凸最適化によるエルゴード的拡散過程のオンライン推定

ONLINE ESTIMATION OF ERGODIC DIFFUSION PROCESSES WITH CONVEX OPTIMIZATION

仲北祥悟（東京大学大学院総合文化研究科）

Abstract. We propose an online parametric estimation method of stochastic differential equations with discrete observations and misspecified modelling based on online gradient descent. Our study provides uniform risk bounds for the estimators over a family of stochastic differential equations. The derivation of the bounds involves three underlying theoretical results: the analysis of the stochastic mirror descent algorithm based on dependent and biased subgradients, the simultaneous exponential ergodicity of classes of diffusion processes, and the proposal of loss functions whose approximated stochastic subgradients are dependent only on the known model and observations.

Keywords: diffusion processes; discrete observations; misspecified models; online gradient descent; simultaneous ergodicity; stochastic differential equations; stochastic mirror descent

1. Introduction

Let us consider the parametric estimation of the following d-dimensional stochastic differential equation (SDE):

$$\mathrm{d}X_t^{a,b} = b\left(X_t^{a,b}\right)\mathrm{d}t + a\left(X_t^{a,b}\right)\mathrm{d}w_t,\ X_0 = x \in \mathbf{R}^d,\ t \geq 0.$$

SDEs describe dynamics with randomness and allow for flexible model structures under mild conditions. Therefore, they are used to model phenomena in broad disciplines such as finance, biology, epidemiology, physics, meteorology, and machine learning. In this study, we propose an online parametric estimation method of b based on discrete observations $\{X_{ih_n}^{a,b}\}_{i=0,\dots,n}$ with $h_n > 0$.

Batch estimation of SDEs with discrete observations is a classical and important problem for statistics of SDEs [see 4, 20, 5, 2, 11, 7, 12, 6]. Notably, some results achieve the asymptotic efficiency of batch estimators with more efficient computational complexities [19, 10, 8, 9].

Online estimation, where the estimator is updated as data are acquired, is also a typical and significant concern in time series data analysis because it is quite useful for real-time decision making. For example, Kalman filtering is one of the most classical online estimation methods based on time series data. However, most studies on the online parametric estimation of SDEs depend on the setting of continuous observations $\{X_t^{a,b}\}_{t\geq 0}$ [17, 1, 16], which is restrictive in real data analysis. Hence, we aim to propose online estimation methods for SDEs with discrete observations.

We provide uniform risk bounds for the parametric estimation of both diffusion and drift coefficients of SDEs with discrete observations and model misspecification via online gradient descent with convex loss functions and their convex approximations. Those bounds give theoretical convergence guarantees of the proposed online estimation method for SDEs with discrete observations, which are the main contribution of our study. To derive the bounds, we combine the three theoretical discussions: (i) model-wise non-asymptotic risk bound for the stochastic mirror descent (SMD) with dependent and biased subgradients; (ii) simultaneous ergodicity and uniform moment bounds for a class of SDEs; and (iii) the proposal of loss functions for the online parametric estimation.

Selecting drift estimation as an example in this section, we set the convex and compact parameter space $\Theta \subset \mathbf{R}^p$ and the triple of measurable functions (b^{m}, M, J) such that $b^{\mathrm{m}}(x, \theta)$

This work was partially supported by JSPS KAKENHI Grant Number JP21K20318.

is the possibly misspecified parametric model, $M(x)$ is a positive semi-definite weight function, and $J(\theta)$ is the regularization term. We set the function

$$\phi(x,y,\theta) := \frac{1}{2}M(x)\left[(y-b^{\mathrm{m}}(x,\theta))^{\otimes 2}\right] + J(\theta).$$

Assume that $\phi(x,y,\theta)$ is convex in θ for all $x,y \in \mathbf{R}^d$ and has measurable elements in the subdifferential for all x,y, and θ. $\{\theta_i; i=1,\dots,n+1\}$ defined by the following online gradient descent algorithm

$$\theta_{i+1} := \mathrm{Proj}_\Theta\left(\theta_i - \frac{h_n}{\sqrt{i}}\partial_\theta\phi\left(X^{a,b}_{(i-1)h_n}, \frac{1}{h_n}\Delta_i X^{a,b}, \theta_i\right)\right),$$

with an arbitrary initial value $\theta_1 \in \Theta$ and a sequence of discrete observations $\{X^{a,b}_{ih_n}; i = 0,\dots,n\}$, is then well-defined as a sequence of random variables by choosing measurable subgradients, where $\Delta_i X^{a,b} = X^{a,b}_{ih_n} - X^{a,b}_{(i-1)h_n}$ and $h_n > 0$ is the discretization step. Note that the learning rate chosen here does not lead to the best convergence but is simple and approximately the best in our study. Our contributions (i) and (iii) provide the following risk bound for the estimator $\bar{\theta}_n := \frac{1}{n}\sum_{i=1}^n \theta_i$ with a fixed (a,b): for some $c>0$,

$$\sup_{\theta\in\Theta}\left(\mathbf{E}^{a,b}_x\left[f^{a,b}\left(\bar{\theta}_n\right)\right] - f^{a,b}(\theta)\right) \le c\left(\frac{\log nh_n^2}{\sqrt{nh_n^2}} + h_n^{\beta/2}\right),$$

where $\beta \in [0,1]$ is a parameter controlling the smoothness of b, $\mathbf{E}^{a,b}_x$ is the expectation over $\{X^{a,b}_t\}_{t\ge 0}$ with $X^{a,b}_0 = x$, $f^{a,b}(\theta) = \int M(\xi)\left[(b^{\mathrm{m}}(\xi,\theta) - b(\xi))^{\otimes 2}\right]\Pi^{a,b}(\mathrm{d}\xi) + J(\theta)$ is the loss function, and $\Pi^{a,b}$ is the invariant probability measure of $X^{a,b}_t$. The contribution (ii) yields the existence of c such that the inequality holds uniformly in $S := \{(a,b)\}$, a class of coefficients of SDEs satisfying the same regularity conditions; hence, the risk bound is uniform in S. Note that $\bar{\theta}_n$ estimates the best $\theta \in \Theta$ (or the quasi-optimal parameter; see [18]) with $b^{\mathrm{m}}(,\theta)$ closest to the true b in the $L^2\left(\Pi^{a,b}\right)$-distance. Moreover, if the model b^{m} correctly specifies b, that is, for all $(a,b) \in S$ there exists θ such that $b = b^{\mathrm{m}}(\cdot,\theta)$, then we can obtain the following bound:

$$\sup_{(a,b)\in S}\mathbf{E}^{a,b}_x\left[f^{a,b}\left(\bar{\theta}_n\right)\right] \le c\left(\frac{\log nh_n^2}{\sqrt{nh_n^2}} + h_n^{\beta/2}\right).$$

