MULTI-STAGE BAYESIAN OPTIMISATION UNDER UNKNOWN STOPPING CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider Bayesian optimisation over multi-stage decision processes with unknown stage-wise transitions and stopping constraints, where a single violation aborts the current run and prevents all downstream observations. Such problems are common in sequential experimental and manufacturing settings, where failed runs are acceptable during learning but remain costly. To explore uncertain but potentially high-value trajectories under this tolerance while limiting the cumulative number of aborted runs, we formulate the problem as a feasibility-indicated Markov decision process and propose `FI-O-VI`, a value iteration method for optimistic lookahead planning that uses a feasibility threshold that decreases over the optimisation budget to provide fine-grained control over the trade-off between exploration and failure. Under standard RKHS regularity and feasibility assumptions, we derive sublinear rates for both cumulative abort count and cumulative value regret. We demonstrate competitive performance against existing baselines on synthetic benchmarks and real-world problems.

1 INTRODUCTION

The optimisation of many complex real-world systems commonly involves a sequence of interdependent decision stages (e.g., (MacLeod et al., 2020; Dai et al., 2024; Kiroirwal et al., 2026; Buckingham et al., 2024)). At each stage, the outcome from the previous stage informs a decision and, together with that decision, produces a new intermediate outcome through a stage-wise (optionally stochastic, e.g., due to process fluctuations, (Langdon et al., 2026)) *transition* mapping. This outcome becomes the input to the next stage and the same update is repeated. This *multi-stage* decision process proceeds sequentially across stages, while the final objective is only observed after this whole sequential decision making (SDM) process is completed. In order to obtain the optimal final objective value when the stage-wise transitions are unknown and expensive to evaluate, Multi-Stage Bayesian optimisation (MSBO) has emerged as a promising approach (Nguyen et al., 2016; Kusakawa et al., 2022; Torresi & Friederich, 2026): by explicitly modelling the transition mappings and making lookahead decisions, MSBO can improve sample efficiency over treating the entire process as a single black box.

Across many such problems, progression through the stages is subject to unknown feasibility conditions, which we term *stopping constraints*. If any of them is violated, the process is immediately aborted: the transition is not realised and the terminal objective is not observed. Such stopping constraints arise across many disciplines: a droplet that fails an intermediate synthesis step and is discarded in autonomous multi-step chemistry (Volk et al., 2023), a device that fails during adaptive sequential reliability qualification before final characterisation (Elhagrasy et al., 2026), and a steel part that cracks during quenching and cannot proceed to final evaluation (Lopez-Garcia et al., 2022). Since a violation costs one run rather than a system fault, such failures are tolerated during learning (Marco et al., 2021; Wakabayashi & Otsuka, 2026; Stenger et al., 2026), and exploring uncertain regions may be necessary to discover high-value feasible trajectories. The goal hence is: **how can one optimise efficiently when violations are allowed during learning but abort the run?**

The presence of stopping constraints induces a trade-off between exploration and feasibility in MSBO. An infeasible decision not only consumes a run but also censors all downstream observations, making aggressive exploration costly. In contrast, conservative strategies reduce the risk of aborting costly runs, but may exclude reaching promising trajectories leading to the optimum. Under a finite budget, the challenge is therefore to explore uncertain trajectories while characterising the number of aborted

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

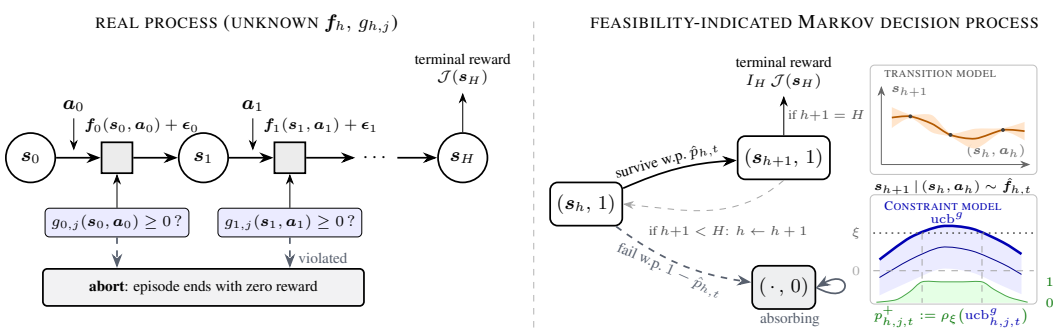


Figure 1: **Left:** multi-stage BO with stopping constraints. At stage h the agent takes an action $\boldsymbol{a}_h$, and the transition $\boldsymbol{s}_{h+1} = \boldsymbol{f}_h(\boldsymbol{s}_h, \boldsymbol{a}_h) + \boldsymbol{\epsilon}_h$ is realised only if every stage constraint is satisfied. A single violation stops all subsequent stages and the episode ends with no reward. **Right:** the equivalent feasibility-indicated MDP: the state is augmented with a survival indicator, and a stage is survived with probability $\hat{p}_{h,t}$, here the optimistic $p^+_{h,t} = \prod_j p^+_{h,j,t}$ computed from the constraint GPs' predictions. FI-O-VI (Section 3.2) solves this MDP by planning optimistically.

runs induced by such exploration. As elaborated in related works at Appendix B, to the best of our knowledge, this combination of unknown stage-wise transitions, unknown stopping constraints, and failure-induced censoring has not been formally studied in MSBO or related areas.

Contributions We formulate MSBO under stopping constraints as a finite-horizon *feasibility-indicated* Markov decision process (FIMDP) by augmenting the state with a binary indicator of whether the process remains active, so that the problem can be approached via dynamic programming (Bellman, 1966). The unknown stage transitions and stopping constraints are modelled by Gaussian process (GP) surrogates (Williams & Rasmussen, 2006), allowing uncertainty about stage-wise feasibility to enter the Bellman recursion explicitly through the augmented transition model. We subsequently propose FI-O-VI (Feasibility-Indicated Optimistic Value Iteration), which plans optimistically with respect to the unknown stopping constraints and stage transitions and uses a feasibility threshold that decreases over the optimisation budget to control how constraint uncertainty affects planning. We show that FI-O-VI yields a sublinear bound on the number of aborted episodes while retaining asymptotic optimality. Our main contributions can be summarised as follows:

- We formalise MSBO under unknown stage-wise stopping constraints (Section 2.1, Figure 1 left) and cast it as a finite-horizon feasibility-indicated MDP (Section 3.1, Figure 1 right), on which the algorithm and the analysis can be carried out.
- We propose FI-O-VI (Section 3.2, Algorithm 1), an optimistic value-iteration method that weights continuation at each stage by comparing optimistic constraint estimates with a feasibility threshold that decreases over the budget (Definition 1). We also study its plug-in variant FI-PoF-VI (Section 3.3), which weights the continuation by the posterior probability of feasibility and performs competitively in practice.
- We characterise the cumulative abort count and cumulative value regret of FI-O-VI, showing that both grow sublinearly under standard RKHS regularity and feasibility assumptions (Theorems 1 and 2). We further show that FI-PoF-VI, despite strong empirical performance, can incur linear regret under certain problem conditions.
- We demonstrate competitive empirical performance on synthetic benchmarks (Figures 2, 6) and on application-driven multi-stage optimisation problems, across cascade depths and different stopping-constraint types.

2 Problem Setting, Regularity and Feasibility Assumptions

2.1 Problem Setting

Illustrated on the left-hand side of Figure 1, we consider an H-stage sequential optimisation problem. At each stage $h \in \{0, 1, \ldots, H-1\}$, after observing the current intermediate outcome $\boldsymbol{s}_h \in \mathcal{S}_h =$

$\mathbb{R}^{d_s}$, a stage-specific decision $\boldsymbol{a}_h \in \mathcal{A}_h$ is selected. An unknown stage-wise transition mapping $\boldsymbol{f}_h : \mathcal{S}_h \times \mathcal{A}_h \to \mathcal{S}_{h+1}$ then produces the next intermediate outcome under additive Gaussian noise $\boldsymbol{\epsilon}_h \sim \mathcal{N}(0, \sigma_h^2 \boldsymbol{I}_{d_s})$:

$$\boldsymbol{s}_{h+1} = \boldsymbol{f}_h(\boldsymbol{s}_h, \boldsymbol{a}_h) + \boldsymbol{\epsilon}_h, \qquad h = 0, \dots, H-1.$$

Each stage may additionally have $m_h \geq 0$ unknown real-valued stopping-constraint functions $g_{h,j} : \mathcal{S}_h \times \mathcal{A}_h \to \mathbb{R}$, which determine whether the process can continue. After a decision is taken at stage h, the corresponding real-valued constraint observations are obtained. If any constraint is violated, i.e., $g_{h,j}(\boldsymbol{s}_h, \boldsymbol{a}_h) < 0$ for some $j \in [m_h]$, the process is aborted before the transition is realised: $\boldsymbol{s}_{h+1}$ is therefore not observed, and no later transition or constraint observation is obtained.[1] Only if all H stages are successfully completed is the terminal outcome $\boldsymbol{s}_H$ observed and a non-negative objective $\mathcal{J} : \mathcal{S}_H \to [0, \overline{r}]$ evaluated.

Let $\pi_h : \mathcal{S}_h \to \mathcal{A}_h$ denote the stage-wise feedback decision rule (a.k.a. the stage policy), such that $\boldsymbol{a}_h = \pi_h(\boldsymbol{s}_h)$, and let $\pi := \{\pi_h\}_{h=0}^{H-1}$ denote the policy of the whole process. For a fixed initial condition $\boldsymbol{s}_0$ [2], the objective is to find a decision rule maximising the expected terminal reward:

$$\pi^\star \in \arg\max_{\pi} \mathbb{E}_{\boldsymbol{\epsilon}_0, \dots, \boldsymbol{\epsilon}_{H-1}} \left[I_H \mathcal{J}(\boldsymbol{s}_H) \mid \boldsymbol{s}_0 \right], \tag{1}$$

where $I_h \in \{0, 1\}$ indicates whether the process remains active upon entering stage h, with $I_0 = 1$ and $I_{h+1} = I_h \mathbb{1}\{g_{h,j}(\boldsymbol{s}_h, \boldsymbol{a}_h) \geq 0, \ \forall j \in [m_h]\}$. We consider a budget of T repeated process executions (episodes), each restarting from the same initial condition $\boldsymbol{s}_0$.

2.2 Regularity and Feasibility Assumptions

Regularity Assumptions Under the frequentist setting (as assumed in Srinivas et al. (2009); Chowdhury & Gopalan (2017)), with GP as the surrogate model of the unknown transition dynamics and constraint functions, we assume their norms are upper bounded in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\kappa$ of a known, normalized positive-definite kernel κ (i.e., $\kappa(\boldsymbol{z}, \boldsymbol{z}) \leq 1, \ \forall \boldsymbol{z} \in \mathcal{S}_h \times \mathcal{A}_h$) : $\|\boldsymbol{f}_h\|_\kappa, \|g_{h,j}\|_\kappa \leq B$ for every output of $\boldsymbol{f}_h$ and every $j \in [m_h]$.

Feasibility Assumptions We make the following assumptions on states reachable through non-aborted trajectories.

Assumption 1 (No dead ends along reachable states)**.** $\exists\, \xi_{\text{reach}} > 0$ *such that* $\forall h \in \{0, \dots, H-1\}$, *if* $I_h = 1$, *then* $\exists\, \boldsymbol{a}_h \in \mathcal{A}_h : \ \min_{j \in [m_h]} g_{h,j}(\boldsymbol{s}_h, \boldsymbol{a}_h) \ \geq \ \xi_{\text{reach}}$.

Assumption 2 (Interior optimum)**.** $\exists\, \xi_{\text{opt}} \in (0, \xi_{\text{reach}}]$ *and an optimal decision rule* $\pi^\star$ *for Eq. 1 such that under* $\pi^\star$, $\forall h \in \{0, \dots, H-1\}$, $\mathbb{P}\big[I_h = 1, \ \min_{j \in [m_h]} g_{h,j}(\boldsymbol{s}_h, \pi_h^\star(\boldsymbol{s}_h)) < \xi_{\text{opt}}\big] = 0$.

Assumption 2 states that the trajectories generated by $\pi^\star$ remain strictly feasible: at every stage reached along these trajectories, all stopping constraints are at least $\xi_{\text{opt}} > 0$. Note that similar interior-optimum conditions are used in BO with failure constraints (Iwazaki et al., 2023), which is specifically necessary for our analysis in the presence of process noise. These assumptions ensure that, for any feasibility threshold $\xi \leq \xi_{\text{opt}}$, the trajectories induced by $\pi^\star$ remain feasible.

3 Methodology

3.1 From the Stopping Constraint to Feasibility-Indicated MDP

To solve this optimisation problem with stopping constraints,[3] we formulate it as a feasibility indicated finite-horizon Markov decision process (FIMDP)[4] whose state carries a binary *feasibility indicator* I_h, with $I_0 = 1$, denoting whether the process remains active. FIMDP's optimal value equals the objective of Eq. 1 under the ground-truth transition functions and stopping constraints. Since f_h and $g_{h,j}$ are unknown in practice, planning is instead performed on the following surrogate FIMDP:

[1] We term this as *pre-transition stopping constraint* and discuss other variants in Section 3.3 .

[2] We follow standard MSBO setting by fixing $\boldsymbol{s}_0$ yet our algorithm can also optimise over $\mathcal{S}_0$.

[3] We follow MSBO scheme and refer to this as an optimisation, while it is more precisely described as an optimal SDM problem given process noise.

[4] A detailed introduction of finite-horizon MDPs is provided in Appendix C.

Surrogate FIMDP We define the surrogate FIMDP as $\hat{\mathcal{M}}_t^{fi} = \langle \mathcal{S}^{fi}, \mathcal{A}, \hat{P}_t^{fi}, R, H \rangle$, illustrated in Figure 1. Its state $\boldsymbol{s}_h^{fi} := (\boldsymbol{s}_h, I_h) \in \mathcal{S}_h^{fi} := \mathcal{S}_h \times \{0, 1\}$ augments the state with the feasibility indicator, where $I_h = 1$ when no stopping constraint has been violated before stage h. Conditional on $I_h = 1$, the augmented transition is a Bernoulli mixture: with the transition-feasibility probability $\hat{p}_{h,t}(\boldsymbol{s}_h, \boldsymbol{a}_h)$, constructed from the posterior of the stopping constraints, the process remains active and the next state follows the posterior predictive distribution induced by the transition model's uncertainty and the process noise; with probability $1 - \hat{p}_{h,t}(\boldsymbol{s}_h, \boldsymbol{a}_h)$ it enters the inactive state:

$$\boldsymbol{s}_{h+1}^{fi} \sim \hat{P}_t^{fi}(\cdot \mid \boldsymbol{s}_h^{fi}, \boldsymbol{a}_h) : \ \boldsymbol{s}_{h+1}^{fi} = \begin{cases} \big(\hat{\boldsymbol{f}}_{h,t}(\boldsymbol{s}_h, \boldsymbol{a}_h) + \boldsymbol{\epsilon}_h, \ 1\big), & I_h = 1, \ \text{w.p. } \hat{p}_{h,t}(\boldsymbol{s}_h, \boldsymbol{a}_h), \\ (\boldsymbol{s}_h, 0), & I_h = 1, \ \text{w.p. } 1 - \hat{p}_{h,t}(\boldsymbol{s}_h, \boldsymbol{a}_h), \\ (\boldsymbol{s}_h, 0), & I_h = 0. \end{cases} \tag{2}$$

where $\hat{\boldsymbol{f}}_{h,t}$ denotes a realisation of the transition GP posterior (Eq. 10). Hence, conditional on the state-action sequence, the probability that the process is active upon entering stage h is $\Pr(I_h = 1 \mid \boldsymbol{s}_{0:h-1}, \boldsymbol{a}_{0:h-1}) = \prod_{k=0}^{h-1} \hat{p}_{k,t}(\boldsymbol{s}_k, \boldsymbol{a}_k)$ assuming independent constraints. The reward is zero at every intermediate stage and is received at the terminal stage only if the process is still active, $R_h \equiv 0$ for $h < H$ and $R_H(\boldsymbol{s}_H^{fi}) = I_H \, \mathcal{J}(\boldsymbol{s}_H)$, so a trajectory that stops before stage H receives zero reward. The specific choice of $\hat{p}_{h,t}$ is described as below.

Choice of Transition Feasibility $\hat{p}_{h,t}$ in Eq. 2 denotes the transition-feasibility probability in the surrogate FIMDP and can be regarded as a design choice. A natural plug-in choice is the Probability-of-Feasibility (PoF), which, assuming independent constraint surrogates, estimates the posterior probability that all constraints are satisfied:

$$p_{h,t}^{\mathrm{PoF}}(\boldsymbol{s}, \boldsymbol{a}) = \prod_{j=1}^{m_h} \Phi\left(\frac{\mu_{h,j,t}^g(\boldsymbol{s}, \boldsymbol{a})}{\sigma_{h,j,t}^g(\boldsymbol{s}, \boldsymbol{a})}\right), \tag{3}$$

where Φ is the standard Gaussian CDF and $\mu_{h,j,t}^g$, $\sigma_{h,j,t}^g$ are the GP posterior mean and standard deviation of constraint $g_{h,j}$ at round t (Eq. 10). On the other hand, the induced per-stage discount due to $p_{h,t}^{\mathrm{PoF}} < 1$ under such a scheme compounds across the stages and hence may suppress promising routes through uncertain regions and degrades long-horizon performance. To establish the optimality guarantee, we instead propose the following optimistic transition feasibility surrogate:

$$p_{h,t}^{+}(\boldsymbol{s}, \boldsymbol{a}) = \prod_{j=1}^{m_h} \rho_{\xi_{h,j,t}}\big(\mathrm{ucb}_{h,j,t}^g(\boldsymbol{s}, \boldsymbol{a})\big), \ \rho_{\xi_{h,j,t}}(u) := \min\Big\{1, \ \max\Big\{0, \ \tfrac{2}{\xi_{h,j,t}}\big(u - \tfrac{\xi_{h,j,t}}{2}\big)\Big\}\Big\}, \tag{4}$$

where $\mathrm{ucb}_{h,j,t}^g = \mu_{h,j,t}^g + \beta_g \sigma_{h,j,t}^g$. The scaling factor β_g and the feasibility threshold $\xi_{h,j,t} > 0$ are algorithm parameters, which are specified in Section 4. The definition of ρ_ξ ensures that whenever the UCB of a constraint exceeds the feasibility threshold then ρ for that threshold is 1; in contrast if any constraint UCB is less than half (which is a deliberate choice that allows the abort number to be upper bounded in Section 4) of the feasibility threshold then $p_{h,t}^+$ is 0. Using this p^+ in the surrogate FIMDP removes the feasibility discount of Eq. 3, weighting all actions equally so long as their UCB exceeds the feasibility threshold, and hence is optimistically feasible. This allows the method to explore closer to the feasibility threshold while also keeping $p_{h,t}^+$ Lipschitz continuous.

3.2 Optimistic Value Iteration

Given that $\hat{p}_{h,t}$ has incorporated the epistemic uncertainty of the constraint model, to further account for the epistemic uncertainty of the transition model, we use an optimistic Bellman backup with a stage-wise exploration bonus (Azar et al., 2017; Chowdhury & Gopalan, 2019). At episode t, let $e_{h,t}(\boldsymbol{s}, \boldsymbol{a}) \geq 0$ denote the exploration bonus associated with uncertainty in the stage transition model (specified in Appendix D.3). Let $V_{h,t}$ and $Q_{h,t}$ denote the state-value function and state-action value function (Q-function) of the surrogate FIMDP, respectively. At an active state, define $\tilde{V}_{h,t}(\boldsymbol{s}) := V_{h,t}(\boldsymbol{s}, 1)$ and $\tilde{Q}_{h,t}(\boldsymbol{s}, \boldsymbol{a}) := Q_{h,t}(\boldsymbol{s}, 1, \boldsymbol{a})$. Assigning the transition bonus to the active branch of Eq. 2, the Bellman recursion is, for $h = H-1, \ldots, 0$,

$$\tilde{Q}_{h,t}(\boldsymbol{s}, \boldsymbol{a}) = \hat{p}_{h,t}(\boldsymbol{s}, \boldsymbol{a})\Big(e_{h,t}(\boldsymbol{s}, \boldsymbol{a}) + \mathbb{E}_{\hat{\boldsymbol{f}}_{h,t}} \mathbb{E}_{\boldsymbol{\epsilon}_h}\big[\tilde{V}_{h+1,t}(\boldsymbol{s}')\big]\Big), \qquad \tilde{V}_{h,t}(\boldsymbol{s}) = \max_{\boldsymbol{a} \in \mathcal{A}_h} \tilde{Q}_{h,t}(\boldsymbol{s}, \boldsymbol{a}), \tag{5}$$

where $\boldsymbol{s}' = \hat{\boldsymbol{f}}_{h,t}(\boldsymbol{s}, \boldsymbol{a}) + \boldsymbol{\epsilon}_h$ is the next state and the terminal condition is $\tilde{V}_{H,t}(\boldsymbol{s}) = \mathcal{J}(\boldsymbol{s})$. Since the inactive state is absorbing and has zero value, marginalising over the feasibility indicator's transition yields the multiplicative factor $\hat{p}_{h,t}(\boldsymbol{s}, \boldsymbol{a})$ in Eq. 5. Hence, $\hat{p}_{h,t}$ acts as a stage-wise weight on both the exploration bonus and the downstream value, which propagate recursively across stages and determines how uncertainty in the stopping constraints affects long-horizon planning.

State-Value Model. To avoid maximising over actions at every evaluation of $\tilde{V}_{h,t}$ in Eq. 5, we follow standard kernelised value-function approximation (Taylor & Parr, 2009; Vakili & Olkhovskaya, 2023) and fit, at each backward step, a kernel ridge regression model $\tilde{V}_{h,t}(\boldsymbol{s}) = \sum_i w_{h,i}\, k_v(\boldsymbol{s}, \boldsymbol{s}_i)$ on a support set of states $\{\boldsymbol{s}_i\}$ that consists of the observed states and a fixed set of state-space coverage points (Appendix D.1.2). With an RBF kernel k_v the expectation in Eq. 5 is available in closed form (Girard et al., 2002), at the last stage, it is instead approximated by quasi-Monte Carlo. The detailed state-value model parameter specification is provided in Appendix H.1.

3.3 Overall Algorithm and Extensions

The overall procedure of FI-O-VI is summarised in Algorithm 1. At each episode t, the GP posteriors of the transitions and the stopping constraints are refitted on all data collected so far, the feasibility thresholds $\xi_{h,j,t}$ are set, and the value functions $\tilde{V}_{h,t}$ are computed by the backward recursion of Eq. 5 with the transition-feasibility probability of Eq. 4. In practice, the scheduled threshold is applied at the stage being executed (Appendix D.2). The episode is then executed in closed loop: at each stage, the action maximising $\tilde{Q}_{h,t}$ at the observed state [5] is executed, the constraint values and the next state are recorded, and a violation aborts the episode, after which no further observation of that episode is obtained. FI-PoF-VI is the same procedure with $\hat{p}_{h,t} = p^{\mathrm{PoF}}_{h,t}$ of Eq. 3 in place of $\hat{p}_{h,t} = p^{+}_{h,t}$ of Eq. 4. The exploration bonus $e_{h,t}$ is specified in Appendix D.3. At deployment time, the exploration bonus is removed and the optimistic feasibility weight is replaced by the plug-in PoF (Appendix H.3).

Post-Transition Stopping Constraints. Another commonly encountered setting (e.g., the setting of Section 5.1) is *post-transition* stopping constraints, where the stage output $\boldsymbol{s}_{h+1}$ is observed before feasibility is assessed, so a violation censors only the subsequent stages and the terminal reward. As elaborated in Appendix E, our algorithms apply to this setting with minimal changes, and the abort-count guarantee carries over when the stage is deterministic.

4 Theoretical Analysis

We study the theoretical properties of FI-O-VI and show that both the cumulative abort count and training value regret grow sublinearly with the number of episodes T.

4.1 Constraint Violation Counts

For the specification of $\xi_{h,j,t}$ in Eq. 4, as we will show soon below, a constant threshold satisfying $0 < \xi_{h,j,t} \le \xi_{\mathrm{opt}}$ suffices for a sublinear abort count and an asymptotically sublinear value regret bound. Under a finite budget T one may prefer a stronger fine-grained control of the abort count in practice, while still maintaining an asymptotic optimality guarantee. To this end, we propose the following annealed schedule for the threshold.

Definition 1 (Annealed feasibility-threshold schedule)**.** *Let $\xi_0 > 0$ and let $c_0 : \mathbb{N} \to \mathbb{R}_{>0}$ and $c_1 : \mathbb{N} \to \mathbb{R}_{\ge 0}$ be non-increasing with $c_0(1) = 1$ and $\lim_{t\to\infty} \big(c_0(t)\xi_0 + c_1(t)\big) < \xi_{\mathrm{opt}}$. For every constraint (h, j) and every episode $t \ge 1$,*

$$\xi_{h,j,t}(\boldsymbol{s}, \boldsymbol{a}) \;:=\; c_0(t)\,\xi_0 \;+\; c_1(t)\,\sigma^{g}_{h,j,t}(\boldsymbol{s}, \boldsymbol{a}). \tag{6}$$

We note that the two terms in Eq. 6 serve distinct purposes. The first plays the role of the constant threshold to control the number of abort, but is no longer required to lie below ξ_{opt} with the annealing

[5] The maximisation over $\mathcal{A}_h$, here and in Eq. 5, is solved numerically by a random screen of candidate actions followed by gradient-based refinement of the best one, see Appendix H.2 for experimental details.

factor $c_0(t)$. The second factor, built on top of Eq. 4 (i.e., $\mu^g_{h,j,t} \geq c_0(t)\xi_0 + (c_1(t) - \beta_g)\sigma^g_{h,j,t}$), introduces a conservative–optimistic mechanism: constraint uncertainty is penalized at the beginning stage of value iteration when $c_1(t) > \beta_g$ and later encouraged when $c_1(t) < \beta_g$.

