The Sample Complexity of Online Reinforcement Learning: A Multi-model Perspective

Michael Muehlebach
Max Planck Institute for Intelligent Systems
Tuebingen, Germany
michaelm@tuebingen.mpg.de

Zhiyu He
Max Planck Institute for Intelligent Systems
Tuebingen, Germany
zhiyu.he@tuebingen.mpg.de

Michael I. Jordan
University of California, Berkeley
Berkeley, USA
jordan@cs.berkeley.edu

Abstract

We study the sample complexity of online reinforcement learning in the general setting of nonlinear dynamical systems with continuous state and action spaces. Our analysis accommodates a large class of dynamical systems ranging from a finite set of nonlinear candidate models to models with bounded and Lipschitz continuous dynamics, to systems that are parametrized by a compact and real-valued set of parameters. In the most general setting, our algorithm achieves a policy regret of $\mathcal{O}(N\epsilon^2 + \ln(m(\epsilon))/\epsilon^2)$, where N is the time horizon, ϵ is a user-specified discretization width, and $m(\epsilon)$ measures the complexity of the function class under consideration via its packing number. In the special case where the dynamics are parametrized by a compact and real-valued set of parameters (such as neural networks, transformers, etc.), we prove a policy regret of $\mathcal{O}(\sqrt{Np})$, where p denotes the number of parameters, recovering earlier sample-complexity results that were derived for *linear time-invariant* dynamical systems. While this article focuses on characterizing sample complexity, the proposed algorithms are likely to be useful in practice, due to their simplicity, their ability to incorporate prior knowledge, and their benign transient behaviors.

1 Introduction

Reinforcement learning describes the situation where a decision-maker chooses actions to control a dynamical system, which is unknown a priori, to optimize a performance measure. At the core of reinforcement learning is the fundamental dilemma between choosing actions that reveal information about the dynamics and choosing actions that optimize performance. These are typically conflicting goals. We consider an online non-episodic setting, where the decision-maker is required to learn continuously and is unable to reset the state of the dynamical system. This further introduces the challenge that the information received by the learner is correlated over time and hence, standard statistical tools cannot be applied directly. Despite these important challenges, we provide a suite of online reinforcement learning algorithms that are relatively straightforward to analyze, while being both practically and theoretically relevant. The algorithms sample from a posterior over the different potential model candidates (or an approximation thereof), apply the corresponding "certainty-equivalent" policies, while carefully introducing enough excitation to ensure that the posterior distribution over models converges sufficiently rapidly.

Preprint. Under review.

We consider three different settings. In the first setting, the decision-maker has access to a finite set of nonlinear candidate models that potentially describe the system dynamics (continuous state and action spaces). This setting is relevant for many practical engineering applications, where the choice of candidate models provides a natural way to incorporate prior knowledge. In this setting our online algorithm achieves a sample complexity of $\mathcal{O}(\ln(N) + \ln(m))$ in terms of policy regret, where N denotes the time horizon and m the number of candidate models. In the second setting, we allow for any class of dynamical system, where the dynamics are given by a bounded set in a normed vector space. This includes, for example, all bounded Lipschitz continuous functions with the supremum norm, or a bounded set of square integrable functions. By applying packing and covering arguments, we can relate the second setting to the first one and derive corresponding policy-regret guarantees that take the form $\mathcal{O}(N\epsilon^2 + \ln(m(\epsilon))/\epsilon^2)$, where ϵ describes the discretization width and $m(\epsilon)$ the packing number, which measures the complexity of the function class [50]. In the third setting, we consider systems that are parametrized by a compact and real-valued set of parameters. This includes the situation where the dynamics are parametrized by neural networks, transformers, or other parametric function approximators, and we obtain a policy regret of $\mathcal{O}(\sqrt{Np})$, where p describes the number of parameters. We further note that in the common situation where our function class is given by a linear combination of nonlinear feature vectors, which also encompasses linear dynamics as a special case, our algorithm is straightforward to implement, as it only requires sampling from a (truncated) Gaussian distribution at every iteration.

Our main contributions are summarized as follows:

- We provide a suite of algorithms with nonasymptotic regret guarantees for online reinforcement learning over continuous state and action spaces with nonlinear dynamics. Numerical results highlight that transients are benign and that our algorithms are likely to be useful in practice.
- Compared to earlier work in the machine learning community [see, e.g., 19, 46], which mainly focused on linear dynamical systems and relied on two-step learning strategies that alternate between least-squares estimation and optimal control design, our work accommodates a much broader class of systems and results in a single-step procedure that unifies control design and identification. Moreover, our analysis is straightforward and recovers the results from earlier work as simple corollaries specialized to linear dynamics.
- Compared to earlier work in the adaptive control community [see, e.g. 5, 25], which focuses on asymptotic stability, boundedness, and deterministic dynamical systems, we consider stochastic dynamical systems and characterize *nonasymptotic performance*. While boundedness (almost surely) cannot be guaranteed in our stochastic setting, where the process noise, for example, has unbounded support, we provide a nonasymptotic bound on the second moment of state trajectories and show that our estimation converges in finite time almost surely. This can be viewed as the stochastic analogue of boundedness and asymptotic convergence.
- Our analysis sheds light on the difference in sample complexity between model-based and model-free reinforcement learning.[1] In the model-based setting, as considered herein, a single iteration provides information about the accuracy of each candidate model, resulting in a regret that scales with $\mathcal{O}(\ln(m))$ in the presence of a finite set of models. In contrast, in the model-free setting, a single iteration provides only information about the feedback policy that is currently applied, resulting in a regret that scales with $\mathcal{O}(m)$ or worse [see, e.g., 33, 22].
- The work provides a powerful *separation* principle that applies to nonlinear dynamics. Indeed, as we will see (e.g., in the proof of Thm. 3.2), our algorithms are based on identifying the best model in the given class and applying a *certainty-equivalent* policy, separating the model identification and the optimal control task. In particular, our algorithms combine optimal certainty-equivalent control with optimal model identification based on the posterior distribution over models. This provides an algorithmic paradigm that contrasts with prevalent approaches based on optimism in the presence of uncertainty and it facilitates policy-regret characterization for systems with continuous states and inputs.

The decision-making problem considered here is central to machine learning and related disciplines, and there has been a great deal of prior work. We provide a short review of recent prior work that is closely related to our approach in the following paragraph; a more detailed review of the literature can be found in App. A.

[1] By model-free reinforcement learning we mean a setting where the decision-maker has only access to a set of feedback policies. The terminology is arguably ill-defined.

Online continuous control is an important benchmark problem in reinforcement learning with continuous states and actions, and is fundamentally harder than the tabular setting with discrete states and actions. Building on the fruitful contributions of learning for control in the linear world [24, 48], the frontier now becomes online control with nonlinear dynamics. Prior work [see, e.g., 29, 11, 37] has shown that sublinear $\mathcal{O}(\sqrt{N})$ regret can be achieved under structural assumptions on the dynamics, e.g., contraction or linear representations of nonlinear features, which are stronger than what is assumed herein. Our approach instead builds on the use of multiple candidate models and aggregation across these models, a situation reminiscent of *multi-model adaptive control*, where the goal is asymptotic stabilization [6]. In contrast, we tackle online reinforcement learning and provide nonasymptotic policy-regret characterizations. Compared to recent work in online switching control [35, 32], we handle a broad class of nonlinear dynamics and achieve a favorable logarithmic regret with respect to the number of models. Some online approaches address nonlinear dynamics with unknown parameters through the paradigm of optimism in the face of uncertainty [2, 3, 18]. In contrast, our multi-model perspective leverages a separation principle that decouples online best model identification and certainty-equivalent control. This separation not only allows computing policies offline to save online computation, but also facilitates explicit characterizations of how policy regret scales with the time horizon, state dimension, and complexity of the function class.

The article is structured as follows: Sec. 2 discusses the problem formulation and presents the main results. Sec. 3 illustrates our analysis, where we focus on the first setting (finite set of models)—the other two settings are similar at a high level and we provide a detailed presentation in the appendix. Sec. 4 provides a short conclusion, while proofs, numerical experiments, and further discussion is included in the appendix.

2 Problem formulation and summary

We consider a reinforcement learning problem where a decision-maker chooses actions $u_k \in \mathbb{R}^{d_{\mathrm{u}}}$ to control a dynamical system $x_{k+1} = f(x_k, u_k) + n_k$, where $x_k \in \mathbb{R}^{d_{\mathrm{x}}}$ denotes the state, $f : \mathbb{R}^{d_{\mathrm{x}}} \times \mathbb{R}^{d_{\mathrm{u}}} \to \mathbb{R}^{d_{\mathrm{x}}}$ the dynamics (unknown to the decision-maker), and $n_k \sim \mathcal{N}(0, \sigma^2 I)$ the process noise. The random variables n_k, $k = 1, \dots$, represent a sequence of independent and identically distributed random variables, and without loss of generality we set $x_1 = 0$. We further denote the Lipschitz constant of f in (x, u) by L.

The decision-maker aims at minimizing the expected loss, $\mathrm{E}[\sum_{k=1}^{N} l(x_k, u_k)]$, where $l : \mathbb{R}^{d_{\mathrm{x}}} \times \mathbb{R}^{d_{\mathrm{u}}} \to \mathbb{R}_{\geq 0}$ captures the stage cost, by learning and applying an appropriate and possibly random feedback policy $u_k = \mu_k(x_k)$.

We consider three different settings. In the first setting (`S1`) the decision-maker has access to m potential (nonlinear) candidate models $F := \{f^1, \dots, f^m\}$, $f^i : \mathbb{R}^{d_{\mathrm{x}}} \times \mathbb{R}^{d_{\mathrm{u}}} \to \mathbb{R}^{d_{\mathrm{x}}}$, $i = 1, \dots, m$ that describe potential system dynamics. Each f^i is L-Lipschitz in (x, u). In the second setting (`S2`), we allow for any class of functions F that is given by a bounded set in a normed vector space, which is therefore much broader and includes, for example, all bounded L-Lipschitz functions with the usual supremum norm or a bounded set of square integrable functions. In the third setting (`S3`), the dynamics are parametrized by the parameter θ, i.e., $F = \{f_\theta(x, u) \mid \theta \in \Omega\}$, where Ω is a compact real-valued set. Without loss of generality, we assume that Ω is contained in a unit ball by scaling the parameters accordingly. This captures the setting where the functions f_θ are represented by neural or transformer architectures, or when f_θ are given by linear combinations of (nonlinear) feature vectors $f_\theta(x, u) = \theta^\top \phi(x, u)$. This also encompasses linear dynamics as a special case. We further assume that the system dynamics f are contained in the set of candidate models, i.e., $f \in F$ for each setting.

This article analyzes the decision-making strategy listed in Alg. 1, which can be easily adapted to the settings `S2`/`S3` (see Alg. 2 in App. D and Alg. 3). The algorithm keeps track of the one-step prediction error, that is,

$$s_k^i = \sum_{j=1}^{k-1} \frac{|x_{j+1} - f^i(x_j, u_j)|^2}{1 + |(x_j, u_j)|^2 / b^2}, \quad f^i \in F,$$

where $b > 0$ is a sufficiently large constant, and $|(x_j, u_j)|$ denotes the ℓ_2-norm of a vector stacking x_j and u_j. The normalization with $1 + |(x_k, u_k)|^2/b^2$ ensures that the variables s_k^i remain bounded even when x_k, u_k become arbitrarily large, while for small x_k, u_k the normalization is close to the identity. This will simplify the subsequent analysis and the resulting statement of the policy-regret

Algorithm 1 Reinforcement learning (S1)

Inputs: $F := \{f^1, \dots, f^m\}$, η, M, $\{\sigma_{\mathrm{u}k}^2\}_{k=1}^{\infty}$

compute $\{\mu^1, \dots, \mu^m\}$ // e.g. by d. p.

for $k = 1, \dots$ **do**

// every Mth step

if $\mathrm{mod}(k-1, M) = 0$ **then**

$s_k^i \leftarrow \sum_{j=1}^{k-1} \frac{|x_{j+1} - f^i(x_j, u_j)|^2}{1+|(x_j,u_j)|^2/b^2}$

$i_k \sim \exp(-\eta s_k^i)/Z$

else

$i_k = i_{k-1}$ //stay with i_{k-1}

end if

//follow policy i_k and add excitation

$u_k = \mu^{i_k}(x_k) + n_{\mathrm{u}k}, \quad n_{\mathrm{u}k} \sim \mathcal{N}(0, \sigma_{\mathrm{u}k}^2 I)$

end for

Algorithm 3 Reinforcement learning (S3)

Inputs: $F = \{f_\theta \mid \theta \in \Omega\}$, η, M, $\{\sigma_{\mathrm{u}k}^2\}_{k=1}^{\infty}$

for $k = 1, \dots$ **do**

// every Mth step

if $\mathrm{mod}(k-1, M) = 0$ **then**

$s_k(\theta) = \sum_{j=1}^{k-1} \frac{|x_{j+1} - f_\theta(x_j, u_j)|^2}{1+|(x_j,u_j)|^2/b^2}$

$\theta_k \sim \exp(-\eta s_k(\theta))\, \mathbb{1}_{\theta\in\Omega}/Z$

compute μ_θ corr. to $f_\theta \in F$ // e.g. by d. p.

else

$\theta_k = \theta_{k-1}$ //stay with i_{k-1}

end if

//follow policy θ_k and add excitation

$u_k = \mu^{\theta_k}(x_k) + n_{\mathrm{u}k}, \quad n_{\mathrm{u}k} \sim \mathcal{N}(0, \sigma_{\mathrm{u}k}^2 I)$

end for

bounds. Our analysis also carries over to $b \to \infty$ and the same regret bounds apply, as is discussed in App. F; however, the constants in the resulting regret guarantees and algorithm parameters become more complex. The sum of squared distances $|x_{j+1} - f^i(x_j, u_j)|^2$ can be interpreted as the negative log-likelihood of model i given the past trajectory $\{x_j, u_j\}_{j=1}^k$, due to the fact that the process noise is Gaussian. Hence, from a Bayesian perspective, the distribution $\exp(-s_k^i)$ represents the probability that model f^i corresponds to f given the past trajectory. The scaling with η implements a softmax (for η large we greedily pick the model that maximizes the posterior, for $\eta \approx 1$ we directly sample from the posterior). Our analysis also applies when n_k is non-zero-mean, since this can be captured by modifying f accordingly, and generalizes to sub-Gaussian process noise.

The algorithm chooses control actions u_k as

$$u_k = \mu^{i_k}(x_k) + n_{\mathrm{u}k},$$

where $n_{\mathrm{u}k} \sim \mathcal{N}(0, \sigma_{\mathrm{u}k}^2 I)$, and i_k is a random variable that is defined in the following way: If $\mathrm{mod}(k-1, M) = 0$, i_k takes the value $i_k = i$ with probability density $p_k^i \sim \exp(-\eta s_k^i)/Z$ (conditional on the past), where Z denotes a normalization constant. If $\mathrm{mod}(k-1, M) \neq 0$, i_k remains fixed, i.e., $i_k = i_{k-1}$. The random variable switches only every Mth step, which ensures that the excitation with $n_{\mathrm{u}k}$ is rich enough, as specified precisely in Ass. 3 below. The feedback policy μ^i describes any policy associated with candidate model f^i, i.e., a policy that achieves the performance

$$\limsup_{N\to\infty} \frac{1}{N} \mathrm{E}\big[\sum_{k=1}^{N} l(x_k^i, \mu^i(x_k^i))\big] = \gamma^i, \tag{1}$$

on the candidate model f^i, where $x_{k+1}^i := f^i(x_k^i, \mu^i(x_k^i)) + n_k$ with $x_1^i = 0$. The policy μ^i can be optimal for model f^i, but this does not necessarily need to be the case. In practice, such a policy can be obtained by solving a Bellman equation through (approximate) dynamic programming [10], or by applying proximal policy optimization in conjunction with an offline simulator. We will consider policy regret as our performance objective, where the policy μ corresponding to the dynamics f represents the benchmark performance.

The reinforcement learning strategy has a very natural interpretation: The strategy selects, at each Mth iteration, the feedback policy μ^{i_k} under which the distribution of i_k follows a softmax function of s_k^i. The system is further excited by adding the random perturbation $n_{\mathrm{u}k}$ to the feedback policy. If persistence of excitation is guaranteed, the estimation will converge at a rate at least $\mathcal{O}(1/k^2)$, which yields a policy regret (compared to the strategy μ corresponding to the dynamics f) that scales logarithmically in the horizon N and the number of candidates m.

We emphasize that our analysis technique translates in straightforward ways to more general situations than the ones described herein. For example, while this article focuses on time-invariant policies, it would be straightforward to also incorporate time-varying policies μ_k^i, and a corresponding finite-horizon benchmark. More precisely, we focus on steady-state performance, where the benchmark

is given by the steady-state performance of policy μ (corresponding to f). However, finite-horizon objectives can be easily accommodated by measuring regret with respect to the optimal finite-horizon policy μ_k; the same nonasymptotic regret bounds would apply. The article also focuses on "naive" excitation signals $n_{\mathrm{u}k}$, sampled from a normal distribution. However, our analysis principle is flexible enough to also incorporate more general type of excitation strategies (e.g., relying on domain-specific knowledge), as long as the excitation has finite second moments and guarantees a persistence condition similar to Ass. 3.

