Snapshots of statistics for SPDEs, optimal control and generative models

Data Science and Computational Statistics Seminar – University of Birmingham

Lukas Trottner
based on joint works with Sören Christensen, Asbjørn Holk, Markus Reiß, Claudia Strauch and Anton Tiepner

11 November 2024

University of Birmingham Kiel University Aarhus University
Humboldt University of Berlin Heidelberg University

UNIVERSITY OF BIRMINGHAM

Overview of current main research interests

1. Data-driven stochastic optimal control
    - if the underlying stochastic process has unknown dynamics, how can we determine a control procedure with sublinear regret?
    - exploration/exploitation tradeoff
2. Statistical aspects of deep generative models
3. Statistics for SPDEs
    - estimate structural breaks in a material from observations of a heat flow that is subject to random perturbations
    - explore methodological connections to change point and image reconstruction methods

A model problem for data-driven optimal control

- consider a d-dimensional Langevin diffusion

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t;$$

  if ergodic: stationary density $\pi \propto \exp(-V(\cdot))$
- we play the following game:
  1. the aim is to keep the process close to a target state, say 0, at minimal long run costs
  2. normally reflect the process in a domain D that we are free to choose:

$$\mathrm{d}X_t^D = -\nabla V(X_t^D)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t + n(X_t^D)\,\mathrm{d}L_t^D$$

  3. costs:

$$J_T(D) = \underbrace{\int_0^T c(X_t^D)\,\mathrm{d}t}_{c \text{ increasing in } |x|} + \underbrace{\kappa L_T^D}_{\text{reflection costs}}$$

A model problem for data-driven optimal control

- consider a d-dimensional Langevin diffusion

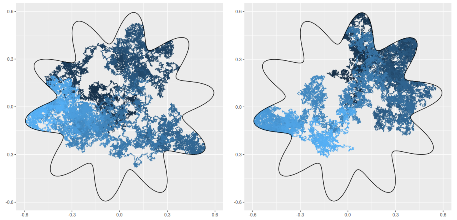
$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t;$$

  if ergodic: stationary density $\pi \propto \exp(-V(\cdot))$
- we play the following game:
  1. the aim is to keep the process close to a target state, say 0, at minimal long run costs
  2. normally reflect the process in a domain D that we are free to choose:

$$\mathrm{d}X_t^D = -\nabla V(X_t^D)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t + n(X_t^D)\,\mathrm{d}L_t^D$$

  3. costs:

$$J_T(D) = \underbrace{\int_0^T c(X_t^D)\,\mathrm{d}t}_{c \text{ increasing in } |x|} + \underbrace{\kappa L_T^D}_{\text{reflection costs}}$$



A model problem for data-driven optimal control

- consider a d-dimensional Langevin diffusion

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t;$$

  if ergodic: stationary density $\pi \propto \exp(-V(\cdot))$
- we play the following game:
  1. the aim is to keep the process close to a target state, say 0, at minimal long run costs
  2. normally reflect the process in a domain D that we are free to choose:

$$\mathrm{d}X_t^D = -\nabla V(X_t^D)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t + n(X_t^D)\,\mathrm{d}L_t^D$$

  3. costs:

$$J_T(D) = \underbrace{\int_0^T c(X_t^D)\,\mathrm{d}t}_{c \text{ increasing in } |x|} + \underbrace{\kappa L_T^D}_{\text{reflection costs}}$$

- Ergodic optimal control: for an admissible domain class Θ determine

$$D^* \in \operatorname*{arg\,min}_{D\in\Theta} \underbrace{\lim_{T\to\infty} \frac{1}{T}\mathbb{E}[J_T(D)]}_{=:J(D)} \quad (\rightsquigarrow \text{shape optimisation problem})$$

- Data-driven optimal control: If V is unknown, determine an estimator $\widehat{D}$ of D^* based on observations of the (controlled) process

Learning the optimal reflection boundary

- Long term average costs are explicitly given by

$$J(D) = \int_D c(x)\pi_D(x)\,\mathrm{d}x + \kappa \int_{\partial D} \pi_D(x)\,\mathcal{H}_{d-1}(\mathrm{d}x),$$

where $\pi_D(x) = \frac{\exp(-V(x))}{\int_D \exp(-V(x))} = \pi(x)/\pi(D)$

⇝ estimator $\hat{\pi}$ of the invariant density of the unreflected process provides plug-in M-estimator

$$\widehat{D} \in \operatorname*{arg\,min}_{D\in\Theta} \frac{\int_D c(x)\hat{\pi}(x)\,\mathrm{d}x + \kappa \int_{\partial D} \hat{\pi}(x)\,\mathcal{H}_{d-1}(\mathrm{d}x)}{\int_D \hat{\pi}(x)\,\mathrm{d}x}$$

