Towards data-driven stochastic control: Learning diffusion dynamics and optimal strategies

International online seminar on SDEs and related topics

Claudia Strauch
joint work with **Sören Christensen**, **Niklas Dexheimer** and **Lukas Trottner**

31 May 2024

Aarhus University Christian-Albrechts-Universität Kiel

AARHUS UNIVERSITY

Motivation: Decision making under uncertainty

More specifically:

- How to optimize the control of **dynamic systems** under **uncertainty**?
- How to find **optimal policies** that **maximize rewards/minimize costs**?

Applications:

- **natural resource management** (⇝ **impulse control**)**:**
  - optimize fishing policies in order to sustainably exploit fish stocks
  - optimize logging and reforestation policies in forestry
- **aerospace guidance** (⇝ **singular control**)**:** spacecraft trajectory optimization to ensure safe and efficient paths
- **game playing:** learn optimal strategies for games like chess, Go, or video games

Decision making under uncertainty

Stochastic control:

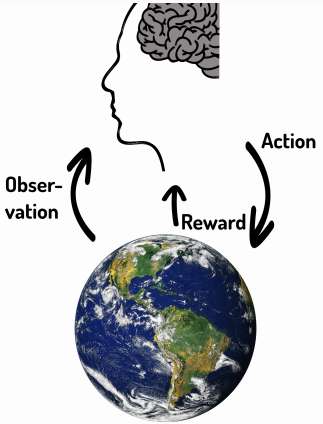
active decision-making agent controls a stochastic process with **known dynamics**

- seeks to find the best possible strategy to achieve a goal by computations, often in a continuous-time setting
- Bellman principle, dynamic programming, HJB equations, ...

Reinforcement learning:

based on the use of algorithms relying on experiences or trials, often in discrete time

- dynamics/rewards do not have to be known ("model-free")
- limited convergence guarantees
- Q-learning, deep learning, AlphaGo, ...



Stochastic control vs. reinforcement learning

	stochastic control	**reinforcement learning**
system dynamics	assumes a concrete mathematical model	can learn from interactions without explicit modeling in a general MDP set-up
knowledge of dynamics	requires full knowledge	can learn from experience
exploration vs. exploitation	focuses on exploiting the known dynamics to optimize control of actions	agents actively explore the environment to gather information and discover optimal policies

C. Rudin et al. (2022): *Interpretable machine learning: Fundamental principles and 10 grand challenges*:

> *"In deep reinforcement learning, policies are defined by deep neural networks, which helps to* **solve complex applied problems**, *but typically the policies are* **essentially impossible to understand or trust**."

On the way to data-driven stochastic control

- the very beginning:

  S. Christensen and C. Strauch (2023). Nonparametric learning for **impulse control** problems. *Ann. Appl. Probab.*, **33**, no. 2, 1369–1387.

- extension to singular control problems and the Lévy framework:

  S. Christensen, C. Strauch and L. Trottner (2024). Learning to reflect: A **unifying approach** for data-driven stochastic control strategies. *Bernoulli*, **30**, no. 3, 2074–2101.

- extension to the multivariate case:

  S. Christensen, A. Holk Thomsen and L. Trottner (2023). Data-driven rules for **multidimensional** reflection problems. `arXiv:2311.06639`.

- refining the statistical analysis/providing minimax optimality:

  S. Christensen, N. Dexheimer and C. Strauch (2023). Data-driven optimal stopping: A **pure exploration** analysis. `arXiv:2312.05880`.

General mathematical framework:

- consider a regular scalar **Itô diffusion process** with dynamics

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \quad t \geqslant 0$$

- **assumptions on the uncontrolled diffusion:** locally Lipschitz, growth condition, and

$$\forall |x| > A: \ \frac{b(x)}{\sigma^2(x)}\,\mathrm{sgn}(x) \leqslant -\gamma$$