Our results are summarized as follows:

Theorem 2.1 *(S1) Let the cost-to-go function corresponding to f and the stage cost l be smooth (see Ass. 1 and 2), the feedback policies μ^i be Lipschitz continuous, and let a persistence of excitation condition be satisfied (see Ass. 3). Then, for a constant learning rate η and $\sigma_{\mathrm{u}k}^2 \sim 1/(d_{\mathrm{u}}k) + \ln(m)/(d_{\mathrm{u}}k^2)$ the policy regret of Alg. 1 is bounded by*

$$\mathrm{E}[\sum_{k=1}^{N} l(x_k, u_k)] - N\gamma \leq c_{\mathrm{r}1}\ln(N) + c_{\mathrm{r}2}\ln(m) + c_{\mathrm{r}3}\sigma^2 d_{\mathrm{x}},$$

for all $N \geq 2M$, where $c_{\mathrm{r}1}, c_{\mathrm{r}2}, c_{\mathrm{r}3}$ are constant, and γ corresponds to the $\mathcal{H}_2$ gain associated with the dynamics f (see (1)*). The precise constants are listed in Thm. 3.2.*

Theorem 2.2 *(S2) Let the set of candidate models F be a bounded set in a normed vector space. Let the cost-to-go function corresponding to f and the stage cost l be smooth (see Ass. 2 and 5), the feedback policies $\mu^{\bar{f}}$ corresponding to an $\bar{f} \in F$ be Lipschitz continuous, and let a persistence of excitation condition be satisfied (see Ass. 4). Then, for all $N \geq 2M$, any $\epsilon > 0$, for a constant learning rate η, and $\sigma_{\mathrm{u}k}^2 \sim 1/(\epsilon^2 d_{\mathrm{u}}k) + \ln(m(\epsilon))/(\epsilon^2 d_{\mathrm{u}}k^2)$, the policy regret of Alg. 2 (see App. D) is bounded by*

$$\mathrm{E}[\sum_{k=1}^{N} l(x_k, u_k)] - N\gamma \leq c_{\mathrm{r}0}N\epsilon^2 + c_{\mathrm{r}1}\ln(N)/\epsilon^2 + c_{\mathrm{r}2}\ln(m(\epsilon))/\epsilon^2 + c_{\mathrm{r}3}\sigma^2 d_{\mathrm{x}},$$

where $m(\epsilon)$ denotes the packing number of the set F. The precise constants are listed in Thm. D.2.

Theorem 2.3 *(S3) Let the set of candidate models F be parametrized by θ, i.e., $F = \{f_\theta(x, u) \mid \theta \in \Omega\}$, where $\Omega \subset \mathbb{R}^p$ is contained in a unit ball of dimension p. Let the cost-to-go function corresponding to f and the stage cost l be smooth (see Ass. 2 and Ass. 7), the feedback policies μ_θ corresponding to each $f_\theta \in F$ be Lipschitz continuous, and let a persistence of excitation condition be satisfied (see Ass. 6). Then, for all $N \geq 2M$, for a constant learning rate η, and $\sigma_{\mathrm{u}k}^2 \sim 1/(d_{\mathrm{u}}k) + p/(d_{\mathrm{u}}k^2)$, the policy regret of Alg. 3 is bounded by*

$$\sum_{k=1}^{N} \mathrm{E}[l(x_k, u_k)] - N\gamma \leq \sqrt{(c_{\mathrm{r}1}\ln(N) + c_{\mathrm{r}2}p)N} + c_{\mathrm{r}3}\sigma^2 d_{\mathrm{x}}.$$

The precise constants are listed in Thm. E.1.

The results characterize precisely how the policy regret scales with the dimension $d_{\mathrm{x}}, d_{\mathrm{u}}$ and the time horizon N. In the setting of Thm. 2.1, we have a finite class of models, and the policy regret scales with $\ln(m)$, which is in line with the literature on online learning [13, 33]. Thm. 2.2 relies on a packing argument, whereby the set F is successively approximated by a finite number of candidate models. The result is stated in full generality; for a specific function class F and packing number $m(\epsilon)$ the right-hand side can be minimized over ϵ (the discretization width). For instance if F consists of the space of bounded L-Lipschitz functions, the packing number $m(\epsilon)$ scales with $d_{\mathrm{x}} \exp((L/\epsilon)^{d_{\mathrm{x}}+d_{\mathrm{u}}})$, which means that the policy regret grows roughly with $N^{(d_{\mathrm{x}}+d_{\mathrm{u}})/(d_{\mathrm{x}}+d_{\mathrm{u}}+2)} = o(N)$, and establishes no-regret learning for a very large class of functions. In the special case where $d_{\mathrm{x}} = d_{\mathrm{u}} = 1$, the right-hand side grows with $\sqrt{N}$. Thm. 2.3 is of direct practical importance, since it provides an algorithm and corresponding regret bound that applies to the typical scenario where the functions F are parametrized, for example by neural networks. In the simplest setting, F consists of linear dynamical systems, which directly recovers well-known results from the literature [e.g., 46, 19]. More precisely, if F consists of linear dynamical systems, the number of parameters is given by $d_{\mathrm{x}}^2 + d_{\mathrm{x}}d_{\mathrm{u}}$, which means that the resulting regret bound scales with $\sqrt{(d_{\mathrm{x}}^2 + d_{\mathrm{x}}d_{\mathrm{u}})N}$.

We conclude the summary by commenting on boundedness of states. In control-theoretic applications (and in the related community) boundedness of solutions and benign transients are a primary concern. We will see that in all our results we can ensure boundedness provided that the stage cost satisfies $l(x,u) \geq \underline{L}_{\mathrm{l}}|x|^2/2$ for a constant $\underline{L}_{\mathrm{l}} > 0$. More precisely, we can guarantee that

$$\underline{L}_{\mathrm{l}}\mathrm{E}[|x_k|^2] \leq 2\mathrm{E}[V(x_k)] \leq c_{\mathrm{b}} \tag{2}$$

for all $k = 1, \ldots,$ along the trajectories of our reinforcement learning algorithm, where V refers to the cost-to-go function corresponding to the dynamics f and policy μ, and $c_{\mathrm{b}} > 0$ is an explicit constant. Due to the fact that the dynamics are Lipschitz continuous and $n_k, n_{\mathrm{u}k}$ are Gaussian, x_k, u_k are in fact sub-Gaussian with mean and second moment bounded by $\sqrt{c_{\mathrm{b}}/\underline{L}_{\mathrm{l}}}$ and $c_{\mathrm{b}}/\underline{L}_{\mathrm{l}}$, respectively, and we can therefore characterize tail probabilities for finite k, as well as for arbitrarily large values of k under ergodicity assumptions on the dynamics arising from μ.

3 Summary of the analysis

This section discusses the technical details and insights that lead to the results presented in Thm. 2.1-2.3. The presentation focuses on the setting `S1`, since, as we will see, the results in setting `S2` and `S3` follow analogously.

3.1 Finite model set-up

This section considers the set-up where F is finite, i.e., $F = \{f^1, \ldots, f^m\}$ and $f \in F$. We denote the cost-to-go function related to the dynamics f and the policy μ by $V : \mathbb{R}^{d_{\mathrm{x}}} \to \mathbb{R}_{\geq 0}$, where V is any function that satisfies the following assumption:

Assumption 1 *(Bellman-type inequality) The cost-to-go function V (corresponding to f and μ) satisfies the following inequality*

$$V(x) \geq \mathrm{E}[l(x,u) + V(f(x,u)+n)] - \gamma - d_{\mathrm{u}}L_{\mathrm{u}}\sigma_{\mathrm{u}}^2, \tag{3}$$

for a constant L_{u} and for all $x \in \mathbb{R}^{d_{\mathrm{x}}}$, where $u = \mu(x) + n_{\mathrm{u}}$, $n \sim \mathcal{N}(0, \sigma^2 I)$, $n_{\mathrm{u}} \sim \mathcal{N}(0, \sigma_{\mathrm{u}}^2 I)$, and the expectation is taken over n and n_{u}.

The rationale behind Ass. 1 is the following: From a dynamic programming point of view computing an optimal policy μ requires solving a corresponding infinite-horizon average-cost-per-stage problem. In general, a corresponding Bellman equation and cost-to-go function might not exist, as for example discussed in [10, Ch. 5]. The formulation via Ass. 1 circumvents these technical difficulties, due to the fact that γ is not required to correspond to the optimal infinite-horizon average cost. Indeed, from a control-theoretic point of view Ass. 1 characterizes a notion of dissipation [51], where V represents a storage function and $-l(x,u) + \gamma$ the supply rate (for $\sigma_{\mathrm{u}} = 0$). Moreover, if a Bellman equation [10, Prop. 5.5.1] and corresponding cost-to-go function exist for the dynamics f, then Ass. 1 is clearly satisfied for the corresponding cost-to-go function (for $\sigma_{\mathrm{u}} = 0$). The additional term $d_{\mathrm{u}}L_{\mathrm{u}}\sigma_{\mathrm{u}}^2$ captures the influence of the excitation n_{u} and is without loss of generality, since for any smooth function $\xi : \mathbb{R}^{d_{\mathrm{u}}} \to \mathbb{R}$ the following applies

$$\mathrm{E}[\xi(u+n_{\mathrm{u}})] = \mathrm{E}\Big[\xi(u) + \nabla\xi(u)^\top n_{\mathrm{u}} + \frac{1}{2}n_{\mathrm{u}}^\top \nabla^2\,\xi(\bar{u})\,n_{\mathrm{u}}\Big] = \mathrm{E}[\xi(u)] + \mathcal{O}(d_{\mathrm{u}}\sigma_{\mathrm{u}}^2),$$

where $n_{\mathrm{u}} \sim \mathcal{N}(0, \sigma_{\mathrm{u}}^2 I)$. The constant L_{u} in Ass. 1 makes the previous bound quantitative.

We will further require the following smoothness conditions:

Assumption 2 *The policies μ^i are L_μ Lipschitz, the stage-cost l is $\bar{L}_{\mathrm{l}}$ smooth, and the cost-to-go function V is $\bar{L}_{\mathrm{V}}$ smooth and satisfies $V(x) \geq -\underline{c}_{\mathrm{V}} + \underline{L}_{\mathrm{V}}|x|^2/2$ for some $\underline{L}_{\mathrm{V}} > 0$ and $\underline{c}_{\mathrm{V}} \geq 0$.*

These smoothness assumptions will be needed to analyze how the cost-to-go V evolves if the feedback policy $\mu^q \neq \mu$ is applied and prevent the state from diverging in finite time. We note that the quadratic lower bound on $V(x)$ is automatically satisfied in view of Ass. 1 if $l(x,u) \geq \underline{L}_{\mathrm{l}}|x|^2/2$ for a constant $\underline{L}_{\mathrm{l}} > 0$. Ass. 1 and 2 are clearly satisfied if f^i are linear functions and l is a positive definite quadratic.

We further require the following assumption:

Assumption 3 *There exists an integer $M > 0$ and two constants $c_\mathrm{e} > 0$ and $b > 0$ such that for any $x_1 \in \mathbb{R}^{d_\mathrm{x}}$, $\sigma_\mathrm{u} > 0$, and $f^i \in F$, $f^i \neq f$,*

$$\frac{1}{M}\sum_{k=1}^{M} \mathrm{E}\Big[\frac{|f^i(x_k,u_k) - f(x_k,u_k)|^2}{1+|(x_k,u_k)|^2/b^2}\Big] \geq d_\mathrm{u} c_\mathrm{e} \sigma_\mathrm{u}^2$$

holds, where $x_{k+1} = f(x_k,u_k) + n_k$, $u_k = \mu^q(x_k) + n_{\mathrm{u}k}$ with $n_k \sim \mathcal{N}(0,\sigma^2 I)$, $n_{\mathrm{u}k} \sim \mathcal{N}(0,\sigma_\mathrm{u}^2 I)$, and $q \in \{1,\dots,m\}$.

The previous assumption specifies persistence of excitation, which guarantees that the estimate i_k of the best candidate model will quickly converge to i^*, where $f^{i^*} = f$. Ass. 3 can be restated in the following equivalent way. For any $\sigma_\mathrm{u} > 0$ and initial condition x_1, the model $f \in F$ is the unique minimizer of the expected one-step prediction error s_k^i, $i \in F$, accumulated over M steps. The integer M corresponds to the time interval by which i_k is updated, see Alg. 1. Ass. 3 includes a normalization with the constant b, which may seem slightly strong compared to the literature [see, e.g., 38, Ch. 8.2], where $b \to \infty$ is usually considered. However, as discussed in App. F our analysis also encompasses the case $b \to \infty$; the resulting constants are more elaborate and we therefore focus our discussion on the situation where b is finite. We further note that in the situation where $f^i \in F$ are linear, that is $f^i(x,u) = A^i x + B^i u$, $\mu^i(x) = K^i x$, Ass. 3 for $b \to \infty$ is straightforward to verify and we obtain, for example, the following bound for $k \geq 2$:

$$\mathrm{E}[|f^i(x_k,u_k) - f(x_k,u_k)|^2] \geq \sigma_\mathrm{u}^2 |B^i - B|_\mathrm{F}^2 + \left(\sigma_\mathrm{u}^2 \underline{\sigma}(W_{k-1}^\mathrm{c}) + \sigma^2\right) |A^i - A + (B^i - B)K^q|_\mathrm{F}^2, \quad (4)$$

where K^q represents the linear feedback controller corresponding to model $f^q \in F$, and W_k^c denotes the controllability Gramian (over k steps):

$$W_k^c = \sum_{j=0}^{k-1} (A + BK^q)^j {}^\top B B^\top (A + BK^q)^j,$$

where $\underline{\sigma}$ denotes the minimum singular value, and $|\cdot|_\mathrm{F}$ the Frobenius norm. Hence, Ass. 3 (for $b \to \infty$) is generically satisfied for linear systems, whereby the constant c_e relates to the controllability of the closed-loop dynamics and the accuracy $|A^i - A|_\mathrm{F}^2$ and $|B^i - B|_\mathrm{F}^2$ of the different candidate models. The previous rationale can be extended to nonlinear dynamical systems, as shown in App. F and Prop. F.2, which highlights that Ass. 3 is satisfied for a broad class of dynamical systems.

Our analysis of Alg. 1 starts by showing that the convergence to the best candidate model is fast, which leads to the logarithmic scaling of the policy regret with N and m. This is summarized with the following proposition:

Proposition 3.1 *Let Ass. 3 be satisfied and let the step size be $\eta \leq \min\{1/(4M\sigma^2), 1/(2ML^2b^2)\}$. Then, the following holds*

$$\Pr(i_k = i) \leq \exp\left(-\frac{d_\mathrm{u} c_\mathrm{e} \eta}{4} \sum_{j=1}^{k-M} \sigma_{\mathrm{u}j}^2\right),$$

for $k = 1, 2, \dots$. Moreover, choose $\sigma_{\mathrm{u}k}^2$ as

$$\sigma_{\mathrm{u}k}^2 = \frac{4}{\eta d_\mathrm{u} c_\mathrm{e} M}\left(\frac{2}{\lceil k/M \rceil} + \frac{\ln(m)}{(\lceil k/M \rceil)^2}\right),$$

where $\lceil \cdot \rceil$ denotes rounding to the next higher integer. Then, it holds that

$$\Pr(i_k \neq i^*) \leq \frac{M^2}{(k-M)^2},$$

for all $k \geq M+1$, where $f^{i^*} = f$.

Proof The proof can be found in App. C.1 and relies on a concentration of measure argument. □

An immediate corollary of the fast convergence rate established with Prop. 3.1 is that the sequence i_k will converge to i^* in finite time (almost surely), where $f^{i^*} = f$. This is discussed in Cor. C.6. As a result of Prop. 3.1, we are now ready to state and prove our first main result that characterizes the policy regret in setting `S1`.