One simple but significant outcome of the above discussion is a non-asymptotic risk guarantee of the following online gradient descent for linear models such that an arbitrary initial value $\theta_1 \in \Theta$,

$$\theta_{i+1} := \mathrm{Proj}_\Theta\left(\theta_i + \frac{1}{\sqrt{i}}\left(\partial_\theta b^{\mathrm{m}}\left(X^{a,b}_{(i-1)h_n},\theta_i\right)\right)\left(\Delta_i X^{a,b} - h_n b^{\mathrm{m}}\left(X^{a,b}_{(i-1)h_n},\theta_i\right)\right)\right),$$

where $b^{\mathrm{m}}(x,\theta)$ is the possibly misspecified parametric model whose components are linear in $\theta \in \Theta$. Note that it corresponds to the case $M(x) = I_d$, $J(\theta) = 0$. As evident, the uniform risk bound for the estimator $\bar{\theta}_n := \frac{1}{n}\sum_{i=1}^n \theta_i$ over a certain family S of the coefficients a,b holds: for some $c>0$,

$$\sup_{(a,b)\in S}\sup_{\theta\in\Theta}\left(\mathbf{E}^{a,b}_x\left[\int \left\|b^{\mathrm{m}}\left(\xi,\bar{\theta}_n\right) - b(\xi)\right\|_2^2 \Pi^{a,b}(\mathrm{d}\xi)\right] - \int \left\|b^{\mathrm{m}}(\xi,\theta) - b(\xi)\right\|_2^2 \Pi^{a,b}(\mathrm{d}\xi)\right)$$
$$\le c\left(\frac{\log nh_n^2}{\sqrt{nh_n^2}} + h_n^{\beta/2}\right).$$

If we assume that b^{m} correctly specifies b, then

$$\sup_{(a,b)\in S}\mathbf{E}^{a,b}_x\left[\int \left\|b^{\mathrm{m}}\left(\xi,\bar{\theta}_n\right) - b(\xi)\right\|_2^2 \Pi^{a,b}(\mathrm{d}\xi)\right] \le c\left(\frac{\log nh_n^2}{\sqrt{nh_n^2}} + h_n^{\beta/2}\right).$$

2. Stochastic mirror descent with dependence and bias

Our first result is an extension of that by Duchi et al. [3], which discusses the SMD algorithm with dependent noises. Specifically, we provide convergence guarantees for the SMD algorithm based on the approximated subgradients of latent loss functions dependent on ergodic noises, which is necessary to view the convergence rate of our estimators discussed in Section 4.

We state the problem and propose the SMD algorithm with approximate subgradients. $(\Omega, \mathcal{A}, P)$ denotes the probability space. $\left(\Xi, \mathcal{B}\left(\mathbf{R}^d\right)|_\Xi\right)$ with $\Xi \in \mathcal{B}\left(\mathbf{R}^d\right)$ is the state space of a latent ergodic process $\{\xi_i; i \in \mathbf{N}\}$ with the invariant probability measure Π on $\left(\Xi, \mathcal{B}\left(\mathbf{R}^d\right)|_\Xi\right)$. We set a compact and convex set $\Theta \in \mathcal{B}\left(\mathbf{R}^p\right)$ as the parameter space.

Let $\{F(\cdot;\xi); \xi \in \Xi\}$ be a family of real-valued convex functions defined on N_Θ, where N_Θ is an open neighbourhood of Θ. We assume a convex function f such that

$$f(\theta) := \int_\Xi F(\theta;\xi)\,\Pi(\mathrm{d}\xi)$$

is finite-valued for all $\theta \in N_\Theta$. We consider the following minimization problem:

$$\min_{\theta\in\Theta} f(\theta).$$

We let $\partial F(\theta;\xi)$ denote the subdifferential of F with respect to θ and assume that there exists a $\left(\mathcal{B}\left(\mathbf{R}^d\right)|_\Xi\right) \otimes \left(\mathcal{B}\left(\mathbf{R}^p\right)|_{N_\Theta}\right)$-measurable function $\mathsf{G}(\theta;\xi)$ such that $\mathsf{G}(\theta;\xi) \in \partial F(\theta;\xi)$ for all $\theta \in \Theta$ and $\xi \in \Xi$.

A prox-function ψ is a differentiable 1-strongly convex function on N_Θ with respect to the norm $\|\cdot\|$. D_ψ is the Bregman divergence generated by ψ such that for all $\theta, \theta' \in \Theta$,

$$D_\psi\left(\theta,\theta'\right) := \psi(\theta) - \psi\left(\theta'\right) - \left\langle \nabla\psi\left(\theta'\right), \theta-\theta'\right\rangle \geq \frac{1}{2}\left\|\theta-\theta'\right\|^2.$$

We consider the SMD algorithm based on the gradients of the approximating functions $H_{i,n}(\cdot)$ for $F(\cdot;\xi_i)$. Let $\{H_{i,n}(\cdot); i = 1,\ldots,n\}$ be a sequence of real-valued random convex functions on N_Θ. Assume that there exists an $\mathcal{A}\otimes\left(\mathcal{B}\left(\mathbf{R}^p\right)|_{N_\Theta}\right)$-measurable random function $\mathsf{K}_{i,n}(\theta)$ such that $\mathsf{K}_{i,n}(\theta) \in \partial H_{i,n}(\theta)$ almost surely (a.s.) for all $\theta \in \Theta$. We define the SMD update: for arbitrary chosen $\theta_1 \in \Theta$,

$$\theta_{i+1} = \operatorname*{argmin}_{\theta\in\Theta}\left\{\left\langle \mathsf{K}_{i,n}(\theta), \theta\right\rangle + \frac{1}{\eta_i} D_\psi(\theta,\theta_i)\right\}, \tag{2.1}$$

where $\{\eta_i\}$ is a sequence of non-increasing positive numbers denoting learning rates.

