In-Context Pure Exploration in Continuous Decision Spaces

Alessio Russo [*1] **Yin-Ching Lee** [*1] **Ryan Welch** [2] **Aldo Pacchiano** [1]

Abstract

In active sequential testing, also termed *pure exploration*, a learner is tasked with the goal to adaptively acquire information so as to identify an unknown ground-truth hypothesis with as few queries as possible. This problem has several motivating applications, including Best-Arm Identification (BAI) in bandits, where actions index hypotheses, and generalized search problems, where strategically chosen queries reveal partial information about a hidden label. In many modern settings, however, the hypothesis, or recommendation space, is *continuous*: for example, identifying a near optimal action in a continuous-armed bandit, localizing an ϵ-ball contained in a target region, or estimating the minimizer of a function from noisy observations. Existing methods are predominantly frequentist and model-specific, while learned approaches have been limited to finite recommendation spaces. We introduce **C-ICPE**, a theory-guided learned model for Bayesian fixed-confidence pure exploration with continuous recommendations. **C-ICPE** meta-trains sequential architectures over a task prior to jointly learn exploration, stopping and recommendations strategies. At inference time, it actively gathers evidence on tasks and identifies an ϵ-optimal recommendation *without* parameter updates.

1. Introduction

Several learning problems are inherently interactive: the learner sequentially performs interventions or stages queries, 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 $1-\delta$.

This regime is well understood in canonical settings with finite decision spaces, including best-arm identification in stochastic multi-armed bandit models [Garivier and Kaufmann, 2016] and best-policy identification in Markov Decision Processes (MDPs) [Puterman, 2014]. In these settings the learner chooses queries (e.g., arms) and outputs an object of interest, often a best action/policy, and the theoretical guarantees have been studied in a range of settings [Degenne et al., 2019, Poiani et al., 2025b, Al Marjani et al., 2021, Russo et al., 2025a].

Despite this progress, practical methods for fixed-confidence pure exploration remain limited when the learner must return a recommendation in a *continuous* space. Existing continuous methods are either frequentist and model-specific [Garivier and Kaufmann, 2021, Takemori et al., 2025, Poiani et al., 2025a, Russo and Pacchiano, 2025] or Bayesian but restricted to Gaussian processes and not optimizing sample complexity [Wilson, 2024]. Even in finite arms, the theory of Bayesian fixed-confidence pure exploration is limited, and results are known only in the finite setting with Gaussian likelihoods/priors [Jang et al., 2024]. No analogous Bayesian theory, nor practical methods, are known for continuous recommendation spaces under general priors.

Recently, [Russo et al., 2025b] proposed In-Context Pure Exploration (ICPE), which meta-trains sequential neural policies for finite active-testing problems. However, ICPE is restricted to finite hypothesis and action sets, and does not address the continuous recommendation setting. Hence, it is currently missing a broadly reusable learned method for Bayesian fixed-confidence pure exploration when the recommendation itself is continuous and performance is optimized over a task prior. We introduce **C-ICPE**, a theory-guided method that learns to collect data, stop, and recommend directly from trajectories in continuous recommendation spaces. This type of (ϵ,δ)-PAC exploration directly addresses problems in experimental sciences such as materials discovery [Liu et al., 2017] and dose-finding [O'Quigley

[1]Faculty of Computing and Data Sciences, Boston University, Boston, MA, USA. [2]Department of Computer Science, Stanford University, Stanford, CA, USA.. Correspondence to: Alessio Russo <arusso2@bu.edu>, Yin-Ching Lee <leewill@bu.edu>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

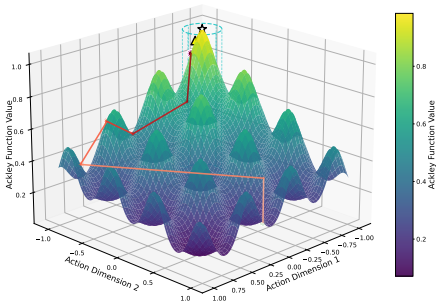


Figure 1. **C-ICPE** is able to identify the global maxima (with ϵ-accuracy and $1-\delta$ confidence) of the inverted Ackley function (with random parameters and observation noise) without gradients while trying to use the least number of data-points.

et al., 1990], where each trial is costly, observations are noisy, and practitioners need not just a good answer but a guarantee that the answer is ϵ-correct with a given confidence. **Contributions.** First, we formulate Bayesian fixed-confidence pure exploration with continuous recommendations, and establish the corresponding Bellman optimality structure. Second, we prove Bayesian (ϵ, δ)-correctness under a local closedness condition that is weaker than the uniqueness assumption in [Russo et al., 2025b], using a novel subdifferential argument. Third, we instantiate this framework into a practical method, **C-ICPE**, to train exploration policies that deploy model-free, and evaluate it on noisy binary search, ϵ-best arm identification on the unit sphere, noisy Ackley minimization, value estimation in Gaussian Processes and a real-world geochemical task where the goal is to locate peak copper concentration from sparse soil measurements [usg, 2004]. To our knowledge, this is the first practical learned framework combining continuous recommendations and fixed-confidence stopping.

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 goal is to collect a history that is sufficiently informative to output a high-quality recommendation $\hat{x} \in \mathcal{X}$, where $\mathcal{X}$ is compact. We refer to $\mathcal{X}$ as the hypothesis or recommendation space. We distinguish $\mathcal{A}$ from $\mathcal{X}$: $\mathcal{A}$ is the query space used to collect information, while $\mathcal{X}$ is the space of objects the learner may return. In many tasks $\mathcal{X} = \mathcal{A}$, but in value-identification tasks $\mathcal{X}$ may instead be a set of possible function values.

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 Section B we state the regularity assumptions used in the theoretical results). Throughout the paper, $x^\star_\theta$ denotes a selected zero-loss target in the recommendation space $\mathcal{X}$, i.e., $L_\theta(x^\star_\theta) = 0$. Depending on the problem, this object may be an optimizer, a best arm, a threshold, or an optimal value. When the zero-loss set is not a singleton, we assume a fixed continuous selector $\theta \mapsto x^\star_\theta$ to ensure regularity.

In many examples, the loss is defined through a function f_θ parametrized by θ. In function optimization problems, we set $\mathcal{X} = \mathcal{A}$, and the goal is to find a point $\hat{x} \in \mathcal{X}$ that optimizes the function. In this case, one can take the risk loss L_θ to be value-gap loss $L_\theta(x) := f_\theta(x) - f_\theta(x^\star_\theta)$, with $x^\star_\theta \in F^\star(\theta)$, where $F^\star(\theta) := \arg\min_{x\in\mathcal{X}} f_\theta(x)$, or a distance loss in the query space $L_\theta(x) := \|x^\star_\theta - x\|$ if the goal is to find x close to a selected optimal point $x^\star_\theta \in F^\star(\theta)$. Other problems include the ϵ-best arm identification problem in multi-armed bandits where $f_\theta(x)$ is the mean reward of arm x, or noisy binary search in a continuum, where the agent observes noisy observations of $\mathrm{sign}(x - x^\star_\theta)$ and the loss is defined in the query space. Problems where the recommendation space $\mathcal{X}$ is not identical to the query space $\mathcal{A}$ include optimal-value identification, where the learner returns a scalar estimate of the optimal value: we take $\mathcal{X} \subset \mathbb{R}$, and set $L_\theta(x) := |x - x^\star_\theta|$ where $x^\star_\theta := \max_{a\in\mathcal{A}} f_\theta(a)$.

Optimization objective. We work in the fixed-confidence ((ϵ, δ)-PAC) regime. A learner is defined by the triplet (π, I, τ): a sampling policy $\pi = (\pi_t)_{t\ge1}$ such that $A_t = \pi_t(H_t)$; a stopping time τ with respect to $\mathcal{F}_t = \sigma(H_t)$, defining when to stop the data acquisition process; an inference rule $I = (I_t)_{t\ge1}$ such that $\hat{x}_\tau = I_\tau(H_\tau)$. The goal of the learner is to adaptively choose queries $A_1, A_2, \dots$ and a stopping time τ, so that the returned $\hat{x}_\tau$ is ϵ-optimal, i.e. $\hat{x}_\tau \in \mathcal{X}_\epsilon(\theta)$, with high probability. Hence, for a given pair $\epsilon > 0, \delta \in (0, 1/2)$, we seek to minimize the (expected) number of queries while ensuring δ-correctness: formally, we solve

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

Our formulation is Bayesian: ν is both a prior over environments and the task distribution used for training and

evaluation. This enables amortized learning across tasks: the models are trained on episodes drawn from ν and transfer to new tasks from the same family without parameter updates. Second, it defines the posterior success probability $q_t(h, x)$ that drives our theory and algorithms. Third, the average-case guarantee under ν is the natural objective for applications where the practitioner has domain knowledge about plausible task distributions. For this setup, we are not aware of analogous Bayes-optimal characterizations for continuous recommendation spaces under general priors, as results are limited to finite settings with Gaussian structure [Jang et al., 2024].

3. Theoretical Background

This section provides the theoretical foundation for **C-ICPE**, where we characterize an optimal learner. Relative to the finite ICPE analysis of Russo et al. [2025b], the continuous setting introduces two technical issues absent in finite spaces: (i) attainment of the Bellman equation over continuous actions requires establishing a semicontinuity chain from the observation model through the posterior predictive to the Q-function, and (ii) the (ϵ, δ)-correctness proof must handle non-singleton dual optima, which we address via a weaker local closedness condition and a subdifferential argument that replaces the uniqueness and monotonicity assumptions of the finite case. Together, these results provide the first theoretical infrastructure for Bayesian pure exploration with continuous recommendations.

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:

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

Here $q_t(h, x)$ is the posterior success probability of recommending x, and $r_t(h)$ is the best posterior success probability achievable from history h. One can show that the maximum is attained and an optimal inference rule is any selector (see Proposition 1 for a proof)

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

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., 2025b]. Introducing a multiplier $\lambda \ge 0$ for the correctness constraint gives the dual

$$\inf_{\lambda \ge 0} \sup_{\pi, I, \tau} V_\lambda(\pi, I, \tau), \tag{2}$$

where

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

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 Lemma 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

$$V_t^\star(h; \lambda) := \sup_{\bar{\pi}} \mathbb{E}^{\bar{\pi}}_{\theta \sim \nu}\left[\lambda r_{\bar{\tau}}(H_{\bar{\tau}}) - (\bar{\tau} - t) \mid H_t = h\right]. \tag{3}$$

Define

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

where the expectation is under the posterior predictive kernel $\bar{P}_t(\cdot \mid h, a)$.

Theorem 3.1 (Bellman equation and greedy deterministic optimality)**.** *Assume the regularity conditions of Section B.3 (Assumption 2 and Assumption 4), then*

$$V_t^\star(h; \lambda) = \max\left\{Q^\star_{t,\text{stop}}(h; \lambda), \sup_{a \in \mathcal{A}} Q^\star_{t,\text{cont}}(h, a; \lambda)\right\}. \tag{4}$$

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.

Unlike the finite case, where Bellman attainment is automatic, the continuous setting requires verifying that the supremum over $a \in \mathcal{A}$ in Eq. (4) is attained. Our proof (Section B.3.3) establishes this through a value-iteration construction that also handles the coupling with the stopping/continue structure, and propagates lower semicontinuity from the observation model through the posterior predictive kernel to the Q-function. This semicontinuity chain is specific to this problem and does not follow from any existing reference by specialization.

Zero duality gap and (ϵ, δ)-correctness. The Bellman theorem characterizes the inner problem for a fixed multiplier λ. We now state when the Lagrangian relaxation is exact. 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 3.2 (Zero duality gap and (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 $g(\lambda) := \inf_{\pi \in \mathcal{T}} \{c(\pi) + \lambda(1 - \delta - \rho(\pi))\}$, $\lambda^\star \in \arg\max_{\lambda \geq 0} g(\lambda)$, and let $\mathcal{S}(\lambda^\star)$ be the set of dual minimizers at $\lambda^\star$. If there exists $\epsilon_0 > 0$ such that*

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

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

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

This theorem improves on the corresponding result in [Russo et al., 2025b] in two ways. First, we replace the assumption that the dual-optimal policy is unique with a weaker 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., 2025b] derives correctness via a monotonicity argument on the optimal cost, our proof (Section B.4) uses a direct subdifferential characterization to show that if all near-optimal policies have $\rho < 1 - \delta$, then every subgradient of the 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 Section B.5.

4. Continuous ICPE: C-ICPE

In this section we describe **C-ICPE**, a practical method based on the theory of the previous section. **C-ICPE** has three components: inference, stopping, and exploration. Each of these are implemented via learned sequential neural architectures, trained end-to-end from interaction data. Once trained, we use **C-ICPE** at deployment time (a.k.a. test-time or inference-time) to perform pure exploration. We now describe the models learned by **C-ICPE** and the training procedure. More details can be found in Section C of the appendix.

Training Protocol. We adopt a similar meta-training protocol as the one used in [Russo et al., 2025b] to train the models. Briefly, we assume access to a simulator over ν from which we can sample trajectories. We use a meta-training, where we sample tasks $\theta \sim \nu$ and assume access to a zero-loss target $x^\star_\theta \in \mathcal{X}$, collect trajectories using **C-ICPE**, store the data in a replay buffer $\mathcal{B}$, and perform off-policy updates: (1) a likelihood update of the parameters of the inference model, (2) a DQN-like update of the critic and (3) an update of the actor based on the learned Q function. After training, we freeze all models; deployment requires no access to $x^\star_\theta$ or the prior. We now discuss the modeling of these components more in detail.

Gaussian inference model. For a history h, the ideal inference rule maximizes the posterior success probability $q_t(h, x) := \mathbb{P}(L_\theta(x) \leq \epsilon | H_t = h)$ over recommendations $x \in \mathcal{X}$. However, computing q_t is not straightforward, as the posterior distribution may have a complex shape. Instead, we train the inference model to learn the posterior law of $x^\star_\theta$ from trajectories and outputs a diagonal Gaussian distribution characterizing the uncertainty around $x^\star_\theta$

$$I_\phi(\cdot|h) = \mathcal{N}\left(\mu_\phi(h), \mathrm{diag}(\sigma^2_\phi(h))\right).$$

The recommendation at stopping is defined as the mean $\hat{x} = \mu_\phi(h)$. The covariance characterizes the uncertainty around this point, and, as shown in Proposition 7, the optimal mean and covariance are the posterior mean and covariance of $x^\star_\theta$ given $H_t = h$. Thus the Gaussian is a moment projection of the posterior law of $x^\star_\theta$. The deployed recommendation $\hat{x} = \mu_\phi(h)$ should therefore be viewed as a tractable approximation to $\arg\max_{x \in \mathcal{X}} q_t(h, x)$, rather than as an exact maximizer of q_t. In Section B.7, we show that μ_ϕ is near-optimal when the posterior uncertainty on $x^\star_\theta$ is small relative to the ϵ-success margin (see Proposition 8). This justification is most direct for localization losses, where $\mathcal{X}_\epsilon(\theta)$ is a neighborhood of $x^\star_\theta$.

We train ϕ using a log-likelihood loss on a batch of partial trajectories $B = (x^\star_i, H_{t_i})_i$ sampled from the buffer $\mathcal{B}$, where $x^\star_i$ is the optimal point for trajectory i and t_i is a timestep sampled uniformly at random for that trajectory

$$\mathcal{L}_{\mathrm{inf}}(B; \phi) = -\sum_{i=1}^{|B|} \log I_\phi\left(x^\star_i | H_{t_i}\right). \tag{5}$$

In the following, we denote by $\bar{\phi}$ the target parameter of the inference model, updated via a Polyak update $\bar{\phi} \leftarrow (1 - \tau_I)\bar{\phi} + \tau_I \phi$ with $\tau_I \in (0, 1)$.

Critic and reward definition. We parametrize the critic by ψ, and model it with two heads: a continuation head $Q_\psi(h, a)$ for $a \in \mathcal{A}$ and a stopping head $Q_\psi(h, a_{\mathrm{stop}})$. We also define the value: let $a_{\mathrm{tgt}}(h')$ be the continuation target action proposed at the next history by the current policy, then the value is defined as

$$V_\psi(h') := \max\left\{Q_\psi(h', a_{\mathrm{stop}}), Q_\psi(h', a_{\mathrm{tgt}}(h'))\right\},$$

Using the definition of the Q function from Theorem 3.1, we learn the paramter ψ using TD-learning. While the ideal reward would be $r_t(h) = \max_x q_t(h, x)$, to better capture the uncertainty around the recommendation $\mu_t(h)$, we use a sampled reward from the target inference model $I_{\bar{\phi}}$:

$$\hat{r}(h, \theta) := \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left\{L_\theta(X^{(k)}) \leq \epsilon\right\}, \qquad X^{(k)} \sim I_{\bar{\phi}}(\cdot|h). \tag{6}$$

Algorithm 1 `C-ICPE`

```
// Training phase
Initialize buffer B, networks Q_ψ, I_φ, actor π.
while Training is not over do
    Sample environment M_θ ~ ν and hypothesis x*_θ; observe Y_1 ~ ρ and set t ← 1.
    repeat
        Execute action A_t ~ π(· | H_t) according to actor π and observe Y_{t+1}.
        Add partial trajectory (H_t, A_t, Y_{t+1}, x*_θ) to B and set t ← t + 1.
    until Q_ψ(H_t, a_stop) ≥ Q_ψ(H_t, A_t).
    In the fixed confidence, update c according to Eq. (10).
    Sample batch B ~ B and update models using L_inf(B; φ) (Eq. (5)) and L_critic(B; ψ) (Eq. (7)); for TD3, train π according to L_act(B; ψ) (Eq. (9)).
end while
// Inference/Deployment phase (models are fixed here)
Sample unknown environment M ~ ν and collect a trajectory H_τ using π (until Q_ψ(H_t, A_t) ≤ Q_ψ(H_t, a_stop)).
Return x̂_τ = μ_φ(H_τ) (recommendation).
```

Conditionally on (h, θ), this is an unbiased estimate of the probability that a sample from the inference distribution lies in $\mathcal{X}_\epsilon(\theta)$. After averaging over the posterior, this reward is $\mathbb{E}_{X\sim I_{\bar{\phi}}(\cdot|h)}[q_t(h, X)] \leq r_t(h)$, so it is a conservative version of the ideal posterior reward and allows the model to capture the current uncertainty in the recommendation rule (Proposition 6 in the appendix makes this precise, as the gap between the practical reward $\widehat{r}_t(h)$ and the ideal Bellman reward $r_t(h)$ is controlled by the second moment of the inference model).

Then, we sample a batch of partial trajectories $B = (H_{t_i}, A_{t_i}, H_{t_i+1}, d_{t_i}, x_i^\star)_i$ from the buffer, where t_i is a uniformly sampled timestep for the i-th trajectory and $d_{t_i} = 1$ when the maximum horizon is reached. Then, we use targets

$$y_i^{\text{stop}} = \hat{r}(H_{t_i}, \theta),$$
$$y_i^{\text{cont}} = -c + d_{t_i}\hat{r}(H_{t_i+1}, \theta) + (1 - d_{t_i})V_{\bar{\psi}}(H_{t_i+1}),$$

where we reparametrized the Lagrange multiplier as a per-step cost $c = 1/\lambda$ and the inner Lagrangian objective remains the same. Then, the critic loss is

$$\mathcal{L}_{\text{critic}}(B; \psi) = \frac{1}{2|B|}\sum_{i=1}^{|B|}\Big[\big(Q_\psi(H_{t_i}, A_{t_i}) - y_i^{\text{cont}}\big)^2 \quad (7)$$
$$+ \big(Q_\psi(H_{t_i}, a_{\text{stop}}) - y_i^{\text{stop}}\big)^2\Big]. \quad (8)$$

The critic also decides when to stop: at rollout time the current policy π first proposes A_t, and the learner stops iff

$$Q_\psi(H_t, a_{\text{stop}}) \geq Q_\psi(H_t, A_t).$$

Hence, at deployment, no access to θ is required as the stopping comparison uses only H_t.

Policy and actors. The policy, or actor rule, decides what continuation action $a \in \mathcal{A}$ to choose next. Depending on the whether $\mathcal{X} = \mathcal{A}$, we propose three possible actor rules.

Thompson Sampling (TS) rule: this rule can be used when the recommendation space and the query space coincide $\mathcal{X} = \mathcal{A}$, and eliminates the need for a separate actor. This rule draws inspiration from classical Thompson Sampling [Thompson, 1933]: the policy is implicitly represented by the inference model, and the actions are sampled according to $A_t \sim I_\phi(\cdot|H_t)$. This is useful when informative queries are themselves plausible recommendations. Early in an episode the exploration is more spread; later, as the posterior target law contracts, TS concentrates near the current recommendation. As target action for the critic we use the mean value of the inference model $a_{\text{tgt}}(h) = \mu_t(h)$.

Top Two Posterior Sampling (TTPS) rule: also this rule can be used when the recommendation space and the query space coincide $\mathcal{X} = \mathcal{A}$. This rule is similar to classical TTPS [Russo, 2016], but we extend it to the continuous case. This rule draws a sample $A_t \sim I_\phi(\cdot|H_t)$ and, with probability $1/2$, it samples until the new sample is farther from the mean $\mu_\phi(H_t)$. The logic is that the posterior mean is the current recommendation, while samples farther from it represent plausible alternatives. TTPS therefore spends some probability mass checking alternatives instead of repeatedly querying near the current mean before the stopping critic is confident. As target action for the critic we use the mean value of the inference model $a_{\text{tgt}}(h) = \mu_t(h)$.

TD3 rule [Fujimoto et al., 2018]: this rule can be used for general recommendation spaces when $\mathcal{A} \neq \mathcal{X}$, and formally tries to solve the Bellman equation in Theorem 3.1. It learns a parametric actor $\pi_\vartheta(h) \in \mathcal{A}$ from the critic. The deterministic actor is trained by

$$\mathcal{L}_{\text{act}}(B; \vartheta) = -\frac{1}{|B|}\sum_i Q_\psi(H_{t_i}, \pi_\vartheta(H_{t_i})). \quad (9)$$

The critic target uses the usual TD3 stabilizers: a target actor, target-action smoothing, and twin critics. In case $\mathcal{X} = \mathcal{A}$ we can use a stochastic TD3 variants, where TD3 learns the mean $\bar{\mu}$ and covariance $\bar{\Sigma}$ of a Gaussian actor. In this case the loss is augmented with an imitation learning loss $\text{KL}(I_\psi(\cdot|h)\|\pi_\vartheta(\cdot|h))$ that provides a rich signal: the actor can increase variance when sampled actions have higher continuation value and shrink it when exploration is no longer useful.

Cost update. We update the per-step cost c by simply performing a gradient step on the dual variable. We sample a fresh batch of trajectories, and estimate the success rate $\hat{p} = \frac{1}{|B|}\sum_{i=1}^{|B|}\mathbf{1}\left\{\mu_\phi(H_\tau^{(i)}) \in \mathcal{X}_\epsilon(\theta^{(i)})\right\}$, and update the cost as follows

$$c \leftarrow \text{Proj}_{[0,1]}\left(c - \eta_c\left((1-\delta) - \hat{p}\right)\right). \quad (10)$$

If empirical success is below $1-\delta$, the cost decreases and trajectories become longer; otherwise, if above the target, the cost increases and stopping becomes more aggressive.

Correctness certification. The zero-duality result from the previous section justifies the ideal Lagrangian objective, but a trained model is still approximate. In Section B.6 we outline how to obtain formal (ϵ, δ)-guarantees on the trained model (see Proposition 5).

5. Empirical Evaluation

We evaluate **C-ICPE** on various benchmarks: noisy binary search, ϵ-best arm identification on the unit sphere, Ackley minimization, and GP max-value estimation. We also validate on on a real-world geochemical exploration task [usg, 2004]. We compare four exploration rules: TS, TTPS, TD3, and uniform sampling. For all experiments we set a maximum sample complexity $t_{\max}$ (details in Section D) and report 95% confidence intervals using bootstrap. We also compare against Bayesian optimization baselines whenever possible: Tree-structured Parzen Estimator (TPE) [Bergstra et al., 2011], Gaussian Process (GP) with UCB or Expected Improvement [Srinivas et al., 2010, Ament et al., 2023], and CMA-ES [Hansen, 2016]. Each is given a fixed sample budget larger than the median stopping time of **C-ICPE**. These methods are not (ϵ, δ)-correct competitors, and optimize a fixed-budget objective. We include them to test whether standard methods already attains the target correctness at comparable budgets. For ϵ-best arm on the sphere, we additionally compare against Lazy Track-and-Stop [Jedra and Proutiere, 2020], an optimal frequentist fixed-confidence baseline.

5.1. Synthetic Benchmarks

We now provide a brief description of the benchmarks, and then discuss the results.

Noisy binary search. The environment parameter θ is drawn uniformly from $[-1,1]^d$, with selector $x^\star(\theta) = \theta$. The agent queries $a \in [-1,1]^d$ and observes, per coordinate, $y_i = \xi_i \cdot \mathrm{sign}(\theta_i - a_i)$, where $\xi_i \in \{-1,+1\}$ are i.i.d. Rademacher random variables with $\mathbb{P}(\xi_i = +1) = 1 - p$. The loss is $L_\theta(x) = \|x-\theta\|_2$, so the ϵ-optimal set is $\mathcal{X}_\epsilon(\theta) = \{x : \|x-\theta\|_2 \le \epsilon\}$. Here $\mathcal{X} = \mathcal{A} = [-1,1]^d$. The difficulty is controlled by the noise rate p and the dimension d: each coordinate provides one bit of corrupted information per query, and the agent must simultaneously localize all d coordinates.