Theorem 3.2 *Let Ass. 1, 2 and 3 be satisfied and choose $\eta \leq \min\{1/(4M\sigma^2), 1/(2ML^2b^2)\}$ and $\sigma_{\mathrm{u}k}^2$ as in Prop. 3.1. Then, the policy regret of Alg. 1 is bounded by*

$$\sum_{k=1}^{N} \mathrm{E}[l(x_k, u_k)] - N\gamma \leq c_{r1} + c_{r2} M \ln(m) + c_{r2} \ln(N),$$

for all $N \geq 2M$, where the constants c_o, c_2 are specified in Lemma C.3, and c_{r1}, c_{r2} are given by

$$c_{r1} = 3c_\alpha M(d_{\mathrm{x}}\sigma^2 \bar{L}_{\mathrm{V}}/2 + c_{\mathrm{o}}) + c_{r2}, \quad c_{r2} = \frac{8c_\alpha(\bar{L}_{\mathrm{V}}L^2 + \bar{L}_{\mathrm{l}} + L_{\mathrm{u}})}{\eta c_{\mathrm{e}}}, \quad c_\alpha = e^{3c_2 M}.$$

Proof (Sketch, details are in App. C.2) The proof relies on using V as a Lyapunov function and performing the following decomposition

$$\mathrm{E}[V(x_{k+1})] = \mathrm{E}[V(x_{k+1})|i_k \neq i^*]\mathrm{Pr}(i_k \neq i^*) + \mathrm{E}[V(x_{k+1})|i_k = i^*]\mathrm{Pr}(i_k = i^*). \tag{5}$$

The first term describes the evolution of $V(x_{k+1})$ when choosing $i_k \neq i$, and in this (unfavorable) situation V may grow at most exponentially. This is captured by the following bound that relies on the continuity assumptions on V (see Lemma C.3)

$$\mathrm{E}[V(x_{k+1})|i_k \neq i^*] \leq c_2 \mathrm{E}[V(x_k)] + \mathcal{O}(\sigma^2 + \sigma_{\mathrm{u}k}^2) - \mathrm{E}[l(x, u_k)|i_k \neq i^*],$$

where the notation $\mathcal{O}$ hides continuity and dimension-related constants. The second term in (5), describes the favorable situation of choosing $i_k = i^*$, where $V(x_{k+1})$ is bounded as a result of the Bellman-type inequality (3). This yields:

$$\mathrm{E}[V(x_{k+1})|i_k = i^*] \leq \mathrm{E}[V(x_k)] + \gamma + \mathcal{O}(\sigma_{\mathrm{u}k}^2) - \mathrm{E}[l(x_k, u_k)|i_k = i^*],$$

where continuity and dimension-related constants are again hidden. By combining the two inequalities we arrive at

$$\mathrm{E}[V(x_{k+1})] \leq \mathrm{E}[V(x_k)](c_2 \mathrm{Pr}(i_k \neq i^*) + 1) + \gamma - \mathrm{E}[l(x_k, u_k)] + \mathcal{O}(\sigma_{\mathrm{u}k}^2 + \mathrm{Pr}(i_k \neq i^*)\sigma^2). \tag{6}$$

From Prop. 3.1, we know that $\mathrm{Pr}(i_k \neq i^*)$ decays at rate $1/k^2$. This means that, roughly speaking, the inequality (6) gives rise to a telescoping sum (see Lemma C.4 for details), which yields

$$\sum_{k=1}^{N} \mathrm{E}[l(x_k, u_k)] - \gamma N \leq \mathcal{O}(\sum_{k=1}^{N} (\sigma_{\mathrm{u}k}^2 + \mathrm{Pr}(i_k \neq i^*)\sigma^2)).$$

The fact that $\mathrm{Pr}(i_k \neq i^*)$ is summable, due to the decay at rate $1/k^2$, and that the sum over $\sigma_{\mathrm{u}k}^2$ evaluates to $\mathcal{O}(\ln(N) + \ln(m))$ establishes the desired result up to constants (these are computed in App. C.2). □

The proof of Thm. 3.2 relies on using V as a Lyapunov function. Provided that the stage cost $l(x, u)$ is bounded below by a quadratic of the type $|x|^2$, we can modify the analysis in straightforward ways to obtain explicit bounds on $\mathrm{E}[V(x_k)]$ and hence on $\mathrm{E}[|x_k|^2]$, uniform over k, which is an important concern in the adaptive control community. Moreover, these bounds require persistence of excitation only over a finite number of steps, since $\mathrm{Pr}(i_k \neq i^*)$ is monotonically decreasing even when Ass. 3 is not satisfied. The details are presented in App. C.4.

3.2 Infinite cardinality

The ideas described in the previous section translate to the situation in which the set of candidate models F is a bounded subset of a normed vector space with norm $\|\cdot\|$. For example, F could represent the set of bounded, L-Lipschitz continuous functions that map from $\mathbb{R}^{d_{\mathrm{x}}} \times \mathbb{R}^{d_{\mathrm{u}}} \to \mathbb{R}^{d_{\mathrm{x}}}$, with $\|\cdot\|$ the supremum norm. Alternatively, F could be a bounded subset of the set of square integrable functions. Our presentation focuses on the main ideas that enable us to apply the arguments from the previous section; the details and formal proofs can be found in App. D.

The decision-making strategy for S2 is listed in Alg. 2 (see App. D). Alg. 2 computes a minimizer $\mathrm{argmin}_{\bar{f} \in F} s_k(\bar{f})$, which will be denoted by f^*, where $s_k(\bar{f})$ denotes the prediction error as before,

$$s_k(\bar{f}) = \sum_{j=1}^{k-1} \frac{|x_{j+1} - \bar{f}(x_j, u_j)|^2}{1 + |(x_k, u_k)|^2/b^2}, \quad \bar{f} \in F.$$

We then construct an ϵ-packing of the set F, denoted by F_k^ϵ by greedily adding functions $f^i \in F$ as long as $\|f^i - \bar{f}\| > \epsilon$ for all $\bar{f} \in F_k^\epsilon$. As a result F_k^ϵ covers F by construction, i.e., for every $f \in F$ there exists $f^i \in F_k^\epsilon$ such that $\|f^i - f\| \leq \epsilon$. The cardinality of F_k^ϵ is bounded by the packing number of F, which is denoted by $m(\epsilon)$. The algorithm then randomly samples i_k as before and applies the feedback policy μ^{i_k} that corresponds to model $f^{i_k} \in F_k^\epsilon$. Clearly, these steps (minimization over $\bar{f} \in F$, constructing the packing, and solving a dynamic programming problem at every iteration) are computationally intractable in general and one would have to resort to approximations in practice. The purpose of Alg. 2 is to provide an upper bound on the *sample complexity* of online reinforcement learning in this very general setting and not to characterize *computational complexity*; see the next subsection for a computationally tractable variant. As before, the key step to our analysis is to ensure that $\Pr(i_k = i)$ decays rapidly for models $f^i \in F_k^\epsilon$ where $\|f^i - f\|$ is large. The fact that the dynamics f are not included in F_k^ϵ is of minor importance, since by construction F_k^ϵ contains f^*, the minimizer of $s_k(f)$. This means that the arguments used in deriving Prop. 3.1 apply in the same way and implies that $\Pr(i_k \notin I_k^*) \leq M^2/(k-M)^2$ as before, where I_k^* denotes the set of models $f^{i*} \in F_k^\epsilon$ that satisfy $\|f^{i^*} - f\| \leq \epsilon$. As a result, the same arguments as in the proof of Thm. 3.2 apply, which yields the statement of Thm. 2.2. The details are presented in App. D.

3.3 Parametric models

The following section discusses the situation where the set of candidate models F is parametrized by a parameter $\theta \in \Omega \subset \mathbb{R}^p$, where Ω is contained in a p-dimensional unit ball, that is,

$$F = \{f_\theta : \mathbb{R}^{d_{\mathrm{x}}} \times \mathbb{R}^{d_{\mathrm{u}}} \to \mathbb{R}^{d_{\mathrm{x}}} \mid \theta \in \Omega\}.$$

The canonical example we have in mind is when f_θ is parametrized with a large neural network, transformer, or state-space architecture, where θ represents the parameters. As in the previous section, we assume that $f \in F$, and without loss of generality, we set $f = f_{\theta=0}$, i.e., the parameters are centered around f.

Alg. 3 has a particularly straightforward interpretation, which also facilitates its implementation in practice. In each iteration, f_{θ_k} is sampled from the posterior distribution over models f_θ, scaled by η. In the special case where $f_\theta(x,u) = \phi(x,u)^\top \theta$ we note that the density $\exp(-\eta s_k(\theta))/Z$ corresponding to the random variable θ_k is Gaussian, with mean and covariance

$$\operatorname*{argmin}_{\theta \in \mathbb{R}^p} \sum_{j=1}^{k-1} \frac{|x_{j+1} - \phi(x_j,u_j)^\top \theta|^2}{1 + |(x_j,u_j)|^2/b^2}, \quad \frac{1}{2\eta}\left(\sum_{j=1}^{k-1} \frac{\phi(x_j,u_j)\phi(x_j,u_j)^\top}{1 + |(x_j,u_j)|^2/b^2}\right)^{-1}.$$

The Gaussian mean and covariance can be efficiently evaluated by running a recursive least squares algorithm, resulting in a per-iteration computational complexity of only $\mathcal{O}(p^2)$. The corresponding computation of the policy μ_θ for the model f_θ is much more challenging, but can, in principle, be done offline with dynamic programming, or in an offline simulation with proximal policy optimization, for example. A notable exception is when $\phi(x,u)$ is linear, in which case the corresponding (steady-state optimal) policy μ_θ is linear and can be computed by solving a Riccati equation in $\mathcal{O}(d_{\mathrm{x}}^3)$ steps. If f_θ has a more general structure, the sampling can, for example, be implemented with Langevin Monte-Carlo [49]. The regret analysis follows the same steps as in Sec. 3.1 and is included in App. E.

4 Conclusion

This article provides policy-regret guarantees for online reinforcement learning with *nonlinear dynamical systems over continuous state and action spaces*. We provide a suite of algorithms and prove that the resulting policy regret over N steps scales as $\mathcal{O}(\ln(N) + \ln(m))$ in a setting where there is a finite class of m models and as $\mathcal{O}(\sqrt{Np})$ in a setting where models are parametrized over a compact real-valued space of dimension p. The results require persistence of excitation, and rely on continuity assumptions on the dynamics, feedback policies, and a corresponding value function.

The results highlight important and fruitful connections between reinforcement learning and control theory. There are numerous exciting future research avenues, including the exploration of an $\mathcal{H}_\infty$ or an output feedback setting, the application to emerging real-world and large-scale infrastructure systems, or the analysis of the model-agnostic case, where the dynamics do not belong to the class of systems known to the learner.

Acknowledgments and Disclosure of Funding

We thank the German Research Foundation and the Max Planck ETH Center for Learning Systems for the support. We also acknowledge funding from the Chair "Markets and Learning," supported by Air Liquide, BNP PARIBAS ASSET MANAGEMENT Europe, EDF, Orange and SNCF, sponsors of the Inria Foundation.

References

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Conference on Learning Theory*, pages 1–26, 2011.

[2] Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *Conference on Uncertainty in Artificial Intelligence*, pages 1–11, 2015.

[3] Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via Lagrangian relaxation. In *International Conference on Machine Learning*, pages 23–31, 2020.

[4] Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119, 2019.

[5] Brian DO Anderson, Thomas S Brinsmead, Franky De Bruyne, Joao Hespanha, Daniel Liberzon, and A Stephen Morse. Multiple model adaptive control. Part 1: Finite controller coverings. *International Journal of Robust and Nonlinear Control*, 10(11-12):909–929, 2000.

[6] Brian DO Anderson and Arvin Dehghani. Challenges of adaptive control–past, permanent and future. *Annual Reviews in Control*, 32(2):123–135, 2008.

[7] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems*, 19, 2006.

[8] Peter L Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Conference on Uncertainty in Artificial Intelligence*, page 35–42, 2009.

[9] Gerald Beer and Michael J. Hoffman. The lipschitz metric for real-valued continuous functions. *Journal of Mathematical Analysis and Applications*, 406(1):229–236, 2013.

[10] Dimitri Bertsekas. *Dynamic Programming and Optimal Control: Volume I*. Athena Scientific, 2017. 4th edition.

[11] Nicholas M Boffi, Stephen Tu, and Jean-Jacques E Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control Conference*, pages 471–483, 2021.

[12] Francesco Borrelli, Alberto Bemporad, and Manfred Morari. *Predictive Control for Linear and Hybrid Systems*. Cambridge University Press, 2017.

[13] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[14] Nicolas Chatzikiriakos and Andrea Iannelli. Sample complexity bounds for linear system identification from a finite set. *IEEE Control Systems Letters*, 8:2751–2756, 2024.

[15] Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems. In *Conference on Learning Theory*, pages 1114–1143, 2021.

[16] Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *International Conference on Machine Learning*, pages 1029–1038, 2018.

[17] Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309, 2019.

[18] Lorenzo Croissant, Marc Abeille, and Bruno Bouchard. Near-continuous time reinforcement learning for continuous state-action spaces. In *International Conference on Algorithmic Learning Theory*, pages 444–498, 2024.

[19] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 31, 2018.

[20] Kenji Doya, Kazuyuki Samejima, Ken-ichi Katagiri, and Mitsuo Kawato. Multiple model-based reinforcement learning. *Neural Computation*, 14(6):1347–1369, 2002.

[21] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

[22] Dylan J Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive decision making. *arXiv preprint arXiv:2312.16730*, 2023.

[23] Elad Hazan, Sham Kakade, and Karan Singh. The nonstochastic control problem. In *International Conference on Algorithmic Learning Theory*, pages 408–421, 2020.

[24] Elad Hazan and Karan Singh. Introduction to online nonstochastic control. *arXiv preprint arXiv:2211.09619*, 2022.

[25] Joao Hespanha, Daniel Liberzon, A Stephen Morse, Brian DO Anderson, Thomas S Brinsmead, and Franky De Bruyne. Multiple model adaptive control. Part 2: switching. *International Journal of Robust and Nonlinear Control*, 11(5):479–496, 2001.

[26] Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6(1):123–158, 2023.

[27] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

[28] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *Mathematics of Operations Research*, 48(3):1496–1521, 2023.

[29] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

[30] Aren Karapetyan, Efe C Balta, Andrea Iannelli, and John Lygeros. Closed-loop finite-time analysis of suboptimal online control. *arXiv preprint arXiv:2312.05607*, 2023.

[31] Taylan Kargin, Sahin Lale, Kamyar Azizzadenesheli, Animashree Anandkumar, and Babak Hassibi. Thompson sampling achieves $\tilde{O}(\sqrt{T})$ regret in linear quadratic control. In *Conference on Learning Theory*, pages 3235–3284, 2022.

[32] Jihun Kim and Javad Lavaei. Online bandit nonlinear control with dynamic batch length and adaptive learning rate. *arXiv preprint arXiv:2410.03230*, 2024.

[33] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[34] Yingying Li, Subhro Das, and Na Li. Online optimal control with affine constraints. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 8527–8537, 2021.

[35] Yingying Li, James A Preiss, Na Li, Yiheng Lin, Adam Wierman, and Jeff S Shamma. Online switching control with stability and regret guarantees. In *Learning for Dynamics and Control Conference*, pages 1138–1151, 2023.

[36] Daniel Liberzon. *Switching in Systems and Control*, volume 190. Springer, 2003.

[37] Yiheng Lin, James A Preiss, Emile Anand, Yingying Li, Yisong Yue, and Adam Wierman. Online adaptive policy selection in time-varying systems: No-regret via contractive perturbations. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Lennart Ljung. *System Identification*. Prentice Hall, second edition, 1999.

[39] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

[40] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020, 2020.

[41] Michael Muehlebach. Adaptive decision-making with constraints and dependent losses. *IFAC-PapersOnLine*, 56(2):5107–5114, 2023.

[42] Kumpati S Narendra and Jeyendran Balakrishnan. Adaptive control using multiple models. *IEEE Transactions on Automatic Control*, 42(2):171–187, 1997.

[43] Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. EPOpt: Learning robust neural network policies using model ensembles. In *International Conference on Learning Representations*, 2017.

[44] James Blake Rawlings, David Q Mayne, Moritz Diehl, et al. *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, 2017.

[45] M.G. Safonov and Tung-Ching Tsao. The unfalsified control concept and learning. *IEEE Transactions on Automatic Control*, 42(6):843–847, 1997.

[46] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online LQR. In *International Conference on Machine Learning*, pages 8937–8948, 2020.

[47] Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pages 3320–3436, 2020.

[48] Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.

[49] Santosh S. Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, 2019.

[50] Martin J. Wainwright. *High-Dimensional Statistics*. Cambridge University Press, 2019.

[51] Jan C Willems. Dissipative dynamical systems. *European Journal of Control*, 13(2-3):134–151, 2007.

[52] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756, 2020.

[53] Peng Zhao, Yu-Hu Yan, Yu-Xiang Wang, and Zhi-Hua Zhou. Non-stationary online learning with memory and non-stochastic control. *Journal of Machine Learning Research*, 24(1):9831–9900, 2023.

A Related work

This article revolves around online reinforcement learning with multiple nonlinear candidate models. We adopt a viewpoint at the intersection of online decision-making, reinforcement learning, and adaptive control. We review representative works along these lines as follows.

Online reinforcement learning is concerned about interacting with an unknown environment to optimize a cumulative performance metric. The initial focus has been on the tabular setting with discrete states and actions [7, 8, 27]. The more challenging scenario of online continuous control attracts increasing attention, where both states and actions are continuous.

