence correctness statement in the discrete ICPE setting.*

C.5.1 Zero Duality Gap via Perturbation Values

In the previous sections we reduced the Bayesian fixed-confidence problem to an optimal-stopping problem with terminal posterior-success reward. We now study when the corresponding Lagrangian relaxation is exact. The main point of this section is that, although the policy space is infinite-dimensional, the duality question can be analyzed through the two-dimensional attainable cost-reward set

$$\mathcal{K} := \{(c(\pi), \rho(\pi)) : \pi \in \mathcal{T}\} \subset [0,\infty) \times [0,1],$$

where

$$c(\pi) := \mathbb{E}^\pi[\tau_\pi], \qquad \rho(\pi) := \mathbb{E}^\pi[r_{\tau_\pi}(H_{\tau_\pi})].$$

The zero duality gap result uses a standard one-dimensional perturbation argument from convex duality (Rockafellar, 1974; Ekeland and Temam, 1999; Rockafellar and Wets, 1998; Borwein and Lewis, 2006). While the general technique idea is not new, what is new is its application to this specific problem studied in this manuscript, where the attainable cost-reward set $\mathcal{K}$ arises from an optimal-stopping control problem over continuous spaces.

After optimizing over the inference rule, the reduced primal problem is

$$P^\star := \inf_{\pi \in \mathcal{T}} c(\pi) \qquad \text{s.t.} \qquad \rho(\pi) \geq 1-\delta. \tag{22}$$

Equivalently, $P^\star = \inf\{c : (c,\rho) \in \mathcal{K},\ \rho \geq 1-\delta\}$. The associated Lagrangian dual value is

$$D^\star := \sup_{\lambda \geq 0} g(\lambda), \qquad g(\lambda) := \inf_{\pi \in \mathcal{T}} \left\{c(\pi) + \lambda(1-\delta-\rho(\pi))\right\}. \tag{23}$$

Equivalently, $g(\lambda) = \inf_{(c,\rho)\in\mathcal{K}}\{c + \lambda(1-\delta-\rho)\}$, and the multiplier satisfies $\lambda \geq 0$. Clearly, by Lagrangian duality, we have that the weak duality easily holds $D^\star \leq P^\star$.

Assumptions. We state some assumptions. The first one is a convexification assumption on the policy class.

Assumption 7 (Time-sharing)**.** *For every $\pi_0, \pi_1 \in \mathcal{T}$ and every $\alpha \in [0,1]$, there exists a policy $\pi_\alpha \in \mathcal{T}$ such that*

$$c(\pi_\alpha) = \alpha c(\pi_1) + (1-\alpha)c(\pi_0), \qquad \rho(\pi_\alpha) = \alpha\rho(\pi_1) + (1-\alpha)\rho(\pi_0).$$

This assumption is natural for randomized sequential policies: before the episode starts, the learner samples an independent Bernoulli random variable and then follows either π_1 or π_0 for the entire episode. We use a-priori randomization over complete policies, rather than only randomized actions at each history, because the latter does not automatically convexify the full stopping-time law. Hence, under this assumption one can easily show that the cost-reward set $\mathcal{K}$ is convex.

Assumption 8 (Strict feasibility)**.** *There exists $\pi_{\mathrm{sf}} \in \mathcal{T}$ such that $\rho(\pi_{\mathrm{sf}}) > 1-\delta$.*

This is the same assumption as in Assumption 5.

The next assumption is not needed for zero duality gap. It is only needed when we want to extract an actual primal-dual saddle point from the zero-gap identity. Unlike global closedness of all bounded slices of $\mathcal{K}$, it only asks for closedness of one near-optimal dual level set around a dual maximizer.

Assumption 9 (Local closedness of near-optimal dual level sets)**.** *Let $\mathcal{G} := \arg\max_{\lambda \geq 0} g(\lambda)$. For every $\lambda^\star \in \mathcal{G}$, there exists $\epsilon_0 > 0$ such that*

$$\mathcal{K}_{\epsilon_0}(\lambda^\star) := \{(c, \rho) \in \mathcal{K} : c + \lambda^\star(1 - \delta - \rho) \leq g(\lambda^\star) + \epsilon_0\}$$

is closed in $\mathbb{R}^2$.

Remark 4 (Sufficient conditions for local closedness)**.** *A simple sufficient condition for Assumption 9 is that the cost-success set $\mathcal{K}$ is closed. Indeed, the map*

$$(c, \rho) \mapsto c + \lambda^\star(1 - \delta - \rho)$$

is continuous, and hence $\mathcal{K}_{\epsilon_0}(\lambda^\star)$ is the intersection of $\mathcal{K}$ with a closed half-space.

More concretely, in a bounded-horizon problem, $\mathcal{K}$ is compact whenever the set of induced strategic measures is compact and the expected cost and terminal-success functionals are continuous. This holds, for example, under compact spaces, weakly continuous kernels, and a continuous terminal-success functional.

Example 1 (Non-unique dual minimizers and multipliers)**.** *Let $1 - \delta = 0.9$ and*

$$\mathcal{K} = \operatorname{conv}\{(0, 0), (1, 0.9), (2, 1)\}.$$

Then

$$g(\lambda) = \min\{0.9\lambda,\ 1,\ 2 - 0.1\lambda\},$$

and therefore

$$g(\lambda) = 1, \qquad \lambda \in [10/9, 10].$$

Thus, the dual-optimal multiplier is not unique. Moreover, at $\lambda^\star = 10$, every point on the segment joining $(1, 0.9)$ and $(2, 1)$ is a dual minimizer. Nevertheless, $\mathcal{K}$ is compact, local closedness holds, $(2, 1)$ is strictly feasible, and $(1, 0.9)$ is primal optimal.

Zero Duality by Perturbation We now show how Assumption 7 and Assumption 8 can be used to prove zero duality gap via a perturbed value. Define the scalar perturbation value function

$$\varphi(u) := \inf\{c : (c, \rho) \in \mathcal{K},\ 1 - \delta - \rho \leq u\}, \qquad u \in \mathbb{R}, \tag{24}$$

with the convention $\inf \emptyset = +\infty$. The original primal value is

$$P^\star = \varphi(0).$$

The variable u is the allowed violation of the confidence constraint. If $u > 0$, the constraint is relaxed; if $u < 0$, the constraint is strengthened.

Lemma 12 (Basic properties of the perturbation value)**.** *Consider Assumption 7 and Assumption 8. Then φ is convex, nonincreasing, and finite on an open interval containing 0. In particular, φ is continuous at 0, and $\partial\varphi(0) \neq \emptyset$.*

Proof. We prove the claims one by one.

Step 1: monotonicity. If $u_1 \leq u_2$, then

$$\{(c, \rho) \in \mathcal{K} : 1 - \delta - \rho \leq u_1\} \subseteq \{(c, \rho) \in \mathcal{K} : 1 - \delta - \rho \leq u_2\}.$$

Therefore

$$\varphi(u_2) \leq \varphi(u_1).$$

Thus φ is nonincreasing.

Step 2: convexity. Using Assumption 7 one can easily show that $\mathcal{K}$ is convex. Take $u_0, u_1 \in \mathbb{R}$, $\alpha \in [0, 1]$, and $\eta > 0$. If $\varphi(u_i) < \infty$, choose $(c_i, \rho_i) \in \mathcal{K}$ such that

$$1 - \delta - \rho_i \leq u_i, \qquad c_i \leq \varphi(u_i) + \eta, \qquad i \in \{0, 1\}.$$

If one of the two values is $+\infty$, the convexity inequality is trivial. By convexity of $\mathcal{K}$,

$$(c_\alpha, \rho_\alpha) \coloneqq \alpha(c_1, \rho_1) + (1-\alpha)(c_0, \rho_0) \in \mathcal{K}.$$

Moreover,

$$1 - \delta - \rho_\alpha = \alpha(1 - \delta - \rho_1) + (1-\alpha)(1 - \delta - \rho_0) \le \alpha u_1 + (1-\alpha)u_0.$$

Hence (c_α, ρ_α) is feasible for $\varphi(\alpha u_1 + (1-\alpha)u_0)$, and so

$$\varphi(\alpha u_1 + (1-\alpha)u_0) \le c_\alpha \le \alpha\varphi(u_1) + (1-\alpha)\varphi(u_0) + \eta.$$

Letting $\eta \downarrow 0$ gives convexity.

Step 3: finiteness near zero. By strict feasibility, there exists $(c_{\mathrm{sf}}, \rho_{\mathrm{sf}}) \in \mathcal{K}$ such that

$$\rho_{\mathrm{sf}} > 1 - \delta.$$

Let

$$\eta_{\mathrm{sf}} \coloneqq \rho_{\mathrm{sf}} - (1 - \delta) > 0.$$

Then

$$1 - \delta - \rho_{\mathrm{sf}} = -\eta_{\mathrm{sf}}.$$

Therefore, for every $u \ge -\eta_{\mathrm{sf}}$, the same point $(c_{\mathrm{sf}}, \rho_{\mathrm{sf}})$ is feasible for $\varphi(u)$. Hence

$$\varphi(u) \le c_{\mathrm{sf}} < \infty, \qquad \forall u \ge -\eta_{\mathrm{sf}}.$$

Also, since $c \ge 0$ for all $(c, \rho) \in \mathcal{K}$, we have $\varphi(u) \ge 0$ whenever $\varphi(u) < \infty$. Thus φ is finite on the open interval $(-\eta_{\mathrm{sf}}, \infty)$, which contains 0.

Step 4: continuity and existence of a subgradient. A proper convex function that is finite on an open interval is continuous on that interval. Since $0 \in \mathrm{int}(\mathrm{dom}\,\varphi)$, the one-dimensional subdifferential $\partial\varphi(0)$ is nonempty. Equivalently, one may take any slope between the left and right derivatives of φ at zero. □

Theorem C.4 (Zero duality gap by perturbation)**.** *Assume Assumption 7 and Assumption 8. Then there exists $\lambda^\star \ge 0$ such that*

$$g(\lambda^\star) = P^\star.$$

Consequently,

$$\inf_{\pi \in \mathcal{T} : \rho(\pi) \ge 1-\delta} c(\pi) = \sup_{\lambda \ge 0} \inf_{\pi \in \mathcal{T}} \{c(\pi) + \lambda(1 - \delta - \rho(\pi))\}. \tag{25}$$

In particular, the Lagrangian dual has no duality gap, and the dual supremum is attained.

Proof. By Lem. 12, choose $s^\star \in \partial\varphi(0)$. Since φ is nonincreasing, every subgradient at zero is nonpositive. Indeed, for $h > 0$, the subgradient inequality gives

$$\varphi(h) \ge \varphi(0) + s^\star h,$$

while monotonicity gives $\varphi(h) \le \varphi(0)$. Therefore $s^\star h \le 0$, and hence $s^\star \le 0$.

Define $\lambda^\star \coloneqq -s^\star \ge 0$. The subgradient inequality gives, for every $u \in \mathbb{R}$,

$$\varphi(u) \ge \varphi(0) + s^\star u = P^\star - \lambda^\star u.$$

Now fix any $(c, \rho) \in \mathcal{K}$ and set $u = 1 - \delta - \rho$. Since (c, ρ) is feasible for $\varphi(u)$, we have $c \ge \varphi(u)$. Thus

$$c \ge \varphi(u) \ge P^\star - \lambda^\star u = P^\star - \lambda^\star(1 - \delta - \rho).$$

Equivalently, $c + \lambda^\star(1 - \delta - \rho) \ge P^\star$. Taking the infimum over all $(c, \rho) \in \mathcal{K}$ gives

$$g(\lambda^\star) = \inf_{(c,\rho)\in\mathcal{K}} \{c + \lambda^\star(1 - \delta - \rho)\} \ge P^\star.$$

By weak duality, we also have $g(\lambda^\star) \le P^\star$. Hence $g(\lambda^\star) = P^\star$. Taking the supremum over $\lambda \ge 0$ proves Eq. (25). □

Remark 5. *Strict feasibility (Assumption 8) is a simple sufficient condition. The exact perturbation condition is lower semicontinuity of φ at the origin by the Fenchel-Moreau theorem. In fact, φ^{**} is the lower semicontinuous convex envelope of φ. Since φ is already convex, equality $\varphi^{**}(0) = \varphi(0)$ holds exactly when φ is lower semicontinuous at 0. Finally, Lem. 12 shows that strict feasibility implies continuity, hence lower semicontinuity, at 0. We state Thm. C.4 under strict feasibility because it is easier to interpret and check. Strict feasibility places 0 in the interior of $\mathrm{dom}\,\varphi$, and finite convex functions are continuous on the interior of their domain. The price is that strict feasibility is stronger than necessary.*

Primal attainment and KKT The zero-gap result above does not require any closedness assumption on $\mathcal{K}$. However, zero duality gap is only a statement about values. To obtain an actual policy that is both primal feasible and dual optimal, we need an attainment condition. The local closedness assumption above is enough: it compactifies a near-optimal dual level set around a dual maximizer, and a feasible minimizing sequence can be extracted inside this compact set.

Lemma 13 (Primal-dual attainment from local closedness)**.** *Assume 7, 8, and 9. Let $\lambda^\star \in \mathcal{G}$ be any dual maximizer. Then there exists $(c^\star, \rho^\star) \in \mathcal{K}$ such that*

$$\rho^\star \geq 1-\delta, \qquad c^\star = P^\star, \qquad c^\star + \lambda^\star(1-\delta-\rho^\star) = g(\lambda^\star).$$

Equivalently, there exists a policy $\pi^\star \in \mathcal{T}$ such that

$$c(\pi^\star) = P^\star, \qquad \rho(\pi^\star) \geq 1-\delta, \qquad \pi^\star \in \mathcal{S}(\lambda^\star).$$

Moreover, $\lambda^\star(1-\delta-\rho(\pi^\star)) = 0$.

Proof. By Thm. C.4, $g(\lambda^\star) = P^\star$. Let $(c_n, \rho_n) \in \mathcal{K}$ be a feasible minimizing sequence for the primal problem, so that

$$\rho_n \geq 1-\delta, \qquad c_n \to P^\star.$$

Define $L_{\lambda^\star}(c,\rho) \coloneqq c + \lambda^\star(1-\delta-\rho)$. Since $\rho_n \geq 1-\delta$ and $\lambda^\star \geq 0$, we have

$$L_{\lambda^\star}(c_n, \rho_n) \leq c_n.$$

On the other hand, by definition of $g(\lambda^\star)$,

$$L_{\lambda^\star}(c_n, \rho_n) \geq g(\lambda^\star) = P^\star.$$

Therefore,

$$P^\star \leq L_{\lambda^\star}(c_n, \rho_n) \leq c_n \to P^\star,$$

and hence $L_{\lambda^\star}(c_n, \rho_n) \to P^\star = g(\lambda^\star)$.

Let $\epsilon_0 > 0$ be given by Assumption 9. For all sufficiently large n,

$$L_{\lambda^\star}(c_n, \rho_n) \leq g(\lambda^\star) + \epsilon_0,$$

so $(c_n, \rho_n) \in \mathcal{K}_{\epsilon_0}(\lambda^\star)$.

We now show that $\mathcal{K}_{\epsilon_0}(\lambda^\star)$ is compact. It is closed by assumption. It is also bounded: for every $(c,\rho) \in \mathcal{K}_{\epsilon_0}(\lambda^\star)$, since $\rho \in [0,1]$, we have $1-\delta-\rho \geq -\delta$, and thus

$$c \leq g(\lambda^\star) + \epsilon_0 + \lambda^\star \delta.$$

Also $c \geq 0$ and $\rho \in [0,1]$. Hence $\mathcal{K}_{\epsilon_0}(\lambda^\star)$ is closed and bounded in $\mathbb{R}^2$, and therefore compact.

Passing to a subsequence, we may assume $(c_n, \rho_n) \to (c^\star, \rho^\star)$ for some $(c^\star, \rho^\star) \in \mathcal{K}_{\epsilon_0}(\lambda^\star) \subseteq \mathcal{K}$. Since $c_n \to P^\star$, we have $c^\star = P^\star$. Since $\rho_n \geq 1-\delta$ for all n, we have $\rho^\star \geq 1-\delta$. By continuity of $L_{\lambda^\star}$,

$$L_{\lambda^\star}(c^\star, \rho^\star) = \lim_n L_{\lambda^\star}(c_n, \rho_n) = P^\star = g(\lambda^\star).$$

Because $(c^\star, \rho^\star) \in \mathcal{K}$, there exists $\pi^\star \in \mathcal{T}$ such that $c(\pi^\star) = c^\star$ and $\rho(\pi^\star) = \rho^\star$. The equality $L_{\lambda^\star}(c^\star, \rho^\star) = g(\lambda^\star)$ implies $\pi^\star \in \mathcal{S}(\lambda^\star)$. Finally, since $c^\star = P^\star$ and $L_{\lambda^\star}(c^\star, \rho^\star) = P^\star$, we obtain

$$\lambda^\star(1-\delta-\rho^\star) = 0.$$

This proves the claim. □

Define the dual-optimal policy set

$$\mathcal{S}(\lambda) \coloneqq \arg\min_{\pi \in \mathcal{T}} \left\{ c(\pi) + \lambda(1-\delta-\rho(\pi)) \right\}.$$

Theorem C.5 (KKT and Bayesian correctness)**.** *Assume 7, 8, and 9. Let $\lambda^\star \in \mathcal{G}$ be any dual maximizer. Then there exists $\pi^\star \in \mathcal{S}(\lambda^\star)$ such that*

$$c(\pi^\star) = P^\star, \qquad \rho(\pi^\star) \geq 1-\delta, \qquad \lambda^\star(1-\delta-\rho(\pi^\star)) = 0.$$

Consequently, if $I_t^\star(h) \in \arg\max_{x \in \mathcal{X}} q_t(h,x)$ is a measurable posterior-optimal inference selector, then

$$\mathbb{P}^{\pi^\star}_{\theta \sim \nu}\left(L_\theta\left(I^\star_{\tau_{\pi^\star}}(H_{\tau_{\pi^\star}}) \right) \leq \epsilon \right) \geq 1-\delta.$$

Proof. By Lem. 13, there exists $\pi^\star \in \mathcal{T}$ such that

$$c(\pi^\star) = P^\star, \qquad \rho(\pi^\star) \geq 1-\delta, \qquad \pi^\star \in \mathcal{S}(\lambda^\star),$$

and

$$\lambda^\star(1-\delta-\rho(\pi^\star)) = 0.$$

It remains only to translate $\rho(\pi^\star) \geq 1-\delta$ into the desired correctness statement.

By definition,

$$\rho(\pi^\star) = \mathbb{E}^{\pi^\star}\left[r_{\tau_{\pi^\star}}(H_{\tau_{\pi^\star}})\right].$$

Since $I_t^\star$ is posterior-optimal, $r_t(h) = q_t(h, I_t^\star(h))$. Hence

$$\rho(\pi^\star) = \mathbb{E}^{\pi^\star}\left[q_{\tau_{\pi^\star}}\left(H_{\tau_{\pi^\star}}, I^\star_{\tau_{\pi^\star}}(H_{\tau_{\pi^\star}})\right)\right].$$

By Lem. 5, the last display equals

$$\mathbb{P}^{\pi^\star}_{\theta\sim\nu}\left(L_\theta\left(I^\star_{\tau_{\pi^\star}}(H_{\tau_{\pi^\star}})\right) \leq \epsilon\right).$$

Since $\rho(\pi^\star) \geq 1-\delta$, the result follows. □

Remark 6 (What is stronger than the previous correctness statement)**.** *The result above separates three issues that were previously entangled: (i) zero duality gap, (ii) attainment of a primal-dual saddle point, (iii) Bayesian-* (ϵ,δ)*-correctness. Time-sharing plus strict feasibility gives zero duality gap and dual attainment. The local closedness assumption is only used to extract an actual feasible dual-optimal policy from the zero-gap identity. Finally, Bayesian correctness follows from the posterior-optimal inference identity.*

Weaknesses and limitations. There are several limitations to keep in mind.

1. First, the theorem is a statement about the randomized, time-sharing closure of the policy class. If one restricts to deterministic policies without ex-ante randomization, $\mathcal{K}$ need not be convex and a Lagrangian duality gap can occur.
2. Second, zero duality gap does not imply that every dual-optimal policy is feasible. The theorem guarantees the existence of a feasible dual-optimal policy under the local closedness assumption. A learned policy that approximately minimizes the Lagrangian at $\lambda^\star$ is not automatically certified by this theorem.
3. Third, local closedness is still an attainment assumption. It is weaker than requiring all bounded slices of $\mathcal{K}$ to be closed, but it is not automatic. Without some local closedness, the primal value may be approached by policies whose cost-reward pairs converge to a boundary point outside $\mathcal{K}$, so no exact optimal policy need exist.

C.6 Approximation and relative sample complexity

We prove Prop. 1. Fix the trained networks and $c > 0$. Let the recommendation $D_t(h) = \hat{x}_\phi(h) \in \mathcal{X}$ be measurable. At each reached history, the actor draws one proposal, the learner then either stops or executes that proposal. We also assume its stopping time $\widehat{\tau}$ has finite expectation. Assume a constrained-optimal learner $\Pi^\star$ exists with $0 < T^\star = \mathbb{E}^{\Pi^\star}\tau^\star < \infty$ and posterior-optimal recommendations. Assume strong duality at $1-\delta$, so that $g(\lambda^\star) = T^\star$.

Fixed-recommendation values. Let $s_t(h) = q_t(h, D_t(h))$. We assume that the fixed-recommendation control problem has bounded measurable Bellman values

$$V_t^\phi(h) = \sup_{\pi,\tau\geq t} \mathbb{E}^{\pi;h}[s_\tau(H_\tau) - c(\tau - t)], \qquad Q_t^\phi(h,a) = -c + \int V_{t+1}^\phi(h,a,y)\bar{P}_t(dy \mid h,a),$$

$$V_t^\phi(h) = \max\{s_t(h), \sup_{a\in\mathcal{A}} Q_t^\phi(h,a)\}.$$

Expectations use the true posterior predictive law, and we write $p = p(\widehat{\Pi})$ and $T = \mathbb{E}\widehat{\tau}$.

Proof of Prop. 1. For a reached history and its proposal a, set

$$z_t(h,a) = |Q_\psi(h, a_{\text{stop}}) - s_t(h)| + |Q_\psi(h,a) - Q_t^\phi(h,a)|,$$
$$b_t(h,a) = V_t^\phi(h) - \max\{s_t(h), Q_t^\phi(h,a)\}.$$

Let $U_t = s_t(H_t)$ if the critic stops, and $U_t = Q_t^\phi(H_t, A_t)$ if it continues. If the critic chooses the worse of the two true values, their difference is at most the sum of its two score errors. Consequently,

$$0 \le V_t^\phi(H_t) - U_t \le b_t(H_t, A_t) + z_t(H_t, A_t).$$

By the trajectory-average definitions in Eq. (12),

$$\mathbb{E}_\nu^{\widehat{\Pi}}\left[\sum_{t=1}^{\widehat{\tau}} \big(b_t(H_t,A_t) + z_t(H_t,A_t)\big)\right] = (e_Q + e_A)T = \kappa T.$$

This weights each error by how often its decision history is encountered.

We justify summing these losses at the random stopping time. Write $V_t = V_t^\phi(H_t)$ and $\tau = \widehat{\tau}$. For deterministic n, the Bellman continuation identity and conditional expectation give

$$\begin{aligned}&\mathbb{E}V_1 - \mathbb{E}[\mathbf{1}\{\tau < n\} s_\tau(H_\tau) + \mathbf{1}\{\tau \ge n\} V_n - c(\tau \wedge n - 1)] \\ &= \mathbb{E}\left[\sum_{t < \tau \wedge n} (V_t - Q_t^\phi(H_t, A_t)) + \mathbf{1}\{\tau < n\}(V_\tau - s_\tau(H_\tau))\right].\end{aligned}$$

All residuals are nonnegative. Bounded convergence for the values and monotone convergence for costs and residuals, using $\mathbb{E}\tau < \infty$, yield

$$\mathbb{E}V_1 - [p - c(T-1)] = \mathbb{E}\sum_{t \le \tau}(V_t - U_t) \le \mathbb{E}\sum_{t \le \tau}(b_t(H_t,A_t) + z_t(H_t,A_t)) = \kappa T.$$

The stopped-posterior identity justifies replacing $\mathbb{E}s_\tau(H_\tau)$ by p.

Now retain the queries and stopping rule of $\Pi^\star$, but return D. This is an admissible fixed-recommendation learner. Its success is $p(\Pi^\star) - e_I$, where e_I is defined in the main text. The Bellman inequality applied to this comparator, with the same bounded-stopping argument, gives

$$\mathbb{E}V_1 \ge p(\Pi^\star) - e_I - c(T^\star - 1) \ge 1 - \delta - e_I - c(T^\star - 1).$$

Combining the last two displays and $p = 1 - \delta + e_p$ gives

$$c(T - T^\star) \le e_I + e_p + \kappa T, \qquad \frac{T - T^\star}{T^\star} \le \frac{\kappa + (e_I + e_p)/T^\star}{c - \kappa}.$$

For the lower bound, strong duality and the actual success of $\widehat{\Pi}$ give

$$T^\star = g(\lambda^\star) \le T + \lambda^\star(1 - \delta - p) = T - \lambda^\star e_p.$$

Indeed, optimizing the recommendation can only improve the success term in the dual infimum, so this inequality applies to the learned decoder as well. Thus $T - T^\star \ge \lambda^\star e_p$. For the upper bound, use $c - \kappa \ge [1 - \lambda^\star(e_c + \kappa)]/\lambda^\star > 0$ in the preceding display. This proves both sides of Prop. 1. □

The inference condition is needed only at the comparator's terminal histories. In a capped problem, the error summands at the imposed terminal return are taken to be zero, since stopping is the only admissible action. A bound at the learned learner's stopping histories alone would miss additional queries used to compensate for a poor decoder. For a forced horizon, the same proof uses the Bellman terminal condition $V_H^\phi = s_H$ and compares with the optimum under that horizon. An imposed terminal return is exact in this comparison and incurs no gate error. The theorem does not identify e_Q with empirical TD loss, or prove that training attains its premises.

C.7 Correctness certification

We prove Prop. 2 for the two procedures used in the main text. The one-shot procedure tests a checkpoint selected before seeing the certification data. The online procedure tests successive checkpoints and may select among those that pass. This second procedure must account for repeated testing and selection.

What is certified. A checkpoint m fixes the complete learner: its inference model, recommendation rule, actor, stopping rule, cost, maximum horizon, feasibility maps, and rollout sampling mode. Its true success probability is

$$p_m = \mathbb{P}^{\pi_m}_{\theta\sim\nu}\left(L_\theta(\hat{x}_m(H_{\tau_m})) \le \epsilon\right).$$

A certification observation is the success or failure of one complete rollout on a fresh task, not one query within a rollout. Each task is drawn independently from ν, with independent observation noise and policy randomization. All returns are evaluated, including those forced by the maximum horizon. Conditional on the frozen checkpoint and the information available before its test, these observations are independent Bernoulli variables with mean p_m.

Throughout, fix $0 < \alpha < \delta < 1/2$. The parameter δ is the desired deployment-error level, while α controls false certification. Let R denote return of a certified checkpoint $\widehat{m}$, and let F denote its failure on a new deployment task. The deployment task and its noise and policy seeds are independent of training and certification. When no checkpoint is returned, the joint event $R \cap F$ does not occur.

One-shot final test. We first select $\widehat{m}$ using training and development data, and then freeze it. Write $\mathcal{F}_{\mathrm{dev}}$ for all information used in this selection. Fix the batch size $n \ge 1$ before drawing the independent certification batch. If K is its number of successes, then, conditionally on $\mathcal{F}_{\mathrm{dev}}$, $K \sim \mathrm{Bin}(n, p_{\widehat{m}})$. To test $H_0 : p_{\widehat{m}} \le 1-\delta$, put $q = 1-\delta$ and define

$$k_\alpha = \min\left\{k \in \{0,\dots,n\} : \mathbb{P}_{Z\sim\mathrm{Bin}(n,q)}(Z \ge k) \le \alpha\right\}.$$

If the set is empty, set $k_\alpha = n+1$, so certification is impossible at that batch size. Equivalently, a passing outcome is possible exactly when $q^n \le \alpha$. We return the checkpoint when $R = \{K \ge k_\alpha\}$; otherwise we abstain, without retrying this final test.

One-shot part of Prop. 2. Condition on $\mathcal{F}_{\mathrm{dev}}$. The selected checkpoint and its success probability $p = p_{\widehat{m}}$ are now fixed, although p need not be known. Write

$$s(p) = \mathbb{P}_{Z\sim\mathrm{Bin}(n,p)}(Z \ge k_\alpha)$$

for the probability that this checkpoint passes the test.

False certification. If $p \le q$, monotonicity of the binomial upper tail and the definition of k_α give $s(p) \le s(q) \le \alpha$. If $p > q$, the checkpoint does not satisfy the null and cannot be falsely certified. Consequently,

$$\mathbb{P}(R,\ p_{\widehat{m}} \le q \mid \mathcal{F}_{\mathrm{dev}}) = \mathbf{1}\{p \le q\}s(p) \le \alpha.$$

Taking expectations over the training and development data proves $\mathbb{P}(R,\ p_{\widehat{m}} \le 1-\delta) \le \alpha$.

Certification followed by deployment failure. Under the same conditioning, certification passes with probability $s(p)$, while the independent deployment fails with probability $1-p$. Thus

$$\mathbb{P}(R \cap F \mid \mathcal{F}_{\mathrm{dev}}) = s(p)(1-p).$$

If $p \le 1-\delta$, passing is rare: this product is at most $\alpha(1-p) \le \alpha$. If $p > 1-\delta$, deployment failure is rare: the product is at most $1-p < \delta$. Since $\alpha < \delta$, both cases give a bound of δ. Taking expectations proves $\mathbb{P}(R \cap F) \le \delta$. □

Power and the training target. The preceding guarantee does not require training to attain its accuracy target. That target instead makes passing likely. When $k_\alpha \le n$, fix $\beta \in (0,1)$ and set

$$q_{\alpha,\beta} = \inf\{p \in [q,1] : s(p) \ge 1-\beta\}.$$

The function s is continuous and nondecreasing, and $s(1) = 1$. Hence the set is nonempty and closed, and every checkpoint with $p_m \ge q_{\alpha,\beta}$ passes with probability at least $1-\beta$. Choosing $1-\delta_{\mathrm{train}} \ge q_{\alpha,\beta}$ therefore provides a target for training, not a proof of attained accuracy. A separate development test may be used to calibrate an additional margin, but none of its observations may be reused in the one-shot final test.

Online testing and its threshold. The online procedure instead tests successive frozen checkpoints and may use earlier test outcomes to decide what to test next. We retain the checkpointwise test based on mixture martingales (Kaufmann and Koolen, 2021). To obtain both conclusions of Prop. 2 after selection among passing checkpoints, use the stricter error tolerance

$$\delta_{\mathrm{deploy}} = \frac{\delta - \alpha}{1 - \alpha}, \qquad q^\star = 1 - \delta_{\mathrm{deploy}} = \frac{1 - \delta}{1 - \alpha}.$$

Thus $q^\star > 1 - \delta$ and $\alpha + (1 - \alpha)\delta_{\mathrm{deploy}} = \delta$. The adjustment leaves room for both a false certificate and the deployment error of a checkpoint meeting the tested threshold. The one-shot proof did not need this adjustment because its checkpoint was selected before the test.

Fresh data for each checkpoint. Let $\mathcal{G}_m$ contain all training and testing information available just before checkpoint m is tested. The checkpoint, its level $\alpha_m \geq 0$, and its mixing distribution Π_m on $[q^\star, 1]$ are chosen using only $\mathcal{G}_m$. The levels satisfy $\sum_{m \geq 1} \alpha_m \leq \alpha$ almost surely; if $\alpha_m = 0$, its test is skipped. Otherwise, hold checkpoint m fixed while evaluating up to $B \geq 1$ fresh independent episodes, with B fixed before testing. Write

$$Z_{m,j} = \mathbf{1}\left\{ L_{\theta^{(m,j)}}\left(\hat{x}_m(H_{\tau_m}^{(m,j)}) \right) \leq \epsilon \right\}, \qquad S_{m,t} = \sum_{j=1}^{t} Z_{m,j}.$$

Conditionally on $\mathcal{G}_m$, the outcomes are i.i.d. $\mathrm{Ber}(p_m)$. Here t counts evaluated episodes, not queries within an episode. Their data are not used to change the checkpoint before its certification decision. Checkpoints tested later may depend on the completed tests; their own episodes must still be fresh.

The online test. For an alternative success probability $r \in [q^\star, 1]$, define

$$L_{m,t}(r) = \left(\frac{r}{q^\star} \right)^{S_{m,t}} \left(\frac{1 - r}{1 - q^\star} \right)^{t - S_{m,t}}, \qquad M_{m,t} = \int_{q^\star}^{1} L_{m,t}(r)\, \Pi_m(dr),$$

with $0^0 = 1$ and $M_{m,0} = 1$. The first quantity is a likelihood ratio comparing success probability r with $q^\star$; the second averages these ratios over the alternatives fixed before testing. It is unrelated to the Gaussian mixture used for inference. We certify checkpoint m as soon as $M_{m,t} \geq 1/\alpha_m$, or finish its test without certification after B episodes. Thus

$$\mathcal{C} = \{ m : \alpha_m > 0,\ \max_{1 \leq t \leq B} M_{m,t} \geq \alpha_m^{-1} \}$$

is the set of passing checkpoints. We may return any measurably selected $\widehat{m} \in \mathcal{C}$; if none passes, we abstain.

Proposition 7 (Training-time certification by checkpointwise mixture martingales)**.** *Under the online protocol above,*

$$\mathbb{P}\left(\exists m \in \mathcal{C} :\ p_m \leq q^\star \right) \leq \alpha.$$

Consequently, any selection from $\mathcal{C}$ satisfies

$$\mathbb{P}(R,\ p_{\widehat{m}} \leq 1 - \delta) \leq \alpha, \qquad \mathbb{P}(R \cap F) \leq \delta.$$

Proof. One checkpoint. Fix m with $\alpha_m > 0$ and condition on $\mathcal{G}_m$. This fixes p_m, α_m, and Π_m. Suppose $p_m \leq q^\star$. For each $r \geq q^\star$, write $L_{m,t+1}(r) = L_{m,t}(r)\ell_{m,t+1}(r)$, where the new factor is $r/q^\star$ after a success and $(1 - r)/(1 - q^\star)$ after a failure. For $\mathcal{F}_{m,t} = \sigma(\mathcal{G}_m, Z_{m,1}, \ldots, Z_{m,t})$,

$$\mathbb{E}[\ell_{m,t+1}(r) \mid \mathcal{F}_{m,t}] = p_m \frac{r}{q^\star} + (1 - p_m)\frac{1 - r}{1 - q^\star} = 1 + \frac{(p_m - q^\star)(r - q^\star)}{q^\star(1 - q^\star)} \leq 1.$$

Thus each $L_{m,t}(r)$ is a nonnegative supermartingale: under the null, its conditional expected value cannot increase. Since Π_m is fixed under this conditioning, Tonelli's theorem gives the same property for $M_{m,t}$, with initial value one.

For completeness, let T_m be the first threshold crossing, capped at B. Since $T_m \leq B$, summing conditional expected increments gives

$$\mathbb{E}[M_{m,T_m} \mid \mathcal{G}_m] = 1 + \sum_{t=0}^{B-1} \mathbb{E}[\mathbf{1}\{T_m > t\}(M_{m,t+1} - M_{m,t}) \mid \mathcal{G}_m] \leq 1.$$

On a crossing, $M_{m,T_m} \geq 1/\alpha_m$; otherwise it remains nonnegative. Hence $\mathbb{P}(m \in \mathcal{C} \mid \mathcal{G}_m) \leq \alpha_m$ whenever $p_m \leq q^\star$. The same bound is trivial for a skipped test.

All checkpoints. Let $A_m = \{m \in \mathcal{C},\ p_m \leq q^\star\}$ be a false certificate. Since p_m is fixed given $\mathcal{G}_m$, the preceding calculation gives $\mathbb{P}(A_m \mid \mathcal{G}_m) \leq \alpha_m$. The union bound and the budget condition now yield

$$\mathbb{P}\left(\bigcup_{m\geq 1} A_m\right) \leq \sum_{m\geq 1} \mathbb{E}[\alpha_m] = \mathbb{E}\left[\sum_{m\geq 1} \alpha_m\right] \leq \alpha.$$

No independence between the trained checkpoints is needed.

Deployment after selection. A returned checkpoint with $p_{\widehat{m}} \leq q^\star$ must be one of the false certificates above. Thus, writing $b = \mathbb{P}(R,\ p_{\widehat{m}} \leq q^\star)$, we have $b \leq \alpha$. Since $q^\star > 1 - \delta$, this also proves the claimed false-certification bound at the nominal threshold $1 - \delta$.

Condition on all training, testing, and selection information. On return, the new deployment fails with probability $1 - p_{\widehat{m}}$. For $p_{\widehat{m}} > q^\star$, this is at most δ_{deploy}; for the remaining returned checkpoints, it is at most one. Therefore

$$\begin{aligned}\mathbb{P}(R \cap F) &\leq \delta_{\text{deploy}} \mathbb{P}(R,\ p_{\widehat{m}} > q^\star) + b \\ &= \delta_{\text{deploy}} \mathbb{P}(R) + (1 - \delta_{\text{deploy}}) b \\ &\leq \delta_{\text{deploy}} + (1 - \delta_{\text{deploy}})\alpha = \delta.\end{aligned}$$

This proves the online part of Prop. 2. □

Remark 7 (What the guarantee certifies)**.** *The bounds permit abstention and do not imply $\mathbb{P}(F \mid R) \leq \delta$. They control the probability of issuing a false certificate and the joint probability of returning a checkpoint and failing on a fresh deployment task. For online testing, the first bound holds simultaneously for all passing checkpoints, so the returned checkpoint need not be the first to pass.*

Remark 8 (Relation to the original ICPE training-time argument)**.** *The pooled argument in Russo et al. (2026) tests the global null that all checkpoints are below the target. Rejecting that null does not by itself identify which checkpoint is correct, or certify a later checkpoint without a persistence assumption. Here each frozen checkpoint has its own test. The budget correction therefore supports selection among the checkpoints that actually pass, without assuming that accuracy improves monotonically during training.*

Remark 9 (Choice of certification budgets)**.** *If at most $M_{\max}$ checkpoints will be tested, a fixed uniform allocation is $\alpha_m = \alpha/M_{\max}$. A fixed allocation favoring later checkpoints is*

$$\alpha_m = \alpha\,\frac{(\gamma-1)\gamma^{m-1}}{\gamma^{M_{\max}} - 1}, \qquad m = 1, \ldots, M_{\max}, \quad \gamma > 1.$$

Both sum to α. The schedule affects the probability of passing, not validity, provided each level is chosen before its test data are observed and the total budget is respected.

Remark 10 (Finite-grid implementation)**.** *A finite mixing law $\Pi_m = \sum_{\ell=1}^{L} \omega_{m,\ell}\delta_{r_{m,\ell}}$, with nonnegative weights summing to one and $r_{m,\ell} \in [q^\star, 1]$, gives*

$$M_{m,t} = \sum_{\ell=1}^{L} \omega_{m,\ell} \left(\frac{r_{m,\ell}}{q^\star}\right)^{S_{m,t}} \left(\frac{1 - r_{m,\ell}}{1 - q^\star}\right)^{t - S_{m,t}}.$$

The grid and weights are fixed before testing checkpoint m. They affect power but do not change the proof. Placing all mass at $r = q^\star$ gives $M_{m,t} = 1$ and cannot produce a certificate.

C.8 Choice of inference model and reward modeling

This subsection clarifies the relation between the ideal posterior quantities used in the Bellman characterization and the inference model used in the implementation. The ideal inference rule maximizes

$$q_t(h, x) \coloneqq \mathbb{P}(L_\theta(x) \leq \epsilon | H_t = h), \qquad r_t(h) \coloneqq \sup_{x \in \mathcal{X}} q_t(h, x),$$

and any measurable maximizer is denoted by

$$I_t^\star(h) \in \arg\max_{x\in\mathcal{X}} q_t(h,x).$$

The implementation uses $I_\phi(\cdot \mid h)$ from Eq. (6) and returns the feasible dominant-component mean $\hat{x}_\phi(h)$. For fixed target parameters, the reward in Eq. (9) has conditional expectation $q_t(h, \hat{x}_{\bar{\phi}}(h))$: condition on $H_t = h$ and average its indicator over $R_t(d\theta \mid h)$. This identity concerns the task–history law, not an arbitrary outcome-reweighted replay law. The sampled reward below is a different terminal objective. Here $\hat{I}_t$ denotes the law of the feasible sample $\hat{X}_t = \mathrm{Proj}_{\mathcal{X}} Z_t$, $Z_t \sim I_\phi(\cdot \mid h)$; its moments need not equal those of the unprojected Gaussian.

Sampled reward. For fixed (h, θ), define the success probability of the implemented stochastic selector by

$$\hat{r}_t(h,\theta) \coloneqq \mathbb{E}_{x\sim\hat{I}_t(\cdot|h)}[\mathbf{1}\{x \in \mathcal{X}_\epsilon(\theta)\}].$$

Its posterior average is

$$\hat{r}_t(h) \coloneqq \mathbb{E}[\hat{r}_t(h,\theta)|H_t = h].$$

Given M indepdent samples $X_t^{(1)}, \ldots, X_t^{(M)} \sim \hat{I}_t(\cdot|h)$, we use

$$\hat{r}_{t,M}(h,\theta) \coloneqq \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{X_t^{(m)} \in \mathcal{X}_\epsilon(\theta)\}.$$

Lemma 14 (Sampled stochastic-selector reward)**.** *Conditionally on* $(H_t = h, \theta)$*,*

$$\mathbb{E}[\hat{r}_{t,M}(h,\theta)|H_t = h,\theta] = \hat{r}_t(h,\theta), \qquad \mathrm{Var}(\hat{r}_{t,M}(h,\theta)|H_t = h,\theta) \le \frac{1}{4M}.$$

Furthermore,

$$\hat{r}_t(h) = \mathbb{E}[q_t(h,\hat{X}_t)|H_t = h] \le r_t(h).$$

Proof. The first claim follows because $\hat{r}_{t,M}$ is the average of M Bernoulli random variables with success probability $\hat{r}_t(h,\theta)$. The variance bound follows from $p(1-p) \le 1/4$. For the posterior identity, $\hat{X}_t$ is conditional independent of θ given $H_t = h$, thus

$$\hat{r}_t(h) = \int_{\mathcal{X}} \mathbb{P}(x \in \mathcal{X}_\epsilon(\theta)|H_t = h)\hat{I}_t(dx|h) = \int_{\mathcal{X}} q_t(h,x)\hat{I}_t(dx|h).$$

The last display is bounded by $\sup_{x\in\mathcal{X}} q_t(h,x) = r_t(h)$. □

The alternative reward $\hat{r}_t(h)$ evaluates a randomized recommendation. It is therefore conservative relative to the ideal deterministic reward $r_t(h)$, which assumes access to the best posterior recommendation.