- minimax optimal adaptive estimators of π (in terms of observation horizon T) can be constructed via kernel estimators under anisotropic Hölder smoothness assumptions

Learning the optimal reflection boundary

- Long term average costs are explicitly given by

$$J(D) = \int_D c(x)\pi_D(x)\,\mathrm{d}x + \kappa \int_{\partial D} \pi_D(x)\,\mathcal{H}_{d-1}(\mathrm{d}x),$$

where $\pi_D(x) = \frac{\exp(-V(x))}{\int_D \exp(-V(x))} = \pi(x)/\pi(D)$

⇝ estimator $\hat{\pi}$ of the invariant density of the unreflected process provides plug-in M-estimator

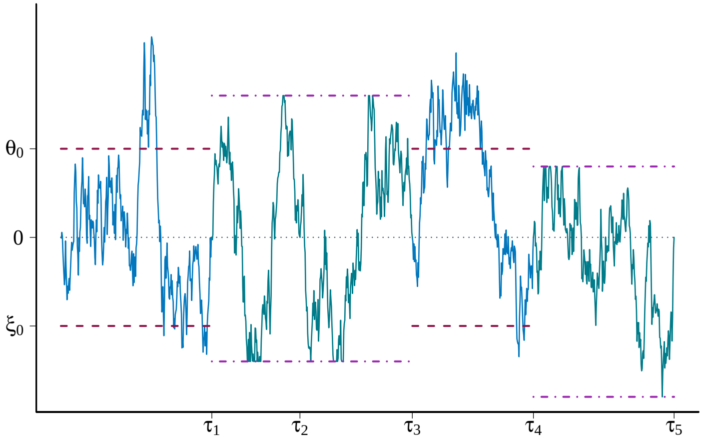
$$\widehat{D} \in \operatorname*{arg\,min}_{D\in\Theta} \frac{\int_D c(x)\hat{\pi}(x)\,\mathrm{d}x + \kappa \int_{\partial D} \hat{\pi}(x)\,\mathcal{H}_{d-1}(\mathrm{d}x)}{\int_D \hat{\pi}(x)\,\mathrm{d}x}$$

- minimax optimal adaptive estimators of π (in terms of observation horizon T) can be constructed via kernel estimators under anisotropic Hölder smoothness assumptions

Problem

Exploration vs. Exploitation

Episodic domain learning



Regret bound for episodic domain learning

Theorem (Christensen, Strauch, T. (2024)[1]; Christensen, Holk Thomsen, T. (2024)[2])

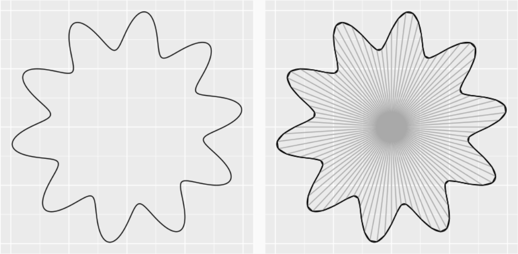
There exists a purely data-driven episodic domain learning strategy $\widehat{Z}$ such that the expected regret per time unit satisfies

