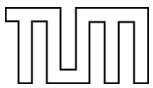


Technische Universität München

Department of Mathematics

Master's Thesis

Kolmogorov semigroups and Stochastic Gradient Monte-Carlo

Roland Halbig

Supervisor: Prof. Dr. Caroline Lasser

Advisor: Prof. Dr. Caroline Lasser

Submission Date: 14.07.2017

Zusammenfassung

Um die zeitliche Entwicklung von Diffusionsgleichungen im Sinne von Itô zu beschreiben, bie-ten sich Kolmogorov Halbgruppen in natürlicher Weise an. Diese Gleichungen stellen eine wichtige Klasse sogenannter Langevin-Gleichungen dar, die sehr häufig für die Simulation von Markov Ketten verwendet werden (sogenannte MCMC Methoden). Es ist längst gezeigt worden, dass diese Gleichungen eng mit dem stochastischen Gradientenverfahren (SGD) zusammenhängen. Diese Masterarbeit stellt in ausführlicher Weise die Theorie der primalen und dualen Kolmogorov Halbgruppe dar. Diese Theorie wird dann dazu verwendet um stochastische Differentialgleichungen zu definieren, welche eine gewünschte Wahrscheinlichkeitsverteilung als Stationärverteilung besitzen. Des Weiteren werden zwei Methoden vorgestellt, mit denen sich diese Gleichung numerisch integrieren lassen. Diese Arbeit enthält außerdem einen neuen Beweis für die Konsistenz des Euler-Integrators. Durch diesen Beweis lassen sich die neuesten Ergebnisse zur Konsistenzordnung wesentlich vereinfachen. Schließlich wird in einigen numerischen Experimenten die Konvergenz verschiedener Ansätze untersucht. Dabei wird ein gewichtetes Modell von Normalverteilungen verwendet um das Unkorrigierte Langevin Verfahren (ULA), den korrigierten Metropolis Langevin Algorithmus (MALA) and das Langevin Abkühlungsverfahren miteinander zu vergleichen.

Contents

Acknowledgements

I would like to thank my academic supervisor Professor Dr. Caroline Lasser for all her support and advice during my thesis. Thank you for your time, your authentic interest in my topic and your encouragement; they made working on this thesis a very enjoyable and memorable experience.

I would also like to acknowledge friends and families who supported me during my Master's program. First and foremost I thank my parents Xaver and Christa Halbig for their constant love and support. You let me always choose my way freely and were there when I needed you, I am very proud of you. I would also like to thank my brother Gerold who has been my best friend in many areas of life. Many thanks also go to my friends: Lio and Mischa for the company in the library; Martina for the bouldering sessions; Sarali for the green, blue and purple moments; Gerold, Richard, Clemens, Julian and Tom for the mountain adventures and my WG for the kitchen evenings. I am happy to have you as my family and friends; you make life colorful and exciting.

Special thanks go to Professor Jonathan Goodman for his kind advice in terms of the infinitesimal variance and to Mathias Kettner GmbH for being flexible with the working times and all the scribbling paper that is now full of mathematics.

Abstract

Kolmogorov semigroups are a natural way of describing time evolution of Itô diffusion equations. These equations are an important class of so-called Langevin equations frequently used in Markov Chain Monte Carlo (MCMC) simulation and have been shown to be closely related to the Stochastic Gradient Descent (SGD). This thesis gives a detailed presentation of the theory for the primal and dual Kolmogorov semigroup. The theory is then used to construct Stochastic Differential Equations with desired stationary distributions. Furthermore, two methods for integrating these equations is introduced. This thesis also contains a novel proof for consistency of the Euler integrator, which significantly simplifies recent results. Some numerical experiments demonstrate the Unadjusted Langevin Algorithm (ULA), the Metropolis Adjusted Langevin Algorithm (MALA) and the Langevin Dynamics Annealing (LDA) algorithm on a Mixture-of-Gaussians model.

Chapter 1

Introduction

Starting around the year 2010, deep neural networks were rapidly popularized also outside of the Artificial Intelligence community. This was due to a number of breakthroughs in Pattern Recognition competitions. These breakthroughs were largely facilitated by the introduction of graphical processing units (GPUs). These made the costly calculations of deep neural networks feasible also for smaller teams of researchers (c.f. [Economist, 2016]).

One of the main ideas of Pattern Recognition and Deep Learning are multi-layer networks. Their principles have been known since the 1960s (e.g. [Rosenblatt, 1961] and [Kelley, 1960]) and they have been subject to continuous research. The backpropagation algorithm used for training these networks is a special variant the of Stochastic Gradient Descent (SGD, c.f. [Dreyfus, 1962]). This algorithm was introduced for Stochastic Approximation Problems in statistics in the early 1950s by [Robbins and Monro, 1951] and [Kiefer and J.Wolfowitz, 1952]. Since then it has been a common denominator to many different areas of study. In Statistics and Machine Learning it is used for estimation of model parameters (e.g. [Hastie et al., 2003]). Physicists use it for the simulation of real-world phenomena (e.g. [Betancourt, 2017]). In Global Optimization and Optimal Control problems SGD methods are used to find efficient solutions (e.g. [Kirkpatrick et al., 1983]).

The setting of a typical Machine learning task for the SGD method can be described as follows (c.f. [Bottou, 2012], [Bertsekas and Tsitsiklis, 2000]): Let X be a random variable with probability distribution $X \sim \mu$. We assume that the distribution μ is unknown to us and we can observe X only through a number of independent realizations $\{X_i\}_{i=1}^N$. We assume that we can approximate μ by a model μ_θ which depends on a parameter θ. Our goal is to estimate θ in such a way that the expected value of an error function $U(\theta, x)$ with respect to this model becomes minimal, i.e.

$$\theta^* = \text{argmin}_\theta \left\{ E_\theta[X] := \frac{1}{N} \sum_{i=1}^{N} U(\theta, X_i) \right\}.$$

If we can show that $\nabla_\theta E_{\theta^*}[X] = 0$, then we can apply a gradient descent method for finding θ^*. If the number of data points N is very large, the computation of $\nabla_\theta E_\theta[X]$

is very costly. Instead, in each update step k, the gradient $\nabla_\theta E_\theta[X]$ is only evaluated for one data point $X^{(k)}$ at a time. In compensation, an iteration consists of one sweep through the whole data set. For each update step we define

$$\theta^{(k+1)} = \theta^{(k)} + h_k \left(\nabla_\theta E_{\theta^{(k)}}[X^{(k)}] + \varepsilon_k(\theta) \right). \tag{1.1}$$

as the **stochastic gradient descent** (SGD). Here $\varepsilon_k(\theta) = E_\theta[X] - E_\theta[X^{(k)}]$ is the stochastic error of the update step k. Other variants of this method use batches of samples or preconditioning techniques. We can show that the SGD method converges under certain conditions on the potential U and the stochastic error ε_k (c.f. [Bertsekas and Tsitsiklis, 2000, Proposition 3]). The stochasticity of equation (1.1) also has some complications: we don't know much about ε_k in the first place so we have to make a number of assumptions here. Also, it is not a trivial task to tell whether the method has converged because we do not know much about the probability distribution of θ. One of the main goals of this thesis is to take a closer look at this.

Around the same time as the SGD, Markov Chain Monte Carlo (MCMC) methods were introduced by [Metropolis et al., 1953] and improved by [Hastings, 1970]. These are methods use Markov chains in order to sample from probability distributions which are stationary with respect to the Markov chain. MCMC methods became very popular for numerical simulations for diffusion equations among theoretical physicists (c.f. [Andrieu et al., 2003]). One problem of the method due to Metropolis and Hastings is that it critically relies on good acceptance rates. It soon became clear that coupling this mehtod with diffusion equations containing the gradient of the energy potential, more efficient sampling mechanisms can be constructed. This idea started around 1978 (e.g. [Rossky et al., 1978]) and matured in the Hybrid Monte Carlo method (c.f. [Duane et al., 1987], [Neal, 1993] and [Roberts and Tweedie, 1996a]). A Langevin diffusion equation has the following form:

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t \tag{1.2}$$

where B_t is a Brownian motion and U is the energy potential. One can use

$$Z_{n+1} = Z_n - \gamma_n[\nabla U(Z_n) + \varepsilon_n] + \sqrt{\gamma_n}\xi_n$$

to generate samples of the stochastic process defined by (1.2). Under certain assumptions on the step size sequence, the samples generated by this algorithm converge to the stationary distribution (c.f. [Pelletier, 1998, Theorem 7]). In this thesis we describe the theory of Kolmogorov semigroups and show that it is a very apt tool for analyzing both diffusion equations and their discretization schemes

We use the following main sources: the basic principles behind our approach first appeared in [Welling and Teh, 2011] as a conjecture, and proofed by [Teh et al., 2014] and [Chen et al., 2015]. Together with [Ma et al., 2015], these are the main sources for

this thesis. Furthermore, we use the following text books: for measure and probability theory and stochastic processes, [Elstrodt, 2009], [Klenke, 2008], [Øksendal, 1998] and [Durrett, 2004]. For functional analysis [Alt, 2006] and [Werner, 2011]. For numerical integration [Kloeden and Platen, 1999] and [Milstein and Tretyakov, 1995] and for semigroup theory [LeVarge, 2003], [Pazy, 1983] and [Butzer and Berens, 1967].

The thesis is organized as follows: first we introduce Itô diffusion equations and describe the most important properties of their solutions in Chapter 2. The Kolmogorov semigroup and its Dual are subject to Chapter 3. The dual semigroup is use to determine the stationary distribution of our diffusion equations in Chapter 4. Chapter 5 will deal with numerical methods for integrating Itô diffusion equations. In Chapter 6 we will give conditions for the step size sequences under which these approximation schemes give consistent estimators for a stationary distribution of the continuous equation. Finally, we will present some simulation results in Chapter 7. Appendix A and B discuss interesting properties of Itô diffusions. Last but not least, in Appendix C an interesting result for the Variance of integration schemes is presented. This result allows to significantly simplify some of the consistency proofs.

Chapter 2

Diffusion equations

In this chapter we will discuss so-called *time-homogeneous Itô diffusions*. These are processes such that the diffusion equation with time in-dependent drift b and diffusion function σ permits a unique solution. We will further discuss some important properties of Itô diffusions. The approach presented here closely follows chapter 7 of [Øksendal, 1998].[1]

Definition 2.1 ([Øksendal, 1998], Def. 7.1.1)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space. An **Itô diffusion** is a stochastic process $X_t(\omega) = X(t, \omega) : [0, \infty) \times \Omega \to \mathbb{R}^n$ which satisfies the stochastic differential equation*

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \quad t \geq s \;\; X_s = x \tag{2.1}$$

*with Brownian Motion B_t. The functions $b : \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ are called **drift** and **diffusion**, respectively. They are assumed to be measurable and to fulfill the following conditions:*

$$|b(x)| + |\sigma(x)| \leq C(1 + |x|), \;\; x \in \mathbb{R}^n \text{ and} \tag{2.2}$$

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq D|x - y|, \;\; x, y \in \mathbb{R}^n. \tag{2.3}$$

We use the absolute value in in the sense of the 1-norm:

$$|b(x)| = \sum_{i=1}^{n} |b(x)| \quad \text{and} \quad |\sigma(x)| = \sum_{i=1}^{n} \sum_{j=1}^{m} |\sigma_{i,j}(x)|.$$

Remark. *We denote the solution to (2.1) as $X_t^{s,x}$. For $s = 0$ we abbreviate X_t^x. For $s = 0$ and $x = 0$, simply X_t.*

Equation (2.1) is a shorthand notation for the integral equation

$$X_t = x + \int_s^t b(X_r)dr + \int_s^t \sigma(X_r)dB_r \tag{2.4}$$

[1] In [Chen et al., 2015], which we will refer to later, Itô diffusion are called Langevin diffusions.

where the integral "$\int_s^t \sigma(X_r)dB_r$" is to be interpreted in the Itô sense: let $n \in \mathbb{N}$ and define a sequence of time steps $s = t_0 < t_1 < \cdots < t_n = t$. Then, for $\omega \in \Omega$,

$$\int_s^t \sigma\left(X(r,\omega)\right) dB_r(\omega) := \lim_{n\to\infty} \sum_{j=0}^{n} \sigma(X(t_j,\omega))[B_{t_{j+1}}(\omega) - B_{t_j}(\omega)]$$

for measurable functions e_j belong to the following class of functions (c.f. [Øksendal, 1998, Chapter 3]):

Definition 2.2. *(c.f. [Øksendal, 1998, Definition 3.1.4]) Let $\mathcal{V}$ be the class of functions $f(t,\omega) : [0,\infty) \times \Omega \to \mathbb{R}$ such that:*

1. *$(t,\omega) \to f(t,\omega)$ is $\mathcal{B}([0,\infty)) \times \mathcal{F}$-measurable.*
2. *$f(t,\omega)$ is $\mathcal{F}_t$-adapted*
3. *$E[\int_s^t f(t,\omega)^2 dt] < \infty$.*

Further details regarding the construction of the Itô integral can be found in chapter 3 of [Øksendal, 1998]. There the reader will find also the definition of **multi-dimensional Itô integrals** used in this thesis. An alternative definition would be the notion of the Stratonovich integral using a midpoint interpolation rule instead of the left point of the time interval $[t_i, t_{i+1}]$ (c.f. [Øksendal, 1998], p. 21).

The σ-algebra $\mathcal{F}_t$ used in Definition 2.2 is the σ-algebra generated by the Brownian Motion B_t. Since for $t \neq s$, B_t and B_s are independent, the σ-algebra

$$\mathcal{F}_t := \sigma(\{B_s(\cdot) : 0 \leq s \leq t\})$$

is a **filtration**, i.e. $\mathcal{F}_s \subset \mathcal{F}_t$ for $0 \leq s < t$. Note that the random component of the Itô diffusion stems from the randomness in the Brownian motion only and it turns out that the Itô diffusion $X_t^{s,x}$ is $\mathcal{F}_t$ measurable for all $t \geq s$ and for all $x \in \mathbb{R}^n$. Figure 2.1 shows three sample paths ($w1$, $w2$ and $w3$) of the standard Brownian motion starting in $B_0 = 0$.

Theorem 2.3 ([Øksendal, 1998], Thm. 5.2.1)**.** *Let $T > 0$ and let Z be a random variable with $E|Z|^2 < \infty$ which is independent of $\mathcal{F}_\infty$. Let $X_t^{s,Z}$ be an Itô diffusion with random start point Z. Then $X_t^{s,Z}$ is the unique solution to (2.1) with random start value $x = Z$ and $X_t^{s,Z}$ is almost surely t-continuous, i.e.*

$$P(\{\omega \in \Omega : t \mapsto X_t^{s,Z(\omega)}(\omega) \ \text{ is continuous}\}) = 1.$$

Furthermore, $X_t^{s,Z}(\omega)$ is adapted to the filtration $\mathcal{F}_t^Z$ generated by Z and $\mathcal{F}_t$ and we have

$$E\left[\int_0^T |X_t^{s,Z}|^2 dt\right] < \infty.$$

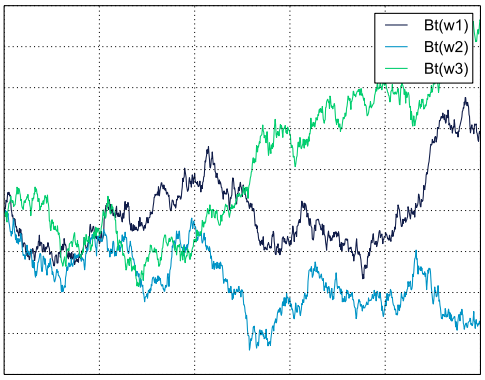


Figure 2.1: Three sampling paths of ths standard Brownian motion

Remark. *Let $Z \in \eta$ be a random variable. We use the notation X_t^Z as in [Øksendal, 1998, Theorem 5.2.1]. The interpretation is as follows: For $\omega \in \Omega$ we have*

$$X_t(\omega) = Z(\omega) + b(X_t(\omega))dt + \sigma(X_t(\omega))dB_t(\omega),$$

i.e. $X(0, \omega) = Z(\omega)$.

Definition 2.4. *(c.f. [Øksendal, 1998, page 9], [Durrett, 2004, 1.3 and 1.3.c], [Klenke, 2008, Definition 1.103]) Let $X_t^Z : \Omega \to \mathbb{R}^n$ be an Itô diffusion and let $\phi : \mathbb{R}^n \to \mathbb{R}$ be a Borel-measurable function. Then the **expectation** of $\phi\left(X_t^Z\right)$ is defined as*

$$E\left[\phi\left(X_t^Z\right)\right] := \int_\Omega \phi\left(X_t^Z(\omega)\right) dP(\omega) = \int_{\mathbb{R}^n} \phi(y)\, \mu_t^Z(dy). \tag{2.5}$$

*The function $\mu_t^Z : \mathcal{B}(\mathbb{R}^n) \to \mathbb{R}^+$, $\mu_t^Z = P \circ (X_t^Z)^{-1}$ is called the **distribution** of X_t^Z and it is a probability measure on $\mathcal{B}(\mathbb{R}^n)$.*

Using this definition we can explain in closer detail what we mean with expressions like $E[\int_0^T X_t^x dt]$ used in Theorem 2.3 above: Let $t \geq 0$. We can apply Fubini's Theorem [Elstrodt, 2009, V. Theorem 2.1 c)] to see:

$$E\left[\int_0^t |X_s^x|^2 ds\right] = \int_0^t E\left[|X_s^x|^2\right] ds < \infty.$$

It follows directly that $E[\,|X_t^x|^2\,] < \infty$ for almost all t. Using Jensen's inequality (c.f. [Klenke, 2008, Theorem 8.19]) and the Cauchy-Schwartz inequality (e.g. [Elstrodt, 2009,

VI. 1.6]), we also get:

$$E\left[\int_0^t |X_s^x| ds\right]^2 \leq \int_\Omega \left|\int_0^t |X_s^x(\omega)|\, ds\right|^2 P(d\omega) = \int_\Omega \left|\int_{\mathbb{R}} \chi_{[0,t]}(s)|\chi_{[0,t]}(s) X_s^x(\omega)|\, ds\right|^2 P(d\omega)$$
$$\leq t \int_\Omega \int_0^t |X_s^x(\omega)|^2\, ds P(d\omega) = tE\left[\int_0^t |X_s^x|^2 dt\right] < \infty.$$

Since the left-hand side is finite, we can apply Fubini's Theorem again and get:

$$E\left[\int_0^t X_s^x ds\right] = \int_0^t E\left[X_s^x\right] ds \quad \text{and} \quad E\left[|X_t^x|\right] < \infty \ \text{ for almost all } t \geq 0.$$

2.1 The Markov Property

If we consider two Itô diffusions starting at the same $x \in \mathbb{R}^n$, but at different times, then in general we have $X_t^{0,x}(\omega) \neq X_{t+h}^{h,x}(\omega)$ for $\omega \in \Omega$. However, it makes sense that if we have two processes with the same initial value but with a different starting time and we let them run for the same amount of time, that their behavior should be quite similar.

Lemma 2.5. *(c.f. [Øksendal, 1998, proof of Theorem. 7.1.2]) For an Itô diffusion X_t^Z and $t, s \geq 0$ we have $X_{t+s}^{0,Z} = X_{t+s}^{s,X_s^Z}$.*

Proof. Let $\omega \in \Omega$. By uniqueness,

$$X_{t+s}^Z(\omega) = Z(\omega) + \int_0^{t+s} b(X_r^Z(\omega))dr + \int_0^{t+s} \sigma(X_r^Z(\omega))dB_r(\omega)$$
$$= X_t^Z(\omega) + \int_t^{t+s} b(X_r^Z(\omega))dr + \int_t^{t+s} \sigma(X_r^Z(\omega))dB_r(\omega) = X_t^{s,X_s^Z}(\omega).$$

□

Lemma 2.6 (c.f. [Øksendal, 1998], p.110)**.** *Itô diffusions are time homogeneous, i.e.*

$$\{X_{t+h}^{t,x}\} =^d \{X_h^{0,x}\}$$

Proof. We have

$$X_{t+h}^{t,x} = x + \int_t^{t+h} b(X_r^{t,x})dr + \int_t^{t+h} \sigma(X_r^{t,x})dB_r = x + \int_0^h b(X_{t+r}^{t,x})dr + \int_0^h \sigma(X_{t+r}^{t,x})d\tilde{B}_r$$

with $\tilde{B}_r = B_{s+r} - B_s$ for $r \geq 0$. Note that the process $Y_h^{0,x} := X_{t+h}^{t,x}$ satisfies the equation

$$dY_h = b(Y_h)dh + \sigma(Y_h)d\tilde{B}_h, \quad h \geq 0 \ \ Y_0 = x$$

which is exactly the equation of $X_h^{0,x}$, with B_h replaced by $\tilde{B}_h$:

$$dX_h = b(X_h)dh + \sigma(X_h)dB_h, \quad h \geq 0 \;\; X_0 = x$$

Since $\{\tilde{B}_r\}_{r\geq 0}$ and $\{B_r\}_{r\geq 0}$ have the same probability distributions, we can conclude that Y_h^x and X_h^x are identical in law by [Øksendal, 1998, Lemma 5.3.1] such that $\{X_{s+h}^{s,x}\}$ and $\{X_h^{0,x}\}$ have the same distribution. □

This result directly implies $E[Y_h^x] = E[X_h^x]$ in the setting of the proof and leads us to the following **Markov Property**:

Theorem 2.7. *(c.f. [Øksendal, 1998, Theorem. 7.1.2]) Let φ be a bounded Borel function $f : \mathbb{R}^n \to \mathbb{R}$. Then for $t, h \geq 0$ and $\omega \in \Omega$,*

$$E\left[\varphi(X_{t+h}^x)|\mathcal{F}_t\right](\omega) = E\left[\varphi(X_h^y)\right]_{y=X_t(\omega)} \tag{2.6}$$

Proof. Let $\omega \in \Omega$. By Lemma 2.5, $X_{t+h}^x = X_{t+h}^{t,X_t^x}$ The function $g : (y, \omega) \to \varphi \circ X_{t+h}^{t,y}(\omega)$ is measurable and integrable (c.f. [Øksendal, 1998, Exercise 7.6]). Therefore, we can approximate g point-wise by measurable functions ϕ and ψ:

$$g(y, \omega) = \lim_{m\to\infty} \sum_{i=1}^{m} \phi_i(y)\psi_i(\omega).$$

By the theorem of bounded convergence (c.f. [Elstrodt, 2009, IV. Theorem 5.2]),

$$\begin{aligned} E[g(X_t^x, \cdot)|\mathcal{F}_t] &= E[\lim_{m\to\infty} \sum_{i=0}^{m} \phi_i(X_t^x)\psi_i|\mathcal{F}_t] = \lim_{m\to\infty} \sum_{i=0}^{m} E[\phi_i(X_t^x)\psi_i|\mathcal{F}_t] \\ &= \lim_{m\to\infty} \sum_{i=0}^{m} \phi_i(X_t^x)E[\psi_i|\mathcal{F}_t] = \lim_{m\to\infty} \sum_{i=0}^{m} E[\phi_i(y)\psi_i|\mathcal{F}_t]_{y=X_t^x} \\ &= E[g(y, \cdot)|\mathcal{F}_t]_{y=X_t^x} \end{aligned}$$

And thus:

$$E[\varphi(X_{t+h}^x)|\mathcal{F}_t] = E[\varphi(X_{t+h}^{t,X_t})|\mathcal{F}_t] = E[g(X_t^x, \cdot)|\mathcal{F}_t] = E[g(y, \cdot)|\mathcal{F}_t]_{y=X_t^x} = E[\varphi(X_{t+h}^{t,y})|\mathcal{F}_t]_{y=X_t}$$

Furthermore note that $X_{t+h}^{t,y}$ is independent of $\mathcal{F}_t$, i.e. by time homogeneity:

$$E[\varphi(X_{t+h}^{t,y})|\mathcal{F}_t]_{y=X_t} = E[\varphi(X_{t+h}^{t,y})]_{y=X_t} = E[\varphi(X_h^y)]_{y=X_t}$$

which completes the proof. □

Chapter 3

Kolmogorov semigroups

Our first goal is to show how we can construct stochastic processes with specific distributions. The procedure is as follows: First we will describe the expectation $E[\phi(X_t^x)]$ for bounded functions $\phi \in C_b(\mathbb{R}^n)$. We will see by Riesz' representation theorem that we can use the expectation to describe how the distributions μ_t^x of X_t^x develop over time. For this we introduce an operator $T(t)\phi(x) = E[\phi(X_t^x)]$. We show that T is strongly right-continuous at the origin $t \downarrow 0$ which will be important for showing existence of solutions to the Cauchy problem in the subsequent chapter below.