- in particular: there exists an **invariant density**

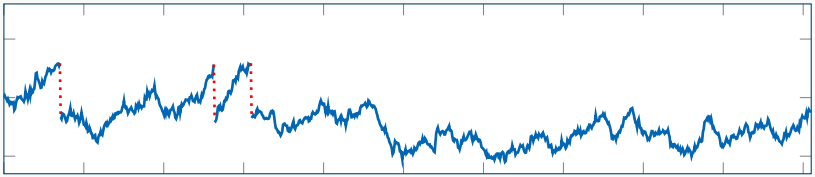
$$\rho(x) \coloneqq \rho_b(x) \coloneqq \frac{1}{C_{b,\sigma}\sigma^2(x)} \exp\left(2\int_0^x \frac{b(y)}{\sigma^2(y)}\mathrm{d}y\right), \quad x \in \mathbb{R}$$

- **notation:** denote by Σ the class of "sufficiently regular" drift functions b

(I) Impulse control problem: Basic framework

Set $K = (\tau_n)_{n\in\mathbb{N}}$, $\tau_1 < \tau_2 < \dots$ an increasing sequence of stopping times, and consider the controlled process $X = X^K$ fulfilling

- on (τ_k, τ_{k+1}), $\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t$,
- $X(\tau_k) = y_0$ fix (in the sequel, $y_0 \equiv 0$).



Goal: Given a payoff function g, choose the intervention times to **maximize the asymptotic growth rate**, i.e., maximize

$$\Phi_b(g) \coloneqq \sup_K \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}_b\left[\sum_{n:\tau_n\leqslant T} g(X^K_{\tau_n-})\right].$$

(I) Solution for impulse control problem (with full knowledge of dynamics)

Denote $\tau_y \coloneqq \inf\{t \geqslant 0 : X_t \geqslant y\}$, $y \in \mathbb{R}$. For each y, the corresponding threshold strategy has value

$$\frac{g(y)}{\xi_b(y)}, \quad y > 0,$$

where, for ρ_b and F_b denoting the invariant density and the associated cdf, respectively,

$$\xi_b(y) \coloneqq \mathbb{E}_b[\tau_y] = 2\int_0^y \frac{F_b(r)}{\rho_b(r)\sigma^2(r)}\mathrm{d}r.$$

Theorem (Alvarez (2004); Helmes et al. (2017))

Under mild assumptions, the value Φ for the impulse control problem is given as the **maximum**

$$\Phi_b(g) = \sup_{y>0} \frac{g(y)}{\xi_b(y)},$$

and the **threshold strategy** $\widehat{K} = (\widehat{\tau}_n)_{n\in\mathbb{N}}$ with $\widehat{\tau}_n = \inf\{t \geqslant \widehat{\tau}_{n-1} : X_t \geqslant y^*\}$, $\widehat{\tau}_0 \coloneqq 0$, for the maximizer $y^* = y_b^*$ of $y \mapsto \frac{g(y)}{\xi_b(y)}$ is optimal.

(II) Singular control problem: Basic framework

As before, we consider a regular scalar **Itô diffusion process** with dynamics

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \quad t \geqslant 0,$$

admitting an invariant density $\rho = \rho_b$. In the set-up of **singular control**, with $Z = (U_t, D_t)_{t\geqslant 0}$, U, D non-decreasing, right-continuous and adapted, the controlled process X^Z fulfills

$$\mathrm{d}X_t^Z = b(X_t^Z)\mathrm{d}t + \sigma(X_t^Z)\mathrm{d}W_t + \mathrm{d}U_t - \mathrm{d}D_t.$$

Goal: Given continuous, nonnegative running cost function c, $q_u, q_d > 0$, **minimize the cost functional**

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T c(X_s^Z)\mathrm{d}s + q_u U_T + q_l D_T\right].$$

(II) Solution for singular control problem (with full knowledge of dynamics)

For each (c, d), the corresponding reflection strategy has value

$$C(c,d) = \frac{1}{\int_c^d \rho_b(x)\mathrm{d}x}\left(\int_c^d c(x)\rho_b(x)\mathrm{d}x + \frac{q_u\sigma^2(c)}{2}\rho_b(c) + \frac{q_d\sigma^2(d)}{2}\rho_b(d)\right).$$

Theorem (Alvarez (2018))

Under some assumptions, the value for the singular control problem is given by

$$V_{\text{sing}} = \min_{(c,d)} C(c,d),$$

and the reflections strategy for the minimizer (c^*, d^*) is optimal.

Central assumption in stochastic control:

The dynamics of the underlying process X are completely known.