ϵ-best arm on the sphere. The environment parameter θ is drawn uniformly on the unit sphere $\mathbb{S}^{d-1}$, with selector $x^\star(\theta) = \theta$. The agent queries $a \in [-1,1]^d$ and observes a noisy linear reward $y = f_\theta(a) + \xi$, where $f_\theta(a) = \theta^\top a$ and $\xi \sim \mathcal{N}(0,\sigma^2)$. The loss is defined via the inner product: $L_\theta(x) = 1 - f_\theta(x)$, so the ϵ-optimal set is $\mathcal{X}_\epsilon(\theta) = \{x : f_\theta(x) \ge 1-\epsilon\}$. Here $\mathcal{X} = \mathcal{A}$, and the difficulty lies in estimating a direction from scalar projections.

Ackley minimization. The agent must locate the global minimizer of a randomly parametrized Ackley function [Naser et al., 2025], a standard multimodal benchmark for global optimization. The parameter is $\theta = (\alpha, \beta, \gamma, \theta^\star)$, where (α, β, γ) control the function shape and $\theta^\star \in [-1,1]^d$ is the global minimizer (selector $x^\star(\theta) = \theta^\star$). The agent queries $a \in [-1,1]^d$ and observes a normalized function evaluation $y = \tilde{f}_{\alpha,\beta,\gamma}(a - \theta^\star) + \xi$, $\xi \sim \mathcal{N}(0,\sigma^2)$. The loss is $L_\theta(x) = \|x - \theta^\star\|_2$. Here $\mathcal{X} = \mathcal{A} = [-1,1]^d$. 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.

GP max-value estimation. A function f is sampled from a Gaussian process $\mathrm{GP}(0, k_{\mathrm{RBF}}(\ell, \sigma_f))$ on $[0,1]^d$, with lengthscale $\ell \sim \mathrm{Unif}[0.05, 0.2]$ and output scale $\sigma_f = 1$. The target is the scalar maximum value $\theta^\star = \max_x f(x)$, with selector $x^\star(\theta) = \theta^\star \in \mathbb{R}$. The agent queries $a \in [0,1]^d$ and observes $y = f(a) + \xi$, $\xi \sim \mathcal{N}(0,\sigma^2)$. The loss is $L_\theta(x) = |x - \theta^\star|$. This is the $\mathcal{X} \neq \mathcal{A}$ setting: the recommendation space $\mathcal{X} \subseteq \mathbb{R}$ is scalar while the action space $\mathcal{A} = [0,1]^d$ is d-dimensional, requiring the TD3 actor to learn an exploration policy decoupled from the inference model. The difficulty is twofold: the agent must both explore to find the region of high function values and estimate the peak value to within ϵ, without knowing the function's lengthscale in advance. For this problem we compare against two non-parametric baselines: (1) uniform sampling over the domain, reporting the trimmed mean of the top-5% observed values; (2) partitioning the domain into uniform bins, sampling uniformly within each bin, and reporting the highest bin average as the value estimate.

Results. Figs. 2-4 report accuracy and sample complexity (and their 95% confidence intervals) at the hardest (ε, σ) per benchmark; full sweeps over (ε, σ, d), experimental details, and robustness to prior misspecification are in Section D. Across all four tasks, C-ICPE with learned exploration (TS, TTPS, or TD3) consistently meets the $1-\delta$ accuracy target, while BO baselines and non-parametric estimators fall well below, confirming that fixed-budget optimization does not yield (ε, δ)-correctness. C-ICPE-uniform is competitive at low d but degrades as dimension increases on Ackley and binary search, where directed exploration matters. On ε-best arm, uniform exploration is optimal by rotational symmetry [Jedra and Proutiere, 2020], and Lazy Track-and-Stop achieves optimal sample complexity by exploiting the linear structure; C-ICPE matches the accuracy target but uses more samples, reflecting the cost of a model-agnostic stopping rule. On GP max-value estimation ($\mathcal{X} \neq \mathcal{A}$), C-ICPE-TD3 meets the target while non-parametric baselines fall short; BO methods are inapplicable here as they return

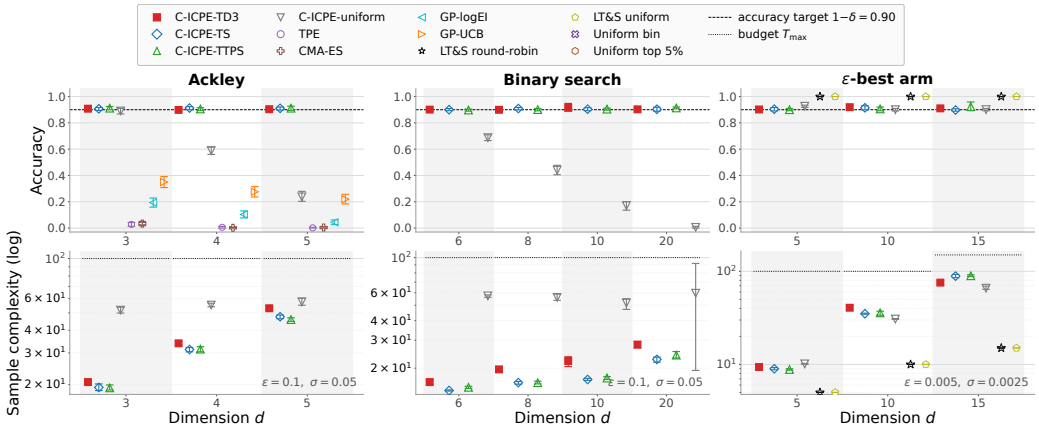


Figure 2. Accuracy (top) and sample complexity (bottom) at the hardest (ε, σ) per benchmark.

locations rather than values. Particularly, in Section B.9 we prove that (Theorem B.5), under an RBF-GP prior with interior regularity, max-value estimation is asymptotically harder than argmax localization: we establish a sample complexity lower bound for value estimation and an upper bound for argmax identification via a two-stage algorithm (T-BAL; Algorithm 2), showing that the geometric structure of the function helps localization but not value estimation.

5.2. Geochemical Exploration

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 with (ϵ, δ)-guarantees. We partition the data into geographic regions, fit a sparse variational Gaussian process 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 Section D.5 for more details.

Results. In Fig. 4 we present the results. In this problem, for sake of simplicity we only test the TD3 and TS actors for ICPE, and compare with respect to classical Bayesian baselines. The results show accuracy and sample complexity on held-out regions: C-ICPE-TD3, C-ICPE-TS achieve the $1-\delta$ accuracy target, while BO baselines fall below the target while using a larger sample budget, demonstrating that both learned exploration and learned stopping contribute on this real-data task. This shows evidence that **C-ICPE** transfers across tasks with genuine distribution shift. In Section D.5 we report the experimental details, and results for different values of ϵ.

6. Discussion, Related Work and Conclusions

Active sequential hypothesis testing (ASHT) provides the broad conceptual umbrella for this paper. In ASHT, a learner adaptively selects experiments and decides when to stop and declare a hypothesis, with the objective of minimizing expected sample size subject to a correctness constraint [Chernoff, 1959, Wald and Wolfowitz, 1948, Ghosh, 1991, Naghshvar and Javidi, 2013, Naghshvar et al., 2012]. A key methodological theme in this literature is that fixed-confidence constraints can be handled via Lagrangian duality. Closely related line of works include Bayesian experimental design and Bayesian active learning, which study adaptive data acquisition when the unknown is drawn from a known prior, typically optimizing expected utility or information gain [Lindley, 1956, Golovin and Krause, 2011, Rainforth et al., 2024], and active learning, which emphasizes selecting informative queries/labels to reduce uncertainty efficiently [Cohn et al., 1996]. Despite the shared emphasis on adaptive measurement and sequential stopping, most classical ASHT results assume substantial knowledge of the observation model: one typically has access to likelihoods (or at least to a parametric family) for every experiment under every hypothesis, enabling explicit likelihood-ratio statistics and model-based allocation [Naghshvar and Javidi, 2013]. While there are efforts toward relaxing this assumption to partial model knowledge [Cecchi and Hegde, 2017], the need for explicit likelihood structure remains a limiting factor for modern continuous environments with complex, history-dependent feedback. In many practical settings, the learner must instead infer both (i) which latent environment/task it is facing and (ii) which actions are informative, using only interaction data and function approximation. This motivates data-driven approaches that preserve the ASHT objective while reducing dependence on fully specified likelihoods.

The most developed special case of ASHT is fixed-confidence pure exploration in bandits, where the hypothesis is the identity of an optimal action. In finite-armed bandits [Lattimore and Szepesvári, 2020], best-arm identification (BAI) at $\epsilon = 0$ is characterized by a mature the-

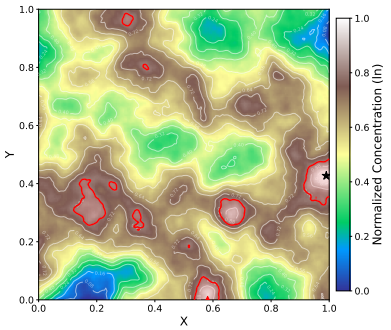


Figure 3. copper concentration in a 2D region in the geochemical exploration task. Red regions indicate concentration of copper within ϵ of the maximum value.

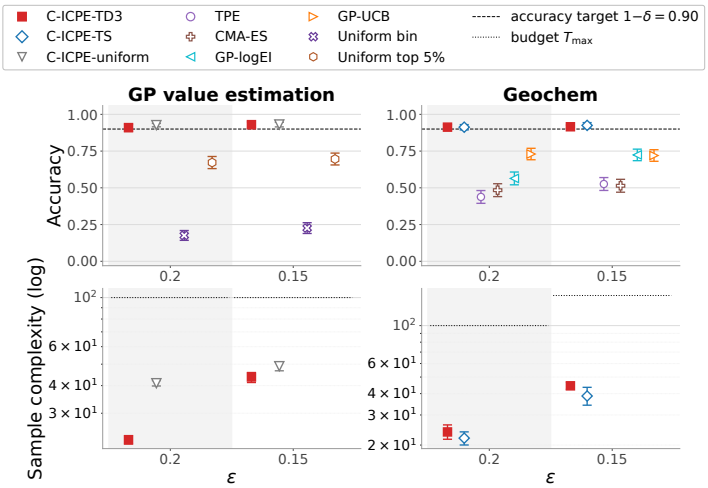


Figure 4. Accuracy (top) and sample complexity (bottom) at the hardest σ per benchmark.

ory: instance-dependent lower bounds quantify the intrinsic complexity of identifying the best arm [Garivier and Kaufmann, 2016, Degenne et al., 2019, Wang et al., 2020, Jedra and Proutiere, 2020, Kocák and Garivier, 2021, Russo and Proutiere, 2023b, Poiani et al., 2025b, Russo et al., 2025a], and a family of algorithms achieves near-optimal sample complexity by coupling adaptive allocation with statistically valid stopping rules [Audibert and Bubeck, 2010, Russo, 2016, Garivier and Kaufmann, 2016, Wang et al., 2021, Jourdan et al., 2022]. These results provide both sharp guidance and strong baselines, but they rely on a finite decision set and model-specific likelihood constructions. Similar themes arise in pure exploration for Markov decision processes, where the goal is to identify an optimal policy with probability at least $1-\delta$ [Al Marjani et al., 2021, Taupin et al., 2023, Al Marjani et al., 2023, Russo and Proutiere, 2023a, Russo and Vannella, 2024]. This literature yields sharp insights into exploration complexity but is likewise developed for finite state-action structure and frequentist guarantees.

Even under well-specified models, moving from $\epsilon = 0$ to (ϵ, δ)-PAC identification in continuous decision spaces can be technically demanding. Several recent works address continuous pure exploration in bandit models [Garivier and Kaufmann, 2021]. Takemori et al. [2025] give a tractable algorithm for continuous-arm linear bandits and Poiani et al. [2025a] derive lower bounds and a Track-and-Stop framework for infinite-answer problems; both are frequentist, model-specific, and require explicit likelihood structure. In MDPs, some work begun to treat (ϵ, δ)-PAC objectives in finite MDP settings for best policy identification [Tirinzoni et al., 2022] and optimal data-collection for policy evaluation [Russo and Pacchiano, 2025], both in the frequentist setting. To our knowledge, no general Bayesian theory is known for continuous recommendation spaces under general priors; current Bayesian fixed-confidence results are limited to finite bandits with Gaussian likelihoods and Gaussian priors [Jang et al., 2024]. While Bayesian ideas drive exploration in finite bandit (with frequentist guarantees), e.g. posterior sampling and top-two methods [Russo and Van Roy, 2014, Russo, 2016, Shang et al., 2020], the most developed Bayesian framework in continuous spaces is Bayesian optimization (BO) [Garnett, 2023], which maintains a posterior over an unknown objective and selects queries via acquisition functions [Hernández-Lobato et al., 2014, Hennig and Schuler, 2012]. BO aims to identify an optimizer, but is typically posed as fixed-budget optimization without a correctness constraint. Wilson [2024] recently introduced a Bayesian (ϵ, δ)-stopping rule for BO, but it is restricted to GP surrogates and does not optimize sample complexity. Despite this progress, no existing method combines three elements: fixed-confidence (ϵ, δ) stopping, continuous recommendations, and a learned exploration and recommender procedure under a Bayesian formulation. **C-ICPE** addresses this gap.

Conclusions. **C-ICPE** is a theory-inspired method for Bayesian fixed-confidence pure exploration with continuous recommendations. On the theoretical side, we establish that the Lagrangian duality and Bellman optimality structure of finite ASHT carries over to continuous spaces under regularity conditions, and prove (ϵ, δ)-correctness under a local closedness assumption that is weaker than the uniqueness condition required in [Russo et al., 2025b]. On the algorithmic side, we show that **C-ICPE** achieves the desired guarantees across different tasks, including a real-world geochemical exploration problem, while using fewer samples than standard optimization baselines. To our knowledge, no prior method combines continuous recommendations, fixed-confidence stopping in a single practical framework. Limitations and broader impact are discussed in Section A.

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.

A. Al Marjani, T. Kocak, and A. Garivier. On the Complexity of All ϵ-Best Arms Identification. In M.-R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak, and G. Tsoumakas, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 13716, pages 317–332, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-26411-5 978-3-031-26412-2. doi: 10.1007/978-3-031-26412-2_20. Series Title: Lecture Notes in Computer Science.

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.

F. Cecchi and N. Hegde. Adaptive active hypothesis testing under limited information. *Advances in Neural Information Processing Systems*, 30, 2017.

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

D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

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.

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.

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.

B. K. Ghosh. A brief history of sequential analysis. *Handbook of sequential analysis*, 1, 1991.

D. Golovin and A. Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011.

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.

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.

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.

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.

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.

T. Kocák and A. Garivier. Best arm identification in spectral bandits. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence*, IJCAI'20, Yokohama, Yokohama, Japan, 2021. ISBN 978-0-9992411-6-5. Number of pages: 7 tex.articleno: 307.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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.

M. Naghshvar, T. Javidi, and K. Chaudhuri. Noisy bayesian active learning. In *2012 50th annual allerton conference on communication, control, and computing (allerton)*, pages 1626–1633. IEEE, 2012.

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. ISSN: 2640-3498.

M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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 and A. Pacchiano. Adaptive exploration for multi-reward multi-policy evaluation. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, editors, *Proceedings of the 42nd international conference on machine learning*, volume 267 of *Proceedings of machine learning research*, pages 52382–52421. PMLR, July 2025.

A. Russo and A. Proutiere. Model-free active exploration in reinforcement learning. In *Advances in neural information processing systems*, volume 36, pages 54740–54753, 2023a.

A. Russo and A. Proutiere. On the sample complexity of representation learning in multi-task bandits with global and

local structure. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 9658–9667, 2023b.

A. Russo and F. Vannella. Multi-reward best policy identification. In *Advances in neural information processing systems*, volume 37, pages 105583–105662, 2024.

A. Russo, Y. Song, and A. Pacchiano. Pure exploration with feedback graphs. In *Proceedings of the 28th international conference on artificial intelligence and statistics*, volume 258 of *Proceedings of machine learning research*, pages 1810–1818. PMLR, 2025a.

A. Russo, R. Welch, and A. Pacchiano. Learning to Explore: An In-Context Learning Approach for Pure Exploration, June 2025b. arXiv:2506.01876 [cs].

D. Russo. Simple Bayesian Algorithms for Best Arm Identification. In *Conference on Learning Theory*, pages 1417–1418. PMLR, June 2016.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4): 1221–1243, 2014.

X. Shang, R. Heide, P. Menard, E. Kaufmann, and M. Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pages 1823–1832. PMLR, 2020.

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. Tirinzoni, A. Al Marjani, and E. Kaufmann. Near instance-optimal pac reinforcement learning for deterministic mdps. In *Advances in neural information processing systems*, volume 35, pages 8785–8798, 2022.

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.

P.-A. Wang, A. Proutiere, K. Ariu, Y. Jedra, and A. Russo. Optimal algorithms for multiplayer multi-armed bandits. In S. Chiappa and R. Calandra, editors, *Proceedings of the twenty third international conference on artificial intelligence and statistics*, volume 108 of *Proceedings of machine learning research*, pages 4120–4129. PMLR, Aug. 2020.

P.-A. Wang, R.-C. Tzeng, and A. Proutiere. Fast Pure Exploration via Frank-Wolfe. In *Advances in Neural Information Processing Systems*, volume 34, pages 5810–5821. Curran Associates, Inc., 2021.

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

Table of Contents for the Appendix

A Limitations and Broader Impact **13**

B Appendix: Theoretical Results **14**

B.1 Problem Modeling . . . 14

B.2 Posterior distribution over the true hypothesis and inference rule optimality . . . 16

B.3 Fixed-confidence setting: formulation and optimal rules . . . 19

B.4 Fixed-confidence setting: (ϵ, δ)-correctness of dual-optimal points . . . 27

B.5 Fixed-confidence setting: zero duality gap via perturbation values . . . 30

B.6 Training-time certification of (ϵ, δ)-correctness . . . 35

B.7 Choice of inference model and reward modeling . . . 38

B.8 Robustness to prior misspecification . . . 42

B.9 Sample Complexity: Value Estimation vs Argmax Localization . . . 44

C Appendix: Algorithms **56**

C.1 History Encoder and Time Pooling Layer . . . 56

C.2 Replay Buffer and Prefix Sampling . . . 56

C.3 Inference Update and Its Regularization . . . 57

C.4 Reward, Critic Update, and Critic Regularization . . . 57

C.5 Actor Rules: TS, TTPS, and TD3 . . . 58

C.6 Cost Update for Fixed Confidence . . . 59

D Appendix: Numerical Results **61**

D.1 Synthetic Benchmarks: description . . . 61

D.2 Synthetic Benchmarks: baselines . . . 64

D.3 Synthetic Benchmarks: numerical results . . . 66

D.4 Synthetic Benchmarks: robustness . . . 75

D.5 Geochemical Exploration: Experimental Details and Numerical Results . . . 79

A. Limitations and Broader Impact

Limitations. *Parametric inference model.* **C-ICPE** models the posterior law of the target $x_\theta^\star$ with a diagonal Gaussian. In the ideal NLL objective, this corresponds to a moment projection: the mean matches $\mathbb{E}[x_\theta^\star|H_t]$ 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 $x_\theta^\star$ 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 the critic evaluates samples from the same inference distribution. Richer posterior families, such as mixtures or 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 $x_\theta^\star$. It is more pronounced in value-gap tasks such as Ackley or 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 our benchmarks this is available in closed form (the shifted minimizer for Ackley, the parameter itself for binary search, the GP maximum for max-value estimation). In general, computing $x^\star(\theta)$ may require numerical optimization, introducing approximation error in the NLL targets. If the selector can only be evaluated approximately or with noise, the inference model may learn a biased posterior, potentially affecting both recommendation quality and stopping calibration. Extending **C-ICPE** to settings where only noisy or approximate selectors are available is an important direction for future work.

Cost calibration. The dual variable c, 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 Section B.8 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. Appendix: Theoretical Results

Roadmap and novelty guide. The theoretical analysis proceeds in four stages. We summarize what is standard and what is new relative to the finite ICPE framework of Russo et al. [2025b].

- **§B.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, has not previously been studied in the pure exploration literature.
- **§B.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. [2025b], 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.
- **§B.4–B.5: Correctness and zero duality gap.** We prove (ϵ, δ)-correctness under a local closedness condition that is strictly weaker than the uniqueness assumption in Russo et al. [2025b]. Our proof uses a subdifferential characterization from Hantoute and López [2008] to derive a contradiction without the monotonicity argument of Russo et al. [2025b]; this is the strongest theorem-level novelty in the appendix. Zero duality gap follows using a perturbation argument [Rockafellar, 1974].
- **§B.6: Model certification.** We introduce a checkpointwise certification protocol based on mixture supermartingales [Kaufmann and Koolen, 2021]. Unlike the pooled approach in Russo et al. [2025b], it tests each frozen checkpoint independently and requires no monotonicity assumption on the training trajectory.
- **§B.7: 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 (Proposition 6). 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 (Propositions 7 and 8).
- **§B.8: Robustness to prior misspecification.** We analyse the robustness to prior misspecification, and what is the predicted impact on sample complexity and accuracy.
- **§B.9: 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 (ϵ, δ)-guarantees.

B.1. Problem Modeling

We specialize to the fixed-confidence ((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).

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 (in our experiments $\mathcal{X} = \mathcal{A}$). For $t \in \mathbb{N}$, define the history space

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

with its product Borel σ-algebra. We also write $\mathcal{H}_\infty \coloneqq \mathcal{Y} \times (\mathcal{A} \times \mathcal{Y})^{\mathbb{N}}$ 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 $(h_t, a) \mapsto P_{\theta,t}(C|h_t, a)$ 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 random history and $\mathcal{F}_t = \sigma(H_t)$. A stopping time τ is defined w.r.t. $(\mathcal{F}_t)_{t\geq 1}$. An inference rule is a sequence of measurable maps $I_t : \mathcal{H}_t \to \mathcal{X}$, and the learner outputs

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

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

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

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

1. *First, if $F(\theta) \coloneqq \arg\max_{x\in\mathcal{X}} f_\theta(x) = \{x^\star(\theta)\}$ 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 $x^\star_{\text{sel}}(\theta) = x^\star(\theta)$ 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 $x^\star_{\text{sel}}(\theta) \coloneqq \arg\min_{x\in F(\theta)} \|x\|^2$ 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 $x^\star_{\text{sel}}(\theta)$, while correctness remains defined through the loss L_θ and the set $\mathcal{X}_\epsilon(\theta)$.

[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.

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.

Fixed-confidence objective. The learner is (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 (ϵ, δ)-correctness:

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

B.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)

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

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

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

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

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

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)**.** *For each state $s < t$, the kernel $p_{\theta,s}(y|h_s, a_s)$ is jointly continuous in (θ, y, h_s, a_s) with strictly positive density. Hence, $(\theta, h_t) \mapsto \ell(\theta, h_t)$ 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 Θ is compact and $\theta \mapsto f$ is continuous and $\theta \mapsto \ell_t(\theta, h')$ is continuous for each $h' \in \mathcal{H}_t$, we have that $\sup_{\theta\in\Theta} |f(\theta)\ell_t(\theta, h_n)| < \infty$, therefore by dominated convergence we have

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

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

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

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 $\mathcal{X}_\epsilon(\theta) = \{x \in \mathcal{X} : L_\theta(x) \le \epsilon\}$. Given a realized history $h \in \mathcal{H}_t$, define the *posterior success probability* of recommending $x \in \mathcal{X}$ as

$$q_t(h, x) := \mathbb{P}_{\theta\sim\nu}^\pi\big(x \in \mathcal{X}_\epsilon(\theta)|H_t = h\big) = R_t\big(\{\theta \in \Theta : L_\theta(x) \le \epsilon\}|h\big), \tag{11}$$

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

$$r_t(h) := \sup_{x\in\mathcal{X}} q_t(h, x). \tag{12}$$

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 $(x, \theta) \mapsto \mathbf{1}\{L_\theta(x) \le \epsilon\}$ 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 $q_t(h, x) = \int_\Theta \mathbf{1}\{L_\theta(x) \le \epsilon\} R_t(\mathrm{d}\theta|h)$ 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 continuous 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 $x_n \to x^\star$. Define $y_n = \mathbf{1}\{x_n \in \mathcal{X}_\epsilon(\theta)\}$. By Fatou's reverse lemma we have

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

  where the last inequality follows from the fact that $\limsup_n y_n \le \mathbf{1}\{x \in \mathcal{X}_\epsilon(\theta)\}$ from the upper semicontinuity. Thus the posterior is upper semicontinuous on $\mathbb{P}_t^\pi$-a.s.
- Define $C_\epsilon = \{(\theta, x) \in \Theta \times \mathcal{X} : L_\theta(x) \le \epsilon\}$. By assumption on L (joint continuity), 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 $\mu(\cdot) := R_t(\cdot|h_n) \otimes \delta_x$. Since $h \mapsto R_t(\cdot|h)$ is weakly continuous (Lemma 2, follows from Assumption 4), by Portmanteau's lemma we have

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

  But $\mu_n(C_\epsilon) = \int_\Theta \mathbf{1}\{L_\theta(x_n) \le \epsilon\} R_t(\mathrm{d}\theta|h_n) = q_t(h_n, x_n)$ and $\mu(C_\epsilon) = \int_\Theta \mathbf{1}\{L_\theta(x) \le \epsilon\} R_t(\mathrm{d}\theta|h) = q_t(h, x)$, 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 (12) is attained (so one may replace $\sup$ by $\max$).