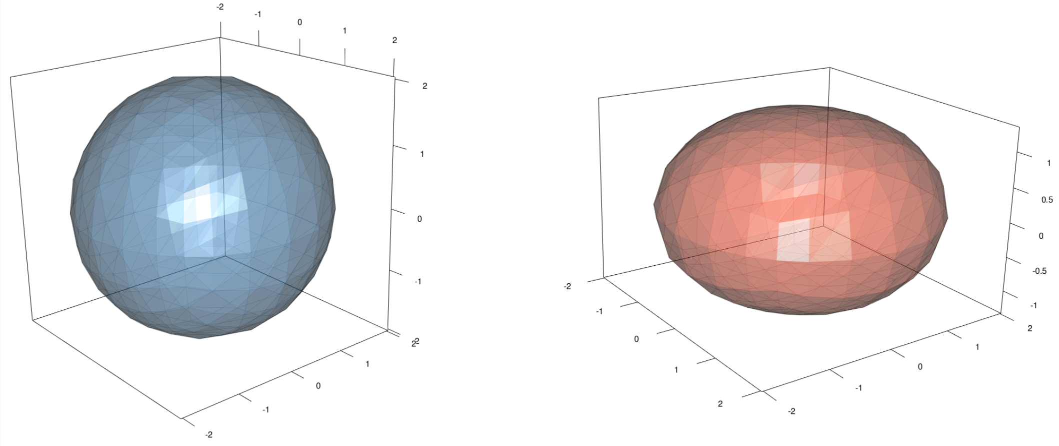
$$\frac{1}{T}\mathbb{E}\Big[\int_0^T c(X_t^{\widehat{Z}})\,\mathrm{d}t + \kappa L_T^{\widehat{Z}}\Big] - J(D^*) \lesssim \begin{cases} \frac{\sqrt{\log T}}{T^{1/3}}, & d = 1, \\ \Big(\frac{(\log T)^2}{T}\Big)^{\frac{1}{3}}, & d = 2, \\ \Big(\frac{\log T}{T}\Big)^{\frac{\overline{\beta}}{3\overline{\beta}+d-2}}, & d \geq 3. \end{cases}$$

[1] Christensen, Strauch and T. (2024). Learning to reflect: A unifying approach for data-driven stochastic control strategies . *Bernoulli.*
[2] Christensen, Holk Thomsen and T. (2024). Data-driven rules for multivariate reflection problems. *SIAM/ASA J. Uncertain. Quantif.*

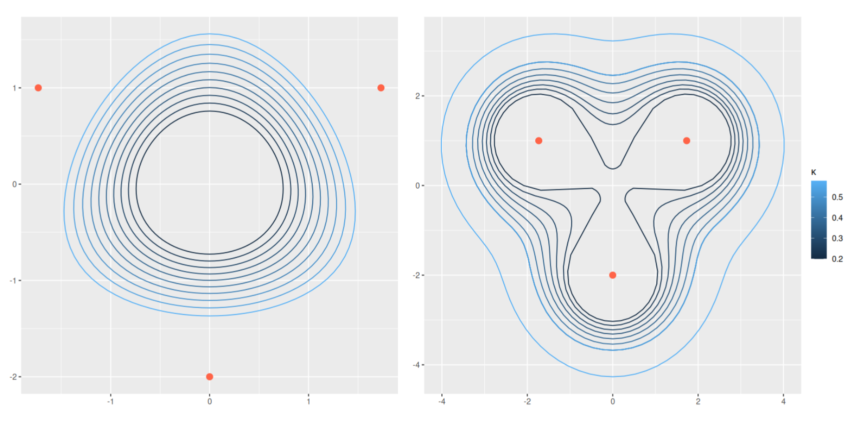
Numerical shape optimisation

- as target domains Θ restrict to strongly star-shaped sets at 0
- for $D \in \Theta$ consider polytope approximation $\widetilde{D}_N$ such that for a sufficiently large number N of spanning points $J(D) \approx J(\widetilde{D}_N) = \tilde{J}(r_1, r_2, \dots, r_N)$
- we derive explicit formulas for $\nabla \tilde{J}(\boldsymbol{r})$, making gradient-based optimisation methods accessible

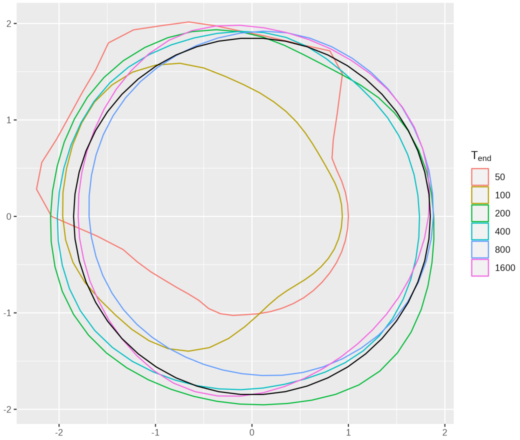




Optimised shapes for Brownian motion with reflection cost $\kappa = 1$ and cost function $c = |\cdot|$ (left) and $c(x, y, z) = \sqrt{x^2 + 5y^2 + z^2}$ (right).