Online continuous control studies optimizing a cumulative cost involving continuous states and actions and Markovian state transitions. At each time step, the stage cost is a function of the state and the input, and the cost is revealed sequentially after the control input is applied. This online nature constitutes the main difference compared with classical optimal control [10] and gives rise to nonasymptotic performance characterizations via regret, i.e., the cumulative performance gap relative to the optimal policy in hindsight. Online control hinges on specifying an appropriate policy class and leveraging effective mechanisms for searching control policies. For instance, the class of linear state feedback policies is often associated with problems involving linear dynamics and quadratic costs. Policies attaining sublinear regret can be found via iterative gradient-based schemes [16, 21, 26] and explore-then-commit pipelines based on the certainty-equivalent principle [19, 39, 46], optimism in the face of uncertainty [1, 17, 3, 18], or Thompson sampling [2, 31]. For linear dynamics and convex stage costs, typical benchmarks are disturbance-action policies represented by linear combinations of states and past disturbances [4, 23, 47, 34, 15, 53]. Growing attention is currently paid to online nonlinear control, for which additional structure (e.g., linear mapping of nonlinear features [29] or matched uncertainty [11]), properties (e.g., contractive perturbation [37] or incremental input-to-state stability [30]), and parameterizations of nonlinear dynamics and policies [2, 18] are required. We refer the readers to [24, 48] for comprehensive reviews.

Our analysis is closely aligned with optimistic methods [2, 18, 29]. These approaches hinge on iteratively refining parametric models with confidence bounds and applying policies associated with the most optimistic model. The regret bounds therein enjoy a square-root dependence on the time horizon and the Eluder dimension [22], which quantifies the hardness of the model class and facilitates transforming model learning errors into performance gaps. In contrast, we offer a fresh multi-model perspective and a separation principle to decouple online best model selection and certainty-equivalent control. This separation allows control policies to be computed offline, thereby alleviating online computational burden. We additionally provide quantitative and accessible persistence of excitation conditions to characterize the discrepancy (i.e., hardness of distinguishing) among candidate models. Compared to characterizations via Eluder dimension, our assumptions are explicit and met for a large set of linear and nonlinear dynamical systems. The resulting policy regret matches the scalings of the aforementioned optimistic methods in terms of time horizon. Furthermore, the policy regret herein features benign dependence on the number of candidate models and problem dimensions.

Multi-model adaptive control emphasizes the versatility of a system to handle diverse operating conditions by switching among multiple candidate models and associated controllers [42, 5, 25, 41, 14]. There is a supervisory controller that tracks the performance of the running controller and, if necessary, applies another more appropriate controller based on switching logic. Oftentimes the switching criterion follows the model (and the corresponding controller) with the smallest estimation error integral [36, 6] or implements performance-based falsification [45]. The above works mainly focus on asymptotic stabilization, whereas this article explores online reinforcement learning characterized by nonasymptotic policy regret measures.

Online control with switching policies is closely related to adaptive control with multiple models. Nonetheless, instead of tackling asymptotic stabilization, online control addresses optimal control from a modern nonasymptotic finite-sample perspective. Specifically, [35, 32] consider regulating a nonlinear dynamical system by iteratively selecting a control input from a finite set of candidate control policies. The key principles are to use the system trajectory driven by the chosen controller as a performance criterion to filter out non-stabilizing controllers and to identify the best stabilizing controller in hindsight via Exp3, a classical multi-armed bandit algorithm. The regret bounds therein scale sublinearly with the time horizon, but grow exponentially with the number of non-stabilizing controllers. In contrast, in the setting with finite candidate models, our algorithm attains a favorable

logarithmic regret in terms of both the time horizon and the number of models. Furthermore, we extend the design and analysis to handle a continuum of nonlinear candidate models lying in a bounded set of function spaces. Our multi-model perspective is also connected to the line of works [20, 43, 40] that use dynamic convex combinations of an ensemble of models to synthesize policies. In contrast, we tackle a challenging non-episodic scenario without state reset and handle more general model classes including parametric families and families with infinite cardinality.

This article leverages a *separation principle*, whereby we dynamically identify the best model within a given class and apply a certainty-equivalent feedback policy. Along the lines of policy extraction, we mention two streams of related works, since these can be readily incorporated into our online reinforcement learning pipelines. One stream is reinforcement learning with *linear function approximation* [52, 28], featuring efficient search for no-regret policies when the transition dynamics and stage costs are approximated by linear representations of feature mappings. Another stream is *model predictive control* [44, 12], where receding-horizon policies are computed in a setting with parameterized dynamics and finite horizons. We envision fruitful advances in these directions will further consolidate our multi-model perspective on online decision-making.

All the aforementioned works provide a comprehensive ground for online decision-making. Nonetheless, achieving sublinear regrets in an online regime encompassing a broad class of nonlinear dynamics models remains a critical challenge. In this article, we adopt a multi-model perspective and provide a suite of algorithms that identify the best candidate model and apply a certainty-equivalent policy, all equipped with nonasymptotic policy-regret guarantees. These guarantees feature favorable scalings with the time horizon, state dimension, and complexity of the function class.

B Numerical example

We present results of a numerical simulation to illustrate our algorithms. To simplify the presentation we consider a linear time-invariant dynamical system of dimension $d_{\mathrm{x}} = 20$ and $d_{\mathrm{u}} = 5$ and apply the two algorithms Alg. 1 and Alg. 3. The stage cost is $l(x, u) = |x|^2 + |u|^2$. The dynamics f (unknown to the decision-maker) consist of five four-dimensional leaky integrators of the type $x^i_{k+1} = 0.8x^i_k + x^{i+1}_k$, $i = 1, \dots, 3$. The dynamics are relatively challenging for control, as there is a lag of five steps until a change in the input affects x^1_k. The above dynamics are compactly written as $f(x, u) = Ax + Bu$, where $x \in \mathbb{R}^{d_{\mathrm{x}}}$ is the state, $u \in \mathbb{R}^{d_{\mathrm{u}}}$ is the input, $A = I_5 \otimes A_0, B = I_5 \otimes B_0$ are system matrices, I_5 is an identity matrix of size 5, $\otimes$ denotes the Kronecker product, and

$$A_0 = \begin{bmatrix} 0.8 & 1 & 0 & 0 \\ 0 & 0.8 & 1 & 0 \\ 0 & 0 & 0.8 & 1 \\ 0 & 0 & 0 & 0.8 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

It is assumed that the elements of the matrices A and B that define the dynamics are unknown with respect to an absolute error of 0.1 and relative error of 20%, which gives rise to a large set of possible models including some open-loop unstable ones. For instance, the jk-th element a_{jk} of A, is known to be in the range $[0.8a_{jk} - 0.1, 1.2a_{jk} + 0.1]$.

B.1 Setting `S1`

Set-up: We generate m candidate models $f^i(x, u) = A^i x + B^i u$ at random, whereby each element of A^i, B^i is randomly drawn from the known parameter range, e.g., the jk-th element a^i_{jk} of A^i is sampled from the uniform distribution over $[0.8a_{jk} - 0.1, 1.2a_{jk} + 0.1]$. The feedback policy μ^i related to candidate model f^i is

$$\mu^i(x) = -K^i x, \quad \text{where} \quad K^i = (I + {B^i}^\top P B^i)^{-1} {B^i}^\top P^i A^i,$$

and $P^i \in \mathbb{R}^{d_{\mathrm{x}} \times d_{\mathrm{x}}}$ is a positive definite matrix satisfying the discrete-time algebraic Riccati equation involving A^i and B^i [10]. The policies μ^i are computed through the built-in `dlqr` command in MATLAB and the settings of Alg. 1 were chosen as specified in Thm. 3.2, that is

$$\eta = 10, \quad \sigma^2_{\mathrm{u}k} = \frac{10}{\eta d_{\mathrm{u}} M} \left(\frac{2}{\lceil k/M \rceil} + \frac{\ln(2m)}{\lceil k/M \rceil)^2} \right), \quad M = 2.$$

Assumptions of Thm. 3.2: Ass. 1-3 are clearly satisfied:

- The cost-to-go function V is given by $V(x) = x^\top P x$, where P satisfies the discrete-time algebraic Riccati equation involving A and B. Hence, Ass. 1 is satisfied with $\gamma = \mathrm{tr}(P)\sigma^2$, $L_\mathrm{u} = \bar{\sigma}(P)$, where $\bar{\sigma}$ denotes the maximum singular value of a matrix.
- Ass. 2 is satisfied with $\underline{c}_\mathrm{V} = 0$, $\underline{L}_\mathrm{V} = \underline{\sigma}(P)$, where $\underline{\sigma}$ is the minimum singular value of a matrix.
- Ass. 3 is satisfied (for any $M > 0$) with
$$c_\mathrm{e} = \min_{i \in \{1,\dots,m\}} |B^i - B|_\mathrm{F}^2,$$
for example, as can be seen from (4). Larger values of c_e can be achieved when choosing M larger and factoring in the controllability Gramian W_k^c.

Please note that the constants $c_\mathrm{e}, \underline{c}_\mathrm{V}, \underline{L}_\mathrm{V}, \gamma$ only appear in the resulting policy-regret bounds and are not needed for running Alg. 1.

Computational complexity: All experiments run on a Laptop (Intel Core i7 processor with 2.30GHz; 32 GB of random access memory) and are executed in a few minutes even when increasing the number of candidate model up to 10,000. The offline computation of the policies μ^i has cost $\mathcal{O}(d_\mathrm{x}^3 + d_\mathrm{u}^2 d_\mathrm{x})$, the online computation in Alg. 1 is $\mathcal{O}(d_\mathrm{x} m + d_\mathrm{u} d_\mathrm{x})$.

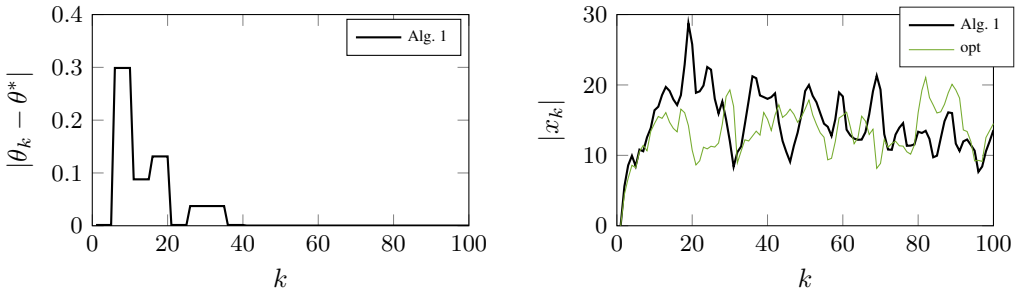
B.2 Setting `S2`

Set-up: The parameter space $\Omega \subset \mathbb{R}^p$ with $p = d_\mathrm{x}^2 + d_\mathrm{x} d_\mathrm{u} = 500$ covers the entire parameter range
$$\Omega = \{(\bar{A}, \bar{B}) \in \mathbb{R}^p \mid \bar{a}_{jk} \in [0.8 a_{jk} - 0.1, 1.2 a_{jk} + 0.1], \quad \bar{b}_{jk} \in [0.8 b_{jk} - 0.1, 1.2 b_{jk} + 0.1]\},$$
where we slightly abuse notation to avoid distinguishing between different ways of stacking vectors and matrices (we will frequently do so in the following as the stacking is clear from context). For a given set of matrices $(\bar{A}, \bar{B}) \in \Omega$ the corresponding feedback controller $\bar{\mu}(x) = -\bar{K}x$ is given as in setting `S1` and requires solving the discrete-time algebraic Riccati equation involving $\bar{A}, \bar{B}$. As before, the feedback policies are computed through the built-in `dlqr` command in MATLAB and the settings of Alg. 3 are chosen as specified in Thm. E.1, that is,
$$\eta = 10, \quad \epsilon = p/T, \quad \sigma_{\mathrm{u}k}^2 = \frac{10}{\eta d_\mathrm{u} M \epsilon} \left(\frac{2}{\lceil k/M \rceil} + \frac{p}{\lceil k/M \rceil^2} \right), \quad M = 5.$$
The posterior distribution over models in Alg. 3 is updated by a recursive least squares algorithm and we set $b \to \infty$. The recursive implementation has the advantage that reasonable estimates of A and B are already provided in the first p steps, which is important for the initial transient behavior.

Assumptions of Thm. E.1: Ass. 2, Ass. 6, and Ass. 7 are satisfied:

- The cost-to-go function is given by $V(x) = x^\top P x$, where P satisfies the discrete-time algebraic Riccati equation involving A and B. One can easily show that Ass. 7 is satisfied by applying Prop. D.1. (As pointed out in [46, Prop. 6], for example, the policies μ_θ are continuously dependent on the system parameters $\theta = (A, B)$.)
- Ass. 7 is satisfied with $\underline{c}_\mathrm{V} = 0$, $\underline{L}_\mathrm{V} = \underline{\sigma}(P)$.
- Ass. 6 is satisfied in view of (4) in Sec. 3, provided that $M = 5$, which ensures that the controllability Gramian W_k^c is full rank for any feedback gain $\bar{K}$. (The dynamics A, B give rise to decoupled four-dimensional leaky integrators, hence the Gramian W_4^c defined in Sec. 3 is guaranteed to be full rank.)

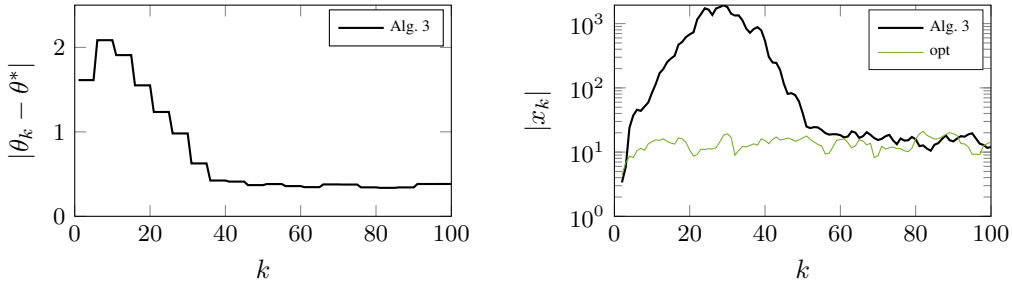
Computational complexity: Sampling the parameter θ_k in Alg. 3 amounts to sampling from a truncated Gaussian, where the mean and covariance of the Gaussian are computed via the recursive least squares algorithm. The computation of mean and covariance can be done in at most $\mathcal{O}(d_\mathrm{x}^2 + d_\mathrm{x} d_\mathrm{u})$ elementary operations at each iteration k. We then sample θ_k by applying rejection sampling (although much more efficient approaches could be applied). The policy μ_{θ_k} is then computed by solving the corresponding discrete-time algebraic Riccati equation, which requires at most $\mathcal{O}(d_\mathrm{x}^3 + d_\mathrm{u}^2 d_\mathrm{x})$ elementary operations.



(a) parameter error

(b) two-norm of state trajectory

Figure 1: The first panel shows the evolution of the parameter error of Alg. 1, while the second panel shows the evolution of the two norm of the states. The green line indicates the performance of the optimal (steady-state) policy on a different realization of n_k. We note that near-optimal steady state performance is reached in about 25 steps.



(a) parameter error

(b) two-norm of state trajectory

Figure 2: The first panel shows the evolution of the parameter error of Alg. 3, while the second panel shows the evolution of the two norm of the states. The green line indicates the performance of the optimal (steady-state) policy on a different realization of n_k. Compared to Fig. 1b the overshoot is larger and the convergence to near-optimal performance requires about 60 iterations.

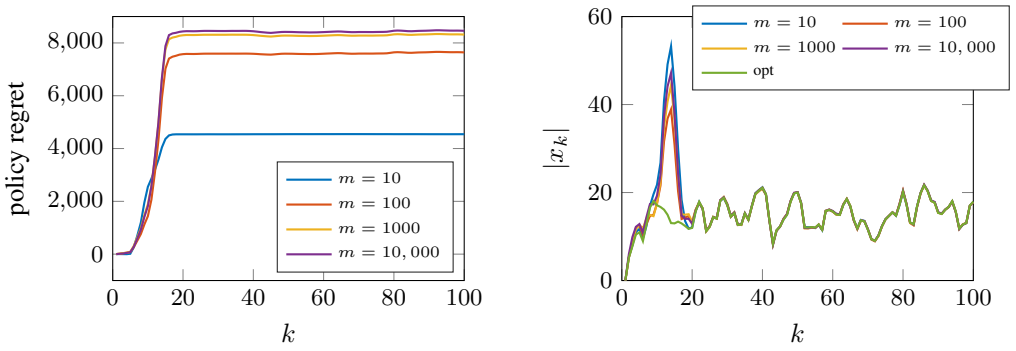
B.3 Results

Simulation results for the setting `S1` are shown in Fig. 1, whereas the results for setting `S2` are shown in Fig. 2. A rapid convergence to near-optimal steady-state behavior can be observed in both cases. We note that the space of parameters in Alg. 3 is uncountable compared to Alg. 1 and therefore Alg. 3 takes about twice as long to converge. Alg. 1 achieves optimal steady-state performance very quickly (in about twenty steps). We therefore believe that Alg. 1 provides an algorithmic paradigm that is applicable to many emerging real-world machine learning and engineering challenges, including the control of intelligent transportation systems or automated supply chains.