Second-moment robustness. The next proposition gives two sufficient conditions under which the sampled reward is close to the ideal one. The first bound uses concentration around the task-level zero-loss target $x_\theta^\star$. The second uses concentration around the posterior-optimal rule $I_t^\star(h)$.

For $z \in \mathcal{X}$, define

$$D_t(h,z) \coloneqq \mathbb{E}[\|\hat{X}_t - z\|^2|H_t = h], \qquad \hat{X}_t \sim \hat{I}_t(\cdot|h).$$

Proposition 8 (Second-moment robustness and ideal-reward gap)**.** *Assume that, for posterior-a.e.* θ*, there exists* $\rho(\theta) > 0$ *such that*

$$B(x_\theta^\star, \rho(\theta)) \cap \mathcal{X} \subseteq \mathcal{X}_\epsilon(\theta).$$

where $B(x,r)$ *is an euclidean ball of radius* r *around* x*. Then*

$$\hat{r}_t(h,\theta) \ge 1 - \frac{D_t(h,x_\theta^\star)}{\rho(\theta)^2}.$$

Moreover, assume $\hat{X}_t$ *is conditionally independent of* θ *given* $H_t = h$*. If* $I_t^\star(h) \in \arg\max_{x\in\mathcal{X}} q_t(h,x)$ *and* $q_t(h,\cdot)$ *is* $L_t(h)$*-Lipschitz on* $B(I_t^\star(h), R_t(h)) \cap \mathcal{X}$*, then*

$$0 \le r_t(h) - \hat{r}_t(h) \le L_t(h)\sqrt{D_t(h,I_t^\star(h))} + \frac{D_t(h,I_t^\star(h))}{R_t(h)^2}.$$

If $q_t(h,\cdot)$ *is globally* $L_t(h)$*-Lipschitz on* $\mathcal{X}$*, the second term is unnecessary.*

Proof. For the first claim, the margin assumption gives

$$\{\|\hat{X}_t - x^\star_\theta\| \le \rho(\theta)\} \subseteq \{\hat{X}_t \in \mathcal{X}_\epsilon(\theta)\}.$$

Therefore

$$1 - \hat{r}_t(h,\theta) \le \mathbb{P}(\|\hat{X}_t - x^\star_\theta\| > \rho(\theta) | H_t = h, \theta).$$

Since the law of $\hat{X}_t$ depends on h but not on θ conditional on h, Markov's inequality gives

$$1 - \hat{r}_t(h,\theta) \le \frac{D_t(h, x^\star_\theta)}{\rho(\theta)^2}.$$

For the second claim, by Lem. 14,

$$\hat{r}_t(h) = \mathbb{E}[q_t(h, \hat{X}_t) | H_t = h].$$

Since $I^\star_t(h)$ maximizes $q_t(h,\cdot)$,

$$r_t(h) - \hat{r}_t(h) = \mathbb{E}[q_t(h, I^\star_t(h)) - q_t(h, \hat{X}_t) | H_t = h] \ge 0.$$

Let

$$E_h := \{\|\hat{X}_t - I^\star_t(h)\| \le R_t(h)\}.$$

On E_h, Lipschitzness gives

$$q_t(h, I^\star_t(h)) - q_t(h, \hat{X}_t) \le L_t(h)\|\hat{X}_t - I^\star_t(h)\|.$$

On E^c_h, the same difference is at most 1. Hence

$$r_t(h) - \hat{r}_t(h) \le L_t(h)\mathbb{E}[\|\hat{X}_t - I^\star_t(h)\| | H_t = h] + \mathbb{P}(E^c_h | H_t = h).$$

Jensen's inequality gives

$$\mathbb{E}[\|\hat{X}_t - I^\star_t(h)\| | H_t = h] \le \sqrt{D_t(h, I^\star_t(h))},$$

and Markov's inequality gives

$$\mathbb{P}(E^c_h | H_t = h) \le \frac{D_t(h, I^\star_t(h))}{R_t(h)^2}.$$

If $q_t(h,\cdot)$ is globally Lipschitz, take $E_h = \mathcal{X}$ and remove the last term. □

The first bound gives the margin interpretation of the sampled reward: if the stochastic selector has small second moment around $x^\star_\theta$, then its samples are likely to be ϵ-optimal. The second bound is different: it compares the stochastic selector to the ideal posterior reward and is small when the selector concentrates around $I^\star_t(h)$ and $q_t(h,\cdot)$ is locally regular. Thus the sampled reward evaluates the implemented selector through the same event used for correctness; it does not assume that the NLL mean is exactly optimal.

For any two feasible sample laws I, J and fixed task, the expected sampled labels differ by at most $\mathrm{TV}(I, J)$, since they are the probabilities of the same measurable set $\mathcal{X}_\epsilon(\theta)$. Thus averaging can smooth the training target; this does not prove that one reward trains a more stable or robust policy.

Gaussian NLL. We next characterize the population Gaussian NLL objective. Let

$$Z := x^\star_\theta$$

be the supervised target, viewed as a random variable under the posterior law $\theta | H_t = h$. For $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{S}^d_{++}$, define

$$\mathcal{L}_h(\mu, \Sigma) := \mathbb{E}[-\log \mathcal{N}(Z; \mu, \Sigma) | H_t = h].$$

Let

$$\mu^{\mathrm{NLL}}_t(h) := \mathbb{E}[Z | H_t = h], \qquad \Sigma^{\mathrm{NLL}}_t(h) := \mathrm{Cov}(Z | H_t = h).$$

Proposition 9 (Gaussian NLL moment projection)**.** *Assume $Z|H_t = h$ has finite second moment and positive definite covariance. Then the unique minimizer of $\mathcal{L}_h(\mu, \Sigma)$ over $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{S}^d_{++}$ is*

$$\mu = \mu^{\mathrm{NLL}}_t(h), \qquad \Sigma = \Sigma^{\mathrm{NLL}}_t(h).$$

If the covariance is restricted to be diagonal, the optimal mean is still $\mu^{\mathrm{NLL}}_t(h)$ and the optimal diagonal entries are the posterior coordinate variances of Z.

Proof. Up to an additive constant,

$$\mathcal{L}_h(\mu, \Sigma) = \frac{1}{2} \log \det \Sigma + \frac{1}{2} \mathbb{E}[(Z - \mu)^\top \Sigma^{-1} (Z - \mu) | H_t = h].$$

For fixed Σ, the second term is minimized at $\mu = \mathbb{E}[Z|H_t = h]$. With this choice, the objective becomes

$$\frac{1}{2} \log \det \Sigma + \frac{1}{2} \mathrm{tr}(\Sigma^{-1} \Sigma_t^{\mathrm{NLL}}(h)).$$

The first-order condition in Σ gives

$$\Sigma = \Sigma_t^{\mathrm{NLL}}(h),$$

and strict convexity in the natural parameters gives uniqueness. The diagonal case follows by the same calculation coordinate-wise. □

Thus Gaussian NLL performs a moment projection of the posterior law of $x_\theta^\star$: it matches the first two posterior moments, or the coordinate variances in the diagonal case. This does not imply that $\mu_t^{\mathrm{NLL}}(h)$ maximizes $q_t(h, \cdot)$ in general. The NLL objective learns the posterior target law, while optimality for stopping is defined by the ϵ-success probability.

Near-optimality of the NLL mean. The NLL objective does not directly maximize $q_t(h, x)$. It learns the posterior mean of the selected target $x_\theta^\star$ under the Gaussian moment projection. The next result gives a simple condition under which this mean is nevertheless close to the ideal rule in posterior success probability. The condition is posterior concentration relative to the margin of the success set, not exact equality between the posterior mean and the maximizer of $q_t(h, \cdot)$.

Proposition 10 (Near-optimality of the NLL mean)**.** *Fix a history h and write*

$$Z \coloneqq x_\theta^\star, \qquad \mu_t^{\mathrm{NLL}}(h) \coloneqq \mathbb{E}[Z|H_t = h], \qquad V_t(h) \coloneqq \mathbb{E}[\|Z - \mu_t^{\mathrm{NLL}}(h)\|^2 | H_t = h].$$

Assume that there exists $\rho > 0$ such that, posterior-a.s., $B(x_\theta^\star, \rho) \cap \mathcal{X} \subseteq \mathcal{X}_\epsilon(\theta)$, where $B(x, r)$ is an Euclidean ball of radius r around x. Let $\hat{\mu}_t(h) \in \mathcal{X}$ be the deployed mean recommendation and assume

$$\|\hat{\mu}_t(h) - \mu_t^{\mathrm{NLL}}(h)\| \le e_t(h).$$

Then

$$0 \le r_t(h) - q_t(h, \hat{\mu}_t(h)) \le \frac{V_t(h) + e_t(h)^2}{\rho^2}.$$

Moreover, if $I_t^\star(h) \in \arg\max_{x \in \mathcal{X}} q_t(h, x)$, $q_t(h, \cdot)$ is $L_t(h)$-Lipschitz on $B(I_t^\star(h), R_t(h)) \cap \mathcal{X}$, and $\|\hat{\mu}_t(h) - I_t^\star(h)\| \le R_t(h)$, then

$$0 \le r_t(h) - q_t(h, \hat{\mu}_t(h)) \le L_t(h) \|\hat{\mu}_t(h) - I_t^\star(h)\|.$$

Proof. The margin assumption implies

$$\{\|\hat{\mu}_t(h) - x_\theta^\star\| \le \rho\} \subseteq \{\hat{\mu}_t(h) \in \mathcal{X}_\epsilon(\theta)\}.$$

Therefore

$$q_t(h, \hat{\mu}_t(h)) \ge 1 - \mathbb{P}(\|\hat{\mu}_t(h) - Z\| > \rho | H_t = h).$$

By Markov's inequality,

$$q_t(h, \hat{\mu}_t(h)) \ge 1 - \frac{\mathbb{E}[\|\hat{\mu}_t(h) - Z\|^2 | H_t = h]}{\rho^2}.$$

Since $\hat{\mu}_t(h)$ is deterministic conditional on $H_t = h$,

$$\mathbb{E}[\|\hat{\mu}_t(h) - Z\|^2 | H_t = h] = V_t(h) + \|\hat{\mu}_t(h) - \mu_t^{\mathrm{NLL}}(h)\|^2 \le V_t(h) + e_t(h)^2.$$

The first claim follows because $r_t(h) \le 1$. For the second claim, use $r_t(h) = q_t(h, I_t^\star(h))$ and Lipschitzness:

$$r_t(h) - q_t(h, \hat{\mu}_t(h)) = q_t(h, I_t^\star(h)) - q_t(h, \hat{\mu}_t(h)) \le L_t(h) \|\hat{\mu}_t(h) - I_t^\star(h)\|.$$

□

In general, $\mu_t^{\mathrm{NLL}}(h) = \mathbb{E}[x_\theta^\star | H_t = h]$ need not maximize $q_t(h, \cdot)$: if the posterior law of $x_\theta^\star$ is multimodal, the mean can lie between modes. The first bound gives the guarantee tied to NLL training. If the posterior uncertainty on $x_\theta^\star$ is small relative to the margin of $\mathcal{X}_\epsilon(\theta)$, then the deployed mean is near-optimal in posterior success probability. The Lipschitz bound is different: it says that any deterministic recommendation close to an ideal maximizer $I_t^\star(h)$ is near-optimal when $q_t(h, \cdot)$ is locally regular. Thus the Lipschitz argument also applies to the NLL mean, but only if one separately controls its distance to $I_t^\star(h)$.

For localization losses $L_\theta(x) = \|x - x_\theta^\star\|$, the margin condition holds with $\rho(\theta) = \epsilon$. For smooth value-gap losses, it follows from a local upper curvature bound near $x_\theta^\star$. For example, if

$$f_\theta(x) - f_\theta(x_\theta^\star) \le \frac{M}{2}\|x - x_\theta^\star\|^2$$

in $B(x_\theta^\star, r_0) \cap \mathcal{X}$ for some $r_0 > 0$, then

$$B\left(x_\theta^\star, \min\left\{r_0, \sqrt{\frac{2\epsilon}{M}}\right\}\right) \cap \mathcal{X} \subseteq \mathcal{X}_\epsilon(\theta).$$

Remark 11 (Exact alignment under symmetric localization)**.** *In special cases the NLL mean is exactly Bayes-optimal. Suppose the success sets are translates of a fixed centrally symmetric convex set, i.e.,*

$$\mathcal{X}_\epsilon(\theta) = \{x \in \mathcal{X} : x - x_\theta^\star \in S_\epsilon\},$$

and suppose $x_\theta^\star | H_t = h$ is Gaussian with mean $\mu_t^{\mathrm{NLL}}(h)$, with $\mu_t^{\mathrm{NLL}}(h) \in \mathcal{X}$. Then

$$\mu_t^{\mathrm{NLL}}(h) \in \arg\max_{x \in \mathcal{X}} q_t(h, x).$$

Indeed, $q_t(h, x)$ is the posterior probability that a Gaussian random variable falls in a translate of S_ϵ, and this probability is maximized when the translate is centered at the Gaussian mean. Without this type of symmetry, the posterior mean, posterior mode, and maximizer of $q_t(h, \cdot)$ can differ.

Correctness of the returned recommendation. The first argument in Prop. 10 applies to any measurable feasible recommendation, including a selected mixture mean. Suppose $\rho > 0$ satisfies $B(x_\theta^\star, \rho) \cap \mathcal{X} \subseteq \mathcal{X}_\epsilon(\theta)$ for ν-almost every task. Put $m_t(h) = \mathbb{E}[x_\theta^\star \mid h]$ and $v_t(h) = \mathbb{E}[\|x_\theta^\star - m_t(h)\|^2 \mid h]$. For any almost-surely finite stopping time,

$$\mathbb{P}(L_\theta(\hat{x}_\phi(H_\tau)) > \epsilon) \le \mathbb{E}\min\left\{1, \frac{v_\tau(H_\tau) + \|m_\tau(H_\tau) - \hat{x}_\phi(H_\tau)\|^2}{\rho^2}\right\}. \tag{26}$$

Indeed, conditional failure is bounded by $\mathbb{P}(\|x_\theta^\star - \hat{x}_\phi(h)\| > \rho \mid h)$. Markov's inequality and the bias–variance identity give the displayed integrand; Lem. 5 then gives the claim.

Suboptimal supervised targets. Zero-loss supervision is not needed by the preceding moment arguments. Let $\tilde{x}_\theta \in \mathcal{X}_\epsilon(\theta)$ be a fixed measurable supervised target with finite conditional second moments. Suppose, for some $\rho > 0$ and ν-almost every task,

$$\|x - \tilde{x}_\theta\| \le \rho, \quad x \in \mathcal{X} \quad \Longrightarrow \quad L_\theta(x) \le \epsilon. \tag{27}$$

The Gaussian NLL characterization then uses the conditional moments of $\tilde{x}_\theta$, and the same margin argument gives, for any measurable feasible recommendation $D(h)$,

$$1 - q_t(h, D(h)) \le \min\left\{1, \frac{\mathbb{E}[\|D(h) - \tilde{x}_\theta\|^2 \mid H_t = h]}{\rho^2}\right\}.$$

Indeed, failure implies distance greater than ρ, so Markov's inequality proves the bound. Evaluating it at an almost-surely finite stopping time and using the stopped-posterior identity gives the corresponding prior-averaged failure bound. Thus the preceding margin-based statements extend by replacing the supervised target and its moments, without changing L_θ or q_t. For uniformly C-Lipschitz losses, $C > 0$, and $L_\theta(\tilde{x}_\theta) \le \epsilon_{\mathrm{sel}} < \epsilon$, Eq. (27) holds with $\rho = (\epsilon - \epsilon_{\mathrm{sel}})/C$, since $L_\theta(x) \le \epsilon_{\mathrm{sel}} + C\|x - \tilde{x}_\theta\|$.

Example 2 (Supervised-target margin and query complexity)**.** *For this illustration, take $\Theta = \mathcal{X} = \mathbb{R}$ and a single available measurement. Fix $v_0, \sigma^2, \epsilon > 0$ and $\delta \in (0, 1/2)$. Let $\theta \sim \mathcal{N}(0, v_0)$ and observe $Y_i = \theta + \xi_i$, where the noises are independent $\mathcal{N}(0, \sigma^2)$ variables, independent of θ. Correctness is measured by $L_\theta(x) = |x - \theta|$, but supervision uses $\tilde{x}_\theta = \theta + b$, with a fixed $0 \le b \le \epsilon$. After $n \ge 0$ measurements, conditional on $\mathcal{F}_n = \sigma(Y_1, \dots, Y_n)$, write*

$$\theta \mid \mathcal{F}_n \sim \mathcal{N}(m_n, v_n), \qquad v_n = (v_0^{-1} + n/\sigma^2)^{-1}, \qquad m_n = v_n \sigma^{-2} \sum_{i=1}^{n} Y_i.$$

An exactly learned Gaussian for the supervised target returns $D_n = m_n + b$. Writing Φ for the standard normal distribution function, its true posterior success is

$$q_b(n) = \mathbb{P}(|D_n - \theta| \le \epsilon \mid \mathcal{F}_n) = \Phi\left(\frac{\epsilon - b}{\sqrt{v_n}}\right) + \Phi\left(\frac{\epsilon + b}{\sqrt{v_n}}\right) - 1. \tag{28}$$

Let $N_b = \inf\{n \ge 0 : q_b(n) \ge 1 - \delta\}$ be the oracle query count, with no horizon cap. It is finite for $b < \epsilon$ and nondecreasing in b, with

$$N_b = \Theta\left(\frac{\sigma^2}{(\epsilon - b)^2}\right) \qquad \text{as } b \uparrow \epsilon,$$

when the other parameters are fixed. At $b = \epsilon$, $N_b = \infty$.

Proof. Completing the square in the Gaussian prior and likelihood gives m_n, v_n. Since $\tilde{x}_\theta = \theta + b$, its conditional law is $\mathcal{N}(m_n + b, v_n)$, and integrating the posterior over $[D_n - \epsilon, D_n + \epsilon]$ gives Eq. (28). This probability depends only on n, so N_b is deterministic. If φ denotes the standard normal density, then

$$\frac{\partial q_b(n)}{\partial b} = \frac{\varphi((\epsilon + b)/\sqrt{v_n}) - \varphi((\epsilon - b)/\sqrt{v_n})}{\sqrt{v_n}} \le 0,$$

which proves monotonicity of the stopping count. For $b < \epsilon$, put $\rho = \epsilon - b$ and $\overline{\Phi} = 1 - \Phi$. Then

$$\overline{\Phi}(\rho/\sqrt{v_n}) \le 1 - q_b(n) \le 2\overline{\Phi}(\rho/\sqrt{v_n}).$$

Writing $z_u = \Phi^{-1}(u)$, the left inequality gives the necessary condition $n \ge \sigma^2(z_{1-\delta}^2/\rho^2 - v_0^{-1})$; the right inequality shows that $n \ge \sigma^2(z_{1-\delta/2}^2/\rho^2 - v_0^{-1})$ is sufficient. Taking nonnegative integer counts proves finiteness and the stated order. At $b = \epsilon$, $q_b(n) = \Phi(2\epsilon/\sqrt{v_n}) - 1/2 < 1/2$ for every finite n, so the stopping threshold cannot be reached. □

The example fixes the observation model and learns the target law exactly: the additional queries arise from the smaller margin of the returned mean. This is not a lower bound for unrestricted recommendations; subtracting the known shift b returns m_n and removes the penalty. In general, changing the supervised target can also change the queries generated by TS, so target loss alone does not order all selectors by efficiency.

C.9 Multimodal posterior stress test

We provide the derivation of Thm. 4.1. We first compute the posterior and the success of each recommendation. We then compare the information collected by the three query rules and show how this gives their expected query counts. The distinction is important: the recommendation determines how much evidence is needed, while the query distribution determines how quickly that evidence is collected.

Throughout the one-dimensional analysis, $m \in (0, 1)$, $0 < \epsilon < m/2$, $\eta \in (0, 1/2)$, and $\delta \in (0, 1/2)$ are fixed. We write $q_\star = 1 - \delta$ and assume

$$0 < w \le \min\{\epsilon/2, (1 - m)/2\}, \qquad 0 < \ell \le \min\{m/2, 1 - m\}.$$

The variables S, U are independent, the initial observation is uninformative, and observations are drawn with fresh Bernoulli randomness conditional on the task and queries. All query and recommendation rules are measurable; ties are resolved in a fixed way. Conditional on the history, query randomization is independent of the task. Gaussian query draws are clipped to $[-1, 1]$. The three oracle rules have no horizon cap; finite expected stopping is proved below, rather than assumed. Constants may depend on the fixed $m, \epsilon, \eta, \delta$, but not on w, ℓ. All logarithms are natural.

C.9.1 Posterior and recommendation

Recall that $\theta = Sm + U$, with S uniform on $\{-1, +1\}$ and U uniform on $[-w, w]$. This experiment uses the exact supervised target $x^\star_\theta = \theta$ and loss $L_\theta(x) = |x - \theta|$. For a realized history h_t, define its likelihood under sign s by

$$\mathsf{L}^s_t(h_t) = \prod_{j=1}^{t-1} g_{s,\ell}(a_j)^{y_{j+1}}(1 - g_{s,\ell}(a_j))^{1-y_{j+1}}, \qquad g_{s,\ell}(a) = \eta + (1-2\eta)e^{-(a-sm)^2/(2\ell^2)}.$$

The initial observation contributes only a sign-independent factor. Let $p_t = \mathbb{P}(S = +1 \mid H_t)$ and $u_t = \max\{p_t, 1 - p_t\}$.

Lemma 15 (Posterior success and Gaussian projection)**.** *For every realized history, the following statements hold.*

First, the posterior law of the target is

$$P_t = p_t \,\mathrm{Unif}(I_+) + (1 - p_t)\,\mathrm{Unif}(I_-), \qquad p_t = \frac{\mathsf{L}^+_t}{\mathsf{L}^+_t + \mathsf{L}^-_t}, \qquad I_\pm = [\pm m - w, \pm m + w].$$

Second, its posterior-success function is

$$q_t(h_t, x) = \frac{p_t}{2w}\,\mathrm{Leb}(I_+ \cap [x - \epsilon, x + \epsilon]) + \frac{1 - p_t}{2w}\,\mathrm{Leb}(I_- \cap [x - \epsilon, x + \epsilon]). \tag{29}$$

Thus $r_t(h_t) = u_t$, attained by the dominant lobe center $c_t = m$ when $p_t \geq 1/2$ and $c_t = -m$ otherwise.

Lastly, the Gaussian NLL projection has mean and variance

$$\mu_t = m(2p_t - 1), \qquad v_t = \frac{w^2}{3} + 4m^2 p_t(1 - p_t).$$

Its mean recommendation has success $q_t(h_t, \mu_t) = \Gamma_\epsilon(u_t)$, where

$$\Gamma_\epsilon(u) = \begin{cases} 0, & 1/2 \leq u \leq b_0, \\ (m/w)u(u - b_0), & b_0 < u < b_1, \\ u, & b_1 \leq u \leq 1, \end{cases} \qquad b_0 = 1 - \frac{\epsilon + w}{2m}, \quad b_1 = 1 - \frac{\epsilon - w}{2m}. \tag{30}$$

Proof. The likelihood depends on the task only through S. Conditional on the history, U therefore remains uniform and independent of S. Query probabilities do not enter the likelihood ratio, since the learner chooses queries without knowing the task. Bayes' rule gives the first claim.

The event $L_\theta(x) \leq \epsilon$ is equivalent to $\theta \in [x - \epsilon, x + \epsilon]$. Integrating its indicator under P_t gives Eq. (29). The gap between the two lobes is $2(m - w) > 2\epsilon$, so a success interval cannot intersect both. The success probability is consequently at most the mass of one lobe. Since $w < \epsilon$, recommending its center covers that entire lobe, proving $r_t = u_t$.

Conditional independence gives $\mathbb{E}[\theta \mid h_t] = m\mathbb{E}[S \mid h_t]$ and $\mathrm{Var}(\theta \mid h_t) = m^2\,\mathrm{Var}(S \mid h_t) + w^2/3$. Since $\mathbb{E}[S \mid h_t] = 2p_t - 1$ and $\mathrm{Var}(S \mid h_t) = 4p_t(1 - p_t)$, these are the displayed Gaussian moments. Their variance is positive because $w > 0$.

It remains to evaluate the success of the mean. By symmetry, consider $p_t = u \geq 1/2$. The mean $m(2u - 1)$ lies between zero and m, and its success interval misses the negative lobe. It begins to intersect the positive lobe when $m(2u - 1) + \epsilon > m - w$, or $u > b_0$, and covers it completely when $m(2u - 1) + \epsilon \geq m + w$, or $u \geq b_1$. Between these thresholds the overlap length is $2m(u - b_0)$. Dividing by $2w$ and multiplying by the lobe probability u gives Eq. (30). □

Evidence required by the recommendation. The center rules stop when $u_t \geq q_\star$. The Gaussian mean needs $\Gamma_\epsilon(u_t) \geq q_\star$, equivalently $u_t \geq u_G$, where

$$u_G = \begin{cases} \left(b_0 + \sqrt{b_0^2 + 4(w/m)q_\star}\right)/2, & q_\star < b_1, \\ q_\star, & q_\star \geq b_1. \end{cases} \tag{31}$$

In particular, $u_G \geq q_\star$. For $m = 0.6, w = 0.02, \epsilon = 0.08$, a center reaches success 0.90 at $u_t = 0.90$, while the Gaussian mean requires $u_t \simeq 0.9483$. This compares the evidence thresholds at a common history; it does not yet compare query counts, since the agents generate different histories. The reference M_t retains the exact lobe weights, means $\pm m$, and component variances $w^2/3$, and therefore returns the same center as the posterior rule.

The thresholds are bounded away from one uniformly in w:

$$q_\star \leq u_G \leq \bar{u} := \max\{q_\star, 1 - \epsilon/(4m)\} < 1. \tag{32}$$

Indeed, $w \leq \epsilon/2$ implies $b_1 \leq 1 - \epsilon/(4m)$, and the mean already attains success u when $u \geq b_1$. This bound will control how far the log odds can move before stopping.

C.9.2 Information collected by the query rules

We next quantify how quickly queries separate the two signs. The posterior log odds $\Lambda_t = \log(p_t/(1-p_t))$, initially zero, satisfy

$$\Lambda_{t+1} - \Lambda_t = Y_{t+1} \log \frac{g_{+,\ell}(A_t)}{g_{-,\ell}(A_t)} + (1 - Y_{t+1}) \log \frac{1 - g_{+,\ell}(A_t)}{1 - g_{-,\ell}(A_t)}. \tag{33}$$

For $s \in \{-1, +1\}$, define

$$d_s(a) = \mathrm{kl}(g_{s,\ell}(a), g_{-s,\ell}(a)), \qquad \mathrm{kl}(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y}.$$

Taking the expectation of Eq. (33) under the true sign gives

$$\mathbb{E}[s(\Lambda_{t+1} - \Lambda_t) \mid H_t, A_t = a, S = s] = d_s(a). \tag{34}$$

Thus $d_s(a)$ measures the expected evidence supplied by query a. To obtain a lower bound on the required query count, we also use the prior-averaged information

$$\begin{aligned} i_p(a) &= p\,\mathrm{kl}(g_{+,\ell}(a), \bar{g}_p(a)) + (1 - p)\,\mathrm{kl}(g_{-,\ell}(a), \bar{g}_p(a)), \\ \bar{g}_p(a) &= p g_{+,\ell}(a) + (1 - p) g_{-,\ell}(a). \end{aligned} \tag{35}$$

This is the expected reduction in the entropy of S after one observation.

For a rule $R \in \{G, M, P\}$, let R_p denote its inference law in Eq. (13) at belief p. Expectations below are over $Z \sim R_p$ and the executed query $A = \mathrm{Proj}_{[-1,1]} Z$; projection leaves P_p unchanged.

Lemma 16 (Information per query)**.** *There are constants $a_0, A_0 > 0$, independent of w, ℓ, such that, for p with $\max\{p, 1 - p\} < \bar{u}$ and either true sign s,*

$$\begin{aligned} a_0 \gamma_R \leq \mathbb{E}_A d_s(A) \leq A_0 \gamma_R, \qquad 0 \leq \mathbb{E}_A i_p(A) \leq A_0 \gamma_R, \\ \gamma_G = \ell/m, \qquad \gamma_M = \gamma_P = \min\{1, \ell/w\}. \end{aligned} \tag{36}$$

The bounds for M, P in fact hold for all $p \in [0, 1]$.

Proof. Where queries are informative. All observation probabilities lie in $[\eta, 1 - \eta]$. For x, y in this interval,

$$2(x - y)^2 \leq \mathrm{kl}(x, y) \leq \frac{(x - y)^2}{y(1 - y)} \leq \frac{(x - y)^2}{\eta(1 - \eta)}.$$

For the lower bound, differentiate twice in x and use $\partial_x^2 \mathrm{kl}(x, y) = 1/[x(1 - x)] \geq 4$; for the upper bound apply $\log z \leq z - 1$ to both logarithms. Hence, for a constant depending only on η,

$$d_s(a),\ i_p(a) \leq C_\eta K_\ell(a), \qquad K_\ell(a) = e^{-(a-m)^2/\ell^2} + e^{-(a+m)^2/\ell^2}, \qquad \int_{\mathbb{R}} K_\ell(a)\, da = 2\sqrt{\pi}\ell. \tag{37}$$

For i_p, apply the same inequality to each sign law and its predictive mixture: the resulting squared differences sum to $p(1 - p)(g_{+,\ell} - g_{-,\ell})^2$.

Let $J_s = [sm - \ell/2, sm + \ell/2]$ and $J = J_+ \cup J_-$. On each interval, the nearby likelihood bump is at least $e^{-1/8}$ and the opposite bump is at most $e^{-49/8}$, because $\ell \leq m/2$. Thus, for either true sign,

$$d_s(a) \geq d_0 := 2(1 - 2\eta)^2 (e^{-1/8} - e^{-49/8})^2 > 0, \qquad a \in J. \tag{38}$$

In particular, a query near either center is informative: it need not be near the actual sign's center. Both intervals lie inside $[-1, 1]$.

Exact posterior and mixture. Each uniform lobe puts mass $\min\{1, \ell/(2w)\}$ in its own J_s. Each Gaussian component of M_p puts mass

$$2\Phi\left(\frac{\sqrt{3}\ell}{2w}\right) - 1 \asymp \min\{1, \ell/w\}$$

in its own J_s, where Φ is the standard normal distribution function. Therefore the lower bound follows from Eq. (38), uniformly in the mixture weights. For the upper bound, the unprojected densities of P_p, M_p are at most $1/(2w)$ and $\sqrt{3}/(\sqrt{2\pi}w)$, respectively. Integrating Eq. (37) gives $C\ell/w$. Both information quantities are also bounded by a constant, so the continuous part contributes at most $C\min\{1, \ell/w\}$.

Single Gaussian. On the stated belief range, G_p has $|\mu_p| \le m$ and variance

$$4m^2\bar{u}(1-\bar{u}) \le v_p \le m^2 + \epsilon^2/12.$$

Its density is consequently at most C/m everywhere and at least a/m on J. For the latter bound, use $|z - \mu_p| \le 2m + \ell/2 \le 9m/4$ in the Gaussian density formula. The interval J has length 2ℓ, so the lower information bound is $a\ell/m$. Integrating Eq. (37) gives the upper bound $C\ell/m$ for the continuous part.

Clipping. The only remaining contribution is from the atoms at $-1, 1$ created by clipping a Gaussian draw. By Eq. (37), their total contribution is at most $Ce^{-(1-m)^2/\ell^2} \le C\ell/(1-m)$, since $\ell \le 1-m$. For G_p this is absorbed in $C\ell/m$, with m fixed. For M_p, when $\ell < w$, use $w \le (1-m)/2$ to absorb it in $C\ell/w$; when $\ell \ge w$, use the constant information bound. Clipping leaves the probability of J unchanged. This completes all bounds in Eq. (36). □

C.9.3 From information to sample complexity

Moment-TS samples G_t and recommends μ_t; Mixture-TS and Posterior-TS sample M_t, P_t and recommend c_t. Each rule updates its own exact posterior and stops at

$$\tau_R = \inf\{t \ge 1 : q_t(H_t, \hat{x}_{R,t}) \ge q_\star\}, \qquad N_R = \tau_R - 1.$$

Thus the center rules have boundary $u_t \ge q_\star$, and Moment-TS has boundary $u_t \ge u_G$.

Proof of Thm. 4.1. Upper bounds and finite stopping. Fix R and condition on $S = s$; write $\mathbb{E}_s$ for this expectation. Before stopping, $u_t < \bar{u}$ by Eq. (32). Equations (34) and (36) give a conditional signed drift of at least $a_0\gamma_R$ per query. Moreover, each log-odds increment has absolute value at most $B_\eta = \log((1-\eta)/\eta)$. Put $B_0 = \log(\bar{u}/(1-\bar{u}))$. The log odds before stopping have absolute value below B_0, so one last observation can overshoot the boundary by at most B_η.

For a deterministic $k \ge 1$, summing only to the bounded time $\tau_R \wedge k$ gives

$$\begin{aligned} a_0\gamma_R\mathbb{E}_s[(\tau_R \wedge k) - 1] &\le \mathbb{E}_s\left[\sum_{t=1}^{k-1} \mathbf{1}\{t < \tau_R\} s(\Lambda_{t+1} - \Lambda_t)\right] \\ &= \mathbb{E}_s[s\Lambda_{\tau_R \wedge k}] \le B_0 + B_\eta. \end{aligned} \tag{39}$$

The equality uses $\Lambda_1 = 0$. Monotone convergence now gives $\mathbb{E}_s N_R \le (B_0 + B_\eta)/(a_0\gamma_R) < \infty$. This also proves almost-sure termination, rather than using it to justify the drift calculation. Since the true posterior success at stopping is at least $q_\star$, the stopped-posterior identity gives prior-averaged correctness at least $q_\star$.

Matching lower bounds. A positive log-odds drift alone supplies no lower bound on the number of queries: a rule can stop at the wrong sign. We instead use the reduction in posterior uncertainty required by either stopping boundary. Define $\mathsf{h}(p) = -p\log p - (1-p)\log(1-p)$, with $0\log 0 = 0$. Bayes' rule and Eq. (35) give

$$\mathsf{h}(p_t) - \mathbb{E}[\mathsf{h}(p_{t+1}) \mid H_t] = \mathbb{E}_A i_{p_t}(A).$$

Sum this identity first to $\tau_R \wedge k$. Bounded convergence for entropy and monotone convergence for the nonnegative information terms give

$$0 < \log 2 - \mathsf{h}(q_\star) \le \log 2 - \mathbb{E}\mathsf{h}(p_{\tau_R}) = \mathbb{E}\sum_{t<\tau_R} \mathbb{E}_A i_{p_t}(A) \le A_0\gamma_R \mathbb{E}N_R. \tag{40}$$

The first inequality follows from $u_{\tau_R} \ge q_\star > 1/2$. Together with Eq. (39), this establishes

$$\frac{\log 2 - \mathsf{h}(q_\star)}{A_0\gamma_R} \le \mathbb{E}N_R \le \frac{B_0 + B_\eta}{a_0\gamma_R}.$$

Substituting $\gamma_G = \ell/m$ and $\gamma_M = \gamma_P = \min\{1, \ell/w\}$ proves the three stated rates, since $\max\{1, w/\ell\} \asymp 1 + w/\ell$.

Comparison with unrestricted querying. Repeatedly querying $+m$ gives, under either sign,

$$d_s(m) \ge 2(1-2\eta)^2(1-e^{-8})^2 =: d_\eta > 0.$$

The same bounded-drift argument with center recommendation and boundary $q_\star$ yields

$$\mathbb{E}N_{+m} \le \frac{\log(q_\star/(1-q_\star)) + B_\eta}{d_\eta}.$$

This is a feasible learner with expected query count bounded independently of w, ℓ. Conversely, on the event of making no query, any learner has success at most $1/2$, because it has seen only an uninformative initial observation and task-independent randomness. Thus every $q_\star$-correct learner satisfies

$$q_\star \le \tfrac{1}{2}\mathbb{P}(N=0) + \mathbb{P}(N \ge 1), \qquad \mathbb{E}N \ge \mathbb{P}(N \ge 1) \ge 2q_\star - 1.$$

Taking the infimum over all admissible learners gives $2q_\star - 1 \le T^\star \le \mathbb{E}N_{+m} \le C$, proving $T^\star \asymp 1$. □

The comparison holds at fixed δ while w, ℓ vary.

C.9.4 The 2^d-lobe construction

The multidimensional construction also separates the effect of a single Gaussian from that of discarding posterior correlations. Let $\mathcal{S}_d = \{-1, +1\}^d$, draw S uniformly from $\mathcal{S}_d$, and independently draw $U \sim \mathrm{Unif}([-w, w]^d)$. Set $\theta = mS + U$, $\mathcal{A} = \mathcal{X} = [-1, 1]^d$, and $L_\theta(x) = \|x - \theta\|_\infty$. Assume $0 < w < \epsilon < m - w$ and $m + w < 1$. The initial observation is uninformative and each query gives a fresh Bernoulli observation with mean

$$g_{s,\ell}(a) = \eta + \frac{1-2\eta}{d}\sum_{j=1}^{d} \exp\left(-\frac{(a_j - ms_j)^2}{2\ell^2}\right). \tag{41}$$

This is the sum of coordinate bumps and reduces to the two-lobe model when $d = 1$. At $a = 0$ the observation law is the same for every sign vector. At $a = ms$, a sign vector r differing in k coordinates has

$$g_{r,\ell}(ms) = \eta + (1-2\eta)\left[1 - \frac{k}{d}(1 - e^{-2m^2/\ell^2})\right].$$

Posterior and Gaussian moments. Put $B_s = ms + [-w, w]^d$. With the likelihood products defined as above, Bayes' rule gives

$$p_t(s) = \frac{\mathsf{L}_t^s}{\sum_r \mathsf{L}_t^r}, \qquad P_t = \sum_s p_t(s)\,\mathrm{Unif}(B_s).$$

Indeed, observations depend on S but not on U, so U remains a product uniform law independent of S conditional on the history. Writing $\overline{s}_t = \sum_s p_t(s)s$, this also gives

$$\mu_t = m\overline{s}_t, \qquad \Sigma_t = \frac{w^2}{3}I_d + m^2\left(\sum_s p_t(s)ss^\top - \overline{s}_t\overline{s}_t^\top\right).$$

A diagonal Gaussian keeps the same mean and the diagonal entries of Σ_t; it need not retain the joint dependence between signs.

Recommendation success. A success cube cannot intersect two different lobes: they differ in some coordinate whose lobe intervals are separated by $2(m-w) > 2\epsilon$. Each center covers its entire lobe. Thus

$$q_t(h,x) = \sum_s p_t(s) \prod_{j=1}^{d} \frac{[\min(ms_j + w, x_j + \epsilon) - \max(ms_j - w, x_j - \epsilon)]_+}{2w},$$
$$r_t(h) = \max_s p_t(s).$$

Let $\widehat{s}_j$ be the marginal majority sign, with $+1$ at a tie, and let $u_j = (1 + |\overline{s}_{t,j}|)/2$. The interval centered at $\mu_{t,j}$ can intersect only the coordinate lobe with sign $\widehat{s}_j$. Its overlap fraction is

$$\rho_\epsilon(u) = \begin{cases} 0, & u \le b_0, \\ (m/w)(u - b_0), & b_0 < u < b_1, \\ 1, & u \ge b_1. \end{cases}$$

Conditional independence of the coordinates of U therefore gives

$$q_t(h, \mu_t) = p_t(\widehat{s}) \prod_{j=1}^{d} \rho_\epsilon(u_j).$$

Consequently, writing $p_t^\star = \max_s p_t(s)$, the recommendation gap is

$$r_t(h) - q_t(h, \mu_t) = \underbrace{p_t^\star - p_t(\widehat{s})}_{\text{joint-mode gap}} + \underbrace{p_t(\widehat{s}) \left(1 - \prod_j \rho_\epsilon(u_j)\right)}_{\text{interpolation gap}}.$$

The first term accounts for marginal majorities that do not identify the most probable joint sign vector; the second accounts for a mean between lobe centers. Both terms are nonnegative. The reference mixture with weights $p_t(s)$, means ms, and covariances $w^2 I_d/3$ returns a most probable center and attains $r_t(h)$, but may require 2^d components.

C.10 Robustness to prior misspecification

We now study deployment under a misspecified task prior. The controller `C-ICPE` is trained under a prior ν on Θ and then frozen. At deployment, the same learned inference network, critic, action rule, cost parameter, and stopping rule are used, but the environment parameter is drawn from a different prior ν'. Thus, conditional on a fixed environment θ, the trajectory law $\mathbb{P}_\theta^{\pi_\phi}$ is unchanged; only the outer averaging measure over θ changes.

Let

$$s(\theta) := \mathbb{P}_\theta^{\pi_\phi}\big(\hat{x}_\phi(H_{\hat{\tau}}) \in \mathcal{X}_\epsilon(\theta)\big) \in [0,1], \qquad t(\theta) := \mathbb{E}_\theta^{\pi_\phi}[\hat{\tau}] \in [0,\infty].$$

For a prior η on Θ, define

$$p^{(\eta)} := \int_\Theta s(\theta)\, \eta(d\theta), \qquad T^{(\eta)} := \int_\Theta t(\theta)\, \eta(d\theta).$$

The training-prior guarantee is

$$p^{(\nu)} \ge 1 - \delta.$$

The goal is to understand how $p^{(\nu')}$ and $T^{(\nu')}$ change when ν is replaced by ν'.