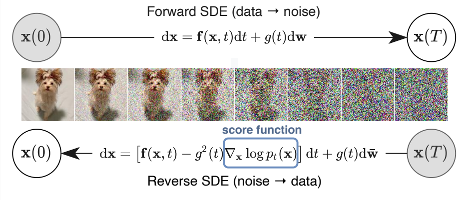


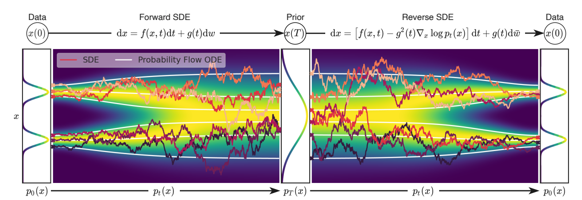
For each κ, we plot the optimized reflection boundaries, where π is a mixture of three Gaussians with means at the points marked in red. Left: Norm cost function, $c = |\cdot|$. Right: Cost function $c(x) = \min\{|x - \mu_1|, |x - \mu_2|, |x - \mu_3|\}$.



Estimates of the optimal shape (black) using kernel estimates after increasing periods of exploration. Notably, after only $T = 150$, the estimated optimal shape has an associated cost only 0.61% higher than the true optimum.

Denoising diffusion models

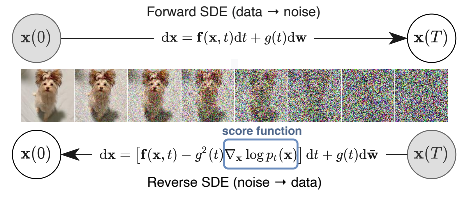


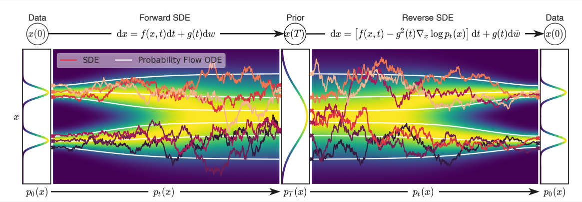


Source: Song et al. (2021). Score based generative modeling through stochastic differential equations. *ICLR*.

- general problem: given iid data $(X_{0,i})_{i=1,\dots,n}$ with unknown distribution p_0, generate a new data sample with (approximately) the same distribution
- denoising diffusion models have demonstrated spectacular generation abilities for vastly different tasks

Denoising diffusion models



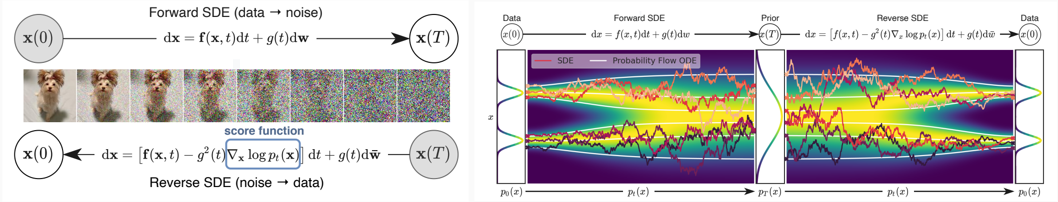


Source: Song et al. (2021). Score based generative modeling through stochastic differential equations. *ICLR*.

Questions

1. are diffusion models minimax learners (in terms of smoothness assumptions on p_0)?
2. how can empirical lack of curse of dimensionality be explained? ⇝ submanifold hypothesis
3. alternative model designs with enhanced theoretical/experimental performance?

Denoising diffusion models



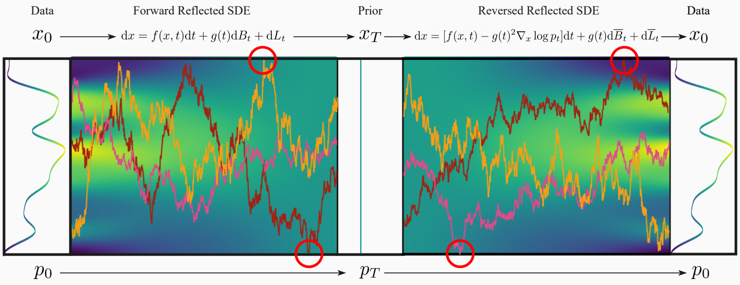
Source: Song et al. (2021). Score based generative modeling through stochastic differential equations. *ICLR*.