To showcase the scalability of our algorithms, we provide comparison results when the number of models m in Alg. 1 is increased from 10 to $10,000$. We perform 40 independent realizations of Alg. 1 for each value of m and show the corresponding policy regret in Fig. 3a (averaged over the 40 realizations). Once again we observe a fast initial transient phase after which the policy regret stabilizes and near-optimal steady-state performance is achieved. Fig. 3b shows the corresponding evolution of the two-norm of the state trajectory on a single realization. The plots highlight that Alg. 1 scales favorably in the number of models.

C Details of Sec. 3.1

We first state and prove two intermediate lemmas that are used in the proof of Prop. 3.1. The two lemmas express the fact that the larger the expected model deviation $f^i - f$ (accumulated over the past steps), the smaller the corresponding probability of selecting model i.



(a) policy regret

(b) The two-norm of the state trajectory

Figure 3: The first panel shows the change in policy regret of Alg. 1 when varying m. The second panel shows the evolution of the two norm of the state trajectory. We note that the behavior is consistent over the different values of m (from 10 to $10,000$, which amounts to three orders of magnitude). The green line indicates the performance of the optimal (steady-state) policy on the *same* realization of n_k.

Lemma C.1 *For any step size $\eta > 0$ it holds that*

$$\Pr(i_k = i) \leq \mathrm{E}[e^{-\eta(s_k^i - s_k^j)}],$$

for all s_k^j (and in particular for $s_k^j = s_k^*$ corresponding to f).

Proof We note that p_k^i is given by

$$p_k^i = \frac{e^{-\eta s_k^i}}{\sum_{j=1}^m e^{-\eta s_k^j}} \leq e^{-\eta(s_k^i - \bar{s}_k)} \leq e^{-\eta(s_k^i - s_k^j)},$$

for all $j \in \{1, \ldots, m\}$, where $\bar{s}_k = \min_{i \in \{1,\ldots,m\}} s_k^i$. In addition, it holds that

$$\Pr(i_k = i) = \mathrm{E}[\mathbb{1}_{i_k = i}] = \mathrm{E}[\mathrm{E}[\mathbb{1}_{i_k=i} | x_1, \ldots, x_k, u_k, n_k]] = \mathrm{E}[p_k^i],$$

where $\mathbb{1}$ denotes the indicator function, which yields the desired result. □

Lemma C.2 *Let*

$$l_k^i := \frac{|x_{k+1} - f^i(x_k, u_k)|^2}{1 + |(x_k, u_k)|^2 / b^2},$$

where $b > 0$ is constant. Let $\mathcal{F}_k$ denote the collection of random variables $x_j, u_j, i_j, n_{j-1}, n_{\mathrm{u}j}$ up to time k. Then, the following bound holds for all $0 < \eta \leq \min\{1/(4\sigma^2), 1/(2L^2b^2)\}$ and for all $1 \leq q \leq k$:

$$\mathrm{E}[e^{-\eta(l_k^i - l_k^*)} | \mathcal{F}_q] \leq \exp\left(-\frac{\eta}{4}\mathrm{E}\left[\frac{|f - f^i|^2}{1 + |(x_k, u_k)|^2/b^2} \middle| \mathcal{F}_q\right]\right),$$

where f stands for $f(x_k, u_k)$ and f^i for $f^i(x_k, u_k)$, and where l_k^* corresponds to the loss of the candidate f.

Proof We note that $|x_{k+1} - f^i(x_k, u_k)|^2$ can be expressed as

$$|f - f^i + n_k|^2 = |f - f^i|^2 + 2n_k^\top (f - f^i) + n_k^\top n_k,$$

and, as a result, $l_k^i - l_k^*$ is given by

$$\frac{|f - f^i|^2 + 2n_k^\top (f - f^i)}{1 + |(x_k, u_k)|^2/b^2}.$$

Hence, conditioned on x_k, u_k, the randomness in $l_k^i - l_k^*$ is solely due to $n_k^\top (f - f^i)$, which describes a sum of d_{x} independent Gaussian random variables, weighted by the components of $f - f^i$. As

a result, we exploit the closed-form expression for the moment generating function of a Gaussian, which yields

$$\mathrm{E}[e^{-\eta(l_k^i - l_k^*)}|x_k, u_k] \leq \exp\left(\frac{2\eta^2\sigma^2|f - f^i|^2}{1 + |(x_k, u_k)|^2/b^2} - \eta\frac{|f - f^i|^2}{1 + |(x_k, u_k)|^2/b^2}\right).$$

Thus, for $\eta \leq 1/(4\sigma^2)$, the following bound holds

$$\mathrm{E}[e^{-\eta(l_k^i - l_k^*)}|\mathcal{F}_q] \leq \mathrm{E}\left[\exp\left(-\frac{\eta|f - f^i|^2/2}{1 + |(x_k, u_k)|^2/b^2}\right)\Big|\mathcal{F}_q\right].$$

As a result of the Lipschitz continuity of f and f^i, the term

$$0 \leq \frac{|f - f^i|^2}{1 + |(x_k, u_k)|^2/b^2} \leq 4L^2b^2 \tag{7}$$

is bounded. When deriving the previous inequality we used the fact that $|f^i(x_k, u_k) - f(x_k, u_k)| \leq |f^i(x_k, u_k) - f^i(0, 0)| + |f(x_k, u_k) - f(0, 0)| \leq 2L|(x_k, u_k)|$ by Lipschitz continuity of f and f^i.[2] We can therefore apply a "Poissonian" inequality [see, e.g., 13, App. A], which yields

$$\mathrm{E}[e^{-\eta(l_k^i - l_k^*)}|\mathcal{F}_q] \leq \exp\left(\frac{(e^{-2\eta L^2b^2} - 1)}{4L^2b^2}\mathrm{E}\left[\frac{|f - f^i|^2}{1 + |(x_k, u_k)|^2/b^2}\Big|\mathcal{F}_q\right]\right),$$

for all $\eta \leq 1/(4\sigma^2)$. The desired result follows from the fact that $(e^{-2\eta L^2b^2} - 1)/(4L^2b^2) \leq -\eta/4$ for all $\eta \leq \min\{1/(4\sigma^2), 1/(2L^2b^2)\}$. □

C.1 Proof of Prop. 3.1

We first consider the iterations $k = k'M + 1$, for $k' = 0, 1, \ldots$. These are the iterations k where the random variable i_k is updated according to the distribution p_k^i (conditional on x_k, u_k). It will be useful to introduce the variables $\bar{l}_{k'}^i$ as follows:

$$\bar{l}_{k'}^i = \sum_{j=1}^{M} l_{k'M+j}^i,$$

which corresponds to a sum of the variables l_k^i over M steps. Let $\mathcal{F}_{k'}$ denote the collection of all random variables $(x_k, u_k, i_k, n_{k-1}, n_{\mathrm{u}k})$ up to time $k = k'M + 1$. We condition on $\mathcal{F}_{k'-1}$ and conclude from Lemma C.2

$$\begin{aligned}
\mathrm{E}[e^{-\eta(\bar{l}_{k'}^i - \bar{l}_{k'}^*)}|\mathcal{F}_{k'-1}] &= \mathrm{E}[e^{-\eta\sum_{j=1}^{M}(l_{k'M+j}^i - l_{k'M+j}^*)}|\mathcal{F}_{k'-1}] \\
&\leq \prod_{j=1}^{M}(\mathrm{E}[e^{-\eta M(l_{k'M+j}^i - l_{k'M+j}^*)}|\mathcal{F}_{k'-1}])^{1/M} \\
&\leq \prod_{j=1}^{M}(e^{-\frac{\eta M}{4}\mathrm{E}[\frac{|f - f^i|^2}{1 + |(x_k, u_k)|^2/b^2}|\mathcal{F}_{k'-1}]})^{1/M} \\
&\leq \exp\left(-\frac{\eta M c_{\mathrm{e}} d_{\mathrm{u}}}{4}\sigma_{\mathrm{u}(k-1)}^2\right),
\end{aligned} \tag{8}$$

where we have used Hölder's inequality for the first inequality, Lemma C.2 for the second inequality, and Ass. 3 for the third inequality. As a result, by unrolling the recursion for $k' - 1$, $k' - 2$, …, we conclude that

$$\mathrm{E}[e^{-\eta(s_k^i - s_k^*)}] \leq \exp\left(-\frac{\eta M c_{\mathrm{e}} d_{\mathrm{u}}}{4}\sum_{j=1}^{k'}\sigma_{\mathrm{u}(Mj)}^2\right).$$

[2] We stated the inequality assuming $f(0, 0) = f^i(0, 0) = 0$. In the more general situation the upper bound $2|f^i(0, 0) - f(0, 0)|^2 + 8L^2b^2$ applies in (7).

By virtue of Lemma C.1, this implies

$$\Pr(i_k = i) \leq \exp\left(-\frac{\eta M c_{\mathrm{e}} d_{\mathrm{u}}}{4} \sum_{j=1}^{k'} \sigma_{\mathrm{u}(Mj)}^2\right).$$

The bound holds in fact also for $k+1, k+2$, until $k+(M-1)$, since, by definition, $i_k = i_{k+1} = \cdots = i_{k+(M-1)}$. This proves the first bound of Prop. 3.1.

It remains to derive the second bound, which is done by approximating the sum over $\sigma_{\mathrm{u}k}^2$ from below. We find

$$\frac{\eta M c_{\mathrm{e}} d_{\mathrm{u}}}{4} \sum_{j=1}^{k'} \sigma_{\mathrm{u}(Mj)}^2 = \sum_{j=1}^{k'} \left(\frac{2}{j} + \frac{\ln(m)}{j^2}\right) \geq \int_1^{k'} \frac{2}{j}\,\mathrm{d}j + \ln(m) \geq 2\ln(k') + \ln(m),$$

for $k' \geq 1$. This concludes that $\Pr(i_k = i) \leq 1/(mk'^2)$, due to the fact that $m \geq 1$. We further note that $k' = (k-1)/M$ by our choice of k. However, i_k remains unchanged for the M next iterations, and therefore

$$\Pr(i_{k+M-1} = i) \leq \frac{M^2}{m(k-1)^2},$$

which holds for all $k \geq 2$ and $i \neq i^*$. This implies $\Pr(i_k = i) \leq M^2/(m(k-M)^2)$ for all $k \geq M+1$ by a change of variables. Applying a union bound yields the second inequality of Prop. 3.1, i.e.,

$$\Pr(i_k \neq i^*) \leq \sum_{i \neq i^*} \Pr(i_k = i) \leq \frac{M^2}{(k-M)^2}.$$

□

C.2 Proof of Thm. 3.2

We will use V as a Lyapunov function and have

$$\mathrm{E}[V(x_{k+1})] = \mathrm{E}[V(x_{k+1})|i_k \neq i^*]\Pr(i_k \neq i^*) + \mathrm{E}[V(x_{k+1})|i_k = i^*]\Pr(i_k = i^*). \tag{9}$$

The first term can be further simplified in view of Lemma C.3, which yields

$$\begin{aligned}\mathrm{E}[V(x_{k+1})|x_k, i_k \neq i^*] \leq c_2 V(x_k) + \bar{L}_{\mathrm{V}} d_{\mathrm{x}} \sigma^2/2 + c_{\mathrm{o}} \\ - \mathrm{E}[l(x,u_k)|x_k, i_k \neq i^*] + (\bar{L}_{\mathrm{V}} L^2 + \bar{L}_{\mathrm{l}}) d_{\mathrm{u}} \sigma_{\mathrm{u}k}^2.\end{aligned} \tag{10}$$

The second term in (9) is bounded as a result of the Bellman-type inequality (3) for the policy μ (the policy that corresponds to V). It will be convenient to rewrite the bound (3) in the following way:

$$\mathrm{E}[V(f(x,u)+n)] \leq V(x) - \mathrm{E}[l(x,u)] + q(x) + d_{\mathrm{u}} L_{\mathrm{u}} \sigma_{\mathrm{u}}^2,$$

where $u = \mu(x) + n_{\mathrm{u}}$, (n_{u}, n) are independent with $n_{\mathrm{u}} \sim \mathcal{N}(0, \sigma_{\mathrm{u}}^2 I)$, $n \sim \mathcal{N}(0, \sigma^2 I)$, and $q(x)$ is chosen such that $q(x) \leq \gamma$ and $-\mathrm{E}[l(x,u)] + q(x) \leq 0$.[3] The function $q(x)$ is introduced to account for the fact that the policy μ might in principle also achieve a running cost $\mathrm{E}[l(x,u)] \leq \gamma$ in the short term, since γ captures only the steady-state performance. As a result, we obtain

$$\mathrm{E}[V(x_{k+1})|i_k = i^*] \leq -\mathrm{E}[l(x_k,u_k)|i_k = i^*] + \mathrm{E}[V(x_k)] + \gamma_k + d_{\mathrm{u}} L_{\mathrm{u}} \sigma_{\mathrm{u}k}^2, \tag{11}$$

where $\gamma_k := \mathrm{E}[q(x_k)]$. By combining (10) and (11) with (9) we arrive at

$$\begin{aligned}\mathrm{E}[V(x_{k+1})] \leq \mathrm{E}[V(x_k)](c_2 \Pr(i_k \neq i^*) + 1) + \gamma_k - \mathrm{E}[l(x_k,u_k)] \\ + \bar{L}_{\mathrm{u}} d_{\mathrm{u}} \sigma_{\mathrm{u}k}^2 + (\bar{L}_{\mathrm{V}} d_{\mathrm{x}} \sigma^2/2 + c_{\mathrm{o}}) \Pr(i_k \neq i^*),\end{aligned}$$

where $\bar{L}_{\mathrm{u}} := \bar{L}_{\mathrm{V}} L^2 + \bar{L}_{\mathrm{l}} + L_{\mathrm{u}}$. As a result of Prop. 3.1, we know that $\Pr(i_k \neq i^*) \leq M^2/(k-M)^2$ for $k \geq M+1$. We further note that $\mathrm{E}[l(x_k,u_k)] \geq \gamma_k$ and $\gamma_k \leq \gamma$ (by our choice of γ_k). We now invoke Lemma C.4 and conclude

$$\sum_{k=1}^{N} (\mathrm{E}[l(x_k,u_k)] - \gamma_k) \leq c_\alpha \bar{L}_{\mathrm{u}} d_{\mathrm{u}} \sum_{k=1}^{N} \sigma_{\mathrm{u}k}^2 + c_\alpha (\bar{L}_{\mathrm{V}} d_{\mathrm{x}} \sigma^2/2 + c_{\mathrm{o}}) \sum_{k=1}^{N} \Pr(i_k \neq i^*),$$

[3] This can by achieved by setting $q(x) = \mathrm{E}[V(f(x,\mu(x)+n_{\mathrm{u}})+n)] - V(x) + \mathrm{E}[l(x,\mu(x)+n_{\mathrm{u}})] - d_{\mathrm{u}}\sigma_{\mathrm{u}}^2 L_{\mathrm{u}}$ for $\gamma \geq \mathrm{E}[l(x,\mu(x)+n_{\mathrm{u}})]$ and $q(x) = \gamma$ otherwise.

where we have used the fact that $V(x_1) = 0$ and the following calculation

$$\prod_{k=1}^{\infty}(c_2\mathrm{Pr}(i_k \neq i^*) + 1) \leq e^{c_2 \sum_{k=1}^{\infty} \mathrm{Pr}(i_k \neq i^*)} \leq e^{3Mc_2} = c_\alpha,$$

due to the fact that

$$\begin{aligned}\sum_{k=1}^{\infty}\mathrm{Pr}(i_k \neq i^*) &\leq 2M - 1 + \sum_{k=2M}^{\infty}\frac{M^2}{(k-M)^2} \\ &\leq 2M + \int_{2M}^{\infty}\frac{M^2}{(k-M)^2}\mathrm{d}k \leq 3M.\end{aligned}$$

Moreover, we bound the sum over $\sigma_{\mathrm{u}k}^2$ as follows

$$\begin{aligned}\sum_{k=1}^{N}\sigma_{\mathrm{u}k}^2 &= \frac{4}{d_{\mathrm{u}}\eta c_{\mathrm{e}}}\sum_{k=1}^{N}\left(\frac{2}{M\lceil k/M\rceil} + \frac{M\ln(m)}{(M\lceil k/M\rceil)^2}\right) \\ &\leq \frac{4}{d_{\mathrm{u}}\eta c_{\mathrm{e}}}\sum_{k=1}^{N}\left(\frac{2}{k} + \frac{M\ln(m)}{k^2}\right) \\ &\leq \frac{8}{d_{\mathrm{u}}\eta c_{\mathrm{e}}}(1 + \ln(N-1) + M\ln(m)).\end{aligned}$$

Combining the previous inequalities and taking advantage of the fact that $\gamma_k \leq \gamma$ yields the desired result. □

C.3 Supporting lemmas in the proof of Thm. 3.2

This section contains two lemmas that support the proof of Thm. 3.2.