Let $\mathcal{R}_{\tau,n}$ with $\tau \in \mathbf{N}_0 (:= \mathbf{N}\cup\{0\})$ be a random function of a sequence of Θ-valued random variables $\{\vartheta_i\}$ such that

$$\mathcal{R}_{\tau,n}(\{\vartheta_i\}) := \sum_{i=1}^{n-\tau}\left(F(\vartheta_i;\xi_{i+\tau}) - H_{i+\tau,n}(\vartheta_i)\right); \tag{2.2}$$

we use the abbreviation $\mathcal{R}_{\tau,n}(\theta') := \mathcal{R}_{\tau,n}(\{\theta'\})$ for non-random $\theta' \in \Theta$. $\mathcal{R}_{\tau,n}$ measures the degrees of discrepancy between $F(\cdot;\xi_i)$ and $H_{i,n}(\cdot)$.

The following decomposition for $\tau \in \mathbf{N}_0$ is useful:

$$\begin{aligned}\sum_{i=1}^{n}\left(f(\theta_i) - f\left(\theta'\right)\right) &= \sum_{i=1}^{n-\tau}\left(f(\theta_i) - f\left(\theta'\right) - F(\theta_i;\xi_{i+\tau}) + F\left(\theta';\xi_{i+\tau}\right)\right)\\ &\quad + \sum_{i=1}^{n-\tau}\left(H_{i+\tau,n}(\theta_i) - H_{i+\tau,n}(\theta_{i+\tau})\right) + \sum_{i=\tau+1}^{n}\left(H_{i,n}(\theta_i) - H_{i,n}\left(\theta'\right)\right)\\ &\quad + \sum_{i=n-\tau+1}^{n}\left(f(\theta_i) - f\left(\theta'\right)\right) + \mathcal{R}_{\tau,n}(\{\theta_i\}) - \mathcal{R}_{\tau,n}\left(\theta'\right),\end{aligned} \tag{2.3}$$

which is a trivial extension to (6.2) by Duchi et al. [3].

We define the Hellinger distance between two probability measures P and Q defined on the common measurable space such that

$$d_{\mathrm{Hel}}(P,Q) := \sqrt{\int \left(\sqrt{\frac{\mathrm{d}P}{\mathrm{d}\mu}} - \sqrt{\frac{\mathrm{d}Q}{\mathrm{d}\mu}}\right)^2 \mathrm{d}\mu}, \tag{2.4}$$

where μ is a measure such that P and Q are absolutely continuous with respect to μ. Such a μ exists; for example, P and Q are absolutely continuous with respect to $\frac{1}{2}(P+Q)$.

Let us consider that $\mathbf{F} := \{\mathcal{F}_i; i \in \mathbf{N}_0\}$ is a filtration such that $\sigma(\xi_j; j \le i) \subset \mathcal{F}_i$ for all $i \in \mathbf{N}_0$ and $\sigma(\theta_j; j \le i+1) \subset \mathcal{F}_i$ for all $i = 0, \dots, n$. Note that $\mathcal{F}_i$-measurability of θ_{i+1} is natural because θ_{i+1} depends on $\xi_1, \dots, \xi_i$ if we do not consider the approximation of $F(\cdot;\xi_i)$ with $H_{i,n}(\cdot)$. We do not determine a concrete $\mathbf{F}$ because appropriate selection depends on applications.

We define the mixing time for ξ_i with respect to the Hellinger distance based on the filtration $\mathbf{F}$: $P_{[i]|\mathbf{F}} := \left\{P^j_{[i]|\mathbf{F}}; j > i\right\}$, $i \in \mathbf{N}_0$ which denotes a family of $P^j_{[i]|\mathbf{F}}$, the conditional distribution of ξ_j given $\mathcal{F}_i$ with $j > i$, and

$$\tau\left(P_{[i]|\mathbf{F}}, \epsilon\right) := \inf\left\{\tau \in \mathbf{N}; d^2_{\mathrm{Hel}}\left(P^{i+\tau}_{[i]|\mathbf{F}}, \Pi\right) \le \epsilon^2\right\}. \tag{2.5}$$

Let us present some assumptions.

(A1) There exists a constant $G > 0$ such that for all $i \in \mathbf{N}$, $\mathcal{F}_{i\wedge n-1}$-measurable Θ-valued random variable ϑ_i,

$$\mathbf{E}\left[\left\|\mathsf{G}(\vartheta_i;\xi_i)\right\|^2_*\right] \le G^2.$$

(A2) There exists a constant $K_n > 0$ such that for all $i = 1, \dots, n$, $\mathcal{F}_{i-1}$-measurable Θ-valued random variables ϑ_i,

$$\mathbf{E}\left[\left\|\mathsf{K}_{i,n}(\vartheta_i)\right\|^2_*\right] \le K_n^2.$$

(A3) The mixing times of $\{\xi_i\}$ are uniform in the sense that there exists a uniform mixing time in expectation $\tau_{\mathbf{E}}\left(P_{|\mathbf{F}}, \epsilon\right) < \infty$ such that for all $\epsilon > 0$,

$$\tau_{\mathbf{E}}\left(P_{|\mathbf{F}}, \epsilon\right) := \inf\left\{\tau \in \mathbf{N}; \sup_{i\in\mathbf{N}_0} \mathbf{E}\left[d^2_{\mathrm{Hel}}\left(P^{i+\tau}_{[i]|\mathbf{F}}, \Pi\right)\right] \le \epsilon^2\right\}.$$

For simplicity, we ignore the dependence of $\tau_{\mathbf{E}}$ on $P_{|\mathbf{F}}$ and use the notation $\tau_{\mathbf{E}}(\epsilon)$.

Furthermore, assume that for the Bregman divergence D_ψ, $\sup_{\theta_1,\theta_2\in\Theta} D_\psi(\theta_1,\theta_2) \le R^2/2$ holds for some $R > 0$.

We obtain a version of Theorem 3.1 by Duchi et al. [3].