What to do if this is not the case?

- Which are the relevant characteristics of X to estimate the optimal level/optimal boundaries?
- How does controlling the process influence the estimation?

Auxiliary question:

How to solve the problem when we use information of an **independent uncontrolled process** to **estimate** the optimal level/boundaries?

(I) Key statistical question for the impulse control problem

- If the diffusion dynamics are unknown: How to estimate

$$\Phi_b(g) = \sup_{y>0} \frac{g(y)}{\xi_b(y)}, \qquad \text{where } \xi_b(y) = 2\int_0^y \frac{F_b(r)}{\rho_b(r)\sigma^2(r)}\mathrm{d}r,$$

and the corresponding maximizer $y^* = \arg\max_{y>0} \frac{g}{\xi_b}(y)$, using a continuous record of observations $(X_s)_{0\leqslant s\leqslant T}$ of the uncontrolled process?

- **Plug-in approach:** Define the nonparametric estimators

$$\widehat{\rho}_T(x) \coloneqq \frac{1}{T}\int_0^T K_T(x - X_s)\mathrm{d}s \quad \text{and} \quad \widehat{F}_T(x) \coloneqq \int_0^T \mathbb{1}\{X_s \in (-\infty, x)\}\mathrm{d}s, \quad x \in \mathbb{R},$$

where $K_T(x) \coloneqq \sqrt{T}K(\sqrt{T}x)$, for a bounded kernel function K with compact support, and use

$$\widehat{\xi}_T(y) \coloneqq 2\int_0^y \frac{\widehat{F}_T(r)}{\widehat{\rho}_T(r)\sigma^2(r) \vee c}\mathrm{d}r \vee c \quad \text{and} \quad \widehat{y}_T \in \arg\max_{y>0} \frac{g}{\widehat{\xi}_T}(y).$$

(I) Bounding the simple regret

How can we bound the **expected loss** when following the threshold strategy with $\widehat{y}_T$ instead of the true maximum y^*, i.e., the **simple regret**

$$\mathbb{E}_b\left[\Phi_b(g) - \frac{g}{\xi_b}(\widehat{y}_T)\right]?$$

Denote

$$\mathcal{G} \coloneqq \left\{ g \in C((0,\infty)) : \left\{ \begin{array}{l} g(0+) < 0, y_1 = \inf\{y > 0 : g(y) > 0\}, \\ \forall b \in \Sigma : \frac{g(y)}{\xi_b(y)} \leqslant \sup_{z\in(0,\zeta]} \frac{g(z)}{\xi_b(z)} \; \forall y > 0 \\ \sup_{z\in[y_1,\zeta]} |g(z)| \leqslant M \end{array} \right. \right\},$$

with $0 < y_1 < \zeta < \infty$ and $0 < M < \infty$, i.e., $\mathcal{G}$ is the class of "sufficiently regular" (bounded, continuous) payoff functions g for which it is known that $y^* \in [y_1, \zeta]$.

(I) Bounding the simple regret

How can we bound the **expected loss** when following the threshold strategy with $\widehat{y}_T$ instead of the true maximum y^*, i.e., the **simple regret**

$$\mathbb{E}_b\left[\Phi_b(g) - \frac{g}{\xi_b}(\widehat{y}_T)\right]?$$

Then, by definition, for any $g \in \mathcal{G}$ and any $p \geqslant 1$,

$$\begin{aligned}
\mathbb{E}_b\left[\left(\Phi_b(g) - \frac{g}{\xi_b}(\widehat{y}_T)\right)^p\right]^{1/p} &\leqslant \mathbb{E}_b\left[\left(\Phi_b(g) - \frac{g}{\xi_b}(\widehat{y}_T) + \frac{g}{\widehat{\xi}_T}(\widehat{y}_T) - \frac{g}{\widehat{\xi}_T}(y^*)\right)^p\right]^{1/p} \\
&\leqslant 2\mathbb{E}_b\left[\sup_{y\in[y_1,\zeta]}\left(\left|\frac{g}{\xi_b}(y) - \frac{g}{\widehat{\xi}_T}(y)\right|\right)^p\right]^{1/p} \\
&\lesssim \sup_{y\in[y_1,\zeta]}\left(\mathbb{E}_b\left[|\widehat{\rho}_T(x) - \rho_b(x)|^p\right]^{1/p} + \mathbb{E}_b\left[\left|\widehat{F}_T(x) - F_b(x)\right|^p\right]^{1/p}\right).
\end{aligned}$$