Definition 3.1. *For any $x \in \mathbb{R}^n$ let X_t^x be an Itô diffusion with distribution μ_t^x. For $t \geq 0$, we define the **Kolmogorov operator** $T(t) = T_X$ as*

$$\begin{aligned} &T(t) : C_b(\mathbb{R}^n) \to C_b(\mathbb{R}^n) \\ &T(t) : \phi \mapsto \int_{\mathbb{R}^n} \phi(y)\, \mu_t^{(\cdot)}(dy) = E[\phi(X_t^{(\cdot)})] \quad \textit{for } \phi \in C_b(\mathbb{R}^n). \end{aligned} \tag{3.1}$$

Theorem 3.2. *The Kolmogorov operator of an Itô diffusion X_t^x is well-defined.*

Proof. The space $C_b(\mathbb{R}^n) = \{\phi \in C(\mathbb{R}^n) : \sup_{x\in\mathbb{R}^n} |\varphi(x)| \leq M < \infty\}$ equipped with the sup-norm $\|\phi\|_\infty := \sup_{x\in R^n} |\phi(x)|$ is a Banach space (c.f. [Alt, 2006, Section 1.2]). Furthermore, for any $x \in \mathbb{R}^n$, any $t \geq 0$ and any $\phi \in C_b(\mathbb{R}^n)$ (i.e. $\sup_{x\in\mathbb{R}^n} \phi(x) \leq M$ for some $0 \leq M < \infty$):

$$[T(t)\phi](x) = E[\phi(X_t^x)] = \int_{\mathbb{R}^n} \phi(y)\mu_t^x(dy) \leq M\mu_t^x(\mathbb{R}^n) = M < \infty$$

such that $T(t)\phi \in C_b(\mathbb{R}^n)$ again. □

Remark. *The approach of defining the Kolmogorov operator is for example applied in [Øksendal, 1998, Definition 7.3.1.] and [Chen et al., 2015, Section 2]. The latter also coined the term "Kolmogorov operator", which originates in the so-called Kolmogorov equations discussed in section 4.1 below.*

Operators like T are the subjects the theory of semigroups deals with. A comprehensive introduction to this topic can be found in [LeVarge, 2003]. The detailed theory has been summarized by e.g. [Pazy, 1983] and [Butzer and Berens, 1967].

3.1 Semigroups

Definition 3.3 (c.f. [LeVarge, 2003], p.3; [Pazy, 1983] Def. 1.1)**.** *Let S be a Banach space with norm $\|\cdot\|_S$. A family of operators $\{T(t) : S \to S : t \geq 0\}$ is called a* ***one-parameter semigroup*** *iff*

- $T(t+s) = T(t)\,T(s)$ *for all* $(t, s \geq 0)$ *and*
- $T(0) = Id_S : \phi \mapsto \phi$ *for all* $\phi \in S$.

A semigroup of operators is called ***linear*** *if all the operators are linear. It is called* ***strongly continuous*** *or a $\mathcal{C}_0$-semigroup if additionally*

- $\lim_{t \downarrow 0} \|T(t)\,\phi - \phi\|_S = 0$ *for all* $\phi \in S$.

Remark. *The equation $T(t+s) = T(t)\,T(s)$ is also known as the* ***Chapman-Kolmogorov equation*** *(c.f. [Klenke, 2008, Definition 14.40]).*

Of course our goal is to show that the operator we defined above forms a $\mathcal{C}_0$-semigroup. We start out with the weak version:

Lemma 3.4. *For any $x \in \mathbb{R}^n$ let X_t^x be an Itô process. The family of Kolmogorov operators $\{T(t) : t \geq 0\}$ as defined in* (3.1) *form a linear semigroup, the* ***Kolmogorov semigroup****.*

Proof. (c.f. [Øksendal, 1998, proof of Theorem 8.1.1]) It is clear that the Kolmogorov operators are linear because the expectation is linear. In order to show the semigroup property, take any $\phi \in C_b(\mathbb{R}^n)$. Then,

$$T(0)\,\phi(x) = E[\phi(X_0^x)] = E[\phi(x)] = \phi(x).$$

Next we will use the Markov property of X_t^x to show the Chapman-Kolmogorov equation for T. In order to do this, we follow [LeVarge, 2003, page 2]: Let $u(t,x) := E[\phi(X_t^x)]$ such that $T(t)\,\phi(x) = u(t,x)$. Then we have

$$\begin{aligned} T(t+s)\,\phi(x) &= E[\phi(X_{t+s}^x)] = E\left[E[\phi(X_{t+s}^x)|\mathcal{F}_s]\right] \\ &= E\left[E[\phi(X_t^y)]_{y=X_s^x}\right] = E[u(t, X_s^x)] \\ &= T(s)\,u(t,x) = T(s)\,T(t)\,\phi(x) \end{aligned}$$

for all $x \in \mathbb{R}^n$. We made use of the tower property of conditional expectations ([Klenke, 2008], Thm. 8.14 (iv) with $\mathcal{G} = \mathcal{F}_s$) and the Markov property of Itô diffusions. □

Theorem 3.5. *The family of Kolmogorov operators is a $\mathcal{C}_0$-semigroup.*

Proof. The above Lemma yields that the Kolmogorov operators form a linear one-parameter semigroup. Now we have to show that for all $\phi \in C_b(\mathbb{R}^n)$,

$$\lim_{t\downarrow 0} \|T(t)\,\phi - \phi\|_\infty = \lim_{t\downarrow 0} \sup_{x\in\mathbb{R}^n} |T(t)\phi(x) - \phi(x)| = 0. \tag{3.2}$$

Let $x \in \mathbb{R}^n$. According to Theorem 2.3, X_t^x is a t-continuous stochastic process almost surely. Now choose $\varepsilon > 0$ arbitrarily. Because of Fatou's Lemma and because X_t^x is almost surely continuous, we get:

$$\begin{aligned}
\liminf_{t\downarrow 0} P(\{\omega : \|X_t^x(\omega) - x\| \le \varepsilon\}) &\ge P(\liminf_{t\downarrow 0}\{\omega : \|X_t^x(\omega) - x\| \le \varepsilon\}) \\
&= P(\{\omega : \exists \delta > 0 \text{ s.t. } \|X_t^x(\omega) - x\| \le \varepsilon \ \ \forall\, 0 \le t \le \delta\}) \\
&\ge P(\{\omega : \forall \varepsilon > 0\ \exists \delta > 0 \text{ s.t. } \|X_t^x(\omega) - x\| \le \varepsilon \ \ \forall\, 0 \le t \le \delta\}) \\
&= 1 \\
\text{and} \quad \lim_{t\downarrow 0} P(\|X_t^x(\omega) - x\| > \varepsilon) &\le \limsup_{t\downarrow 0} P(\|X_t^x(\omega) - x\| > \varepsilon) \le 0
\end{aligned}$$

which is the definition of convergence of X_t^x to x in probability (c.f. [Klenke, 2008, Definition 6.2. i)]). According to a Corollary of Slutzky's theorem ([Klenke, 2008, Theorem 13.18 and Corollary 13.19]), it also converges in distribution. By [Klenke, 2008, Definition 13.17], this means that $\mu_t^x \to_w \mu_0^x$ as $t \downarrow 0$ weakly. By [Klenke, 2008, Definition 13.12]:

$$\lim_{t\downarrow 0} \sup_{x\in\mathbb{R}^n} |T(t)\phi(x) - \phi(x)| = \lim_{t\downarrow 0} \sup_{x\in\mathbb{R}^n} |E[\phi(X_t^x)] - E[\phi(X_0^x)]| = 0$$

for all continuous bounded functions $\phi \in C_b(\mathbb{R}^n)$. Finally, let $Z \in \mathbb{R}^n$ be a random variable with distribution η (possibly $\eta = x$). Then we have

$$\left|E[\phi(X_t^Z)] - E[\phi(Z)]\right| \le \sup_{x\in\mathbb{R}^n} |E[\phi(X_t^x)] - E[\phi(x)]| = \|T(t)\,\phi - \phi\|_\infty \to 0.$$

This proofs that the Kolmogorov semigroup is indeed strongly continuous. □

3.2 The Dual Semigroup

The governing dynamics of the expectation $E[\phi(X_t^x)]$ stem from changes in the probability distributions μ_t^x of X_t^x. From the Markov property discussed in Theorem 2.7, the Kolmogorov semigroup is also known as the **Markovian (transition) semigroup** (e.g. [Prato and Zabczyk, 2003], [Prato, 2004]). Therefore, if $\mu_t^{s,x}$ permits a density $p_t^{s,x}$, we can also view it as a transition kernel $f(s,x,t,y) := p_t^{s,x}(y)$. It describes how $X_t^{s,x}$ transitions from the state $x \in \mathbb{R}^n$ at time $s > 0$ to the state $y \in \mathbb{R}^n$ at time $t \ge s$.

The dynamics of the transition measures are most naturally describable in terms of the dual semigroup. In order to define it, however, will need the following notions first:

Definition 3.6. *([Elstrodt, 2009, VII. Section 5 and Theorem 1.9]) Let $M(\mathbb{R}^n)$ be the set of* ***Borel measures*** *on $\mathbb{R}^n$. Let*

$$\mu^+(A) := \sup\{\mu(B) : B \in \mathcal{B}(\mathbb{R}^n), B \subset A\}$$
$$\text{and} \quad \mu^-(A) := -\inf\{\mu(B) : B \in \mathcal{B}(\mathbb{R}^n), B \subset A\}$$

be the ***positive*** *and the* ***negative variation*** *of $\mu \in M(\mathbb{R}^n)$, respectively [1]. Then the* ***total variation*** *of μ is defined as*

$$\|\mu\| := \mu^+(\mathbb{R}^n) + \mu^-(\mathbb{R}^n), \tag{3.3}$$

Furthermore, the set of ***finite Radon measures*** *on $\mathcal{B}(\mathbb{R}^n)$ is the set of inner regular measures of finite variation ([Elstrodt, 2009, VIII Definition 1.1]):*

$$\mathcal{M}(\mathbb{R}^n) = \left\{\mu \in M(\mathbb{R}^n) : \|\mu\| < \infty,\ \mu(A) = \sup_{K \subset A, K \in \mathcal{K}} \mu(K) \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^n)\right\}. \tag{3.4}$$

where $\mathcal{K}$ is the set of compact measurable sets. We also define the ***non-negative Radon measures*** *as $\mathcal{M}_+ := \{\mu \in \mathcal{M} : \mu \geq 0\}$.*

Remarks. *(a) The total variation is a norm on $M(\mathbb{R}^n)$ and the tuple $(\mathcal{M}, \|\cdot\|)$ is a* ***Banach space*** *(c.f. [Elstrodt, 2009, VII. Theorem 1.14]).*

(b) For a positive measure $\mu \geq 0$, we have $||\mu|| = \sup\{\mu(A) : A \in \mathcal{A}\}$. This follows directly from the definition of $\mu^+(\mathbb{R}^n)$.

(c) The measures in $\mathcal{M}$ are also regarded to as $rca(\mathbb{R}^n)$ (e.g. [Alt, 2006, p.185]).

(d) For measures $\mu \in \mathcal{M}$ with a density f, the total variation can also be defined via the $L^1(\mathbb{R}^n, \lambda)$ norm (c.f. [Elstrodt, 2009, VII. Example 1.13]):

$$||\mu|| = ||f||_{L^1(\mathbb{R}^n,\lambda)} = \int_{\mathbb{R}^n} |f(x)|\, \lambda(dx) \tag{3.5}$$

The following lemma will be very important for showing continuity of the dual semigroup:

Lemma 3.7. *All Radon measures in $\mathcal{M}(\mathbb{R}^n)$ are regular. This means they are outer and inner regular (c.f. [Elstrodt, 2009, VIII. Definition 1.1]):*

$$\mu(A) = \sup\{\mu(K) : K \subset A, K \in \mathcal{K}\}$$
$$= \inf\{\mu(O) : A \subset O, O \in \mathcal{O}\}$$

where $\mathcal{O}$ is the set of open measurable sets. In particular, all ***probability measures*** *$\mathcal{M}_1 := \{\mu \in M(\mathbb{R}^n) : \mu \geq 0 \wedge \|\mu\| = 1\}$ are regular and $\mathcal{M}_1 \subset \mathcal{M}$.*

[1] Positive and negative variation are usually defined using the **Hahn decomposition** (c.f. [Elstrodt, 2009, VII. Theorem 1.8]): let $A \in \mathcal{B}(\mathbb{R}^n)$. A has got a positive part P and a negative part N such that P and N are disjoint and $A = P \cup N$. Furthermore, $\mu(A) := \mu^+(P) - \mu^-(N)$ with $\mu^+(A) := \mu(A \cap P)$ and $\mu^-(A) := -\mu(A \cap N)$ (c.f. [Elstrodt, 2009, VII. Section 3]).

Proof. Note that $\mathbb{R}^n$ is a separable Banach space. By [Elstrodt, 2009, Appendix A.22], it is a Polish space. Since all measures in $\mathcal{M}$ are finite Borel measures, Ulam's Theorem [Elstrodt, 2009, VIII. Theorem 1.16] states that they are also regular. $\mathcal{M}_1 \subset \mathcal{M}$ follows immediately from the definitions. □

We now characterize the **domain** of the dual Kolmogorov operator, which is given by the **Riesz' representation theorem**:

Theorem 3.8. *([Elstrodt, 2009, VIII. Theorem 2.12]) The dual space of $C_b(\mathbb{R}^n)$ is the set $\mathcal{M}$ of finite Radon measures $\mathcal{M}$ equipped with the total variation norm.*

The dual space of $C_b(\mathbb{R}^n)$ is defined as the space of all linear forms $I : C_b(\mathbb{R}^n) \to \mathbb{R}$. Now Riesz' representation theorem tells us that for every $\phi \in C_b(\mathbb{R}^n)$ and any positive[2] linear form I there is a unique finite Radon measure μ such that we can express it with the **dual bracket** $I[\phi] = \langle \phi, \mu \rangle := \int_{\mathbb{R}^n} \phi(y)\, \mu(dy)$. In particular, if Z is a random variable, then the linear form $\phi \mapsto E[\phi(X_t^Z)]$ can uniquely be expressed as

$$E[\phi(X_t^Z)] = \int_{\mathbb{R}^n} \phi(y)\, \mu_t^Z(dy) = \left\langle \phi, \mu_t^Z \right\rangle .$$

Definition 3.9. *Let $\mathcal{M}$ be the set of finite Radon measures and $\eta \in \mathcal{M}$. Then for $t \geq 0$, the **dual Kolmogorov operator** is defined as*

$$T^*(t) : \mathcal{M} \to \mathcal{M}, \quad \langle \phi, T^*(t)\, \eta \rangle := \langle T(t)\, \phi, \eta \rangle \tag{3.6}$$

Theorem 3.10. *(c.f. [Butzer and Berens, 1967, Proposition 1.4.3]) The family of dual Kolmogorov operators $T^* = \{T^*(t) : t \geq 0\}$ is a semigroup.*

Proof. We first have to show that $T^*(t)$ is well-defined for all $t \geq 0$. If $\eta \in \mathcal{M}_1$, then according to [Klenke, 2008, Theorem 1.104] there exists a random variable $Z \sim \eta$ and

$$\begin{aligned}\langle T(t)\phi, \eta \rangle &= \int_{\mathbb{R}^n} [T(t)\phi](y)\eta(dy) = \int_{\mathbb{R}^n} E[\phi(X_t^y)]\eta(dy) = \int_{\mathbb{R}^n}\int_\Omega \phi(X_t^y(\omega))P(d\omega)\eta(dy) \\ &= \int_\Omega \int_\Omega \phi(X_t^{Z(\tilde{\omega})}(\omega))P(d\omega)P(d\tilde{\omega}).\end{aligned}$$

From [Elstrodt, 2009, V. Theorem 1.5] we know that there exists a unique product measure $\tilde{P} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^+$ such that $\tilde{P}(A \times B) = P \otimes P(A \times B) = P(A)P(B)$. This product measure is also a probability measure. By [Klenke, 2008, Theorem 1.104] there also exists a random variable $\tilde{X}_t$ such that:

$$\int_\Omega \int_\Omega \phi(X_t^{Z(\tilde{\omega})}(\omega))P(d\omega)P(d\tilde{\omega}) = \int_\Omega \int_\Omega \phi(\tilde{X}_t^Z(\omega, \tilde{\omega}))(P \otimes P)(d\omega \times d\tilde{\omega}) = E\left[\phi(\tilde{X}_t^Z)\right]$$

[2]Here we mean "positive" in the sense that $\phi \geq 0 \Rightarrow E[\phi(X_t^x)] \geq 0$ (c.f. [Elstrodt, 2009, VIII. §2 1.]).

where $\tilde{X}_t^Z : \Omega \times \Omega \to \mathbb{R}^n$ is a random variable with distribution $\tilde{\mu}_t^Z$ such that the equation $\langle T(t)\phi, \eta\rangle = \langle \phi, \tilde{\mu}_t^Z\rangle$ holds and $T^*(t)\eta = \tilde{\mu}_t^Z$ in the weak sense.

If $\eta \in \mathcal{M}_+$, we can normalize it as $\eta = \|\eta\|\bar{\eta}$ with $\bar{\eta} \in \mathcal{M}_1$ and

$$\langle \phi, T^*(t)\eta\rangle = \|\eta\|\langle T(t)\phi, \bar{\eta}\rangle = \|\eta\|\langle \phi, \tilde{\mu}_t^{\bar{\eta}}\rangle$$

such that $\|\eta\|\mu_t^{\bar{\eta}} \in \mathcal{M}_+$. If however $\eta \in \mathcal{M}$, then the positive part η^+ and the negative part with inverse sign, $-\eta^-$, both are in $\mathcal{M}_+$. and by linearity of the linear form, $T^*(t)$ is well-defined.

The semigroup properties basically follow from the semigroup properties of T:

$$\begin{aligned} \langle \phi, T^*(0)\eta\rangle &= \langle T(0)\phi, \eta\rangle = \langle \phi, \eta\rangle \\ \text{and} \quad \langle \phi, T^*(t)T^*(s)\eta\rangle &= \langle T(s)T(t)\phi, \eta\rangle = \langle T(t+s)\phi, \eta\rangle = \langle \phi, T^*(t+s)\eta\rangle \end{aligned}$$

for all $\phi \in C_b(\mathbb{R}^n)$. □

If $\eta \in \mathcal{M}_1$, then we can also give a direct characterization of $T^*(t)$:

Theorem 3.11. *Let $Z \sim \eta$ be a random variable and let $T^*(t)$ and be the dual Kolmogorov operator. Then we have*

$$T^*(t)\eta = \mu_t^Z. \tag{3.7}$$

Proof. Let $t \geq 0$, $\phi \in C_b(\mathbb{R}^n)$ and $A \in \mathcal{B}(\mathbb{R}^n)$. This result follows from the Markov property of Itô diffusions (Theorem 2.7):

$$\begin{aligned} \langle \phi, T^*(t)\eta\rangle = \langle T(t)\phi, \eta\rangle &= \int_{\mathbb{R}^n} E[\phi(X_t^x)]\eta(dx) = \int_{\mathbb{R}^n} E[\phi(X_t^x)]|_{x=Z(\omega)}P(d\omega) \\ &= \int_{\mathbb{R}^n} E[\phi(X_t^x)]|_{x=X_0^Z(\omega)}P(d\omega) = \int_{\mathbb{R}^n} E[\phi(X_t^Z)|\mathcal{F}_0](\omega)P(d\omega) \\ &= E\left[E[\phi(X_t^Z)|\mathcal{F}_0]\right] = E[\phi(X_t^Z)] = \langle \phi, \mu_t^Z\rangle. \end{aligned}$$

The last steps are due to how the conditional expectation is defined ($E[E[\phi(X_t^Z)|\mathcal{F}_0]] = E[\phi(X_t^Z)]$, c.f. [Klenke, 2008, Definition 8.11]). Thus we have

$$\langle \phi.T^*(t)\eta\rangle = \langle \phi, \mu_t^Z\rangle \text{ for all } \phi \in C_b(\mathbb{R}^n). \tag{3.8}$$

According to [Elstrodt, 2009, VIII. Theorem 4.6], we have indeed $T^*(t)\eta = \mu_t^Z$.

For further details about this entity, also refer to Appendix A. □

It turns out we can even show strong continuity for the dual Kolmogorov operator:

Theorem 3.12. *The dual semigroup of the Kolmogorov operator is strongly continuous.*

Proof. Let η be a measure in $\mathcal{M}_1$ and $Z \sim \eta$. Let X_t^Z be an Itô diffusion with distribution μ_t^Z. According to [Elstrodt, 2009, VII. Theorem 1.8] we can express the total variation as

$$\begin{aligned}\|\mu_t^Z - \eta\| &= (\mu_t^Z - \eta)^+(\mathbb{R}^n) + (\mu_t^Z - \eta)^-(\mathbb{R}^n) = (\mu_t^Z - \eta)(P) - (\mu_t^Z - \eta)(N) \\ &= \mu_t^Z(P) - \mu_t^Z(N) - \eta(P) + \eta(N) = \int_{\mathbb{R}^n} (\chi_P - \chi_N)\mu_t^Z - \int_{\mathbb{R}^n} (\chi_P - \chi_N)\eta\end{aligned}$$

with Hahn decomposition $P \dot\cup N = \mathbb{R}^n$ and indicator function χ. We can rewrite these integrals as dual brackets. In particular for the first integral this implies that we can express it using the primal Kolmogorov using Theorem 3.11:

$$\int_{\mathbb{R}^n} (\chi_P - \chi_N)\mu_t^Z = \langle \chi_P - \chi_N, \mu_t^Z \rangle = \langle \chi_P - \chi_N, T^*(t)\eta \rangle = \langle T(t)(\chi_P - \chi_N), \eta \rangle.$$

We can use this to bound the total variation distance of μ_t^Z and η in the following way:

$$\|\mu_t^Z - \eta\| = \langle T(t)(\chi_P - \chi_N), \eta \rangle - \langle \chi_P - \chi_N, \eta \rangle \leq \|T(t)(\chi_P - \chi_N) - (\chi_P - \chi_N)\|_\infty \|\eta\|$$

Clearly, η is of finite total variation. Since $(\chi_P - \chi_N) \in C_b(\mathbb{R}^n)$ is bounded we can use strong continuity of $T(t)$ and see:

$$\lim_{t \downarrow 0} \|\mu_t^Z - \eta\| = 0 \quad \text{for all } \eta \in \mathcal{M}_1.$$

This shows the strong continuity of T^* for all $\eta \in \mathcal{M}_1$. Now let η be a signed measure $\eta = \eta^+ + \eta^-$ with $\eta^+, (-\eta^-) \in \mathcal{M}_1$ and two random variables $Z^+ \sim \eta^+$ and $Z^- \sim (-\eta^-)$. We can use the linearity of the dual bracket and apply the Kolmogorov operator to η:

$$\langle \phi, T^*(t)\eta \rangle = \langle T(t)\phi, \eta^+ \rangle - \langle T(t)\phi, -\eta^- \rangle = \langle \phi, T^*(t)\eta^+ \rangle - \langle \phi, T^*(t)(-\eta^-) \rangle = \langle \phi, \mu_t^{Z^+} - \mu_t^{Z^-} \rangle.$$

In the light of [Elstrodt, 2009, VIII. Theorem 4.6] and Theorem 3.11, this implies that $T^*(t)\eta = \mu_t^{Z^+} - \mu_t^{Z^-}$. Convergence in the total variation norm follows immediately:

$$\begin{aligned}\lim_{t \downarrow 0} \|T^*(t)\eta - \eta\| &= \lim_{t \downarrow 0} \|T^*(t)\eta^+ - T^*(t)\eta^- - \eta^+ + (-\eta^-)\| \\ &\leq \lim_{t \downarrow 0} \|T^*(t)\eta^+ - \eta^+\| + \lim_{t \downarrow 0} \|T^*(t)\eta^- - (-\eta^-)\| = 0.\end{aligned}$$

If in general $\eta \in \mathcal{M}$, then we can scale η^+ and $(-\eta^-)$ as we demonstrated in the proof of Theorem 3.10 and apply the same considerations to the scaled probability measures. The scaling factor can be pulled outside the linear form and because η^+ and $(-\eta^-)$ are finite, do not hinder our convergence argument. Hence, the dual Kolmogorov operator is strongly continuous. □

Remark. *This statement is stronger for the Kolmogorov operator than the standard result of semigroup theory (c.f. [Pazy, 1983, Theorem 10.4.], [Butzer and Berens, 1967, Corollary 1.4.8.])*[3].