Theorem 2.1. *Under (A1)–(A3), for any $\epsilon > 0$ and $\theta' \in \Theta$,*

$$\begin{aligned}\mathbf{E}\left[\sum_{i=1}^n \left(f(\theta_i) - f\left(\theta'\right)\right)\right] &\le 2\sqrt{2}GRn\epsilon + \sqrt{2}\left(\tau_{\mathbf{E}}(\epsilon) - 1\right)K_n^2 \sum_{i=1}^n \eta_i + \frac{R^2}{2\eta_n} + \frac{K_n^2}{2}\sum_{i=1}^n \eta_i \\ &\quad + \left(\tau_{\mathbf{E}}(\epsilon) - 1\right)GR + \mathbf{E}\left[\mathcal{R}_{\tau_{\mathbf{E}}(\epsilon)-1,n}(\{\theta_i\}) - \mathcal{R}_{\tau_{\mathbf{E}}(\epsilon)-1,n}\left(\theta'\right)\right].\end{aligned}$$

This upper bound is the same as that in Theorem 3.1 by Duchi et al. [3] except for the residuals, which immediately disappear if $F(\cdot;\xi_i) = H_{i,n}(\cdot)$, and the constant factor $\sqrt{2}$ of the second term on the right hand side. Assumptions (A1) and (A2) on the subgradients are weaker than Assumption A in their study; therefore, this result includes a generalization of that by Duchi et al. [3] in the sense of achieving the same bound except for the constant factor with a weaker condition.

3. Simultaneous ergodicity of classes of diffusion processes

We discuss the simultaneous ergodicity of a family of diffusion processes $X_t^{a,b}(x)$, defined by the following SDE:

$$\mathrm{d}X_t^{a,b}(x) = b\left(X_t^{a,b}(x)\right)\mathrm{d}t + a\left(X_t^{a,b}(x)\right)\mathrm{d}w_t, X_0^{a,b}(x) = x, \tag{3.1}$$

where $b : \mathbf{R}^d \to \mathbf{R}^d$ and $a : \mathbf{R}^d \to \mathbf{R}^d \otimes \mathbf{R}^d$ are non-random functions, $x \in \mathbf{R}^d$ is a non-random vector, and w_t is a d-dimensional Wiener process. The transition kernel is denoted as $P_t^{a,b} : \mathbf{R}^d \times \mathcal{B}\left(\mathbf{R}^d\right) \to [0,1]$ for all $t > 0$. For simplicity, we occasionally use the notation $X_t^{a,b} = X_t^{a,b}(x)$ when no confusion can arise.

In this section, we illustrate the simultaneous ergodicity and of a family of diffusion processes. The simultaneous ergodicity of a family of diffusion processes refers to the ergodicity such that the rate of convergence is uniform in the family. They enable us to validate that the risk bounds by Theorem 2.1 hold uniformly in families with such properties.

3.1. Local Dobrushin condition.

For the local Dobrushin condition, we set the following time-homogeneous versions of the conditions in Menozzi et al. [15].

(H_α^a) There exist constants $\kappa_0 \geq 1$ and $\alpha \in (0,1]$ such that for all $x, y, \xi \in \mathbf{R}^d$

$$\kappa_0^{-1}\|\xi\|_2^2 \leq \langle a^{\otimes 2}(x)\,\xi, \xi\rangle \leq \kappa_0\|\xi\|_2^2,$$

and

$$\|a(x) - a(y)\|_F \leq \kappa_0\|x-y\|_2^\alpha.$$

(H_β^b) b is measurable, and there exist constants $\kappa_1 > 0$ and $\beta \in [0,1]$ such that for all $x, y \in \mathbf{R}^d$,

$$\|b(0)\|_2 \leq \kappa_1,\ \|b(x) - b(y)\|_2 \leq \kappa_1\left(\|x-y\|_2^\beta \vee \|x-y\|_2\right).$$

Under (H_α^a) and (H_β^b), the SDE has a unique weak solution.

Let ρ be a nonnegative smooth function with support in the unit ball of $\left(\mathbf{R}^d, \|\cdot\|_2\right)$ and $\int_{\mathbf{R}^d}\rho(x)\,\mathrm{d}x = 1$. Define $\rho_\epsilon(x) := \epsilon^{-d}\rho\left(\epsilon^{-1}x\right)$ for $\epsilon \in (0,1]$ and $b_\epsilon(x) := b * \rho_\epsilon(x) = \int_{\mathbf{R}^d} b(y)\,\rho_\epsilon(x-y)\,\mathrm{d}y$. The following then holds:

$$\left\|\|\nabla_x b_1\|_2\right\|_\infty := \sup_{x\in\mathbf{R}^d}\|\nabla_x^n b_1(x)\|_2 \leq \kappa_1 \mathrm{vol}\left(B_1(\mathbf{0})\right)\sup_{x:\|x\|_2\leq 1}\|\nabla_x^n \rho(x)\|_2 \tag{3.2}$$

[see (1.9) of 15]. Let $\varphi_t^{(\epsilon)}(x)$, $t \geq 0$ be a deterministic flow $\dot{\varphi}_t^{(\epsilon)}(x) := b_\epsilon(\varphi_t^{(\epsilon)}(x))$, $\varphi_0(x) = x$.

The following Aronson-type estimates for the transition density function of X_t hold.

Theorem 3.1 (a corollary of Theorem 1.2 by [15])**.** *Under (H_α^a) and (H_β^b), for any $T > 0$, $t \in (0,T)$ and $x \in \mathbf{R}^d$, the unique weak solution $X_t^{a,b}(x)$ admits a density $p_t^{a,b}(x,y)$, which is continuous in $x, y \in \mathbf{R}^d$. Moreover, $p_t^{a,b}$ has the following properties:*

(i) *(Two-sided density bounds) there exist constants $\lambda_0 \in (0,1]$ and $C_0 \geq 1$ depending only on $(T, \alpha, \beta, \kappa_0, \kappa_1, d)$ such that for all $t \in (0,T)$ and $x, y \in \mathbf{R}^d$,*

$$\frac{1}{C_0 t^{d/2}}\exp\left(-\frac{\left\|y - \varphi_t^{(1)}(x)\right\|_2^2}{\lambda_0 t}\right) \leq p_t^{a,b}(x,y) \leq \frac{C_0}{t^{d/2}}\exp\left(-\frac{\lambda_0\left\|y - \varphi_t^{(1)}(x)\right\|_2^2}{t}\right);$$

(ii) *(Gradient estimate in x) there exist constants $\lambda_1 \in (0,1]$ and $C_1 \geq 1$ depending only on $(T, \alpha, \beta, \kappa_0, \kappa_1, d)$ such that for all $t \in (0,T)$ and $x, y \in \mathbf{R}^d$,*

$$\left\|\nabla_x p_t^{a,b}(x,y)\right\|_2 \leq \frac{C_1}{t^{(d+1)/2}}\exp\left(-\frac{\lambda_1\left\|y - \varphi_t^{(1)}(x)\right\|_2^2}{t}\right).$$

C_j and λ_j are completely determined by $(T, \alpha, \beta, \kappa_0, \kappa_1, d)$; hence, for SDEs satisfying (H_α^a) and H_β^b for the same parameters, the density estimates are uniform across those models.