For a normalized kernel, $\sigma^g_{h,j,t}(\boldsymbol{z}) \leq 1$. Define the schedule *conservation transition time* $t_{\text{reach}} := \min\{t \geq 1 : c_0(t)\xi_0 + c_1(t) \leq \xi_{\text{reach}}\}$ and *optimisation transition time* as $t_{\text{opt}} := \min\{t \geq 1 : c_0(t)\xi_0 + c_1(t) \leq \xi_{\text{opt}}\}$. Both are finite by Definition 1 and has $t_{\text{reach}} \leq t_{\text{opt}}$. Setting $\beta_g = B$ (so that $\text{ucb}^g_{h,j,t} \geq g_{h,j}$ by Lemma 1), we have the following abort count upper bound of FI-O-VI.

Theorem 1 (Abort count of FI-O-VI)**.** *Let $A_T = \sum_{t=1}^{T} \mathbb{1}\{I^t_H = 0\}$ be the number of aborted episodes. Suppose the conditions of Section 2.2 and Assumption 1 hold. Then FI-O-VI with the feasibility-threshold schedule of Definition 1 satisfies, for every realized trajectory and every $T \geq 1$,*

$$A_T \leq \min\left\{T,\ t_{\text{reach}} - 1 + \frac{32MB^2\gamma_T}{c_0(T)^2\xi_0^2\log(1+\lambda^{-1})}\right\}, \tag{7}$$

where $M = \sum_{h=0}^{H-1} m_h$, λ is the constraint models' regulariser of Eq. 10, and γ_T is the maximal information gain from T observations. In particular, $A_T = o(T)$ whenever $\gamma_T/c_0(T)^2 = o(T)$.

The proof sketch is as follows. After the schedule $c_0(t)\xi_0 + c_1(t)$ falls below ξ_{reach}, any abort must occur at an executed action with non-negligible constraint posterior uncertainty: for some (h, j), $\sigma^g_{h,j,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h) > c_0(t)\xi_0/(4B)$. Thus every abort at $t \geq t_{\text{reach}}$ consumes part of the corresponding information-gain budget, while each earlier episode can abort at most once and together they contribute the additive $t_{\text{reach}} - 1$. Note that the sublinearity condition $\gamma_T/c_0(T)^2 = o(T)$ can be satisfied for standard kernels defined in bounded space, for instance, a polynomial decay $c_0(t) = t^{-\alpha}$ with a kernel-dependent rate α (Appendix D.2). For unbounded space, the analysis of γ_T is much more involved (Flynn et al., 2026). However, in our case, we can always fix a bounded region and consider projections of the noisy output, our proof remains the same and the standard analysis of γ_T applies. Theorem 1 also motivates the annealing mechanism, as for fixed ξ_0 to preserve the optimality (requiring $\xi_0 \leq \xi_{\text{opt}}$), the violation numbers can be considerate especially when problem scale increase (M or γ_T increase), annealing allows a more conservative exploration strategy at the beginning while remaining optimistic to preserve the optimality, as we will show in the next section.

4.2 Value Regret Rate

Let $v^\star = \mathbb{E}\left[I_H \mathcal{J}(\boldsymbol{s}_H) \mid \boldsymbol{s}_0, \pi^\star\right]$ be the optimal value where the expectation taken over the process noise. In episode t, let $r^t = I^t_H \mathcal{J}(\boldsymbol{s}^t_H)$ and define the realized cumulative regret as $R_T = \sum_{t=1}^{T}(v^\star - r^t)$. For the GP transition model, set $\underline{\sigma} = \min_h \sigma_h$ and $\overline{\sigma} = \max_h \sigma_h$. For confidence level $\delta \in (0, 1)$, set $\beta^f_t = B + \frac{\overline{\sigma}}{\sqrt{\lambda}}\sqrt{2(\gamma_{t-1} + \log(2D_f/\delta))}$ with $D_f = Hd_s$. Writing $\boldsymbol{\sigma}^f_{h,t}$ for the vector of transition mapping posterior's standard deviations, define the theoretical instantiation of the exploration bonus as:

$$\psi(x) = 2\Phi(x/2) - 1, \quad e^{\text{TV}}_{h,t}(\boldsymbol{z}) = \overline{r}\,\psi\left(\frac{\beta^f_t + 1}{\sigma_h}\left\|\boldsymbol{\sigma}^f_{h,t}(\boldsymbol{z})\right\|_2\right). \tag{8}$$

The following theorem gives a high-probability asymptotic regret bound for FI-O-VI:

Theorem 2 (No-Regret of FI-O-VI)**.** *Suppose the conditions of Theorem 1 and Assumption 2 hold. For every $T \geq 1$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$R_T = O\left(\frac{\gamma_T}{c_0(T)^2} + (\beta^f_T + 1)\sqrt{T\gamma_T}\right). \tag{9}$$

Consequently, for fixed H and state-action dimensions, $R_T = o(T)$ whenever $\gamma_T/c_0(T)^2 = o(T)$ and $(\beta^f_T + 1)^2\gamma_T = o(T)$.

The first rate in Eq. 9 captures the cost of aborted episodes, while the second arises from uncertainty in the transition model. Random fluctuations in the realized return contribute only a lower-order $\sqrt{T}$ term and are therefore omitted from the displayed rate. A finite-T bound, including this lower-order term and all constants, is given in Eq. 63.

Comparison with FI-PoF-VI. In FI-PoF-VI, from the construction Eq. 3, one notices that even when $g_{h,j}(\boldsymbol{s}, \pi_h^\star(\boldsymbol{s})) \geq \xi_{\text{opt}}$ for every $j \in [m_h]$, the uncertainty term $\sigma^g_{h,j,t}(\boldsymbol{s}, \pi_h^\star(\boldsymbol{s}))$ can make $p^{\text{PoF}}_{h,t} < 1$, which may prevent exploration of an optimal action. By contrast, in FI-O-VI, as long as the optimistic condition holds $\text{ucb}^g_{h,j,t}(\boldsymbol{s}_h, \boldsymbol{a}_h) \geq \xi_{h,j,t}(\boldsymbol{s}_h, \boldsymbol{a}_h)$, we will have $p^+_{h,t}(\boldsymbol{s}_h, \pi_h^\star(\boldsymbol{s}_h)) = 1$, without requiring $\sigma^g_{h,j,t} = 0$. In Section F.3, we construct a concrete theoretical example where FI-PoF-VI has $A_T = 0$ and $R_T = \Omega(T)$, while FI-O-VI still satisfies $R_T = o(T)$ with high probability. This phenomenon is also observed empirically in our experiments (Figure 9).

5 Experiments

We empirically validate our approach in this section. For comparison baselines, since none of the existing MSBO handles constraints, we extend existing unconstrained approaches to our setting by applying the standard feasibility-weighting strategies from the constrained BO literature. Specifically, we compare our algorithms with four baselines. 1) FB-CUCB: a fully black box approach which discards the cascade and models the terminal objective with a single GP over the concatenated actions, using UCB (Srinivas et al., 2009) weighted by the probability of feasibility (Gardner et al., 2014), in an open loop manner. 2) BOFN-CEI: an extension of the Monte Carlo sampling in Astudillo & Frazier (2021), which propagates reparameterized rollouts through the per-stage GPs and optimises all H actions jointly, weighted by the sampled feasibility of each rollout (Letham et al., 2019; Balandat et al., 2020). 3) Cascade-CEI: Cascade BO (Kusakawa et al., 2022), which iteratively replans across stages, optimising the same feasibility-weighted Monte Carlo objective as 2) but restricted to the remaining stages given the realized state. 4) Random Feasible (RF): which at each stage h re-draws a random sequence of remaining actions that is predicted feasible (i.e., from $\{\boldsymbol{a}_{h:H-1} : \prod_{k \geq h} p^{\text{PoF}}_{k,t}(\boldsymbol{s}_k, \boldsymbol{a}_k) \geq 1/2\}$) and executes its first action. Appendix H.2 provides detailed descriptions of each baseline and their corresponding parameter specification used in experiments. [6] All cascade-aware methods use the same transition and constraint GP surrogates (Appendix H.1).

Experimental Setup. The process noise is $\sigma_h = 0.02$ at every stage of every synthetic instance. The initial design consists of $n_{\text{init}} = 2(d_s + d_a)$ completed episodes. Each algorithm is run on the same 20 seeds, each of which determines the initial design and the process-noise draws shared by all methods. Following the reinforcement-learning convention, performance is measured by the evaluation reward of the deployable policy of each method (Section H.3), estimated every 5 episodes from 200 rollouts (for a given seed, every method sees the same process-noise realisations).

5.1 Synthetic Experiments

We first evaluate the methods on synthetic multi-stage problems that vary the decision horizon, the stage transitions, the constraints and the state and action dimensions. Since this problem setting has no established benchmarks, we construct eight instances from well-established single-stage benchmark functions and GP sample paths (Section G.1).

Figure 2 shows the median evaluation reward and interquartile range, with dashed lines indicating approximate optimal values computed by dynamic programming on a discretised state grid. Figure 7 shows the final evaluation rewards across seeds. In general, FI-O-VI and FI-PoF-VI generally outperform the forward planning baselines. On several problems, FI-PoF-VI improves faster in the early episodes, which is expected because the initial feasibility thresholds in the ξ_t schedule make FI-O-VI deliberately more conservative during the early stage of optimisation. This behaviour is consistent with the role of optimistic feasibility in the regret analysis of FI-O-VI, whereas the same guarantee is not established for the plug-in FI-PoF-VI variant.

Figure 8 illustrates the state-action pairs selected by FI-O-VI on the scalar-state problems. As optimisation proceeds, the samples concentrate in feasible high-value regions, with most aborts occurring near the constraint boundary. The decreasing frequency of aborts is also visible in Figure 6 frequently across problems. Specifically, Figure 9 compares the two methods at the final stage of GMM, where the optimum lies near the constraint boundary. In the illustration, FI-O-VI continues

[6] Our code is available at https://anonymous.4open.science/r/_CMSBO_Review-1C52/.

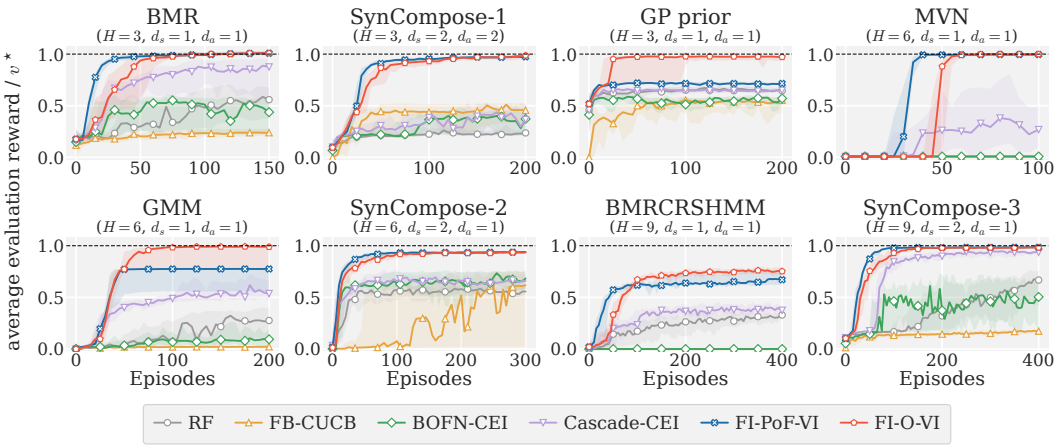


Figure 2: Evaluation reward of the deployed policy on the eight synthetic benchmark problems, as a fraction of the optimal value $v^\star$ of the instance (dashed line). Results are reported using median and the interquartile range.

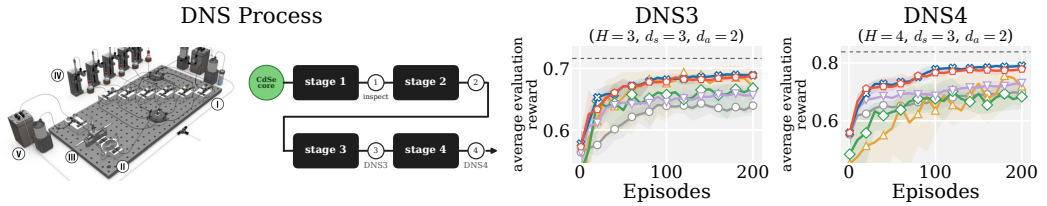


Figure 3: Droplet nanocrystal synthesis. Left: the AlphaFlow droplet reactor, adapted from Volk et al. (2023)*. DNS3 contains three deposition cycles and DNS4 contains four. Right: average evaluation reward of the deployed policy, shown as the median over 20 seeds with the interquartile band. The dashed line marks the approximate optimum†. Colours and markers are as in Figure 2.

to consider this region plausible under its optimistic feasibility treatment and eventually reaches the optimum after several boundary evaluations. In contrast, the plug-in feasibility probability used by FI-PoF-VI discourage actions near the uncertain boundary and the method stops querying closer to the optimum. As a result, FI-PoF-VI converges to a lower-value solution on most seeds (as also seen in Figure 7). The measured runtimes of all methods are reported in Appendix I.

Sensitivity and Ablation Study We further investigate the sensitivity of FI-O-VI to the initial feasibility threshold ξ_0 and the transition episode t_{tr}. To assess the effect of annealing, we also conduct an ablation that replaces the schedule with a fixed feasibility threshold (Section H.4.5). Smaller ξ_0 or earlier t_{tr} generally lead to more aborts. Delaying the transition reduces aborts but can also delay reward improvement (as seen on MVN). Final rewards are similar across most settings, although the smallest ξ_0 tested gives a lower reward on several problems. Removing annealing increases early aborts and slows reward improvement on several problems. These results support an initially conservative schedule, which limits early aborts while allowing more exploration later.

Droplet Nanocrystal Synthesis (DNS). Nanocrystals can be grown by repeatedly injecting small amounts of chemicals into a liquid droplet. Each injection cycle grows a shell around the particles,

*Source image licensed under CC BY 4.0.

†For each problem, the approximate optimum is obtained by evaluating 5,000 randomly sampled recipes and then applying local optimisation to the best one on the noise-free cascade.

‡For each problem, the approximate optimum is the best recipe found by a random search followed by local optimisation.

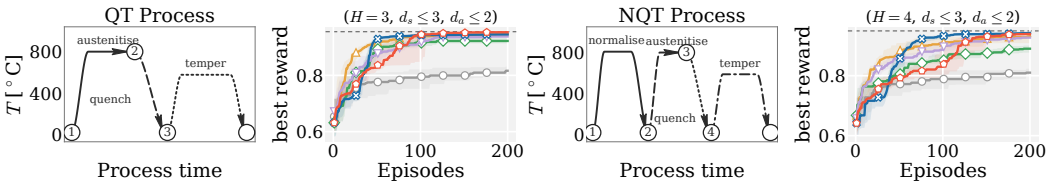


Figure 4: Two heat-treatment cascades and the best reward found on each. Process panels: the thermal cycle of the best-known recipe with one line style per operation. Operation k starts at circle k and its arrowhead enters the next circle, where the part is inspected and scrapped if it fails the check. Reward panels: reward of the best recipe executed so far, median over 20 seeds with the interquartile band. The dashed line marks the approximate optimum[‡]. Colours and markers are as in Figure 2.

and the resulting absorption spectrum provides a measure of their growth. We consider two problems based on the process studied by Volk et al. (2023): DNS3 has three cycles and DNS4 has four (Figure 3). Each stage repeats the same five-injection cycle, with the volumes of two precursor chemicals chosen separately at each stage. The state consists of three features of the latest absorption spectrum: the wavelength of the first absorption peak, its peak-to-valley ratio and its intensity. The goal is to maximise the red shift of the final peak, which indicates shell growth, while keeping the droplet viable. Before each cycle, the proposed recipe must pass a viability check based mainly on predicted peak-intensity retention and peak detectability. If it fails, synthesis stops and the next spectrum and all later observations are unavailable (the pre-transition setting). Further details of the experimental setting are provided in Appendix G.2.1. Figure 3 shows that `FI-PoF-VI` and `FI-O-VI` achieve the highest median evaluation rewards on both problems.

Recipe Search for the Heat Treatment of Steel. Heat treatment adjusts a steel part's hardness through a sequence of furnace and quench operations. We study two deterministic recipe-search problems based on standard routes for a representative chromium–molybdenum low-alloy steel, using published metallurgical models. QT comprises austenitising (Lee & Lee, 2008; Andrews, 1965), quenching (Collins et al., 2023; Maynier et al., 1977), and tempering (Kang & Lee, 2014), with $H{=}3$ and five recipe variables. NQT adds normalising (a preliminary heating-and-air-cooling operation), giving $H{=}4$ and seven recipe variables. Each furnace operation is controlled by its temperature and holding time, and quenching by its cooling rate. The state records grain size, phase fractions and hardness. After each operation, an inspection checks grain size, quench hardness and cracking, or final hardness, and a failed inspection stops all further processing (the post-transition constraint setting). The terminal reward combines closeness to the target hardness with a toughness index (Appendix G.2.2). Figure 4 shows that both VI methods achieve competitive rewards, and Table 5 the trade-off between reward and the number of scrapped parts.

6 Conclusion

We studied multi-stage Bayesian optimisation with unknown stopping constraints, where a violation terminates the process and censors all downstream observations. We proposed `FI-O-VI`, which combines optimistic value iteration with an annealed feasibility-threshold schedule, and its plug-in variant `FI-PoF-VI`. Under RKHS regularity and feasibility assumptions, we proved sublinear bounds on the cumulative abort count and cumulative value regret of `FI-O-VI`. Across synthetic benchmarks and application-driven multi-stage problems, our methods performed competitively against existing baselines over different cascade depths and stopping-constraint types.

Limitations and Future Work. Our analysis establishes regret guarantees for pre-transition constraints with nonzero Gaussian transition noise, while the experiments also examine deterministic post-transition problems. Extending the regret analysis to these settings and accounting for state-value approximation error remain open directions. For larger problems, future work will investigate scalable transition models (e.g., Bayesian Neural Networks) and state-value models, together with the balance between exploration and aborts under a finite learning budget.

AI USE STATEMENT

Generative AI tools were used in the preparation of this submission, under the direction of the authors, for: (1) language editing of text written by the authors; (2) drafting and checking individual proof steps, with every proof subsequently verified line by line by the authors; (3) searching for related work, which the authors read and selected; (4) assisting in the implementation of the code, the synthetic benchmarks and the two real-life problems, whose design decisions were made by the authors, and in producing the figures from the experimental results. The research question, the method and the theoretical claims originate with the authors; generative AI tools served as a sounding board in discussing design choices but were not used to generate the core ideas or the citations. All AI-assisted material was reviewed and verified by the authors, who take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

The code to reproduce this work, including the configurations of every experiment, the scripts that generate every figure and table, and a Jupyter notebook demonstrating the proposed method, is available at https://anonymous.4open.science/r/_CMSBO_Review-1C52/.

REFERENCES

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

KW Andrews. Empirical formulae for the calculation of some transformation temperatures. *J. Iron Steel Inst.*, pp. 721–727, 1965.

Candelieri Antonio. Sequential model based optimization of partially defined functions under unknown constraints. *Journal of Global Optimization*, 79(2):281–303, 2021.

Raul Astudillo and Peter Frazier. Bayesian optimization of composite functions. In *International Conference on Machine Learning*, pp. 354–363. PMLR, 2019.

Raul Astudillo and Peter Frazier. Bayesian optimization of function networks. *Advances in neural information processing systems*, 34:14463–14475, 2021.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pp. 263–272. PMLR, 2017.

François Bachoc, Céline Helbert, and Victor Picheny. Gaussian process optimization with failures: classification and convergence proof. *Journal of Global Optimization*, 78(3):483–506, 2020.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538, 2020.

SS Barsov and Vladimir V Ulyanov. Estimates of the proximity of gaussian measures. In *Doklady Mathematics*, volume 34, pp. 462–466, 1987.

Richard Bellman. Dynamic programming. *science*, 153(3731):34–37, 1966.

Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine learning*, 112(10):3713–3747, 2023.

Maximilian Bloor, Tom Savage, Calvin Tsay, Antonio del Rio Chanona, and Max Mowbray. Gaussian process q-learning for finite-horizon markov decision processes. In *Reinforcement Learning Conference*, 2025.

Poompol Buathong, Jiayue Wan, Raul Astudillo, Samuel Daulton, Maximilian Balandat, and Peter I Frazier. Bayesian optimization of function networks with partial evaluations. *arXiv preprint arXiv:2311.02146*, 2023.

Jack M Buckingham, Ivo Couckuyt, and Juergen Branke. Bayesian optimization for non-convex two-stage stochastic optimization problems. *arXiv preprint arXiv:2408.17387*, 2024.

Elliot Chane-Sane, Pierre-Alexandre Leziart, Thomas Flayols, Olivier Stasse, Philippe Souères, and Nicolas Mansard. Cat: Constraints as terminations for legged locomotion reinforcement learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13303–13310. IEEE, 2024.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in neural information processing systems*, 31, 2018.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017.

Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3197–3205. PMLR, 2019.

Joshua Collins, Martina Piemonte, Mark Taylor, Jonathan Fellowes, and Ed Pickering. A rapid, open-source cct predictor for low-alloy steels, and its application to compositionally heterogeneous material. *Metals*, 13(7):1168, 2023.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14156–14170. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a36b598abb934e4528412e5a2127b931-Paper.pdf.

Tianwei Dai, Sriram Vijayakrishnan, Filip T Szczypiński, Jean-François Ayme, Ehsan Simaei, Thomas Fellowes, Rob Clowes, Lyubomir Kotopanov, Caitlin E Shields, Zhengxue Zhou, et al. Autonomous mobile robots for exploratory synthetic chemistry. *Nature*, 635(8040):890–897, 2024.

Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pp. 465–472, 2011.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Youssef A Elhagrasy, Ian Hill, et al. Adaptive sequential test planning for multi-mechanism reliability qualification via bayesian monte carlo tree search. *IEEE Transactions on Reliability*, 2026.

Christian Fiedler, Johanna Menn, Lukas Kreisköther, and Sebastian Trimpe. On safety in safe bayesian optimization. *arXiv preprint arXiv:2403.12948*, 2024.

Hamish Flynn, Joe Watson, Ingmar Posner, and Jan Peters. Posterior sampling reinforcement learning with gaussian processes for continuous control: Sublinear regret bounds for unbounded state spaces. *arXiv preprint arXiv:2603.08287*, 2026.

Jose P Folch, Calvin Tsay, Robert M Lee, Behrang Shafei, Weronika Ormaniec, Andreas Krause, Mark van der Wilk, Ruth Misener, and Mojmír Mutnỳ. Transition constrained bayesian optimization via markov decision processes. *Advances in Neural Information Processing Systems*, 37:88194–88235, 2024.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kilian Q Weinberger, and John P Cunningham. Bayesian optimization with inequality constraints. In *Icml*, volume 2014, pp. 937–945, 2014.

Michael A Gelbart, Jasper Snoek, and Ryan P Adams. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*, 2014.

Agathe Girard, Carl Rasmussen, Joaquin Q Candela, and Roderick Murray-Smith. Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. *Advances in neural information processing systems*, 15, 2002.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11216–11235, 2024.

Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. Bayesian optimization with censored response data. *arXiv preprint arXiv:1310.1947*, 2013.

Carl Hvarfner, Erik Orm Hellsten, and Luigi Nardi. Vanilla bayesian optimization performs great in high dimensions. *arXiv preprint arXiv:2402.02229*, 2024.

Shogo Iwazaki, Shion Takeno, Tomohiko Tanabe, and Mitsuru Irie. Failure-aware gaussian process optimization with regret bounds. *Advances in Neural Information Processing Systems*, 36:24388–24400, 2023.

Shogo Iwazaki, Tomohiko Tanabe, Mitsuru Irie, Shion Takeno, Kota Matsui, and Yu Inatsu. No-regret bayesian optimization with stochastic observation failures. In *AISTATS*, pp. 415–423, 2025.

Jing Jingzhe, Fan Zheyi, Szu Hui Ng, and Qingpei Hu. Local constrained bayesian optimization. *arXiv preprint arXiv:2603.07965*, 2026.

Singon Kang and Seok-Jae Lee. Prediction of tempered martensite hardness incorporating the composition-dependent tempering parameter in low alloy steels. *Materials Transactions*, 55(7): 1069–1072, 2014.

Andrea Karlova, Rishabh Kabra, Daniel Augusto de Souza, and Brooks Paige. Cobalt: Censored optimization and bayesian active learning techniques. In *Forty-Second Annual Conference on Uncertainty in Artificial Intelligence*, 2026.

Saksham Kiroirwal, Julius Pfrommer, and Jürgen Beyerer. Joint parameter and state-space bayesian optimization: Using process expertise to accelerate manufacturing optimization. *Procedia CIRP*, 146:565–570, 2026.

Akshay Kudva, Wei-Ting Tang, and Joel A Paulson. Multi-objective bayesian optimization for networked black-box systems: A path to greener profits and smarter designs. *Optimization of Sustainable Process Systems: Multiscale Models and Uncertainties*, pp. 63–89, 2026.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

Shunya Kusakawa, Shion Takeno, Yu Inatsu, Kentaro Kutsukake, Shogo Iwazaki, Takashi Nakano, Toru Ujihara, Masayuki Karasuyama, and Ichiro Takeuchi. Bayesian optimization for cascade-type multistage processes. *Neural Computation*, 34(12):2408–2431, 2022.

Becky Langdon, Gabriel D Patrón, Chrysoula D Kappatou, Robert M Lee, Behrang Shafei, Jixiang Qing, Ruth Misener, Mark van der Wilk, and Calvin Tsay. Meta-learning for sample-efficient Bayesian optimisation of fed-batch processes. *arXiv preprint arXiv:2605.05382*, 2026.