3.3 Density semigroups

In practical applications, we are especially interested in random variables having a density function. We need the following definitions:

Definition 3.13. *(c.f. [Elstrodt, 2009, VII. Definition 2.1]) Let λ be the Lebesgue measure. A measure $\mu \in \mathcal{M}(\mathbb{R}^n)$ is said to be **absolutely continuous** with respect to λ, if*

$$\mu(A) = 0 \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^n) \text{ with } \lambda(A) = 0.$$

We define the set of all probability distributions having a density as

$$\mathcal{M}_1^{\ll}(\mathbb{R}^n) := \{\mu \in \mathcal{M}_1(\mathbb{R}^n) : \mu \ll \lambda\}$$

and the space of probability densities as

$$L_1^1(\mathbb{R}^n) := \{\rho : \rho \text{ is the density to some } \nu \in \mathcal{M}_1^{\ll}(\mathbb{R}^n)\} = \{\rho \in L^1(\mathbb{R}^n, \lambda) : \|\rho\|_1 = 1\}.$$

For diffusion processes having a density we can use the density version of the dual Kolmogorov operator:

Definition 3.14. *Let $Z \sim \eta \in \mathcal{M}_1^{\ll}$ with density ρ and let $X_t^Z \sim \mu_t^Z \in \mathcal{M}_1^{\ll}$ be an Itô diffusion which density p_t^Z. Then, for $t \geq 0$, the **Kolmogorov density operators** are defined as*

$$T^{\ll}(t) : L_1^1(\mathbb{R}^n) \to L_1^1(\mathbb{R}^n), \quad \langle \phi, T^{\ll}(t)\rho \rangle_1 := \langle \phi, T^*(t)\eta \rangle \tag{3.9}$$

where $\langle \phi, \rho \rangle_1 := \int_{\mathbb{R}^n} |\phi(y)\rho(y)|\lambda(dy) \leq \|\phi\|_\infty \|\rho\|_1$ (c.f. [Werner, 2011, Theorem I.1.10]).

Theorem 3.15. *The family of Kolmogorov density operators $T^{\ll} := \{T^{\ll}(t) : t \geq 0\}$ is a strongly continuous linear semigroup.*

Proof. We first show that the $T^{\ll}(t)$ are well-defined for $t \geq 0$. According to the Radon-Nikodym theorem [Elstrodt, 2009, VII. Theorem 2.3] the densities f of measures $\mu \in \mathcal{M}_1^{\ll}$ are given by the non-negative integrable functions $L^1(\mathbb{R}^n)$. Furthermore, we have $\|f\|_1 = \|\mu\| = 1$ and we write $f \in L_1^1(\mathbb{R}^n)$. Since $\mathcal{M}_1^{\ll}$ is a Banach space, also the

[3] **Proposition:** Let S be a Banach space and let T be a $\mathcal{C}_0$-semigroup with infinitesimal generator A. Then the dual semigroup T^*, restricted to the closure $\text{clos}(\mathcal{D}_{A^*})$ is a $\mathcal{C}_0$-semigroup on this closure. If S is reflexive, i.e. $S^* = S$, then $\text{clos}(\mathcal{D}_{A^*}) = \mathcal{D}_{A^*}$.

space $L_1^1(\mathbb{R}^n)$ is Banach and the Kolmogorov density operator is well-defined.

The semigroup property directly translates from the the semigroup property of T^*. It remains to show strong continuity. For this note the equality of norms which we have already seen in equation (3.5). Let $\mu \in \mathcal{M}_1^{\ll}$ be an absolutely continuous probability measure with density $\rho \in L_1^1(\mathbb{R}^n)$. Furthermore, let $\mathbb{R}^n = P \cup N$ be the Hahn-decomposition of the signed measure $T^*(t)\,\mu - \mu$. Then according to [Elstrodt, 2009, VII. Theorem 1.9] for $t \geq 0$ we can rewrite the total variation norm as

$$\begin{aligned}
\|T^*(t)\,\mu - \mu\| &= [T^*(t)\,\mu - \mu](\mathbb{R}^n \cap P) - [T^*(t)\,\mu - \mu](\mathbb{R}^n \cap N) \\
&= \int_{\mathbb{R}^n \cap P} [T^{\ll}(t)\,\rho - \rho]\,dy - \int_{\mathbb{R}^n \cap N} [T^{\ll}(t)\,\rho - \rho]\,dy \\
&= \int_{\mathbb{R}^n} |T^{\ll}(t)\,\rho - \rho|\,dy = \|T^{\ll}(t)\rho - \rho\|_1
\end{aligned}$$

and strong continuity of $T^{\ll}$ follows directly from strong continuity of T^*. □

Chapter 4

The Fokker Planck Equation

The Kolmogorov operator allows us to describe the time-evolution of the expectation $E[\phi(X_t^x)]$ for continuous functions $\phi \in C_0(\mathbb{R}^n)$ under the dynamics of an Itô diffusion X_t^x. In this chapter we want to characterize these dynamics via a certain (deterministic!) partial differential equation. An important concept for this is the **infinitesimal generator** of a Kolmogorov semigroup. We start our derivation with the **Itô formula**. It is an explicit formula for the chain rule of Itô diffusions:

Theorem 4.1 (c.f. [Øksendal, 1998], Thm. 4.2.1)**.** *Let X_t^x be an Itô diffusion, i.e. an homogeneous stochastic process solving* (2.1)*, and let $h(x) = (h_1(x), \ldots, h_m(x)$ be a $\mathcal{C}^2$ map $h : \mathbb{R}^n \to \mathbb{R}^m$. Then the process $h(X_t^x)$ is again an Itô process whose k-th component is*

$$dh_k(X_t^x) = \sum_{i=1}^{n} \partial_i h_k(X_t^x) dX_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \partial_i \partial_j h_k(X_t^x) dX_i dX_j, \ k = 1, \ldots, m. \tag{4.1}$$

Here we have $dB_i dB_j = \delta_{i,j} dt$ and $(dt)^2 = dB_i dt = dt dB_i = 0$ (c.f. [Steele, 2000, Section 8.4., Table 8.2: Box Algebra]). As a consequence of the Itô formula we aim to express the expectation of the transformed process explicitly. We follow the same argumentation as [Øksendal, 1998, pp.118]:

Let the function $h : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^2$-mapping. First of all we insert $dX_t = b(X_t)dt + \sigma(X_t)dB_t$ into the Itô formula for h (4.1) above, resulting in

$$dh(X_t^x) = \sum_i \partial_i h(X_t^x) b(X_t^x) dt + \frac{1}{2} \sum_{i,j} \partial_i \partial_j h(X_t^x) (\sigma\sigma^T)_{i,j}(X_t^x) dt + \sum_{i,j} \partial_i h(X_t^x) \sigma_{i,j}(X_t^x) dB_j$$

Definition 4.2. *We define the **generator** of an Itô diffusion as $A : C^2(\mathbb{R}^n) \to C(\mathbb{R}^n)$,*

$$(Ah)(x) := \sum_{i=1}^{n} \partial_i h(x) b_i(x) + \frac{1}{2} \sum_{i,j=1}^{n} \partial_i \partial_j h(x) (\sigma\sigma^T)_{i,j}(x). \tag{4.2}$$

Now we can use it to simplify the formula above

$$dh(X_t^x) = (Ah)(X_t^x)\,dt + \sum_{i,j} \partial_i h(X_t^x)\sigma_{i,j}(X_t^x)\,dB_j.$$

and write it in the integral notation, additionally taking expectations:

$$E[h(X_t^x)] = h(x) + E\left[\int_0^t (Ah)(X_t^x)ds\right] + E\left[\int_0^t \sum_{i,j} \partial_i h(X_t^x)\sigma_{i,j}(X_t^x)dB_j\right] \tag{4.3}$$

What happens with the expectation of the Itô integral on the far right? We know that the function $(t,\omega) \to \sum_{i,j} \partial_i h(X_t^x(\omega))\sigma_{i,j}(X_t^x(\omega))$ is measurable. If we can also assume that it is square integrable and adapted to the filtration $\mathcal{F}_t = \sigma(\{B_s | 0 \le s \le t\})$, then by [Øksendal, 1998, Definition 3.1.4 and Theorem 3.2.1 (iii)] its expectation is zero.

In fact, for functions in $C_b^2(\mathbb{R}^n)$, the expectation of (4.3) is given by the well known **Dynkin formula**:

Theorem 4.3. *(c.f. [Øksendal, 1998, Theorem 7.4.1] for $\phi \in C_0^2(\mathbb{R}^n)$ and [Dynkin, 1965, Corollary to Theorem 5.1.] for $\phi \in C_b^2(\mathbb{R}^n)$) Let $\tau > 0$ be a stopping time and let $\phi \in \mathcal{C}_b^2(\mathbb{R}^n)$.[1] Then for an Itô diffusion X_t^x with $E[\tau] < \infty$,*

$$E[\phi(X_\tau^x)] = \phi(x) + E\left[\int_0^\tau A\phi(X_s^x)ds\right]. \tag{4.4}$$

This formula holds for all **stopping times** τ. These are random variables in $[0,\infty]$ such that $\{\tau \le t\} \in \mathcal{F}_t$ for all $t \in [0,\infty)$ ([Klenke, 2008, Definition 9.15]). Of course any fixed $t \ge 0$ is also a stopping time because then for any $s \in [0,\infty)$ the set $\{t \le s\}$ is either empty or equal to Ω.

4.1 Infinitesimal Generators

In this section we want to formalize the motivation:

Definition 4.4 (c.f. [LeVarge, 2003], Def. 4.1)**.** *Let the family $T := \{T(t) : t \ge 0\}$ be a semigroup of linear operators on a Banach space S. We call*

$$\mathcal{D}_\mathcal{A} = \left\{\phi \in S : \lim_{t \downarrow 0} \frac{T(t) - T(0)}{t}\,\phi \text{ exists}\right\}$$

*the **domain** of an operator $\mathcal{A}$ and define the **infinitesimal generator** $\mathcal{A}$ of T as*

$$\mathcal{A}\,\phi := \lim_{t \downarrow 0} \frac{T(t) - T(0)}{t}\,\phi \text{ for all } \phi \in \mathcal{D}_\mathcal{A}. \tag{4.5}$$

[1] The function space $C_b^2(\mathbb{R}^n)$ will be defined in Section 4.6. Basically these are all twice differentiable functions with bounded derivatives up to order 2.

Remark. *Loosely speaking, $\mathcal{A}\,\phi \cong T'(0)\,\phi$, the time differential of T at time 0. Of course the domain $\mathcal{D}_A \subset S$ of A is still to be determined.*

Before we further characterize the domain $\mathcal{D}_A$ in the following section, let us state the following important result.

Proposition 4.5. *(c.f. [Butzer and Berens, 1967, Proposition 1.1.4], [LeVarge, 2003, Theorems 5.4 and 5.5]) Let T be a C_0-semigroup on a Banach space S with infinitesimal generator $\mathcal{A}$. Then for all $t \geq 0$,*

a) $\phi \in \mathcal{D}_\mathcal{A} \Rightarrow T(t)\,\phi \in \mathcal{D}_\mathcal{A}$,

b) $\frac{d}{dt}T(t)\,\phi = \mathcal{A}\,T(t)\,\phi = T(t)\,\mathcal{A}\,\phi$ for all $\phi \in \mathcal{D}_\mathcal{A}$ and

c) $\mathcal{D}_\mathcal{A}$ is dense in S, i.e., $\overline{\mathcal{D}_\mathcal{A}} = S$, and $\mathcal{A}$ is a closed operator.

The Kolmogorov operators $T(t)$, $T^*(t)$ and $T^{\ll}(t)$ are continuous linear operators, which implies that they are also bounded operators (c.f. [Alt, 2006, Lemma 3.1]). Their infinitesimal generators $\mathcal{A}$, $\mathcal{A}^*$ and $\mathcal{A}^{\ll}$, however, are not bounded in general. For a certain class of functions $\phi \in \mathcal{D}_\mathcal{A} \subset C_b(\mathbb{R}^n)$, we can give an explicit formula for the $\mathcal{A}$ and $\mathcal{A}^{\ll}$.

Definition 4.6. *(c.f. [Alt, 2006, Section 1.4]) The space of functions with continuous derivatives up to order m is defined as*

$$C^m(\mathbb{R}^n) := \{f \in C(\mathbb{R}^n) :\ \partial^s f \in C(\mathbb{R}^n) \ \text{ for every multi-index } |s| \leq m\}. \tag{4.6}$$

Equipped with the following norm, this space is complete:

$$\|f\|_{C^m} := \sum_{|s|\leq m} \|\partial^s f\|_\infty. \tag{4.7}$$

Furthermore, we define $C_b^m(\mathbb{R}^n) := \{f \in C^m(\mathbb{R}^n) :\ \|f\|_{C^m} < \infty\}$.

We want to give an explicit for $\mathcal{A}$ and need the following lemma:

Lemma 4.7. *Let A be the operator defined in (4.2) above. Let $Z \sim \eta \in \mathcal{M}_1$ be a random variable with $E[|Z|^2] < \infty$. Then for any $\phi \in C_b^2(\mathbb{R}^n)$, we have:*

$$E[|A\phi(Z)|] < \infty.$$

If additionally $E[|Z|^4] < \infty$, then also $E[|A\phi(Z)|^2] < \infty$.

Proof. We can bound the operator A in the following way:

$$\begin{aligned}
|\mathcal{A}\phi(x)| &\leq \sum_{i=1}^n |b(x)_i||\partial_i\phi(x)| + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n |(\sigma(x)\sigma(x)^T)_{i,j}||\partial_i\partial_j\phi(x)| \\
&\leq nC(1+|x|)|\nabla\phi(x)| + nC(1+|x|)^2|\nabla^2\phi(x)| \\
&\leq nC(2+2|x|+|x|^2)\|\phi\|_{C_b^2(\mathbb{R}^n)}.
\end{aligned}$$

For the expectation we get:

$$E[|A\phi(Z)|] = \int_{\mathbb{R}^n} |A\phi(y)|\eta(dy) \leq nC \int_{\mathbb{R}^n} [2 + 2|y| + |y|^2] \|\phi\|_{C_b^2(\mathbb{R}^n)} \, \eta(dy)$$
$$= nC \left(2 + 2E[|Z|] + E[|Z|^2]\right) \|\phi\|_{C_b^2(\mathbb{R}^n)} < \infty.$$

The second statement directly follows by squaring the bound above. □

This lemma will be useful for showing what we were obviously aiming at all the time:

Theorem 4.8 ([Øksendal, 1998], Theorem 7.3.3)**.** *The infinitesimal generator of the Kolmogorov semigroup restricted to $C_b^2(\mathbb{R}^n)$ is equal to the operator A from equation (4.2). I.e. for $\phi \in C_b^2(\mathbb{R}^n)$ $\mathcal{A}$ can be expressed as*

$$\mathcal{A}\phi(x) = \sum_{i=1}^{n} b_i(x) \frac{\partial \phi}{\partial x_i}(x) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\sigma\sigma^T)_{i,j}(x) \frac{\partial^2 \phi}{\partial x_i \partial x_j}(x). \tag{4.8}$$

Proof. Let X_t^x be an Itô diffusion. An informal derivation of this result has already been given in the motivation section of this chapter above. As we have seen it led us to Dynkin's formula (4.4) applied for a stopping time $\tau = t \geq 0$, functions $\phi \in C_0^2(\mathbb{R}^n)$ and an operator A of the form of the right-hand side of equation (4.8). Dynkin's formula can easily be extended to functions $C_b^2(\mathbb{R}^n)$, such that

$$E[\phi(X_t^x)] = \phi(x) + E\left[\int_0^t (A\phi)(X_s^x) ds\right].$$

It remains to show that (4.8) indeed is the infinitesimal generator of the Kolmogorov semigroup T. In order to do this we need to rearrange this equation. Then we multiply by $1/t$ and take the limit $t \downarrow 0$.

$$\mathcal{A}\phi(x) = \lim_{t \downarrow 0} \frac{E[\phi(X_t^x)] - \phi(x)}{t} = \partial_t E\left[\int_0^t (A\phi)(X_s^x)\, ds\right].$$

Now we calculate the expression on the right hand side for $\phi \in \mathcal{C}_b^2$:

$$\int_0^t \int_{\mathbb{R}^n} |A\phi(y)| \ \mu_s^x(dy) ds \leq t \sup_{s \in (0,t]} \int_{\mathbb{R}^n} |A\phi(y)| \ \mu_s^x(dy) = t \sup_{s \in (0,t]} E[|A\phi(X_t^Z)|] < \infty. \tag{4.9}$$

because we assumed that X_t^Z has got a density for $t > 0$ and thus we can apply Lemma 4.7. Furthermore, by Fubini's theorem ([Elstrodt, 2009, V. Theorem 2.1 (c)]), and by the fundamental theorem of integral calculus ([Elstrodt, 2009, VII. Theorem 4.14 (a)]),

$$\mathcal{A}\phi(x) = \partial_t E\left[\int_0^t A\phi(X_s^x) ds\right] = \partial_t \int_0^t \int_{\mathbb{R}^n} A\phi(y) \mu_s^x(dy) ds = \int_{\mathbb{R}^n} Ah(y) \mu_0^x(dy) = A\phi(x)$$

which shows that A indeed is the infinitesimal generator of T. □

In the light of this proposition we can also state the density generator $\mathcal{A}^{\ll}$ explicitly:

Proposition 4.9. *The weak* infinitesimal generator of T^* is equal to the adjoint operator of $\mathcal{A}$. The infinitesimal generator $\mathcal{A}^{\ll}$ of $T^{\ll}$ is the adjoint of $\mathcal{A}$ in the $L^1(\mathbb{R}^n)$-sense.*

Proof. The first statement is given in [Butzer and Berens, 1967, Corollary 1.4.5]. To see $\mathcal{A}^{\ll} = (\mathcal{A})^*$, observe the following: let $Z \sim \eta$ a random variable with density ρ and let X_t^Z be an Itô diffusion. Then for all $\phi \in C_0^2(\mathbb{R}^n)$,

$$\langle \phi, \mathcal{A}^{\ll}\rho\rangle_1 = \langle \mathcal{A}\phi, \rho\rangle_1 = \int_{\mathbb{R}^n} \mathcal{A}\,\phi(y)\rho(y)dy = \int_{\mathbb{R}^n} \phi(y)\,(\mathcal{A})^*\,\rho(y)dy = \langle \phi, (\mathcal{A})^*\rho\rangle\,.$$

□

Corollary 4.10. *(c.f. [Øksendal, 1998, Exercise 8.3]) Let X_t^Z be an Itô diffusion with density $p_t^x \in L_1^1(\mathbb{R}^n) \cap C_b^2(\mathbb{R}^n)$ and restrict the infinitesimal generator $\mathcal{A}$ to the same function space. Then $\mathcal{A}^{\ll}$ is given by*

$$\mathcal{A}^{\ll}p_t^x(y) = -\sum_i \frac{\partial(b_i p_t^x)}{\partial y_i}(y) + \frac{1}{2}\sum_{i,j}\frac{\partial^2((\sigma\sigma^T)_{i,j}p_t^x)}{\partial y_i \partial y_j}(y). \tag{4.10}$$

Proof. The formula follows from integration by parts of equation (4.8): Let $\phi \in C_0(\mathbb{R}^n) \cap C_b^2(\mathbb{R}^n)$. According to proposition 4.9, $\langle \phi, A^{\ll}\rho\rangle_1 = \langle \phi, (\mathcal{A})^*\rho\rangle_1 = \langle \mathcal{A}\,\phi, \rho\rangle_1$:

$$\begin{aligned}
\langle \mathcal{A}\,\phi, \rho\rangle_1 &= \int_{\mathbb{R}^n}\sum_i b_i(x)\partial_i\phi(x)p_t^x dx + \frac{1}{2}\int_{\mathbb{R}^n}\sum_{i,j}(\sigma\sigma^T)_{i,j}(x)\partial_i\partial_j\phi(x)p_t^x dx \\
&= -\int_{\mathbb{R}^n}\sum_i \partial_i(b_i p_t^x)(x)\phi(x)dx - \frac{1}{2}\int_{\mathbb{R}^n}\sum_{i,j}\partial_i((\sigma\sigma^T)_{i,j}p_t^x)(x)\partial_j\phi(x)dx \\
&= -\int_{\mathbb{R}^n}\sum_i \partial_i(b_i p_t^x)(x)\phi(x)dx + \frac{1}{2}\int_{\mathbb{R}^n}\sum_{i,j}\partial_i\partial_j((\sigma\sigma^T)_{i,j}p_t^x)(x)\phi(x)dx \\
&= \langle \phi, \mathcal{A}^{\ll}\,\rho\rangle_1\,.
\end{aligned}$$

The boundary integrals in this process vanish all because p_t^x and ϕ approach zero as $|x| \to \infty$. □

4.2 Abstract Cauchy Problems

As we have seen, the expectation $t \mapsto E[\phi(X_t^x)]$ develops continuously as a function of time (c.f. also [LeVarge, 2003, Cor. 5.3]). Another way of describing the time-evolution of semigroup-operators is in the form of a initial value problem, the so-called *abstract Cauchy problem.*

Definition 4.11 (c.f. [LeVarge, 2003], 2.3)**.** *Let T be the Kolmogorov semigroup and let $\mathcal{A}$ its infinitesimal generator. Let $\phi \in \mathcal{D}_{\mathcal{A}}$ and define $u(t,x) = E[\phi(X_t^x)]$. Then the* ***abstract Cauchy problem*** *is the following homogeneous initial value problem:*

$$\begin{aligned} \frac{d}{dt}u(t) &= \mathcal{A}\,u(t), \quad t > 0 \\ u(0) &= \phi \end{aligned} \tag{4.11}$$

We can explain the importance of this problem like this: One can even show that every C_0-semigroup is uniquely determined by its infinitesimal generator (c.f. [LeVarge, 2003, Theorem 5.7]). On one side of the coin, the Kolmogorov semigroup T describes how the expectation of $\phi(X_t^x)$ evolves with the path of the Itô process X_t^Z via the relation $T(t)u(s) = u(t+s)$. On the other side, equation (4.11) gives an alternative way by taking a closer look on what happens at an infinitesimal time-step using the generator $\mathcal{A}$. It was Andrey Kolmogorov who first devised a time evolution equation for continuous stochastic processes (c.f. [Kolmogoroff, 1931, Chapter 4, pp.438]. That's why the abstract Cauchy problem for the Kolmogorov semigroup is also referred to as the **Kolmogorov backward equation**.

