Geometric convergence of slice sampling

Dissertation

for the award of the degree

“Doctor rerum naturalium” (Dr.rer.nat.)

of the Georg-August-Universität Göttingen

within the doctoral program

“Mathematical Sciences”

of the Georg-August University School of Science
(GAUSS)

submitted by

Viacheslav Natarovskii

from Moscow, Russia

Göttingen, 2021

Thesis Committee:

Jun.-Prof. Dr. Daniel Rudolf
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Axel Munk
Institut für Mathematische Stochastik, Universität Göttingen

Members of the Examination Board:

Reviewer:
Jun.-Prof. Dr. Daniel Rudolf
Institut für Mathematische Stochastik, Universität Göttingen

Second Reviewer:
Prof. Dr. Axel Munk
Institut für Mathematische Stochastik, Universität Göttingen

Further Members of the Examination Board:

Jun.-Prof. Dr. Björn Sprungk
Fakultät für Mathematik und Informatik, Technische Universität Bergakademie Freiberg

Prof. Dr. Dominic Schuhmacher
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Gerlind Plonka-Hoch
Institut für Numerische und Angewandte Mathematik, Universität Göttingen

Prof. Dr. Max Wardetzky
Institut für Numerische und Angewandte Mathematik, Universität Göttingen

Date of the oral examination: 14.10.2021

Acknowledgements

I want to thank my supervisor Dr. Daniel Rudolf for his huge support during my PhD. He successfully directed me towards my defense and inspired me during difficult times. He taught me a lot of things and definitely contributed greatly to my scientific career. Important to mention is that Daniel always found time for me and was available every time I needed his support in both professional and private concerns.

During my PhD I had the pleasure to work with the former PostDoc Björn Sprungk, who definitely accelerated and improved the work in our small team. I really appreciate that it was always so easy to just step by his office to chat. In 2020, Björn accepted a Junior Professor position at Freiberg University and left Göttingen, but I am very happy that we remained in contact and continued to work together.

I would like to thank all members of the Institute for Mathematical Stochastics including the IT Support and secretaries. It was a great pleasure to work here. I also would like to thank Jonas Lueg, Johannes Wieditz, Marco Seiler, Thomas Staudt and Gytis Kulaitis for fruitful discussions and personal support.

For the financial support I want to thank DAAD (German Academic Exchange Service) for making it possible for me to come to Germany and stay here over four years, but also the Research Training Group 2088 for supporting me with collecting my scientific experience outside of Göttingen.

Finally and also most importantly, I would like to thank my beloved wife Johanna Zastrow who was always by my side and supported me. This work would have never been done without you!

Preface

In Bayesian statistics sampling w.r.t. a posterior distribution, which is given through a prior and a likelihood function, is a challenging task. The generation of exact samples is in general quite difficult, since the posterior distribution is often known only up to a normalizing constant. A standard way to approach this problem is a Markov chain Monte Carlo (MCMC) algorithm for approximate sampling w.r.t. the target distribution. In this cumulative dissertation geometric convergence guarantees are given for two different MCMC methods: simple slice sampling and elliptical slice sampling.

First, for the simple slice sampler we show Wasserstein contraction and a lower bound of the spectral gap, depending on certain properties of the density of the target distribution. This leads to an explicit upper bound of the total variation distance between the distribution of the n-th step of the Markov chain and a limit distribution, which in particular yields quantitative geometric convergence guarantees. Furthermore, we show that our estimates cannot be improved in general.

Second, for the elliptical slice sampler under weak assumptions on the density of the target distribution we show geometric ergodicity of the corresponding Markov chain in terms of total variation distance. Moreover, we discuss limitations of our result and provide a “tail-shift” modification for scenarios where our assumptions are not satisfied.

This cumulative dissertation is based on two publications: [Natarovskii et al., 2021b] and [Natarovskii et al., 2021a], which are listed in the addenda at the end of this document. The first article, which can be found in Chapter A, addresses the aforementioned results concerning simple slice sampling, whereas the second paper (see Chapter B) provides geometric convergence guarantees for the elliptical slice sampling.

The outline of this dissertation is as following: We start with a broad overview over general slice sampling algorithms in Chapter 1. Afterwards, in Chapter 2, we present paper A, where we firstly discuss existing results in the literature concerning simple slice sampling in Section 2.1. Then we summarize the main results of paper A in Section 2.2. In Section 2.3 we discuss those results and suggest possible future research. Finally, a discussion about my own contribution to paper A is given in Section 2.4. Thereafter, in Chapter 3, we present paper B. First, we provide a literature review on elliptical slice sampling in Section 3.1 as well as a brief summary of the main results of paper B in Section 3.2. We discuss these results, suggest some possible ways for the future research in Section 3.3 and present the results concerning reversibility of elliptical slice sampling on Hilbert spaces that have not yet been published in Section 3.3.1. Finally, a statement about my own contribution to paper B is given in Section 3.4.

Contents

List of symbols

$\mathbb{N}$	Set of positive integers
$\mathbb{R}$	Set of real numbers
$\mathcal{B}(G)$	Borel σ-algebra of G
$\lvert\cdot\rvert$	Absolute value
$\|\cdot\|$	Euclidean norm
$\|\cdot\|_\infty$	L^∞ norm
$\|\mu-\nu\|_{\mathrm{tv}}$	Total variation distance between measures μ and ν
$W(\mu,\nu)$	Wasserstein distance between measures μ and ν
λ_d	d-dimensional Lebesgue measure
$\mathcal{N}(x,\Sigma)$	Gaussian distribution with mean x and covariance matrix Σ
I_d	$d \times d$ identity matrix
$\mathbb{P}(A)$	Probability of event A
$\mathbb{E}(X)$	Expected value of random variable X

CHAPTER 1

Introduction

In Bayesian statistics extracting knowledge from the posterior distribution through sampling is a common task. Generation of exact samples is usually difficult due to several facts, such as for example that the density of the posterior distribution is often known only up to a normalizing constant. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(G, \mathcal{B}(G))$ be a measurable space for some $G \subseteq \mathbb{R}^d$. Suppose in addition that $\varrho : G \to [0, \infty)$ is an unnormalized density function of the posterior (or target) distribution μ on G w.r.t. a σ-finite probability measure μ_0, which we call prior or reference measure, that is

$$\mu(A) = \frac{\int_A \varrho(x)\mu_0(\mathrm{d}x)}{\int_{\mathbb{R}^d} \varrho(x)\mu_0(\mathrm{d}x)}, \qquad A \in \mathcal{B}(G).$$

The standard approach for approximate sampling w.r.t. μ is Markov chain Monte Carlo, abbreviated as MCMC. In these methods a Markov chain is constructed such that the distribution of the n-th step of the chain converges in some sense to the target distribution μ when n goes to ∞. For measuring the error of an MCMC algorithm we use in this dissertation the total variation distance between two probability measures ν_1 and ν_2 on G, which can be defined as

$$\|\nu_1(\cdot) - \nu_2(\cdot)\|_{\mathrm{tv}} := \sup_{A \in \mathcal{B}(G)} |\nu_1(A) - \nu_2(A)|.$$

Let the MCMC method be determined by a Markov chain $(X_n)_{n\in\mathbb{N}}$ and an initial distribution ν on G, that is $X_1 \sim \nu$, then the error of the MCMC algorithm in terms of total variation distance is given by

$$\|\mathbb{P}(X_n \in \cdot \mid X_1 \sim \nu) - \mu(\cdot)\|_{\mathrm{tv}} \,,$$

where $\mathbb{P}(X_n \in \cdot \mid X_1 \sim \nu)$ denotes the distribution of the n-th step of the Markov chain with initial distribution ν.

The Metropolis-Hastings algorithm [Metropolis et al., 1953; Hastings, 1970] is the most famous transition mechanism which leads to an MCMC method. We introduce here its simple version with the symmetric Gaussian proposal, which is also often called random walk Metropolis (RWM). Suppose $\mu_0 = \lambda_d$ is the d-dimensional Lebesgue measure and let $\mathcal{N}(x, \sigma^2 I_d)$ be the symmetric d-dimensional Gaussian distribution centered in $x \in \mathbb{R}^d$ for some fixed parameter $\sigma > 0$. Then the transition of the random walk Metropolis is defined in the following way:

Algorithm 1.0.1 Random walk Metropolis

For the reference measure $\mu_0 = \lambda_d$ the transition from the current state $X_n = x$ to the next state X_{n+1} is given by:

```
draw a proposal $Y \sim \mathcal{N}(x, \sigma^2 I_d)$, call the result $y$;
calculate the acceptance probability $\alpha \leftarrow \min\left\{1, \frac{\varrho(y)}{\varrho(x)}\right\}$;
draw $U$ uniformly on $[0, 1]$, call the result $u$;
if $u \leq \alpha$ then
    accept and set $X_{n+1} \leftarrow y$;
else
    reject and set $X_{n+1} \leftarrow x$;
end if
```

It is known that under weak regularity assumptions the random walk Metropolis converges to the target distribution for any σ (see e.g. [Roberts and Tweedie, 1996]). However, in practice it is very important to pick a good step-size parameter. The intuition behind this is that if the variance of the proposal is too big, then the Markov chain will converge slowly, since it will change its position very rarely. On the other hand, if σ is too small, then almost all proposals will be accepted, but the Markov chain will then be very unlikely to leave a small neighborhood around the starting point, and therefore again will converge slowly to the target distribution. This means that every time Algorithm 1.0.1 is used, its step-size must be tuned to achieve the optimal performance. For this in practice usually one has to find a σ, such that empirical expected acceptance ratio, which is defined as the number of accepted proposals divided by the number of all proposals through one run of the Markov chain, is close to 0.234. This number comes from the paper [Gelman et al., 1997], where the authors show that if the target distribution is a high dimensional Gaussian, then the step-size parameter that is optimal in some sense corresponds to the value 0.234 of the empirical expected acceptance ratio. The first disadvantage of this approach is that the optimality of the value was shown only for the Gaussian case. But even if we suppose that for other posterior distributions the same value of the acceptance ratio should be targeted, then the tuning procedure can still be computationally very expensive, since it requires to run the algorithm multiple times to find the optimal proposal variance.

Therefore, slice sampling [Neal, 2003], which originates to auxiliary variables methods [Edwards and Sokal, 1988; Besag and Green, 1993; Damlen et al., 1999], is of great interest, since it does not usually require anything to be tuned in contrast to Metropolis-Hastings. We start here with the definition of the ideal slice sampler. For that let $G(t) := \{x \in G \mid \varrho(x) \geq t\}$ be the level set of the density ϱ for any $t \in (0, \|\varrho\|_\infty)$ and $\mu_{0,t}$ denote the reference measure restricted to the level set $G(t)$, that is

$$\mu_{0,t}(A) := \frac{\mu_0(A \cap G(t))}{\mu_0(G(t))}, \qquad A \in \mathcal{B}(G).$$

Then the ideal slice sampler can be defined in the following way:

Algorithm 1.0.2 Ideal slice sampling

For reference measure μ_0 the transition from the current state $X_n = x$ to the next one X_{n+1} is given by:

```
1: draw $T_n$ uniformly on $[0, \varrho(x)]$, call the result $t$;
2: draw $X_{n+1}$ from the distribution $\mu_{0,t}$.
```

In general the last step of the ideal slice sampling algorithm can be very difficult to implement. Therefore, several so-called hybrid slice samplers were proposed, where the second step is replaced by running one step of some Markov chain, which is implementable and whose Markov kernel $H_t(x, \cdot)$ on $G(t)$, where x is the current state and t is the chosen level, is reversible w.r.t. $\mu_{0,t}$, that is,

$$\int_A H_t(x, B)\, \mu_{0,t}(\mathrm{d}x) = \int_B H_t(x, A)\, \mu_{0,t}(\mathrm{d}x), \qquad \forall A, B \in \mathcal{B}(G), t \in (0, \|\varrho\|)_\infty.$$

Examples of these algorithms are hybrid slice sampling with the stepping-out and shrinkage procedure [Neal, 2003], elliptical slice sampling [Murray et al., 2010] and also latent slice sampling introduced in the recent work of [Li and Walker, 2020]. Algorithmically hybrid slice sampler can be described in the following way:

Algorithm 1.0.3 Hybrid slice sampling

For reference measure μ_0 the transition from the current state $X_n = x$ to the next one X_{n+1} is given by:

```
1: draw $T_n$ uniformly on $[0, \varrho(x)]$, call the result $t$;
2: draw $X_{n+1}$ from the distribution $H_t(x, \cdot)$.
```

Intuitively it is clear that the performance of the hybrid algorithm cannot be better in comparison to the ideal slice sampling since the auxiliary Markov chain for the last step of the algorithm samples w.r.t. $\mu_{0,t}$ only approximately (for more details see [Łatuszyński and Rudolf, 2014]). For this reason it is still very important to analyze the convergence of the ideal slice sampling, since this will allow us to understand the limitations of the performance of more practical hybrid slice sampling methods.

In the first half of this dissertation, and in particular in paper A, we focus on the *simple slice sampler*, which corresponds to Algorithm 1.0.2 with $\mu_0 = \lambda_d$ being a d-dimensional Lebesgue measure. More precisely, it can be described algorithmically in the following way:

Algorithm 1.0.4 Simple slice sampling

For the reference measure $\mu_0 = \lambda_d$ the transition from the current state $X_n = x$ to the next one X_{n+1} is given by:

```
1: draw $T_n$ uniformly on $[0, \varrho(x)]$, call the result $t$;
2: draw $X_{n+1}$ uniformly on the level set $G(t)$.
```

In this case $\mu_{0,t}$ becomes a uniform distribution on the level set $G(t)$, and we assume that direct sampling w.r.t. this distribution is possible for any $t > 0$. Under a boundedness condition of ϱ the uniform ergodicity of simple slice sampler was shown in [Mira and Tierney, 2002] and also qualitative and quantitative results about geometric ergodicity were proven in [Roberts and Rosenthal, 1999]. In paper A, which is summarized in Chapter 2, under weak assumptions on the function ϱ we provide an explicit lower bound of the spectral gap of simple slice sampling, which in particular leads to the quantitative geometric convergence result in terms of total variation distance. Moreover, we show Wasserstein contraction for a class of rotational invariant log-concave densities.

Another big class of slice sampling algorithms are the ones with non-Lebesgue reference measure. Examples of them are: polar slice sampler [Roberts and Rosenthal, 2002] and elliptical slice sampling [Murray et al., 2010]. The second half of this dissertation is concentrated on the latter algorithm. Elliptical slice sampler corresponds to the hybrid slice sampling algorithm for the reference measure $\mu_0 = \mathcal{N}(0, C)$ being a d-dimensional Gaussian distribution centered in 0 for some non-degenerate covariance matrix C and the Markov chain on the level set $G(t)$, which can be defined algorithmically in the following way:

Algorithm 1.0.5 Elliptical slice sampling

For the reference measure $\mu_0 = \mathcal{N}(0, C)$ the transition from the current state $X_n = x$ to the next one X_{n+1} is given by:

```
draw $T_n$ uniformly on $[0, \varrho(x)]$, call the result $t$;
draw $W \sim \mu_0 = \mathcal{N}(0, C)$, call the result $w$;
draw $\Theta$ uniformly on $[0, 2\pi]$, call the result $\theta$;
$\theta_{\min} \leftarrow \theta - 2\pi$;
$\theta_{\max} \leftarrow \theta$;
while $\varrho(\cos(\theta)x + \sin(\theta)w) < t$ do
    if $\theta < 0$ then
        $\theta_{\min} \leftarrow \theta$;
    else
        $\theta_{\max} \leftarrow \theta$;
    end if
    draw $\Theta$ uniformly on $[\theta_{\min}, \theta_{\max}]$, set the result to $\theta$;
end while
$X_{n+1} \leftarrow \cos(\theta)x + \sin(\theta)w$.
```

In [Murray et al., 2010] numerical experiments have shown good performance in several scenarios, however, to our knowledge there were no geometric convergence guarantees. In paper B, which is summarized in Chapter 3, under weak assumptions on the density ϱ we provide qualitative geometric convergence result in terms of total variation distance for elliptical slice sampling.

CHAPTER 2

Quantitative spectral gap estimate and Wasserstein contraction of simple slice sampling

Paper A is presented in this chapter. It provides a literature review on simple slice sampling in Section 2.1 as well as a brief summary of the main results of paper A in Section 2.2. We discuss those results and present an outlook for future research in Section 2.3. Finally, a statement about my own contribution to paper A is given in Section 2.4.

2.1 Literature review

In this section we provide a brief review of the literature concerning simple slice sampling, its convergence properties as well as the connection between spectral gap and Wasserstein distance, which is the main topic of paper A.

The simple slice sampler described in Chapter 1 is a Markov chain Monte Carlo algorithm for approximate sampling w.r.t. some probability distribution, which originates in the auxiliary variables methods [Edwards and Sokal, 1988; Besag and Green, 1993; Damlen et al., 1999]. Several modifications of the algorithm (which we call hybrid slice sampling in Chapter 1) were built upon the simple slice sampler, such as e.g. slice sampling with stepping-out and shrinkage procedure [Neal, 2003], hit-and-run slice sampling [Rudolf and Ullrich, 2018], slice sampling particle belief propagation [Muller et al., 2013], factor slice sampler [Tibbits et al., 2014] and even parallel multivariate slice sampling [Tibbits et al., 2011], which benefits hugely from simultaneous evaluations on different processor units. Intuitively it is clear that any hybrid slice sampler cannot perform better than the simple slice sampler, since it only simulates the ideal uniform distribution on the slice. Therefore, it is of big importance to investigate the convergence of simple slice sampling for a better understanding of the limitations of the hybrid slice sampling. Indeed, uniform convergence of simple slice sampling is shown in [Mira and Tierney, 2002] and qualitative and quantitative results about geometric convergence are provided in [Roberts and Rosenthal, 1999]. However, before we wrote paper A little was known about the spectral gap of the algorithm. In particular, apart from general implications from uniform and geometric ergodicity from [Mira and Tierney, 2002; Roberts and Rosenthal, 1999] there were no explicit estimate of the spectral gap for simple slice sampler. It is an important spectral characteristic of any algorithm, since a lower bound of the spectral gap together with reversibility leads for example to geometric ergodicity [Roberts and Rosenthal, 1997], to central limit theorem [Kipnis and Varadhan, 1986] and also to estimation of the asymptotic variance [Flegal and Jones, 2010]. Moreover, an explicit lower bound of the spectral gap, which we provide in paper A, implies an explicit upper bound of the total variation distance between the n-th step of the algorithm and the target distribution [Novak and Rudolf, 2014] and an explicit upper bound of the sample average [Rudolf, 2012]. In addition to that, as it was shown in [Łatuszyński and Rudolf, 2014], a spectral gap estimate of simple slice sampling leads to a lower and an upper bound of spectral gap of certain hybrid slice samplers. Finally, Wasserstein contraction of simple slice sampling is alone of interest, but for us, it deserves additional attention, since an explicit estimate of the Wasserstein contraction coefficient implies an explicit lower bound of the spectral gap [Ollivier, 2009].

2.2 Main results

In this section we provide necessary notation and a brief summary over the main results in paper A. For more details we refer to A/*Section 1 (Introduction)*.

In paper A geometric convergence guarantees are given for the simple slice sampler as well as an explicit lower bound of the spectral gap. First, the Wasserstein contraction of the simple slice sampling with the contraction coefficient of at most $\frac{d}{d+1}$, where d denotes the dimension, is shown for log-concave rotational invariant densities, which also implies an explicit lower bound of $\frac{1}{d+1}$ of the spectral gap. Afterwards we proved the fact that the spectral gaps for two different densities coincide, if the volumes of level sets, or level-set functions, are equal. This led to the definition of the classes Λ_k (for any $k \in \mathbb{N}$) of appropriate level-set functions, where one can use the already proven Wasserstein contraction and achieve a lower bound of $\frac{1}{k+1}$ of the spectral gap. Finally, in one example we show that the Wasserstein contraction coefficient is equal to $\frac{d}{d+1}$, which means that our estimate cannot be improved in general. To formulate the main results more formally, we need to define some notation. For convenience of the reader we use the notation from paper A here, which slightly differs from the one used in Chapter 1.

