Safely Learning Controlled Stochastic Dynamics

Luc Brogat-Motte [1, 2] Alessandro Rudi [3] Riccardo Bonalli [1]

[1] Laboratoire des Signaux et Systèmes, CNRS, CentraleSupélec, Université Paris-Saclay [2] Istituto Italiano di Tecnologia [3] SDA Bocconi, Bocconi University

In a Nutshell

Motivation: Reliable control requires accurate models, yet many real systems (robots, medical devices, industrial plants) have partially unknown dynamics. Learning them demands interaction through control actions, but unsafe inputs can damage the system or its environment.

Problem: Learn stochastic dynamics while exploring *only within a safe region*.

Challenge: Learning requires exploration, but safety requires knowledge of the dynamics.

Approach: An iterative kernel-based method that jointly learns dynamics and the set of safe controls.

Results: Provably safe exploration and finite-sample guarantees for continuous-time stochastic systems with Sobolev-regular dynamics, validated on a nonlinear benchmark.

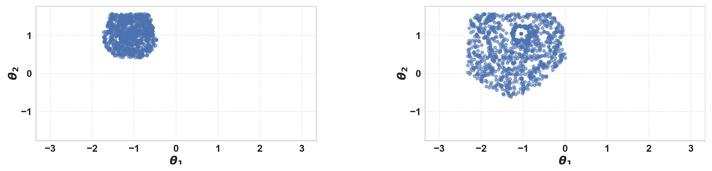


Figure 1. Control coverage after 1000 iterations for $\varepsilon = \xi \in \{0.1, 0.3\}$ (left: 0.1, right: 0.3).

Problem Setting

Controlled stochastic system

We consider a continuous-time system governed by a nonlinear SDE:

$$dX_t = b(X_t, u(t, X_t))\,dt + a(X_t, u(t, X_t))\,dW_t, \quad X_0 \sim p_0,$$

where b, a describe drift and diffusion terms, and W_t is an n-dimensional Brownian motion.

Control parameterization

Controls $u_\theta(t, x)$ are parameterized by $\theta \in D \subset \mathbb{R}^m$, where D denotes a compact parameter domain. Each θ induces a stochastic process X^{u_θ}.

Safe control

A function $g : \mathbb{R}^n \to \mathbb{R}$ defines safe ($g(x) \geq 0$) and unsafe ($g(x) < 0$) regions. The probability of being safe at time t and its minimum over horizon T are

$$s(\theta, t) = \mathbb{P}\big(g(X_t^{u_\theta}) \geq 0\big), \quad s^\infty(\theta, T) = \inf_{0 \leq t \leq T} s(\theta, t).$$

A control is ε-safe up to T if $s^\infty(\theta, T) \geq 1 - \varepsilon$.

Safe data collection

Only ε-safe controls are executed:

$$(\theta_k,\ X^{u_{\theta_k}}(w_i^k, t_\ell)), \quad k = 1{:}K,\ i = 1{:}Q,\ \ell = 1{:}M_k.$$

Learning problem

We aim to estimate the state density map

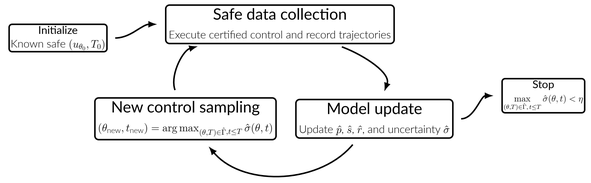
$$p_\theta(t, x) = \text{density of } X_t^{u_\theta} \quad \text{for } (\theta, t, x) \in D \times [0, T_{\max}] \times \mathbb{R}^n,$$

while ensuring safe data collection.

Proposed Method

We introduce an iterative method for **safe exploration and learning** of stochastic dynamics.

Starting from *a single a-priori known safe control*, the method alternates between *safe data collection*, *model update*, and *new control sampling*, progressively expanding the certified safe region.



Dynamics model

From Q trajectories under control u_θ, we estimate the state density and predictive uncertainty:

$$\hat{p}_\theta(t, x) \quad \text{and} \quad \hat{\sigma}(\theta, t),$$

using kernel-based models.

Safety model

The *estimated safety probability*

$$\hat{s}(\theta, t) = \int_{g(x) \geq 0} \hat{p}_\theta(t, x)\,dx$$

quantifies how likely the system is to remain in the safe region under control u_θ at time t.