Proposition 1 (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 $\mathbb{P}^\pi_{\theta\sim\nu}\big(I_t(H_t) \in \mathcal{X}_\epsilon(\theta)\big)$ 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,

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

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*

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

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

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

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

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

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

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

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

B.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 $I = (I_t)_{t\ge1}$, and stopping time τ. The objective is

$$\inf_{\pi, I, \tau} \mathbb{E}^{\pi}_{\theta\sim\nu}[\tau] \qquad \text{s.t.} \qquad \mathbb{P}^{\pi}_{\theta\sim\nu}\big(I_\tau(H_\tau) \in \mathcal{X}_\epsilon(\theta)\big) \;\ge\; 1 - \delta, \quad \mathbb{E}^{\pi}_{\theta\sim\nu}[\tau] < \infty. \tag{13}$$

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 $\mathcal{F}_t = \sigma(H_t)$.

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, stopping time τ, and 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. By the tower rule and $\hat{x}_\tau$ being $\sigma(H_\tau)$-measurable,

$$\mathbb{P}^{\pi}_{\theta\sim\nu}(\hat{x}_\tau \in \mathcal{X}_\epsilon(\theta)) = \mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}[\mathbf{1}\{\hat{x}_\tau \in \mathcal{X}_\epsilon(\theta)\} \mid H_\tau]\right] = \mathbb{E}^{\pi}\left[q_\tau(H_\tau, \hat{x}_\tau)\right].$$

□

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 $\lambda \ge 0$, 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 Lemma 5, this can be written as

$$V_\lambda(\pi, I, \tau) = \lambda(1 - \delta) + \mathbb{E}^{\pi}\left[\tau - \lambda q_\tau\big(H_\tau, I_\tau(H_\tau)\big)\right]. \tag{14}$$

The Lagrangian dual of (13) is then

$$\sup_{\lambda\ge0} \inf_{\pi, I, \tau} V_\lambda(\pi, I, \tau). \tag{15}$$

B.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 Proposition 1.)

Since τ is adapted, plugging the optimal inference rule into Eq. (14) yields the simplified dual objective

$$\sup_{\lambda \geq 0} \inf_{\pi, \tau} \lambda(1-\delta) + \mathbb{E}^\pi \left[\tau - \lambda r_\tau(H_\tau)\right]. \tag{16}$$

B.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 $\bar{\pi}$, (i) $\bar{\tau} = \tau$ a.s., and (ii) the stopped history $H_{\bar{\tau}}$ has the same distribution as H_τ under π. In particular, for every $\lambda \geq 0$, $V_\lambda(\pi, I, \tau) = V_\lambda(\bar{\pi}, I, \bar{\tau})$.*

Proof. Since τ is a stopping time w.r.t. $\mathcal{F}_t = \sigma(H_t)$, the event $\{\tau = t\}$ belongs to $\mathcal{F}_t$; hence there exists a measurable set $S_t \subset \mathcal{H}_t$ such that $\{\tau = t\} = \{H_t \in S_t\}$. Define $\bar{\pi}$ as follows: at time t, given history $h \in \mathcal{H}_t$,

$$\bar{\pi}_t(a_{\text{stop}}|h) = \mathbf{1}\{h \in S_t\}, \qquad \bar{\pi}_t(\cdot|h) = \pi_t(\cdot|h) \text{ on } \mathcal{A} \text{ when } h \notin S_t.$$

Then $\{A_t = a_{\text{stop}}\} = \{H_t \in S_t\} = \{\tau = t\}$, so $\bar{\tau} = \tau$ a.s. Moreover, on the event $\{\tau > t\}$ the action distribution and observation kernel coincide with those under π, so the induced law of $(H_t)_{t \leq \tau}$ is the same; in particular $H_{\bar{\tau}}$ under $\bar{\pi}$ has the same distribution as H_τ under π. Hence, one can easily show that the equality $V_\lambda(\pi, I, \tau) = V_\lambda(\bar{\pi}, I, \bar{\tau})$ follows. □

B.3.3. Optimal Policy

Lemma 6 shows that (for fixed λ) the inner problem in Eq. (16) 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 $-\lambda\, r_t(h)$.

Define the optimal cost-to-go (for fixed λ) from a history $h \in \mathcal{H}_t$ as

$$V_t^\star(h; \lambda) := \inf_{\bar{\pi} = (\bar{\pi}_i)_{i \geq t}} \mathbb{E}_{\theta \sim \nu}^{\bar{\pi}} \left[\sum_{s=t}^{\bar{\tau}-1} 1 - \lambda r_{\bar{\tau}}(H_{\bar{\tau}}) \middle| H_t = h \right], \tag{17}$$

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., 2025b], 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., 2025b], a standard decomposition yields the Bellman optimality relation

$$V_t^\star(h; \lambda) = \min\left\{ Q_{t,\text{stop}}^\star(h; \lambda), \inf_{a \in \mathcal{A}} Q_{t,\text{cont}}^\star(h, a; \lambda) \right\}. \tag{18}$$

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*

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

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

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

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 (Lemma 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 $W_t^{(0)}(h) := 0$ for all t, h. Recursively for $n \geq 0$, define

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

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

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

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 $U_t(h) := \sup_{n \geq 0} W_t^{(n)}(h)$ 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 Lemma 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 $\mu_m := \delta_{a_m} \otimes \bar{P}_t(\cdot \mid h, a_m)$ and $\mu := \delta_a \otimes \bar{P}_t(\cdot \mid h, a)$. By Lemma 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 $\mu_{m_k} := \delta_{(h_{m_k}, a_{m_k})} \otimes \bar{P}_t(\cdot \mid h_{m_k}, a_{m_k})$ and $\mu := \delta_{(h,a)} \otimes \bar{P}_t(\cdot \mid h, a)$. 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.

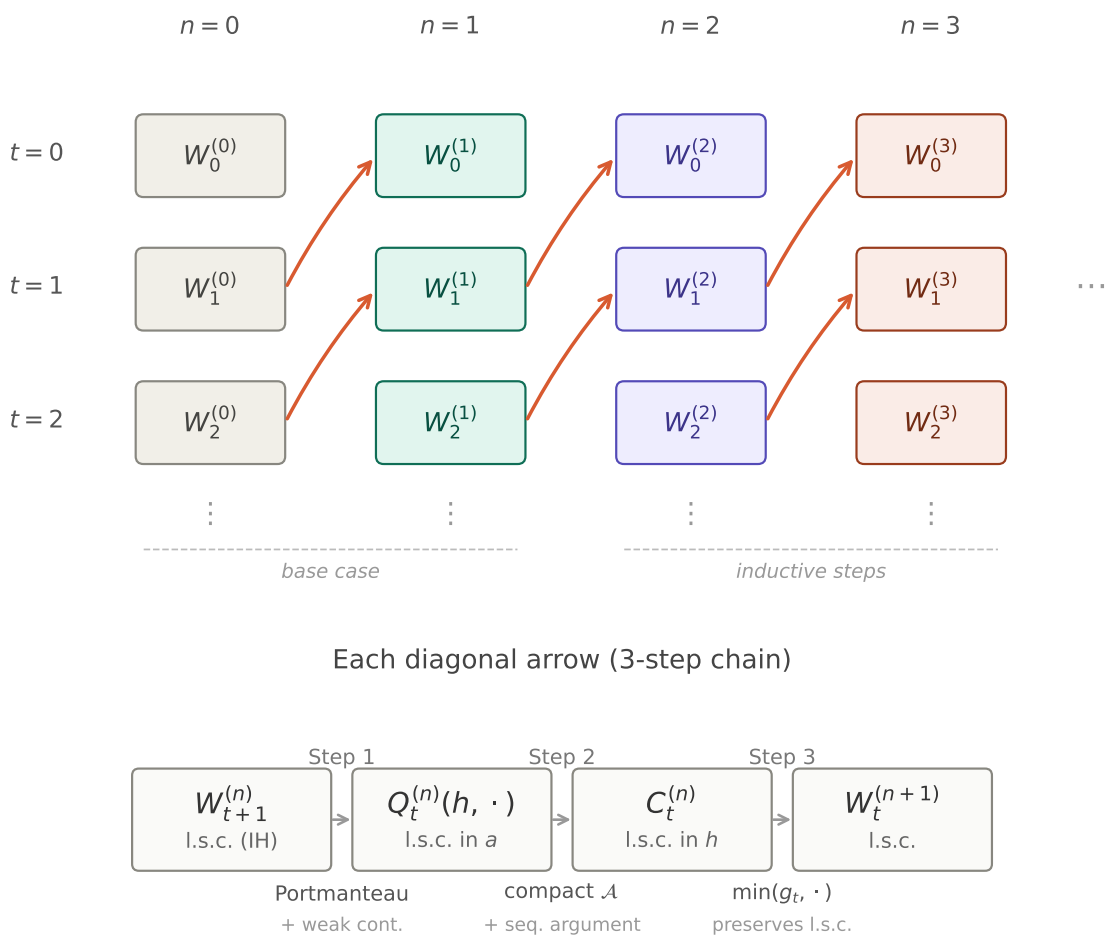


Figure 5. Induction diagram used in the proof of Lemma 8.

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*

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

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 Lemma 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 $m_n = \inf_a Q_t^{(n)}(h,a) \le \inf_a Q_t^U(h,a) = C_t^U(h)$, and thus $m \le C_t^U(h)$. 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

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

Therefore $\liminf_k m_{n_k} = m \ge Q_t^{(i_0)}(h,a)$. Since this holds for any i_0, we have $m \ge \sup_{i_0} Q_t^{(i_0)}(h,a) = Q_t^U(h,a) \ge C_t^U(h)$, 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 B.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 $U_t(h) \le g_t(h)$ since $U_t(h) = \min\{g_t(h), C_t^U(h)\}$.
2. if π continues with some action a, then

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

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

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

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

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 $g_\tau \ge 0$, 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 $U_t(h) \le \lim_{N\to\infty} J_N^\pi(h) := J^\pi(h)$, and taking infimum over π we get $U_t(h) \le W_t^\star(h)$.

Step 2 (equality). From Lemma 9, $Q_t^U(h,a)$ is l.s.c. on $\mathcal{A}$ for each h, and $\mathcal{A}$ is compact. Furthemore $Q_t^U(h,a)$ 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 $g_t(h) \le C_t^U(h)$, 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

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

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

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

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 2 (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 (18), 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 (18), provided the actor approximates the minimizer of $Q^\star_{t,\mathrm{cont}}(h,\cdot)$ and the critic estimates are calibrated.

B.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*

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

and the shaped return

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

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 3 (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. □

B.4. Fixed-confidence setting: (ϵ,δ)-correctness of dual-optimal points

In this section we provide fixed-confidence guarantees for the continuous ICPE setting. Compared to the correctness argument in [Russo et al., 2025b], 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. [2025b] 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., 2025b]), 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, $g(\lambda) \geq -\delta\lambda$. Let (c,ρ) be a point satisfying Assumption 5. Then $\rho > 1-\delta$, and thus $g(\lambda) \leq c - \lambda\eta$ for $\eta = -(1-\delta-\rho)$. Since $\eta > 0$, and $-\lambda\delta \leq g(\lambda) \leq c - \lambda\eta$, 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 $g(\lambda^\star) = \max_{\lambda\geq 0} g(\lambda)$. □

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., 2025b] 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 4 (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 a lower bound on any subdifferential $d \in \partial f(x)$.

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

Theorem B.2 ((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*

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

where $I^\star_t(h) \in \arg\max_h q_t(h,x)$ 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 Lemma 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 $\rho \le 1-\delta-\alpha$ for optimal policies. Under (**HYP**), we claim that there exists $\alpha > 0$ such that $\rho \le 1-\delta-\alpha$ 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 $c \le g(\lambda^\star)+\epsilon+\lambda^\star\delta$, 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 $\rho \le 1-\delta-\alpha/2$. Always under (**HYP**), we now claim that for near-optimal points we actually have $\rho \le 1-\delta-\alpha/2$. 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

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

Since $\rho_n \in [0,1]$ and $c_n \le g(\lambda^\star)+1/n+\lambda^\star\delta$, 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

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

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

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 $1-\delta-\alpha/2 \ge \rho$ and $c \le g(0)+\epsilon$. Consider then the following two cases:

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

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

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

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 $f(\lambda) = -g(\lambda)$. From Lemma 11 then f is convex. Since $\lambda^\star > 0$, it lies in the interior of $\mathrm{dom}(f)$. Then, by Proposition 4 , 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 2. *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.*

B.5. Fixed-confidence setting: 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

$$P^\star := \inf_{\pi \in \mathcal{T}} c(\pi) \qquad \text{s.t.} \qquad \rho(\pi) \geq 1-\delta. \tag{19}$$

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

$$D^\star := \sup_{\lambda \geq 0} g(\lambda), \qquad g(\lambda) := \inf_{\pi \in \mathcal{T}} \{c(\pi) + \lambda(1-\delta-\rho(\pi))\}. \tag{20}$$

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 $\mathbb{R}^2$.

B.5.1. 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

$$\varphi(u) := \inf\{c : (c,\rho) \in \mathcal{K},\ 1-\delta-\rho \leq u\}, \qquad u \in \mathbb{R}, \tag{21}$$

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 $\mathcal{K}$,

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

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

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

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 B.3 (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,

$$\inf_{\pi\in\mathcal{T}:\rho(\pi)\ge 1-\delta} c(\pi) = \sup_{\lambda\ge 0}\inf_{\pi\in\mathcal{T}} \left\{c(\pi) + \lambda(1-\delta-\rho(\pi))\right\}. \tag{22}$$

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

Proof. By Lemma 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 $\lambda^\star := -s^\star \ge 0$. 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. (22). □

Remark 3. *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, Lemma 12 shows that strict feasibility implies continuity, hence lower semicontinuity, at 0. We state Theorem B.3 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.*

B.5.2. 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 Theorem B.3, $g(\lambda^\star) = P^\star$. 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

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

Theorem B.4 (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

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

Proof. By Lemma 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

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

By Lemma 5, the last display equals

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

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

Remark 4 (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.

B.6. Training-time certification of (ϵ, δ)-correctness

We now give a training-time correctness guarantee for ICPE that treats the learned policy, stopping rule, and recommender as a black box. The guarantee is designed for the practical workflow used in our experiments: every few training epochs we freeze the current model, evaluate it on a finite batch of fresh task realizations, and certify whether the checkpoint satisfies the (ϵ, δ)-guarantees.

The key point is that certification must account for repeated testing. A fixed-level certificate applied repeatedly at many checkpoints is not valid without correction. We therefore assign a certification budget α_m to checkpoint m, with total budget at most $\alpha \in (0, 1)$, and certify each checkpoint using a checkpointwise mixture martingale.

Fix a target success threshold

$$q^\star = 1 - \delta,$$

and an additional parameter $\alpha \in (0, 1)$ that controls the probability of ever falsely certifying a checkpoint.

Protocol. We index certification checkpoints by $m \geq 1$. In the experiments, for example, checkpoint m may correspond to every 2500 training epochs. At checkpoint m, let

$$w_m = (\phi_m, \psi_m, \vartheta_m)$$

denote the frozen parameters of the inference model I_ϕ, the critic Q_ψ and the actor π_ϑ. We denote by I_m, π_m, and τ_m, respectively, the induced inference rule, sampling policy, and stopping time. The probability of success of checkpoint m is

$$p_m \coloneqq \mathbb{P}^{\pi_m}_{\theta\sim\nu}(I_m(H_{\tau_m}) \in \mathcal{X}_\epsilon(\theta))\,.$$

Thus p_m is the probability of success of the actual recommender rule used by the frozen checkpoint.

For each checkpoint m, after w_m is fixed, we draw i.i.d. a batch of B evaluation episodes. For $j = 1, \ldots, B$, let $\theta^{(m,j)} \sim \nu$ be the task in the j-th evaluation episode for checkpoint m, and let $H^{(m,j)}_{\tau_m}$ be the stopped history obtained by running the frozen snapshot w_m on that task. Define the success indicator

$$Z_{m,j} \coloneqq \mathbf{1}\Big\{I_m\Big(H^{(m,j)}_{\tau_m}\Big) \in \mathcal{X}_\epsilon(\theta^{(m,j)})\Big\}\,.$$

We write

$$S_{m,t} \coloneqq \sum_{j=1}^{t} Z_{m,j}, \qquad t = 0, 1, \ldots, B.$$

Hence, For each checkpoint m, conditional on the training history before evaluating checkpoint m and conditional on the frozen checkpoint w_m, the variables

$$Z_{m,1}, \ldots, Z_{m,B}$$

are independent and identically distributed as $\mathrm{Ber}(p_m)$. Moreover, the certification batch at checkpoint m is not reused for training updates before the certification decision for checkpoint m is made.

Checkpointwise mixture martingale. At checkpoint m, we test the null hypothesis

$$H_{0,m}: \quad p_m \leq q^\star.$$

Let Π_m be a probability distribution on $[q^\star, 1]$, chosen before observing the certification batch at checkpoint m. For $r \in [q^\star, 1]$, define

$$L_{m,t}(r) \coloneqq \left(\frac{r}{q^\star}\right)^{S_{m,t}} \left(\frac{1-r}{1-q^\star}\right)^{t-S_{m,t}}, \qquad t = 0, 1, \ldots, B,$$

with the usual convention $0^0 = 1$. The checkpointwise mixture martingale is

$$M_{m,t} \coloneqq \int_{q^\star}^{1} L_{m,t}(r)\,\Pi_m(dr), \qquad M_{m,0} = 1.$$

In implementation, Π_m may be a finite grid distribution, in which case no numerical integration is needed. If

$$\Pi_m = \sum_{\ell=1}^{L} \omega_{m,\ell}\, \delta_{r_{m,\ell}}, \qquad \sum_{\ell=1}^{L} \omega_{m,\ell} = 1,$$

then

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

Certification budgets. Let $(\alpha_m)_{m\geq 1}$ be nonnegative certification budgets, chosen predictably before observing the certification batch at checkpoint m, and satisfying

$$\sum_{m\geq 1} \alpha_m \leq \alpha.$$

We declare checkpoint m certified if

$$\max_{1\leq t\leq B} M_{m,t} \geq \alpha_m^{-1}.$$

Let

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

be the set of certified checkpoints.

Proposition 5 (Training-time certification by checkpointwise mixture martingales)**.** *Suppose that the certification budgets satisfy $\sum_{m\geq 1} \alpha_m \leq \alpha$ almost surely. Then*

$$\mathbb{P}(\forall m \in \mathcal{C},\ p_m > q^\star) \geq 1-\alpha.$$

Equivalently, every certified checkpoint is (ϵ,δ)-correct, and, after training, any adaptively selected checkpoint

$$\hat{m} \in \mathcal{C}$$

is (ϵ,δ)-correct with confidence at least $1-\alpha$.

Proof. Fix a checkpoint m. We condition on the training history before the certification batch for checkpoint m, and on the frozen checkpoint w_m. Under this conditioning, p_m, Π_m, and α_m are fixed, and

$$Z_{m,1},\ldots,Z_{m,B} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p_m).$$

Assume the null hypothesis

$$H_{0,m} : p_m \leq q^\star$$

holds. Fix any $r \in [q^\star, 1]$. We show that $\{L_{m,t}(r)\}_{t=0}^{B}$ is a nonnegative supermartingale under $H_{0,m}$. Let $\mathcal{F}_{m,t}$ be the sigma-field generated by the training history, the frozen checkpoint w_m, and the first t certification outcomes

$$Z_{m,1},\ldots,Z_{m,t}.$$

For $t < B$,

$$\frac{L_{m,t+1}(r)}{L_{m,t}(r)} = \begin{cases} r/q^\star, & Z_{m,t+1} = 1, \\ (1-r)/(1-q^\star), & Z_{m,t+1} = 0. \end{cases}$$

Therefore,

$$\mathbb{E}\left[\frac{L_{m,t+1}(r)}{L_{m,t}(r)} \,\middle|\, \mathcal{F}_{m,t}\right] = p_m \frac{r}{q^\star} + (1-p_m)\frac{1-r}{1-q^\star}.$$

A direct calculation gives

$$p_m \frac{r}{q^\star} + (1-p_m)\frac{1-r}{1-q^\star} = 1 + \frac{(p_m - q^\star)(r-q^\star)}{q^\star(1-q^\star)}.$$

Since $p_m \le q^\star$ and $r \ge q^\star$, the final term is nonpositive. Hence

$$\mathbb{E}\left[\frac{L_{m,t+1}(r)}{L_{m,t}(r)} \,\middle|\, \mathcal{F}_{m,t}\right] \le 1.$$

Thus

$$\mathbb{E}[L_{m,t+1}(r) \mid \mathcal{F}_{m,t}] \le L_{m,t}(r),$$

so $\{L_{m,t}(r)\}_{t=0}^{B}$ is a nonnegative supermartingale.

Because $M_{m,t}$ is a mixture of the nonnegative supermartingales $L_{m,t}(r)$, Tonelli's theorem gives

$$\mathbb{E}[M_{m,t+1} \mid \mathcal{F}_{m,t}] = \int_{q^\star}^{1} \mathbb{E}[L_{m,t+1}(r) \mid \mathcal{F}_{m,t}]\,\Pi_m(dr) \le \int_{q^\star}^{1} L_{m,t}(r)\,\Pi_m(dr) = M_{m,t}.$$

Therefore $\{M_{m,t}\}_{t=0}^{B}$ is a nonnegative supermartingale with $M_{m,0} = 1$. By Ville's inequality,

$$\mathbb{P}\left(\max_{1\le t\le B} M_{m,t} \ge \alpha_m^{-1} \,\middle|\, \text{past}, w_m, H_{0,m}\right) \le \alpha_m.$$

Let

$$A_m \coloneqq \left\{p_m \le q^\star \text{ and } \max_{1\le t\le B} M_{m,t} \ge \alpha_m^{-1}\right\}$$

be the event that checkpoint m is falsely certified. From the conditional bound above,

$$\mathbb{P}(A_m \mid \text{past}) \le \alpha_m.$$

Taking expectations,

$$\mathbb{P}(A_m) \le \mathbb{E}[\alpha_m].$$

By the union bound,

$$\mathbb{P}(\exists m \ge 1 :\ A_m) \le \sum_{m\ge 1} \mathbb{P}(A_m) \le \mathbb{E}\left[\sum_{m\ge 1} \alpha_m\right] \le \alpha.$$

Thus, with probability at least $1-\alpha$, no checkpoint with $p_m \le q^\star$ is certified. Equivalently,

$$p_m > q^\star \qquad \forall m \in \mathcal{C}.$$

Taking $q^\star = 1-\delta$ gives the stated (ϵ, δ)-correctness guarantee. □

Remark 5 (What the guarantee certifies)**.** *The guarantee is simultaneous over all certified checkpoints:*

$$\mathbb{P}(\forall m \in \mathcal{C},\ p_m > 1-\delta) \ge 1-\alpha.$$

Thus the final returned model need not be the first certified model. It may be any checkpoint selected after training, as long as it belongs to $\mathcal{C}$. The result does not certify checkpoints that were never certified, nor does it imply that all checkpoints after the first certified checkpoint are correct.

Remark 6 (Relation to the original ICPE training-time argument)**.** *The original finite-hypothesis ICPE argument pools evidence across checkpoints and tests a global null of the form*

$$\sup_m p_m \le q^\star.$$

That pooled martingale can show that training has produced evidence against the hypothesis that all checkpoints are bad. However, by itself it does not certify an arbitrary later checkpoint unless one adds a persistence or monotonicity assumption on the sequence (p_m), as done in [Russo et al., 2025b].

The checkpointwise martingale above is different. It tests each frozen checkpoint separately, spends error probability across checkpoints, and therefore certifies the checkpoints that actually pass the test. This matches the practical workflow in which we may check a model, fail to certify it, continue training, and later certify a different checkpoint. No monotonicity or persistence assumption on the training trajectory is required.

Remark 7 (Choice of certification budgets)**.** *If all checkpoints are treated symmetrically and a maximum number $M_{\max}$ of checks is fixed in advance, the uniform allocation*

$$\alpha_m = \frac{\alpha}{M_{\max}}$$

is the simplest choice. If later checkpoints are expected to be better, one can use the exponentially back-loaded allocation

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

This preserves the same correctness theorem because the proof only uses

$$\sum_m \alpha_m \le \alpha.$$

The choice of schedule affects power, not validity. Larger α_m makes checkpoint m easier to certify, so back-loading the budget gives more power to later checkpoints at the expense of earlier ones.