Suppose $G \subseteq \mathbb{R}^d$ and $\varrho : G \to (0, \infty)$ is an unnormalized density of the probability distribution π w.r.t. the Lebesgue measure, that is

$$\pi(A) = \frac{\int_A \varrho(x)\,\mathrm{d}x}{\int_G \varrho(x)\,\mathrm{d}x}, \qquad A \in \mathcal{B}(G).$$

Simple slice sampler, which is a Markov chain Monte Carlo algorithm for approximate sampling w.r.t. target distribution π, is defined in the following way:

Algorithm 2.2.1 Simple slice sampling

The transition from the current state $X_n = x$ to the next one X_{n+1} is given by:

```
1: draw $T_n$ uniformly on $[0, \varrho(x)]$, call the result $t$;
2: draw $X_{n+1}$ uniformly on the corresponding level set
       $$G(t) := \{x \in G \mid \varrho(x) \geq t\}.$$
```

Let U_ϱ be the transition kernel of a Markov chain generated by the simple slice sampling of a distribution π with the unnormalized density ϱ, that is

$$U_\varrho(x, A) := \mathbb{P}(X_2 \in A \mid X_1 = x), \qquad x \in G, A \in \mathcal{B}(G).$$

For the Wasserstein contraction let us define the Wasserstein distance w.r.t. the Euclidean norm $|\cdot|$ between two measures μ, ν by

$$W(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{G \times G} |x - y| \,\mathrm{d}\gamma(x, y),$$

where the infimum is taken over all couplings of μ and ν, that is, over all measures on the product space $G \times G$ with marginals μ and ν. Note that only in paper A and in this chapter the Euclidean norm is denoted by $|\cdot|$, whereas in the rest of this dissertation $\|\cdot\|$ is used.

Our first main result (A/*Theorem 2.1*) shows Wasserstein contraction for rotational invariant log-concave unnormalized densities. More formally, for some constant $R \in (0, \infty]$ and some strictly increasing and convex function $\varphi : [0, R) \to \mathbb{R}$ for the unnormalized density $\varrho : \{x \in \mathbb{R}^d : |x| \leq R\} \to (0, \infty)$, defined as

$$\varrho(x) := \exp(-\varphi(|x|)),$$

holds

$$W(U_\varrho(x, \cdot), U_\varrho(y, \cdot)) \leq \frac{d}{d+1}|x - y|, \qquad \forall x, y \in \mathbb{R}^d \text{ with } |x| \leq R, |y| \leq R. \tag{2.1}$$

It is important to notice that we allow R to be equal to ∞ here, which corresponds to the densities defined on the whole $\mathbb{R}^d$.

The second main result (A/*Theorem 3.10*) is a lower bound of the spectral gap. As it is common in the literature with the same letter U_ϱ we denote the corresponding Markov operator, that is

$$U_\varrho f(x) := \int_G f(y) U_\varrho(x, \mathrm{d}y), \qquad x \in G.$$

Then the spectral gap of the Markov operator U_ϱ is defined by

$$\mathrm{gap}_\pi(U_\varrho) := 1 - \|U_\varrho\|_{L_2^0(\pi) \to L_2^0(\pi)},$$

where $L_2^0(\pi) := \{f : G \to \mathbb{R} \mid \int_G |f|^2 \mathrm{d}\pi < \infty\}$. Directly from (2.1) by applying Theorem 1.5 from [Chen and Wang, 1994] it follows that $\mathrm{gap}_\pi(U_\varrho) \geq \frac{1}{d+1}$ under the same assumptions as in A/*Theorem 2.1*. Furthermore, we were able to enlarge the class of distributions where the spectral gap estimate can be achieved by defining an appropriate class of level-set functions, which is defined as a d-dimensional volume of the level set. More formally, for a fixed density ϱ we define a level-set function $\ell_\varrho : (0, \|\varrho\|_\infty) \to [0, \infty)$ by $\ell_\varrho(t) := \lambda_d(G(t))$. We say that a level-set function ℓ_ϱ belongs to the class Λ_k for some natural number k, if ℓ_ϱ is strictly increasing and the function $g(s) := \ell_\varrho^{-1}(s^k)$, where ℓ_ϱ^{-1} denotes the inverse function, is log-concave, that is $\log g$ is concave. The idea behind this is that the latter two requirements allow us to construct a density function on $\mathbb{R}^k$ with the same level-set function, which appears to be the crucial spectral parameter of simple slice sampling, such that all assumptions of A/*Theorem 2.1* are satisfied. Thus, for any unnormalized density ϱ with the level-set function $\ell_\varrho \in \Lambda_k$ we have that

$$\mathrm{gap}_\pi(U_\varrho) \geq \frac{1}{k+1}.$$

2.3 Discussion and outlook

In this section we discuss the results of paper A as well as possible directions of future research.

Paper A provides an explicit lower bound for the spectral gap of the corresponding Markov operator, which is in general a crucial characteristic of any algorithm, since it implies geometric ergodicity and even an explicit estimate for the upper bound of the total variation distance between the n-th step of the algorithm and the target distribution π. However, an important assumption of the simple slice sampler is that sampling a uniform distribution on an arbitrary level set is possible, which can be very challenging especially in high-dimensional settings. Therefore, other methods, such as e.g. slice sampling with stepping-out and shrinkage procedure [Neal, 2003] and hit-and-run slice sampling [Rudolf and Ullrich, 2018], that simulate the uniform distribution on the slice are usually used in practice. Our estimate also implies explicit lower and upper bounds of the spectral gap for certain hybrid slice samplers by applying the results from [Łatuszyński and Rudolf, 2014], where the authors show that the spectral gap of hybrid slice sampler is smaller than the spectral gap of the simple slice sampler, but not much smaller. The spectral gap approach can also be used for other slice sampling algorithms, such as for example the elliptical slice sampler [Murray et al., 2010]. In this particular case it would be interesting to find some equivalent of the level-set function used in paper A, such that certain properties would lead to an explicit spectral gap estimate.

2.4 Own contribution

My main contribution to paper A is proving the Wasserstein contraction for rotational invariant log-concave densities (A/*Theorem 2.1*) as well as suggesting the classes Λ_k of the level-set functions (A/*Definition 3.9*). Together with Daniel Rudolf we built a theory for proving the fact that spectral gaps of the corresponding Markov operators are equal if the level-set functions coincide (A/*Corollary 3.7*). Furthermore, with Björn Sprungk we were able to finalize the proof of the spectral gap result (A/*Theorem 3.10*) and to formulate the properties of the classes Λ_k in A/*Section 3.2.1 (Properties of the class* Λ_k*)*. I also came up with the illustrative examples and prepared corresponding pictures.

CHAPTER 3

Geometric convergence of elliptical slice sampling

In this chapter we present paper B: First, we provide a literature review on elliptical slice sampling in Section 3.1 as well as a brief summary of the main results of paper B in Section 3.2. We discuss those results, suggest some ideas for possible directions of future research in Section 3.3 and present the results concerning reversibility of elliptical slice sampling on Hilbert spaces that have not yet been published in Section 3.3.1. Finally, my own contribution to paper B is discussed in Section 3.4.

3.1 Literature review

In this section we provide a brief summary of the literature concerning elliptical slice sampling and its convergence, which we address in paper B.

The elliptical slice sampler which was described in Chapter 1 is a Markov chain Monte Carlo method for approximate sampling w.r.t. some probability distribution, which is given through an unnormalized density w.r.t. a Gaussian reference measure. The algorithm was introduced in [Murray et al., 2010] and belongs to the hybrid slice sampling family described in Chapter 1. It is based on the one hand on the preconditioned Crank-Nicolson (pCN) Metropolis [Neal, 1999; Cotter et al., 2013; Rudolf and Sprungk, 2018] and on the other hand on the shrinkage procedure firstly introduced in [Neal, 2003]. In the original paper [Murray et al., 2010] numerical experiments of elliptical slice sampler were performed on a number of applications, such as Gaussian regression, Gaussian process classification and Log Gaussian Cox process. Moreover, the authors provide arguments for the reversibility of the algorithm. Elliptical slice sampler is widely used in practice, since in contrast to Metropolis-Hastings and pCN Metropolis no tuning is required. Many other sampling algorithms are based upon the elliptical slice sampling, such as elliptical slice sampling with expectation propagation [Fagan et al., 2016], the boomerang sampler [Bierkens et al., 2020], pseudo-marginal slice sampling [Murray and Graham, 2016] and generalized elliptical slice sampling [Nishihara et al., 2014]. Therefore, convergence of elliptical slice sampling is of great interest. However, apart from reversibility arguments and good performance in numerical experiments for several specific scenarios provided in [Murray et al., 2010], there were no geometric convergence guarantees before we wrote paper B. There under weak assumptions on the density we indeed prove geometric convergence of elliptical slice sampling in terms of total variation distance between the n-th step of the algorithm and the target distribution. This was done by deriving a small set and a Lyapunov function, which led to geometric ergodicity by using standard theorems for Markov chains [Meyn and Tweedie, 2009; Hairer and Mattingly, 2011].

3.2 Main results

In this section we provide necessary notation and a brief summary over the main results in paper B. For more details we refer to B.1/*Section 1 (Introduction)*.

In Bayesian statistics the elliptical slice sampler is a Markov chain Monte Carlo method for approximate sampling of a posterior distribution with Gaussian prior introduced in [Murray et al., 2010], which is widely used especially since no tuning is required. In our paper under weak assumptions on the posterior density we show geometric convergence in terms of total variation distance. For more formal presentation of the main result we need to provide some notations.

Let $\varrho : \mathbb{R}^d \to (0,\infty)$ be the unnormalized density of the posterior distribution μ w.r.t. Gaussian prior $\mu_0 = \mathcal{N}(0, C)$ for some non-degenerate covariance matrix C, that is

$$\mu(A) = \frac{\int_A \varrho(x)\,\mu_0(\mathrm{d}x)}{\int_{\mathbb{R}^d} \varrho(x)\,\mu_0(\mathrm{d}x)}, \qquad A \in \mathcal{B}(\mathbb{R}^d).$$

The elliptical slice sampler is an MCMC method for approximate sampling w.r.t. the target distribution μ, which works as follows:

Algorithm 3.2.1 Elliptical slice sampling

The transition from the current state $X_n = x$ to the next one X_{n+1} is given by:

```
draw $W \sim \mu_0 = \mathcal{N}(0, C)$, call the result $w$;
draw $T_x$ uniformly on $[0, \varrho(x)]$, call the result $t$;
draw $\Theta$ uniformly on $[0, 2\pi]$, call the result $\theta$;
$\theta_{\min} \leftarrow \theta - 2\pi$;
$\theta_{\max} \leftarrow \theta$;
while $\varrho(\cos(\theta)x + \sin(\theta)w) < t$ do
    if $\theta < 0$ then
        $\theta_{\min} \leftarrow \theta$;
    else
        $\theta_{\max} \leftarrow \theta$;
    end if
    draw $\Theta$ uniformly on $[\theta_{\min}, \theta_{\max}]$, set the result to $\theta$;
end while
$X_{n+1} \leftarrow \cos(\theta)x + \sin(\theta)w$.
```

Under weak assumptions on the unnormalized density ϱ we were able to show geometric convergence of the elliptical slice sampling. We say that ϱ satisfies B.1/*Assumption 2.1* if

- it is bounded away from 0 and ∞ on any compact set and
- there exist constants $\alpha > 0$ and $R > 0$, such that

  $$\left\{y \in \mathbb{R}^d : \|y\| \leq \alpha \|x\|\right\} \subseteq \left\{y \in \mathbb{R}^d : \varrho(y) \geq \varrho(x)\right\}, \qquad \forall x \in \mathbb{R}^d \text{ with } \|x\| > R,$$

  where $\|\cdot\|$ denotes the Euclidean norm.

Our main result of the paper (B.1/*Theorem 2.2*) is that under B.1/*Assumption 2.1* described above there exist constants $K > 0$ and $\gamma \in (0, 1)$, such that

$$\|\mathbb{P}(X_{n+1} \in \cdot \mid X_1 = x) - \mu(\cdot)\|_{\text{tv}} \leq K(1 + \|x\|)\gamma^n, \qquad \forall n \in \mathbb{N}, \forall x \in \mathbb{R}^d, \tag{3.1}$$

where $\|\cdot\|_{\text{tv}}$ denotes a total variation distance. This provides geometric convergence of elliptical slice sampling, since the distribution of X_{n+1} converges exponentially fast to μ in terms of total variation distance. Finally, it is important to notice that we do not provide explicit estimates of constants K and γ, and therefore we see our main result as a qualitative convergence result.

Showing that $V(x) := \|x\|$ is a Lyapunov function as well as that any compact set is small led to the geometric convergence of elliptical slice sampling by applying standard theorems for Markov chains (see for example Chapter 15 in [Meyn and Tweedie, 2009] or [Hairer and Mattingly, 2011]).

3.3 Discussion and outlook

In this section we discuss the results from paper B and suggest some directions for possible future research on elliptical slice sampling. At the end, we present the results concerning reversibility of elliptical slice sampling on Hilbert spaces that have not yet been published.

In paper B we show geometric convergence of elliptical slice sampling in terms of total variation distance. However, as we see in B.1/*Section 4.3 (Volcano Density and Limitations of the Result)* our assumption is not satisfied for a volcano distribution, since the density has no tails, but numerical experiments show that elliptical slice sampler performs well in this case. This suggests that our condition is not necessary and can be improved. We suppose that the good tail behavior of the target distribution μ and not of the unnormalized density ϱ is the crucial requirement for the existence of the Lyapunov function and therefore for the geometric convergence. The second part of the theory, the small set condition, can also be possibly improved by showing that any compact set G is small not w.r.t. the measure λ_G, which is the d-dimensional Lebesgue measure restricted to G, as it was shown in B.1/*Lemma 3.4*, but w.r.t. the measure $\mu_{0,G}$, which is the reference measure μ_0 restricted to G, or w.r.t. μ_G, which is the target distribution μ restricted to G. Both improvements could lead to a less restrictive assumption on the density ϱ.

In B.2/*Section 3 ("Tail-Shift" Modification)* of the supplementary material we propose a modification where a small part of the prior is shifted to the density function, which in the case of volcano distribution and for logistic regression introduces exponential tails and allows us to apply our main theorem. Nonetheless, proving geometric convergence of the unmodified setting for scenarios where our assumption is not satisfied but numerical experiments show good performance of the elliptical slice sampling would be interesting.

As we already emphasized, big advantage of elliptical slice sampling is that no tuning is required. However, the price for this is the need of evaluating the density function multiple times within one step of the Markov chain. In our numerical experiments for volcano distribution the density function was evaluated on average 1.5 times per iteration of elliptical slice sampler, and the runtime of the program was approximately 1.5 times longer for elliptical slice sampling than for the similarly performing pCN Metropolis, which had to be tuned firstly. This seems to be a good price, but nonetheless it would be interesting to find some theoretical estimates of the average number of density evaluations in general cases.

Another fact which we experienced in the numerical experiments for the volcano distribution (and also for some other rotational invariant distributions with exponential tails which did not appear in the paper) is that the performance of the elliptical slice sampling in terms of effective sample size appears to be dimension-independent (see B.1/*Figure 2*). This is very promising, since then one might be able to apply the technique similar to the one used in [Hairer et al., 2014] and in [Rudolf and Sprungk, 2018] and extend elliptical slice sampling to infinite-dimensional Hilbert spaces. For that one would have to derive a dimension-independent upper bound of the total variation distance. Unfortunately, the constants in the bound provided in paper B are neither tractable due to limitations of the proof technique nor are they dimension-independent. Therefore, to achieve a dimension-independent upper bound of the total variation distance one could try to apply a technique similar to the one used in paper A, namely providing a lower bound of the spectral gap by showing Wasserstein contraction. The most challenging part in this approach is that one needs to come up with a suitable coupling. Unfortunately, we managed to do this only for one trivial scenario, where the target distribution is Gaussian. It is also important to deal with reversibility of the algorithm on infinite dimensional spaces. Since the arguments for reversibility are only sketched in [Murray et al., 2010], it is not clear whether they can be extended to an infinite-dimensional setting. Therefore, we finish the discussion section by suggesting a proof of reversibility of the extended algorithm on general Hilbert spaces.

3.3.1 On reversibility of elliptical slice sampling in Hilbert spaces

Checking reversibility is an important step towards verifying theoretical convergence of any algorithm. In this section under certain assumptions on the density we show reversibility of the elliptical slice sampler on a possibly infinite-dimensional Hilbert space. Before presenting the main result we define the setting.

Let $\mathcal{H}$ be a separable Hilbert space. Consider the measurable space $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$, where $\mathcal{B}(\mathcal{H})$ denotes the corresponding Borel σ-algebra. Let $\mu_0 = \mathcal{N}(0, C)$ be a Gaussian measure on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$, where $C \colon \mathcal{H} \to \mathcal{H}$ is a nonsingular covariance operator on $\mathcal{H}$. (That is, C is a linear bounded, self-adjoint and positive trace class operator with $\ker C = \{0\}$.) Let μ be a probability measure of interest on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ and assume that $\varrho \colon \mathcal{H} \to [0, \infty)$ satisfies

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(x) = \frac{\varrho(x)}{\int_{\mathcal{H}} \varrho(y) \mu_0(\mathrm{d}y)}, \qquad x \in \mathcal{H},$$

that is, ϱ is the unnormalized density of μ w.r.t. μ_0. For $t \geq 0$ let

$$\mathcal{H}(t) := \{x \in \mathcal{H} \colon \varrho(x) \geq t\}$$

be the (super-) level set of ϱ w.r.t. level t and for $x, y \in \mathcal{H}$ define the ellipse

$$E(x, y) := \{x \cos\theta + y \sin\theta \colon \theta \in [0, 2\pi)\} \subset \mathcal{H}.$$

We extend the elliptical slice sampling to Hilbert spaces by simply performing exactly the same steps as in Algorithm 3.2.1. The main result of this section is formulated in the following theorem.

Theorem 3.3.1. *Suppose that density ϱ satisfies that for any $x, y \in \mathcal{H}$ and for any $t \in [0, \|\varrho\|_\infty]$ the set $E(x, y) \cap \mathcal{H}(t)$ is a disjoint union of finitely many continuous curves on $\mathcal{H}$. Then the transition kernel of the elliptical slice sampler is reversible w.r.t. μ.*

Before we prove the main theorem we provide some technical lemmas. In the following we work with random variables usually mapping from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ either to $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ or $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Furthermore, $I \colon \mathcal{H} \to \mathcal{H}$ denotes the identity. We start with a simple technical lemma.

Lemma 3.3.2. *Let X and Y be independent random variables each distributed according to $\mu_0 = \mathcal{N}(0, C)$. For any $\theta \in [0, 2\pi)$ let $T_\theta \colon \mathcal{H} \times \mathcal{H} \to \mathcal{H} \times \mathcal{H}$ be given by*

$$T_\theta(x, y) = (x \cos\theta + y \sin\theta, x \sin\theta - y \cos\theta).$$

Then

$$\mathbb{E}(F(X, Y)) = \mathbb{E}(F(T_\theta(X, Y))) \tag{3.2}$$

for any measurable function $F \colon \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ for which one of the expectations exists.

Proof. By the fact that $X, Y \sim \mu_0 = \mathcal{N}(0, C)$ are independent, we have that the random vector $\begin{pmatrix} X \\ Y \end{pmatrix}$ on $\mathcal{H} \times \mathcal{H}$ is distributed according to $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} C & 0 \\ 0 & C \end{pmatrix}\right)$. Note that

$$T_\theta(x,y)^t = \begin{pmatrix} \cos\theta I & \sin\theta I \\ \sin\theta I & -\cos\theta I \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Thus, by the linear transformation theorem for Gaussian measures, see Proposition 1.2.3 in [Da Prato and Zabczyk, 2002], we obtain that the vector $T_\theta(X,Y)^t$ is distributed according to

$$\begin{aligned} &N\left(\begin{pmatrix} \cos\theta I & \sin\theta I \\ \sin\theta I & -\cos\theta I \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \cos\theta I & \sin\theta I \\ \sin\theta I & -\cos\theta I \end{pmatrix} \begin{pmatrix} C & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} \cos\theta I & \sin\theta I \\ \sin\theta I & -\cos\theta I \end{pmatrix}\right) \\ &\qquad = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} C & 0 \\ 0 & C \end{pmatrix}\right). \end{aligned}$$

Hence, the distributions of (X,Y) and $T_\theta(X,Y)$ coincide, such that (3.2) holds. □

By λ_1 we denote the 1-dimensional Lebesgue measure. We define a probability measure $U_{t,E(x,y)} \colon \mathcal{B}(\mathcal{H}) \to [0,1]$ which can be interpreted as a distribution on the ellipse intersected with the level set. It is given as the uniform distribution on the angle parametrized intersection of ellipse and level set, that is,

$$U_{t,E(x,y)}(A) := \frac{\lambda_1\left(\{\theta \in [0,2\pi) \colon x\cos\theta + y\sin\theta \in A \cap \mathcal{H}(t)\}\right)}{\lambda_1\left(\{\theta \in [0,2\pi) \colon x\cos\theta + y\sin\theta \in \mathcal{H}(t)\}\right)}, \quad A \in \mathcal{B}(\mathcal{H}).$$

Note that there is a one-to-one correspondence between $[0,2\pi)$ and $E(x,y)$. (This leads to the fact that $U_{t,E(x,y)}$ depends only on x,y through the ellipse $E(x,y)$.) To simplify the notation, the denominator of $U_{t,E(x,y)}$ is denoted by $c_t(x,y)$, such that we can write

$$U_{t,E(x,y)}(A) = \frac{1}{c_t(x,y)} \int_0^{2\pi} \mathbf{1}_{A\cap\mathcal{H}(t)}(x\cos\theta + y\sin\theta)\,\mathrm{d}\theta.$$

Note that $U_{t,E(x,y)}$ is not the uniform distribution on $E(x,y) \cap \mathcal{H}(t)$. Let us add here two simple observations which are useful later.