Lemma 3.2. *Under (H^b_β), the following holds:*

$$\left\|\varphi^{(1)}_t(x)-x\right\|_2 \le \kappa_1 t\left(2+\|x\|_2^\beta \vee \|x\|_2\right)\exp\left(\left\|\|\nabla_x b_1\|_2\right\|_\infty t\right).$$

We verify the local Dobrushin condition using Theorem 3.1 and Lemma 3.2. Note that we omit the explicit dependence of coefficients on ρ in the statements because we only need consider a fixed ρ.

Proposition 3.3. *For fixed $T_1, T_2 > 0$ with $T_1 < T_2$ and compact and convex $K \subset \mathbf{R}^d$, there exists a constant $\delta > 0$ dependent only on $(T_1, T_2, \alpha, \beta, \kappa_0, \kappa_1, d, K)$ such that for all $t \in (T_1, T_2)$,*

$$\sup_{x,y\in K}\left\|P^{a,b}_{2t}(x,\cdot)-P^{a,b}_{2t}(y,\cdot)\right\|_{\mathrm{TV}} \le 2-\delta.$$

3.2. **Lyapunov-type condition.** In addition to (H^b_β) and (H^a_α), we also set the following drift condition for exponential ergodicity:

(L^b_γ) There exist constants $\gamma \ge 0$ and $\varkappa_1 > 0$ such that for all $x \in \mathbf{R}^d$,

$$\langle b(x), x\rangle \le -\varkappa_1^{-1}\|x\|_2^{1+\gamma} + \varkappa_1.$$

We define the operator $\mathcal{L}^{a,b}$ such that for all $f \in \mathcal{C}^2(\mathbf{R}^d)$,

$$\mathcal{L}^{a,b} f(x) := \langle b(x), \partial_x f(x)\rangle + \frac{1}{2}\operatorname{tr}\left(a^{\otimes 2}(x)\,\partial_x^2 f(x)\right). \tag{3.3}$$

Let $\mathbf{E}^{a,b}_x$ denote the expectation with respect to the weak solution for fixed a, b, and x.

Proposition 3.4. *Under (H^a_α), (H^b_β), and (L^b_γ), for all $(\gamma,\nu) \in \mathbf{R}^2_+$ such that $\gamma = 0$ and $\nu \in \left(0, 2\varkappa_1^{-1}/\kappa_0\right)$ or arbitrary $\gamma > 0$ and $\nu > 0$, there exist positive constants $E_1, E_2 > 0$ dependent only on $(\gamma, \nu, \kappa_0, \varkappa_1, d)$ such that for any $h > 0$ and $x \in \mathbf{R}^d$,*

$$\mathbf{E}^{a,b}_x\left[V\left(X^{a,b}_h\right)\right] - V(x) \le -\left(1-e^{-E_1 h}\right)V(x) + \frac{E_2\left(1-e^{-E_1 h}\right)}{E_1},$$

where $V := \exp\left(\nu\sqrt{1+\|x\|_2^2}\right)$.

The next corollary follows immediately.

Corollary 3.5. *Under the same assumptions as Proposition 3.4, we have*

$$\sup_{t\ge 0}\mathbf{E}^{a,b}_x\left[\exp\left(\nu\sqrt{1+\left\|X^{a,b}_t\right\|_2^2}\right)\right] \le \exp\left(\nu\sqrt{1+\|x\|_2^2}\right) + \frac{E_2}{E_1}.$$

For any $m \ge 0$, we also have

$$\sup_{t\ge 0}\mathbf{E}^{a,b}_x\left[\left\|X^{a,b}_t\right\|_2^m\right] \le \frac{m!}{\nu^m}\left(\exp\left(\nu\sqrt{1+\|x\|_2^2}\right) + \frac{E_2}{E_1}\right).$$

3.3. **Harris-type theorem.** The following exponential ergodicity with uniform constants is an immediate consequence of Theorems 2.6.1 and 2.6.3 and Corollary 2.8.3 by Kulik [13].

Theorem 3.6. *Under the assumptions (H^a_α), (H^b_β), and (L^b_γ) with $\gamma \ge 0$, there exists a unique invariant probability measure $\Pi^{a,b}$ such that for all $t \ge 0$ and $x \in \mathbf{R}^d$,*

$$\left\|P^{a,b}_t(x,\cdot) - \Pi^{a,b}(\cdot)\right\|_{\mathrm{TV}} \le c_1\exp\left(-t/c_2\right)\left(V(x)+c_3\right),$$

where $V(x) := \exp\left(\nu\sqrt{1+\|x\|_2^2}\right)$, and $c_1, c_2, c_3, \nu > 0$ are positive constants dependent only on $(\alpha, \beta, \gamma, \kappa_0, \kappa_1, \varkappa_1, d)$.

Theorem 3.6 leads to the simultaneous exponential ergodicity of the d-dimensional diffusion processes defined by the SDEs satisfying (H^a_α), (H^b_β), and (L^b_γ) with the same constants $(\alpha, \beta, \gamma, \kappa_0, \kappa_1, \varkappa_1)$.

4. Estimation of stochastic differential equations

We consider the estimation of the unknown drift coefficient $b : \mathbf{R}^d \to \mathbf{R}^d$ of the following SDE based on discrete observations of X_t (estimation of the unknown diffusion coefficient is quite parallel):

$$\mathrm{d}X_t^{a,b} = b\left(X_t^{a,b}\right)\mathrm{d}t + a\left(X_t^{a,b}\right)\mathrm{d}w_t,\ X_0 = x, \tag{4.1}$$

where $x \in \mathbf{R}^d$ is a deterministic initial value and w_t is a d-dimensional Wiener process. We do not necessarily aim to estimate the optimal parameters by considering b to be included in the statistical model; rather, we consider misspecified modelling and estimate the quasi-optimal parameter [18] to know the model closest to b in the sense of the L^2-distances with respect to invariant probability measures.