Questions

1. are diffusion models minimax learners (in terms of smoothness assumptions on p_0)?
2. how can empirical lack of curse of dimensionality be explained? ⇝ submanifold hypothesis
3. alternative model designs with better theoretical/experimental justification?

- Oko et al. (2023, *ICML*) and Tang and Yang (2024, *AISTATS*) develop statistical theory for "vanilla" diffusion models

Denoising reflected diffusion models

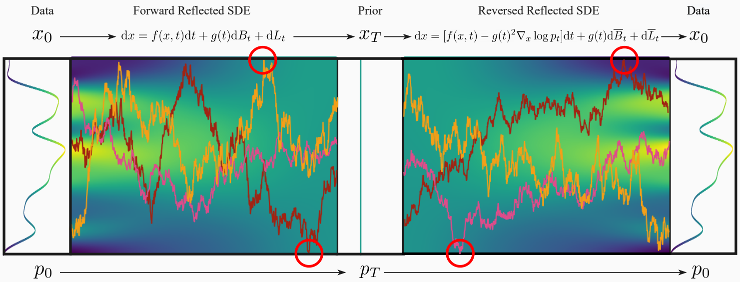


Source: Lou and Ermon (2023). Reflected Diffusion Models. *ICML*.

Questions

1. are diffusion models minimax learners (in terms of smoothness assumptions on p_0)?
2. how can empirical lack of curse of dimensionality be explained? ⇝ submanifold hypothesis
3. alternative model designs with better theoretical/experimental justification?

Denoising reflected diffusion models



Source: Lou and Ermon (2023). Reflected Diffusion Models. *ICML*.

Questions

1. are reflected diffusion models minimax learners (in terms of smoothness assumptions on p_0)?
2. how can empirical lack of curse of dimensionality be explained? ⇝ submanifold hypothesis
3. alternative model designs with enhanced theoretical/experimental justification?

Modelling with symmetric reflected forward model

- we choose as a forward model a normally reflected diffusion on a bounded domain D:

$$\mathrm{d}X_t = \nabla f(X_t)\,\mathrm{d}t + \sqrt{2f(X_t)}\,\mathrm{d}W_t + n(X_t)\,\mathrm{d}L_t, \quad X_0 \sim p_0, \quad f \geq f_{\min} > 0.$$

- exponentially fast convergence to invariant distribution $\mathcal{U}(D)$
- backwards dynamics determined by f and score

$$s^\circ(x,t) = \nabla \log p_t(x),$$

where

$$p_t(x) = \sum_{j=0}^{\infty} \mathrm{e}^{-\lambda_j t}\langle p_0, e_j\rangle_{L^2} e_j(x), \quad (\lambda_j, e_j)_j \text{ eigenpairs of } -\nabla\cdot f\nabla \text{ with Neumann bound. cond.}$$

⇝ calibrate deep neural network class $\mathcal{S}$ that allows approximation with desired accuracy

⇝ denoising score matching estimator:

$$\hat{s} \in \operatorname*{arg\,min}_{s\in\mathcal{S}} \frac{1}{n}\sum_{i=1}^{n}\int_{\underline{T}}^{\overline{T}}\int_D |s(y,t) - \nabla_y \log p_t(X_{0,i}, y)|^2 p_t(X_{0,i}, y)\,\mathrm{d}y\,\mathrm{d}t.$$

Denoising reflected diffusion models are minimax learners

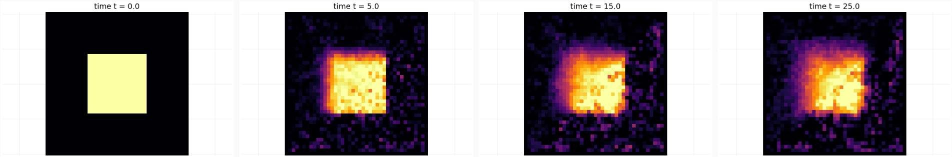
Theorem (Holk, Strauch and T. (2024))

Suppose that $p_0 = \tilde{p}_0 + \alpha$ for some $0 \leq \tilde{p}_0 \in H_c^k(D)$ and $\alpha > 0$, where $k > d/2$. Then, there exists a class of feed forward ReLU neural networks $\mathcal{S}$, with explicit size constraints in terms of n, d and s, such that

$$\mathbb{E}\Big[\operatorname{TV}(p_0, \overleftarrow{p}^{\hat{s}}_{\overline{T}-\underline{T}})\Big] \lesssim n^{-\frac{k}{2k+d}} (\log n)^3 (\log\log n)^{1/2},$$

where $\overline{T} \asymp \log n$ and $\underline{T} \asymp n^{-2k/((2-k/d)\wedge 1)(2k+d)}$.