Theorem 4.12. *The Kolmogorov backward equation has got a unique t-differentiable solution $u \in \mathcal{D}_{\mathcal{A}}$.*

Proof. We already showed that T forms a C_0-semigroup. According to the Hille-Yosida Theorem ([LeVarge, 2003, Theorem 5.9]), $\mathcal{A}$ is a closed operator with nonempty resolvent set. According to [Pazy, 1983, 1. Corollary 2.5] $\mathcal{A}$ is densely defined. Therefore, by [Pazy, 1983, 4. Thm. 1.3], the abstract Cauchy problem has got a unique solution. Furthermore, this solution is t-differentiable on $[0,\infty)$ for every initial value $\phi \in \mathcal{D}_{\mathcal{A}}$. This solution is in $\mathcal{D}_{\mathcal{A}}$ because by Proposition 4.5 b, $\mathcal{A}u(t,\cdot) = \mathcal{A}T(t)\phi = T(t)\mathcal{A}\phi$. □

This solution is unique just as Itô diffusions are unique. We can also follow the same argument as above for the dual Kolmogorov operator:

Definition 4.13. *Let $Z \sim \eta \in \mathcal{D}_{\mathcal{A}^*}$ be a real-valued random variable. Let T^* be the dual Kolmogorov semigroup and $\mathcal{A}^*$ its infinitesimal generator. If μ_t^Z the distribution of the Itô diffusion X_t^Z, then the* ***dual Cauchy problem*** *is the following homogeneous initial value problem:*

$$\begin{aligned} \frac{d}{dt}\mu_t^Z &= \mathcal{A}^*\,\mu_t^Z, \quad t > 0 \\ u(0) &= \eta. \end{aligned} \tag{4.12}$$

If η permits a density $\rho \in \mathcal{D}_{\mathcal{A}^*}$, then the measures in this problem can be replaced by their densities and it is then called the ***Kolmogorov forward equation*** *(e.g. [Øksendal, 1998, Exercise 8.3]) or the* ***Fokker-Planck equation*** *(e.g. [Risken, 1984])*

Theorem 4.14. *The dual Cauchy problem and the Fokker-Planck equation have got unique t-differentiable solutions in $\mathcal{D}_{\mathcal{A}^*}$ and $\mathcal{D}_{\mathcal{A}^\ll}$, respectively.*

Proof. Since both T^* and $T^\ll$ are strongly continuous semigroups, the same reasoning as in Theorem 4.12 applies here. □

If the distribution η permit a densities ρ, then the density of the Itô diffusion p_t^Z also exists. But what if $T \sim \delta_x$ with $x \in \mathbb{R}^n$?

Theorem 4.15. *In the setting of Theorem 4.14, assume $Z \sim \delta_x$ for $x \in \mathbb{R}^n$. Let X_t^x be an Itô diffusion started at x. If for $t > 0$, X_t^x has got a density p_t^x which is continuously t-differentiable, then Theorem 4.14 still holds true for the Fokker-Planck equation.*

Proof. Let $x \in \mathbb{R}^n$ and $Z \sim \delta_x$, let $X_t^x = X_t^Z$ be an Itô diffusion with density $p_t^x \in H^{1,2}(\mathbb{R}^n)$ which is $t - differentiable$ for all $t > 0$.

Let $\phi \in C_0^2$. By the Kolmogorov backward equation (4.11) and Proposition 4.5 b, we get

$$\partial_t u(t,x) = \mathcal{A}u(t,x) = \mathcal{A}T(t)\phi(x) = T(t)\mathcal{A}\phi(x) = E[\mathcal{A}\phi(X_t^x)] = \langle \mathcal{A}\,\phi, p_t^x\rangle_1 = \langle \phi, \mathcal{A}^\ll p_t^x\rangle_1.$$

Now we approach the problem from another angle: The map $y \mapsto \phi(y)p_t^x(y)$ is Lebesgue integrable because ϕ is bounded. The map $t \mapsto \phi(y)p_t^x(y)$ is differentiable because we assumed that $p_t^x \in C_0^{1,2}(\mathbb{R}^n)$ for $t > 0$. Furthermore:

$$\int_{\mathbb{R}^n} \sup_{0<s\le t} |\phi\, p_s^x|(y)dy \le \|\phi\|_\infty \int_{\mathbb{R}^n} (\sup_{0<s\le t} p_s^x)(y)dy \le \|\phi\|_\infty < \infty,$$

because p_t^x is a probability density for all $t > 0$. We can now apply [Klenke, 2008, Lemma 6.28] and we can exchange the time differentiation with the integral:

$$\partial_t u(t,x) = \partial_t E[\phi(X_t^x)] = \partial_t \int_{\mathbb{R}^n} \phi(y)p_t^x(y)dy = \int_{\mathbb{R}^n} \phi(y)\partial_t\, p_t^x(y)dy = \langle \phi, \partial_t p_t^x\rangle_1 .$$

Combing these two results we get:

$$\langle \phi, \partial_t p_t^x - \mathcal{A}^\ll p_t^x\rangle_1 = 0 \quad \text{for all } \phi \in C_b^2(\mathbb{R}^n) \text{ and } t > 0. \tag{4.13}$$

By a $\mathcal{C}_0^2$ version of the fundamental lemma of Calculus of Variations (c.f. [Alt, 2006], 2.21), the Fokker-Planck equation holds. (In the proof of the Lemma one can do all calculations if we only demand equation (4.13) above for $\mathcal{C}_0^2$ functions. This is due to the fact that $\mathcal{C}_0^2 \subset \mathcal{C}_0^\infty$.) □

It is not easy to see if X_t^x has got a density for $t > 0$. We would argue that it is a reasonable conclusion based on the fact that Itô diffusions are generated by Brownian motions, which do have a density. However, we could not prove this and so we state the fact as a conjecture in Appendix B and assume it holds making the following assumption:

Assumption 1. *We assume that the Itô diffusion X_t^x has got a density p_t^x for all $x \in \mathbb{R}^n$ and $t > 0$.*

4.3 Stationary distributions of Itô diffusions

In the following theorem we will use the Fokker-Planck equation to derive specific diffusion equations which have a stationary distribution which we want to sample from:

Theorem 4.16. *([Ma et al., 2015, Theorem 1]) Assume an Itô diffusion X_t^x with density $p_t^x \in L_1^1(\mathbb{R}^n) \cap C_b^2(\mathbb{R}^n)$ solving of the SDE*

$$\begin{aligned} X_0 &= x \\ dX_t &= -[D(X_t) + Q(X_t)]\nabla H(X_t)dt + \Gamma(X_t)dt + \sqrt{2D(X_t)}dB_t \\ &\quad \text{with } \Gamma_i(y) := \sum_j \partial_j [D_{i,j} + Q_{i,j}](y). \end{aligned}$$

Assume D is a positive semi-definite drift matrix and Q is a skew-symmetric diffusion matrix. Then X_t has got a stationary distribution $\propto \exp\{-H(y)\}$. If D is positive definite, then the stationary distribution is unique.

Proof. We insert the choice of b and σ into the density version of the Fokker-Planck equation. Writing $H_i' = \partial_{x_i} H$ we have

$$\partial_t p_t^x = \sum_i \partial_i (\sum_j [D_{i,j} + Q_{i,j}] H_j' \, p_t^x) - \sum_i \partial_i \left(\Gamma_i p_t^x\right) + \sum_{i,j} \partial_i \partial_j [D_{i,j} p_t^x]$$

Let us decompose the second summand using the product rule:

$$\begin{aligned} \sum_i \partial_i \left(\Gamma_i p_t^x\right) &= \sum_i \partial_i \Gamma_i p_t^x + \sum_i \Gamma_i \partial_i p_t^x \\ &= \sum_{i,j} p_t^x \partial_i \partial_j [D_{i,j} + Q_{i,j}] + \sum_{i,j} \partial_i [D_{i,j} + Q_{i,j}] \partial_j p_t^x \\ &= \sum_{i,j} \partial_j (p_t^x \partial_i D_{i,j}) + \sum_{i,j} \partial_j \left(p_t^x \partial_i Q_{i,j}\right). \end{aligned}$$

These sums can be further split up into

$$\begin{aligned} \sum_{i,j} \partial_j (p_t^x \partial_i D_{i,j}) &= \sum_{i,j} \partial_i \partial_j (D_{i,j} p_t^x) - \sum_{i,j} \partial_i [D_{i,j} \partial_j p_t^x] \\ \sum_{i,j} \partial_j (p_t^x \partial_i Q_{i,j}) &= -\sum_{i,j} \partial_i [Q_{i,j} \partial_j p_t^x], \end{aligned}$$

since Q is skew-symmetric. Inserting this into the equation above we get:

$$\begin{aligned} \partial_t p_t^x &= \sum_{i,j} \partial_i \left([D_{i,j} + Q_{i,j}] H_j' \, p_t^x\right) + \sum_{i,j} \partial_i ([D_{i,j} + Q_{i,j}] \partial_j p_t^x) \\ &= \sum_{i,j} \partial_i \left([D_{i,j} + Q_{i,j}][H_j' \, p_t^x + \partial_j p_t^x]\right) \end{aligned} \tag{4.14}$$

Now for the stationary distribution $p_t^x := \pi^x =\propto e^{-H}$, we get $\partial_t \pi^x = 0$, i.e.

$$\sum_{i,j} \partial_i \left([D_{i,j} + Q_{i,j}][H'_j \, \pi^x + \partial_j \pi^x] \right) = 0 \tag{4.15}$$

Now if D is positive definite, we say, then the operator $L\pi^x := \sum_{i,j} \partial_i([D_{i,j} + Q_{i,j}][H'_j \, \pi^x + \partial_j \pi^x])$ is **elliptic** (c.f. [Gilbarg and Trudinger, 1998, p.31]). According to Theorem [Gilbarg and Trudinger, 1998, Theorem 6.8], the equation (4.15) has got a unique solution in $C_b^2(\mathbb{R}^n)$. Since the π^x with $\partial_j \pi^x = -\pi^x \partial_j H$ solves (4.15), we get $\pi =\propto \exp\{-H(y)\}$. □

Remarks.

(a) Uniqueness in the theorem above also holds if the process can be shown to be ergodic (c.f. [Ma et al., 2015, Theorem 1]).

(b) For the completeness of the framework, c.f. [Ma et al., 2015, Theorem 2].

Chapter 5

Numerical Integration

We have seen in the preceding chapter how we can device Langevin diffusion equations which have a Boltzmann distribution as a stationary distribution. In this chapter we want to construct Markov chains with this stationary distribution as the limit. The methods we describe here are basic for Monte Carlo Markov Chain (MCMC) simulation.

If we recall the Langevin diffusion equation (2.1) we see that it contains a deterministic and a stochastic part:

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t.$$

While the solution of an initial value problem of the former can be solved by various integration methods (e.g. [Hanke-Bourgeois, 2009]), the stochastic part needs to be approximated via stochastic simulation. We say that we **sample** from a distribution of a random process, if for a collection of $\omega_i \in \Omega$ determine a sample path $t \mapsto X(t, \omega_i$ $(i = 1, \dots M)$. The mean value of the paths can, for example be used to describe the evolution of the mean value for each $t_l \geq 0$, $l = 0, \dots L$ (c.f. [Klenke, 2008, Weak Law of Large Numbers, Theorem 5.14]).

In this chapter we will assume that $m = n$, i.e. $\sigma : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ and we will mostly only consider fixed starting points $x \in \mathbb{R}^n$. The results presented here, however, can also be shown for a starting distribution $\eta \sim \mathcal{M}$.

5.1 The Euler-Maruyama Method

Definition 5.1. *(c.f. [Kloeden and Platen, 1999, Sections 9.1, 10.2 and 14.1])*
Let $0 = t_0 < t_1 < \cdots < t_L = T$ be a discretization of the time interval $[0, T]$. The ***Euler-Maruyama*** *discretization scheme is defined as*

$$\hat{X}^x_{t_0} = x \tag{5.1}$$

$$\hat{X}^x_{t_{l+1}} = \hat{X}^x_{t_l} + b(\hat{X}^x_{t_l})(t_{l+1} - t_l) + \sigma(\hat{X}^x_{t_l})(B_{t_{l+1}} - B_{t_l}) \quad (l = 0, \dots L) \tag{5.2}$$

*where $B_{t_i} \in \mathbb{R}^n$ are realizations of a n-dimensional standard Brownian motion and $\sigma(\hat{X}^x_{t_l}) \in \mathbb{R}^{n\times n}$ is a $n \times n$-dimensional matrix. Whenever the context is un-ambiguous, we will simply write $\hat{X}_l = \hat{X}_{t_l} = \hat{X}^x_{t_l}$. Finally, the **step sizes** of the time discretization above are defined as:*

$$h_l := t_{l+1} - t_l \quad \textit{and} \quad h := \max_{1\le l\le L} h_l.$$

It can be shown that the increments are independently distributed Gaussian random variables $(B_{l+1} - B_l) \sim N(0, h_l 1_{n\times n})$(c.f. [Kloeden and Platen, 1999, Section 9.1]). If we want to simulate these Brownian motion increments in some numerical method, we can draw from the standard normal distribution and scale these draws by $\sqrt{t}$:

Proof. Let B_t be a Brownian Motion and let $\xi \sim \mathcal{N}(0,1)$. Then:
$P(\sqrt{t}\xi \le x) = P\left(\xi \le \frac{x}{\sqrt{t}}\right) = F_{0,1}\left(\frac{x}{\sqrt{t}}\right) = F_{0,\sqrt{t}}(x) = P(B_t \le x)$. □

This allows us to rewrite equation (5.1) above using standard normal Gaussian realizations $\xi_l \sim \mathcal{N}(0, 1_{n\times n})$ $(l = 0, \dots L-1)$:

$$\hat{X}_0 = x \tag{5.3}$$

$$\hat{X}_{l+1} = \hat{X}_l + h_l\, b(\hat{X}_l) + \sqrt{h_l}\, \sigma(\hat{X}_l)\, \xi_l \quad (l = 1, \dots L). \tag{5.4}$$

Definition 5.2. *(c.f. [Milstein and Tretyakov, 1995, p.4], [Kloeden and Platen, 1999, Sections 9.6 and 9.7]) We say a discretization scheme is of **mean-square order of accuracy** K, if the approximations $\hat{X}_l$ satisfy*

$$\left(E\left|X^x_{t_l} - \hat{X}^x_{t_l}\right|^2\right)^{1/2} \le C\, h^K \quad \textit{for all } l = 1, \dots L. \tag{5.5}$$

*In this definitions the constant C needs to be independent of h and l. Similarly, we say it is of **strong order of accuracy** K, if*

$$E\left|X^x_{t_l} - \hat{X}^x_{t_l}\right| \le C\, h^K. \tag{5.6}$$

*We also speak of mean-square and strong **order of convergence**, respectively. If the approximation is of strong order K, we also write*

$$\hat{X}^x_{t_l} = X^x_{t_l} + \mathcal{O}(h^K) \quad \Leftrightarrow \quad E\left|\hat{X}^x_{t_l} - X^x_{t_l}\right| \in \mathcal{O}(h^K)$$

as $h > 0$ approaches 0 from above. If not otherwise determined, $\mathcal{O}$ is used with the 1-norm for vectors and matrices: $|b(x)| = \sum_{i=1}^n |b_i(x)|$ and $|\sigma(x)| = \sum_{i=1}^n \sum_{j=1}^m |\sigma_{i,j}(x)|$ for $b : \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times n}$, respectively.

Lemma 5.3. *(c.f. [Milstein and Tretyakov, 1995, p.4] and [Kloeden and Platen, 1999, Theorem 10.2.2]) The Euler-Maruyama scheme is of mean-square order of accuracy* $1/2$.

It follows directly from Jensen's inequality (c.f. [Klenke, 2008, Theorem 8.19]) that

$$\left(E\left|X_{t_l}^x - \hat{X}_{t_l}^x\right|\right)^2 \leq E\left[|X_{t_l}^x - \hat{X}_{t_l}^x|^2\right]$$

such that strong order of accuracy follows directly form mean-square order of accuracy.

The stochastic equivalent of the deterministic Euler Method is of order $1/2$ only, compared to an order of 1 (e.g. [Hanke-Bourgeois, 2009, Theorem 74.1]). This fact is a consequence of the stochasticity of the process. It is most visible in the Itô isometry ([Øksendal, 1998, Corollary 3.1.7]) and can well be illustrated in a short example (c.f. also the proof of [Kloeden and Platen, 1999, Theorem 10.2.2]):

$$E\left[\left|\int_{t_l}^{t_{l+1}} \sigma(X_r)dB_r\right|^2\right] = E\left[\int_{t_l}^{t_{l+1}} \sigma(X_r)^2 dr\right] = \int_{t_l}^{t_{l+1}} E\left[\sigma(X_r)^2\right] dr \leq h_l \sup_{r\in[t_l,t_{l+1}]} E\left|\sigma(X_r)\right|^2.$$

More than the absolute accuracy of an approximation method we are interested in its capabilities of approximating the distribution of a diffusion process.

Definition 5.4. *(c.f. [Milstein and Tretyakov, 1995, p.6], [Kloeden and Platen, 1999, Section 9.7]) We say a discretization scheme* $\hat{X}$ *is of* ***(weak) order of accuracy*** K*, if*

$$\left|E\left[\phi(X_{t_l}^x)\right] - E[\phi(\hat{X}_{t_l}^x)]\right| \leq Ch^K \quad \text{for all } \phi \in C_b^1(\mathbb{R}^n) \tag{5.7}$$

and for all $l = 1, \ldots, L$ *with a constant* C *independent of* h *and* l.

By the triangle inequality and the Lipschitz property of functions $\phi \in C_b^1(\mathbb{R}^n)$, methods of strong order of accuracy K are also of weak order of accuracy K. In the case of the Euler-Maruyama method, however, we can make an even stronger statement:

Lemma 5.5. *(c.f. [Kloeden and Platen, 1999, Theorem 14.1.5]) Let* $t \in [0, T]$ *and let* X_t^x *be an Itô diffusion for which all moments exist. Assume further that* $bC_b^2(\mathbb{R}^n)$, $\sigma \in C_b^2(\mathbb{R}^n, \mathbb{R}^n)$ *and* $z^T\sigma\sigma^T(x)z \geq \lambda|z|^2$ *for* $x, z \in \mathbb{R}^n$ *and* $\lambda > 0$. *Then*

$$\left|E\left[\phi(X_t^x)\right] - E\left[\phi(\hat{X}_t^x)\right]\right| \leq Ch \quad \text{for all } \phi \in C_b^4(\mathbb{R}^n). \tag{5.8}$$

For $\phi \in C_b^2(\mathbb{R}^n)$, *the Euler-Maruyama discretization is of weak order* $1/2$.

So the Euler-Maruyama discretization is of weak order $1/2$, but in general it ought to behave quite better in the weak sense of approximation.

5.2 The Milstein Method

As we have seen the limitations of the Euler method stem from the stochasticity of our Itô process. If we could somehow account for this fact, we could improve the approximation order. The following approach is able to do this introducing a correction term for the stochastic integral. An informal derivation of the idea behind the correction approach can be found in [Rouah, 2013]. The idea is basically of using Itô Taylor expansions of different degrees for the coefficients b and σ.

Definition 5.6. *(c.f. [Kloeden and Platen, 1999, p.346]) For a stochastic differential equation as in* (2.1)*, the* ***Milstein*** *discretization is defined as component-wise as*

$$\begin{aligned}\hat{X}_{k,t_0} &= x_k \\ \hat{X}_{k,t_{l+1}} &= \hat{X}_{k,t_l} + h_l b(\hat{X}_{t_l})_k + \sqrt{h_l}\,\left(\sigma(\hat{X}_{t_l})\xi_l\right)_k \\ &\quad + \frac{1}{2}\,h_l\left(\sum_{j=1}^{n}\sum_{r=1}^{n}\sum_{i=1}^{n}\frac{\partial\sigma_{k,j}}{\partial x_r}(\hat{X}_{t_l})\;\sigma(\hat{X}_{t_l})_{r,i}\right)\;(\xi_{j,l}^2-1)\quad (l=0,\dots L-1).\end{aligned}$$

In Theorem 10.3.5 [Kloeden and Platen, 1999] give conditions under which the Milstein scheme's strong convergence is of order 1, i.e.

$$E[|X_T^Z - Y^\delta(T)|] \le D\delta.$$

These conditions are differentiability conditions of the diffusion term σ and are covered, for example but not only, if we restrict ourselves to $b \in C_b^1(\mathbb{R}^n)$ and $\sigma \in C_b^4(\mathbb{R}^{n\times n})$.