Lemma 3.3.3. *For any $x,y \in \mathcal{H}$ and any $\theta \in [0,2\pi)$ we have $E(x,y) = E(T_\theta(x,y))$. Furthermore, for any $t \geq 0$ we have $c_t(x,y) = c_t(T_\theta(x,y))$.*

Proof. Note that the function $\theta' \mapsto x\cos\theta' + y\sin\theta'$ is 2π-periodic. Using this and the angle sum identities of trigonometric functions gives

$$E(T_\theta(x,y)) = \{x\cos(\theta-\theta') + y\sin(\theta-\theta')\colon \theta' \in [0,2\pi)\} = E(x,y).$$

Again by the aforementioned periodicity we have for any $b \in \mathbb{R}$ that

$$c_t(x,y) = \int_b^{b+2\pi} \mathbf{1}_{\mathcal{H}(t)}(x\cos\theta' + y\sin\theta')\,\mathrm{d}\theta'.$$

With $b = \theta - 2\pi$ and the angle sum identities of trigonometric functions we obtain

$$\begin{aligned} c_t(T_\theta(x,y)) &= \int_0^{2\pi} \mathbf{1}_{\mathcal{H}(t)}(x\cos(\theta-\theta') + y\sin(\theta-\theta'))\,\mathrm{d}\theta' \\ &= \int_{\theta-2\pi}^{\theta} \mathbf{1}_{\mathcal{H}(t)}(x\cos\theta' + y\sin\theta')\,\mathrm{d}\theta' = c_t(x,y). \end{aligned}$$ □

Let $(H_{t,E(x,y)})_{x,y\in\mathcal{H},t>0}$ be a family of transition kernels, where $H_{t,E(x,y)}$ is a transition kernel on $E(x,y)\cap\mathcal{H}(t)$, which corresponds to the steps 3-14 of Algorithm 3.2.1. For convenience, we extend the definition of the transition kernel $H_{t,E(x,y)}$ on $(\mathcal{H},\mathcal{B}(\mathcal{H}))$. We set

$$\overline{H}_{t,E(x,y)}(z,A) = \begin{cases} 0 & z \notin E(x,y)\cap\mathcal{H}(t), \\ H_{t,E(x,y)}(z, A\cap\mathcal{H}(t)\cap E(x,y)) & z \in E(x,y)\cap\mathcal{H}(t) \end{cases}.$$

In the following we write $H_{t,E(x,y)}$ for $\overline{H}_{t,E(x,y)}$ and consider $H_{t,E(x,y)}$ as extension on $(\mathcal{H},\mathcal{B}(\mathcal{H}))$.

Now the transition kernel of the elliptical slice sampler based on $(H_{t,E(x,y)})_{x,y\in\mathcal{H},t>0}$ is given as

$$H(x,A) := \frac{1}{\varrho(x)}\int_0^{\varrho(x)}\int_{\mathcal{H}} H_{t,E(x,y)}(x,A)\mu_0(\,\mathrm{d}y)\,\mathrm{d}t, \qquad x\in\mathcal{H}, A\in\mathcal{B}(\mathcal{H}).$$

We provide a criterion for H being reversible w.r.t. μ.

Lemma 3.3.4. *Suppose that for any $x,y\in\mathcal{H}$ and $t>0$ the transition kernel $H_{t,E(x,y)}$ is reversible w.r.t. $U_{t,E(x,y)}$, that is, for all $A,B\in\mathcal{H}$ we have*

$$\int_A H_{t,E(x,y)}(z,B)U_{t,E(x,y)}(\,\mathrm{d}z) = \int_B H_{t,E(x,y)}(z,A)U_{t,E(x,y)}(\,\mathrm{d}z). \tag{3.3}$$

Then transition kernel H is reversible w.r.t. μ.

Proof. The goal is to show that for any $A, B \in \mathcal{H}$ holds

$$\int_A H(x,B)\varrho(x)\mu_0(\,\mathrm{d}x) = \int_B H(x,A)\varrho(x)\mu_0(\,\mathrm{d}x). \tag{3.4}$$

We start with the left-hand side and by using that $c_t(x,y) = \int_0^{2\pi} \mathbf{1}_{\mathcal{H}(t)}(x\cos\theta + y\sin\theta)\,\mathrm{d}\theta$ we obtain

$$\begin{aligned}
&\int_A H(x,B)\varrho(x)\mu_0(\,\mathrm{d}x) = \int_0^\infty \int_{\mathcal{H}}\int_{\mathcal{H}} \mathbf{1}_{\mathcal{H}(t)\cap A}(x)H_{t,E(x,y)}(x,B)\mu_0(\,\mathrm{d}x)\mu_0(\,\mathrm{d}y)\,\mathrm{d}t\\
&= \int_0^\infty \int_0^{2\pi}\int_{\mathcal{H}}\int_{\mathcal{H}} \frac{\mathbf{1}_{\mathcal{H}(t)}(x\cos\theta + y\sin\theta)\mathbf{1}_{\mathcal{H}(t)\cap A}(x)H_{t,E(x,y)}(x,B)}{c_t(x,y)}\mu_0(\,\mathrm{d}x)\mu_0(\,\mathrm{d}y)\,\mathrm{d}\theta\,\mathrm{d}t\\
&= \int_0^\infty \int_0^{2\pi} \mathbb{E}(F_{t,\theta}(X,Y))\,\mathrm{d}\theta\,\mathrm{d}t,
\end{aligned}$$

where X, Y are independent μ_0-distributed random variables and

$$F_{t,\theta}(x,y) = \frac{\mathbf{1}_{\mathcal{H}(t)}(x\cos\theta + y\sin\theta)\mathbf{1}_{\mathcal{H}(t)\cap A}(x)H_{t,E(x,y)}(x,B)}{c_t(x,y)}.$$

By using (3.2) and

$$F_{t,\theta}(T_\theta(x,y)) = \frac{\mathbf{1}_{\mathcal{H}(t)}(x)\mathbf{1}_{\mathcal{H}(t)\cap A}(x\cos\theta + y\sin\theta)H_{t,E(T_\theta(x,y))}(x,B)}{c_t(T_\theta(x,y))},$$

with the fact from Lemma 3.3.3 that $E(T_\theta(x,y)) = E(x,y)$ as well as $c_t(T_\theta(x,y)) = c_t(x,y)$ we have

$$\begin{aligned}
&\int_0^{2\pi} \mathbb{E}(F_{t,\theta}(T_\theta(X,Y)))\,\mathrm{d}\theta\\
&= \int_0^{2\pi}\int_{\mathcal{H}}\int_{\mathcal{H}(t)} \frac{\mathbf{1}_{\mathcal{H}(t)\cap A}(x\cos\theta + y\sin\theta)H_{t,E(x,y)}(x\cos\theta + y\sin\theta, B)}{c_t(x,y)}\mu_0(\mathrm{d}x)\mu_0(\mathrm{d}y)\,\mathrm{d}\theta\\
&= \int_{\mathcal{H}}\int_{\mathcal{H}(t)}\int_A H_{t,E(x,y)}(z,B)U_{t,E(x,y)}(\,\mathrm{d}z)\mu_0(\mathrm{d}x)\mu_0(\mathrm{d}y).
\end{aligned}$$

Altogether we obtain

$$\int_A H(x,B)\varrho(x)\mu_0(\,\mathrm{d}x) = \int_0^\infty \int_{\mathcal{H}}\int_{\mathcal{H}(t)}\int_A H_{t,E(x,y)}(z,B)U_{t,E(x,y)}(\,\mathrm{d}z)\mu_0(\mathrm{d}x)\mu_0(\mathrm{d}y)\,\mathrm{d}t \tag{3.5}$$

Hence, by (3.3) arguing backwards by the same arguments as for deriving (3.5) we obtain the reversibility identity (3.4). □

Now for some $a, b \in \mathbb{R}, a < b$ and $I \subseteq [a, b]$ we introduce an auxiliary Markov chain $(Y_n)_{n\in\mathbb{N}}$ on I with transition kernel $Q^I_{[a,b]}$, which is a generalization of the steps 3–13 of Algorithm 3.2.1, where one imitates the uniform distribution on I. One step of this Markov chain, which we call shrinkage procedure, is defined algorithmically as follows:

Algorithm 3.3.1 Shrinkage procedure

The transition from the current state $Y_n = \theta_0 \in I$ to the next state Y_{n+1} is given by:

```
draw $\Theta$ uniformly on $[\theta_0, \theta_0 + (b - a)]$, call the result $\theta$;
define an interval $[\theta_{\min}, \theta_{\max}] := [\theta - (b - a), \theta]$;
while $\theta \notin I$ and $\theta + b - a \notin I$ and $\theta - b + a \notin I$ do
    if $\theta < \theta_0$ then
        $\theta_{\min} \leftarrow \theta$;
    else
        $\theta_{\max} \leftarrow \theta$;
    end if
    draw $\Theta$ uniformly on $[\theta_{\min}, \theta_{\max}]$, set the result to $\theta$;
end while
$Y_{n+1} \leftarrow \theta + (b - a)\mathbf{1}_{\{\theta<a\}} - (b - a)\mathbf{1}_{\{\theta>b\}}$.
```

Notice that thanks to the last step the returning value of the algorithm is always in I.

We require some further notation. Namely, for $x, y \in \mathcal{H}$ and $t > 0$ let

$$I_t(x, y) := \{\theta \in [0, 2\pi) \colon x \cos\theta + y \sin\theta \in \mathcal{H}(t)\}.$$

For convenience of the reader we provide the transition mechanism of the elliptical slice sampler for given $x, y \in \mathcal{H}$ and $t > 0$ on $E(x, y) \cap \mathcal{H}(t)$, determined by a transition kernel $H_{t,E(x,y)}$ as follows:

Algorithm 3.3.2 Shrinkage procedure of elliptical slice sampling

Let $Z_0 \in E(x, y) \cap \mathcal{H}(t)$ be the current state, which in particular means that there exists a unique $\theta_0 \in [0, 2\pi)$, such that $Z_0 = x \cos\theta_0 + y \sin\theta_0$. Then, the next state $Z_1 \in E(x, y) \cap \mathcal{H}(t)$ is chosen by

```
draw $\Theta$ according to the transition kernel $Q^{I_t(x,y)}_{[0,2\pi]}(\theta_0, \cdot)$, call the result $\theta$;
$Z_1 \leftarrow x \cos\theta + y \sin\theta$.
```

Firstly, we need a following auxiliary Lemma.

Lemma 3.3.5. *Let $a, b \in \mathbb{R}$, such that $a < b$, and assume that $I = I_1 \cup \cdots \cup I_k \subseteq [a, b]$ consists of $k \in \mathbb{N}$ disjoint intervals $I_j = [a_j, b_j]$ for $j = 1, \dots, k$ with*

$$a \leq a_1 < b_1 \leq \cdots \leq a_k < b_k \leq b.$$

Then there exists a number $n \in \mathbb{N}$ as well as constants $\alpha_i \in [0, \infty)$ and sets $S_i \subseteq I$ for $i = 1, \dots, n$ such that the transition kernel $Q^I_{[a,b]}$ described in Algorithm 3.3.1 can be represented as

$$Q^I_{[a,b]}(\theta_0, A) = \sum_{i=1}^{n} \alpha_i \mathbf{1}_{S_i}(\theta_0) U_{S_i}(A), \qquad \theta_0 \in I, A \subseteq [a, b], \tag{3.6}$$

where $\mathbf{1}_{S_i}$ denotes the indicator function of S_i and

$$U_S(A) := \frac{\lambda_1(S \cap A)}{\lambda_1(S)}, \qquad A, S \subseteq [a, b].$$

Proof. We perform the proof using induction by k.

First, let $k = 1$, which means that there is only one interval $I = I_1$. Then (3.6) holds for $n = 1, \alpha_1 = 1, S_1 = I_1$.

Now suppose, that for any number $j < k$ and for any real numbers $a < b$ the statement (3.6) holds and consider the case with exactly k intervals. For further steps set $J_2, \dots, J_k$ to be the intervals between $I_1, \dots, I_k$, that is

$$J_2 := (b_1, a_2), \dots, J_k := (b_{k-1}, a_k),$$

and define $J_1 := (a, a_1) \cup (b_k, b)$ to be the union of the left and right remaining parts. Notice that J_1 may be empty. By performing one step w.r.t. transition kernel $Q^I_{[a,b]}(\theta_0, \cdot)$ we distinguish two cases: sample according to the whole set, i.e. w.r.t U_I, or according to some subsets of I, when at least one of the intervals $I_1, \dots, I_k$ is cut off within steps 3–10 of Algorithm 3.3.1.

1. Sampling w.r.t. the whole set is possible, if either I is reached right at the first trial, which happens with probability $\frac{\lambda_1(I)}{b-a}$, or if one J_l for some $l \in \{1, \dots, k\}$ is explored before reaching I, which occurs with probability $\frac{\lambda_1(J_l)}{b-a} \cdot \frac{\lambda_1(I)}{b-a-\lambda_1(J_l)}$. Thus, with probability

$$p := \frac{\lambda_1(I)}{b-a} + \sum_{l=1}^{k} \frac{\lambda_1(J_l)}{b-a} \cdot \frac{\lambda_1(I)}{b-a-\lambda_1(J_l)}$$

one samples w.r.t. U_I.

2. Now suppose that we explore J_l and J_m for some $l \in \{1, \dots, k\}$ and some $m \in \{l+1, \dots, k\}$, which means that at least one interval is cut off. This occurs with probability

$$p_{l,m} := \frac{\lambda_1(J_l)}{b-a} \cdot \frac{\lambda_1(J_m)}{b-a-\lambda_1(J_l)} + \frac{\lambda_1(J_m)}{b-a} \cdot \frac{\lambda_1(J_l)}{b-a-\lambda_1(J_m)}.$$

If $\theta_0 \in I_l \cup \dots \cup I_{m-1}$, then we cut off $[a, a_l] \cup [b_{m-1}, b]$ and proceed with the algorithm on the truncated interval $[a_l, b_{m-1}]$. Thus, for any $\theta_0 \in I_l \cup \dots \cup I_{m-1}$ with probability $p_{l,m}$ one samples w.r.t. $Q_{[a_l,b_{m-1}]}^{I_l \cup \dots \cup I_{m-1}}$.

If on the other hand $\theta_0 \in I_1 \cup \dots I_{l-1} \cup I_m \cup \dots \cup I_k$, then we cut off the interval $[b_{l-1}, a_m]$ and proceed with the algorithm on the truncated interval $[a, b - \delta_{l,m}]$, where $\delta_{l,m} := a_m - b_{l-1}$. Here the points on $[b_{l-1}, a_m]$ of the original interval are collapsed into one point b_{l-1} and points on $[a_m, b]$ are shifted to the left by $\delta_{l,m}$. Thus, for any $\theta_0 \in I_1 \cup \dots I_{l-1} \cup I_m \cup \dots \cup I_k$ with probability $p_{l,m}$ one samples w.r.t. $Q_{[a,b-\delta_{l,m}]}^{I_1 \cup \dots I_{l-1} \cup I_m - \delta_{l,m} \cup \dots \cup I_k - \delta_{l,m}}$.

Putting everything together we get

$$\begin{aligned} Q_{[a,b]}^{I}(\theta_0, A) &= p \cdot U_I(A) \\ &\quad + \sum_{l=1}^{k} \sum_{m=l+1}^{k} p_{l,m} \mathbf{1}_{I_l \cup \dots \cup I_{m-1}}(\theta_0) Q_{[a_l,b_{m-1}]}^{I_l \cup \dots \cup I_{m-1}}(\theta_0, A \cap [a_l, b_{m-1}]) \\ &\quad + \sum_{l=1}^{k} \sum_{m=l+1}^{k} p_{l,m} \mathbf{1}_{I_1 \cup \dots I_{l-1} \cup I_m \cup \dots \cup I_k}(\theta_0) Q_{[a,b-\delta_{l,m}]}^{I_1 \cup \dots I_{l-1} \cup I_m - \delta_{l,m} \cup \dots \cup I_k - \delta_{l,m}}(r_{l,m}(\theta_0), R_{l,m}(A)), \end{aligned}$$

where

$$r_{l,m}(\theta) := \begin{cases} \theta, & \theta \le b_{l-1} \\ b_{l-1}, & \theta \in (b_{l-1}, a_m] \\ \theta - \delta_{l,m}, & \theta > a_m \end{cases} \qquad \text{and} \qquad R_{l,m}(A) := \{ r_{l,m}(\theta) \mid \theta \in A \}.$$

Finally, notice that in both cases, where we cut off at least one segment, we have less than k intervals, and therefore by using the induction hypothesis we can represent both Q kernels as in (3.6). This, together with the fact that the inverse function $r_{l,m}^{-1}$ exists for $\theta_0 \in I_1 \cup \dots I_{l-1} \cup I_m \cup \dots \cup I_k$ and the inverse function $R_{l,m}^{-1}$ exists for $A \subseteq [a, b_{l-1}] \cup (a_m, b)$, clearly leads to the existence of $n \in \mathbb{N}$ as well as constants $\alpha_i \in [0, \infty)$ and sets $S_i \subseteq I$ for $i = 1, \dots, n$ such that (3.6) holds for $Q_{[a,b]}^{I}$, where I consists of exactly k intervals. □

Now we are ready to prove the statement of the main theorem:

Proof of Theorem 3.3.1. The transition kernel of the elliptical slice sampler is given by

$$H(x,B) = \frac{1}{\varrho(x)} \int_0^{\varrho(x)} \int_{\mathcal{H}} H_{t,E(x,y)}(x,B)\mu_0(\,\mathrm{d}y)\,\mathrm{d}t.$$

Therefore, it is sufficient to check (3.3). First, notice that due to the assumption of the theorem we always have a finite number of intervals in the intersection $\mathcal{H}(t) \cap E(x,y)$. Therefore, we can apply Lemma 3.3.5 and get that for any $x, y \in \mathcal{H}$ and $t > 0$ there exists a number $n_t(x,y) \in \mathbb{N}$ as well as constants $\alpha_{i,t}(x,y) \in (0,\infty)$ and sets $S_{i,t}(x,y) \in \mathcal{B}(\mathcal{H}(t) \cap E(x,y))$ for $i = 1, \dots, n$ such that the transition kernel $H_{t,E(x,y)}$ described in Algorithm 3.3.2 with $z \in \mathcal{H}(t) \cap E(x,y)$ and $A \in \mathcal{B}(\mathcal{H})$ can be represented as

$$H_{t,E(x,y)}(z,A) = \sum_{i=1}^{n} \alpha_{i,t}(x,y) \mathbf{1}_{S_{i,t}(x,y)}(z) \widetilde{U}_{S_{i,t}(x,y)}(A), \tag{3.7}$$

where

$$\widetilde{U}_{S_{i,t}(x,y)}(A) := \frac{\lambda_1(\{\theta \in [0,2\pi) \colon x\cos\theta + y\sin\theta \in S_{i,t}(x,y) \cap A\})}{\lambda_1(\{\theta \in [0,2\pi) \colon x\cos\theta + y\sin\theta \in S_{i,t}(x,y)\})}.$$

Then after setting

$$c_{i,t}(x,y) := \lambda_1(\{\theta \in [0,2\pi) \colon x\cos\theta + y\sin\theta \in S_{i,t}(x,y)\})$$

with $i = 1, \dots, n$ we have

$$\begin{aligned}
&\int_A H_{t,E(x,y)}(z,B) U_{t,E(x,y)}(\,\mathrm{d}z) \\
&= \frac{1}{c_t(x,y)} \int_0^{2\pi} \mathbf{1}_A(x\cos\theta + y\sin\theta) H_{t,E(x,y)}(x\cos\theta + y\sin\theta, B)\,\mathrm{d}\theta \\
&= \sum_{i=1}^{n} \frac{\alpha_{i,t}(x,y)}{c_t(x,y)} \int_0^{2\pi} \mathbf{1}_A(x\cos\theta + y\sin\theta) \mathbf{1}_{S_{i,t}(x,y)}(x\cos\theta + y\sin\theta) \widetilde{U}_{S_{i,t}(x,y)}(B)\,\mathrm{d}\theta \\
&= \sum_{i=1}^{n} \frac{\alpha_{i,t}(x,y)}{c_t(x,y)} c_{i,t}(x,y) \widetilde{U}_{S_{i,t}(x,y)}(A) \widetilde{U}_{S_{i,t}(x,y)}(B),
\end{aligned}$$

which is symmetric in A and B. Thus, by arguing backwards one obtains the desired equation (3.3). □

3.4 Own contribution

My main contribution to paper B is the idea of the proof of the existence of the small set in B.1/*Section 3.3 (Small Set)*, which was successfully applied together with Daniel Rudolf. Furthermore, I proved the technical B.1/*Lemma 3.3*, which shows the existence of a Lyapunov function, which together with the small set condition implied the main geometric convergence result. Moreover, I showed that the main assumption is satisfied in most examples described in B.1/*Section 4 (Illustrative Examples)* and B.2/*Section 2 (Further Example from the Exponential Family)* from the supplementary material. Finally, I wrote a Python code for numerical experiments described in B.1/*Section 4.3 (Volcano Density and Limitations of the Result)* and prepared all pictures.