Remark 8 (Finite-grid implementation)**.** *In practice we use a finite grid*

$$q^\star \le r_1 < \cdots < r_L \le 1$$

with weights $\omega_1, \ldots, \omega_L$. Then

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

which is easy to compute and remains a valid mixture martingale. The grid should place mass on plausible alternatives above $q^\star$, for example $r_\ell \in \{0.91, 0.92, 0.93, 0.94, 0.95, 0.97\}$ when $q^\star = 0.9$. Placing mass closer to $q^\star$ improves power for small margins but requires larger batches; placing mass farther above $q^\star$ improves power when the checkpoint is substantially better than the target.

B.7. 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

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

and any measurable maximizer is denoted by

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

The implementation does not parameterize $I_t^\star$ directly. Instead, it uses a stochastic selector $\hat{I}_t(\cdot|h)$, represented by a diagonal Gaussian base distribution

$$\hat{I}_t(\cdot|h) = \mathcal{N}(\mu_t(h), \Sigma_t(h)).$$

The recommendation is the mean $\mu_t(h)$, while the critic evaluates samples from $\hat{I}_t(\cdot|h)$.

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

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

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 reward $\hat{r}_t(h)$ evaluates the stochastic selector we actually use. 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

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

Proposition 6 (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

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

Therefore

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

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

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

For the second claim, by Lemma 14,

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

Since $I_t^\star(h)$ maximizes $q_t(h, \cdot)$,

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

Let

$$E_h \coloneqq \{\|\hat{X}_t - I_t^\star(h)\| \leq R_t(h)\}.$$

On E_h, Lipschitzness gives

$$q_t(h, I_t^\star(h)) - q_t(h, \hat{X}_t) \leq L_t(h)\|\hat{X}_t - I_t^\star(h)\|.$$

On E_h^c, the same difference is at most 1. Hence

$$r_t(h) - \hat{r}_t(h) \leq L_t(h)\mathbb{E}[\|\hat{X}_t - I_t^\star(h)\||H_t = h] + \mathbb{P}(E_h^c|H_t = h).$$

Jensen's inequality gives

$$\mathbb{E}[\|\hat{X}_t - I_t^\star(h)\||H_t = h] \leq \sqrt{D_t(h, I_t^\star(h))},$$

and Markov's inequality gives

$$\mathbb{P}(E_h^c|H_t = h) \leq \frac{D_t(h, I_t^\star(h))}{R_t(h)^2}.$$

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 $x_\theta^\star$, 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 $I_t^\star(h)$ 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.

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

$$Z \coloneqq x_\theta^\star$$

be the selected zero-loss 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

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

Let

$$\mu_t^{\mathrm{NLL}}(h) \coloneqq \mathbb{E}[Z|H_t = h], \qquad \Sigma_t^{\mathrm{NLL}}(h) \coloneqq \mathrm{Cov}(Z|H_t = h).$$

Proposition 7 (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*

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

If the covariance is restricted to be diagonal, the optimal mean is still $\mu_t^{\mathrm{NLL}}(h)$ 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 8 (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:

$$r_t(h)-q_t(h,\hat\mu_t(h)) = q_t(h,I_t^\star(h))-q_t(h,\hat\mu_t(h)) \le L_t(h)\|\hat\mu_t(h)-I_t^\star(h)\|.$$

□

The proposition clarifies what is, and is not, implied by the Gaussian NLL. 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

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

near $x_\theta^\star$, then

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

Remark 9 (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.,*

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

and suppose $x^\star_\theta | H_t = h$ 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.

B.8. 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

$$s(\theta) := \mathbb{P}_\theta^{\pi_\phi}\big(\mu_\phi(H_{\hat\tau}) \in \mathcal{X}_\epsilon(\theta)\big) \in [0,1], \qquad t(\theta) := \mathbb{E}_\theta^{\pi_\phi}[\hat\tau] \in [0,\infty].$$

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

$$p^{(\nu)} \geq 1 - \delta.$$

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 15 (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

$$p^{(\eta)} = \int_\Theta \mathbb{P}_\theta^{\pi_\phi}\big(\mu_\phi(H_{\hat\tau}) \in \mathcal{X}_\epsilon(\theta)\big)\,\eta(d\theta) = \int_\Theta s(\theta)\,\eta(d\theta).$$

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 9 (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 $|p' - p| \le \sqrt{\chi^2(\nu'\|\nu)\,\mathrm{Var}_\nu(s)}$.. Hence $p \ge 1-\delta$ implies*

$$p' \ge 1 - \delta - \sqrt{\chi^2(\nu'\|\nu)\delta}.$$

(c) *For arbitrary priors, $|p' - p| \le \mathrm{TV}(\nu', \nu)$. Consequently,*

$$p' \ge 1 - \delta - \mathrm{TV}(\nu', \nu).$$

By Pinsker's inequality, this also gives

$$p' \ge 1 - \delta - \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\nu'\|\nu)}$$

whenever $D_{\mathrm{KL}}(\nu'\|\nu) < \infty$, and the analogous bound with the KL arguments reversed.

(d) *Let $G \subseteq \Theta$ be measurable. If $s(\theta) \ge 1-\eta$ on G, then*

$$p' \ge (1-\eta)\nu'(G).$$

Moreover, if $p \ge 1-\delta$ and

$$G_\eta := \{\theta : s(\theta) \ge 1-\eta\},$$

then

$$\nu(G_\eta) \ge 1 - \frac{\delta}{\eta}, \qquad p' \ge (1-\eta)\nu'(G_\eta).$$

Proof. Let $e(\theta) = 1 - s(\theta) \in [0,1]$.

For (a), if $w = d\nu'/d\nu$, then

$$1 - p' = \mathbb{E}_{\nu'}[e] = \mathbb{E}_\nu[we] \le \|w\|_\infty \mathbb{E}_\nu[e] = C(1-p).$$

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

$$|p' - p| \le \sqrt{\mathbb{E}_\nu[(w-1)^2]}\,\sqrt{\mathrm{Var}_\nu(s)} = \sqrt{\chi^2(\nu'\|\nu)\,\mathrm{Var}_\nu(s)}.$$

Since $0 \le s \le 1$, $s^2 \le s$, hence

$$\mathrm{Var}_\nu(s) \le p(1-p) \le 1-p.$$

If $p \ge 1-\delta$, then $\mathrm{Var}_\nu(s) \le \delta$, giving the displayed bound.

For (c), since $s \in [0,1]$,

$$|p' - p| = \left| \int s\, d(\nu' - \nu) \right| \le \mathrm{TV}(\nu', \nu).$$

Pinsker's inequality gives the KL consequences.

For (d),

$$p' = \int_G s\, d\nu' + \int_{G^c} s\, d\nu' \ge (1-\eta)\nu'(G).$$

If $G = G_\eta$, then $G_\eta^c = \{e > \eta\}$. Since $\mathbb{E}_\nu[e] \le \delta$, Markov's inequality gives

$$\nu(G_\eta^c) \le \frac{\delta}{\eta}.$$

□

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 10 (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 $T' \le CT$.*

(b) *If $\nu' \ll \nu$, $\chi^2(\nu'\|\nu) < \infty$, and $t \in L^2(\nu)$, then*

$$|T' - T| \le \sqrt{\chi^2(\nu'\|\nu)\,\mathrm{Var}_\nu(t)}.$$

(c) *If $\hat{\tau} \le T_{\max}$ almost surely under every θ, then*

$$|T' - T| \le T_{\max}\,\mathrm{TV}(\nu', \nu).$$

(d) *If $\hat{\tau} \le T_{\max}$ almost surely and $t(\theta) \le \tau_0$ on a measurable set $F \subseteq \Theta$, then*

$$T' \le \tau_0 + (T_{\max} - \tau_0)\nu'(F^c).$$

Proof. The first claim follows from

$$T' = \mathbb{E}_\nu[wt] \le \|w\|_\infty \mathbb{E}_\nu[t] = CT.$$

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 $t \le \tau_0$ on F and $t \le T_{\max}$ everywhere. □

Remark 10 (Beta–Uniform shifts)**.** *For one-dimensional shifts with $\nu = \mathrm{Unif}(0,1)$ and $\nu' = \mathrm{Beta}(\alpha,\beta)$, the constants in Proposition 9 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 $\alpha, \beta \ge 1$, 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.

B.9. Sample Complexity: Value Estimation vs Argmax Localization

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. To compare the sample complexity of argmax localization vs max-value estimation, we use the fact that under regularity assumptions we approximately have $r \sim \sqrt{\epsilon}$. This follows from a Taylor's expansion

$$F(X) = F(X^\star) + \frac{1}{2}(X - X^\star)^\top \nabla^2 F(X^\star)(X - X^\star) + o(\|X - X^\star\|^2).$$

from which we find $F(X) - F(X^\star) \approx c\|X - X^\star\|_2^2$ for a suitable constant. Therefore, localizing to radius r gives an error of roughly cr^2 on the value. Therefore, for ϵ accuracy on the value, we can take the radius to be $r \sim \sqrt{\epsilon}$.

Then, we have the following main result.

Theorem B.5 (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 β_η is defined in Lemma 17. 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 $\epsilon > 0$,*
$$T^{\mathrm{Bayes}}_{\mathrm{val},\nu}(\epsilon,\delta) \ge c\,\frac{\sigma^2}{\epsilon^2}\log\frac{1}{\delta} - c'.$$

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.

Proof. Set $\alpha \in (\beta_\eta, \delta)$. Since $\beta_\eta < \alpha$, Lemma 17 provides deterministic constants $\rho, \mu, L, M, \Gamma > 0$ and an event $\mathcal{E}_\delta$ with $\mathbb{P}(\mathcal{E}_\alpha) \ge 1-\delta$.

*Part 1.*We apply Theorem B.7 applies with $r = \sqrt{\epsilon}$, giving

$$T^{\mathrm{Bayes}}_{\mathrm{arg},\nu}(\sqrt{\epsilon},\delta) \le B + O\left(\sigma^2\epsilon^{-3/2}\log\log\frac{1}{\epsilon}\right).$$

Part 2. By Theorem B.6, with $C_- = \exp(-d/(2\ell_-^2))$ and $C_+ = \exp(d/(4\ell_-^2))$,

$$T^{\text{Bayes}}_{\text{val},\nu}(\epsilon,\delta) \geq \frac{\sigma^2}{C_+^2}\left[\frac{C_-^2\, z^2_{1-\delta/2}}{\epsilon^2} - 1\right]_+ .$$

For fixed δ and small ϵ, using $z^2_{1-\delta/2} \geq c_0 \log(1/\delta)$, this gives

$$T^{\text{Bayes}}_{\text{val},\nu}(\epsilon,\delta) \geq c\,\frac{\sigma^2}{\epsilon^2}\log\frac{1}{\delta} - c'.$$

Separation.

$$\frac{T^{\text{Bayes}}_{\text{val},\nu}(\epsilon,\delta)}{T^{\text{Bayes}}_{\text{arg},\nu}(\sqrt{\epsilon},\delta)} \geq \frac{c\,\sigma^2\epsilon^{-2}\log(1/\delta) - c'}{B + C\,\sigma^2\,\epsilon^{-3/2}\log\log(1/\epsilon)} \longrightarrow \infty \qquad \text{as } \epsilon \downarrow 0,$$

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.

B.9.1. Max-value estimation complexity

To derive a lower bound on $T_{\text{val},\nu}(\epsilon,\delta)$, 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).

Theorem B.6. *Consider the problem of estimating $F^\star = \max_{x\in D} F(x)$ where $D = [0,1]^d$, $F \mid \Lambda \sim \text{GP}(0, k_\Lambda)$ 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 $\xi_t \sim \mathcal{N}(0,1)$. If the algorithm outputs $\hat{v}$ at some stopping time τ satisfying $\mathbb{P}_{\Lambda,F}(|\hat{v} - F^\star| \leq \epsilon) \geq 1-\delta$, then we say that the algorithm is (ϵ,δ)-correct. Then, for any (ϵ,δ)-correct algorithm we have that*

$$\inf_{\mathcal{A}:(\epsilon,\delta)-\textit{correct}} \mathbb{E}[\tau] \geq \frac{\sigma^2}{C_+^2}\left[\frac{C_-^2 z^2_{1-\delta/2}}{\epsilon^2} - 1\right]_+ ,$$

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 $X = \frac{\text{Cov}(X,Y)}{\text{Var}(Y)}Y + W$, 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

$$\Theta_\ell \coloneqq \frac{1}{s_\ell}\int_D F(x)\mathrm{d}x, \qquad s_\ell^2 \coloneqq \int_{D\times D} k_\ell(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

[2] Inverse of the CDF of a standard normal distribution.

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 $\theta' \neq \theta$ we have

$$V(\theta') \geq c_\ell(x_\theta)\theta' + W_\ell(x_\theta) = V(\theta) + c_\ell(x_\theta)(\theta' - \theta).$$

Similarly,

$$V(\theta) \geq c_\ell(x_{\theta'})\theta + W_\ell(x_{\theta'}) = V(\theta') + c_\ell(x_{\theta'})(\theta - \theta').$$

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 $k_\ell(x,y) \leq 1 \Rightarrow c_\ell(x) \leq 1/s_\ell$ and $k_\ell(x,y) \geq \exp\left(-d/(2\ell^2)\right) \Rightarrow c_\ell(x) \geq \exp\left(-d/(2\ell^2)\right)/s_\ell$. Since $s_\ell^2 \geq \exp\left(-d/(2\ell^2)\right)$ and $s_\ell^2 \leq 1$, we find

$$\underbrace{\exp\left(-d/2\ell_-^2)\right)}_{=:C_-} \leq c_\ell(x) \leq \underbrace{\frac{1}{\exp\left(-d/(4\ell_-^2)\right)}}_{=:C_+}.$$

Then, from the bounds above on V we find

$$V(\theta') \geq V(\theta) + C_-(\theta' - \theta), \qquad V(\theta) \geq V(\theta') - C_+(\theta' - \theta),$$

leading to

$$C_-(\theta' - \theta) \leq V(\theta') - V(\theta) \leq C_+(\theta' - \theta).$$

Therefore V is bi-Lipschitz, and thus invertible. Choosing $u = V(\theta)$ and $v = V(\theta')$, we obtain

$$|V^{-1}(u) - V^{-1}(v)| \leq \frac{1}{C_-}|u - v|.$$

Therefore, with $\hat{v} = V(\hat{\Theta})$ and $F^\star = V(\Theta)$ we get

$$|\hat{\Theta} - \Theta| \leq \frac{1}{C_-}|\hat{v} - F^\star|.$$

Hence, assuming the algorithm has access to W_ℓ, we can obtain a lower bound on the sample complexity by using Lemma 16. For any (ϵ', δ)-algorithm that estimates Θ, we can choose $\epsilon' = \epsilon/C_-$ to obtain $|V^{-1}(\hat{\Theta}) - V^{-1}(\Theta)| \leq \epsilon'$ at the stopping time. Therefore,

$$\inf_{\mathcal{A}:(\epsilon,\delta)-\text{correct}} \mathbb{E}[\tau] \geq \frac{\sigma^2}{C_+^2}\left[\frac{C_-^2 z_{1-\delta/2}^2}{\epsilon^2} - 1\right]_+,$$

□

Lemma 16 (Scalar estimation lemma)**.** *Let $\Theta \sim \mathcal{N}(0,1)$. Suppose to observe $Z_t = C_t\Theta + \xi_t$ with $|C_t| \le C$ chosen adaptively and $\xi_t \sim \mathcal{N}(0,\sigma^2)$. 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*

$$\inf_{\mathcal{A}:(\epsilon,\delta)-correct} \mathbb{E}[\tau] \ge \frac{\sigma^2}{C^2}\left[\frac{z_{1-\delta/2}^2}{\epsilon^2} - 1\right]_+,$$

where $z_{1-\delta/2} = \Phi^{-1}(1-\delta/2)$.

Proof. Let $\mathcal{H}_t$ denote the filtration history after t observations. Since everything is Gaussian, also the posterior law $\Theta \mid \mathcal{H}_t$ is Gaussian, of parameter (m_t, v_t). In particular, the precision is

$$P_t = v_t^{-1} = 1 + \frac{1}{\sigma^2}\sum_{s=1}^{t} C_s^2.$$

After stopping, for any estimate u we have

$$\mathbb{P}(|u-\Theta| \le \epsilon \mid \mathcal{H}_\tau) = \mathbb{P}(\Theta \in [u-\epsilon, u+\epsilon] \mid \mathcal{H}_\tau).$$

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 (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 concave, the argument is increasing and concave in P_τ, then $s \mapsto \Phi(\epsilon\sqrt{s})$ is concave, we also obtain

$$1-\delta \le 2\Phi\left(\epsilon\sqrt{\mathbb{E}[P_\tau]}\right) - 1.$$

Taking the inverse, and defining $z_{1-\delta/2} = \Phi^{-1}(1-\delta/2)$, we find

$$z_{1-\delta/2} \le \epsilon\sqrt{\mathbb{E}[P_\tau]}.$$

Hence, we are just let with bounding P_τ :

$$P_\tau \le 1 + \frac{1}{\sigma^2}C^2\tau,$$

from which we get

$$z_{1-\delta/2} \le \epsilon\sqrt{1 + \frac{C^2}{\sigma^2}\mathbb{E}[\tau]}.$$

Therefore

$$\mathbb{E}[\tau] \ge \frac{\sigma^2}{C^2}\left[\frac{z_{1-\delta/2}^2}{\epsilon^2} - 1\right]_+$$

□

B.9.2. 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

$$T_{\arg,\nu}(r,\delta) = \inf_{\mathcal{A}} \mathbb{E}_{\Lambda,F,\xi}[\tau],$$

where the infimum is over all sequential algorithms $\mathcal{A}$ that output $\hat{X}$ and satisfy

$$\mathbb{P}_{\Lambda,F,\xi}\left(\inf_{X^\star\in D^\star}\|\hat{X}-X^\star\|\le r\right)\ge 1-\delta.$$

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 teh 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

$$\mathcal{I}_\eta=\{\mathrm{dist}(D^\star,\partial D)\ge\eta\},\qquad \beta_\eta=\mathbb{P}_{\Lambda,F}(\mathcal{I}_\eta^c).$$

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 11. *An RBF-GP on D is a.s. C^∞, and its restriction to the interior $\mathrm{int}(D)$ is a.s. a Morse function (every critical point has invertible Hessian, all critical values are distinct, and there are finitely many critical points). If the maximizer is not attained at the boundary, then it is unique almost surely.*

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 17 (Regularity under the RBF prior)**.** *Define $B(x,\rho)$ to be the ball centered around x of radius ρ with some norm $\|\cdot\|$. 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 $\mathbb{P}_{\Lambda,F}(\mathcal{E}_\alpha)\ge 1-\alpha$, and on $\mathcal{E}_\alpha$ the following properties hold:*

$$B(X^\star,\rho_\alpha)\subset D,$$
$$\mu_\alpha I\preceq-\nabla^2F(x)\preceq L_\alpha I\qquad\forall x\in B(X^\star,\rho_\alpha),$$
$$\sup_{x\in B(X^\star,\rho_\alpha)}\|\nabla^3F(x)\|_{\mathrm{op}}\le M_\alpha,$$

and

$$F^\star-\sup_{x\notin B(X^\star,\rho_\alpha/2)}F(x)\ge\Gamma_\alpha.$$

Proof. Let $\gamma=\alpha-\beta_\eta$.

Under $\mathcal{I}_\eta$ the maximizer $X^\star$ is in the interior, and unique almost surely. Define $\lambda^\star=\lambda_{\min}(-\nabla^2F(X^\star))$: under $\mathcal{I}_\eta$ we have that $\lambda^\star>0$ almost surely, hence, there exists $\mu_\alpha>0$ s.t. $\mathbb{P}(\mathcal{I}_\eta\bigcap\{\lambda^\star/2<\mu_\alpha\})\le\gamma/5$.

Next, we use the fact that $F\in C^\infty$ and D is compact to obtain an upper bound on the Hessian:

$$\sup_{x\in D}\|\nabla^2F(x)\|_{\mathrm{op}}\le L_0<\infty.$$

Therefore, there exists $L_\alpha<\infty$ such that $\mathbb{P}(L_\alpha<L_0)\le\gamma/5$. With a similar reasoning, we also obtain $\sup_{x\in D}\|\nabla^3F(x)\|_{\mathrm{op}}\le M_0$, and thus there exists $M_\alpha<\infty$ such that $\mathbb{P}(M_\alpha<M_0)\le\gamma/5$.

Next, by continuity and the margin condition of $\mathcal{I}_\eta$, the following quantity exists and is strictly positive almost surely

$$\rho_0(\omega)=\sup\left\{q<\eta:\sup_{X\in B(X^\star,q)}\|\nabla^2F(X)-\nabla^2F(X^\star)\|_{\mathrm{op}}\le\frac{\lambda^\star}{2}\right\},$$

where ω denotes a realization of F. Therefore, there exists $\rho_\alpha>0$ such that $\mathbb{P}(\mathcal{I}_\eta\bigcap\{\rho_0<\rho_\alpha\})\le\gamma/5$.

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 from Lemma 17

▷ **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 $K \leftarrow \lceil \frac{4L}{3\mu}\log\frac{\rho}{2r}\rceil, \quad q \leftarrow 1-\frac{3\mu}{4L}, \quad e_0 \leftarrow \rho/2$
for $k = 0, \dots, K-1$ **do**
    $e_k \leftarrow e_0\, q^k$
    $s_k \leftarrow \min\left\{\rho/8,\ \sqrt{3\mu e_k/(4\sqrt{d}\,M)}\right\}$
    $n_k \leftarrow \lceil 64d\cdot\sigma^2\mu^{-2}e_k^{-2}s_k^{-2}\log(2dK/\delta_1)\rceil$
    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 $\hat{X} = x_K$

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 $\mathbb{P}(\mathcal{I}_\eta \bigcap \{\Gamma(\rho_\alpha) < \Gamma_\alpha\}) \le \gamma/5$.

Define

$$\mathcal{E} = \mathcal{I}_\eta \cap \{\lambda^\star/2 \ge \mu_\alpha\} \cap \{L_\alpha \ge L_0\} \cap \{M_\alpha \ge M_0\} \cap \{\rho_0 \ge \rho_\alpha\} \cap \{\Gamma(\rho_\alpha) \ge \Gamma_\alpha\}.$$

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 $\mathbb{P}(\mathcal{E}^c) \le \beta_\eta + 5\frac{\gamma}{5} = \alpha$. 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

$$-\nabla^2 F(x) \succeq -\nabla^2 F(X^\star) - \frac{\lambda^\star}{2}I \succeq \frac{\lambda^\star}{2}I \succeq \mu_\alpha I.$$

□

Argmax upper bound. Under the event $\mathcal{I}_\eta$, we want to construct an algorithm that localizes $X^\star$ to within radius r with probability $1-\delta$. The idea is to show an upper bound on the minimal lower bound of the type

$$T_{\arg,\nu}(r,\delta) := \inf_{\mathcal{A}} \mathbb{E}[\tau] \le B_\alpha + O(\sigma^2 r^{-\gamma}\log(1/r)),$$

for some $\gamma > 0$.

Under $\mathcal{I}_\eta$, we have the guarantees from Lemma 17, and thus

$$F(X) = F(X^\star) + \frac{1}{2}(X - X^\star)^\top \nabla^2 F(X^\star)(X - X^\star) + o(\|X - X^\star\|^2).$$

from which we find $F(X) - F(X^\star) \approx \frac{\mu_\alpha}{2}\|X - X^\star\|_2^2$. Therefore, localizing to radius r gives an error of roughly $\frac{\mu_\alpha r^2}{2}$ on the value. Therefore, for ϵ accuracy on the value, we can take the radius to be $r \sim \sqrt{\epsilon}$

Therefore, under $\mathcal{I}_\eta$ if $\beta_\eta < 1$, and

$$\liminf_{\epsilon \to 0} \frac{T_{\mathrm{val},\nu}(\epsilon, \delta)}{T_{\mathrm{arg},\nu}(\sqrt{\epsilon}, \delta)} \to \infty$$

one can argue that the problem of estimating the max-value is intrinsically harder than the problem of estimating the argmax as the accuracy radius decreases. In particular, considering the lower bound on the max-value estimation problem Theorem B.6, and the proposed upper bound on $T_{\mathrm{arg},\nu}(r, \delta)$, we obtain that

$$\frac{T_{\mathrm{val},\nu}(\epsilon, \delta)}{T_{\mathrm{arg},\nu}(\sqrt{\epsilon}, \delta)} \geq \frac{\Omega(\epsilon^{-2})}{O(\epsilon^{-\gamma/2} \log\log(1/\sqrt{\epsilon}))},$$

which diverges for $\gamma \in (0, 4)$ as $\epsilon \to 0$.

So value estimation is 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.

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 Lemma 17. 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 analysed gradient descent when $x_0 \in B(X^\star, \rho)$.