Change estimation for a stochastic heat equation



- Stochastic heat equation

$$\mathrm{d}X(t) = \Delta_\vartheta X(t)\,\mathrm{d}t + \mathrm{d}W(t), \quad \Delta_\vartheta = \nabla \cdot \vartheta \nabla,$$

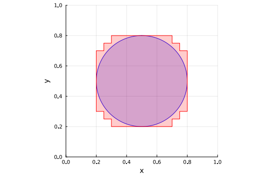
driven by space-time white noise $\dot{W}(t,x)$ and broken diffusivity

$$\vartheta(x) = \vartheta_- \mathbf{1}_{\Lambda_-}(x) + \vartheta_+ \mathbf{1}_{\Lambda_+}(x), \quad x \in [0,1]^d = \Lambda_- \uplus \Lambda_+.$$

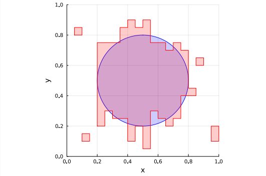
- special case for $d = 1$: $\Lambda_+ = (\tau, 1]$ with change point τ

Estimation approach via local observations

- tile space with hypercubes $\mathrm{Sq}(\alpha)$ of side length δ and aim for estimation of minimal tiling $\Lambda_+^{\updownarrow}$



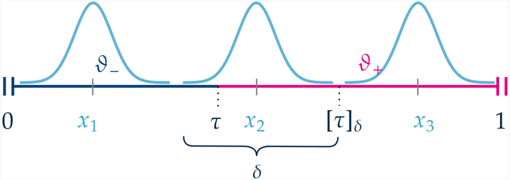
$\Lambda_+^{\updownarrow}$



$\widehat{\Lambda}_+$

- observations are local in space and continuous in time ($t \in [0, T]$, T fixed):

$$X_{\delta,\alpha}(t) = \langle X(t), K_{\delta,\alpha} \rangle, \qquad \text{where } K_{\delta,\alpha}(x) = \delta^{-d/2} K((x - x_\alpha)/\delta),$$
$$X^{\Delta}_{\delta,\alpha}(t) = \langle X(t), \Delta K_{\delta,\alpha} \rangle$$



Simultaneous M-estimator

- local observations yield modified local log-likelihoods $\ell_{\delta,\alpha}(\vartheta_-, \vartheta_+, \Lambda_+)$, where $\Lambda_+ \in \mathcal{A}$ for a family of tiling sets $\mathcal{A} \ni \Lambda_+^{\updownarrow}$

$\rightsquigarrow$ $(\hat{\vartheta}_-, \hat{\vartheta}_+, \widehat{\Lambda}_+) \in \arg\max_{(\vartheta_-,\vartheta_+,\Lambda_+)\in[\underline{\vartheta},\overline{\vartheta}]^2\times\mathcal{A}} \sum_\alpha \ell_{\delta,\alpha}(\vartheta_-, \vartheta_+, \Lambda_+)$

- in the 1D case we furthermore adjust $\ell_{\delta,\alpha}$ to account for the error induced by a constant approximation of ϑ on change point interval $\rightsquigarrow$ fundamentally important to obtain optimal convergence rates for $\vartheta^0_\pm$

jump height $\eta := |\vartheta_+^0 - \vartheta_-^0|$

Theorem (Reiß, Strauch and T., 2023)

(i) non-vanishing signal: $|\eta| \geq \underline{\eta} > 0$ for all $\delta \in 1/\mathbb{N}$. Then,

$$|\hat{\tau} - \tau^0| = \mathcal{O}_{\mathbb{P}}(\delta) \quad \text{and} \quad |\hat{\vartheta}_\pm - \vartheta_\pm^0| = \mathcal{O}_{\mathbb{P}}(\delta^{3/2}).$$