Bibliography

Besag, J. and Green, P. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B*, pages 25–37.

Bierkens, J., Grazzi, S., Kamatani, K., and Roberts, G. (2020). The boomerang sampler. In *International Conference on Machine Learning*, pages 908–918. PMLR.

Chen, M. and Wang, F. (1994). Application of coupling method to the first eigenvalue on manifold. *Sci. China Ser. A*, 37(1):1–14.

Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. (2013). MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, pages 424–446.

Da Prato, G. and Zabczyk, J. (2002). *Second order partial differential equations in Hilbert spaces*, volume 293. Cambridge University Press.

Damlen, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344.

Edwards, R. and Sokal, A. (1988). Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review D*, 38(6):2009.

Fagan, F., Bhandari, J., and Cunningham, J. P. (2016). Elliptical slice sampling with expectation propagation. In *UAI*.

Flegal, J. and Jones, G. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.*, 38(2):1034–1070.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.

Hairer, M. and Mattingly, J. C. (2011). Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pages 109–117. Springer.

Hairer, M., Stuart, A. M., and Vollmer, S. J. (2014). Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24(6):2455–2490.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Kipnis, C. and Varadhan, S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19.

Łatuszyński, K. and Rudolf, D. (2014). Convergence of hybrid slice sampling via spectral gap. *arXiv preprint arXiv:1409.2709*.

Li, Y. and Walker, S. G. (2020). A Latent Slice Sampling Algorithm. *arXiv preprint arXiv:2010.08509*.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press.

Mira, A. and Tierney, L. (2002). Efficiency and convergence properties of slice samplers. *Scand. J. Statist.*, 29(1):1–12.

Muller, O., Yang, M. Y., and Rosenhahn, B. (2013). Slice sampling particle belief propagation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1129–1136.

Murray, I., Adams, R., and MacKay, D. (2010). Elliptical slice sampling. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 541–548.

Murray, I. and Graham, M. (2016). Pseudo-marginal slice sampling. In *Artificial Intelligence and Statistics*, pages 911–919.

Natarovskii, V., Rudolf, D., and Sprungk, B. (2021a). Geometric convergence of elliptical slice sampling. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7969–7978. PMLR.

Natarovskii, V., Rudolf, D., and Sprungk, B. (2021b). Quantitative spectral gap estimate and Wasserstein contraction of simple slice sampling. *The Annals of Applied Probability*, 31(2):806–825.

Neal, R. M. (1999). Regression and classification using Gaussian process priors. *J. M. Bernardo et al., editors, Bayesian Statistics*, 6:475–501.

Neal, R. M. (2003). Slice sampling. *The annals of statistics*, 31(3):705–767.

Nishihara, R., Murray, I., and Adams, R. P. (2014). Parallel MCMC with generalized elliptical slice sampling. *The Journal of Machine Learning Research*, 15(1):2087–2112.

Novak, E. and Rudolf, D. (2014). Computation of expectations by Markov chain Monte Carlo methods. In et al., S. D., editor, *Extraction of Quantifiable Information from Complex Systems*, volume 102 of *Lecture Notes in Computational Science and Engineering*, pages 397–411. Springer, Cham.

Ollivier, Y. (2009). Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.*, 256(3):810–864.

Roberts, G. and Rosenthal, J. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab.*, 2:no. 2, 13–25.

Roberts, G. and Rosenthal, J. (1999). Convergence of slice sampler Markov chains. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(3):643–660.

Roberts, G. and Rosenthal, J. (2002). The polar slice sampler. *Stoch. Models*, 18(2):257–280.

Roberts, G. and Tweedie, R. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110.

Rudolf, D. (2012). Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Mathematicae*, 485:1–93.

Rudolf, D. and Sprungk, B. (2018). On a generalization of the preconditioned Crank–Nicolson Metropolis algorithm. *Foundations of Computational Mathematics*, 18(2):309–343.

Rudolf, D. and Ullrich, M. (2018). Comparison of hit-and-run, slice sampler and random walk Metropolis. *Journal of Applied Probability*, 55(4):1186–1202.

Tibbits, M. M., Groendyke, C., Haran, M., and Liechty, J. C. (2014). Automated factor slice sampling. *Journal of Computational and Graphical Statistics*, 23(2):543–563.

Tibbits, M. M., Haran, M., and Liechty, J. C. (2011). Parallel multivariate slice sampling. *Statistics and Computing*, 21(3):415–430.

Curriculum Vitae

Viacheslav
Natarovskii

✉vnataro@uni-goettingen.de

Education

Since 2017 **PhD Student**, *Georg-August-Universität Göttingen*, (Supervisors: Jun.-Prof. Dr. Daniel Rudolf, Prof. Dr. Axel Munk).

2017 **Specialist (equivalent to M.Sc.) in Mathematics**, *Lomonosov Moscow State University*, Thesis title: *Nonsymmetric stochastic systems with synchronization*, (Supervisor: Assistant Prof. Dr. Anatoly Manita).

2011 **Secondary education**, *Lyceum "The Second School"*, Moscow, Russia.

Publications

[1] **Natarovskii, V.**, Rudolf, D. and Sprungk, B. (2021). Quantitative spectral gap estimate and Wasserstein contraction of simple slice sampling. *The Annals of Applied Probability*

[2] **Natarovskii, V.**, Rudolf, D. and Sprungk, B. (2021). Geometric convergence of elliptical slice sampling. *Proceedings of the 38th International Conference on Machine Learning*

Conference talks and posters

2021 **International Conference on Machine Learning 2021**, *online*, talk + poster.

2020 **WOPS 2020**, *Hong Kong (online)*, talk.

2020 **Bernoulli-IMS One World Symposium 2020**, *online*, talk.

2019 **15. Doktorandentreffen Stochastik**, *Darmstadt*, talk.

2018 **LMS Invited Lecture Series and CRISM Summer School in Computational Statistics 2018**, *Warwick*, poster.

Addenda

Here we provide two published articles: [Natarovskii et al., 2021b] in A and [Natarovskii et al., 2021a] in B. Notice that due to specific format restrictions of the journal the second paper is presented here in two parts: the article itself (B.1) and the supplementary material (B.2). A summary of the articles is given in the following.

- **Quantitative spectral gap estimate and Wasserstein contraction of simple slice sampling**
  Viacheslav Natarovskii, Daniel Rudolf and Björn Sprungk
  The Annals of Applied Probability
  Vol. 31, No. 2, 806–825
  2021

  Abstract: We prove Wasserstein contraction of simple slice sampling for approximate sampling w.r.t. distributions with log-concave and rotational invariant Lebesgue densities. This yields, in particular, an explicit quantitative lower bound of the spectral gap of simple slice sampling. Moreover, this lower bound carries over to more general target distributions depending only on the volume of the (super-)level sets of their unnormalized density.

- **Geometric convergence of elliptical slice sampling**
  Viacheslav Natarovskii, Daniel Rudolf and Björn Sprungk
  Proceedings of the 38th International Conference on Machine Learning
  Volume 139 of *Proceedings of Machine Learning Research*, pages 7969-7978
  2021

  Abstract: For Bayesian learning, given likelihood function and Gaussian prior, the elliptical slice sampler, introduced by Murray, Adams and MacKay 2010, provides a tool for the construction of a Markov chain for approximate sampling of the underlying posterior distribution. Besides of its wide applicability and simplicity its main feature is that no tuning is required. Under weak regularity assumptions on the posterior density we show that the corresponding Markov chain is geometrically ergodic and therefore yield qualitative convergence guarantees. We illustrate our result for Gaussian posteriors as they appear in Gaussian process regression, as well as in a setting of a multi-modal distribution. Remarkably, our numerical experiments indicate a dimension-independent performance of elliptical slice sampling even in situations where our ergodicity result does not apply.

CHAPTER A

Quantitative spectral gap estimate and Wasserstein contraction of simple slice sampling

The Annals of Applied Probability
2021, Vol. 31, No. 2, 806–825
https://doi.org/10.1214/20-AAP1605
© Institute of Mathematical Statistics, 2021

QUANTITATIVE SPECTRAL GAP ESTIMATE AND WASSERSTEIN CONTRACTION OF SIMPLE SLICE SAMPLING

BY VIACHESLAV NATAROVSKII[1,*], DANIEL RUDOLF[1,†] AND BJÖRN SPRUNGK[2]

[1] *Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, [*] vnataro@uni-goettingen.de; [†] daniel.rudolf@uni-goettingen.de*

[2] *Faculty of Mathematics and Computer Science, Technische Universität Bergakademie Freiberg, bjoern.sprungk@math.tu-freiberg.de*

We prove Wasserstein contraction of simple slice sampling for approximate sampling w.r.t. distributions with log-concave and rotational invariant Lebesgue densities. This yields, in particular, an explicit quantitative lower bound of the spectral gap of simple slice sampling. Moreover, this lower bound carries over to more general target distributions depending only on the volume of the (super-)level sets of their unnormalized density.

1. Introduction. A challenging problem in Bayesian statistics and computational science is sampling w.r.t. distributions which are only known up to a normalizing constant. Assume that $G \subseteq \mathbb{R}^d$ and $\rho : G \to (0, \infty)$ is integrable w.r.t. to the Lebesgue measure. The goal is to sample w.r.t. the distribution determined by ρ, say π, that is,

$$\pi(A) = \frac{\int_A \rho(x)\,\mathrm{d}x}{\int_G \rho(x)\,\mathrm{d}x}, \qquad A \in \mathcal{B}(G).$$

Here $\mathcal{B}(G)$ denotes the Borel σ-algebra. In most cases this can only be done approximately and the idea is to construct a (time-homogeneous) Markov chain $(X_n)_{n\in\mathbb{N}}$ which has π as limit distribution, that is, for increasing n the distribution of X_n converges to π. Slice sampling methods provide auxiliary variable Markov chains for doing this and several different versions have been proposed and investigated [2, 7, 10–12, 14, 15, 20, 21]. In particular also Metropolis–Hastings algorithms can be considered as such methods; see [7, 25]. In the underlying work we investigate simple slice sampling which works as follows:[1]

ALGORITHM 1.1. Given the current state $X_n = x \in G$ the simple slice sampling algorithm generates the next Markov chain instance X_{n+1} by the following two steps:

1. Draw T_n uniformly distributed in $[0, \rho(x)]$, call the result t.
2. Draw X_{n+1} uniformly distributed on

$$G(t) := \{x \in G \mid \rho(x) \geq t\},$$

the (super-) level set of ρ at t.

The charm of this algorithmic approach lies certainly in the empirically attestable and intuitively reasonable well-behaving convergence properties of the corresponding Markov chain. Indeed, robust convergence properties are also established theoretically. Mira and Tierney in [12] prove uniform ergodicity under boundedness conditions on G and ρ. Roberts and Rosenthal [20] provide qualitative statements about geometric ergodicity under weak assumptions

Received March 2019; revised March 2020.
MSC2020 subject classifications. Primary 65C40; secondary 60J22, 62D99, 65C05.
Key words and phrases. Slice sampling, spectral gap, Wasserstein contraction.

[1] It is straightforward to verify that π is a stationary distribution of the simple slice sampler.

as well as prove quantitative estimates of the total variation distance of the difference of the distribution of X_n and π under a condition on the initial state. However, less is known about the spectral gap. Namely, beyond the general implications [19, 22] from uniform and geometric ergodicity of the results of [12, 20] there is, to our knowledge, no explicit estimate of the spectral gap of simple slice sampling available. Let U_ρ be the transition operator/kernel of a Markov chain generated by simple slice sampling of a distribution π with (unnormalized) density ρ. The spectral gap is defined by

$$\mathrm{gap}_\pi(U_\rho) := 1 - \|U_\rho\|_{L_2^0(\pi)\to L_2^0(\pi)},$$

where $L_2^0(\pi)$ is the space of functions $f\colon G \to \mathbb{R}$ with zero mean and finite variance (i.e., $\mathbb{E}_\pi(f) := \int_G f\,\mathrm{d}\pi = 0$; $\|f\|_{2,\pi}^2 := \int_G |f|^2\,\mathrm{d}\pi < \infty$). A spectral gap, that is, $\mathrm{gap}_\pi(U_\rho) > 0$, leads to desirable robustness and convergence properties. For example, it is well known that a spectral gap implies geometric ergodicity [9, 19], and since U_ρ is reversible, it also implies a central limit theorem (CLT) for all $f \in L_2(\pi)$; see [8]. In addition to that it allows the estimation of the CLT asymptotic variance [6]. In particular, an explicit lower bound of $\mathrm{gap}_\pi(U_\rho)$ leads to quantitative estimates of the total variation distance and a mean squared error bound of Markov chain Monte Carlo. More precisely, it is well known (see, for instance, [17], Lemma 2) that

$$\|\nu U_\rho^n - \pi\|_{\mathrm{tv}} \le \bigl(1 - \mathrm{gap}_\pi(U_\rho)\bigr)^n \left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi} - 1\right\|_{2,\pi},$$

where $\|\nu - \mu\|_{\mathrm{tv}} := \sup_{A\in\mathcal{B}(G)} |\nu(A) - \mu(A)|$ denotes the total variation distance, $\nu = \mathbb{P}_{X_1}$ and $\nu U_\rho^n = \mathbb{P}_{X_{n+1}}$. Moreover, in [22] it is shown for the sample average that

$$\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^n f(X_j) - \mathbb{E}_\pi(f)\right|^2 \le \frac{2}{n\cdot \mathrm{gap}_\pi(U_\rho)} + \frac{c_p\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi} - 1\|_\infty}{n^2\cdot \mathrm{gap}_\pi(U_\rho)},$$

for any $p > 2$ and any $f\colon G \to \mathbb{R}$ with $\|f\|_p^p = \int_G |f|^p\,\mathrm{d}\pi \le 1$, where c_p is an explicit constant which depends only on p.

The crucial drawback of simple slice sampling is that the second step in the algorithm is difficult to perform, in particular, in high-dimensional scenarios. However, in [15] and the more recent papers [13, 14, 16, 26, 27] efficient slice sampling algorithms are designed, which mimic (to some extent) simple slice sampling. Already [15] constructs a number of algorithms which perform a single Markov chain step on the chosen level set instead of sampling the uniform distribution. We call those methods hybrid slice sampler. For us the motivation to study simple slice sampling is twofold:

1. There is to our knowledge no quantitative statement about the spectral gap available and for simple slice sampling one would expect particularly good dependence on the dimension which we to some extent verify.
2. In the recent work of [10] it is proven that certain hybrid slice sampler, in terms of spectral gap, are, on the one hand, worse than simple slice sampling but on the other hand not much worse. Hence knowledge of the spectral gap of simple slice sampling might carry over to estimates of the spectral gap of hybrid slice samplers, in particular to those suggested in [15].

Now let us explain the main results of the underlying work. For this let the Wasserstein distance w.r.t. the Euclidean norm $|\cdot|$ of probability measures ν, μ on $(G, \mathcal{B}(G))$ be given by

$$W(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{G\times G} |x - y|\,\mathrm{d}\gamma(x,y),$$

where $\Gamma(\mu, \nu)$ is the set of all couplings of μ and ν. The set of couplings is defined by all measures on $G \times G$ with marginals μ and ν.

First main result (Theorem 2.1): For a rotational invariant and log-concave (unnormalized) density ρ defined either on Euclidean balls or the whole $\mathbb{R}^d$ we show in Theorem 2.1 Wasserstein contraction of simple slice sampling, that is, for all $x, y \in G \subseteq \mathbb{R}^d$ we have

$$W\big(U_\rho(x, \cdot), U_\rho(y, \cdot)\big) \leq \left(1 - \frac{1}{d+1}\right)|x - y|.$$

This has a number of useful consequences. It is well known (see, for instance, [23], Section 2) that this implies

$$W\big(\nu U_\rho^n, \pi\big) \leq \left(1 - \frac{1}{d+1}\right)^n W(\nu, \pi) \tag{1}$$

for any initial distribution ν on G. In addition to that by [4], Theorem 1.5 (see also [18], Proposition 30), it implies $\text{gap}_\pi(U_\rho) \geq 1/(d+1)$. Two simple examples which satisfy the assumptions of Theorem 2.1 are given by $\rho(x) = \exp(-|x|)$ and $\rho(x) = \exp(-|x|^2/2)$ where $G = \mathbb{R}^d$. For the former one Roberts and Rosenthal in [21] argue with empirical experiments that simple slice sampling "does not mix rapidly in higher dimensions". Indeed, we observe theoretically that for increasing dimension the performance of simple slice sampling gets worse, however, we disagree to some extent to their statement, since the dependence on the dimension is moderate. Namely, from (1) we obtain for any initial distribution that for $W(\nu U_\rho^n, \pi) \leq \varepsilon$ with $\varepsilon \in (0, 1)$ we need

$$n \geq (d+1) \log\big(\varepsilon^{-1} W(\nu, \pi)\big),$$

which increases only linearly in d.

Second main result (Theorem 3.10): Based on the fact that in the second step of Algorithm 1.1 we sample w.r.t. the uniform distribution on the (super-)level set $G(t)$, one can conjecture that its geometric shape does not matter. However, its "size" or volume should matter.[2] To this end, we define the level-set function $\ell_\rho \colon (0, \infty) \to [0, \infty)$ of $\rho \colon G \to (0, \infty)$, with $G \subseteq \mathbb{R}^d$, by $\ell_\rho(t) := \lambda_d(G(t))$ for $t \in (0, \infty)$, where λ_d denotes the d-dimensional Lebesgue measure. The idea is now, to identify certain "nice" properties of ℓ_ρ which lead to spectral gap estimates. Here, we propose classes Λ_k, with $k \in \mathbb{N}$, of level-set functions containing all continuous $\ell \colon (0, \infty) \to [0, \infty)$ satisfying, that:

- ℓ is strictly decreasing on the open interval
$$\operatorname{supp} \ell := \big(0, \sup\{t \in (0, \infty) \mid \ell(t) > 0\}\big)$$
(which implies the existence of the inverse ℓ^{-1} on $(0, \|\ell\|_\infty)$ with $\|\ell\|_\infty := \sup_{s \in (0,\infty)} \ell(s)$), and
- the function $g \colon (0, \|\ell\|_\infty^{1/k}) \to \operatorname{supp} \ell$, given by $g(s) = \ell^{-1}(s^k)$ is log-concave (i.e., $\log g$ is concave).

In Theorem 3.10 we then show that, if for an unnormalized density $\rho \colon G \to (0, \infty)$ we have $\ell_\rho \in \Lambda_k$ for a $k \in \mathbb{N}$, then

$$\text{gap}_\pi(U_\rho) \geq \frac{1}{k+1}. \tag{2}$$

A crucial tool in the proof of Theorem 3.10 is the equality of the spectral gap of U_ρ and the spectral gap of the transition operator of the "level Markov chain" $(T_n)_{n \in \mathbb{N}}$ defined within

[2] This is already observed in [20, 21].