Theorem B.7 (Sample Complexity of T-BAL)**.** *Consider the event $\mathcal{E}_\alpha$ in Lemma 17. Set $r > 0, \delta \in (0, 1), \alpha \in (\beta_\eta, \delta)$ and $\delta_0 = \delta_1 = (\delta - \alpha)/2$. Then, T-BAL (Algorithm 2) satisfies $\mathbb{P}_{F,\Lambda}(\|\hat{X} - X^\star\| \leq r) \geq 1 - \delta$, using at-most*

$$B_\alpha + O\left(\frac{\sigma^2}{\mu^2}\max(2, L/\mu)\left[\frac{d^2}{\rho^2 r^2} + \frac{d^{5/2}M}{\mu r^3}\right]\log\frac{Ld\log(\rho/r)}{\mu\delta_1}.\right)$$

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 Proposition 11. 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.

<u>*One step gradient ascent.*</u> 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

$$\nabla F(x) = \phi(1) - \phi(0) = \int_0^1 \phi'(t)\mathrm{d}t = (x - X^\star)\underbrace{\int_0^1 \nabla^2 F(x^\star + t(x - X^\star))\mathrm{d}t}_{=:-A(x)}.$$

Then

$$x^+ - X^\star = x - \frac{1}{L}(x - X^\star)A(x) - X^\star = (I - \frac{A(x)}{L})(x - X^\star).$$

The matrix $(I - \frac{A(x)}{L})$ has eigenvalues in $[0, 1 - \mu/L]$ for $x \in B(X^\star, \rho)$. Therefore,

$$\|x^+ - X^\star\| \leq \left(1 - \frac{\mu}{L}\right)\|x - X^\star\|$$

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

$$\|\hat{g}_k - \nabla F(x_k)\| \leq \frac{\mu e_k}{4},$$

where $e_k = e_0 q^k$. Then

$$\begin{aligned}
\|x_{k+1} - X^\star\| &= \|x_k + \frac{1}{L}\hat{g}_k - X^\star\|, \\
&= \|x_k + \frac{1}{L}\nabla F(x_k) + \frac{1}{L}\hat{g}_k - \frac{1}{L}\nabla F(x_k) - X^\star\|, \\
&\leq \|x_k + \frac{1}{L}\nabla F(x_k) - X^\star\| + \frac{1}{L}\|\hat{g}_k - \nabla F(x_k)\|, \\
&\leq \left(1 - \frac{\mu}{L}\right)\|x_k - X^\star\| + \frac{\mu e_k}{4L}, \\
&\leq \left(1 - \frac{3\mu}{4L}\right) e_k.
\end{aligned}$$

So noisy gradients still guarantee convergence as long as we can show $\|\hat{g}_k - \nabla F(x_k)\| \leq \frac{\mu e_k}{4}$.

Finite-difference gradient bound: bias term. We now bound $\|\hat{g}_k - \nabla F(x_k)\|$ through a bias-variance decomposition:

$$\|\hat{g}_k - \nabla F(x_k)\| \leq \|\hat{g}_k - \mathbb{E}[\hat{g}_k]\| + \|\mathbb{E}[\hat{g}_k] - \nabla F(x_k)\|.$$

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 Lemma 17.

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 $|\phi'''| \leq M$, and therefore

$$\left|\frac{\phi(s_k) - \phi(-s_k)}{2s_k} - \phi'(0)\right| \leq \frac{s_k^2 M}{6},$$

leading to

$$\|\mathbb{E}[\hat{g}_k] - \nabla F(x_k)\| \leq \frac{\sqrt{d}s_k^2 M}{6}.$$

Setting $\frac{\sqrt{d}s_k^2 M}{6} \leq \mu e_k/8$ yields

$$s_k \leq \sqrt{\frac{3\mu e_k}{4\sqrt{d}M}}$$

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. Since we need $e_K \le r$, and $e_K = e_0 q^K$, we have

$$K = \left\lceil \frac{\log(e_0/r)}{\log(1/q)} \right\rceil.$$

Since $\log(1/q) - \log q = -\log(1 - 3\mu/4L) \ge 3\mu/4L$, we have

$$K \le 4L\frac{\log(e_0/r)}{3\mu} \le 4L\frac{\log(\rho/(2r))}{3\mu},$$

where we used $e_0 \le \rho/2$. Then, summing the number of samples from $k = 0, \dots, K-1$, we obtain

$$\sum_{k=0}^{K-1} 2dn_k = \sum_{k=0}^{K-1} \frac{128d^2 \cdot \sigma^2}{\mu^2 e_k^2 s_k^2} \log \frac{2dK}{\delta_1}.$$

Use that $\frac{1}{s_k^2} \le \frac{64}{\rho^2} + \frac{4\sqrt{d}M}{3\mu e_k}$ and that $e_k = e_0 q^k$ with $q \in (0,1)$:

$$\begin{aligned}\sum_{k=0}^{K-1} 2dn_k &\le \sum_{k=0}^{K-1} \left[\frac{128 \cdot 64 \cdot d^2 \cdot \sigma^2}{\mu^2 e_k^2 \rho^2} + \frac{512 d^{5/2} M \cdot \sigma^2}{3\mu^3 e_k^3}\right] \log \frac{2dK}{\delta_1}, \\ &\lesssim \log \frac{2dK}{\delta_1} \sum_{k=0}^{K-1} \left[\frac{1}{e_k^2} + \frac{1}{e_k^3}\right].\end{aligned}$$

Regarding the series,

$$\sum_{k=0}^{K-1} \frac{1}{e_k^p} = \frac{1}{e_0^p} \sum_{k=0}^{K-1} \frac{1}{q^{pk}} \leq e_0^{-p} \frac{1-q^{-pK}}{1-q^{-p}} \leq \frac{e_0^{-p} q^{-pK}}{q^{-p}-1} = \frac{e_K^{-p}}{q^{-p}-1}.$$

Now, note that since K is the smallest integer achieving $e_K = e_0 q^K \leq r$, we have $e_0 q^{K-1} \geq r$, and thus $e_K = q e_{K-1} \geq qr \Rightarrow (qr)^{-p} \geq e_K^{-p}$. We obtain

$$\sum_{k=0}^{K-1} \frac{1}{e_k^p} \leq \frac{(qr)^{-p}}{q^{-p}-1} = \frac{r^{-p}}{1-q^p}.$$

To lower bound the denominator, recall that q is of the form $(1-x)$. Using that $(1-x)^p \leq e^{-xp}$ for $x \in (0,1)$, we have $1-q^p \geq 1-e^{-p3\mu/(4L)}$.

Since $1-e^{-t} \geq t/2$ for $t \in (0,1)$, for p sufficiently small we obtain $1-q^p \geq \frac{3p\mu}{8L}$. If p is large, such that $t \geq 1$, then $1-e^{-t} \geq 1-e^{-1} \geq 1/2$. Therefore, $\frac{1}{1-q^p} \leq \max(2, \frac{8L}{3p\mu})$, and

$$\sum_{k=0}^{K-1} \frac{1}{e_k^p} \leq r^{-p} \max\left(2, \frac{8L}{3p\mu}\right).$$

Since $p \in \{2,3\}$ we also have $\sum_{k=0}^{K-1} \frac{1}{e_k^p} \leq r^{-p} \max\left(2, \frac{8L}{6\mu}\right)$.

Then, we conclude that the second phase sample complexity is upper bounded by

$$\begin{aligned}\sum_{k=0}^{K-1} 2dn_k &\leq \frac{128\sigma^2}{\mu^2} \cdot \max\left(2, \frac{8L}{6\mu}\right) \cdot \left[\frac{64d^2}{\rho^2 r^2} + \frac{4d^{5/2}M}{3\mu r^3}\right] \log \frac{2dK}{\delta_1}, \\ &\leq \frac{128\sigma^2}{\mu^2} \cdot \max\left(2, \frac{8L}{6\mu}\right) \cdot \left[\frac{64d^2}{\rho^2 r^2} + \frac{4d^{5/2}M}{3\mu r^3}\right] \log \frac{8Ld\log(\rho/(2r))}{3\mu\delta_1}.\end{aligned}$$

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 Lemma 17). 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 11. *Consider Algorithm 2, and let $\delta_0 \in (0,1)$. Under $\mathcal{E}_\alpha$ (see Lemma 17), the first phase samples $B_\alpha \leq 2^d h^{-d} \lceil 128\sigma^2\Gamma^{-2}\log(2^{d+1}h^{-d}/\delta_0)\rceil$ 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 Lemma 17, 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 $g \in \mathcal{D}$ 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 Lemma 17 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 $x_0 \in B(X^\star, \rho/2)$ and

$$\hat{F}(x_0) \geq \hat{F}(g^\star) \geq F^\star - \frac{L}{2}\left(\min\{\rho/8, \sqrt{3\Gamma/(2L)}\}\right)^2 - \frac{\Gamma}{8},$$

Lastly, the number of points sampled depends on the number of points in $\mathcal{G}$ is bounded by $(1/h+1)^d$. Since $h \leq 1$, then $|\mathcal{G}| \leq 2^d h^{-d}$. □

C. 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 Section B.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.

C.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.

The readout is not simply the last hidden state. 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 $Q_\psi(H_t, a)$.

C.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.

There is one convention that matters. 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 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.

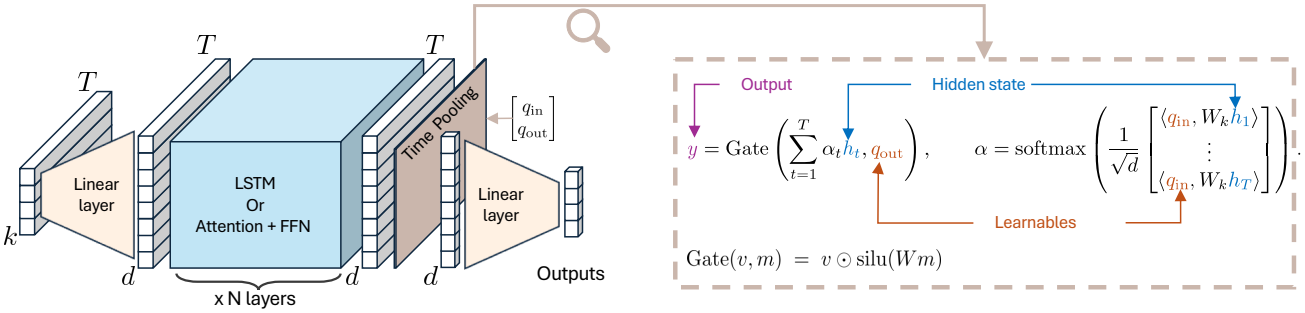


Figure 6. Sequential encoder and query-pooling readout used by the inference network I_ϕ, the critic Q_ψ, and the learned TD3 actors.

C.3. Inference Update and Its Regularization

For a prefix h and target $x_\theta^\star$, the inference model outputs a diagonal Gaussian

$$I_\phi(\cdot|h) = \mathcal{N}(\mu_\phi(h), \mathrm{diag}(\sigma_\phi^2(h))).$$

The ideal objective in the main text is the negative log-likelihood in Eq. (5). 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

$$\mathcal{L}_{\mathrm{inf}}(\phi) = \mathbb{E}_{(h,x_\theta^\star)\sim\mathcal{B}}\left[\rho_\tau\big(\ell_\phi(h, x_\theta^\star)\big) + \alpha_{\mathrm{anc}}\,\mathrm{SmoothL1}\left(\mu_\phi(h), x_\theta^\star\right)\right]. \tag{23}$$

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 $Z \sim I_\phi(\cdot|h)$,

$$\mathcal{L}_{\mathrm{sph}}(\phi) = -\mathbb{E}\left[\left\langle \frac{Z}{\|Z\|_2}, \frac{x_\theta^\star}{\|x_\theta^\star\|_2} \right\rangle\right]. \tag{24}$$

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.

C.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. Since the ideal reward $r_t(h) = \max_x q_t(h, x)$ is not available, we use the target inference model to estimate how likely the implemented stochastic selector is to be already ϵ-correct:

$$\hat r_{\bar\phi,m}(h, \theta) = \frac{1}{m}\sum_{k=1}^{m} \mathbf{1}\left\{L_\theta\left(X^{(k)}\right) \le \epsilon\right\}, \qquad X^{(k)} \sim I_{\bar\phi}(\cdot|h). \tag{25}$$

The reason for sampling is that the stopping decision should depend on posterior concentration, not only on the posterior mean. If $\mu_\phi(h)$ is close to $x_\theta^\star$ but $\sigma_\phi(h)$ is still large, stopping is risky. The sampled reward makes this visible to the critic, and this is exactly the gap controlled by Proposition 6.

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 it is the target inference mean plus small smoothing noise, projected to

the feasible action set. 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

$$y_{\text{stop}}(h, \theta) = \hat{r}_{\bar{\phi},m}(h, \theta), \tag{26}$$

$$y_{\text{cont}}(h, a, h', d, \theta) = -c(1-d) + d\,\hat{r}_{\bar{\phi},m}(h', \theta) + \gamma(1-d)V_{\bar{\psi}}(h'). \tag{27}$$

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

$$\mathcal{L}_Q(\psi) = \frac{1}{2}\mathbb{E}_{\mathcal{B}}\left[M\sum_{j=1}^{2}\left(Q_{\psi,j}(h, a) - y_{\text{cont}}\right)^2\right] + \frac{w_{\text{stop}}}{2}\mathbb{E}_{\mathcal{B}}\left[\sum_{j=1}^{2}\left(Q_{\psi,j}(h, a_{\text{stop}}) - y_{\text{stop}}\right)^2\right], \tag{28}$$

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 is exactly the learned version of the Bellman comparison in Theorem 3.1.

C.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

$$A_t = \mu_{\phi}(H_t) + \sigma_{\phi}(H_t) \odot \xi_t, \qquad \xi_t \sim \mathcal{N}(0, I). \tag{29}$$

At evaluation in greedy mode we set $\xi_t = 0$, so the action is the projected posterior mean. This choice is deliberately simple. 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 inference mean, with only the small target-smoothing noise described above.

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 posterior mean, provided the current sample is too close to the mean. In Euclidean tasks the distance is $\|z - \mu_{\phi}(H_t)\|_2$; in ϵ-best-arm it is $1 - \langle z, \mu_{\phi}(H_t)\rangle$. The intuition is that the posterior mean is the current 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 inference mean plus small smoothing noise.

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

$$a^{+}(h') = \tanh(\pi_{\bar{\vartheta}}(h')) + \zeta, \qquad \zeta \sim \mathcal{N}(0, \sigma_{\text{tgt}}^2 I),$$

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. (23); for ϵ-best-arm also use Eq. (24).
  Estimate sampled rewards with $I_{\bar{\phi}}$ using Eq. (25) and update Q_ψ using Eqs. (26) to (28).
  If using TD3, update π_ϑ on the delayed actor schedule using Eqs. (30) and (31).
  Polyak-update $I_{\bar{\phi}}$, $Q_{\bar{\psi}}$, and, when present, $\pi_{\bar{\vartheta}}$; update c with Eq. (32).
end while

// Deployment phase
Freeze the learned networks and 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 $\hat{x} = \mu_\phi(H_t)$.
  end if
  Execute A_t, observe Y_{t+1}, and update H_{t+1}.
end for
return $\hat{x} = \mu_\phi(H_{T_{\max}+1})$.

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

$$\mathcal{L}^{\mathrm{det}}_{\mathrm{act}}(\vartheta) = -\mathbb{E}_{h\sim\mathcal{B}}\left[Q_{\psi,1}\left(h, \tanh(\pi_\vartheta(h))\right)\right]. \tag{30}$$

The actor uses the first critic for the policy gradient, while the target uses the first critic.

When $\mathcal{A} = \mathcal{X}$ and the actor is Gaussian, we regularize the TD3 actor toward the inference distribution,

$$\mathcal{L}^{\mathrm{KL}}_{\mathrm{act}}(\vartheta) = -\mathbb{E}_{h\sim\mathcal{B}}\left[Q_{\psi,1}(h, A_\vartheta(h))\right] + \beta_{\mathrm{KL}}\mathbb{E}_{h\sim\mathcal{B}}\left[\mathrm{KL}\left(I_\phi(\cdot|h)\|\pi_\vartheta(\cdot|h)\right)\right], \qquad A_\vartheta(h) \sim \pi_\vartheta(\cdot|h). \tag{31}$$

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.

C.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

$$D_i = \begin{cases} \|\mu_\phi(H^{(i)}_{\tau_i}) - x^\star_{\theta_i}\|_2, & \text{Euclidean tasks,} \\ 1 - \left\langle \mu_\phi(H^{(i)}_{\tau_i}), x^\star_{\theta_i} \right\rangle, & \epsilon\text{-best-arm tasks,} \end{cases}$$

and use the smooth accuracy proxy

$$\hat{p} = \frac{1}{B}\sum_{i=1}^{B} \sigma\left(\frac{\epsilon - D_i}{\kappa\epsilon}\right).$$

The update is

$$c \leftarrow \mathrm{Proj}_{[0,1]}\left(c - \eta_c\big((1-\delta) - \hat{p}\big)\right). \tag{32}$$

Thus, if the observed correctness is below $1 - \delta$, 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. (10).

D. Appendix: Numerical Results

In this section we present more details on the numerical results. We refer the reader to the code for more details (see the `README.md` file), especially regarding the hyperparameters. We now present the synthetic benchmarks with additional numerical results. These additional results include sweeps over various values of ϵ, σ, as well as checking robustness to prior misspecification. We conclude with details and additional results regarding the geochemical exploration task.

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 12 (On testing the inference model)**.** *Several of our synthetic localization benchmarks fall close to the symmetric case described in Section B.7 (after Proposition 8): 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.*

D.1. Synthetic Benchmarks: description

The synthetic benchmarks in Section 5 are designed to isolate different aspects of the continuous fixed-confidence problem. Binary search is the cleanest localization problem: every query returns a noisy comparison with the unknown 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. Finally, GP max-value estimation is included because it is the case where the query space and the recommendation space are genuinely 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

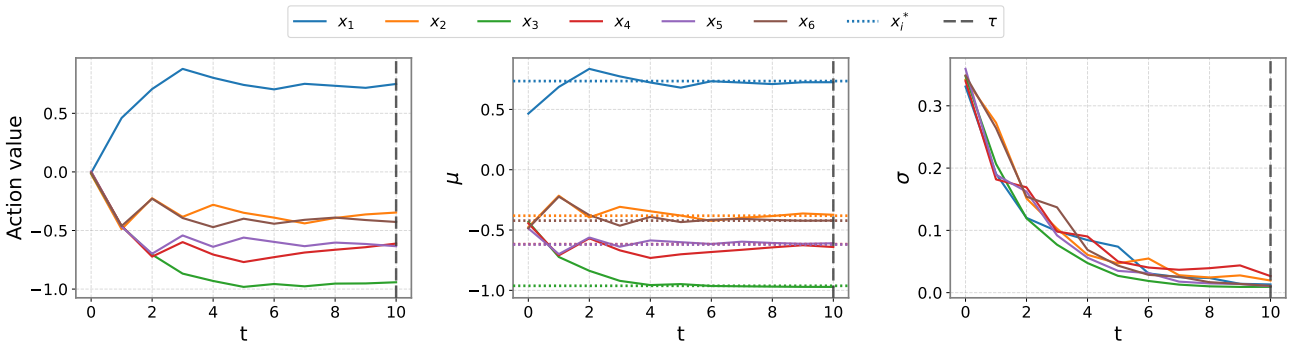


Figure 7. Binary search: visualization of how **C-ICPE** explores in 6 dimensions. From left to right: the query action (left), posterior mean (middle), and posterior standard deviation (right) along an exploration trajectory in the noisy binary search problem.

vectors to the unit sphere. For GP max-value estimation the loss is the scalar absolute error $L_\theta(x) = |x - v_\theta^\star|$. Confidence intervals are computed with hierarchical bootstrap over test episodes and random seeds.

D.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,

$$Y_{t,i} = \xi_{t,i} \operatorname{sign}(\theta_i - A_{t,i}), \qquad \mathbb{P}(\xi_{t,i} = 1) = 1 - p, \qquad \mathbb{P}(\xi_{t,i} = -1) = p, \tag{33}$$

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.

D.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

$$Y_t = \theta^\top A_t + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \sigma^2). \tag{34}$$

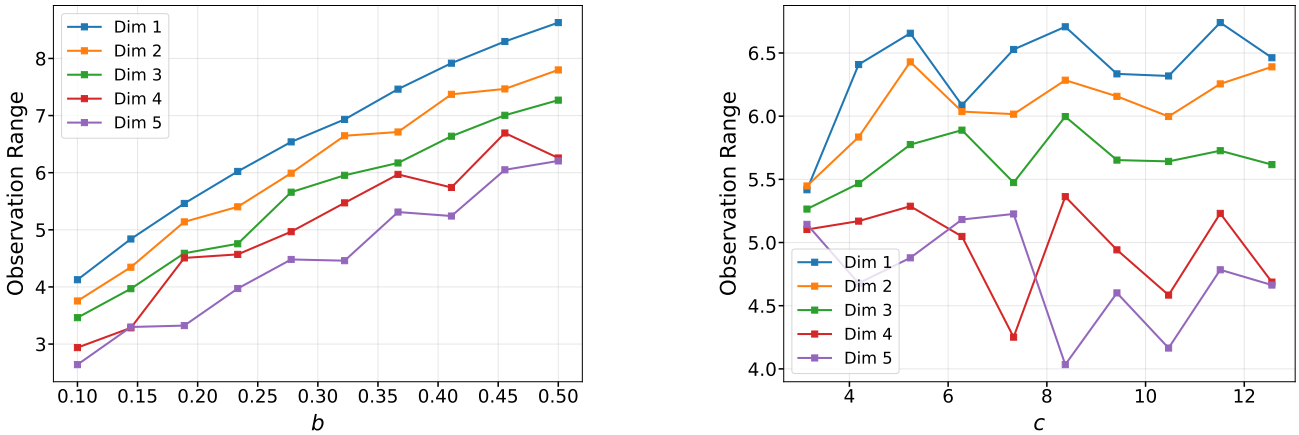
The recommendation is correct if $\theta^\top \hat{x} \geq 1 - \epsilon$. Although the implementation stores the enclosing action bounds as $[-1, 1]^d$, the inference mean, posterior samples, TTPS candidates, and uniform baseline actions are projected to $\mathbb{S}^{d-1}$ for this benchmark. This projection is important: Euclidean uncertainty is not the right object near the sphere, and two vectors with the same direction but different norms should not be treated as different hypotheses. 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. However, note that in this problem the optimal exploration strategy is uniform [Jedra and Proutiere, 2020]. Therefore, it shows to what degree **C-ICPE** is able to learn a good inference model.

D.1.3. Ackley minimizer identification.

The Ackley task is a shifted and randomly parametrized global-optimization problem. The task parameter is

$$\theta = (a, b, c, \theta^\star), \qquad \theta^\star \sim \mathrm{Unif}([-1, 1]^d),$$



(a) *(b)*

Figure 8. Effect of Ackley function's parameters on output range across multiple dimensions: (a) range vs b; (b) range vs c.

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 $u = A_t - \theta^\star$, the unnormalized Ackley value is

$$F_{a,b,c}(u) = a + e - a \exp\left(-b\sqrt{\frac{1}{d}\sum_{j=1}^{d} u_j^2}\right) - \exp\left(\frac{1}{d}\sum_{j=1}^{d}\cos(cu_j)\right). \tag{35}$$

We observe the sign-inverted and normalized value (more on this in the next page)

$$Y_t = 1 - 2\frac{F_{a,b,c}(A_t - \theta^\star)}{Z_{\mathrm{norm}}(b, c, d)} + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \sigma^2), \tag{36}$$

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: $Z_{\mathrm{norm}}(b, c, d) = \pi - 0.21 \cdot D + 9.68 \cdot b + 0.04 \cdot c$.

D.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

$$f \sim \mathrm{GP}(0, k_{\mathrm{RBF}}(\ell, \sigma_f)), \qquad \ell \sim \mathrm{Unif}[0.05, 0.2], \qquad \sigma_f = 1,$$

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:

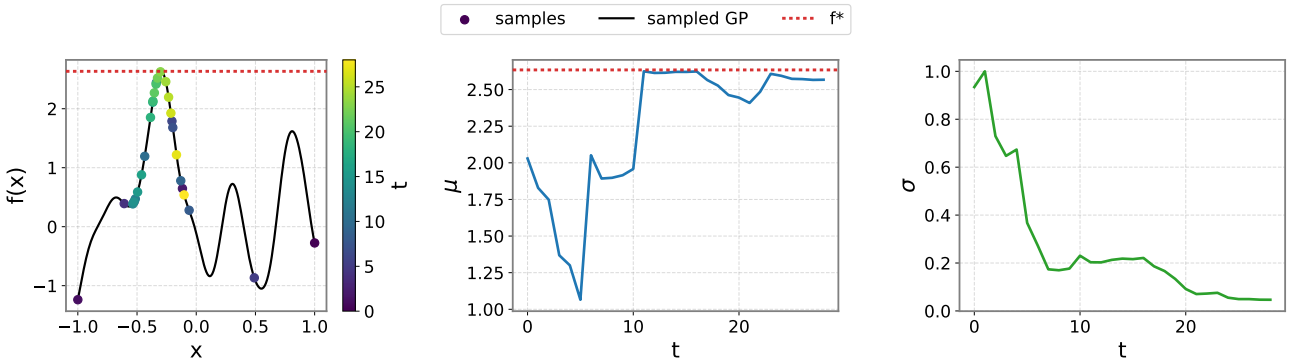


Figure 9. Visualization of how **C-ICPE** explores in the max-value estimation problem. From left to right: the query action (left), posterior mean (middle), and posterior standard deviation (right) along an exploration trajectory in the GP value estimation problem. Darker samples are queried earlier in the trajectory, while lighter samples are queried later.

$$Y_t = f(A_t) + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \sigma^2). \tag{37}$$

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 $|\hat{x} - v_\theta^\star| \leq \epsilon$.