Assumption 10 (Measurable performance profiles)**.** *The maps $s : \Theta \to [0,1]$ and $t : \Theta \to [0,\infty]$ are Borel measurable. Moreover, $t \in L^1(\eta)$ for every prior η considered below.*

Lemma 17 (Prior shift as reweighting)**.** *Under Assumption 10, correctness and stopping time under any prior η are given by*

$$p^{(\eta)} = \mathbb{E}_{\theta\sim\eta}[s(\theta)], \qquad T^{(\eta)} = \mathbb{E}_{\theta\sim\eta}[t(\theta)].$$

If $\eta \ll \nu$ with density ratio $w_\eta = d\eta/d\nu$, then

$$p^{(\eta)} = \mathbb{E}_\nu[w_\eta s], \qquad T^{(\eta)} = \mathbb{E}_\nu[w_\eta t].$$

Proof. Under prior η, the joint law factors as

$$\eta(d\theta)\,\mathbb{P}_\theta^{\pi_\phi}(dh).$$

Applying Tonelli's theorem to the success indicator gives

$$p^{(\eta)} = \int_\Theta \mathbb{P}_\theta^{\pi_\phi}\big(\hat{x}_\phi(H_{\hat{\tau}}) \in \mathcal{X}_\epsilon(\theta)\big)\,\eta(d\theta) = \int_\Theta s(\theta)\,\eta(d\theta).$$

The identity for $T^{(\eta)}$ follows similarly from Tonelli applied to the nonnegative random variable $\hat{\tau}$. The reweighting identities follow by the change of measure $d\eta = w_\eta d\nu$. □

The next proposition gives several complementary ways of transferring the training-prior guarantee $p^{(\nu)} \geq 1-\delta$ to a deployment prior ν'. The density-ratio and χ^2 bounds are useful when $\nu' \ll \nu$; the total-variation and good-set bounds do not require absolute continuity.

Proposition 11 (Success probability under prior shift)**.** *Assume Assumption 10, and write $p = p^{(\nu)}$, $p' = p^{(\nu')}$. Then:*

(a) *If $\nu' \ll \nu$ and $C = \|d\nu'/d\nu\|_\infty$, then $1-p' \leq C(1-p)$. In particular, $p \geq 1-\delta$ implies $p' \geq 1 - C\delta$.*

(b) *If $\nu' \ll \nu$ and $\chi^2(\nu'\|\nu) < \infty$, then $|p'-p| \leq \sqrt{\chi^2(\nu'\|\nu)\,\mathrm{Var}_\nu(s)}$.. Hence $p \geq 1-\delta$ implies*

$$p' \geq 1 - \delta - \sqrt{\chi^2(\nu'\|\nu)\delta}.$$

(c) *For arbitrary priors, $|p'-p| \leq \mathrm{TV}(\nu',\nu)$. Consequently,*

$$p' \geq 1 - \delta - \mathrm{TV}(\nu',\nu).$$

By Pinsker's inequality, this also gives

$$p' \geq 1 - \delta - \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\nu'\|\nu)}$$

whenever $D_{\mathrm{KL}}(\nu'\|\nu) < \infty$, and the analogous bound with the KL arguments reversed.

(d) *Let $G \subseteq \Theta$ be measurable. If $s(\theta) \geq 1-\eta$ on G, then*

$$p' \geq (1-\eta)\nu'(G).$$

Moreover, if $p \geq 1-\delta$ and

$$G_\eta := \{\theta : s(\theta) \geq 1-\eta\},$$

then

$$\nu(G_\eta) \geq 1 - \frac{\delta}{\eta}, \qquad p' \geq (1-\eta)\nu'(G_\eta).$$

Proof. Let $e(\theta) = 1 - s(\theta) \in [0,1]$.

For (a), if $w = d\nu'/d\nu$, then

$$1 - p' = \mathbb{E}_{\nu'}[e] = \mathbb{E}_\nu[we] \leq \|w\|_\infty \mathbb{E}_\nu[e] = C(1-p).$$

For (b), since $\mathbb{E}_\nu[w] = 1$,

$$p' - p = \mathbb{E}_\nu[(w-1)s] = \mathbb{E}_\nu[(w-1)(s-p)].$$

Cauchy–Schwarz yields

$$|p'-p| \leq \sqrt{\mathbb{E}_\nu[(w-1)^2]}\,\sqrt{\mathrm{Var}_\nu(s)} = \sqrt{\chi^2(\nu'\|\nu)\,\mathrm{Var}_\nu(s)}.$$

Since $0 \leq s \leq 1$, $s^2 \leq s$, hence

$$\mathrm{Var}_\nu(s) \leq p(1-p) \leq 1-p.$$

If $p \geq 1 - \delta$, then $\mathrm{Var}_\nu(s) \leq \delta$, giving the displayed bound.

For (c), since $s \in [0, 1]$,

$$|p' - p| = \left| \int s \, d(\nu' - \nu) \right| \leq \mathrm{TV}(\nu', \nu).$$

Pinsker's inequality gives the KL consequences.

For (d),

$$p' = \int_G s \, d\nu' + \int_{G^c} s \, d\nu' \geq (1 - \eta)\nu'(G).$$

If $G = G_\eta$, then $G_\eta^c = \{e > \eta\}$. Since $\mathbb{E}_\nu[e] \leq \delta$, Markov's inequality gives

$$\nu(G_\eta^c) \leq \frac{\delta}{\eta}.$$

□

The good-set formulation is often the most faithful explanation of empirical robustness. Average correctness under ν implies that the controller is accurate on a large ν-measure set of environments. Deployment remains accurate whenever ν' continues to place most of its mass on that same set.

We now state the analogous bounds for the stopping time. Since t is not bounded by one, the bounds require either density-ratio control, a second-moment assumption, or a bounded horizon.

Proposition 12 (Stopping time under prior shift)**.** *Assume Assumption 10, and write $T = T^{(\nu)}$, $T' = T^{(\nu')}$. Then:*

(a) *If $\nu' \ll \nu$ and $C = \|d\nu'/d\nu\|_\infty$, then $T' \leq CT$.*

(b) *If $\nu' \ll \nu$, $\chi^2(\nu'\|\nu) < \infty$, and $t \in L^2(\nu)$, then*

$$|T' - T| \leq \sqrt{\chi^2(\nu'\|\nu)\,\mathrm{Var}_\nu(t)}.$$

(c) *If $\hat{\tau} \leq T_{\max}$ almost surely under every θ, then*

$$|T' - T| \leq T_{\max}\,\mathrm{TV}(\nu', \nu).$$

(d) *If $\hat{\tau} \leq T_{\max}$ almost surely and $t(\theta) \leq \tau_0$ on a measurable set $F \subseteq \Theta$, then*

$$T' \leq \tau_0 + (T_{\max} - \tau_0)\nu'(F^c).$$

Proof. The first claim follows from

$$T' = \mathbb{E}_\nu[wt] \leq \|w\|_\infty \mathbb{E}_\nu[t] = CT.$$

For the second, use

$$T' - T = \mathbb{E}_\nu[(w - 1)t] = \mathbb{E}_\nu[(w - 1)(t - T)]$$

and apply Cauchy–Schwarz. For the third, $t/T_{\max} \in [0, 1]$, so the total-variation bound applies. For the fourth, split

$$T' = \int_F t \, d\nu' + \int_{F^c} t \, d\nu'$$

and use $t \leq \tau_0$ on F and $t \leq T_{\max}$ everywhere. □

Remark 12 (Beta–Uniform shifts)**.** *For one-dimensional shifts with $\nu = \mathrm{Unif}(0, 1)$ and $\nu' = \mathrm{Beta}(\alpha, \beta)$, the constants in Prop. 11 can be evaluated in closed form. For example,*

$$\chi^2(\nu'\|\nu) = \frac{B(2\alpha - 1, 2\beta - 1)}{B(\alpha, \beta)^2} - 1$$

when $\alpha, \beta > 1/2$, and is infinite otherwise. The essential supremum $\|d\nu'/d\nu\|_\infty$ is finite exactly when $\alpha, \beta \geq 1$, with the usual interior-mode formula when $\alpha, \beta > 1$. The reverse KL used to index the robustness tables is

$$D_{\mathrm{KL}}(\mathrm{Unif}\|\mathrm{Beta}(\alpha, \beta)) = \log B(\alpha, \beta) + \alpha + \beta - 2.$$

These constants are useful for interpreting the experimental tables, but the robustness mechanism itself is entirely captured by the profile bounds above.

The results above are deliberately environment-agnostic and help identify generic failure modes: robustness can fail only when the deployment prior emphasizes regions where the frozen controller is inaccurate or slow, or when ν' leaves the support region on which the controller was effectively trained.

C.11 Sample Complexity: Value Estimation vs Argmax Localization

The same observation model can support different recommendation tasks. Here we compare their optimal query complexities, separately from the numerical comparison of C-ICPE and BO value estimators. In this subsection we study the sample complexity of estimating the maximum value of a gaussian process $F \sim \mathrm{GP}(0, k_\ell)$ on $D = [0,1]^d$, and consider an RBF kernel $k_\ell(x, x') := \exp\left(-\frac{\|x-x'\|^2}{2\ell^2}\right)$ with $x, x' \in D$.

We let the unknown lengthscale be random:

$$\Lambda \sim \nu, \qquad \mathrm{supp}(\nu) \subseteq [\ell_-, \ell_+] \subset (0, \infty).$$

The learner observes

$$Y_t = F(a_t) + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \sigma^2),$$

where $a_t \in D$ is chosen adaptively from the past history and $\sigma > 0$ is known.

Define

$$F^\star = \max_{x \in D} F(x), \qquad D^\star = \arg\max_{x \in D} F(x),$$

and write $X^\star$ whenever the maximizer is unique.

Argmax localization complexity. For $r > 0$ and $\delta \in (0,1)$, define the Bayesian argmax-localization complexity

$$T_{\mathrm{arg},\nu}(r, \delta) = \inf_{\mathcal{A}} \mathbb{E}_{\Lambda,F,\xi}[\tau],$$

where the infimum is over all sequential algorithms $\mathcal{A}$ that output $\hat{X}$ and satisfy

$$\mathbb{P}_{\Lambda,F,\xi}\left(\inf_{X^\star \in D^\star} \|\hat{X} - X^\star\| \le r\right) \ge 1 - \delta.$$

Max-value estimation complexity. Similarly, define the Bayesian max-value-estimation complexity

$$T_{\mathrm{val},\nu}(\epsilon, \delta) = \inf_{\mathcal{A}} \mathbb{E}_{\Lambda,F,\xi}[\tau],$$

where the infimum is over all sequential algorithms that output $\hat{v}$ and satisfy

$$\mathbb{P}_{\Lambda,F,\xi}\left(|\hat{v} - F^\star| \le \epsilon\right) \ge 1 - \delta.$$

All probabilities and expectations are under the joint hierarchical law of (Λ, F), the observation noise, and any internal randomness of the algorithm.

Main result. At a nondegenerate interior maximum, the local value gap $F(X^\star) - F(X)$ is of order $\|X - X^\star\|_2^2$. Thus localization to radius $r = O(\sqrt{\epsilon})$ controls value regret to order ϵ, but does not itself estimate the numerical peak height. We compare these two accuracy requirements below.

Theorem C.6 (Bayesian value–argmax separation)**.** *Consider the hierarchical RBF-GP model: $\Lambda \sim \nu$ with $\mathrm{supp}(\nu) \subseteq [\ell_-, \ell_+] \subset (0, \infty)$, and $F \mid \Lambda = \ell \sim \mathrm{GP}(0, k_\ell)$ on $D = [0,1]^d$, with observations $Y_t = F(a_t) + \xi_t$, $\xi_t \sim \mathcal{N}(0, \sigma^2)$.*

Fix $\delta \in (0,1)$. Suppose there exists $\eta > 0$ such that $\beta_\eta = \mathbb{P}_{\Lambda,F}(\mathcal{I}_\eta^c) < \delta$, where $\mathcal{I}_\eta$ is defined in Definition 1. Then:

1. ***(Argmax upper bound.)** There exist constants $B < \infty$ and $C < \infty$, depending on ν, d, δ, σ but not on ϵ, such that for all sufficiently small $\epsilon > 0$,*

$$T^{\mathrm{Bayes}}_{\mathrm{arg},\nu}(\sqrt{\epsilon}, \delta) \le B + C\,\sigma^2\,\epsilon^{-3/2} \log\log \frac{1}{\epsilon}.$$

2. ***(Value lower bound.)*** *There exist constants $c > 0$ and $c' < \infty$, depending on ν, d, δ, σ but not on ϵ, such that for all $\epsilon > 0$,*

$$T^{\mathrm{Bayes}}_{\mathrm{val},\nu}(\epsilon, \delta) \geq c\,\frac{\sigma^2}{\epsilon^2}\log\frac{1}{\delta} - c'.$$

Consequently,

$$\lim_{\epsilon \to 0} \frac{T^{\mathrm{Bayes}}_{\mathrm{val},\nu}(\epsilon, \delta)}{T^{\mathrm{Bayes}}_{\mathrm{arg},\nu}(\sqrt{\epsilon}, \delta)} = \infty.$$

Thus, under the high-probability interior regularity condition, max-value estimation is asymptotically harder than argmax localization in the fully Bayesian hierarchical RBF model. The limit holds at fixed δ satisfying $\beta_\eta < \delta$; it is not a claim uniform over arbitrarily small δ on a domain with boundary.

Proof. Set $\alpha \in (\beta_\eta, \delta)$. Since $\beta_\eta < \alpha$, Lem. 19 provides deterministic constants $\rho, \mu, L, M, \Gamma > 0$ and an event $\mathcal{E}_\alpha$ with $\mathbb{P}(\mathcal{E}_\alpha) \geq 1 - \alpha$.

Part 1. We apply Thm. C.8 with $r = \sqrt{\epsilon}, \delta_0 = \delta_1 = (\delta - \alpha)/2$, giving

$$T^{\mathrm{Bayes}}_{\mathrm{arg},\nu}(\sqrt{\epsilon}, \delta) \leq B + O\left(\sigma^2 \epsilon^{-3/2} \log\log\frac{1}{\epsilon}\right).$$

Part 2. By Thm. C.7, with $C_- = \exp(-d/(2\ell_-^2))$ and $C_+ = \exp(d/(4\ell_-^2))$,

$$T^{\mathrm{Bayes}}_{\mathrm{val},\nu}(\epsilon, \delta) \geq \frac{\sigma^2}{C_+^2}\left[\frac{C_-^2\, z^2_{1-\delta/2}}{\epsilon^2} - 1\right]_+.$$

For fixed δ and small ϵ, using $z^2_{1-\delta/2} \geq c_0 \log(1/\delta)$, this gives

$$T^{\mathrm{Bayes}}_{\mathrm{val},\nu}(\epsilon, \delta) \geq c\,\frac{\sigma^2}{\epsilon^2}\log\frac{1}{\delta} - c'.$$

Separation.

$$\frac{T^{\mathrm{Bayes}}_{\mathrm{val},\nu}(\epsilon, \delta)}{T^{\mathrm{Bayes}}_{\mathrm{arg},\nu}(\sqrt{\epsilon}, \delta)} \geq \frac{c\,\sigma^2\epsilon^{-2}\log(1/\delta) - c'}{B + C\,\sigma^2\,\epsilon^{-3/2}\log\log(1/\epsilon)} \longrightarrow \infty \qquad \text{as } \epsilon \downarrow 0,$$

since $\epsilon^{-2} \gg \epsilon^{-3/2}\log\log(1/\epsilon)$. □

So value estimation can be asymptotically harder than argmax localization. Even at the radius where localizing the argmax would "in principle" tell you the value to accuracy ϵ, the localization itself is cheaper than directly estimating the value. The reason is simple: the algorithm exploits the geometry (gradient information, smoothness) to localize at a fast rate, but it does not help with the problem of estimating the height of a function.

C.11.1 Max-value estimation complexity

To derive a lower bound on $T_{\mathrm{val},\nu}(\epsilon, \delta)$, we convert the problem of estimating the maximum into the problem of estimating a parameter Θ. We note that the lower bound is not tight, but for our purpose (of showing that the argmax localization is easier) this is not important, as our goal is to show that even this approximate lower bound still yields an harder problem. In particular, we assume the learner has access to a particular quantity W_ℓ that appears in the proof. We obtain the following result (note that the constants are not optimized, and the only goal is to show the dependency on ϵ).

The constants below are not optimized. For fixed dimension, prior, noise, and confidence, the positive coefficient of ϵ^{-2} suffices for the asymptotic separation from argmax localization.

Theorem C.7. *Consider the problem of estimating $F^\star = \max_{x\in D} F(x)$ where $D = [0,1]^d, F \mid \Lambda \sim \mathrm{GP}(0, k_\Lambda)$ with kernel $k_\ell(x,y) = \exp(-\|x-y\|_2^2/(2\ell^2))$ and $\Lambda \sim \nu$ with continuous support in $[\ell_-, \ell_+] \subset (0,\infty)$. Consider any sequential algorithm $\mathcal{A}$ that in each round selects a_t and observes $Y_t = F(a_t) + \xi_t$, where $\xi_t \sim \mathcal{N}(0,\sigma^2)$. If the algorithm outputs $\hat{v}$ at some stopping time τ satisfying $\mathbb{P}_{\Lambda,F}(|\hat{v} - F^\star| \le \epsilon) \ge 1-\delta$, then we say that the algorithm is Bayesian (ϵ,δ)-correct. Then, for any Bayesian (ϵ,δ)-correct algorithm we have that*

$$\inf_{\mathcal{A}:(\epsilon,\delta)-\textit{correct}} \mathbb{E}[\tau] \ge \frac{\sigma^2}{C_+^2}\left[\frac{C_-^2 z_{1-\delta/2}^2}{\epsilon^2} - 1\right]_+,$$

where $z_{1-\delta/2} = \Phi^{-1}(1-\delta/2)$[2], $C_- = \exp(-d/(2\ell_-^2))$ and $C_+ = \exp(d/(4\ell_-^2))$.

Proof. The proof relies on converting the problem into that of estimating a parameter Θ. We use the property of independent Gaussian r.v.: for jointly Gaussian X, Y we have that $X = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}Y + W$, where W is an independent zero-mean Gaussian.

In this case, we take $Y = \Theta_\ell$ and $X = F$, where we define Θ_ℓ as

$$\Theta_\ell \coloneqq \frac{1}{s_\ell}\int_D F(x)\mathrm{d}x, \qquad s_\ell^2 \coloneqq \int_{D\times D} k_\ell(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Computing the covariance, we obtain

$$\begin{aligned}
\mathrm{Cov}(F(x),\Theta_\ell) &= \mathbb{E}[F(x)\Theta_\ell], \\
&= \frac{1}{s_\ell}\mathbb{E}\left[F(x)\int_D F(y)\mathrm{d}y\right], \\
&= \frac{1}{s_\ell}\int_D \mathbb{E}\left[F(x)F(y)\right]\mathrm{d}y, \\
&= \frac{1}{s_\ell}\int_D k_\ell(x,y)\,\mathrm{d}y =: c_\ell(x).
\end{aligned}$$

Therefore, we can rewrite the observation using that $F(x) = c_\ell(x)\Theta_\ell + W_\ell(x)$ for some GP W_ℓ with 0-mean, and thus

$$Y_t = [c_\ell(a_t)\Theta_\ell + W_\ell(a_t)] + \xi_t.$$

From the expression of $F(x)$ we observe that

$$V(\theta) = \max_x [c_\ell(x)\theta + W_\ell(x)],$$

is increasing in θ. Denote a maximizer by x_θ, then, for $\theta' > \theta$ we have

$$V(\theta') \ge c_\ell(x_\theta)\theta' + W_\ell(x_\theta) = V(\theta) + c_\ell(x_\theta)(\theta'-\theta).$$

Similarly,

$$V(\theta) \ge c_\ell(x_{\theta'})\theta + W_\ell(x_{\theta'}) = V(\theta') + c_\ell(x_{\theta'})(\theta-\theta').$$

We now derive bounds on c_ℓ. From the definition

$$c_\ell(x) = \frac{1}{s_\ell}\int_D k_\ell(x,y)\,\mathrm{d}y,$$

since D is compact, and k_ℓ is the RBF kernel, we have that $k_\ell(x,y) \le 1 \Rightarrow c_\ell(x) \le 1/s_\ell$ and $k_\ell(x,y) \ge \exp\left(-d/(2\ell^2)\right) \Rightarrow c_\ell(x) \ge \exp\left(-d/(2\ell^2)\right)/s_\ell$. Since $s_\ell^2 \ge \exp\left(-d/(2\ell^2)\right)$ and $s_\ell^2 \le 1$, we find

$$\underbrace{\exp\left(-d/(2\ell_-^2)\right)}_{=:C_-} \le c_\ell(x) \le \underbrace{\frac{1}{\exp\left(-d/(4\ell_-^2)\right)}}_{=:C_+}.$$

Then, from the bounds above on V we find

$$V(\theta') \ge V(\theta) + C_-(\theta'-\theta), \qquad V(\theta) \ge V(\theta') - C_+(\theta'-\theta),$$

[2]Inverse of the CDF of a standard normal distribution.

leading to

$$C_-(\theta' - \theta) \le V(\theta') - V(\theta) \le C_+(\theta' - \theta).$$

Therefore V is bi-Lipschitz, and thus invertible. Choosing $u = V(\theta)$ and $v = V(\theta')$, we obtain

$$|V^{-1}(u) - V^{-1}(v)| \le \frac{1}{C_-}|u - v|.$$

Therefore, with $\hat{v} = V(\hat{\Theta})$ and $F^\star = V(\Theta)$ we get

$$|\hat{\Theta} - \Theta| \le \frac{1}{C_-}|\hat{v} - F^\star|.$$

Reveal (Λ, W_Λ) to the learner. Conditional on this side information, the normalized latent coefficient Θ_Λ is still standard normal and hence is independent of the revealed pair. Subtracting $W_\Lambda(a_t)$ leaves observations $c_\Lambda(a_t)\Theta_\Lambda + \xi_t$, with predictable coefficients bounded by C_+. Since V is onto $\mathbb{R}$, a value estimate $\hat{v}$ yields $\hat{\Theta} = V^{-1}(\hat{v})$ and

$$|\hat{\Theta} - \Theta_\Lambda| = |V^{-1}(\hat{v}) - V^{-1}(F^\star)| \le |\hat{v} - F^\star|/C_-.$$

A Bayesian (ϵ, δ)-correct value estimator therefore gives an $(\epsilon/C_-, \delta)$-correct scalar estimator. Apply Lem. 18, allowing independent initial side information, to this easier problem. Therefore,

$$\inf_{\mathcal{A}:(\epsilon,\delta)-\text{correct}} \mathbb{E}[\tau] \ge \frac{\sigma^2}{C_+^2}\left[\frac{C_-^2 z_{1-\delta/2}^2}{\epsilon^2} - 1\right]_+,$$

□

Lemma 18 (Scalar estimation lemma)**.** *Let $\Theta \sim \mathcal{N}(0,1)$ and let any initial side information and learner randomization be independent of Θ and the observation noises. Observe $Z_t = C_t\Theta + \xi_t$, where $|C_t| \le C$, $C > 0$, is measurable before observation t, and the noises are independent $\mathcal{N}(0, \sigma^2)$ variables, independent of Θ and the initial side information, with $\sigma > 0$. Let τ be a stopping time such that at τ the algorithm outputs $\hat{\Theta}$ satisfying $\mathbb{P}(|\hat{\Theta} - \Theta| \le \epsilon) \ge 1 - \delta$ for $\epsilon > 0, \delta \in (0,1)$. Then*

$$\inf_{\mathcal{A}:(\epsilon,\delta)-\textit{correct}} \mathbb{E}[\tau] \ge \frac{\sigma^2}{C^2}\left[\frac{z_{1-\delta/2}^2}{\epsilon^2} - 1\right]_+,$$

where $z_{1-\delta/2} = \Phi^{-1}(1 - \delta/2)$.

Proof. Let $\mathcal{H}_t$ contain the initial side information, learner randomization, and the first t observations. At a realized history, the predictable coefficients are fixed and the conditional density is proportional to

$$\exp\left(-\frac{\theta^2}{2} - \frac{1}{2\sigma^2}\sum_{s=1}^{t}(Z_s - C_s\theta)^2\right).$$

Thus $\Theta \mid \mathcal{H}_t$ is Gaussian with parameters (m_t, v_t), even though the adaptively collected variables need not be jointly Gaussian. Its precision is

$$P_t = v_t^{-1} = 1 + \frac{1}{\sigma^2}\sum_{s=1}^{t} C_s^2.$$

If $\mathbb{E}\tau = \infty$, the lower bound is immediate. Otherwise $\tau < \infty$ almost surely, and the posterior identity holds on each event $\{\tau = t\}$. For any $\mathcal{H}_\tau$-measurable estimate u,

$$\mathbb{P}(|u - \Theta| \le \epsilon \mid \mathcal{H}_\tau) = \mathbb{P}(\Theta \in [u - \epsilon, u + \epsilon] \mid \mathcal{H}_\tau).$$

This quantity is maximized when the interval is centered in m_τ, therefore

$$\mathbb{P}(|u - \Theta| \le \epsilon \mid \mathcal{H}_\tau) \le \Phi\left(\frac{\epsilon}{\sqrt{v_\tau}}\right) - \Phi\left(-\frac{\epsilon}{\sqrt{v_\tau}}\right) = 2\Phi\left(\frac{\epsilon}{\sqrt{v_\tau}}\right) - 1.$$

Then, for any Bayesian (ϵ, δ)-PAC algorithm we have

$$1 - \delta \le \mathbb{E}[\mathbb{P}(|u - \Theta| \le \epsilon \mid \mathcal{H}_\tau)] \le 2\mathbb{E}[\Phi(\epsilon\sqrt{P_\tau})] - 1.$$

Since Φ is nondecreasing and concave on $[0,\infty)$, $s \mapsto \Phi(\epsilon\sqrt{s})$ is concave for $s \geq 0$. Jensen's inequality therefore gives

$$1 - \delta \leq 2\Phi\left(\epsilon\sqrt{\mathbb{E}[P_\tau]}\right) - 1.$$

Taking the inverse, and defining $z_{1-\delta/2} = \Phi^{-1}(1-\delta/2)$, we find

$$z_{1-\delta/2} \leq \epsilon\sqrt{\mathbb{E}[P_\tau]}.$$

Hence, we are just let with bounding P_τ :

$$P_\tau \leq 1 + \frac{1}{\sigma^2}C^2\tau,$$

from which we get

$$z_{1-\delta/2} \leq \epsilon\sqrt{1 + \frac{C^2}{\sigma^2}\mathbb{E}[\tau]}.$$

Therefore

$$\mathbb{E}[\tau] \geq \frac{\sigma^2}{C^2}\left[\frac{z_{1-\delta/2}^2}{\epsilon^2} - 1\right]_+$$

□

C.11.2 ARGMAX LOCALIZATION COMPLEXITY

We now study the problem of locating the argmax of a GP. Recall that for $r > 0$ and $\delta \in (0,1)$, define the Bayesian argmax-localization complexity

$$T_{\mathrm{arg},\nu}(r,\delta) = \inf_{\mathcal{A}} \mathbb{E}_{\Lambda,F,\xi}[\tau],$$

where the infimum is over all sequential algorithms $\mathcal{A}$ that output $\hat{X}$ and satisfy

$$\mathbb{P}_{\Lambda,F,\xi}\left(\inf_{X^\star \in D^\star} \|\hat{X} - X^\star\| \leq r\right) \geq 1 - \delta.$$

Our goal is to provide a meaningful upperbound on $T_{\mathrm{arg},\nu}(r,\delta)$. We provide an algorithm, T-BAL (Two-Stage Bayesian Argmax Localization; see Algorithm 2), that locates the argmax of a GP with a finite number of samples. The algorithm works in two phases: we first locate the nice region where the argmax lies, and then perform gradient ascent on that region. The analysis relies on the argmax being away from the boundary of D. Therefore, we introduce the following regularity event.

Definition 1 (Interior regularity event)**.** For $\eta > 0$, let $\mathrm{dist}(D^\star, \partial D) = \inf_{X \in D^\star} \mathrm{dist}(X, \partial D)$ and define

$$\mathcal{I}_\eta = \{\mathrm{dist}(D^\star, \partial D) \geq \eta\}, \qquad \beta_\eta = \mathbb{P}_{\Lambda,F}(\mathcal{I}_\eta^c).$$

One can show that on $\mathcal{I}_\eta$, the set of maximizer is a singleton almost surely (i.e., $D^\star = \{X^\star\}$), and similarly one can show that the Hessian is non-degenerate in $X^\star$. In fact, we have the following.

Remark 13. *An RBF process has a C^∞ version on a neighborhood of D. Almost surely, its critical points in D are nondegenerate, finite in number, and have distinct values. If the maximum is not attained on the boundary, its maximizer is unique.*

Justification of Remark 13. Fix $\ell > 0$. The Gaussian spectral density of the RBF kernel is strictly positive and has moments of every order. The derivative fields therefore have smooth versions. For $v \in \mathbb{S}^{d-1}$, the vector $(\nabla F(x), \nabla^2 F(x)v)$ is nondegenerate: a vanishing variance would force the polynomial $i\langle a,\omega\rangle - \langle b,\omega\rangle\langle v,\omega\rangle$ to vanish for every ω, hence $a = b = 0$. Its densities are uniformly bounded over the compact parameter set $D \times \mathbb{S}^{d-1}$. On the event that this field has Lipschitz constant at most M, a zero implies an $O(Mh)$ value at a point of an h-net. There are $O(h^{-(2d-1)})$ net points and each small-ball probability is $O_M(h^{2d})$. Letting $h \downarrow 0$, then $M \uparrow \infty$, rules out degenerate critical points.

For distinct points at distance at least $1/n$, apply the same argument to $(\nabla F(x), \nabla F(y), F(x) - F(y))$. Its $2d+1$ coordinates are nondegenerate, since the corresponding distinct-point exponential

polynomials are linearly independent on the full spectral support. A net of the $2d$-dimensional parameter set again gives a probability $O_M(h)$ of a zero. Taking the countable union over n rules out equal critical values. Nondegenerate critical points are isolated; compactness and continuity of the gradient make their number finite. An interior maximizer is a critical point, so distinct critical values give uniqueness when all maximizers are interior. These properties hold conditionally for every $\ell > 0$ and hence under the hierarchical lengthscale prior as well. □

We work under $\mathcal{I}_\eta$: this not only allows the maximizer to be unique, but also allows us to be at-least at a distance η from the boundary, where it is more degenerate. Then, we can show the existence of the following constants.

Lemma 19 (Regularity under the RBF prior)**.** *Define $B(x, \rho)$ to be the ball centered around x of radius ρ with the Euclidean norm $\|\cdot\|_2$. Fix $\eta > 0$ and suppose $\beta_\eta < 1$. For every $\alpha \in (\beta_\eta, 1)$ there exist deterministic constants $\rho_\alpha, \mu_\alpha, L_\alpha, M_\alpha, \Gamma_\alpha > 0$ and an event $\mathcal{E}_\alpha \subseteq \mathcal{I}_\eta$ such that $\mathbb{P}_{\Lambda,F}(\mathcal{E}_\alpha) \geq 1 - \alpha$, and on $\mathcal{E}_\alpha$ the following properties hold:*

$$B(X^\star, \rho_\alpha) \subset D,$$
$$\mu_\alpha I \preceq -\nabla^2 F(x) \preceq L_\alpha I \qquad \forall x \in B(X^\star, \rho_\alpha),$$
$$\sup_{x \in B(X^\star, \rho_\alpha)} \|\nabla^3 F(x)\|_{\mathrm{op}} \leq M_\alpha,$$

and

$$F^\star - \sup_{x \notin B(X^\star, \rho_\alpha/2)} F(x) \geq \Gamma_\alpha.$$

Proof. Let $\gamma = \alpha - \beta_\eta$.

Under $\mathcal{I}_\eta$ the maximizer $X^\star$ is in the interior, and unique almost surely. Define $\lambda^\star = \lambda_{\min}(-\nabla^2 F(X^\star))$: under $\mathcal{I}_\eta$ we have that $\lambda^\star > 0$ almost surely, hence, there exists $\mu_\alpha > 0$ s.t. $\mathbb{P}(\mathcal{I}_\eta \bigcap \{\lambda^\star/2 < \mu_\alpha\}) \leq \gamma/5$.

Next, we use the fact that $F \in C^\infty$ and D is compact to obtain an upper bound on the Hessian:

$$\sup_{x \in D} \|\nabla^2 F(x)\|_{\mathrm{op}} \leq L_0 < \infty.$$

Therefore, there exists $L_\alpha < \infty$ such that $\mathbb{P}(L_\alpha < L_0) \leq \gamma/5$. With a similar reasoning, we also obtain $\sup_{x \in D} \|\nabla^3 F(x)\|_{\mathrm{op}} \leq M_0$, and thus there exists $M_\alpha < \infty$ such that $\mathbb{P}(M_\alpha < M_0) \leq \gamma/5$.

Next, by continuity and the margin condition of $\mathcal{I}_\eta$, the following quantity exists and is strictly positive almost surely

$$\rho_0(\omega) = \sup \left\{ q < \eta : \sup_{X \in B(X^\star, q)} \|\nabla^2 F(X) - \nabla^2 F(X^\star)\|_{\mathrm{op}} \leq \frac{\lambda^\star}{2} \right\},$$

where ω denotes a realization of F. Therefore, there exists $\rho_\alpha > 0$ such that $\mathbb{P}(\mathcal{I}_\eta \bigcap \{\rho_0 < \rho_\alpha\}) \leq \gamma/5$.

Define then $\Gamma(\rho_\alpha) = F^\star - \sup_{X \notin B(X^\star, \rho_\alpha/2)} F(X)$: under $\mathcal{I}_\eta$ this is strictly positive since the supremum does not attain $F^\star$. Hence, there exists $\Gamma_\alpha > 0$ such that $\mathbb{P}(\mathcal{I}_\eta \bigcap \{\Gamma(\rho_\alpha) < \Gamma_\alpha\}) \leq \gamma/5$.

Define

$$\mathcal{E} = \mathcal{I}_\eta \cap \{\lambda^\star/2 \geq \mu_\alpha\} \cap \{L_\alpha \geq L_0\} \cap \{M_\alpha \geq M_0\} \cap \{\rho_0 \geq \rho_\alpha\} \cap \{\Gamma(\rho_\alpha) \geq \Gamma_\alpha\}.$$

Then, since

$$\mathcal{E}^c = \mathcal{I}_\eta^c \cup \{\lambda^\star/2 < \mu_\alpha\} \cup \cdots = \mathcal{I}_\eta^c \cup (\mathcal{I}_\eta \cap \{\lambda^\star/2 < \mu_\alpha\}) \cup \cdots$$

we have $\mathbb{P}(\mathcal{E}^c) \leq \beta_\eta + 5\frac{\gamma}{5} = \alpha$. Hence, under $\mathcal{E}$ we have that $B(X^\star, \rho_\alpha) \subset D$, and all the properties follow quite immediately. We only show the lower bound on the Hessian: for any $X \in B(X^\star, \rho_\alpha)$:

$$-\frac{\lambda^\star}{2} I \preceq \nabla^2 F(x) - \nabla^2 F(X^\star) \preceq \frac{\lambda^\star}{2} I.$$

Hence

$$-\nabla^2 F(x) \succeq -\nabla^2 F(X^\star) - \frac{\lambda^\star}{2} I \succeq \frac{\lambda^\star}{2} I \succeq \mu_\alpha I.$$

□

Algorithm 2 Two-stage Bayesian argmax localization (T-BAL)

Require: Target radius r, confidence δ, noise level σ^2, constants ρ, μ, L, M, Γ at some $\alpha \in (\beta_\eta, \delta)$ from Lem. 19; set $\delta_0 = \delta_1 = (\delta - \alpha)/2$.

▷ **Stage 1: Coarse grid search**

Set $h \leftarrow \min\{\rho/8, \sqrt{3\Gamma/(2L)}\}$
Construct h-net $\mathcal{G}$ of D with $|\mathcal{G}| \le C_d h^{-d}$
Set $n_0 \leftarrow \lceil 128\sigma^2\Gamma^{-2}\log(2|\mathcal{G}|/\delta_0)\rceil$
for each $g \in \mathcal{G}$ **do**
    Query $F(g)$ exactly n_0 times; compute sample mean $\hat{F}(g)$
end for
$x_0 \leftarrow \arg\max_{g\in\mathcal{G}} \hat{F}(g)$

▷ **Stage 2: Local finite-difference gradient ascent**

Set $q \leftarrow 1 - 3\mu/(4L), e_0 \leftarrow \rho/2$ and $K \leftarrow \max\left\{0, \left\lceil \frac{\log(\rho/(2r))}{-\log q} \right\rceil\right\}$
for $k = 0, \dots, K-1$ **do**
    $e_k \leftarrow e_0\, q^k$
    $s_k \leftarrow \min\left\{\rho/8,\ \sqrt{3\mu e_k/(4\sqrt{d}\,M)}\right\}$
    $n_k \leftarrow \lceil 64d \cdot \sigma^2\mu^{-2}e_k^{-2}s_k^{-2}\log(2dK/\delta_1)\rceil$
    If any $x_k \pm s_k e_j$ lies outside D, return $\mathrm{Proj}_D(x_k)$ and terminate.
    for $j = 1, \dots, d$ **do**
        Query $F(x_k + s_k e_j)$ and $F(x_k - s_k e_j)$ each n_k times
        $\hat{g}_{k,j} \leftarrow \frac{\bar{Y}(x_k+s_k e_j) - \bar{Y}(x_k - s_k e_j)}{2s_k}$
    end for
    $x_{k+1} \leftarrow x_k + \frac{1}{L}\hat{g}_k$
end for
return $\hat{X} = \mathrm{Proj}_D(x_K)$

Argmax upper bound. On $\mathcal{E}_\alpha$, the Hessian bounds imply

$$\frac{\mu_\alpha}{2}\|X - X^\star\|_2^2 \le F(X^\star) - F(X) \le \frac{L_\alpha}{2}\|X - X^\star\|_2^2$$

inside the regular basin. The upper curvature constant L_α therefore determines a sufficient radius for controlling value regret. We now prove the localization bound used in Thm. C.6.

Analysis of T-BAL (Algorithm 2). We provide now an algorithm for argmax localization in Gaussian processes. The algorithm, Two-Stage Bayesian Argmax Localization (T-BAL), outlined in Algorithm 2, works in two phases. In the first phase we try to find a point inside $B(X^\star, \rho/2)$, where ρ is described in Lem. 19. In the second phase, assuming we are inside the above ball, we perform gradient ascent to find the maximum using finite differences to approximate the gradients.

The second phase uses gradient ascent, starting from $x_0 \in B(X^\star, \rho/2)$.

Theorem C.8 (Sample Complexity of T-BAL)**.** *Consider the event $\mathcal{E}_\alpha$ in Lem. 19. Set $r \in (0, \rho/2), \delta \in (0,1), \alpha \in (\beta_\eta, \delta)$ and $\delta_0 = \delta_1 = (\delta - \alpha)/2$. Then, T-BAL (Algorithm 2) satisfies $\mathbb{P}_{F,\Lambda}(\|\hat{X} - X^\star\| \le r) \ge 1 - \delta$, using at-most*

$$B_\alpha + 2dK + O\left(\frac{\sigma^2}{\mu^2}\max(2, L/\mu)\left[\frac{d^2}{\rho^2 r^2} + \frac{d^{5/2}M}{\mu r^3}\right]\log\frac{Ld\log(\rho/r)}{\mu\delta_1}.\right)$$

samples, where B_α is an appropriate finite constant for each α that does not depend on r.

Proof. In this proof we consider the analysis of the second stage, while the first stage and the constant B_α are provided in Prop. 13. The idea is to prove a bound on gradient ascent. We first show that one step of the ideal gradient ascent brings us closer to $X^\star$. We then find the value of s_k and n_k to compute the approximate gradient up to the desired accuracy. After that, we estimate how many iterations we need, and compute the total number of required samples.

One step gradient ascent. Let $x^+ = x + \frac{1}{L}\nabla F(x)$.

Define $\phi(t) = \nabla F(X^\star + t(x - X^\star))$. Since $\nabla F(X^\star) = 0$, using the fundamental theorem of calculus we have

$$\nabla F(x) = \phi(1) - \phi(0) = \int_0^1 \phi'(t)\mathrm{d}t = \underbrace{\int_0^1 \nabla^2 F(X^\star + t(x - X^\star))\mathrm{d}t}_{=:-A(x)}(x - X^\star).$$

Then

$$x^+ - X^\star = x - \frac{1}{L}A(x)(x - X^\star) - X^\star = (I - \frac{A(x)}{L})(x - X^\star).$$

The matrix $(I - \frac{A(x)}{L})$ has eigenvalues in $[0, 1 - \mu/L]$ for $x \in B(X^\star, \rho)$. Therefore,

$$\|x^+ - X^\star\| \le \left(1 - \frac{\mu}{L}\right)\|x - X^\star\|$$

so the contraction factor is $q_0 = 1 - \mu/L$.

Noisy gradients. Suppose we do not have access to the exact gradient, but only to a noisy gradient $\hat{g}_k$ in round k. Assume the noisy gradient satisfy

$$\|\hat{g}_k - \nabla F(x_k)\| \le \frac{\mu e_k}{4},$$

where $e_k = e_0 q^k$. Then

$$\begin{aligned}
\|x_{k+1} - X^\star\| &= \|x_k + \frac{1}{L}\hat{g}_k - X^\star\|, \\
&= \|x_k + \frac{1}{L}\nabla F(x_k) + \frac{1}{L}\hat{g}_k - \frac{1}{L}\nabla F(x_k) - X^\star\|, \\
&\le \|x_k + \frac{1}{L}\nabla F(x_k) - X^\star\| + \frac{1}{L}\|\hat{g}_k - \nabla F(x_k)\|, \\
&\le \left(1 - \frac{\mu}{L}\right)\|x_k - X^\star\| + \frac{\mu e_k}{4L}, \\
&\le \left(1 - \frac{3\mu}{4L}\right)e_k.
\end{aligned}$$

So noisy gradients still guarantee convergence as long as we can show $\|\hat{g}_k - \nabla F(x_k)\| \le \frac{\mu e_k}{4}$.