Algorithm 1.1. This statement is provided in Lemma 3.3. Observe, that in the formulation of the second main result we did not impose any uni-modality, log-concavity or rotational invariance assumption on ρ. It is allowed that the d-variate function ρ has more than one mode, the only requirement is that the corresponding level-set function belongs to Λ_k. In many cases, for $k = d$ this is satisfied, however, also $k < d$ is possible; see Example 3.15. It contains the special case where ρ is assumed to be the density of the d-variate standard normal distribution, which leads to $\ell_\rho \in \Lambda_{\lfloor d/2 \rfloor}$. In that case for large d the lower bound from (2) improves the spectral gap estimate of Theorem 2.1 roughly by a factor of 2. We also consider a d-variate "volcano density", where we show that this leads to a level-set function in Λ_1, such that the corresponding spectral gap of simple slice sampling is independent of the dimension satisfying the lower bound $1/2$.

The outline of the paper is as follows. In the next section we provide basic notation and prove our main result w.r.t. the Wasserstein contractivity. Then, in Section 3 we state and discuss the necessary operator theoretic definitions and investigate the important relation between the Markov chains $(X_n)_{n\in\mathbb{N}}$ and $(T_n)_{n\in\mathbb{N}}$ generated by the simple slice sampling algorithm. There we also prove the main theorem about the lower bound of the spectral gap and illustrate the result after a discussion about the sets Λ_k by examples.

2. Wasserstein contraction. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the common probability space on which all random variables are defined. The sequence of random variables $(X_n)_{n\in\mathbb{N}}$ determined by Algorithm 1.1 provides a Markov chain on G, that is, for all $A \in \mathcal{B}(G)$ it satisfies (almost surely)

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \dots, X_n) = U_\rho(X_n, A),$$

where the transition kernel of simple slice sampling $U_\rho \colon G \times \mathcal{B}(G) \to [0, 1]$ is given by

$$U_\rho(x, A) = \frac{1}{\rho(x)} \int_0^{\rho(x)} U_t(A)\, \mathrm{d}t.$$

Here U_t denotes the uniform distribution on the level set

$$G(t) := \{x \in \mathbb{R}^d \mid \rho(x) \geq t\},$$

thus, $U_t(A) = \frac{\lambda_d(A\cap G(t))}{\lambda_d(G(t))}$ for $t > 0$. Note that by construction the transition kernel U_ρ is reversible w.r.t. π, that is,

$$\int_B U_\rho(x, A)\pi(\mathrm{d}x) = \int_A U_\rho(x, B)\pi(\mathrm{d}x), \quad A, B \in \mathcal{B}(G).$$

In particular, this implies that π is a stationary distribution of U_ρ. Further, by $B_R^{(d)}$ we denote the d-dimensional closed Euclidean ball with radius $R > 0$ around zero and by $\mathring{B}_R^{(d)}$ its interior. For log-concave rotational invariant unnormalized densities we formulate now our Wasserstein contraction result of the simple slice sampler.

THEOREM 2.1. *For $R \in (0, \infty]$ let $\varphi \colon [0, R) \to \mathbb{R}$ be a strictly increasing and convex function on $[0, R)$. Define $\rho \colon \mathring{B}_R^{(d)} \to (0, \infty)$ by $\rho(x) := \exp(-\varphi(|x|))$. Then, for any $x, y \in \mathring{B}_R^{(d)}$ we have*

$$W\big(U_\rho(x, \cdot), U_\rho(y, \cdot)\big) \leq \left(1 - \frac{1}{d+1}\right)\big||x| - |y|\big|. \tag{3}$$

Before we prove the result let us provide some comments on it.

REMARK 2.2. Let us emphasize here that we allow $R = \infty$, which leads to $\mathring{B}_R = \mathbb{R}^d$. Moreover, we remark that since on the right-hand side of (3) we have the absolute value of the difference of the Euclidean norm of x and y an immediate consequence by the triangle inequality is

$$W\big(U_\rho(x,\cdot), U_\rho(y,\cdot)\big) \le \left(1 - \frac{1}{d+1}\right)|x-y|.$$

EXAMPLE 2.3. Let $\varphi\colon [0,\infty) \to \mathbb{R}$ be given as $\varphi(s) = s^2/2$. This gives $\rho(x) = \exp(-|x|^2/2)$ which leads to π being a multivariate standard normal density. With $R = \infty$ and the convexity of φ we obtain (3).

For the proof of Theorem 2.1 we need the following auxiliary result.

LEMMA 2.4. *With $G = \mathring{B}_R^{(d)}$ let $\rho\colon G \to (0,\infty)$ be given as in Theorem 2.1. Then, for any $x, y \in G$ we have*

$$W\big(U_\rho(x,\cdot), U_\rho(y,\cdot)\big) \le \frac{d}{d+1} \cdot \frac{1}{\lambda_d(B_1^{(d)})^{1/d}} \int_0^1 \big|\ell_\rho\big(r\rho(x)\big)^{1/d} - \ell_\rho\big(r\rho(y)\big)^{1/d}\big|\,\mathrm{d}r,$$

where $\ell_\rho\colon (0,\infty) \to [0,\infty)$ is the level-set function defined by $\ell_\rho(t) := \lambda_d(G(t))$.

PROOF. Since φ is strictly increasing and convex it is continuous and thus injective. Moreover, note that the image of φ satisfies $\varphi([0,R)) = [-\log\|\rho\|_\infty, -\log\inf\rho)$. Here $\|\rho\|_\infty := \sup_{x\in\mathring{B}_R^{(d)}} \rho(x)$ and $\inf\rho$ is an abbreviation of $\inf_{x\in\mathring{B}_R^{(d)}} \rho(x)$ with the convention $\log 0 := -\infty$. Hence, there exists the inverse

$$\varphi^{-1}\colon \big[-\log\|\rho\|_\infty, -\log\inf\rho\big) \to [0,R).$$

In the case $\inf\rho = 0$ the inverse φ^{-1} is defined on $[-\log\|\rho\|_\infty, \infty)$. In the case $\inf\rho > 0$ we extend the inverse φ^{-1} to $[-\log\|\rho\|_\infty, \infty)$ by setting

$$\varphi^{-1}(t) := \sup\{s \in [0,R)\colon \varphi(s) \le t\}, \quad t \in \big[-\log\|\rho\|_\infty, \infty\big).$$

Note that by this extension we do not change φ^{-1} in $[-\log\|\rho\|_\infty, -\log\inf\rho)$ and obtain

$$\varphi^{-1}(t) = R \quad \forall t \ge -\log\inf\rho.$$

For simplicity of the notation we write ℓ for ℓ_ρ. Observe that

$$G(t) = \{x \in \mathring{B}_R^{(d)} \mid |x| \le \varphi^{-1}(\log t^{-1})\} = B^{(d)}_{(\ell(t)/\lambda_d(B_1^{(d)}))^{1/d}}, \quad t \in \big(0, \|\rho\|_\infty\big),$$

since $\ell(t) = \lambda_d(G(t)) = \varphi^{-1}(\log t^{-1})^d \lambda_d(B_1^{(d)})$. Thus, U_t denotes the uniform distribution on the Euclidean ball around the origin with radius $(\ell(t)/\lambda_d(B_1^{(d)}))^{1/d}$. Now it is straightforward to verify that $u_{t,s}\colon \mathcal{B}(G^2) \to [0,1]$ determined by

$$u_{t,s}(A \times B) := \frac{1}{\lambda_d(B_1^{(d)})} \int_{B_1^{(d)}} \mathbf{1}_A\bigg(\bigg(\frac{\ell(t)}{\lambda_d(B_1^{(d)})}\bigg)^{1/d} z\bigg) \mathbf{1}_B\bigg(\bigg(\frac{\ell(s)}{\lambda_d(B_1^{(d)})}\bigg)^{1/d} z\bigg)\,\mathrm{d}z,$$

where $A, B \in \mathcal{B}(G)$, is a coupling of U_t and U_s. For example, we have

$$\begin{aligned} u_{t,s}(A \times G) &= \frac{1}{\lambda_d(B_1^{(d)})} \int_{B_1^{(d)}} \mathbf{1}_A\bigg(\bigg(\frac{\ell(t)}{\lambda_d(B_1^{(d)})}\bigg)^{1/d} z\bigg)\,\mathrm{d}z \\ &= \frac{1}{\ell(t)} \int_{G(t)} \mathbf{1}_A(y)\,\mathrm{d}y = U_t(A). \end{aligned}$$

Further, note that $c\colon G^2 \times \mathcal{B}(G^2) \to [0,1]$ determined by

$$c(x, y, A \times B) := \int_0^1 u_{r\rho(x), r\rho(y)}(A \times B)\,\mathrm{d}r$$

is a Markovian coupling of $U_\rho(x,\cdot)$ and $U_\rho(y,\cdot)$, that is, $c(x,y,A\times G) = U_\rho(x,A)$ and $c(x,y,G\times B) = U_\rho(y,B)$ for all $x,y \in G$ and $A, B \in \mathcal{B}(G)$. Indeed, since

$$u_{t,s}(A \times G) = U_t(A), \qquad u_{t,s}(G \times B) = U_s(B)$$

we get, for example,

$$c(x,y,A\times G) = \int_0^1 U_{r\rho(x)}(A)\,\mathrm{d}r = \frac{1}{\rho(x)} \int_0^{\rho(x)} U_t(A)\,\mathrm{d}t = U_\rho(x,A).$$

Summarized, for arbitrary $x, \widetilde{x} \in G$ and $A, B \in \mathcal{B}(G)$ we obtain

$$c(x,\widetilde{x},A\times B) = \frac{1}{\lambda_d(B_1^{(d)})} \int_0^1 \int_{B_1^{(d)}} \mathbf{1}_A\biggl(\biggl(\frac{\ell(r\rho(x))}{\lambda_d(B_1^{(d)})}\biggr)^{1/d} z\biggr) \mathbf{1}_B\biggl(\biggl(\frac{\ell(r\rho(\widetilde{x}))}{\lambda_d(B_1^{(d)})}\biggr)^{1/d} z\biggr)\,\mathrm{d}z\,\mathrm{d}r.$$

Using the Markovian coupling we obtain for arbitrary $x, \widetilde{x} \in G$ that

$$\begin{aligned} W\bigl(U_\rho(x,\cdot), U_\rho(\widetilde{x},\cdot)\bigr) &\le \int_{G^2} |y - \widetilde{y}|\, c(x,\widetilde{x}, \mathrm{d}y\,\mathrm{d}\widetilde{y}) \\ &= \frac{1}{\lambda_d(B_1^{(d)})} \int_0^1 \int_{B_1^{(d)}} \biggl| \biggl(\frac{\ell(r\rho(x))}{\lambda_d(B_1^{(d)})}\biggr)^{1/d} - \biggl(\frac{\ell(r\rho(\widetilde{x}))}{\lambda_d(B_1^{(d)})}\biggr)^{1/d} \biggr| |z|\,\mathrm{d}z\,\mathrm{d}r \\ &= \frac{\lambda_d(B_1^{(d)})}{\lambda_d(B_1^{(d)})^{1+1/d}} \cdot \frac{d}{d+1} \int_0^1 \bigl|\ell\bigl(r\rho(\widetilde{x})\bigr)^{1/d} - \ell\bigl(r\rho(x)\bigr)^{1/d}\bigr|\,\mathrm{d}r, \end{aligned}$$

which finishes the proof. □

REMARK 2.5. In the previous proof we used the coupling $u_{t,s} \in \Gamma(U_t, U_s)$ for $s, t \in (0, \|\rho\|_\infty)$. In the setting of Lemma 2.4 observe that for $d = 1$ it is related to the optimal Hoeffding–Fréchet coupling. This optimality property also holds for arbitrary $d > 1$, which is justified as follows. We derive an upper bound for $W(U_t, U_s)$ by $u_{t,s}$,

$$\begin{aligned} W(U_t, U_s) &\le \int_{G\times G} |x - y|\,\mathrm{d}u_{t,s}(x,y) \\ &= \biggl| \biggl(\frac{\ell(t)}{\lambda_d(B_1^{(d)})}\biggr)^{1/d} - \biggl(\frac{\ell(s)}{\lambda_d(B_1^{(d)})}\biggr)^{1/d} \biggr| \int_{B_1^{(d)}} |z| \frac{\mathrm{d}z}{\lambda_d(B_1^{(d)})} \\ &= \biggl| \biggl(\frac{\ell(t)}{\lambda_d(B_1^{(d)})}\biggr)^{1/d} - \biggl(\frac{\ell(s)}{\lambda_d(B_1^{(d)})}\biggr)^{1/d} \biggr| \frac{d}{d+1}, \end{aligned}$$

where we used $\int_{B_1^{(d)}} |z|\mathrm{d}z = \frac{d}{d+1}\lambda_d(B_1^{(d)})$. To derive a lower bound of $W(U_t, U_s)$ we apply the Kantorovich–Rubinstein duality formula of the Wasserstein distance (see, e.g., [29], Chapter 1.2,) w.r.t. U_t and U_s. It is given by

$$W(U_t, U_s) = \sup_{\|g\|_{\mathrm{Lip}} \le 1} \biggl| \int_G g(z)\bigl(U_t(\mathrm{d}z) - U_s(\mathrm{d}z)\bigr) \biggr|,$$

where $\|g\|_{\mathrm{Lip}} := \sup_{x,y\in G} \frac{|g(x)-g(y)|}{|x-y|}$ for $g\colon G \to \mathbb{R}$. (The supremum is taken over Lipschitz continuous functions with Lipschitz constant less or equal to 1.) Considering $h(z) := |z|$ and noting $\|h\|_{\mathrm{Lip}} \leq 1$ as well as

$$\int_G |z| U_t(\mathrm{d}z) = \left(\frac{\ell(t)}{\lambda_d(B_1^{(d)})}\right)^{1/d} \int_{B_1^{(d)}} |z| \frac{\mathrm{d}z}{\lambda_d(B_1^{(d)})} = \left(\frac{\ell(t)}{\lambda_d(B_1^{(d)})}\right)^{1/d} \frac{d}{d+1}$$

then yields

$$W(U_t, U_s) \geq \left|\int_G h(z)\big(U_t(\mathrm{d}z)\big) - U_s(\mathrm{d}z)\big)\right| = \left|\left(\frac{\ell(t)}{\lambda_d(B_1^{(d)})}\right)^{1/d} - \left(\frac{\ell(s)}{\lambda_d(B_1^{(d)})}\right)^{1/d}\right| \frac{d}{d+1}.$$

Hence

$$W(U_t, U_s) = \left|\left(\frac{\ell(t)}{\lambda_d(B_1^{(d)})}\right)^{1/d} - \left(\frac{\ell(s)}{\lambda_d(B_1^{(d)})}\right)^{1/d}\right| \frac{d}{d+1},$$

which implies that $u_{t,s}$ is an optimal coupling.

Now we provide the proof of Theorem 2.1.

PROOF OF THEOREM 2.1. Again, for ℓ_ρ we write ℓ. To verify the claim of the theorem by Lemma 2.4 it is sufficient to show that

$$\frac{1}{\lambda_d(B_1^{(d)})^{1/d}} \int_0^1 \left|\ell\big(r\rho(x)\big)^{1/d} - \ell\big(r\rho(y)\big)^{1/d}\right| \mathrm{d}r \leq \big||x| - |y|\big| \quad \forall x, y \in \mathring{B}_R^{(d)}.$$

Then, by the extended inverse φ^{-1} derived in the proof of Lemma 2.4 we have

$$\ell(t) = \lambda_d\big(B_1^{(d)}\big)\big(\varphi^{-1}(-\log t)\big)^d, \quad t \in \big(0, \|\rho\|_\infty\big]. \tag{4}$$

Here also note that by the definition of ρ we have $\varphi(0) = -\log\|\rho\|_\infty$. The representation (4) yields for any $r \in (0, 1]$ and $x \in \mathring{B}_R^{(d)}$ that

$$\ell\big(r\rho(x)\big)^{1/d} = \lambda_d\big(B_1^{(d)}\big)^{1/d} \varphi^{-1}\big(\varphi(|x|) - \log r\big),$$

which leads to

$$\begin{aligned}
&\lambda_d\big(B_1^{(d)}\big)^{-1/d} \int_0^1 \left|\ell\big(r\rho(x)\big)^{1/d} - \ell\big(r\rho(y)\big)^{1/d}\right| \mathrm{d}r \\
&\quad = \int_0^1 \left|\varphi^{-1}\big(\varphi(|x|) - \log r\big) - \varphi^{-1}\big(\varphi(|y|) - \log r\big)\right| \mathrm{d}r.
\end{aligned}$$

We now show that for any $r \in (0, 1]$ and any $s, \widetilde{s} \in [0, R)$ we have

$$\left|\varphi^{-1}\big(\varphi(s) - \log r\big) - \varphi^{-1}\big(\varphi(\widetilde{s}) - \log r\big)\right| \leq |s - \widetilde{s}|,$$

which immediately yields the assertion of the theorem.

For this let $s, \widetilde{s} \in [0, R)$ and assume without loss of generality that $s \leq \widetilde{s}$. Define for arbitrary fixed $s \in [0, R)$ the value $r_{\min}(s)$ by

$$\varphi(s) - \log r_{\min}(s) = -\log \inf \rho.$$

Hence

$$\varphi^{-1}\big(\varphi(s) - \log r\big) = R \quad \forall r \leq r_{\min}(s).$$

Moreover, we set

$$s'(r) := \varphi^{-1}\big(\varphi(s) - \log r\big) \in [0, R) \quad \forall r > r_{\min}(s)$$

and since φ is continuous and increasing we have

$$\varphi\big(s'(r)\big) = \varphi(s) - \log r \geq \varphi(s), \quad s \leq s'(r).$$

The same arguments lead to

$$\varphi^{-1}\big(\varphi(\widetilde{s}) - \log r\big) = R \quad \forall r \leq r_{\min}(\widetilde{s})$$

and

$$\varphi\big(\widetilde{s}'(r)\big) = \varphi(\widetilde{s}) - \log r \geq \varphi(\widetilde{s}), \qquad \widetilde{s} \leq \widetilde{s}'(r)$$

for

$$\widetilde{s}'(r) := \varphi^{-1}\big(\varphi(\widetilde{s}) - \log r\big) \in [0, R) \quad \forall r > r_{\min}(\widetilde{s}).$$

Note, that due to $s \leq \widetilde{s}$ we have $\varphi(s) \leq \varphi(\widetilde{s})$ and, thus, $r_{\min}(\widetilde{s}) \leq r_{\min}(s)$. We distinguish three cases w.r.t. $r \in (0, 1]$:

1. Assume $r \leq r_{\min}(\widetilde{s})$: Here $\varphi^{-1}(\varphi(s) - \log r) = \varphi^{-1}(\varphi(\widetilde{s}) - \log r) = R$ and

$$0 = \big|\varphi^{-1}\big(\varphi(s) - \log r\big) - \varphi^{-1}\big(\varphi(\widetilde{s}) - \log r\big)\big| \leq |s - \widetilde{s}|.$$

2. Assume $r > r_{\min}(s)$: Here

$$\big|\varphi^{-1}\big(\varphi(s) - \log r\big) - \varphi^{-1}\big(\varphi(\widetilde{s}) - \log r\big)\big| = \big|s'(r) - \widetilde{s}'(r)\big|$$

with $s'(r), \widetilde{s}'(r) \in [0, R)$. We now exploit the convexity of φ on $[0, R)$ which is equivalent to

$$R_\varphi(u, v) := \frac{\varphi(u) - \varphi(v)}{u - v}, \quad u, v \in [0, R),$$

being increasing in u for fixed v and vice versa (since R_φ is symmetric).

Hence, since $s \leq s'(r)$ and $\widetilde{s} \leq \widetilde{s}'(r)$, we obtain

$$\begin{aligned}\frac{\varphi(s'(r)) - \varphi(\widetilde{s}'(r))}{s'(r) - \widetilde{s}'(r)} &\geq \frac{\varphi(s) - \varphi(\widetilde{s})}{s - \widetilde{s}} \\ &= \frac{(\varphi(s) - \log r) - (\varphi(\widetilde{s}) - \log r)}{s - \widetilde{s}} = \frac{\varphi(s'(r)) - \varphi(\widetilde{s}'(r))}{s - \widetilde{s}}\end{aligned}$$

which implies

$$\big|s'(r) - \widetilde{s}'(r)\big| \leq |s - \widetilde{s}|. \tag{5}$$

3. Assume $r_{\min}(\widetilde{s}) \leq r < r_{\min}(s)$: Here[3]

$$\big|\varphi^{-1}\big(\varphi(s) - \log r\big) - \varphi^{-1}\big(\varphi(\widetilde{s}) - \log r\big)\big| = \big|\widetilde{s}'(r) - R\big|.$$

By the fact that φ is increasing and convex it is continuous, such that there exists an $\hat{s} \in [0, R)$ with $s \leq \hat{s} \leq \widetilde{s}$ satisfying

$$-\log \inf \rho = \varphi(\hat{s}) - \log r$$

and, hence, $\hat{s}'(r) = R$. By employing the same reasoning as in (5) using the convexity of φ we have that

$$\big|\widetilde{s}'(r) - R\big| \leq |\widetilde{s} - \hat{s}| \leq |s - \widetilde{s}|.$$

This finishes the proof. □

[3] This case only occurs if $\lim_{t \uparrow R} \varphi(t) = -\log \inf \rho < \infty$. In that situation define $\varphi(R) := -\log \inf \rho$ and observe that with this extension φ is increasing and convex on $[0, R]$.