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, ϵ-best-arm, 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 $\mathcal{A} \neq \mathcal{X}$.

D.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 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, 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-based Bayesian optimization, 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 for (ϵ, δ)-correctness. To make the comparison conservative, we give them budgets tied to the empirical stopping time of the corresponding **C-ICPE** variant.

They are then evaluated under the same success criterion $L_\theta(\hat{x}) \leq \epsilon$. This comparison asks whether a generic optimizer,

Environment	d	ε	$t_{\max}$
ε-Best-Arm	15	0.005	150
GP-value estimation	1	0.2	60
Geochemical	2	0.15	150

Table 1. Sample budget $t_{\max}$ for setting where $t_{max} \neq 100$

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.

D.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 $x_\theta^\star$ 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.

D.3.1. Noisy Binary Search

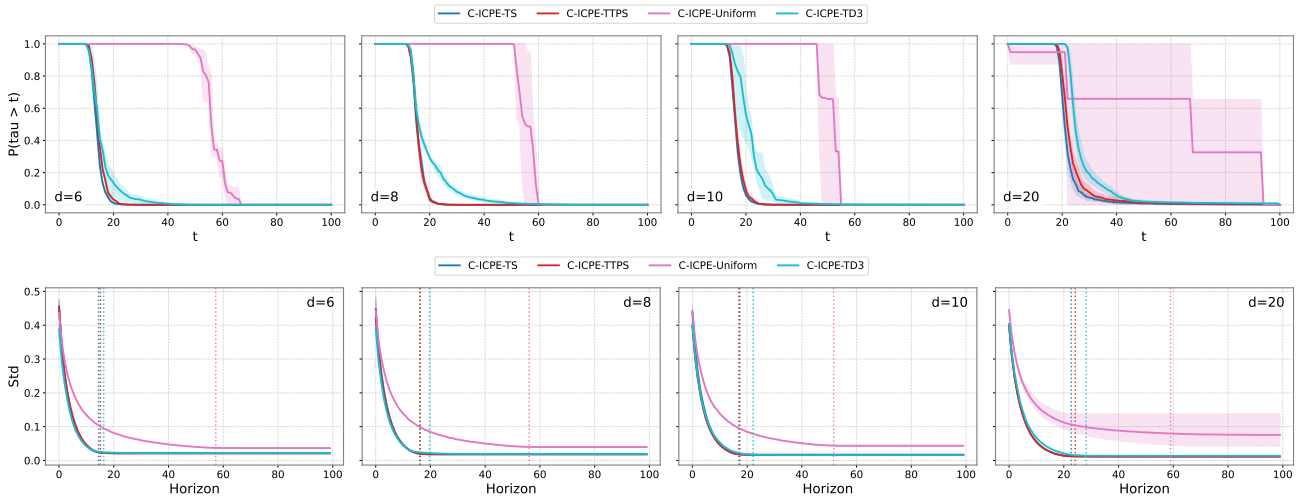


Figure 10. Results for Binary Search Problem with fixed confidence $\delta = 0.1$ and $N = 100$ across different dimensions at the most hardest (ε, σ) setting: (top) survival function of τ; (bottom) inference uncertainty convergence.

Tables 2 and 3 report accuracy and sample complexity across all (d, ε, σ) configurations; Figure 10 shows the survival function of τ and the convergence of the inference model's standard deviation. In Fig. 7 we also show ho **C-ICPE** explores in 5 dimensions, depicting the queries chosen by the actor, the posterior mean, and the posterior standard deviation over timesteps.

Accuracy. All three active actors (TS, TTPS, TD3) meet the $1-\delta = 0.90$ target across every configuration tested, with mean accuracy between 0.895 and 0.916. The confidence intervals confirm that the target is met reliably: even the lower bounds remain at or above 0.886. **C-ICPE-Uniform** matches the active methods at $d = 6$ with $\varepsilon = 0.2$ (accuracy ≥ 0.901) but degrades sharply as either d increases or ε decreases. At $(\varepsilon, \sigma) = (0.1, 0.05)$, accuracy drops from 0.688 at $d = 6$ to 0.440 at $d = 8$, 0.171 at $d = 10$, and 0.006 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** and **C-ICPE-TTPS** achieve comparable sample complexity across all settings: at $d = 20$, $(\varepsilon, \sigma) = (0.1, 0.05)$, **C-ICPE-TS** stops in 22.7 queries on average and **C-ICPE-TTPS** in 24.2, both with tight confidence intervals. **C-ICPE-TD3** is competitive at moderate dimensions ($d \leq 10$) but exhibits higher mean stopping times and substantially wider confidence intervals at $d = 20$ (e.g., 35.8 $[19.2, 67.6]$ at $\varepsilon = 0.2$, $\sigma = 0.05$), suggesting that the learned actor is less stable in high dimensions. Sample complexity scales sublinearly in d for the active actors: **C-ICPE-TS** increases from 11.1 ($d = 6$) to 16.4 ($d = 20$) at $(\varepsilon, \sigma) = (0.2, 0.05)$.

Stopping behavior. The survival functions (Figure 10, top) corroborate the sample complexity results. At $d \leq 10$, **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. **C-ICPE-Uniform** has a heavy-tailed survival function that extends to the horizon, and at $d = 20$ it rarely stops before $t_{\max}$. The inference standard deviation (bottom row) confirms that the active actors' posteriors contract rapidly, reaching near-zero uncertainty before the median stopping time (dashed vertical

lines), while **C-ICPE-Uniform** at $d = 20$ retains high residual uncertainty throughout the episode.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
6	C-ICPE-TD3	0.895 [.886,.902]	0.898 [.886,.908]	0.903 [.893,.914]	**0.901** [.889,.912]
	C-ICPE-TS	**0.904** [.895,.912]	**0.907** [.899,.913]	**0.909** [.899,.920]	0.900 [.893,.909]
	C-ICPE-TTPS	0.897 [.890,.905]	0.903 [.893,.913]	0.900 [.893,.907]	0.898 [.889,.905]
	C-ICPE-uniform	0.901 [.894,.908]	0.904 [.893,.914]	0.854 [.835,.874]	0.688 [.666,.709]
8	C-ICPE-TD3	0.903 [.897,.911]	0.898 [.890,.907]	0.901 [.891,.912]	0.899 [.890,.906]
	C-ICPE-TS	0.905 [.892,.917]	0.898 [.889,.905]	**0.916** [.902,.930]	**0.909** [.899,.918]
	C-ICPE-TTPS	0.903 [.896,.911]	**0.905** [.893,.915]	0.907 [.895,.917]	0.901 [.891,.909]
	C-ICPE-uniform	**0.907** [.897,.918]	0.894 [.886,.901]	0.724 [.693,.756]	0.440 [.406,.477]
10	C-ICPE-TD3	0.901 [.892,.908]	0.898 [.889,.904]	0.904 [.893,.914]	**0.915** [.891,.948]
	C-ICPE-TS	0.896 [.886,.903]	**0.903** [.890,.916]	**0.916** [.905,.925]	0.904 [.892,.915]
	C-ICPE-TTPS	**0.902** [.895,.910]	0.900 [.891,.906]	0.901 [.891,.911]	0.904 [.897,.912]
	C-ICPE-uniform	0.873 [.853,.891]	0.860 [.847,.873]	0.493 [.471,.514]	0.171 [.136,.201]
20	C-ICPE-TD3	0.897 [.888,.905]	0.905 [.896,.913]	**0.925** [.908,.940]	0.903 [.895,.912]
	C-ICPE-TS	**0.911** [.893,.931]	0.903 [.892,.914]	0.901 [.891,.910]	0.903 [.889,.919]
	C-ICPE-TTPS	0.896 [.887,.903]	**0.907** [.897,.917]	0.900 [.890,.909]	**0.913** [.906,.921]
	C-ICPE-uniform	0.662 [.630,.698]	0.079 [.001,.136]	0.036 [.022,.049]	0.006 [.000,.013]

Table 2. Binary search: accuracy (mean and 95% CI) for every (d, ε, σ) configuration.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
6	C-ICPE-TD3	**10.4** [10.2,10.6]	15.5 [15.2,15.9]	13.3 [12.5,14.2]	16.3 [15.7,16.8]
	C-ICPE-TS	11.1 [11.0,11.3]	**15.3** [15.2,15.4]	**12.0** [11.9,12.0]	**14.4** [14.3,14.6]
	C-ICPE-TTPS	11.3 [11.0,11.6]	16.2 [16.1,16.3]	12.4 [12.1,12.6]	15.1 [14.9,15.4]
	C-ICPE-uniform	36.3 [36.0,36.7]	54.1 [52.4,56.9]	60.0 [58.2,62.3]	57.3 [56.1,58.1]
8	C-ICPE-TD3	**11.8** [11.7,11.9]	**17.3** [17.0,17.5]	15.8 [15.0,16.8]	19.7 [19.1,20.2]
	C-ICPE-TS	12.0 [11.5,12.4]	17.4 [16.9,17.7]	**13.0** [12.6,13.4]	16.2 [15.8,16.5]
	C-ICPE-TTPS	12.3 [12.2,12.4]	17.8 [17.6,18.0]	13.4 [13.2,13.6]	**16.2** [15.9,16.5]
	C-ICPE-uniform	44.0 [43.4,44.7]	62.1 [61.6,62.7]	62.1 [60.0,64.4]	56.0 [53.8,58.8]
10	C-ICPE-TD3	**12.7** [12.5,12.8]	19.4 [18.5,20.7]	18.2 [16.2,20.4]	22.2 [20.5,23.6]
	C-ICPE-TS	12.8 [12.7,12.8]	**18.5** [18.1,18.8]	**14.0** [13.8,14.1]	**17.0** [16.7,17.3]
	C-ICPE-TTPS	13.4 [12.9,13.8]	18.8 [18.7,18.9]	14.1 [14.0,14.2]	17.3 [17.0,17.7]
	C-ICPE-uniform	47.6 [45.2,49.3]	68.1 [67.2,68.9]	59.2 [57.9,60.4]	51.7 [47.1,55.0]
20	C-ICPE-TD3	35.8 [19.2,67.6]	33.0 [27.2,42.2]	26.7 [23.1,32.3]	28.2 [26.9,29.5]
	C-ICPE-TS	**16.4** [16.2,16.6]	**23.3** [23.0,23.5]	**17.6** [17.3,17.9]	**22.7** [21.8,23.5]
	C-ICPE-TTPS	16.4 [16.1,16.6]	24.4 [24.1,24.7]	18.7 [18.0,19.4]	24.2 [23.2,25.5]
	C-ICPE-uniform	64.0 [62.0,66.0]	47.5 [29.5,62.4]	81.1 [71.6,88.2]	59.6 [19.4,91.9]

Table 3. Binary search: sample complexity (mean and 95% CI) for every (d, ε, σ) configuration.

D.3.2. ϵ-BEST ARM PROBLEM

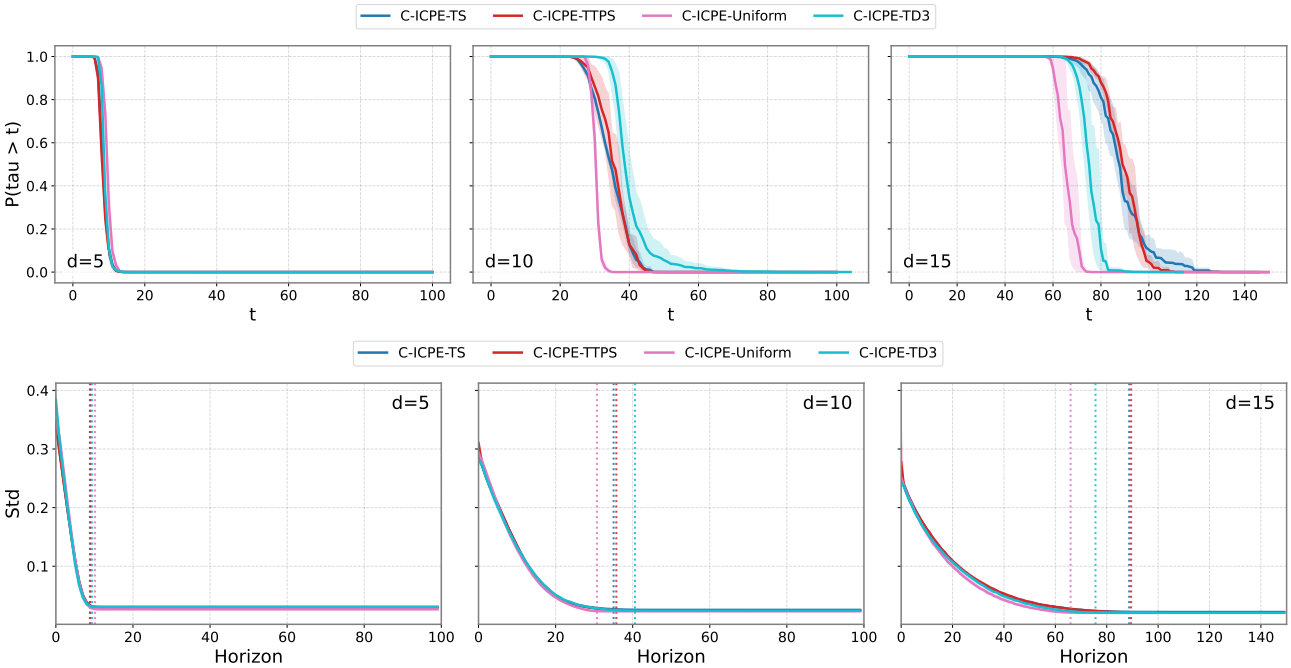


Figure 11. Results for ϵ-Best-Arm Identification Problem with fixed confidence $\delta = 0.005$ and $N = 100/150$ across different dimensions at the most hardest (ε, σ) setting: (top) survival function of τ; (bottom) inference uncertainty convergence.

Tables 4 and 5 report accuracy and sample complexity; Figure 11 shows the survival function and inference uncertainty convergence.

Accuracy. All **C-ICPE** variants meet the $1 - \delta$ target across every (d, ε, σ) configuration, with mean accuracy between 0.897 and 0.939. Notably, **C-ICPE-Uniform** also meets the target throughout (accuracy 0.897–0.936), 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.

Sample complexity. The gap between **C-ICPE** and LT&S is substantial and grows with dimension. LT&S queries exactly d samples in every configuration, one per canonical direction, achieving the minimum for a full-rank linear estimator. **C-ICPE** uses 2–6× more queries: at $d = 15$, $(\varepsilon, \sigma) = (0.005, 0.0025)$, **C-ICPE-TS** stops at 88.8 and **C-ICPE-TTPS** at 89.5, compared to 15.0 for LT&S. This gap reflects the cost of a model-agnostic stopping rule: **C-ICPE** does not know the observation model is linear and must learn when to stop from interaction data alone. Among **C-ICPE** variants, **C-ICPE-Uniform** is competitive with and sometimes more sample-efficient than the active actors. At $d = 15$, $(\varepsilon, \sigma) = (0.005, 0.0025)$, **C-ICPE-Uniform** stops at 65.9 while **C-ICPE-TS** requires 88.8 and **C-ICPE-TTPS** 89.5. This inversion occurs because the problem's symmetry makes directed exploration unnecessary, and the overhead of posterior-driven action selection, which occasionally concentrates queries in already well-estimated directions, slows convergence relative to uniform coverage. We also note that the current noise levels are low relative to ε, placing the problem in a regime where LT&S 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.

Stopping behavior. The survival functions (Figure 11, top) show that all **C-ICPE** actors have similar stopping profiles at $d = 5$, where episodes terminate within $t \approx 10$. At $d = 15$, the curves spread: **C-ICPE-Uniform** stops earlier (median ≈ 60) than **C-ICPE-TS** and **C-ICPE-TTPS** (median $\approx$ 75–85), again reflecting the advantage of isotropic coverage in this symmetric problem. The inference standard deviation (bottom row) converges at comparable rates across all actors, confirming that the posterior contracts uniformly regardless of the exploration strategy, the sample complexity differences are driven by stopping calibration, not by differences in information acquisition.

d	Method	$\varepsilon = 0.02$		$\varepsilon = 0.005$	
		$\sigma = 0.005$	$\sigma = 0.01$	$\sigma = 0.00125$	$\sigma = 0.0025$
5	C-ICPE-TD3	0.925 [.907,.943]	0.915 [.905,.926]	0.898 [.890,.907]	0.902 [.893,.912]
	C-ICPE-TS	0.939 [.912,.958]	0.939 [.929,.947]	0.908 [.890,.930]	0.904 [.891,.917]
	C-ICPE-TTPS	0.934 [.922,.944]	0.938 [.930,.944]	0.902 [.891,.911]	0.900 [.893,.909]
	C-ICPE-uniform	0.936 [.925,.946]	0.931 [.919,.943]	0.922 [.914,.928]	0.927 [.918,.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]
10	C-ICPE-TD3	0.920 [.908,.931]	0.901 [.891,.909]	0.908 [.896,.919]	0.919 [.906,.932]
	C-ICPE-TS	0.902 [.893,.911]	0.909 [.897,.920]	0.908 [.892,.922]	0.915 [.899,.930]
	C-ICPE-TTPS	0.917 [.907,.926]	0.904 [.897,.911]	0.911 [.900,.921]	0.906 [.892,.920]
	C-ICPE-uniform	0.911 [.901,.921]	0.905 [.895,.913]	0.905 [.894,.916]	0.900 [.890,.910]
	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	0.904 [.894,.913]	0.906 [.896,.915]	0.909 [.898,.920]	0.910 [.902,.917]
	C-ICPE-TS	0.918 [.898,.937]	0.918 [.900,.932]	0.908 [.895,.921]	0.898 [.889,.907]
	C-ICPE-TTPS	0.907 [.894,.919]	0.907 [.894,.919]	0.910 [.887,.934]	0.923 [.893,.959]
	C-ICPE-uniform	0.897 [.887,.905]	0.902 [.889,.913]	0.906 [.893,.919]	0.901 [.892,.910]
	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 4. ϵ-best arm: accuracy (mean and 95% CI) 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	7.4 [7.3,7.5]	7.5 [7.4,7.5]	9.4 [9.2,9.6]	9.4 [9.3,9.5]
	C-ICPE-TS	7.3 [7.2,7.5]	7.6 [7.5,7.6]	8.9 [8.6,9.1]	9.0 [8.8,9.2]
	C-ICPE-TTPS	7.3 [7.1,7.4]	7.5 [7.4,7.5]	8.7 [8.6,8.9]	8.8 [8.7,9.0]
	C-ICPE-uniform	8.3 [8.2,8.4]	8.2 [8.0,8.4]	10.0 [9.9,10.1]	10.1 [10.0,10.3]
	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	21.7 [21.1,22.6]	21.9 [21.6,22.3]	37.3 [35.3,39.0]	40.6 [38.2,43.3]
	C-ICPE-TS	20.0 [19.5,20.7]	23.9 [22.8,25.3]	31.4 [30.1,33.7]	35.0 [34.7,35.5]
	C-ICPE-TTPS	20.7 [20.4,21.0]	22.8 [22.6,23.0]	32.7 [30.9,34.6]	35.8 [33.8,37.3]
	C-ICPE-uniform	22.1 [21.9,22.3]	22.4 [22.2,22.5]	31.1 [30.8,31.4]	30.7 [30.5,31.1]
	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	46.9 [46.5,47.2]	48.6 [48.1,49.2]	71.4 [70.2,72.6]	75.7 [73.4,78.9]
	C-ICPE-TS	46.5 [46.0,47.1]	53.8 [53.1,54.7]	74.8 [73.3,76.3]	88.8 [85.4,92.7]
	C-ICPE-TTPS	45.2 [44.0,47.0]	52.7 [50.9,54.1]	78.9 [74.6,87.0]	89.5 [87.7,91.7]
	C-ICPE-uniform	44.5 [43.3,45.8]	45.1 [43.6,46.9]	68.1 [65.6,70.8]	65.9 [64.1,69.4]
	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 5. ϵ-best arm: sample complexity (mean and 95% CI) for every (d, ε, σ) configuration.

D.3.3. ACKLEY MINIMIZATION

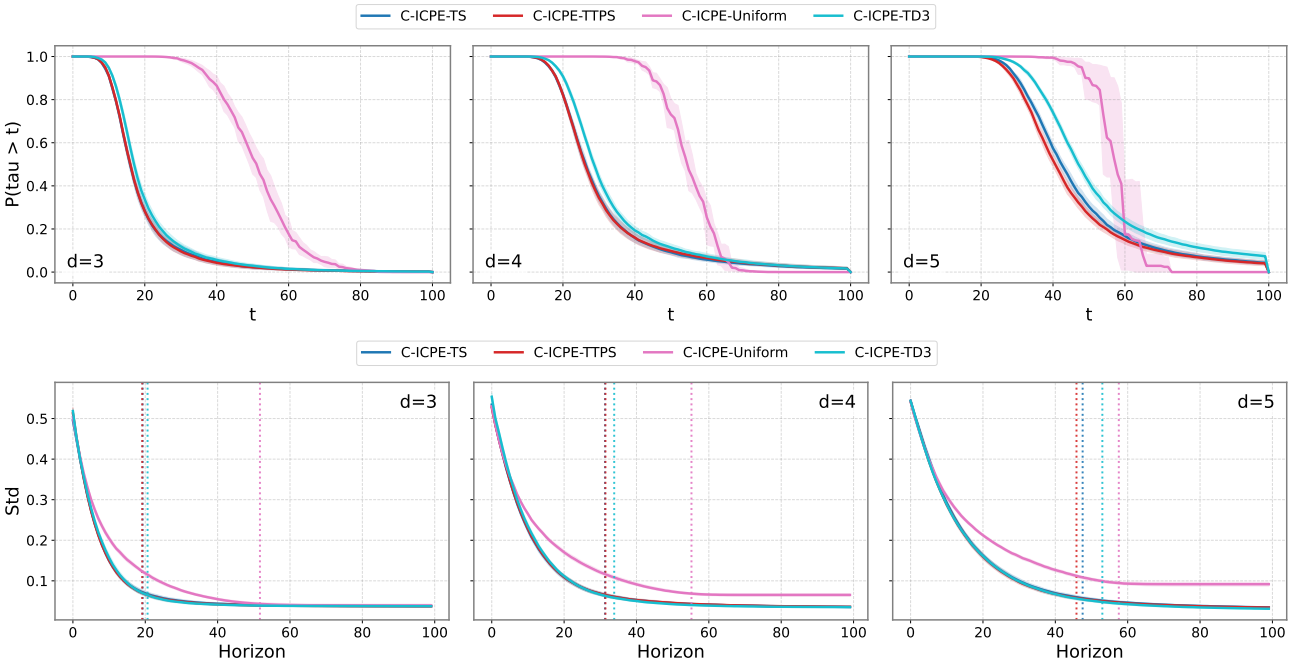


Figure 12. Results for Ackley function with fixed confidence $\delta = 0.1$ and $N = 100$ across different dimensions at the most hardest (ε, σ) setting: (top) survival function of τ; (bottom) inference uncertainty convergence.

Tables 7 and 6 report accuracy and sample complexity; Figure 12 shows the survival function and inference uncertainty convergence.

Accuracy. All active **C-ICPE** variants meet the $1 - \delta = 0.90$ target across every (d, ε, σ) configuration, with mean accuracy between 0.896 and 0.928. **C-ICPE-Uniform** meets the target at $d = 3$ (accuracy ≥ 0.890) but degrades as dimension increases: at $(\varepsilon, \sigma) = (0.1, 0.05)$, accuracy drops from 0.890 at $d = 3$ to 0.589 at $d = 4$ and 0.237 at $d = 5$. Unlike the ε-best arm problem, the Ackley function has no symmetry that makes uniform exploration competitive: the multimodal landscape and flat outer region require directed queries to distinguish the global minimizer from local optima. The Bayesian optimization baselines fail across the board. GP-UCB achieves the highest accuracy among them (0.344–0.426 at $d = 3$) but remains far below the 0.90 target even at the easiest configuration. TPE and CMA-ES are near zero at $d \geq 4$. These methods optimize a fixed-budget objective without a stopping rule and are not designed for (ε, δ)-correctness; the comparison confirms that standard continuous optimization does not yield fixed-confidence guarantees at comparable sample budgets.

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 47.5 and **C-ICPE-TTPS** at 45.8, compared to 57.5 for **C-ICPE-Uniform**. **C-ICPE-TS** and **C-ICPE-TTPS** achieve similar sample complexity throughout, while **C-ICPE-TD3** is slightly higher (e.g., 53.0 at the same setting). Sample complexity grows with dimension for all methods: **C-ICPE-TS** increases from 14.9 ($d = 3$) to 34.4 ($d = 5$) at $(\varepsilon, \sigma) = (0.2, 0.05)$. Increasing σ at fixed ε consistently raises sample complexity, as expected from the noisier observations.