5.3 Kolmogorov Integrators

The approximation schemes above take the perspective of discretization the diffusion equation. Another approach would be to approximate the Kolmogorov operator T. Here we follow the approach of (c.f. [Chen et al., 2015, Appendix B]): Assume we have a function $\phi \in C_b^\infty(\mathbb{R}^n)$, sufficiently smooth such that by Proposition b (b)

$$\frac{d^k}{dt^k}u(t,x) = \frac{d^{k-1}}{dt^{k-1}}\frac{d}{dt}u(t,x) = \mathcal{A}\,\frac{d^{k-1}}{dt^{k-1}}u(t,x) = \mathcal{A}^k\,u(t,x)$$

for $k \ge 1$ and almost all $x \in \mathbb{R}^n$. We can exchange the differential operators $\frac{d}{dt}$ and $\mathcal{A}$ because of the smoothness of ϕ and because $\mathcal{A}\phi$ is bounded in expectation (c.f. Lemma 4.7). Now we develop u around the time point $t = 0$:

$$u(h,x) = \phi(x) + \sum_{k=1}^{\infty}\frac{h^k}{k!}\frac{d^k}{dt^k}\,u(t,x)\big|_{t=0} = \phi(x) + \sum_{k=1}^{\infty}\frac{h^k}{k!}\;\mathcal{A}^k\,u(t,x)\big|_{t=0} = e^{hA}\phi(x).$$

Of course the expression $u(t,x) = e^{t\mathcal{A}}\phi(x)$ is only formal because in general we won't have $\phi \in C_b^\infty(\mathbb{R}^n)$. One can show that this expression makes sense in particular for functions $\phi \in \mathcal{D}_\mathcal{A}$ such that $\mathcal{A}$ is bounded (c.f. [LeVarge, 2003, Theorems 4.4 and 5.7]). Also, even when $\mathcal{A}$ is un-bounded, one can find a *Yosida approximation* $\mathcal{A}_\lambda$ of the generator s$\mathcal{A}$ such that

$$u(t,x) = \lim_{\lambda\to\infty} e^{t\mathcal{A}_\lambda}\phi(x).$$

for $x \in \mathbb{R}^n$ (c.f. [LeVarge, 2003, Section 5.10]). On the other hand, we can also use the considerations above to talk about approximating $e^{\mathcal{A}t}$:

Definition 5.7. *Let T be a Kolmogorov semigroup with infinitesimal generator $\mathcal{A}$. A **Kolmogorov integrator** is an approximation operator $\hat{P}$ of T. A Kolmogorov integrator is of order $k \geq 0$ if*

$$\hat{P}_h\phi = e^{h\mathcal{A}}\phi + \mathcal{O}(h^{k+1}) \quad \Leftrightarrow \quad \sup_{Z\in L^4(\mathbb{R}^n,P)} E\left[\left|e^{h\mathcal{A}}\phi(Z) - \hat{P}_h(Z)\phi\right|\right] \in \mathcal{O}(h^{k+1}) \text{ as } h \downarrow 0.$$

Theorem 5.8. *Let $Z \in \eta$ be a real random variable with $E[|Z|^3] < \infty$. Assume $\phi \in C_b^4(\mathbb{R}^n)$. Then the Euler integrator is a Kolmogorov integrator of order 1.*

Proof. Let $Z \sim \eta \in \mathcal{M}$ be a random variable and let $\hat{X}_h^Z$ be the first step of Euler-Maruyama approximation given in equation (5.1). Then the **Euler integrator** is defined as:

$$\hat{P}(h)\phi(Z) := E\left[\phi(\hat{X}_h^Z)\right]. \tag{5.9}$$

We develop ϕ around $x \in \mathbb{R}^n$. According to the n-dimensional Taylor expansion formula (c.f. [Forster, 2013, §7, Theorem 1]), there is some $\theta \in [0,1]$ such that:

$$\begin{aligned}\phi(x + h\,b(x) + \sqrt{h}\,\sigma(x)\xi) =& \phi(x) + \nabla\phi(x)^T\left[h\,b(x) + \sqrt{h}\,\sigma(x)\xi\right] \\ &+ \frac{1}{2}\left[h\,b(x) + \sqrt{h}\,\sigma(x)\xi\right]^T \nabla^2\phi(x)\left[h\,b(x) + \sqrt{h}\,\sigma(x)\xi\right] \\ &+ \sum_{|\alpha|=3} \frac{D^\alpha\phi\left(x + \theta[h\,b(x) + \sqrt{h}\,\sigma(x)\xi]\right)}{\alpha!}\left[h\,b(x) + \sqrt{h}\,\sigma(x)\xi\right]^\alpha\end{aligned}$$

Note that Z is independent of ξ and b, $\nabla\phi$, $\nabla^2\phi$ and σ all are measurable functions, thus all mixed terms vanish in expectation, e.g.

$$\begin{aligned} E[\sqrt{h}\,\nabla\phi(Z)^T\sigma(Z)\,\xi] &= 0 \\ E\left[h\,\sqrt{h}\,(b^T\nabla^2\phi\,\sigma)(Z)\,\xi\right] &= 0 \\ \text{and setting } \tilde{Z} := Z + \theta[h\,b(Z) + \sqrt{h}\,\sigma(Z)\xi&: \\ E\left[|D^\alpha\phi\left(\tilde{Z}\right)\sigma(Z)\xi|\right] \leq E\left[|D^\alpha\phi\left(\tilde{Z}\right)|\right] E\left[|\sigma(Z)^\alpha\xi^\alpha|\right] &= 0\end{aligned}$$

for all multi-indices α with $|\alpha| = 3$. Furthermore, since $\|D^\alpha \_phi\|_{C_b^3(\mathbb{R}^n)} \leq \infty$ and $E[|Z|^3] < \infty$, when we insert these results into the definition of the Euler integrator and, after some algebra[1], we get:

$$\begin{aligned}\hat{P}(h)\phi(Z) &= E[\phi(Z + h\, b(Z) + \sqrt{h}\, \sigma(Z)\xi)] \\ &= E[\phi(Z)] + hE\left[\nabla\phi(Z)^T b(Z) + \frac{1}{2} Tr\left[(\sigma\sigma^T)\nabla^2\phi(Z)\right]\right] + \mathcal{O}\left(h^2\right) \\ &= E[\phi(Z)] + hE[\mathcal{A}\phi(Z)] + \mathcal{O}\left(h^2\right)\end{aligned}$$

The same result can also be obtained by numerical integration of the Kolmogorov backward equation: we have $\mathcal{A}^2\phi \in C_b(\mathbb{R}^n)$ for $\phi \in C_b^4(\mathbb{R}^n)$ and can bound the first order derivative of the Taylor formula:

$$|\partial_t E[\mathcal{A}\phi(X_s^Z)]| = |\partial_t \langle T(s)\mathcal{A}\phi, \eta\rangle| = |\langle \partial_t T(s)\mathcal{A}\phi, \eta\rangle| = |\langle T(s)\mathcal{A}^2\phi, \eta\rangle| \leq \|\mathcal{A}^2\phi\|_{C_b(\mathbb{R}^n)}.$$

for $s \geq 0$ using Lemma 4.7 and Proposition 4.5 b to check the conditions of the Differentiation lemma [Klenke, 2008, Theorem 6.28]. Via the Taylor formula

$$E[\mathcal{A}\phi(X_h^Z)] = E[\mathcal{A}\phi(Z)] + h\partial_h E[\mathcal{A}\phi(X_s^Z)] \text{ for some } s \in [0, h]$$

we arrive at the following integration rule:

$$T(h)\phi(Z) = \phi(Z) + E\left[\int_0^h \mathcal{A}\phi(X_s^Z)ds\right] = \phi(Z) + h\, E[\mathcal{A}\phi(Z)] + \mathcal{O}(h^2).$$

Together we get that the Euler integrator is of order 1, because

$$E\left[|P(h)\phi(Z) - T(h)\phi(Z)|\right] = \mathcal{O}(h^2).$$

□

Remark. *The notion of Kolmogorov Integrators clearly is weaker than the notion of weak convergence. Note in particular that in weak convergence we demand that the condition $|E[\phi(X_{t_l}^x)] - E[\phi(\hat{X}_{t_l}^x)]| \leq Ch^K$ holds for all $l = 0, \ldots L$. Whereas here we only describe dependence with respect to one step of size h.*

---

[1] $E\left[(\sigma\xi)^T\nabla^2\phi(\sigma\xi)\right] = E\left[\sum_i (\sigma\xi)_i (\nabla^2\phi\sigma\xi)_i\right] = E\left[\sum_i (\sum_l \sigma_{i,l}\xi_l)(\sum_k \sum_j [\nabla^2\phi]_{i,j}\sigma_{j,k}\xi_k)\right] =$
$= \sum_i \sum_l \sum_k \sum_j \sigma_{i,l}[\nabla^2\phi]_{i,j}\sigma_{j,k}E[\xi_l\xi_k] = \sum_i \sum_j (\sum_k \sigma_{i,k}\sigma_{j,k})\,[\nabla^2\phi]_{i,j} =$
$= \sum_i \sum_j (\sigma\sigma^T)_{i,j}[\nabla^2\phi]_{i,j} = Tr\left((\sigma\sigma^T)\nabla^2\phi\right) = \sum_i \sum_j (\sigma\sigma^T)_{i,j}\partial_i\partial_j\phi$

5.4 Noisy Integration

We will now come back to the stochastic regime of Big Data applications. We will see that the stochasticity is only a minor change in the formulae presented above. However, it does have a significant impact on the approximation behavior of the stochastic integrators.

Note that another important aspect of sub-sampling in Hamiltonian Monte Carlo is discussed in [Betancourt, 2015]. He makes a more general point that sub-sampling in HMC is a bad idea with respect to scalability. On the other hand, as [Chen et al., 2015] notes, in real world applications we do not care so much about convergence in the strong sense, but in the weak sense (c.f. [Chen et al., 2015, Appendix J]).

Follow int [Chen et al., 2015], we want to think of noisy integration as a two step process: We first split up equation (2.1) into $L \in \mathbb{N}$ initial value problems

$$\begin{aligned} \mathrm{d}X_t^1 &= b^l(X_t^1)\mathrm{d}t + \sigma(X_t^1)\mathrm{d}B_t, \quad t \in [0,h], & X_0^1 &= x \\ \mathrm{d}X_t^l &= b^l(X_t^l)\mathrm{d}t + \sigma(X_t^l)\mathrm{d}B_t, \quad t \in [(l-1)h, lh], & X_{(l-1)h}^l &= X_{(l-1)h}^{l-1} \quad (l = 2, \dots, L) \end{aligned}$$

where X_s^l be the solution of equation of the l-th equation at time $s \in [(l-1)h, lh]$ and with $b^l = b$ for all $l = 1, \dots L$. Since each of the equations above do have an unique solution (c.f. Theorem 2.3), we have that the composed solution

$$X_t^{e,x} := \sum_{l=1}^{L} X_t^l \, 1_{\{t \in [(l-1)h, lh]\}}, \quad X_0^1 = x \tag{5.10}$$

is equal to the solution of (2.1). Now as a second step we assume that the functions b_l $(l = 1, \dots L)$ cannot be evaluated exactly, but are subject to some noise such that $\tilde{b}_l \approx b$. The corresponding equations

$$\tilde{X}_0^1 = x \tag{5.11}$$

$$\tilde{X}_{(l-1)h}^l = \tilde{X}_{(l-1)h}^{l-1} \quad \text{for } l = 2, \dots L \tag{5.12}$$

$$\text{and} \quad d\tilde{X}_t^l = \tilde{b}_l(\tilde{X}_t^l)\mathrm{d}t + \sigma(\tilde{X}_t^l)\mathrm{d}B_t, \quad \text{for } t \in [(l-1)h, lh] \text{ and } l = 1, \dots L. \tag{5.13}$$

for do now no longer have the same composed solution as the original SDE. However, we can get at least weak uniqueness, if we assume that $\tilde{b}_l$ are unbiased estimates of b. What this means exactly and how this can be shown will be discussed in the subsequent chapters. First, we need to glue these equations back together again:

Definition 5.9. *For $l = 1, \dots L$ and $s \in [(l-1)h, lh]$, we denote the solution to these approximated equations as $\tilde{X}_s^l$ and the* ***composed approximated solution*** *as*

$$\tilde{X}_t^e := \sum_{l=1}^{L} \tilde{X}_t^l \, 1_{\{t \in [(l-1)h, lh]\}}.$$

For $l = 1, \dots L$ and $t \in [0, Lh]$ and for a function ϕ we will define the ***noisy Kolmogorov operators*** *and the* ***noisy Kolmogorov integrator*** *as*

$$\tilde{T}_l(t)\phi(x) := E[\phi(\tilde{X}_t^e)] \tag{5.14}$$

$$\text{and} \quad \tilde{P}_l(h)\phi(x) := E[\phi(\hat{X}_{lh}^l)], \tag{5.15}$$

respectively. Here $\hat{X}_{lh}^l$ denotes an approximate solution to the $l-th$ noisy differential equation $\tilde{X}_{lh}^l$.

A noisy integrator will introduce a second type of error to the integration process, such that we can split up the approximation error into

$$\begin{aligned} E\left[\left|T(t)\phi(x) - \tilde{P}(t)\phi(x)\right|\right] &\leq E\left[\left|T(t)\phi(x) - \tilde{T}(t)\phi(x)\right|\right] + E\left[\left|\tilde{T}(t)\phi(x) - \tilde{P}(t)\phi(x)\right|\right] \\ &=: \varepsilon_{1,t} + \varepsilon_{2,t}. \end{aligned}$$

We will call $\varepsilon_{1,t}$ the **discretization error** and we will call $\varepsilon_{2,t}$ the **sampling error** or the **estimation error**.

Chapter 6

Consistency

The left-hand side of the weak order of convergence condition (5.7) is also called the **Bias** of the estimator $E\left[\phi(\hat{X}_h^x)\right]$ for the true value for the expectation $E\left[\phi(\hat{X}_h^x)\right]$ (e.g. [Lehmann and Casella, 1998, p.5]). On the other hand, the left-hand side of the mean-square order of convergence (5.5) is the square root of the so-called **Mean Square Error**, or **MSE** for short (e.g. [Hastie et al., 2003, p.24]).

This chapter is devoted to show how we can construct integrators which converge to a limit distribution (i.e. the MSE vanishes) and which also approach the correct target distribution (i.e. the Bias vanishes). We will show that both conditions are sufficient to show that an integrator is consistent estimator for the expectation.

We start off by deriving bounds for both, Bias and MSE, in the case of first-order Kolmogorov integrators. Finally, we will show the existence of stationary distributions and give examples for so-called consistent estimators. Throughout this chapter we will follow ideas presented in [Chen et al., 2015] and [Teh et al., 2014].

6.1 The Poisson Equation

One of the main building blocks in the paper of [Chen et al., 2015] is the **Poisson equation**. We will establish a very basic result for it and then use it to derive bounds for the Bias and the MSE.

Definition 6.1. *Let X_t^Z be an Itô diffusion with infinitesimal generator $\mathcal{A}$ and stationary distribution μ, i.e. $T^*(t)\mu = \mu$ for all $t \geq 0$. Let*

$$\bar{\phi} := \int_{\mathbb{R}^n} \phi(y)\mu(dy) = \langle \phi, \mu \rangle$$

*be the expectation of a function $\phi \in C_b(\mathbb{R}^n)$ under the law of μ. Further let $\hat{X}_t^Z$ be an approximate solution for the diffusion equation of X_t^Z. Then the **Poisson equation** is*

defined as the problem of finding a function ψ such that

$$\mathcal{A}\psi(\hat{X}_{t_l}^Z) = \phi(\hat{X}_{t_l}^Z) - \bar{\phi}, \quad l = 1, \dots L. \tag{6.1}$$

So the Poisson equation gives a link between an Itô diffusion, its stationary distribution and its integrator. We define an estimator for $\bar{\phi}$ as

$$\hat{\phi} = \frac{1}{S_L} \sum_{l=1}^{L} h_l \, \phi(\hat{X}_{t_l}), \tag{6.2}$$

where h_l is the step size in the l-th iteration and $S_L := \sum_{l=1}^{L} h_l$ is the total integrated time. If we multiply the equations 6.1 by h_l, sum them up and divide by S_L, we can give the following equivalent formulation:

$$\frac{1}{S_L} \sum_{l=1}^{L} h_l \mathcal{A}\psi(\hat{X}_{t_l}^Z) = \frac{1}{S_L} \sum_{l=1}^{L} \phi(\hat{X}_{t_l}) - \bar{\phi} = \hat{\phi} - \bar{\phi}. \tag{6.3}$$

It is interesting under which condition to the function ψ, equation (6.3) has got a (unique) solution.

The Poisson equation is a standard tool in "averaging, homogenization and ergodic theory" ([Mattingly et al., 2010, Section 4.1]) which can also be used to analyze the "long-term average"(at the same place). In the latest branch of theorems about stationary distributions, a common assumption is to assume a solution of the Poisson equation which is bounded by a Lyapunov type function (e.g. [Teh et al., 2014] and [Chen et al., 2015]). Here we want to follow an approach that is presented in [Mattingly et al., 2010].

Theorem 6.2. *([Mattingly et al., 2010, Assumption 1 and Theorem 4.1]) Let X_t^Z be an Itô diffusion with infinitesimal generator $\mathcal{A}$. Given Sobolev functions $\phi \in W_b^k(\mathbb{R}^n)$ (a weak variant of $C_b^k(\mathbb{R}^n)$) with $k \in \mathbb{N} \cup \{0\}$. We assume that the matrix-valued function $\sigma(x)\sigma(x)^T$ is uniformly positive definite, i.e. there exists a $\alpha > 0$ such that for all $z \in \mathbb{R}^n$ and all $x \in T^n$:*

$$z^T \sigma(x)\sigma(x)^T z \geq \alpha |z|^2.$$

Then there is a unique solution $\psi \in W_b^{k+2}(\mathbb{R}^n)$ to equation (6.3).

Since this theorem is only valid on the n-dimensional torus, we need to make a slight modification on this theorem and make the following assumptions:

Assumption 2. *In the following we assume that*

a) the matrix-valued function $\sigma(x)\sigma(x)^T$ is uniformly positive definite, i.e. there exists a $\alpha > 0$ such that for all $x, z \in \mathbb{R}^n$ such that $z^T \sigma(x)\sigma(x)^T z \geq \alpha |z|^2$

b) Theorem 6.2 also holds for $x \in \mathbb{R}^n$ instead T^n and for functions $\phi \in C_b^2(\mathbb{R}^n)$ instead of $W_b^k(\mathbb{R}^n)$.

c) the unique solution ψ asserted by Theorem 6.2 is in $C_b^4(\mathbb{R}^n)$ instead of $W_b^4(\mathbb{R}^n)$.

Note in particular that for constant $\sigma(x) = \sigma$, the matrix $\sigma\sigma^T$ is uniform positive definite. Now, the following proposition will pave the way for the Poisson equation, connecting it to a first order integrator:

Proposition 6.3. *(c.f. [Chen et al., 2015, Proof of Theorem 3]) Assume assumptions 2 (a) through (c) hold and let $\hat{P}$ be first order Kolmogorov integrator and let $\phi \in C_b^4(\mathbb{R}^n)$. Then,*

$$\begin{aligned}\hat{\phi} - \bar{\phi} = &\frac{1}{S_L}\sum_{l=1}^{L}(E[\psi(\hat{X}_{t_l})] - \psi(\hat{X}_{t_l})) + \frac{1}{S_L}\left(E[\psi(\hat{X}_{t_{L+1}})] - \psi(\hat{X}_0))\right) \\ &- \frac{1}{S_L}\sum_{l=1}^{L}(h_{l+1} - h_l)\mathcal{A}\psi(\hat{X}_{t_l}) - \frac{1}{S_L}h_1\mathcal{A}\psi(\hat{X}_0) + \mathcal{O}\left(\frac{\sum_{l=1}^{L+1} h_l^2}{S_L}\right).\end{aligned} \tag{6.4}$$

Proof. Let P_h be an first order integrator. Then we can expand

$$\begin{aligned}E[\psi(\hat{X}_{t_l})] &= P(h_l)\hat{X}_{t_{l-1}} = e^{h_l\mathcal{A}}\psi(\hat{X}_{t_{l-1}}) + \mathcal{O}(h_l^2) \\ &= (1 + h_l\mathcal{A})\,\psi(\hat{X}_{t_{l-1}}) + \mathcal{O}(h_l^2).\end{aligned}$$

We reorder the equation, sum over all l from 1 to $L+1$ and divide by S_L:

$$\frac{1}{S_L}\sum_{l=1}^{L+1} h_l\mathcal{A}\psi(\hat{X}_{t_{l-1}}) = \frac{1}{S_L}\sum_{l=1}^{L+1}\left(E[\psi(\hat{X}_{t_l})] - \psi(\hat{X}_{t_{l-1}})\right) + \mathcal{O}\left(\frac{\sum_{l=1}^{L+1} h_l^2}{S_L}\right)$$

Now we can rewrite the sum on the right-hand side as

$$\sum_{l=1}^{L+1}(E[\psi(\hat{X}_{t_l})] - \psi(\hat{X}_{t_{l-1}})) = \sum_{l=1}^{L}(E[\psi(\hat{X}_{t_l})] - \psi(\hat{X}_{t_l})) + E[\psi(\hat{X}_{t_{L+1}})] - \psi(\hat{X}_0)$$

The left hand side can further be transformed using the Poisson equation (6.3):

$$\begin{aligned}\frac{1}{S_L}\sum_{l=1}^{L+1} h_l\mathcal{A}\psi(\hat{X}_{t_{l-1}}) &= \frac{1}{S_L}\sum_{l=1}^{L} h_{l+1}\mathcal{A}\psi(\hat{X}_{t_l}) + \frac{1}{S_L}h_1\mathcal{A}\psi(\hat{X}_0) \\ &= \hat{\phi} - \bar{\phi} + \frac{1}{S_L}\sum_{l=1}^{L}(h_{l+1} - h_l)\mathcal{A}\psi(\hat{X}_{t_l}) + \frac{1}{S_L}h_1\mathcal{A}\psi(\hat{X}_0)\end{aligned}$$

Finally, putting everything together, we get (6.4). □

6.2 Error Bounds

Before we give the theorem for bounds of Bias and MSE, we need one more lemma. For convenience we write "$a \lesssim b$" if there is a constant $C > 1$ such that $a \leq C\, b$.

Lemma 6.4. *Let Z be a real-valued random variable with $E[Z^k] < \infty$ for $k \in \mathbb{N}$. Then the Euler-Maruyama approximation $\hat{X}_t^Z$ also has got $E[|\hat{X}_t^Z|^k] < \infty$ for all $l = 0, \dots L$. In particular this holds if $Z \sim \delta_x$ for $x \in \mathbb{R}^n$.*

Proof. We have

$$\begin{aligned} E[|\hat{X}_{l+1}^Z|^k] &= E[|\hat{X}_l^Z + h_l b(\hat{X}_l^Z) + \sqrt{h_l}\sigma(\hat{X}_l^Z)\xi_l|^k] \\ &\lesssim E[|\hat{X}_l^Z|^k] + h_l^k E[|b(\hat{X}_l^Z)|^k] + \sqrt{h_l} E[|\sigma(\hat{X}_l^Z)\xi_l|^k]. \end{aligned}$$

Now if $l = 0$, $E[|\hat{X}_0^Z|^k] = E[|Z|^k] < \infty$ and

$$\begin{aligned} E[|\hat{X}_1^Z|^k] &\lesssim E[|Z|^k] + h_l^k E[|b(Z)|^k] + \sqrt{h_l} E[|\sigma(Z)\xi_l|^k] \\ &\lesssim E[|Z|^k] + h_l^k E[|1+Z|^k] + \sqrt{h_l} E[|1+Z|^k |\xi_l|^k] < \infty. \end{aligned}$$

Using the same expansions and induction over $l = 1, \dots L$ with random variables $\hat{X}_l^{\hat{X}_{l-1}}$ instead Z gives the result. □

Theorem 6.5. *(c.f. [Chen et al., 2015, Theorem 5]) Let Z be a real-valued random variable with $E[Z^4] < \infty$ and let X_t^Z be an Itô diffusion, let $\phi \in C_b^4(\mathbb{R}^n)$ and let $\hat{P}$ be the Euler integrator. Then we can bound the Bias and the MSE via:*

$$\left| E[\hat{\phi}] - \bar{\phi} \right| = \mathcal{O}\left(\frac{1}{S_L} + \frac{h}{S_L} + \frac{\sum_{l=1}^{L+1} h_l^2}{S_L} \right) \tag{6.5}$$

$$\text{and} \quad E\left[|\, \hat{\phi} - \bar{\phi}\, |^2 \right] = \mathcal{O}\left(\frac{1}{S_L} + \frac{1}{S_L^2} + \left(\frac{\sum_{l=1}^{L+1} h_l^2}{S_L} \right)^2 \right) \tag{6.6}$$

Proof. Since for $\phi \in C_b^4(\mathbb{R}^n)$ and $E[|Z|^4] < \infty$, P_h be is a 1st order integrator. Then by taking expectations in (6.4) we get

$$\begin{aligned} \left| E\left[\hat{\phi} - \bar{\phi} \right] \right| \leq & \frac{1}{S_L} \sum_{l=1}^{L} \left| E\left[E[\psi(\hat{X}_{t_l})] - \psi(\hat{X}_{t_l}) \right] \right| + \frac{1}{S_L} \left| E[\psi(\hat{X}_{t_{L+1}})] - \psi(\hat{X}_0)) \right| \\ & + \frac{1}{S_L} \sum_{l=1}^{L} |(h_{l+1} - h_l)| E\left[\left| \mathcal{A}\psi(\hat{X}_{t_l}) \right| \right] + \frac{1}{S_L} h_1 E\left[\left| \mathcal{A}\psi(\hat{X}_0) \right| \right] + \mathcal{O}\left(\frac{\sum_{l=1}^{L+1} h_l^2}{S_L} \right). \end{aligned} \tag{6.7}$$

The first summand is equal to zero. The second can be bounded by S_L^{-1} because $\|\psi\|_\infty < \infty$ according to assumption 2 c. For bounding the third summand note that according

to Lemma 6.4, $E[|\hat{X}_{t_l}|^2] < \infty$. According to Lemma 4.7, this implies that $E[|\mathcal{A}\phi(\hat{X}_{t_l})|]$ is bounded for all $l = 1, \dots L$. Finally the fourth summand is bounded because $x \in \mathbb{R}^n$ is fixed and $|x| < \infty$. Hence, we conclude assertion (6.5).