It is fair to ask whether the estimate can be improved. The following example answers this question. Namely, in any dimension we find a parameterized family of unnormalized densities for which (3) holds with equality.

EXAMPLE 2.6. Let $\alpha > 0$ be an arbitrary parameter. With the notation of Theorem 2.1 set $R = \infty$ and $\varphi(s) = \alpha s$ on $[0, \infty)$. The function φ is strictly increasing and concave on $[0, \infty)$. Hence, for $\rho\colon \mathbb{R}^d \to (0, \infty)$ with $\rho(x) = \exp(-\alpha|x|)$ the estimate of (3) is true. Further observe that $G(t) = B^{(d)}_{(\log t^{-1})/\alpha}$. For $x, y \in \mathbb{R}^d$ we use again the Kantorovich–Rubinstein duality formula of the Wasserstein distance w.r.t. $U_\rho(x, \cdot)$ and $U_\rho(y, \cdot)$, that is,

$$W\big(U_\rho(x,\cdot), U_\rho(y,\cdot)\big) = \sup_{\|g\|_{\mathrm{Lip}} \le 1} \left| \int_{\mathbb{R}^d} g(z)\big(U_\rho(x, \mathrm{d}z) - U_\rho(y, \mathrm{d}z)\big) \right|, \tag{6}$$

where $\|g\|_{\mathrm{Lip}} := \sup_{x,y\in\mathbb{R}^d} \frac{|g(x)-g(y)|}{|x-y|}$ for $g\colon \mathbb{R}^d \to \mathbb{R}$. We argue as in Remark 2.5 and set $h(z) = |z|$. Note that this function satisfies $\|h\|_{\mathrm{Lip}} \le 1$ as well as

$$\begin{aligned}
\int_{\mathbb{R}^d} h(z) U_\rho(x, \mathrm{d}z) &= \frac{1}{\rho(x)} \int_0^{\rho(x)} \int_{B^{(d)}_{(\log t^{-1})/\alpha}} |z| \frac{\mathrm{d}z}{\lambda_d(B^{(d)}_{(\log t^{-1})/\alpha})} \,\mathrm{d}t \\
&= \frac{1}{\rho(x)} \int_0^{\rho(x)} \int_{B_1^{(d)}} \frac{\log t^{-1}}{\alpha} \cdot |z| \frac{\mathrm{d}z}{\lambda_d(B_1^{(d)})} \,\mathrm{d}t \\
&= \frac{d}{(d+1)\alpha} \cdot \frac{1}{\rho(x)} \int_0^{\rho(x)} \log t^{-1} \,\mathrm{d}t = \frac{d}{(d+1)\alpha}\big(-\log \rho(x) - 1\big) \\
&= \frac{d}{(d+1)\alpha}\big(\alpha|x| - 1\big),
\end{aligned}$$

where we again used the fact that $\int_{B_1^{(d)}} |z| \mathrm{d}z = \frac{d}{d+1} \lambda_d(B_1^{(d)})$. Hence, by (6), employing the function h we get a lower bound of $W(U_\rho(x,\cdot), U_\rho(y,\cdot))$, which coincides with the upper bound (3). Thus, the Markovian coupling $c(x, y, \cdot) \in \Gamma(U_\rho(x,\cdot), U_\rho(y,\cdot))$ constructed in Lemma 2.4 is in this scenario optimal and

$$W\big(U_\rho(x,\cdot), U_\rho(y,\cdot)\big) = \left(1 - \frac{1}{d+1}\right)\big||x| - |y|\big|, \quad x, y \in \mathbb{R}^d.$$

This establishes that the inequality stated in Theorem 2.1 can, in general, not be improved.

3. Spectral gap estimate. In this section we investigate spectral gap properties of the Markov operator induced by the transition kernel U_ρ of the Markov chain $(X_n)_{n\in\mathbb{N}}$. For this we need further definitions. By $L_2(\pi)$ we denote the Hilbert space of functions $f\colon G \to \mathbb{R}$ with finite norm $\|f\|_{2,\pi} := (\int_G |f|^2 \,\mathrm{d}\pi)^{1/2}$. By the reversibility of U_ρ we have that π is a stationary distribution. The transition kernel U_ρ can be extended to a linear operator $U_\rho\colon L_2(\pi) \to L_2(\pi)$ defined by

$$U_\rho f(x) := \int_G f(y) U_\rho(x, \mathrm{d}y), \quad x \in G.$$

It is well known that a general Markov operator is self-adjoint on $L_2(\pi)$ iff the corresponding transition kernel is reversible w.r.t. π; see, for example, [22], Lemma 3.9. We denote the (mean) functional $\mathbb{E}_\pi\colon L_2(\pi) \to \mathbb{R}$ by $\mathbb{E}_\pi(f) := \int_G f \,\mathrm{d}\pi$ and note that this can be extended to a bounded linear operator $\mathbb{E}_\pi\colon L_2(\pi) \to L_2(\pi)$ with $\mathbb{E}_\pi(f) \equiv \int_G f \,\mathrm{d}\pi$. With this notation the spectral gap of U_ρ is determined by the operator norm of $U_\rho - \mathbb{E}_\pi$, that is, it is given by

$$\mathrm{gap}_\pi(U_\rho) := 1 - \|U_\rho - \mathbb{E}_\pi\|_{L_2(\pi)\to L_2(\pi)}.$$

Further let $L_2^0(\pi)$ be the set of functions $f \in L_2(\pi)$ with $\mathbb{E}_\pi(f) = 0$. Using the normed linear space $L_2^0(\pi)$ it is well known that $\|U_\rho\|_{L_2^0(\pi)\to L_2^0(\pi)} = \|U_\rho - \mathbb{E}_\pi\|_{L_2(\pi)\to L_2(\pi)}$ (see, e.g., [22], Lemma 3.16) such that

$$\mathrm{gap}_\pi(U_\rho) = 1 - \|U_\rho\|_{L_2^0(\pi)\to L_2^0(\pi)}.$$

An immediate consequence of Theorem 2.1, for example, by applying [18], Proposition 30, is the following.

COROLLARY 3.1. *Assume that φ satisfies the conditions formulated in Theorem 2.1 and $\rho(x) = \exp(-\varphi(|x|))$. Then*

$$\mathrm{gap}_\pi(U_\rho) \ge \frac{1}{d+1}.$$

The aim of this section is to extend and improve the previous estimate to a larger class of density functions which are not necessarily log-concave and rotational invariant.

For this, in addition to the Markov chain $(X_n)_{n\in\mathbb{N}}$, the auxiliary variable Markov chain $(T_n)_{n\in\mathbb{N}}$ also determined by Algorithm 1.1 is useful. In the next section we introduce the corresponding transition kernel, provide a relation to U_ρ and investigate further properties of $(T_n)_{n\in\mathbb{N}}$.

3.1. *Auxiliary variable Markov chain.* The sequence of auxiliary random variables $(T_n)_{n\in\mathbb{N}}$ from Algorithm 1.1 provides also a Markov chain. In contrast to $(X_n)_{n\in\mathbb{N}}$ the Markov chain $(T_n)_{n\in\mathbb{N}}$ is defined on $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$, with $\mathbb{R}^+ := (0,\infty)$ and the transition kernel is given by

$$Q_\rho(t,B) = \frac{1}{\lambda_d(G(t))}\int_{G(t)} \frac{\lambda_1(B\cap[0,\rho(x)])}{\rho(x)}\,\mathrm{d}x, \qquad B \in \mathcal{B}(\mathbb{R}^+).$$

Recall that the level-set function of ρ is given by $\ell_\rho(t) = \lambda_d(G(t))$ and define a probability measure μ on $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ by

$$\mu(B) := \frac{\int_B \ell_\rho(t)\,\mathrm{d}t}{\int_0^\infty \ell_\rho(r)\,\mathrm{d}r}, \qquad B \in \mathcal{B}(\mathbb{R}^+).$$

From [10], Lemma 1, it follows that the transition kernel Q_ρ is reversible w.r.t. μ. For the convenience of the reader we prove this fact in our setting.

LEMMA 3.2. *The transition kernel Q_ρ on $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ is reversible w.r.t. μ.*

PROOF. For any $A, B \in \mathcal{B}(\mathbb{R}^+)$ we have

$$\begin{aligned}\int_B Q_\rho(t,A)\mu(\mathrm{d}t) &= \int_B \frac{1}{\lambda_d(G(t))}\int_{G(t)} \frac{\lambda_1(A\cap[0,\rho(x)])}{\rho(x)}\,\mathrm{d}x \frac{\ell_\rho(t)\,\mathrm{d}t}{\int_0^\infty \ell_\rho(r)\,\mathrm{d}r}\\ &= \int_0^\infty \mathbf{1}_B(t)\int_G \frac{\mathbf{1}_{G(t)}(x)}{\rho(x)}\int_0^\infty \mathbf{1}_A(s)\mathbf{1}_{[0,\rho(x)]}(s)\,\mathrm{d}s \frac{\mathrm{d}x\,\mathrm{d}t}{\int_0^\infty \lambda_d(G(r))\,\mathrm{d}r}.\end{aligned}$$

Using the fact that $\mathbf{1}_{G(s)}(x) = \mathbf{1}_{[0,\rho(x)]}(s)$ we have

$$\int_B Q_\rho(t,A)\mu(\mathrm{d}t) = \int_0^\infty \int_G \int_0^\infty \mathbf{1}_A(s)\mathbf{1}_B(t)\frac{\mathbf{1}_{G(t)}(x)\mathbf{1}_{G(s)}(x)}{\rho(x)} \frac{\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t}{\int_0^\infty \lambda_d(G(r))\,\mathrm{d}r}.$$

Note that the right-hand side of the previous equation is symmetric in A and B, such that we can change their roles and argue backwards. This leads to

$$\int_B Q_\rho(t, A)\mu(\mathrm{d}t) = \int_A Q_\rho(t, B)\mu(\mathrm{d}t),$$

which finishes the proof. □

Now we present a relation of the spectral gap of U_ρ to the spectral gap of Q_ρ. Here we need the Hilbert space $L_2(\mu)$, which consists of functions $h\colon \mathbb{R}^+ \to \mathbb{R}$ with finite $\|h\|_{2,\mu} := (\int_{\mathbb{R}^+} |h|^2 \mu(\mathrm{d}t))^{1/2}$. To state the spectral gap of Q_ρ let $\mathbb{E}_\mu\colon L_2(\mu) \to \mathbb{R}$ be the (mean) functional given by $\mathbb{E}_\mu h := \int_{\mathbb{R}^+} h \, \mathrm{d}\mu$, which we consider as linear operator mapping $L_2(\mu)$ functions to constant ones. Then, the spectral gap of Q_ρ is given by the operator norm

$$\mathrm{gap}_\mu(Q_\rho) := 1 - \|Q_\rho - \mathbb{E}_\mu\|_{L_2(\mu)\to L_2(\mu)},$$

where the transition kernel Q_ρ is extended to the self-adjoint Markov operator $Q_\rho\colon L_2(\mu) \to L_2(\mu)$ defined by

$$Q_\rho h(t) := \int_{\mathbb{R}^+} h(s) Q_\rho(t, \mathrm{d}s), \quad t \in \mathbb{R}^+.$$

Note that the self-adjointness here comes (again as for U_ρ) by the fact that Q_ρ is reversible. With this notation we obtain:

LEMMA 3.3. *The spectral gaps of Q_ρ and U_ρ coincide, that is,* $\mathrm{gap}_\pi(U_\rho) = \mathrm{gap}_\mu(Q_\rho)$.

PROOF. Define the linear operators $V\colon L_2(\mu) \to L_2(\pi)$ and $V^*\colon L_2(\pi) \to L_2(\mu)$ by

$$(Vg)(x) := \frac{1}{\rho(x)} \int_0^{\rho(x)} g(t)\,\mathrm{d}t, \quad g \in L_2(\mu),$$
$$(V^* f)(t) := \frac{1}{\lambda_d(G(t))} \int_{G(t)} f(x)\,\mathrm{d}x, \quad f \in L_2(\pi).$$

Now we show that V^* is the adjoint operator of V, that is, $\langle Vg, f\rangle_\pi = \langle g, V^* f\rangle_\mu$, where $\langle\cdot,\cdot\rangle_\pi$ and $\langle\cdot,\cdot\rangle_\mu$ are the inner products of $L_2(\pi)$ and $L_2(\mu)$, respectively. We have

$$\begin{aligned}\langle Vg, f\rangle_\pi &= \int_G (Vg)(x) f(x)\pi(\mathrm{d}x) = \int_G \frac{1}{\rho(x)} \int_0^{\rho(x)} g(t)\,\mathrm{d}t f(x) \frac{\rho(x)}{\int_G \rho(y)\,\mathrm{d}y}\,\mathrm{d}x \\ &= \int_G \int_0^\infty \mathbf{1}_{[0,\rho(x)]}(t) g(t) f(x)\,\mathrm{d}t \frac{\mathrm{d}x}{\int_G \rho(y)\,\mathrm{d}y}.\end{aligned}$$

Further we use the fact that $\mathbf{1}_{[0,\rho(x)]}(t) = \mathbf{1}_{G(t)}(x)$, that $\int_G \rho(y)\,\mathrm{d}y = \int_0^\infty \ell_\rho(r)\,\mathrm{d}r$ and change the order of the integrals. Finally, we have

$$\begin{aligned}\langle Vg, f\rangle_\pi &= \int_0^\infty g(t) \int_G f(x)\mathbf{1}_{G(t)}(x)\,\mathrm{d}x \frac{\mathrm{d}t}{\int_0^\infty \ell_\rho(r)\,\mathrm{d}r} \\ &= \int_0^\infty g(t) \frac{1}{\lambda_d(G(t))} \int_{G(t)} f(x)\,\mathrm{d}x \frac{\ell_\rho(t)\,\mathrm{d}t}{\int_0^\infty \ell_\rho(r)\,\mathrm{d}r} \\ &= \int_0^\infty g(t)(V^* f)(t)\mu(\mathrm{d}t) = \langle g, V^* f\rangle_\mu.\end{aligned}$$

Furthermore, we have $U_\rho = VV^*$ and $Q_\rho = V^*V$. Now, define $S\colon L_2(\mu) \to L_2(\pi)$ and $S^*\colon L_2(\pi) \to L_2(\mu)$ by

$$S(g) := \int_0^\infty g(t)\mu(\mathrm{d}t), \quad g \in L_2(\mu),$$
$$S^*(f) := \int_G f(x)\pi(\mathrm{d}x), \quad f \in L_2(\pi).$$

Also, note here that S^* is the adjoint operator of S, as well as, $\mathbb{E}_\pi = SS^*$ and $\mathbb{E}_\mu = S^*S$. Define $R := V - S$ and the adjoint $R^* = V^* - S^*$. By the fact that also $\mathbb{E}_\pi = SV^* = VS^*$ we have

$$RR^* = (V - S)(V^* - S^*) = VV^* - \mathbb{E}_\pi = U_\rho - \mathbb{E}_\pi.$$

Similarly, by $\mathbb{E}_\mu = V^*S = S^*V$ we obtain $R^*R = Q_\rho - \mathbb{E}_\mu$. Now using the well-known fact (see, e.g., [5], Proposition 2.7) that

$$\|R\|_{L_2(\mu)\to L_2(\pi)} = \|R^*\|_{L_2(\pi)\to L_2(\mu)}$$

the statement of the lemma follows by

$$\|RR^*\|_{L_2(\pi)\to L_2(\pi)} = \|R\|^2_{L_2(\mu)\to L_2(\pi)} = \|R^*\|^2_{L_2(\pi)\to L_2(\mu)} = \|R^*R\|_{L_2(\mu)\to L_2(\mu)}$$

and the definition of the spectral gap. □

REMARK 3.4. Similar arguments as in the previous proof have been used in [28], Section 4.2, in a finite state space setting as well as in [10, 24, 25].

Now we argue that the transition kernel Q_ρ (and therefore also the Markov operator) only depends on ρ via its level-set function ℓ_ρ.

LEMMA 3.5. *For an unnormalized density $\rho\colon G \to \mathbb{R}^+$ we have for any $t \in \mathbb{R}^+$ that*

$$Q_\rho(t, B) = \frac{1}{\ell_\rho(t)} \int_t^\infty \frac{\lambda_1(B \cap [0, r])}{r} \mathrm{d}(-\ell_\rho)(r), \quad B \in \mathcal{B}(\mathbb{R}^+),$$

where on the right-hand side we use the Lebesgue–Stieltjes integral w.r.t. $-\ell_\rho$.

PROOF. Let $g\colon (t, \ell_\rho(0)) \to \mathbb{R}^+$ with $g(r) = \lambda_1(B \cap [0, r])/r$ and note that the pushforward measure $\rho_*\lambda_d$ on $\mathbb{R}_+$ is defined by

$$\rho_*\lambda_d(B) := \lambda_d \circ \rho^{-1}(B) = \lambda_d(\rho^{-1}(B)), \quad B \in \mathcal{B}(\mathbb{R}^+).$$

Hence for any $r, s \in \mathbb{R}^+$ with $r < s$ we have

$$\begin{aligned}\rho_*\lambda_d((r, s]) &= \lambda_d(\{x \in G(t) : r < \rho(x) \le s\}) \\ &= \lambda_d(\{x \in G(t) : r < \rho(x)\}) - \lambda_d(\{x \in G(t) : s < \rho(x)\}) \\ &= -(\ell_\rho(s+) - \ell_\rho(r+)),\end{aligned}$$

where $\ell_\rho(t+)$ denotes the right limit at $t \in \mathbb{R}^+$ of the left-continuous level-set function. Thus, $\rho_*\lambda_d$ is the Lebesgue–Stieltjes measure associated to the monotone nondecreasing function

$-\ell_\rho\colon \mathbb{R}_+ \to (-\infty, 0]$ (see, e.g., [1], Section 1.3.2), and we obtain with a change of variable (see [3], Theorem 3.6.1, page 190) that

$$\begin{aligned} Q_\rho(t,B) &= \frac{1}{\ell_\rho(t)} \int_{G(t)} \frac{\lambda_1(B\cap[0,\rho(x)])}{\rho(x)}\,\mathrm{d}x \\ &= \frac{1}{\ell_\rho(t)} \int_{G(t)} g\big(\rho(x)\big)\lambda_d(\mathrm{d}x) \\ &= \frac{1}{\ell_\rho(t)} \int_t^{\ell_\rho(0)} g(r)\rho_*\lambda_d(\mathrm{d}r) \\ &= \frac{1}{\ell_\rho(t)} \int_t^{\infty} \frac{\lambda_1(B\cap[0,r])}{r}\mathrm{d}(-\ell_\rho)(r). \end{aligned}$$ □

REMARK 3.6. For a given $\rho\colon G\to\mathbb{R}^+$ with continuously differentiable level-set function ℓ_ρ the previous result can be stated as

$$Q_\rho(t,B) = -\frac{1}{\ell_\rho(t)} \int_t^{\infty} \frac{\lambda_1(B\cap[0,r])}{r}\ell'_\rho(r)\,\mathrm{d}r, \qquad B\in\mathcal{B}(\mathbb{R}^+).$$

An immediate consequence of Lemma 3.3 and Lemma 3.5 is the following important result.

COROLLARY 3.7. *Let $d,\widetilde{d}\in\mathbb{N}$ and $G\subseteq\mathbb{R}^d$ as well as $\widetilde{G}\subseteq\mathbb{R}^{\widetilde{d}}$. Further let $\rho\colon G\to\mathbb{R}^+$ and $\widetilde{\rho}\colon\widetilde{G}\to\mathbb{R}^+$ satisfying $\ell_\rho(t)=\ell_{\widetilde{\rho}}(t)$ for all $t\in\mathbb{R}^+$. Then*

$$Q_\rho(t,B) = Q_{\widetilde{\rho}}(t,B), \qquad t\in\mathbb{R}^+, B\in\mathcal{B}(\mathbb{R}^+),$$

and

$$\mathrm{gap}_\pi(U_\rho) = \mathrm{gap}_\mu(Q_\rho) = \mathrm{gap}_\mu(Q_{\widetilde{\rho}}) = \mathrm{gap}_{\widetilde{\pi}}(U_{\widetilde{\rho}}),$$

where $\widetilde{\pi}$ denotes the distribution induced by $\widetilde{\rho}$.