Stopping behavior. The survival functions (Figure 12, 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 $t = 80$. 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.

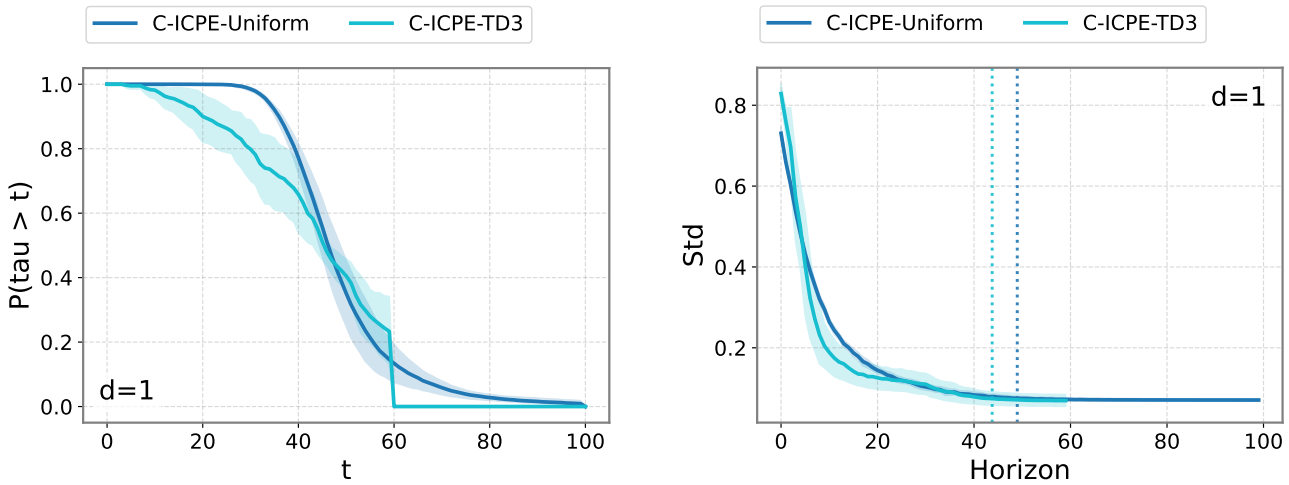
d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
3	C-ICPE-TD3	15.3 [14.6,15.9]	18.9 [18.2,19.6]	18.9 [18.2,19.8]	20.6 [19.9,21.3]
	C-ICPE-TS	**14.9** [14.5,15.5]	18.8 [18.1,19.7]	**17.5** [16.9,18.0]	19.3 [18.5,20.2]
	C-ICPE-TTPS	15.4 [14.8,16.1]	**18.7** [18.1,19.4]	18.0 [17.3,18.7]	**19.2** [18.6,19.9]
	C-ICPE-uniform	37.6 [34.4,40.2]	47.9 [46.3,49.4]	50.5 [48.6,52.2]	51.7 [49.8,53.7]
4	C-ICPE-TD3	25.9 [25.0,26.7]	32.8 [31.8,33.9]	31.5 [30.5,32.6]	33.9 [33.1,34.7]
	C-ICPE-TS	24.4 [23.2,25.8]	**31.1** [29.7,32.6]	**27.5** [26.3,28.9]	**31.3** [30.4,32.1]
	C-ICPE-TTPS	**24.2** [22.7,25.8]	32.2 [31.0,33.6]	27.9 [27.0,29.0]	31.4 [30.4,32.4]
	C-ICPE-uniform	62.4 [60.3,64.6]	59.4 [56.8,61.4]	57.8 [57.3,58.3]	55.2 [54.0,56.5]
5	C-ICPE-TD3	38.1 [37.0,39.3]	50.7 [49.3,52.3]	44.1 [42.2,46.2]	53.0 [51.9,54.3]
	C-ICPE-TS	**34.4** [33.3,35.9]	**45.8** [44.2,47.1]	38.5 [37.5,39.5]	47.5 [46.3,48.6]
	C-ICPE-TTPS	35.6 [33.4,38.0]	47.0 [45.7,48.1]	**37.7** [36.9,38.9]	**45.8** [44.8,46.9]
	C-ICPE-uniform	70.0 [63.9,79.2]	58.2 [57.1,59.7]	61.0 [55.3,68.6]	57.5 [55.1,60.1]

Table 6. Ackley: sample complexity (mean and 95% CI) for every (d, ε, σ) configuration.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.1$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.025$	$\sigma = 0.05$
3	C-ICPE-TD3	**0.915** [.902,.927]	0.904 [.894,.914]	0.898 [.888,.906]	0.908 [.899,.918]
	C-ICPE-TS	0.896 [.886,.904]	0.899 [.890,.909]	0.899 [.887,.906]	0.907 [.895,.917]
	C-ICPE-TTPS	0.902 [.894,.913]	**0.907** [.894,.919]	**0.911** [.900,.920]	**0.912** [.899,.924]
	C-ICPE-uniform	0.902 [.886,.915]	0.905 [.890,.918]	0.906 [.894,.914]	0.890 [.868,.910]
	TPE	0.098 [.072,.124]	0.136 [.106,.166]	0.016 [.005,.027]	0.028 [.014,.042]
	CMA-ES	0.086 [.061,.111]	0.124 [.095,.153]	0.020 [.008,.032]	0.034 [.018,.050]
	GP-logEI	0.322 [.281,.363]	0.356 [.314,.398]	0.218 [.182,.254]	0.194 [.159,.229]
	GP-UCB	0.400 [.357,.443]	0.426 [.383,.469]	0.344 [.302,.386]	0.350 [.308,.392]
4	C-ICPE-TD3	0.907 [.895,.919]	**0.906** [.895,.914]	0.899 [.888,.908]	0.899 [.891,.907]
	C-ICPE-TS	0.907 [.899,.918]	0.904 [.892,.916]	**0.914** [.905,.924]	**0.913** [.900,.923]
	C-ICPE-TTPS	0.898 [.886,.908]	0.898 [.888,.908]	0.908 [.897,.919]	0.906 [.894,.915]
	C-ICPE-uniform	**0.908** [.897,.919]	0.739 [.713,.765]	0.735 [.712,.762]	0.589 [.560,.616]
	TPE	0.048 [.029,.067]	0.084 [.060,.108]	0.006 [.000,.013]	0.006 [.000,.013]
	CMA-ES	0.020 [.008,.032]	0.056 [.036,.076]	0.002 [.000,.006]	0.002 [.000,.006]
	GP-logEI	0.270 [.231,.309]	0.206 [.171,.241]	0.132 [.102,.162]	0.104 [.077,.131]
	GP-UCB	0.344 [.302,.386]	0.308 [.267,.349]	0.336 [.295,.377]	0.276 [.237,.315]
5	C-ICPE-TD3	0.904 [.888,.918]	0.904 [.891,.916]	0.907 [.892,.920]	0.904 [.891,.916]
	C-ICPE-TS	**0.909** [.900,.920]	**0.912** [.899,.921]	0.913 [.905,.925]	0.911 [.897,.921]
	C-ICPE-TTPS	0.905 [.893,.915]	0.904 [.889,.915]	**0.928** [.916,.939]	**0.912** [.895,.927]
	C-ICPE-uniform	0.690 [.618,.756]	0.416 [.401,.437]	0.428 [.332,.538]	0.237 [.203,.279]
	TPE	0.030 [.015,.045]	0.042 [.024,.060]	0.000 [.000,.000]	0.002 [.000,.006]
	CMA-ES	0.014 [.004,.024]	0.040 [.023,.057]	0.002 [.000,.006]	0.004 [.000,.010]
	GP-logEI	0.154 [.122,.186]	0.074 [.051,.097]	0.068 [.046,.090]	0.044 [.026,.062]
	GP-UCB	0.344 [.302,.386]	0.214 [.178,.250]	0.284 [.244,.324]	0.220 [.184,.256]

Table 7. Ackley: accuracy (mean and 95% CI) for every (d, ε, σ) configuration.

D.3.4. GP MAX-VALUE ESTIMATION



(a) *(b)*

Figure 13. Results for GP value estimation with fixed confidence $\delta = 0.1$ and $N = 100$ across different dimensions at the most hardest (ε, σ) setting: (a) survival function of τ; (b) inference uncertainty convergence.

Tables 8 and 9 report accuracy and sample complexity; Figure 13 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 $\mathcal{X} = \mathcal{A}$. In Fig. 9 we also show ho **C-ICPE** explores in this problem, depicting the queries chosen by the actor, the posterior mean, and the posterior standard deviation over timesteps.

Accuracy. Both **C-ICPE** variants meet the $1-\delta$ target across all (ε, σ) configurations. **C-ICPE-TD3** achieves 0.906–0.930 and **C-ICPE-Uniform** 0.901–0.931; the two methods are comparable in accuracy, with **C-ICPE-Uniform** slightly higher at the noisier settings (e.g., 0.927 vs. 0.910 at $\varepsilon = 0.2$, $\sigma = 0.1$). The non-parametric baselines fail to meet the target: Uniform bin achieves 0.130–0.230 and Uniform top 5% reaches 0.648–0.762. Bayesian optimization baselines are not applicable to this task, as they return locations rather than value estimates.

Sample complexity. **C-ICPE-TD3** uses substantially fewer samples than **C-ICPE-Uniform**, particularly at the larger tolerance. At $(\varepsilon, \sigma) = (0.2, 0.05)$, **C-ICPE-TD3** stops at 13.7 queries on average compared to 31.7 for **C-ICPE-Uniform** — a 2.3× reduction. At $(\varepsilon, \sigma) = (0.2, 0.1)$ the ratio is 1.8× (22.7 vs. 41.0). The gap narrows at $\varepsilon = 0.15$: 39.9 vs. 41.6 at $\sigma = 0.0375$ and 43.7 vs. 49.0 at $\sigma = 0.075$. This pattern indicates that the learned exploration policy provides the largest benefit when the tolerance is generous enough that a well-chosen sequence of queries can resolve the max-value quickly, whereas at tighter tolerances both methods require extensive coverage and the advantage of directed exploration diminishes.

Stopping behavior. The survival function (left plot in Fig. 13) 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. 13) 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.

d	Method	$\varepsilon = 0.2$		$\varepsilon = 0.15$	
		$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.0375$	$\sigma = 0.075$
1	C-ICPE-TD3	**0.906** [.878,.928]	0.910 [.883,.932]	**0.920** [.890,.942]	0.930 [.903,.948]
	C-ICPE-uniform	0.902 [.889,.914]	**0.927** [.915,.938]	0.901 [.887,.916]	**0.931** [.915,.946]
	Uniform bin	0.130 [.100,.160]	0.176 [.143,.209]	0.230 [.193,.267]	0.226 [.189,.263]
	Uniform top 5%	0.648 [.606,.690]	0.672 [.631,.713]	0.762 [.725,.799]	0.696 [.656,.736]

Table 8. GP value estimation: accuracy (mean and 95% CI) for every (d, 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	**13.7** [12.3,15.4]	**22.7** [21.8,23.8]	**39.9** [36.6,43.4]	**43.7** [41.3,45.8]
	C-ICPE-uniform	31.7 [30.6,32.8]	41.0 [39.8,42.1]	41.6 [39.5,43.9]	49.0 [46.7,51.2]

Table 9. GP value estimation: sample complexity (mean and 95% CI) for every (d, ε, σ) configuration.

D.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. 14 to 17 report accuracy and confidence intervals for all actor variants at the most challenging (ε, σ) configuration and highest dimensionality per benchmark.

D.4.1. Noisy Binary Search

The training prior is Uniform$[-1, 1]^{20}$, corresponding to $\alpha = \beta = 1$ (white box). In Fig. 14 we report the results. **C-ICPE-TS** and **C-ICPE-TTPS** are robust across the full grid of Beta priors: accuracy remains above 0.83 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. **C-ICPE-TD3** matches or exceeds the other actors when the deployment prior is concentrated ($\alpha, \beta \geq 3$, reaching 0.943), but degrades sharply when either parameter is small: at $\alpha = 0.5, \beta = 7$ accuracy drops to 0.612. 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.

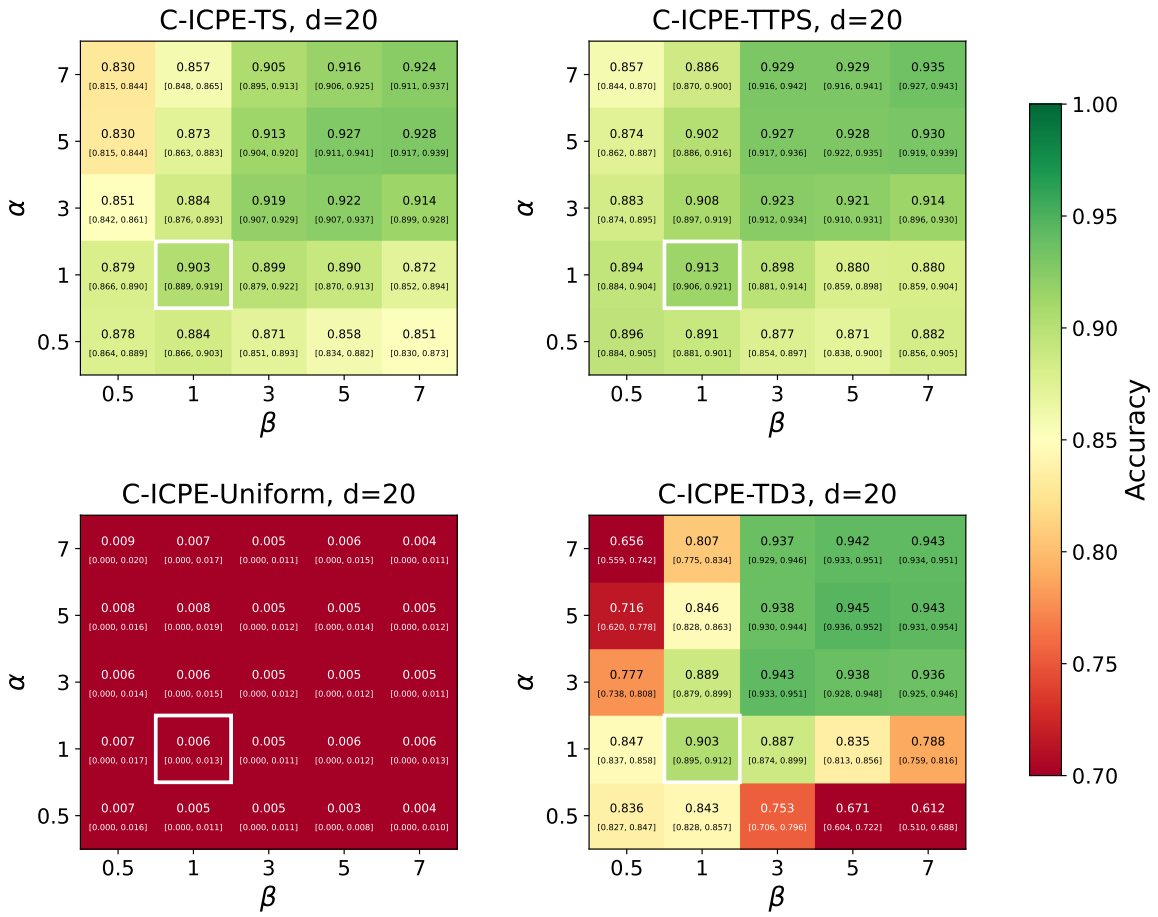


Figure 14. Robustness to prior misspecification on the 20D noisy binary search problem ($\varepsilon = 0.1$ and $\sigma = 0.05$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.

D.4.2. ϵ-BEST ARM PROBLEM

We report results in Fig. 15. All four **C-ICPE** variants are remarkably stable across the full Beta prior grid at $d = 15$: accuracy varies by less than 0.04 across all (α, β) configurations for each actor. This robustness is a direct consequence of the rotational symmetry of the problem. Since θ is drawn on $\mathbb{S}^{d-1}$ and the loss $L_\theta(x) = 1 - \theta^\top x$ is invariant to orthogonal transformations, the intrinsic difficulty of each task instance does not depend on the location of θ on the sphere. Reweighting the prior therefore has little effect on the distribution of problem difficulty, unlike binary search or Ackley where boundary effects create heterogeneous difficulty across the parameter space. Notably, **C-ICPE-Uniform** performs well here ($\approx$ 0.89–0.90 uniformly), consistent with the observation that isotropic exploration is optimal [Jedra and Proutiere, 2020]. **C-ICPE-TTPS** achieves the highest accuracy overall (0.92–0.94), suggesting that the challenger mechanism provides a modest benefit even in a setting where uniform exploration is already near-optimal.

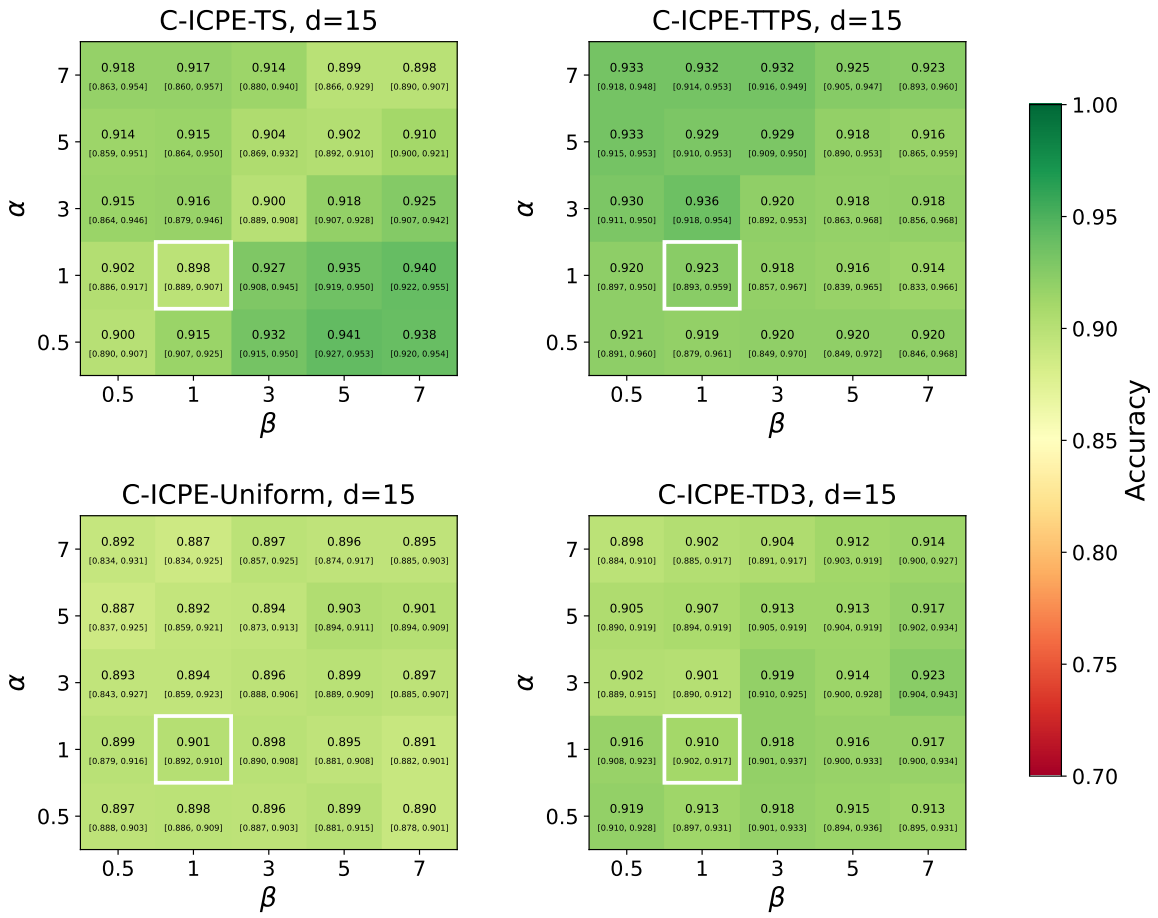


Figure 15. Robustness to prior misspecification on the 15D ε-best-arm identification problem ($\varepsilon = 0.005$ and $\sigma = 0.0025$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.

D.4.3. ACKLEY MINIMIZATION

Results are reported in Fig. 16. The Ackley benchmark at $d = 5$ exhibits the strongest sensitivity to prior misspecification among all tasks. At the training prior ($\alpha = \beta = 1$), **C-ICPE-TS** achieves 0.911 and **C-ICPE-TTPS** 0.912. 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, with **C-ICPE-TS** reaching 0.975 and **C-ICPE-TD3** 0.963. The gains reflect the structure of the Ackley function: 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 or β 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, degrade. 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. **C-ICPE-Uniform** fails across the board (accuracy ≤ 0.45), confirming that directed exploration is essential for this multimodal benchmark regardless of the prior.

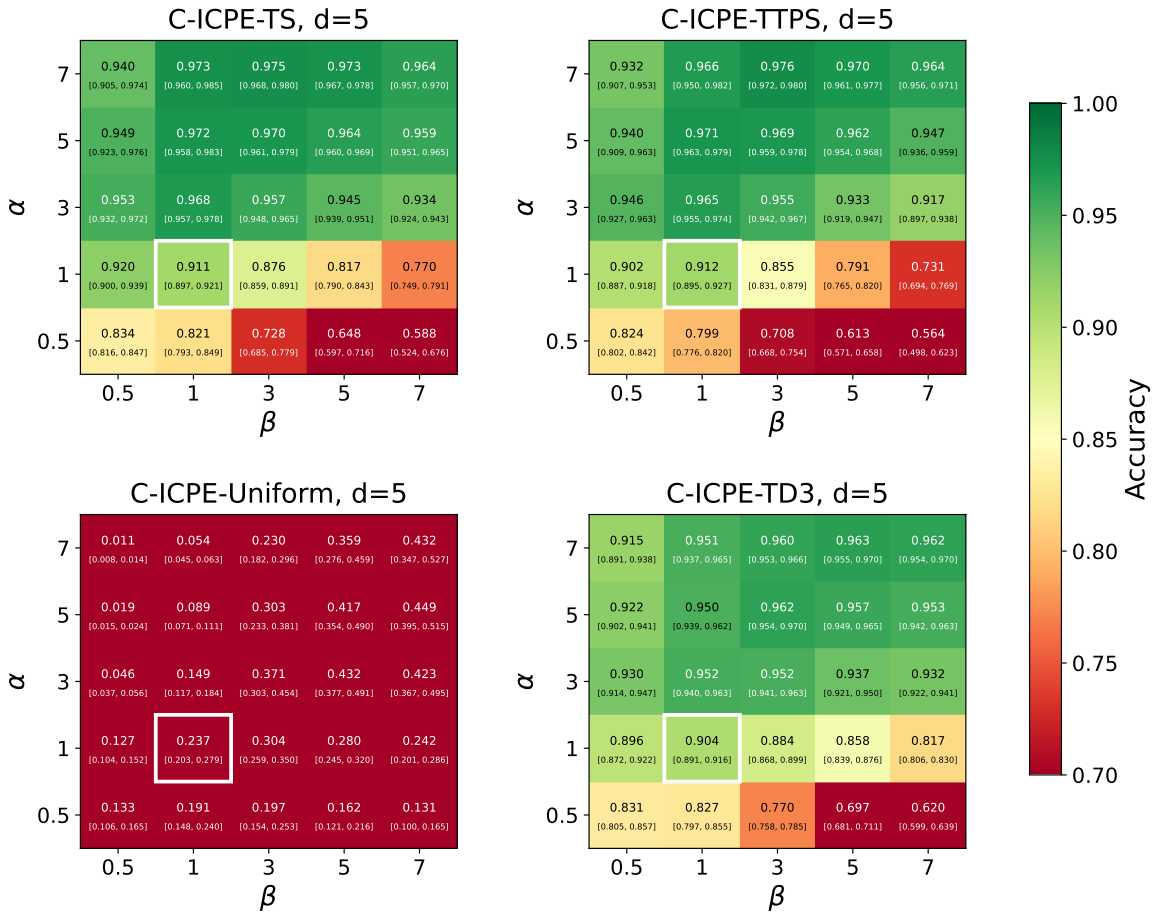


Figure 16. Robustness to prior misspecification on the 5D Ackley function ($\varepsilon = 0.1$ and $\sigma = 0.05$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.

D.4.4. GP MAX-VALUE ESTIMATION

We report results in Fig. 17. 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 0.992 at $(\alpha, \beta) = (7, 0.5)$ and remains above 0.94 throughout the upper-left triangle of the grid. However, it degrades when β is large (0.772 at $\alpha = 0.5, \beta = 7$; 0.817 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 0.837 and 0.965 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 comparable (0.920 vs. 0.913), 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.