Regarding the MSE, from proposition 6.3 we get $E\left[|\,\hat{\phi} - \bar{\phi}\,|^2\right]$

$$\begin{aligned}
\lesssim & E\left[\left(\frac{\sum_{l=1}^{L} E[\psi(\hat{X}_{t_l})] - \psi(\hat{X}_{t_l})}{S_L}\right)^2\right] + E\left[\left(\frac{E[\psi(\hat{X}_{t_{L+1}})] - \psi(\hat{X}_0)}{S_L}\right)^2\right] \\
& + E\left[\left(\frac{1}{S_L}\sum_{l=1}^{L}(h_{l+1} - h_l)\mathcal{A}\psi(\hat{X}_{t_l})\right)^2\right] + E\left[\left(\frac{h_1\,\mathcal{A}\,\psi(\hat{X}_0)}{S_L}\right)^2\right] + \mathcal{O}\left(\left[\frac{\sum_{l=1}^{L+1} h_l^2}{S_L}\right]^2\right)
\end{aligned}$$

We want to deal with the first summand the last. The second summand can be bounded by S_l^{-2} because $\psi \in C_b^2(\mathbb{R}^n)$ by assumption 2 c and we can apply Lemma C.2. Similarly as in the case of the Bias, we can bound summand number three and four because $E[|Z|^4] < \infty$ and with to Lemma 6.4, $E[|\hat{X}_{t_l}|^4] < \infty$. According to Lemma 4.7, this implies that $E[|\mathcal{A}\phi(\hat{X}_{t_l})|^2]$ is bounded for all $l = 1, \dots L$. Note that in the third summand we can bound $E[(S_L^{-1}\sum_{l=1}^{L}(h_{l+1} - h_l)\mathcal{A}\psi(\hat{X}_{t_l}))^2] \lesssim S_L^{-2}\sum_{l=1}^{L} h_l^2 \leq S_L^{-1}$. Finally, we calculate for the first summand:

$$E\left[\left(\frac{\sum_{l=1}^{L} E[\psi(\hat{X}_{t_l})] - \psi(\hat{X}_{t_l})}{S_L}\right)^2\right] \lesssim \frac{1}{S_L^2}\sum_{l=1}^{L}\mathrm{Var}\left[\psi\left(\hat{X}_{t_l}^{\hat{X}_{t_{l-1}}}\right)\right] = \mathcal{O}\left(\frac{1}{S_L^2}\sum_{l=1}^{L} h_l\right) = \mathcal{O}\left(\frac{1}{S_L}\right),$$

because ψ has got continuous derivatives up to order 4 by assumption 2 c and $\mathrm{Var}[\psi(\hat{X}_{t_l}^{\hat{X}_{t_{l-1}}})] = \mathcal{O}(h_l)$ by Proposition C.4 (note that the Euler integrator if of weak order 1 if $\psi \in C_b^4(\mathbb{R}^n)$, c.f. Lemma 5.5). This shows (6.6) and the proof is complete. □

If we want to get similar results for the noisy settings, we apply the following approximation using the noisy operators and noisy integrators from Definition 5.9 (c.f. also [Chen et al., 2015]):

$$\begin{aligned}
E\left[\phi(\tilde{X}_t^e)\right] &= \tilde{T}_L(h) \circ \cdots \circ \tilde{T}_1(h) \\
&= \tilde{P}_L(h) \circ \cdots \circ \tilde{P}_1(h) \\
&= E\left[\phi(\hat{X}_t^l)\right].
\end{aligned}$$

For the following theorem we further need to give the "noise" in $\tilde{T}$ a name and we define the **noise operator**

$$\Delta V_l := \mathcal{A} - \tilde{\mathcal{A}}_l, \quad (l = 1, \dots L) \tag{6.8}$$

where $\tilde{\mathcal{A}}_l$ is the infinitesimal generator of the noisy operator $\tilde{T}$.

Theorem 6.6. *Let Z be a real-valued random variable with $E[Z^4] < \infty$ and let X_t^Z be an Itô diffusion, let $\phi \in C_b^4(\mathbb{R}^n)$ and let $\tilde{P}$ be the noisy Euler integrator. Assume that the noise operator and is absolutely and in square bounded in expectation, i.e.*

$$E[|\Delta V_l\, \psi(Z)|] \leq E[\|\Delta V_l\, \psi\|_\infty] \leq E[\|\Delta V_l\,\|] < \infty$$
$$\text{and } E[|\Delta V_l\, \psi(Z)|^2] \leq E[\|(\Delta V_l\, \psi)^2\|_\infty] \leq E[\|\Delta V_l^2\|] < \infty$$

for all $l = 1, \dots L$, all $\psi \in C_b^4(\mathbb{R}^n)$ and all square integrable random variables Z (we use the classical operator norm $\|\Delta V\| := \sup_\psi \|\Delta V\, \psi\|_\infty$). Then we can bound the Bias and the MSE via:

$$\left|E\left[\tilde{\phi} - \bar{\phi}\right]\right| = \mathcal{O}\left(\frac{1}{S_L} + \frac{h}{S_L} + \frac{\sum_{l=1}^{L+1} h_l E[\|\Delta V_l\|]}{S_L} + \frac{\sum_{l=1}^{L+1} h_l^2}{S_L}\right)$$
$$\text{and} \quad E\left[\left(\tilde{\phi} - \bar{\phi}\right)^2\right] = \mathcal{O}\left(\frac{1}{S_L} + \frac{1}{S_L^2} + \frac{\sum_{l=1}^{L+1} h_l^2 E[\|\Delta V_l\|^2]}{S_L^2} + \left(\frac{\sum_{l=1}^{L+1} h_l^2}{S_L}\right)^2\right).$$

Proof. In the proofs of Proposition 6.3 above, we set $\tilde{T}_l = T + \Delta V_l$ before summing over all $l = 1, \dots L$.

$$\begin{aligned} E[\psi(\tilde{X}_{t_l})] &= \left(1 + h_l \tilde{\mathcal{A}}\right) \psi(\hat{X}_{t_{l-1}}) + \mathcal{O}(h_l^2) \\ &= (1 + h_l \mathcal{A} + h_l \Delta V_l)\, \psi(\hat{X}_{t_{l-1}}) + \mathcal{O}(h_l^2). \end{aligned}$$

As before, we reorder the equation, sum over all l from 1 to $L+1$ and divide by S_L:

$$\frac{1}{S_L}\sum_{l=1}^{L+1} h_l \mathcal{A}\psi(\hat{X}_{t_{l-1}}) = \frac{1}{S_L}\sum_{l=1}^{L+1}\left(E[\psi(\hat{X}_{t_l})] - \psi(\hat{X}_{t_{l-1}})\right) - \frac{1}{S_L}\sum_{l=1}^{L+1} h_l \Delta V_l \psi(\hat{X}_{t_{l-1}}) + \mathcal{O}\left(\frac{\sum_{l=1}^{L+1} h_l^2}{S_L}\right).$$

We immediately see that the resulting equation for $\tilde{\phi} - \bar{\phi}$ will only be differ by a factor of $\frac{1}{S_L}\sum_{l=1}^{L+1} h_l E[\Delta V_l \psi(\hat{X}_{t_{l-1}})]$. Additionally inserting this factor into the proof of the bias 6.5, we get

$$\left|\frac{1}{S_L}\sum_{l=1}^{L+1} h_l E[\Delta V_l \psi(\hat{X}_{t_{l-1}})]\right| \leq \frac{1}{S_L}\sum_{l=1}^{L+1} h_l \left|E[\Delta V_l \psi(\hat{X}_{t_{l-1}})]\right| = \mathcal{O}\left(\frac{\sum_{l=1}^{L+1} h_l E[\|\Delta V_l\|]}{S_L}\right)$$

because ψ is bounded by Assumption 2. By the same argumentation, for the MSE we get an additional term of

$$\frac{1}{S_L^2} E\left[\left(\sum_{l=1}^{L+1} h_l \Delta V_l \psi(\hat{X}_{t_{l-1}})\right)^2\right] = \frac{1}{S_L^2} E\left[\left(\sum_{l=1}^{L+1} h_l \Delta V_l \psi(\hat{X}_{t_{l-1}})\right)^2\right] \lesssim \frac{\sum_{l=1}^{L+1} h_l^2 E\left[|\Delta V_l|^2\right]}{S_L^2}.$$

Using convexity of the square function, we arrive at the desired formula for the MSE. □

Remark. *The above Theorems are only given for integrators of order $K = 1$. For higher orders please refer to [Chen et al., 2015, Theorems 2, 3 and 5]. We took most of the ideas of the proof from this source. The last step using the variance of the approximation given in Appendix C, however, is our own result and significantly simplifies the proof in [Chen et al., 2015, Theorems 3 and 5].*

6.3 Consistent Estimators

In this section we want to use the results we obtained above. For this we need the following important concept:

Definition 6.7. *([Lehmann and Casella, 1998, Definition 8.1]) Let $X_1, \dots X_n$ be i.i.d. random variables following the law μ. Let g be a statistics which is to be estimated. We say that a statistics $\hat{g}_n = \hat{g}_n(X_1, \dots, X_n)$ is a* ***consistent estimator*** *for g, if*

$$\hat{g}_n \to^{\mu} g \quad \text{in measure.} \tag{6.9}$$

Lemma 6.8. *(c.f. [Lehmann and Casella, 1998, Theorem 8.2]) A sufficient condition for consistency of an estimator $\hat{g}_n$ is the following:*

$$E[\hat{g}_n - g]^2 \to 0 \tag{6.10}$$

$$\text{and} \quad [E[\hat{g}_n] - g] \to 0, \quad Var[\hat{g}_n] \to 0 \quad \text{as } n \to \infty. \tag{6.11}$$

If we can show these conditions for $\hat{\phi} - \bar{\psi} = \frac{1}{S_L} \sum_{l=1}^{L} h_l \, \phi(\hat{X}_{t_l}), - \langle \phi, \mu \rangle$, then we can conclude $\hat{\phi} \to \bar{\phi}$ in probability. We expressed the bounds on the Bias and the MSE in terms of S_L^{-1} and, $E[\|\Delta\|]$ and $S_L^{-1} \sum_{l=1}^{L+1} h_l^2$ which leads us to the following two assumptions on the step sequence and the noise operator:

Assumption 3. *The step sizes $(h_l)_{l \in \mathbb{N}}$ are such that the following two conditions hold:*

$$S_L := \sum_{l=1}^{L} h_l \to \infty \quad \text{and} \quad \frac{\sum_{l=1}^{L+1} h_l^2}{S_L} \to 0 \quad \text{as } L \to \infty.$$

Remark. *There is a certain resemblance of this sequence of step sizes with* ***Simulated Annealing****: they need to be cooled down, yet not too fast. In fact, any Itô diffusion equation can be rearranged to behave as a simulated annealing algorithm. Let for example $dX_t = D\nabla U(X_t)dt + \sqrt{2D}dB_t$ a diffusion equation with canonical density $\propto e^{-U(x)}$. Then $dX_t := \tilde{D}\nabla\tilde{U}(X_t)dt + \sqrt{2\tilde{D}/\beta}dB_t$ has got the canonical density $\propto e^{-\beta\tilde{U}(x)}$ with $\tilde{D} := \beta D$ and $\tilde{U} = \beta U$ (c.f [Pelletier, 1998, 1.1], [Robert and Casella, 2004, p. 202], [Chow et al., 2009]).*

Assumption 4. *Let $\tilde{T}_l$ be a noisy Kolmogorov operator. We assume that we can express $\tilde{T}_l$ as $\tilde{T}_l = T + \Delta V_l$. We assume that*

$$E[\|\Delta V_l\|] = 0 \quad \textit{and} \quad E[\|\Delta V_l^2\|] < \infty \quad \textit{for all } l = 1, \dots L.$$

The first condition in ΔV_l ensures that the operator $\tilde{T}$ is not biased. [Chen et al., 2015] use an operator of the form $\Delta V_l = (\nabla_\theta \tilde{U}_l - \nabla_\theta U) \cdot \nabla_p$ and also assume that it is unbiased. We can use these two assumptions in order to show:

Theorem 6.9. *Let Z be a real-valued random variable with $E[Z^4] < \infty$ and let X_t^Z be an Itô diffusion, let $\phi \in C_b^4(\mathbb{R}^n)$ and let $\tilde{P}$ be the noisy Euler integrator. Under Assumptions 3 and 4 we have*

$$\hat{\phi} \to^P \bar{\phi} \textit{ and } \tilde{\phi} \to^P \bar{\phi} \quad \textit{in probability as } L \to \infty.$$

Proof. The proof is an immediate consequence of the bounds on the Bias and the MSE. Together with Lemma 6.8 they show that the noisy Euler integrator is consistent. □

This theorem is important because if we want to device practical samplers which have some desired probability density as a stationary density π, we want the sampler (or integrator) to converge to the correct π. This is what the theorem gives us.

Remark. *An important theorem which also motivates the choice of $\hat{\phi} = S_L^{-1} \sum_{l=1}^L \phi(\hat{X}_l)$ is the Krylov-Bogoliubov theorem. Together with the properties of the integrator and the bounds for the MSE, it is an alternative way for showing existence of a stationary distribution and shall be stated here for completeness:*

Theorem 6.10. *([Prato and Zabczyk, 2003, 3.1 and Theorem 3.1.1]) Assume that $T(t)$ is a Feller semigroup, i.e. it fulfills $\|T(t)\phi\|_\infty \leq \|\phi\|_\infty$ and $T(t)\phi \in C_b(\mathbb{R}^n)$. If for some $\nu \in \mathcal{M}_1(\mathbb{R}^n)$ we have $\frac{1}{S_n} \int_0^{S_n} T^*(t)\nu dt \to \mu$ weakly as $n \to \infty$, then μ is an invariant measure for $T(t)$, $t \geq 0$.*

Chapter 7

Numerical Experiments

There is a wide variety of distances for probability distributions which have been introduced by diverse fields such as measure theory, statistics, information theory, or Machine Learning. A comprehensive list of distances and their relations to each other can be found, e.g. in [Gibbs and Su, 2002]. One of the most comprehensive distance metrics in the one-dimensional case is the **Kolmogorov-Smirnov distance**. For two distribution functions F and G, it is defined as (c.f. [Gibbs and Su, 2002]):

$$d_K(F, G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|$$

The Kolmogorov-Smirnov distance cannot be used to bound the total variation norm because it is bounded by the discrepancy which in turn is bounded by the total variation norm (c.f. [Gibbs and Su, 2002]). However, we can use it to metricize weak convergence of random variables (c.f. [Gibbs and Su, 2002, Theorem 6]) and it is still more meaningful than Bias or MSE, because they strongly depend on the test function of choice.

Dealing with data sets, the following empirical variant of this distance is important: Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of a random variable with distribution function F. The **empirical distribution function** of F is defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}.$$

By the *Glivenko-Cantelli Theorem* $\hat{F}_n$ is a consistent estimator of the distribution function F (c.f. [Robert and Casella, 2004], Chapter 1.6.2). The **Kolmogorov-Smirnov test statistics**, or **KS-statistics**, for $\hat{F}_n$ is defined as (c.f. [Robert and Casella, 2004], Section 12.2.2):

$$K(\hat{F}_n, F) := \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \max_{1 \leq i \leq n} \max \left\{ \frac{i}{n} - F\left(X_{i:n}\right), F\left(X_{i:n}\right) - \frac{i-1}{n} \right\}, \quad (7.1)$$

where $X_{i:n}$ is the order statistics with $X_{1:n} = \min_{1 \leq i \leq n} X_i$ and $X_{i:n} = \min\{X_i : X_i > X_{j-1:n}\}$ for $1 < j \leq n$ (c.f. [Georgii, 2009, Section 8.3]). The order statistics is a reordering of X_i in ascending order. Of course, the KS-statistics can also be used to compare two empirical distribution functions with each other.

7.1 Sampling

Let $\rho(x)$ be the density of the distribution η we want to draw samples from. We can rewrite ρ as a Boltzmann density

$$\rho(x) \propto \exp\{\log \rho(x)\} =: \exp\{-U(x)\}$$

with $U(x) = -\log \rho(x)$. As we have seen in Theorem 4.16 above, ρ is the stationarity distribution to the following Itô diffusion equation (among others):

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t.$$

If we simply draw realizations from this equation with constant step sizes $h > 0$ using the Euler-Maruyama method, this sampler is called the **Unadjusted Langevin Algorithm**, or **ULA** for short. As [Roberts and Tweedie, 1996a] describe in their paper, the Euler-Maruyama discretization of the Langevin diffusion may not have the right stationary distribution ρ. It is known to converge to the wrong distribution in general such that the resulting Markov chain will be biased in distribution (c.f. [Roberts and Tweedie, 1996a] for an one-dimensional diffusions).

One way of handling the bias of the ULA-sampler is the **Metropolis-Hastings correction** (c.f. [Robert and Casella, 2004, Chapter 7]). In the context of diffusion equations this method is also known as *Metropolis Adjusted Langevin Algorithm*, or **MALA** for short. Let the **proposal density** $q(x|X)$ be given by one ULA-step with constant step size h (c.f. [Roberts and Tweedie, 1996a, Chapter 3]):

$$q(Y|X) = \mathcal{N}(X - \nabla U(X), \sqrt{2h}1). \tag{7.2}$$

In MALA we want to accept a proposal with the **acceptance probability** $\alpha(Y|X)$, otherwise keep the old value for the next sample.

$$\alpha(Y|X) := \min\left\{1, \frac{\rho(Y)q(X|Y)}{\rho(X)q(Y|X)}\right\}.$$

The acceptance probability is chosen in such a way that together with certain conditions on the proposal kernel (namely η-irreducibility and aperiodicity) it ensures that the resulting Markov chain converges to the correct posterior distribution (c.f. [Robert and Casella, 2004, Theorem 7.2 and 7.4, and Lemma 7.6] and [Roberts and Tweedie, 1996b]).

The performance of the MALA is critically linked to high acceptance rates. The closer the proposal gets to the target distribution, the better. But even with a well-chosen proposal density, calculating the acceptance probability is very cost-intensive.

Another resource-efficient algorithm can be constructed using Assumption 3 the results of Theorem 6.9:

Definition 7.1. *The **Langevin Dynamics Annealing** or **LDA** algorithm be defined as the ULA algorithm with decreasing step-sizes with the following schedule (c.f. [Welling and Teh, 2011, p.2]):*

$$h_l := \frac{a}{(b+l)^\gamma}. \tag{7.3}$$

Note that (7.3) *fulfills the assumptions 3 necessary for consistency.*

An additional option to the LDA algorithm is to delay the schedule by a factor M and to use

$$\frac{a}{(b+c(l))^\gamma}$$

*with a counter $c(l)$ that is only incremented every M-th step. We call this algorithm **Delayed Langevin Dynamics Annealing** or **DLDA**.*

7.2 Stochastic Gradient Monte Carlo

In order to apply the algorithms presented above to stochastic error functions, we need to give a short introduction to one of the most important concepts of parameter estimation in Statistics (a.k.a. **Statistical Learning**). Lets assume that our probability density p depends on an additional parameter θ:

$$p(x) = p(x|\theta).$$

The **Maximum A Posteriori estimate**, or **MAP estimate** ([Hastie et al., 2003], [Bishop, 2006]) aims at finding the parameter θ for which a set of independent identically distributed data observations D with $d_i \in D$ and $d_i \sim p(\cdot|\theta)$ is most probable:

$$\max_\theta P(\theta|D) \tag{7.4}$$

We can use Bayes' Theorem (e.f. [Klenke, 2008, Theorem 8.7]) and apply a logarithm to the maximization problem, we easily derive the standard formulation for most statistical learning problems ([Chen et al., 2014]).

$$\max_\theta \left\{ \log P(\theta) + \sum_{i=1}^{N} \log P(d_i|\theta) \right\} \tag{7.5}$$

The first term in this formula is called the (logarithmic) *Prior* of the parameter θ, or the *regularization term*; the second term is called the *Log-likelihood function* of the parameter θ ([Bishop, 2006], p.22).

7.3 Experiment: Mixture of Gaussians

One of the simplest densities to sample from is the one-dimensional mixture of Gaussians density

$$\rho(x) = \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{m} \omega_j e^{-(x-\mu_j)^2/2} = \frac{1}{\sqrt{2\pi}} \left[\sum_{j=2}^{m} \omega_j e^{-(x-\mu_j)^2/2} + \left(1 - \sum_{j=2}^{m} \omega_j\right) e^{-(x-\mu_1)^2/2} \right]$$

with $\omega_j \in [0,1]$ and $\sum_{j=1}^{m} \omega_j = 1$. Now in order to express ρ in terms of a Boltzmann density and we get $\rho(x) = Z^{-1} e^{-U(x)}$ with $Z = 1$ and

$$U(x) = -\log \rho(x) = -\log \left(\sum_{j=1}^{m} \omega_j e^{-(x-\mu_j)^2/2} \right) + \frac{1}{2} \log(2\pi)$$

with partial derivatives

$$\frac{\partial}{\partial x_j} U(x) = \frac{\omega_j (x - \mu_j) e^{-(x-\mu_j)^2/2}}{\rho(x)} \text{ for } j = 1, \dots m.$$

Note that we have $|\nabla U(x)| \leq C(1 + |x|)$ because

$$|\partial_{x_j} U(x)| = \left| \frac{x - \mu_j}{1 + \sum_{i \neq j} \frac{\omega_i}{\omega_j} e^{-(x-\mu_i)^2/2 + (x-\mu_j)^2/2}} \right| \leq |\mu_j|(1 + |x|).$$

Let F be the cumulative distribution function of a standard Gaussian random variable. Then a multi-modal Gaussian distributed random variable X has got the cumulative distribution function

$$G(x) := P(X \leq x) = \sum_{j=1}^{m} \omega_j \; F(x - \mu_j)$$

which we will use to calculate the Kolmogorov-Smirnov statistics.

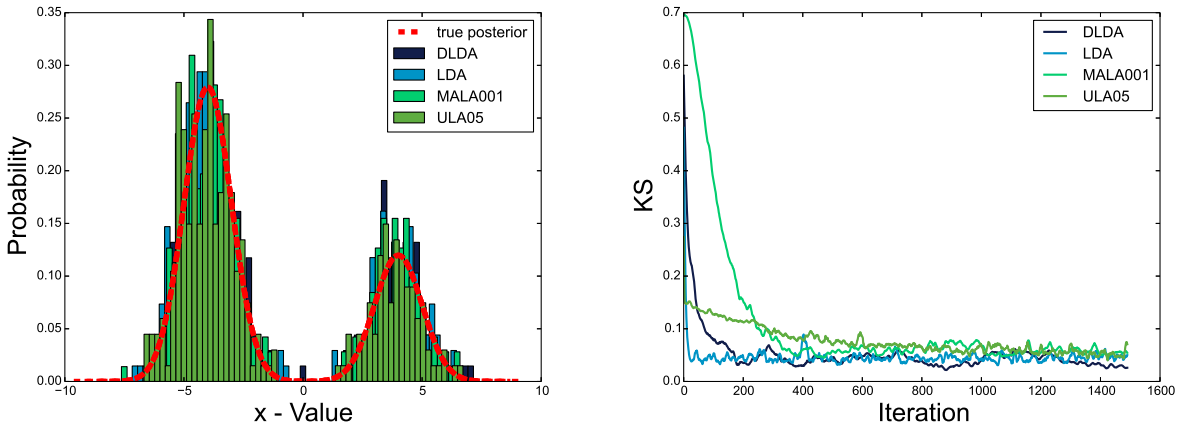
For the stochastic Gradient experiments we make the following assumptions and definitions:

We want to assume that the prior $p(\theta)$ is constant such that it does not influence the optimization procedure. Also, we will only optimize for $\theta = (\mu_1, \dots \mu_n)$ since searching for optimal ω requires additional boundary conditions. As above we define *objective function*, or *potential* $U : \mathbb{R}^n \to \mathbb{R}$ with respect to θ ([Chen et al., 2014]) as

$$U(\theta) = -\sum_{i=1}^{N} \log P(d_i|\theta) \tag{7.6}$$

and instead of *maximizing* (7.5) we want to *minimize* it. The gradient of U can be calculated as follows:

$$\frac{\partial}{\partial \mu_j} U(x) = -\frac{\partial}{\partial x_j} U(x).$$



(a) Sampling histogram (b) Kolmogorov-Smirnov statistics (KS)

Figure 7.1: The Bi-modal Gaussians experiment

Simulation Results

We simulated the Bi-modal Gaussian model for each of the four samplers ULA, MALA, LDA and DLDA. If different step sizes are used, we append it without decimal dot to the sampler name, for example "ULA05" for the ULA algorithm with constant step-size 0.5. The parameters for the LDA schedules will be announced for each experiment.