(ii) vanishing signal: $\eta = o(\delta)$, $\delta^{3/2} = o(\eta)$ and $\vartheta_\pm^0 \longrightarrow \vartheta_*$, then

$$\frac{\eta^2}{\delta^3} \frac{T\|K'\|_{L^2}^2}{2\vartheta^*}(\hat{\tau} - \tau^0) \xrightarrow{\mathrm{d}} \operatorname*{arg\,min}_{h\in\mathbb{R}} \left\{B^{\leftrightarrow}(h) + \frac{|h|}{2}\right\}, \qquad \text{as } \delta \to 0.$$

jump height $\eta := |\vartheta_+^0 - \vartheta_-^0|$

Theorem (Reiß, Strauch and T., 2023)

(i) non-vanishing signal: $|\eta| \geq \underline{\eta} > 0$ for all $\delta \in 1/\mathbb{N}$. Then,

$$|\hat{\tau} - \tau^0| = \mathcal{O}_{\mathbb{P}}(\delta) \quad \text{and} \quad |\hat{\vartheta}_\pm - \vartheta_\pm^0| = \mathcal{O}_{\mathbb{P}}(\delta^{3/2}).$$

(ii) vanishing signal: $\eta = o(\delta)$, $\delta^{3/2} = o(\eta)$ and $\vartheta_\pm^0 \longrightarrow \vartheta_*$, then

$$\frac{\eta^2}{\delta^3} \frac{T\|K'\|_{L^2}^2}{2\vartheta^*}(\hat{\tau} - \tau^0) \xrightarrow{\mathrm{d}} \operatorname*{arg\,min}_{h\in\mathbb{R}} \left\{ B^{\leftrightarrow}(h) + \frac{|h|}{2} \right\}, \qquad \text{as } \delta \to 0.$$

Theorem (Tiepner and T., 2024)

Suppose that the number of tiles intersecting $\partial\Lambda_+^0$ is of order $\delta^{-d+\beta}$, $\beta \in (0,1]$. Then,

$$\mathbb{E}\big[\operatorname{vol}_d(\widehat{\Lambda}_+ \triangle \Lambda_+^0)\big] \lesssim \delta^\beta.$$

- Λ_+^0 graph of a β-Hölder function $\rightsquigarrow \mathbb{E}\big[\operatorname{vol}_d(\widehat{\Lambda}_+ \triangle \Lambda_+^0)\big] \lesssim \delta^\beta$
- Λ_+^0 convex $\rightsquigarrow \mathbb{E}\big[\operatorname{vol}_d(\widehat{\Lambda}_+ \triangle \Lambda_+^0)\big] \lesssim \delta$

jump height $\eta := |\vartheta_+^0 - \vartheta_-^0|$

Theorem (Reiß, Strauch and T., 2023)

(i) non-vanishing signal: $|\eta| \geq \underline{\eta} > 0$ for all $\delta \in 1/\mathbb{N}$. Then,

$$|\hat{\tau} - \tau^0| = \mathcal{O}_{\mathbb{P}}(\delta) \quad \text{and} \quad |\hat{\vartheta}_\pm - \vartheta_\pm^0| = \mathcal{O}_{\mathbb{P}}(\delta^{3/2}).$$

(ii) vanishing signal: $\eta = o(\delta)$, $\delta^{3/2} = o(\eta)$ and $\vartheta_\pm^0 \longrightarrow \vartheta_*$, then

$$\frac{\eta^2}{\delta^3}\frac{T\|K'\|_{L^2}^2}{2\vartheta^*}(\hat{\tau} - \tau^0) \xrightarrow{\text{d}} \arg\min_{h\in\mathbb{R}}\left\{B^{\leftrightarrow}(h) + \frac{|h|}{2}\right\}, \qquad \text{as } \delta \to 0.$$

Theorem (Tiepner and T., 2024)

Suppose that the number of tiles intersecting $\partial\Lambda_+^0$ is of order $\delta^{-d+\beta}$, $\beta \in (0,1]$. Then,

$$\mathbb{E}\left[\operatorname{vol}_d(\widehat{\Lambda}_+ \triangle \Lambda_+^0)\right] \lesssim \delta^\beta.$$

- Λ_+^0 graph of a β-Hölder function $\rightsquigarrow \mathbb{E}\left[\operatorname{vol}_d(\widehat{\Lambda}_+ \triangle \Lambda_+^0)\right] \lesssim \delta^\beta$
- Λ_+^0 convex $\rightsquigarrow \mathbb{E}\left[\operatorname{vol}_d(\widehat{\Lambda}_+ \triangle \Lambda_+^0)\right] \lesssim \delta$

Thank you for your attention!