(I) Bounding the simple regret (general case)

Proposition (Christensen & CS (23); Christensen, Dexheimer, CS (23))

For any $T > 0$, $p \geqslant 1$, it holds

$$\sup_{b\in\Sigma}\sup_{x\in\mathbb{R}}\left(\mathbb{E}_b\left[|\widehat{\rho}_T(x)-\rho_b(x)|^p\right]^{1/p}+\mathbb{E}_b\left[\left|\widehat{F}_T(x)-F_b(x)\right|^p\right]^{1/p}\right)\lesssim T^{-1/2}\left(p^{1/2}+pT^{-1/2}\right).$$

Consequently,

$$\sup_{b\in\Sigma}\sup_{g\in\mathcal{G}}\mathbb{E}_b\left[\left(\Phi_b(g)-\frac{g}{\xi_b}(\widehat{y}_T)\right)^p\right]^{1/p}\leqslant T^{-1/2}\left(p^{1/2}+pT^{-1/2}\right).$$

- bound is **nonasymptotic**, estimator achieves **parametric rate**, fully **data-driven**
- explicit bound for **all moments**

Previous bound holds true for **general payoff functions**, can we improve this under more specific assumptions?

Assumption A (margin condition, Tsybakov noise condition)

Let $\Delta_0 \in (0,1)$, $n \in \mathbb{N}$, $\eta, \beta > 0$, and let f be a continuous function on $(0,\infty)$ fulfilling $\sup_{x\in(0,\infty)} f(x) < \infty$. We say that f satisfies Assumption A if there exist $x_1, \dots, x_n \in (0,\infty)$ such that $f(x_i) = \sup_{x\in(0,\infty)} f(x) < \infty$ for all $i = 1, \dots, n$ and

$$\forall 0 < \Delta \leqslant \Delta_0, \quad \mathcal{X}_f(\Delta) \subseteq \bigcup_{i=1}^{n} \left(x_i - \frac{1}{2}\eta\Delta^\beta,\ x_i + \frac{1}{2}\eta\Delta^\beta \right),$$

where

$$\mathcal{X}_f(\Delta) \coloneqq \left\{ x \in (0,\infty) : \sup_{y\in(0,\infty)} f(y) - f(x) \leqslant \Delta \right\}, \quad \Delta > 0.$$

$\rightsquigarrow$ β intuitively measures the **difficulty of identifying the true maximiser**

(I) Bounding the simple regret: Margin condition

For each (sufficiently regular) $b \in \Sigma$, define the class of payoff functions

$$\mathcal{G}_b(\beta) \coloneqq \left\{ g \in \mathcal{G} : \frac{g}{\xi_b} \text{ satisfies Assumption A} \right\}.$$

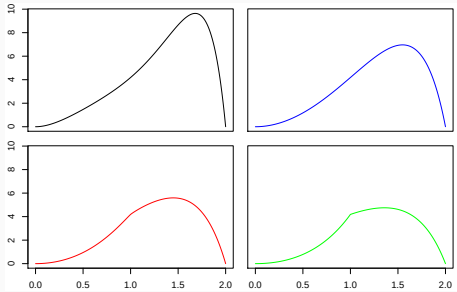


Figure 1: Plots of the function $g(x) = (1 - |1 - x|^{1/\beta})\xi_b(x)$ for $\beta = 0.25, 0.5, 0.75, 1$, with $b(x) = -x/2$.