All expectations and noise probabilities in the following finite-difference calculation are conditional on F, Λ and the past before stage k. The sampled locations are then fixed and the new noises are independent. The induction $\|x_k - X^\star\| \le e_k \le \rho/2$ and $s_k \le \rho/8$ keeps the domain guard inactive on the success event. *Finite-difference gradient bound: bias term.* We now bound $\|\hat{g}_k - \nabla F(x_k)\|$ through a bias-variance decomposition:

$$\|\hat{g}_k - \nabla F(x_k)\| \le \|\hat{g}_k - \mathbb{E}[\hat{g}_k]\| + \|\mathbb{E}[\hat{g}_k] - \nabla F(x_k)\|.$$

We begin with the bias term $\|\mathbb{E}[\hat{g}_k] - \nabla F(x_k)\|$. Note that the noisy gradients are computed as follow for each direction e_j:

$$\hat{g}_{k,j} \leftarrow \frac{\bar{Y}(x_k + s_k e_j) - \bar{Y}(x_k - s_k e_j)}{2s_k},$$

where $\bar{Y}$ is an average over n_k samples of the values observed. The idea is to bound $|\mathbb{E}[\hat{g}_{k,j}] - \partial_j F(x_k)|$ using Taylor series and the fact that $\nabla^3 F$ is bounded in $B(X^\star, \rho)$ by Lem. 19.

Then, fix a direction e_j and let $\phi(t) = F(x_k + te_j)$. Then

$$\begin{aligned}
\phi(s_k) &= \phi(0) + \phi'(0)s_k + \frac{1}{2}\phi''(0)s_k^2 + \frac{s_k^3}{6}\phi'''(\xi_+), \\
\phi(-s_k) &= \phi(0) - \phi'(0)s_k + \frac{1}{2}\phi''(0)s_k^2 - \frac{s_k^3}{6}\phi'''(\xi_-),
\end{aligned}$$

where $\xi_+ \in (0, s_k), \xi_- \in (-s_k, 0)$. Then

$$\frac{\phi(s_k) - \phi(-s_k)}{2s_k} = \phi'(0) + \frac{s_k^2}{12}[\phi'''(\xi_+) + \phi'''(\xi_-)].$$

Hence, for s_k sufficiently small, $s_k < \rho/2$ (we choose $s_k < \rho/8$) we have $|\phi'''| \le M$, and therefore

$$\left|\frac{\phi(s_k) - \phi(-s_k)}{2s_k} - \phi'(0)\right| \le \frac{s_k^2 M}{6},$$

leading to

$$\|\mathbb{E}[\hat{g}_k] - \nabla F(x_k)\| \le \frac{\sqrt{d} s_k^2 M}{6}.$$

Setting $\frac{\sqrt{d} s_k^2 M}{6} \le \mu e_k/8$ yields

$$s_k \le \sqrt{\frac{3\mu e_k}{4\sqrt{d}M}}$$

and we set $s_k = \min\left\{\rho/8, \sqrt{\frac{3\mu e_k}{4\sqrt{d}M}}\right\}$. Furthermore, note that for this choice of s_k we can guarantee that we stay inside the nice region $B(X^\star, \rho)$: since $\|x_k - X^\star\| \le e_k \le \rho/2$ and $s_k \le \rho/8$, then

$$\|x_k \pm s_k e_j - X^\star\| \le \frac{5\rho}{8} < \rho.$$

Finite-difference gradient bound: variance term. We now bound the variance term $\|\hat{g}_k - \mathbb{E}[\hat{g}_k]\|$. We have

$$\hat{g}_{k,j} - \mathbb{E}[\hat{g}_{k,j}] = \frac{\bar{\xi}(x_k + s_k e_j) - \bar{\xi}(x_k - s_k e_j)}{2s_k}, \qquad \bar{\xi}(x) = \frac{1}{n_k}\sum_{i=1}^{n_k} \xi_i.$$

Therefore, $\mathrm{Var}(\hat{g}_{k,j}) = \frac{2\sigma^2}{4s_k^2 n_k}$. Then

$$\mathbb{P}(|\hat{g}_{k,j} - \mathbb{E}[\hat{g}_{k,j}]| > \mu e_k/(8\sqrt{d})) \le 2\exp\left(-\frac{\mu^2 e_k^2 s_k^2 n_k}{64d \cdot \sigma^2}\right).$$

Set the right hand-side smaller than $\delta_1/(dK)$, where K is the total number of iterations $k = 0, \dots, K-1$, to obtain

$$n_k \ge \frac{64d \cdot \sigma^2}{\mu^2 e_k^2 s_k^2} \log\frac{2dK}{\delta_1}.$$

Then, a union bound over $j = 1, \dots, d$ and $k = 0, \dots, K-1$, yields

$$\mathbb{P}(\exists j \in \{1, \dots, d\}, k \in \{0, \dots, K-1\} : |\hat{g}_{k,j} - \mathbb{E}[\hat{g}_{k,j}]| > \mu e_k/(8\sqrt{d})) \le \delta_1.$$

Hence, with probability $1 - \delta_1$ we have that

$$\|\hat{g}_k - \nabla F(x_k)\| \le \frac{\mu e_k}{8} + \frac{\mu e_k}{8} = \frac{\mu e_k}{4},$$

which is what we wanted to show.

Sample complexity. For $0 < r < e_0$, the chosen K is the smallest integer with $e_0 q^K \le r$. Hence $qr < e_K \le r$ and

$$K \le 1 + \frac{4L}{3\mu}\log(e_0/r).$$

For $p = 2, 3$,

$$\sum_{k=0}^{K-1} e_k^{-p} = \frac{e_K^{-p} - e_0^{-p}}{q^{-p} - 1} \le \frac{r^{-p}}{1 - q^p} \le \frac{4L}{3\mu} r^{-p},$$

where $1 - q^p \ge 1 - q = 3\mu/(4L)$. Since $\lceil u \rceil \le 1 + u$ and $s_k^{-2} \le 64/\rho^2 + 4\sqrt{d}M/(3\mu e_k)$,

$$\begin{aligned}\sum_{k=0}^{K-1} 2dn_k &\le 2dK + \frac{128d^2\sigma^2}{\mu^2}\log\frac{2dK}{\delta_1}\sum_{k=0}^{K-1} e_k^{-2} s_k^{-2} \\ &\le 2dK + \frac{128\sigma^2}{\mu^2}\frac{4L}{3\mu}\left[\frac{64d^2}{\rho^2 r^2} + \frac{4d^{5/2}M}{3\mu r^3}\right]\log\frac{2dK}{\delta_1}.\end{aligned}$$

This implies the stated bound, with the logarithm understood for $0 < r < \rho/2$. For fixed noise and regularity constants, $2dK = O(\log(1/r))$ is dominated by the displayed noisy refinement term as $r \downarrow 0$.

Connecting everything together. In phase 1 we have probablity δ_0 of failure, while in phase 2 we have probability δ_1 of failure. Since we work under the event $\mathcal{E}_\alpha$ with failure probability α, we have $\mathbb{P}(\text{failure}) \leq \delta_0 + \delta_1 + \alpha = \delta$. □

The first phase analysis is provided in the following proposition. We construct an h-net $\mathcal{G}$ of D that depends on the geometric of the problem (see Lem. 19). For each point in $\mathcal{G}$, we sample $F(g)$ exactly n_0 times, such that we have good concentration, and we return the point that achieves the maximum.

Proposition 13. *Consider Algorithm 2, and let $\delta_0 \in (0,1)$. Under $\mathcal{E}_\alpha$ (see Lem. 19), the first phase samples $B_\alpha \leq C_d h^{-d}\lceil 128\sigma^2\Gamma^{-2}\log(2C_d h^{-d}/\delta_0)\rceil$, where $C_d = (\sqrt{d}+2)^d$ queries. Furthermore, we have that $x_0 \in B(X^\star, \rho/2)$ with probability $1-\delta_0$.*

Proof. Consider the event $\mathcal{E}_\alpha$ in Lem. 19, and omit the subscript α for simplicity. We construct an h-net $\mathcal{G}$ of D such that for every $x \in D$ there exists $g \in \mathcal{G}$ satisfying $\|x - g\| \leq h$. Hence, there exists a point $g^\star \in \mathcal{G}$ satisfying $\|X^\star - g^\star\| \leq h$.

At each grid point g take n_0 samples and compute the sample mean $\hat{F}(g)$: by the Gaussian tail bound, we have

$$\mathbb{P}(|\hat{F}(g) - F(g)| > \Gamma/8) \leq 2\exp\left(-\frac{n_0\Gamma^2}{128\sigma^2}\right).$$

Choosing $n_0 = \frac{128\sigma^2}{\Gamma^2}\log\left(\frac{2|\mathcal{G}|}{\delta_0}\right)$, for some $\delta_0 \in (0,1)$, and taking a union bound over g, we obtain

$$\mathbb{P}(\exists g \in \mathcal{G} : |\hat{F}(g) - F(g)| > \Gamma/8) \leq \delta_0.$$

We now choose h small enough such that a lower bound on $\hat{F}(g^\star)$ upper bounds a valid upper bound on $\hat{F}(g)$ for $g \notin B(X^\star, \rho/2)$. Under the event $\mathcal{E} = \{\forall g \in \mathcal{G} : |\hat{F}(g) - F(g)| \leq \Gamma/8\}$, we have that for $g \notin B(X^\star, \rho/2)$

$$\hat{F}(g^\star) \geq F(g^\star) - \Gamma/8, \qquad \text{and} \qquad \hat{F}(g) \leq F(g) + \Gamma/8 \leq F^\star - \Gamma + \Gamma/8 = F^\star - \frac{7}{8}\Gamma,$$

where we used the fact from Lem. 19 that $F^\star - F(g) \geq \Gamma$ for $g \notin B(X^\star, \rho/2)$. To construct a lower bound on $F(g^\star)$, we use the gradient properties of F in $B(X^\star, \rho)$. To that aim, we need to ensure $g^\star$ is sufficiently inside the ball, that is, choose h small enough.

We require $h \leq \rho/2$, and for simplicity, we just set $h \leq \rho/8$. Then, $g^\star \in B(X^\star, \rho/2)$, nd since in the ball the function is smooth, with $\nabla F(X^\star) = 0$, we have

$$F(g^\star) \geq F(X^\star) - \frac{L}{2}\|X^\star - g^\star\|^2 \geq F^\star - \frac{L}{2}h^2.$$

Hence, we require

$$F^\star - \frac{L}{2}h^2 - \frac{\Gamma}{8} \geq F^\star - \frac{7}{8}\Gamma,$$

which is satisfied if $h^2 \leq \frac{3}{2L}\Gamma$. Hence, any point outside $B(X^\star, \rho/2)$ cannot upper bound $\hat{F}(g^\star)$.

Therefore, $x_0 = \arg\max_{g\in\mathcal{G}} \hat{F}(g)$ satisfies (ties between distinct grid means have probability zero under the independent positive-variance Gaussian noise) $x_0 \in B(X^\star, \rho/2)$ and

$$\hat{F}(x_0) \geq \hat{F}(g^\star) \geq F^\star - \frac{L}{2}\left(\min\{\rho/8, \sqrt{3\Gamma/(2L)}\}\right)^2 - \frac{\Gamma}{8},$$

Take a Cartesian grid with coordinate spacing at most $h/\sqrt{d}$. It is a Euclidean h-net and has at most $(\sqrt{d}/h + 2)^d \leq C_d h^{-d}$ points for $h \leq 1$. Multiplying by the stated integer replication count proves the bound. □

D Appendix: Algorithms

This appendix describes the implementation of C-ICPE used in the experiments. We keep the notation of the main text: H_t is the current history, $I_\phi(\cdot|H_t)$ is the inference distribution over the target $x^\star_\theta$, $Q_\psi(H_t, a)$ is the critic for a continuation action $a \in \mathcal{A}$, and $Q_\psi(H_t, a_{\text{stop}})$ is the value of stopping. The implementation follows the Lagrangian view in Sec. C.3: the inference model learns a stochastic selector, the critic learns the stop/continue Bellman comparison, and the actor rule proposes the next continuation action. The main practical point is that all three objects are trained from replay. For this reason we use target networks, conservative critic targets, and simple regularizers that keep early noisy estimates from determining the stopping boundary.

D.1 History Encoder and Time Pooling Layer

The inference network, the critic, and the learned TD3 actors use the same sequential template, although their parameters are separate. Each interaction step is embedded as a token by concatenating the query and the next observation, $u_s = [A_s; Y_{s+1}]$, and passing it through a small embedding network. The resulting sequence $(e_1, \dots, e_t)$ is processed by a causal sequential backbone. In the code this backbone can be an LSTM, an attention stack, or a recurrent/linear-attention variant. Padding is masked throughout, so replay batches can contain histories of different lengths without leaking future observations.

After obtaining hidden states $h_1, \dots, h_t \in \mathbb{R}^d$, we use a query-conditioned pooling layer. Given a query q, it computes

$$\alpha_i(q, H_t) = \frac{\exp(\langle W_q q, W_k h_i\rangle/\sqrt{d})}{\sum_{j=1}^{t} \exp(\langle W_q q, W_k h_j\rangle/\sqrt{d})}, \qquad v(q, H_t) = \sum_{i=1}^{t} \alpha_i(q, H_t) W_v h_i,$$

and then applies a feature-wise gate, $z(q, H_t) = v(q, H_t) \odot \text{silu}(W_m q)$. This is useful because the relevant part of the history depends on what the model is asked to do. For inference and stopping, the query is a time embedding, since we need a representation of the current prefix. For the continuation critic, the query is the candidate action a, so the same history can be read differently when evaluating different future measurements. The critic therefore has two readouts from the same encoded history: a time-conditioned readout for $Q_\psi(H_t, a_{\text{stop}})$ and an action-conditioned readout for $Q_\psi(H_t, a)$.

D.2 Replay Buffer and Prefix Sampling

Training is off-policy. A rollout samples a task $\theta \sim \nu$, interacts with the corresponding environment until the critic stops or the maximum horizon is reached, and stores the whole trajectory together with $x^\star_\theta$ and a flag indicating whether the trajectory ended by executing the stop action. Minibatches are built by first sampling trajectories and then sampling prefixes inside them. The prefix sampler uses a mixture of uniformly sampled prefixes, prefixes around the current average stopping time, and terminal prefixes. This gives the inference network examples from all time scales, but also gives the critic many prefixes close to the point where the decision changes from continue to stop.

If a trajectory terminates because the agent stops at time t, the training prefix for that terminal state is H_t, because no new observation is collected after the stop action. If the trajectory terminates only

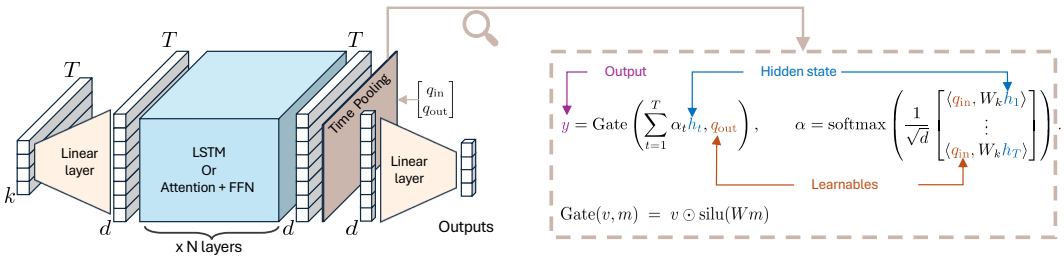


Figure 6: Sequential encoder and query-pooling readout used by the inference network I_ϕ, the critic Q_ψ, and the learned TD3 actors.

because the maximum horizon is reached, the last collected observation is included. This makes replay consistent with deployment: stopping is a decision made before paying for the next sample.

D.3 Inference Update and Its Regularization

For a prefix h and target $x_\theta^\star$, the inference model outputs the Gaussian mixture in Eq. (6). Weights are obtained from logits and variances are positive. The ordinary per-target mixture loss is evaluated stably as

$$-\log I_\phi(x_\theta^\star \mid h) = -\operatorname{logsumexp}_j\{\log \omega_{\phi,j}(h) + \log \varphi_{\phi,j}(x_\theta^\star \mid h)\}.$$

At stopping, all variants use Eq. (7), including the feasibility map and the fixed tie-break. The following robust single-Gaussian loss is the $J = 1$ variant.

The ideal objective in the main text is the negative log-likelihood in Eq. (8). In the implementation we use a robust version. Define the per-coordinate averaged NLL

$$\ell_\phi(h, x_\theta^\star) = -\frac{1}{d_{\mathcal{X}}} \sum_{j=1}^{d_{\mathcal{X}}} \log \mathcal{N}\big((x_\theta^\star)_j; \mu_{\phi,j}(h), \sigma_{\phi,j}^2(h)\big),$$

and

$$\rho_\tau(u) = \min\{u, 0\} + \tau \log\left(1 + \frac{[u]_+}{\tau}\right), \qquad [u]_+ = \max\{u, 0\}.$$

For Euclidean recommendation problems we train with

$$\mathcal{L}_{\mathrm{inf}}(\phi) = \mathbb{E}_{(h,x_\theta^\star)\sim\mathcal{B}}\left[\rho_\tau\big(\ell_\phi(h, x_\theta^\star)\big) + \alpha_{\mathrm{anc}}\,\mathrm{SmoothL1}\left(\mu_\phi(h), x_\theta^\star\right)\right]. \tag{42}$$

The robust transform is a simple way of saying that early bad prefixes should not dominate the variance head. A standard Gaussian NLL can become very large when the model underestimates uncertainty for one minibatch, and this can push the log-variance to extremes. The logarithmic tail keeps the ranking of ordinary examples, but reduces the influence of rare outliers. The SmoothL1 anchor is also important: the final decision is the mean $\mu_\phi(H_\tau)$, so we want the mean to remain a good deterministic recommendation even when the Gaussian still has large uncertainty. We additionally clamp the predicted log-standard deviations to fixed lower and upper bounds for numerical stability.

For the ϵ-best-arm problem, the target is directional. The magnitude of a vector is not meaningful once arms are represented on the sphere, and correctness is measured by cosine distance. Therefore recommendations and samples are normalized before they are evaluated. In this case we keep the robust NLL and replace the Euclidean anchor by a spherical alignment term. With $Z \sim I_\phi(\cdot|h)$,

$$\mathcal{L}_{\mathrm{sph}}(\phi) = -\mathbb{E}\left[\left\langle \frac{Z}{\|Z\|_2}, \frac{x_\theta^\star}{\|x_\theta^\star\|_2} \right\rangle\right]. \tag{43}$$

The reason is that an Euclidean anchor would penalize harmless radial errors, while the bandit loss only cares about the direction.

A target copy $I_{\bar\phi}$ is maintained by Polyak averaging. The critic never uses the online inference model inside its TD targets; it uses $I_{\bar\phi}$. This separation is important because the reward itself is learned through I_ϕ, and bootstrapping from a rapidly moving reward makes the stop/continue comparison unstable.

D.4 Reward, Critic Update, and Critic Regularization

The critic learns two quantities from the same replay prefixes. The stopping head learns the value of recommending now, and the continuation head learns the value of paying for one more observation and then acting optimally. We use the recommendation reward $\hat r(h, \theta) = \mathbf{1}\{L_\theta(\hat x_{\bar\phi}(h)) \le \epsilon\}$ from Eq. (9). The following sampled reward is an alternative, not the reward of the deterministic returned recommendation:

$$\hat r_{\bar\phi,m}(h, \theta) = \frac{1}{m}\sum_{k=1}^{m} \mathbf{1}\left\{L_\theta\left(X^{(k)}\right) \le \epsilon\right\}, \qquad X^{(k)} = \mathrm{Proj}_{\mathcal{X}}\, Z^{(k)}, \quad Z^{(k)} \overset{\mathrm{iid}}{\sim} I_{\bar\phi}(\cdot \mid h). \tag{44}$$

This alternative estimates success averaged over inference samples. It differs from the success of $\hat{x}_{\bar{\phi}}(h)$, even when its mean is feasible; see Sec. C.8.

Let d be the terminal flag for a replay transition, and let $a^+(h')$ be the target continuation action at the next prefix. This target action depends on the actor rule. For TS and TTPS the main-text specification uses $a^+(h') = \hat{x}_{\bar{\phi}}(h')$ without sampling. For TD3 it is the target actor action, again with small target-policy smoothing noise. The smoothing noise is much smaller than the posterior sampling noise used during rollouts.

When using twin critics, we scalarize target values by the conservative minimum

$$\bar{Q}_{\bar{\psi}}(h, a) = \min\{Q_{\bar{\psi},1}(h, a), Q_{\bar{\psi},2}(h, a)\},$$

and define

$$V_{\bar{\psi}}(h') = \max\left\{\bar{Q}_{\bar{\psi}}(h', a_{\text{stop}}), \bar{Q}_{\bar{\psi}}(h', a^+(h'))\right\}.$$

The TD targets are

$$y_{\text{stop}}(h, \theta) = \hat{r}(h, \theta), \tag{45}$$

$$y_{\text{cont}}(h, a, h', d, \theta) = -c + d\,\hat{r}(h', \theta) + (1 - d)V_{\bar{\psi}}(h'). \tag{46}$$

Thus the stopping head is directly supervised by the current confidence, while the continuation head is supervised by the gain from collecting the next observation. The action-head loss is applied only to prefixes that did not already stop, because a stopped transition has no genuine continuation action. The stopping head is trained on every prefix, since stopping is a valid action at every prefix. With $M = 1$ for non-stopped replay transitions and $M = 0$ for stopped transitions, the critic loss is

$$\mathcal{L}_Q(\psi) = \frac{1}{2}\mathbb{E}_{\mathcal{B}}\left[M\sum_{j=1}^{2}\left(Q_{\psi,j}(h, a) - y_{\text{cont}}\right)^2\right] + \frac{w_{\text{stop}}}{2}\mathbb{E}_{\mathcal{B}}\left[\sum_{j=1}^{2}\left(Q_{\psi,j}(h, a_{\text{stop}}) - y_{\text{stop}}\right)^2\right], \tag{47}$$

with the obvious single-critic version when twin critics are disabled. We also optionally clip the bootstrap values in a minibatch to moderate quantiles before taking the maximum in $V_{\bar{\psi}}$. It prevents a few very optimistic target values from propagating through replay and moving the stopping boundary too early.

At rollout time the default stopping test is

$$Q_\psi(H_t, a_{\text{stop}}) \geq Q_\psi(H_t, A_t),$$

where both quantities are scalarized by the conservative twin rule. This approximates the Bellman comparison in Thm. 3.1 using one proposal rather than an exact action maximization.

D.5 Actor Rules: TS, TTPS, and TD3

The actor rule only chooses continuation actions. The stop action is handled by the critic comparison above. We use three actor rules depending on the relation between the query space $\mathcal{A}$ and the recommendation space $\mathcal{X}$.

Thompson sampling. When $\mathcal{A} = \mathcal{X}$, we can use the inference distribution itself as the actor. The TS rule samples

$$J_t \sim \text{Cat}(\omega_\phi(H_t)), \qquad A_t = \text{Proj}_{\mathcal{A}}\left(\mu_{\phi,J_t}(H_t) + \sigma_{\phi,J_t}(H_t) \odot \xi_t\right), \quad \xi_t \sim \mathcal{N}(0, I). \tag{48}$$

A greedy mode instead queries $\hat{x}_\phi(H_t)$; it is a distinct rule from stochastic TS and must be identified separately. Early in a task, the learned posterior is broad and TS explores. Later, when the inference distribution contracts, the same rule automatically becomes exploitative. No separate actor is trained, which removes one source of approximation error and is appropriate when informative queries are themselves plausible recommendations. In the critic target, the next action for TS is the target recommendation, as in the main text.

Top-two posterior sampling. TTPS is also inference-based and therefore also assumes $\mathcal{A} = \mathcal{X}$. It first draws a Thompson sample Z_1. With probability $1/2$ this sample is used directly. With the remaining probability, the rule draws a challenger Z_2 and keeps the sample that is farther from the current recommendation, provided the current sample is too close to the recommendation. In

Algorithm 3 Implementation of C-ICPE

Initialize replay buffer $\mathcal{B}$, inference network I_ϕ, critic Q_ψ, actor rule $\mathrm{Act} \in \{\mathrm{TS}, \mathrm{TTPS}, \mathrm{TD3}\}$, target networks $I_{\bar{\phi}}, Q_{\bar{\psi}}$, and cost c.
// Training phase
while training is not over **do**
  Sample a batch of tasks $\theta \sim \nu$ and initialize their histories H_1.
  for $t = 1, \ldots, T_{\max}$ **do**
    Propose continuation actions A_t using TS, TTPS, or the TD3 actor with its current exploration schedule.
    Stop tasks satisfying $Q_\psi(H_t, a_{\mathrm{stop}}) \geq Q_\psi(H_t, A_t)$ after the warmup and minimum-time gates.
    Execute A_t on the remaining tasks, observe Y_{t+1}, and append (A_t, Y_{t+1}) to the histories.
  end for
  Store the completed trajectories, stop flags, terminal times, and targets $x^\star_\theta$ in $\mathcal{B}$.
  Sample replay prefixes and update I_ϕ with Eq. (8), or the stated $J = 1$ regularized variant. Evaluate the target recommendation with Eq. (9) and update Q_ψ using Eqs. (45) to (47).
  If using TD3, update π_ϑ on the delayed actor schedule using Eqs. (49) and (50).
  Polyak-update $I_{\bar{\phi}}, Q_{\bar{\psi}}$, and, when present, $\pi_{\bar{\vartheta}}$; update c with Eq. (51).
end while

// Deployment phase
Select and freeze a complete checkpoint using development data; apply the independent one-shot certification test.
If certification fails, **return** ABSTAIN; otherwise initialize a fresh test task.
for $t = 1, \ldots, T_{\max}$ **do**
  Propose A_t using the selected actor rule.
  if $Q_\psi(H_t, a_{\mathrm{stop}}) \geq Q_\psi(H_t, A_t)$ **then**
    return $\hat{x} = \hat{x}_\phi(H_t)$.
  end if
  Execute A_t, observe Y_{t+1}, and update H_{t+1}.
end for
return $\hat{x} = \hat{x}_\phi(H_{T_{\max}})$.

Euclidean tasks the distance is $\|z - \hat{x}_\phi(H_t)\|_2$; in ϵ-best-arm it is $1 - \langle z, \hat{x}_\phi(H_t)\rangle$. The intuition is that the selected component mean is the recommendation, while farther posterior samples represent plausible alternatives. TTPS spends part of the sampling budget checking those alternatives before the critic decides that stopping is safe. As for TS, the target action used by the critic is the target recommendation.

TD3 actor. TD3 is the actor rule used when we want to learn the continuation action from the critic, and especially when $\mathcal{A}$ and $\mathcal{X}$ are different objects. A target actor $\pi_{\bar{\vartheta}}$ is used in the critic target, while the online actor π_ϑ is updated on a delayed schedule. The target action is

$$a^+(h') = \tanh(\pi_{\bar{\vartheta}}(h')) + \zeta, \qquad \zeta \sim \mathcal{N}(0, \sigma^2_{\mathrm{tgt}} I),$$

with clipping or normalization depending on the domain. The target-policy smoothing noise ζ prevents the critic from learning sharp artificial peaks in action space, and the delayed actor update prevents the actor from chasing a critic that is still changing after every minibatch.

For a deterministic actor, the main update is

$$\mathcal{L}^{\mathrm{det}}_{\mathrm{act}}(\vartheta) = -\mathbb{E}_{h\sim\mathcal{B}}\left[Q_{\psi,1}\left(h, \tanh(\pi_\vartheta(h))\right)\right]. \tag{49}$$

The actor uses the first critic for the policy gradient, while twin-critic targets use the conservative minimum defined above.

When $\mathcal{A} = \mathcal{X}$ and the actor is Gaussian, we regularize the TD3 actor toward the inference distribution,

$$\mathcal{L}^{\mathrm{KL}}_{\mathrm{act}}(\vartheta) = -\mathbb{E}_{h\sim\mathcal{B}}\left[Q_{\psi,1}(h, A_\vartheta(h))\right] + \beta_{\mathrm{KL}}\mathbb{E}_{h\sim\mathcal{B}}\left[\mathrm{KL}\left(I_\phi(\cdot|h)\|\pi_\vartheta(\cdot|h)\right)\right], \qquad A_\vartheta(h) \sim \pi_\vartheta(\cdot|h). \tag{50}$$

During training we also perform random exploration to encourage parametric exploration (with small probability we sample a random action), and void collapsing of the actor.

D.6 Cost Update for Fixed Confidence

The scalar cost c is the implemented Lagrange tradeoff between confidence and sample complexity. We update it from the empirical stopped success rate. For a batch of completed rollouts, let

$$D_i = \begin{cases} \|\hat{x}_\phi(H_{\tau_i}^{(i)}) - x_{\theta_i}^\star\|_2, & \text{Euclidean tasks,} \\ 1 - \left\langle \hat{x}_\phi(H_{\tau_i}^{(i)}), x_{\theta_i}^\star \right\rangle, & \epsilon\text{-best-arm tasks,} \end{cases}$$

and define the optional smooth accuracy proxy below. The main-text update instead uses the empirical success $\hat{p} = B^{-1} \sum_i \mathbf{1}\{L_{\theta_i}(\hat{x}_\phi(H_{\tau_i}^{(i)})) \le \epsilon\}$. The alternative is

$$\hat{p}_{\text{smooth}} = \frac{1}{B} \sum_{i=1}^{B} \sigma\left(\frac{\epsilon - D_i}{\kappa\epsilon}\right).$$

The update is

$$c \leftarrow \text{Proj}_{[0,1]}\left(c - \eta_c\big((1 - \delta_{\text{train}}) - \hat{p}\big)\right). \tag{51}$$

Thus, if the observed correctness is below $1 - \delta_{\text{train}}$, the cost decreases and continuing becomes cheaper, so trajectories become longer. If correctness is above the target, the cost increases and stopping becomes more aggressive. The sign of the update is still exactly the dual intuition in Eq. (11).

E Appendix: Numerical Results

This appendix is organized as follows. The remainder of this introduction describes the protocol shared by all experiments: certification, the choice of δ_{train}, the ablations, compute, and confidence intervals. Sec. E.1 describes the synthetic benchmarks and Sec. E.2 the baselines and sample budgets. Sec. E.3 reports, for each synthetic benchmark and every (d, ε, σ) configuration, accuracy, sample complexity, certification p-values and the critic-free ablation. Sec. E.4 studies robustness to prior misspecification, and Sec. E.5 presents the geochemical task. Hyperparameters are listed in the `README.md` file of the code.

Certification protocol. For each seed and configuration, we freeze one checkpoint selected on development rollouts and apply the one-shot test of Sec. C.7 with $\delta = 0.1$ and $\alpha = 0.01$. We run $n = 20{,}000$ fresh tasks $\theta \sim \nu$, one rollout each, independent of training and of the evaluation episodes below, and certify the checkpoint if its number of successes K satisfies $K \geq k_{\text{test}}$. Here $k_{\text{test}} = \kappa(n, 1-\delta, \alpha) = 18{,}099$, where $\kappa(n, p, \alpha)$ is the smallest k with $\mathbb{P}(K \geq k) \leq \alpha$ for $K \sim \text{Bin}(n, p)$. The certification tables report the exact binomial p-values across the three seeds; a configuration is certified for every seed when the largest p-value is below α.

Choice of δ_{train}. The cost update in Eq. (11) targets accuracy $1 - \delta_{\text{train}}$ rather than $1 - \delta$, so that the test above has high power. Let $q^\star(n, k, \beta)$ be the smallest q with $\mathbb{P}(K \geq k) \geq \beta$ for $K \sim \text{Bin}(n, q)$, and fix the power $\beta = 0.99$. First, a checkpoint with accuracy at least $1 - \delta_{\text{ctrl}} = q^\star(n, k_{\text{test}}, \beta) = 0.9097$ passes the certification test with probability at least 0.99. Second, we add a margin for the gap between the training-time accuracy estimate and the checkpoint's true accuracy: at the looser level $\alpha_{\text{ctrl}} = 0.05$, let $k_{\text{ctrl}} = \kappa(n, 1 - \delta_{\text{ctrl}}, \alpha_{\text{ctrl}}) = 18{,}261$ and $1 - \delta_{\text{train}} = q^\star(n, k_{\text{ctrl}}, \beta) = 0.9176$. This gives $\delta = 0.1 > \delta_{\text{ctrl}} = 0.0903 > \delta_{\text{train}} = 0.0824$, and we use $\delta_{\text{train}} = 0.0824$ in all experiments. The per-query costs c of the certified checkpoints are reported in Tabs. 2, 6, 10, 14, 18 and 22.

Ablations and reference variants. Each ablation changes one component of `C-ICPE` and keeps the others, so that differences in accuracy and sample complexity can be attributed to that component.

- *Query rule.* `C-ICPE-Uniform` keeps the inference network, critic, and cost update, but queries uniformly at random; it measures the value of active exploration. On the sphere, `C-ICPE-RoundRobin` instead queries the canonical basis in round-robin order, as `LT&S round-robin` does; it tests whether the learned inference and stopping match a specialized fixed-confidence method when given the same queries.
- *Inference family.* `C-ICPE-TS-Mixture` replaces the single Gaussian with a mixture of Gaussians. Motivated by Thm. 4.1, it tests whether representing several posterior modes reduces the number of queries.
- *Stopping rule.* The variants marked '+CF' keep the trained inference network and actor, but replace the critic with a threshold on the inference network's own belief. At each step we draw 2000 samples $X^{(k)} \sim I_\phi(\cdot \mid H_t)$, let d_k be the distance between $X^{(k)}$ and the recommendation $\hat{x}_\phi(H_t)$, estimate $\hat{\mathbb{P}}(d > \varepsilon) = \frac{1}{2000} \sum_{k=1}^{2000} \mathbf{1}\{d_k > \varepsilon\}$, and stop once $\hat{\mathbb{P}}(d > \varepsilon) \leq \delta_{\text{train}}$. This isolates the role of the critic, which estimates the success of the returned recommendation from data, as in the stopping value of Thm. 3.1.
- *Recommendation.* On GP value estimation, `C-ICPE-TD3+GP` keeps the queries collected by `C-ICPE-TD3` and replaces the inference network with a GP fitted to them at test time; it separates the quality of the queries from that of the learned recommendation.
- *Reference rules.* On multi-lobe localization, `Moment-TS` and `Posterior-TS` use the exact posterior, projected to a Gaussian or kept exact; they separate the effect of the Gaussian family from neural approximation error.

The results of each ablation are discussed in the corresponding benchmark subsection of Sec. E.3 and in Sec. E.5.

Computational resources. All experiments were run on NVIDIA V100 GPU or NVIDIA L40S GPU. For the synthetic benchmarks, each C-ICPE training run takes approximately 12 hours. With 3

random seeds, at least 4 (ε, σ) configurations, and 3 dimensionalities per benchmark, the total training budget per synthetic task is $12 \times 3 \times 4 \times 3 = 432$ GPU-hours. For the geochemical exploration task, each training run takes approximately 70 hours; with 3 seeds, 2 ε configurations, and a single dimensionality ($d = 2$), the total is $70 \times 3 \times 2 = 420$ GPU-hours.

Confidence intervals via hierarchical bootstrap. To account for variability across both task instances and trajectory randomness, we report 95% confidence intervals computed via hierarchical bootstrap (Efron, 1992). For each trained model (seed), we sample 300 test environments $\theta \sim \nu$ and collect 15 independent trajectories per environment. The total variance of a statistic $\hat{\mu}$ (e.g., accuracy) decomposes as

$$\mathrm{Var}(\hat{\mu}) = \underbrace{\mathrm{Var}_s\big(\mathbb{E}[\hat{\mu} \mid s]\big)}_{\text{between-seed}} + \underbrace{\mathbb{E}_s\big[\mathrm{Var}_\theta\big(\mathbb{E}[\hat{\mu} \mid s, \theta]\big)\big]}_{\text{between-environment}} + \underbrace{\mathbb{E}_{s,\theta}\big[\mathrm{Var}(\hat{\mu} \mid s, \theta)\big]}_{\text{within-environment}},$$

where the first term captures variability due to training, the second due to which tasks are drawn from ν, and the third due to observation noise and policy stochasticity within a fixed task. A single bootstrap replicate is constructed by (i) resampling seeds with replacement, then (ii) for each resampled seed, resampling environments with replacement, then (iii) for each resampled environment, resampling trajectories with replacement, and computing the statistic on the resampled dataset. This three-level resampling preserves all components of variance. We draw 10,000 bootstrap replicates and report the 2.5% and 97.5% percentiles as the confidence interval.

Remark 14 (On testing the inference model)**.** *Several of our synthetic localization benchmarks fall close to the symmetric case described in Sec. C.8 (after Prop. 10): the loss is a distance to a selected target $x^\star_\theta$, so $\mathcal{X}_\epsilon(\theta)$ is a ball, interval, or cap around this target. In the optimization benchmarks, such as the value estimation task and the geochemical task, the loss is instead induced by value gaps and the success sets need not be symmetric or convex. These experiments therefore test the method beyond the setting where the Gaussian NLL mean has an exact Bayes-optimality interpretation.*

E.1 Synthetic Benchmarks: description

The synthetic benchmarks in Sec. 5 are designed to isolate different aspects of the continuous fixed-confidence problem. In binary search, every query returns a noisy comparison with the target. The ϵ-best-arm problem keeps the same idea of identifying $x^\star_\theta$, but changes the geometry to a sphere and makes the loss directional rather than Euclidean. Ackley minimization adds nuisance parameters and a multimodal response surface, so the agent has to learn an exploration rule that is not purely local. Multi-Lobe localization aims to test the effect of multimodal posterior. Finally, GP max-value estimation is included because it is the case where the query space and the recommendation space are different: the agent queries a location, but it recommends a scalar value. This is the setting where a TD3 actor is necessary, since posterior samples from the inference network are no longer valid actions.

Common protocol. Each training episode starts by sampling a fresh task parameter θ from the task prior ν. The agent then observes a sequential history $H_t = (A_1, Y_2, \ldots, A_{t-1}, Y_t)$ and either stops or selects a new action. We use a maximum horizon $t_{\max}$; if the learned stopping rule does not stop before this horizon, the episode is truncated and the final recommendation is still evaluated. For the synthetic experiments reported in the survival plots and correctness tables, we use $t_{\max} = 100$ unless otherwise stated.

All fixed-confidence runs use $\delta = 0.1$. The reported accuracy is

$$\widehat{\mathrm{Acc}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{L_{\theta_i}(\hat{x}_i) \le \epsilon\right\},$$

and the goal is to achieve accuracy at least $1 - \delta$ while minimizing the expected stopping time $\mathbb{E}[\tau]$. For Euclidean localization tasks we use $L_\theta(x) = \|x - x^\star_\theta\|_2$. For ϵ-best-arm we use the directional loss $L_\theta(x) = 1 - \langle x, x^\star_\theta \rangle$ after normalizing both vectors to the unit sphere. For GP max-value estimation the loss is the scalar absolute error $L_\theta(x) = |x - v^\star_\theta|$. Confidence intervals are computed with hierarchical bootstrap over test episodes and random seeds.

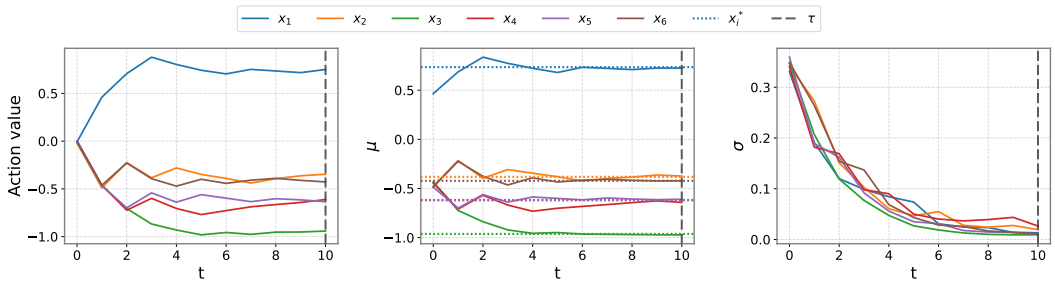


Figure 7: Binary search: visualization of how C−ICPE explores in 6 dimensions. From left to right: the query action (left), posterior mean (middle), and posterior standard deviation (right) along an exploration trajectory in the noisy binary search problem.

E.1.1 Noisy binary search.

In binary search the unknown target is $x_\theta^\star = \theta \in [-1, 1]^d$, sampled uniformly. A query $a \in [-1, 1]^d$ returns one noisy comparison per coordinate,

$$Y_{t,i} = \xi_{t,i} \operatorname{sign}(\theta_i - A_{t,i}), \qquad \mathbb{P}(\xi_{t,i} = 1) = 1 - p, \qquad \mathbb{P}(\xi_{t,i} = -1) = p, \tag{52}$$

independently over i and t. We use this problem because the statistically useful action is interpretable: a good policy should place queries near the current posterior median in each coordinate and shrink the feasible region. This makes binary search a sanity check for the inference network and critic. If the inference model does not contract its posterior, or if the critic cannot recognize when the posterior radius is below ϵ, the method will fail even in this simple setting. Since $\mathcal{A} = \mathcal{X} = [-1, 1]^d$, TS and TTPS can act directly by sampling from the learned posterior over the target.

E.1.2 ϵ-best-arm identification on the sphere.

For the continuous ϵ-best-arm problem, the task parameter is a direction $x_\theta^\star = \theta \in \mathbb{S}^{d-1}$ sampled by normalizing a standard Gaussian vector. The agent queries a vector a and observes

$$Y_t = \theta^\top A_t + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \sigma^2). \tag{53}$$

The recommendation is correct if $\theta^\top \hat{x} \geq 1 - \epsilon$. Although the implementation stores the enclosing action bounds as $[-1, 1]^d$, the inference mean, posterior samples, TTPS candidates, and uniform baseline actions are projected to $\mathbb{S}^{d-1}$ for this benchmark. We project onto the sphere so that vectors with the same direction are the same hypothesis. The reason for this benchmark is that each observation is a scalar projection. The agent must choose directions that disambiguate the posterior over θ, while the stopping rule must reason in terms of cosine error rather than Euclidean error.