We apply the discussion in Section 2 to present model-wise risk bounds for parametric estimation via online subgradient descent, which is obtained by setting $\psi(\cdot) = \|\cdot\|_2^2/2$ and $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_2$, and that in Section 3 to render those upper bounds uniform with respect to SDEs in certain classes. The notation for the classes of coefficients are as follows: let $\alpha, \gamma, \kappa_0, \kappa_1, \varkappa_1 > 0$, $\beta \geq 0$, $\varpi := (\alpha, \beta, \gamma, \kappa_0, \kappa_1, \varkappa_1)$, and S_ϖ be the class of coefficients such that

$$S_\varpi := \left\{ (a,b) \;\middle|\; \begin{array}{l} a \text{ satisfies } (H_\alpha^a) \text{ and } b \text{ satisfies } (H_\beta^b) \text{ and } (L_\gamma^b) \\ \text{with the same constants } \alpha, \beta, \gamma, \kappa_0, \kappa_1, \varkappa_1 \end{array} \right\}. \tag{4.2}$$

We finally obtain the risk bounds that uniformly hold for all $(a,b) \in S_\varpi$ with fixed ϖ by combining Theorems 2.1 and 3.6 and Corollary 3.5.

Our estimation is based on discrete observations $\{X_{ih_n}^{a,b}\}_{i=0,\dots,n}$ for a sample size of $n \in \mathbf{N}$ and the discretization step $h_n \in (0,1]$. For abbreviation, we use the notation $\Delta_i X = X_{ih_n}^{a,b} - X_{(i-1)h_n}^{a,b}$ for all $i = 1, \dots, n$. In addition, we write $\mathbf{E}_x^{a,b}$ and $X_t^{a,b}$ simply as $\mathbf{E}$ and X_t in cases wherein no confusion can arise.

4.1. Estimation with general loss functions.

Let $\Xi = \mathbf{R}^d$, $\Theta \in \mathcal{B}(\mathbf{R}^p)$ be the compact convex parameter space, N_Θ be an open neighbourhood of Θ, and $R := \sup\{\|\theta - \theta'\|_2 ; \theta, \theta' \in \Theta\}$. The loss function on Θ is defined with unknown b and a known triple (b^{m}, M, J) of functions with Borel-measurable elements : (1) the parametric model $b^{\mathrm{m}} : \mathbf{R}^d \times N_\Theta \to \mathbf{R}^d$; (2) the weight function $M : \mathbf{R}^d \to \mathbf{R}^d \otimes \mathbf{R}^d$, which is positive semi-definite for all $\xi \in \mathbf{R}^d$; and (3) the regularization term $J : N_\Theta \to \mathbf{R}$.

We define a function $F : N_\Theta \times \mathbf{R}^d \to \mathbf{R}$ such that

$$F(\theta;\xi) = F^b(\theta;\xi) := \frac{1}{2} M(\xi)\left[\left(b^{\mathrm{m}}(\xi,\theta) - b(\xi)\right)^{\otimes 2}\right] + J(\theta). \tag{4.3}$$

We consider the minimization problem of the following loss function on Θ:

$$f^{a,b}(\theta) := \int F^b(\theta;\xi)\,\Pi^{a,b}(\mathrm{d}\xi), \tag{4.4}$$

where $\Pi^{a,b}$ is the invariant probability measure of $X_t^{a,b}$.

Clearly, $F(\theta; x)$, which depends on the unknown coefficient b, is unknown. Hence, we consider the approximated loss functions based on discrete observations and observe the performance of the estimator given by online gradient descents.

The sampled loss functions are given by the h_n-skeleton of $X_t^{a,b}$:

$$F(\theta;\xi_i) = \frac{1}{2} M(\xi_i)\left[\left(b^{\mathrm{m}}(\xi_i,\theta) - b(\xi_i)\right)^{\otimes 2}\right] + J(\theta), \tag{4.5}$$

where $\xi_i = X_{(i-1)h_n}^{a,b}$. $H_{i,n}(\theta)$, a random function on Θ, should be sufficiently close to F; hence, we set

$$\begin{aligned} H_{i,n}(\theta) := \frac{1}{2h_n^2} M\left(X_{(i-1)h_n}\right)\left[\left(\Delta_i X - h_n b^{\mathrm{m}}\left(X_{(i-1)h_n},\theta\right)\right)^{\otimes 2} - \left(\Delta_i X - h_n b\left(X_{(i-1)h_n}\right)\right)^{\otimes 2}\right] \\ + J(\theta). \end{aligned} \tag{4.6}$$

A simple computation leads to the equality

$$\begin{aligned}&H_{i,n}(\theta)-F(\theta;\xi_i)\\&=\frac{1}{h_n}M\left(X_{(i-1)h_n}\right)\left[b\left(X_{(i-1)h_n}\right)-b^{\mathrm{m}}\left(X_{(i-1)h_n},\theta\right),\Delta_iX-h_nb\left(X_{(i-1)h_n}\right)\right].\end{aligned}$$

Let $\mathbf{F}=\{\mathcal{F}_i;i\in\mathbf{N}_0\}$, where $\mathcal{F}_i=\sigma(X_t;t\le ih_n)$. We set the following assumptions on F, $H_{i,n}$, and (b^{m},M,J).