(I) Bounding the simple regret under the margin condition

Theorem (Christensen, Dexheimer, CS (23))

For any $T > 0$, $\beta > 0$, $p \geqslant 1$,

$$\sup_{b \in \Sigma} \sup_{g \in \mathcal{G}_b(\beta)} \mathbb{E}_b \left[\left(\Phi_b(g) - \frac{g}{\xi_b}(\widehat{y}_T) \right)^p \right] \leqslant \inf_{0 < \alpha < \beta \wedge 1} C_1^{\frac{p}{1-\alpha}} T^{-\frac{p}{2-2\alpha}} \left(\left(\frac{p}{1-\alpha} \right)^{\frac{p}{2-2\alpha}} + \left(\frac{p}{1-\alpha} \right)^{\frac{p}{1-\alpha}} T^{-\frac{p}{2-2\alpha}} \right).$$

In particular,

$$\sup_{b \in \Sigma} \sup_{g \in \mathcal{G}_b(\beta)} \mathbb{E}_b \left[\Phi_b(g) - \frac{g}{\xi_b}(\widehat{y}_T) \right] \in \mathrm{O}\left(\Psi(\beta, T) \right),$$

where

$$\Psi(\beta, T) \coloneqq \begin{cases} T^{-\frac{1}{2-2\beta}}, & 0 < \beta < 1, \\ \exp\left(-C_2 T\right), & \beta \geqslant 1. \end{cases}$$

(I) PAC bounds for the simple regret under the margin condition

Since we can explicitly control **all moments of the simple regret**, we also obtain the following **PAC bounds**.

Corollary (Christensen, Dexheimer, CS (23))

For any $\delta \in (0, e^{-1}]$, $\varepsilon \in (0,1)$, the uniform PAC bound

$$\sup_{b \in \Sigma} \sup_{g \in \mathcal{G}_b(\beta)} \mathbb{P}_b \left(\Phi_b(g) - \frac{g}{\xi_b}(\widehat{y}_T) \geqslant \varepsilon \right) \leqslant \delta$$

holds, if $\beta \in (0,1)$, for any

$$T \geqslant \frac{4 C_1^2 e^{2-2\beta} \log(\delta^{-1})}{(1-\beta)\varepsilon^{2-2\beta}},$$

and, if $\beta \geqslant 1$, for any

$$T \geqslant \frac{4 C_1^2}{\log(2)} \log(2\delta^{-1}) \log(e\varepsilon^{-1}).$$

(II) Using estimators (singular control)

Given some **invariant density estimator** $\widehat{\rho}_T$, use

$$\widehat{C}_T(c,d) \coloneqq \frac{1}{\int_c^d \widehat{\rho}_T(x)\mathrm{d}x}\left(\int_c^d c(x)\widehat{\rho}_T(x)\mathrm{d}x + \frac{q_u\sigma^2(c)}{2}\widehat{\rho}_T(c) + \frac{q_d\sigma^2(d)}{2}\widehat{\rho}_T(d)\right)$$

and

$$\widehat{(c,d)}_T \in \arg\min_{(c,d)} \widehat{C}_T(c,d).$$

Proposition (Christensen, CS, Trottner (24))

Assume that we have a data-driven estimator $\widehat{\rho}_T$ for ρ. Then,

$$\begin{aligned}\mathbb{E}_b\left[V_{\text{sing}} - C(\widehat{(c,d)}_T)\right] &\leqslant 2\mathbb{E}_b\left[\max_{(c,d)}\left|C(c,d) - \widehat{C}_T(c,d)\right|\right] \\ &\lesssim \mathbb{E}_b\left[\|\widehat{\rho}_T - \rho_b\|_{L^\infty}\right].\end{aligned}$$

$\Rightarrow$ **need nonparametric bounds for** $\mathbb{E}_b\left[\|\widehat{\rho}_T - \rho_b\|_{L^\infty}\right]$

Exploration vs. exploitation

- previous results require estimation based on data of the **uncontrolled** process

$\Rightarrow$ high associated costs from **not controlling** the process

- estimation of optimal boundary based on **controlled process** cannot be expected to converge

Problem

Exploration vs. exploitation!

