Journal of Machine Learning Research 24 (2023) 1-38 Submitted 6/22; Revised 3/23; Published 4/23

Concentration analysis of multivariate elliptic diffusions

Lukas Trottner TROTTNER@MATH.AU.DK
Department of Mathematics
Aarhus University
Aarhus, Denmark

Cathrine Aeckerle-Willems AECKERLE@UNI-MANNHEIM.DE
Department of Economics
University of Mannheim
Mannheim, Germany

Claudia Strauch STRAUCH@MATH.AU.DK
Department of Mathematics
Aarhus University
Aarhus, Denmark

Editor: Gabor Lugosi

Abstract

We prove concentration inequalities and associated PAC bounds for both continuous- and discrete-time additive functionals for possibly unbounded functions of multivariate, nonreversible diffusion processes. Our analysis relies on an approach via the Poisson equation allowing us to consider a very broad class of subexponentially ergodic, multivariate diffusion processes. These results add to existing concentration inequalities for additive functionals of diffusion processes which have so far been only available for either bounded functions or for unbounded functions of processes from a significantly smaller class. We demonstrate the power of these exponential inequalities by two examples of very different areas. Considering a possibly high-dimensional, parametric, nonlinear drift model under sparsity constraints we apply the continuous-time concentration results to validate the restricted eigenvalue condition for Lasso estimation, which is fundamental for the derivation of oracle inequalities. The results for discrete additive functionals are applied for an investigation of the unadjusted Langevin MCMC algorithm for sampling of moderately heavy tailed densities π. In particular, we provide PAC bounds for the sample Monte Carlo estimator of integrals $\pi(f)$ for polynomially growing functions f that quantify sufficient sample and step sizes for approximation within a prescribed margin with high probability.

Keywords: Concentration inequality, elliptic diffusions, Lasso estimation, MCMC, PAC bounds

1. Introduction

Concentration inequalities for additive functionals belong to the fundamental probabilistic tools in statistics and related areas such as statistical learning and reinforcement learning since they allow exact quantification of the deviation of estimators from a given target. In particular, concentration inequalities for independent data such as Hoeffding, Bernstein and McDiarmid inequalities are of central importance for deriving PAC guarantees in classifi-

cation and regression contexts (see, e.g., Devroye et al. (1996); Wainwright (2019)). While such questions have been well understood for decades in classical settings for independent or strongly mixing data—see also the recent investigations of Bernstein and Hoeffding inequalities and related applications in statistical learning for Markov chains with spectral gap in Jiang et al. (2018); Fan et al. (2021)—the general picture for additive functionals of diffusion processes is less clear. Particularly when it comes to unbounded functionals, whose deviation properties around their ergodic mean are fundamentally important in a multitude of applications, useful results are rather scarce. Important achievements in this direction can be found in Cattiaux and Guillin (2008); Gao et al. (2014), where for a restricted class of reversible diffusion processes exponential inequalities are derived by means of functional inequalities. While these results are mathematically elegant and explicitly quantify the contribution of the asymptotic variance, they come at the price of structural constraints on the diffusion coefficients which are hard to verify and often inappropriate for specific applications.

The goal of this paper is therefore to derive usable exponential concentration inequalities for unbounded functionals, both for continuous as well as discrete multivariate diffusion data, under comparatively weak assumptions on the coefficients and the speed of ergodicity. With our particular focus on applications, we translate these inequalities into PAC bounds for the approximation task and demonstrate their usefulness in specific high-dimensional applications to (i) penalized drift estimation under sparsity constraints, where we extend results for the classical Ornstein–Uhlenbeck model in Gaïffas and Matulewicz (2019); Ciołek et al. (2020) to more flexible parametrized models with relaxed ergodicity assumptions, and (ii) performance guarantees for unadjusted Langevin MCMC algorithms for heavy-tailed target sampling, which is a setting that substantially differs from the related pioneering work Dalalyan (2017); Durmus and Moulines (2017) for strongly log-concave targets. Here, for a given quantity of interest π and a sample based estimator $\widehat{\pi}_t$ with $t \in \mathbb{T}$ —where $\mathbb{T} = [0, \infty)$ or $\mathbb{T} = \Delta\mathbb{N}_0$ for some sampling distance $\Delta > 0$, depending on whether continuous or discrete data is available—, we say that $\widehat{\pi}_t$ satisfies an (ε, δ)-PAC bound for $t \geq T(\varepsilon, \delta) \in \mathbb{T}$, given $\varepsilon > 0, \delta \in (0, 1)$, if

$$\forall t \geq T(\varepsilon, \delta), \quad \mathbb{P}\big(|\widehat{\pi}_t - \pi| \leq \varepsilon\big) \geq 1 - \delta,$$

i.e., given a sample length of at least $T(\varepsilon, \delta)$, $\widehat{\pi}_t$ approximates the target π within an ε-margin with probability at least $1 - \delta$. Such results are statistically much more insightful than upper bounds on the mean deviation, which do not reveal detailed information on the distribution of the loss.

In our particular context, the objectives are exponential inequalities and associated PAC bounds of sample mean estimators of the quantity $\pi = \mu(f) \coloneqq \int f(x)\, \mu(\mathrm{d}x)$, where μ is the stationary distribution of a subexponentially ergodic elliptic diffusion $\mathbf{X}$ and f is a polynomially growing function. That is, we provide an in-depth analysis of the deviations around π of $\widehat{\pi}_t = t^{-1/2}\mathbb{G}_t(f)$, where

$$\mathbb{G}_t(f) \coloneqq \frac{1}{\sqrt{t}} \int_0^t f(X_s)\, \mathrm{d}s, \tag{1}$$

given continuous data $(X_s)_{0\le s\le t}$, and of $\widehat{\pi}_{n\Delta} = (n\Delta)^{-1/2}\mathbb{G}_{n,\Delta}(f)$, where

$$\mathbb{G}_{n,\Delta}(f) \coloneqq \frac{1}{\sqrt{n\Delta}}\sum_{k=1}^{n} f(X_{k\Delta})\Delta, \tag{2}$$

given discrete data $(X_{k\Delta})_{k=1,\dots,n}$, as well as their burned-in versions. Since our specific framework is what sets this paper apart from related studies such as Gao et al. (2014), we will now introduce both the class of processes we are working with as well as the *Poisson equation* and its solution studied in Pardoux and Veretennikov (2001), which is at the heart of our theoretical analysis based on *martingale approximation*.

Basic framework Consider a d-dimensional elliptic diffusion that is given as the weak solution to the SDE

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \tag{3}$$

where $b\colon \mathbb{R}^d \to \mathbb{R}^d$ is a locally Lipschitz drift vector such that $\|b(x)\| \lesssim 1 + \|x\|^{q'}$ for some $q' \ge 0$ and $\sigma\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a uniformly continuous, bounded and locally Lipschitz $d\times d$-matrix-valued function such that $a \coloneqq \sigma\sigma^\top$ is uniformly elliptic, i.e.,

$$\langle a(x)\eta/\|\eta\|, \eta/\|\eta\|\rangle \ge \lambda, \quad x\in\mathbb{R}^d, \eta\in\mathbb{R}^d\setminus\{0\},$$

for some constant $\lambda > 0$. We denote by $(\mathbf{X}, (\mathbb{P}^x)_{x\in\mathbb{R}^d})$ the Markovian weak solution of (3) such that under $\mathbb{P}^x$ the process $\mathbf{X}$ solves (3) with initial condition $X_0 = x$ and has continuous paths almost surely. Note that $\mathbf{X}$ has the Feller property, cf. (Stroock and Varadhan, 2006, Corollary 11.1.5), and is therefore Borel right such that it falls into the general framework for stability of Markov processes. Without loss of generality, we may assume that there exists a family of shift operators $(\theta_t)_{t\ge0}$ for $\mathbf{X}$, that is, $X_t\circ\theta_s = X_{t+s}$ for any $s,t\ge0$. Let $\lambda_-, \lambda_+, \Lambda$ be the tightest constants such that, for any $x\neq 0$,

$$0 < \lambda_- \le \langle a(x)x/\|x\|, x/\|x\|\rangle \le \lambda_+, \quad \operatorname{tr}(a(x))/d \le \Lambda,$$

where our assumptions guarantee that such constants always exist since we may always choose $\Lambda = d^{-1}\sup_{x\in\mathbb{R}^d}\operatorname{tr}(a(x)) < \infty$, $\lambda_- = \lambda$ and $\lambda_+ = \sup_{x\in\mathbb{R}^d}\|\sigma(x)\|^2 < \infty$.

Our subsequent analysis substantially relies on the following growth condition on the drift,

$(\mathscr{A}(q))$ if $\|x\| \ge M_0$, then $\langle b(x), x/\|x\|\rangle \le -\mathfrak{r}\|x\|^{-q}$,

where $q\in[-1,1), M_0\ge 0, \mathfrak{r} > 0$. For $q = 0$, this condition equals the standard ergodicity condition in many recent investigations of multivariate diffusion processes exploiting the exponential β-mixing property. As will be discussed in Section 2, the case $q > 0$ corresponds to a subexponential ergodic behaviour of the diffusion.

Our approach to deviation inequalities is driven by the martingale approximation technique, which has been employed for the same purpose in the literature under more restrictive structural assumptions. Aeckerle-Willems and Strauch (2021) study concentration inequalities in the context of scalar exponentially ergodic diffusions in the regime $q = 0$ with polynomially growing drift, and in Nickl and Ray (2020), multivariate diffusions with unit

diffusion matrix and periodic Lipschitz drift are considered. Galtchouk and Pergamenshchikov (2007) essentially treat the scalar dissipative case with $q = -1$. All of these papers put a special emphasis on uniformicity of the concentration inequalities with respect to the diffusion coefficients in order to apply them to statistical minimax estimation problems. Moreover, the martingale approximation is employed in Mattingly et al. (2010) for providing L^2 convergence guarantees of Monte Carlo estimators for well-behaved SDEs on the torus based on samples obtained by numerical approximation schemes.

Central to the martingale approximation technique is the existence of a solution to the Poisson equation

$$Lu = f, \tag{4}$$

for appropriate functions f where, given $u \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$ having weak partial derivatives up to second order belonging to $L^1_{\mathrm{loc}}(\mathbb{R}^d)$,

$$Lu(x) = \langle b(x), \nabla u(x)\rangle + \frac{1}{2}\mathrm{tr}\big(a(x)D^2u(x)\big), \quad x \in \mathbb{R}^d,$$

is a second order local operator. Note that, on the domain $\mathcal{C}^2_0(\mathbb{R}^d)$, L is the infinitesimal generator of the diffusion process. In the scalar case, (4) has an explicit $\mathcal{C}^2$-solution, which is used in Aeckerle-Willems and Strauch (2021) to obtain sup-norm moment bounds for empirical processes that are uniform over a class of SDE coefficients. Such results can then be employed for minimax optimal sup-norm adaptive drift estimation as demonstrated in Aeckerle-Willems and Strauch (2022).

For multivariate diffusions, such explicit solutions are not obtainable in general such that one needs to deal with the Poisson equation in a more abstract manner. In Pardoux and Veretennikov (2001), the authors demonstrate that in our framework, for any $f\colon \mathbb{R}^d \to \mathbb{R}$ such that $|f(x)| \leq \mathfrak{L}(1+\|x\|^\eta)$ for some finite constants $\mathfrak{L} > 0, \eta \geq 0$, there exists a solution $u[f] \in \bigcap_{p>1} \mathcal{W}^{2,p}_{\mathrm{loc}}(\mathbb{R}^d)$ that is unique in the local Sobolev space $\mathcal{W}^{2,p}_{\mathrm{loc}}(\mathbb{R}^d)$ for any $p > d$. This solution is given as

$$u[f](x) = \int_0^\infty \mathbb{E}^x[-f(X_t)]\,\mathrm{d}t, \quad x \in \mathbb{R}^d,$$

i.e., $u[f](x)$ is expressed as the potential of $-f$ under $\mathbb{P}^x$. Therefore, for such f we denote

$$L^{-1}[f](x) \coloneqq \int_0^\infty \mathbb{E}^x[-f(X_t)]\,\mathrm{d}t, \quad x \in \mathbb{R}^d,$$

such that $LL^{-1}[f] = f$, $\boldsymbol{\lambda}$-a.e., where $\boldsymbol{\lambda}$ denotes the Lebesgue measure on $\mathbb{R}^d$. The Sobolev regularity of $L^{-1}[f]$ is an essential property for our purposes, since it allows us to apply the Itō–Krylov formula for martingale approximation. This approach will enable us to conclude the desired deviation inequalities from moment bounds for the martingale approximation.

Outline and main results In Section 2, we collect some essential known facts on the subexponentially ergodic nature of the diffusion $\mathbf{X}$ implied by the drift condition ($\mathscr{A}(q)$) and put them into a form suited to our needs. In Section 3.1, we present our first main result, the concentration inequality for the continuous-time scaled additive functional $\mathbb{G}_t(f)$ for polynomially growing f (Theorem 3) which is based on our derivation of the martingale

approximation $\mathbb{G}_t(f)$ and bounds on the solution to the Poisson equation and its gradient going back to Pardoux and Veretennikov (2001). We translate these inequalities into stationary and non-stationary PAC bounds in Corollary 8 and 9, respectively. In Section 3.2, we then proceed to derive explicit deviation bounds in terms of the sampling frequency Δ and number of observations n for the discrete scaled additive functional $\mathbb{G}_{n,\Delta}(f)$ by combining Theorem 3 with an approximation argument, see Theorem 11. As for the continuous data, we use this result to infer PAC bounds for the sample mean estimator and its burn-in version. In Section 4.1, we apply the continuous-time results to the problem of estimating the coefficients in a possibly high-dimensional drift model of the form $b_{\theta_0} = \sum_{j=1}^N \theta_j \psi_j$, $\theta_0 = (\theta_1, ..., \theta_N) \in \mathbb{R}^N$, given a dictionary $(\psi_j)_{1\le j\le N}$ of Lipschitz continuous functions $\psi_j \colon \mathbb{R}^d \to \mathbb{R}^d$ under sparsity constraints on the coefficients via a Lasso approach. Our concentration inequality is the key to showing that the central restricted eigenvalue condition is in place, which then in turn yields oracle inequalities in line with those well known in the classical regression context. Finally, Section 4.2 is devoted to an application of our discrete deviation results, where we study the convergence properties of the unadjusted Langevin algorithm for moderately heavy-tailed target distributions π, in terms of sufficient sample and step size conditions for sampling within an ε-margin in total variation as well as for ensuring an (ε, δ)-PAC bound of the sample Monte Carlo estimator of a given target integral $\pi(f)$, again for polynomially bounded functions f.

2. Subexponential ergodicity of the diffusion

We now give an exact quantification of the stability of $\mathbf{X}$, which underlies the arguments from Pardoux and Veretennikov (2001) and also plays the central technical role in our approach. For details on terms from Markov stability theory such as petite sets or Harris recurrence, we refer to Douc et al. (2009).

Define $q_+ = q \vee 0$. For $q \in (0,1)$, choose $\iota = \iota(q) > 0$ small enough such that $\mathfrak{r} > \iota\lambda_+(1-q)/2$. In this framework, it was shown in (Douc et al., 2009, Proposition 5.1, Theorem 5.4) as a refinement of results in Malyshkin (2000) that $\mathbf{X}$ possesses a unique invariant distribution μ and that there exists some constant $C(q)$ such that, for $V_q(x) \coloneqq \exp(\iota\|x\|^{1-q})$, we have

$$\left\|\mathbb{P}^x(X_t \in \cdot) - \mu\right\|_{\mathrm{TV}} \le C(q) V_q(x) (1+t)^{\frac{2q}{1+q}} \mathrm{e}^{-(\iota' t)^{(1-q)/(1+q)}}, \quad x \in \mathbb{R}^d, t \ge 0, \tag{5}$$

with $\iota' = \iota'(q) \coloneqq \iota^{(1+q)/(1-q)}(1+q)(\mathfrak{r} - \lambda_+\iota(1-q)/2)$ and $\|\nu\|_{\mathrm{TV}} \coloneqq \sup_{\|f\|_\infty \le 1} |\nu(f)|$ for a signed finite measure ν. Thus, for $q \in (0,1)$, $\mathbf{X}$ is subexponentially ergodic. In case $q \in [-1, 0]$, $\mathbf{X}$ is exponentially ergodic, i.e., for some constants $C(0)$, ι and ι' (not explicitly related to the constants $C(q), \iota(q), \iota'(q)$ from above),

$$\left\|\mathbb{P}^x(X_t \in \cdot) - \mu\right\|_{\mathrm{TV}} \le C(0) \mathrm{e}^{\iota\|x\|} \mathrm{e}^{-\iota' t}, \quad x \in \mathbb{R}^d, t \ge 0, \tag{6}$$

cf. (Pardoux and Veretennikov, 2001, Proposition 1). Moreover, (Douc et al., 2009, Theorem 5.3) and (Pardoux and Veretennikov, 2001, Proposition 1) establish that $V_{q+} \in L^1(\mu)$ (here, $V_0(\cdot) = \exp(\iota\|\cdot\|)$), i.e.,

$$\mathbb{E}^\mu[V_{q_+}(X_0)] = \int_{\mathbb{R}^d} \exp(\iota\|x\|^{1-q_+})\, \mu(\mathrm{d}x) < \infty. \tag{7}$$

It will be central for us to trade off subexponential ergodicity at a slower temporal rate with a less punishing penalty function. Let $\widetilde{V}_\alpha(x) \coloneqq 1 + \|x\|^\alpha$ for $\alpha \geq 0$. Then, for $\zeta > 0$ and $\gamma > 2(1+\zeta)$, Proposition 1 in Pardoux and Veretennikov (2001) demonstrates that

$$\left\|\mathbb{P}^x(X_t \in \cdot) - \mu\right\|_{\mathrm{TV}} \leq C(\gamma,\zeta)\widetilde{V}_\gamma(x)(1+t)^{-(1+\zeta)}, \quad x \in \mathbb{R}^d, t \geq 0,$$

i.e., polynomial convergence with polynomial penalty function whose degree depends on the degree of the temporal rate that can be freely chosen. We will need to make use of polynomial convergence with respect to a stronger norm than the total variation norm considered above. To this end, let H_1, H_2 be a pair of Young functions on $\mathbb{R}_+$, which are in particular invertible and satisfy

$$xy \leq H_1(x) + H_2(y), \quad x, y \geq 0, \tag{8}$$

and let $\mathcal{I}$ be the family of pairs of inverse Young functions augmented by $(\mathbf{1}, \mathrm{Id})$ and $(\mathrm{Id}, \mathbf{1})$. The prototypical example for such pairs are $H_1(x) = x^p/p, H_2(x) = x^q/q$ with p, q conjugate Hölder exponents such that $1/p + 1/q = 1$, in which case (8) is simply Young's inequality. More generally, one may pair any convex function with its Legendre transform to obtain (8).