Reset model

To ensure repeated and independent exploration episodes, we introduce a *reset mechanism*. We assume a *reset region* $h(x) \geq 0$ from which the system can safely return to the initial distribution. We estimate the probability of reaching this region under control u_θ at time t as

$$\hat{r}(\theta, t) = \int_{h(x) \geq 0} \hat{p}_\theta(t, x)\,dx.$$

Safe data collection

At each iteration we consider the set $\hat{\Gamma}$ of control–horizon pairs (θ, T) whose lower confidence bounds (computed from $\hat{\sigma}$) stay above the thresholds $1 - \varepsilon$ and $1 - \xi$ up to horizon T:

$$(\theta, T) \in \hat{\Gamma} \iff \text{LCB}\big(\hat{s}(\theta, [0, T])\big) \geq 1 - \varepsilon, \quad \text{LCB}\big(\hat{r}(\theta, T)\big) \geq 1 - \xi.$$

The next query is selected by maximizing predictive uncertainty within this feasible set:

$$(\theta_{\text{new}}, t_{\text{new}}) = \arg\max_{(\theta,T)\in\hat{\Gamma}, t\leq T} \hat{\sigma}(\theta, t).$$

This **selects the most uncertain point inside the certified safe region** $\hat{\Gamma}$, gradually expanding it while preserving high-probability safety guarantees.

Assumptions

- **Initial safe set:** $S_0 \subset D \times [0, T_{\max}]$, $s(\theta, t) \geq 1 - \varepsilon$.
- **Initial reset set:** $R_0 \subset D \times [0, T_{\max}]$, $r(\theta, t) \geq 1 - \xi$.
- **Sobolev-regular dynamics:** $p \in H^\nu(\mathbb{R}^{n+m+1})$, $\nu > \max(n, m+1)/2$, uniform in x and (θ, t).

Safety and Estimation Guarantees

Our method provides *formal safety and estimation guarantees* with high probability $1 - \delta$.

Certified high-probability safety

At all iterations, the maintained set $\hat{\Gamma}$ contains only (ε, ξ)–safe controls:

$$s^\infty(\theta, T) \geq 1 - \varepsilon, \qquad r(\theta, T) \geq 1 - \xi \quad \forall (\theta, T) \in \hat{\Gamma}.$$

Since only controls from $\hat{\Gamma}$ are executed, safety holds throughout training, and $\hat{\Gamma}$ also defines a certified safe region for deployment.

Estimation and sample complexity

For all $(\theta, T) \in \hat{\Gamma}, t \leq T$,

$$\|\hat{p}_\theta(t, \cdot) - p_\theta(t, \cdot)\|_{L^\infty} \leq c_1\eta, \quad |\hat{s}(\theta, t) - s(\theta, t)| \leq c_2\eta, \quad |\hat{r}(\theta, t) - r(\theta, t)| \leq c_3\eta,$$

with accuracy η achieved (up to log factors) using

$$Q \gtrsim N^{\frac{2\nu+n}{2\nu-n}}, \qquad N = \mathcal{O}\big(\eta^{-\frac{2}{1-\alpha}}\big), \quad \alpha > \frac{m+1}{m+1+2\nu}.$$

Smoother dynamics (larger ν) yield *faster convergence* and *lower sample complexity*.

Numerical Experiments

The method is tested on a nonlinear 2D stochastic system to assess safe and efficient exploration.

System

$$dX_t = V_t\,dt, \qquad dV_t = u(t, X_t, V_t)\,dt + a(X_t)\,dW_t,$$

with diffusion $a(X) = Ae^{-\|X - X_c\|^2/2\sigma^2}$, safe region $(-10, 10)^2$, and reset disk at the origin.

Controls: Parameterized by two acceleration angles (θ_1, θ_2) with fixed magnitude v, followed by a feedback phase steering to the reset region.

Experimental setup: We run 1000 iterations, starting from a single known safe control, under safety and reset thresholds $\varepsilon = \xi \in \{0.1, 0.3\}$. *Fig. 1:* corresponding control coverage; *Fig. 2:* sample trajectories.

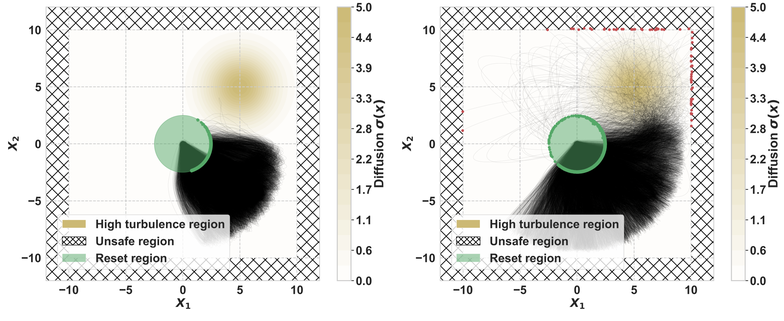


Figure 2. Sample trajectories after 1000 iterations for $\varepsilon = \xi \in \{0.1, 0.3\}$ (left: 0.1, right: 0.3).

Key takeaways

- **(i) Safe exploration.** All executed trajectories satisfied required safety and reset constraints.
- **(ii) Tunable safety-exploration trade-off.** By tuning the safety and reset thresholds (ε, ξ), users directly control the balance between conservative exploration and fast unsafe exploration.