- **our approach:** divide time axis into **exploration periods of length** S_T (estimate optimal intervention time/boundary based on uncontrolled process) and **exploitation periods** (control the process based on estimates to reduce the costs)
- optimal strategy must **balance effects** by determining exploration/exploitation ratio

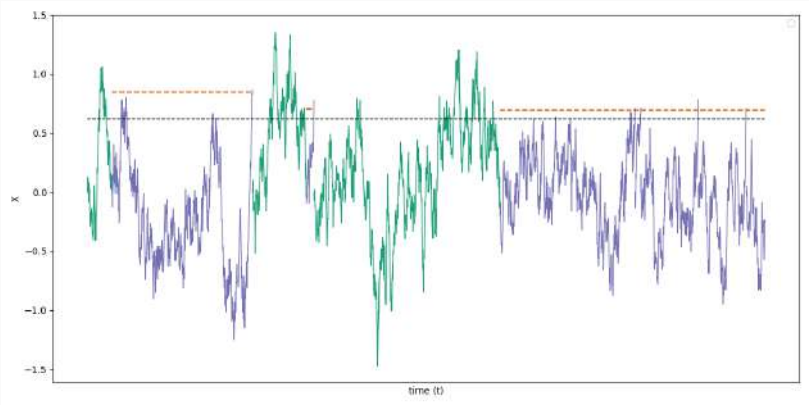


Figure 2: Illustration of a path controlled by data-driven impulse strategy with drift function $b(x) = -2x$, diffusion coefficient $\sigma(x) = 1$, payoff function $g(x) = 0.7 - |1 - x|$. The exploration periods are marked green, the exploitation periods purple, and the estimated thresholds orange. The optimal threshold is marked with the black dashed line.

(I) Upper bound for the cumulative regret

- for the **cumulative regret of the data-driven strategy**, we obtain

$$\Phi_b(g)\,T - \mathbb{E}_b\left[\sum_{n:\tau_n \leqslant T} g(\widehat{y}_{\tau_n})\right] \lesssim S_T + T\psi(S_T),$$

  where ψ is an upper bound for the **simple regret** of the proposed estimator

- choosing

$$S_T \sim T\psi(S_T) = \begin{cases} T^{2/3}, & \text{general case,} \\ T^{\frac{2-2\beta}{3-2\beta}}, & \text{margin condition with } \beta \in (0,1), \\ \log T, & \text{margin condition with } \beta \geqslant 1 \end{cases}$$

  asymptotically yields the best result of a cumulative regret of order

$$\Phi_b(g) - \frac{1}{T}\mathbb{E}_b\left[\sum_{n:\tau_n \leqslant T} g(\widehat{y}_{\tau_n})\right] \lesssim \begin{cases} T^{-1/3}, & \text{general case,} \\ T^{-\frac{1}{3-2\beta}}, & \text{margin condition with } \beta \in (0,1), \\ \frac{\log T}{T}, & \text{margin condition with } \beta \geqslant 1 \end{cases}$$

(I) What about minimax optimality of the upper bounds on the regret?

We consider the **general case**.

Theorem (Christensen, Dexheimer, CS (23))

- (**simple regret**) For any loss function ℓ, there exist constants $c_{\ell,1}, c_{\ell,2} > 0$ such that, for large enough T,

$$\inf_{\widetilde{y}_T} \sup_{b \in \Sigma} \sup_{g \in \mathcal{G}} \mathbb{E}_b \left[\ell \left(c_{\ell,1} T^{1/2} \left(\Phi_b(g) - \frac{g(\widetilde{y}_T)}{\xi_b(\widetilde{y}_T)} \right) \right) \right] \geqslant c_{\ell,2},$$

  where the infimum extends over all estimators $\widetilde{y}_T$.

- (**cumulative regret**) There exists a constant $c_3 > 0$ such that, for large enough $T > 0$,

$$\inf_{K} \sup_{b \in \Sigma} \sup_{g \in \mathcal{G}} \left(\Phi_b(g)\, T - \mathbb{E} \left[\sum_{n : \tau_n \leqslant T} g(X^K_{\tau_{n-}}) \right] \right) \geqslant c_3 \sqrt{T},$$

  where the infimum extends over all impulse control strategies wrt $\mathcal{G}$.