Thus, the above corollary tells us that the spectral gap of simple slice sampling is entirely determined by the level-set function $\ell_\rho\colon\mathbb{R}^+\to[0,\infty)$ of the (unnormalized) target density ρ and does, for instance, not necessarily depend on the dimension of G. In particular, Corollary 3.7 allows us to extend the spectral gap result of Corollary 3.1 to much larger classes of target distributions as we explain in detail in the next subsection.

3.2. *Spectral gap result.* Corollary 3.7 implies that the lower bound for the spectral gap of simple slice sampling of rotational invariant and log-concave (unnormalized) target densities also holds for other target densities which share the same level-set function. Thus, our idea is to identify convenient classes of target densities $\rho\colon G\to[0,\infty)$, with $G\subseteq\mathbb{R}^d$, which possess the same level-set function as a rotational invariant and log-concave unnormalized density $\widetilde{\rho}\colon\widetilde{G}\to[0,\infty)$, with $\widetilde{G}\subseteq\mathbb{R}^{\widetilde{d}}$. We illustrate this approach first by an example and formalize it rigorously afterwards.

EXAMPLE 3.8. We consider a bimodal distribution π on the set

$$G = \big(m_0 + \mathring{B}^{(d)}_{\sqrt{\log 16}}\big) \cup \mathring{B}^{(d)}_{\sqrt{\log 4}} \subset \mathbb{R}^d$$

with $m_0=(5,0,\dots,0)\in\mathbb{R}^d$ given by the unnormalized density

$$\rho(x) = \max\left\{\exp\left(-\frac{1}{2}|x|^2\right), \exp\left(-\frac{1}{4}|x-m_0|^2\right)\right\} - \frac{1}{2}.$$

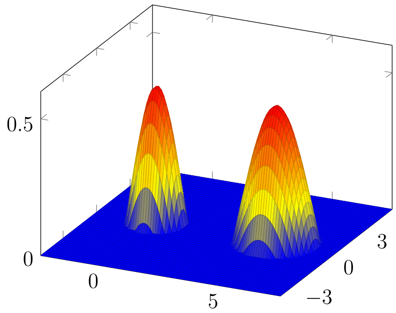

FIG. 1. *Plot of ρ from Example 3.8 for $d = 2$.*

Notice that ρ is positive on G. Here it is worth mentioning that in particular in such scenarios an efficient implementation of simple slice sampling is challenging and we are at this point merely interested in theoretical properties. By construction, the level sets of ρ consist of two disjoint balls, that is, we have

$$G(t) = \big(m_0 + \mathring{B}^{(d)}_{\sqrt{\log(1/2+t)^{-4}}}\big) \cup \mathring{B}^{(d)}_{\sqrt{\log(1/2+t)^{-2}}}, \quad t \in [0, 1/2).$$

This leads to

$$\ell_\rho(t) = (2^{d/2} + 4^{d/2})\lambda_d(B_1^{(d)})(\log(1/2+t)^{-1})^{d/2}, \quad t \in [0, 1/2).$$

In Figure 1 and Figure 2 we provide an illustration of ρ and ℓ_ρ for $d = 2$. Straightforwardly one obtains the inverse of ℓ_ρ given by $\ell_\rho^{-1}\colon (0, \ell_\rho(0)) \to (0, 1/2)$ with

$$\ell_\rho^{-1}(s) = \exp\bigg(-\bigg(\frac{s}{(2^{d/2} + 2^d)\lambda_d(B_1^{(d)})}\bigg)^{2/d}\bigg) - 1/2.$$

Now, for $k \in \mathbb{N}$ we can define rotational invariant unnormalized densities

$$\widetilde{\rho}^{(k)}\colon B^{(k)}_{(\ell_\rho(0)/\lambda_k(B_1^{(k)}))^{1/k}} \to (0, \infty)$$

by

$$\widetilde{\rho}^{(k)}(y) := \ell_\rho^{-1}(\lambda_k(B_1^{(k)})|y|^k)$$

which have the same level-set function as ρ, that is, $\ell_\rho(t) = \ell_{\widetilde{\rho}^{(k)}}(t)$ for all $t \in (0, 1/2)$. Note that the dimension of the domain of $\widetilde{\rho}^{(k)}$ is k, whereas for ρ it is d and d does not need to

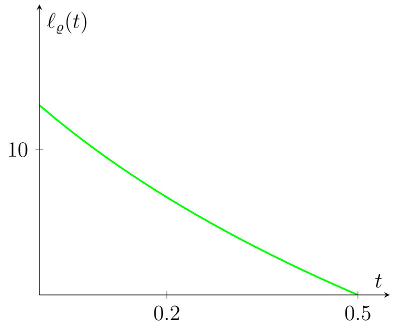

FIG. 2. *Plot of ℓ_ρ of Example 3.8 for $d = 2$.*

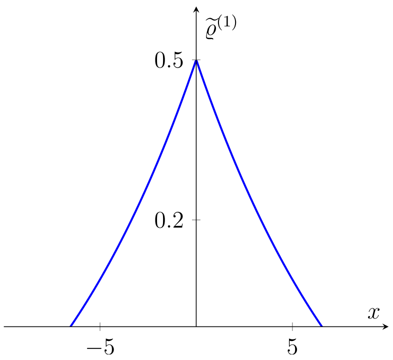

FIG. 3. *Plot of* $\widetilde{\rho}^{(1)}$.

coincide with k. In Figure 3 and Figure 4 we display $\widetilde{\rho}^{(k)}$ for $k=1$, $k=2$ and $d=2$. By Corollary 3.7 we can conclude that the spectral gaps of U_ρ and $U_{\widetilde{\rho}^{(k)}}$ are the same. Moreover, the auxiliary densities $\widetilde{\rho}^{(k)}$ are of the form $\widetilde{\rho}^{(k)}(x) = \exp(-\varphi_k(|x|))$ on their domain, where

$$\varphi_k(s) := -\log \ell^{-1}(s^k) = -\log\Big(\exp\Big(-\Big(\frac{s^k}{(2^{d/2}+2^d)\lambda_d(B_1^{(d)})}\Big)^{2/d}\Big) - 1/2\Big)$$

for all $s \in [0, (\ell_\rho(0)/\lambda_k(B_1^{(k)})^{1/k})$. Thus, for $k \geq \lceil \frac{d}{2} \rceil$ the function φ_k is strictly increasing and convex, that is, the unnormalized density $\widetilde{\rho}^{(k)}$ satisfies the assumptions of Theorem 2.1 and Corollary 3.1, respectively. Hence, we can conclude that simple slice sampling of the bimodal target π on $\mathbb{R}^d$ given by ρ has a spectral gap of at least

$$\mathrm{gap}_\pi(U_\rho) \geq \frac{1}{\lceil \frac{d}{2} \rceil + 1}.$$

The previous example suggests the definition of the following classes of level-set functions.

DEFINITION 3.9. A continuous function $\ell\colon (0,\infty) \to [0,\infty]$ belongs to the class Λ_k with $k \in \mathbb{N}$ if:

1. ℓ is strictly decreasing on its open support

$$\operatorname{supp} \ell := \big(0, \sup\{t \in (0,\infty) \mid \ell(t) > 0\}\big),$$

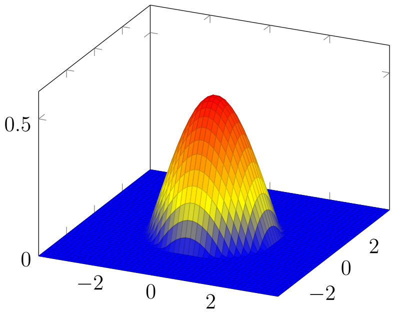

FIG. 4. *Plot of* $\widetilde{\rho}^{(2)}$.

which implies the existence of the inverse ℓ^{-1} on $\ell(\operatorname{supp}\ell) = (0, \|\ell\|_\infty)$ with

$$\|\ell\|_\infty := \sup_{s\in(0,\infty)} = \lim_{t\to 0+} \ell(t) = \ell(0+),$$

2. the function $g\colon (0, \|\ell\|_\infty^{1/k}) \to \operatorname{supp}\ell$ given by $g(s) := \ell^{-1}(s^k)$ is log-concave, that is, $\log g$ is concave.

The main result of this section is then as follows.

THEOREM 3.10. *For an unnormalized density $\rho\colon G \to \mathbb{R}^+$ assume that its level-set function $\ell_\rho \in \Lambda_k$ for $k \in \mathbb{N}$. Then*

$$\operatorname{gap}_\pi(U_\rho) \ge \frac{1}{k+1}.$$

PROOF. The idea here is to construct an unnormalized density $\widetilde{\rho}^{(k)}\colon \mathbb{R}^k \to \mathbb{R}^+$ such that $\ell_\rho = \ell_{\widetilde{\rho}^{(k)}}$ and $\widetilde{\rho}^{(k)}$ satisfies the assumptions of Theorem 2.1. The statement then follows by Corollary 3.1 and Corollary 3.7. To this end, we define $\widetilde{\rho}^{(k)}\colon \mathring{B}_{R_k}^{(k)} \to \mathbb{R}^+$ with $R_k := (\|\ell_\rho\|_\infty/\lambda_k(B_1^{(k)}))^{1/k}$ by

$$\widetilde{\rho}^{(k)}(x) := \ell_\rho^{-1}\bigl(\lambda_k(B_1^{(k)})|x|^k\bigr), \quad |x| < R_k.$$

By construction we have for any $t \in (0, \infty)$

$$\ell_{\widetilde{\rho}^{(k)}}(t) = \lambda_k\bigl(\{x \in \mathbb{R}^k : \widetilde{\rho}^{(k)}(x) \ge t\}\bigr) = \ell_\rho(t).$$

Next, we observe that $\widetilde{\rho}^{(k)}(x) = \exp(-\varphi_k(|x|))$ for $|x| < R_k$ with

$$\varphi_k(s) := -\log \ell_\rho^{-1}\bigl(\lambda_k(B_1^{(k)})s^k\bigr), \quad s \in [0, R_k).$$

Since ℓ_ρ belongs to Λ_k, we know that $s \mapsto \log \ell_\rho^{-1}(s^k)$ is concave. This yields the convexity of φ_k on $[0, R_k)$. Moreover, $\ell_\rho \in \Lambda_k$ implies that also ℓ_ρ^{-1} is strictly decreasing on $[0, \|\ell_\rho\|_\infty)$. Thus, the mapping $s \mapsto \log \ell_\rho^{-1}(s^k)$ is strictly decreasing and, therefore, φ_k is strictly increasing. Hence, the unnormalized density $\widetilde{\rho}^{(k)}$ satisfies the assumptions of Theorem 2.1 which finishes the proof. □

Notice that the lower the number k of the class Λ_k the larger the lower bound of the spectral gap. Subsequently, we provide some (sufficient) characterizations of the classes Λ_k.

3.2.1. *Properties of the class Λ_k.* The requirements of a level-set function to belong to the class Λ_k are not easy to check. We provide some auxiliary tools. The following is a trivial consequence of the definition of Λ_k.

PROPOSITION 3.11. *If $\ell \in \Lambda_k$ for $k \in \mathbb{N}$ and $c > 0$, then $c \cdot \ell \in \Lambda_k$.*

Now a sufficient condition for being in Λ_1 is stated.

PROPOSITION 3.12. *If $\ell\colon (0, \infty) \to [0, \infty)$ is strictly decreasing and concave, then $\ell \in \Lambda_1$.*

PROOF. Since ℓ is strictly decreasing and concave we have that ℓ^{-1} is concave. Then $\log \ell^{-1}$ is log-concave and $\ell \in \Lambda_1$. □

Assuming smoothness of ℓ the previous result can be extended and provides a characterisation of Λ_k.

PROPOSITION 3.13. *Let* $\ell\colon (0,\infty) \to [0,\infty)$ *be continuously differentiable on its open support* $\operatorname{supp}\ell$ *with* $\ell'(t) < 0$. *Define the function* $\psi\colon \operatorname{supp}\ell \to [0,\infty)$ *by* $\psi(t) := \frac{t\ell'(t)}{\ell(t)^{1-1/k}}$ *for* $k \in \mathbb{N}$. *Then*

$$\ell \in \Lambda_k \quad \Longleftrightarrow \quad \psi \text{ is decreasing}.$$

PROOF. The function ℓ is strictly decreasing on $\operatorname{supp}\ell$, since $\ell'(t) < 0$ on that interval. This implies that the inverse $\ell^{-1}\colon [0, \|\ell\|_\infty) \to \operatorname{supp}\ell$ exists and is strictly decreasing. Define the function $\varphi_k\colon [0, \|\ell\|_\infty^{1/k}) \to \mathbb{R}$ with $\varphi_k(s) := -\log \ell^{-1}(s^k)$. Observe that φ_k is strictly increasing and by the inverse mapping theorem continuously differentiable on $\operatorname{supp}\ell$. We have

$$\varphi_k'(s) = -\frac{\mathrm{d}}{\mathrm{d}s}\log \ell^{-1}(s^k) = -\frac{1}{\ell^{-1}(s^k)}\left(\frac{\mathrm{d}}{\mathrm{d}s}\ell^{-1}(s^k)\right) = -\frac{1}{\ell^{-1}(s^k)}\frac{ks^{k-1}}{\ell'(\ell^{-1}(s^k))}.$$

Given the assumptions we have that $\ell \in \Lambda_k$ if and only if φ_k is convex. The latter is equivalent to φ_k' being increasing. Note that for $s \in [0, \|\ell\|_\infty^{1/k})$

$$\begin{aligned}\frac{ks^{k-1}}{\ell'(\ell^{-1}(s^k))} &= k\frac{s^k}{s\ell'(\ell^{-1}(s^k))} = k\frac{\ell(\ell^{-1}(s^k))}{s\ell'(\ell^{-1}(s^k))} = k\frac{\ell(\ell^{-1}(s^k))}{(\ell(\ell^{-1}(s^k)))^{1/k}\ell'(\ell^{-1}(s^k))}\\ &= k\frac{(\ell(\ell^{-1}(s^k)))^{1-1/k}}{\ell'(\ell^{-1}(s^k))}.\end{aligned}$$

Hence, with $h(t) := -\frac{\ell^{1-1/k}(t)}{t\ell'(t)}$ we obtain $\varphi_k'(s) = k \cdot h(\ell^{-1}(s^k))$, which leads to the fact that

$$\varphi_k' \text{ increasing} \quad \Longleftrightarrow \quad h \text{ decreasing}.$$

However, the latter is equivalent to the fact that the mapping $t \mapsto \frac{t\ell'(t)}{\ell(t)^{1-1/k}}$ is decreasing, since $\frac{\ell(t)^{1-1/k}}{t\ell'(t)} < 0$ on $\operatorname{supp}\ell$. □

REMARK 3.14. Roberts and Rosenthal [20] derived convergence results of simple slice sampling given the assumption that $t \mapsto t\ell'(t)$ is decreasing which corresponds to the sufficient condition for $\ell \in \Lambda_1$. In particular, they write "However, it is surprising that this same bound[4] applies to any density ρ such that $t\ell'(t)$ is nonincreasing".[5] We also observe this surprising fact, but w.r.t. the spectral gap. In contrast to their result, in general, we do not require the existence of the first derivative from the level-set function. Moreover, our result for Λ_k with $k > 1$ has no analogues in the work of Roberts and Rosenthal. To emphasize this we consider in Section 3.2.2 an example of a level-set function which is in Λ_2 but not in Λ_1.

[4] They provide a quantitative bound of $\|U_\rho^n(x,\cdot) - \pi(\cdot)\|_{\mathrm{tv}}$ for any continuously differentiable ℓ as in Proposition 3.13.

[5] For the formulas we adapted their statement to our notation, namely in their work our ρ is π and our ℓ is denoted by Q.

3.2.2. *Further examples.* We illustrate in two more examples the advantages of Theorem 3.10 compared to Theorem 2.1.

EXAMPLE 3.15. For $\alpha > 0$ and $\gamma > 0$ let $\rho^{(d)}: \mathbb{R}^d \to \mathbb{R}^+$ be given by $\rho^{(d)}(x) = \exp(-\alpha|x|^\gamma)$. By Proposition 3.11 it is sufficient to consider

$$\ell(t) := \left(\frac{\log t^{-1}}{\alpha}\right)^{d/\gamma} = c_1 \ell_{\rho^{(d)}}(t), \quad t \in (0,\infty),$$

with $c_1 = \lambda_d(B_1^{(d)})$. The function ℓ is strictly decreasing and $\log \ell^{-1}(s^k) = -\alpha s^{\gamma \frac{k}{d}}$. Thus, for any $\gamma \geq 1$ and $k = d$ it is concave on $(0,\infty)$, such that for this parameters $\ell \in \Lambda_d$ and by Theorem 3.10

$$\text{gap}(U_{\rho^{(d)}}) \geq \frac{1}{d+1}.$$

However, we notice that $\log \ell^{-1}(s^k) = -\alpha s^{\gamma k/d}$ is concave for $k \geq \lceil d/\gamma \rceil$. Otherwise, for $k < \lceil d/\gamma \rceil$ it is convex. Thus, we have that $\ell_\rho \in \Lambda_{\lceil d/\gamma \rceil}$ but if $d < \gamma$, then $\ell_\rho \notin \Lambda_{\lceil d/\gamma \rceil - 1}$. For instance, for $\gamma = d/2$ we have that $\ell_\rho \in \Lambda_2$ and $\ell_\rho \notin \Lambda_1$. Hence, Theorem 3.10 tells us that for this class of target densities

$$\text{gap}(U_{\rho^{(d)}}) \geq \frac{1}{\lceil d/\gamma \rceil + 1} \geq \frac{1}{d+1}.$$

In the following we consider a “volcano” density.

EXAMPLE 3.16. Let $\rho^{(d)}: \mathbb{R}^d \to \mathbb{R}^+$ be given by $\rho^{(d)}(x) = e^{-|x|^{2d}+2|x|^d}$. In contrast to Example 3.15 here we have more than a single peak. For $d = 2$ the density is plotted in Figure 5. It is easy to see that $\ell_{\rho^{(d)}}$ is proportional to the strictly decreasing function $\ell: (0,\infty) \to [0,\infty)$ given by

$$\ell(t) := \begin{cases} 1 + \sqrt{1 + \log t^{-1}}, & 0 < t \leq 1, \\ 2\sqrt{1 + \log t^{-1}}, & 1 < t \leq e, \end{cases}$$

such that by Proposition 3.11 it is sufficient to consider ℓ. This leads to

$$\ell^{-1}(s) = \begin{cases} e^{1-\frac{s^2}{4}}, & 0 \leq s < 2, \\ e^{-s^2+2s}, & s \geq 2, \end{cases}$$

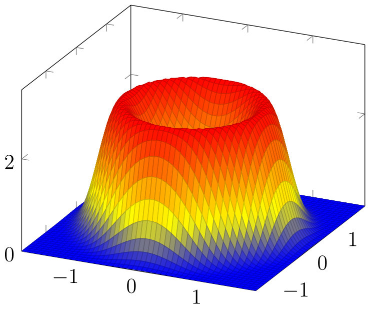

FIG. 5. *Plot of* $\rho^{(2)}(x) = e^{-|x|^4+2|x|^2}$.

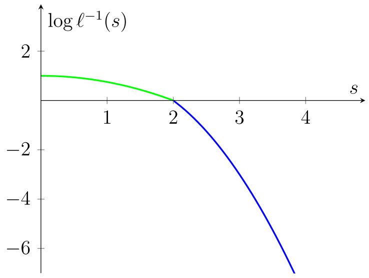

FIG. 6. *Plot of* $\log \ell^{-1}(s)$ *from Example* 3.16.

and we have that $\log \ell^{-1}(s)$ is concave; see also Figure 6. Hence $\ell \in \Lambda_1$ for arbitrary d and Theorem 3.10 implies

$$\operatorname{gap}_\pi(U_{\rho^{(d)}}) \geq \frac{1}{2}.$$

Acknowledgements. Viacheslav Natarovskii thanks the DFG Research Training Group 2088 for their support. Daniel Rudolf thanks Andreas Eberle for fruitful discussions on this topic and acknowledges support of the Felix–Bernstein-Institute for Mathematical Statistics in the Biosciences (Volkswagen Foundation) and the Campus laboratory AIMS. Björn Sprungk thanks the DFG for supporting this research within project 389483880 and the Research Training Group 2088.