Lemma C.3 *Let Ass. 2 be satisfied. Then, there exist two constants $c_2, c_{\mathrm{o}} \geq 0$ such that*

$$\mathrm{E}[V(f(x,\mu^i(x)+n_{\mathrm{u}})+n)] \leq c_2 V(x) - \mathrm{E}[l(x,\mu^i(x)+n_{\mathrm{u}})] + c_{\mathrm{o}} + (\bar{L}_{\mathrm{V}}L^2+\bar{L}_{\mathrm{l}})d_{\mathrm{u}}\sigma_{\mathrm{u}}^2 + \frac{\bar{L}_{\mathrm{V}}}{2}d_{\mathrm{x}}\sigma^2,$$

for all $x \in \mathbb{R}^{d_{\mathrm{x}}}$, $\sigma_{\mathrm{u}} > 0$, and $i \in \{1,\ldots,m\}$, where the constant c_2 is given by

$$c_2 = (8L^2\bar{L}_{\mathrm{V}}(1+L_\mu)^2 + 2\bar{L}_{\mathrm{l}}(1+2L_\mu^2))/\underline{L}_{\mathrm{V}},$$

and c_{o} can be expressed as an explicit function of $\max_{i\in[m]}|\mu^i(0)|$, $V(0)$, $|\nabla V(0)|$, $l(0,0)$, $|\nabla l(0,0)|$, $|f(0,\mu(0))|$, and $\underline{c}_{\mathrm{V}}$. The random variables n_{u} and n are independent and satisfy $n \sim \mathcal{N}(0,\sigma^2 I)$, $n_{\mathrm{u}} \sim \mathcal{N}(0,\sigma_{\mathrm{u}}^2 I)$.

Proof We exploit smoothness of V to bound $\mathrm{E}[V(f(x,\mu^i(x)+n_{\mathrm{u}})+n)]$ by

$$\mathrm{E}[V(f(x,\mu^i(x)+n_{\mathrm{u}})] + \frac{\bar{L}_{\mathrm{V}}}{2}d_{\mathrm{x}}\sigma^2,$$

where we used the fact that the term linear in n vanishes in expectation. We further note that the term $V(f(x,\mu^i(x)+n_{\mathrm{u}}))$ can be bounded in a similar way:

$$V(f(x,\mu^i(x)+n_{\mathrm{u}})) \leq V(f(x,\mu^i(x))) + \nabla V(f(x,\mu^i(x)))^{\top}\nabla_u f(\xi)n_{\mathrm{u}} + \frac{\bar{L}_{\mathrm{V}}}{2}|\nabla_u f(\xi)n_{\mathrm{u}}|^2,$$

where we applied the mean value theorem to rewrite $f(x,\mu^i(x)+n_{\mathrm{u}}) - f(x,\mu^i(x))$ as $\nabla_u f(\xi)n_{\mathrm{u}}$ for some ξ (dependent on n_{u}). By applying Young's inequality and taking advantage of the fact that $\nabla_u f$ is bounded above we arrive at

$$\mathrm{E}[V(f(x,\mu^i(x)+n_{\mathrm{u}}))] \leq V(f(x,\mu^i(x))) + \frac{1}{2\bar{L}_{\mathrm{V}}}|\nabla V(f(x,\mu^i(x)))|^2 + \bar{L}_{\mathrm{V}}L^2 d_{\mathrm{u}}\sigma_{\mathrm{u}}^2.$$

Due to smoothness, V is guaranteed to satisfy

$$|\nabla V(x)| \leq c_{\mathrm{o}1} + \bar{L}_{\mathrm{V}}|x|, \quad V(x) \leq c_{\mathrm{o}2} + \bar{L}_{\mathrm{V}}|x|^2,$$

where $c_{\mathrm{o}1} = |\nabla V(0)|$, and the constant $c_{\mathrm{o}2} \geq 0$ is similarly related to $|\nabla V(0)|$ and $V(0)$. As a result, we obtain the following upper bound on $\mathrm{E}[V(f(x, \mu^i(x) + n_{\mathrm{u}}))]$:

$$\mathrm{E}[V(f(x, \mu^i(x) + n_{\mathrm{u}}))] \leq c_{\mathrm{o}2} + \frac{c_{\mathrm{o}1}^2}{\bar{L}_{\mathrm{V}}} + 2\bar{L}_{\mathrm{V}}|f(x, \mu^i(x))|^2 + \bar{L}_{\mathrm{V}} L^2 d_{\mathrm{u}} \sigma_{\mathrm{u}}^2.$$

The fact that $f(x, \mu^i(x))$ is $L(1 + L_\mu)$ Lipschitz can be used to conclude that $|f(x, \mu^i(x))|^2 \leq c_{\mathrm{o}3} + 2L^2(1 + L_\mu)^2|x|^2$, where $c_{\mathrm{o}3} = 2|f(0, \mu(0))|^2$, which, in turn, yields the following upper bound

$$\mathrm{E}[V(f(x, \mu^i(x) + n_{\mathrm{u}}))] \leq c_{\mathrm{o}2} + \frac{c_{\mathrm{o}1}^2}{\bar{L}_{\mathrm{V}}} + 2\bar{L}_{\mathrm{V}} c_{\mathrm{o}3} + 4\bar{L}_{\mathrm{V}} L^2 (1 + L_\mu)^2 |x|^2 + \bar{L}_{\mathrm{V}} L^2 d_{\mathrm{u}} \sigma_{\mathrm{u}}^2.$$

We further note that the fact that l is $\bar{L}_{\mathrm{l}}$ smooth and $l(x, u) \geq 0$ implies $l(x, u) \leq c_{\mathrm{o}4} + \bar{L}_{\mathrm{l}}(|x|^2 + |u|^2)$ and therefore

$$\mathrm{E}[l(x, \mu^i(x) + n_{\mathrm{u}})] \leq c_{\mathrm{o}5} + \bar{L}_{\mathrm{l}} d_{\mathrm{u}} \sigma_{\mathrm{u}}^2 + \bar{L}_{\mathrm{l}}(1 + 2L_\mu^2)|x|^2,$$

where $c_{\mathrm{o}5} \geq 0$ is related to $\max_{i \in [m]} |\mu^i(0)|$, $l(0, 0)$, and $|\nabla l(0, 0)|$. Combining the previous two inequalities results in

$$\mathrm{E}[V(f(x, \mu^i(x) + n_{\mathrm{u}}))] \leq c_{\mathrm{o}6} + \frac{c_2 \underline{L}_{\mathrm{V}}}{2}|x|^2 - \mathrm{E}[l(x, \mu^i(x) + n_{\mathrm{u}})] + (\bar{L}_{\mathrm{V}} L^2 + \bar{L}_{\mathrm{l}}) d_{\mathrm{u}} \sigma_{\mathrm{u}}^2,$$

where $c_{\mathrm{o}6} \geq 0$ is constant and can be expressed as a function of $\max_{i \in [m]} |\mu^i(0)|$, $V(0)$, $|\nabla V(0)|$, $l(0, 0)$, $|\nabla l(0, 0)|$, and $|f(0, \mu(0))|$. The results follows by inserting $\underline{L}_{\mathrm{V}}|x|^2/2 \leq V(x) + \underline{c}_V$ in the previous inequality. □

Lemma C.4 *Let the sequence*

$$V_{k+1} \leq (1 + \alpha_k) V_k + g_k^+ - g_k^-, \quad V_k \geq 0,$$

be given, where $k = 1, 2, \ldots$, $\alpha_k \geq 0$ $g_k^+ \geq 0$, and $g_k^- \geq 0$ are arbitrary sequences such that $c_\alpha := \prod_{k=1}^{\infty}(1 + \alpha_k) < \infty$. Then, the following holds for all $N \geq 1$

$$\sum_{j=1}^{N} g_j^- \leq c_\alpha \left(\sum_{j=1}^{N} g_j^+ + V_1 \right).$$

Proof By unrolling the linear difference equation we obtain

$$\begin{aligned} V_{N+1} &\leq \prod_{k=1}^{N}(1 + \alpha_k) V_1 + \sum_{i=1}^{N}(g_i^+ - g_i^-) \prod_{j=i+1}^{N}(1 + \alpha_j) \\ &\leq c_\alpha \left(V_1 + \sum_{i=1}^{N} g_i^+ \right) - \sum_{i=1}^{N} g_i^-, \end{aligned}$$

where we exploited the fact that $\prod_{k=1}^{N}(1 + \alpha_k) < c_\alpha < \infty$. □

C.4 Finite second moment

Corollary C.5 *Let Ass. 2 be satisfied, let $\sigma_{\mathrm{u}k}^2$ be as in Prop. 3.1, let $l(x, u) \geq \underline{L}_{\mathrm{l}}|x|^2/2$ for some constant $\underline{L}_{\mathrm{l}} > 0$, and $\eta \leq \min\{1/(4M\sigma^2), 1/(2ML^2b^2)\}$. Let Ass. 3 be satisfied for at least the first*

$$k_0 := \left\lceil M \left(1 + \sqrt{2\bar{L}_{\mathrm{V}} c_2 / \underline{L}_{\mathrm{l}}} \right) \right\rceil$$

steps. Then, it holds that

$$\mathrm{E}[V(x_k)] \leq \max\{c_3, c_4\}, \quad \forall k \geq 1,$$

with

$$\begin{aligned} c_3 &= c_2^{k_0} k_0 (\bar{L}_{\mathrm{V}} d_{\mathrm{x}} \sigma^2 + c_{\mathrm{o}} + (\bar{L}_{\mathrm{V}} L^2 + \bar{L}_{\mathrm{l}}) d_{\mathrm{u}} \sigma_{\mathrm{u}1}^2), \\ c_4 &= \frac{2\bar{L}_{\mathrm{V}}}{\underline{L}_{\mathrm{l}}} (\gamma + (\bar{L}_{\mathrm{V}} L^2 + \bar{L}_{\mathrm{l}} + L_{\mathrm{u}}) d_{\mathrm{u}} \sigma_{\mathrm{u}1}^2 + \bar{L}_{\mathrm{V}} d_{\mathrm{x}} \sigma^2 + c_{\mathrm{o}}). \end{aligned}$$

Proof We conclude from Prop. 3.1 that $\Pr(i_k \neq i^*)$ is bounded by

$$\Pr(i_k \neq i^*) \leq \frac{M^2}{(k-M)^2} \leq \frac{\underline{L}_\mathrm{l}}{2\bar{L}_\mathrm{V} c_2}, \tag{12}$$

for all $k \geq k_0$. It is important to note that persistence of excitation is only required to hold for k_0 steps, as, by our choice of η, $\Pr(i_k \neq i^*)$ is monotonically decreasing (see proof of Prop. 3.1). By Lemma C.3 we conclude that over the first k_0 steps the following holds

$$\mathrm{E}[V(x_{k+1})] \leq c_2 \mathrm{E}[V(x_k)] + \bar{L}_\mathrm{V} d_\mathrm{x} \sigma^2 + c_\mathrm{o} + (\bar{L}_\mathrm{V} L^2 + \bar{L}_\mathrm{l}) d_\mathrm{u} \sigma_{\mathrm{u}k}^2,$$

which implies that

$$\mathrm{E}[V(x_k)] \leq c_2^{k_0} k_0 (\bar{L}_\mathrm{V} d_\mathrm{x} \sigma^2 + c_\mathrm{o} + (\bar{L}_\mathrm{V} L^2 + \bar{L}_\mathrm{l}) d_\mathrm{u} \sigma_{\mathrm{u}1}^2),$$

for all $k \leq k_0 + 1$, where we have exploited that $\sigma_{\mathrm{u}k}$ is monotonically decreasing.

By following the same reasoning (case distinction between $i_k = i^*$ and $i_k \neq i^*$) as in the proof of Thm. 3.2 we arrive at

$$\mathrm{E}[V(x_{k+1})] \;\leq\; \mathrm{E}[V(x_k)](\underline{L}_\mathrm{l}/(2\bar{L}_\mathrm{V}) + 1) + \gamma - \mathrm{E}[l(x_k, u_k)] + \bar{L}_\mathrm{u} d_\mathrm{u} \sigma_{\mathrm{u}1}^2 + \bar{L}_\mathrm{V} d_\mathrm{x} \sigma^2 + c_\mathrm{o},$$

for all $k \geq k_0$, where we have used inequality (12) to bound $\Pr(i_k \neq i^*)$, $\gamma_k \leq \gamma$, and the fact that $\sigma_{\mathrm{u}k}$ is decreasing. The constant $\bar{L}_\mathrm{u}$ is given by $\bar{L}_\mathrm{u} = \bar{L}_\mathrm{V} L^2 + \bar{L}_\mathrm{l} + L_\mathrm{u}$. Due to the fact that l is bounded below by a quadratic we conclude that $l(x, u) \geq \underline{L}_\mathrm{l}/\bar{L}_\mathrm{V} V(x)$ for all $x \in \mathbb{R}^{d_\mathrm{x}}$, which can be used to simplify the above inequality:

$$\mathrm{E}[V(x_{k+1})] \leq \mathrm{E}[V(x_k)](1 - \underline{L}_\mathrm{l}/(2\bar{L}_\mathrm{V})) + \gamma + \bar{L}_\mathrm{u} d_\mathrm{u} \sigma_{\mathrm{u}1}^2 + \bar{L}_\mathrm{V} d_\mathrm{x} \sigma^2 + c_\mathrm{o}.$$

This readily implies

$$\mathrm{E}[V(x_k)] \leq 2\frac{\bar{L}_\mathrm{V}}{\underline{L}_\mathrm{l}} (\gamma + \bar{L}_\mathrm{u} d_\mathrm{u} \sigma_{\mathrm{u}1}^2 + \bar{L}_\mathrm{V} d_\mathrm{x} \sigma^2 + c_\mathrm{o}),$$

for all $k \geq k_0$, which yields the desired result. □

C.5 Convergence in finite time

Corollary C.6 *(Finite time convergence) Let the assumptions of Prop. 3.1 be satisfied. Then, almost surely, $\{i_k\}_{k=1}^\infty$ converges to i^* in finite time, that is,*

$$\Pr(\sup_{i_k \neq i^*} k < \infty) = 1.$$

Proof We conclude form Prop. 3.1 that $\Pr(i_k \neq i^*) \leq M^2/(k-M)^2$ for all $k \geq M+1$. This implies for any $j \geq M+1$

$$\Pr(\sup_{i_k \neq i^*} k > j) \leq \sum_{k=j}^{\infty} \Pr(i_k \neq i^*),$$

where the right-hand side is bounded above by

$$\sum_{k=j}^{\infty} \frac{M^2}{(k-M)^2} \leq \frac{M^2}{(j-M)^2} + \int_j^\infty \frac{M^2}{(k-M)^2} \mathrm{d}k \leq \frac{M^2}{j-M} \left(1 + \frac{1}{j-M}\right).$$

Hence, the right-hand side converges to zero for large j, which yields the desired result. □

D Details of Sec. 3.2

In order to provide regret guarantees, we will slightly modify Ass. 3 from setting `S1` as follows.

Assumption 4 *There exists an integer $M > 0$ and a constant $c_\mathrm{e} > 0$ such that for all $x_1 \in \mathbb{R}^{d_\mathrm{x}}$, $\sigma_\mathrm{u} > 0$, and $f^1, f^2 \in F$,*

$$\frac{1}{M} \sum_{k=1}^{M} \mathrm{E}\Big[\frac{|f^1(x_k, u_k) - f^2(x_k, u_k)|^2}{1 + |(x_k, u_k)|^2/b^2}\Big] \geq d_\mathrm{u} c_\mathrm{e} \sigma_\mathrm{u}^2 \|f^1 - f^2\|^2,$$

holds, where $x_{k+1} = f(x_k, u_k) + n_k$, $u_k = \hat{\mu}(x_k) + n_{\mathrm{u}k}$ with $n_k \sim \mathcal{N}(0, \sigma^2 I)$ and $u_k \sim \mathcal{N}(0, \sigma_\mathrm{u}^2 I)$, and where $\hat{\mu}$ is any policy corresponding to a model $f \in F$.