(D1) For all $\xi,\xi'\in\mathbf{R}^d$,

$$\frac{1}{2}M(\xi)\left[\left(b^{\mathrm{m}}(\xi,\theta)-\xi'\right)^{\otimes2}\right]+J(\theta)$$

is convex with respect to $\theta\in N_\Theta$. Moreover, there exist positive constants $G>0$ and $\check{K}>0$, a $\mathcal{B}(\mathbf{R}^d)\otimes(\mathcal{B}(\mathbf{R}^p)|_{N_\Theta})$-measurable function G, and an $\mathcal{F}_i\otimes(\mathcal{B}(\mathbf{R}^p)|_{N_\Theta})$-measurable random function K such that $\mathsf{G}(\theta;\xi)\in\partial F(\theta;\xi)$ and $\mathsf{K}_{i,n}(\theta)\in\partial H_{i,n}(\theta)$ a.s. for all $\xi\in\mathbf{R}^d$, $\theta\in\Theta$, and $i=1,\dots,n$, and

$$\sup_{(a,b)\in S_\varpi}\sup_{i\in\mathbf{N}}\mathbf{E}_x^{a,b}\left[\|\mathsf{G}(\vartheta_i;\xi_i)\|_2^2\right]\le G,\quad\sup_{(a,b)\in S_\varpi}\sup_{i=1,\dots,n}\mathbf{E}_x^{a,b}\left[\|\mathsf{K}_{i,n}(\vartheta_i)\|_2^2\right]\le\frac{\check{K}}{h_n}.$$

for all $n\in\mathbf{N}$ and sequence of $\mathcal{F}_{i\wedge n-1}$-measurable Θ-valued random variables ϑ_i.

(D2) There exists a constant $\zeta>0$ such that for all $x\in\mathbf{R}^d$ and $\theta\in N_\Theta$,

$$\|b^{\mathrm{m}}(x,\theta)\|_2\le\zeta\left(1+\|x\|_2^\zeta\right),\ \|M(x)\|_2\le\zeta.$$

The following proposition provides the bound for the residual terms $\mathcal{R}_{\tau,n}$.

Proposition 4.1. *Assume that (D2) holds. There exists a constant $c>0$ dependent only on $(\beta,\kappa_0,\kappa_1,\zeta,d)$ such that for any sequence $\{\vartheta_i;i=1,\dots,n\}$ of $\mathcal{F}_{i-1}$-measurable Θ-valued random variables ϑ_i, $\tau\in\mathbf{N}_0$, and $n\in\mathbf{N}$,*

$$\left|\mathbf{E}_x^{a,b}\left[\sum_{i=1}^{n-\tau}\left(F(\vartheta_i;\xi_{i+\tau})-H_{i+\tau,n}(\vartheta_i)\right)\right]\right|\le cnh_n^{\beta/2}\left(1+\sup_{t\ge0}\mathbf{E}_x^{a,b}\left[\left\|X_t^{a,b}\right\|_2^c\right]\right).$$

We obtain our main result on the drift estimation using learning rates whose optimality is attributable to Duchi et al. [3]; we ignore the influence of G, $\check{K}$, and R.

Theorem 4.2. *Assume that $h_n\in(0,1]$, $\log nh_n^2\ge1$, and (D1)–(D2) hold. Under the update rule (2.1) with $\eta_i:=\eta h_n/\sqrt{i}$ and fixed $\eta>0$, for any $x\in\mathbf{R}^d$, there exists a positive constant $c>0$ dependent only on $\left(\varpi,\zeta,\eta,G,\check{K},R,d,x\right)$ such that*

$$\sup_{(a,b)\in S_\varpi}\sup_{\theta\in\Theta}\mathbf{E}_x^{a,b}\left[\sum_{i=1}^n\left(f^{a,b}(\theta_i)-f^{a,b}(\theta)\right)\right]\le c\left(\sqrt{\frac{n}{h_n^2}}\log nh_n^2+nh_n^{\beta/2}\right).$$

When we assume a usual identifiability condition of the quasi-optimal parameter, that is, the optimal point $\theta_0^{a,b}$ of $f^{a,b}$, Theorem 4.2 yields the rate of convergence.

Corollary 4.3. *Assume that the same assumptions as in Theorem 4.2 hold, the update rule (2.1) holds with $\eta_i:=\eta h_n/\sqrt{i}$ and $\eta>0$, and for all $(a,b)\in S_\varpi$, there exist $\chi^{a,b}>0$ and $\theta_0^{a,b}\in\Theta$ such that*

$$\frac{\chi^{a,b}}{2}\left\|\theta-\theta_0^{a,b}\right\|_2^2\le f^{a,b}(\theta)-f^{a,b}\left(\theta_0^{a,b}\right).$$

(i) *There exists a positive constant $c>0$ dependent only on $\left(\varpi,\zeta,\eta,G,\check{K},R,d,x\right)$ such that*

$$\sup_{(a,b)\in S_\varpi}\mathbf{E}_x^{a,b}\left[\frac{\chi^{a,b}}{2}\left\|\bar{\theta}_n-\theta_0^{a,b}\right\|_2^2\right]\le c\left(\frac{\log nh_n^2}{\sqrt{nh_n^2}}+h_n^{\beta/2}\right).$$

(ii) *If $nh_n^2\to\infty$ and $\sup_{n\in\mathbf{N}}nh_n^{2+\beta}<\infty$, then $\bar{\theta}_n-\theta_0^{a,b}=\mathcal{O}_P\left(\sqrt[4]{\frac{nh_n^2}{(\log nh_n^2)^2}}\right)$.*

(iii) *If* $nh_n^2 \to \infty$ *and* $\sup_{n\in\mathbf{N}} nh_n^{2+\beta/4\rho} < \infty$ *for some* $\rho \in (0, 1/4)$, *then* $\bar{\theta}_n - \theta_0^{a,b} = \mathcal{O}_P\left(\left(nh_n^2\right)^{\rho}\right)$.

4.2. **Estimation with least-square loss functions.** We now consider the least-square-type loss functions [e.g., see 14] for SDEs with drift coefficients linear in the parameters. The target loss function is

$$f^{a,b}(\theta) = \int \frac{1}{2} \left\| b^{\mathrm{m}}(\xi, \theta) - b(\xi) \right\|_2^2 \Pi^{a,b}(\mathrm{d}\xi). \tag{4.7}$$

It corresponds to the case $M(x) = I_d$ and $J(\theta) = 0$ for all $x \in \mathbf{R}^d$ and $\theta \in N_\Theta$. Hence, $H_{i,n}(\theta)$ is given as

$$H_{i,n}(\theta) := \frac{1}{2h_n^2} \left\| \Delta_i X - h_n b^{\mathrm{m}}\left(X_{(i-1)h_n}, \theta\right) \right\|_2^2 - \frac{1}{2h_n^2} \left\| \Delta_i X - h_n b\left(X_{(i-1)h_n}\right) \right\|_2^2. \tag{4.8}$$

We set the following assumption:

(D2') $b^{\mathrm{m}}(x, \theta)$ is in $\mathcal{C}^1\left(\mathbf{R}^d \times N_\Theta\right)$ and each component is linear in $\theta \in N_\Theta$ for all $x \in \mathbf{R}^d$, and there exists a constant $\zeta > 0$ such that for all $x \in \mathbf{R}^d$ and $\theta \in N_\Theta$,

$$\left\| b^{\mathrm{m}}(x, \theta) \right\|_2 \le \zeta\left(1 + \|x\|_2^{\zeta}\right), \ \left\| \partial_\theta b^{\mathrm{m}}(x, \theta) \right\|_F \le \zeta\left(1 + \|x\|_2^{\zeta}\right).$$

Under (D2'), $H_{i,n}(\theta)$ is a.s. convex in θ and its gradient is given as

$$\mathsf{K} := -\frac{1}{h_n} \left(\partial_\theta b^{\mathrm{m}}\right)\left(X_{(i-1)h_n}, \theta\right)\left(\Delta_i X - h_n b^{\mathrm{m}}\left(X_{(i-1)h_n}, \theta\right)\right). \tag{4.9}$$

By choosing $\eta_i = h_n/\sqrt{i}$, we obtain a simple update rule of the online gradient descent:

$$\theta_{i+1} := \mathrm{Proj}_\Theta\left(\theta_i + \frac{1}{\sqrt{i}} \left(\partial_\theta b^{\mathrm{m}}\right)^\top \left(X_{(i-1)h_n}, \theta_i\right)\left(\Delta_i X - h_n b^{\mathrm{m}}\left(X_{(i-1)h_n}, \theta_i\right)\right)\right).$$

Lemma 4.4. *Assume that (D2') holds. There exists a constant* $c > 0$ *dependent only on* (ϖ, ζ, d, x) *such that for all* $i = 1, \ldots, n$, $n \in \mathbf{N}$, $\mathcal{F}_{i-1}$*-measurable* Θ*-valued random variables* ϑ_i,

$$\sup_{(a,b)\in S_\varpi} \mathbf{E}_x^{a,b}\left[\left\| \frac{1}{h_n}\left(\partial_\theta b_{i-1}^{\mathrm{m}}(\vartheta_i)\right)^\top \left(\Delta_i X - h_n b_{i-1}^{\mathrm{m}}(\vartheta_i)\right)\right\|_2^2\right] \le \frac{c}{h_n}.$$

The existence of G dependent only on (ϖ, ζ, d, x) in (D1) is more obvious. Hence, we obtain the following simple but useful corollary:

Corollary 4.5. *Under* $h_n \in (0, 1]$, $\log nh_n^2 \ge 1$, *the update rule* (2.1) *with* $\eta_i = \eta h_n/\sqrt{i}$ *and* $\eta > 0$, *(D1) and (D2'),* $\bar{\theta}_n := \frac{1}{n}\sum_{i=1}^n \theta_i$ *has the uniform risk bound such that*

$$\sup_{(a,b)\in S_\varpi} \sup_{\theta\in\Theta} \mathbf{E}_x^{a,b}\left[f^{a,b}\left(\bar{\theta}_n\right) - f^{a,b}(\theta)\right] \le c\left(\frac{\log nh_n^2}{\sqrt{nh_n^2}} + h_n^{\beta/2}\right),$$

where $c > 0$ *is a constant dependent only on* $(\varpi, \zeta, \eta, R, d, x)$.

References

[1] Bhudisaksang, T. and Cartea, A. (2021). Online drift estimation for jump-diffusion processes. *Bernoulli*, 27(4):2494–2518.

[2] Bibby, B. M. and Sørensen, M. (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, 1:17–39.

[3] Duchi, J. C., Agarwal, A., Johansson, M., and Jordan, M. L. (2012). Ergodic mirror descent. *SIAM Journal of Optimization*, 22(4):1549–1578.

[4] Florens-Zmirou, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20:547–557.

[5] Genon-Catalot, V. and Jacod, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 29(1):119–151.

[6] Gobet, E., Hoffmann, M., and Reiß, M. (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *The Annals of Statistics*, 32(5):2223–2253.
[7] Hoffmann, M. (1999). Adaptive estimation in diffusion processes. *Stochastic Processes and their Applications*, 79:135–163.
[8] Kaino, Y. and Uchida, M. (2018a). Hybrid estimators for stochastic differential equations from reduced data. *Statistical Inference for Stochastic Processes*, 21:435–454.
[9] Kaino, Y. and Uchida, M. (2018b). Hybrid estimators for small diffusion processes based on reduced data. *Metrika*, 81:745–773.
[10] Kaino, Y., Uchida, M., and Yoshida, Y. (2017). Hybrid estimation for an ergodic diffusion process based on reduced data. *Bulletin of Informatics and Cybernetics*, 49:89–118.
[11] Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics*, 24(2):211–229.
[12] Kessler, M. and Sørensen, M. (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, 5:299–314.
[13] Kulik, A. (2017). *Ergodic Behavior of Markov Processes*. De Gruyter, Berlin, Boston.
[14] Masuda, H. (2005). Simple estimators for parametric Markovian trend of ergodic processes based on sampled data. *Journal of the Japanese Statistical Society*, 35(2):147–170.
[15] Menozzi, S., Pesce, A., and Zhang, X. (2021). Density and gradient estimates for non degenerate Brownian SDEs with unbounded measurable drift. *Journal of Differential Equations*, 272:330–369.
[16] Sharrock, L. and Kantas, N. (2022). Joint online parameter estimation and optimal sensor placement for the partially observed stochastic advection-diffusion equation. *SIAM/ASA Journal on Uncertainty Quantification*, 10(110):55–95.
[17] Surace, S. C. and Pfister, J.-P. (2019). Online maximum-likelihood estimation of the parameters of partially observed diffusion processes. *IEEE Transactions on Automatic Control*, 64(7):2814–2829.
[18] Uchida, M. and Yoshida, N. (2011). Estimation for misspecified ergodic diffusion processes from discrete observations. *ESAIM: Probability and Statistics*, 15:270–290.
[19] Uchida, M. and Yoshida, N. (2012). Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Processes and their Applications*, 122(8):2885–2924.
[20] Yoshida, N. (1992). Estimation for diffusion processes from discrete observation. *Journal of Multivariate Analysis*, 41(2):220–242.