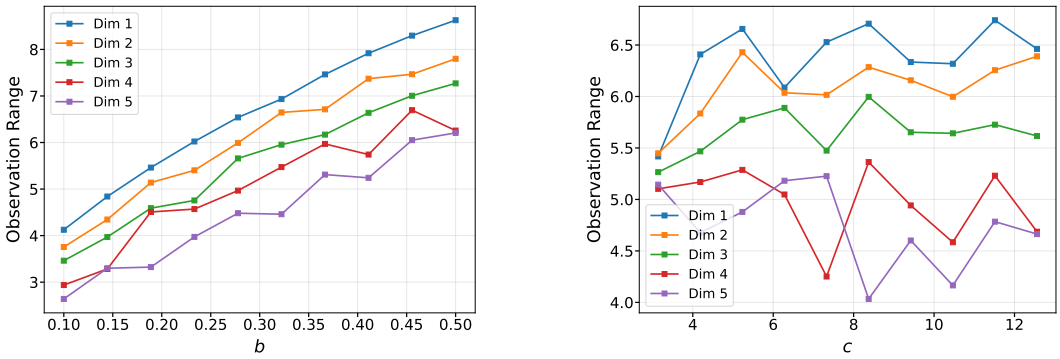
This experiment also lets us compare to a specialized frequentist fixed-confidence baseline. Lazy Track-and-Stop uses the known linear observation structure and an analytic generalized-likelihood-ratio stopping rule. In our implementation it queries canonical directions and keeps a least-squares estimate of θ, stopping only after both the likelihood-ratio condition and the spectral coverage condition are satisfied. This is not a general baseline for all our tasks, but it is a useful reference point on the one benchmark where a specialized fixed-confidence method is available. Round-robin sampling of an orthonormal basis is order-optimal in this problem (Jedra and Proutiere, 2020), so it tests how well C−ICPE learns inference and stopping.

E.1.3 Ackley minimizer identification.

The Ackley task is a shifted and randomly parametrized global-optimization problem. The task parameter is

$$\theta = (a, b, c, \theta^\star), \qquad \theta^\star \sim \mathrm{Unif}([-1, 1]^d),$$

where $\theta^\star$ is the global minimizer and (a, b, c) control the shape of the response surface. In the reported runs we fix $a = 10$ and sample $b \sim \mathrm{Unif}[0.1, 0.5]$ and $c \sim \mathrm{Unif}[\pi, 4\pi]$. Given $u = A_t - \theta^\star$,



(a) (b)

Figure 8: Effect of Ackley function's parameters on output range across multiple dimensions: (a) range vs b; (b) range vs c.

the unnormalized Ackley value is

$$F_{a,b,c}(u) = a + e - a \exp\left(-b\sqrt{\frac{1}{d}\sum_{j=1}^{d} u_j^2}\right) - \exp\left(\frac{1}{d}\sum_{j=1}^{d}\cos(cu_j)\right). \tag{54}$$

We observe the sign-inverted and normalized value (more on this in the next page)

$$Y_t = 1 - 2\frac{F_{a,b,c}(A_t - \theta^\star)}{Z_{\text{norm}}(b,c,d)} + \xi_t, \qquad \xi_t \sim \mathcal{N}(0,\sigma^2), \tag{55}$$

so that larger observations are better and the target remains the minimizer $\theta^\star$. The nuisance parameters are not provided to the agent. Thus, across episodes, `C-ICPE` must infer not only where the optimum is but also how observations should be interpreted for that episode.

Ackley is included because it is deliberately hostile to naive local search. The function has many oscillations near the optimum and a broad outer region where observations can be weakly informative. The active policy therefore has to balance broad exploration with local refinement, and the critic has to stop based on whether the inferred minimizer is accurate, not based on whether the last observed function value was high.

Ackley function output normalization. The Ackley function's global minimum is always at the origin with a value of 0, but the maximum value within our defined recommendation space $\mathcal{X}$ depends on the function parameters and dimensionality. Figs. 8a and 8b show how the values of b and c affect the function output ranges. Larger b values consistently increase the range, while c has less significant effect on the output ranges. From the figures, we also see that the ranges depend on dimensionality, where lower dimensions tend to have larger ranges. The issue with varying output ranges is that the influence of noise can vary across different priors sampled, and we want the noise effect to be on the same scale. Additionally, without normalization, `C-ICPE` must not only learn the relative patterns from H_t but also account for the scale differences across different H_t. For these reasons, we derive a normalization constant empirically from multiple samples across different b, c values and dimensionalities: $Z_{\text{norm}}(b,c,d) = \pi - 0.21 \cdot d + 9.68 \cdot b + 0.04 \cdot c$.

E.1.4 GP MAX-VALUE ESTIMATION

The GP benchmark separates the action and recommendation spaces. At the beginning of an episode we sample a latent function

$$f \sim \text{GP}(0, k_{\text{RBF}}(\ell, \sigma_f)), \qquad \ell \sim \text{Unif}[0.05, 0.2], \qquad \sigma_f = 1,$$

on $[0,1]^d$, with $d = 1$ in the implementation. The sample path is generated on a dense grid using circulant embedding. Queries are continuous points $A_t \in [0,1]^d$; the observed value is obtained by linear interpolation in one dimension or bilinear interpolation in two dimensions, followed by Gaussian noise:

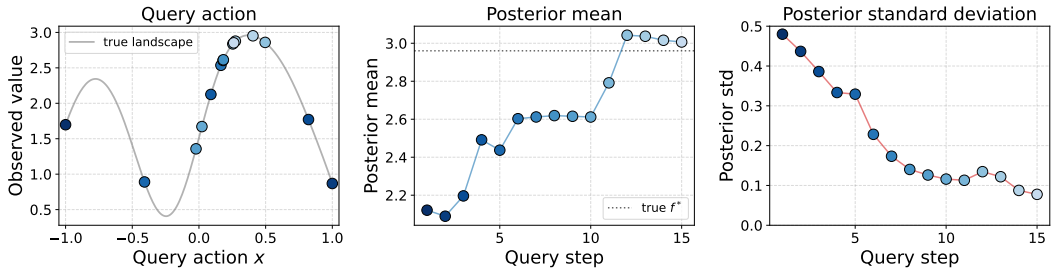


Figure 9: Visualization of how C−ICPE explores in the max-value estimation problem. From left to right: the query action (left), posterior mean (middle), and posterior standard deviation (right) along an exploration trajectory in the GP value estimation problem. Darker samples are queried earlier in the trajectory, while lighter samples are queried later.

$$Y_t = f(A_t) + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \sigma^2). \tag{56}$$

The target is the scalar maximum value

$$v_\theta^\star = \max_{u \in [0,1]^d} f(u), \qquad \mathcal{X} \subseteq \mathbb{R}, \qquad \mathcal{A} = [0,1]^d,$$

where the maximum is computed on the same grid used to generate the episode. The final recommendation is the inference mean for this scalar value, and success is $|\hat{x} - v_\theta^\star| \le \epsilon$.

The reason for using max-value estimation rather than another argmax-localization task is that it tests the decoupling of inference and exploration. In binary search and Ackley, a posterior sample of $x_\theta^\star$ is itself a reasonable query, so TS and TTPS can be implemented directly from the inference distribution. For GP value estimation this would be meaningless: a sample from the inference model is a scalar value, not a point in $[0,1]^d$. Therefore the action must be learned through the critic. We use the TD3 actor for this benchmark because the critic can assign value to a query according to how much it is expected to improve the future estimate of $v_\theta^\star$, even though the query is not itself a recommendation. This experiment is consequently the main empirical check that C−ICPE handles the general $\mathcal{A} \neq \mathcal{X}$.

E.1.5 Multi-Lobe Localization

The multi-lobe task is a multimodal localization problem in which the exact posterior and posterior-success function can be computed in closed form. This allows us to separate the intrinsic effect of the Gaussian projection from neural approximation error, and to test whether multimodality primarily increases the number of queries required to resolve the correct posterior mode. The task parameter is

$$\theta = mS + U, \qquad S \sim \mathrm{Unif}(\{-1,1\}^d), \qquad U \sim \mathrm{Unif}([-w,w]^d), \tag{57}$$

so θ sits at one of 2^d equally likely vertices mS of a d-dimensional hypercube of half-width m, with small jitter w. In the reported experiments, we fix $m = 0.6$, $w = 0.02$. Given an arm $A_t \in [-1,1]^d$, the agent receives a Bernoulli observation

$$Y_t \mid S = s, A_t = a \sim \mathrm{Bernoulli}\big(g_{s,\ell}(a)\big), \qquad g_{s,\ell}(a) := \eta + (1-2\eta)\,\frac{1}{d}\sum_{j=1}^{d} \exp\left(-\frac{(a_j - s_j m)^2}{2\ell^2}\right), \tag{58}$$

where $\eta = 0.2$, and $\ell \in \{0.15, 0.3\}$ which controls the localization of the informative query regions. An action returns observation 1 with probability $1-\eta$ exactly at sm, decaying toward a floor of η as action moves away from sm.

We score success under the L_∞ distance between the recommended location and θ, with $\varepsilon = 0.08$. Since $w < \epsilon < m - w$, success depends only on identifying the correct lobe. Multi-lobe is included because its true posterior over θ is exactly multimodal by construction. Without any query, all 2^d lobes are equally likely, so the Bayes-optimal belief is a mixture over 2^d well-separated modes. A diagonal single-Gaussian belief can only track a unimodal summary of this mixture, where the mean could sits between lobes. However, a mixture-of-Gaussians can represent each candidate lobe explicitly and collapse mass onto the correct one as evidence accumulates.

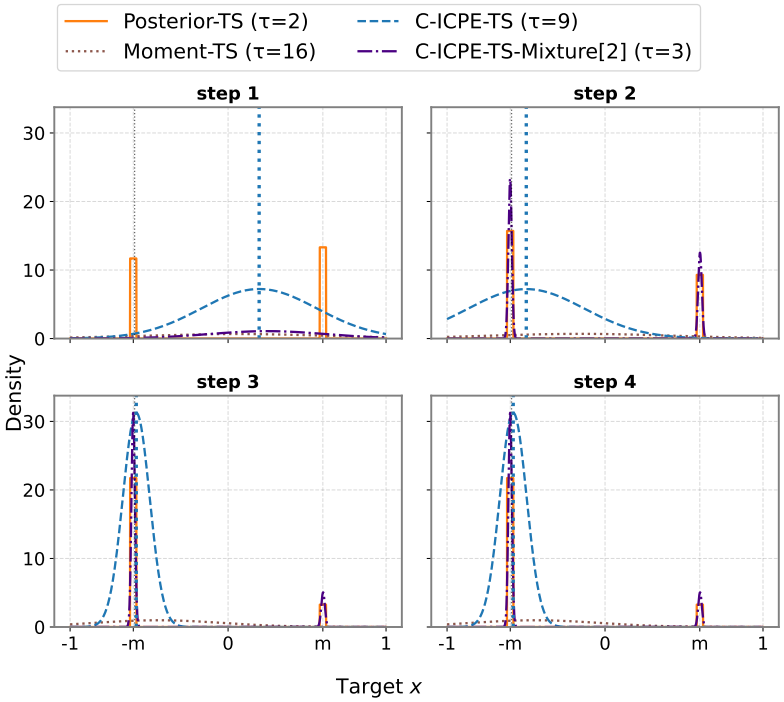


Figure 10: Evolution of the belief over the target θ in the multi-lobe localization problem ($d = 1$, $\ell = 0.15$) for a single episode, with the true target at $-m$ (black dotted line). The legend gives each method's stopping time τ; after stopping, a method's belief is held fixed. The vertical dotted blue line is the `C-ICPE-TS`'s posterior mean.

E.2 Synthetic Benchmarks: Baselines

The most important baseline is `C-ICPE-Uniform`, which keeps the same inference network, critic, stopping rule, replay buffer, and fixed-confidence cost update as `C-ICPE`, but replaces the learned active query rule by uniform exploration. This isolates the value of active experimentation: if `C-ICPE` improves over `C-ICPE-Uniform`, the gain cannot be explained by the inference model alone, because both methods use the same form of inference and the same stopping mechanism.

For the ε-best-arm identification problem, to mirror `LT&S round-robin`'s query strategy, we implement another baseline, `C-ICPE-RoundRobin`, in which we replace the actor with a round-robin policy that queries the canonical basis vectors in turn. This isolates the effect of the query policy and lets us study whether the critic can reliably decide when to stop and whether inference can recover $x^\star_\theta$.

For tasks with $\mathcal{A} = \mathcal{X}$, we evaluate TS and TTPS because they use the learned posterior in the most direct way. TS samples a plausible target and queries it. TTPS keeps the posterior mean as the current recommendation and intentionally samples a plausible challenger that is sufficiently different. This is useful when many posterior samples are small perturbations around the current mean: such samples do not test the remaining uncertainty, whereas a challenger query can reveal whether another region is still plausible. For GP value estimation, TS and TTPS are not the right action rules for the reason described above, so we use TD3.

For all benchmarks, we set the sample budget to $t_{\max} = 100$ by default. If the trained `C-ICPE` policy failed to reach the target $(1 - \delta)$-accuracy on rollouts, we extended the sample budget by 50. For easier setting, such as low dimension, we instead used a smaller budget. The settings in which $t_{max} \neq 100$ are listed in Table 1 , all other configurations use $t_{\max} = 100$.

We also compare against standard fixed-budget optimization methods implemented through Optuna (Akiba et al., 2019): `TPE`, `GP-logEI`, and `CMA-ES`. We also compare with `GP-UCB` via BoTorch (Balandat et al., 2020). These baselines do not have a learned stopping rule and are not optimized

Environment	Method	d	ε/ℓ	$t_{\max}$
ε-Best-Arm	all	15	0.005	150
GP-value estimation	all	1	0.2	60
Geochemical	all	2	0.15	150
Multi-Lobe	`C-ICPE-TS`, `Moment-TS`	2	0.15	300

Table 1: Sample budget $t_{\max}$ for settings where $t_{\max} \neq 100$.

for (ϵ, δ)-correctness. To make the comparison conservative, each baseline receives a fixed budget equal to the mean stopping time of the most sample-efficient `C-ICPE` variant in each configuration, rounded up. They are then evaluated under the same success criterion $L_\theta(\hat{x}) \leq \epsilon$. This comparison asks whether a generic optimizer, with a fixed budget equal to `C-ICPE`'s expected sample complexity, already reaches the fixed-confidence target.

- **TPE** (Bergstra et al., 2011): splits observations into good and bad groups based on a quantile threshold and models the objective by building a density estimator for each group, then selects candidates that maximize the ratio of good-to-bad.
- **CMA-ES** (Hansen, 2016): an evolutionary algorithm that samples a population of candidates and iteratively updates a multivariate Gaussian distribution by adapting its covariance matrix based on the successful candidates.
- **GP-logEI** (Ament et al., 2023): updates the Matérn kernel's hyperparameters by maximizing the marginal log-likelihood on the past observations, and uses log expected improvement as the acquisition function.
- **GP-UCB** (Srinivas et al., 2010): a variant of GP-based Bayesian Optimization with a Matérn kernel that uses upper confidence bound as the acquisition function.
- **Uniform bin**: partitions the query space into $\left\lceil \sqrt{\mathbb{E}[\tau]} \right\rceil$ bins, where $\mathbb{E}[\tau]$ is the corresponding `C-ICPE` variant's expected sample complexity, and uniformly sample within each bin. We compute the average value of each bin and report the maximum.
- **Uniform top 5%**: queries uniformly and return the mean of the top 5% values. The number of query matches the corresponding `C-ICPE` variant's expected sample complexity.
- **Lazy Track-and-Stop round robin** (Jedra and Proutiere, 2020): queries the canonical basis in a round-robin fashion. The method maintains a least square estimate $\hat{\theta}_t$, and stops whenever

$$Z_t \;\geq\; \beta(\delta, t) \quad \text{and} \quad \min_j N_t(j) \;\geq\; \max\Big(c,\; \tfrac{\rho(\delta,t)}{\|\hat{\theta}_t\|^2}\Big),$$

  where Z_t is the generalized likelihood ratio (GLR) for the ε_t best-arm hypothesis evaluated against the worst-case competitor on the ε_t-boundary. This measures how much the current guess is better than the closest alternative. $N_t(j)$ is the number of pulls of the j-th element of the canonical basis, and $\beta(\delta, t)$, $\rho(\delta, t)$, ε_t are the threshold rule, spectral threshold, and gap-relaxation threshold. Lastly c is a constant defined in (Jedra and Proutiere, 2020). See also (Jedra and Proutiere, 2020) for more details. Upon stopping, the agent recommends $\hat{a}_t = \hat{\theta}_t / \|\hat{\theta}_t\|$.
- **Lazy Track-and-Stop uniform** (Jedra and Proutiere, 2020): Same as Lazy Track-and-Stop round robin except it samples the basis uniformly.
- **Posterior-TS**: maintains the true posterior over lobe identity, $\pi_t(s) := P(S = s \mid H_t)$ for $s \in \{-1, +1\}^d$, with uniform prior $\pi_1(s) = 2^{-d}$, updated after each Bernoulli observation (A_t, Y_t) by Bayes' rule,

$$\pi_{t+1}(s) \;\propto\; \pi_t(s)\,\big[Y_t\, g_s(A_t) + (1 - Y_t)\,(1 - g_s(A_t))\big], \qquad s \in \{-1, +1\}^d.$$

  At each step it samples $\tilde{\theta}_t \sim R_t(\cdot \mid H_t)$ from this posterior and queries $A_t = \tilde{\theta}_t$, stopping at

$$\tau_{\text{post}} := \inf\{t : u_t \geq q^\star\}, \qquad u_t := \max_s \pi_t(s),$$

where u_t is the posterior mass on the currently most likely lobe and $q^\star = 1 - \delta$. Upon stopping, it recommends that lobe's center, $\hat{a}_t = m\,\hat{s}_t$, $\hat{s}_t := \arg\max_s \pi_t(s)$. Since it acts directly on the exact u_t with no approximation, Posterior-TS upper-bounds what any method restricted to this query process can achieve.

- **Moment-TS**: uses the identical posterior-sampling query rule and belief p_t as Posterior-TS, but collapses the belief to its Gaussian moment projection before acting,

$$\mu_t = m(2p_t - 1), \qquad \sigma_t^2 = \frac{w^2}{3} + 4m^2 p_t(1 - p_t),$$

querying $A_t = \mathrm{Proj}_{[-1,1]}(Z_t)$ with $Z_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$. Because a point estimate can only be judged by its own ε-correctness probability, its stopping rule replaces u_t with the success probability of recommending the mean,

$$\tau_G := \inf\{t : \Gamma_\varepsilon(u_t) \geq q^\star\}, \qquad \Gamma_\varepsilon(u) = \begin{cases} 0 & u \leq b_0 \\ \dfrac{m}{w}\,u\,(u - b_0) & b_0 < u < b_1 \\ u & u \geq b_1, \end{cases}$$

with $b_0 = 1 - \frac{\varepsilon + w}{2m}$ and $b_1 = 1 - \frac{\varepsilon - w}{2m}$. Upon stopping, it recommends the mean, $\hat{a}_t = \mu_t$. Since $\Gamma_\varepsilon(u) \leq u$ everywhere and $\Gamma_\varepsilon(u) = 0$ for $u \leq b_0$, this stopping rule is provably stricter than Posterior-TS's. The mean cannot be trusted as a recommendation until the belief has moved onto a lobe far enough to clear the geometric floor b_0, independent of $q^\star$. Therefore, Moment-TS isolates the intrinsic sample-complexity cost of mean-based recommendation under a multimodal target, separate from any neural approximation error.

E.3 Synthetic Benchmarks: Numerical Results

We now present detailed accuracy and sample complexity results across all benchmarks, sweeping over (ε, σ, d) configurations. Tables report mean accuracy and sample complexity with 95% confidence intervals for every method and parameter combination. To complement these aggregate statistics, we examine the stopping behavior of each actor through two diagnostics: (i) the survival function $P(\tau > t)$, which reveals how quickly the learned critic commits to stopping, and (ii) the standard deviation of the inference model over the horizon, which tracks how rapidly the posterior uncertainty around $x^\star_\theta$ contracts. Across all benchmarks, `C-ICPE` with learned exploration (TS, TTPS, or TD3) consistently meets the $1-\delta$ accuracy target while `C-ICPE-Uniform` degrades as dimension increases, particularly on Ackley and binary search where directed exploration is essential.

E.3.1 Noisy Binary Search

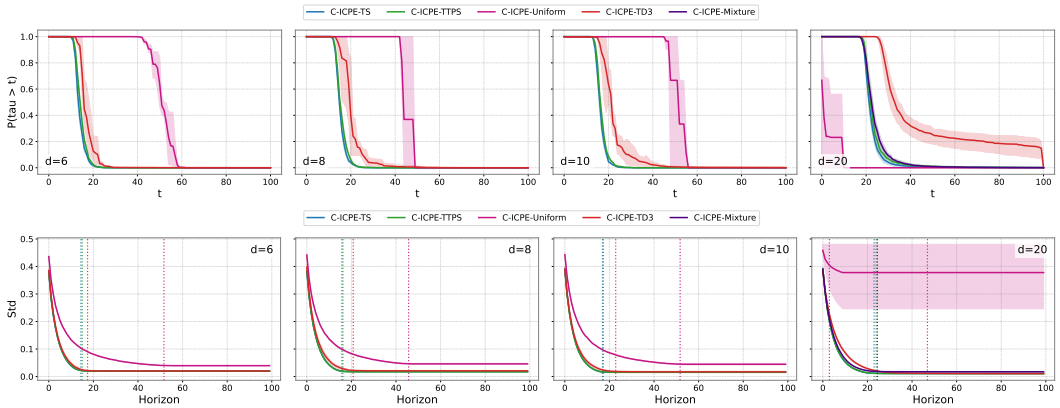


Figure 11: Results for Binary Search Problem with fixed confidence $\delta = 0.1$ and $N = 100$ across different dimensions at the hardest (ε, σ) setting: (top) survival function of τ; (bottom) inference uncertainty convergence.

Tables 3 and 4 report accuracy and sample complexity across all (d, ε, σ) configurations; Fig. 11 shows the survival function of τ and the convergence of the inference model's standard deviation. For simplicity, we tested `C-ICPE-TS-Mixture` on the highest dimension ($d = 20$). We also evaluate `C-ICPE-TS+CF`, a critic-free ablation of the stopping rule: it uses the same trained inference network and actor, but stops once the inference network places mass at least $1 - \delta_{\text{train}}$ on the ε-ball around its recommendation $\hat{x}$, instead of using the trained critic. This isolates the contribution of the critic to stopping. Finally, Fig. 7 illustrates how `C-ICPE` explores in $d = 6$, showing the queries chosen by the actor and the evolution of the posterior mean and standard deviation over time. We also report the cost multiplier at the end of training for every (d, ε, σ) configuration in Table 2.

Accuracy. All three active actors (TS, TTPS, TD3) meet the $1 - \delta = 0.90$ target across every configuration tested, with mean accuracy between 0.909 and 0.951. `C-ICPE-TS-Mixture` also meets the 0.90 target across all (ε, σ). The confidence intervals, and certifications, confirm that the target is met reliably. `C-ICPE-Uniform` matches the active methods at $d = 6$ with $\varepsilon = 0.2$ (accuracy ≥ 0.916) but degrades sharply as either d increases or ε decreases. At $(\varepsilon, \sigma) = (0.1, 0.05)$, accuracy drops from 0.594 at $d = 6$ to 0.276 at $d = 8$, 0.176 at $d = 10$, and 0.000 at $d = 20$. This confirms that passive exploration cannot accumulate sufficient directional information per coordinate to localize the target within the allowed horizon in high dimensions.

Sample complexity. Among active methods, `C-ICPE-TS`, `C-ICPE-TTPS`, `C-ICPE-TS-Mixture` achieve comparable sample complexity across all settings: at $d = 20$, $(\varepsilon, \sigma) = (0.1, 0.05)$, `C-ICPE-TS` stops in 23.0 queries on average, `C-ICPE-TTPS` in 23.9, and `C-ICPE-TS-Mixture` in 24.5, all three with tight confidence intervals. `C-ICPE-TD3` is competitive at moderate dimensions ($d \leq 10$) or larger ε (0.2) but exhibits higher mean stopping times and substantially wider confidence intervals at $d = 20$ (e.g., 46.9 at $(\varepsilon, \sigma) = (0.1, 0.05)$),

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
6	C-ICPE-TD3	0.0354 [0.0282,0.0427]	0.0025 [0.0024,0.0026]	0.0329 [0.0317,0.0341]	0.0198 [0.0150,0.0246]
	C-ICPE-TS	0.0323 [0.0295,0.0352]	0.0030 [0.0014,0.0046]	0.0633 [0.0563,0.0704]	0.0335 [0.0313,0.0356]
	C-ICPE-TTPS	0.0225 [0.0191,0.0260]	0.0019 [0.0007,0.0031]	0.0492 [0.0441,0.0543]	0.0274 [0.0195,0.0353]
	C-ICPE-uniform	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]
8	C-ICPE-TD3	0.0119 [0.0110,0.0128]	0.0026 [0.0020,0.0032]	0.0187 [0.0147,0.0228]	0.0142 [0.0131,0.0154]
	C-ICPE-TS	0.0297 [0.0185,0.0410]	0.0022 [0.0020,0.0023]	0.0600 [0.0558,0.0641]	0.0213 [0.0153,0.0273]
	C-ICPE-TTPS	0.0188 [0.0125,0.0252]	0.0009 [0.0008,0.0010]	0.0528 [0.0523,0.0534]	0.0134 [0.0053,0.0215]
	C-ICPE-uniform	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]
10	C-ICPE-TD3	0.0090 [0.0068,0.0112]	0.0055 [0.0035,0.0076]	0.0159 [0.0137,0.0181]	0.0161 [0.0141,0.0181]
	C-ICPE-TS	0.0314 [0.0216,0.0411]	0.0026 [0.0004,0.0048]	0.0445 [0.0359,0.0530]	0.0082 [0.0048,0.0117]
	C-ICPE-TTPS	0.0250 [0.0103,0.0396]	0.0009 [0.0006,0.0011]	0.0348 [0.0311,0.0385]	0.0069 [0.0040,0.0098]
	C-ICPE-uniform	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]
20	C-ICPE-TD3	0.0283 [0.0243,0.0323]	0.0000 [0.0000,0.0001]	0.0000 [0.0000,0.0000]	0.0008 [-0.0008,0.0023]
	C-ICPE-TS	0.0147 [0.0053,0.0242]	0.0001 [0.0000,0.0003]	0.0278 [0.0247,0.0309]	0.0088 [0.0081,0.0095]
	C-ICPE-TTPS	0.0091 [0.0054,0.0127]	0.0001 [0.0000,0.0002]	0.0197 [0.0127,0.0266]	0.0138 [0.0074,0.0201]
	C-ICPE-uniform	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]
	C-ICPE-TS-Mixture	0.0014 [0.0008,0.0020]	0.0004 [-0.0002,0.0009]	0.0022 [0.0013,0.0031]	0.0016 [-0.0005,0.0038]

Table 2: Binary search: final cost multiplier (mean and 95% CI across seeds) for every (d, ε, σ) configuration.

suggesting that the learned actor is less stable in more diffuclt settings. Sample complexity scales sublinearly in d for the active actors: `C-ICPE-TS` increases from 10.9 ($d = 6$) to 16.1 ($d = 20$) at $(\varepsilon, \sigma) = (0.2, 0.05)$.

Stopping behavior. The survival functions (Fig. 11, top) corroborate the sample complexity results. Across all dimensions, `C-ICPE-TS` and `C-ICPE-TTPS` exhibit sharp transitions: $P(\tau > t)$ drops from 1 to 0 within a narrow window, indicating that the critic identifies a consistent stopping point. The inference standard deviation (bottom row) confirms that the active actors' posteriors contract rapidly, reaching near-zero uncertainty before the mean stopping time (dashed vertical lines), while `C-ICPE-Uniform` at $d = 20$ retains high residual uncertainty throughout the episode.

Certification. Table 5 reports, for every configuration, a one-sided exact binomial test of $H_0 : p \leq 0.9$ against $H_1 : p > 0.9$ ($1 - \delta = 0.9$ is the certification target) at $n = 20000$ independent trials, $\alpha = 0.01$: rejecting H_0 means the method's true success probability is statistically confirmed to exceed the target. All three active methods (`C-ICPE-TD3`, `C-ICPE-TS`, `C-ICPE-TTPS`) certify at every (d, ε, σ) configuration tested, with p-values never exceeding 10^{-5} and frequently below 10^{-15}. `C-ICPE-TS-Mixture` also shows a low p-values below 10^{-9}. `C-ICPE-Uniform`'s certifiable region shrinks steadily with dimension: at $d = 6$ it fails only at the harder $\varepsilon = 0.1$ settings ($p = 1.0$); by $d = 8$ and $d = 10$ it certifies only at the single easiest configuration ($\varepsilon = 0.2, \sigma = 0.05$); by $d = 20$ it fails everywhere ($p = 1.0$ at all four configurations). This shows passive exploration method's certifiable region contracts with problem difficulty until it vanishes entirely, while directed exploration continues to certify by wide margins regardless.

Critic-free stopping. `C-ICPE-TS+CF` matches or improves on `C-ICPE-TS` in every configuration. It attains higher accuracy (0.923–0.945 vs. 0.909–0.921) with similar sample complexity, and it certifies in every configuration, with every seed rejecting H_0 at $p \leq 6.2 \times 10^{-13}$ (Table 5). Its p-values are far smaller than those of `C-ICPE-TS` (e.g., 1.1×10^{-93} vs. 8.4×10^{-10} at $d = 6$, $(\varepsilon, \sigma) = (0.1, 0.025)$), which reflects its larger margin above the target. In binary search, the posterior is unimodal and the inference network is well calibrated, so its own belief already provides a reliable stopping signal.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
6	C-ICPE-TD3	0.914 [.906,.922]	0.915 [.908,.923]	0.927 [.918,.935]	0.932 [.909,.956]
	C-ICPE-TS	0.915 [.906,.922]	0.918 [.911,.927]	0.913 [.905,.919]	0.920 [.912,.925]
	C-ICPE-TS+CF	**0.935** [.929,.941]	**0.923** [.915,.931]	**0.942** [.940,.945]	**0.934** [.928,.940]
	C-ICPE-TTPS	0.916 [.909,.923]	0.919 [.911,.925]	0.920 [.911,.929]	0.917 [.910,.924]
	C-ICPE-uniform	0.918 [.910,.926]	0.916 [.906,.924]	0.717 [.703,.731]	0.594 [.571,.618]
8	C-ICPE-TD3	0.910 [.899,.920]	0.917 [.907,.927]	**0.946** [.922,.963]	**0.941** [.917,.965]
	C-ICPE-TS	0.914 [.904,.922]	0.914 [.907,.923]	0.919 [.909,.927]	0.917 [.910,.925]
	C-ICPE-TS+CF	**0.931** [.925,.938]	**0.924** [.918,.931]	0.945 [.941,.949]	0.940 [.936,.944]
	C-ICPE-TTPS	0.919 [.912,.925]	0.915 [.908,.922]	0.915 [.906,.923]	0.914 [.907,.921]
	C-ICPE-uniform	0.916 [.909,.922]	0.773 [.760,.786]	0.424 [.393,.452]	0.276 [.249,.306]
10	C-ICPE-TD3	0.915 [.908,.923]	0.911 [.905,.918]	**0.951** [.942,.957]	0.922 [.912,.930]
	C-ICPE-TS	0.911 [.903,.917]	0.909 [.901,.916]	0.913 [.905,.922]	0.919 [.910,.927]
	C-ICPE-TS+CF	**0.938** [.935,.941]	**0.926** [.918,.935]	0.942 [.942,.943]	**0.934** [.931,.938]
	C-ICPE-TTPS	0.919 [.911,.925]	0.919 [.911,.927]	0.916 [.909,.922]	0.918 [.909,.926]
	C-ICPE-uniform	0.920 [.911,.928]	0.723 [.710,.735]	0.355 [.328,.385]	0.176 [.136,.220]
20	C-ICPE-TD3	0.919 [.910,.927]	0.919 [.909,.930]	0.933 [.917,.947]	0.925 [.917,.931]
	C-ICPE-TS	0.917 [.907,.926]	0.915 [.908,.922]	0.915 [.907,.921]	0.921 [.911,.930]
	C-ICPE-TS+CF	**0.926** [.919,.933]	**0.923** [.919,.927]	**0.937** [.931,.943]	**0.931** [.925,.937]
	C-ICPE-TTPS	0.914 [.904,.923]	0.917 [.909,.923]	0.917 [.908,.925]	0.920 [.911,.929]
	C-ICPE-uniform	0.800 [.786,.814]	0.286 [.165,.383]	0.000 [.000,.000]	0.000 [.000,.000]
	C-ICPE-TS-Mixture	0.916 [.908,.924]	0.913 [.904,.922]	0.915 [.908,.921]	0.916 [.909,.924]

Table 3: Binary search: accuracy (mean and 95% CI) for every (d, ε, σ) configuration. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
6	C-ICPE-TD3	11.1 [11.0,11.2]	16.3 [16.1,16.4]	14.0 [13.7,14.3]	17.4 [16.4,19.1]
	C-ICPE-TS	10.9 [10.8,10.9]	15.7 [15.6,15.8]	**11.4** [11.4,11.5]	14.4 [14.2,14.6]
	C-ICPE-TS+CF	**10.8** [10.5,11.1]	**15.1** [14.9,15.3]	11.4 [11.3,11.5]	**14.0** [14.0,14.1]
	C-ICPE-TTPS	11.2 [11.2,11.3]	16.2 [16.1,16.4]	11.9 [11.8,12.0]	15.0 [14.9,15.1]
	C-ICPE-uniform	38.6 [38.3,38.9]	55.6 [54.9,56.2]	48.3 [47.5,49.2]	51.6 [50.5,53.2]
8	C-ICPE-TD3	12.1 [11.9,12.2]	18.0 [17.9,18.2]	16.8 [15.5,17.7]	20.8 [19.9,22.2]
	C-ICPE-TS	11.9 [11.9,12.0]	17.5 [17.4,17.6]	12.5 [12.3,12.6]	15.7 [15.7,15.8]
	C-ICPE-TS+CF	**11.8** [11.6,11.9]	**16.9** [16.7,17.0]	**12.4** [12.3,12.5]	**15.6** [15.5,15.7]
	C-ICPE-TTPS	12.5 [12.4,12.6]	18.1 [18.0,18.2]	13.0 [13.0,13.1]	16.2 [16.1,16.3]
	C-ICPE-uniform	45.1 [44.8,45.3]	50.0 [49.5,50.5]	44.3 [43.9,44.9]	45.7 [43.9,47.4]
10	C-ICPE-TD3	13.1 [13.1,13.2]	19.5 [19.3,19.7]	18.4 [17.6,19.1]	22.8 [22.1,23.5]
	C-ICPE-TS	12.9 [12.8,13.0]	18.9 [18.8,19.0]	13.4 [13.3,13.5]	16.9 [16.6,17.1]
	C-ICPE-TS+CF	**12.9** [12.8,13.0]	**18.2** [18.0,18.4]	**13.4** [13.3,13.5]	**16.5** [16.4,16.7]
	C-ICPE-TTPS	13.5 [13.4,13.7]	19.7 [19.6,19.8]	14.0 [13.9,14.0]	17.3 [17.2,17.4]
	C-ICPE-uniform	52.0 [51.7,52.3]	53.9 [53.3,54.8]	50.8 [50.0,52.2]	51.7 [47.7,55.3]
20	C-ICPE-TD3	19.5 [17.9,22.2]	27.2 [26.9,27.6]	39.1 [38.3,40.0]	46.9 [42.9,50.8]
	C-ICPE-TS	**16.1** [16.0,16.2]	23.4 [23.3,23.6]	17.5 [17.2,17.9]	23.0 [22.4,23.7]
	C-ICPE-TS+CF	16.1 [15.9,16.3]	**23.1** [22.8,23.4]	17.8 [17.3,18.2]	22.0 [21.6,22.4]
	C-ICPE-TTPS	16.5 [16.4,16.7]	24.3 [24.0,24.5]	18.4 [17.9,18.8]	23.9 [22.8,25.4]
	C-ICPE-uniform	71.9 [70.9,73.0]	68.5 [55.7,78.0]	**5.3** [0.0,10.7]	**2.9** [0.0,7.2]
	C-ICPE-TS-Mixture	18.4 [18.3,18.5]	26.8 [26.4,27.2]	19.8 [19.6,20.1]	24.5 [24.0,25.0]

Table 4: Binary search: sample complexity (mean and 95% CI) for every (d, ε, σ) configuration. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
6	C-ICPE-TD3	4.9×10^{-16} [5.5×10^{-22},1.5×10^{-15}]	5.4×10^{-10} [8.4×10^{-14},1.6×10^{-9}]	7.0×10^{-14} [8.1×10^{-38},2.1×10^{-13}]	2.4×10^{-12} [1.4×10^{-281},6.0×10^{-12}]
	C-ICPE-TS	1.9×10^{-7} [1.2×10^{-20},5.6×10^{-7}]	4.3×10^{-13} [5.9×10^{-15},1.3×10^{-12}]	8.4×10^{-10} [6.2×10^{-13},2.5×10^{-9}]	6.5×10^{-12} [2.5×10^{-20},2.0×10^{-11}]
	C-ICPE-TS+CF	$\mathbf{1.7 \times 10^{-40}}$ [3.1×10^{-88},5.2×10^{-40}]	$\mathbf{2.1 \times 10^{-13}}$ [8.2×10^{-37},6.2×10^{-13}]	$\mathbf{1.1 \times 10^{-93}}$ [1.1×10^{-132},2.7×10^{-93}]	$\mathbf{6.1 \times 10^{-65}}$ [1.2×10^{-85},1.8×10^{-64}]
	C-ICPE-TTPS	2.1×10^{-8} [1.2×10^{-13},6.3×10^{-8}]	8.4×10^{-10} [2.4×10^{-19},2.5×10^{-9}]	9.1×10^{-12} [2.8×10^{-18},2.7×10^{-11}]	8.4×10^{-12} [1.5×10^{-14},1.7×10^{-11}]
	C-ICPE-uniform	3.9×10^{-11} [7.1×10^{-15},1.2×10^{-10}]	6.9×10^{-8} [1.7×10^{-23},2.1×10^{-7}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]
8	C-ICPE-TD3	7.1×10^{-6} [1.8×10^{-18},2.1×10^{-5}]	4.4×10^{-17} [7.8×10^{-25},8.5×10^{-17}]	1.0×10^{-36} [6.3×10^{-247},3.0×10^{-36}]	6.1×10^{-22} [3.6×10^{-269},1.8×10^{-21}]
	C-ICPE-TS	1.2×10^{-8} [4.9×10^{-20},3.7×10^{-8}]	6.3×10^{-11} [5.6×10^{-17},1.9×10^{-10}]	9.1×10^{-15} [3.4×10^{-22},2.7×10^{-14}]	1.3×10^{-15} [1.9×10^{-16},2.2×10^{-15}]
	C-ICPE-TS+CF	$\mathbf{1.5 \times 10^{-55}}$ [4.5×10^{-69},4.4×10^{-55}]	$\mathbf{1.5 \times 10^{-25}}$ [4.9×10^{-32},4.6×10^{-25}]	$\mathbf{1.1 \times 10^{-105}}$ [5.8×10^{-130},3.2×10^{-105}]	$\mathbf{1.1 \times 10^{-89}}$ [2.1×10^{-108},3.4×10^{-89}]
	C-ICPE-TTPS	1.0×10^{-7} [2.2×10^{-24},3.0×10^{-7}]	8.8×10^{-16} [2.8×10^{-18},2.2×10^{-15}]	2.3×10^{-14} [1.1×10^{-37},7.0×10^{-14}]	5.4×10^{-9} [1.8×10^{-21},1.6×10^{-8}]
	C-ICPE-uniform	2.1×10^{-11} [3.6×10^{-13},6.2×10^{-11}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]
10	C-ICPE-TD3	1.1×10^{-8} [8.6×10^{-15},3.2×10^{-8}]	7.2×10^{-13} [2.4×10^{-19},2.1×10^{-12}]	$\mathbf{1.4 \times 10^{-126}}$ [7.8×10^{-190},4.2×10^{-126}]	2.1×10^{-13} [5.0×10^{-48},6.2×10^{-13}]
	C-ICPE-TS	1.5×10^{-9} [7.0×10^{-14},4.5×10^{-9}]	3.0×10^{-10} [7.0×10^{-14},4.8×10^{-10}]	4.7×10^{-12} [1.6×10^{-17},1.4×10^{-11}]	2.5×10^{-10} [3.6×10^{-25},7.6×10^{-10}]
	C-ICPE-TS+CF	$\mathbf{1.3 \times 10^{-62}}$ [3.1×10^{-73},3.9×10^{-62}]	$\mathbf{4.4 \times 10^{-24}}$ [7.8×10^{-39},1.3×10^{-23}]	2.5×10^{-67} [4.0×10^{-114},7.5×10^{-67}]	$\mathbf{3.7 \times 10^{-57}}$ [2.9×10^{-68},1.0×10^{-56}]
	C-ICPE-TTPS	4.8×10^{-11} [6.2×10^{-20},1.4×10^{-10}]	8.1×10^{-17} [2.7×10^{-25},2.4×10^{-16}]	7.1×10^{-18} [9.6×10^{-21},1.6×10^{-17}]	4.2×10^{-15} [9.6×10^{-21},1.3×10^{-14}]
	C-ICPE-uniform	4.4×10^{-17} [1.0×10^{-24},1.3×10^{-16}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]
20	C-ICPE-TD3	4.9×10^{-14} [5.6×10^{-29},1.5×10^{-13}]	1.3×10^{-13} [1.6×10^{-17},3.6×10^{-13}]	7.7×10^{-15} [4.5×10^{-107},1.5×10^{-14}]	1.2×10^{-21} [2.9×10^{-52},3.7×10^{-21}]
	C-ICPE-TS	4.1×10^{-14} [7.6×10^{-21},1.2×10^{-13}]	6.3×10^{-11} [2.2×10^{-15},1.9×10^{-10}]	3.3×10^{-16} [1.6×10^{-37},9.9×10^{-16}]	2.2×10^{-21} [1.9×10^{-34},6.0×10^{-21}]
	C-ICPE-TS+CF	$\mathbf{4.6 \times 10^{-35}}$ [1.5×10^{-38},1.4×10^{-34}]	$\mathbf{5.3 \times 10^{-18}}$ [7.3×10^{-26},1.6×10^{-17}]	$\mathbf{3.2 \times 10^{-43}}$ [1.7×10^{-69},9.6×10^{-43}]	$\mathbf{2.2 \times 10^{-32}}$ [6.8×10^{-70},6.6×10^{-32}]
	C-ICPE-TTPS	2.4×10^{-8} [6.7×10^{-18},7.2×10^{-8}]	4.9×10^{-7} [8.5×10^{-17},1.5×10^{-6}]	3.6×10^{-9} [1.9×10^{-19},1.1×10^{-8}]	5.4×10^{-9} [1.8×10^{-40},1.6×10^{-8}]
	C-ICPE-uniform	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]
	C-ICPE-TS-Mixture	2.7×10^{-9} [4.4×10^{-16},8.0×10^{-9}]	3.7×10^{-12} [2.3×10^{-21},1.0×10^{-11}]	1.0×10^{-10} [7.9×10^{-24},3.0×10^{-10}]	5.1×10^{-12} [1.2×10^{-19},1.4×10^{-11}]

Table 5: Binary search: certification p-value ($H_0 : p \leq 0.9$, one-sided exact binomial test on $n = 20000$ independent trials, significance level $\alpha = 0.01$) for every (d, ε, σ) configuration: mean across seeds, with the [min, max] range across seeds in brackets. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

E.3.2 ϵ-BEST ARM PROBLEM

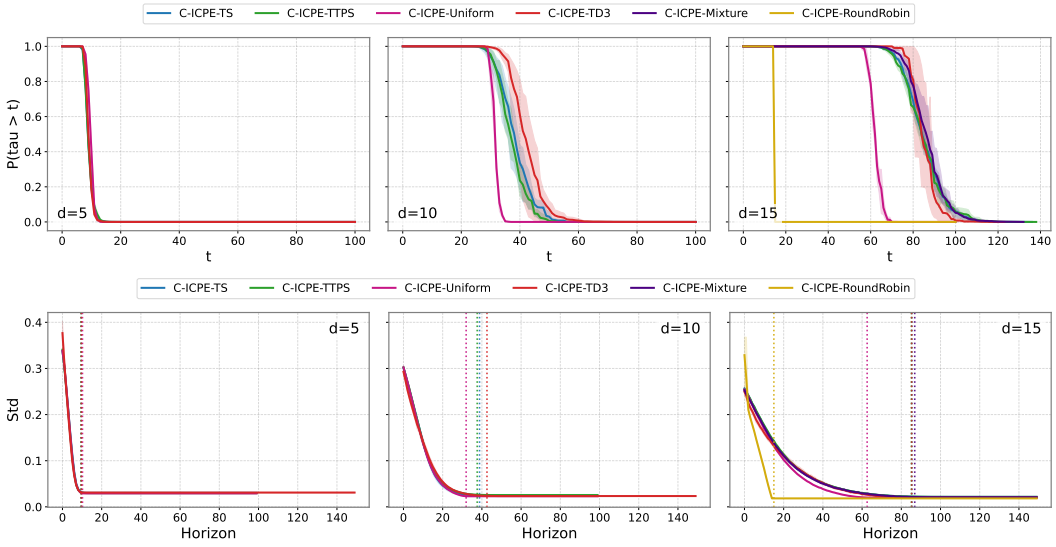


Figure 12: Results for ϵ-Best-Arm Identification Problem with fixed confidence $\delta = 0.1, \epsilon = 0.005$ and $N = 100/150$ across different dimensions at the hardest (ε, σ) setting: (top) survival function of τ; (bottom) inference uncertainty convergence.