Seok-Jae Lee and Young-Kook Lee. Prediction of austenite grain growth during austenitization of low alloy steels. *Materials & Design*, 29(9):1840–1844, 2008.

Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. Constrained bayesian optimization with noisy experiments. 2019.

Zihan Li and Jonathan Scarlett. Regret bounds for noise-free cascaded kernelized bandits. *arXiv preprint arXiv:2211.05430*, 2022.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

Ricardo Daniel Lopez-Garcia, Israel Medina-Juárez, and Araceli Maldonado-Reyes. Effect of quenching parameters on distortion phenomena in aisi 4340 steel. *Metals*, 12(5):759, 2022.

Congwen Lu and Joel A Paulson. No-regret bayesian optimization with unknown equality and inequality constraints using exact penalty functions. *IFAC-PapersOnLine*, 55(7):895–902, 2022.

Congwen Lu and Joel A Paulson. No-regret constrained bayesian optimization of noisy and expensive hybrid models using differentiable quantile function approximations. *Journal of Process Control*, 131:103085, 2023.

Benjamin P MacLeod, Fraser GL Parlane, Thomas D Morrissey, Florian Häse, Loïc M Roch, Kevan E Dettelbach, Raphaell Moreira, Lars PE Yunker, Michael B Rooney, Joseph R Deeth, et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6(20): eaaz8867, 2020.

Alonso Marco, Dominik Baumann, Majid Khadiv, Philipp Hennig, Ludovic Righetti, and Sebastian Trimpe. Robot learning with crash constraints. *IEEE Robotics and Automation Letters*, 6(2): 1439–1446, 2021.

James P Materkowski and George Krauss. Tempered martensite embrittlement in sae 4340 steel. *Metallurgical Transactions A*, 10(11):1643–1651, 1979.

Ph Maynier, B Jungmann, and J Dollet. Creusot–loire system for the prediction of the mechanical properties of low alloy steel products. *Hardenability concepts with applications to steel*, pp. 518–545, 1977.

Abhijit Mazumdar, Rafal Wisniewski, and Manuela L Bujorianu. Safe reinforcement learning for constrained markov decision processes with stochastic stopping time. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pp. 1923–1928. IEEE, 2024.

Nikola Milosevic, Leonard Franz, Daniel Haeufle, Georg Martius, Nico Scherf, and Pavel Kolev. Stochastic decision horizons for constrained reinforcement learning. *arXiv preprint arXiv:2602.04599*, 2026.

Henry B Moss, Sebastian W Ober, and Victor Picheny. Inducing point allocation for sparse gaussian processes in high-throughput bayesian optimisation. In *International Conference on Artificial Intelligence and Statistics*, pp. 5213–5230. PMLR, 2023.

Thanh Dai Nguyen, Sunil Gupta, Santu Rana, Vu Nguyen, Svetha Venkatesh, Kyle J Deane, and Paul G Sanders. Cascade bayesian optimization. In *Australasian Joint Conference on Artificial Intelligence*, pp. 268–280. Springer, 2016.

Joel A Paulson and Congwen Lu. Cobalt: Constrained bayesian optimization of computationally expensive grey-box models exploiting derivative information. *Computers & Chemical Engineering*, 160:107700, 2022.

Victor Picheny, Joel Berkeley, Henry B Moss, Hrvoje Stojic, Uri Granta, Sebastian W Ober, Artem Artemev, Khurram Ghani, Alexander Goodall, Andrei Paleyes, et al. Trieste: Efficiently exploring the depths of black-box functions with tensorflow. *arXiv preprint arXiv:2302.08436*, 2023.

Jixiang Qing, Henry B Moss, Tom Dhaene, and Ivo Couckuyt. PF^{2}ES: Parallel feasible pareto frontier entropy search for multi-objective bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2565–2588. PMLR, 2023.

Amirhossein Roknilamouki, Arnob Ghosh, Ming Shi, Fatemeh Nourzad, Eylem Ekici, and Ness B Shroff. Provably efficient rl for linear mdps under instantaneous safety constraints in non-convex feature spaces. *arXiv preprint arXiv:2502.18655*, 2025.

Ming Shi, Yingbin Liang, and Ness Shroff. A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints. In *International Conference on Machine Learning*, pp. 31243–31268. PMLR, 2023.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

David Stenger, Paul Brunzema, Johanna Menn, Alexander von Rohr, Angela P Schoellig, and Sebastian Trimpe. A decade of bayesian optimization for controller tuning and robot learning: Tutorial, review, and future prospects. *arXiv preprint arXiv:2609.09403*, 2026.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International conference on machine learning*, pp. 997–1005. PMLR, 2015.

Yanan Sui, Vincent Zhuang, Joel Burdick, and Yisong Yue. Stagewise safe bayesian optimization with gaussian processes. In *International conference on machine learning*, pp. 4781–4789. PMLR, 2018.

Hao Sun, Ziping Xu, Meng Fang, Zhenghao Peng, Jiadong Guo, Bo Dai, and Bolei Zhou. Safe exploration by solving early terminated mdp. *arXiv preprint arXiv:2107.04200*, 2021.

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

Gavin Taylor and Ronald Parr. Kernelized value function approximation for reinforcement learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 1017–1024, 2009.

Guy Tennenholtz, Nadav Merlis, Lior Shani, Shie Mannor, Uri Shalit, Gal Chechik, Assaf Hallak, and Gal Dalal. Reinforcement learning with a terminator. *Advances in Neural Information Processing Systems*, 35:35696–35709, 2022.

Luca Torresi and Pascal Friederich. Multi-stage bayesian optimisation for dynamic decision-making in self-driving labs. *Digital Discovery*, 5(4):1900–1912, 2026.

Sattar Vakili and Julia Olkhovskaya. Kernelized reinforcement learning with order optimal regret bounds. *Advances in Neural Information Processing Systems*, 36:4225–4247, 2023.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 82–90. PMLR, 2021.

Amanda A Volk, Robert W Epps, Daniel T Yonemoto, Benjamin S Masters, Felix N Castellano, Kristofer G Reyes, and Milad Abolhasani. Alphaflow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nature Communications*, 14(1):1403, 2023.

Alexander Von Rohr, David Stenger, Dominik Scheurenberg, and Sebastian Trimpe. Local bayesian optimization for controller tuning with crash constraints. *at-Automatisierungstechnik*, 72(4): 281–292, 2024.

Yuki K Wakabayashi and Takuma Otsuka. Bayesian optimization for self-driving materials laboratories: From algorithms to physics-informed workflows. *arXiv preprint arXiv:2608.26016*, 2026.

Shengbo Wang and Ke Li. Constrained bayesian optimization under partial observations: Balanced improvements and provable convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15607–15615, 2024.

Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3274–3307. PMLR, 2022.

Honghao Wei, Xin Liu, and Lei Ying. Safe reinforcement learning with instantaneous constraints: The role of aggressive exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21708–21716, 2024.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Nuoya Xiong, Yihan Du, and Longbo Huang. Provably safe reinforcement learning with step-wise violation constraints. *Advances in Neural Information Processing Systems*, 36:54341–54353, 2023.

Wenjie Xu, Yuning Jiang, Bratislav Svetozarevic, and Colin Jones. Constrained efficient global optimization of expensive black-box functions. In *International Conference on Machine Learning*, pp. 38485–38498. PMLR, 2023.

Dasol Yoon, Poompol Buathong, Chia-Hao Lee, Yujia Zhang, David A Muller, and Peter I Frazier. Scalable bayesian optimization of composite functions for image-based inverse problems in materials characterization. *arXiv preprint arXiv:2609.02126*, 2026.

Xingyu Zhou and Bo Ji. On kernelized multi-armed bandits with constraints. *Advances in neural information processing systems*, 35:14–26, 2022.

APPENDIX CONTENTS

A Notation **17**

B Related Work **18**

C Preliminaries **19**

D Algorithm Details **20**

D.1 KRR based Value Model 20

D.2 Algorithmic Specification of ξ_t 22

D.3 Algorithmic Specification of $e_{h,t}$, β_f, β_g 22

E Post-Transition Stopping Constraints **23**

F Proofs **24**

F.1 Proof of the Abort Count Bound 24

F.2 Proof of the Value Regret Bound 27

F.3 Comparison with FI-PoF-VI 33

G Benchmark Problems **35**

G.1 Synthetic Problems 35

G.2 Real-Life Problems 37

H Experimental Details **40**

H.1 Model Selection 40

H.2 Comparison Baselines and Parameter Settings 40

H.3 Performance Evaluation 42

H.4 Additional Experimental Results 42

I Runtime Analysis **44**

A NOTATION

Table 1 lists the notation of the main text, grouped by where it is introduced.

Table 1: Notation.

Symbol	Meaning or convention	Section
Process		
H	number of stages; stages are indexed $h \in \{0, \dots, H-1\}$	Section 2.1
$\boldsymbol{s}_h \in \mathcal{S}_h \subseteq \mathbb{R}^{d_s}$	state entering stage h; $\boldsymbol{s}_0$ the fixed initial state, $\boldsymbol{s}_H$ the terminal state	Section 2.1
$\boldsymbol{a}_h \in \mathcal{A}_h \subseteq \mathbb{R}^{d_a}$	action (decision) taken at stage h	Section 2.1
$\boldsymbol{f}_h$, $\boldsymbol{\epsilon}_h$	unknown stage transition, $\boldsymbol{s}_{h+1} = \boldsymbol{f}_h(\boldsymbol{s}_h, \boldsymbol{a}_h) + \boldsymbol{\epsilon}_h$, with process noise $\boldsymbol{\epsilon}_h \sim \mathcal{N}(0, \sigma_h^2 \boldsymbol{I}_{d_s})$	Section 2.1
$g_{h,j}$, m_h	unknown stopping constraints of stage h, $j \in [m_h]$; the stage aborts if $g_{h,j}(\boldsymbol{s}_h, \boldsymbol{a}_h) < 0$ for some j	Section 2.1
$I_h \in \{0, 1\}$	survival indicator: $I_h = 1$ if no constraint was violated at stages $0, \dots, h-1$	Section 2.1
$\mathcal{J} : \mathcal{S}_H \to [0, \bar{r}]$	terminal objective, received only if $I_H = 1$	Section 2.1
π_h, π, $\pi^\star$	stage decision rule $\boldsymbol{a}_h = \pi_h(\boldsymbol{s}_h)$, the policy $\pi = \{\pi_h\}$, and an optimal policy	Section 2.1
$v^\star$	optimal value $\mathbb{E}[I_H \mathcal{J}(\boldsymbol{s}_H) \mid \boldsymbol{s}_0, \pi^\star]$	Section 2.1
T, t	episode budget and episode index; n_{init} initial-design episodes	Section 2.1
d_{sa}	$d_s + d_a$, the input width of a stage model	Section 2.1
ξ_{reach}, ξ_{opt}	feasibility margins: every reachable alive state admits an action with margin ξ_{reach}, and $\pi^\star$ keeps margin ξ_{opt}	Assumptions 1, 2
Surrogate models		
κ, $\mathcal{H}_\kappa$, B	kernel, its RKHS, and the bound $\|\boldsymbol{f}_h\|_\kappa, \|g_{h,j}\|_\kappa \le B$	Section 2.2
$\boldsymbol{z} = (\boldsymbol{s}_h, \boldsymbol{a}_h)$	generic input of a stage model	Section 2.2
λ	likelihood variance (regulariser) of a GP model, Eq. 10; σ_h^2 for transitions, λ_g for constraints	Section 2.2
$\mu^f_{h,\ell,t}$, $\sigma^f_{h,\ell,t}$	posterior mean and standard deviation of output ℓ of the stage-h transition model at episode t	Section 2.2
$\mu^g_{h,j,t}$, $\sigma^g_{h,j,t}$	posterior mean and standard deviation of constraint model (h, j) at episode t	Section 2.2
$\text{ucb}^g_{h,j,t}$, $\text{lcb}^g_{h,j,t}$	$\mu^g_{h,j,t} \pm \beta_g \sigma^g_{h,j,t}$	Section 2.2
β_g, β_f	confidence scales of the constraint and transition models	Appendix D.3
$\hat{\boldsymbol{f}}_{h,t}$, $\widehat{P}_{h,t}$	a realisation of the transition posterior, and the posterior predictive kernel of $\boldsymbol{s}_{h+1}$ given $(\boldsymbol{s}_h, \boldsymbol{a}_h)$; $P_h^\star$ the true kernel	Sections 3.1, 3.2
γ_T	maximal information gain of T observations	Section 2.2
Φ	standard normal cumulative distribution function	Section 2.2
Feasibility-indicated MDP and planning		
$\boldsymbol{s}^{fi}_h = (\boldsymbol{s}_h, I_h)$	augmented state; $\mathcal{S}^{fi}_h = \mathcal{S}_h \times \{0, 1\}$	Section 3.1
$\hat{\mathcal{M}}^{fi}_t$, $\hat{P}^{fi}_t$	surrogate FIMDP at episode t and its transition kernel, Eq. 2	Section 3.1
R_h	stage reward of the FIMDP: $R_h \equiv 0$ for $h < H$, $R_H = I_H \mathcal{J}(\boldsymbol{s}_H)$	Section 3.1
$\hat{p}_{h,t}$	transition-feasibility probability of stage h in the surrogate FIMDP, Eq. 2; a design choice	Section 3.1
$p^{\text{PoF}}_{h,t}$	the plug-in choice, the probability of feasibility of Eq. 3 (`FI-PoF-VI`)	Section 3.1
$p^+_{h,t}$, $p^+_{h,j,t}$	the optimistic choice and its per-constraint factor, $p^+_{h,t} = \prod_j p^+_{h,j,t}$, Eq. 4 (`FI-O-VI`)	Section 3.1
ρ_ξ	ramp $\rho_\xi(u) = \min\{1, \max\{0, \frac{2}{\xi}(u - \frac{\xi}{2})\}\}$	Eq. 4
$\xi_{h,j,t}$	feasibility threshold of constraint (h, j) at episode t	Definition 1
$V_{h,t}$, $Q_{h,t}$	value and action-value functions of the surrogate FIMDP	Section 3.1

Symbol	Meaning or convention	Section
$\tilde{V}_{h,t}, \tilde{Q}_{h,t}$	their restrictions to active states, $\tilde{V}_{h,t}(\boldsymbol{s}) = V_{h,t}(\boldsymbol{s}, 1)$, computed by Eq. 5	Section 3.1
$e_{h,t}$	exploration bonus of the transition model; $e^{\mathrm{TV}}_{h,t}$ its theoretical instantiation	Section 3.2, Appendix D.3
$\overline{r}$	upper bound of $\mathcal{J}$, used to cap the backup	Section 3.1
Threshold schedule and analysis		
$\xi_0, c_0(t), c_1(t)$	level and the two non-increasing schedules of Definition 1, $\xi_{h,j,t} = c_0(t)\xi_0 + c_1(t)\sigma^g_{h,j,t}$	Section 4
$\alpha, \zeta, t_{\mathrm{tr}}$	parameters of the implemented schedule $c_0(t) = t^{-\alpha}$, $c_1(t) = \beta_g 2^{1-s^\zeta}$, $s = (t-1)/(t_{\mathrm{tr}}-1)$ (Appendix D.2); $\bar{\alpha}_\kappa$ the admissible upper limit of α	Section 4
$t_{\mathrm{reach}}, t_{\mathrm{opt}}$	first episodes at which $c_0(t)\xi_0 + c_1(t)$ falls below ξ_{reach} and ξ_{opt}	Section 4
M	total number of constraints, $\sum_h m_h$	Section 4
A_T	number of aborted episodes among the first T	Section 4
R_T	cumulative value regret over T episodes	Section 4
δ	confidence level of the high-probability statements	Section 4
ψ	$\psi(x) = 2\Phi(x/2) - 1$, the total-variation factor of $e^{\mathrm{TV}}_{h,t}$	Section 4

B Related Work

Constrained Bayesian Optimisation (CBO) A large body of work studies BO with real-valued unknown constraints (Gardner et al., 2014; Gelbart et al., 2014; Qing et al., 2023). When violations are permitted, existing approaches provide regret and cumulative-violation guarantees via, e.g., penalty-based formulations (Lu & Paulson, 2022), primal-dual methods (Zhou & Ji, 2022), optimistic constraint modeling (Xu et al., 2023), and local penalty-function surrogates (Jingzhe et al., 2026). When violations must be avoided, safe BO methods (Sui et al., 2015; 2018; Berkenkamp et al., 2023; Fiedler et al., 2024) typically require feasible initial samples and progressively expand the reachable feasible region. More closely related to our setting, failure-aware CBO considers evaluations that may fail, where failure information may be handled through censored-regression models (Hutter et al., 2013; Karlova et al., 2026), classifiers (Bachoc et al., 2020; Antonio, 2021), or latent GP models thresholded on constraint values observed at successful evaluations (Marco et al., 2021; Wang & Li, 2024; Von Rohr et al., 2024) or stochastic failures Iwazaki et al. (2025). Notably, Iwazaki et al. (2023) introduce an annealing exclusion region centred on each observed failure and provide both a regret bound and a bound on the number of failures under an interior-optimum condition. These settings nevertheless do not consider multi-stage processes.

Multi-Stage and Function-Network Bayesian Optimisation. A growing line of work exploits intermediate observations by modelling systems as networks of interconnected functions rather than a single black box (Astudillo & Frazier, 2019; Nguyen et al., 2016; Astudillo & Frazier, 2021; Kusakawa et al., 2022; Buathong et al., 2023; Yoon et al., 2026). Among these, multi-stage BO considers sequential chains whose stage outputs serve as inputs to later stages (Nguyen et al., 2016; Kusakawa et al., 2022), including settings with noisy or partial intermediate observations (Torresi & Friederich, 2026; Kiroirwal et al., 2026), where a sample may be actively paused or discarded by the optimiser, while existing regret analyses focus on noise-free unconstrained chains (Li & Scarlett, 2022). Constrained extensions have considered constrained function networks (Kudva et al., 2026), constrained composite objectives (Paulson & Lu, 2022; Lu & Paulson, 2023), and BO under known transition constraints (Folch et al., 2024). Closest to our construction, Volk et al. (2023) plan reagent sequences in an autonomous multi-step synthesis by enumerating four-step action sequences and multiplying the predicted reward by the product of a classifier's per-step viability probabilities, suppressing the reward once that product falls below a fixed value. The multiplicative weight is the plug-in survival weight that `FI-PoF-VI` instantiates with GP feasibility in the backup of Section 3.1, while the cut on the accumulated product has no counterpart in either of our variants. Their state is the recent action history rather than a learned transition, the constraint enters only as a binary

label, optimism is placed on the predicted reward rather than on feasibility, and neither the number of terminated runs nor the regret is analysed.

Constrained Reinforcement Learning. Our SDM is also closely related to reinforcement learning with constraints (see, e.g., (Garcıa & Fernández, 2015; Gu et al., 2024) for reviews). The standard constrained MDP formulation (Altman, 2021) constrains policies through an expected cumulative cost budget, with optimistic exploration attaining sublinear regret together with sublinear or zero budget excess (Efroni et al., 2020; Liu et al., 2021; Wei et al., 2022). More closely related to ours are instantaneous or state-wise constraints, which require feasibility at every decision step (Shi et al., 2023; Xiong et al., 2023; Wei et al., 2024; Roknilamouki et al., 2025). Some of these methods avoid violations using prior safe actions or policies, while others permit violations during learning and bound their cumulative amount. However, constraint violation is not itself an episode-termination mechanism.

Termination has also been considered in constrained RL, and the device our FIMDP uses, a state augmented with an absorbing "aborted" indicator, is the standard one of this line of work; what is new here is what the indicator's transition depends on, an unknown per-stage constraint learned from censored feedback, and the optimistic transition feasibility placed on it. Sun et al. (2021) transform a CMDP into an early-terminated MDP that enters an absorbing state once the accumulated cost exceeds a prescribed budget. Chane-Sane et al. (2024); Milosevic et al. (2026) instead convert instantaneous violation signals into stochastic termination or state-action continuation probabilities that attenuate future reward or bootstrapping. In these approaches, the stopping rule is constructed from the underlying cost or violation signal, rather than being an unknown feasibility boundary learned under censored feedback. Closer to ours, Tennenholtz et al. (2022) study a tabular MDP in which an external observer terminates the episode with a probability determined by the cumulative sum of unknown state–action costs, which are inferred from termination and non-termination feedback. Their stopping is therefore stochastic and history-dependent, rather than triggered deterministically by violation of the current-stage constraint. Mazumdar et al. (2024) consider a reach-avoid setting that stops upon reaching known goal or forbidden sets under unknown transitions. However, these formulations do not consider unknown real-valued stage-wise constraints whose violation censors the next transition and all downstream observations.

In summary, our setting lies at the intersection of MSBO, CBO, and constrained RL, yet this specific problem setting is not addressed by the related BO or RL formulations above.

C Preliminaries

Finite-Horizon MDP. A finite-horizon Markov decision process (MDP) (Sutton et al., 1998) is a tuple $\langle \mathcal{S}_h, \mathcal{A}_h, P_h, R_h, H\rangle$ indexed by the stage $h = 0, \dots, H-1$, where H is the horizon, $\mathcal{S}_h \subseteq \mathbb{R}^{d_s}$ and $\mathcal{A}_h \subseteq \mathbb{R}^{d_a}$ are the state and action spaces of stage h, $P_h(\cdot \mid \boldsymbol{s}_h, \boldsymbol{a}_h)$ is the transition kernel from stage h to stage $h+1$, $R_h(\boldsymbol{s}_h, \boldsymbol{a}_h, \boldsymbol{s}_{h+1})$ is the stage reward for $h < H$, and $R_H(\boldsymbol{s}_H)$ is the terminal reward. A policy is a collection of stage-wise decision rules $\pi = \{\pi_h\}_{h=0}^{H-1}$ with $\pi_h : \mathcal{S}_h \to \mathcal{A}_h$, as in Section 2.1. Its value and action-value functions are

$$V_h^\pi(\boldsymbol{s}) = \mathbb{E}\Big[\sum\nolimits_{k=h}^{H-1} R_k\big(\boldsymbol{s}_k, \pi_k(\boldsymbol{s}_k), \boldsymbol{s}_{k+1}\big) + R_H(\boldsymbol{s}_H) \,\Big|\, \boldsymbol{s}_h = \boldsymbol{s}\Big],$$
$$Q_h^\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{s}' \sim P_h(\cdot \mid \boldsymbol{s}, \boldsymbol{a})}\big[R_h(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') + V_{h+1}^\pi(\boldsymbol{s}')\big],$$

where the expectation is over the trajectory generated by $\{P_k\}$ under π. The optimal value functions $V_h^\star = \max_\pi V_h^\pi$ and $Q_h^\star$ satisfy the Bellman optimality equations (Bellman, 1966)

$$Q_h^\star(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{s}' \sim P_h(\cdot \mid \boldsymbol{s}, \boldsymbol{a})}\big[R_h(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') + V_{h+1}^\star(\boldsymbol{s}')\big],$$
$$V_h^\star(\boldsymbol{s}) = \max_{\boldsymbol{a} \in \mathcal{A}_h} Q_h^\star(\boldsymbol{s}, \boldsymbol{a}), \qquad V_H^\star(\boldsymbol{s}) = R_H(\boldsymbol{s}),$$

so that $V_h^\star$ is obtained by backward recursion. We consider the episodic setting in which the process is executed for T episodes; in episode t a policy π^t is selected from the data gathered in earlier episodes.

Gaussian Process Surrogates. The unknown stage-wise mappings $\boldsymbol{f}_h$ and stopping constraints $g_{h,j}$ are modelled by Gaussian process (GP) surrogates. We write $\boldsymbol{z} = (\boldsymbol{s}_h, \boldsymbol{a}_h) \in \mathcal{S}_h \times \mathcal{A}_h$ for a

generic input of a stage-h model. Given a kernel κ, a likelihood variance $\lambda > 0$ and observations $\{(\boldsymbol{z}_i, y_i)\}_{i=1}^n$, the posterior mean and variance are

$$\mu_n(\boldsymbol{z}) = \boldsymbol{k}_n(\boldsymbol{z})^\top (\boldsymbol{K}_n + \lambda \boldsymbol{I})^{-1} \boldsymbol{y}_n, \qquad \sigma_n^2(\boldsymbol{z}) = \kappa(\boldsymbol{z}, \boldsymbol{z}) - \boldsymbol{k}_n(\boldsymbol{z})^\top (\boldsymbol{K}_n + \lambda \boldsymbol{I})^{-1} \boldsymbol{k}_n(\boldsymbol{z}), \quad (10)$$

with $\boldsymbol{K}_n = [\kappa(\boldsymbol{z}_i, \boldsymbol{z}_l)]_{i,l \le n}$, $\boldsymbol{k}_n(\boldsymbol{z}) = [\kappa(\boldsymbol{z}_i, \boldsymbol{z})]_{i \le n}$ and $\boldsymbol{y}_n = [y_i]_{i \le n}$ (Williams & Rasmussen, 2006). The transition models use $\lambda = \sigma_h^2$ and the constraint models a fixed $\lambda = \lambda_g$ (Section H.1). At round t, $\mu_{h,\ell,t}^f$ and $\sigma_{h,\ell,t}^f$ denote the posterior mean and standard deviation of the ℓ-th output of the stage-h transition model, with $\boldsymbol{\sigma}_{h,t}^f(\boldsymbol{z}) = (\sigma_{h,1,t}^f(\boldsymbol{z}), \dots, \sigma_{h,d_s,t}^f(\boldsymbol{z}))^\top$, and $\mu_{h,j,t}^g$, $\sigma_{h,j,t}^g$ those of the constraint model (h, j). The true and the posterior-predictive transition kernels are

$$P_h^\star(\cdot \mid \boldsymbol{z}) = \mathcal{N}\big(\boldsymbol{f}_h(\boldsymbol{z}), \sigma_h^2 \boldsymbol{I}_{d_s}\big), \qquad \widehat{P}_{h,t}(\cdot \mid \boldsymbol{z}) = \mathcal{N}\Big(\boldsymbol{\mu}_{h,t}^f(\boldsymbol{z}), \operatorname{diag}\Big((\boldsymbol{\sigma}_{h,t}^f(\boldsymbol{z}))^2\Big) + \sigma_h^2 \boldsymbol{I}_{d_s}\Big),$$

and for a bounded measurable function U, $P_h^\star U$ and $\widehat{P}_{h,t} U$ denote its integration against these kernels. For a scalar model with kernel κ and likelihood variance λ, the maximal information gain (Srinivas et al., 2009) of n observations is $\gamma_n = \frac{1}{2} \sup_{\boldsymbol{z}_{1:n}} \log\det(\boldsymbol{I} + \lambda^{-1} \boldsymbol{K}_n)$.