Following earlier work on discrete and continuous-time Markov models Douc et al. (2004, 2008); Jarner and Roberts (2002); Tuominen and Tweedie (1994); Fort and Roberts (2005), such inverse Young functions are used in Douc et al. (2009) for subgeometrically ergodic Markov models to quantify the trade-off between speed of convergence and strength of the underlying *f-norm*, which we introduce next. For a measurable function $f \geq 1$ and a signed measure ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, its f-norm is defined by

$$\|\nu\|_f \coloneqq \sup_{|g| \leq f} |\nu(g)|.$$

In particular, $\|\cdot\|_{\mathrm{TV}} = \|\cdot\|_{\mathbf{1}}$. Let us also define the δ-delayed first hitting time of a set $B \in \mathcal{B}(\mathbb{R}^d)$ by $\tau_B(\delta) = \inf\{t \geq \delta : X_t \in B\}$, for $\delta \geq 0$, with $\tau_B = \tau_B(0)$. Moreover, we say that B with $\mu(B) > 0$ is accessible, since μ is a maximal irreducibility measure of $\mathbf{X}$. Also note that μ as the invariant distribution of a Feller process is maximal Harris, such that in particular for any accessible set B we have $\mathbb{P}^x(\tau_B(\delta) < \infty) = 1$ for any $x \in \mathbb{R}^d$, $\delta \geq 0$.

With the techniques from Douc et al. (2009), we obtain the following result on polynomial f-norm convergence and modulated moments, whose proof is given in Appendix A. This explicit ergodicity result is central both for our derivation of the concentration inequality for continuous data and for its subsequent discrete extension. It will turn out that appropriate choices for the pairing of Young functions to optimize the trade-off between convergence rate and strength of the f-norm will be essential when dealing with polynomially bounded test functions. We therefore truly need the full generality of the statement, which underlines the power of the approach in Douc et al. (2009) for concrete applications.

Proposition 1 *Let $\gamma \geq 1 + q$ and $q \in (-1, 1)$. Then, there exist functions $r_{\gamma,q}(t) \sim (1+t)^{(\gamma-(1+q))/(1+q)}$ and $f_{\gamma,q} \sim \widetilde{V}_{\gamma-(1+q)}$ such that, for any pair of inverse Young functions $\Psi = (\Psi_1, \Psi_2) \in \mathcal{I}$ and some constant $C(\Psi)$, we have*

$$(\Psi_1(r_{\gamma,q}(t)) \vee 1)\|\mathbb{P}^x(X_t \in \cdot) - \mu\|_{\mathbf{1} \vee \Psi_2 \circ f_{\gamma,q}} \leq C(\Psi)\widetilde{V}_\gamma(x), \quad x \in \mathbb{R}^d, t \geq 0, \tag{9}$$

and, for any accessible set $B \in \mathcal{B}(\mathbb{R}^d)$ and any $\delta > 0$, there exists a constant $c(\Psi) > 0$ such that

$$\mathbb{E}^x\Big[\int_0^{\tau_B(\delta)} \Psi_1(r_{\gamma,q}(t))\Psi_2(f_{\gamma,q}(X_t))\,\mathrm{d}t\Big] \le c(\Psi)\widetilde{V}_\gamma(x), \quad x \in \mathbb{R}^d. \tag{10}$$

Moreover, if $q = -1$ and $\gamma > 0$, then, for any $\alpha \in (0, \mathfrak{r}\gamma)$, there exist a function $r_\alpha(t) \sim \exp(-\alpha t)$ and $f_\gamma \sim \widetilde{V}_\gamma$ such that (1) and (9) are true with $f_{\gamma,q}$ and $r_{\gamma,q}$ replaced by f_γ and r_α, respectively.

Let us note that, for any $\eta \ge 0$,

$$\sup_{t\ge 0} \mathbb{E}^x\big[\|\widetilde{X}_t\|^\eta\big] \le c(\eta)\big(1 + \|x\|^\eta\big), \quad x \in \mathbb{R}^d,$$

where $\widetilde{X}_t = X_{\tau^{-1}(t)}$ for the time change $\tau(t) \coloneqq \int_0^t \|\sigma^\top(X_s)X_s/\|X_s\|\|^2\,\mathrm{d}s$, cf. (Pardoux and Veretennikov, 2001, Proposition 1). Setting $\Psi_1 = \mathbf{1}$ and $\Psi_2 = \mathrm{Id}$ in (9), it follows for the process on its unchanged time scale that, for any $\eta > 0$,

$$\sup_{t\ge 0} \mathbb{E}^x\big[\|X_t\|^\eta\big] \le \mathfrak{C}(\eta)(1 + \|x\|^{\eta+1+q}), \quad x \in \mathbb{R}^d. \tag{11}$$

3. Concentration of additive diffusion functionals

Recall the definition of the scaled additive functionals $\mathbb{G}_t(f)$ and $\mathbb{G}_{n,\Delta}(f)$ from (1) and (2), respectively. Motivated by the existence of a regular solution to the Poisson equation for polynomially bounded functions, we study deviations of $\mathbb{G}_t(f)$ and its discrete version $\mathbb{G}_{n,\Delta}(f)$ for functions f belonging to the function class $\mathcal{F}(\eta, \mathfrak{L})$ given by

$$\mathcal{F}(\eta, \mathfrak{L}) \coloneqq \big\{\widetilde{f} - \mu(\widetilde{f}) : \widetilde{f} \in \mathcal{G}(\eta, \mathfrak{L})\big\},$$

for

$$\mathcal{G}(\eta, \mathfrak{L}) \coloneqq \big\{f\colon \mathbb{R}^d \to \mathbb{R} : |f(x)| \le \mathfrak{L}(1 + \|x\|^\eta),\, x \in \mathbb{R}^d\big\},$$

for some finite constant $\mathfrak{L} > 0, \eta \ge 0$.

There is a vast amount of literature on concentration inequalities for path integrals of general Markov processes. The most powerful results are generally established under the assumption of functional inequalities such as Poincaré or log-Sobolev. However, the elliptic diffusions considered in this paper generally do not satisfy such rather strong functional inequalities. In this regard, Cattiaux and Guillin (2008) establish concentration inequalities for bounded functionals under a so-called weak Poincaré inequality, which is demonstrated to be equivalent to an α-mixing assumption on the process, cf. (Cattiaux and Guillin, 2008, Proposition 3.4). Recall that a stationary Markov process $(Y_t)_{t\ge 0}$ with natural filtration $(\mathcal{F}_t)_{t\ge 0}$ and initial distribution ν is said to be α-mixing if the mixing coefficient $\alpha_\nu(t) \coloneqq \sup_{s\ge 0}\sup_{A\in\mathcal{F}_s, B\in\mathcal{F}_{s+t}} |\mathbb{P}^\nu(A\cap B) - \mathbb{P}^\nu(A)\mathbb{P}^\nu(B)|$ tends to zero as $t \to \infty$. It follows from (5), (6) and (7) that the stationary β-mixing coefficient $\beta(t) \coloneqq \int_{\mathbb{R}^d} \|\mathbb{P}^x(X_t \in \cdot) - \mu\|_{\mathrm{TV}}\,\mu(\mathrm{d}x)$ of our diffusion process satisfies

$$\beta(t) \le c\exp\big(-\iota'' t^{(1-q_+)/(1+q_+)}\big),$$

for any $\iota'' \in (0, \iota')$ and some constant c depending on ι'', i.e., the stationary diffusion $\mathbf{X}$ is subexponentially β-mixing. Consequently, using the well-known fact that $\alpha_\mu(t) \le \beta(t)$, (Cattiaux and Guillin, 2008, Proposition 3.9) yields the following result for *bounded* f.

Theorem 2 *(Cattiaux and Guillin, 2008, Proposition 3.9) For $\iota'' \in (0, \iota')$, define*

$$c(q, \iota'') := \frac{1}{2}\left(\frac{1+q_+}{1-q_+}\right)^{1/(1-q_+)} \left(\frac{(1-q_+)\iota''}{1+q_+}\right)^{(1+q_+)/(2(1-q_+))}. \tag{12}$$

For any such ι'', there exists a constant $\mathfrak{c} > 0$ such that, for all $f \in \mathcal{F}(0, \mathfrak{L})$ and $(u, t) \in \mathbb{R}_+^2$ such that

$$\mathfrak{c}(1+q_+)(1-q_+)^{-(1-q_+)/2} \leq u < \big(c(q, \iota'')\lfloor t \rfloor/\sqrt{t}\big)^{1-q_+}, \tag{13}$$

it holds

$$\mathbb{P}^\mu\Big(|\mathbb{G}_t(f)| > 2\mathfrak{L}\big(c(q, \iota'')^{-1} u^{\frac{1}{1-q_+}} + t^{-1/2}\big)\Big) \leq 2\mathrm{e}^{-u}.$$

In the above result, the restriction on u in (13) is explained by the proof technique that makes use of general moment bounds for discrete α-mixing sequences from Rio (2017). This approach requires the integral to be divided into a finite number of blocks with a carefully chosen length that determines the degree of mixing of the block sequence.

In the following, we add to this result by allowing polynomially growing integrands f. It is well-known that dropping the boundedness assumption poses major challenges in deriving concentration inequalities, some of which have been elegantly solved in Gao et al. (2014) for symmetric Markov processes satisfying (strong) functional inequalities. It should also be noted that in (Cattiaux and Guillin, 2008, Section 3.2) some arguments are provided how conclusions for unbounded integrands f can be drawn from Theorem 2 by employing a truncation technique. However, there appears to be a gap in the proposed strategy, which prevents it from being applicable for $u > 0$ such that $u/\sqrt{t}$ is small. Since our ultimate focus is on applications of our concentration inequalities to the inference of PAC bounds for $t^{-1/2}\mathbb{G}_t(f)$, we do not further pursue an approach relying on discrete mixing results, but employ a different technique that is embedded more naturally in the continuous framework.

3.1 Continuous observations

Our main result for continuous observations is the following exponential concentration bound for polynomially bounded functions.

Theorem 3 *There exists a constant $\mathfrak{W}$, depending on q, η and the diffusion coefficients b and σ, such that, for any $p \geq 2$, $t > 0$ and $f \in \mathcal{F}(\eta, \mathfrak{L})$, we have*

$$\|\mathbb{G}_t(f)\|_{L^p(\mathbb{P}^\mu)} \leq \mathfrak{L}\mathfrak{W} p^{\frac{1}{2} + \frac{\eta+q'+q+1}{1-q_+}}. \tag{14}$$

As a consequence, for any $t > 0$,

$$\mathbb{P}^\mu\Big(|\mathbb{G}_t(f)| > \mathrm{e}\mathfrak{L}\mathfrak{W} u^{\frac{1}{2} + \frac{\eta+q'+q+1}{1-q_+}}\Big) \leq \mathrm{e}^{-u}, \quad u \geq 2. \tag{15}$$

The proof will be given by combining a sequence of technical lemmas that we develop in the following. An interpretation of the result will be stated later in Remark 7 since this requires making explicit reference to the proof. The first result that we need are bounds on the L^p-norms of the invariant measure μ which are implied by its subexponential tails.

Lemma 4 *For all $p \geq 1$, it holds*

$$\mathbb{E}^{\mu}\big[\|X_0\|^p\big]^{1/p} \leq c_{q_+} p^{1/(1-q_+)},$$

where

$$c_{q_+} = \mathrm{e}^{\mathrm{e}/2+(1-q_+)/12}((1-q_+)\iota\mathrm{e})^{-1/(1-q_+)}\sqrt{\frac{2\pi}{1-q_+}}\mathbb{E}^{\mu}[V_{q+}(X_0)].$$

Proof Let $a_{q_+} = ((1-q_+)\iota\mathrm{e})^{-1/(1-q_+)}$. Using Markov's inequality, it follows that, for any $u \geq 1$,

$$\mathbb{P}^{\mu}\big(\|X_0\| \geq \mathrm{e}^{1/(1-q_+)}a_{q_+}u\big) \leq \mathbb{E}^{\mu}[V_{q_+}(X_0)]\exp\big(-u^{1-q_+}/(1-q_+)\big),$$

with $\mathbb{E}^{\mu}[V_{q_+}(X_0)] < \infty$ due to (7). The assertion now follows from (Foucart and Rauhut, 2013, Proposition 7.13). ∎

Next, we state the martingale approximation of the additive functional $\mathbb{G}_t(f)$ for polynomially bounded f with the help of the Itō–Krylov formula, which extends the usual Itō formula for diffusion processes from functions with $\mathcal{C}^2$-regularity to functions with slightly weaker Sobolev regularity. This is necessary in light of the regularity of the solution to the Poisson equation that we described in Section 2.

Lemma 5 *For any $f \in \mathcal{F}(\eta, \mathfrak{L})$, we have a decomposition*

$$\mathbb{G}_t(f) = \frac{1}{\sqrt{t}}\mathbb{M}_t(f) + \frac{1}{\sqrt{t}}\mathbb{R}_t(f),$$

where $(\mathbb{M}_t(f))_{t\geq 0}$ is a continuous square-integrable $\mathbb{P}^{\mu}$-martingale and both $f \mapsto \mathbb{M}_\cdot(f)$ and $f \mapsto \mathbb{R}_\cdot(f)$ are linear. Moreover, there exists a global constant $c \geq 1$ such that, for any $p \geq 1, t \geq 0$

$$\mathbb{E}^{\mu}\big[|\mathbb{M}_t(f)|^p\big]^{1/p} \leq c\lambda_+^{1/2}p^{1/2}\sqrt{t}\big\|\|\nabla L^{-1}[f]\|\big\|_{L^{p\vee 2}(\mu)}, \tag{16}$$

and

$$\mathbb{E}^{\mu}\big[|\mathbb{R}_t(f)|^p\big]^{1/p} \leq 2\|L^{-1}[f]\|_{L^p(\mu)}. \tag{17}$$

Proof For any $p \geq 1$, $L^{-1}[f] \in \mathcal{W}^{2,p}_{\mathrm{loc}}(\mathbb{R}^d)$, the coefficients b, σ are locally bounded, $\sigma\sigma^\top$ is uniformly positive definite and (7) guarantees $\mathbb{E}^{\mu}[\int_0^t \|b(X_s)\|^2\,\mathrm{d}s] < \infty$ since $\|b\| \lesssim \widetilde{V}_{q'}$. Thus, we can apply the Itō–Krylov formula (cf. (Krylov, 2009, Theorem 2.10.1)) to obtain the $\mathbb{P}^{\mu}$-a.s. identities

$$\begin{aligned} L^{-1}[f](X_t) &= L^{-1}[f](X_0) + \int_0^t \big(\nabla L^{-1}[f](X_s)\big)^\top \sigma(X_s)\,\mathrm{d}W_s + \int_0^t LL^{-1}[f](X_s)\,\mathrm{d}s \\ &= L^{-1}[f](X_0) + \int_0^t \big(\nabla L^{-1}[f](X_s)\big)^\top \sigma(X_s)\,\mathrm{d}W_s + \int_0^t f(X_s)\,\mathrm{d}s, \quad t \geq 0. \end{aligned}$$

Here, the second equality follows from $LL^{-1}[f] = f$, $\boldsymbol{\lambda}$-a.e., and $\mathbb{P}^x(X_t \in \cdot) \ll \boldsymbol{\lambda}$ for any $(t,x) \in (0,\infty)\times\mathbb{R}^d$, which implies

$$\mathbb{E}^x\Big[\Big|\int_0^t LL^{-1}[f](X_s)\,\mathrm{d}s - \int_0^t f(X_s)\,\mathrm{d}s\Big|\Big] \leq \int_0^t\int_{\mathbb{R}^d}|LL^{-1}[f](y) - f(y)|p_s(x,y)\,\mathrm{d}y\,\mathrm{d}s = 0.$$

Thus, $\int_0^t LL^{-1}[f](X_s)\,\mathrm{d}t = \int_0^t f(X_s)\,\mathrm{d}s$ $\mathbb{P}^x$-a.s. for any $x \in \mathbb{R}^d$ and hence also $\mathbb{P}^\mu$-a.s. follows. Consequently,

$$\mathbb{G}_t(f) = \frac{1}{\sqrt{t}}\int_0^t f(X_s)\,\mathrm{d}s = \frac{1}{\sqrt{t}}\mathbb{M}_t(f) + \frac{1}{\sqrt{t}}\mathbb{R}_t(f), \quad t \geq 0,$$

where

$$\mathbb{M}_t(f) = -\int_0^t \big(\nabla L^{-1}[f](X_s)\big)^\top \sigma(X_s)\,\mathrm{d}W_s, \quad t \geq 0,$$

and

$$\mathbb{R}_t(f) = L^{-1}[f](X_t) - L^{-1}[f](X_0), \quad t \geq 0.$$

The square-integrable martingale property of $\mathbb{M}_\cdot(f)$ follows from

$$\|\nabla L^{-1}[f](x)\| \lesssim \widetilde{V}_{\eta+q'+q+1}(x), \quad x \in \mathbb{R}^d,$$

which is demonstrated in the proof of Lemma 6, such that (7) implies that

$$\int_0^t \mathbb{E}^\mu\big[\big\|\big(\nabla L^{-1}[f](X_s)\big)^\top \sigma(X_s)\big\|^2\big]\,\mathrm{d}s \lesssim t\lambda_+\mu\big(\widetilde{V}_{2(\eta+q'+q+1)}\big) < \infty, \quad t \geq 0.$$

The bound (17) is now an immediate consequence of stationarity under $\mathbb{P}^\mu$ and Minkowski's inequality. Moreover, using the Burkholder–Davis–Gundy inequality for continuous martingales started in 0 in the form given in (Barlow and Yor, 1982, Proposition 4.2), it follows that, for some $c \geq 1$ and any $p \geq 1$,

$$\begin{aligned}\mathbb{E}^\mu\big[|\mathbb{M}_t(f)|^p\big] &\leq c^p p^{p/2}\mathbb{E}^\mu\big[\langle \mathbb{M}_\cdot(f)\rangle_t^{p/2}\big] \\ &= c^p p^{p/2}\mathbb{E}^\mu\Big[\Big(\int_0^t \|\sigma^\top(X_s)\nabla L^{-1}[f](X_s)\|^2\,\mathrm{d}s\Big)^{p/2}\Big] \\ &\leq c^p p^{p/2}\lambda_+^{p/2}\mathbb{E}^\mu\Big[\Big(\int_0^t \|\nabla L^{-1}[f](X_s)\|^2\,\mathrm{d}s\Big)^{p/2}\Big].\end{aligned} \tag{18}$$

Consequently, for $p \geq 2$, first using Jensen's inequality and then Fubini together with stationarity gives

$$\begin{aligned}\mathbb{E}^\mu\big[|\mathbb{M}_t(f)|^p\big] &\leq c^p p^{p/2}\lambda_+^{p/2}t^{p/2}\mathbb{E}^\mu\Big[\frac{1}{t}\int_0^t \|\nabla L^{-1}[f](X_s)\|^p\,\mathrm{d}s\Big] \\ &= c^p p^{p/2}\lambda_+^{p/2}t^{p/2}\big\|\|\nabla L^{-1}[f]\|\big\|_{L^p(\mu)}^p,\end{aligned}$$

and hence

$$\mathbb{E}^\mu\big[|\mathbb{M}_t(f)|^p\big]^{1/p} \leq c\lambda_+^{1/2}p^{1/2}\sqrt{t}\big\|\|\nabla L^{-1}[f]\|\big\|_{L^p(\mu)}, \quad t \geq 0.$$

In case $p \in [1,2)$, we get from (18) with another application of Jensen's inequality and Fubini

$$\begin{aligned}\mathbb{E}^\mu\big[|\mathbb{M}_t(f)|^p\big] &\leq c^p p^{p/2}\lambda_+^{p/2}\mathbb{E}^\mu\Big[\frac{1}{t}\int_0^t \big\|\nabla L^{-1}[f](X_s)\big\|^2\,\mathrm{d}s\Big]^{p/2} \\ &= c^p p^{p/2}\lambda_+^{p/2}t^{p/2}\big\|\|\nabla L^{-1}[f]\|\big\|_{L^2(\mu)}^p.\end{aligned}$$

Thus, for any $p \geq 1$, (16) follows. ■

In view of (16) and (17), to exploit the martingale approximation we need concrete bounds on the solution of the Poisson equation $L^{-1}[f]$ and its gradient $\nabla L^{-1}[f]$. This is the content of the next lemma, which can essentially be obtained from combining Lemma 4 with the Sobolev estimates from Pardoux and Veretennikov (2001) and Bogachev et al. (2018). For later reference and some clarification concerning the role of the drift growth, we give a full proof that simplifies some arguments from Pardoux and Veretennikov (2001) thanks to Proposition 1.