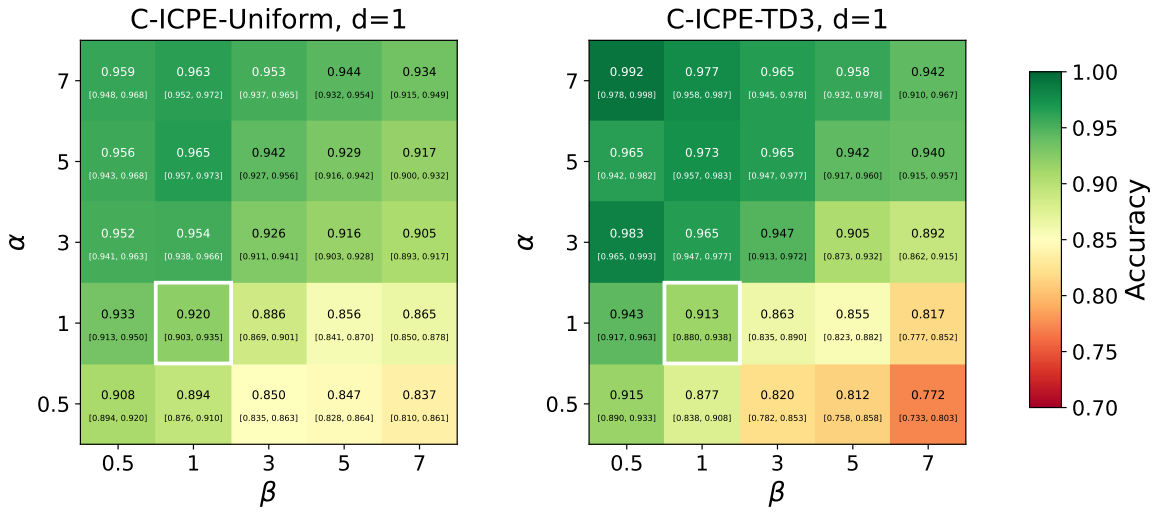


Figure 17. Robustness to prior misspecification on the GP value estimation problem ($\varepsilon = 0.15$ and $\sigma = 0.075$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.

D.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.

D.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. 19 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. 19 [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.

D.5.2. Region Partitioning and GP Fitting

We partition the geochemical survey data into 17 geographic regions, each covering approximately $1^\circ \times 2^\circ$ 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. 20 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:

$$f \sim \mathrm{GP}(0, \sigma_f^2\, k_{\text{Matérn-3/2}}(\cdot,\cdot;\ell_1,\ell_2)), \qquad y_i = f(\mathbf{s}_i) + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \sigma_n^2), \tag{38}$$

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

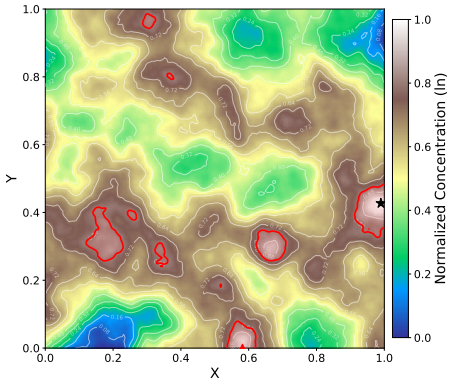


Figure 18. Example copper concentration in a 2D region in the geochemical exploration task. Red regions indicate concentration of copper within ϵ of the maximum value.

concentration, and $(\ell_1, \ell_2, \sigma_f, \sigma_n)$ 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. 21 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).

D.5.3. Ground Truth Construction

For each fitted SVGP, we evaluate the posterior mean on a dense 200×200 grid over the normalized domain $[0, 1]^2$. The ground truth target is defined as:

$$\theta^\star = \arg\max_{\mathbf{s} \in \text{grid}} \mu_{\text{GP}}(\mathbf{s}), \tag{39}$$

i.e., the grid location with the highest posterior mean copper concentration.

D.5.4. Task Prior and Train/Test Split

The 17 regions are split into training and evaluation sets. 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 the corresponding fitted GP as the unknown function. The agent queries 2D locations $a \in [0, 1]^2$ and observes noisy evaluations $y = \mu_{\text{GP}}(a) + \xi$, $\xi \sim \mathcal{N}(0, \sigma_n^2)$, where σ_n is the fitted noise standard deviation for that region. The goal is to identify the location $\theta^\star$ of peak concentration to within ϵ 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. **(ϵ, δ)-correct identification on realistic functions.** The GP posterior means are spatially structured, non-stationary (due to irregular sampling), and vary in smoothness across regions, a substantial departure from the synthetic benchmarks.
2. **Robustness to distribution shift.** The evaluation regions have different hyperparameters $(\ell_1, \ell_2, \sigma_f, \sigma_n)$ from the training regions, so the agent must generalize across spatial correlation structures it has not encountered during meta-training.

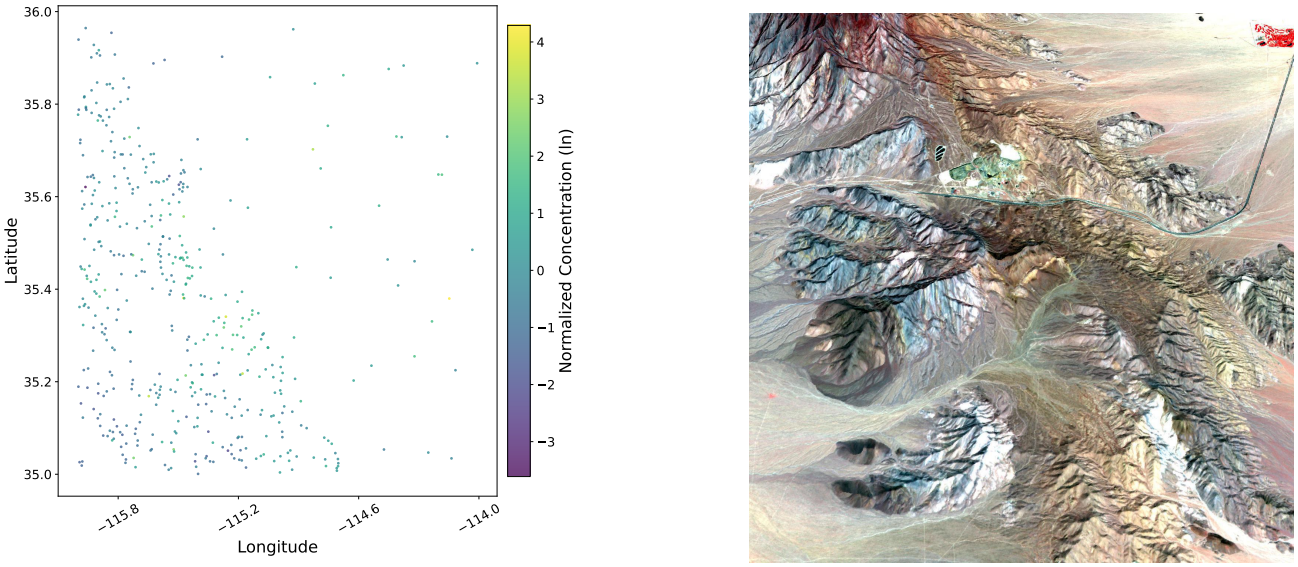


Figure 19. **Left:** Normalized log-copper concentration in the Kingman region (southeastern California / southern Nevada). Each point is a soil sample from the USGS National Geochemical Survey [usg, 2004]; the red star marks the location of peak GP posterior mean. **Right:** ASTER satellite image of 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].

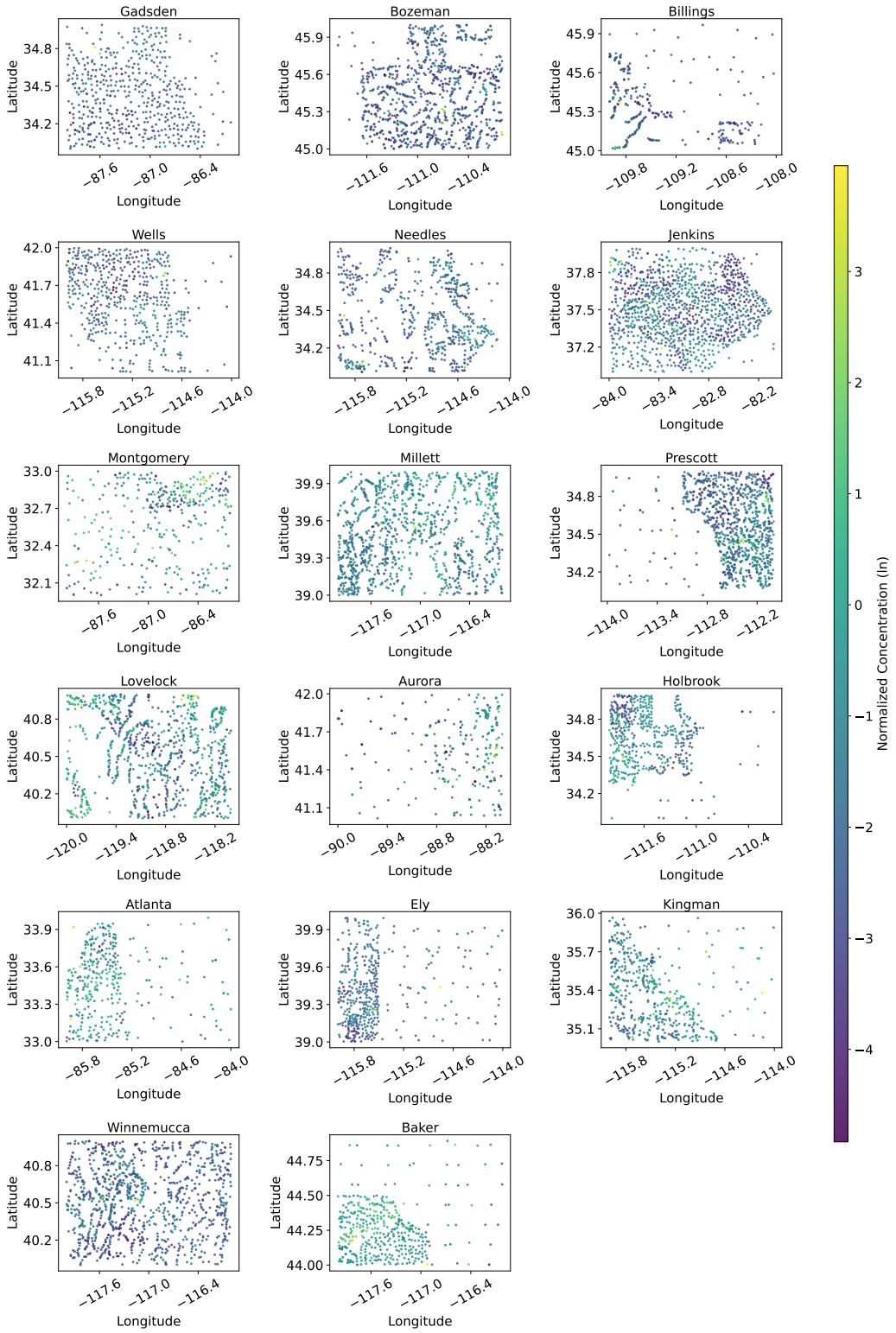


Figure 20. All 17 geographic regions used in the geochemical experiment. Each panel shows soil sample locations colored by normalized log-copper concentration. Regions are split into training and evaluation sets; evaluation regions have spatial structure not seen during meta-training.

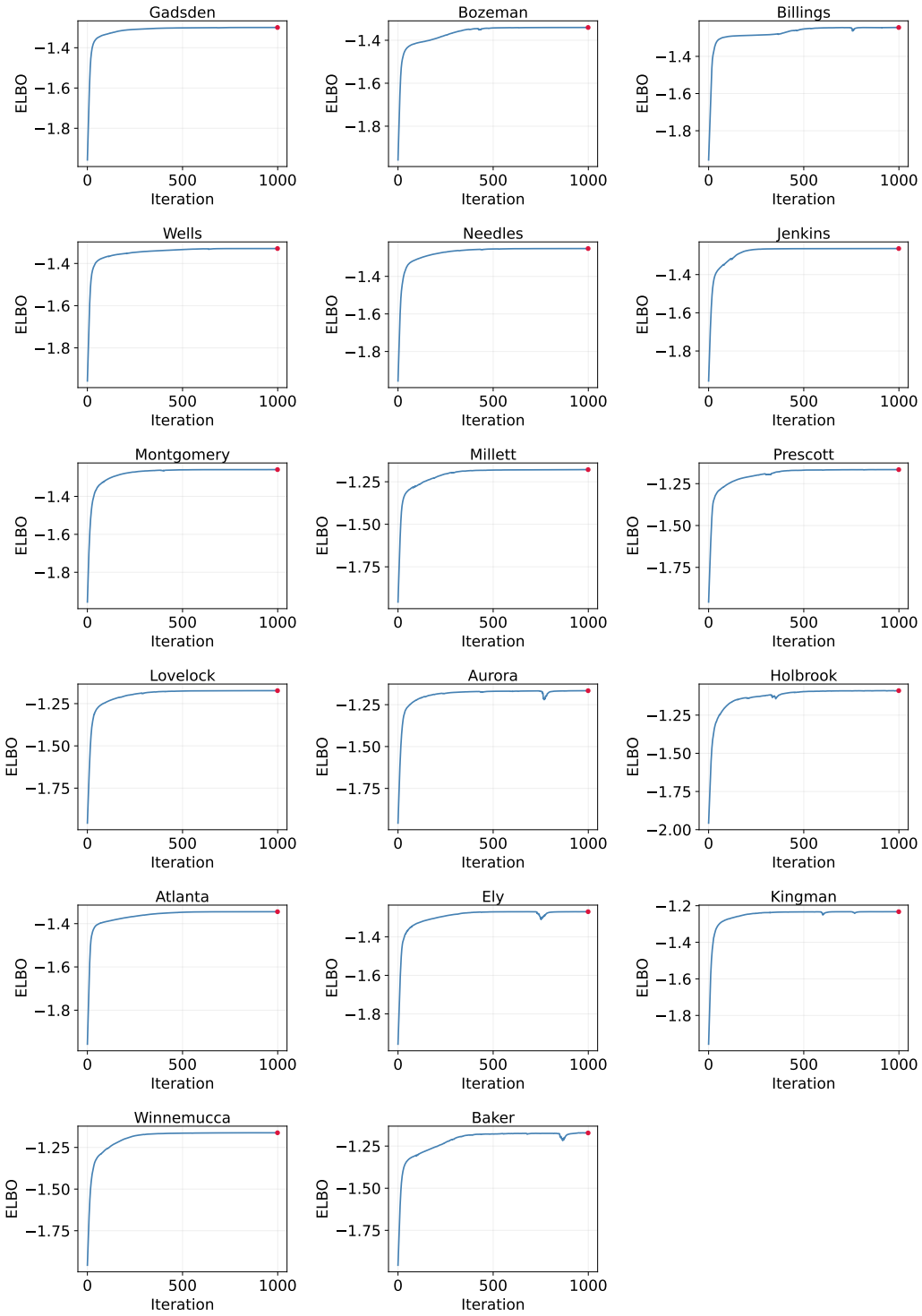


Figure 21. ELBO training curves for sparse variational GP fits across all 17 regions (1000 Adam iterations, $M = 500$ inducing points). All regions converge smoothly, indicating well-behaved GP fits.

D.5.5. Numerical Results

Baselines. We compare C-ICPE-TD3 and C-ICPE-TS against four black-box optimization baselines that do not possess a stopping rule for (ε, δ)-correct identification: TPE [Bergstra et al., 2011], CMA-ES [Hansen, 2016], GP-logEI [Ament et al., 2023], and GP-UCB [Srinivas et al., 2010]. Since these methods optimize a fixed-budget objective and have no principled mechanism for adaptive stopping, we allocate each a fixed sample budget calibrated to the mean sample complexity of C-ICPE-TS: $N = 22$ for $\varepsilon = 0.2$ and $N = 39$ for $\varepsilon = 0.15$. After exhausting this budget, each baseline returns the best-observed location as its recommendation. This protocol is deliberately generous: the baselines receive as many samples as `C-ICPE-TS` typically needs on average, yet bear no cost for deciding when to stop.

		$\varepsilon = 0.2$ $\sigma = 0$	$\varepsilon = 0.15$ $\sigma = 0$
2	C-ICPE-TD3	**0.913** [.894,.931]	0.916 [.904,.927]
	C-ICPE-TS	0.913 [.894,.930]	**0.925** [.905,.944]
	TPE	0.438 [.394,.482]	0.526 [.482,.570]
	CMA-ES	0.484 [.440,.528]	0.514 [.470,.558]
	GP-logEI	0.564 [.520,.608]	0.724 [.685,.763]
	GP-UCB	0.730 [.691,.769]	0.720 [.681,.759]

Table 10. Geochem: accuracy (mean and 95% CI) for every (d, ε, σ) configuration.

		$\varepsilon = 0.2$ $\sigma = 0$	$\varepsilon = 0.15$ $\sigma = 0$
2	C-ICPE-TD3	23.9 [21.6,26.2]	44.4 [42.4,46.5]
	C-ICPE-TS	**22.0** [20.0,23.9]	**38.7** [34.2,43.5]

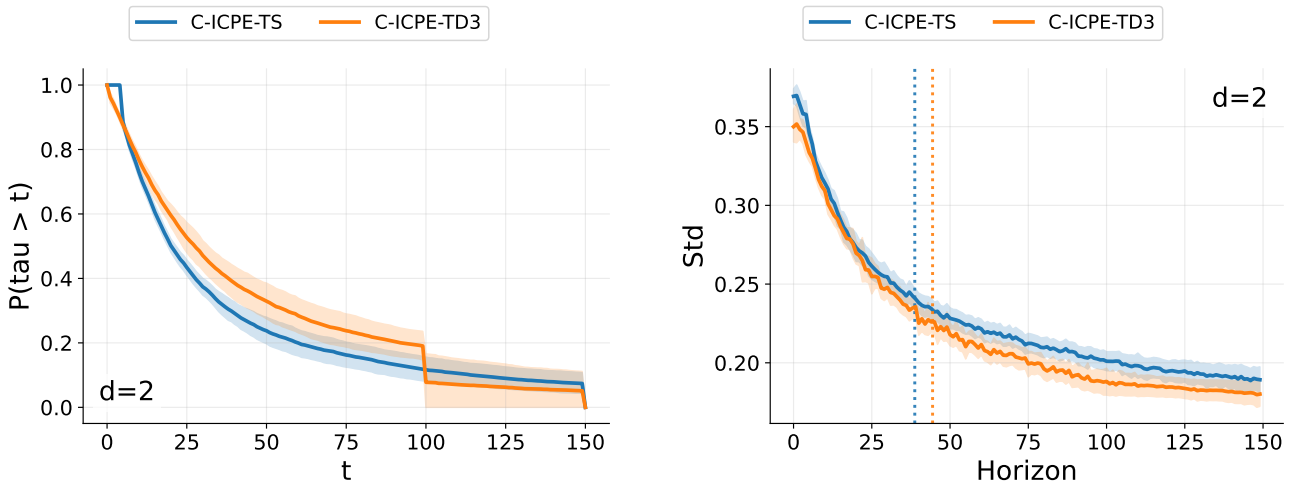
Table 11. Geochem: sample complexity (mean and 95% CI) for every (d, ε, σ) configuration.

Accuracy and sample complexity. Table 10 reports identification accuracy (mean and 95% CI over held-out regions) at the two tolerance levels, with $\delta = 0.1$ and maximum horizon $N = 150$. Both C-ICPE variants exceed the $1 - \delta = 0.90$ correctness target in every configuration. At $\varepsilon = 0.2$, C-ICPE-TD3 and C-ICPE-TS are tied at 0.913 $[.894, .931]$ and 0.913 $[.894, .930]$, respectively. At the harder $\varepsilon = 0.15$, C-ICPE-TS pulls ahead with 0.925 $[.905, .944]$ versus 0.916 $[.904, .927]$ for C-ICPE-TD3. Among the baselines, GP-UCB is the strongest, reaching 0.730 $[.691, .769]$ at $\varepsilon = 0.2$ and 0.720 $[.681, .759]$ at $\varepsilon = 0.15$, still roughly 19–20 percentage points below C-ICPE despite receiving a comparable sample budget. GP-logEI performs comparably to GP-UCB at $\varepsilon = 0.15$ (0.724) but falls to 0.564 at $\varepsilon = 0.2$. TPE and CMA-ES remain below 0.53 in both settings, indicating that gradient-free search without a surrogate is ineffective on these spatially structured surfaces.

Table 11 reports sample complexity for the two C-ICPE variants. C-ICPE-TS is more sample-efficient in both settings: 22.0 $[20.0, 23.9]$ versus 23.9 $[21.6, 26.2]$ at $\varepsilon = 0.2$, and 38.7 $[34.2, 43.5]$ versus 44.4 $[42.4, 46.5]$ at $\varepsilon = 0.15$. The gap widens at the tighter tolerance, suggesting that Thompson sampling's implicit exploration adapts more efficiently to the difficulty of each region.

Survival function and uncertainty convergence. Figure 22a displays the survival function $\mathbb{P}(\tau > t)$ at the hardest setting ($\varepsilon = 0.15$, $d = 2$). Both variants exhibit a rapid initial decline. The shaded confidence bands for C-ICPE-TD3 are noticeably slightly wider, reflecting higher variance in stopping times, consistent with the wider confidence interval in Table 11.

Figure 22b shows the posterior standard deviation of the recommendation as a function of the horizon. Both methods converge to approximately $\mathrm{Std} \approx 0.19$ by $t = 150$. The vertical dashed lines mark each method's median stopping time; C-ICPE-TS stops earlier than C-ICPE-TD3, and at both stopping points the standard deviation has already dropped below 0.25. This confirms that C-ICPE-TS stops at a point where the inference model's posterior is sufficiently concentrated, rather than stopping prematurely. However, it's interesting to note that the posterior variance has an overall larger decrease with TD3, confirming that C-ICPE-TD3 is learning a good explorative policy. The slightly larger stopping time may then be due to the fact that training was stopped early, and one could have trained for longer for better performance of C-ICPE-TD3.



(a) *(b)*

Figure 22. Results for geochemical problem with fixed confidence $\delta = 0.1$ and $N = 150$ across different dimensions at the most hardest ε setting: (a) survival function of τ; (b) inference uncertainty convergence.

Robustness to prior misspecification. Figure 23 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). C-ICPE-TS maintains accuracy between 0.90 and 0.93 across the entire 5×5 grid, with no discernible degradation even at extreme configurations such as $(\alpha, \beta) = (0.5, 7)$ or $(7, 0.5)$. C-ICPE-TD3 is similarly stable in the upper portion of the grid ($\alpha \geq 3$), but shows a mild decline to 0.90 at $(\alpha, \beta) = (0.5, 5)$ and $(0.5, 7)$. Across all 25 configurations, every cell remains at or above 0.90, satisfying the $1 - \delta$ correctness target. This degree of robustness is notably stronger than what is observed on the Ackley benchmark (Figure 16), where boundary-skewed priors cause substantial degradation.

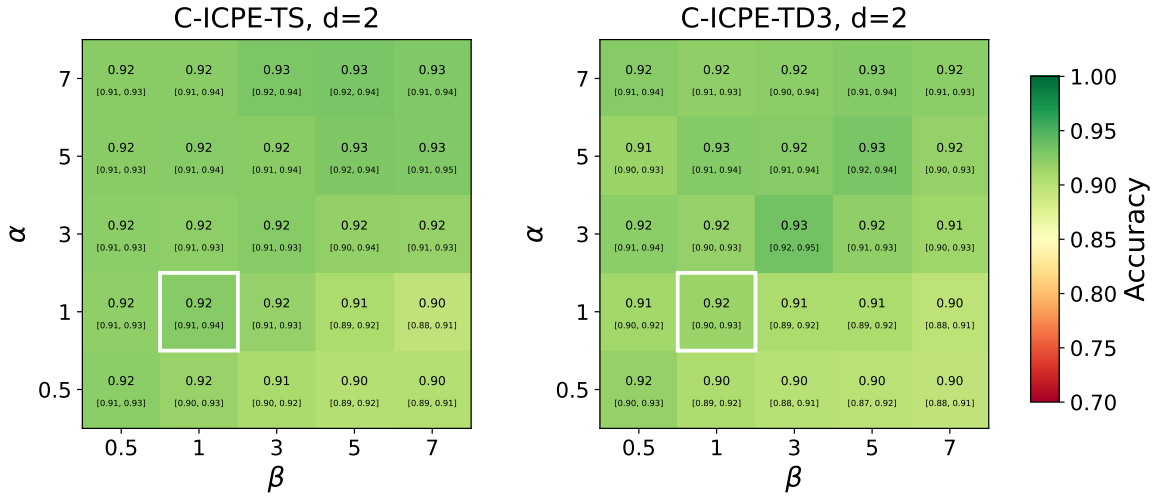


Figure 23. Robustness to prior misspecification on the Geochemical exploration ($\varepsilon = 0.15$). Each heatmap reports the mean accuracy and the confidence intervals under varying Beta prior hyperparameters $(\alpha, \beta) \in \{0.5, 1, 3, 5, 7\}$. The white box indicates the matched prior ($\alpha = \beta = 1$) during training.

Taken together, these results demonstrate that C-ICPE transfers to a real-data task involving genuine distribution shift: the evaluation regions have spatial correlation structures, lengthscales, and noise levels not seen during meta-training, yet both variants maintain (ε, δ)-correctness while using fewer samples than fixed-budget baselines that fail to meet the accuracy target. This validates **C-ICPE** as a practical tool for sequential experimental design.