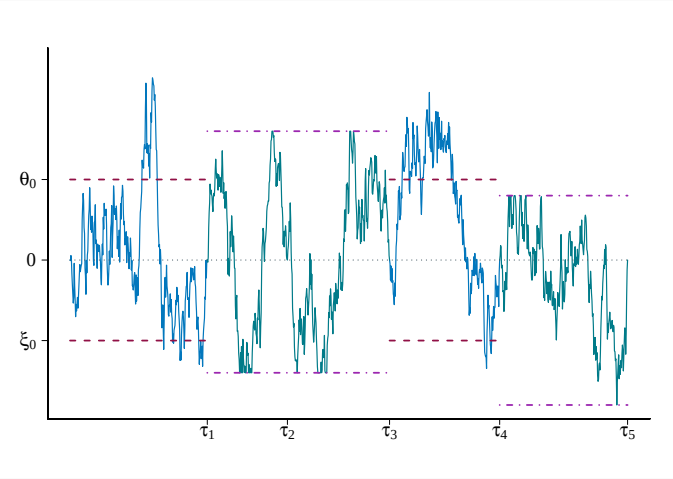


Figure 3: A path controlled using a data-driven reflection strategy with exploration (blue) and exploitation (turquoise) periods. The estimated optimal reflection boundaries are represented by purple lines.

Theorem (Christensen, CS, Trottner (24))

Assume that we have a data-driven estimator $\widehat{\rho}_T$ for ρ with

$$\mathbb{E}_b^0 \left[\|\widehat{\rho}_T - \rho_b\|_{L^\infty} \right] \in \mathrm{O}\left(\sqrt{\log T / T} \right).$$

If $S_T \approx T^{2/3}$, the regret is of order $\mathrm{O}\big(\sqrt{\log T}\, T^{-1/3}\big)$.

- **missing piece for our data-driven strategy:** estimator $\widehat{\rho}_T$ of ρ with sup-**norm rate** $\mathrm{O}(\sqrt{\log T / T})$
- **assumption:** continuous record $X^T = (X_t)_{t \in [0,T]}$ available
- **classical candidate: kernel density estimator**

$$\widehat{\rho}_{h,T}(x) := \frac{1}{hT} \int_0^T K\left(\frac{x - X_t}{h} \right) \mathrm{d}t, \quad x \in \mathbb{R}$$

Controlling the risk of estimators

Two approaches:

① make use of specific structure of diffusions by employing **local time** and **continuous martingale techniques** (Aeckerle-Willems & CS (2021))

② use **mixing properties** to control the long-time transitional behaviour and **heat kernel bounds** on the transition density for the short-time behaviour (Dexheimer, CS, Trottner (2022))

↝ both approaches allow to handle deviation inequalities and moment bounds for suprema of empirical processes of the form

$$\sup_{g\in\mathcal{G}} \Big| \underbrace{\frac{1}{\sqrt{T}} \int_0^T g(X_s)\mathrm{d}s}_{=:\mathbb{G}_T(g)} \Big|, \quad \mathcal{G} \subset L^\infty(\mathbb{R}),$$

via Talagrand's **generic chaining** device

Putting the pieces together

Theorem (Christensen, CS, Trottner (24))

Assume that we have a data-driven estimator $\widehat{\rho}_T$ for ρ with

$$\mathbb{E}_b^0 \left[\|\widehat{\rho}_T - \rho_b\|_{L^\infty} \right] \in \mathrm{O}\left(\sqrt{\log T / T} \right).$$

If $S_T \approx T^{2/3}$, the regret is of order $\mathrm{O}\big(\sqrt{\log T}\, T^{-1/3}\big)$.

Under standard assumptions on drift b and diffusion coefficient guaranteeing

(a) **exponential ergodicity** of X,

(b) **heat kernel bound** on semigroup, i.e., $\sup_{x,y\in\mathbb{R}} p_t(x,y) \lesssim 1/\sqrt{t}$, $t \in (0,1)$,

it follows from general sup-norm estimation results for Markov processes (Dexheimer, CS, Trottner (2022)) that for any bounded, open set D and $h_T \sim \log^2 T/\sqrt{T}$,

$$\mathbb{E}_b^0 \left[\|\widehat{\rho}_{h_T,T} - \rho_b\|_{L^\infty(D)} \right] \in \mathrm{O}\left(\sqrt{\log T / T} \right) \quad \checkmark .$$