Lemma 6 *Let $p \geq 1$. There exist constants $\mathfrak{U}(q,\eta), \mathfrak{V}(q,q',\eta)$ (independent of p) such that, for any $f = \widetilde{f} - \mu(\widetilde{f}) \in \mathcal{F}(\eta, \mathfrak{L})$,*

$$\big\|L^{-1}[f]\big\|_{L^p(\mu)} \leq \mathfrak{L}\mathfrak{U}(q,\eta) p^{\frac{\eta+q+1}{1-q_+}} \tag{19}$$

and

$$\big\|\|\nabla L^{-1}[f]\|\big\|_{L^p(\mu)} \leq \mathfrak{L}\mathfrak{V}(q,q',\eta) p^{\frac{\eta+q'+q+1}{1-q_+}}. \tag{20}$$

Proof By a slight adjustment to the proof of Proposition 4.1 in Bogachev et al. (2018)[1], we obtain for any $r > d$

$$\big\|\nabla L^{-1}[f](x)\big\| \lesssim \big(1 + \sup_{y \in B(x,1)} |b(y)|\big) \|L^{-1}[f]\|_{L^r(B(x,1))} + \|f\|_{L^r(B(x,1))}. \tag{21}$$

Therefore, using Hölder's inequality and the growth condition on the drift b,

$$\big\|\|\nabla L^{-1}[f]\|\big\|_{L^p(\mu)} \lesssim \|(1 + \|\cdot\|^{q'})\|_{L^{2p}(\mu)} \big\|\|L^{-1}[f]\|_{L^r(B(\cdot,1))}\big\|_{L^{2p}(\mu)} + \big\|\|f\|_{L^r(B(\cdot,1))}\big\|_{L^p(\mu)}. \tag{22}$$

If $q > -1$, let $\gamma > 2(1+q)$. Then we can calculate as in the proof of (Pardoux and Veretennikov, 2001, Theorem 1) to obtain

$$\begin{aligned}
|L^{-1}[f](x)| &\leq \int_0^\infty \int_{\mathbb{R}^d} |\widetilde{f}(y)| |p_t(x,y) - \rho(y)| \,\mathrm{d}y \,\mathrm{d}t \\
&\leq \int_0^\infty \Big(\int_{\mathbb{R}^d} |p_t(x,y) - \rho(y)| \,\mathrm{d}y \Big)^{1/2} \Big(\int_{\mathbb{R}^d} |\widetilde{f}(y)|^2 (p_t(x,y) + \rho(y)) \,\mathrm{d}y \Big)^{1/2} \mathrm{d}t \Big) \\
&= \int_0^\infty \big(\|P_t(x,\cdot) - \mu\|_{\mathrm{TV}} \big)^{1/2} \Big(\int_{\mathbb{R}^d} |\widetilde{f}(y)|^2 (p_t(x,y) + \rho(y)) \,\mathrm{d}y \Big)^{1/2} \mathrm{d}t \Big) \\
&\leq \mathfrak{L} C'(\gamma,q) (\widetilde{V}_\gamma(x))^{1/2} \int_0^\infty (1+t)^{-\frac{\gamma-(1+q)}{1+q}} \Big(\int_{\mathbb{R}^d} \widetilde{V}_\eta(y)^2 (p_t(x,y) + \rho(y)) \,\mathrm{d}y \Big)^{1/2} \mathrm{d}t \\
&\leq \mathfrak{L} C(\eta,\gamma,q) (\widetilde{V}_\gamma(x))^{1/2} (1 + \|x\|^{\eta + (1+q)/2}) \int_1^\infty t^{-(\gamma-(1+q))/(1+q)} \,\mathrm{d}t \\
&= \mathfrak{L} C'(\eta,\gamma,q) \widetilde{V}_{\eta + (1+q+\gamma)/2}(x),
\end{aligned}$$

1. as the authors point out, the gradient bounds derived in (Pardoux and Veretennikov, 2001, Theorem 1) are only valid in case of bounded drift

where we used Cauchy–Schwarz for the second inequality and (9) for the third inequality. The last inequality arises from (7) and (11). A similar calculation, using exponential ergodicity with arbitrary polynomial penalty $\widetilde{V}_\gamma$ in case $q = -1$ with any $\gamma > 0$, shows that the above estimate remains valid for $q = -1$. By the strong Markov property, we have for any $R > 0$ and $\tau_R \coloneqq \tau_{\overline{B(0,R)}}$,

$$L^{-1}[f](x) = \mathbb{E}^x\big[L^{-1}[f](X_{\tau_R})\big] + \mathbb{E}^x\Big[\int_0^{\tau_R} f(X_t)\,\mathrm{d}t\Big].$$

By the above, u is locally bounded and thus the first term is bounded for any $R > 0$. For the second term, we can employ the Itō formula argument from (Pardoux and Veretennikov, 2001, Theorem 2) to improve this bound to $|L^{-1}[f]| \lesssim \mathfrak{L}\widetilde{V}_{\eta+1+q}$. Alternatively, apart from the case $\eta = 0$, $q = -1$, we may simply note that, by setting $\Psi_1 = \mathbf{1}$ and $\Psi_2 = \mathrm{Id}$, (10) yields that for any $\delta > 0$

$$\Big|\mathbb{E}^x\Big[\int_0^{\tau_R} f(X_t)\,\mathrm{d}t\Big]\Big| \le \mathbb{E}^x\Big[\int_0^{\tau_R(\delta)} |f(X_t)|\,\mathrm{d}t\Big] \le \mathfrak{L}C(\eta, q, R, \delta)\widetilde{V}_{\eta+1+q}(x), \quad x \in \mathbb{R}^d.$$

Here we used that $|f| \lesssim \mathfrak{L}\widetilde{V}_\eta$ and that $B(0,R)$ is accessible by $\boldsymbol{\lambda}$-irreducibility of $\mathbf{X}$ implied by uniform positive definiteness of $\sigma\sigma^\top$, cf. (Stramer and Tweedie, 1997, Theorem 2.3). Thus,

$$|L^{-1}[f](x)| \le C\mathfrak{L}\widetilde{V}_{\eta+1+q}(x), \quad x \in \mathbb{R}^d,$$

follows for some constant C depending on η and q. Consequently, for any $p \ge 1$, Lemma 4 yields

$$\|L^{-1}[f]\|_{L^p(\mu)} \le C\mathfrak{L}\|\widetilde{V}_{\eta+1+q}\|_{L^p(\mu)} \le C\mathfrak{L}c_q^{\eta+q+1}p^{\frac{\eta+q+1}{1-q_+}}.$$

Using that

$$|\widetilde{V}_{\eta+q+1}(x+y)| \le 2^{\eta+q}(\widetilde{V}_{\eta+q+1}(x)+\widetilde{V}_{\eta+q+1}(y)) \le 2^{\eta+q}(2+\widetilde{V}_{\eta+q+1}(x)), \quad x \in \mathbb{R}^d, y \in B(0,1),$$

it also follows that

$$\big\|\|L^{-1}[f]\|_{L^r(B(\cdot,1))}\big\|_{L^{2p}(\mu)} \le C\mathfrak{L}2^{\eta+q+1}\|\widetilde{V}_{\eta+q+1}\|_{L^{2p}(\mu)} \le \mathfrak{L}C2^{\eta+q+1}c_q^{\eta+q+1}(2p)^{\frac{\eta+q+1}{1-q_+}}.$$

Moreover, $|\widetilde{f}(x+y)| \le \mathfrak{L}2^\eta(2+\widetilde{V}(x))$ for $y \in B(0,1)$ implies

$$\begin{aligned}\|\|f\|_{L^r B(\cdot,1)}\|_{L^p(\mu)} &\le \mathfrak{L}2^\eta\big(2 + (1+\boldsymbol{\lambda}(B(0,1)))\|\widetilde{V}_\eta\|_{L^p(\mu)}\big)\\ &\le \mathfrak{L}2^\eta(1 \vee \boldsymbol{\lambda}(B(0,1)))(1 + c_{q+}^\eta p^{\eta/(1-q_+)}),\end{aligned}$$

such that (22) allows us to conclude that

$$\big\|\|\nabla L^{-1}[f]\|\big\|_{L^p(\mu)} \le \mathfrak{L}\mathfrak{V}(q, q', \eta)p^{\frac{\eta+q'+q+1}{1-q_+}}.$$

■

We are now ready to infer Theorem 3 from the previous results.

Proof [Proof of Theorem 3] The moment bounds (14) are an immediate consequence of the combined statements of Lemma 5 and Lemma 6. By Markov's inequality, (14) implies (15). ■

Remark 7 *It would be desirable that the concentration rate provided by Theorem 3 matches the rate in Theorem 2 for the bounded case $\eta = 0$ and the rates for polynomially growing integrands for scalar, exponentially ergodic diffusions with at most linear drift (i.e., $d = 1, q = 0, q' = 1$) from (Aeckerle-Willems and Strauch, 2021, Proposition 7). The reason for the gap in the rate can be traced down to the Sobolev estimates* (21)*, where the gradient $\nabla L^{-1}[f]$ is bounded in terms of $L^{-1}[f]$. In contrast, the strategy in Aeckerle-Willems and Strauch (2021), see also Galtchouk and Pergamenshchikov (2007), works in the other direction. That is, by exploiting the explicit solution of the Poisson equation in $d = 1$, tight pointwise bounds on the gradient $\nabla L^{-1}[f]$ are established first, which are then used to bound the remainder term $L^{-1}[f](X_t) - L^{-1}[f](X_0) = \int_{X_0}^{X_t} \nabla L^{-1}[f](x)\,\mathrm{d}x$ in the martingale approximation. Such a strategy is not feasible in the multivariate setting since $L^{-1}[f]$ is not explicitly known. Improving our concentration result Theorem 3 would therefore require tighter estimates on the solution of the Poisson equation and its gradient than those that can be achieved with the ideas from Pardoux and Veretennikov (2001). This is a challenging and interesting question for future research.*

Stationary and non-stationary PAC bounds As an immediate consequence of Theorem 2 and Theorem 3, we can derive the following quantitative version of the ergodic theorem and a stationary PAC bound for (sub-) geometric diffusions. Let us define the rate function

$$\varsigma(\eta, q, q') \coloneqq \begin{cases} \frac{1}{1-q_+}, & \text{if } \eta = 0, \\ \frac{1}{2} + \frac{\eta+q'+q+1}{1-q_+}, & \text{if } \eta > 0, \end{cases}$$

and the sample length function

$$\Psi(\varepsilon, \delta) \coloneqq \begin{cases} \left(\frac{c(q,\iota'')^{-1} (\log(2/\delta))^{1/(1-q_+)} + 1}{1 \wedge \varepsilon/(2\mathfrak{L})} \right)^2, & \text{if } \eta = 0, \\ \left(\frac{\mathrm{e}\mathfrak{L}\mathfrak{W}(\log(1/\delta))^{\frac{1}{2} + \frac{\eta+q'+q+1}{1-q_+}}}{\varepsilon} \right)^2, & \text{if } \eta > 0, \end{cases}$$

with $c(q, \iota'')$ defined in (12) and $\mathfrak{W}$ denoting the constant established in Theorem 3.

Corollary 8 *Let $f \in \mathcal{G}(\eta, \mathfrak{L})$, $\varepsilon > 0$ and $\delta \in (0,1)$ such that $\delta < 2\exp(-\mathfrak{c}(1+q_+)(1-q_+)^{-(1-q_+)/2})$ if $\eta = 0$ and $\delta < \mathrm{e}^{-2}$ if $\eta > 0$. Then, for $t \geq \Psi(\varepsilon, \delta)$, it holds*

$$\mathbb{P}^\mu\Big(\Big| \frac{1}{t} \int_0^t f(X_s)\,\mathrm{d}s - \mu(f) \Big| \leq \varepsilon \Big) \geq 1 - \delta. \tag{23}$$

Moreover, for any increasing sequence $(t_n)_{n\in\mathbb{N}} \subset \mathbb{R}_+$ with $\inf_{n\in\mathbb{N}}(t_n - t_{n-1}) > 0$, it holds for any $\delta_0 > 0$

$$\lim_{n\to\infty} \sqrt{t_n} (\log t_n)^{-(\varsigma(\eta,q,q')+\delta_0)} \Big| \frac{1}{t_n} \int_0^{t_n} f(X_s)\,\mathrm{d}s - \mu(f) \Big| = 0, \quad \mathbb{P}^\mu\text{-a.s.} \tag{24}$$

If f is bounded, we even have, for any $\widetilde{q} \in (q_+, 1)$,

$$\lim_{t\to\infty} \sqrt{t}(\log t)^{-\frac{1}{1-\widetilde{q}}} \Big| \frac{1}{t}\int_0^t f(X_s)\,\mathrm{d}s - \mu(f)\Big| = 0, \quad \mathbb{P}^\mu\text{-}a.s.$$

Proof The first two assertions immediately follow from Theorem 2 and Theorem 3. For any $\varepsilon > 0$ and $\delta > 0$, there exists $t(\varepsilon) \geq \mathrm{e}$ such that, for any $t \geq t(\varepsilon)$, we have

$$\mathfrak{L}\{(\mathrm{e}\mathfrak{W}) \vee (c(q,\iota'')^{-1} + 1)\}(2\log t)^{-\frac{\delta_0}{1-q_+}} \leq \varepsilon.$$

Consequently, by Theorem 2 and Theorem 3, it follows that for

$$U_t \coloneqq \sqrt{t}(2\log t)^{-(\varsigma(\eta,q,q')+\delta_0)}\Big(\frac{1}{t}\int_0^t f(X_s)\,\mathrm{d}s - \mu(f)\Big)$$

and $t \geq t(\varepsilon)$ such that, in case $\eta = 0$, additionally $2\log t \geq \mathfrak{c}(1+q_+)(1-q_+)^{-(1-q_+)/2}$,

$$\mathbb{P}^\mu(|U_t| > \varepsilon) \leq \mathbb{P}^\mu\Big(|\mathbb{G}_t(f)| > \mathfrak{L}\{(\mathrm{e}\mathfrak{W}) \vee (c(q,\iota'')^{-1}+1)\}(2\log t)^{\varsigma(\eta,q,q')}\Big) \leq t^{-2}.$$

Thus,

$$\mathbb{P}^\mu(|U_t| > \varepsilon) \leq \mathbf{1}_{[0,t(\varepsilon))} + t^{-2}\mathbf{1}_{[t(\varepsilon),\infty)} =: g_\varepsilon(t), \quad t > 0.$$

Since $g_\varepsilon \in L^1(\mathbb{R}_+)$ and is decreasing, it follows for $a \coloneqq \inf_{n\in\mathbb{N}}(t_{n+1} - t_n) > 0$

$$\infty > \int_{t_n}^\infty g_\varepsilon(t)\,\mathrm{d}t \geq \sum_{m\geq n}(t_{m+1}-t_m)g_\varepsilon(t_{m+1}) \geq a\sum_{m\geq n+1} g_\varepsilon(t_m) \geq a\sum_{m\geq n+1}\mathbb{P}^\mu(|U_{t_m}| > \varepsilon).$$

Hence, for any $\varepsilon > 0$, $\sum_{n\in\mathbb{N}}\mathbb{P}^\mu(|U_{t_n}| > \varepsilon) < \infty$ such that Borel–Cantelli implies that $\lim_{n\to\infty} U_{t_n} = 0$, $\mathbb{P}^\mu$-a.s., which gives (24). This argument is borrowed from the proof of Lemma 3.1 in Bosq (1997). By the same lemma, it follows from the above that we even have convergence along any sequence $(\widetilde{t}_n)_{n\in\mathbb{N}}$, $\mathbb{P}^\mu$-a.s., provided that the map $t \mapsto U_t$ is uniformly continuous $\mathbb{P}^\mu$-a.s. This can be easily verified when f is bounded (see, e.g., the proof of Proposition 4.3 in Bosq (1997)), which proves the last assertion. ■

To get a non-stationary PAC bound, we consider the burn-in sample average

$$\mathbb{H}_{v,t}(f) \coloneqq \frac{1}{\sqrt{t}}\mathbb{G}_t(f)\circ\theta_v = \frac{1}{t}\int_v^{v+t} f(X_s)\,\mathrm{d}s, \quad t > 0, v \geq 0,$$

with burn-in length v. Our naming convention follows the MCMC literature, where a standard procedure of dealing with non-stationary simulation procedures is to run the simulation algorithm for a certain amount of time before collecting samples, which is usually referred to as the burn-in.

Corollary 9 *Let $\varepsilon > 0, \delta \in (0,1)$ such that $\delta < 2\mathrm{e}^{-2}$ if $\eta > 0$ and $\delta < 4\exp(-\mathfrak{c}(1+q_+)(1-q_+)^{-(1-q_+)/2})$ if $\eta = 0$. Let also ν be some probability distribution such that $V_{q_+} \in L^1(\nu)$.*

Choose some $\iota'' \in (0, \iota')$*, and define* $C \coloneqq C(q_+)c(\iota'')\|V_{q_+}\|_{L^1(\nu)}$*, where* $c(\iota'')$ *is some constant such that*

$$\forall t \geq 1: \quad (1+t)^{\frac{2q_+}{1+q_+}} \mathrm{e}^{-(\iota' t)^{(1-q_+)/(1+q_+)}} \leq c(\iota'')\mathrm{e}^{-(\iota'' t)^{(1-q_+)/(1+q_+)}}.$$

Then, for $t \geq \Psi(\varepsilon, \delta/2)$ *and burn-in length* $v \geq 1 \vee (\log(2C/\delta))^{(1+q_+)/(1-q_+)}/\iota''$*, we have, for any* $f \in \mathcal{G}(\eta, \mathfrak{L})$*,*

$$\mathbb{P}^\nu(|\mathbb{H}_{v,t}(f) - \mu(f)| \leq \varepsilon) \geq 1 - \delta.$$

Proof Under the given assumptions, (5), (6) imply

$$\|\mathbb{P}^x(X_t \in \cdot) - \mu\|_{\mathrm{TV}} \leq C\frac{V_{q_+}(x)}{\|V_{q_+}\|_{L^1(\nu)}}\mathrm{e}^{-(\iota'' t)^{(1-q_+)/(1+q_+)}}. \tag{25}$$

Define $g(y) \coloneqq \mathbb{P}^y(|t^{-1}\int_0^t f(X_s)\,\mathrm{d}s - \mu(f)| > \varepsilon)$. By the Markov property, (25) and the magnitude of the burn-in v, for any $x \in \mathbb{R}^d$,

$$\begin{aligned}
&\Big|\mathbb{P}^x(|\mathbb{H}_{v,t}(f) - \mu(f)| > \varepsilon) - \mathbb{P}^\mu\Big(\Big|\frac{1}{t}\int_0^t f(X_s)\,\mathrm{d}s - \mu(f)\Big| > \varepsilon\Big)\Big| \\
&\quad = \big|\mathbb{E}^x[g(X_v)] - \mu(g)\big| \leq \|\mathbb{P}^x(X_v \in \cdot) - \mu\|_{\mathrm{TV}} \leq C\frac{V_{q_+}(x)}{\|V_{q_+}\|_{L^1(\nu)}}\mathrm{e}^{-(\iota'' v)^{(1-q_+)/(1+q_+)}} \\
&\quad \leq \frac{V_{q_+}(x)}{\|V_{q_+}\|_{L^1(\nu)}}\,\frac{\delta}{2}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\Big|\mathbb{P}^\nu(|\mathbb{H}_{v,t}(f) - \mu(f)| > \varepsilon) - \mathbb{P}^\mu\Big(\Big|\frac{1}{t}\int_0^t f(X_s)\,\mathrm{d}s - \mu(f)\Big| > \varepsilon\Big)\Big| \\
&\quad \leq \int_{\mathbb{R}^d}\Big|\mathbb{P}^x(|\mathbb{H}_{v,t}(f) - \mu(f)| > \varepsilon) - \mathbb{P}^\mu\Big(\Big|\frac{1}{t}\int_0^t f(X_s)\,\mathrm{d}s - \mu(f)\Big| > \varepsilon\Big)\Big|\,\nu(\mathrm{d}x) \leq \frac{\delta}{2}.
\end{aligned}$$

Consequently, using $t \geq \Psi(\varepsilon, \delta/2)$ and (23), it follows by the triangle inequality that $\mathbb{P}^\nu(|\mathbb{H}_{v,t}(f) - \mu(f)| > \varepsilon) \leq \delta$. ■

3.2 Discrete observations

We now derive concentration inequalities for discrete observations from our continuous observation results by using the approximation strategy from Galtchouk and Pergamenshchikov (2013). In Galtchouk and Pergamenshchikov (2013), only bounded functions f and scalar exponentially ergodic diffusions in the quite strong regime ($\mathscr{A}(q)$) with $q = -1$ are considered, which in particular implies sub-Gaussian tails of the invariant density. We demonstrate how this can be extended to the multivariate case for unbounded functions f under less restrictive ergodicity assumptions. For this purpose, the following technical key result from Galtchouk and Pergamenshchikov (2013) is of central importance.