Algorithm 2 Reinforcement learning (S2)

Inputs: F, η, M, $\{\sigma_{\mathrm{u}k}^2\}_{k=1}^{\infty}$, ϵ
for $k = 1, \dots$ **do**
  // every Mth step
  if $\mathrm{mod}(k-1, M) = 0$ **then**
    $f^* \leftarrow \mathrm{argmin}_{\bar{f} \in F}\, s_k(\bar{f})$
    $F^\epsilon \leftarrow \mathrm{greedyCover}(F, f^*, \epsilon)$
    $s_k(f^i) \leftarrow \sum_{j=1}^{k-1} \frac{|x_{j+1} - f^i(x_j, u_j)|^2}{1+|(x_k,u_k)|^2/b^2}$
    $i_k \sim \exp(-\eta s_k(f^i))/Z,\ f^i \in F^\epsilon$
    compute μ^{i_k} corr. to $f^{i_k} \in F^\epsilon$ // e.g. by d.p.
  else
    $i_k = i_{k-1}$ //stay with i_{k-1}
  end if
  //follow policy i_k and add excitation
  $u_k = \mu^{i_k}(x_k) + n_{\mathrm{u}k}, \quad n_{\mathrm{u}k} \sim \mathcal{N}(0, \sigma_{\mathrm{u}k}^2 I)$
end for

Algorithm 4 greedyCover(F, f^*, ϵ)

$F^\epsilon \leftarrow \{f^*\}$
$S \leftarrow \{\bar{f} \in F \mid \|\bar{f} - f^i\| > \epsilon, \forall f^i \in F^\epsilon\}$
while $S \neq \{\}$ **do**
  // pick an element from S
  $F^\epsilon \leftarrow F^\epsilon \cup \{\bar{f}\},\ \bar{f} \in S$
  $S \leftarrow \{\bar{f} \in F \mid \|\bar{f} - f^i\| > \epsilon, \forall f^i \in F^\epsilon\}$
end while
return F^ϵ

The assumption is stated for a finite $b > 0$ even though it can be relaxed to $b \to \infty$, and the same policy-regret guarantees apply although with more elaborate constants, see App. F for further discussion. For $b \to \infty$ the assumption describes persistence of excitation as used in system identification and statistics, [see, e.g., 38, Ch. 8.2]. From a maximum-likelihood point of view, Ass. 8 ensures that the dynamics f correspond to a unique non-degenerate minimum of the one-step prediction error, accumulated over M steps. The assumption is generically satisfied if the models $f \in F$ are linear and $\|\cdot\|$ denotes the Lipschitz-norm [see, e.g., 9], as highlighted in Sec. 3. A similar reasoning applies to nonlinear systems, see Sec. F.

We will further strengthen the Bellman-inequality from Ass. 1 to ensure that the steady-state performance γ of the policy μ is stable under small policy changes that arise from models $f^i \in F$ close to f. This notion of stability requires μ to optimize the corresponding Q-function. This is made precise as follows.

Assumption 5 *(Bellman-type inequality) For all small enough $\xi > 0$ there exists a cost-to-go function V (corresponding to f and μ) satisfying the following inequality:*

$$V(x) \geq \mathrm{E}[l(x, \mu^i(x) + n_{\mathrm{u}}) + V(f(x, \mu^i(x) + n_{\mathrm{u}}) + n)] - \gamma - L_{\mathrm{u}} d_{\mathrm{u}} \sigma_{\mathrm{u}}^2 - L_\mu \xi^2,$$

for all policies μ^i corresponding to $\|f^i - f\| < \xi$, for all $x \in \mathbb{R}^{d_{\mathrm{x}}}$, where L_{u}, $L_\mu > 0$ are constant, $n \sim \mathcal{N}(0, \sigma^2 I)$, and $n_{\mathrm{u}} \sim \mathcal{N}(0, \sigma_{\mathrm{u}}^2 I)$.

The following proposition provides a sufficient condition for Ass. 5 to hold. In particular, the proposition applies to the class of linear dynamical systems with a quadratic, positive definite stage cost, where all assumptions are satisfied [see also Prop. 6 in 46].

Proposition D.1 *Let Ass. 1 and Ass. 2 be satisfied and fix $x \in \mathbb{R}^{d_{\mathrm{x}}}$. If, in addition,*

$$l(x,u) \geq \underline{L}_l |x|^2/2, \quad \mu(x) \in \operatorname*{argmin}_{u \in \mathbb{R}^{d_{\mathrm{u}}}} \mathrm{E}[l(x,u) + V(f(x,u)+n)], \quad \textit{and} \quad \|\mu^i - \mu\|_{\mathrm{op}} \leq L'_\mu \xi,$$

holds for all policies μ^i corresponding to $\|f^i - f\|_{\mathrm{op}} < \xi$ and all $\xi > 0$ small enough, then Ass. 5 is satisfied for x and all σ_{u} small enough, where $\underline{L}_{\mathrm{l}}, L'_\mu > 0$ are constant and $n \sim \mathcal{N}(0, \sigma^2 I)$. The Lipschitz-norm $\|\cdot\|_{\mathrm{op}}$ is defined for any Lipschitz-continuous function $q : \mathbb{R}^{d_{\mathrm{x}}} \to \mathbb{R}^d$ as

$$\|q\|_{\mathrm{op}} := \max\Big\{|q(0)|, \sup_{x_1, x_2 \in \mathbb{R}^{d_{\mathrm{x}}}} \frac{|q(x_1) - q(x_2)|}{|x_1 - x_2|}\Big\}.$$

Proof Let $c_{\mathrm{s}}/L_\mu'^2$ denote the smoothness constant of $\mathrm{E}[l(x,u) + V(f(x,u)+n)]$ in u in a neighborhood of $u = \mu(x)$. From smoothness and the fact that $\mu(x)$ is a minimizer we conclude

$$\begin{aligned}\mathrm{E}[l(x,\mu^i(x)) + V(f(x,\mu^i(x)) + n)] &\leq \mathrm{E}[l(x,\mu(x)) + V(f(x,\mu(x)) + n)] + c_{\mathrm{s}} |\mu^i(x) - \mu(x)|^2/(2L'_\mu) \\ &\leq V(x) + \gamma + c_{\mathrm{s}} |\mu^i(x) - \mu(x)|^2/(2L_\mu'^2), \end{aligned} \tag{13}$$

where Ass. 1 has been used for the second step (where we set $\sigma_{\mathrm{u}} = 0$). We further note that

$$|\mu^i(x) - \mu(x)| = |\mu^i(0) - \mu(0) + \mu^i(x) - \mu(x) - (\mu^i(0) - \mu(0))| \leq L'_\mu(\xi + \xi|x|),$$

due to the fact that $\|\mu^i - \mu\|_{\mathrm{op}} \leq L'_\mu \xi$. We multiply (13) by $1 + 4c_{\mathrm{s}}\xi^2/\underline{L}_{\mathrm{l}}$, define $\tilde{V} := (1 + 4c_{\mathrm{s}}\xi^2/\underline{L}_{\mathrm{l}})\, V$, and arrive at

$$\begin{aligned}\mathrm{E}[l(x, \mu^i(x)) + \tilde{V}(f(x, \mu^i(x)) + n)] \leq \tilde{V}(x) + (1 + 4c_{\mathrm{s}}\xi^2/\underline{L}_{\mathrm{l}})\gamma - 2\xi^2 c_{\mathrm{s}}|x|^2 \\ + c_{\mathrm{s}}(1 + 4c_{\mathrm{s}}\xi^2/\underline{L}_{\mathrm{l}})\xi^2 + c_{\mathrm{s}}(1 + 4c_{\mathrm{s}}\xi^2/\underline{L}_{\mathrm{l}})\xi^2|x|^2,\end{aligned}$$

where we have used the fact that $l(x, u) \geq \underline{L}_{\mathrm{l}}|x|^2/2$. We choose $\xi^2 \leq \underline{L}_{\mathrm{l}}/(4c_{\mathrm{s}})$ and rearrange terms. This results in

$$\mathrm{E}[l(x, \mu^i(x)) + \tilde{V}(f(x, \mu^i(x)) + n)] \leq \tilde{V}(x) + \gamma + 2c_{\mathrm{s}}(2\gamma/\underline{L}_{\mathrm{l}} + 1)\xi^2 + c_{\mathrm{s}}(-1 + 4c_{\mathrm{s}}\xi^2/\underline{L}_{\mathrm{l}})\xi^2|x|^2,$$

where the last term is non-positive. We therefore conclude that the inequality in Ass. 5 holds for $\tilde{V}$ with $L_\mu = 2c_{\mathrm{s}}(2\gamma/\underline{L}_{\mathrm{l}} + 1)$, $L_{\mathrm{u}} = c_{\mathrm{s}}/(2L'^2_\mu)$, and a small enough σ_{u}. □

We are now ready to prove the main result of this section:

Theorem D.2 *Let Ass. 2, Ass. 4, and Ass. 5 be satisfied and choose η and $\sigma_{\mathrm{u}k}$ as*

$$\eta = \min\left\{\frac{1}{4M\sigma^2}, \frac{1}{2ML^2b^2}\right\}, \quad \sigma^2_{\mathrm{u}k} = \frac{4}{\eta c_{\mathrm{e}} d_{\mathrm{u}} M\epsilon^2}\left(\frac{2}{\lceil k/M\rceil} + \frac{\ln(m(\epsilon))}{(\lceil k/M\rceil)^2}\right).$$

Then, the policy regret of Alg. 2 is bounded by

$$\sum_{k=1}^{N} \mathrm{E}[l(x_k, u_k)] - N\gamma \leq c_{\mathrm{r}1}\frac{3\ln(N) + M\ln(m(\epsilon))}{\epsilon^2} + L_\mu N\epsilon^2 + c_{\mathrm{r}2}$$

for all $N \geq 2M$, where $m(\epsilon)$ denotes the packing number of F for a packing of size ϵ and where the constants $c_{\mathrm{r}1}$ and $c_{\mathrm{r}2}$ are given by

$$c_{\mathrm{r}1} = \frac{8c_\alpha(\bar{L}_{\mathrm{V}}L^2 + \bar{L}_{\mathrm{l}} + L_{\mathrm{u}})}{\eta c_{\mathrm{e}}}, \quad c_\alpha = e^{3c_2 M}, \quad c_{\mathrm{r}2} = 3Mc_\alpha(\bar{L}_{\mathrm{V}} d_{\mathrm{x}}\sigma^2/2 + c_{\mathrm{o}}).$$

Proof The proof follows Thm. 3.2. At every iteration k we denote by I^*_k the set of models $f^{i^*} \in F^\epsilon_k$ that satisfy $\|f^{i^*} - f\| \leq \epsilon$. We then conclude from the same reasoning as in Prop. 3.1 that

$$\Pr(i_k \notin I^*_k) \leq \frac{M^2}{(k - M)^2},$$

for all $k \geq M + 1$. We make therefore the case distinction $i_k \in I^*_k$ and $i_k \notin I^*_k$, which then yields by the same arguments (see (9))

$$\sum_{k=1}^{N}(\mathrm{E}[l(x_k, u_k)] - \gamma_k) \leq NL_\mu\epsilon^2 + c_\alpha\bar{L}_{\mathrm{u}}d_{\mathrm{u}}\sum_{k=1}^{N}\sigma^2_{\mathrm{u}k} + c_\alpha(\bar{L}_{\mathrm{V}}d_{\mathrm{x}}\sigma^2/2 + c_{\mathrm{o}})\sum_{k=1}^{N}\Pr(i_k \notin I^*_k),$$

where there is an additional error term, due to the fact that f^{i^*} and f could be different (although $\|f^{i^*} - f\| \leq \epsilon$ for any $i^* \in I^*_k$, by construction of I^*_k). The desired result follows from the previous inequality. However, compared to Thm. 3.2 we used the slightly more conservative bound

$$\sum_{k=1}^{N}\sigma^2_{\mathrm{u}k} \leq \frac{8}{d_{\mathrm{u}}\eta c_{\mathrm{e}}\epsilon^2}(3\ln(N) + M\ln(m(\epsilon))),$$

which applies as long as $N \geq 2$, and simplifies the resulting constants. □

E Details of Sec. 3.3

For deriving the regret bound we will slightly adapt the persistence of excitation condition Ass. 3 from Sec. 3. The motivation is analogous to Ass. 4 and we refer the reader to App. D and App. F for further discussion.

Assumption 6 *There exists an integer $M > 0$ and a constant $c_\mathrm{e} > 0$ such that*

$$\frac{1}{M}\sum_{k=1}^{M} \mathrm{E}\Big[\frac{|f_\theta(x_k,u_k) - f(x_k,u_k)|^2}{1+|(x_k,u_k)|^2/b^2}\Big] \geq d_\mathrm{u} c_\mathrm{e} \sigma_\mathrm{u}^2 |\theta|^2,$$

for all $\theta \in \Omega$ and all $x_1 \in \mathbb{R}^{d_\mathrm{x}}$, where $x_{k+1} = f(x_k,u_k) + n_k$, $u_k = \mu_\theta(x_k) + n_{\mathrm{u}k}$, and $n_k, n_{\mathrm{u}k}$ are independent random variables that satisfy $n_{\mathrm{u}k} \sim \mathcal{N}(0,\sigma_\mathrm{u}^2 I)$, $n_k \sim \mathcal{N}(0,\sigma^2 I)$.

The second assumption, which will be important, is a strengthened version of the Bellman-inequality from Ass. 1. The assumption ensures that the steady-state performance γ of the policy μ is stable under small policy changes that arise from models $f_\theta \in F$ that are close to f. The sufficient condition provided by Prop. D.1 applies here in the same way ($\|f_\theta - f\|_\mathrm{op}$ reduces to $|\theta|$) and we therefore conclude that the assumption below is, for example, satisfied for linear dynamical systems with a quadratic, positive definite stage cost.

Assumption 7 *(Bellman-type inequality) For all small enough $\xi > 0$, there exists a cost to go function V (corresponding to f and μ) satisfying the following inequality:*

$$V(x) \geq \mathrm{E}[l(x,\mu_\theta(x)+n_\mathrm{u}) + V(f(x,\mu_\theta(x)+n_\mathrm{u})+n)] - \gamma - L_\mathrm{u} d_\mathrm{u} \sigma_\mathrm{u}^2 - L_\mu \xi^2,$$

for all policies μ_θ with $|\theta| < \xi$, for all $x \in \mathbb{R}^{d_\mathrm{x}}$, where $L_\mathrm{u}, L_\mu > 0$ are constant, $n \sim \mathcal{N}(0,\sigma^2 I)$, and $n_\mathrm{u} \sim \mathcal{N}(0,\sigma_\mathrm{u}^2 I)$.

We now prove the main result characterizing policy regret for the setting `S3`.

Theorem E.1 *Let Ass. 2, Ass. 6, and Ass. 7 be satisfied and choose η and $\sigma_{\mathrm{u}k}^2$ as*

$$\eta \leq \min\Big\{\frac{1}{4M\sigma^2}, \frac{1}{2ML^2b^2}\Big\}, \quad \sigma_{\mathrm{u}k}^2 = \frac{4}{\eta c_\mathrm{e} d_\mathrm{u} M \epsilon^2}\left(\frac{2}{\lceil k/M\rceil} + \frac{p}{(\lceil k/M\rceil)^2}\right).$$

Then, for all $N \geq 2M$ there exists a large enough p, such that the policy regret of Alg. 3 is bounded by

$$\sum_{k=1}^{N} \mathrm{E}[l(x_k,u_k)] - N\gamma \leq 2\sqrt{c_{\mathrm{r}1}(3\ln(N)+Mp)N} + c_{\mathrm{r}2},$$

where the constants $c_{\mathrm{r}1}$ and $c_{\mathrm{r}3}$ are given by

$$c_{\mathrm{r}1} = \frac{8c_\alpha L_\mu(\bar{L}_\mathrm{V} L^2 + \bar{L}_\mathrm{l} + L_\mathrm{u})}{\eta c_\mathrm{e}}, \quad c_\alpha = e^{3c_2 M}, \quad c_{\mathrm{r}2} = 3Mc_\alpha(\bar{L}_\mathrm{V} d_\mathrm{x}\sigma^2/2 + c_\mathrm{o}),$$

with

$$\epsilon^2 = \sqrt{\frac{c_{\mathrm{r}1}(3\ln(N)+Mp)}{L_\mu^2 N}}.$$

Proof We first argue that the reasoning in Lemma C.1 applies in a very similar way to setting `S3`. To that extent, we first define the random variable p_k as follows

$$p_k = \frac{\int_{\Omega\setminus\{\theta:|\theta|\leq\epsilon\}} e^{-\eta(s_k(\theta)-\bar{s}_k)}\mathrm{d}\theta}{\int_\Omega e^{-\eta(s_k(\theta)-\bar{s}_k)}\mathrm{d}\theta},$$

where $\mathrm{d}\theta$ denotes the Lebesgue measure, and $\bar{s}_k = \min_{\theta\in\Omega} s_k(\theta)$. However, compared to the discrete setting, where the denumerator was simply bounded below by unity, the situation is more delicate. More precisely, we bound the denumerator from below by $|B_\delta| e^{-\eta h(\delta)}$, where $h(\delta) := \max_{\theta\in B_\delta} s_k(\theta) - \bar{s}_k$ and B_δ denotes a ball of radius δ with volume $|B_\delta|$ centered at a minimizer of $s_k(\theta)$. From the smoothness of $s_k(\theta)$ we conclude that $h(\delta) = \mathcal{O}(\delta^2)$ for small δ. Due to our

normalization, Ω is contained in a ball of unit radius and we have $|B_\delta|/|\Omega| \geq |B_\delta|/|B_1| \geq \delta^p$ where p refers to the dimension of Ω. Hence we arrive at the following lower bound

$$\int_\Omega e^{-\eta(s_k(\theta)-\bar{s}_k)}\mathrm{d}\theta \geq |\Omega|\delta^p e^{-h(\delta)} \gtrsim |\Omega|e^{-p},$$

where the second inequality arises from carefully choosing δ in order to balance the the term δ^p and $e^{-h(\delta)}$. This yields the following bound on p_k (which resembles the discrete setting)

$$p_k \leq \frac{e^p}{|\Omega|}\int_{\Omega\setminus\{\theta:|\theta|\leq\epsilon\}} e^{-\eta(s_k(\theta)-s_k^*)}\mathrm{d}\theta,$$