Tables 7 and 8 report accuracy and sample complexity, and Fig. 12 shows the survival function and inference uncertainty convergence. For simplicity, we tested `C-ICPE-TS-Mixture` and `C-ICPE-RoundRobin` on the highest dimension ($d = 15$). As in binary search, we also evaluate the critic-free ablation `C-ICPE-TS+CF`. We also report the cost multiplier at the end of training for every (d, ε, σ) configuration in Table 6.

Accuracy. All `C-ICPE` variants meet the $1-\delta$ target across every (d, ε, σ) configuration, with mean accuracy between 0.914 and 0.948. Notably, `C-ICPE-Uniform` also meets the target throughout (accuracy 0.913–0.931), consistent with the rotational symmetry of the problem: since the loss $L_\theta(x) = 1-\theta^\top x$ is invariant to orthogonal transformations, no query direction is intrinsically more informative than another, and isotropic exploration is in general optimal. `Lazy Track-and-Stop` achieves perfect accuracy (1.000) in all configurations by exploiting the known linear observation structure and an analytic likelihood-ratio stopping rule. Both `C-ICPE-TS-Mixture` and `C-ICPE-RoundRobin` meet the 0.9 target across all configs, `C-ICPE-RoundRobin` especially achieve 1.0 accuracy at $(\varepsilon = 0.005)$.

Sample complexity. The gap between `C-ICPE` and `Lazy Track-and-Stop` is substantial and grows with dimension. `Lazy Track-and-Stop` queries exactly d samples in every configuration, one per canonical direction, which is the minimum for a full-rank linear estimator. `C-ICPE` uses 2–6$\times$ more queries: at $d = 15$ and $(\varepsilon, \sigma) = (0.005, 0.0025)$, `C-ICPE-TS` and `C-ICPE-TTPS` stop at 85.4 and 85.1 queries, respectively, compared to 15.0 for `Lazy Track-and-Stop`. The current noise levels are low relative to ε, placing the problem in a regime where `Lazy Track-and-Stop` resolves the target direction in a single pass of d orthogonal queries. At higher noise levels, where multiple measurement rounds are necessary, we expect the relative performance of `C-ICPE` to improve. The ablations clarify where this gap comes from. `C-ICPE-Uniform` is competitive with, and sometimes more sample-efficient than, the active-actor variants: in the same setting, it stops at 62.6 queries. This inversion arises because the problem's symmetry makes directed exploration unnecessary, while posterior-driven action selection occasionally concentrates queries in already well-estimated directions, slowing convergence relative to uniform coverage. Since the posterior is not multimodal here, the Gaussian mixture offers no benefit: `C-ICPE-TS-Mixture` stops at 86.8 queries. In contrast, `C-ICPE-RoundRobin`, which uses the same round-robin query strategy as `LT&S round-robin`, matches `Lazy Track-and-Stop`, with sample complexity equal to d. Notably, at $d = 15$ and $\varepsilon = 0.02$, it stops after 14.8 queries while attaining an accuracy of ≈ 0.96,

above the $1 - \delta = 0.9$ target it was trained for. With the query policy held fixed, this shows that the critic learns precisely when to stop and the inference network provides accurate recommendations.

Stopping behavior. The survival functions (Fig. 12, top) show that all C−ICPE actors have nearly identical stopping profiles at $d = 5$, where episodes terminate within $t \approx 10$. As dimension grows, the curves separate. At $d = 10$, C−ICPE−Uniform stops earliest (≈ 33), followed by C−ICPE−TS and C−ICPE−TTPS (≈ 40), with C−ICPE−TD3 latest and most variable. At $d = 15$, C−ICPE−RoundRobin stops almost deterministically at $t \approx 15$, C−ICPE−Uniform at ≈ 62, and the remaining actors at ≈ 85, again reflecting the advantage of isotropic coverage in this symmetric problem. The inference standard deviation (bottom row) explains this ordering: at $d = 15$, the posterior contracts fastest under round-robin queries, followed by uniform sampling, and slowest under the posterior-driven actors. In each case, the stopping time closely tracks the point at which the standard deviation plateaus. The differences in sample complexity are therefore driven by how quickly each query policy acquires information.

Certification. Table 9 reports the certification p-values for every (d, ε, σ) configuration. Unlike binary search, C−ICPE−Uniform's certifiable region does not shrink with dimension here: it rejects H_0 at every configuration tested, $d = 5$ through $d = 15$, with no failures anywhere in the table, which is consistent with the rotational symmetry of ε-best arm making passive exploration competitive, unlike the harder, less symmetric tasks that follow. At $d = 15$, C−ICPE−RoundRobin achieves the strongest certification by a wide margin, with $p < 10^{-915}$ at both $\varepsilon = 0.005$ settings. All other active methods (C−ICPE−TD3, C−ICPE−TS, C−ICPE−TTPS, C−ICPE−TS−Mixture) also certify comfortably everywhere, with no configuration in this table showing a failure for any method.

Critic-free stopping. Unlike in binary search, removing the critic degrades correctness. C−ICPE−TS+CF falls below the 0.90 target in 6 of the 12 configurations, with accuracy as low as 0.804 at $d = 10$, $(\varepsilon, \sigma) = (0.005, 0.0025)$, whereas C−ICPE−TS meets the target everywhere (0.914–0.942). This is not explained by stopping earlier since the sample complexity of C−ICPE−TS+CF is similar to that of C−ICPE−TS, differing by at most about 12%. Its results are also much less consistent across seeds, with wide confidence intervals (e.g., $[0.702, 0.905]$ in the configuration above). Certification makes this observation sharper where every seed of C−ICPE−TS+CF rejects H_0 in only 2 of the 12 configurations, and at $d = 10$ no configuration certifies across all seeds (Table 9), while every seed of C−ICPE−TS certifies in every configuration. In this problem, the inference network's belief is not a calibrated estimate of the success probability, so thresholding its mass on the ε-ball is unreliable. As a contrast, the critic, which is trained directly on recommendation outcomes, corrects for this miscalibration.

d	Method	$\varepsilon = 0.02$		$\varepsilon = 0.005$	
		$\sigma = 0.005$	$\sigma = 0.01$	$\sigma = 0.00125$	$\sigma = 0.0025$
5	C-ICPE-TD3	0.1128 [0.1101,0.1155]	0.1096 [0.1081,0.1111]	0.0317 [0.0228,0.0405]	0.0333 [0.0262,0.0405]
	C-ICPE-TS	0.1413 [0.1142,0.1684]	0.1212 [0.1199,0.1225]	0.0152 [0.0142,0.0162]	0.0121 [0.0085,0.0156]
	C-ICPE-TTPS	0.1415 [0.1124,0.1707]	0.1196 [0.1148,0.1244]	0.0175 [0.0116,0.0235]	0.0119 [0.0084,0.0154]
	C-ICPE-uniform	0.1146 [0.1128,0.1165]	0.1129 [0.1103,0.1154]	0.0682 [0.0664,0.0700]	0.0679 [0.0670,0.0689]
10	C-ICPE-TD3	0.0345 [0.0335,0.0355]	0.0210 [0.0081,0.0340]	0.0001 [0.0000,0.0003]	0.0086 [0.0035,0.0137]
	C-ICPE-TS	0.0007 [0.0004,0.0011]	0.0001 [-0.0001,0.0003]	0.0002 [0.0000,0.0004]	0.0001 [0.0000,0.0002]
	C-ICPE-TTPS	0.0023 [0.0011,0.0034]	0.0000 [0.0000,0.0000]	0.0002 [-0.0002,0.0006]	0.0002 [-0.0002,0.0006]
	C-ICPE-uniform	0.0211 [0.0150,0.0271]	0.0190 [0.0150,0.0229]	0.0000 [0.0000,0.0001]	0.0001 [0.0000,0.0001]
15	C-ICPE-TD3	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0001]	0.0001 [0.0000,0.0002]
	C-ICPE-TS	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0001]	0.0000 [0.0000,0.0001]
	C-ICPE-TTPS	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0001 [0.0000,0.0001]
	C-ICPE-uniform	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0001]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0001]
	C-ICPE-TS-Mixture	0.0315 [-0.0082,0.0711]	0.0358 [0.0124,0.0592]	0.0351 [-0.0030,0.0731]	0.0441 [0.0370,0.0513]

Table 6: ϵ-best arm: final cost multiplier (mean and 95% CI across seeds) for every (d, ε, σ) configuration.

d	Method	$\varepsilon = 0.02$		$\varepsilon = 0.005$	
		$\sigma = 0.005$	$\sigma = 0.01$	$\sigma = 0.00125$	$\sigma = 0.0025$
5	C-ICPE-TD3	0.921 [.912,.931]	0.924 [.916,.931]	0.914 [.904,.923]	0.920 [.911,.927]
	C-ICPE-TS	0.942 [.931,.951]	0.933 [.922,.944]	0.917 [.906,.928]	0.923 [.915,.928]
	C-ICPE-TS+CF	0.925 [.898,.951]	0.936 [.921,.951]	0.891 [.859,.923]	0.914 [.896,.932]
	C-ICPE-TTPS	0.948 [.934,.963]	0.928 [.918,.936]	0.925 [.913,.936]	0.927 [.917,.937]
	C-ICPE-uniform	0.930 [.920,.938]	0.931 [.923,.938]	0.923 [.916,.931]	0.917 [.909,.923]
	LT&S round-robin	**1.000** [1.000,1.000]	**1.000** [1.000,1.000]	**1.000** [1.000,1.000]	**1.000** [1.000,1.000]
	LT&S uniform	1.000 [1.000,1.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]
10	C-ICPE-TD3	0.927 [.921,.935]	0.920 [.908,.932]	0.920 [.913,.926]	0.925 [.913,.936]
	C-ICPE-TS	0.920 [.910,.929]	0.920 [.911,.926]	0.928 [.918,.936]	0.923 [.915,.931]
	C-ICPE-TS+CF	0.908 [.891,.924]	0.895 [.882,.909]	0.889 [.879,.898]	0.804 [.702,.905]
	C-ICPE-TTPS	0.924 [.916,.931]	0.921 [.911,.931]	0.932 [.923,.938]	0.933 [.921,.943]
	C-ICPE-uniform	0.913 [.906,.921]	0.922 [.914,.928]	0.917 [.910,.924]	0.923 [.911,.935]
	LT&S round-robin	**1.000** [1.000,1.000]	**1.000** [1.000,1.000]	**1.000** [1.000,1.000]	**1.000** [1.000,1.000]
	LT&S uniform	1.000 [1.000,1.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]
15	C-ICPE-TD3	0.917 [.909,.924]	0.925 [.918,.931]	0.927 [.917,.936]	0.925 [.918,.931]
	C-ICPE-TS	0.917 [.908,.925]	0.914 [.903,.924]	0.922 [.905,.938]	0.934 [.925,.942]
	C-ICPE-TS+CF	0.949 [.920,.977]	0.927 [.895,.958]	0.892 [.880,.905]	0.898 [.856,.939]
	C-ICPE-TTPS	0.914 [.905,.925]	0.919 [.910,.927]	0.938 [.923,.953]	0.936 [.926,.945]
	C-ICPE-uniform	0.918 [.911,.925]	0.915 [.905,.923]	0.921 [.909,.932]	0.920 [.912,.928]
	C-ICPE-RoundRobin	0.963 [.911,1.000]	0.966 [.935,1.000]	**1.000** [1.000,1.000]	**1.000** [1.000,1.000]
	C-ICPE-TS-Mixture	0.930 [.923,.936]	0.920 [.912,.928]	0.920 [.907,.933]	0.925 [.917,.931]
	LT&S round-robin	**1.000** [1.000,1.000]	**1.000** [1.000,1.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]
	LT&S uniform	1.000 [1.000,1.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]

Table 7: ϵ-best arm: accuracy (mean and 95% CI) for every (d, ε, σ) configuration. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

d	Method	$\varepsilon = 0.02$		$\varepsilon = 0.005$	
		$\sigma = 0.005$	$\sigma = 0.01$	$\sigma = 0.00125$	$\sigma = 0.0025$
5	C-ICPE-TD3	7.3 [7.3,7.4]	7.5 [7.4,7.5]	9.5 [9.3,9.6]	9.6 [9.5,9.7]
	C-ICPE-TS	8.3 [8.3,8.4]	8.0 [7.7,8.3]	9.3 [9.2,9.5]	9.5 [9.3,9.7]
	C-ICPE-TS+CF	8.0 [7.8,8.3]	7.9 [7.6,8.1]	9.6 [8.8,10.4]	9.8 [9.6,9.9]
	C-ICPE-TTPS	8.4 [8.2,8.6]	8.0 [8.0,8.1]	9.3 [9.1,9.5]	9.5 [9.4,9.6]
	C-ICPE-uniform	8.5 [8.4,8.7]	8.3 [8.2,8.3]	10.3 [10.2,10.4]	10.2 [10.2,10.3]
	LT&S round-robin	**5.0** [5.0,5.0]	**5.0** [5.0,5.0]	**5.0** [5.0,5.0]	**5.0** [5.0,5.0]
	LT&S uniform	5.0 [5.0,5.0]	5.0 [5.0,5.0]	5.0 [5.0,5.0]	5.0 [5.0,5.0]
10	C-ICPE-TD3	21.5 [21.3,21.6]	22.6 [22.1,23.1]	35.9 [34.7,37.2]	42.6 [38.7,44.7]
	C-ICPE-TS	21.1 [20.4,21.8]	24.1 [23.5,24.8]	35.1 [34.0,35.9]	38.8 [36.2,40.3]
	C-ICPE-TS+CF	21.5 [20.9,22.1]	23.0 [21.5,24.5]	35.7 [31.7,39.6]	34.1 [29.0,39.1]
	C-ICPE-TTPS	20.8 [20.1,21.3]	23.2 [22.5,23.8]	33.4 [32.6,34.2]	37.6 [36.6,39.2]
	C-ICPE-uniform	22.1 [21.9,22.4]	22.4 [22.0,22.6]	31.7 [31.5,32.0]	32.0 [31.7,32.3]
	LT&S round-robin	**10.0** [10.0,10.0]	**10.0** [10.0,10.0]	**10.0** [10.0,10.0]	**10.0** [10.0,10.0]
	LT&S uniform	10.0 [10.0,10.0]	10.0 [10.0,10.0]	10.0 [10.0,10.0]	10.0 [10.0,10.0]
15	C-ICPE-TD3	46.5 [46.1,47.0]	49.7 [49.4,49.9]	73.8 [72.1,75.4]	85.3 [81.9,90.1]
	C-ICPE-TS	46.3 [45.7,47.0]	55.3 [54.6,56.0]	76.8 [73.6,79.7]	85.4 [84.6,86.3]
	C-ICPE-TS+CF	50.0 [46.3,53.8]	56.9 [54.5,59.4]	77.8 [66.6,89.0]	87.4 [85.1,89.7]
	C-ICPE-TTPS	45.4 [45.1,45.6]	56.0 [54.9,57.2]	75.6 [72.8,78.8]	85.1 [84.4,86.1]
	C-ICPE-uniform	44.0 [43.9,44.1]	44.7 [44.2,45.1]	62.0 [61.5,62.3]	62.6 [61.9,63.1]
	C-ICPE-RoundRobin	**14.8** [14.6,15.0]	**14.8** [14.6,15.0]	**15.0** [15.0,15.0]	**15.0** [15.0,15.0]
	C-ICPE-TS-Mixture	47.1 [46.7,47.5]	57.5 [56.8,58.3]	74.3 [73.4,75.0]	86.8 [85.4,89.0]
	LT&S round-robin	15.0 [15.0,15.0]	15.0 [15.0,15.0]	15.0 [15.0,15.0]	15.0 [15.0,15.0]
	LT&S uniform	15.0 [15.0,15.0]	15.0 [15.0,15.0]	15.0 [15.0,15.0]	15.0 [15.0,15.0]

Table 8: ϵ-best arm: sample complexity (mean and 95% CI) for every (d, ε, σ) configuration. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

d	Method	$\varepsilon = 0.02$		$\varepsilon = 0.005$	
		$\sigma = 0.005$	$\sigma = 0.01$	$\sigma = 0.00125$	$\sigma = 0.0025$
5	C-ICPE-TD3	2.5×10^{-19} [1.4×10^{-39},7.5×10^{-19}]	1.8×10^{-11} [1.1×10^{-32},5.3×10^{-11}]	8.6×10^{-11} [7.6×10^{-21},2.6×10^{-10}]	3.4×10^{-7} [2.8×10^{-39},1.0×10^{-6}]
	C-ICPE-TS	8.4×10^{-78} [1.8×10^{-131},2.5×10^{-77}]	7.3×10^{-32} [1.8×10^{-88},2.2×10^{-31}]	6.5×10^{-17} [3.7×10^{-40},1.9×10^{-16}]	$\mathbf{6.0 \times 10^{-19}}$ [1.1×10^{-37},1.8×10^{-18}]
	C-ICPE-TS+CF	1.5×10^{-1} [9.7×10^{-136},4.5×10^{-1}]	3.0×10^{-34} [7.1×10^{-128},9.1×10^{-34}]	3.7×10^{-1} [4.3×10^{-2},1.0×10^{0}]	3.3×10^{-1} [1.3×10^{-40},1.0×10^{0}]
	C-ICPE-TTPS	$\mathbf{9.4 \times 10^{-85}}$ [1.1×10^{-287},2.8×10^{-84}]	$\mathbf{2.2 \times 10^{-43}}$ [6.8×10^{-57},6.7×10^{-43}]	$\mathbf{7.5 \times 10^{-19}}$ [2.2×10^{-37},2.2×10^{-18}]	5.1×10^{-15} [3.3×10^{-47},1.5×10^{-14}]
	C-ICPE-uniform	1.7×10^{-12} [5.1×10^{-76},5.1×10^{-12}]	2.0×10^{-38} [4.7×10^{-51},5.8×10^{-38}]	3.3×10^{-12} [1.4×10^{-34},1.0×10^{-11}]	4.0×10^{-11} [7.9×10^{-24},1.2×10^{-10}]
10	C-ICPE-TD3	$\mathbf{2.9 \times 10^{-45}}$ [2.0×10^{-78},8.8×10^{-45}]	2.5×10^{-4} [2.3×10^{-41},7.5×10^{-4}]	1.6×10^{-20} [3.7×10^{-27},4.9×10^{-20}]	$\mathbf{3.4 \times 10^{-17}}$ [7.4×10^{-60},5.6×10^{-17}]
	C-ICPE-TS	8.2×10^{-21} [5.7×10^{-36},2.5×10^{-20}]	5.2×10^{-6} [2.2×10^{-31},1.6×10^{-5}]	1.2×10^{-44} [5.5×10^{-56},1.8×10^{-44}]	2.1×10^{-13} [9.6×10^{-43},6.2×10^{-13}]
	C-ICPE-TS+CF	3.3×10^{-1} [3.9×10^{-35},1.0×10^{0}]	6.7×10^{-1} [2.4×10^{-2},1.0×10^{0}]	6.9×10^{-1} [8.0×10^{-2},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]
	C-ICPE-TTPS	1.6×10^{-29} [3.0×10^{-66},3.1×10^{-29}]	8.4×10^{-10} [1.1×10^{-38},2.5×10^{-9}]	$\mathbf{1.2 \times 10^{-55}}$ [3.0×10^{-96},2.9×10^{-55}]	4.7×10^{-12} [1.8×10^{-109},1.4×10^{-11}]
	C-ICPE-uniform	2.2×10^{-16} [3.1×10^{-20},6.6×10^{-16}]	$\mathbf{6.5 \times 10^{-18}}$ [5.6×10^{-29},2.0×10^{-17}]	3.7×10^{-21} [4.2×10^{-26},9.6×10^{-21}]	4.9×10^{-14} [4.7×10^{-65},1.5×10^{-13}]
15	C-ICPE-TD3	1.3×10^{-9} [1.5×10^{-13},3.9×10^{-9}]	1.3×10^{-19} [7.4×10^{-29},3.8×10^{-19}]	2.7×10^{-22} [7.0×10^{-51},5.5×10^{-22}]	3.5×10^{-13} [1.1×10^{-36},1.1×10^{-12}]
	C-ICPE-TS	2.1×10^{-13} [7.4×10^{-29},6.2×10^{-13}]	1.9×10^{-10} [9.8×10^{-23},5.6×10^{-10}]	4.6×10^{-11} [2.3×10^{-47},1.4×10^{-10}]	1.1×10^{-9} [3.0×10^{-96},3.4×10^{-9}]
	C-ICPE-TS+CF	1.8×10^{-28} [1.4×10^{-277},5.3×10^{-28}]	3.3×10^{-1} [2.8×10^{-185},1.0×10^{0}]	9.9×10^{-1} [9.8×10^{-1},1.0×10^{0}]	6.7×10^{-1} [1.2×10^{-65},1.0×10^{0}]
	C-ICPE-TTPS	4.7×10^{-13} [1.5×10^{-19},1.3×10^{-12}]	6.4×10^{-25} [7.1×10^{-47},1.7×10^{-24}]	5.1×10^{-21} [1.5×10^{-174},1.5×10^{-20}]	2.4×10^{-47} [1.3×10^{-170},7.1×10^{-47}]
	C-ICPE-uniform	4.0×10^{-16} [7.4×10^{-30},1.2×10^{-15}]	2.2×10^{-16} [2.3×10^{-30},6.6×10^{-16}]	2.1×10^{-11} [1.1×10^{-37},6.2×10^{-11}]	2.3×10^{-25} [2.3×10^{-28},6.0×10^{-25}]
	C-ICPE-RoundRobin	$\mathbf{5.6 \times 10^{-44}}$ [$< 10^{-915}$,1.1×10^{-43}]	$\mathbf{1.1 \times 10^{-54}}$ [$< 10^{-915}$,3.4×10^{-54}]	$\mathbf{< 10^{-915}}$ [$< 10^{-915}$,$< 10^{-915}$]	$\mathbf{< 10^{-915}}$ [$< 10^{-915}$,$< 10^{-915}$]
	C-ICPE-TS-Mixture	4.8×10^{-33} [2.1×10^{-51},1.5×10^{-32}]	5.9×10^{-24} [7.3×10^{-40},1.7×10^{-23}]	9.9×10^{-18} [3.2×10^{-46},3.0×10^{-17}]	5.1×10^{-15} [1.8×10^{-40},1.5×10^{-14}]

Table 9: ϵ-best arm: certification p-value ($H_0 : p \leq 0.9$, one-sided exact binomial test on $n = 20000$ independent trials, significance level $\alpha = 0.01$) for every (d, ε, σ) configuration: mean across seeds, with the [min, max] range across seeds in brackets. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

E.3.3 ACKLEY MINIMIZATION

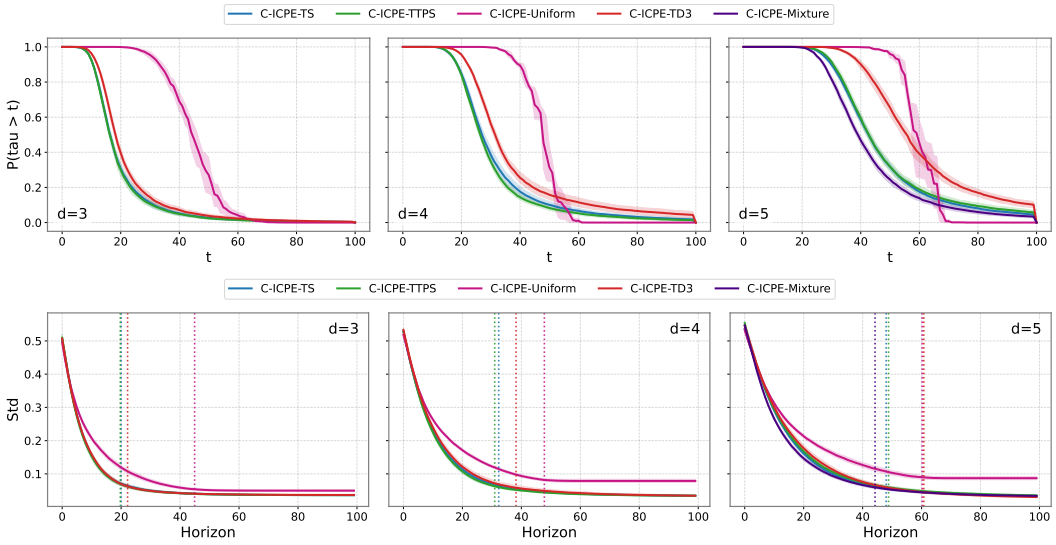


Figure 13: Results for Ackley function with fixed confidence $\delta = 0.1$ and $N = 100$ across different dimensions at the hardest (ε, σ) setting: (top) survival function of τ; (bottom) inference uncertainty convergence.

Tables 11 and 12 report accuracy and sample complexity, and Fig. 13 shows the survival function and inference uncertainty convergence. For simplicity, we evaluate `C-ICPE-TS-Mixture` in the highest dimension, $d = 5$. As in the previous problems, we also evaluate the critic-free ablation `C-ICPE-TS+CF`. Each BO baseline receives a fixed budget equal to the mean stopping time of the most sample-efficient `C-ICPE` variant in each configuration, rounded up; to assess how the baselines scale with additional queries, we also report results with a budget of 100 queries. We also report the cost multiplier at the end of training for every (d, ε, σ) configuration in Table 10.

Accuracy. All active-actor `C-ICPE` variants meet the $1 - \delta = 0.90$ target across every (d, ε, σ) configuration, with mean accuracy between 0.910 and 0.935. In contrast, `C-ICPE-Uniform` falls short of the target at $(\varepsilon, \sigma) = (0.1, 0.05)$, and its accuracy degrades sharply with dimension, from 0.807 at $d = 3$ to 0.456 at $d = 4$ and 0.293 at $d = 5$. Unlike the ε-best-arm problem, the Ackley function has no symmetry that makes uniform exploration competitive: its multimodal landscape and flat outer region require directed queries to distinguish the global minimizer from local optima. The baselines fail across the board. `GP-UCB` achieves the highest accuracy among them, yet reaches only 0.308–0.386 at $d = 3$, far below the target even in the easiest configuration, while `TPE` and `CMA-ES` are near zero for $d \geq 4$. A larger budget does not close the gap: at $d = 5$, the best fixed-budget baseline, $\texttt{GP-UCB}_{T=100}$, reaches only 0.388–0.500. These methods optimize a fixed-budget objective without a stopping rule and are not designed for (ε, δ)-correctness; the results indicate that standard black-box optimization does not provide fixed-confidence guarantees, even with substantially more queries than `C-ICPE` requires.

Sample complexity. Active actors use roughly half the samples of `C-ICPE-Uniform` across all dimensions. At $d = 5$, $(\varepsilon, \sigma) = (0.1, 0.05)$, `C-ICPE-TS` stops at 48.0 and `C-ICPE-TTPS` at 48.8, compared to 60.2 for `C-ICPE-Uniform`. `C-ICPE-TS` and `C-ICPE-TTPS` achieve similar sample complexity throughout, while `C-ICPE-TD3` is higher (e.g., 60.8 at the same setting). Sample complexity grows with dimension for all methods: `C-ICPE-TS` increases from 15.0 ($d = 3$) to 35.4 ($d = 5$) at $(\varepsilon, \sigma) = (0.2, 0.05)$. Increasing σ at fixed ε consistently raises sample complexity, as expected from the noisier observations. Notably, `C-ICPE-TS-Mixture` shows the lowest sample complexity across all settings at $d = 5$ among all the methods. Although the goal is to localize a single global optimum, the Ackley function has many local optima, which makes the posterior over the optimum's location highly multimodal. `C-ICPE-TS-Mixture` captures this multimodality better, as reflected in its lowest sample complexity.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
3	C-ICPE-TD3	0.0191 [0.0131,0.0251]	0.0012 [0.0008,0.0015]	0.0034 [-0.0006,0.0073]	0.0012 [-0.0002,0.0026]
	C-ICPE-TS	0.0464 [0.0331,0.0598]	0.0075 [0.0025,0.0125]	0.0047 [-0.0029,0.0124]	0.0001 [0.0000,0.0001]
	C-ICPE-TTPS	0.0289 [0.0177,0.0402]	0.0008 [-0.0001,0.0017]	0.0002 [0.0002,0.0003]	0.0001 [0.0000,0.0002]
	C-ICPE-uniform	0.0002 [0.0000,0.0003]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]
4	C-ICPE-TD3	0.0003 [-0.0001,0.0006]	0.0001 [0.0000,0.0002]	0.0001 [0.0001,0.0001]	0.0001 [0.0000,0.0002]
	C-ICPE-TS	0.0008 [-0.0001,0.0016]	0.0002 [0.0001,0.0002]	0.0033 [0.0001,0.0065]	0.0001 [0.0000,0.0002]
	C-ICPE-TTPS	0.0005 [0.0001,0.0009]	0.0001 [0.0000,0.0003]	0.0016 [-0.0015,0.0048]	0.0000 [0.0000,0.0001]
	C-ICPE-uniform	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]
5	C-ICPE-TD3	0.0001 [0.0000,0.0003]	0.0005 [-0.0004,0.0013]	0.0009 [-0.0005,0.0024]	0.0001 [-0.0001,0.0003]
	C-ICPE-TS	0.0003 [0.0001,0.0005]	0.0001 [0.0001,0.0002]	0.0003 [0.0001,0.0005]	0.0001 [-0.0001,0.0003]
	C-ICPE-TTPS	0.0003 [0.0002,0.0003]	0.0002 [0.0001,0.0003]	0.0004 [-0.0001,0.0009]	0.0002 [0.0002,0.0003]
	C-ICPE-uniform	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]
	C-ICPE-TS-Mixture	0.0002 [0.0001,0.0004]	0.0001 [0.0000,0.0001]	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0001]

Table 10: Ackley: final cost multiplier (mean and 95% CI across seeds) for every (d, ε, σ) configuration.

Stopping behavior. The survival functions (Fig. 13, top) show a clear separation between active and passive exploration. At $d = 3$, `C-ICPE-TS` and `C-ICPE-TTPS` exhibit sharp transitions around $t \approx 15$–25, while `C-ICPE-Uniform` has a heavy tail extending past $t = 40$. At $d = 5$, the active actors stop around $t \approx 35$–50 while `C-ICPE-Uniform` rarely stops before $t = 60$ and retains substantial probability mass near the horizon. The inference standard deviation (bottom row) reveals a qualitative difference from binary search: the posterior uncertainty plateaus around 0.05–0.1 rather than converging to zero. This reflects the inherent difficulty of the Ackley landscape, the multimodal structure and observation noise prevent the inference model from achieving the same posterior concentration as in the unimodal binary search setting. Nevertheless, the critic learns to stop at an appropriate uncertainty level that is sufficient for ε-correctness.

Certification. Table 13 reports the certification p-values for every (d, ε, σ) configuration. Every active method rejects H_0 by many orders of magnitude at every configuration, with p-values typically below 10^{-10} , which far beyond the $\alpha = 0.01$ threshold, confirming that the accuracies in Table 11 reflect a certified margin above the 0.9 target rather than a favorable point estimate at finite sample size. `C-ICPE-Uniform` is consistent with its accuracy results: it also certifies at the easier $\varepsilon = 0.2$ configurations ($p < 10^{-15}$), but at $\varepsilon = 0.1$, where its point accuracy already falls well below 0.9, the test fails outright ($p = 1.0$ across all seeds), correctly reflecting the absence of any statistical evidence for a 0.9 success rate there.

Critic-free stopping. For Ackley proble, the effect of removing the critic depends on dimension. At $d = 3$ and $d = 4$, `C-ICPE-TS+CF` exceeds the target in every configuration (0.927–0.968), with sample complexity within about 9% of `C-ICPE-TS`, and it certifies everywhere. However, at $d = 5$, it falls below the target in the two noisier configurations, reaching 0.894 at $(\varepsilon, \sigma) = (0.2, 0.1)$ and 0.898 at $(0.1, 0.05)$, where it also stops earlier than `C-ICPE-TS` (44.8 vs. 49.0 and 45.0 vs. 48.0 queries). Certification confirms these failures (Table 13), at $(0.2, 0.1)$ no seed rejects H_0, and at $(0.1, 0.05)$ only some seeds do, whereas `C-ICPE-TS` certifies in every configuration with every seed at $p \leq 8.5 \times 10^{-11}$. As the problem becomes harder, the inference network's belief becomes overconfident and a fixed threshold on it stops too early, while the trained critic can adjust the stopping rule to maintain correctness.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
3	C-ICPE-TD3	0.917 [.907,.925]	0.913 [.903,.921]	0.913 [.904,.921]	0.919 [.912,.927]
	C-ICPE-TS	0.913 [.901,.924]	0.915 [.905,.924]	0.922 [.914,.929]	0.925 [.915,.933]
	C-ICPE-TS+CF	**0.968** [.956,.980]	**0.948** [.945,.950]	**0.960** [.956,.964]	**0.949** [.947,.951]
	C-ICPE-TTPS	0.916 [.905,.925]	0.918 [.910,.926]	0.921 [.915,.929]	0.919 [.909,.927]
	C-ICPE-uniform	0.923 [.915,.932]	0.915 [.900,.928]	0.863 [.835,.889]	0.807 [.789,.820]
	TPE	0.098 [.072,.124]	0.138 [.108,.168]	0.014 [.004,.024]	0.024 [.011,.037]
	CMA-ES	0.074 [.051,.097]	0.100 [.074,.126]	0.016 [.005,.027]	0.018 [.006,.030]
	GP-logEI	0.330 [.289,.371]	0.320 [.279,.361]	0.168 [.135,.201]	0.170 [.137,.203]
	GP-UCB	0.386 [.343,.429]	0.364 [.322,.406]	0.316 [.275,.357]	0.308 [.267,.349]
	TPE$_{T=100}$	0.848 [.816,.880]	0.808 [.773,.843]	0.690 [.649,.731]	0.640 [.598,.682]
	CMA-ES$_{T=100}$	0.620 [.577,.663]	0.594 [.551,.637]	0.532 [.488,.576]	0.480 [.436,.524]
	GP-logEI$_{T=100}$	0.820 [.786,.854]	0.734 [.695,.773]	0.696 [.656,.736]	0.568 [.525,.611]
	GP-UCB$_{T=100}$	0.690 [.649,.731]	0.654 [.612,.696]	0.706 [.666,.746]	0.652 [.610,.694]
4	C-ICPE-TD3	0.917 [.907,.925]	0.921 [.911,.930]	0.923 [.911,.933]	0.910 [.895,.925]
	C-ICPE-TS	0.923 [.910,.935]	0.918 [.910,.927]	0.921 [.912,.930]	0.920 [.909,.931]
	C-ICPE-TS+CF	**0.950** [.940,.960]	**0.927** [.920,.934]	**0.943** [.935,.952]	**0.935** [.919,.950]
	C-ICPE-TTPS	0.924 [.915,.931]	0.924 [.913,.933]	0.930 [.921,.939]	0.918 [.910,.928]
	C-ICPE-uniform	0.850 [.830,.866]	0.684 [.662,.707]	0.604 [.572,.638]	0.456 [.426,.485]
	TPE	0.058 [.037,.079]	0.072 [.049,.095]	0.006 [.000,.013]	0.008 [.000,.016]
	CMA-ES	0.030 [.015,.045]	0.044 [.026,.062]	0.006 [.000,.013]	0.000 [.000,.000]
	GP-logEI	0.204 [.169,.239]	0.174 [.141,.207]	0.106 [.079,.133]	0.074 [.051,.097]
	GP-UCB	0.324 [.283,.365]	0.286 [.246,.326]	0.246 [.208,.284]	0.208 [.172,.244]
	TPE$_{T=100}$	0.596 [.553,.639]	0.510 [.466,.554]	0.232 [.195,.269]	0.180 [.146,.214]
	CMA-ES$_{T=100}$	0.474 [.430,.518]	0.362 [.320,.404]	0.198 [.163,.233]	0.154 [.122,.186]
	GP-logEI$_{T=100}$	0.562 [.518,.606]	0.334 [.293,.375]	0.356 [.314,.398]	0.230 [.193,.267]
	GP-UCB$_{T=100}$	0.602 [.559,.645]	0.560 [.516,.604]	0.594 [.551,.637]	0.534 [.490,.578]
5	C-ICPE-TD3	0.919 [.910,.929]	0.915 [.896,.933]	0.921 [.911,.931]	0.919 [.905,.933]
	C-ICPE-TS	0.917 [.904,.928]	**0.926** [.914,.935]	0.925 [.913,.933]	0.919 [.908,.930]
	C-ICPE-TS+CF	0.924 [.914,.934]	0.894 [.892,.896]	**0.929** [.914,.944]	0.898 [.888,.909]
	C-ICPE-TTPS	0.923 [.913,.932]	0.920 [.905,.934]	0.925 [.910,.936]	0.912 [.901,.922]
	C-ICPE-uniform	0.786 [.766,.801]	0.523 [.497,.548]	0.490 [.429,.555]	0.293 [.268,.317]
	C-ICPE-TS-Mixture	**0.935** [.923,.945]	0.920 [.907,.931]	0.914 [.901,.924]	**0.920** [.910,.930]
	TPE	0.014 [.004,.024]	0.020 [.008,.032]	0.000 [.000,.000]	0.002 [.000,.006]
	CMA-ES	0.010 [.001,.019]	0.026 [.012,.040]	0.000 [.000,.000]	0.000 [.000,.000]
	GP-logEI	0.106 [.079,.133]	0.072 [.049,.095]	0.042 [.024,.060]	0.030 [.015,.045]
	GP-UCB	0.294 [.254,.334]	0.182 [.148,.216]	0.178 [.144,.212]	0.160 [.128,.192]
	TPE$_{T=100}$	0.282 [.243,.321]	0.228 [.191,.265]	0.040 [.023,.057]	0.026 [.012,.040]
	CMA-ES$_{T=100}$	0.268 [.229,.307]	0.172 [.139,.205]	0.032 [.017,.047]	0.022 [.009,.035]
	GP-logEI$_{T=100}$	0.248 [.210,.286]	0.126 [.097,.155]	0.148 [.117,.179]	0.070 [.048,.092]
	GP-UCB$_{T=100}$	0.500 [.456,.544]	0.414 [.371,.457]	0.450 [.406,.494]	0.388 [.345,.431]

Table 11: Ackley: accuracy (mean and 95% CI) for every (d, ε, σ) configuration. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. Unless marked $T = 100$, baselines use a budget equal to the mean stopping time of the most sample-efficient `C-ICPE` variant, rounded up. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