Lemma 10 *(Galtchouk and Pergamenshchikov, 2013, Proposition A.1) Let a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_j)_{j=1,\dots,n}, \mathbb{P})$ a random vector $(\mathcal{X}_j)_{j=1,\dots,n}$ be given, such that for all $j \in \{1,\dots,n\}$, $\mathcal{X}_j$ is $\mathcal{F}_j$-measurable and in $L^p(\mathbb{P})$ for some $p \geq 2$. Then, for*

$$b_{j,n}(p) \coloneqq \Big(\mathbb{E}\Big[\Big(|\mathcal{X}_j| \sum_{k=j}^{n} |\mathbb{E}[\mathcal{X}_k|\mathcal{F}_j]|\Big)^{p/2}\Big]\Big)^{2/p}, \quad j = 1,\dots,n,$$

we have

$$\Big\| \sum_{j=1}^{n} \mathcal{X}_j \Big\|_{L^p(\mathbb{P})} \leq \Big(2p \sum_{j=1}^{n} b_{j,n}(p)\Big)^{1/2}.$$

Let $\Delta = \Delta_n \in (0,1]$ be some fixed sampling distance, and suppose that we have partial observations $(X_{\Delta k})_{k=1,\dots,n}$ of the subexponentially ergodic diffusion process $\mathbf{X}$ satisfying the coefficient assumptions from Section 2. Recall that

$$\mathbb{G}_{n,\Delta}(f) = \frac{1}{\sqrt{n\Delta}} \sum_{k=1}^{n} f(X_{k\Delta})\Delta$$

denotes the discretized version of the scaled additive functional $\mathbb{G}_{n\Delta}(f)$. Then, for fixed $f = \widetilde{f} - \mu(\widetilde{f})$, we may write

$$\mathbb{G}_{n,\Delta}(f) = \mathbb{G}_{n\Delta}(f) + \frac{1}{\sqrt{n\Delta}} \mathbb{A}_{n,\Delta}, \tag{26}$$

with discretization error

$$\mathbb{A}_{n,\Delta} \coloneqq \sum_{k=1}^{n} \int_{(k-1)\Delta}^{k\Delta} (\widetilde{f}(X_{k\Delta}) - \widetilde{f}(X_t))\,\mathrm{d}t.$$

With our results from Section 3, it is now clear that we must analyze the concentration of $\mathbb{A}_{n,\Delta}$ around 0 to obtain concentration inequalities for the discrete additive functional $\mathbb{G}_{n,\Delta}(f)$. To do so for unbounded functionals $\widetilde{f}$, we exploit polynomial f-norm convergence from Proposition 1.

Theorem 11 *Let $\eta_1, \eta_2, \eta_3 \geq 0$ and $\widetilde{f} \in \mathcal{G}(\eta_1, \mathfrak{L}) \cap \mathcal{W}^{2,p}_{\mathrm{loc}}(\mathbb{R}^d)$, $p \geq d$, with $\nabla \widetilde{f} \in L^{2d}_{\mathrm{loc}}(\mathbb{R}^d)$ such that $\|\nabla \widetilde{f}(x)\| \lesssim 1 + \|x\|^{\eta_2}$ and, for all $i,j = 1,\dots,n$, $|\partial_{x_i,x_j} \widetilde{f}(x)| \lesssim 1 + \|x\|^{\eta_3}$. Define $\alpha = \alpha(q', \eta_2, \eta_3) \coloneqq (q' + \eta_2) \vee \eta_3$. In case $q > -1$, let $\widetilde{\gamma} > 1 + q$, $r \geq 1$ such that $\widetilde{\gamma} - (1+q) > r(\alpha \vee (1+q)/(r-1))$. If $q = -1$, set $\widetilde{\gamma} = \alpha$. Then, for $f = \widetilde{f} - \mu(\widetilde{f})$, there exists a constant $\mathfrak{D}$ that is independent of n, p, Δ such that, for any $p \geq 2$,*

$$\|\mathbb{G}_{n,\Delta}(f)\|_{L^p(\mathbb{P}^\mu)} \leq \mathfrak{D}\Big(\sqrt{n}\Delta^{3/2} + \Delta p^{\frac{\max\{(\widetilde{\gamma}+2\alpha+1-q_+)/2, \eta_2+1-q_+\}}{1-q_+}} + p^{\frac{1}{2}+\frac{\eta+q'+q+1}{1-q_+}}\Big) =: \Phi(n,\Delta,p).$$

Consequently,

$$\mathbb{P}^\mu(|\mathbb{G}_{n,\Delta}(f)| > \mathrm{e}\Phi(n,\Delta,u)) \leq \mathrm{e}^{-u}, \quad u \geq 2.$$

Proof Let us write $t_k = k\Delta$ and $a \lesssim b$ if $a \le Cb$ for some constant C independent of p, n, Δ. The Itō–Krylov formula gives for $t \in [0, t_k]$

$$\begin{aligned}\widetilde{f}(X_{k\Delta}) - \widetilde{f}(X_t) &= \int_t^{t_k} L\widetilde{f}(X_s)\,\mathrm{d}s + \int_t^{t_k} \nabla\widetilde{f}(X_s)^\top \sigma(X_s)\,\mathrm{d}W_s \\ &= \mu(L\widetilde{f})(t - t_k) + \Phi_k(t) + \omega_k(t),\end{aligned}$$

where $\Phi_k(t) \coloneqq \int_t^{t_k} \phi(X_s)\,\mathrm{d}s$ for $\phi(y) = L\widetilde{f}(y) - \mu(L\widetilde{f})$ and $\omega_k(t) \coloneqq \int_t^{t_k} \nabla\widetilde{f}(X_s)^\top \sigma(X_s)\,\mathrm{d}W_s$. Thus, setting $\mathcal{X}_k = \int_{t_{k-1}}^{t_k} \Phi_k(t)\,\mathrm{d}t$ and $\chi_k = \int_{t_{k-1}}^{t_k} \omega_k(t)\,\mathrm{d}t$, we have

$$\mathbb{A}_{n,\Delta} = \mu(L\widetilde{f})\frac{n\Delta^2}{2} + \sum_{k=1}^n \mathcal{X}_k + \sum_{k=1}^n \chi_k. \tag{27}$$

The polynomial bounds on the gradient and the Hessian of $\widetilde{f}$ together with $\sup_{x\in\mathbb{R}^d}\|\sigma(x)\| < \infty$ and $\|b(x)\| \lesssim 1 + \|x\|^{q'}$ imply that $|L\widetilde{f}(x)| \lesssim 1 + \|x\|^\alpha$ for $\alpha \coloneqq (q' + \eta_2) \vee \eta_3$. Suppose first $q > -1$, and let $\widetilde{\gamma} > 1 + q$ and $r > 1$ such that $\widetilde{\gamma} - (1+q) > r(\alpha \vee (1+q)/(r-1))$. This implies that $(\widetilde{\gamma} - (1+q))/r > \alpha$ and $(\widetilde{\gamma} - (1+q))/(s(1+q)) > 1$ for $s = r/(r-1)$. Thus, if we choose the inverse Young functions $\Psi_1(x) = (sx)^{1/s}$ and $\Psi_2(x) = (rx)^{1/r}$, it follows for $f_{\widetilde{\gamma},q}(x) = \|x\|^{\widetilde{\gamma}-(1+q)}$ that $|L\widetilde{f}| \lesssim \mathbf{1} \vee \Psi_2 \circ f_{\widetilde{\gamma},q}$. Proposition 1 then yields

$$\|\mathbb{P}^x(X_t \in \cdot) - \mu\|_{\mathbf{1}\vee\Psi_2\circ f_{\widetilde{\gamma},q}} \lesssim \widetilde{V}_{\widetilde{\gamma}}(x)(1+t)^{-(\widetilde{\gamma}-(1+q))/(s(1+q))}, \quad x \in \mathbb{R}^d, t \ge 0. \tag{28}$$

Let $\mathbb{F} = (\mathcal{F}_t)_{t\ge0}$ be the natural filtration of $(X_t)_{t\ge0}$. Then, using the Markov property and (28), we obtain for $t > u$

$$\begin{aligned}|\mathbb{E}^\mu[\phi(X_t)|\mathcal{F}_u]| &= |\mathbb{E}^{X_u}[\phi(X_{t-u})]| = |\mathbb{E}^{X_u}[L\widetilde{f}(X_{t-u})] - \mu(L\widetilde{f})| \\ &\lesssim \|\mathbb{P}^{X_u}(X_{t-u} \in \cdot) - \mu\|_{\mathbf{1}\vee\Psi_2\circ f_{\widetilde{\gamma},q}} \lesssim \widetilde{V}_{\widetilde{\gamma}}(X_u)(1 + (t-u))^{-(\widetilde{\gamma}-(1+q))/(s(1+q))}.\end{aligned}$$

For $k > j$, this gives

$$\begin{aligned}\mathbb{E}^\mu[\mathcal{X}_k|\mathcal{F}_{t_j}] &\lesssim \widetilde{V}_{\widetilde{\gamma}}(X_{t_j}) \int_{t_{k-1}}^{t_k}\int_t^{t_k} (1 + (u - t_j))^{-(\widetilde{\gamma}-(1+q))/(s(1+q))}\,\mathrm{d}u\,\mathrm{d}t \\ &\le \widetilde{V}_{\widetilde{\gamma}}(X_{t_j})\Delta^2(1 + (k-1-j)\Delta)^{-(\widetilde{\gamma}-(1+q))/(s(1+q))}.\end{aligned}$$

Hence, for $j < n$,

$$\sum_{k=j+1}^n |\mathbb{E}^\mu[\mathcal{X}_k|\mathcal{F}_{t_j}]| \lesssim \widetilde{V}_{\widetilde{\gamma}}(X_{t_j})\Delta^2 \int_0^\infty (1 + \Delta t)^{-(\widetilde{\gamma}-(1+q))/(s(1+q))}\,\mathrm{d}t = \widetilde{V}_{\widetilde{\gamma}}(X_{t_j})\Delta\tfrac{s(1+q)}{\widetilde{\gamma}-(1+q)(1+s)},$$

where we used that $(\widetilde{\gamma}-(1+q))/(s(1+q)) > 1$. Consequently, letting $b_{j,n}(p)$ be the functional from Lemma 10, it follows for $j < n$ from the Cauchy–Schwarz inequality, stationarity and Lemma 4

$$b_{j,n}(p) \le \|\mathcal{X}_j\|^2_{L^{2p}(\mathbb{P}^\mu)} \Big\| \sum_{k=j+1}^n |\mathbb{E}^\mu[\mathcal{X}_k|\mathcal{F}_{t_j}]| \Big\|_{L^p(\mathbb{P}^\mu)} \lesssim \Delta\|\mathcal{X}_j\|^2_{L^{2p}(\mathbb{P}^\mu)} \|\widetilde{V}_{\widetilde{\gamma}}(X_0)\|_{L^p(\mathbb{P}^\mu)}$$

$$\lesssim \Delta p^{\frac{\widetilde{\gamma}}{1-q_+}} \|\mathcal{X}_j\|^2_{L^p(\mathbb{P}^\mu)}.$$

In case $q = -1$, we simply observe that Proposition 1 implies that there exists $\beta > 0$ such that $\|\mathbb{P}^x(X_t \in \cdot) - \mu\|_{\widetilde{V}_\alpha} \lesssim \widetilde{V}_\alpha(x)\exp(-\beta t)$ and hence, proceeding as above, we end up with

$$b_{j,n}(p) \lesssim \Delta \|\mathcal{X}_j\|^2_{L^{2p}(\mathbb{P}^\mu)} \|\widetilde{V}_\alpha(X_0)\|_{L^p(\mathbb{P}^\mu)} \lesssim \Delta p^{\frac{\alpha}{1-q_+}} \|\mathcal{X}_j\|^2_{L^p(\mathbb{P}^\mu)}.$$

Now, stationarity under $\mathbb{P}^\mu$, Hölder's inequality together with Fubini and Lemma 4 yield

$$\begin{aligned}\|\mathcal{X}_j\|^p_{L^p(\mu)} &= \mathbb{E}^\mu\Big[\Big(\int_0^\Delta \int_t^\Delta \phi(X_s)\,\mathrm{d}s\,\mathrm{d}t\Big)^p\Big] \leq \Delta^{2(p-1)} \int_0^\Delta \int_t^\Delta \mathbb{E}^\mu\big[|\phi(X_s)|^p\big]\,\mathrm{d}s\,\mathrm{d}t \\ &= \Delta^{2p}\|\phi(X_0)\|^p_{L^p(\mathbb{P}^\mu)} \leq c^p \Delta^{2p} p^{p\alpha/(1-q_+)},\end{aligned}$$

for some constant $c > 0$. Thus, we obtain

$$b_{j,n}(p) \lesssim \Delta^3 p^{(\widetilde{\gamma}+2\alpha)/(1-q_+)},$$

and hence by Lemma 10

$$\Big\|\sum_{k=1}^n \mathcal{X}_k\Big\|_{L^p(\mathbb{P}^\mu)} \lesssim \sqrt{n\Delta^3} p^{(\widetilde{\gamma}+2\alpha+1-q_+)/(2(1-q))}. \tag{29}$$

Let us now treat $\sum_{k=1}^n \chi_k$. As in the proof of Lemma 5, we obtain by the Burkholder–Davis–Gundy inequality, (Barlow and Yor, 1982, Proposition 4.2) and Lemma 4 that

$$\mathbb{E}^\mu[|\omega_k|^p] \leq c^p p^{p/2} \lambda_+^p \Delta^{p/2} \|\nabla \widetilde{f}\|^p_{L^p(\mu)} \leq C(\eta_2)^p \lambda_+^p \Delta^{p/2} p^{p/2+p\eta_2/(1-q_+)}.$$

Therefore, with Hölder's inequality,

$$\mathbb{E}^\mu[|\chi_k|^p] \leq \Delta^{p-1} \int_{t_{k-1}}^{t_k} \mathbb{E}^\mu[|\omega_k(t)|^p]\,\mathrm{d}t \leq C(\eta_2)^p \lambda_+^p \Delta^{3p/2} p^{p/2+p\eta_2/(1-q_+)}.$$

Let $b^\chi_{n,p}$ be the functional from Lemma 10 with respect to $(\chi_k)_{k=1,\dots,n}$ and $(\mathcal{F}_{t_k})_{k=1,\dots,n}$, and note that, for $k > j$ and $t \in [t_j, t_k]$, $\mathbb{E}^\mu[\omega_k(t)|\mathcal{F}_{t_j}] = 0$ since $(\int_0^t \nabla \widetilde{f}(X_s)^\top \sigma(X_s)\,\mathrm{d}W_s)_{t\geq 0}$ is an $\mathbb{F}$-martingale. Thus,

$$b^\chi_{j,n}(p) = \mathbb{E}^\mu[|\chi_j|^p]^{2/p} \lesssim \Delta^3 p^{(2\eta_2+(1-q_+))/(1-q_+)}.$$

Consequently, by Lemma 10,

$$\Big\|\sum_{k=1}^n \chi_k\Big\|_{L^p(\mathbb{P}^\mu)} \lesssim \sqrt{n}\Delta^{3/2} p^{(\eta_2+1-q_+)/(1-q_+)}. \tag{30}$$

Taking into account that $\mu(|L\widetilde{f}|) < \infty$, (27), (29) and (30) imply that

$$\Big\|\frac{1}{\sqrt{n\Delta}}\mathbb{A}_{n\Delta}\Big\|_{L^p(\mathbb{P}^\mu)} \lesssim \sqrt{n}\Delta^{3/2} + \Delta p^{\frac{\max\{(\widetilde{\gamma}+2\alpha+1-q_+)/2,\eta_2+1-q_+\}}{1-q_+}}.$$

Plugging this bound into (26) and using Theorem 3, it follows that there exists some constant $\mathfrak{D}$ that is independent of n, p, Δ such that, for $p \geq 1$,

$$\Big\|\mathbb{G}_{n,\Delta}(f)\Big\|_{L^p(\mathbb{P}^\mu)} \leq \mathfrak{D}\Big(\sqrt{n}\Delta^{3/2} + \Delta p^{\frac{\max\{(\widetilde{\gamma}+2\alpha+1-q_+)/2,\eta_2+1-q_+\}}{1-q_+}} + p^{\frac{1}{2}+\frac{\eta+q'+q+1}{1-q_+}}\Big).$$

Markov's inequality now yields the asserted concentration inequality. ∎

PAC bounds Similarly to Corollary 8 and Corollary 9, we can derive PAC bounds for the discrete ergodic average and its burn-in version. The proof is identical and therefore omitted.