The general setup is this: in each experiment for each sampler we simulate $m = 300$ independent Markov chains all starting in $x_0 = 0$ for $N = 1500$ steps.

Bi-modal Gaussians

We chose the following model:

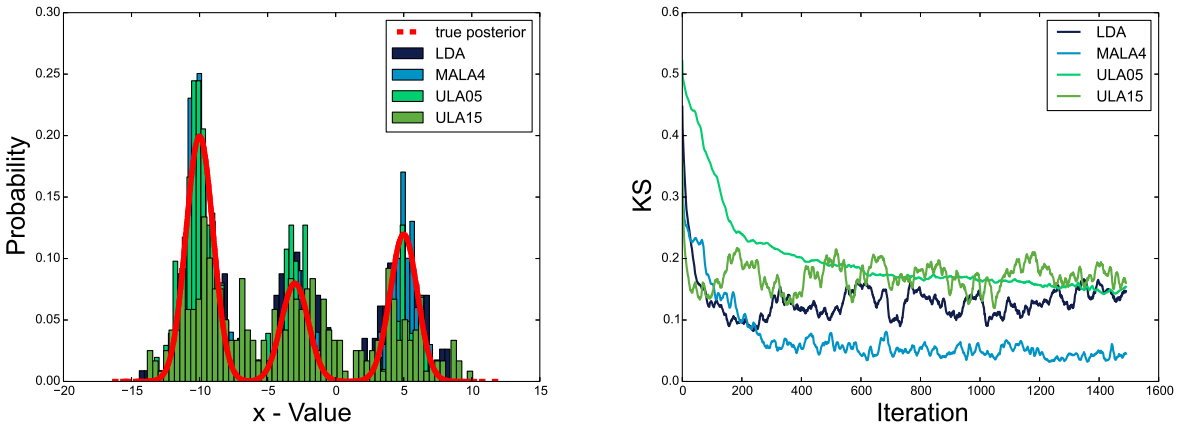
$$\mu_1 = 4 = -\mu_2 \quad \text{and } \omega_1 = 0.7,\ \omega_2 = 0.3$$

For the LDA and DLDA algorithm the following parameter choices of cooling schedule proved to be quite good[1]:

$$a = 1, \quad b = 5, \quad \gamma = 1, \quad M = 20.$$

We see from Figure 7.1 that the algorithm converges very well to the correct distribution for all samplers.

[1]Kolmogorov-Smirnov statistics for the LDA after 500 iterations and evaluated on 120 sampling paths averaged over 30 runs each: 0.063, c.f. appendix E



(a) Histogram comparison

(b) Kolmogorov-Smirnov statistics (KS)

Figure 7.2: The Tri-modal Gaussians experiment

Tri-modal Gaussians

The situation for the tri-modal Gaussian model is a bit different: We chose the following model:

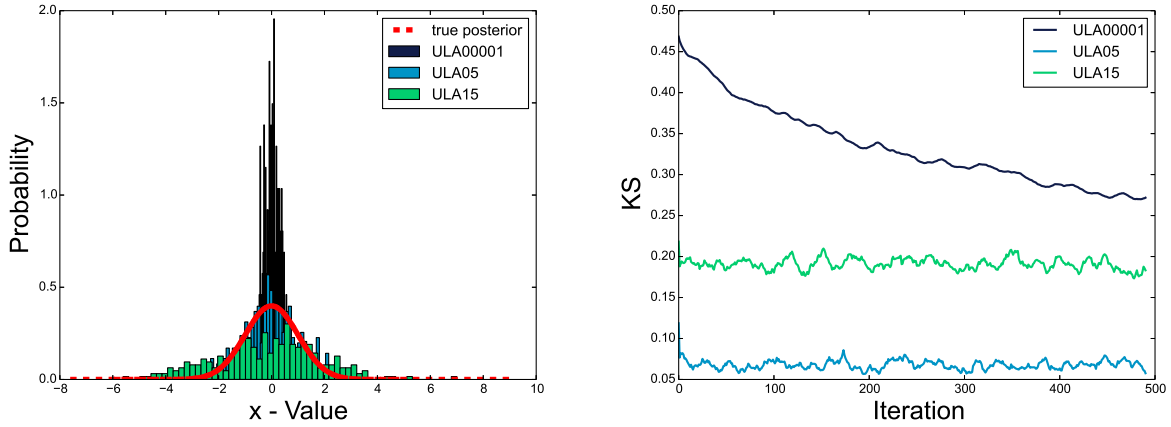
$$\mu_1 = -10, \quad \mu_2 = -3, \quad \mu_3 = 5 \quad \text{and } \omega_1 = 0.5, \quad \omega_2 = 0.2, \quad \omega_3 = 0.2.$$

The cooling schedules in this setup look quite different. Not only is the optimal choice of parameters somewhat reversed, but good results for the bi-modal model give poor results in the tri-modal case and vice versa (c.f. appendix E):

$$a = 2, \quad b = 1, \quad \gamma = 0.7, M = 20$$

We see from Figure 7.2 that the algorithm converges quite well to the correct distribution for all samplers. Comparing the two Figures 7.2b with 7.1b, however, we see that it is is significantly more difficult to obtain accurate samples for the tri-modal case compared to the bi-modal Gaussians. One of the reasons can be interpreted from Figure 7.2a: while ULA with step size 1.5 overemphesizes the regions between the modes, the LDA has got some difficulty escaping the area around the mode at -3. The reason for this is that it is the annealing: it becomes more and more improbable that the LDA leaves a mode to go to another if there is a through in between.

It is interesting to view the plots for MALA and DLDA (Figure 7.4): While the DLDA is not effective at all, for both small ($h = 0.05$) and medium ($h = 1.5$) step sizes, the algorithm converges very slowly although it has very high acceptance rates. Only for a step size of 4, it gives good results (c.f. Figure 7.2).



(a) Sampling histogram

(b) Kolmogorov-Smirnov statistics (KS)

Figure 7.3: The uni-modal Gaussians experiment

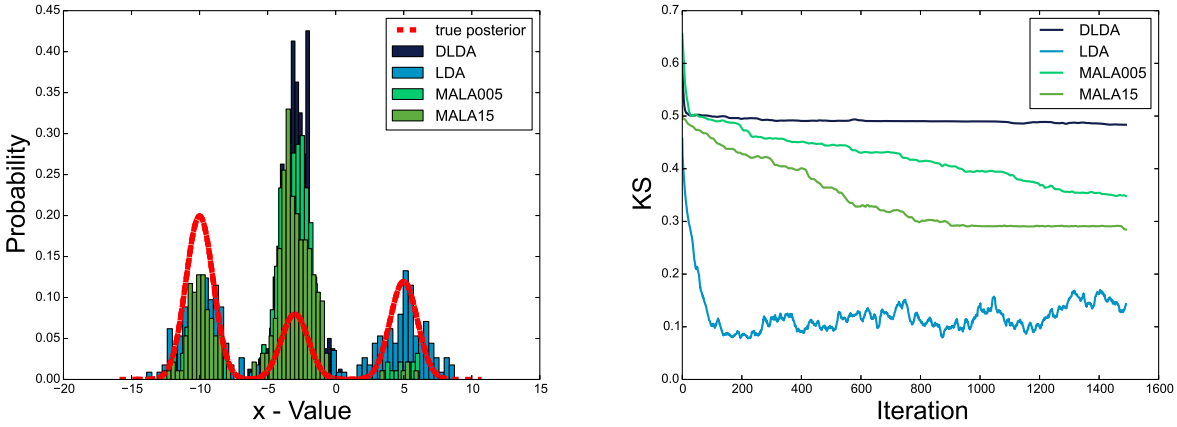
Exploring the State Space

The reason for the behaviour of MALA in the case above comes from the fact that MALA is very slow in exploring the state space. We can observe the same behaviour for the ULA in the uni-modal Gaussian case: If the step-size is too small, then accurate sampling takes a lot of time. If it is too large, then the algorithm is fundamentally biassed. We can observe this behaviour in Figures 7.3a and 7.3b.

In cases like this other methods can be used, for example **Hamiltonian Monte Carlo** methods (e.g. [Andrieu et al., 2003, Section 3.6.1] or [Sanz-Serna, 2014, Section 8]). These are a class of higher-order methods that traverse the sampling space much more efficient and yield better proposals. Methods like these could not be surveyed in this thesis, but they easily fit in the framework presented here. Modern presentations of these higher-order methods can be found for example in [Chen et al., 2015], [Ma et al., 2015], [Bussi and Parrinello, 2007].

Measure of Convergence

It would be a good idea to use the Kolmogorov-Smirnov statistics as a measure for convergence. We can compare the Kolmogorov-Smirnov statistics of iteration l with iteration $l - 10$ for all m Markov chains in Figure 7.5. In the plot it seems as if ULA with step size 0.5 shows the same behaviour as the other two methods from iteration 200. However, as we can clearly see in Figure 7.2b, in reality it approches the target distribution much slower. Therefore, it is not a very good method to test for convergence and only works when the target distribution is known analytically.



(a) Histogram comparison (b) Kolmogorov-Smirnov statistics (KS)

Figure 7.4: The Tri-modal Gaussians experiment

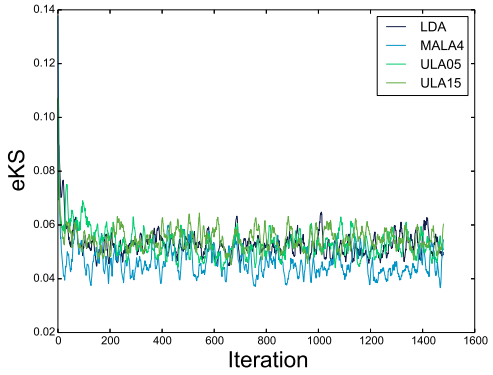


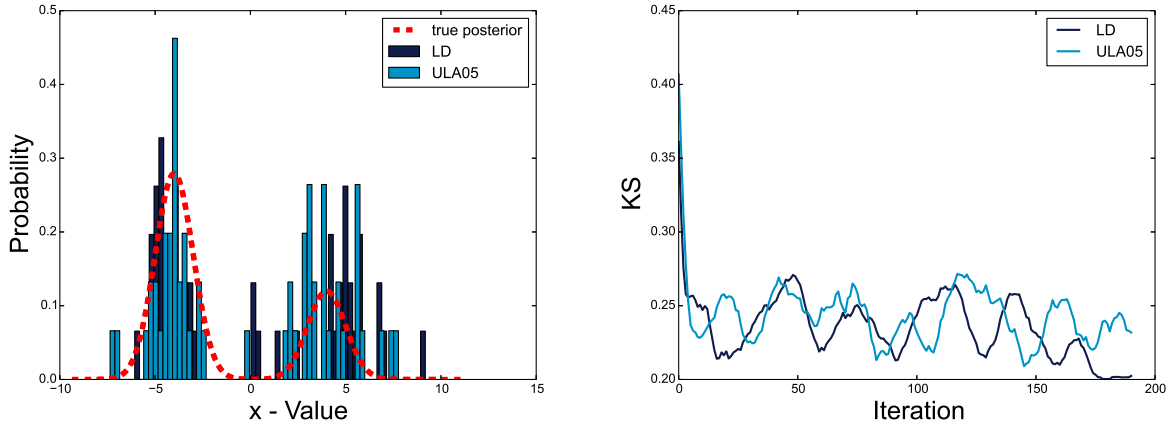
Figure 7.5: Tri-modal Gaussians: comparison of the KS-statistics with the 10th previous step

Stochastic Bi-modal Gaussians

In the stochastic experiment, we drew $m = 5000$ samples $D = \{d_i\}_{i=1}^m$ from the distribution $P(x|\mu_1, \dots \mu_m) \propto \sum_{j=1}^m \omega_j e^{-(x-\mu_j)^2/2}$ with $\mu_1 = -\mu_2 = 4$.

According to equation (7.5), using constant a prior independent of μ, we have

$$P(\theta|D) \propto \prod_{i=1}^{N} P(d_i|\theta)$$



(a) Histogram comparison

(b) Kolmogorov-Smirnov statistics (KS)

Figure 7.6: Joint distributions of μ: Stochastic Gradient Monte Carlo Gaussians

Using the stochastic gradient with a batch size of 50, we have for each iteration $l = 1, \dots L$:

$$\nabla \tilde{U}_l(\mu) = -C \log P(\theta|D) = -C \sum_{i=1}^{N} P(d_i|\theta).$$

The simulation results for 30 independent chains and 180 iterations are visible in Figure 7.6. The figures show the joint distributions of μ_1 and μ_2 and it is cisible that both exact modes are approximated well. The quite large Kolmogorov-Smirnov statistics greater than $1/2$ which is probably due to the small batch size. Too small batches put a bias on the stochastic gradient such that Theorem 6.9 does not hold.

Chapter 8

Conclusion

In this thesis we gave a full description of the theory behind modern Stochastic Gradient Markov chain Monte Carlo methods. To this end we introduced semigroup theory and applied it to the Kolmogorov operator and its dual and dual density operators. We showed how the infinitesimal generator of the Kolmogorov semigroup can be used to study time behavior of the Markov chain in the abstract Cauchy problem. We gave an explicit description of the generators and used them to derive stochastic equations which have the canonical distribution as stationary distribution. Finally, we saw how simple numerical integration schemes can be derived and we gave conditions under which these schemes approach the correct stationary distribution of the continuous equation. The last chapter showed results and limitations of the theory presented here.

The most interesting aspect of this work was to bring together Stochastic Gradient Descent and MCMC methods. Unfolding the theory of semigroups of Itô diffusions such that the Fokker-Planck equation can applied proofed to be full of intricacies. However, the theory is sound and it is worth being studied because it teaches many subtleties of the nature of stochastic processes.

Future research could focus on a more in depth study of the interrelation of Stoachstic Gradient Descent methods with the Simulated Annealing algorithm. It would also be interesting to review different convergence tests, as the MSE does not give much information about stationarity of the distribution and the Kolmogorov-Smirnov statistics does not bound the total variation norm. Finally, the study of higher-order integrators would be the next step to continue the work of this thesis. This is a very active field of research, partially undergone by Google and other companies. The complete description of diffusion equations discussed in this work opens the door to plentiful new integration, sampling and optimization methods which will most probably be the subject of many interesting publications in the next couple of years.

Appendix A

Random Starting Points

There are two different ways we can interpret the expression " X_t^Z ": The first interpretation is the one presented in [Øksendal, 1998, Theorem 5.2.1] and which we followed until here: X_t^Z is the solution to the differential equation

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \;\; t \geq 0 \;\text{ with initial value } X_0 = Z.$$

In integral notation, this reads:

$$X_t^Z(\omega) = Z(\omega) + \int_0^t b(X_s^Z(\omega))ds + \int_0^t \sigma(X_s^Z(\omega))dB_s(\omega).$$

Another way of starting an Itô diffusion process at a random distribution is the one presented in [Durrett, 2004] for Brownian motions: We start a process for each $\tilde{\omega} \in \Omega$:

$$d\tilde{X}_t = b(\tilde{X}_t)dt + \sigma(\tilde{X}_t)dB_t, \;\; t \geq 0 \;\text{ with initial value } \tilde{X}_0 = Z(\tilde{\omega}).$$

Which makes $\tilde{X}_t^Z$ a random variable on the product space $\Omega \times \Omega$:

$$\tilde{X}_t^Z : (\omega, \tilde{\omega}) \mapsto X_t^{Z(\tilde{\omega})}(\omega).$$

The distribution of this random variable can be gotten by averaging over all $\tilde{\omega}$ (c.f. [Durrett, 2004, p.4]):

$$\tilde{\mu}_t^Z(A) = \int_\Omega \mu_t^{Z(\tilde{\omega})}(A)P(d\tilde{\omega})$$

for $A \in \mathcal{F}$. Of course, we would like to know if we can say that the two distributions are equal, i.e. whether $\mu_t^Z = \tilde{\mu}_t^Z$, or in other words if

$$E[\phi(X_t^Z)] = \int_\Omega E[\phi(X_t^{Z(\tilde{\omega})})]P(d\tilde{\omega}) = \int_{\Omega\times\Omega} \phi\left(X_t^{Z(\tilde{\omega})}(\omega)\right) P \otimes P(d\tilde{\omega} \times d\omega) = E[\phi(\tilde{X}_t^Z)]$$

for $\phi \in C_b(\mathbb{R}^n)$. This is true by Theorem 3.11:

$$E[\phi(X_t^Z)] = \langle \phi, T^*(t)\eta \rangle = \langle T(t)\phi, \eta \rangle = E[\phi(\tilde{X}_t^Z)].$$

This statement holds for all $\phi \in C_b(\mathbb{R}^n)$, and therefore [Elstrodt, 2009, VIII. Theorem 4.6] guarantees that indeed $\mu_t^Z = \tilde{\mu}_t^Z$.

Appendix B

Absolute Continuity

One more thing we would like to show which follows from the smoothness of the density of the Brownian motion:

Conjecture. *Let X_t^x be an Itô diffusion such that its initial distribution is the Dirac distribution δ_x. Then we have $\mu_t^x \ll \lambda$ for all $t > 0$.*

The idea behind this claim is that the distribution of the Brownian Motion is absolutely continuous. From the construction of the Itô integral follows the following procedure:

1. The Brownian differences are absolutely continuous: $B_t - B_s \sim \mathcal{N}(0, \sqrt{t-s})$
2. σ and X_t are measurable such that the distribution of $Y_{n,j}$ is absolutely continuous with
$$Y_{n,j}(t,\omega) = \sigma(X(t_j,\omega))\chi[t_j, t_{j+1})(t)(B_{t_{j+1}} - B_{t_j})(\omega)$$
3. The distribution of $Y_n := \sum_{j\geq 0} Y_{n,j}$ is absolutely continuous
4. The limit $Y := \lim_{n\to\infty} Y_n$ has got an absolutely continuous distribution.

$$Y(t,\omega) := \lim_{n\to\infty} \underbrace{\sum_{j\geq 0} \underbrace{\sigma(X(t_j,\omega))\chi[t_j,t_{j+1})(t) \quad \underbrace{(B_{t_{j+1}} - B_{t_j})}_{\text{absolutely continuous}} \quad (\omega)}_{Y_{n,j}\text{measurable}}}_{Y_n \text{ absolutely continuous?}}$$

While step 1. is obvious, we could not show the remaining steps. However, it sounds reasonable that at least for continuous σ, μ_t^x is not a point mass any more for $t > 0$.

Another idea was that if $Y_n \to X$ in probability, then also $Y_n^{-1}(A)$ should converge to $X^{-1}(A)$. This can be further formalized into the following Lemma:

Lemma B.1. *Let $(\Omega, \mathcal{B}, P)$ be a probability space and let Y_n be a sequence of random variables with distributions $\mu_n \ll \lambda$, i.e. absolutely continuous with respect to the Lebesgue measure. If*

a) $Y_n \to^P X$ in probability with $X \sim \mu$ and $E[|X|^2] < \infty$,

b) For all $A \in \mathcal{B}(\mathbb{R}^n)$ with $\lambda(A) = 0$ we have

$$\begin{aligned} & P\left(X^{-1}(A) \triangle X^{-1}(X(Y_n^{-1}(A)))\right) \to 0 \\ \text{and} \quad & P\left(Y_n^{-1}(A) \triangle X^{-1}(X(Y_n^{-1}(A)))\right) \to 0 \quad \text{as } n \to \infty. \end{aligned}$$

Then μ is absolutely continuous w.r.t. λ, i.e. $\mu \ll \lambda$.

Proof. For any $A \in \mathcal{B}$, we can split up $\mu(A)$ in the following fashion:

$$\mu(A) \leq \left|\mu(A) - \mu(X(Y_n^{-1}(A)))\right| + \left|\mu(X(Y_n^{-1}(A)))\right|$$

Now let $A \in \mathcal{B}$ be a set of zero Lebesgue measure $\lambda(A) = 0$. Regarding the first integral we can use the Cauchy-Schwartz inequality to calculate

$$\begin{aligned} \left|\mu(A) - \mu(X(Y_n^{-1}(A)))\right| &= \left|P \circ X^{-1}(A) - P \circ X^{-1}(X(Y_n^{-1}(A)))\right| \\ &= \left|\int_\Omega [1_{X^{-1}(A)} - 1_{X^{-1}(X(Y_n^{-1}(A)))}](\omega) P(d\omega)\right| \\ &\leq 1^{1/2} \cdot \left(\int_\Omega [1_{X^{-1}(A)} - 1_{X^{-1}(X(Y_n^{-1}(A)))}]^2(\omega) P(d\omega)\right)^{1/2} \\ &= P\left[X^{-1}(A) \triangle X^{-1}(X(Y_n^{-1}(A)))\right] \to 0 \end{aligned}$$

by assumption B.1 b as $n \to \infty$. Concerning the second term, we know that $\mu_n(A) = 0$ and, using the same argument as above,

$$\begin{aligned} \left|\mu(X(Y_n^{-1}(A)))\right| &= \left|P \circ Y_n^{-1}(A) - P \circ X^{-1}(X(Y_n^{-1}(A)))\right| \\ &\leq P\left[Y_n^{-1}(A) \triangle X^{-1}(X(Y_n^{-1}(A)))\right] \to 0 \end{aligned}$$

also by assumption B.1 b as $n \to \infty$. Hence, we get $\mu(A) = 0$ for all $A \in \mathcal{A}$ with Lebesgue measure zero. □

Appendix C

Infinitesimal Variance

The Brownian motion B_t has got a covariance of $E[B_t\, B_s] = \min\{t, s\}$ and a variance of $E[B_t^2] = t$ (c.f. [**?**, p.2]). What do we know about the variance of an Itô diffusion? A first answer to this question is given infinitesimally in [**?**, Chapter 5, Example 1.1]: Let X_t^i be the $i-th$ component of X_t and let b and σ be continuous functions. Then we have

$$\begin{aligned} &\frac{d}{dt}\, E[X_t]\big|_{t=0} = b(x) \\ \text{and} \quad &\frac{d}{dt}\, \left(E[X_t^i X_t^j] - E[X_t^i]E[X_t^j]\right)\big|_{t=0} = (\sigma\sigma^T)_{i,j}(x) \end{aligned}$$

This is why b can also be called **infinitesimal drift** and $\sigma\sigma^T$ is also called the **infinitesimal covariance**. [Goodman, 2013] develops this idea further: If $h > 0$ is a step-size, then we can write

$$\begin{aligned} E[X_{s+t} - X_s|\mathcal{F}_s] &= b(X_s)\; t + \mathcal{O}(t^2) \\ \text{and} \quad E[|X_{s+t} - X_s|^2|\mathcal{F}_s] &= (\sigma\sigma^T)(X_s)\; t + \mathcal{O}(t^2). \end{aligned}$$

In the lecture notes, these results are only given in one dimension, but similar ones should be able to derive in higher dimensions as well. They can be derived from two identities given in [**?**, p.201] below equation (47) and follow from the definition of the quadratic variation (c.f. [**?**, Section 2; summary]).