Bellman Recursions of the FIMDP and the Information Gain. With the true transitions and constraints, the optimal Bellman recursion of the FIMDP of Section 3.1 is

$$\begin{aligned} V_H^\star(\boldsymbol{s}) &= \mathcal{J}(\boldsymbol{s}), \\ Q_h^\star(\boldsymbol{s}, \boldsymbol{a}) &= \mathbb{1}\{g_{h,j}(\boldsymbol{z}) \ge 0,\ \forall j \in [m_h]\}\, (P_h^\star V_{h+1}^\star)(\boldsymbol{z}), \\ V_h^\star(\boldsymbol{s}) &= \max_{\boldsymbol{a} \in \mathcal{A}_h} Q_h^\star(\boldsymbol{s}, \boldsymbol{a}), \qquad v^\star = V_0^\star(\boldsymbol{s}_0). \end{aligned} \quad (11)$$

For `FI-O-VI`, the empirical Bellman recursion is

$$\begin{aligned} \widetilde{V}_{H,t}(\boldsymbol{s}) &= \mathcal{J}(\boldsymbol{s}), \\ \widetilde{Q}_{h,t}(\boldsymbol{s}, \boldsymbol{a}) &= \min\Big\{ p_{h,t}^+(\boldsymbol{z}) \Big[(\widehat{P}_{h,t} \widetilde{V}_{h+1,t})(\boldsymbol{z}) + e_{h,t}^{\mathrm{TV}}(\boldsymbol{z}) \Big], \overline{r} \Big\}, \\ \widetilde{V}_{h,t}(\boldsymbol{s}) &= \max_{\boldsymbol{a} \in \mathcal{A}_h} \widetilde{Q}_{h,t}(\boldsymbol{s}, \boldsymbol{a}), \end{aligned} \quad (12)$$

where $p_{h,t}^+$ uses Eq. 4 with the schedule in Eq. 6, and the bonus term is given in Eq. 8. The clip at $\overline{r}$ is an analysis device that keeps all true and optimistic values in $[0, \overline{r}]$; Eq. 5 and the implementation do not clip. Without the clip the recursion gives $\widetilde{V}_{h,t} \le (H - h + 1)\, \overline{r}$, and every bound below holds with $\overline{r}$ replaced by that constant wherever the range of $\widetilde{V}_{h+1,t}$ enters. In addition, for a scalar model with likelihood variance λ we write

$$\Gamma_n = \frac{2\gamma_n}{\log(1 + \lambda^{-1})},$$

the bound on the sum of posterior variances over any n queries (Srinivas et al., 2009, Lemma 5.3).

D Algorithm Details

D.1 KRR based Value Model

We provide below the closed form of the expectation in Eq. 5 under the transition posterior (Section D.1.1) and the construction of the support set of the value model of Section 3.2 (Section D.1.2).

D.1.1 Closed form of the expected continuation value

The expectation of an RBF kernel under a Gaussian input is again an RBF kernel with inflated lengthscales (Girard et al., 2002; Deisenroth & Rasmussen, 2011): for $\hat{\boldsymbol{f}}_h \sim \mathcal{N}(\boldsymbol{\mu}, \operatorname{diag}(\boldsymbol{\sigma}^2))$ and $k_v(\boldsymbol{s}, \boldsymbol{s}') = \sigma_v^2 \exp\big(-\frac{1}{2} \sum_d (s_d - s'_d)^2 / \ell_{v,d}^2\big)$,

$$\tilde{k}_v(\boldsymbol{\mu}, \boldsymbol{s}_i) \;:=\; \mathbb{E}_{\hat{\boldsymbol{f}}_h}\big[k_v(\hat{\boldsymbol{f}}_h, \boldsymbol{s}_i)\big] = \sigma_v^2 \prod_{d=1}^{d_s} \frac{\ell_{v,d}}{\sqrt{\ell_{v,d}^2 + \sigma_d^2}} \exp\Big(- \frac{(\mu_d - s_{i,d})^2}{2(\ell_{v,d}^2 + \sigma_d^2)} \Big), \quad (13)$$

Algorithm 1 FI-O-VI (FI-PoF-VI differs only in the continuation weight of line 5)

Require: horizon H, budget T, initial state $\boldsymbol{s}_0$, initial data $\mathcal{D}_0$, terminal objective $\mathcal{J}$, feasibility-threshold schedule (Definition 1), confidence scales β_f, β_g

1: **for** $t = 1, \dots, T$ **do**
2: fit the posteriors $\hat{\boldsymbol{f}}_{h,t}$, $\hat{g}_{h,j,t}$ on $\mathcal{D}_{t-1}$ and set the feasibility thresholds $\xi_{h,j,t}$ of the schedule ▷ Eq. 10, Definition 1; in practice applied at the stage being executed, Section D.2
BACKWARD OPTIMISTIC VALUE ITERATION
3: $\tilde{V}_{H,t} \leftarrow \mathcal{J}$
4: **for** $h = H-1, \dots, 0$ **do**
5: $\hat{p}_{h,t}(\boldsymbol{s},\boldsymbol{a}) \leftarrow p^{+}_{h,t}(\boldsymbol{s},\boldsymbol{a}) = \prod_{j \le m_h} \rho_{\xi_{h,j,t}}\big(\mathrm{ucb}^{g}_{h,j,t}(\boldsymbol{s},\boldsymbol{a})\big)$ ▷ FI-PoF-VI: $\Phi(\mu^{g}_{h,j,t}/\sigma^{g}_{h,j,t})$
6: $\tilde{Q}_{h,t}(\boldsymbol{s},\boldsymbol{a}) \leftarrow \hat{p}_{h,t}(\boldsymbol{s},\boldsymbol{a})\Big(e_{h,t}(\boldsymbol{s},\boldsymbol{a}) + \mathbb{E}_{\hat{\boldsymbol{f}}_{h,t}} \mathbb{E}_{\boldsymbol{\epsilon}_h}\big[\tilde{V}_{h+1,t}\big(\hat{\boldsymbol{f}}_{h,t}(\boldsymbol{s},\boldsymbol{a}) + \boldsymbol{\epsilon}_h\big)\big]\Big)$
7: $\tilde{V}_{h,t} \leftarrow \mathrm{KRR}\left\{\big(\boldsymbol{s}_i, \max_{\boldsymbol{a} \in \mathcal{A}_h} \tilde{Q}_{h,t}(\boldsymbol{s}_i,\boldsymbol{a})\big)\right\}_{i \le N^{t}_{\mathrm{syn}}}$ ▷ support states: Section D.1.2
GREEDY EXECUTION FROM $\boldsymbol{s}_0$
8: $\boldsymbol{s}_{0,t} \leftarrow \boldsymbol{s}_0$
9: **for** $h = 0, \dots, H-1$ **do**
10: execute $\boldsymbol{a}^{t}_{h} \in \arg\max_{\boldsymbol{a} \in \mathcal{A}_h} \tilde{Q}_{h,t}(\boldsymbol{s}^{t}_{h},\boldsymbol{a})$ and observe $\{g_{h,j}(\boldsymbol{s}^{t}_{h},\boldsymbol{a}^{t}_{h})\}_{j \le m_h}$
11: **if** $\min_j g_{h,j}(\boldsymbol{s}^{t}_{h},\boldsymbol{a}^{t}_{h}) < 0$ **then**
12: **abort**: $I_{h+1} \leftarrow 0$; the next state and $\mathcal{J}$ are unobserved
13: **break**
14: observe $\boldsymbol{s}^{t}_{h+1} = \boldsymbol{f}_h(\boldsymbol{s}^{t}_{h},\boldsymbol{a}^{t}_{h}) + \boldsymbol{\epsilon}_h$
15: **if** $I_H = 1$ **then** observe $\mathcal{J}(\boldsymbol{s}^{t}_{H})$; append the episode's observations to $\mathcal{D}_t$

where $\sigma^2_d = (\sigma^{f}_{h,d,t})^2 + \sigma^2_h$ collects the transition posterior variance and the process noise, which enter the same Gaussian. Since $\tilde{V}_{h+1}(\boldsymbol{s}) = \sum_i w_{h+1,i}\, k_v(\boldsymbol{s},\boldsymbol{s}_i)$ is linear in the kernel, the expectation in Eq. 5 follows by linearity,

$$\mathbb{E}_{\hat{\boldsymbol{f}}_h} \mathbb{E}_{\boldsymbol{\epsilon}_h}\big[\tilde{V}_{h+1}\big(\hat{\boldsymbol{f}}_h(\boldsymbol{s},\boldsymbol{a}) + \boldsymbol{\epsilon}_h\big)\big] = \sum\nolimits_i w_{h+1,i}\, \tilde{k}_v\big(\boldsymbol{\mu}^{f}_{h,t}(\boldsymbol{s},\boldsymbol{a}),\, \boldsymbol{s}_i\big), \tag{14}$$

so the backup is evaluated in closed form and needs no sampling.

D.1.2 Support Point Selection

The KRR value model $\tilde{V}_h$ is fitted on a support set of states with three layers, each serving one purpose.

Observed states. Every state observed at stage h in previous episodes, so that $\tilde{V}_h$ is accurate where the policy goes.

Space-filling coverage. A fixed scrambled Sobol design of N_{Sobol} points over the stage's declared state box $[\boldsymbol{\ell}_h, \boldsymbol{u}_h]$ (Appendix H.1), identical in every episode, so that the regression does not extrapolate where no state has been observed.

We add $N_{\text{V-dpp}}$ adaptive support points using greedy maximum-a-posteriori inference for a quality-weighted determinantal point process (DPP) (Kulesza & Taskar, 2012; Chen et al., 2018). At stage h, the candidate set consists of $N_{\mathrm{cand}} = 500$ points sampled uniformly from $[\boldsymbol{\ell}_h, \boldsymbol{u}_h]$ once with a fixed seed. For episode $t > 1$, the quality assigned to a candidate $\boldsymbol{s}$ is $|\tilde{V}^{(t-1)}_h(\boldsymbol{s})|$. In the first episode, all candidates are assigned equal quality. Similarity is defined by a squared-exponential kernel with lengthscale $\ell_{\mathrm{DPP},h} = w_h N^{-1/d_s}$, where w_h is the median side length of the stage-h state box and $N = 50$. This scale corresponds to the characteristic spacing of N points in a d_s-dimensional box. The greedy selection is conditioned on the observed-state and Sobol support points already included in the support set. The adaptive points therefore favour regions with large estimated value while remaining diverse with respect to the existing support points. Related quality-weighted DPP constructions have been used for inducing-point placement (Moss et al., 2023) and for subsampling data in Gaussian-process Q-function models (Bloor et al., 2025). Here, we use the DPP to select synthetic inputs for the KRR value model.

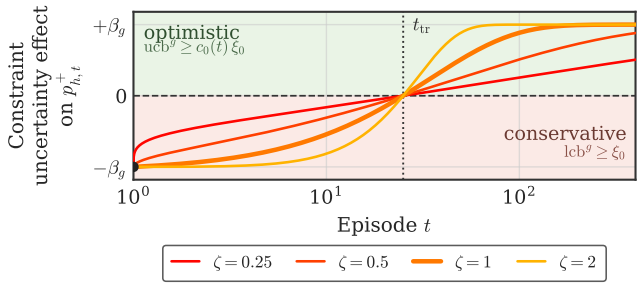


Figure 5: **Effect of constraint uncertainty on the continuation weight.** The coefficient $\beta_g - c_1(t)$ changes from $-\beta_g$ at $t = 1$, through zero at $t = t_{\mathrm{tr}}$, towards $+\beta_g$. The corresponding condition therefore changes from an LCB criterion to a posterior-mean criterion and then to a UCB criterion. The figure uses $t_{\mathrm{tr}} = 25$; ζ controls the rate of this change.

D.2 Algorithmic Specification of ξ_t

Implemented Schedule. We instantiate Definition 1 with

$$c_0(t) = t^{-\alpha}, \qquad c_1(t) = \beta_g\, 2^{\,1-s^\zeta}, \qquad s = \frac{t-1}{t_{\mathrm{tr}}-1}, \tag{15}$$

Substituting these choices into Eq. 6 gives the implemented feasibility threshold

$$\xi_{h,j,t}(\boldsymbol{s},\boldsymbol{a}) = t^{-\alpha}\xi_0 + \beta_g\, 2^{\,1-s^\zeta}\sigma^g_{h,j,t}(\boldsymbol{s},\boldsymbol{a}), \qquad s = \frac{t-1}{t_{\mathrm{tr}}-1}.$$

Here, $\xi_0 > 0$ sets the initial base level of the feasibility threshold, while α controls how this level decreases across episodes. The episode t_{tr} marks the change from conservative to optimistic treatment of constraint uncertainty, while $\zeta > 0$ controls how rapidly this change occurs. Since both $c_0(t)$ and $c_1(t)$ converge to zero, the limiting condition in Definition 1 is satisfied for any $\xi_0 > 0$.

To see the effect of this schedule, recall that a constraint receives unit feasibility weight when $\mathrm{ucb}^g_{h,j,t} \geq \xi_{h,j,t}$. Substituting the schedule gives

$$\mu^g_{h,j,t} + \big(\beta_g - c_1(t)\big)\sigma^g_{h,j,t} \geq c_0(t)\xi_0. \tag{16}$$

At $t = 1$, $c_1(1) = 2\beta_g$, so this condition is equivalent to $\mathrm{lcb}^g_{h,j,1} \geq \xi_0$. At $t = t_{\mathrm{tr}}$, $c_1(t) = \beta_g$, and the posterior uncertainty cancels, leaving $\mu^g_{h,j,t} \geq t_{\mathrm{tr}}^{-\alpha}\xi_0$. After t_{tr}, $c_1(t) < \beta_g$, so posterior uncertainty contributes optimistically. As t increases, the condition approaches $\mathrm{ucb}^g_{h,j,t} \geq 0$. When $\zeta = 1$, $c_1(t)$ halves every $t_{\mathrm{tr}} - 1$ episodes. Figure 5 shows how the coefficient $\beta_g - c_1(t)$ evolves across episodes for different values of ζ.

We choose β_g, ξ_0, t_{tr}, and ζ empirically, and set $\alpha = \bar{\alpha}_\kappa/2$ using the admissible rate derived below. The scheduled feasibility threshold is used at the stage being executed, while subsequent stages in the lookahead use the fixed threshold 0.01 specified in Section H.2.

Admissible Decay Rates. Since $c_0(T) = T^{-\alpha}$, the abort bound of Theorem 1 and the first term of the regret bound of Theorem 2 both contain $\gamma_T/c_0(T)^2 = \gamma_T T^{2\alpha}$. For the linear and squared-exponential kernels, the information-gain bounds of Srinivas et al. (2009) grow at most polylogarithmically with T, so this term is sublinear for $0 < \alpha < 1/2$. For a Matérn-ν kernel, Vakili et al. (2021) give $\gamma_T = \widetilde{O}\big(T^{d_{sa}/(2\nu+d_{sa})}\big)$, which gives $0 < \alpha < \nu/(2\nu + d_{sa})$. Table 2 summarises these upper limits, where $d_{sa} = d_s + d_a$ is the state-action dimension. In the experiments, we use $\alpha = \bar{\alpha}_\kappa/2$, the midpoint of the corresponding admissible range.

D.3 Algorithmic Specification of $e_{h,t}$, β_f, β_g

To specify a practical exploration bonus, we keep the form of the bonus used in the analysis of Section 4.2, Eq. 8, which is proportional to the posterior standard deviation of the transition model,

Table 2: Admissible decay rates $\bar{\alpha}_\kappa$ of the annealed schedule (Definition 1) for standard kernels, from the information-gain rates γ_T of Srinivas et al. (2009) (linear, squared exponential) and Vakili et al. (2021) (Matérn); $d_{sa} = d_s + d_a$.

Kernel κ	γ_T	$\bar{\alpha}_\kappa$
Linear	$O(d_{sa} \log T)$	$1/2$
Squared exponential	$O\big((\log T)^{d_{sa}+1}\big)$	$1/2$
Matérn-ν	$\tilde{O}\big(T^{d_{sa}/(2\nu+d_{sa})}\big)$	$\nu/(2\nu+d_{sa})$

$e_{h,t}(\boldsymbol{s},\boldsymbol{a}) \propto \|\boldsymbol{\sigma}^f_{h,t}(\boldsymbol{s},\boldsymbol{a})\|$, and fix its scale from known quantities:

$$e_{h,t}(\boldsymbol{s},\boldsymbol{a}) = \beta_f\, \bar{r}\, \frac{\big\|\boldsymbol{\sigma}^f_{h,t}(\boldsymbol{s},\boldsymbol{a})\big\|}{\sigma^{\text{prior}}_{h,t}}, \qquad \sigma^{\text{prior}}_{h,t} := \big\|\big(\theta^{1/2}_{h,1,t}, \ldots, \theta^{1/2}_{h,d_s,t}\big)\big\|, \tag{17}$$

where $\bar{r}$ is the reward bound of Section 2.1 and $\theta_{h,i,t}$ is the kernel variance of the i-th output coordinate of stage-h's transition model at episode t, so that $\sigma^{\text{prior}}_{h,t} = \sup_{\boldsymbol{z}} \|\boldsymbol{\sigma}^f_{h,t}(\boldsymbol{z})\|$ is the prior standard deviation of the transition model and $e_{h,t} = \beta_f \bar{r}$ at an input about which nothing is known. Following common practice (Srinivas et al., 2009; Curi et al., 2020) the scaling coefficients β_f and β_g are set empirically.

E Post-Transition Stopping Constraints

In the post-transition setting, stage h first produces $\boldsymbol{s}_{h+1} = \boldsymbol{f}_h(\boldsymbol{s}_h, \boldsymbol{a}_h) + \boldsymbol{\epsilon}_h$, after which the stopping constraints $c_{h,j} : \mathcal{S}_{h+1} \to \mathbb{R}$ are evaluated: $I_{h+1} = I_h\, \mathbb{1}\Big\{c_{h,j}(\boldsymbol{s}_{h+1}) \geq 0,\ \forall j \in [m_h]\Big\}$. Thus a violation censors all later stages and the terminal reward, but not $\boldsymbol{s}_{h+1}$ itself.

FI-O-VI in the Post-Transition Setting. The constraint value $c_{h,j}(\boldsymbol{s}_{h+1})$ is observed for every executed decision and is modelled as a function of $\boldsymbol{z}_h = (\boldsymbol{s}_h, \boldsymbol{a}_h)$ as before. Since the transition is observed regardless of whether the stage survives, the exploration bonus is no longer weighted by feasibility, and the Bellman backup becomes

$$\widetilde{Q}_{h,t}(\boldsymbol{s},\boldsymbol{a}) = e_{h,t}(\boldsymbol{s},\boldsymbol{a}) + p^+_{h,t}(\boldsymbol{s},\boldsymbol{a})\widehat{P}_{h,t}\widetilde{V}_{h+1,t}(\boldsymbol{s},\boldsymbol{a}), \qquad \widetilde{V}_{h,t}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\mathcal{A}_h} \widetilde{Q}_{h,t}(\boldsymbol{s},\boldsymbol{a}). \tag{18}$$

Compared with Eq. 5, an action with $p^+_{h,t} = 0$ may still attain a large value through its transition bonus. When selecting the action actually applied to the process, we therefore restrict the maximisation to actions with $p^+_{h,t} > 0$. During training, this ensures that every executed action satisfies $\text{ucb}^g_{h,j,t} > \xi_{h,j,t}/2$ for all stopping constraints, which is the condition used in the abort-count analysis. We impose the same restriction when evaluating the deployed policy.

Deterministic Stages. If $\boldsymbol{\epsilon}_h \equiv \mathbf{0}$, the stopping event is determined by the fixed function of the preceding state–action pair $g_{h,j}(\boldsymbol{s},\boldsymbol{a}) := c_{h,j}(\boldsymbol{f}_h(\boldsymbol{s},\boldsymbol{a}))$, so the abort-count analysis reduces to the pre-transition case. For a linear specification $c_{h,j}(\boldsymbol{s}) = \boldsymbol{w}^\top_{h,j}\boldsymbol{s} + b_{h,j}$, with the known offset $b_{h,j}$ absorbed into the constant prior mean, $g_{h,j} = \sum_i w_{h,j,i} f_{h,i}$ and $\|g_{h,j}\|_\kappa \leq \|\boldsymbol{w}_{h,j}\|_1 B$. Hence Theorem 1 applies with $\|\boldsymbol{w}_{h,j}\|_1 B$ in place of B. For nonlinear $c_{h,j}$, the same conclusion holds if the regularity assumption is imposed directly on the composite $g_{h,j}$.

The regret guarantee of Theorem 2 does not follow from the current proof because its total-variation transition bound degenerates when $\sigma_h = 0$. The same issue arises for deterministic pre-transition stages, so this is a limitation of the analysis rather than of the post-transition formulation, and does not imply that FI-O-VI is not asymptotically optimal in the deterministic setting. We speculate that an alternative analysis could instead control transition error through the smoothness of the optimistic value function. In particular, the feasibility weight of Eq. 4 is piecewise linear and the fitted KRR value function is Lipschitz in the state, yielding a stage-wise Lipschitz bound for the backup in Eq. 18. We leave a formal analysis of this approach, as well as the stochastic noise setting to future work.

F PROOFS

F.1 PROOF OF THE ABORT COUNT BOUND

Lemma 1. *For any stage h, constraint index j, and episode t,*

$$|g_{h,j}(\boldsymbol{z}) - \mu^g_{h,j,t}(\boldsymbol{z})| \le B\sigma^g_{h,j,t}(\boldsymbol{z}). \tag{19}$$

Proof. Write $g = g_{h,j}$ and abbreviate $\mu_n = \mu^g_{h,j,t}$ and $\sigma_n = \sigma^g_{h,j,t}$. Let $\boldsymbol{K}_n = [\kappa(\boldsymbol{z}_i, \boldsymbol{z}_l)]_{i,l}$ be the Gram matrix. Denote $\boldsymbol{k}_n(\boldsymbol{z}) = (\kappa(\boldsymbol{z}_1, \boldsymbol{z}), \dots, \kappa(\boldsymbol{z}_n, \boldsymbol{z}))^\top$, $\boldsymbol{g}_n = (g(\boldsymbol{z}_1), \dots, g(\boldsymbol{z}_n))^\top$ and $\boldsymbol{A}_n = (\boldsymbol{K}_n + \lambda \boldsymbol{I})^{-1}$. Recall that

$$\mu_n(\boldsymbol{z}) = \boldsymbol{k}_n(\boldsymbol{z})^\top \boldsymbol{A}_n \boldsymbol{g}_n,$$
$$\sigma_n^2(\boldsymbol{z}) = \kappa(\boldsymbol{z}, \boldsymbol{z}) - \boldsymbol{k}_n(\boldsymbol{z})^\top \boldsymbol{A}_n \boldsymbol{k}_n(\boldsymbol{z}).$$

By the reproducing property, we have

$$\mu_n(\boldsymbol{z}) = \sum_{i=1}^n \left[\boldsymbol{A}_n \boldsymbol{k}_n(\boldsymbol{z})\right]_i \langle g,\, \kappa(\boldsymbol{z}_i, \cdot)\rangle_\kappa$$
$$= \left\langle g,\, \sum_{i=1}^n \left[\boldsymbol{A}_n \boldsymbol{k}_n(\boldsymbol{z})\right]_i \kappa(\boldsymbol{z}_i, \cdot) \right\rangle_\kappa.$$

Define the residual evaluation representer as

$$r_{n,\boldsymbol{z}} = \kappa(\boldsymbol{z}, \cdot) - \sum_{i=1}^n \left[\boldsymbol{A}_n \boldsymbol{k}_n(\boldsymbol{z})\right]_i \kappa(\boldsymbol{z}_i, \cdot).$$

Using $\boldsymbol{A}_n \boldsymbol{K}_n \boldsymbol{A}_n = \boldsymbol{A}_n - \lambda \boldsymbol{A}_n^2$ gives

$$\|r_{n,\boldsymbol{z}}\|_\kappa^2 = \kappa(\boldsymbol{z}, \boldsymbol{z}) - 2\boldsymbol{k}_n(\boldsymbol{z})^\top \boldsymbol{A}_n \boldsymbol{k}_n(\boldsymbol{z}) + \boldsymbol{k}_n(\boldsymbol{z})^\top \boldsymbol{A}_n \boldsymbol{K}_n \boldsymbol{A}_n \boldsymbol{k}_n(\boldsymbol{z})$$
$$= \sigma_n^2(\boldsymbol{z}) - \lambda \boldsymbol{k}_n(\boldsymbol{z})^\top \boldsymbol{A}_n^2 \boldsymbol{k}_n(\boldsymbol{z}) \le \sigma_n^2(\boldsymbol{z}).$$

By the Cauchy–Schwarz inequality, we have

$$g(\boldsymbol{z}) - \mu_n(\boldsymbol{z}) = \langle g, r_{n,\boldsymbol{z}}\rangle_\kappa \le \|g\|_\kappa \|r_{n,\boldsymbol{z}}\|_\kappa \le B\sigma_n(\boldsymbol{z}),$$

which proves Eq. 19. □

Lemma 2. *Under Assumption 1, for every $t \ge t_{\text{reach}}$, every action $\boldsymbol{a}_h^t$ selected by FI-O-VI at a reached alive state $\boldsymbol{s}_h^t$ satisfies*

$$p^+_{h,t}(\boldsymbol{s}_h^t, \boldsymbol{a}_h^t) > 0, \quad \text{ucb}^g_{h,j,t}(\boldsymbol{s}_h^t, \boldsymbol{a}_h^t) > \frac{\xi_{h,j,t}(\boldsymbol{s}_h^t, \boldsymbol{a}_h^t)}{2} \ge \frac{c_0(t)\xi_0}{2} \tag{20}$$

for every $j \in [m_h]$.