Corollary 12 *Let $\eta_1, \eta_2, \eta_3 \geq 0$ and $f \in \mathcal{G}(\eta_1, \mathfrak{L}) \cap \mathcal{W}^{2,p}_{\mathrm{loc}}(\mathbb{R}^d)$, $p \geq d$, with $\nabla f \in L^{2d}_{\mathrm{loc}}(\mathbb{R}^d)$ such that $\|\nabla f(x)\| \lesssim 1 + \|x\|^{\eta_2}$ and, for all $i, j = 1, \dots, d$, $|\partial_{x_i, x_j} f(x)| \lesssim 1 + \|x\|^{\eta_3}$. Define $\alpha, \widetilde{\gamma}$ as in Theorem 11, and denote*

$$\varrho = \varrho(\alpha, \eta_2, \widetilde{\gamma}, q) \coloneqq \frac{\max\{(\widetilde{\gamma} + 2\alpha + 1 - q_+)/2, \eta_2 + 1 - q_+\}}{1 - q_+}$$

and

$$\widetilde{\varsigma} = \widetilde{\varsigma}(\eta_1, q, q') \coloneqq \frac{1}{2} + \frac{\eta + q' + q + 1}{1 - q_+}.$$

For $\varepsilon > 0$, $\delta \in (0, \mathrm{e}^{-2})$, suppose that $\Delta < \varepsilon/(3\mathrm{e}\mathfrak{D})$ and

$$n \geq \Psi(\Delta, \varepsilon, \delta) \coloneqq \frac{1}{\Delta} \left(\frac{3\mathrm{e}\mathfrak{D} \max\left\{ \Delta(\log(1/\delta))^{\varrho}, (\log(1/\delta))^{\widetilde{\varsigma}} \right\}}{\varepsilon} \right)^2.$$

Then,

$$\mathbb{P}^{\mu}\Big(\Big| \frac{1}{n} \sum_{k=1}^{n} f(X_{k\Delta}) - \mu(f) \Big| \leq \varepsilon \Big) \geq 1 - \delta.$$

Moreover, let the discrete burn-in estimator be given by

$$\mathbb{H}_{m,n,\Delta}(f) \coloneqq \frac{1}{\sqrt{n\Delta}} \mathbb{G}_{n,\Delta}(f) \circ \theta_{m\Delta} = \frac{1}{n} \sum_{k=m+1}^{n+m} f(X_{k\Delta}).$$

Then, given the constants ι'', C from Corollary 9 and some initial distribution ν such that $V_{q_+} \in L^1(\nu)$, for any $n \geq \Psi(\Delta, \varepsilon, \delta/2)$ and burn-in length $m \geq 1 \vee \Delta^{-1}(\log(2C/\delta))^{\frac{1+q_+}{1-q_+}}/\iota''$, it holds

$$\mathbb{P}^{\nu}\Big(|\mathbb{H}_{m,n,\Delta}(f) - \mu(f)| \leq \varepsilon \Big) \geq 1 - \delta.$$

4. Applications

We now demonstrate the usefulness of our probabilistic results in two concrete applications. While exponential inequalities are important for a multitude of statistical problems (e.g., in the context of adaptive nonparametric estimation or for the verification of uniform convergence results), we will focus in Section 4.1 on the analysis of a high-dimensional diffusion model under sparsity constraints, which in particular necessitates the use of inequalities for *unbounded* functions. Specifically, we will see that Theorem 3 allows us to derive non-asymptotic error bounds for penalised estimators, which, to the best of our knowledge, are so far only available for Ornstein–Uhlenbeck processes. In Section 4.2, we use our discrete concentration results from Section 3.2 to derive explicit convergence guarantees for an MCMC algorithm designed to sample from target densities with subexponential tails.

4.1 Lasso estimation for parametrized drift coefficients

As opposed to the now very well understood high-dimensional discrete models (cf., e.g., Bühlmann and van de Geer (2011) or Wainwright (2019)), for which a wealth of estimation algorithms including corresponding theoretical results are available, there are still few in-depth studies of estimation problems for high-dimensional continuous-time processes. Important references in this context are Gaïffas and Matulewicz (2019) and Ciołek et al. (2020), who investigate drift estimation in a high-dimensional Ornstein–Uhlenbeck (OU) model under sparsity constraints. A remarkable feature is that the restricted eigenvalue property, which usually has to be verified explicitly in discrete models such as linear regression, is already implied by the ergodicity in the specified diffusion model. This finding is based on the use of sufficiently sharp probabilistic tools in the form of concentration inequalities suited to the model: while Gaïffas and Matulewicz (2019) provide a proof based on functional inequalities allowing to cover only the reversible case, Ciołek et al. (2020) use Malliavin calculus methods to show that the restricted eigenvalue property is satisfied in the general ergodic OU case. At the same time, they point out (cf. their Remark 4.4) that other mathematical methods are needed for proving such concentration phenomena in more general diffusion models.

Motivated by the considerations in Pokern et al. (2009), we outline in this section how our results from Section 3.1 can be used to study more general high-dimensional diffusion models. Suppose that the data $X^T = (X_t)_{0\leq t\leq T}$ has been generated by the following Itō SDE,

$$\mathrm{d}X_t = b_0(X_t)\,\mathrm{d}t + \sigma_0(X_t)\,\mathrm{d}W_t, \tag{31}$$

$W = (W_t)_{t\geq 0}$ a standard d-dimensional Brownian motion. The diffusion matrix σ_0 is assumed to be known and we wish to estimate the drift vector b_0. Suppose that both σ_0 and b_0 are globally Lipschitz, that σ_0 is bounded, that $a_0 \coloneqq \sigma_0\sigma_0^\top$ is uniformly elliptic, i.e.,

$$\exists \lambda_-, \lambda_+ > 0\ \forall x, \eta \in \mathbb{R}^d: \quad \lambda_-\|\eta\|^2 \leq \langle \eta, a_0(x)\eta\rangle \leq \lambda_+\|\eta\|^2,$$

and that the drift condition

($\mathscr{L}(q)$) there exists $M_0, \mathfrak{r} > 0$ such that

$$\forall \|x\| > M_0: \langle b_0(x), x/\|x\|\rangle \leq -\mathfrak{r}\|x\|^{-q}$$

is satisfied for some $q \in [-1, 1)$, such that the process falls into the ergodic framework of Section 2. Denote by μ_0 and ρ_0 its invariant measure and the corresponding invariant density, respectively. We further assume that X^T is the stationary solution, i.e., $X_0 \sim \mu_0$, and denote $\mathbb{P} \coloneqq \mathbb{P}^{\mu}$.

Denote by $\mathbf{P}_b$ the law of Y^T, where $Y^T = (Y_t)_{0\leq t\leq T}$ is the strong solution to the SDE $\mathrm{d}Y_t = b(Y_t)\,\mathrm{d}t + \sigma_0(Y_t)\,\mathrm{d}W_t$, $Y_0 = X_0$. Then, the Radon–Nikodym derivative of $\mathbf{P}_b$ with respect to $\mathbf{P}_0$ is given as

$$\frac{d\mathbf{P}_b}{d\mathbf{P}_0}(X^T) = \exp\left(-\frac{1}{2}\int_0^T b^\top(X_t)a_0^{-1}(X_t)b(X_t)\,\mathrm{d}t + \int_0^T b^\top(X_t)a_0^{-1}(X_t)\,\mathrm{d}X_t\right)$$

(see (Liptser and Shiryaev, 2001, Section 7.6.4)). Given the data X^T, one can derive the time scaled negative of the log likelihood functional for the unknown drift b. This functional is given by

$$\mathcal{L}_T(b) = \frac{1}{2T}\int_0^T \Big(b^\top(X_t)a_0^{-1}(X_t)b(X_t)\,\mathrm{d}t - 2b^\top(X_t)a_0^{-1}(X_t)\,\mathrm{d}X_t\Big). \tag{32}$$

The log likelihood function (32) for b is unbounded from below in general if the data is finite, $T<\infty$. However, letting $T\to\infty$, (32) tends to a functional whose unique minimizer is b_0. More precisely, it is shown in Lemma 6.1 in Pokern et al. (2009) that $\mathcal{L}_T(b)$ converges a.s. towards the functional

$$\mathcal{L}_\infty(b) = \int_{\mathbb{R}^d}\Big(\tfrac{1}{2}b^\top(x)a_0^{-1}(x)(b(x)-b_0(x))\Big)\rho_0(x)\,\mathrm{d}x.$$

In order to regularize (32), Pokern et al. (2009) suggest to assume a *parametric* structure of the drift coefficient. For the class of generalised OU processes fulfilling the linear SDE

$$\mathrm{d}X_t = -\boldsymbol{A}X_t\,\mathrm{d}t + \sigma\,\mathrm{d}W_t,\quad t\ge 0, \tag{33}$$

$\boldsymbol{A}$ and σ some $d\times d$-matrices and W a d-dimensional Brownian motion, this assumption is obviously satisfied. A more general, but still treatable class of processes is obtained as follows: Given a system $(\psi_j)_{1\le j\le N}$ of Lipschitz continuous basis functions $\psi_j\colon\mathbb{R}^d\to\mathbb{R}^d$, introduce

$$\mathcal{V} \coloneqq \Big\{b_\theta(\cdot)=\sum_{j=1}^N\theta_j\psi_j(\cdot),\ \theta\in\mathbb{R}^N\Big\}.$$

Let $\boldsymbol{\psi}(\cdot)\coloneqq(\psi_1(\cdot),\dots,\psi_N(\cdot))$ be the dictionary matrix and define for $x\in\mathbb{R}^d$, $\boldsymbol{\Psi}(x)\coloneqq(\sigma_0^{-1}(x)\boldsymbol{\psi}(x))^\top\sigma_0^{-1}(x)\boldsymbol{\psi}(x)$. Let us also define the matrices

$$\overline{\boldsymbol{\Psi}}_T \coloneqq \frac{1}{T}\int_0^T\boldsymbol{\Psi}(X_s)\,\mathrm{d}s = (\overline{\psi}_{ij,T})_{1\le i,j\le N}\quad\text{and}\quad \overline{\boldsymbol{\Psi}}_\infty\coloneqq\mathbb{E}[\boldsymbol{\Psi}(X_0)] = (\overline{\psi}_{ij,\infty})_{1\le i,j\le N}$$

with entries

$$\begin{aligned}\overline{\psi}_{ij,T} &\coloneqq \frac{1}{T}\int_0^T\big\langle\psi_i(X_s),a_0^{-1}(X_s)\psi_j(X_s)\big\rangle\,\mathrm{d}s,\\ \overline{\psi}_{ij,\infty} &\coloneqq \int_{\mathbb{R}^d}\big\langle\psi_i(x),a_0^{-1}(x)\psi_j(x)\big\rangle\rho_0(x)\,\mathrm{d}x,\quad i,j=1,\dots,N.\end{aligned}$$

We impose the following assumptions on the dictionary:

($\mathscr{L}1$) There exist $\mathfrak{L}>0$ and $\eta\in[0,1]$ such that the maximal eigenvalue of $\boldsymbol{\Psi}(x)$ satisfies

$$\lambda_{\max}(\boldsymbol{\Psi}(x))\le\mathfrak{L}(1+\|x\|^{2\eta}),\quad x\in\mathbb{R}^d;$$

($\mathscr{L}2$) the random matrix $\overline{\boldsymbol{\Psi}}_T$ is positive definite $\mathbb{P}$-a.s.

Assumption ($\mathscr{L}1$) allows a maximal polynomial drift of order η of the basis functions, where $\eta \in [0,1]$ is consistent with their assumed Lipschitz continuity. Assumption ($\mathscr{L}2$) is a necessary technical condition on the positive semidefinite matrix $\overline{\mathbf{\Psi}}_T$ that we need for the penalized MLE to be well-defined. It can be verified given sufficient smoothness of the dictionary, see Example 1. Moreover, since by stationarity $\overline{\mathbf{\Psi}}_\infty = \mathbb{E}[\overline{\mathbf{\Psi}}_T]$, ($\mathscr{L}2$) implies that $\overline{\mathbf{\Psi}}_\infty$ is positive definite. Therefore, if we denote the minimal eigenvalue of $\overline{\mathbf{\Psi}}_\infty$ by $\lambda_{\min}(\overline{\mathbf{\Psi}}_\infty) =: \mathfrak{e}_\infty$, then $\mathfrak{e}_\infty > 0$. Let us also set $\mathfrak{D}_\infty := \max_{i=1,\dots,N} \overline{\psi}_{ii,\infty}$.

We now give an example of a dictionary that satisfies the above assumptions and can be used to model drifts satisfying the drift condition ($\mathscr{L}(q)$).

Example 1 *Let*

$$E_i = \mathbf{1}_{1+\lfloor\!\lfloor (i-d^2\lfloor\!\lfloor i/d^2\rfloor\!\rfloor)/d\rfloor\!\rfloor, 1+(i-1) \bmod d}, \quad i = 1,\dots,nd^2,$$

where $\mathbf{1}_{k,l}$ *is the* $d \times d$ *matrix whose* (k,l)*-th entry is* 1 *and all other entries are* 0*, and* $\lfloor\!\lfloor x \rfloor\!\rfloor = \max\{z \in \mathbb{Z} : z < x\}$*. Set then, for* $\widetilde{q}_i \in [-1,1)$ *and* $\widetilde{\alpha}_i > 0$*,*

$$\psi_i(x) = E_i x(\widetilde{\alpha}_i + \|x\|)^{-(\widetilde{q}_i+1)}, \quad \text{where } \widetilde{q}_i = \widetilde{q}_j \text{ and } \widetilde{\alpha}_i = \widetilde{\alpha}_j \text{ if } \lfloor\!\lfloor i/d^2 \rfloor\!\rfloor = \lfloor\!\lfloor j/d^2 \rfloor\!\rfloor,$$

which is nothing else but saying that any $b \in \mathcal{V}$ *can be written as*

$$b_\theta(x) := \sum_{i=1}^N \theta_i \psi_i(x) = \sum_{i=1}^n A_i(\theta) x(\alpha_i + \|x\|)^{-(q_i+1)}, \quad x \in \mathbb{R}^d,$$

where $N = nd^2$*,*

$$(A_i(\theta))_{k,l} = \theta_{(i-1)d^2+(k-1)d+l}, \quad i = 1,\dots,n \text{ and } k,l = 1,\dots,d,$$

and $q_i = \widetilde{q}_{1+d^2\lfloor\!\lfloor i/d^2\rfloor\!\rfloor}, \alpha_i = \widetilde{\alpha}_{1+d^2\lfloor\!\lfloor i/d^2\rfloor\!\rfloor}$*. Suppose that* $q_i < q_j$ *for* $i > j$*, the matrices* $A_i(\theta_0)$ *corresponding to the true value* θ_0 *are symmetric, and that there exists* $k_0 \in \{1,\dots,n\}$ *such that* $\lambda_{\max}(A_{k_0}(\theta_0)) < 0$ *and, for all* $k_0 < k \le n$*, it holds* $\lambda_{\max}(A_k(\theta_0)) = 0$*. Then, it follows from the Courant–Fischer theorem that, for any* $x \neq 0$*,*

$$\begin{aligned}\langle b_{\theta_0}(x), x/\|x\|\rangle &= \sum_{i=1}^n \|x\|(\alpha_i + \|x\|)^{-(1+q_i)} \langle x/\|x\|, A_i(\theta_0) x/\|x\|\rangle \\ &\le \sum_{i=1}^{k_0} \|x\|(\alpha_i + \|x\|)^{-(1+q_i)} \lambda_{\max}(A_i(\theta_0)).\end{aligned}$$

This implies that there exists $M_0, c > 0$ *such that, for* $\mathfrak{r} = -c\lambda_{\max}(A_{k_0}(\theta_0)) > 0$*, the drift condition* ($\mathscr{L}(q)$) *is satisfied for* $q = q_{k_0}$*. Let* μ *be the invariant distribution of the associated diffusion* $\mathbf{X}$*. Also note that* ($\mathscr{L}1$) *holds for* $\eta = (-q_1)_+$*.*

To see that ($\mathscr{L}2$) *is satisfied, note first that, for any* $\theta \neq 0$*, there exists some* $j \in \{1,\dots,d\}$ *such that* $x \mapsto (b_\theta(x))_j$ *is analytic and not identical to zero on* $\mathbb{R}^d \setminus \{0\}$*. Consequently,* $(b_\theta(\cdot)_j)^{-1}(\{0\}) \setminus \{0\} = \{x \neq 0 : (\boldsymbol{\psi}(x)\theta)_j = 0\}$ *is contained in a countable union of smooth manifolds of dimension* $d-1$*, i.e., in a countable union of smooth hypersurfaces. Assume now that* $\overline{\mathbf{\Psi}}_T$ *is not positive definite a.s. Since* $\theta^\top \overline{\mathbf{\Psi}}_T \theta = \int_0^T \|(\sigma_0^{-1}\boldsymbol{\psi})(X_s)\theta\|^2 \,\mathrm{d}s$

and, moreover, the matrix is positive semidefinite, the paths of $\mathbf{X}$ *are continuous and* $\mathbb{P}^\mu(X_0 = 0) = 0$*, this implies that there exists a measurable set* $\Omega_0 \subset \{X_0 \neq 0\}$ *with* $\mathbb{P}^\mu(\Omega_0) > 0$ *such that, for any* $\omega \in \Omega_0$*, the whole path* $(X_s(\omega))_{s\in[0,T]}$ *is contained in* $(b_\theta(\cdot)_j)^{-1}(\{0\})$ *for some* $\theta \neq 0$ *and* $j \in \{1, \dots, d\}$*. It follows from above that on* Ω_0*, the process stays in some smooth hypersurface for a strictly positive amount of time. Such path behaviour is however impossible a.s. for an elliptic diffusion process. Thus,* $\overline{\mathbf{\Psi}}_T$ *must be positive definite* $\mathbb{P}$*-a.s.*

Note that the above example includes the OU models investigated in Gaïffas and Matulewicz (2019); Ciołek et al. (2020) as a special case. Under $\mathbf{P}_{b_0}$, the above parametrisation yields the functional

$$\begin{aligned}
&\mathcal{L}_T(\theta)\\
&= -\frac{1}{2T}\Big(2\int_0^T \big(\sigma_0^{-1}b_\theta\big)^\top(X_t)\,\mathrm{d}W_t - \int_0^T \big\|\sigma_0^{-1}(b_\theta - b_{\theta_0})(X_t)\big\|^2\,\mathrm{d}t + \int_0^T \big\|\big(\sigma_0^{-1}b_{\theta_0}\big)(X_t)\big\|^2\,\mathrm{d}t\Big)\\
&= \frac{1}{2}\theta^\top\overline{\mathbf{\Psi}}_T\theta - \theta^\top\overline{h},
\end{aligned}$$

$\overline{h}$ denoting the vector with components

$$\overline{h}_i = \frac{1}{T}\int_0^T \big\langle \psi_i(X_s), a_0^{-1}(X_s)\,\mathrm{d}X_s\big\rangle, \quad i = 1, \dots, N.$$

Using almost sure positive definiteness of $\overline{\mathbf{\Psi}}_T$, it follows that on a set of full $\mathbb{P}$-measure, the MLE is the unique minimizer of $\mathcal{L}_T(\cdot)$, given by

$$\widehat{\theta}_{\mathrm{MLE}} \coloneqq \overline{\mathbf{\Psi}}_T^{-1}\overline{h}.$$

While this approach yields a well-defined estimator, the MLE will perform quite inaccurately in high-dimensional settings.