where we have also replaced $\bar{s}_k$ with s_k^* due to the fact that $\bar{s}_k$ is a minimum. Following the same reasoning as in Lemma C.1 and Prop. 3.1 yields therefore

$$\Pr(|\theta_k| > \epsilon) \leq \frac{e^p}{|\Omega|}\int_{\Omega\setminus\{\theta:|\theta|\leq\epsilon\}} \mathrm{E}[e^{-\eta(s_k(\theta)-s_k^*)}]\mathrm{d}\theta \leq e^p \exp\left(-\frac{c_\mathrm{e} d_\mathrm{u}\eta}{4}\epsilon^2 \sum_{j=1}^{k-M}\sigma_{\mathrm{u}j}^2\right),$$

where Fubini's theorem has been used in the first step to interchange expectation and integration. In addition, due to the modification of $\sigma_{\mathrm{u}k}^2$ compared to Thm. D.2 (where now $m(\epsilon)$ is replaced by e^p), we find that

$$\Pr(|\theta_k| \geq \epsilon) \leq \frac{M^2}{(k-M)^2},$$

for all $k \geq M+1$. We apply the same reasoning as in the proof of Thm. 3.2, where we now have the case distinction $\Pr(|\theta_k| > \epsilon)$ and $\Pr(|\theta_k| \leq \epsilon)$ (corresponding to $\Pr(i_k \neq i^*)$ and $\Pr(i_k = i^*)$). This concludes that

$$\sum_{k=1}^N (\mathrm{E}[l(x_k,u_k)] - \gamma_k) \leq L_\mu N\epsilon^2 + c_\alpha \bar{L}_\mathrm{u} d_\mathrm{u}\sum_{k=1}^N \sigma_{\mathrm{u}k}^2 + c_\alpha(\bar{L}_\mathrm{V} d_\mathrm{x}\sigma^2/2 + c_\mathrm{o})\sum_{k=1}^N \Pr(|\theta_k| > \epsilon),$$

where, as before,

$$\sum_{k=1}^N \Pr(|\theta_k| > \epsilon) \leq 3M.$$

We further note that the sum over $\sigma_{\mathrm{u}k}^2$ yields

$$\sum_{k=1}^N \sigma_{\mathrm{u}k}^2 \leq \frac{8}{d_\mathrm{u}\eta c_\mathrm{e}\epsilon^2}(1+\ln(N)+Mp) \leq \frac{8}{d_\mathrm{u}\eta c_\mathrm{e}\epsilon^2}(3\ln(N)+Mp),$$

where $N \geq 2M \geq 2$ (and therefore $\ln(N) \geq 1/2$) has been used in the second step. Consequently the regret is bounded by

$$\sum_{k=1}^N \mathrm{E}[l(x_k,u_k)] - N\gamma \leq L_\mu N\epsilon^2 + c_{\mathrm{r}1}\frac{3\ln(N)+Mp}{L_\mu\epsilon^2} + c_{\mathrm{r}2}.$$

The choice of ϵ achieves an optimal trade-off between the first two terms, which yields the desired result. □

F Relaxing persistence of excitation

This section discusses the situation when $b \to \infty$. We slightly modify Ass. 3 to the following:

Assumption 8 *There exists an integer $M > 0$ and two constants $c_\mathrm{e} > 0$ and $b > 0$ such that for any $x_1 \in \mathbb{R}^{d_\mathrm{x}}$, $\sigma_\mathrm{u} > 0$, and $f^i \in F$, $f^i \neq f$,*

$$\frac{1}{M}\sum_{k=1}^M \mathrm{E}\Big[|f^i(x_k,u_k) - f(x_k,u_k)|^2\Big] \geq c_\mathrm{e}(d_\mathrm{u}\sigma_\mathrm{u}^2 + d_\mathrm{x}\sigma^2)$$

holds, where $x_{k+1} = f(x_k,u_k) + n_k$, $u_k = \mu^q(x_k) + n_{\mathrm{u}k}$ with $n_k \sim \mathcal{N}(0,\sigma^2 I)$, $n_{\mathrm{u}k} \sim \mathcal{N}(0,\sigma_\mathrm{u}^2 I)$, and $q \in \{1,\ldots,m\}$.

We now derive a variant of Lemma C.2 that relies on the fact that over finite time x_k and u_k are sub-Gaussian random variables. This is summarized as follows.

Lemma F.1 *Let Ass. 8 be satisfied and let*

$$l_k^i = |x_{k+1} - f^i(x_k, u_k)|^2,$$

for $k = 1, 2, \ldots$, where x_k, u_k denotes the trajectory resulting from Alg. 1, and $\sigma_{\mathrm{u}k}^2$ is monotonically decreasing. Let $k' \geq 0$ be an integer and define $k = k'M + 1$ (i.e., k is a time instance where i_k switches). Then, the following bound holds for all $0 < \eta \leq \min\{1/(4M\sigma^2), \eta_0\}$ and all $j = k, \ldots, k + M - 1$

$$\mathrm{E}[e^{-\eta \sum_{j=k}^{k+M-1}(l_j^i - l_j^*)}|x_k] \leq \exp\left(-\frac{\eta c_{\mathrm{e}}}{4} \sum_{j=k}^{k+M-1} (d_{\mathrm{u}}\sigma_{\mathrm{u}j}^2 + d_{\mathrm{x}}\sigma^2)\right),$$

where

$$\eta_0 = \frac{1}{4M(\sigma^2 d_{\mathrm{x}} + \sigma_{\mathrm{u}1}^2 d_{\mathrm{u}})} \cdot \frac{c_{\mathrm{e}}}{128M^2(2L^{2M}(1+L_\mu)^{2M})^2},$$

and f_j, f_j^i is shorthand notation for $f(x_j, u_j)$ and $f^i(x_j, u_j)$, respectively.

Proof Without loss of generality we set $k = 1$ and $k' = 0$ (the proof follows exactly the same steps for $k' > 0$). We first note that the random variables x_j and u_j for $j \geq k$ are Lipschitz continuous functions of the noise variables $\{n_q, n_{\mathrm{u}q}\}_{q=1}^{M-1}$. We define the random variable $X_j := |f(x_j, u_j) - f^i(x_j, u_j)|/\sqrt{2}$ and note that X is sub-Gaussian with variance proxy

$$\tilde{\sigma}^2 := 2L^{2M}(1+L_\mu)^{2M} M(\sigma_{\mathrm{u}1}^2 d_{\mathrm{u}} + \sigma^2 d_{\mathrm{x}}),$$

due to the fact that there are at most M steps between x_1 and x_j. We will simplify the notation by introducing the following variables

$$\tilde{L} := 2L^{2M}(1+L_\mu)^{2M}, \sigma_{\mathrm{e}}^2 := M(\sigma_{\mathrm{u}1}^2 d_{\mathrm{u}} + \sigma^2 d_{\mathrm{x}}),$$

such that $\tilde{\sigma}^2 = \tilde{L}\sigma_{\mathrm{e}}^2$ and $128M^2\tilde{\sigma}^2\eta_0 = c_{\mathrm{e}}/(4\tilde{L})$. The previous result exploits the fact that a L_{v}-Lipschitz function of a set of p_{v} independent standard Gaussian random variables is sub-Gaussian with variance proxy $p_{\mathrm{v}}L_{\mathrm{v}}$ [see, e.g., 50, Ch.2.3]. By following the same argument as in Prop. 3.1 and Lemma. C.2 we arrive at

$$\begin{aligned}
\mathrm{E}[e^{-\eta \sum_{j=1}^{M}(l_j^i - l_j^*)}] &\leq \left(\prod_{j=1}^{M} \mathrm{E}[e^{-\eta M(l_j - l_j^*)}]\right)^{1/M} \\
&\leq \left(\prod_{j=1}^{M} \mathrm{E}[e^{-\eta M|f_j - f_j^i|^2/2}]\right)^{1/M} \\
&\leq \left(\prod_{j=1}^{M} \mathrm{E}[e^{-\eta M X_j^2}]\right)^{1/M},
\end{aligned}$$

where we used the shorthand notation f_j, f_j^i as in the statement of the lemma and the fact $\eta \leq 1/(4M\sigma^2)$. The random variables X_j are sub-Gaussian with variance proxy $\tilde{\sigma}^2$ and therefore $X_j^2 - \mathrm{E}[X_j^2]$ are sub-Exponential with parameter $16\tilde{\sigma}^2$. As a result, we can simplify the previous inequality to

$$\mathrm{E}[e^{-\eta \sum_{j=1}^{M}(l_j^i - l_j^*)}] \leq \left(\prod_{j=1}^{M} e^{-\eta M \mathrm{E}[X_j^2] + \eta\tilde{\sigma}^2 c_{\mathrm{e}}/(4\tilde{L})}\right)^{1/M},$$

since $128M\eta\tilde{\sigma}^2 \leq c_{\mathrm{e}}/(4\tilde{L})$ by our choice of η_0. As a result of Ass. 8 we infer

$$\sum_{j=1}^{M} \mathrm{E}[X_j^2] \geq M c_{\mathrm{e}}(d_{\mathrm{u}}\sigma_{\mathrm{u}1}^2 + d_{\mathrm{x}}\sigma^2)/2 = c_{\mathrm{e}}\sigma_{\mathrm{e}}^2/2,$$

and therefore

$$\sum_{j=1}^{M} \mathrm{E}[X_j^2] - \frac{c_\mathrm{e}\tilde{\sigma}^2}{4\tilde{L}} \geq \sigma_\mathrm{e}^2(\frac{c_\mathrm{e}}{2} - \frac{c_\mathrm{e}}{4}) = \frac{\sigma_\mathrm{e}^2 c_\mathrm{e}}{4}.$$

This establishes

$$\mathrm{E}[e^{-\eta \sum_{j=1}^{M}(l_j^i - l_j^*)}] \leq e^{-\eta c_\mathrm{e}\sigma_\mathrm{e}^2/4},$$

and yields the desired result. □

The conclusion from the setting with linear dynamics can be generalized to nonlinear systems as follows. In order to simplify the presentation we consider the situation where the process noise is absent ($\sigma = 0$); the same rationale applies when $\sigma > 0$. The following result demonstrates that Ass. 3 is generic and holds for a broad class of nonlinear dynamics. This result also highlights a close connection between controllability and the required notion of persistence of excitation.

Proposition F.2 *Let $x \in \mathbb{R}^{d_\mathrm{x}}$ and $q \in \{1, \ldots, m\}$ be fixed. Then, there exists a constant $c_\mathrm{e}' > 0$ such that*

$$\mathrm{E}[|f^i(x_k, u_k) - f(x_k, u_k)|^2] \geq c_\mathrm{e}' d_\mathrm{u} \sigma_\mathrm{u}^2$$

for all small enough $\sigma_\mathrm{u} > 0$ if either of the two inequalities are satisfied

$$|f^i(\bar{x}_k, \mu^q(\bar{x}_k)) - f(\bar{x}_k, \mu^q(\bar{x}_k))|^2 > 0, \qquad \underline{\sigma}(W_{k-1}^\mathrm{c})\, |A_k^i - A_k|_\mathrm{F}^2 + |B_k^i - B_k|_\mathrm{F}^2 > 0,$$

where x_k is defined recursively via $x_1 = x$, $x_{j+1} = f(x_j, \mu^q(x_j) + n_{\mathrm{u}j})$ with $n_{\mathrm{u}j} \sim \mathcal{N}(0, \sigma_{\mathrm{u}j}^2 I)$ $j = 1, \ldots, k-1$ and

$$\begin{aligned}
A_k &:= \frac{\partial}{\partial x} f(x, \mu^q(x))\bigg|_{x=\bar{x}_k}, \quad A_k^i := \frac{\partial}{\partial x} f^i(x, \mu^q(x))\bigg|_{x=\bar{x}_k}, \\
B_k &:= \frac{\partial}{\partial u} f(x, u)\bigg|_{x=\bar{x}_k, u=\mu^q(\bar{x}_k)}, \quad B_k^i := \frac{\partial}{\partial u} f(x, u)\bigg|_{x=\bar{x}_k, u=\mu^q(\bar{x}_k)}, \\
W_k^\mathrm{c} &:= \sum_{j=1}^{k-1} A_{k-1} A_{k-2} \ldots A_{j+1} B_j B_j^\top A_{j+1}^\top \ldots A_{k-2}^\top A_{k-1}^\top.
\end{aligned}$$

Moreover, $\bar{x}_k$ corresponds to the noise-free trajectory and is defined via $\bar{x}_1 = x, \bar{x}_{j+1} = f(\bar{x}_j, \mu^q(\bar{x}_j))$.

Proof We start by considering the situation where $f^i(\bar{x}_k, \mu^q(\bar{x}_k)) \neq f(\bar{x}_k, \mu^q(\bar{x}_k))$. We note that

$$\mathrm{E}[|f^i(x_k, \mu^q(x_k) + n_{\mathrm{u}k}) - f(x_k, \mu^q(x_k) + n_{\mathrm{u}k})|^2]$$

continuously depends on σ_u and converges to $|f^i(\bar{x}_k, \mu^q(\bar{x}_k)) - f(\bar{x}_k, \mu^q(\bar{x}_k))|^2 > 0$ as $\sigma_\mathrm{u} \to 0$. Hence the desired inequality is clearly satisfied for all small enough $\sigma_\mathrm{u} > 0$.

Next we consider the situation where $f^i(\bar{x}_k, \mu^q(\bar{x}_k)) = f(\bar{x}_k, \mu^q(\bar{x}_k))$ and apply Taylor's theorem as follows:

$$f(x_k, \mu^q(x_k) + n_{\mathrm{u}k}) = f(\bar{x}_k, \mu^q(\bar{x}_k)) + A_k(\bar{x}_k - x_k) + B_k n_{\mathrm{u}k} + o(\bar{x}_k - x_k, n_{\mathrm{u}k}),$$

where o is a continuous function that satisfies $o(\xi)/|\xi| \to 0$ for $|\xi| \to 0$. We therefore conclude

$$|f^i(x_k, u_k) - f(x_k, u_k)| = \Big|(A_k^i - A_k)(x_k - \bar{x}_k) + (B_k^i - B_k) n_{\mathrm{u}k} + o(\bar{x}_k - x_k, n_{\mathrm{u}k})\Big|,$$

where we slightly abused notation to redefine the reminder term (we will frequently do so throughout the remainder of the proof). We further apply Taylor's theorem to express $x_k - \bar{x}_k$ as

$$x_k - \bar{x}_k = \sum_{j=1}^{k-1} A_{k-1} \ldots A_{j+1} B_j n_{\mathrm{u}j} + o(n_{\mathrm{u}1}, \ldots, n_{\mathrm{u}k-1}).$$

By combining the previous two equations, squaring, and taking expectations, we arrive at

$$\mathrm{E}[|f^i(x_k, u_k) - f(x_k, u_k)|^2] \geq (|A_k^i - A_k|_\mathrm{F}^2 \underline{\sigma}(W_{k-1}^\mathrm{c}) + |B_k^i - B_k|_\mathrm{F}^2)\sigma_\mathrm{u}^2 - o(\sigma_\mathrm{u}^2),$$

where we took advantage of the fact that $n_{\mathrm{u}1}, \ldots, n_{\mathrm{u}k}$ are mutually independent. We further used the following reasoning: i) independence between $n_{\mathrm{u}i}$ and $n_{\mathrm{u}j}$, $i \neq j$, concludes

$$\mathrm{E}[o(n_{\mathrm{u}i}) n_{\mathrm{u}j}^{\top}] = \mathrm{E}[o(n_{\mathrm{u}i}) \mathrm{E}[n_{\mathrm{u}j}^{\top} \mid n_{\mathrm{u}i}]] = 0.$$

ii) for $i = j$ we have

$$\mathrm{E}[o(|n_{\mathrm{u}i}|^2)] = \int_{\mathbb{R}^{d_{\mathrm{u}}}} o(|\xi|^2) \frac{1}{(\sqrt{2\pi}\sigma_{\mathrm{u}})^{d_{\mathrm{u}}}} e^{-|\xi|^2/(2\sigma_{\mathrm{u}}^2)} \mathrm{d}\xi \leq \underbrace{\int_{\mathbb{R}^{d_{\mathrm{u}}}} o(|\xi|^2) \frac{2^q}{\sqrt{2\pi}^{d_{\mathrm{u}}}} \frac{q!}{|\xi|^{2q}} \mathrm{d}\xi}_{=\mathrm{const.}} \sigma_{\mathrm{u}}^{2q-d_{\mathrm{u}}},$$

for any integer $q \geq 0$ large enough, where we have bounded the exponential using $e^{-\xi} \leq q!/\xi^q$ for all $\xi \geq 0$. This implies that $\mathrm{E}[o(|n_{\mathrm{u}i}|^2)] = o(\sigma_{\mathrm{u}}^2)$ (in fact $\mathrm{E}[o(|n_{\mathrm{u}i}|^2)]$ decays much faster for small σ_{u}), which leads to the desired result. □