Proof. Fix (t, h) with $I_h^t = 1$. We first prove that, for any reached alive state $\boldsymbol{s}_h^t$, there is an action $\bar{\boldsymbol{a}}_h^t \in \mathcal{A}_h$ such that $p^+_{h,t}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t) = 1$. In fact, the reachability part of Assumption 1 provides an action $\bar{\boldsymbol{a}}_h^t \in \mathcal{A}_h$ such that

$$g_{h,j}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t) \ge \xi_{\text{reach}} \ge \xi_{h,j,t}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t), \quad \forall j \in [m_h],$$

where the second inequality follows from $\xi_{h,j,t}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t) \le c_0(t)\xi_0 + c_1(t) \le \xi_{\text{reach}}$ for $t \ge t_{\text{reach}}$. The exact confidence inequality in Lemma 1 implies, for every $j \in [m_h]$,

$$\text{ucb}^g_{h,j,t}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t) = \mu^g_{h,j,t}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t) + B\sigma^g_{h,j,t}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t)$$
$$\ge g_{h,j}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t) \ge \xi_{h,j,t}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t).$$

By the definition of the ramp in Eq. 4, $\rho_\xi(u) = 1$ whenever $u \ge \xi$. Hence every factor of the transition feasibility $p^+_{h,t}$ at the witness equals one, and therefore

$$p^+_{h,t}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t) = \prod_{j=1}^{m_h} \rho_{\xi_{h,j,t}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t)}\left(\text{ucb}^g_{h,j,t}(\boldsymbol{s}_h^t, \bar{\boldsymbol{a}}_h^t)\right) = 1. \tag{21}$$

Now we prove the claim. Define

$$W_{h,t}(\boldsymbol{s},\boldsymbol{a}) := (\widehat{P}_{h,t}\widetilde{V}_{h+1,t})(\boldsymbol{s},\boldsymbol{a}) + e^{\mathrm{TV}}_{h,t}(\boldsymbol{s},\boldsymbol{a}).$$

Since the terminal reward is nonnegative, backward induction in Eq. 12 gives $\widetilde{V}_{h+1,t} \geq 0$. Since $\widehat{P}_{h,t}$ is a probability kernel and $e^{\mathrm{TV}}_{h,t} > 0$, it follows that

$$W_{h,t}(\boldsymbol{s},\boldsymbol{a}) \geq e^{\mathrm{TV}}_{h,t}(\boldsymbol{s},\boldsymbol{a}) > 0 \qquad \forall \boldsymbol{a} \in \mathcal{A}_h. \tag{22}$$

Since $I^t_h = 1$, Eqs. 21 and 22 yield

$$\widetilde{Q}_{h,t}(\boldsymbol{s}^t_h, \bar{\boldsymbol{a}}^t_h) = \min\left\{W_{h,t}(\boldsymbol{s}^t_h, \bar{\boldsymbol{a}}^t_h), \bar{r}\right\} > 0.$$

The executed action $\boldsymbol{a}^t_h$ is a maximizer of this objective, and therefore

$$\widetilde{Q}_{h,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h) > 0,$$

which forces $p^+_{h,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h) > 0$. For the selected action,

$$p^+_{h,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h) = \prod_{j=1}^{m_h} \rho_{\xi_{h,j,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h)}\left(\mathrm{ucb}^g_{h,j,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h)\right) > 0.$$

From Eq. 4, $\rho_{\xi_{h,j,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h)}(u) > 0$ holds if and only if $u > \xi_{h,j,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h)/2$. Therefore

$$\mathrm{ucb}^g_{h,j,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h) > \frac{\xi_{h,j,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h)}{2} \qquad \forall j \in [m_h].$$

The construction in Eq. 6 proves the final inequality in Eq. 20. □

Lemma 3. *If an episode $t \geq t_{\mathrm{reach}}$ aborts at stage h for violation of constraint $j \in [m_h]$,*

$$\sigma^g_{h,j,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h) > \frac{c_0(t)\xi_0}{4B}. \tag{23}$$

Proof. Let $\boldsymbol{z}_t = (\boldsymbol{s}^t_h, \boldsymbol{a}^t_h)$ be the state–action pair at which the abort occurs. Since stage h is reached while the episode is still alive, Lemma 2 implies that

$$\mathrm{ucb}^g_{h,j,t}(\boldsymbol{z}_t) > \frac{\xi_{h,j,t}(\boldsymbol{z}_t)}{2}. \tag{24}$$

Violation of constraint j implies that $g_{h,j}(\boldsymbol{z}_t) < 0$. By Lemma 1, we have

$$\begin{aligned}\mathrm{ucb}^g_{h,j,t}(\boldsymbol{z}_t) &= \mu^g_{h,j,t}(\boldsymbol{z}_t) + B\sigma^g_{h,j,t}(\boldsymbol{z}_t) \\ &\leq g_{h,j}(\boldsymbol{z}_t) + 2B\sigma^g_{h,j,t}(\boldsymbol{z}_t) \\ &< 2B\sigma^g_{h,j,t}(\boldsymbol{z}_t).\end{aligned} \tag{25}$$

Combining Eqs. 24 and 25 gives

$$\frac{\xi_{h,j,t}(\boldsymbol{z}_t)}{2} < \mathrm{ucb}^g_{h,j,t}(\boldsymbol{z}_t) < 2B\sigma^g_{h,j,t}(\boldsymbol{z}_t).$$

Dividing by $2B > 0$ proves $\sigma^g_{h,j,t}(\boldsymbol{z}_t) > \xi_{h,j,t}(\boldsymbol{z}_t)/(4B)$, and the lower bound follows from Eq. 6. □

Proof of Theorem 1. Separate the post-admission aborts from the earlier episodes by defining

$$\mathcal{A}^+_T := \{t \in [T] : t \geq t_{\mathrm{reach}},\ I^t_H = 0\}.$$

Then $A_T \leq \min\{T, t_{\mathrm{reach}} - 1\} + |\mathcal{A}^+_T|$. For each $t \in \mathcal{A}^+_T$, let $\tau_t \in \{0, \ldots, H-1\}$ be the abort stage. At the executed pair $(\boldsymbol{s}^t_{\tau_t}, \boldsymbol{a}^t_{\tau_t})$, at least one constraint is negative. If several constraints are negative, choose one of them arbitrarily and denote its index by j_t. Thus

$$g_{\tau_t, j_t}(\boldsymbol{s}^t_{\tau_t}, \boldsymbol{a}^t_{\tau_t}) < 0. \tag{26}$$

For each constraint pair (h, j), define

$$\mathcal{A}_{h,j}(T) = \{t \in \mathcal{A}_T^+ : (\tau_t, j_t) = (h, j)\}, \quad N_{h,j}(T) = |\mathcal{A}_{h,j}(T)|.$$

Every post-admission aborted episode is assigned to exactly one pair (h, j). Therefore the sets $\mathcal{A}_{h,j}(T)$ are disjoint and their union is $\mathcal{A}_T^+$, which gives

$$|\mathcal{A}_T^+| = \sum_{h=0}^{H-1} \sum_{j=1}^{m_h} N_{h,j}(T). \tag{27}$$

We now bound $N_{h,j}(T)$ for a fixed pair (h, j). Let $n_h(T)$ be the number of episodes among $[T]$ that reach stage h while still alive. Sort these episodes in increasing order as

$$t_{h,1} < \cdots < t_{h,n_h(T)},$$

and denote their executed inputs by

$$\boldsymbol{z}_{h,q} := (\boldsymbol{s}_h^{t_{h,q}}, \boldsymbol{a}_h^{t_{h,q}}), \quad q \in [n_h(T)].$$

Write $\widetilde{\sigma}^g_{h,j,q-1}$ for the posterior standard deviation after the first $q-1$ observations in this sequence and before observing the value at $\boldsymbol{z}_{h,q}$. By construction,

$$\widetilde{\sigma}^g_{h,j,q-1}(\boldsymbol{z}_{h,q}) = \sigma^g_{h,j,t_{h,q}}(\boldsymbol{s}_h^{t_{h,q}}, \boldsymbol{a}_h^{t_{h,q}}).$$

Every $t \in \mathcal{A}_{h,j}(T)$ reaches stage h, so it has a unique index $q(t)$ satisfying $t = t_{h,q(t)}$. Moreover, the assigned constraint j is negative by Eq. 26. Applying Lemma 3 and using $t \le T$ gives

$$\widetilde{\sigma}^g_{h,j,q(t)-1}(\boldsymbol{z}_{h,q(t)}) > \frac{\xi_{h,j,t}(\boldsymbol{z}_{h,q(t)})}{4B} \ge \frac{c_0(t)\xi_0}{4B} \ge \frac{c_0(T)\xi_0}{4B}. \tag{28}$$

Squaring and summing Eq. 28 over the assigned episodes yields

$$\begin{aligned} N_{h,j}(T)\frac{c_0(T)^2\xi_0^2}{16B^2} &< \sum_{t\in\mathcal{A}_{h,j}(T)} \left(\widetilde{\sigma}^g_{h,j,q(t)-1}(\boldsymbol{z}_{h,q(t)})\right)^2 \\ &\le \sum_{q=1}^{n_h(T)} \left(\widetilde{\sigma}^g_{h,j,q-1}(\boldsymbol{z}_{h,q})\right)^2. \end{aligned} \tag{29}$$

For a normalized kernel, the sum of posterior variances is bounded by the maximal information gain as (Srinivas et al., 2009, Lemma 5.3)

$$\sum_{q=1}^{n_h(T)} \left(\widetilde{\sigma}^g_{h,j,q-1}(\boldsymbol{z}_{h,q})\right)^2 \le \frac{2\gamma_{n_h(T)}}{\log(1+\lambda^{-1})} = \Gamma_{n_h(T)}. \tag{30}$$

Since $n_h(T) \le T$ and maximal information gain is nondecreasing, $\Gamma_{n_h(T)} \le \Gamma_T$. Combining Eqs. 29 and 30 gives

$$N_{h,j}(T) \le \frac{16B^2\Gamma_T}{c_0(T)^2\xi_0^2} = \frac{32B^2\gamma_T}{c_0(T)^2\xi_0^2\log(1+\lambda^{-1})}. \tag{31}$$

Finally, there are $M = \sum_{h=0}^{H-1} m_h$ scalar constraint models. Summing Eq. 31 over all (h, j) in Eq. 27, adding the pre-admission episodes, and using $A_T \le T$ proves the final bound

$$A_T \le \min\left\{T,\ t_{\text{reach}} - 1 + \frac{32MB^2\gamma_T}{c_0(T)^2\xi_0^2\log(1+\lambda^{-1})}\right\}. \tag{32}$$

This is exactly Eq. 7. □

Theorem 1 bounds the abort count once the schedule is fixed. Rearranging Eq. 32 for the threshold scale reads the bound in the opposite direction, turning a target abort rate at a finite budget into a sufficient condition on $c_0(T)\xi_0$.

F.2 Proof of the Value Regret Bound

Lemma 4. *For every $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $h \in [H]$, $\ell \in [d_s]$, $t \in [T]$, and $\boldsymbol{z} \in \mathcal{Z}_h$,*

$$|f_{h,\ell}(\boldsymbol{z}) - \mu^f_{h,\ell,t}(\boldsymbol{z})| \le \beta_t \sigma^f_{h,\ell,t}(\boldsymbol{z}), \tag{33}$$

where

$$\beta_t = B + \frac{\overline{\sigma}}{\sqrt{\lambda}}\sqrt{2\left(\gamma_{t-1} + \log\frac{D_f}{\delta}\right)}.$$

Proof. The proof follows Chowdhury & Gopalan (2017) with a few adaptations. In particular, fix a stage h and coordinate ℓ, and write $f = f_{h,\ell}$. Denote the canonical feature map by $\varphi(\boldsymbol{z}) = \kappa(\boldsymbol{z}, \cdot)$. Define the covariance operator as

$$V_n = \lambda I_{\mathcal{H}_\kappa} + \sum_{i=1}^{n} \varphi(\boldsymbol{z}_i) \otimes \varphi(\boldsymbol{z}_i). \tag{34}$$

and also define $S_n = \sum_{i=1}^n \epsilon_i \varphi(\boldsymbol{z}_i)$ where $\epsilon_i \sim \mathcal{N}(0, \sigma_h^2)$. Using the same argument as Chowdhury & Gopalan (2017), we obtain that

$$|f(\boldsymbol{z}) - \mu_n(\boldsymbol{z})| \le \left(B + \frac{1}{\sqrt{\lambda}}\|S_n\|_{V_n^{-1}}\right)\sigma_n(\boldsymbol{z}), \tag{35}$$

where $\|u\|^2_{V_n^{-1}} = \langle u, V_n^{-1}u\rangle_\kappa$. The self-normalized kernel martingale inequality (Chowdhury & Gopalan, 2017, Theorem 1) implies that, for any $\eta \in (0,1)$, with probability at least $1-\eta$, simultaneously for all $n \ge 0$,

$$\|S_n\|_{V_n^{-1}} \le \overline{\sigma}\sqrt{2\left(\gamma_n + \log(1/\eta)\right)}. \tag{36}$$

Substitution into Eq. 35 gives,

$$|f(\boldsymbol{z}) - \mu_n(\boldsymbol{z})| \le \left[B + \frac{\overline{\sigma}}{\sqrt{\lambda}}\sqrt{2\left(\gamma_n + \log(1/\eta)\right)}\right]\sigma_n(\boldsymbol{z}). \tag{37}$$

Apply Eq. 37 to each of the $D_f = Hd_s$ coordinates with $\eta = \delta/D_f$. A union bound shows that all scalar bounds hold with probability at least $1 - D_f\eta = 1-\delta$, which proves the claim as required.

□

Lemma 5. *Let P and Q be probability measures on the same space $\mathcal{M}$. Then every measurable function $U : \mathcal{M} \to [0, \overline{r}]$ satisfies*

$$|PU - QU| \le \overline{r}\ \mathrm{TV}(P, Q), \tag{38}$$

where $\mathrm{TV}(P,Q) = \sup_{A\in\mathcal{M}} |P(A) - Q(A)|$ is the total variation distance, and

$$PU = \int_{\mathcal{M}} U(x)\, P(dx), \quad QU = \int_{\mathcal{M}} U(x)\, Q(dx).$$

Proof. For each $x \in \mathcal{M}$, the bound $0 \le U(x) \le \overline{r}$ gives the layer-cake representation

$$U(x) = \int_0^{\overline{r}} \mathbb{1}\{U(x) > u\}\, du. \tag{39}$$

Since the integrand is nonnegative and measurable, Tonelli's theorem implies that

$$\begin{aligned} PU &= \int_{\mathcal{M}} \int_0^{\overline{r}} \mathbb{1}\{U(x) > u\}\, du\, P(dx) \\ &= \int_0^{\overline{r}} P\big(\{x \in \mathcal{M} : U(x) > u\}\big)\, du. \end{aligned} \tag{40}$$

For each u, denote the superlevel set by $A_u = \{x \in \mathcal{M} : U(x) > u\}$. Applying the triangle inequality, we have

$$\begin{aligned} |PU - QU| &= \left| \int_0^{\overline{r}} \big(P(A_u) - Q(A_u)\big)\, du \right| \\ &\le \int_0^{\overline{r}} |P(A_u) - Q(A_u)|\, du \\ &\le \int_0^{\overline{r}} \mathrm{TV}(P, Q)\, du = \overline{r}\ \mathrm{TV}(P, Q), \end{aligned}$$

which proves Eq. 38. □

In particular, when P and Q are Gaussians we have the following explicit expression for the total variation distance.

Lemma 6 (Barsov & Ulyanov, 1987, Theorem 1)**.** *Let $d \ge 1$, $\boldsymbol{m}_1, \boldsymbol{m}_2 \in \mathbb{R}^d$, and $\sigma > 0$. Then*

$$\mathrm{TV}\big(\mathcal{N}(\boldsymbol{m}_1, \sigma^2 \boldsymbol{I}_d), \mathcal{N}(\boldsymbol{m}_2, \sigma^2 \boldsymbol{I}_d)\big) = 2\Phi\left(\frac{\|\boldsymbol{m}_1 - \boldsymbol{m}_2\|_2}{2\sigma}\right) - 1, \tag{41}$$

where $\Phi(x)$ denotes the standard normal cumulative distribution function.

Lemma 7. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, simultaneously for all $h \in [H]$, $t \in [T]$, and $\boldsymbol{z} \in \mathcal{Z}_h$,*

$$\mathrm{TV}\left(P_h^\star(\cdot \mid \boldsymbol{z}), \widehat{P}_{h,t}(\cdot \mid \boldsymbol{z})\right) \le \psi\left(\frac{\beta_t + 1}{\sigma_h} \|\boldsymbol{\sigma}_{h,t}^f(\boldsymbol{z})\|_2\right). \tag{42}$$

Consequently, every measurable $U : \mathcal{S}_{h+1} \to [0, \overline{r}]$ satisfies

$$\left|(P_h^\star U)(\boldsymbol{z}) - (\widehat{P}_{h,t} U)(\boldsymbol{z})\right| \le \overline{r}\, \psi\left(\frac{\beta_t + 1}{\sigma_h} \|\boldsymbol{\sigma}_{h,t}^f(\boldsymbol{z})\|_2\right). \tag{43}$$

Proof. Fix $(h, t, \boldsymbol{z})$, and define

$$\boldsymbol{d} = \boldsymbol{f}_h(\boldsymbol{z}) - \boldsymbol{\mu}_{h,t}^f(\boldsymbol{z}).$$

By Lemma 4, for each $\ell \in [d_s]$, with probability at least $1 - \delta$, it holds $|d_\ell| \le \beta_t \sigma_{h,\ell,t}^f(\boldsymbol{z})$. Squaring and summing over the coordinates yields

$$\|\boldsymbol{d}\|_2 \le \beta_t \|\boldsymbol{\sigma}_{h,t}^f(\boldsymbol{z})\|_2. \tag{44}$$

For the GP sample $\hat{\boldsymbol{f}}_{h,t}$, define the centered variable

$$\boldsymbol{\eta} = \hat{\boldsymbol{f}}_{h,t}(\boldsymbol{z}) - \boldsymbol{\mu}_{h,t}^f(\boldsymbol{z}).$$

Condition on the data $\mathcal{D}_t$, it holds

$$\boldsymbol{\eta} \mid \mathcal{D}_t \sim \mathcal{N}\Big(\boldsymbol{0}, [\mathrm{diag}(\boldsymbol{\sigma}_{h,t}^f(\boldsymbol{z}))]^2\Big).$$

For a fixed realization of $\boldsymbol{\eta}$, let $Q_{\boldsymbol{\eta}} = \mathcal{N}\left(\boldsymbol{\mu}_{h,t}^f(\boldsymbol{z}) + \boldsymbol{\eta}, \sigma_h^2 \boldsymbol{I}_{d_s}\right)$. Conditional on $\boldsymbol{\eta}$, the sampled transition $\hat{\boldsymbol{f}}_{h,t}(\boldsymbol{z}) + \boldsymbol{\epsilon}_h$ has law $Q_{\boldsymbol{\eta}}$. Thus, integrating over the GP posterior gives

$$\widehat{P}_{h,t}(\cdot \mid \boldsymbol{z}) = \mathbb{E}_{\boldsymbol{\eta}} Q_{\boldsymbol{\eta}}.$$

Recall that $P_h^\star(\cdot \mid \boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{\mu}_{h,t}^f(\boldsymbol{z}) + \boldsymbol{d}, \sigma_h^2 \boldsymbol{I}_{d_s}\right)$. For every measurable set A,

$$|P_h^\star(A \mid \boldsymbol{z}) - \mathbb{E}_{\boldsymbol{\eta}} Q_{\boldsymbol{\eta}}(A)| \le \mathbb{E}_{\boldsymbol{\eta}} |P_h^\star(A \mid \boldsymbol{z}) - Q_{\boldsymbol{\eta}}(A)|.$$

Taking the supremum over A and moving the supremum inside the expectation by Jensen's inequality yields

$$\begin{aligned} \mathrm{TV}\Big(P_h^\star(\cdot \mid \boldsymbol{z}), \widehat{P}_{h,t}(\cdot \mid \boldsymbol{z})\Big) &\le \mathbb{E}_{\boldsymbol{\eta}}\, \mathrm{TV}(P_h^\star(\cdot \mid \boldsymbol{z}), Q_{\boldsymbol{\eta}}) \\ &= \mathbb{E}_{\boldsymbol{\eta}} \psi\left(\frac{\|\boldsymbol{d} - \boldsymbol{\eta}\|_2}{\sigma_h}\right), \end{aligned} \tag{45}$$

where the last equality follows from Lemma 6.

Writing ϕ for the standard normal density, $\psi'(x) = \phi(x/2) > 0$ and $\psi''(x) = -(x/4)\phi(x/2) \le 0$ for $x \ge 0$. Thus ψ is increasing and concave on $[0, \infty)$. Consequently,

$$\mathbb{E}_{\boldsymbol{\eta}}\psi\left(\frac{\|\boldsymbol{d}-\boldsymbol{\eta}\|_2}{\sigma_h}\right) \le \mathbb{E}_{\boldsymbol{\eta}}\psi\left(\frac{\|\boldsymbol{d}\|_2+\|\boldsymbol{\eta}\|_2}{\sigma_h}\right) \le \psi\left(\frac{\|\boldsymbol{d}\|_2+\mathbb{E}\|\boldsymbol{\eta}\|_2}{\sigma_h}\right), \tag{46}$$

where the first inequality uses the triangle inequality and monotonicity of ψ, and the second is Jensen's inequality. Moreover, the Cauchy–Schwarz inequality implies that

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\eta}\|_2 &\le \sqrt{\mathbb{E}\|\boldsymbol{\eta}\|_2^2} = \sqrt{\sum_{\ell=1}^{d_s}\mathbb{E}\eta_\ell^2} \\ &= \sqrt{\sum_{\ell=1}^{d_s}[\sigma^f_{h,\ell,t}(\boldsymbol{z})]^2} \\ &= \|\boldsymbol{\sigma}^f_{h,t}(\boldsymbol{z})\|_2. \end{aligned} \tag{47}$$

Combining Eq. 44 – Eq. 47 proves Eq. 42. Finally, because $0 \le U \le \overline{r}$, Lemma 5 gives

$$\left|(P^\star_h U)(\boldsymbol{z}) - (\widehat{P}_{h,t}U)(\boldsymbol{z})\right| \le \overline{r}\ \mathrm{TV}\Big(P^\star_h(\cdot \mid \boldsymbol{z}), \widehat{P}_{h,t}(\cdot \mid \boldsymbol{z})\Big),$$

which, together with Eq. 42, proves Eq. 43. □

Lemma 8. *Suppose Assumption 2 holds. For any $\delta \in (0,1)$, with probability at least $1-\delta$, for every $t \in [T] \cap \{t : t \ge t_{\mathrm{opt}}\}$,*

$$\widetilde{V}_{0,t}(\boldsymbol{s}_0) \ge v^\star. \tag{48}$$

Proof. First of all, for a fixed $\delta \in (0,1)$, we apply Lemma 7, therefore, with probability at least $1-\delta$,

$$\left|(P^\star_h U)(\boldsymbol{z}) - (\widehat{P}_{h,t}U)(\boldsymbol{z})\right| \le e^{\mathrm{TV}}_{h,t}(\boldsymbol{z}) \tag{49}$$

simultaneously for all $h, t, \boldsymbol{z}$ and every measurable $U : \mathcal{S}_{h+1} \to [0, \overline{r}]$. We work on this simultaneous event in the following.