Our concern is to investigate the estimation of b_θ in the large N/large T regime. More precisely, we want to study the statistical properties of penalized estimators $\widehat{\theta}_T$, defined as

$$\widehat{\theta}_T = \arg\min_{\theta\in\mathbb{R}^N}\{\mathcal{L}_T(\theta) + \lambda\|\theta\|_1\}, \tag{34}$$

$\lambda > 0$ some regularisation parameter. Strictly speaking, since positive definiteness of $\overline{\mathbf{\Psi}}_T$ holds only a.s., this estimator may only be well defined in an almost sure sense, but by an appropriate restriction of the underlying probability space we can and will assume that it is well-defined everywhere without loss of generality. Denote

$$\|\theta_1 - \theta_2\|_{L^2}^2 \coloneqq \frac{1}{T}\int_0^T \big\|\sigma_0^{-1}(b_{\theta_1} - b_{\theta_2})(X_t)\big\|^2\,\mathrm{d}t = (\theta_1 - \theta_2)^\top\overline{\mathbf{\Psi}}_T(\theta_1 - \theta_2), \quad \theta_1, \theta_2 \in \mathbb{R}^N.$$

Then, for any $\theta \in \mathbb{R}^N$,

$$\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2}^2 \leq \big\|\theta - \theta_0\big\|_{L^2}^2 + \frac{2}{T}\int_0^T \Big(\sigma_0^{-1}\Big(b_{\widehat{\theta}_T} - b_\theta\Big)\Big)^\top(X_t)\,\mathrm{d}W_t + 2\lambda\Big(\|\theta\|_1 - \big\|\widehat{\theta}_T\big\|_1\Big). \tag{35}$$

In order to obtain error bounds for the Lasso estimator $\widehat{\theta}_T$, the martingale part appearing on the rhs of (35) needs to be controlled which is usually done by means of Bernstein's inequality for continuous martingales. Another important part of the derivation of error bounds is the verification of the restricted eigenvalue condition which in our setting amounts in showing that

$$\inf_{\theta\in\mathcal{S}_1(s),\eta\in\mathcal{S}_2(s,\theta)} \frac{\|\theta-\eta\|_{L^2}^2}{\|\theta-\eta\|^2} \quad \text{is bounded away from 0 with high probability,}$$

where, for $\|\theta\|_0 \coloneqq \sum_i \mathbf{1}_{\{\theta_i\neq 0\}}$, fixed $c_0 > 0$ and $\mathcal{I}_s(\theta)$ denoting a set of coordinates of s largest elements of θ,

$$\begin{aligned}\mathcal{C}(s,c_0) &\coloneqq \Big\{\zeta\in\mathbb{R}^N : \|\zeta\|_1 \leq (1+c_0)\big\|\zeta_{|\mathcal{I}_s(\zeta)}\big\|_1\Big\},\\ \mathcal{S}_1(s) &\coloneqq \big\{\theta\in\mathbb{R}^N : \|\theta\|_0 = s\big\} \quad \text{and} \quad \mathcal{S}_2(s,\theta) \coloneqq \big\{\eta\in\mathbb{R}^N : \theta-\eta\in\mathcal{C}(s,c_0)\big\}.\end{aligned}$$

To start with, we will demonstrate how our previous general developments can be used to verify these assumptions. In fact, our error bounds for the Lasso estimator formulated below are based on the following direct application of Theorem 3.

Lemma 13 *There exists a constant $\mathfrak{W} > 0$ such that, for any vectors $\zeta\in\mathbb{R}^N$ with $\|\zeta\|\leq 1$ and $R \geq 2/\sqrt{T}$,*

$$\mathbb{P}\Big(\big|\zeta^\top\big(\overline{\mathbf{\Psi}}_\infty - \overline{\mathbf{\Psi}}_T\big)\zeta\big| > R\Big) \leq \exp\left(-\left(\frac{\sqrt{T}R}{\mathrm{e}\mathfrak{L}\mathfrak{W}}\right)^{\kappa(q,\eta)}\right), \quad \textit{where } \kappa(q,\eta) \coloneqq \frac{2(1-q_+)}{6\eta+2q+3-q_+}. \tag{36}$$

Proof Observe first that it suffices to prove the lemma for $\|\zeta\| = 1$. Fix any such ζ and set $\widetilde{f}_\zeta(x) = \zeta^\top\mathbf{\Psi}(x)\zeta$ and $f_\zeta = \widetilde{f}_\zeta - \mu_0(\widetilde{f}_\zeta)$. By assumption $(\mathscr{L}1)$, we have for any $x\in\mathbb{R}^d$

$$|\widetilde{f}_\zeta(x)| = \|\sigma_0^{-1}(x)\boldsymbol{\psi}(x)\zeta\|^2 \leq \|\sigma_0^{-1}(x)\boldsymbol{\psi}(x)\|^2 = \lambda_{\max}(\mathbf{\Psi}(x)) \leq \mathfrak{L}(1+\|x\|^{2\eta}).$$

Moreover, using $\|\sigma_0(x)\| = \|\sigma_0(x)^\top\|$,

$$\begin{aligned}\max_{i=1,\dots,N}\|\psi_i(x)\| &\leq \sqrt{\lambda_+}\max_{i=1,\dots,n}\|\sigma_0^{-1}(x)\psi_i(x)\| = \sqrt{\lambda_+}\max_{i=1,\dots,N}\|\sigma_0^{-1}(x)\boldsymbol{\psi}(x)e_i\|\\ &\leq \sqrt{\lambda_+}\|\sigma_0^{-1}(x)\boldsymbol{\psi}(x)\| \leq \sqrt{\mathfrak{L}\lambda_+}(1+\|x\|^\eta),\end{aligned}$$

such that $\|b_{\theta_0}(x)\| \leq \sqrt{\mathfrak{L}\lambda_+}\|\theta_0\|_1(1+\|x\|^\eta)$ follows. Consequently, Theorem 3 implies that there exists some constant $\mathfrak{W}$ independent of ζ such that

$$\mathbb{P}\Big(\big|\zeta^\top\big(\overline{\mathbf{\Psi}}_\infty - \overline{\mathbf{\Psi}}_T\big)\zeta\big| > R\Big) = \mathbb{P}\big(T^{-1/2}|\mathbb{G}_T(f_\zeta)| > R\big) \leq \exp\left(-\left(\frac{\sqrt{T}R}{\mathrm{e}\mathfrak{L}\mathfrak{W}}\right)^{\kappa(q,\eta)}\right).$$

■

We are now ready to verify the restricted eigenvalue property and state deviation bounds for the martingale term.

Proposition 14 *(a) For any $\varepsilon_0 \in (0,1)$ and $\forall T \geq T_0(\varepsilon_0, s, c_0, \mathfrak{LW})$, it holds*

$$\mathbb{P}\bigg(\inf_{\theta\in\mathcal{S}_1(s),\eta\in\mathcal{S}_2(s,\theta)} \frac{\|\theta-\eta\|_{L^2}^2}{\|\theta-\eta\|^2} \geq \frac{\mathfrak{e}_\infty}{2}\bigg) \geq 1-\varepsilon_0,$$

where

$$T_0(\varepsilon_0, s, c_0, c) \coloneqq \Big\{\log\Big(21^{2s}\Big(d \wedge \big(\tfrac{\mathrm{e}d}{2s}\big)^{2s}\Big)\Big) - \log\varepsilon_0\Big\}^{\frac{2}{\kappa(q,\eta)}} \cdot \frac{18^2(c_0+2)^2\mathrm{e}^2c^2}{\mathfrak{e}_\infty^2}$$

(b) For $s, c_0 > 0$, define the event

$$\begin{aligned}\mathscr{E}(s,c_0) \coloneqq \bigg\{\inf_{\theta-\eta\in\mathcal{C}(s,c_0)} \frac{\|\theta-\eta\|_{L^2}^2}{\|\theta-\eta\|^2} \geq \frac{\mathfrak{e}_\infty}{2}\bigg\} \cap \bigg\{\max_{i=1,\dots,N} \overline{\psi}_{ii,T} \leq \mathfrak{D}_\infty + \frac{\mathfrak{e}_\infty}{2}\bigg\} \\ \cap \bigg\{\sup_{\theta\neq\eta\in\mathbb{R}^N} \frac{\frac{1}{T}\int_0^T \big(\sigma_0^{-1}\big(b_\theta - b_\eta\big)\big)^\top(X_t)\,\mathrm{d}W_t}{\|\theta-\eta\|_1} \leq \frac{\lambda}{2}\bigg\}.\end{aligned} \tag{37}$$

Then, for any $\varepsilon_0 \in (0,1)$, $T \geq T_0(\frac{\varepsilon_0}{3}, s, c_0, \mathfrak{LW})$ and

$$\lambda \geq \sqrt{\frac{4(2\mathfrak{D}_\infty + \mathfrak{e}_\infty)}{T} \cdot \log\bigg(\frac{6N}{\varepsilon_0}\bigg)},$$

it holds $\mathbb{P}(\mathscr{E}(s,c_0)) \geq 1-\varepsilon_0$.

Proof Introduce $\mathcal{K}(s) \coloneqq \big\{\zeta \in \mathbb{R}^N \setminus \{0\} : \|\zeta\|_0 \leq s\big\}$. Using Lemmata F.1 and F.3 of Basu and Michailidis (2015), it follows

$$\sup_{\zeta\in\mathcal{C}(s,c_0)} \frac{\zeta^\top\big(\overline{\mathbf{\Psi}}_\infty - \overline{\mathbf{\Psi}}_T\big)\zeta}{\|\zeta\|^2} \leq 3(c_0+2) \sup_{\zeta\in\mathcal{K}(2s)} \frac{\zeta^\top\big(\overline{\mathbf{\Psi}}_\infty - \overline{\mathbf{\Psi}}_T\big)\zeta}{\|\zeta\|^2}$$

Furthermore, for any subset $E \subset \mathbb{R}^N$,

$$\inf_{\zeta\in E} \frac{\|\zeta\|_{L^2}^2}{\|\zeta\|^2} = \inf_{\zeta\in E:\|\zeta\|\leq 1} \zeta^\top\overline{\mathbf{\Psi}}_T\zeta$$

and for $\zeta \neq 0$,

$$\frac{\|\zeta\|_{L^2}^2}{\|\zeta\|^2} = \frac{\zeta^\top\overline{\mathbf{\Psi}}_\infty\zeta}{\|\zeta\|^2} - \frac{\zeta^\top\big(\overline{\mathbf{\Psi}}_\infty - \overline{\mathbf{\Psi}}_T\big)\zeta}{\|\zeta\|^2} \geq \lambda_{\min}(\overline{\mathbf{\Psi}}_\infty) - \frac{\zeta^\top\big(\overline{\mathbf{\Psi}}_\infty - \overline{\mathbf{\Psi}}_T\big)\zeta}{\|\zeta\|^2}.$$

The proof of Lemma F.2 in Basu and Michailidis (2015) allows to deduce from (36) that, for any $R \geq 2/\sqrt{T}$,

$$\mathbb{P}\bigg(\sup_{\zeta\in\mathcal{K}(s),\|\zeta\|\leq 1} \big|\zeta^\top\big(\overline{\mathbf{\Psi}}_\infty - \overline{\mathbf{\Psi}}_T\big)\zeta\big| > 3R\bigg) \leq 21^s\bigg(d \wedge \Big(\frac{\mathrm{e}d}{s}\Big)^s\bigg) \exp\bigg(-\Big(\frac{\sqrt{T}R}{\mathrm{e}\mathfrak{LW}}\Big)^{\kappa(q,\eta)}\bigg).$$

Thus,

$$\begin{aligned}\mathbb{P}\bigg(\inf_{\zeta\in\mathcal{C}(s,c_0)}\frac{\|\zeta\|_{L^2}^2}{\|\zeta\|^2}>\frac{\mathfrak{e}_\infty}{2}\bigg)&\geq\mathbb{P}\bigg(\sup_{\zeta\in\mathcal{C}(s,c_0),\|\zeta\|\leq 1}\big|\zeta^\top(\overline{\mathbf{\Psi}}_\infty-\overline{\mathbf{\Psi}}_T)\zeta\big|\leq\frac{\mathfrak{e}_\infty}{2}\bigg)\\&\geq\mathbb{P}\bigg(\sup_{\zeta\in\mathcal{K}(2s),\|\zeta\|\leq 1}\big|\zeta^\top(\overline{\mathbf{\Psi}}_\infty-\overline{\mathbf{\Psi}}_T)\zeta\big|\leq\frac{\mathfrak{e}_\infty}{6(c_0+2)}\bigg)\\&\geq 1-21^{2s}\bigg(d\wedge\Big(\frac{\mathrm{e}d}{2s}\Big)^{2s}\bigg)\exp\bigg(-\Big(\frac{\sqrt{T}\mathfrak{e}_\infty}{18(c_0+2)\mathrm{e}\mathfrak{LW}}\Big)^{\kappa(q,\eta)}\bigg),\end{aligned}$$

resulting in the asserted condition on the sample size T. For proving part (b), note first that the relation

$$\bigg\{\max_{i=1,\dots,N}\big|\overline{\psi}_{ii,T}-\overline{\psi}_{ii,\infty}\big|>\frac{\mathfrak{e}_\infty}{2}\bigg\}\subset\bigg\{\sup_{\zeta\in\mathcal{C}(s,c_0)}\frac{\big|\zeta^\top(\overline{\mathbf{\Psi}}_T-\overline{\mathbf{\Psi}}_\infty)\zeta\big|}{\|\zeta\|^2}>\frac{\mathfrak{e}_\infty}{2}\bigg\}$$

in particular implies that, for $T\geq T_0(\frac{\varepsilon_0}{3},s,c_0,\mathfrak{LW})$,

$$\mathbb{P}\bigg(\max_{i=1,\dots,N}\overline{\psi}_{ii,T}>\mathfrak{D}_\infty+\frac{\mathfrak{e}_\infty}{2}\bigg)\leq\frac{\varepsilon_0}{3}.$$

It remains to control the deviation of the martingale term. Given $\theta,\eta\in\mathbb{R}^N$, we write

$$\frac{2}{T}\int_0^T\big(\sigma_0^{-1}(b_\theta-b_\eta)\big)^\top(X_t)\,\mathrm{d}W_t=2(\theta-\eta)^\top(\varepsilon_{1,T},\dots,\varepsilon_{N,T})^\top,$$

where

$$\varepsilon_{i,T}\coloneqq\frac{1}{T}\int_0^T\big(\sigma_0^{-1}\psi_i\big)^\top(X_s)\,\mathrm{d}W_s,\quad i=1,\dots,N.\tag{38}$$

Note that the quadratic variation of this continuous martingale is given by

$$\begin{aligned}\langle\varepsilon_i\rangle_T&=\frac{1}{T^2}\int_0^T\big(\sigma_0^{-1}\psi_i\big)^\top(X_s)\big(\sigma_0^{-1}\psi_i\big)(X_s)\,\mathrm{d}s\\&=\frac{1}{T^2}\int_0^T\big\langle\psi_i,a_0^{-1}\psi_i\big\rangle(X_s)\,\mathrm{d}s=\frac{1}{T}\overline{\psi}_{ii,T}.\end{aligned}$$

Using Bernstein's inequality for continuous martingales and taking into account the condition on λ, we thus obtain for for $T\geq T_0(\frac{\varepsilon_0}{3},s,c_0,\mathfrak{LW})$

$$\begin{aligned}&\mathbb{P}\bigg(\sup_{\theta\neq\eta\in\mathbb{R}^N}\frac{\frac{2}{T}\int_0^T\big(\sigma_0^{-1}(b_\theta-b_\eta)\big)^\top(X_t)\,\mathrm{d}W_t}{\|\theta-\eta\|_1}>\lambda\bigg)\\&\quad\leq\mathbb{P}\bigg(\sup_{\theta\neq\eta}\frac{2(\theta-\eta)^\top(\varepsilon_{1,T},\dots,\varepsilon_{N,T})}{\|\theta-\eta\|_1}>\lambda,\max_{i=1,\dots,N}\overline{\psi}_{ii,T}<\mathfrak{D}_\infty+\frac{\mathfrak{e}_\infty}{2}\bigg)\\&\qquad+\mathbb{P}\bigg(\max_{i=1,\dots,N}\overline{\psi}_{ii,T}>\mathfrak{D}_\infty+\frac{\mathfrak{e}_\infty}{2}\bigg)\end{aligned}$$

$$\leq \sum_{i=1}^{N} \mathbb{P}\Big(2|\varepsilon_{i,T}| > \lambda, \langle 2\varepsilon_i\rangle_T \leq \frac{4}{T}\Big(\mathfrak{D}_\infty + \frac{\mathfrak{e}_\infty}{2}\Big)\Big) + \frac{\varepsilon_0}{3}$$
$$\leq 2N \exp\Big(-\frac{T\lambda^2}{8\mathfrak{D}_\infty + 4\mathfrak{e}_\infty}\Big) + \frac{\varepsilon_0}{3} \leq \frac{2\varepsilon_0}{3}.$$

■

On the basis of the given key deviation inequalities, the machinery of high-dimensional statistics now allows the derivation of oracle inequalities. Our proof strategy follows the one developed in Ciołek et al. (2020).

Theorem 15 (oracle inequality) *Assume that we are given a continuous record of observations of the solution of* (31)*, where* $b_0 = b_{\theta_0} \in \mathcal{V}$ *with* $\|\theta_0\|_0 \leq s_0$*. Fix* $\gamma > 0$ *and* $\varepsilon_0 \in (0,1)$*, and consider the Lasso estimator* $\widehat{\theta}_T$ *introduced in* (34)*. Then, for*

$$\lambda \geq 2\sqrt{\frac{(2\mathfrak{D}_\infty + \mathfrak{e}_\infty)}{T} \cdot \log\Big(\frac{6N}{\varepsilon_0}\Big)} \quad \text{and} \quad T \geq T_0\Big(\tfrac{\varepsilon_0}{3}, s_0, 3 + \tfrac{4}{\gamma}, \mathfrak{LW}\Big), \tag{39}$$

with probability at least $1 - \varepsilon_0$*, we have*

$$\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2}^2 \leq (1+\gamma) \inf_{\theta \in \mathbb{R}^N : \|\theta\|_0 \leq s_0} \Big\{ \big\|\theta - \theta_0\big\|_{L^2}^2 + \frac{9(2+\gamma)^2}{2\gamma(1+\gamma)\mathfrak{e}_\infty} \|\theta\|_0 \lambda^2 \Big\}. \tag{40}$$

Furthermore, for any λ *fulfilling* (39) *and* $T \geq T_0\big(\frac{\varepsilon_0}{3}, s_0, 3, \mathfrak{LW}\big)$*, with probability at least* $1 - \varepsilon_0$*,*

$$\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2}^2 \leq \lambda^2 \cdot \frac{18 s_0}{\mathfrak{e}_\infty}. \tag{41}$$

By specifying λ as proposed in (39), the previous result implies that an upper bound of order $(s_0 \log N)/T$ on the squared L^2 risk of the Lasso estimator $\widehat{\theta}_T$ holds with high probability. In particular, in terms of the rate of convergence, our techniques give the same results as the concentration inequalities tailored to the specific OU model used in Gaïffas and Matulewicz (2019) and Ciołek et al. (2020), respectively.