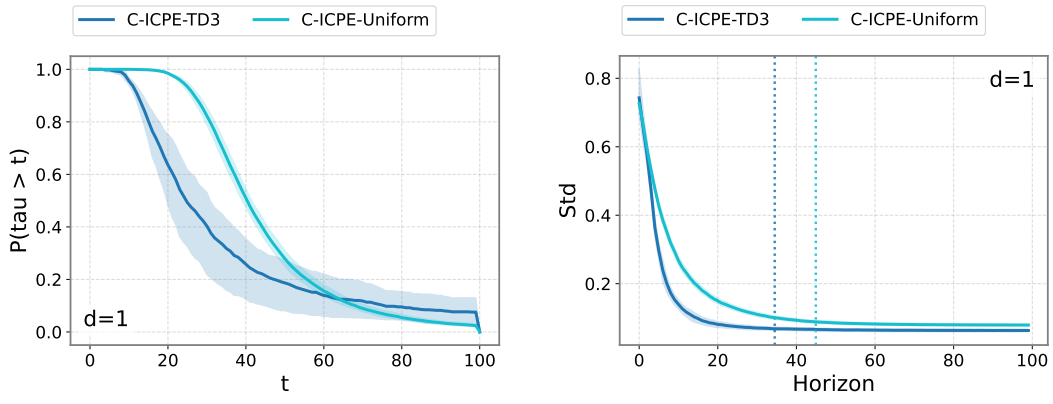
d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
3	C-ICPE-TD3	16.5 [15.9,17.1]	20.4 [19.7,21.0]	20.2 [19.5,21.1]	22.2 [21.4,22.9]
	C-ICPE-TS	**15.0** [14.4,15.6]	**18.8** [18.0,19.7]	**17.3** [16.8,17.8]	20.0 [19.3,20.9]
	C-ICPE-TS+CF	16.3 [16.0,16.6]	20.1 [19.7,20.5]	17.8 [17.2,18.3]	20.4 [19.7,21.1]
	C-ICPE-TTPS	15.5 [15.0,16.1]	19.4 [18.7,20.2]	18.0 [17.4,18.6]	**19.7** [19.1,20.4]
	C-ICPE-uniform	38.7 [37.9,39.5]	51.3 [49.6,53.4]	44.4 [42.8,46.3]	44.9 [43.1,46.5]
4	C-ICPE-TD3	27.2 [26.3,28.3]	35.8 [34.8,36.9]	32.8 [31.6,34.1]	38.2 [36.5,40.0]
	C-ICPE-TS	24.4 [23.6,25.2]	31.6 [30.2,33.1]	28.0 [27.2,29.2]	32.3 [31.0,33.8]
	C-ICPE-TS+CF	25.0 [24.5,25.5]	31.3 [30.6,32.1]	28.3 [27.0,29.5]	32.0 [31.0,33.1]
	C-ICPE-TTPS	**24.1** [23.1,25.2]	**31.3** [30.4,32.3]	**26.6** [25.8,27.5]	**31.0** [30.3,31.9]
	C-ICPE-uniform	55.8 [54.4,57.2]	53.5 [52.4,54.8]	48.6 [47.0,50.1]	47.8 [46.7,49.3]
5	C-ICPE-TD3	42.7 [41.4,44.1]	58.9 [57.8,60.1]	52.3 [50.6,53.6]	60.8 [59.0,62.4]
	C-ICPE-TS	35.4 [33.8,37.2]	49.0 [47.8,50.2]	38.7 [37.7,40.0]	48.0 [47.1,48.9]
	C-ICPE-TS+CF	34.4 [33.9,34.9]	44.8 [44.1,45.6]	39.1 [38.3,39.8]	45.0 [43.4,46.6]
	C-ICPE-TTPS	34.0 [32.9,35.5]	48.2 [46.7,49.6]	38.6 [36.9,40.3]	48.8 [47.4,50.2]
	C-ICPE-uniform	68.7 [66.4,70.6]	66.9 [65.5,68.0]	62.1 [58.3,65.3]	60.2 [58.6,61.3]
	C-ICPE-TS-Mixture	**31.4** [30.4,32.2]	**43.0** [41.9,44.6]	**35.3** [34.3,36.5]	**44.2** [43.3,45.1]

Table 12: Ackley: sample complexity (mean and 95% CI) for every (d, ε, σ) configuration. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
3	C-ICPE-TD3	1.1×10^{-10} [2.5×10^{-13},3.0×10^{-10}]	4.4×10^{-7} [9.9×10^{-16},1.3×10^{-6}]	9.4×10^{-9} [1.2×10^{-25},2.8×10^{-8}]	2.4×10^{-11} [1.3×10^{-28},7.3×10^{-11}]
	C-ICPE-TS	1.1×10^{-14} [1.4×10^{-18},3.3×10^{-14}]	1.7×10^{-12} [1.0×10^{-24},5.1×10^{-12}]	3.2×10^{-21} [2.9×10^{-31},9.6×10^{-21}]	2.5×10^{-21} [4.9×10^{-50},7.6×10^{-21}]
	C-ICPE-TS+CF	$\mathbf{6.1 \times 10^{-221}}$ [$< 10^{-300}$,1.8×10^{-220}]	$\mathbf{8.1 \times 10^{-117}}$ [1.0×10^{-155},2.4×10^{-116}]	$\mathbf{4.7 \times 10^{-179}}$ [3.2×10^{-267},1.4×10^{-178}]	$\mathbf{2.2 \times 10^{-117}}$ [1.2×10^{-182},6.6×10^{-117}]
	C-ICPE-TTPS	7.4×10^{-11} [4.6×10^{-23},2.2×10^{-10}]	2.8×10^{-18} [9.3×10^{-28},8.3×10^{-18}]	1.5×10^{-29} [2.7×10^{-32},4.2×10^{-29}]	1.3×10^{-17} [1.6×10^{-37},3.7×10^{-17}]
	C-ICPE-uniform	1.5×10^{-17} [2.9×10^{-21},3.7×10^{-17}]	8.2×10^{-16} [1.4×10^{-21},1.5×10^{-15}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]
4	C-ICPE-TD3	1.4×10^{-10} [4.9×10^{-20},4.1×10^{-10}]	2.9×10^{-14} [9.3×10^{-28},8.4×10^{-14}]	5.9×10^{-14} [7.9×10^{-33},1.8×10^{-13}]	1.0×10^{-20} [5.3×10^{-28},3.1×10^{-20}]
	C-ICPE-TS	6.0×10^{-19} [4.6×10^{-46},1.8×10^{-18}]	2.3×10^{-14} [2.8×10^{-23},7.0×10^{-14}]	2.8×10^{-11} [1.5×10^{-54},8.5×10^{-11}]	1.0×10^{-25} [6.4×10^{-27},2.7×10^{-25}]
	C-ICPE-TS+CF	$\mathbf{1.7 \times 10^{-95}}$ [4.7×10^{-157},5.2×10^{-95}]	1.7×10^{-12} [4.2×10^{-37},5.1×10^{-12}]	$\mathbf{3.8 \times 10^{-77}}$ [1.0×10^{-166},1.1×10^{-76}]	$\mathbf{1.4 \times 10^{-29}}$ [3.6×10^{-128},4.2×10^{-29}]
	C-ICPE-TTPS	2.6×10^{-20} [3.1×10^{-29},7.8×10^{-20}]	$\mathbf{6.5 \times 10^{-18}}$ [5.6×10^{-42},2.0×10^{-17}]	7.0×10^{-35} [4.0×10^{-42},1.9×10^{-34}]	6.5×10^{-21} [9.1×10^{-34},1.9×10^{-20}]
	C-ICPE-uniform	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]
5	C-ICPE-TD3	2.4×10^{-11} [5.4×10^{-35},7.3×10^{-11}]	6.0×10^{-19} [7.3×10^{-26},1.8×10^{-18}]	2.8×10^{-12} [9.5×10^{-26},8.5×10^{-12}]	1.4×10^{-14} [2.4×10^{-17},4.0×10^{-14}]
	C-ICPE-TS	4.7×10^{-12} [1.4×10^{-26},1.4×10^{-11}]	$\mathbf{4.1 \times 10^{-23}}$ [1.2×10^{-31},7.6×10^{-23}]	2.8×10^{-17} [3.7×10^{-40},8.5×10^{-17}]	4.1×10^{-16} [3.6×10^{-32},9.9×10^{-16}]
	C-ICPE-TS+CF	5.9×10^{-14} [6.6×10^{-55},1.8×10^{-13}]	8.8×10^{-1} [7.2×10^{-1},1.0×10^{0}]	5.0×10^{-13} [3.7×10^{-75},1.5×10^{-12}]	6.7×10^{-1} [1.2×10^{-4},1.0×10^{0}]
	C-ICPE-TTPS	1.1×10^{-9} [1.7×10^{-28},3.4×10^{-9}]	1.2×10^{-16} [3.0×10^{-28},3.6×10^{-16}]	$\mathbf{1.4 \times 10^{-30}}$ [6.5×10^{-41},4.2×10^{-30}]	$\mathbf{2.2 \times 10^{-18}}$ [1.7×10^{-33},6.7×10^{-18}]
	C-ICPE-uniform	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]
	C-ICPE-TS-Mixture	$\mathbf{5.3 \times 10^{-38}}$ [5.5×10^{-56},1.6×10^{-37}]	1.8×10^{-18} [3.6×10^{-34},5.4×10^{-18}]	1.3×10^{-15} [1.7×10^{-24},4.0×10^{-15}]	4.9×10^{-16} [2.1×10^{-38},1.5×10^{-15}]

Table 13: Ackley: certification p-value ($H_0 : p \leq 0.9$, one-sided exact binomial test on $n = 20000$ independent trials, significance level $\alpha = 0.01$) for every (d, ε, σ) configuration: mean across seeds, with the [min, max] range across seeds in brackets. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

E.3.4 GP MAX-VALUE ESTIMATION



(a) (b)

Figure 14: Results for GP value estimation with fixed confidence $\delta = 0.1$ and $N = 100$ across different dimensions at the hardest (ε, σ) setting: (a) survival function of τ; (b) inference uncertainty convergence.

Tables 15 and 16 report accuracy and sample complexity, and Fig. 14 shows the survival function and inference uncertainty convergence. This is the $\mathcal{X} \neq \mathcal{A}$ setting: the agent queries locations $a \in [0,1]^d$ but recommends a scalar value estimate $\hat{x} \in \mathbb{R}$. Only `C-ICPE-TD3` and `C-ICPE-Uniform` are evaluated, since TS and TTPS require $\mathcal{X} = \mathcal{A}$. As in the previous problems, we also evaluate the critic-free ablation `C-ICPE-TD3+CF`. For comparison, we also evaluate black-box optimization baselines. Each baseline sequentially selects query points $a_1, \ldots, a_T$ and observes the corresponding noisy responses $y_1, \ldots, y_T$, where the budget T is set to the mean sample complexity of `C-ICPE`. After T queries, we fit a GP with a Matérn kernel ($\nu = 5/2$) to $\{(a_i, y_i)\}_{i=1}^{T}$ and recommend the maximum of the posterior mean, $\hat{x} = \max_a \mu_T(a)$. To separate the quality of `C-ICPE`'s queries from that of its inference network, we also apply the same post hoc procedure to `C-ICPE`'s query history, fitting a GP to the collected data and recommending the maximum of its posterior mean; we refer to this variant as `C-ICPE+GP`. Finally, Fig. 9 illustrates how `C-ICPE` explores this problem, showing the actor's queries and the evolution of the posterior mean and standard deviation over time. We also report the cost multiplier at the end of training for every (ε, σ) configuration in Table 14.

Accuracy. Both `C-ICPE` variants meet the $1 - \delta = 0.9$ target across all (ε, σ) configurations. `C-ICPE-TD3` achieves 0.900–0.927 and `C-ICPE-Uniform` 0.913–0.924, so the two are comparable in accuracy. With a comparable number of queries, `GP-logEI` and `TPE` attain higher accuracy, at 0.978–0.996 and 0.958–0.978, respectively. However, fitting a post hoc GP to `C-ICPE`'s own query history closes most of this gap, where `C-ICPE-TD3+GP` reaches 0.922–0.948, and `C-ICPE-Uniform+GP` reaches 0.960–0.971, on par with `TPE`. This indicates that `C-ICPE`'s queries are informative and that the remaining gap lies mainly in the inference network's recommendation. `GP-UCB` falls just short of the target in two of the four settings (0.890–0.902), and `CMA-ES` fails in all of them (0.722–0.748). The naive baselines perform far worse: Uniform bin achieves only 0.094–0.204 and Uniform top 5% 0.612–0.728.

Sample complexity. `C-ICPE-TD3` requires fewer queries than `C-ICPE-Uniform` in every configuration, by a factor of 1.30–1.75×. At $\varepsilon = 0.2$, it stops after 22.3 and 20.6 queries at $\sigma = 0.05$ and $\sigma = 0.1$, compared to 30.6 and 36.0 for `C-ICPE-Uniform`. The advantage persists at $\varepsilon = 0.15$ (24.5 vs. 38.0 at $\sigma = 0.0375$), though it narrows in the noisiest setting (34.5 vs. 44.9 at $\sigma = 0.075$). This indicates that the learned exploration policy selects more informative queries than uniform sampling, allowing `C-ICPE` to resolve the maximum value with fewer samples while still meeting the accuracy target. This difference in query counts also explains why `C-ICPE-Uniform+GP` attains higher accuracy than `C-ICPE-TD3+GP` (Table 15) since its post hoc GP is fit to more data.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.15$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.0375$	$\sigma = 0.075$
1	C-ICPE-TD3	0.0064 [0.0033,0.0094]	0.0070 [0.0058,0.0081]	0.0080 [0.0072,0.0088]	0.0060 [0.0036,0.0083]
	C-ICPE-uniform	0.0074 [0.0074,0.0075]	0.0061 [0.0059,0.0064]	0.0067 [0.0065,0.0069]	0.0053 [0.0052,0.0054]

Table 14: GP value estimation: final cost multiplier (mean and 95% CI across seeds) for every (ε, σ) configuration.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.15$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.0375$	$\sigma = 0.075$
1	C-ICPE-TD3	0.925 [.886,.957]	0.900 [.877,.919]	0.906 [.879,.930]	0.927 [.902,.947]
	C-ICPE-TD3+CF	0.893 [.861,.926]	0.902 [.888,.916]	0.871 [.855,.886]	0.859 [.831,.886]
	C-ICPE-uniform	0.924 [.913,.932]	0.923 [.913,.934]	0.913 [.895,.929]	0.915 [.902,.926]
	C-ICPE-TD3+GP	0.948 [.923,.968]	0.922 [.893,.944]	0.935 [.912,.953]	0.947 [.911,.973]
	C-ICPE-uniform+GP	0.965 [.955,.973]	0.960 [.952,.968]	0.971 [.964,.978]	0.969 [.962,.976]
	TPE	0.972 [.958,.986]	0.958 [.940,.976]	0.972 [.958,.986]	0.978 [.965,.991]
	CMA-ES	0.730 [.691,.769]	0.722 [.683,.761]	0.726 [.687,.765]	0.748 [.710,.786]
	GP-logEI	**0.984** [.973,.995]	**0.978** [.965,.991]	**0.986** [.976,.996]	**0.996** [.990,1.000]
	GP-UCB	0.902 [.876,.928]	0.890 [.863,.917]	0.900 [.874,.926]	0.890 [.863,.917]
	Uniform bin	0.178 [.144,.212]	0.176 [.143,.209]	0.094 [.068,.120]	0.204 [.169,.239]
	Uniform top 5%	0.616 [.573,.659]	0.612 [.569,.655]	0.620 [.577,.663]	0.728 [.689,.767]

Table 15: GP value estimation: accuracy (mean and 95% CI) for every (d, ε, σ) configuration. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

Stopping behavior. The survival function (left plot in Fig. 14) shows that `C-ICPE-TD3` stops earlier than `C-ICPE-Uniform`, with the bulk of episodes terminating between $t = 20$ and $t = 60$. Both methods exhibit gradual transitions rather than the sharp drops observed in binary search, reflecting the greater variability in task difficulty under the GP prior (functions with short lengthscales require more queries to resolve the peak value). The inference standard deviation (right plot in Fig. 14) starts high (≈ 0.8) and contracts rapidly in the first $20 - 40$ queries, then plateaus. This residual uncertainty is consistent with the difficulty of estimating a function's global maximum from noisy pointwise observations: even after localizing the region of high values, the precise peak height remains uncertain until sufficient samples accumulate near the optimum.

Certification. Table 17 reports the certification p-values for every (d, ε, σ) configuration. Both `C-ICPE-TD3` and `C-ICPE-Uniform` reject H_0 at $\alpha = 0.01$ for every seed and configuration, with the largest p-value across all seeds at 1.2×10^{-9} and mean p-values as low as 4.4×10^{-29}. The +GP variants yield far smaller p-values (10^{-41} to below 10^{-300}), reflecting their higher accuracy.

Critic-free stopping. For GP value estimation problem, removing the critic breaks correctness. `C-ICPE-TD3+CF` stops 22–38% earlier than `C-ICPE-TD3` (14.0–23.1 vs. 20.7–34.4 queries) and falls below the target in three of the four configurations, with accuracy as low as 0.859. It fails certification in every configuration (Table 17), whereas `C-ICPE-TD3` certifies everywhere. Here the inference network's belief over the maximum value is overconfident, so thresholding its mass on the ε-ball stops before enough evidence has been collected while the trained critic corrects for this.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.15$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.0375$	$\sigma = 0.075$
1	C-ICPE-TD3	22.3 [21.0,23.6]	20.7 [17.7,23.3]	24.5 [22.1,27.3]	34.4 [26.1,41.4]
	C-ICPE-TD3+CF	**14.0** [12.4,15.6]	**16.2** [14.7,17.7]	**15.2** [12.4,18.0]	**23.1** [18.3,27.8]
	C-ICPE-uniform	30.6 [29.7,31.4]	36.0 [35.4,36.6]	38.0 [36.9,39.2]	44.9 [43.5,46.4]
	C-ICPE-TD3+GP	22.1 [20.5,23.9]	19.9 [16.9,23.0]	23.5 [20.2,27.3]	34.3 [28.9,39.3]
	C-ICPE-uniform+GP	30.6 [29.9,31.2]	36.2 [35.4,37.0]	38.1 [36.6,39.7]	44.4 [43.6,45.3]

Table 16: GP value estimation: sample complexity (mean and 95% CI) for every (d, ε, σ) configuration. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.15$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.0375$	$\sigma = 0.075$
1	C-ICPE-TD3	2.4×10^{-28} [4.9×10^{-32},7.0×10^{-28}]	4.1×10^{-16} [1.8×10^{-18},1.2×10^{-15}]	2.9×10^{-11} [1.7×10^{-24},8.5×10^{-11}]	8.9×10^{-13} [4.4×10^{-16},2.6×10^{-12}]
	C-ICPE-TD3+CF	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	6.7×10^{-1} [2.8×10^{-2},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]	1.0×10^{0} [1.0×10^{0},1.0×10^{0}]
	C-ICPE-uniform	6.6×10^{-19} [2.5×10^{-20},1.8×10^{-18}]	4.4×10^{-29} [5.9×10^{-37},1.3×10^{-28}]	4.0×10^{-10} [8.6×10^{-15},1.2×10^{-9}]	3.3×10^{-12} [2.7×10^{-25},1.0×10^{-11}]
	C-ICPE-TD3+GP	1.3×10^{-103} [3.2×10^{-144},4.0×10^{-103}]	6.1×10^{-41} [7.3×10^{-65},1.8×10^{-40}]	1.0×10^{-60} [1.9×10^{-119},3.1×10^{-60}]	1.4×10^{-81} [1.6×10^{-138},4.3×10^{-81}]
	C-ICPE-uniform+GP	$\mathbf{6.1 \times 10^{-242}}$ [4.6×10^{-259},1.8×10^{-241}]	$\mathbf{3.1 \times 10^{-201}}$ [2.0×10^{-215},9.3×10^{-201}]	$\mathbf{2.5 \times 10^{-303}}$ [$< 10^{-300}$,7.6×10^{-303}]	$\mathbf{5.3 \times 10^{-286}}$ [1.3×10^{-298},1.2×10^{-285}]

Table 17: GP value estimation: certification p-value ($H_0 : p \leq 0.9$, one-sided exact binomial test on $n = 20000$ independent trials, significance level $\alpha = 0.01$) for every (d, ε, σ) configuration: mean across seeds, with the [min, max] range across seeds in brackets. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

E.3.5 MULTI-LOBE LOCALIZATION

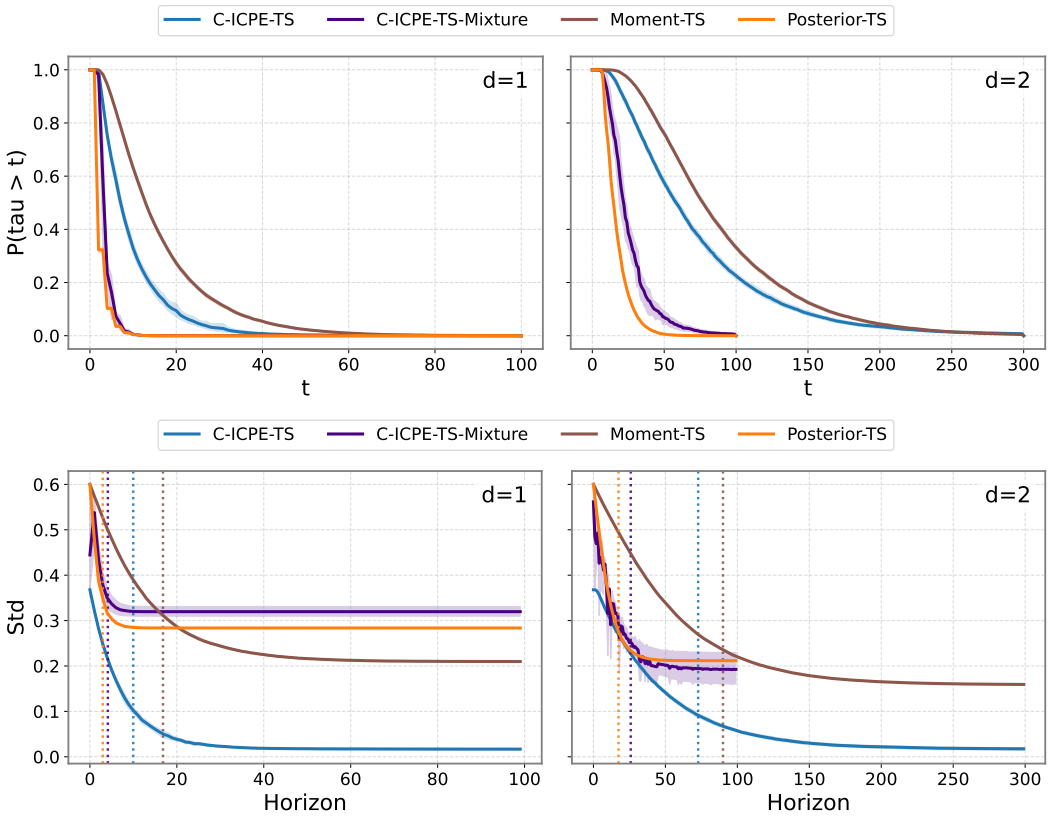


Figure 15: Results for Multi-Lobe localization with fixed confidence $\delta = 0.1$ and $N = 100$ across different dimensions at the hardest (ε, ℓ) setting: (a) survival function of τ; (b) inference uncertainty convergence.

Tables 19 and 20 report accuracy and sample complexity, and Fig. 15 shows the survival function and inference uncertainty convergence. As in the previous problems, we also evaluate critic-free ablations, both `C-ICPE-TS+CF` and `C-ICPE-TS-Mixture+CF`. Fig. 10 visualizes how the belief of each method evolves over the first few steps of a single episode at $d = 1$. We also report the cost multiplier at the end of training for every (d, ℓ) configuration in Table 18.

Accuracy. Table 19 reports accuracy for every (d, ℓ) configuration. At $d = 1$, all four methods clear the target (0.911–0.969), and accuracy does not degrade as ℓ shrinks from 0.3 to 0.15. With only two well-separated lobes, even collapsing the belief to a single Gaussian (`C-ICPE-TS`, `Moment-TS`) preserves enough information to identify the correct lobe. At $d = 2$, this no longer holds within the default budget. At $\ell = 0.15$, `C-ICPE-TS` and `Moment-TS` reach the target only when the maximum horizon is tripled to $t_{\max} = 300$ (Table 1). At $\ell = 0.3$, `Moment-TS` falls just short of the target (0.896). In contrast, the two methods that retain a multimodal belief meet the target in every configuration with the default $t_{\max} = 100$: at $d = 2$, $\ell = 0.15$, `C-ICPE-TS-Mixture` achieves 0.919 and `Posterior-TS` 0.921. A diagonal single-Gaussian belief is thus not merely a worse approximation of a multimodal posterior, as the number of competing modes grows and they become harder to separate, it requires a substantially larger query budget to identify the correct one.

Sample complexity. Table 20 shows that sample complexity grows with dimension for every method, and grows as ℓ decreases for every method except `Posterior-TS`. `Posterior-TS` makes 3.0 queries at both ℓ values in $d = 1$ and 17.4 at both in $d = 2$. Since it always queries near the center of its currently most probable lobe, each query stays within $w = 0.02$ of the true hypothesis, so the resulting observation remains close to maximally informative regardless of ℓ. The

d	Method	$\ell = 0.3$, $m = 0.6$	$\ell = 0.15$, $m = 0.6$
1	C-ICPE-TS	0.0297 [0.0245,0.0349]	0.0049 [-0.0012,0.0111]
	C-ICPE-TS-Mixture	0.0526 [0.0316,0.0736]	0.0883 [0.0824,0.0942]
2	C-ICPE-TS	0.0000 [0.0000,0.0000]	0.0000 [0.0000,0.0000]
	C-ICPE-TS-Mixture	0.0004 [0.0000,0.0009]	0.0009 [-0.0007,0.0025]

Table 18: Multi-lobe localization: final cost multiplier (mean and 95% CI across seeds) for every (d, ℓ) configuration.

single-Gaussian methods are far more sensitive to ℓ, since their queries are centered at the belief mean rather than at a specific lobe: in $d = 1$, halving ℓ nearly doubles the queries of `Moment-TS` (8.8 to 16.8) and `C-ICPE-TS` (5.6 to 10.0), and in $d = 2$ `Moment-TS` again doubles (45.3 to 90.0). `C-ICPE-TS-Mixture`, which can represent each lobe directly and query near the favored one, behaves much like Posterior-TS. Its queries increase by only 17–20% as ℓ decreases, and it is the closest method to Posterior-TS in every configuration. At $d = 2$, $\ell = 0.15$, it requires 25.9 queries, compared to 72.8 for `C-ICPE-TS` and 90.0 for `Moment-TS`. This is the central result of the multi-lobe task: all methods reach the target accuracy, but a multimodal belief does so with 3–3.5$\times$ fewer queries than a single Gaussian, approaching the efficiency of the exact posterior.

Stopping behavior. The survival functions (Fig. 15, top) mirror the sample-complexity results. At $d = 1$, `C-ICPE-TS-Mixture` and `Posterior-TS` stop almost immediately, with nearly all episodes terminating by $t \approx 10$; `C-ICPE-TS` follows more slowly, and `Moment-TS` retains a long tail past $t = 40$. At $d = 2$, the gap widens, `C-ICPE-TS` and `Moment-TS` decay gradually, with roughly 20% of episodes still running at $t = 100$ and some continuing past $t = 250$. `C-ICPE-TS-Mixture` closely mirrows `Posterior-TS`'s survival fuction with the default horizon. The inference standard deviation (bottom) shows a qualitative difference between the two belief types. The multimodal methods plateau quickly and at a higher level (≈ 0.28–0.32 at $d = 1$ and ≈ 0.2 at $d = 2$), since they retain mass on competing lobes. The single-Gaussian methods instead contract slowly toward much lower values, as the belief commits to a single region, and they stop while their standard deviation is still decreasing.

Certification. Table 21 reports the certification p-values for every (d, ℓ) configuration. At $d = 1$, all four methods reject H_0 for every seed, with $p \leq 3.0 \times 10^{-6}$ throughout, consistent with their accuracy margins in Table 19. At $d = 2$, the only certification failure is `Moment-TS` at $\ell = 0.3$ ($p = 0.15$), where no seed rejects H_0 at $\alpha = 0.01$, consistent with its below-target accuracy of 0.896. All other methods certify in every configuration.

Critic-free stopping. The two ablations behave very differently as it strongly depends on how well the belief represents the posterior. `C-ICPE-TS+CF` falls below the target in every configuration (0.839–0.854), with confidence intervals entirely below 0.90, while stopping 20–25% earlier than `C-ICPE-TS` (58.0 vs. 72.8 queries at $d = 2$, $\ell = 0.15$). A single Gaussian struggles to represent both lobes, so its belief concentrates on one of them and becomes overconfident; thresholding its mass on the ε-ball therefore stops before the correct lobe has been identified. Certification test confirms this, `C-ICPE-TS+CF` fails in every configuration (Table 21). `C-ICPE-TS-Mixture+CF`, whose belief can place mass on each lobe, remains much closer to its critic-based counterpart: it exceeds the target at $d = 1$ (0.942–0.944) and is borderline at $d = 2$ (0.901 and 0.898), with sample complexity within about one query of `C-ICPE-TS-Mixture`. It certifies for every seed at $d = 1$, but at $d = 2$ only some seeds reject H_0, whereas both `C-ICPE-TS` and `C-ICPE-TS-Mixture` certify in every configuration for every seed. The critic is thus most important when the belief is misspecified, and it still provides a margin of safety when the belief is expressive enough to represent the posterior.

d	Method	$\ell = 0.3$ $m = 6$	$\ell = 0.15$ $m = 6$
1	C-ICPE-TS	0.915 [.908,.923]	0.915 [.903,.925]
	C-ICPE-TS+CF	0.839 [.828,.850]	0.843 [.838,.847]
	C-ICPE-TS-Mixture	0.911 [.903,.918]	0.919 [.911,.926]
	C-ICPE-TS-Mixture+CF	0.944 [.940,.947]	0.942 [.936,.948]
	Moment-TS	**0.969** [.966,.972]	**0.967** [.963,.969]
	Posterior-TS	0.942 [.938,.945]	0.941 [.937,.945]
2	C-ICPE-TS	0.921 [.909,.932]	0.915 [.906,.921]
	C-ICPE-TS+CF	0.851 [.833,.869]	0.854 [.853,.855]
	C-ICPE-TS-Mixture	0.912 [.903,.923]	0.919 [.905,.931]
	C-ICPE-TS-Mixture+CF	0.901 [.869,.933]	0.898 [.880,.916]
	Moment-TS	0.896 [.888,.904]	**0.943** [.938,.946]
	Posterior-TS	**0.921** [.914,.928]	0.921 [.913,.928]

Table 19: Multi-lobe localization: accuracy (mean and 95% CI) for every (d, ℓ) configuration. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

d	Method	$\ell = 0.3$ $m = 6$	$\ell = 0.15$ $m = 6$
1	C-ICPE-TS	5.6 [5.4,5.8]	10.0 [9.4,10.6]
	C-ICPE-TS+CF	4.2 [4.1,4.3]	7.7 [7.2,8.2]
	C-ICPE-TS-Mixture	3.5 [3.2,3.8]	4.1 [3.6,4.4]
	C-ICPE-TS-Mixture+CF	4.5 [3.5,5.5]	4.6 [4.0,5.2]
	Moment-TS	8.8 [8.7,8.9]	16.8 [16.6,17.0]
	Posterior-TS	**3.0** [2.9,3.0]	**3.0** [2.9,3.0]
2	C-ICPE-TS	55.2 [41.7,70.2]	72.8 [71.1,74.7]
	C-ICPE-TS+CF	45.0 [13.8,76.1]	58.0 [57.1,59.0]
	C-ICPE-TS-Mixture	21.6 [17.8,28.0]	25.9 [22.6,28.5]
	C-ICPE-TS-Mixture+CF	20.4 [17.3,23.5]	24.5 [22.2,26.8]
	Moment-TS	45.3 [44.8,45.7]	90.0 [89.2,91.1]
	Posterior-TS	**17.4** [17.2,17.6]	**17.4** [17.3,17.6]

Table 20: Multi-lobe localization: sample complexity (mean and 95% CI) for every (d, ℓ) configuration. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

d	Method	$\ell = 0.3$ $m = 6$	$\ell = 0.15$ $m = 6$
1	C-ICPE-TS	9.9×10^{-7} $[8.6 \times 10^{-15}, 3.0 \times 10^{-6}]$	1.5×10^{-11} $[1.3 \times 10^{-22}, 4.5 \times 10^{-11}]$
	C-ICPE-TS+CF	1.0×10^{0} $[1.0 \times 10^{0}, 1.0 \times 10^{0}]$	1.0×10^{0} $[1.0 \times 10^{0}, 1.0 \times 10^{0}]$
	C-ICPE-TS-Mixture	2.4×10^{-9} $[3.6 \times 10^{-13}, 7.0 \times 10^{-9}]$	1.6×10^{-10} $[4.3 \times 10^{-18}, 4.8 \times 10^{-10}]$
	C-ICPE-TS-Mixture+CF	3.0×10^{-129} $[2.9 \times 10^{-143}, 9.0 \times 10^{-129}]$	2.0×10^{-89} $[8.7 \times 10^{-116}, 5.9 \times 10^{-89}]$
	Moment-TS	$\mathbf{3.2 \times 10^{-290}}$ $[2.2 \times 10^{-321}, 9.6 \times 10^{-290}]$	$\mathbf{2.5 \times 10^{-264}}$ $[2.7 \times 10^{-291}, 7.6 \times 10^{-264}]$
	Posterior-TS	6.6×10^{-90} $[5.9 \times 10^{-105}, 2.0 \times 10^{-89}]$	3.0×10^{-88} $[1.2 \times 10^{-103}, 9.1 \times 10^{-88}]$
2	C-ICPE-TS	2.7×10^{-22} $[7.9 \times 10^{-36}, 5.5 \times 10^{-22}]$	8.6×10^{-12} $[1.7 \times 10^{-24}, 2.0 \times 10^{-11}]$
	C-ICPE-TS+CF	1.0×10^{0} $[1.0 \times 10^{0}, 1.0 \times 10^{0}]$	1.0×10^{0} $[1.0 \times 10^{0}, 1.0 \times 10^{0}]$
	C-ICPE-TS-Mixture	1.1×10^{-5} $[3.5 \times 10^{-18}, 3.2 \times 10^{-5}]$	2.6×10^{-12} $[3.4 \times 10^{-80}, 7.2 \times 10^{-12}]$
	C-ICPE-TS-Mixture+CF	3.3×10^{-1} $[1.7 \times 10^{-23}, 1.0 \times 10^{0}]$	6.6×10^{-1} $[1.9 \times 10^{-10}, 1.0 \times 10^{0}]$
	Moment-TS	1.5×10^{-1} $[1.7 \times 10^{-2}, 3.5 \times 10^{-1}]$	$\mathbf{3.3 \times 10^{-101}}$ $[4.5 \times 10^{-107}, 8.3 \times 10^{-101}]$
	Posterior-TS	$\mathbf{1.4 \times 10^{-22}}$ $[1.1 \times 10^{-26}, 4.3 \times 10^{-22}]$	4.2×10^{-23} $[6.4 \times 10^{-27}, 1.3 \times 10^{-22}]$

Table 21: Multi-lobe localization: certification p-value ($H_0 : p \leq 0.9$, one-sided exact binomial test on $n = 20000$ independent trials, significance level $\alpha = 0.01$) for every (d, ℓ) configuration: mean across seeds, with the [min, max] range across seeds in brackets. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

E.4 Synthetic Benchmarks: robustness

When training C−ICPE, the task prior ν is uniform over the parameter space. We investigate robustness to prior misspecification by evaluating frozen C−ICPE models under Beta(α, β) deployment priors with varying concentration parameters. Figs. 16 to 20 report accuracy and confidence intervals for all actor variants at the most challenging (ε, σ) configuration and highest dimensionality per benchmark.

E.4.1 Noisy Binary Search

The training prior is $\mathrm{Uniform}[-1, 1]^{20}$, corresponding to $\alpha = \beta = 1$ (white box). In Fig. 16 we report the results. C−ICPE−TS, C−ICPE−TTPS, and C−ICPE−TS−Mixture are robust across the full grid of Beta priors: accuracy remains above 0.85 in all configurations, even under substantial distributional shift ($\alpha = 0.5, \beta = 7$ or vice versa). Performance improves mildly when both $\alpha, \beta \geq 3$, since concentrated priors place more mass in the interior of $[-1, 1]^d$, where localization is easier. Notably, C−ICPE−TS−Mixture perform better at $\beta = 0.5$, especially when $\alpha \geq 3$, C−ICPE−TS and C−ICPE−TTPS drops below 0.9 while C−ICPE−TS−Mixture maintains accuracy ≥ 0.93. C−ICPE−TD3 matches or exceeds the other actors when the deployment prior is concentrated ($\alpha, \beta \geq 3$, reaching 0.963), but degrades sharply when either parameter is small: accuracy drops to 0.487 ($\alpha = 0.5, \beta = 7$) or 0.370 ($\alpha = 7, \beta = 0.5$). Small α or β produces a U-shaped Beta distribution that concentrates mass near the boundary of the domain, where targets are harder to disambiguate and the learned actor generalizes poorly. C−ICPE−Uniform fails entirely in $d = 20$ (accuracy < 0.01), confirming that passive exploration cannot localize a target in high-dimensional binary search within the allowed horizon.

E.4.2 ε-best-arm problem

Fig. 17 reports accuracy at $d = 15$ across the full Beta prior grid. Robustness varies considerably across actors. C−ICPE−RoundRobin attains perfect accuracy in every configuration, and C−ICPE−TD3 is nearly unaffected by misspecification, with accuracy between 0.921 and 0.936. The structure of the problem helps explain this stability, since the loss $L_\theta(x) = 1 - \theta^\top x$ is invariant to orthogonal transformations, the intrinsic difficulty of a task instance does not depend on the direction of θ, unlike binary search or Ackley, where boundary effects make difficulty heterogeneous across the parameter space. C−ICPE−TS, C−ICPE−TTPS, and C−ICPE−TS−Mixture are moderately sensitive, with accuracy ranging from roughly 0.89 to 0.96, and fall slightly below the 0.9 target in a few configurations: C−ICPE−TS and C−ICPE−TTPS when both α and β are large, and C−ICPE−TS−Mixture when α is large and β is small. C−ICPE−Uniform is the least robust, degrading from 0.920 under the matched prior to 0.795 at $(\alpha, \beta) = (7, 7)$. Since the problem's difficulty is direction-independent, these differences likely stem from the shift in the distribution of θ relative to the training prior rather than from harder task instances, and a fixed, basis-aligned query policy such as round-robin is the least affected by this shift.

E.4.3 Ackley minimization

Results are reported in Fig. 18. The Ackley benchmark at $d = 5$ exhibits the strongest sensitivity to prior misspecification among all tasks. C−ICPE−TS, C−ICPE−TTPS, C−ICPE−TD3, and C−ICPE−TS−Mixture exhibits similar pattern on the robustness. At the training prior ($\alpha = \beta = 1$), they achieve $\approx$ 0.91–0.92. When both $\alpha, \beta \geq 3$, i.e., the deployment prior concentrates mass toward the interior of $[-1, 1]^d$, all active methods improve substantially, where accruacy icnreases to ≈ 0.98. The gains reflect the structure of the Ackley function, it targets near the center of the domain sit in a region of higher curvature where observations are more informative, making identification easier.

Conversely, when either α is small (0.5), the Beta prior becomes U-shaped or boundary-skewed, placing significant mass on targets near the edges of $[-1, 1]^d$. In the flat outer region of the Ackley function, observations carry little signal, and the learned policies, trained under a uniform prior that rarely produces such extreme configurations, accuracy degrades to 0.521–0.656. The effect is most pronounced for $\alpha = 0.5, \beta = 7$. This asymmetry between interior and boundary targets is specific to the Ackley geometry and is absent in the rotationally symmetric ε-best arm problem.

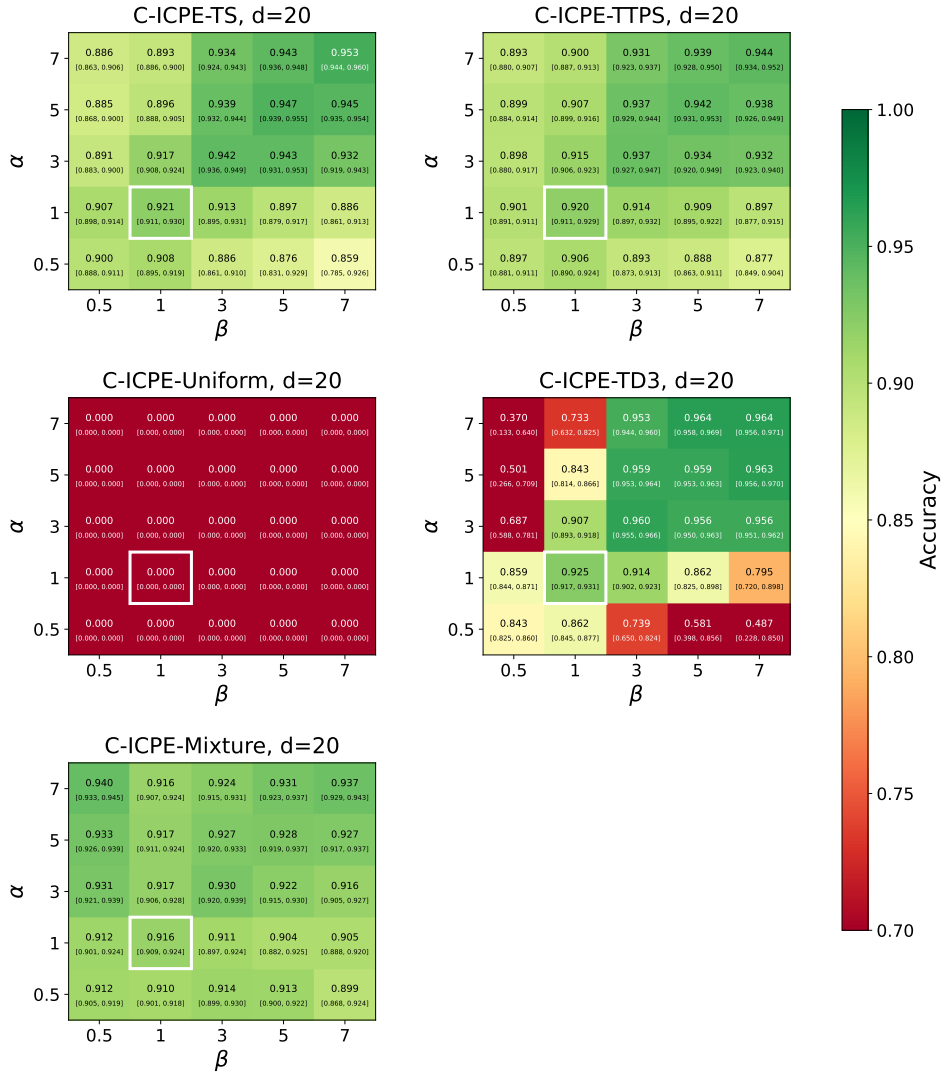


Figure 16: Robustness to prior misspecification on the 20D noisy binary search problem ($\varepsilon = 0.1$ and $\sigma = 0.05$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.