At any live state, the interior optimal policy satisfies

$$V_H^{\pi^\star}(\boldsymbol{s}) = \mathcal{J}(\boldsymbol{s}), \quad V_h^{\pi^\star}(\boldsymbol{s}) = (P^\star_h V^{\pi^\star}_{h+1})(\boldsymbol{s}, \pi^\star_h(\boldsymbol{s})).$$

The constraint indicator equals one by Assumption 2. We prove that

$$\widetilde{V}_{h,t}(\boldsymbol{s}_h) \ge V_h^{\pi^\star}(\boldsymbol{s}_h), \tag{50}$$

almost surely on $\{I_h = 1\}$, by backward induction on h. At $h = H$, the two terminal values are identical

$$\widetilde{V}_{H,t}(\boldsymbol{s}) = \mathcal{J}(\boldsymbol{s}) = V_H^{\pi^\star}(\boldsymbol{s}).$$

Now assume that

$$\widetilde{V}_{h+1,t}(\boldsymbol{s}') \ge V_{h+1}^{\pi^\star}(\boldsymbol{s}')$$

almost surely at the live next states generated by $\pi^\star$. Fix a live state $\boldsymbol{s}$ generated by this policy, set $\boldsymbol{a} = \pi^\star_h(\boldsymbol{s})$, and write $\boldsymbol{z} = (\boldsymbol{s}, \boldsymbol{a})$. By Assumption 2,

$$g_{h,j}(\boldsymbol{z}) \ge \xi_{\mathrm{opt}} \ge c_0(t)\xi_0 + c_1(t) \ge \xi_{h,j,t}(\boldsymbol{z}), \quad \forall j \in [m_h]. \tag{51}$$

By Lemma 1, we have

$$\mathrm{ucb}^g_{h,j,t}(\boldsymbol{z}) = \mu^g_{h,j,t}(\boldsymbol{z}) + B\sigma^g_{h,j,t}(\boldsymbol{z}) \ge g_{h,j}(\boldsymbol{z}) \ge \xi_{h,j,t}(\boldsymbol{z}).$$

Thus every ramp factor equals one and $p^+_{h,t}(\boldsymbol{z}) = 1$. Taking $U = \widetilde{V}_{h+1,t}$ in Eq. 49 gives

$$\begin{aligned}(\widehat{P}_{h,t}\widetilde{V}_{h+1,t})(\boldsymbol{z}) + e^{\mathrm{TV}}_{h,t}(\boldsymbol{z}) &\geq (P^\star_h \widetilde{V}_{h+1,t})(\boldsymbol{z}) \\ &\geq (P^\star_h V^{\pi^\star}_{h+1})(\boldsymbol{z}) = V^{\pi^\star}_h(\boldsymbol{s}). \end{aligned} \tag{52}$$

The second inequality follows from the induction hypothesis and positivity of $P^\star_h$. Since the transition feasibility $p^+_{h,t}$ equals one and $V^{\pi^\star}_h(\boldsymbol{s}) \leq \overline{r}$, clipping the left hand side of Eq. 52 at $\overline{r}$ preserves the inequality. Therefore

$$\widetilde{V}_{h,t}(\boldsymbol{s}) = \max_{\boldsymbol{a}' \in \mathcal{A}_h} \widetilde{Q}_{h,t}(\boldsymbol{s}, \boldsymbol{a}') \geq \widetilde{Q}_{h,t}(\boldsymbol{s}, \boldsymbol{a}) \geq V^{\pi^\star}_h(\boldsymbol{s}),$$

which completes the induction. At $h = 0$, $V^{\pi^\star}_0(\boldsymbol{s}_0) = v^\star$ since $\pi^\star$ is optimal. This proves Eq. 48. □

Lemma 9. *Fix $\delta \in (0,1)$. For every reached transition (t,h) with $I^t_{h+1} = 1$, define*

$$Z_{t,h} = (P^\star_h \widetilde{V}_{h+1,t}(\cdot, 1))(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h) - \widetilde{V}_{h+1,t}(\boldsymbol{s}^t_{h+1}, 1). \tag{53}$$

With probability at least $1-\delta$, every $t \in [T]$ with $I^t_H = 1$ satisfies

$$\widetilde{V}_{0,t}(\boldsymbol{s}^t_0, 1) - r^t \leq 2 \sum_{h=0}^{H-1} e^{\mathrm{TV}}_{h,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h) + \sum_{h=0}^{H-1} Z_{t,h}. \tag{54}$$

Proof. Similar to Lemma 8, we work on the following event which holds with probability at least $1-\delta$,

$$\left|(\widehat{P}_{h,t}U)(\boldsymbol{z}) - (P^\star_h U)(\boldsymbol{z})\right| \leq e^{\mathrm{TV}}_{h,t}(\boldsymbol{z}).$$

Fix $t \in [T]$ such that $I^t_H = 1$. At stage h, the executed action $\boldsymbol{a}^t_h$ maximizes the planner's action value. Set $\boldsymbol{z}^t_h = (\boldsymbol{s}^t_h, \boldsymbol{a}^t_h)$. It follows that

$$\begin{aligned}\widetilde{V}_{h,t}(\boldsymbol{s}^t_h) &= \widetilde{Q}_{h,t}(\boldsymbol{s}^t_h, \boldsymbol{a}^t_h) \\ &\leq (\widehat{P}_{h,t}\widetilde{V}_{h+1,t})(\boldsymbol{z}^t_h) + e^{\mathrm{TV}}_{h,t}(\boldsymbol{z}^t_h). \end{aligned} \tag{55}$$

Combining it with Lemma 7 gives

$$\begin{aligned}\widetilde{V}_{h,t}(\boldsymbol{s}^t_h) &\leq (P^\star_h \widetilde{V}_{h+1,t})(\boldsymbol{z}^t_h) + 2e^{\mathrm{TV}}_{h,t}(\boldsymbol{z}^t_h) \\ &= \widetilde{V}_{h+1,t}(\boldsymbol{s}^t_{h+1}) + Z_{t,h} + 2e^{\mathrm{TV}}_{h,t}(\boldsymbol{z}^t_h). \end{aligned} \tag{56}$$

Summing Eq. 56 over $h = 0, \ldots, H-1$ telescopes, we obtain that

$$\widetilde{V}_{0,t}(\boldsymbol{s}^t_0) - \widetilde{V}_{H,t}(\boldsymbol{s}^t_H) \leq \sum_{h=0}^{H-1} Z_{t,h} + 2\sum_{h=0}^{H-1} e^{\mathrm{TV}}_{h,t}(\boldsymbol{z}^t_h).$$

Since $I^t_H = 1$, the episode reaches the terminal stage and $r^t = \mathcal{J}(\boldsymbol{s}^t_H) = \widetilde{V}_{H,t}(\boldsymbol{s}^t_H)$, which proves Eq. 54. □

Proof of Theorem 2. Fix $\delta \in (0,1)$. We first work on the event where Lemma 7 holds, which has probability at least $1 - \delta/2$. Define

$$\mathcal{T}^+_T = \{t \in [T] : t \geq t_{\mathrm{opt}}\}.$$

We partition $[T]$ into three parts: (i) the episodes before t_{opt}, (ii) the completed episodes in $\mathcal{T}^+_T$, (iii) the aborted episodes in $\mathcal{T}^+_T$. Every episode before t_{opt} contributes at most $\overline{r}$ to regret. For $t \in \mathcal{T}^+_T$ with $I^t_H = 1$, Lemmas 8 and 9 imply that

$$v^\star - r^t \leq 2 \sum_{h=0}^{H-1} e^{\mathrm{TV}}_{h,t}(\boldsymbol{z}^t_h) + \sum_{h=0}^{H-1} Z_{t,h}.$$

If $I_H^t = 0$, then $r^t = 0$ and $v^\star - r^t \le \overline{r}$. Define $n_{\mathrm{opt}}(T) = \min\{T, t_{\mathrm{opt}} - 1\}$ and $A_T^+ = \sum_{t\in\mathcal{T}_T^+} \mathbb{1}\{I_H^t = 0\}$. Summing over the three classes of episodes yields

$$\begin{aligned} R_T &= \Bigg(\sum_{\substack{t\in[T]:\\ t<t_{\mathrm{opt}}}} + \sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=1}} + \sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=0}} \Bigg)(v^\star - r^t) \\ &\le \overline{r} n_{\mathrm{opt}}(T) + \sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=1}} (v^\star - r^t) + \overline{r} A_T^+ \\ &\le \overline{r} n_{\mathrm{opt}}(T) + \overline{r} A_T^+ + 2 \sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=1}} \sum_{h=0}^{H-1} e_{h,t}^{\mathrm{TV}}(\boldsymbol{z}_h^t) + \sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=1}} \sum_{h=0}^{H-1} Z_{t,h}. \end{aligned} \tag{57}$$

For the last term, we write as

$$\sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=1}} \sum_{h=0}^{H-1} Z_{t,h} = \sum_{t\in\mathcal{T}_T^+} \sum_{h=0}^{H-1} I_{h+1}^t Z_{t,h} - \sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=0}} \sum_{h=0}^{H-1} I_{h+1}^t Z_{t,h}.$$

Since $|Z_{t,h}| \le \overline{r}$, and each aborted episode contains at most H survived transitions, it follows that

$$- \sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=0}} \sum_{h=0}^{H-1} I_{h+1}^t Z_{t,h} \le H\overline{r} A_T^+.$$

Substituting into Eq. 57 gives the following decomposition

$$R_T \le \overline{r} n_{\mathrm{opt}}(T) + \underbrace{(H+1)\overline{r} A_T^+}_{\text{(I)}} + \underbrace{2 \sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=1}} \sum_{h=0}^{H-1} e_{h,t}^{\mathrm{TV}}(\boldsymbol{z}_h^t)}_{\text{(II)}} + \underbrace{\sum_{t\in\mathcal{T}_T^+} \sum_{h=0}^{H-1} I_{h+1}^t Z_{t,h}}_{\text{(III)}}. \tag{58}$$

We bound these terms in order.

(I). Since $t_{\mathrm{opt}} \ge t_{\mathrm{reach}}$, the post-admission argument in the proof of Theorem 1 applies to every abort counted by A_T^+. Hence

$$A_T^+ \le \min\left\{ T - n_{\mathrm{opt}}(T),\ \frac{32MB^2\gamma_T}{c_0(T)^2 \xi_0^2 \log(1+\lambda^{-1})} \right\}.$$

Consequently,

$$\text{(I)} \le (H+1)\overline{r} \min\left\{ T - n_{\mathrm{opt}}(T),\ \frac{32MB^2\gamma_T}{c_0(T)^2 \xi_0^2 \log(1+\lambda^{-1})} \right\}. \tag{59}$$

(II). Let ϕ denote the standard normal density. For every $x \ge 0$,

$$\psi(x) = 2\Phi(x/2) - 1 = 2\int_0^{x/2} \phi(u)\, du \le \frac{x}{\sqrt{2\pi}}.$$

The last inequality follows from $\phi(u) \le \phi(0) = 1/\sqrt{2\pi}$. Substituting the argument of ψ from Eq. 8 into this inequality gives

$$\begin{aligned} e_{h,t}^{\mathrm{TV}}(\boldsymbol{z}) &= \overline{r}\, \psi\left(\frac{\beta_t^f + 1}{\sigma_h} \|\boldsymbol{\sigma}_{h,t}^f(\boldsymbol{z})\|_2 \right) \\ &\le \frac{\overline{r}(\beta_t^f + 1)}{\sigma_h \sqrt{2\pi}} \|\boldsymbol{\sigma}_{h,t}^f(\boldsymbol{z})\|_2 \\ &\le \frac{\overline{r}(\beta_T^f + 1)}{\underline{\sigma}\sqrt{2\pi}} \|\boldsymbol{\sigma}_{h,t}^f(\boldsymbol{z})\|_2. \end{aligned}$$

The standard posterior variance bound (Srinivas et al., 2009, Lemma 5.3) gives,

$$\sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=1}} [\sigma^f_{h,\ell,t}(\boldsymbol{z}_h^t)]^2 \le \Gamma_T.$$

Summing this inequality over the $D_f = Hd_s$ transition coordinates yields

$$\sum_{\substack{t\in\mathcal{T}_T^+:\\ I_H^t=1}} \sum_{h=0}^{H-1} \|\boldsymbol{\sigma}^f_{h,t}(\boldsymbol{z}_h^t)\|_2^2 \le D_f\Gamma_T. \tag{60}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}\text{(II)} &\le \frac{2\overline{r}(\beta_T^f+1)}{\underline{\sigma}\sqrt{2\pi}}\sqrt{H\big(T-n_{\mathrm{opt}}(T)\big)}\sqrt{D_f\Gamma_T}\\ &= \frac{2\overline{r}(\beta_T^f+1)}{\underline{\sigma}\sqrt{\pi\log(1+\lambda^{-1})}}\sqrt{HD_f\big(T-n_{\mathrm{opt}}(T)\big)\gamma_T}.\end{aligned} \tag{61}$$

(III). Let $\prec$ denote the lexicographic order on the pairs $(t,h)\in[T]\times[H]$. Let $\mathcal{F}_{t,h}$ be the σ-algebra generated by $\{(\boldsymbol{s}_{h'}^{t'},\boldsymbol{\epsilon}_{h'}^{t'})\}_{(t',h')\prec(t,h)}$ and $\boldsymbol{s}_h^t$. By construction, on $\{I_{h+1}^t=1\}$,

$$\boldsymbol{s}_{h+1}^t \mid \mathcal{F}_{t,h} \sim P_h^\star(\cdot\mid \boldsymbol{z}_h^t).$$

Set $X_{t,h} = \mathbb{1}\{t\in\mathcal{T}_T^+\}I_{h+1}^t Z_{t,h}$, with $X_{t,h}=0$ when $I_{h+1}^t=0$. Since I_{h+1}^t, $\boldsymbol{z}_h^t$, and $\widetilde{V}_{h+1,t}$ are $\mathcal{F}_{t,h}$-measurable, the preceding conditional law yields

$$\mathbb{E}[X_{t,h}\mid\mathcal{F}_{t,h}] = \mathbb{1}\{t\in\mathcal{T}_T^+\}I_{h+1}^t\left[(P_h^\star\widetilde{V}_{h+1,t})(\boldsymbol{z}_h^t) - \int \widetilde{V}_{h+1,t}(\boldsymbol{s}')P_h^\star(d\boldsymbol{s}'\mid\boldsymbol{z}_h^t)\right] = 0.$$

Thus the lexicographically ordered partial sums of $X_{t,h}$ form a martingale. Moreover, $|X_{t,h}|\le\overline{r}$, and only $H\big(T-n_{\mathrm{opt}}(T)\big)$ increments can be nonzero. If $T=n_{\mathrm{opt}}(T)$, then (III) $=0$. Otherwise, by the Azuma–Hoeffding inequality, for every $x>0$,

$$\mathbb{P}\Big(\sum_{t\in\mathcal{T}_T^+}\sum_{h=0}^{H-1} X_{t,h}\ge x\Big) \le \exp\left(-\frac{x^2}{2H(T-n_{\mathrm{opt}}(T))\overline{r}^2}\right).$$

Taking $x=\overline{r}\sqrt{2H\big(T-n_{\mathrm{opt}}(T)\big)\log(2/\delta)}$ makes the right-hand side $\delta/2$. Thus, with probability at least $1-\delta/2$,

$$\text{(III)} \le \overline{r}\sqrt{2H\big(T-n_{\mathrm{opt}}(T)\big)\log(2/\delta)}. \tag{62}$$

The decomposition in Eq. 58 holds with probability at least $1-\delta/2$. The bounds for terms (I) and (II) are deterministic, while the martingale bound for term (III) fails with probability at most $\delta/2$. A union bound therefore shows that all three bounds hold simultaneously with probability at least $1-\delta$. Substituting Eqs. 59 – Eq. 62 into Eq. 58 gives

$$\begin{aligned}R_T \le{}& \overline{r}n_{\mathrm{opt}}(T) + (H+1)\overline{r}\min\left\{T-n_{\mathrm{opt}}(T),\ \frac{32MB^2\gamma_T}{c_0(T)^2\xi_0^2\log(1+\lambda^{-1})}\right\}\\ &+\frac{2\overline{r}(\beta_T^f+1)}{\underline{\sigma}\sqrt{\pi\log(1+\lambda^{-1})}}\sqrt{HD_f\big(T-n_{\mathrm{opt}}(T)\big)\gamma_T}\\ &+\overline{r}\sqrt{2H\big(T-n_{\mathrm{opt}}(T)\big)\log(2/\delta)}.\end{aligned} \tag{63}$$

Since $n_{\mathrm{opt}}(T)\le t_{\mathrm{opt}}-1$ is a finite constant, this implies the dominant rate in Eq. 9. □

F.3 Comparison with FI-PoF-VI

Consider the special case where $H = 1$, $\mathcal{S}_0 = \{\boldsymbol{s}_0\}$ and $\mathcal{A}_0$ is a finite set with at least two actions. Fix a feasible action $\boldsymbol{a}^{\mathrm{b}}$ that is *not* optimal, hence there is a positive gap

$$\eta^{\mathrm{b}} = v^\star - Q_0^\star(\boldsymbol{s}_0, \boldsymbol{a}^{\mathrm{b}}) > 0.$$

Write $\boldsymbol{z} = (\boldsymbol{s}_0, \boldsymbol{a})$. Assume that, for some fixed $\varrho \in [0, 1)$,

$$\left|\kappa(\boldsymbol{z}, \boldsymbol{z}^{\mathrm{b}})\right| \le \varrho, \quad \forall \boldsymbol{a} \in \mathcal{A}_0 \setminus \{\boldsymbol{a}^{\mathrm{b}}\}.$$

For FI-PoF-VI, write

$$p_{0,t}^{\mathrm{PoF}}(\boldsymbol{z}) = \prod_{j=1}^{m_0} \Phi\left(\frac{\mu_{0,j,t}^g(\boldsymbol{z})}{\sigma_{0,j,t}^g(\boldsymbol{z})}\right), \quad p_{\mathrm{b}} = p_{0,1}^{\mathrm{PoF}}(\boldsymbol{z}^{\mathrm{b}}).$$

Define the following quantity which acts as an upper bound on the PoF of an unvisited action

$$\overline{p}_\varrho = \prod_{j=1}^{m_0} \Phi\left(\frac{\varrho g_{0,j}(\boldsymbol{z}^{\mathrm{b}})}{\sqrt{1-\varrho^2}}\right).$$

Theorem 3. *Assume that* $0 < \underline{r} \le \mathcal{J}(\boldsymbol{s}) \le \overline{r}$ *for any* $\boldsymbol{s} \in \mathcal{S}_1$, *and*

$$\underline{r}\, p_{\mathrm{b}} > 2\overline{r}\, \overline{p}_\varrho. \tag{64}$$

- *For FI-PoF-VI, if* $\boldsymbol{a}^{\mathrm{b}}$ *is selected initially, then it will be selected in every episode. In particular, for* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$,

$$R_T^{\mathrm{PoF}} \ge \eta^{\mathrm{b}} T - \overline{r}\sqrt{\frac{T}{2}\log\frac{1}{\delta}}. \tag{65}$$

- *For FI-O-VI, choose the schedule of Definition 1 with* $c_0(t) = t^{-1/4}$. *For every* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$,

$$R_T^{\mathrm{O}} = O\left(\sqrt{T}\log(1+T)\right). \tag{66}$$

Proof. We prove the claims for FI-PoF-VI and FI-O-VI separately.

Linear Regret for FI-PoF-VI. We first prove by induction that if $\boldsymbol{a}^{\mathrm{b}}$ is selected initially, then it will be selected in every following episode. Suppose this is true for the first t episodes. It follows that the Gram matrix is $\boldsymbol{K}_n = \mathbf{1}_n \mathbf{1}_n^\top$, and it holds that

$$(\boldsymbol{K}_n + \lambda \boldsymbol{I})^{-1}\mathbf{1}_n = \frac{\mathbf{1}_n}{\lambda + n}, \text{ and } \mathbf{1}_n^\top(\boldsymbol{K}_n + \lambda \boldsymbol{I})^{-1}\mathbf{1}_n = \frac{n}{\lambda + n}.$$

Substituting into Eq. 10 gives

$$\begin{cases} \mu_{0,j,t}^g(\boldsymbol{z}^{\mathrm{b}}) = \dfrac{t}{\lambda + t} g_{0,j}(\boldsymbol{z}^{\mathrm{b}}), \\ [\sigma_{0,j,t}^g(\boldsymbol{z}^{\mathrm{b}})]^2 = 1 - \dfrac{t}{\lambda + t} = \dfrac{\lambda}{\lambda + t}, \end{cases}$$

which implies that

$$\frac{\mu_{0,j,t}^g(\boldsymbol{z}^{\mathrm{b}})}{\sigma_{0,j,t}^g(\boldsymbol{z}^{\mathrm{b}})} = \frac{t\, g_{0,j}(\boldsymbol{z}^{\mathrm{b}})}{\sqrt{\lambda(\lambda + t)}}. \tag{67}$$

The feasibility condition implies that $g_{0,j}(\boldsymbol{z}^{\mathrm{b}}) \ge 0$ for every j. Define a function $\theta(t) = t/\sqrt{t + \lambda}$. A direct computation shows that

$$\theta'(t) = \frac{\lambda + t/2}{(\lambda + t)^{3/2}} > 0.$$

Thus the quotient in Eq. 67 is increasing in t. Since Φ is increasing and its values are nonnegative, it follows that $p_{0,t}^{\mathrm{PoF}}(\boldsymbol{z}^{\mathrm{b}})$ is increasing and

$$p_{0,t}^{\mathrm{PoF}}(\boldsymbol{z}^{\mathrm{b}}) \geq p_{\mathrm{b}}.$$

For any other $\boldsymbol{a} \neq \boldsymbol{a}^{\mathrm{b}}$, write $\boldsymbol{z} = (\boldsymbol{s}_0, \boldsymbol{a})$, and $k_{\boldsymbol{a}} = \kappa(\boldsymbol{z}, \boldsymbol{z}^{\mathrm{b}})$. A similar computation gives

$$\begin{cases} \mu_{0,j,t}^{g}(\boldsymbol{z}) = \dfrac{t}{\lambda + t} k_{\boldsymbol{a}} g_{0,j}(\boldsymbol{z}^{\mathrm{b}}), \\ [\sigma_{0,j,t}^{g}(\boldsymbol{z})]^2 = 1 - \dfrac{t}{\lambda + t} k_{\boldsymbol{a}}^2. \end{cases}$$

By assumption, we have

$$|\mu_{0,j,t}^{g}(\boldsymbol{z})| = \frac{t}{\lambda + t} |k_{\boldsymbol{a}}| g_{0,j}(\boldsymbol{z}^{\mathrm{b}}) \leq \varrho g_{0,j}(\boldsymbol{z}^{\mathrm{b}}), \quad [\sigma_{0,j,t}^{g}(\boldsymbol{z})]^2 \geq 1 - \varrho^2 > 0.$$

Together they imply that

$$p_{0,t}^{\mathrm{PoF}}(\boldsymbol{z}) \leq \prod_{j=1}^{m_0} \Phi\left(\frac{\varrho g_{0,j}(\boldsymbol{z}^{\mathrm{b}})}{\sqrt{1 - \varrho^2}} \right) = \overline{p}_{\varrho}. \tag{68}$$

Recall that the iteration for `FI-PoF-VI` is

$$\widetilde{Q}_{0,t}^{\mathrm{PoF}}(\boldsymbol{z}) = \min\left\{ p_{0,t}^{\mathrm{PoF}}(\boldsymbol{z}) \left[(\widehat{P}_{0,t} \mathcal{J})(\boldsymbol{z}) + e_{0,t}^{\mathrm{TV}}(\boldsymbol{z}) \right], \overline{r} \right\}. \tag{69}$$

It follows that every unvisited action satisfies

$$\widetilde{Q}_{0,t}^{\mathrm{PoF}}(\boldsymbol{z}) \leq \overline{p}_{\varrho}(\overline{r} + \overline{r}) = 2\overline{r}\,\overline{p}_{\varrho}. \tag{70}$$

In contrast, action $\boldsymbol{a}^{\mathrm{b}}$ gives

$$\widetilde{Q}_{0,t}^{\mathrm{PoF}}(\boldsymbol{z}^{\mathrm{b}}) \geq \min\{ p_{0,t}^{\mathrm{PoF}}(\boldsymbol{z}^{\mathrm{b}}) \underline{r}, \overline{r} \} = p_{0,t}^{\mathrm{PoF}}(\boldsymbol{z}^{\mathrm{b}}) \underline{r} \geq \underline{r}\, p_{\mathrm{b}} > 2\overline{r}\,\overline{p}_{\varrho}. \tag{71}$$

Thus at episode $t+1$, `FI-PoF-VI` will also select $\boldsymbol{a}^{\mathrm{b}}$. By induction, $\boldsymbol{a}^{\mathrm{b}}$ is selected at every episode.

Denote $v^{\mathrm{b}} = Q_0^{\star}(\boldsymbol{z}^{\mathrm{b}})$. By construction, the rewards of `FI-PoF-VI` are realizations of the bounded random variable $r_t = \mathcal{J}(\boldsymbol{f}_0(\boldsymbol{z}^{\mathrm{b}}) + \boldsymbol{\epsilon}_0^t)$ with mean v^{b}. In particular,

$$R_T^{\mathrm{PoF}} = \sum_{t=1}^{T} \left[(v^{\star} - v^{\mathrm{b}}) - (r_t - v^{\mathrm{b}}) \right] = \eta^{\mathrm{b}} T - \sum_{t=1}^{T} (r_t - v^{\mathrm{b}}). \tag{72}$$

Taking expectations gives $\mathbb{E}[R_T^{\mathrm{PoF}}] = \eta^{\mathrm{b}} T$. Since $0 \leq r_t \leq \overline{r}$, applying Hoeffding's inequality, we obtain that

$$\mathbb{P}\left(\sum_{t=1}^{T} (r_t - v^{\mathrm{b}}) \geq x \right) \leq \exp\left(-\frac{2x^2}{T\overline{r}^2} \right).$$

Setting $x = \overline{r}\sqrt{(T/2)\log(1/\delta)}$, we have

$$\mathbb{P}\left(R_T^{\mathrm{PoF}} < \eta^{\mathrm{b}} T - \overline{r}\sqrt{\frac{T}{2} \log \frac{1}{\delta}} \right) \leq \delta,$$

which is equivalent to Eq. 65.

Sublinear Regret for `FI-O-VI`. The assumptions satisfy the conditions of Theorem 2 with $H = 1$ and $\mathcal{S}_0 = \{\boldsymbol{s}_0\}$. Since $\mathcal{A}_0$ is finite, $\mathrm{rank}(\boldsymbol{K}_n) \leq |\mathcal{A}_0|$ and $\mathrm{tr}(\boldsymbol{K}_n) = n$. Let $\nu_1, \ldots, \nu_{|\mathcal{A}_0|}$ be the eigenvalues of $\boldsymbol{K}_n$. Using the concavity of the logarithm, we have

$$\begin{aligned} \frac{1}{2} \log\det(\boldsymbol{I} + \lambda^{-1} \boldsymbol{K}_n) &= \frac{1}{2} \sum_{i=1}^{|\mathcal{A}_0|} \log\left(1 + \frac{\nu_i}{\lambda}\right) \\ &\leq \frac{|\mathcal{A}_0|}{2} \log\left(1 + \frac{\sum_{i=1}^{|\mathcal{A}_0|} \nu_i}{\lambda |\mathcal{A}_0|}\right) \\ &= \frac{|\mathcal{A}_0|}{2} \log\left(1 + \frac{n}{\lambda |\mathcal{A}_0|}\right). \end{aligned}$$

Therefore the maximal information gain satisfies

$$\gamma_n \le \frac{|\mathcal{A}_0|}{2}\log\left(1+\frac{n}{\lambda|\mathcal{A}_0|}\right) = O(\log(1+n)). \tag{73}$$

Applying Theorem 2 and substituting $c_0(T) = T^{-1/4}$ yields, with probability at least $1-\delta$,

$$\begin{aligned} R_T^{\mathrm{O}} &= O\left(\frac{\gamma_{T+1}}{c_0(T)^2} + (\beta_{T+1}^f + 1)\sqrt{T\gamma_{T+1}}\right) \\ &= O\left(\sqrt{T}\log(1+T) + \sqrt{\log(1+T)}\sqrt{T\log(1+T)}\right) \\ &= O\left(\sqrt{T}\log(1+T)\right), \end{aligned}$$

which proves Eq. 66.