Proof [Proof of Theorem 15] Recall the definition of the event $\mathscr{E}(s, c_0)$ according to (37), and let $s \geq s_0$. On $\mathscr{E}(s, c_0)$, the basic inequality (35) implies for any $\theta \in \mathbb{R}^N$ that

$$\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2}^2 + \lambda\big\|\widehat{\theta}_T - \theta\big\|_1 \leq \big\|\theta - \theta_0\big\|_{L^2}^2 + 2\lambda\Big(\big\|\theta\big\|_1 - \big\|\widehat{\theta}_T\big\|_1 + \big\|\widehat{\theta}_T - \theta\big\|_1\Big)$$
$$\leq \big\|\theta - \theta_0\big\|_{L^2}^2 + 4\lambda\big\|\widehat{\theta}_T|_{\mathrm{supp}(\theta)} - \theta\big\|_1. \tag{42}$$

Assume now that $\theta \in \mathbb{R}^N$ fulfills $\|\theta\|_0 \leq s_0$, and consider the event

$$4\lambda\big\|\widehat{\theta}_T|_{\mathrm{supp}(\theta)} - \theta\big\|_1 > \gamma\big\|\theta - \theta_0\big\|_{L^2}^2. \tag{43}$$

If this does not occur, (40) holds true since we obtain from (42) that

$$\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2}^2 \leq (1+\gamma)\big\|\theta - \theta_0\big\|_{L^2}^2.$$

Otherwise, if (43) holds, we have on $\mathscr{E}(s, c_0)$ that

$$\begin{aligned}\lambda\big\|\widehat{\theta}_T - \theta\big\|_1 &\leq \big\|\theta - \theta_0\big\|_{L^2}^2 + 4\lambda\big\|\widehat{\theta}_T|_{\mathrm{supp}(\theta)} - \theta\big\|_1 \\ &\leq \frac{4\lambda}{\gamma}\big\|\widehat{\theta}_T|_{\mathrm{supp}(\theta)} - \theta\big\|_1 + 4\lambda\big\|\widehat{\theta}_T|_{\mathrm{supp}(\theta)} - \theta\big\|_1\end{aligned}$$

i.e., $\widehat{\theta}_T - \theta \in \mathcal{C}(s, c_0)$ for the choice $c_0 = 3 + \frac{4}{\gamma}$, such that, after using Cauchy–Schwarz,

$$\big\|\widehat{\theta}_T|_{\mathrm{supp}(\theta)} - \theta\big\|_1 \leq \big\|\widehat{\theta}_T - \theta\big\| \cdot \sqrt{\|\theta\|_0} \leq \sqrt{\frac{2\|\theta\|_0}{\mathfrak{e}_\infty}}\big\|\widehat{\theta}_T - \theta\big\|_{L^2}.$$

Summing up,

$$\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2}^2 \leq \big\|\theta - \theta_0\big\|_{L^2}^2 + 3\lambda\sqrt{\frac{2\|\theta\|_0}{\mathfrak{e}_\infty}}\Big(\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2} + \big\|\theta - \theta_0\big\|_{L^2}\Big).$$

Applying the Young inequalities

$$\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2} \leq \frac{ax}{2} + \big\|\widehat{\theta}_T - \theta_0\big\|_{L^2}^2 \cdot \frac{1}{2ax}, \quad \big\|\theta - \theta_0\big\|_{L^2} \leq \frac{ax}{2} + \big\|\theta - \theta_0\big\|_{L^2}^2 \cdot \frac{1}{2ax},$$

with $a = (2+\gamma)/(2\gamma)$ and $x = 3\lambda\sqrt{2\|\theta\|_0/\mathfrak{e}_\infty}$, we finally obtain

$$\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2}^2 \leq (1+\gamma)\bigg(\big\|\theta - \theta_0\big\|_{L^2}^2 + \frac{9(2+\gamma)^2}{2\gamma(1+\gamma)\mathfrak{e}_\infty}\ \lambda^2\|\theta\|_0\bigg).$$

For the proof of (41), note that, taking $\theta = \theta_0$, (42) implies that, on $\mathscr{E}(s_0, c_0)$,

$$\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2}^2 + \lambda\big\|\widehat{\theta}_T - \theta_0\big\|_1 \leq 4\lambda\big\|\widehat{\theta}_T|_{\mathrm{supp}(\theta_0)} - \theta_0\big\|_1.$$

Now, since $\widehat{\theta}_T - \theta_0 \in \mathcal{C}(s_0, 3)$ on $\mathscr{E}(s_0, c_0)$,

$$\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2}^2 \leq 3\lambda\big\|\widehat{\theta}_T|_{\mathrm{supp}(\theta_0)} - \theta_0\big\|_1 \leq 3\lambda\sqrt{\frac{2s_0}{\mathfrak{e}_\infty}}\big\|\widehat{\theta}_T - \theta_0\big\|_{L^2},$$

which already gives the asserted inequality. ■

Remark 16 *While our results are non-asymptotic, we do face a restriction in that the constant $\mathfrak{W}$ appearing in the lower bound for the required sample size (see* (39)*) is not explicit. However, it appears to be very demanding to work out explicit constants in a general framework. In the spirit of Pokern et al. (2009), our arguments could also be carried out for the more restricted class of* reversible *diffusion processes by assuming a parametric form of the potential function and then considering a parametrized drift function $b_\theta(x) = \frac{1}{2}\operatorname{div}(a_0(x)) - \frac{1}{2}a_0(x)\nabla V_\theta(x)$ for $V_\theta \in \mathcal{V}$. Although functional inequalities (e.g., of Poincaré-type) are applicable in this reversible framework, the control of the constants involved still constitutes a fundamental challenge.*

We conclude this section by briefly categorising the results and sketching potential future research. Note first that Theorem 2.7 in Dexheimer and Strauch (to appear) provides a lower bound on the Frobenius norm for the estimation of the matrix $\boldsymbol{A}$ in the d-dimensional OU model (33) with $\sigma = \mathbb{I}_d$ over the class of s_0-sparse matrices. Translating the number of parameters into our framework, the lower bound is of order $s_0 \log(N/s_0)/T$. Compared to the upper bound of order $(s_0 \log N)/T$, there is thus only a logarithmic gap, appearing in this very form also in Gaïffas and Matulewicz (2019) and Ciołek et al. (2020). As demonstrated in Dexheimer and Strauch (to appear) in the context of drift estimation for Lévy-driven OU processes, the key to eliminating the logarithmic gap lies in a refined deviation inequality for the stochastic error (in our context specified as $\varepsilon_{i,T}$ as defined in (38)). In fact, the combination of concentration inequalities in the sense of Lemma 13 (which is a rather straightforward consequence of our general Theorem 3) with the techniques from Section 3.2 of Dexheimer and Strauch (to appear) can be expected to allow the derivation of minimax optimal penalized estimators also for general diffusion models.

4.2 MCMC for moderately heavy-tailed targets

In general, Markov chain Monte Carlo (MCMC) is a collective term for algorithms relying on ergodicity of Markov chains that are (i) easy to simulate and (ii) specifically designed such that their invariant distribution approximates a given target density, for which samples are to be obtained. These algorithms have a long and rich history. At this point, we cannot give a detailed account of the literature which would do justice to the field, but only want to point out its fundamental importance in connected areas such as Bayesian optimization or inverse problems in high dimensional contexts, where the posterior distribution becomes the target. Other than the fundamental theoretical work in Dalalyan (2017); Durmus and Moulines (2017) that will be discussed below, we refer to Dalalyan and Karagulyan (2019); Durmus et al. (2019); Durmus and Moulines (2019); Erdogdu et al. (2018); Erdogdu and Hosseinzadeh (2021); Teh et al. (2016); Vollmer et al. (2016) for some recent contributions that motivated our study. Our particular interest lies on the so called *Unadjusted Langevin Algorithm* (ULA), which we describe next.

Suppose that we are given a target density $\pi \propto \exp(-U(x))$ for some continuously differentiable function $U\colon \mathbb{R}^d \to \mathbb{R}$, which is usually referred to as the *potential*. Let us also assume that ∇U is L-Lipschitz continuous such that the (unadjusted or overdamped) *Langevin diffusion*

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \quad t \geq 0,$$

has a unique strong solution, which is a Feller Markov process with invariant distribution

$$\pi(\mathrm{d}x) = \frac{1}{\int_{\mathbb{R}^d} \exp(-U(y))\,\mathrm{d}y} \exp(-U(x))\,\mathrm{d}x, \quad x \in \mathbb{R}^d.$$

To obtain samples with approximate distribution π and to approximate integrals $\pi(f)$ for π-integrable functions f via the corresponding Monte Carlo estimator, in practice one needs to discretize the SDE to make simulation procedures feasible. The ULA uses a simple Euler discretization scheme as numerical SDE approximation, where the Euler discretization with step size Δ is the Markov chain given by the stochastic difference equation

$$\vartheta^{(\Delta)}_{n+1} = \vartheta^{(\Delta)}_n - \Delta\nabla U(\vartheta^{(\Delta)}_n) + \sqrt{2\Delta}\xi_{n+1}, \quad n \in \mathbb{N}_0, \quad \vartheta^{(\Delta)}_0 \overset{\mathrm{d}}{=} X_0,$$

where $(\xi_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. standard normal random variables on $\mathbb{R}^d$ independent of $\vartheta_0^{(\Delta)}$. Sampling such a chain is computationally efficient, provided the gradient ∇U can be cheaply evaluated. By considering the time-inhomogeneous Markov process given as the strong solution to the SDE

$$\mathrm{d}Z_t^{(\Delta)} = b(\mathbf{Z}^{(\Delta)}, t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \quad t \geq 0, \quad Z_0^{(\Delta)} = X_0,$$

with non-anticipatory drift

$$b(\boldsymbol{z}, t) = -\sum_{k=0}^{\infty} \nabla U(z_{k\Delta})\mathbf{1}_{[k\Delta,(k+1)\Delta)}(t), \quad (\boldsymbol{z}, t) \in \mathcal{C}(\mathbb{R}_+, \mathbb{R}^d) \times \mathbb{R}_+,$$

it is straightforward to show that the laws of $(\vartheta_n^{(\Delta)})_{n\in\mathbb{N}_0}$ and $(Z_{n\Delta}^{(\Delta)})_{n\in\mathbb{N}_0}$ coincide.

It has been observed in the literature Dalalyan (2017); Durmus and Moulines (2017) that for potentials U that are either strongly convex—i.e., π is strongly log-concave—or that are convex and superexponential outside some ball, explicit requirements on the step length Δ and sample size n can be formulated to guarantee sampling with ε-precision in total variation or Wasserstein distance. For strongly log-concave densities, the natural connection to the gradient descent in a convex setting is pointed out in Dalalyan (2017).

We now apply our previous results to obtain PAC bounds and related suggestions for sample size n, burn-in m and discretization step Δ for the ULA Monte Carlo estimator of polynomially growing functions for moderately heavy-tailed target densities π such that

$$\exists \iota > 0, q \in (0,1) \quad \text{such that} \quad \int_{\mathbb{R}^d} \exp(\iota\|x\|^{1-q})\,\pi(\mathrm{d}x) < \infty.$$

As follows from (7), this is the case if $-\nabla U$ satisfies $(\mathscr{A}(q))$ with $q \in (0,1)$, i.e., there exists some $M_0, \mathfrak{r} > 0$ such that

$(\mathscr{U}(q))$ $\langle \nabla U(x), x/\|x\|\rangle \geq \mathfrak{r}\|x\|^{-q}, \quad \|x\| \geq M_0.$

This setting differs substantially from the (strongly) convex setting in Dalalyan (2017); Durmus and Moulines (2017), whose assumptions imply $(\mathscr{U}(q))$ with $q \in [-1, 0)$ and therefore, in particular, require the targets to have exponential moments, i.e., light tails. Heavy-tailed target densities implied by our assumption $q \in (0,1)$ become relevant, e.g., in Bayesian inverse problems with heavy-tailed noise or prior. As the following result demonstrates, the Euler discretized Markov chain $\boldsymbol{\vartheta}^{(\Delta)}$ under $(\mathscr{U}(q))$ inherits the subexponential ergodic behaviour from the original Langevin diffusion $\mathbf{X}$, provided that U does not grow too fast. The proof is a straightforward application of the results from Douc et al. (2004)—which is the discrete-time counterpart to Douc et al. (2009)—and is postponed to Appendix A. Let $(\mathbf{P}^x)_{x\in\mathbb{R}^d}$ be a family of probability measures such that $\boldsymbol{\vartheta}^{(\Delta)}$ is started in x under $\mathbf{P}^x$.

Proposition 17 *Let $q \in (0,1)$. Suppose that U satisfies $(\mathscr{U}(q))$ and, moreover, for some $M_1 > 0$, $\|\nabla U(x)\| \leq \|x\|^\beta$ for $\beta \leq (1-q)/2$. Then, for any $\Delta > 0$ in case $\beta < (1-q)/2$ or any $\Delta \leq \mathfrak{r}$ in case $\beta = (1-q)/2$, there exists an invariant probability measure $\pi_{(\Delta)}$ for the chain $\boldsymbol{\vartheta}^{(\Delta)}$ and there are constants $c = c(q,\Delta) > 0$ and $\widetilde{c} = \widetilde{c}(q,\Delta)$ such that, for $f_q(x) \sim (1+\|x\|)^{-2q}\exp(\widetilde{c}\|x\|^{1-q})$ and $r_q(n) \sim n^{-2q/(1+q)}\exp(cn^{(1-q)/(1+q)})$, we have for any $x \in \mathbb{R}^d$ and pairs of inverse Young functions $(\Psi_1, \Psi_2) \in \mathcal{I}$*

$$\lim_{n\to\infty} \Psi_1(r_q(n))\big\|\mathbf{P}^x(\vartheta_n^{(\Delta)} \in \cdot) - \pi_{(\Delta)}\big\|_{\Psi_2\circ f_q} = 0.$$

This convergence behaviour is in line and in fact states more precisely the findings from (Roberts and Tweedie, 1996, Section 3) for the ULA in $d = 1$ for the model class of sub-Weibull distributions. It is vital to note that π and $\pi_{(\Delta)}$ do not coincide, so even if the ULA converges at subgeometric rate for fixed step size Δ, we need to choose Δ appropriately small to obtain useful approximations. We make this tuning parameter choice precise in the following.

Typical potentials satisfying $(\mathscr{U}(q))$—such as $U(x) \propto \|x\|^{1-q}$ outside some ball centered around 0—are not convex, and their gradient may converge at infinity. In fact, if we have $\lim_{\|x\|\to\infty}\|\nabla U(x)\| = 0$, then (Roberts and Tweedie, 1996, Theorem 2.4) implies that the Langevin diffusion $\mathbf{X}$ is not exponentially ergodic. Hence, we cannot expect π to have exponentially decaying tails. Therefore, in contrast to the usually encountered potentials exhibiting some degree of convexity, it is quite natural for our purposes to assume that ∇U is bounded under $(\mathscr{U}(q))$ for $q \in (0, 1)$. This makes it easy to prove the following result quantifying convergence of ULA to the target π and the performance of the ULA Monte Carlo estimator with burn-in m,

$$\mathbb{H}^{\boldsymbol{\vartheta}^{(\Delta)}}_{m,n,\Delta}(f) \coloneqq \frac{1}{n}\sum_{k=m+1}^{m+n} f\big(\vartheta^{(\Delta)}_{k\Delta}\big),$$

based on our results from Section 3.2 and the Girsanov argument underlying the total variation convergence result from Dalalyan (2017) for strongly convex potentials. Denote by $\mathbb{P}^{x,n\Delta}_{\mathbf{X}}$ and $\mathbb{P}^{x,n\Delta}_{\mathbf{Z}^{(\Delta)}}$ the laws of $(X_t)_{t\in[0,\Delta n]}$ and $(Z^{(\Delta)}_t)_{t\in[0,n\Delta]}$, respectively, under $\mathbb{P}^x$.

Proposition 18 *Suppose that $U \in \mathcal{C}^1(\mathbb{R}^d)$ has an L-Lipschitz continuous and bounded gradient that satisfies $(\mathscr{U}(q))$ for some $q \in (0, 1)$.*

(i) For any $\Delta \in (0, 1]$ and initial distribution ν such that $V_q \in L^1(\nu)$, it holds for any $n \in \mathbb{N}$,

$$\big\|\mathbf{P}^{\nu}\big(\vartheta^{(\Delta)}_n \in \cdot\big) - \pi\big\|_{\mathrm{TV}} \le \mathfrak{C}\nu(V_q)\exp\big(-(\iota'' n\Delta)^{(1-q)/(1+q)}\big) + \sqrt{\frac{(1 + \|\|\nabla U(\cdot)\|^2_\infty\|_\infty)dL^2 n\Delta^2}{2}}, \tag{44}$$

for some constant $\mathfrak{C} > 0$ and $\iota'' \in (0, \iota^{(1+q)/(1-q)}(1 + q)(\mathfrak{r} - \iota(1 - q)))$ for some $\iota < \mathfrak{r}/(1 - q)$.

(ii) Let $\eta_1, \eta_2, \eta_3 \ge 0$, $f \in \mathcal{G}(\eta_1, \mathfrak{L}) \cap \mathcal{W}^{2,p}_{\mathrm{loc}}(\mathbb{R}^d)$, $p \ge d$, with $\nabla f \in L^{2d}_{\mathrm{loc}}(\mathbb{R}^d)$ such that $\|\nabla f(x)\| \lesssim 1 + \|x\|^{\eta_2}$ and for all $i, j = 1, \dots, d$, $|\partial_{x_i,x_j} f(x)| \lesssim 1 + \|x\|^{\eta_3}$. Let also $C, \iota'', \alpha, \widetilde{\gamma}, \widetilde{\varsigma} = \widetilde{\varsigma}(\eta_1, q, 0), \varrho = \varrho(\alpha, \eta_2, \widetilde{\gamma}, q)$ be the constants from Corollary 12, adapted to the specific parameters of the Langevin diffusion. Then, for Δ satisfying both $\Delta < \min\{1, \varepsilon/(3\mathrm{e}\mathfrak{D}), (\log(1/\delta))^{\widetilde{\varsigma}-\varrho}\}$ and

$$\Delta \le \frac{(\delta\varepsilon)^2}{2(1 + \|\|\nabla U(\cdot)\|_\infty\|_\infty)dL^2\big((\log(4/\delta))^{2\widetilde{\varsigma}} + \varepsilon^2\big(2 + (\log(4C/\delta))^{(1+q)/(1-q)}/\iota''\big)\big)},$$

sample size

$$n = n(\Delta, \varepsilon, \delta) = \big\lceil \Psi(\Delta, \varepsilon, \delta/4)\big\rceil$$

and burn-in

$$m = m(\Delta, \varepsilon, \delta) = \lceil \Delta^{-1} (\log(4C/\delta))^{(1+q)/(1-q)} / \iota'' \rceil,$$

it holds for any initial distribution ν such that $V_q \in L^1(\nu)$ that

$$\mathbf{P}^\nu \big(\big| \mathbb{H}^{\boldsymbol{\vartheta}^{(\Delta)}}_{m,n,\Delta}(f) - \pi(f) \big| \leq \varepsilon \big) \geq 1 - \delta.$$

Proof As in the proof of Lemma 2 in Dalalyan (2017), see also Dalalyan and Tsybakov (2012), using L-Lipschitz continuity of ∇U and Girsanov's theorem, it follows that the Kullback–Leibler divergence of $\mathbb{P}^{x,n\Delta}_{\mathbf{X}}$ wrt $\mathbb{P}^{x,n\Delta}_{\mathbf{Z}^{(\Delta)}}$ fulfills

$$\mathrm{KL}\big(\mathbb{P}^{x,n\Delta}_{\mathbf{X}} \big\| \mathbb{P}^{x,n\Delta}_{\mathbf{Z}^{(\Delta)}} \big) \leq \frac{L^2 \Delta^3}{12} \sum_{k=0}^{n-1} \mathbb{E}^x \big[\big\| \nabla U(Z^{(\Delta)}_{k\Delta}) \big\|^2 \big] + \frac{dL^2 n \Delta^2}{4}.$$