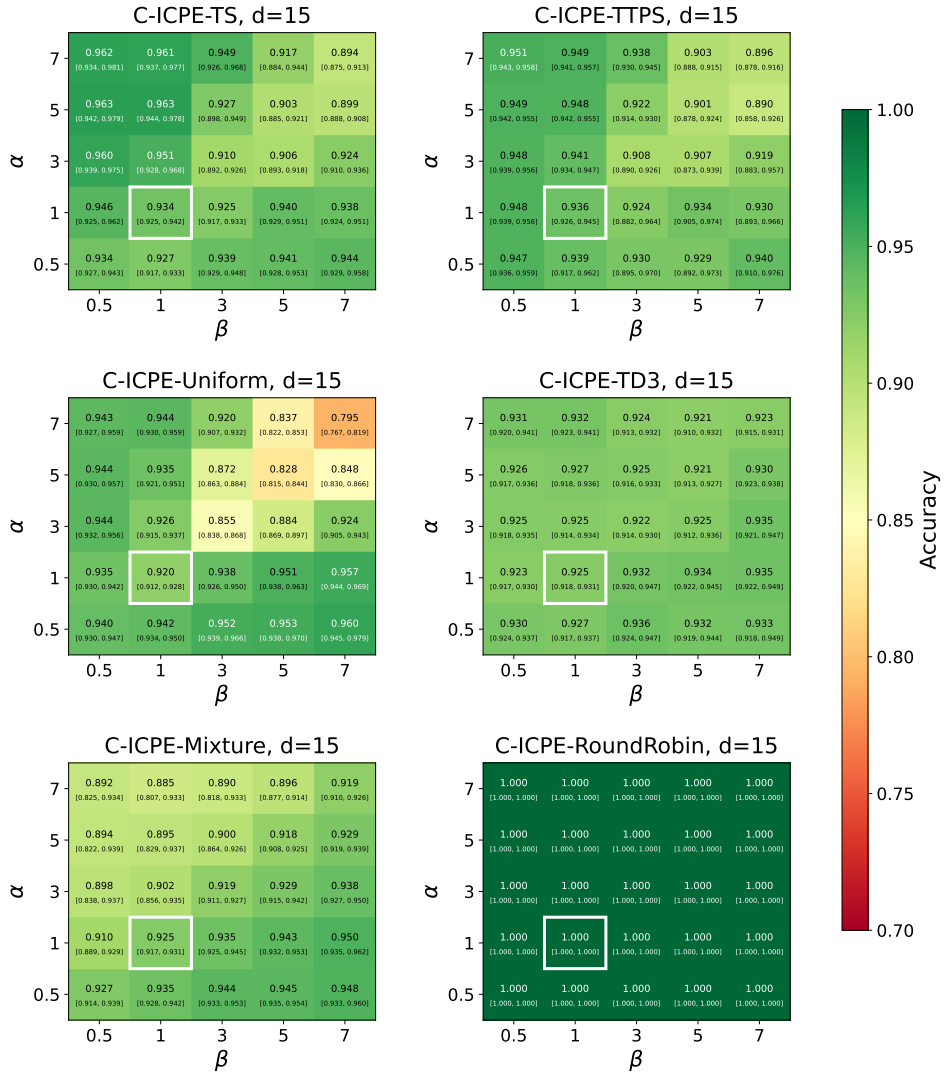


Figure 17: Robustness to prior misspecification on the 15D ε-best-arm identification problem ($\varepsilon = 0.005$ and $\sigma = 0.0025$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.

C-ICPE-Uniform fails across the board (accuracy ≤ 0.544), confirming that directed exploration is essential for this multimodal benchmark regardless of the prior.

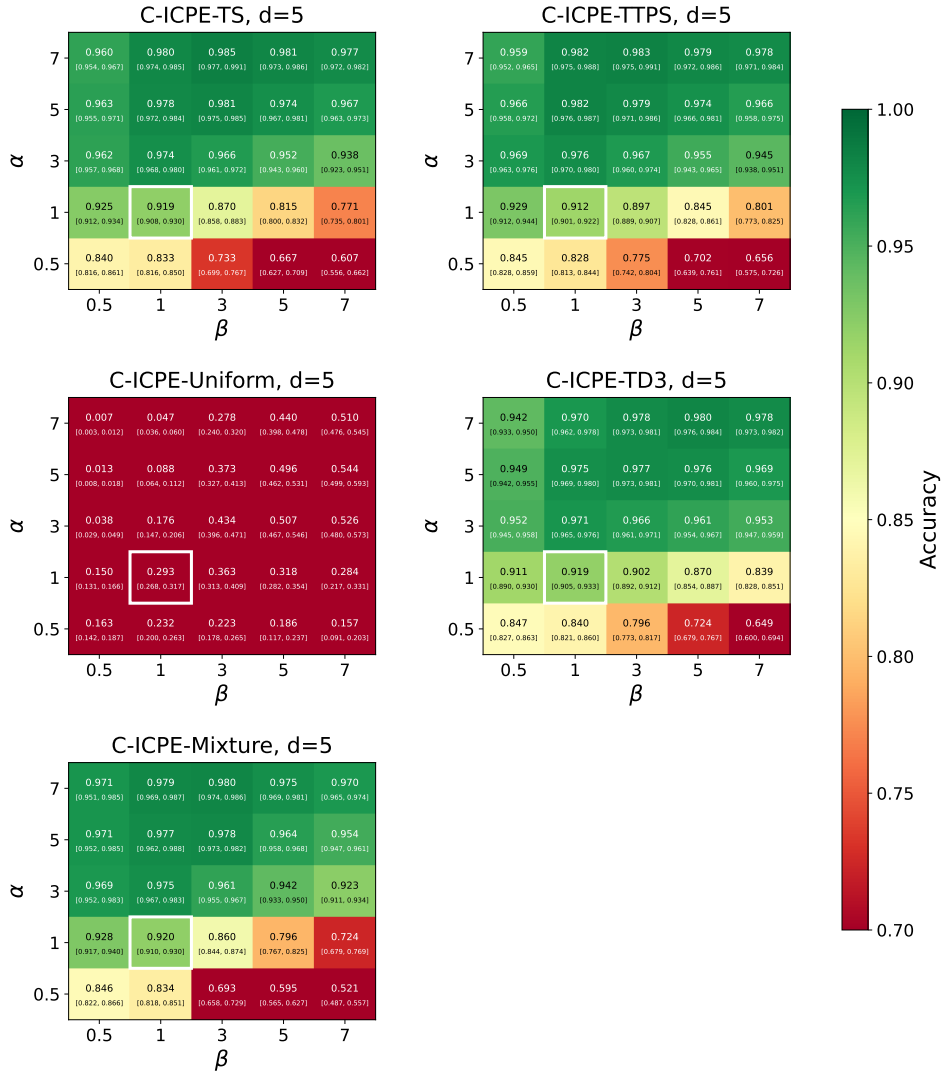


Figure 18: Robustness to prior misspecification on the 5D Ackley function ($\varepsilon = 0.1$ and $\sigma = 0.05$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.

E.4.4 GP MAX-VALUE ESTIMATION

We report results in Fig. 19. This is the $\mathcal{X} \neq \mathcal{A}$ setting ($d = 1$), so only C-ICPE-Uniform and C-ICPE-TD3 are applicable. C-ICPE-TD3 achieves the highest accuracy when the deployment prior is well-matched or concentrated with small β: it reaches 0.972 at $(\alpha, \beta) = (7, 0.5)$ and remains above 0.947 throughout the upper-left triangle of the grid. However, it degrades when β is large (0.770 at $\alpha = 0.5, \beta = 7$; 0.791 at $\alpha = 1, \beta = 7$), indicating that the learned exploration policy is sensitive to prior shifts that alter the distribution of task difficulty. C-ICPE-Uniform, by contrast, is more robust: accuracy stays between 0.835 and 0.958 across the entire grid, with a milder gradient from upper-left to lower-right. Because the uniform actor does not depend on a learned exploration policy, its performance varies only through the stopping criterion and inference model, both of which appear stable under moderate prior shift. At the training prior ($\alpha = \beta = 1$), the two methods are

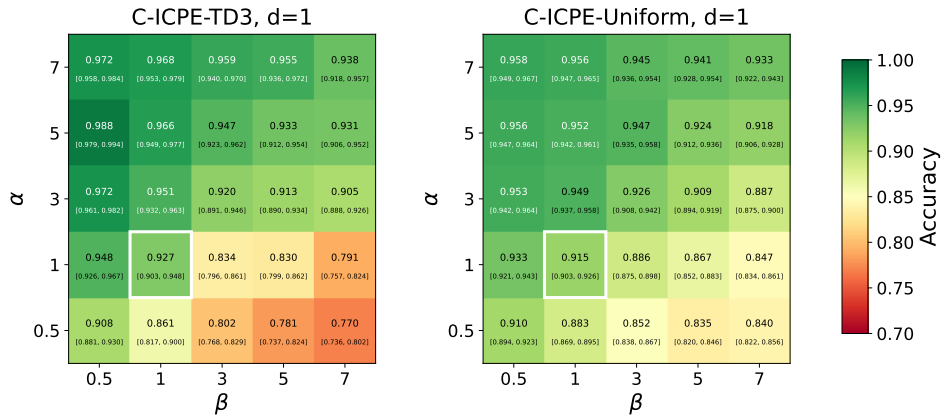


Figure 19: Robustness to prior misspecification on the GP value estimation problem ($\varepsilon = 0.15$ and $\sigma = 0.075$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.

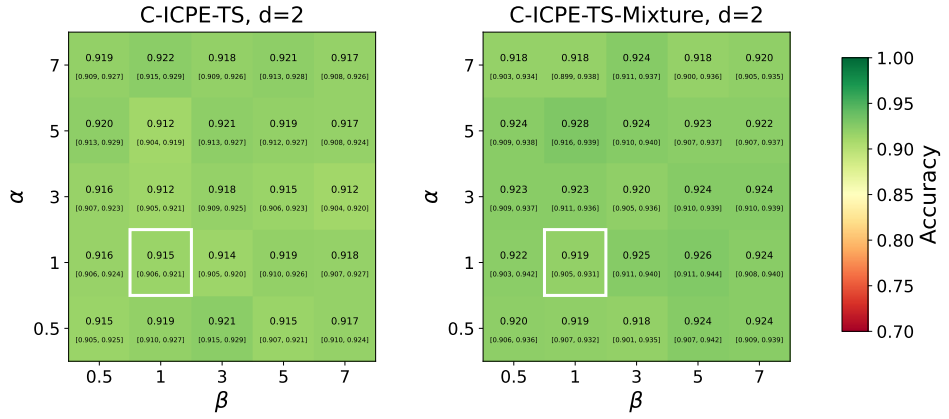


Figure 20: Robustness to prior misspecification on the Multi-Lobe localization problem ($\varepsilon = 0.08$ and $\ell = 0.15$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.

comparable (0.927 vs. 0.915), but `C-ICPE-TD3` offers a clear advantage when the deployment prior concentrates mass on tasks where directed exploration helps most ($\alpha \geq 3$, $\beta \leq 1$). The overall pattern suggests that the TD3 actor learns an exploration strategy well-adapted to the training distribution but with limited extrapolation to deployment priors that shift the typical task structure.

E.4.5 Multi-Lobe Localization

Results are reported in Fig. 20. Unlike Ackley, multi-lobe localization shows essentially no sensitivity to prior misspecification for either architecture. With larger horizon $t_{max} = 300$, `C-ICPE-TS` flat between 0.912 and 0.922 and `C-ICPE-TS-Mixture` flat between 0.918 and 0.928 across the grid, both matching their accuracy at the trained prior ($\alpha = \beta = 1$) almost exactly. This follows directly from the environment's structure where the observation model $g_{s,\ell}(a)$ depends only on the lobe sign S, never on the within-lobe offset U, so which lobe is identifiable is entirely independent of how U is distributed. Therefore, misspecifying its prior cannot make the localization problem itself any easier or harder.

E.5 Geochemical Exploration: Experimental Details and Numerical Results

The geochemical exploration task is a stylized version of a real problem in mineral exploration: given a budget of field samples (each requiring physical collection, transport, and laboratory analysis), identify the most promising location for further investigation. In practice, each sample costs hundreds to thousands of dollars, and field campaigns are logistically constrained. A method that can identify the target location with (ϵ, δ)-guarantees while minimizing the number of samples has direct economic value.

E.5.1 Dataset and Motivation

We use data from the USGS National Geochemical Survey (usg, 2004), which provides soil and sediment measurements of element concentrations across the continental United States. We focus on copper (Cu) concentrations, which are of direct interest in mineral exploration: copper deposits are spatially heterogeneous, and identifying regions of peak concentration from sparse, noisy field measurements is a costly sequential decision problem.

The dataset contains point measurements at irregularly spaced locations, each reporting the concentration of multiple elements. We extract copper concentration values and apply a log-transform followed by z-score normalization per region, yielding standardized log-concentrations that serve as observations.

As a concrete example, Fig. 22 shows the Kingman region in southeastern California and southern Nevada. This region contains the Mountain Pass rare earth mine (35.5°N, 115.5°W), an open-pit mine of rare earth elements in the Mojave Desert. A satellite image of the mine and surrounding terrain, acquired by the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) instrument on NASA's Terra spacecraft on March 28, 2010, is shown in Fig. 22 (NASA/GSFC/METI/ERSDAC/JAROS, and U.S./Japan ASTER Science Team, 2010). The region exhibits spatially varying copper concentrations with a clear peak near the mining district, making it a representative example of the kind of localization problem `C-ICPE` is designed to solve.

E.5.2 Region Partitioning and GP Fitting

We partition the geochemical survey data into 17 geographic regions, each covering approximately $1° \times 2°$ in latitude and longitude. The regions span diverse geological settings across the western and southeastern United States: Gadsden, Bozeman, Billings, Wells, Needles, Jenkins, Montgomery, Millett, Prescott, Lovelock, Aurora, Holbrook, Atlanta, Ely, Kingman, Winnemucca, and Baker. Fig. 23 shows all 17 regions with sample locations colored by normalized log-copper concentration. The regions vary substantially in sample density (from $\sim$50 to $\sim$500 measurements), spatial structure, and concentration range, providing a diverse task distribution for meta-training.

For each region, we fit a sparse variational Gaussian process (SVGP) with a Matérn-3/2 ARD kernel and Gaussian likelihood. The model is:

$$f \sim \mathrm{GP}(0, \sigma_f^2\, k_{\text{Matérn-3/2}}(\cdot, \cdot\,; \ell_1, \ell_2)), \qquad y_i = f(\mathbf{s}_i), \tag{59}$$

where $\mathbf{s}_i \in \mathbb{R}^2$ are UTM coordinates (normalized to $[0,1]^2$ for numerical stability), y_i is the standardized log-copper concentration, and (ℓ_1, ℓ_2) are per-region hyperparameters learned by maximizing the variational evidence lower bound (ELBO). We use $M = 500$ inducing points initialized via k-means clustering of the observation locations, and optimize for 1000 Adam iterations at learning rate 0.01.

Fig. 24 shows the ELBO training curves for all 17 regions. All regions converge smoothly, confirming that the SVGP fits are well-behaved. The fitted hyperparameters, in particular the lengthscales (ℓ_1, ℓ_2), vary across regions, reflecting different spatial correlation structures: some regions exhibit short-range variability (small lengthscales) while others have smoother concentration surfaces (large lengthscales).

E.5.3 Ground Truth Construction

The fitted hyperparameters define a distribution over surfaces for each region. Each episode draws a sample g from the GP prior in Eq. (59) with the region's lengthscales scaled by a factor of 3,

and rescales it to $[0, 1]$ to obtain the task surface f. The ground truth target is the location of peak concentration,

$$\theta^\star = \arg\max_{x \in [0,1]^2} f(x), \tag{60}$$

so that $f(\theta^\star) = 1$. Although the agent recommends a location, correctness is measured by the value at that location rather than its distance to $\theta^\star$:

$$L_\theta(x) = f(\theta^\star) - f(x) = 1 - f(x), \tag{61}$$

and a recommendation $\hat{x}$ is ε-correct if $L_\theta(\hat{x}) \leq \varepsilon$. Unlike the distance-based loss used for Ackley, several spatially separated locations can therefore be ε-correct, as shown in Fig. 21.

E.5.4 Task Prior and Train/Test Split

The 17 regions are split into 12 training and 5 evaluation regions. Training regions define the task prior ν: during meta-training, each episode samples a region uniformly from the training set and presents `C-ICPE` with a surface sampled from that region's fitted GP prior (see above). The agent queries locations $a \in [0, 1]^2$ and observes evaluations $y = f(a)$. The goal is to recommend a location $\hat{x}$ with $L_\theta(\hat{x}) \leq \varepsilon$ with probability at least $1 - \delta$.

Evaluation is performed on held-out regions whose spatial structure, lengthscales, and concentration patterns were not seen during training. This tests two properties simultaneously:

1. **Bayesian (ϵ, δ)-correct identification on realistic functions.** The task surfaces are spatially structured and vary in smoothness.
2. **Robustness to distribution shift.** The evaluation regions have different hyperparameters (ℓ_1, ℓ_2) from the training regions, so the agent must generalize across spatial correlation structures it has not encountered during meta-training.

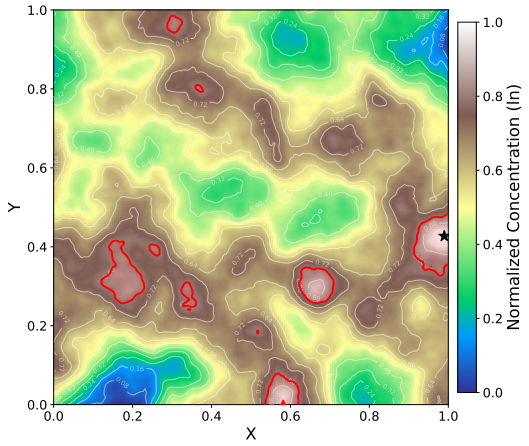


Figure 21: Example copper concentration in a 2D region in the geochemical exploration task. Red regions indicate concentration of copper within ϵ of the maximum value.

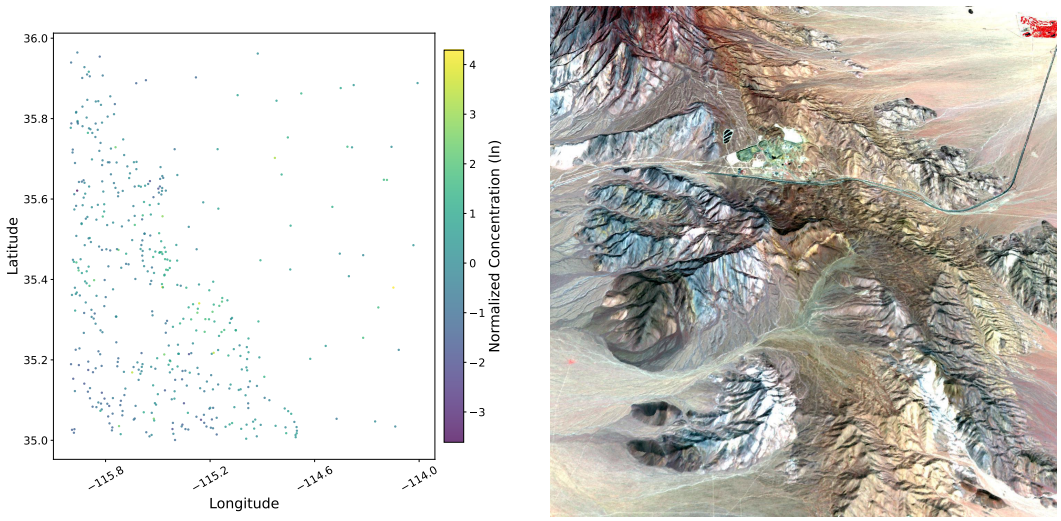


Figure 22: **Left:** Normalized log-copper concentration in the Kingman region (southeastern California / southern Nevada). Each point is a soil sample from the USGS National Geochemical Survey (usg, 2004); the red star marks the location of peak GP posterior mean. **Right:** ASTER satellite image of the Mountain Pass rare earth mine (35.5°N, 115.5°W) within this region, acquired March 28, 2010. The mine area is visible as the light-colored open pit in the upper center of the image. Credit: NASA/GSFC/METI/ERSDAC/JAROS, and U.S./Japan ASTER Science Team (NASA/GSFC/METI/ERSDAC/JAROS, and U.S./Japan ASTER Science Team, 2010).

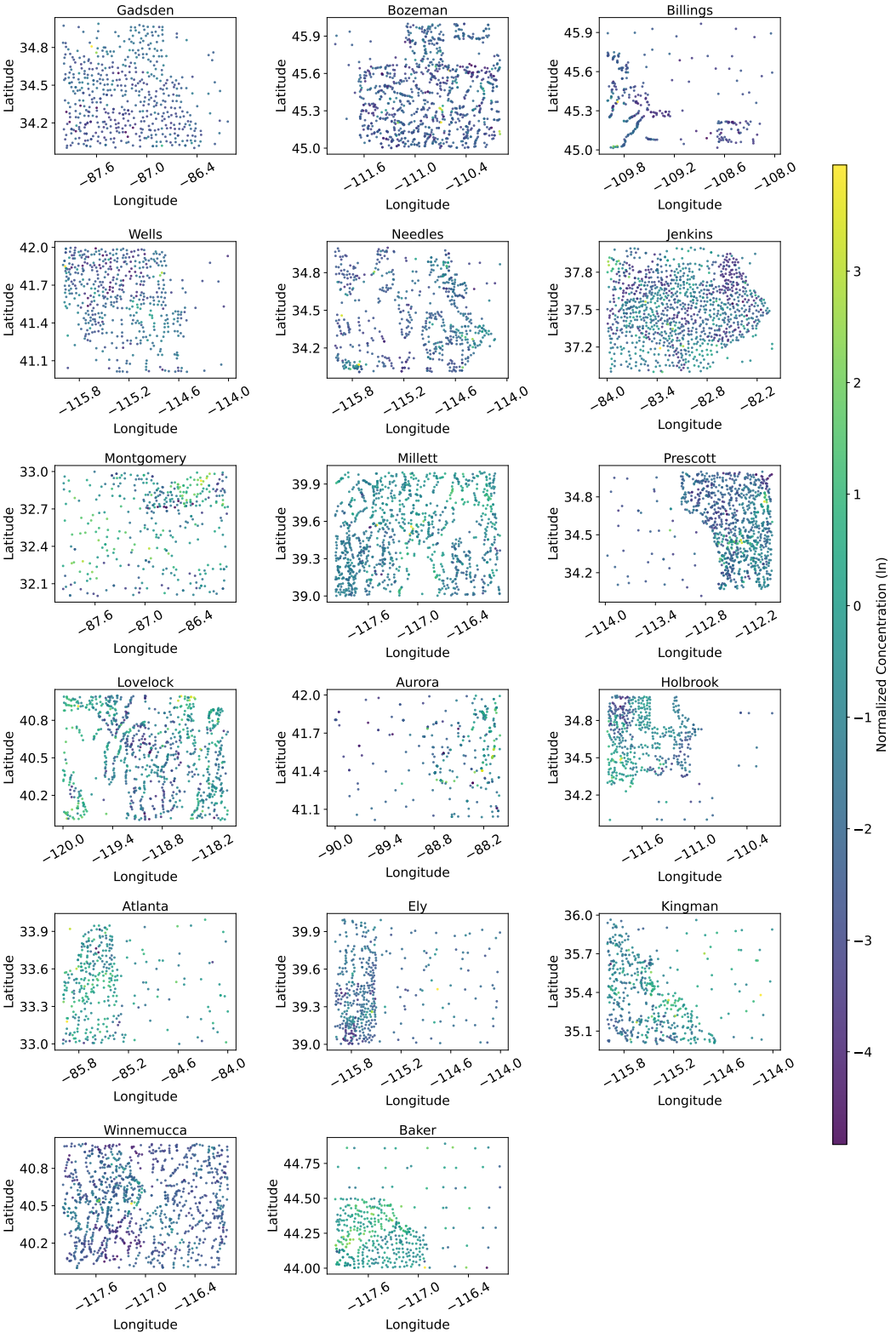


Figure 23: All 17 geographic regions used in the geochemical experiment. Each panel shows soil sample locations colored by normalized log-copper concentration. Regions are split into 12 training and 5 evaluation regions; evaluation regions have spatial structure not seen during meta-training.

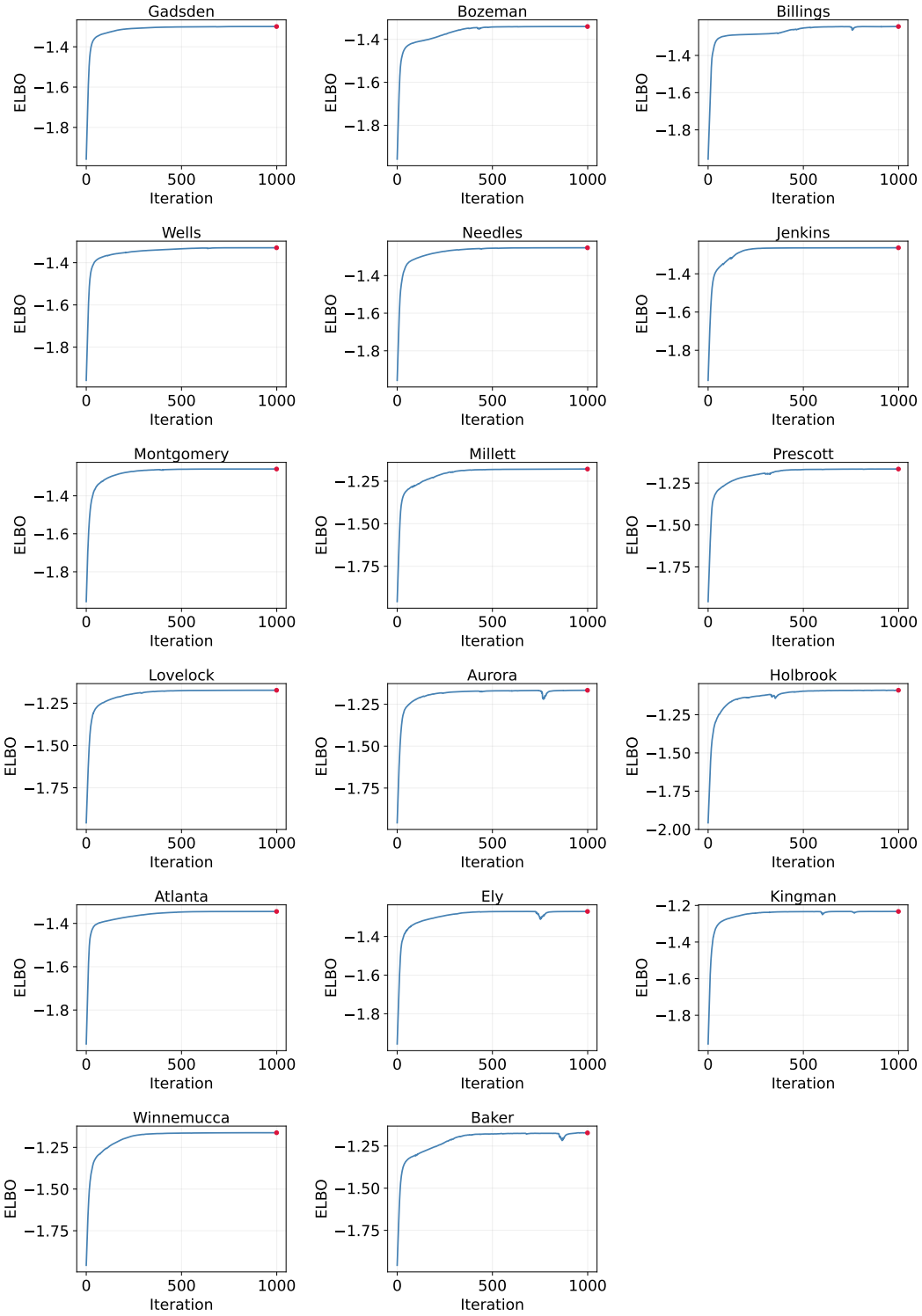


Figure 24: ELBO training curves for sparse variational GP fits across all 17 regions (1000 Adam iterations, $M = 500$ inducing points). All regions converge smoothly, indicating well-behaved GP fits.

d	Method	$\varepsilon = 0.2$ $\sigma = 0$	$\varepsilon = 0.15$ $\sigma = 0$
2	C-ICPE-TD3	0.0089 [0.0087,0.0091]	0.0052 [0.0047,0.0057]
	C-ICPE-TS	0.0098 [0.0094,0.0102]	0.0062 [0.0059,0.0065]
	C-ICPE-TS-Mixture	0.0171 [0.0152,0.0190]	0.0145 [0.0141,0.0148]

Table 22: Geochem: final cost multiplier (mean and 95% CI across seeds) for every ε configuration.

d	Method	$\varepsilon = 0.2$ $\sigma = 0$	$\varepsilon = 0.15$ $\sigma = 0$
2	C-ICPE-TD3	0.920 [.910,.930]	0.920 [.910,.931]
	C-ICPE-TS	0.921 [.911,.928]	0.919 [.906,.931]
	C-ICPE-TS+CF	0.966 [.958,.974]	0.953 [.952,.954]
	C-ICPE-TS-Mixture	0.920 [.908,.931]	0.925 [.916,.935]
	TPE	0.436 [.392,.480]	0.462 [.418,.506]
	CMA-ES	0.488 [.444,.532]	0.472 [.428,.516]
	GP-logEI	0.566 [.523,.609]	0.648 [.606,.690]
	GP-UCB	0.722 [.683,.761]	0.700 [.660,.740]
	TPE$_{T=100}$	0.942 [.928,.956]	0.890 [.871,.909]
	CMA-ES$_{T=100}$	0.894 [.875,.913]	0.764 [.738,.790]
	GP-logEI$_{T=100}$	**0.983** [.975,.991]	**0.980** [.971,.989]
	GP-UCB$_{T=100}$	0.866 [.845,.887]	0.791 [.765,.816]

Table 23: Geochem: accuracy (mean and 95% CI) for every (d, ε, σ) configuration. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. Unless marked $T = 100$, baselines use a budget equal to the mean stopping time of the most sample-efficient `C-ICPE` variant, rounded up. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

E.5.5 Numerical Results

Baselines. We compare `C-ICPE-TD3`, `C-ICPE-TS`, and `C-ICPE-TS-Mixture` against four black-box optimization baselines: `TPE` Bergstra et al. (2011), `CMA-ES` Hansen (2016), `GP-logEI` Ament et al. (2023), and `GP-UCB` Srinivas et al. (2010). These methods optimize a fixed-budget objective and lack a stopping rule for (ε, δ)-correct identification. Each baseline receives a fixed budget equal to the mean stopping time of the most sample-efficient `C-ICPE` variant in each configuration, rounded up; to test whether the baselines benefit from additional queries, we also report results with a budget of 100 queries. After exhausting the budget, each baseline recommends the best-observed location. As in the synthetic benchmarks, we also evaluate the critic-free ablation `C-ICPE-TS+CF`. We also report the cost multiplier at the end of training for every ε configuration in Table 22.

Accuracy and sample complexity. Table 23 reports identification accuracy (mean and 95% CI) at the two ε values, with $\delta = 0.1$. All three `C-ICPE` variants exceed the $1 - \delta = 0.90$ target in every configuration, with accuracy of 0.919–0.925 and tight confidence intervals. However, they differ instead in sample complexity (Table 24): `C-ICPE-TS-Mixture` requires the fewest queries, 21.3 at $\varepsilon = 0.2$ and 31.6 at $\varepsilon = 0.15$, compared to 41.5 and 42.3 for `C-ICPE-TS` and `C-ICPE-TD3` at $\varepsilon = 0.15$, a reduction of about 25% on the harder setting. Under a matched budget, all baselines fall well short of the target. `GP-UCB` is the strongest, reaching 0.722 at $\varepsilon = 0.2$ and 0.700 at $\varepsilon = 0.15$, followed by `GP-logEI` (0.566 and 0.648), while `TPE` and `CMA-ES` remain below 0.5 in both settings. With a fixed budget of 100 queries, 2.4–4.7$\times$ the sample complexity of `C-ICPE`, the baselines improve substantially. `GP-logEI` reaches 0.980–0.983, exceeding `C-ICPE`, and `TPE` reaches 0.942 at $\varepsilon = 0.2$ but still fall below 0.9 at $\varepsilon = 0.15$, while `CMA-ES` and `GP-UCB` still fall short of the target at both ε values. The baselines can therefore reach high accuracy on these spatially

d	Method	$\varepsilon = 0.2$ $\sigma = 0$	$\varepsilon = 0.15$ $\sigma = 0$
2	C-ICPE-TD3	24.4 [23.0,25.9]	42.3 [39.5,45.3]
	C-ICPE-TS	24.9 [23.5,26.4]	41.5 [38.8,43.9]
	C-ICPE-TS+CF	63.0 [60.6,65.5]	91.5 [88.8,94.2]
	C-ICPE-TS-Mixture	**21.3** [20.0,22.6]	**31.6** [29.6,33.5]

Table 24: Geochem: sample complexity (mean and 95% CI) for every (d, ε, σ) configuration. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

d	Method	$\varepsilon = 0.2$ $\sigma = 0$	$\varepsilon = 0.15$ $\sigma = 0$
2	C-ICPE-TD3	9.8×10^{-17} $[1.5 \times 10^{-19}, 1.9 \times 10^{-16}]$	2.5×10^{-10} $[3.8 \times 10^{-44}, 7.6 \times 10^{-10}]$
	C-ICPE-TS	8.0×10^{-17} $[2.2 \times 10^{-23}, 2.4 \times 10^{-16}]$	1.1×10^{-10} $[4.3 \times 10^{-22}, 2.6 \times 10^{-10}]$
	C-ICPE-TS+CF	$\mathbf{2.8 \times 10^{-292}}$ $[1.1 \times 10^{-307}, 8.3 \times 10^{-292}]$	$\mathbf{1.4 \times 10^{-195}}$ $[1.1 \times 10^{-216}, 3.0 \times 10^{-195}]$
	C-ICPE-TS-Mixture	5.9×10^{-14} $[7.9 \times 10^{-36}, 1.8 \times 10^{-13}]$	4.1×10^{-20} $[4.3 \times 10^{-33}, 1.2 \times 10^{-19}]$

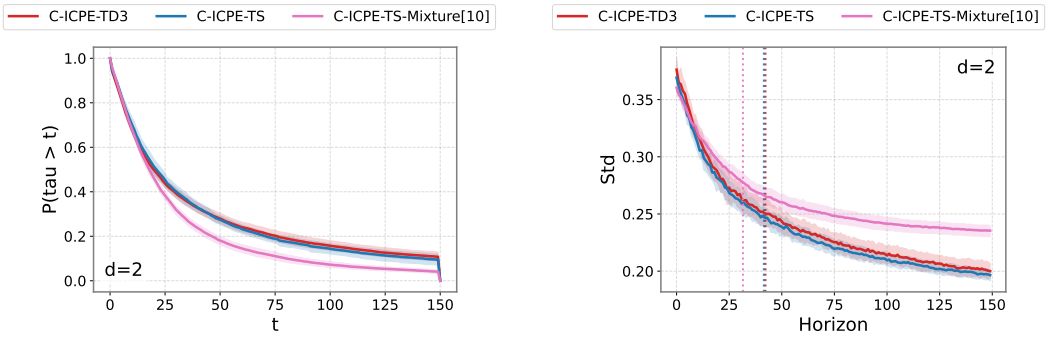
Table 25: Geochem: certification p-value ($H_0 : p \leq 0.9$, one-sided exact binomial test on $n = 20000$ independent trials, significance level $\alpha = 0.01$) for every (d, ε, σ) configuration: mean across seeds, with the [min, max] range across seeds in brackets. The number inside the square bracket indicates the number of components for `C-ICPE-TS-Mixture`. '+CF' denotes the critic-free stopping-rule ablation: the same trained network and actor, but the stop decision is taken directly from the inference network's own belief (stopping once it places mass $\geq 1 - \delta_{\text{train}}$ on the ε-ball around $\hat{x}$) instead of the trained critic.

structured surfaces given enough queries, but they require several times more samples than `C-ICPE` and provide no mechanism for deciding when to stop.

Survival function and uncertainty convergence. Fig. 25a shows the survival function $\mathbb{P}(\tau > t)$ at the hardest setting ($\varepsilon = 0.15$, $d = 2$). All three variants exhibit a rapid initial decline, with about half of the episodes terminating by $t \approx 20$. Beyond this point, `C-ICPE-TS-Mixture` stops noticeably faster and has a much thinner tail, which accounts for its lower mean sample complexity. For all methods, a small fraction of episodes (roughly 10% for `C-ICPE-TS` and `C-ICPE-TD3` and 5% for `C-ICPE-TS-Mixture`) reach the maximum horizon of $N = 150$ without stopping.

Fig. 25b shows the posterior standard deviation of the recommendation as a function of the horizon, with the dotted vertical lines marking each method's mean stopping time. `C-ICPE-TS` and `C-ICPE-TD3` behave nearly identically, stopping at $t \approx 42$ when their standard deviation has decreased to ≈ 0.25 and continuing to contract to ≈ 0.20 by $t = 150$. `C-ICPE-TS-Mixture` stops earlier ($t \approx 32$) and plateaus at a higher level of ≈ 0.24. As in the multi-lobe problem, the posterior here is multimodal, since several regions can be ε-correct, as illustrated in Fig. 21. The higher plateau reflects the mixture retaining mass on these candidate regions, so its level is not directly comparable to that of a single-Gaussian belief.

Certification. Table 25 reports the certification p-value for both (ε, σ) configurations evaluated. All three methods reject H_0 comfortably at both settings, with p-values ranging from 10^{-10} to 10^{-20} and no failures anywhere in the table. `C-ICPE-TS-Mixture` attains the smallest p-value at the harder $\varepsilon = 0.15$ setting (4.1×10^{-20}), while `C-ICPE-TS` is marginally ahead at $\varepsilon = 0.2$; given



(a) (b)

Figure 25: Results for geochemical problem with fixed confidence $\delta = 0.1$ and $N = 150$ across different dimensions at the hardest ε setting: (a) survival function of τ; (b) inference uncertainty convergence.

both are already many orders of magnitude past $\alpha = 0.01$, this ordering reflects a modest difference in margin rather than a meaningful certification gap between methods.

Critic-free stopping. On geochemical problem, removing the critic affects efficiency rather than correctness. `C-ICPE-TS+CF` exceeds the target (0.966 and 0.953) and certifies for every seed (Table 25), but it requires 2.2–2.5× more queries than `C-ICPE-TS` (63.0 vs. 24.9 at $\varepsilon = 0.2$ and 91.5 vs. 41.5 at $\varepsilon = 0.15$). Its belief is thus underconfident: under the value-based loss, several spatially separated locations can be ε-correct, while the critic-free rule counts only the belief mass near $\hat{x}$, which likely underestimates the probability of success. The critic avoids this and stops as soon as the recommendation is reliable. Together with the previous problems, this shows that the critic corrects miscalibration of the belief in both directions: it prevents premature stopping when the belief is overconfident and unnecessary queries when it is underconfident.

Robustness to prior misspecification. Fig. 26 reports accuracy under misspecified Beta priors $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}^2$ on the normalized $[0, 1]^2$ domain, with $\varepsilon = 0.15$. The matched prior used during meta-training corresponds to $\alpha = \beta = 1$ (uniform). All three methods maintain accuracy between ≈ 0.90 and ≈ 0.93 across the entire grid, with no significant degradation even at extreme configurations such as $(\alpha, \beta) = (0.5, 7)$ or $(7, 0.5)$. This degree of robustness is notably stronger than what is observed on the Ackley benchmark (Fig. 18), where boundary-skewed priors cause substantial degradation.

Taken together, `C-ICPE` transfers to surrogate tasks built from held-out regions of real survey data: it remains certified, while the BO baselines fall below the target at matched budget and exceed it only with 100 queries.

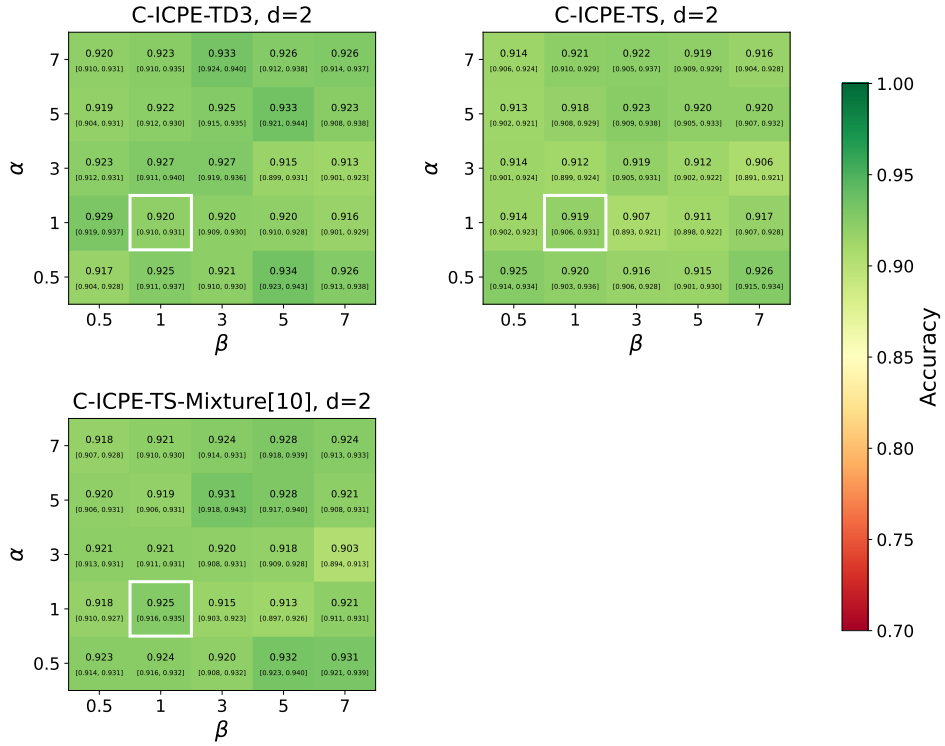


Figure 26: Robustness to prior misspecification on the Geochemical exploration ($\varepsilon = 0.15$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.