□

G Benchmark Problems

G.1 Synthetic Problems

We evaluate the methods on eight synthetic cascades, described below. Following Kusakawa et al. (2022), we place the outputs of the stage-transition functions on the common scale $[0,1]$ before adding process noise. For functions not already defined on this scale, we determine a fixed linear transformation from feasible points sampled from the state and action domains. The same transformation is then used throughout the experiment. Process noise may therefore move a realised state outside $[0,1]$. Unless stated otherwise, each stopping constraint has the form $g_h(\boldsymbol{s},\boldsymbol{a}) = q_h(\boldsymbol{s},\boldsymbol{a}) - \tau_h$, where q_h is a fixed sample from a zero-mean GP with unit kernel variance and τ_h controls the feasible region. These functions are fixed across methods and experimental seeds but remain unknown to the algorithms.

Branin–McCormick–Rosenbrock Cascade (BMR, $H=3$, $d_s=d_a=1$) BMR is a three-stage cascade whose transitions are the negated Branin, McCormick, and negated Rosenbrock functions, respectively. At each stage, the state and action are mapped to the function's original domain, and the output is scaled as described above. Actions are restricted to $\mathcal{A}_h = [0.25, 0.75]$ (to ensure that the next state retains a significant dependence on the current state). Each stage has one stopping constraint sampled from a GP with lengthscale 0.25 and threshold $\tau_h = 0.5$. Starting from $s_0 = 0.5$, the terminal reward is

$$\mathcal{J}(s_H) = \exp\left(-\frac{(s_H-0.7633)^2}{2(0.1)^2}\right) + 0.75\exp\left(-\frac{(s_H-0.3887)^2}{2(0.1)^2}\right).$$

SynCompose-1 ($H=3$, $d_s=d_a=2$) SynCompose-1 is a three-stage cascade with two-dimensional states and actions. Each transition maps the four-dimensional input $(\boldsymbol{s}_h, \boldsymbol{a}_h)$ to a two-dimensional next state. The component functions are $(\text{RotHypEllip}, -\text{Zakharov})$ at $h=0$, $(-\text{Rosenbrock}, \text{RotHypEllip})$ at $h=1$, and $(\text{Zakharov}, -\text{DixonPrice})$ at $h=2$. As in BMR, the inputs are mapped to each function's original domain. SynCompose-1 also applies a fixed, component-specific transformation to the four inputs. The outputs are scaled as described above. Every stage has one stopping constraint sampled from a GP with lengthscale 0.3 and threshold $\tau_h = 0.5$. The transition functions use $\boldsymbol{a}_h \in [0,1]^2$. For each stopping constraint, the action input is $\tilde{\boldsymbol{a}}_h = \frac{1}{4}\mathbf{1} + \frac{1}{2}\boldsymbol{a}_h \in [0.25, 0.75]^2$. Starting from $\boldsymbol{s}_0 = (0.5, 0.5)$, the terminal reward is

$$\mathcal{J}(\boldsymbol{s}_H) = \exp\left[-\frac{1}{2}(\boldsymbol{s}_H-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{s}_H-\boldsymbol{\mu})\right],$$

where $\boldsymbol{\mu} = (0.714, 0.456)^\top$, the marginal standard deviations are 0.233 and 0.160, and the correlation is -0.06.

Gaussian-Process Prior Cascade ($\mathcal{GP}$-prior, $H = 3$, $d_s = d_a = 1$) This problem is a three-stage scalar cascade. Each transition is a fixed sample from a zero-mean GP with a Matérn-$5/2$ kernel, unit variance, and lengthscales 0.5 for the state and 1.0 for the action. The longer action lengthscale ensures that the transition retains a significant dependence on the current state. Transition outputs are scaled as described above. At every stage, continuation requires two independently sampled GP constraints to be nonnegative; both have lengthscale 0.25 and threshold $\tau_h = 0.2$. Actions lie in $[0, 1]$, and every episode starts from $s_0 = 0.5$. The terminal reward is $\mathcal{J}(s_H) = \min\{\max\{s_H, 0\}, 1\}$, which equals s_H on $[0, 1]$ and remains bounded when process noise moves the terminal state outside this interval. The optimal trajectory passes close to the constraint boundaries.

Multivariate Normal PDF (MVN, $H = 6$, $d_s = d_a = 1$) Following Nguyen et al. (2016), MVN is a six-stage scalar cascade. The transition at each stage is given by

$$f_h(s_h, a_h) = \exp\left(-\frac{\|(s_h, a_h) - \boldsymbol{\mu}_h\|^2}{2(0.15)^2}\right).$$

The stage centres are $\boldsymbol{\mu}_0 = (0.5, 0.251)$, $\boldsymbol{\mu}_1 = (1, 0.342)$, $\boldsymbol{\mu}_2 = (1, 0.681)$, $\boldsymbol{\mu}_3 = (1, 0.549)$, $\boldsymbol{\mu}_4 = (1, 0.256)$, and $\boldsymbol{\mu}_5 = (1, 0.460)$. Thus, after the first stage, a high terminal reward requires the state to remain near the peak value 1 while the action follows the successive centres. Every stage has one stopping constraint $g_h(s_h, a_h) = q_h(s_h, a_h) + 0.5$, where q_h is a fixed zero-mean, unit-variance Matérn-$5/2$ GP sample with lengthscale 0.5. Actions lie in $[0, 1]$, and every episode starts from $s_0 = 0.5$. The terminal reward is $\mathcal{J}(s_H) = \max\{s_H, 0\}$, which differs from s_H only when process noise makes the terminal state negative.

Gaussian-Mixture PDF (GMM, $H = 6$, $d_s = d_a = 1$). GMM extends MVN by allowing several Gaussian components in each transition:

$$f_h(s_h, a_h) = \sum_{m=1}^{5} w_{h,m} \exp\left(-\frac{\|(s_h, a_h) - \boldsymbol{\mu}_{h,m}\|^2}{2\sigma_{h,m}^2}\right).$$

Two narrow components form two trajectories through all six stages, while a broad component centred at $(0.5, 0.5)$ gives a small response away from them. The remaining two components have zero weight except at stage $h = 2$, where they create two additional peaks. The complete weights, centres, and standard deviations are provided in the released benchmark specification. Every stage has one stopping constraint $g_h(s_h, a_h) = q_h(s_h, a_h) + 0.7$, where q_h is a fixed zero-mean, unit-variance Matérn-$5/2$ GP sample with lengthscale 0.3. Actions lie in $[0, 1]$, and every episode starts from $s_0 = 0.5$. The terminal reward is the same two-component Gaussian mixture used for BMR.

SynCompose-2 ($H = 6$, $d_s = 2$, $d_a = 1$). SynCompose-2 is a six-stage cascade with two-dimensional states and scalar actions. Each transition maps the three-dimensional input $(\boldsymbol{s}_h, a_h)$ to a two-dimensional next state. The component functions are $(\text{RotHypEllip}, \text{Trid})$ at $h = 0$, $(\text{RotHypEllip}, -\text{DixonPrice})$ at $h = 1$, $(-\text{RotHypEllip}, \text{RotHypEllip})$ at $h = 2$, $(-\text{Zakharov}, -\text{Rosenbrock})$ at $h = 3$, $(-\text{SumPow}, -\text{Trid})$ at $h = 4$, and $(-\text{Rosenbrock}, -\text{SumPow})$ at $h = 5$. The inputs and outputs are transformed as in SynCompose-1. Only stages $h = 1, 3, 5$ have a stopping constraint, sampled from a GP with lengthscale 0.3 and threshold $\tau_h = 0.5$. The transition functions use $a_h \in [0, 1]$, while each stopping constraint uses $\tilde{a}_h = \frac{1}{4} + \frac{1}{2}a_h \in [0.25, 0.75]$. Starting from $\boldsymbol{s}_0 = (0.5, 0.5)$, the terminal reward has the same bivariate Gaussian form as in SynCompose-1, with $\boldsymbol{\mu} = (0.386, 0.273)^\top$, marginal standard deviations 0.179 and 0.140, and correlation 0.57.

BMRCRSHMM ($H = 9$, $d_s = d_a = 1$). BMRCRSHMM is a nine-stage scalar cascade. Its first three transitions use the same function sequence as BMR. The transition functions are negated Branin at $h = 0$, McCormick at $h = 1$, negated Rosenbrock at $h = 2$, Six-Hump Camel at $h = 3$, negated Rosenbrock at $h = 4$, Styblinski–Tang at $h = 5$, negated Himmelblau at $h = 6$, negated McCormick at $h = 7$, and McCormick at $h = 8$. At each stage, the state and action are mapped to the function's original domain, and the output is scaled as described above. Actions are restricted to

$\mathcal{A}_h = [0.25, 0.75]$. Every stage has one stopping constraint sampled from a GP with lengthscale 0.25 and threshold $\tau_h = 0.5$. Every episode starts from $s_0 = 0.5$, and the terminal reward is the same two-component Gaussian mixture used for BMR.

SynCompose-3 ($H = 9$, $d_s = 2$, $d_a = 1$). SynCompose-3 extends the construction of SynCompose-2 to nine stages. The component functions are $(\text{DixonPrice}, \text{Zakharov})$ at $h = 0$, $(\text{Sphere}, -\text{SumPow})$ at $h = 1$, $(\text{Rosenbrock}, -\text{Trid})$ at $h = 2$, $(-\text{Rosenbrock}, -\text{SumPow})$ at $h = 3$, $(\text{Rosenbrock}, \text{SumPow})$ at $h = 4$, $(\text{RotHypEllip}, -\text{SumPow})$ at $h = 5$, $(\text{Rosenbrock}, -\text{RotHypEllip})$ at $h = 6$, $(\text{Zakharov}, -\text{RotHypEllip})$ at $h = 7$, and $(-\text{Rosenbrock}, -\text{Trid})$ at $h = 8$. The inputs and outputs are transformed as in SynCompose-1. Only stages $h = 1, 4, 7$ have a stopping constraint, sampled from a GP with lengthscale 0.3 and threshold $\tau_h = 0.5$. The transition functions use $a_h \in [0, 1]$, while each stopping constraint uses $\tilde{a}_h = \frac{1}{4} + \frac{1}{2} a_h \in [0.25, 0.75]$. Starting from $\boldsymbol{s}_0 = (0.5, 0.5)$, the terminal reward has the same bivariate Gaussian form as in SynCompose-1, with $\boldsymbol{\mu} = (0.349, 0.744)^\top$, marginal standard deviations 0.197 and 0.248, and correlation -0.45.

G.2 Real-Life Problems

G.2.1 Droplet Nanocrystal Synthesis

Benchmark Definition. DNS3 and DNS4 model the colloidal atomic layer deposition of a CdS shell on a CdSe nanocrystal core in the droplet reactor of Volk et al. (2023). A droplet receives deposition cycles, each a fixed sequence of five injections (oleylamine, sodium sulfide, formamide, cadmium acetate, oleylamine), and its absorption spectrum is measured in line. One stage is one cycle. DNS4 runs four cycles, which is the full droplet of 20 injections in their volume and time study, and DNS3 runs the first three. On their platform every injection has a volume of 1 to 10 μL and a reaction time of 40 to 400 s, both mapped linearly to $[0, 1]$. The decision $\boldsymbol{a}_h \in [0, 1]^2$ is the volume of the two precursor injections, sodium sulfide and cadmium acetate. The reaction times and the three remaining injections are fixed at the midpoint of their ranges. The state $\boldsymbol{s}_{h+1} = (\text{AP}_{h+1}, \text{PV}_{h+1}, \text{API}_{h+1})$ is the spectrum measured after the cycle of stage $h \in \{0, \ldots, H-1\}$: the wavelength of the first absorption peak, the peak-to-valley ratio, and the peak intensity, placed on the fixed scales $[470, 560]$ nm, $[1, 2.5]$, and $[0, 0.045]$. The initial state is their 480 nm core, $(\text{AP}_0, \text{PV}_0, \text{API}_0) = (478\,\text{nm}, 1.68, 0.041)$.

Transition Model. The transition is the digital twin released with Volk et al. (2023)[7], three ensembles of 200 multilayer perceptrons fitted to their laboratory campaign on the 480 nm core. For every injection they predict the change of AP since the previous injection, PV, and API, from the cycle index, the position within the cycle, the volumes and times of the current and the three previous injections, and, for PV and API, the last measured value. Each target was passed through a Yeo–Johnson transform, and the ensemble prediction is the mean over the members within one standard deviation of the member median. Our re-implementation reproduces the released sampler to numerical precision. Since the memory of the twin spans four injections, the first injections of a cycle also depend on the last three injections of the previous cycle. We fix these three remembered injections at the midpoint of their ranges, which makes a cycle a function of the incoming spectrum and its own decision. On random recipes this changes the wavelength after a cycle by about 1.5 nm (one standard deviation) with a small bias, a fraction of the 2 to 3 nm change per cycle, and leaves the intensity retention used by the constraint practically unchanged. The twin is deterministic. As in the synthetic benchmarks, we add process noise of standard deviation 0.02 and observation noise of 10^{-4} on the scaled state.

Restriction of the Recipe. Their study searches the volume and the time of all five injections. On that ten-dimensional recipe the twin extrapolates. A greedy search under our constraint reaches a peak of 555 nm along recipes with maximal injections in the late cycles, whereas the campaign released with the twin never exceeded 533 nm, and the ensemble members disagree more on those recipes than on random ones. A benchmark on the full recipe would therefore reward exploiting the

[7] https://github.com/AbolhasaniLab/AlphaFlow.

twin outside its data. With the two precursor volumes the same search reaches 537 nm, close to the 538 nm of the droplet executed at the end of their laboratory campaign.

Inspections and Reward. A cycle is executed only when its margin is nonnegative. A violation loses the droplet, so neither the spectrum after that cycle nor any later cycle is observed, which is the pre-transition setting of Section 2.1. The margin of stage h is

$$g_h(\boldsymbol{s}_h, \boldsymbol{a}_h) = 2\min\left\{\log\frac{\mathrm{API}_{h+1}}{\mathrm{API}_h} - \log\tau_h,\ \frac{\mathrm{PV}_{h+1} - 1.2}{0.5},\ \frac{\mathrm{AP}_{h+1} - 450}{50}\right\},$$

where $(\mathrm{AP}_{h+1}, \mathrm{PV}_{h+1}, \mathrm{API}_{h+1})$ is the prediction of the twin for the cycle. The first term is a material budget. The peak intensity is the concentration proxy behind their viability rule that the absorbance at 350 nm stay above 0.03, and the floors $(\tau_0, \ldots, \tau_3) = (0.44, 0.63, 0.63, 0.73)$ are the tenth percentile of the retention of random recipes at each cycle. They differ by cycle because the first cycle removes about half of the intensity for every recipe. The second term is a continuous version of their rule that a first absorption peak be detectable, and the third is the 450 nm floor of their sampler. Under these floors 70% of random recipes complete four cycles and 73% complete three. Their binary viability label also covers the loss of a droplet at phase separation, for which no continuous measurement was released, and it is not used. The terminal reward is the red shift of the peak,

$$J(\boldsymbol{s}_H) = \frac{\mathrm{AP}_H - 478}{70}.$$

Reference Reward. The reference reward is the best recipe found by a random search followed by greedy local search on the noise-free cascade, which gives 0.716 on DNS3 and 0.838 on DNS4. These are reference values rather than known optima.

Runs and Results. Each seed starts from eight completed recipes drawn from fixed pools of about 2,100 recipes that survived a screen of 3,000 uniformly sampled ones. The deployed policy is evaluated every 10 episodes on 200 rollouts. Table 3 reports the final results.

Table 3: Median average evaluation reward and number of droplets lost after 200 episodes over 20 seeds. Figure 3 shows the interquartile bands for reward.

	DNS3 (H=3)		DNS4 (H=4)	
method	evaluation reward	lost	evaluation reward	lost
RF	0.639	11.5	0.696	18
FB-CUCB	0.671	13.5	0.719	44
BOFN-CEI	0.667	68	0.683	80.5
Cascade-CEI	0.656	24.5	0.730	33
FI-PoF-VI	**0.689**	**3**	**0.790**	**7**
FI-O-VI	0.688	44	0.778	72

G.2.2 Recipe Search for the Heat Treatment of Steel

Benchmark Definition. QT and NQT model a deterministic heat-treatment sequence for a representative chromium–molybdenum low-alloy steel. We fix its composition to 0.40C, 0.85Mn, 0.25Si, 0.95Cr, and 0.20Mo in weight percent, corresponding to representative values within the SAE 4140 composition ranges. The simulated part is 9 mm thick. QT contains austenitising, quenching, and tempering. NQT adds normalising before these three operations. Table 4 gives the recipe ranges, states, and post-operation inspections. Recipe coordinates are mapped to $[0, 1]$. The cooling rate is mapped logarithmically, while the other coordinates are mapped linearly. Measured states are placed on fixed scales: grain-size number on $[2, 12]$, phase fractions and toughness on $[0, 1]$, as-quenched hardness on $[150, 700]$ HV, and tempered hardness on $[100, 700]$ HV.

Transition Models. Let T_C denote the temperature of an operation in degrees Celsius and T_K the same temperature in kelvin, and let $f_{\mathrm{aus}} = \min\{\max\{(T_C - Ae_1)/(Ae_3 - Ae_1), 0\}, 1\}$ be the austenitised fraction, where Ae_1 and Ae_3 follow Andrews (1965). We use $d_L =$

$76671\exp[-89098/(RT_K)]t^{0.211}\,\mu$m, adapted from the grain-growth model of Lee & Lee (2008). QT sets the austenite grain diameter to $\max\{d_L, 3\,\mu\text{m}\}$. For the normalising stage of NQT, the starting diameter is $d_0 = \max\{0.5 f_{\text{aus}} d_L + (1 - f_{\text{aus}})60, 3\}\,\mu$m. The following austenitising stage grows this grain according to $d^p = d_0^p + [76671\exp(-89098/(RT_K))]^p t$, where $p = 1/0.211$. Diameters are converted to the grain-size number $G = 2\log_2(254/d) + 1$, where a larger G denotes a finer grain.

Quenching and Tempering. At quenching, G is rounded to the nearest half-integer and supplied with the cooling rate to the continuous-cooling-transformation model of Collins et al. (2023). Let x_M, x_B, and x_{FP} be its martensite, bainite, and ferrite–pearlite fractions, and let H_M, H_B, and H_{FP} be their hardnesses from Maynier et al. (1977). The quench produces

$$f_M = f_{\text{aus}} x_M, \qquad H_q = f_{\text{aus}} \frac{x_M H_M + x_B H_B + x_{FP} H_{FP}}{x_M + x_B + x_{FP}} + (1 - f_{\text{aus}})200.$$

The next state is (f_M, H_q, G). Tempering uses the model of Kang & Lee (2014),

$$\text{TP} = T_K \left(\log_{10} t + k_0 + \sum_i k_i X_i \right), \qquad H_{TM} = \left(1542.97 - \frac{25.31}{X_C} \right) \exp(-1.23 \times 10^{-4} \text{TP}).$$

The final hardness combines H_{TM} with a softened non-martensitic contribution. With T_C the tempering temperature, the toughness index is

$$K_{\text{tough}} = \left[1 - 0.5 \exp\left(- \left(\frac{T_C - 320}{60} \right)^2 \right) \right] \left[0.6 + 0.4 \min\left\{ \max\left\{ \frac{G - 4}{6}, 0 \right\}, 1 \right\} \right] \times \left[0.5 + 0.5 \min\{\max\{f_M, 0\}, 1\} \right].$$

The first factor represents tempered-martensite embrittlement near 320 °C (Materkowski & Krauss, 1979). The other two account for grain refinement and the martensite fraction. The complete expression for the final-hardness blend is given in the released simulator.

Inspections and Reward. A stage continues when its scaled inspection margin is nonnegative. NQT uses $(G_0 - 7)/5$ after normalising, and both chains use $(G - 4)/5$ after austenitising. After quenching, the margin is

$$\min\left\{ \frac{H_q - 380}{200}, \log_{10}\left(\frac{r_{\text{crit}}}{r \max\{f_M, 0.05\}} \right) \right\}, \qquad r_{\text{crit}} = 2837(9)^{-2.2}\ ^\circ\text{C/s},$$

where r is the cooling rate. The tempering margin is $(100 - |H_{\text{fin}} - 330|)/200$. The terminal reward is

$$J(\boldsymbol{s}_H) = \max\left\{ 1 - \frac{|H_{\text{fin}} - 330|}{100}, 0 \right\} K_{\text{tough}}.$$

The process noise is zero. Inspections occur after each operation, and a failed inspection prevents all later operations and the terminal reward from being observed. For learning, each constraint is represented as the deterministic composite of the operation and its inspection.

Each seed starts from eight completed recipes drawn from fixed pools obtained by screening 3,000 uniformly sampled recipes, of which about a third survive. Table 5 reports the final results.

Table 4: Stages of the two chains. NQT is QT with the normalising operation in front; the austenitising grain of NQT grows from the normalised grain (Sellars form of the growth law, $d^p = d_0^p + [A\exp(-Q/RT)]^p t$ with $p = 1/0.211$), that of QT from a fresh structure.

operation	recipe $\boldsymbol{a}_h$ (range)	measured state $\boldsymbol{s}_{h+1}$	inspection
normalise (NQT only)	T_C (780 to 980 °C), t (600 to 3600 s)	grain size G_0	$G_0 \geq 7$
austenitise	T_C (780 to 1000 °C), t (600 to 7200 s)	grain size G, austenitised fraction f_γ	$G \geq 4$
quench	cooling rate (0.5 to 100 °C/s, log-uniform)	martensite f_M, hardness H_q, G	$H_q \geq 380$ HV, no crack
temper	T_C (200 to 680 °C), t (1800 to 14400 s)	hardness H_{fin}, toughness index K_{tough}	$\lvert H_{\text{fin}} - 330\rvert \leq 100$ HV

Table 5: Median best reward and number of parts scrapped after 200 episodes over 20 seeds. Figure 4 shows the interquartile bands for reward.

	QT (H=3)		NQT (H=4)	
method	best reward	scrapped	best reward	scrapped
RF	0.816	9.5	0.810	11.5
FB-CUCB	0.946	**3**	0.932	**4**
BOFN-CEI	0.922	78	0.890	45
Cascade-CEI	0.938	34	0.927	32.5
FI-PoF-VI	0.945	13	**0.943**	17
FI-O-VI	**0.953**	20	0.939	57

H Experimental Details

H.1 Model Selection

Each scalar component of a stage transition or constraint is modelled by an independent GP with a Matérn-$5/2$ kernel and automatic relevance determination. We re-estimate the kernel hyperparameters after every episode by maximum a posteriori (MAP) estimation. For the transition model at stage h, the likelihood variance is fixed to the known process noise variance σ_h^2, and a constant prior mean is estimated jointly with the kernel hyperparameters. On the noise-free heat-treatment problems, we instead fix the likelihood variance to 10^{-4} for numerical stability, as we do for every constraint model. Following the implementation of Kusakawa et al. (2022), we restrict the kernel variance to $[\sigma_0^2/2,\ 2\sigma_0^2]$ and each lengthscale to $[1/20,\ 1/2]$ of the width of its input coordinate. Here σ_0 is set from the declared output scale and equals 1 after the synthetic outputs are scaled as described in Section G.1. Within these intervals, we use a uniform prior on the kernel variance and the dimension-scaled log-normal lengthscale hyperprior of Hvarfner et al. (2024),[8] The constraint prior mean is $-\tau_h$ on the synthetic problems and the mean initial-design margin on the heat-treatment problems. All remaining constraint hyperparameters are estimated by MAP under the same bounds, and all methods use the same fitted models.

Value Model. We approximate each value function $\tilde{V}_h$ by KRR with an automatic-relevance-determination RBF kernel. Inputs are scaled coordinatewise to $[0, 1]$ over the current support set, and targets are standardised. For each stage, the kernel variance, lengthscales and ridge parameter are selected by leave-one-out cross-validation (Williams & Rasmussen, 2006). We compare 20 isotropic settings with $\ell_v \in [0.1\bar{d}, 3\bar{d}]$, where $\bar{d}$ is the median pairwise distance between support points, $\sigma_v^2 \in [0.1, 5]$, and $\lambda_v \in [10^{-6}, 10^{-1}]$. We then refine the individual lengthscales for at most two sweeps using multiplicative factors $\{1/4, 1/2, 2, 4\}$.

For intermediate stages with pre-transition constraints, the expectation of the fitted value function is available in closed form (Section D.1.1). Otherwise, we use randomised Sobol points that remain fixed within each planning round: 128 for the synthetic problems and 32 for the heat-treatment problems. In the censored pre-transition setting, we also replace negative approximate backups by zero before maximising over actions, consistent with the nonnegativity of the exact recursion.

H.2 Comparison Baselines and Parameter Settings

We detail our extensions of the existing approaches below. For post-transition constraints, each baseline learns whether an action taken from the current state will lead to a feasible next state. BOFN-CEI, Cascade-CEI, and Random Feasible use these predicted probabilities when selecting actions. Following its original formulation, FB-CUCB instead models the minimum constraint margin along the entire trajectory.

[8] We implement the algorithm in BoTorch (Balandat et al., 2020) where this hyperprior is the default. The algorithm is however library agnostic and can be implemented as well on either `TensorFlow` (e.g. Picheny et al., 2023) and `Jax`.