For (general) stationary, exponentially β-mixing Markov processes with inv. distribution μ, i.e., $\beta(t) = \int \|P_t(x,\cdot) - \mu\|_{TV}\, \mu(\mathrm{d}x) \lesssim \exp(-\kappa t)$, we obtain for $m_T \leqslant T/4, \tau \in [m_T, 2m_T]$,

$$\left(\mathbb{E}_\mu\left[\sup_{g\in\mathcal{G}} |\mathbb{G}_T(g)|^p\right]\right)^{1/p} \leqslant C_1 \int_0^\infty \log \mathcal{N}\big(u, \mathcal{G}, \tfrac{2m_T}{\sqrt{T}} d_\infty\big)\mathrm{d}u + C_2 \int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}, d_{\mathbb{G},\tau})}\mathrm{d}u$$
$$+ 4 \sup_{g\in\mathcal{G}} \left(\frac{2m_T}{\sqrt{T}} \|g\|_\infty c_1 p + \|g\|_{\mathbb{G},\tau} c_2 \sqrt{p} + \frac{1}{2}\|g\|_\infty c_\kappa \sqrt{T} \mathrm{e}^{-\frac{\kappa m_T}{p}}\right),$$

where $d_{\mathbb{G},\tau}(f,g) = \mathrm{Var}(\mathbb{G}_\tau(f-g))$. With the decomposition

$$\mathbb{E}_\mu\left[\left\|\widehat{\rho}_{h,T} - \rho\right\|_{L^\infty(D)}\right] = \|\mathbb{E}_\mu[\widehat{\rho}_{h,T}(\cdot)] - \rho\|_{L^\infty(D)} + \mathbb{E}_\mu\left[\left\|\widehat{\rho}_{h,T} - \mathbb{E}_\mu[\widehat{\rho}_{h,T}(\cdot)]\right\|_{L^\infty(D)}\right]$$
$$=: \boldsymbol{B} + \boldsymbol{V},$$

we can use the general result to bound the stochastic error $\boldsymbol{V}$ via

$$\boldsymbol{V} = \frac{1}{\sqrt{T}h}\mathbb{E}_\mu\left[\sup_{g\in\mathcal{G}} |\mathbb{G}_T(g)|\right], \quad \mathcal{G} = \left\{K\big(\tfrac{x-\cdot}{h}\big) - \mu\big(K\big(\tfrac{x-\cdot}{h}\big)\big) : x \in D \cap \mathbb{Q}\right\}.$$

Data-driven approaches to stochastic control

Simple recipe:

(A) Identify a stochastic control problem which admits an explicit solution;

(B) identify the dependence on the underlying diffusion dynamics;

(C) use a plug-in approach to obtain a data-driven estimate.

What makes the problems studied so far tractable?

Estimation problem boils down to nonparametric analysis of invariant density/cdf estimators for scalar diffusions.

$\hookrightarrow$ parametric convergence rate

$\hookrightarrow$ bandwidth of kernel density estimators can be chosen independently of the smoothness

$\hookrightarrow$ specific for invariant density estimation of scalar diffusion processes from continuous observations

Summary

- we develop **data-driven strategies** for the determination of **optimal intervention times/boundaries** appearing in stochastic optimal control problems associated to diffusion processes under uncertainty on the underlying dynamics
- central to the approach are efficient nonparametric estimators of the cost/payoff-representing functionals
- we show improved results for the simple regret under more specific assumptions (**margin condition**) on the payoff functions
- **optimality of the results on the simple regret** is verified by providing matching minimax lower bounds
- we provide a solution for the **emerging exploration-exploitation dilemma**

Thank you for your attention!

Further references

C. Aeckerle-Willems and C. Strauch (2021). Concentration of scalar ergodic diffusions and some statistical implications. *Ann. Inst. Henri Poincaré Probab. Stat.*, **57**, no. 4, 1857–1887.

L. H. R. Alvarez (2004). A class of solvable impulse control problems. *Appl. Math. Optim.*, **49**, no. 3, 265–295.

N. Dexheimer, C. Strauch and L. Trottner (2022). Adaptive invariant density estimation for continuous-time mixing Markov processes under sup-norm risk. *Ann. Inst. Henri Poincaré Probab. Stat.*, **58**, no. 4, 2029–2064.

K. L. Helmes, R. H. Stockbridge and C. Zhu (2017). Continuous inventory models of diffusion type: long-term average cost criterion. *Ann. Appl. Probab.*, **27**, no. 3, 1831–1885.