Using Itô 's formula [**?**, Section 2.10, (10.2)], it should be possible to derive similar results for the transformed process $\phi(X_t)$:[1]

$$E[|\phi(X_{s+t}) - \phi(X_s)|^2|\mathcal{F}_s] = [(D\phi)^T\, (D\phi)](X_s)\; Var[X_{s+t}|F_s] + O(t^2).$$

Unfortunately this identity was not found anywhere in the literature and a proof would exceed the scope of this thesis. Instead in the following we present the somewhat weaker result:

$$E[|\phi(X_t) - \phi(X_0)|^2|\mathcal{F}_0] \in \mathcal{O}(t).$$

[1] Great thanks to Prof. Jonathan Goodman for his excellent advice!

We start out with two lemmata necessary for the proof and finally show this equation for the exact process and for a weak first order integrator. The idea of this proof is taken partly from the proof of [Chen et al., 2015, Theorem 3].

Lemma C.1. *Let $Z \in \mathbb{R}^n$ be a real valued random variable with $E[Z^2] < \infty$. Let X_t^Z be an Itô process and let $\phi \in C_b^2(\mathbb{R}^n)$. Then,*

$$E\left[\left(E[\phi(X_t^Z)] - \phi(Z)\right)^2\right] \in \mathcal{O}(t^2)$$

Proof. Let $\omega \in \Omega$. We apply Dynkin's formula (4.4) for $\tau = t$. According to Lemma 4.7, we further have $E[|\mathcal{A}\phi(X_s^Z)|] < \infty$ for all $0 \leq t \leq s$ and therefore $\sup_{0\leq s\leq t} E[|\mathcal{A}\,\phi(X_s^{Z(\omega)})|] < \infty$. For the same reason we can apply Fubini's theorem:

$$\begin{aligned}\left|E[\phi(X_t^{Z(\omega)})] - \phi(Z(\omega))\right| &= \left|E\left[\int_0^t \mathcal{A}\,\phi(X_s^{Z(\omega)})\,ds\right]\right| \leq E\left[\int_0^t \left|\mathcal{A}\,\phi(X_s^{Z(\omega)})\right|\,ds\right] \\ &= \int_0^t E\left[\left|\mathcal{A}\,\phi(X_s^{Z(\omega)})\right|\right]\,ds \leq t \sup_{0\leq s\leq t} E\left[\left|\mathcal{A}\,\phi(X_s^{Z(\omega)})\right|\right] < \infty\end{aligned}$$

When we square this equation, we get $\left(E[\phi(X_t^Z(\omega))] - \phi(Z(\omega))\right)^2 \in \mathcal{O}(t^2)$. Since this is true for all $\omega \in \Omega$, we can take expectations and obtain the desired result. □

Lemma C.2. *Let $Z \in \mathbb{R}^n$ be a real valued random variable with $E[Z^2] < \infty$, let $\hat{X}_t^Z$ be the Euler-Maruyama approximation of an Itô process and let $\phi \in C_0^2(\mathbb{R}^n)$. Then,*

$$E\left[\left(\phi(\hat{X}_t^Z) - \phi(Z)\right)^2\right] \in \mathcal{O}(t)$$

Proof. For the Euler-Maruyama approximation (5.3), we have

$$\hat{X}_t^Z - Z = t\,b(Z) + \sqrt{t}\,\sigma(Z)\xi_{0,1}$$

where $\xi_{0,1}$ is a standard Gaussian random variable. We develop ϕ around the point Z using the multidimensional Taylor expansion (c.f. [Forster, 2013, §7, Theorem 1]):

$$\phi(\hat{X}_t^Z) - \phi(Z) = D\phi(Z)\left[t\,b(Z) + \sqrt{t}\,\sigma(Z)\xi_{0,1}\right] + \mathcal{O}(t)$$

and thus: $$\begin{aligned}E\left[\left(\phi(\hat{X}_t^Z) - \phi(Z)\right)^2\right] &\lesssim t\,\|D\phi(Z)\|^2\,E[|\sigma(Z)|^2] + t^2\,\|D\phi(Z)\|^2\,E[|b(Z)|^2] + \mathcal{O}(t^2) \\ &= \mathcal{O}(t),\end{aligned}$$

because $\|D\phi(Z)\| \leq C\|\phi\|_\infty < \infty$, because $b(x)$ and $\sigma(x)$ are bounded by $|x|$ and because $E[|Z|^2] < \infty$. □

Lemma C.3. *Let $Z \in \mathbb{R}^n$ be a real valued random variable with $E[Z^2] < \infty$. Let X_t^Z be an Itô diffusion and let $\phi \in C^2(\mathbb{R}^n)$. Then:*

$$Var\left[\phi\left(X_t^Z\right)\right] \in \mathcal{O}(t)$$

Proof. Let $\hat{X}_t^Z$ be the Euler-Maruyama approximation of the Itô diffusion X_t^Z. Then

$$\begin{aligned}
&\operatorname{Var}\left[\phi\left(X_t^Z\right)\right] = E\left[\left(E\left[\phi\left(X_t^Z\right)\right] - \phi\left(X_t^Z\right)\right)^2\right] \\
&\lesssim E\left[\left(E\left[\phi\left(X_t^Z\right)\right] - \phi\left(Z\right)\right)^2\right] + E\left[\left(\phi\left(Z\right) - \phi\left(\hat{X}_t^Z\right)\right)^2\right] + E\left[\left(\phi\left(\hat{X}_t^Z\right) - \phi\left(X_t^Z\right)\right)^2\right]
\end{aligned}$$

Now we apply the Lemmata C.2 and C.1 to yield orders of $\mathcal{O}(t^2)$ and $\mathcal{O}(t)$, respectively. For the third expectation we make the following argument: since $\phi \in C_0^2(\mathbb{R}^n)$ is Lipschitz continuous, there exists a constant $L > 0$ such that $|\phi(\hat{X}_t^Z) - \phi(X_t^Z)| \leq L|\hat{X}_t^Z - X_t^Z|$. Since the Euler-Maruyama discretization is of mean square order of accuracy $1/2$ (c.f. 5.3), we can bound the third summand by LCt. This shows the assertion is true. □

Proposition C.4. *Let $Z \in \mathbb{R}^n$ be a real valued random variable with $E[Z^2] < \infty$. Let $\hat{X}_t^Z$ a step of an approximation scheme of weak order* 1*. Then, for all functions $\phi \in C^2(\mathbb{R}^n)$:*

$$Var\left[\phi\left(\hat{X}_t^Z\right)\right] \in \mathcal{O}(t).$$

Proof. We can expand the variance as follows:

$$\begin{aligned}
\operatorname{Var}\left[\phi\left(\hat{X}_t^Z\right)\right] &= E\left[\left(E[\phi(\hat{X}_t^Z)] - \phi(\hat{X}_t^Z)\right)^2\right] \\
&\lesssim E\left[\left(E[\phi(\hat{X}_t^Z)] - E\left[\phi(X_t^Z)\right]\right)^2\right] + \operatorname{Var}\left[\phi\left(X_t^Z\right)\right] + E\left[\left(\phi(X_t^Z) - \phi(\hat{X}_t^Z)\right)^2\right].
\end{aligned}$$

Since $\hat{X}_t^Z$ is an approximation scheme of weak order 1, the first summand is $\in \mathcal{O}(t^2)$. The second summand can be bounded $\in \mathcal{O}(t)$ as we showed in Lemma C.3 above and the last term is in $\mathcal{O}(t)$ as we have seen in Lemma C.2. □

Appendix D

Implementation of MALA

The acceptance probability is chosen in such a way that it ensures that the resulting Markov process satisfies the **detailed balance** condition: Let $\alpha(Y|X) < 1$, then $\alpha(X|Y) = 1$ and

$$\alpha(Y|X)q(Y|X)\rho(X) = \frac{\rho(Y)q(X|Y)}{\rho(X)q(Y|X)}q(Y|X)\rho(X) = q(X|Y)\rho(Y) = \alpha(X|Y)q(X|Y)\rho(Y).$$

Together with η-irreducibility (c.f. [Robert and Casella, 2004, Definition 6.13]), the detailed balance condition ensures that the resulting process converges weakly in law (c.f. [Robert and Casella, 2004, Theorem 7.2 and Theorem 7.4]). If we can also show that process is aperiodic (c.f. [Robert and Casella, 2004, Definition 6.23]), then the convergence is even given in the total variation distance (c.f. [Robert and Casella, 2004, Lemma 7.6]).

Algorithm D.1 (Logarithmic MALA)**.** *Let ρ be the density of our target distribution. We define the Hamiltonian function as $H(y) := -\log \rho(y)$. By Theorem 4.16, the following Diffusion equation has got ρ as its target density:*

$$dX_t = -\nabla H(X_t)dt + \sqrt{2}dB_t = \nabla \log \rho(X_t)dt + \sqrt{2}dB_t.$$

We discretize the diffusion equation using the Euler-Maruyama scheme. Let $X \sim \eta$ be a random variable and let $\xi \sim \mathcal{N}(0,1)$ be Gaussian normal. Now define the proposal Y by

$$Y := X - h\nabla H(X) + \sqrt{2h}\xi$$

Then the acceptance probability is defined as

$$\alpha(Y|X) := \frac{\rho(Y)q(X|Y)}{\rho(X)q(Y|X)} \Rightarrow \log \alpha(Y|X) = -H(Y) + H(X) + \log \frac{q(X|Y)}{q(Y|X)}$$

Now since $Y|X \sim \mathcal{N}(X - h\nabla H(X), \sqrt{2h})$ and $X|Y \sim \mathcal{N}(Y - h\nabla H(Y), \sqrt{2h})$, we have

$$\log \frac{q(X|Y)}{q(Y|X)} = -\frac{1}{4h}\|X - Y + h\nabla H(Y)\|_2^2 + \frac{1}{4h}\|Y - X + h\nabla H(X)\|_2^2.$$

We accept the proposal if $\log U \leq \min\{0, \log \alpha(Y|X)\}$ with $U \sim U[0,1]$. Otherwise we draw another proposal. Then $Y \sim \rho$.

Appendix E

Parameter Study

Bi-modal Gaussian: LDA

Optimization for LDA algorithm. Kolmogorov-Smirnov distance after 500 iterations. Sample: 120 independent Markov chains started at 0.

a	b	γ	KS	a	b	γ	KS	a	b	γ	KS	a	b	γ	KS
10.0	5.0	0.51	0.170	5.0	1.0	1.0	0.235	0.5	5.0	0.75	0.104	0.1	1.0	1.0	0.527
10.0	5.0	0.75	0.091	5.0	0.2	0.51	0.087	0.5	5.0	1.0	0.185	0.1	0.2	0.51	0.149
10.0	5.0	1.0	0.202	5.0	0.2	0.75	0.230	0.5	1.0	0.51	0.114	0.1	0.2	0.75	0.271
10.0	1.0	0.51	0.565	5.0	0.2	1.0	0.120	0.5	1.0	0.75	0.102	0.1	0.2	1.0	0.532
10.0	1.0	0.75	0.151	1.0	5.0	0.51	0.146	0.5	1.0	1.0	0.073	0.01	5.0	0.51	0.599
10.0	1.0	1.0	0.133	1.0	5.0	0.75	0.102	0.5	0.2	0.51	0.085	0.01	5.0	0.75	0.689
10.0	0.2	0.51	0.804	**1.0**	**5.0**	**1.0**	**0.063**	0.5	0.2	0.75	0.117	0.01	5.0	1.0	0.699
10.0	0.2	0.75	0.207	1.0	1.0	0.51	0.133	0.5	0.2	1.0	0.086	0.01	1.0	0.51	0.641
10.0	0.2	1.0	0.349	1.0	1.0	0.75	0.070	0.1	5.0	0.51	0.098	0.01	1.0	0.75	0.698
5.0	5.0	0.51	0.118	1.0	1.0	1.0	0.174	0.1	5.0	0.75	0.348	0.01	1.0	1.0	0.699
5.0	5.0	0.75	0.269	1.0	0.2	0.51	0.176	0.1	5.0	1.0	0.606	0.01	0.2	0.51	0.636
5.0	5.0	1.0	0.190	1.0	0.2	0.75	0.162	0.1	1.0	0.51	0.087	0.01	0.2	0.75	0.683
5.0	1.0	0.51	0.154	1.0	0.2	1.0	0.204	0.1	1.0	0.75	0.201	0.01	0.2	1.0	0.699
5.0	1.0	0.75	0.077	0.5	5.0	0.51	0.111								

Tri-modal Gaussian: LDA

Optimization for LDA algorithm. Kolmogorov-Smirnov distance after 500 iterations. Sample: 120 independent Markov chains started at 0.

a	b	γ	KS	a	b	γ	KS	a	b	γ	KS	a	b	γ	KS	a	b	γ	KS
5.0	5.0	0.6	0.281	2.0	5.0	0.6	0.312	1.0	5.0	0.6	0.300	0.5	5.0	0.6	0.501	0.1	5.0	0.6	0.500
5.0	5.0	0.7	0.247	2.0	5.0	0.7	0.300	1.0	5.0	0.7	0.300	0.5	5.0	0.7	0.500	0.1	5.0	0.7	0.503
5.0	5.0	0.8	0.263	2.0	5.0	0.8	0.300	1.0	5.0	0.8	0.501	0.5	5.0	0.8	0.502	0.1	5.0	0.8	0.516
5.0	5.0	1.0	0.372	2.0	5.0	1.0	0.500	1.0	5.0	1.0	0.501	0.5	5.0	1.0	0.500	0.1	5.0	1.0	0.700
5.0	1.0	0.6	0.188	2.0	1.0	0.6	0.300	1.0	1.0	0.6	0.386	0.5	1.0	0.6	0.300	0.1	1.0	0.6	0.500
5.0	1.0	0.7	0.164	2.0	1.0	0.7	0.224	1.0	1.0	0.7	0.300	0.5	1.0	0.7	0.500	0.1	1.0	0.7	0.506
5.0	1.0	0.8	0.224	2.0	1.0	0.8	0.394	1.0	1.0	0.8	0.300	0.5	1.0	0.8	0.500	0.1	1.0	0.8	0.500
5.0	1.0	1.0	0.265	2.0	1.0	1.0	0.336	1.0	1.0	1.0	0.500	0.5	1.0	1.0	0.500	0.1	1.0	1.0	0.502
5.0	0.5	0.6	0.375	2.0	0.5	0.6	0.187	1.0	0.5	0.6	0.300	0.5	0.5	0.6	0.500	0.1	0.5	0.6	0.500
5.0	0.5	0.7	0.379	2.0	0.5	0.7	0.250	1.0	0.5	0.7	0.433	0.5	0.5	0.7	0.501	0.1	0.5	0.7	0.505
5.0	0.5	0.8	0.287	**2.0**	**0.5**	**0.8**	**0.141**	1.0	0.5	0.8	0.440	0.5	0.5	0.8	0.500	0.1	0.5	0.8	0.502
5.0	0.5	1.0	0.179	2.0	0.5	1.0	0.223	1.0	0.5	1.0	0.411	0.5	0.5	1.0	0.663	0.1	0.5	1.0	0.303

Tri-modal Gaussian: ULA

Optimization for ULA algorithm. Kolmogorov-Smirnov distance after 500 iterations. Sample: 120 independent Markov chains started at 0.

dt	KS	dt	KS	dt	KS
0.01	0.492	0.46	0.275	0.86	0.145
0.06	0.401	0.51	0.131	0.91	0.142
0.11	0.392	0.56	0.233	0.96	0.115
0.16	0.382	0.61	0.133	1.0	0.119
0.21	0.368	0.66	0.162	1.5	0.168
0.26	0.347	0.71	0.180	2.0	0.358
0.31	0.304	**0.76**	**0.098**	2.5	0.380
0.36	0.307	0.81	0.147	3.0	0.291
0.41	0.281	0.86	0.145	3.5	-100000.000

Bibliography

[Alt, 2006] Alt, H. W. (2006). *Lineare Funktionalanalysis*. Springer London, 5 edition.

[Andrieu et al., 2003] Andrieu, C., de Freitas, N., Doucet, A., and J.Jordan, M. (2003). An introduction to mcmc for machine learning.

[Bertsekas and Tsitsiklis, 2000] Bertsekas, D. P. and Tsitsiklis, J. N. (2000). Gradient convergence in gradient methods with errors.

[Betancourt, 2015] Betancourt, M. (2015). The fundamental incompatibility of scalable hamiltonian monte carlo and naive data subsampling. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 533–540, Lille, France. PMLR.

[Betancourt, 2017] Betancourt, M. (2017). The convergence of markov chain monte carlo methods: From the metropolis method to hamiltonian monte carlo.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[Bottou, 2012] Bottou, L. (2012). Stochastic gradient tricks. In Montavon, G., Orr, G. B., and Müller, K.-R., editors, *Neural Networks, Tricks of the Trade, Reloaded*, Lecture Notes in Computer Science (LNCS 7700), pages 430–445. Springer.

[Bussi and Parrinello, 2007] Bussi, G. and Parrinello, M. (2007). Accurate sampling using langevin dynamics.

[Butzer and Berens, 1967] Butzer, P. L. and Berens, H. (1967). *Semi-Groups of Operators and Approximation*. Springer New York.

[Chen et al., 2015] Chen, C., Ding, N., and Carin, L. (2015). On the convergence of stochastic gradient mcmc algorithms with high order integrators.

[Chen et al., 2014] Chen, T., B.Fox, E., and Guestrin, C. (2014). Stochastic gradient Hamiltonian Monte Carlo. In *Proc. International Conference on Machine Learning*.

[Chow et al., 2009] Chow, S.-N., Yang, T.-S., and Zhou, H. (2009). Global optimization by intermittent diffusion.

[Dreyfus, 1962] Dreyfus, S. (1962). The numerical solution of variational problems. *https://www.researchgate.net/publication/256244271_The_numerical_solution_of_variational_problems*.

[Duane et al., 1987] Duane, S., A.D.Kennedy, Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo.

[Durrett, 2004] Durrett, R. (2004). *Probability*. Cambridge University Press, 3rd edition.

[Dynkin, 1965] Dynkin, E. B. (1965). *Markov Processes - I*. Springer.

[Economist, 2016] Economist, T. (2016). From not working to neural networking. *https://www.economist.com/news/special-report/21700756-artificial-intelligence-boom-based-old-idea-modern-twist-not*.

[Elstrodt, 2009] Elstrodt, J. (2009). *Maß- und Integrationstheorie*. Springer-Verlag, Berlin, Heidelberg, 4. edition.

[Forster, 2013] Forster, O. (2013). *Analysis 2*. Springer, 10th edition.

[Georgii, 2009] Georgii, H. O. (2009). *Stochastic*. de Gruyter, 4th edition.

[Gibbs and Su, 2002] Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics.

[Gilbarg and Trudinger, 1998] Gilbarg, D. and Trudinger, N. S. (1998). *Elliptic Partial Differential Equations of Second Order*. Springer.

[Goodman, 2013] Goodman, J. (2013). Lecture notes on stochastic caclulus, week 7 - diffusion processes. *http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2013/resources.html*.

[Hanke-Bourgeois, 2009] Hanke-Bourgeois, M. (2009). *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Vieweg+Teubner, 3rd edition.

[Hastie et al., 2003] Hastie, T., Tibshirani, R., and Friedman, J. H. (2003). *The Elements of Statistical Learning*. Springer, corrected edition.

[Hastings, 1970] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57.

[Kelley, 1960] Kelley, H. J. (1960). Gradient theory of optimal flight paths. *https://arc.aiaa.org/doi/10.2514/8.5282*.

[Kiefer and J.Wolfowitz, 1952] Kiefer, J. and J.Wolfowitz (1952). Stochastic estimation of the maximum of a regression function.

[Kirkpatrick et al., 1983] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. 220 No. 4598.

[Klenke, 2008] Klenke, A. (2008). *Probability Theory: A Comprehensive Course*. Springer London.

[Kloeden and Platen, 1999] Kloeden, P. E. and Platen, E. (1999). *Numerical solution of stochastic differential equations*. Applications of mathematics. Springer, Berlin, New York.

[Kolmogoroff, 1931] Kolmogoroff, A. (1931). Über die analytischen methoden in der wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104(1):415–458.

[Lehmann and Casella, 1998] Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer, 2nd edition.

[LeVarge, 2003] LeVarge, S. L. (2003). Semigroups of linear operators.

[Ma et al., 2015] Ma, Y.-A., Chen, T., and Fox, E. (2015). A complete recipe for stochastic gradient mcmc. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 2917–2925. Curran Associates, Inc.

[Mattingly et al., 2010] Mattingly, J. C., Stuart, A. M., and Tretyakov, M. V. (2010). Convergence of numerical time-averaging and stationary measures via poisson equations.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 Number 6.

[Milstein and Tretyakov, 1995] Milstein, G. N. and Tretyakov, M. V. (1995). *Numerical Integration of Stochastic Differential Equations*. Springer.

[Neal, 1993] Neal, R. M. (1993). Probabilistic inference using markov chain monte carlo methods.

[Pazy, 1983] Pazy, A. (1983). *Semigroups of Linear Operators and Applications to Partial Differential Equations*, volume 44 of *Applied Mathematical Sciences*. Springer, New York.

[Pelletier, 1998] Pelletier, M. (1998). Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *The Annals of Applied Probability*, 8.

[Prato, 2004] Prato, G. D. (2004). *Kolmogorov Equations for Stochastic PDEs*. Birkhäuser Verlag.

[Prato and Zabczyk, 2003] Prato, G. D. and Zabczyk, J. (2003). *Ergodicity for Infinite Dimensional Systems*. Cambridge University Press.

[Risken, 1984] Risken, H. (1984). *The Fokker-Planck Equation*. Springer.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407.

[Robert and Casella, 2004] Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

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

[Roberts and Tweedie, 1996b] Roberts, G. O. and Tweedie, R. L. (1996b). Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83(1):95.

[Rosenblatt, 1961] Rosenblatt, F. (1961). Principles of neurodynamics: Perceptions and the theory of brain mechanism.

[Rossky et al., 1978] Rossky, P. J., Doll, J. D., and Friedman, H. L. (1978). Brownian dynamics as smart monte carlo simulation.

[Rouah, 2013] Rouah, F. D. (2013). Euler and Milstein Discretization.

[Sanz-Serna, 2014] Sanz-Serna, J. M. (2014). *Markov chain Monte Carlo and numerical differential equations.*, pages 39–88. Cham: Springer; Firenze: Fondazione CIME.

[Steele, 2000] Steele, J. M. (2000). *Stochastic Calculus and Financial Applications*. Springer.

[Teh et al., 2014] Teh, Y. W., Thiéry, A. H., and Vollmer, S. J. (2014). Consistency and fluctuations for stochastic gradient Langevin dynamics. *submitted*.

[Welling and Teh, 2011] Welling, M. and Teh, Y. W. (2011). Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning*.

[Werner, 2011] Werner, D. (2011). *Funktionalanalysis*. Springer, 7th edition.

[Øksendal, 1998] Øksendal, B. K. (1998). *Stochastic differential equations : an introduction with applications*. Springer, Berlin, Heidelberg, New York. Corrected second printing 2000.

Master's Thesis

Kolmogorov semigroups and Stochastic Gradient Monte-Carlo

Written by Roland Halbig

Supervisor: Prof. Dr. Caroline Lasser

Submission Date: 14.07.2017