FI-O-VI and FI-PoF-VI. Both methods use the transition bonus of Eq. 17 with $\beta_f = 1$ and the KRR value model of Section H.1. At each stage, the support set of Section D.1.2 contains 40 Sobol points, 10 value-weighted DPP points, and the observed states. `FI-O-VI` uses the optimistic transition feasibility $p^+_{h,t}$ of Eq. 4 with $\beta_g = 1$ and the schedule of Eq. 15, setting $\xi_0 = 0.02$, $\zeta = 2$, $t_{\mathrm{tr}} = \lfloor T/4 \rfloor$, and $\alpha = \frac{1}{2}\bar{\alpha}_\kappa$ for $\nu = 5/2$. The scheduled feasibility threshold is applied at the stage being executed, while later stages in the lookahead use the fixed threshold 0.01. We use Algorithm 1 for pre-transition constraints and the modification in Section E for post-transition constraints. If no candidate action has a positive transition feasibility under the scheduled threshold, the candidates are re-scored under the fixed threshold 0.01, and failing that `FI-O-VI` executes the action with the largest worst-constraint lower confidence bound. `FI-PoF-VI` replaces $p^+_{h,t}$ by the plug-in probability $\prod_j \Phi(\mu^g_{h,j}/\sigma^g_{h,j})$ and otherwise uses the same transition bonus and value model. The common action search screens 500 points sampled uniformly from $[0.01, 0.99]^{d_a}$ and refines the best point with up to 20 L-BFGS steps. For post-transition constrained execution, `FI-O-VI` uses SLSQP with the same 20-step limit. These settings match the optimisation budget used for the baselines.

FB-CUCB. FB-CUCB treats the complete action sequence $\boldsymbol{x} = (\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{H-1}) \in \prod_{h=0}^{H-1} \mathcal{A}_h$ as a single input and does not model the intermediate states. It fits two independent GPs on this space, both with a Matérn-5/2 kernel with automatic relevance determination, normalised inputs, standardised outputs, and marginal-likelihood fitting. The reward GP is trained only on episodes that reach the terminal reward $J(\boldsymbol{s}_H)$. The constraint GP is trained on every episode using the smallest margin observed before the episode ends, which is negative exactly when the episode is stopped.

The acquisition is an upper confidence bound weighted by the probability of feasibility, $\alpha_\beta(\boldsymbol{x}) = \big(\mu_J(\boldsymbol{x}) + \beta\sigma_J(\boldsymbol{x})\big)\Phi(\mu_g(\boldsymbol{x})/\sigma_g(\boldsymbol{x}))$ (Srinivas et al., 2009; Gardner et al., 2014). We set $\beta = 1$ during optimisation and $\beta = 0$ for evaluation. Because the method has no state-transition model, it selects the whole sequence at the beginning of an episode and executes it open loop. It uses the shared search budget of 500 screened sequences followed by up to 20 L-BFGS steps.

BOFN-CEI. BOFN-CEI extends the function-network formulation of Astudillo & Frazier (2021) to the constrained cascade, with one GP for each transition and constraint. At each episode, $M = 32$ randomised Sobol points are drawn and held fixed during optimisation. For an action sequence $(\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{H-1})$, point m generates the trajectory

$$\boldsymbol{s}^{(m)}_{h+1} = \boldsymbol{\mu}^f_h(\boldsymbol{s}^{(m)}_h, \boldsymbol{a}_h) + \sqrt{\big(\boldsymbol{\sigma}^f_h(\boldsymbol{s}^{(m)}_h, \boldsymbol{a}_h)\big)^2 + \sigma^2_h}\, \boldsymbol{z}^{(m)}_{h+1}, \qquad \boldsymbol{s}^{(m)}_0 = \boldsymbol{s}_0,$$

which includes both uncertainty in the transition model and the known process noise. The constraint posterior gives the probability of feasibility at each simulated stage. The acquisition targets the nested expectation

$$\alpha(\boldsymbol{a}_{0:H-1}) = \mathbb{E}_f\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left[\big(J(\boldsymbol{s}_H) - J^\star(f)\big)^+ \prod_{h=0}^{H-1} \prod_{j\in[m_h]} \Phi\left(\frac{\mu^g_{h,j}(\boldsymbol{s}_h, \boldsymbol{a}_h)}{\sigma^g_{h,j}(\boldsymbol{s}_h, \boldsymbol{a}_h)}\right)\right]\right],$$
$$J^\star(f) = \max_{t\in\mathcal{C}} \mathbb{E}_{\boldsymbol{\epsilon}}\Big[J(\boldsymbol{s}_H) \mid f, \boldsymbol{a}^{(t)}\Big],$$

where $\mathcal{C}$ contains the completed episodes. We estimate this acquisition using the 32 sampled paths. For simplicity, each transition-model sample is paired with one Monte Carlo sample of the process noise rather than an inner average over several samples. The same paired sample is used for the candidate and the completed sequences when estimating $J^\star(f)$, following noisy expected improvement (Letham et al., 2019). BOFN-CEI optimises all H actions jointly and executes the resulting sequence open loop. For evaluation, it drops the improvement term and maximises the feasibility-weighted terminal reward. It uses the shared budget of 500 screened sequences followed by up to 20 L-BFGS steps.

Cascade-CEI. CASCADE-CEI follows Kusakawa et al. (2022), extended with the stopping constraints. It uses the same constrained expected-improvement acquisition and the same $M = 32$ approximation as BOFN-CEI. The difference is that CASCADE-CEI operates in closed loop: after observing the state at stage h, it re-optimises the remaining actions and executes only the first one.

The same closed-loop rule is used for evaluation. Because $\mathcal{J}(\boldsymbol{s}_H)$ is a nonlinear function of the terminal state rather than the final GP output, we sample the final stage together with the preceding stages instead of using the closed-form final-stage expected improvement of the original method.

Random Feasible. At stage h, Random Feasible draws $N = 500$ candidate action sequences for the remaining stages and propagates them from the observed state using the transition posterior means. It accepts candidates whose predicted probability of completing the remaining chain, $\prod_{k=h}^{H-1} \prod_{j \in [m_k]} \Phi(\mu^g_{k,j}/\sigma^g_{k,j})$, is at least $1/2$. One accepted candidate is selected uniformly at random, its first action is executed, and the procedure is repeated at the next stage. If none qualifies, the candidate with the largest completion probability is used. For evaluation, the method uses the shared closed-loop recommender that maximises the feasibility-weighted predicted terminal reward.

H.3 Performance Evaluation

On the synthetic problems, we evaluate the policy recommended by each method after the shared initial design and every 5 optimisation episodes. At each evaluation, the models learned up to that point are held fixed and the policy's expected return is estimated from 200 fresh rollouts. A rollout that violates any constraint receives zero reward. All methods are evaluated using the same process-noise realisations, and the evaluation rollouts are not added to the training data.

For evaluation, each method uses its corresponding planning procedure. `FI-O-VI` and `FI-PoF-VI` refit their value models without the exploration bonus and use the posterior probability of feasibility. CASCADE-CEI and RANDOM-FEASIBLE replan after each observed transition to maximise the feasibility-weighted expected terminal reward. BOFN-CEI maximises the same quantity once over all H actions, following its open-loop formulation. FB-CUCB also selects a complete action sequence, using the posterior mean of its terminal reward model weighted by the probability of feasibility.

The two heat-treatment problems are deterministic, so repeated evaluation rollouts are unnecessary. For these problems, we report the highest terminal reward among the completed recipes after each episode.

H.4 Additional Experimental Results

H.4.1 Cumulative Abort Count

Figure 6 shows how aborted episodes accumulate during the synthetic experiments reported in Figure 2. The abort rate of `FI-O-VI` decreases with the episode budget on every problem. `FI-O-VI` aborts fewer episodes than BOFN-CEI and CASCADE-CEI on every problem while attaining higher rewards, whereas `FI-PoF-VI` and RANDOM-FEASIBLE abort less often than `FI-O-VI` but achieve lower rewards on most problems.

H.4.2 Final Evaluation Reward

Figure 7 shows the distribution of the final evaluation reward over the 20 repetitions for each method. Its medians are the endpoints of the curves in Figure 2, and the spread shows on how many seeds a method ends in a poor mode, as `FI-PoF-VI` does on most seeds of GMM.

H.4.3 Empirical Results Illustration

Figure 8 shows where `FI-O-VI` spends its evaluations on the five instances with a scalar state, for which the stage landscape can be drawn on the (s, a) plane. SynCompose-1, SynCompose-2 and SynCompose-3 have a two-dimensional state and are therefore not shown.

Figure 9 compares the decisions made by `FI-O-VI` and `FI-PoF-VI` at the final stage of GMM as optimisation proceeds.

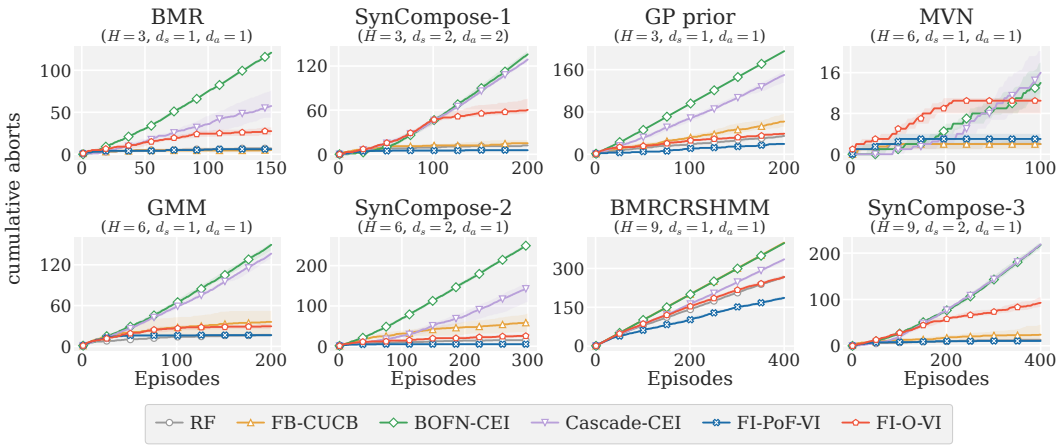


Figure 6: Cumulative aborted episodes on the eight synthetic cascades of Figure 2, in the same panel order (median and interquartile band).

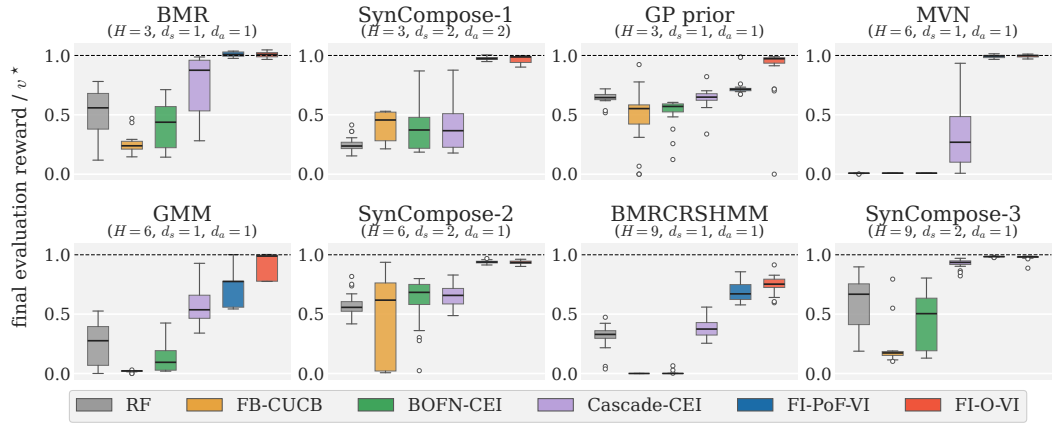


Figure 7: Final evaluation reward of the recommended policy as a fraction of the optimal value $v^\star$, on the eight synthetic cascades of Figure 2 in the same panel order. Each box holds 20 points, one per seed: the evaluation reward at the last episode, averaged over its 200 rollouts.

H.4.4 Sensitivity to the Process Noise

Figure 10 reports, for every method, the final evaluation reward as a fraction of the optimal value $v^\star(\sigma_h)$ at the noise level in force, together with the cumulative aborts, on BMR and GMM for $\sigma_h \in \{0, 0.01, 0.02, 0.05, 0.1\}$ over 10 seeds; the 0.02 level is the main campaign. The optimal value itself falls with the noise, from 0.99 to 0.45 on GMM. `FI-O-VI` stays at the optimum without process noise, the regime the regret analysis does not cover, remains within 0.2 of it up to $\sigma_h = 0.05$, and falls further at the largest level. Its abort count grows with the noise because it continues to reach the later stages, whereas the abort counts of the forward planners fall because they stop passing the early ones.

H.4.5 Sensitivity Studies

Every arm below is `FI-O-VI` with one setting of the feasibility-threshold schedule changed from the values of Section H.1, run on the eight synthetic instances for 10 seeds each. The reference arm is the published setting, $\xi_0 = 2\text{e}{-2}$ and $t_{\text{tr}} = T/4$; the last arm replaces the schedule by a constant threshold $\xi = 2\text{e}{-2}$ at every episode, so it isolates the annealing from the rest of the algorithm.

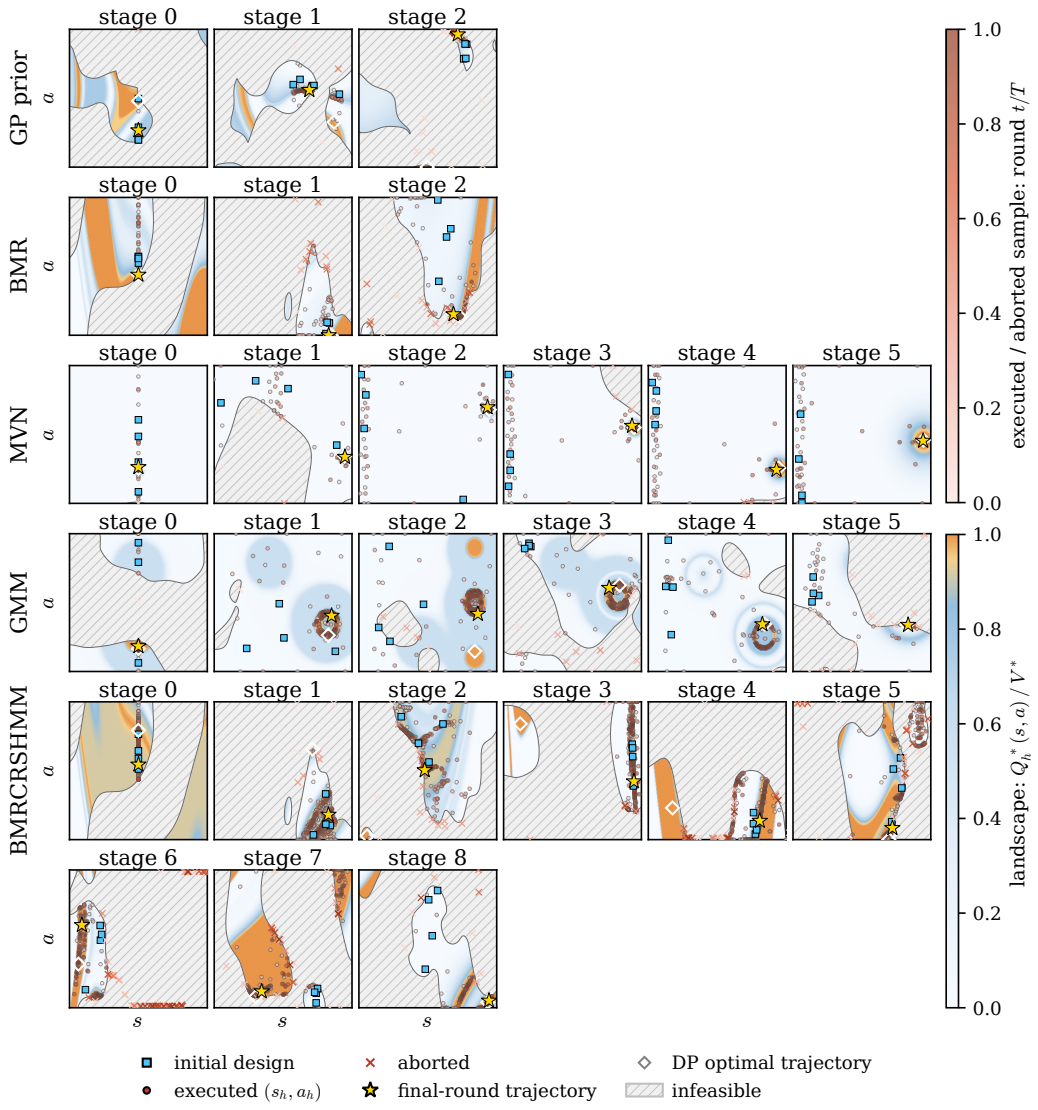


Figure 8: State–action pairs selected by FI-O-VI in one repetition on the five scalar-state problems. The shading shows the normalised optimal value $Q_h^\star(s,a)/v^\star$, and the hatched region is infeasible. The markers show the initial design, the selected pairs coloured by episode, aborted stages, the final trajectory, and the optimal trajectory. As optimisation proceeds, the selected pairs concentrate in feasible, high-value regions, with most aborts occurring near the constraint boundary.

I Runtime Analysis

Compute Setup. All experiments were run on an internal SLURM cluster without GPUs; every run, for every method, was allocated 4 CPU cores and 16 GB of RAM.

Per-Episode Runtime. Table 6 reports the wall-clock time per optimisation episode of every method on the eight synthetic instances, averaged over the 20 seeds, split into the refitting of the stage GPs, which is common to all methods, and the method-specific planning cost; for the two FI methods the latter is further split into value-model fitting and action selection. Stage-model fitting is the largest cost on the three problems with a two-dimensional state, while the additional value-model fitting and action-selection costs remain moderate across the experiments.

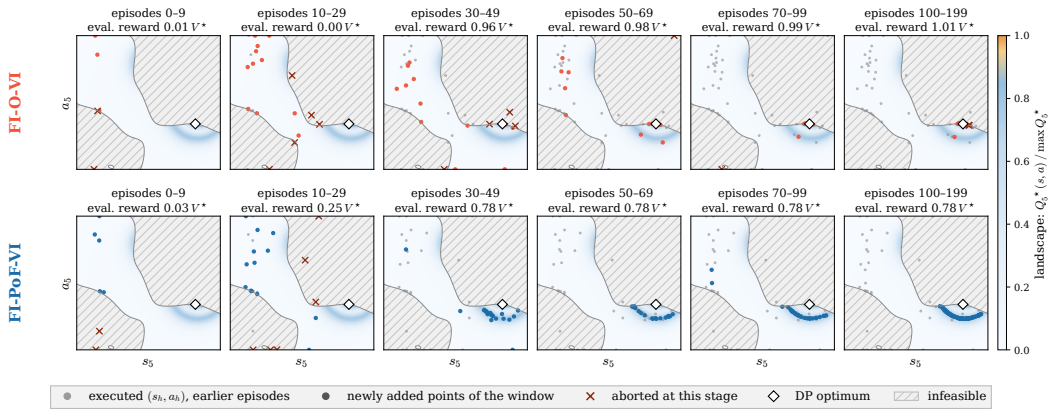


Figure 9: Decisions selected by `FI-O-VI` and `FI-PoF-VI` at the final stage of GMM in one of the 20 repetitions. The shading shows the true $Q_5^\star(s,a)$, and the hatched region is infeasible. Coloured points mark decisions from the indicated episodes, grey points show earlier decisions, crosses denote aborts at this stage, and the diamond marks the optimum. Each title gives the evaluation reward at the end of the corresponding interval.

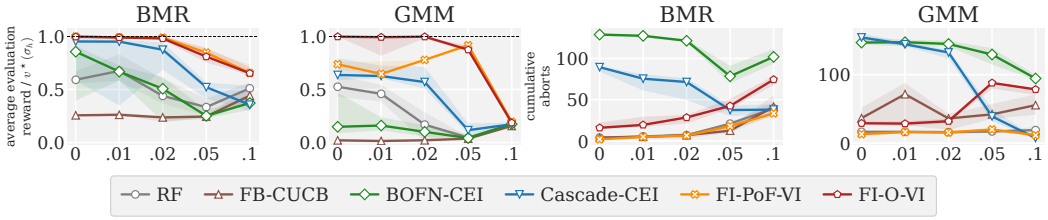


Figure 10: Sensitivity to the process noise σ_h on BMR and GMM: final evaluation reward as a fraction of the optimal value $v^\star(\sigma_h)$ at that noise level (left pair) and cumulative aborted episodes (right pair); median over 10 seeds with the interquartile band.

Table 6: Wall-clock time per optimisation episode in seconds (mean ± standard deviation across the 20 seeds, evaluation rollouts excluded).

	H	d_s	d_a	Stage-model fit	Planning (baselines)				FI-PoF-VI		FI-O-VI	
				(GPs)	RF	FB-CUCB	BOFN-CEI	CASCADE-CEI	Value model fit	Planning	Value model fit	Planning
GP prior	3	1	1	7.6±1.5	0.11±0.08	0.73±0.09	0.26±0.07	0.62±0.13	1.7±0.1	0.40±0.05	1.5±0.2	0.69±0.40
BMR	3	1	1	0.64±0.12	0.07±0.01	0.76±0.16	0.39±0.07	0.83±0.11	1.2±0.2	0.28±0.06	1.4±0.3	1.6±0.3
SynCompose-1	3	2	2	7.2±3.0	0.09±0.03	2.9±4.9	0.33±0.12	0.85±0.81	2.1±0.4	0.25±0.04	2.0±0.4	0.89±0.18
MVN	6	1	1	1.1±0.2	0.23±0.04	1.5±0.3	0.42±0.08	2.0±3.9	1.8±0.4	1.2±0.3	1.9±0.4	1.4±0.3
GMM	6	1	1	1.3±0.3	0.17±0.04	1.7±0.4	0.46±0.15	1.6±0.3	3.7±0.8	0.70±0.18	4.0±1.0	2.4±1.0
SynCompose-2	6	2	1	14±6	0.23±0.11	2.4±3.3	0.43±0.12	2.0±1.1	9.5±4.4	0.66±0.34	10±7	1.3±0.6
BMRCRSHMM	9	1	1	2.4±0.7	0.22±0.04	3.0±1.1	0.81±0.21	3.3±0.7	12±3	1.6±0.3	12±3	5.0±1.0
SynCompose-3	9	2	1	22±5	0.46±0.04	3.9±1.1	0.97±0.33	4.5±1.1	21±4	1.2±0.3	17±5	2.5±0.6

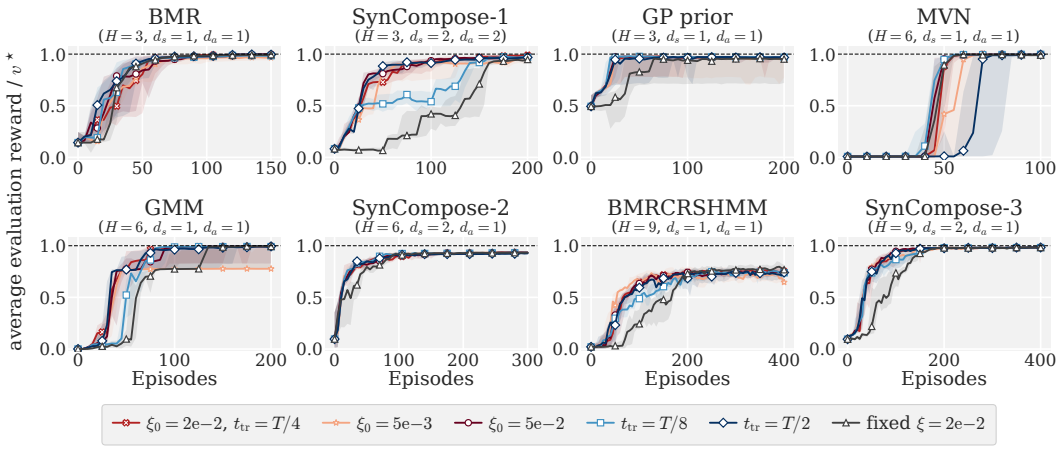


Figure 11: Evaluation reward under the feasibility-threshold schedules, as a fraction of $v^\star$, in the panel order of Figure 2 (median over 10 seeds, interquartile band).

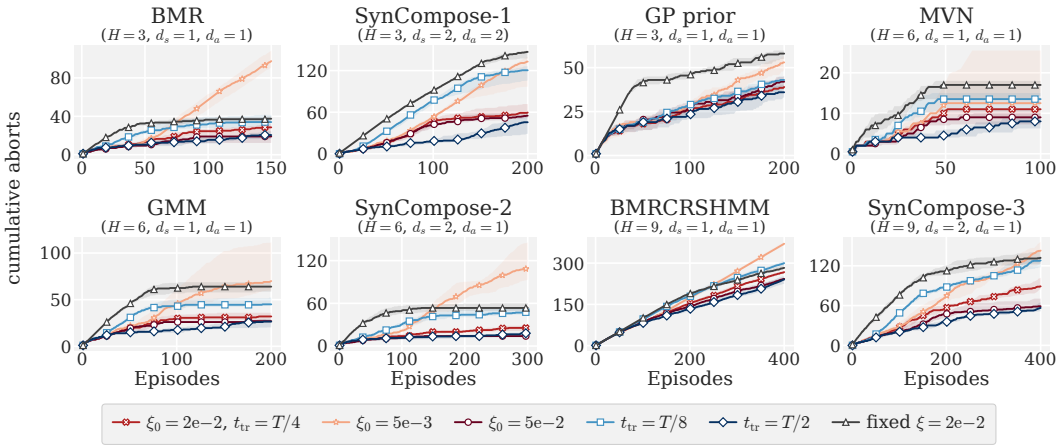


Figure 12: Cumulative aborted episodes of the same arms in the same panel order.