Thus, using $\| \|\nabla U(\cdot)\| \|_\infty \leq \sqrt{d} \| \|\nabla U(\cdot)\|_\infty \|_\infty$ and Pinsker's inequality, it follows

$$\big\| \mathbb{P}^x(X_{n\Delta} \in \cdot) - \mathbf{P}^x\big(\vartheta^{(\Delta)}_n \in \cdot \big) \big\|_{\mathrm{TV}} \leq \big\| \mathbb{P}^{x,n\Delta}_{\mathbf{X}} - \mathbb{P}^{x,n\Delta}_{\mathbf{Z}^{(\Delta)}} \big\|_{\mathrm{TV}} \leq \sqrt{\frac{(1 + \| \|\nabla U(\cdot)\|^2_\infty \|_\infty) dL^2 n \Delta^2}{2}}. \tag{45}$$

By triangle inequality, subexponential convergence in (5) with the parameters adapted to the Langevin diffusion and (45), we immediately obtain (44). Moreover, for Δ given as in part (ii), the choice $n = n(\Delta, \varepsilon, \delta)$ and $m = m(\Delta, \varepsilon, \delta)$, if we define

$$g_\varepsilon((x_t)_{t \in [0,(n+m)\Delta]}) = \mathbf{1}_{(\varepsilon, \infty)} \Big(\Big| \frac{1}{n} \sum_{k=m+1}^{n+m} (f(x_{k\Delta}) - \pi(f)) \Big| \Big),$$

for a path $(x_t)_{t \in [0,(n+m)\Delta]} \in \mathcal{C}([0, (n+m)\Delta], \mathbb{R}^d)$, it follows from (45) that

$$\begin{aligned}
&|\mathbb{P}^\nu(|\mathbb{H}_{m,n,\Delta}(f) - \pi(f)| > \varepsilon) - \mathbf{P}^\nu(|\mathbb{H}^{\boldsymbol{\vartheta}^{(\Delta)}}_{n,m,\Delta}(f) - \pi(f)| > \varepsilon)| \\
&\quad = \big| \mathbb{E}^\nu \big[g_\varepsilon\big((X_t)_{t \in [0,(n+m)\Delta]} \big) \big] - \mathbb{E}^\nu \big[g_\varepsilon\big((Z^{(\Delta)}_t)_{t \in [0,(n+m)\Delta]} \big) \big] \big| \\
&\quad \leq \int_{\mathbb{R}^d} \big\| \mathbb{P}^{x,(n+m)\Delta}_{\mathbf{X}} - \mathbb{P}^{x,(n+m)\Delta}_{\mathbf{Z}^{(\Delta)}} \big\|_{\mathrm{TV}} \, \nu(\mathrm{d}x) \\
&\quad \leq \sqrt{\frac{(1 + \| \|\nabla U(\cdot)\|^2_\infty \|_\infty) dL^2 (n+m) \Delta^2}{2}} \\
&\quad \leq \Bigg(\frac{(1 + \| \|\nabla U(\cdot)\|^2_\infty \|_\infty) dL^2 \Delta}{2} \Big(2 + \frac{(\log(4/\delta))^{2\widetilde{\varsigma}}}{\varepsilon^2} + (\log(4C/\delta))^{(1+q)/(1-q)} / \iota'' \Big) \Bigg)^{1/2} \\
&\quad \leq \delta/2.
\end{aligned}$$

Statement (ii) now follows from Corollary 12 and triangle inequality. ■

The above result gives lower bounds on the required step length, sample size and burn-in for an ε-precise integral approximation of $\pi(f)$ with probability at least $1-\delta$ for polynomially bounded f with polynomially bounded weak derivative and Hessian. These are summarized in Table 1. An obvious application of this result are explicit finite sample guarantees for MCMC moment approximations of the target π.

	step length Δ	sample size n	burn-in m
ε-prec. sampling	$\frac{\varepsilon^2}{d(\log(\mathfrak{C}/\varepsilon))^{(1+q)/(1-q)})}$	$\frac{d(\log(\mathfrak{C}/\varepsilon))^{2(1+q)/(1-q)}}{\varepsilon^2}$	–
(ε,δ)-PAC bound	$\frac{(\delta\varepsilon)^2}{d(\log(1/\delta))^{2(\eta_1+(q+3)/2)/(1-q)}}$	$\frac{d\mathfrak{D}^2(\log(1/\delta))^{(4(\eta_1+(q+3)/2))/(1-q)}}{\delta^2\varepsilon^4}$	$\frac{d(\log(1/\delta))^{2(\eta_1+q+2)/(1-q)}}{(\delta\varepsilon)^2}$

Table 1: Order of sufficient sampling frequency Δ, sample size n and burn-in m for (ε,δ)-PAC bounds and sampling within ε-TV margin

Remark 19 *It should be noted that the exact dimensional dependence of $\mathfrak{D}$ is not clear, which, similarly to the previous section, is an effect of unspecified constants in the ergodicity and Sobolev bounds used for the derivation of the concentration inequalities. Overcoming this issue is highly non-trivial and subject of ongoing research efforts. In contrast, the convex, respectively strongly convex, settings in Durmus and Moulines (2017); Dalalyan (2017) give rise to Poincaré, respectively log-Sobolev, inequalities with explicit constants such that the investigated required number of iterations for sampling within an ε-margin in total variation can be made explicit in terms of the dimension in these papers. According to the above, the simulation grid should be significantly finer and the sample size significantly larger to obtain exact PAC-guarantees compared to the case when one would just be interested in sampling with ε-precision in total variation. Here, the dependence on the level ε is a natural correspondence to the sample sizes (and hence necessary number of gradient evaluations) found in Dalalyan (2017); Durmus and Moulines (2017).*

Our results yield explicit and useful guarantees for a sampling scenario that is quite different from what is usually encountered in the theoretical MCMC literature. Still, we expect that the dependence of (Δ, n, m) on δ for the PAC bounds can be improved in the sense that the δ^2-dependency is likely too strict. Its occurrence is explained by our strategy to control the total variation distance between the law of the Langevin diffusion $\mathbf{X}$ and its numerical approximation $\mathbf{Z}^{(\Delta)}$ in terms of their KL-divergence using Pinsker's inequality. This leads to a suboptimal bound on the total variation distance, causing the additional dependence on δ^2. We are not aware of any other approaches in the MCMC literature to control this loss on the path level. This issue can be possibly circumvented by deriving concentration inequalities for $\mathbb{H}^{\boldsymbol{\vartheta}^{(\Delta)}}_{m,n,\Delta}(f)$ around its mean directly and lift these to concentration inequalities of $\mathbb{H}^{\boldsymbol{\vartheta}^{(\Delta)}}_{m,n,\Delta}(f)$ around the target $\pi(f)$ by establishing appropriate bias estimates. This is the strategy pursued in (Durmus and Moulines, 2015, Proposition 18)—an earlier preprint version of Durmus and Moulines (2017)—where the authors infer a sufficient sample size $n \sim d\log(1/\delta)/\varepsilon^4$ and sampling frequency $\Delta \sim \varepsilon^2/d$ for the ULA MC estimator of the integral $\pi(f)$ for a strictly log-concave density (in particular, $q=-1$) and bounded f. Since we focus on applications that can be treated with our theoretical results from Section 3.2, we do not push further the issue of improving our bounds in the setting of a heavy-tailed target π and unbounded integrands f. Instead, we leave it open as an interesting question for future research.

Acknowledgments

CS and LT gratefully acknowledge financial support of Carlsberg Foundation Young Researcher Fellowship grant CF20-0640 "Exploring the potential of nonparametric modelling of complex systems via SPDEs".

Appendix A. Remaining proofs

Proof [Proof of Proposition 1] By (Douc et al., 2009, Proposition 1), every compact set is petite and any skeleton chain is irreducible. Moreover, if we let $V \in \mathcal{C}^2(\mathbb{R}^d)$ such that $V = \|x\|^\gamma$ for $\|x\| \geq M_0$ and $V \geq 1$, and we can show that LV is locally bounded and

$$LV(x) \lesssim -\phi \circ V(x)(1 + o(1)), \quad \|x\| \geq M_0, \tag{46}$$

for $\phi(x) = \mathfrak{r}\gamma x^{(\gamma-(1+q))/\gamma}$ which is increasing, differentiable and concave on $(0, \infty)$, it will follow from (Douc et al., 2009, Theorem 3.4) that, for any $\varepsilon \in (0, 1)$, the condition $\mathbf{D}(C_\varepsilon, V, \phi_\varepsilon, a_\varepsilon)$ is satisfied for $\phi_\varepsilon = (1 - \varepsilon)\phi$, $C_\varepsilon = \overline{B(0, M_\varepsilon)}$ for $M_\varepsilon \geq M_0$ large enough and $a_\varepsilon = \sup_{\|x\| \leq M_\varepsilon} |LV(x) + \phi_\varepsilon \circ V(x)|$. This then implies the result using Theorem 3.2 and Proposition 4.6 from Douc et al. (2009). (Note that in the notation of Douc et al. (2009), $f^* = \phi_\varepsilon \circ V \sim f_{\gamma,q}$, $H^{-1}_{\phi_\varepsilon}(t) = (1 + (1 + q)(1 - \varepsilon)t/\gamma)^{\gamma/(1+q)}$ for $q \in (-1, 1)$ and $H^{-1}_{\phi_\varepsilon}(t) = \exp(-\mathfrak{r}\gamma(1 - \varepsilon)t)$ for $q = -1$, hence $r_*(t) = \phi_\varepsilon \circ H^{-1}_{\phi_\varepsilon}(t) \sim r_{\gamma,q}(t)$.) Since b, σ are locally bounded and L is a local operator, it is immediate that LV is locally bounded as well. Further, for $\|x\| \geq M_0$, $(\mathscr{A}(q))$ implies

$$\langle b(x), \nabla V(x)\rangle = \gamma\|x\|^{\gamma-1}\langle b(x), x/\|x\|\rangle \leq -r\gamma\|x\|^{\gamma-1-q} = -\phi \circ V(x),$$

and the assumptions on the diffusion matrix yield

$$\begin{aligned} |\mathrm{tr}\big(a(x)D^2V(x)\big)| &= \Big| \sum_{i,j=1}^{d} a_{i,j}(x)\big(\mathbf{1}_{\{i=j\}}\gamma\|x\|^{\gamma-2} + \gamma(\gamma-2)x_ix_j\|x\|^{\gamma-4}\big)\Big| \\ &\leq (\Lambda\gamma d + \gamma|\gamma-2|\lambda_+)\|x\|^{\gamma-2} = o(\phi \circ V(x)). \end{aligned}$$

This gives (46) and therefore the result. ∎

Proof [Proof of Proposition 17] Let $P^{(\Delta)}(x, B) = \mathbf{P}^x(\vartheta^{(\Delta)}_n \in B)$, $(x, B) \in \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$, be the transition kernel of the Markov chain $\boldsymbol{\vartheta}^{(\Delta)}$. Since $P^{(\Delta)}(x, \cdot) = \mathcal{N}(x - h\nabla U(x), 2h\mathbb{I}_d)$, it follows from classical Meyn–Tweedie arguments (cf. (Hansen, 2003, Theorem 3.1) for the precise statement) that $P^{(\Delta)}$ is an aperiodic and $\boldsymbol{\lambda}$-irreducible Markov kernel and that all compact sets are small and hence petite. Let $\Phi(x) = x - \Delta\nabla U(x)$ such that we may write $\vartheta^{(\Delta)}_{n+1} = \Phi(\vartheta^{(\Delta)}_n) + \sqrt{2}\Delta\xi_{n+1}$. By our assumptions on the gradient ∇U, we can choose $M \geq M_0 \vee M_1 \vee 1$ large enough such that, for $\|x\| \geq M$, we have

$$\begin{aligned} \|\Phi(x)\|^2 &= \|x\|^2 - 2\Delta\langle x, \nabla U(x)\rangle + \Delta^2\|\nabla U(x)\|^2 \\ &\leq \|x\|^2 - 2\Delta\mathfrak{r}\|x\|^{1-q} + \Delta^2\|x\|^{2\beta} \\ &\leq \|x\|^2\big(1 - \Delta\mathfrak{r}\|x\|^{-(1+q)}\big) \end{aligned}$$

$$\leq \big(\|x\|\big(1 - \tfrac{\mathfrak{r}\Delta}{2}\|x\|^{-(1+q)}\big)\big)^2.$$

Hence, Assumption 3.4 from Douc et al. (2004) is fulfilled with $R_0 = M, \rho = 1+q, r = \mathfrak{r}\Delta/2$. Moreover, since the noise $(\xi_n)_{n\in\mathbb{N}}$ is i.i.d. Gaussian, Assumption 3.3 from Douc et al. (2004) is satisfied for any $z_0 > 0$ and $\gamma_0 = 1$. It thus follows from (Douc et al., 2004, Theorem 3.3) that their central drift condition $\mathbf{D}(\phi, V, C)$ holds for $\phi(v) = cv(1 + \log v)^{-2q/(1-q)}$, $V(x) = \mathrm{e}^{z\|x\|^{1-q}}$ and the compact set $C = \overline{B(0, \widetilde{M})}$, for some $c, z > 0$ and $\widetilde{M} \geq M$. Consequently, for $H_\phi(t) = \int_1^t 1/\phi(v)\,\mathrm{d}v$, we have $r_q \sim \phi \circ H_\phi^{-1}$ and $f_q \sim \phi \circ V$ for appropriate choices of the constants $c(q, \Delta), \widetilde{c}(q, \Delta)$. Since $P^{(\Delta)}$ is irreducible and aperiodic and C is petite, we can now apply (Douc et al., 2004, Theorem 2.8) to prove the claim. ■

References

C. Aeckerle-Willems and C. Strauch. Concentration of scalar ergodic diffusions and some statistical implications. *Ann. Inst. Henri Poincaré Probab. Stat.*, 57(4):1857–1887, 2021.

C. Aeckerle-Willems and C. Strauch. Sup-norm adaptive drift estimation for multivariate nonreversible diffusions. *Ann. Statist.*, 50(6):3484–3509, 2022.

M. T. Barlow and M. Yor. Semimartingale inequalities via the Garsia-Rodemich-Rumsey lemma, and applications to local times. *J. Functional Analysis*, 49(2):198–229, 1982.

S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567, 2015.

V. I. Bogachev, M. Röckner, and S. V. Shaposhnikov. The Poisson equation and estimates for distances between stationary distributions of diffusions. *J. Math. Sci. (N.Y.)*, 232(3, Problems in mathematical analysis. No. 92 (Russian)):254–282, 2018.

D. Bosq. Parametric rates of nonparametric estimators and predictors for continuous time processes. *Ann. Statist.*, 25(3):982–1000, 1997.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

P. Cattiaux and A. Guillin. Deviation bounds for additive functionals of Markov processes. *ESAIM Probab. Stat.*, 12:12–29, 2008.

G. Ciołek, D. Marushkevych, and M. Podolskij. On Dantzig and Lasso estimators of the drift in a high dimensional Ornstein-Uhlenbeck model. *Electron. J. Stat.*, 14(2):4395–4420, 2020.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):651–676, 2017.

A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Process. Appl.*, 129(12):5278–5311, 2019.

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443, 2012.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

N. Dexheimer and C. Strauch. On Lasso and Slope drift estimators for Lévy-driven Ornstein–Uhlenbeck processes. *Bernoulli*, to appear.

R. Douc, G. Fort, E. Moulines, and P. Soulier. Practical drift conditions for subgeometric rates of convergence. *Ann. Appl. Probab.*, 14(3):1353–1377, 2004.

R. Douc, A. Guillin, and E. Moulines. Bounds on regeneration times and limit theorems for subgeometric Markov chains. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(2):239–257, 2008.

R. Douc, G. Fort, and A. Guillin. Subgeometric rates of convergence of f-ergodic strong Markov processes. *Stochastic Process. Appl.*, 119(3):897–923, 2009.

A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. arXiv:1507.05021v1, 2015.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

A. Durmus, S. Majewski, and B. a. Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *J. Mach. Learn. Res.*, 20:Paper No. 73, 46, 2019.

M. A. Erdogdu and R. Hosseinzadeh. On the convergence of langevin monte carlo: The interplay between tail growth and smoothness. In M. Belkin and S. Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1776–1822. PMLR, 15–19 Aug 2021.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. *Advances in Neural Information Processing Systems*, 31, 2018.

J. Fan, B. Jiang, and Q. Sun. Hoeffding's inequality for general Markov chains and its applications to statistical learning. *J. Mach. Learn. Res.*, 22:Paper No. 139, 35, 2021.

G. Fort and G. O. Roberts. Subgeometric ergodicity of strong Markov processes. *Ann. Appl. Probab.*, 15(2):1565–1589, 2005.

S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

S. Gaïffas and G. Matulewicz. Sparse inference of the drift of a high-dimensional Ornstein-Uhlenbeck process. *J. Multivariate Anal.*, 169:1–20, 2019.

L. Galtchouk and S. Pergamenshchikov. Uniform concentration inequality for ergodic diffusion processes. *Stochastic Process. Appl.*, 117(7):830–839, 2007.

L. Galtchouk and S. Pergamenshchikov. Uniform concentration inequality for ergodic diffusion processes observed at discrete times. *Stochastic Process. Appl.*, 123(1):91–109, 2013.

F. Gao, A. Guillin, and L. Wu. Bernstein-type concentration inequalities for symmetric Markov processes. *Theory Probab. Appl.*, 58(3):358–382, 2014.

N. R. Hansen. Geometric ergodicity of discrete-time approximations to multivariate diffusions. *Bernoulli*, 9(4):725–743, 2003.

S. F. Jarner and G. O. Roberts. Polynomial convergence rates of Markov chains. *Ann. Appl. Probab.*, 12(1):224–247, 2002.

B. Jiang, Q. Sun, and J. Fan. Bernstein's inequality for general markov chains, 2018. arXiv:1805.10721.

N. V. Krylov. *Controlled diffusion processes*, volume 14 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2009. Translated from the 1977 Russian original by A. B. Aries, Reprint of the 1980 edition.

R. S. Liptser and A. N. Shiryaev. *Statistics of random processes. I*, volume 5 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, expanded edition, 2001. General theory, Stochastic Modelling and Applied Probability.

M. N. Malyshkin. Subexponential estimates for the rate of convergence to the invariant measure for stochastic differential equations. *Teor. Veroyatnost. i Primenen.*, 45(3):489–504, 2000.

J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM J. Numer. Anal.*, 48(2): 552–577, 2010.

R. Nickl and K. Ray. Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist.*, 48(3):1383–1408, 2020.

E. Pardoux and A. Y. Veretennikov. On the Poisson equation and diffusion approximation. I. *Ann. Probab.*, 29(3):1061–1085, 2001.

Y. Pokern, A. M. Stuart, and E. Vanden-Eijnden. Remarks on drift estimation for diffusion processes. *Multiscale Model. Simul.*, 8(1):69–95, 2009.

E. Rio. *Asymptotic theory of weakly dependent random processes*, volume 80 of *Probability Theory and Stochastic Modelling*. Springer, Berlin, 2017.

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

O. Stramer and R. L. Tweedie. Existence and stability of weak solutions to stochastic differential equations with non-smooth coefficients. *Statist. Sinica*, 7(3):577–593, 1997.

D. W. Stroock and S. R. S. Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin, 2006. Reprint of the 1997 edition.

Y. W. Teh, A. H. Thiery, and S. J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.*, 17:Paper No. 7, 33, 2016.

P. Tuominen and R. L. Tweedie. Subgeometric rates of convergence of f-ergodic Markov chains. *Adv. in Appl. Probab.*, 26(3):775–798, 1994.

S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. Exploration of the (non-)asymptotic bias and variance of stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.*, 17:Paper No. 159, 45, 2016.

M. J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. A non-asymptotic viewpoint.