REFERENCES

[1] ATHREYA, K. B. and LAHIRI, S. N. (2006). *Measure Theory and Probability Theory*. *Springer Texts in Statistics*. Springer, New York. MR2247694

[2] BESAG, J. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B* **55** 25–37. MR1210422

[3] BOGACHEV, V. I. (2007). *Measure Theory. Vol. I, II*. Springer, Berlin. MR2267655 https://doi.org/10.1007/978-3-540-34514-5

[4] CHEN, M. F. and WANG, F. Y. (1994). Application of coupling method to the first eigenvalue on manifold. *Sci. China Ser. A* **37** 1–14. MR1308707

[5] CONWAY, J. B. (1985). *A Course in Functional Analysis*. *Graduate Texts in Mathematics* **96**. Springer, New York. MR0768926 https://doi.org/10.1007/978-1-4757-3828-5

[6] FLEGAL, J. M. and JONES, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.* **38** 1034–1070. MR2604704 https://doi.org/10.1214/09-AOS735

[7] HIGDON, D. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Amer. Statist. Assoc.* **93** 585–595.

[8] KIPNIS, C. and VARADHAN, S. R. S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.* **104** 1–19. MR0834478

[9] KONTOYIANNIS, I. and MEYN, S. P. (2012). Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probab. Theory Related Fields* **154** 327–339. MR2981426 https://doi.org/10.1007/s00440-011-0373-4

[10] ŁATUSZYŃSKI, K. and RUDOLF, D. (2014). Convergence of hybrid slice sampling via spectral gap. arXiv preprint, arXiv:1409.2709.

[11] MIRA, A., MØLLER, J. and ROBERTS, G. O. (2001). Perfect slice samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 593–606. MR1858405 https://doi.org/10.1111/1467-9868.00301

[12] MIRA, A. and TIERNEY, L. (2002). Efficiency and convergence properties of slice samplers. *Scand. J. Stat.* **29** 1–12. MR1894377 https://doi.org/10.1111/1467-9469.00267

[13] MULLER, O., YANG, M. Y. and ROSENHAHN, B. (2013). Slice sampling particle belief propagation. In *Proceedings of the IEEE International Conference on Computer Vision* 1129–1136.

[14] MURRAY, I., ADAMS, R. and MACKAY, D. (2010). Elliptical slice sampling. *J. Mach. Learn. Res. Workshop Conf. Proc.* **9** 541–548.

[15] Neal, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. With discussions and a rejoinder by the author. MR1994729 https://doi.org/10.1214/aos/1056562461

[16] Nishihara, R., Murray, I. and Adams, R. P. (2014). Parallel MCMC with generalized elliptical slice sampling. *J. Mach. Learn. Res.* **15** 2087–2112. MR3231603

[17] Novak, E. and Rudolf, D. (2014). Computation of expectations by Markov chain Monte Carlo methods. In *Extraction of Quantifiable Information from Complex Systems. Lecture Notes in Computational Science and Engineering* **102** 397–411. Springer, Cham. MR3329348 https://doi.org/10.1007/978-3-319-08159-5_20

[18] Ollivier, Y. (2009). Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.* **256** 810–864. MR2484937 https://doi.org/10.1016/j.jfa.2008.11.001

[19] Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Probab.* **2** 13–25. MR1448322 https://doi.org/10.1214/ECP.v2-981

[20] Roberts, G. O. and Rosenthal, J. S. (1999). Convergence of slice sampler Markov chains. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 643–660. MR1707866 https://doi.org/10.1111/1467-9868.00198

[21] Roberts, G. O. and Rosenthal, J. S. (2002). The polar slice sampler. *Stoch. Models* **18** 257–280. MR1904830 https://doi.org/10.1081/STM-120004467

[22] Rudolf, D. (2012). Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Math.* **485** 1–93. MR2977521 https://doi.org/10.4064/dm485-0-1

[23] Rudolf, D. and Schweizer, N. (2018). Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli* **24** 2610–2639. MR3779696 https://doi.org/10.3150/17-BEJ938

[24] Rudolf, D. and Ullrich, M. (2013). Positivity of hit-and-run and related algorithms. *Electron. Commun. Probab.* **18** no. 49, 8. MR3078012 https://doi.org/10.1214/ECP.v18-2507

[25] Rudolf, D. and Ullrich, M. (2018). Comparison of hit-and-run, slice sampler and random walk Metropolis. *J. Appl. Probab.* **55** 1186–1202. MR3899935 https://doi.org/10.1017/jpr.2018.78

[26] Tibbits, M. M., Groendyke, C., Haran, M. and Liechty, J. C. (2014). Automated factor slice sampling. *J. Comput. Graph. Statist.* **23** 543–563. MR3215824 https://doi.org/10.1080/10618600.2013.791193

[27] Tibbits, M. M., Haran, M. and Liechty, J. C. (2011). Parallel multivariate slice sampling. *Stat. Comput.* **21** 415–430. MR2806618 https://doi.org/10.1007/s11222-010-9178-z

[28] Ullrich, M. (2014). Rapid mixing of Swendsen–Wang dynamics in two dimensions. *Dissertationes Math.* **502** 64. MR3222829 https://doi.org/10.4064/dm502-0-1

[29] Villani, C. (2003). *Topics in Optimal Transportation. Graduate Studies in Mathematics* **58**. Amer. Math. Soc., Providence, RI. MR1964483 https://doi.org/10.1090/gsm/058

CHAPTER B

Geometric convergence of elliptical slice sampling

Notice that due to specific format restrictions of the journal this paper is presented here in two parts: the article itself (B.1) and the supplementary material (B.2).

B.1 The article

Geometric Convergence of Elliptical Slice Sampling

Viacheslav Natarovskii[1] Daniel Rudolf[1] Björn Sprungk[2]

Abstract

For Bayesian learning, given likelihood function and Gaussian prior, the elliptical slice sampler, introduced by Murray, Adams and MacKay 2010, provides a tool for the construction of a Markov chain for approximate sampling of the underlying posterior distribution. Besides of its wide applicability and simplicity its main feature is that no tuning is required. Under weak regularity assumptions on the posterior density we show that the corresponding Markov chain is geometrically ergodic and therefore yield qualitative convergence guarantees. We illustrate our result for Gaussian posteriors as they appear in Gaussian process regression, as well as in a setting of a multi-modal distribution. Remarkably, our numerical experiments indicate a dimension-independent performance of elliptical slice sampling even in situations where our ergodicity result does not apply.

1. Introduction

Probabilistic modeling provides a versatile tool in the analysis of data and allows for statistical inference. In particular, in Bayesian approaches one is able to quantify model and prediction uncertainty by extracting knowledge from the posterior distribution through sampling. The generation of exact samples w.r.t. the posterior distribution is usually quite difficult, since it is in most scenarios only known up to a normalizing constant. Let $\varrho\colon \mathbb{R}^d \to (0,\infty)$ be determined by a likelihood function given some data (which we omit in the following for simplicity) as mapping from the parameter space into the non-negative reals and let $\mu_0 = \mathcal{N}(0,C)$ be a Gaussian prior distribution on $\mathbb{R}^d$ with non-degenerate covariance matrix C, such that the posterior distribution μ on $\mathbb{R}^d$ takes the form

$$\mu(\mathrm{d}x) = \frac{\varrho(x)}{Z}\mu_0(\mathrm{d}x), \quad Z := \int_{\mathbb{R}^d} \varrho(x)\,\mu_0(\mathrm{d}x). \tag{1}$$

For convenience, we abbreviate the former relation between the measures $\mu(\mathrm{d}x)$ and $\varrho(x)\mu_0(\mathrm{d}x)$ as $\mu(\mathrm{d}x) \propto \varrho(x)\mu_0(\mathrm{d}x)$.

A standard approach for generating approximate samples w.r.t. μ is given by Markov chain Monte Carlo. The idea is to construct a Markov chain, which has μ as its stationary and limit distribution[1]. For this purpose in machine learning (and computational statistics in general) Metropolis-Hastings algorithms and slice sampling algorithms (which include Gibbs sampling) are classical tools, see, e.g., (Neal, 1993; Andrieu et al., 2003; Neal, 2003).

Murray, Adams and MacKay in (Murray et al., 2010) introduced the elliptical slice sampler. On the one hand it is based on a Metropolis-Hastings method suggested by Neal (Neal, 1999) (nowadays also known as preconditioned Crank-Nicolson Metropolis (Cotter et al., 2013; Rudolf & Sprungk, 2018)) and on the other hand it is a modification of slice sampling with stepping-out and shrinkage (Neal, 2003). Elliptical slice sampling is illustrated in (Murray et al., 2010) on a number of applications, such as Gaussian regression, Gaussian process classification and a Log Gaussian Cox process. Apart from its simplicity and wide applicability the main advantage of the suggested algorithm is that it performs well in practice and no tuning is necessary. In addition to that in many scenarios it appears as a building block and/or influenced methodological development of sampling approaches (Fagan et al., 2016; Hahn et al., 2019; Bierkens et al., 2020; Murray & Graham, 2016; Nishihara et al., 2014).

However, despite the arguments for being reversible w.r.t. the desired posterior in (Murray et al., 2010) there is, to our knowledge, no theory guaranteeing indeed convergence of the corresponding Markov chain. Under a tail and a weak boundedness assumption on ϱ we derive a small set and Lyapunov function which imply geometric ergodicity by standard theorems for Markov chains on general state spaces, see e.g. chapter 15 in (Meyn & Tweedie, 2009)

[1] Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Göttingen, Germany [2] Faculty of Mathematics and Computer Science, Technische Universität Bergakademie Freiberg, Germany. Correspondence to: Viacheslav Natarovskii <vnataro@uni-goettingen.de>, Daniel Rudolf <daniel.rudolf@uni-goettingen.de>, Björn Sprungk <bjoern.sprungk@math.tu-freiberg.de>.

Proceedings of the 38^th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

[1] Limit distribution in the sense that for $n \to \infty$ the distribution of the nth random variable of the Markov chain converges to μ.

and/or (Hairer & Mattingly, 2011).

Before we state our ergodicity result in Section 2 we provide the algorithm and introduce notation as well as basic facts. Afterwards we state the detailed analysis, in particular, the strategy of proof as well as verify the two crucial conditions of having a Lyapunov function and a sufficiently large small set. In Section 4 we illustrate the applicability of our theoretical result in a fully Gaussian and multi-modal scenario. Additionally, we compare elliptical with simple slice sampling and different Metropolis-Hastings algorithms numerically. The experiments indicate dimension-independent statistical efficiency of elliptical slice sampling which will be the content of future research.

2. Convergence of Elliptical Slice Sampling

We start with stating the transition mechanism/kernel of elliptical slice sampling in algorithmic form and provide our notation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space of all subsequently used random variables. For $a, b \in \mathbb{R}$ with $a < b$ let $\mathcal{U}[a,b]$ be the uniform distribution on $[a,b]$ and let $\mathcal{B}(\mathbb{R}^d)$ be the Borel σ-algebra of $\mathbb{R}^d$. Furthermore, the Euclidean ball with radius $R > 0$ around $x \in \mathbb{R}^d$ is denoted by $B_R(x)$ and the Euclidean norm is given by $\| \cdot \|$.

2.1. Transition Mechanism

We use the function $p : \mathbb{R}^d \times \mathbb{R}^d \times [0, 2\pi] \to \mathbb{R}^d$ defined as

$$p(x, w, \theta) := \cos(\theta)\, x + \sin(\theta)\, w, \tag{2}$$

where, for fixed $x, w \in \mathbb{R}^d$, the map $\theta \mapsto p(x, w, \theta)$ describes an ellipse in $\mathbb{R}^d$ with conjugate diameters x, w. Furthermore, for $t \geq 0$ let

$$G_t := \{x \in \mathbb{R}^d \colon \varrho(x) \geq t\},$$

be the (super-)level set of ϱ w.r.t. t. Using this notation a single transition of elliptical slice sampling from $x \in \mathbb{R}^d$ to y is presented in Algorithm 1. Here $y \in \mathbb{R}^d$ is considered as a realization of a random variable Y_x. Let us denote the transition kernel which corresponds to elliptical slice sampling by $E \colon \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ and for $A \in \mathcal{B}(\mathbb{R}^d)$ observe that

$$E(x, A) = \frac{1}{\varrho(x)} \int_0^{\varrho(x)} \int_{\mathbb{R}^d} E_{x,w,t}(A)\, \mu_0(\mathrm{d}w)\mathrm{d}t,$$

where

$$E_{x,w,t}(A) := \mathbb{P}(Y_x \in A \mid W = w, T_x = t) \tag{3}$$

is determined by steps 3-14 of Algorithm 1. These steps of the algorithm determine the sampling mechanism on G_t intersected with the ellipse by using a suitable adaptation of the shrinkage procedure, see (Neal, 2003; Murray et al., 2010). Let $(X_n)_{n \in \mathbb{N}}$ be a Markov chain generated by Algorithm 1, that is, a Markov chain on $\mathbb{R}^d$ with transition kernel E. Then, for any $n \in \mathbb{N}$, $A \in \mathcal{B}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$ we have

$$\mathbb{P}(X_{n+1} \in A \mid X_1 = x) = E^n(x, A), \tag{4}$$

where E^n is iteratively defined as

$$E^{n+1}(x, A) = \int_{\mathbb{R}^d} E^n(z, A) E(x, \mathrm{d}z), \tag{5}$$

with $E^0(x, A) = \mathbb{1}_A(x)$ denoting the indicator function of the set A.

Algorithm 1 Elliptical Slice Sampler

```
input current state $x \in \mathbb{R}^d$
output next state $y$ as realization of a random variable $Y_x$
 1: draw $W \sim \mu_0$, call the result $w$;
 2: draw $T_x \sim \mathcal{U}[0, \varrho(x)]$, call the result $t$;
 3: draw $\Theta \sim \mathcal{U}[0, 2\pi]$, call the result $\theta$;
 4: $\theta_{\min} \leftarrow \theta - 2\pi$
 5: $\theta_{\max} \leftarrow \theta$
 6: while $p(x, w, \theta) \notin G_t$ do
 7:     if $\theta < 0$ then
 8:         $\theta_{\min} \leftarrow \theta$
 9:     else
10:         $\theta_{\max} \leftarrow \theta$
11:     end if
12:     draw $\Theta \sim \mathcal{U}[\theta_{\min}, \theta_{\max}]$, set the result to $\theta$;
13: end while
14: $y \leftarrow p(x, w, \theta)$
```

2.2. Main Result

Before we formulate the theorem, we state the assumptions which eventually imply the convergence result.

Assumption 2.1. *The function* $\varrho \colon \mathbb{R}^d \to (0, \infty)$ *satisfies the following properties:*

1. *It is bounded away from* 0 *and* ∞ *on any compact set.*
2. *There exists an* $\alpha > 0$ *and* $R > 0$*, such that*

$$B_{\alpha\|x\|}(0) \subseteq G_{\varrho(x)} \qquad \textit{for } \|x\| > R.$$

The boundedness condition from below and above of ϱ on compact sets is relatively weak and appears frequently in qualitative proofs for geometric ergodicity of Markov chain algorithms, see e.g. (Roberts & Tweedie, 1996). The second condition tells us that ϱ has a sufficiently nice tail behavior. It is satisfied if the tails are rotational invariant and monotone decreasing, e.g., like $\exp(-\kappa\|x\|)$ for arbitrary $\kappa > 0$. For examples of ϱ which satisfy Assumption 2.1 we refer to Section 4.

For stating the geometric ergodicity of elliptical slice sampling we introduce the total variation distance of two probability measures π, ν on $\mathbb{R}^d$ as

$$\|\pi - \nu\|_{\text{tv}} := \sup_{\|f\|_\infty \le 1} \left| \int_{\mathbb{R}^d} f(x)(\pi(\mathrm{d}x) - \nu(\mathrm{d}x)) \right|,$$

where $\|f\|_\infty := \sup_{x \in \mathbb{R}^d} |f(x)|$ for $f \colon \mathbb{R}^d \to \mathbb{R}$.

Theorem 2.2. *For elliptical slice sampling under Assumption 2.1 there exist constants $C > 0$ and $\gamma \in (0,1)$, such that*

$$\|E^n(x,\cdot) - \mu\|_{\text{tv}} \le C(1 + \|x\|)\gamma^n, \quad \forall n \in \mathbb{N}, \forall x \in \mathbb{R}^d. \tag{6}$$

Remark 2.3. A transition kernel which satisfies an inequality as in (6) is called geometrically ergodic, since the distribution of X_{n+1}, given that the initial state $X_1 = x$, converges exponentially/geometrically fast to μ. Here, the right-hand side depends on x only via the term $1 + \|x\|$. We view this result as a qualitative statement telling us about exponential convergence of the Markov chain whereas we do not care too much about the constants $C > 0$ and $\gamma \in (0,1)$. The main reason behind this is, that the employed technique of proof does usually not provide sharp bounds on γ and C, particularly regarding their dependence on the dimension d.

3. Detailed Analysis

For proving geometric ergodicity for Markov chains on general state spaces we employ a standard strategy, which consists of the verification of a suitable small set as well as a drift or Lyapunov condition, see e.g. chapter 15 in (Meyn & Tweedie, 2009) or (Hairer & Mattingly, 2011). More precisely we use a consequence of the Harris ergodic theorem as formulated in (Hairer & Mattingly, 2011), which provides a relatively concise introduction and proof of a geometric ergodicity result for Markov chains.

3.1. Strategy of Proof

To formulate the convergence theorem we need the notion of a Lyapunov function and a small set. For this let $P \colon \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ be a generic transition kernel.

We call a function $V \colon \mathbb{R}^d \to [0,\infty)$ *Lyapunov function* of P with $\delta \in [0,1)$ and $L \in [0,\infty)$ if for all $x \in \mathbb{R}^d$ holds

$$PV(x) := \int_{\mathbb{R}^d} V(y)\, P(x, \mathrm{d}y) \le \delta V(x) + L. \tag{7}$$

Furthermore, a set $S \in \mathcal{B}(\mathbb{R}^d)$ is a *small set w.r.t.* P and a non-zero finite measure ν on $\mathbb{R}^d$, if

$$P(x,A) \ge \nu(A), \qquad \forall x \in S, A \in \mathcal{B}(\mathbb{R}^d).$$

With this terminology we can state a consequence of Theorem 1.2 in (Hairer & Mattingly, 2011), which we justify for the convenience of the reader in the supplementary material.

Proposition 3.1. *Suppose that for a transition kernel P there is a Lyapunov function $V \colon \mathbb{R}^d \to [0,\infty)$ with $\delta \in [0,1)$ and $L \in [0,\infty)$ ((7) is satisfied). Additionally, for some constant $R > 2L/(1-\delta)$ let*

$$S_R := \{x \in \mathbb{R}^d \colon V(x) \le R\} \tag{8}$$

be a small set w.r.t. P and a non-zero measure ν on $\mathbb{R}^d$. Then, there is a unique stationary distribution $\mu_\star$ on $\mathbb{R}^d$, that is, for all $A \in \mathcal{B}(\mathbb{R}^d)$

$$\mu_\star P(A) := \int_{\mathbb{R}^d} P(x,A)\mu_\star(\mathrm{d}x) = \mu_\star(A),$$

and there exist constants $\gamma \in (0,1)$ as well as $C < \infty$ such that

$$\|P^n(x,\cdot) - \mu_\star\|_{\text{tv}} \le C(1 + V(x))\gamma^n, \quad \forall n \in \mathbb{N}, \forall x \in \mathbb{R}^d.$$

(Here P^n is the n-step transition kernel defined as in (5).)

From the arguments of reversibility of elliptical slice sampling w.r.t. μ derived in (Murray et al., 2010) we know already that μ is a stationary distribution w.r.t. the transition kernel E. The idea is now to first detect a suitable Lyapunov function V of E satisfying (7) for $P = E$ and a $\delta \in [0,1)$ and $L \in [0,\infty)$ and, having this, proving that the corresponding set S_R from (8) is a small set w.r.t. E and a suitable measure ν.

3.2. Lyapunov Function

Besides the usefulness of a Lyapunov function in the context of geometric convergence of Markov chains as in Proposition 3.1 it arises to derive certain stability properties, e.g., it crucially appears in the perturbation theory of Markov chains in measuring the difference of transition kernels (Rudolf & Schweizer, 2018; Medina-Aguayo et al., 2020).

We start with the following abstract proposition inspired by Lemma 3.2 in (Hairer et al., 2014), see also Proposition 3 in (Hosseini & Johndrow, 2018).

Proposition 3.2. *Let P be a transition kernel on $\mathbb{R}^d$ such that for $Y_x \sim P(x,\cdot)$, $x \in \mathbb{R}^d$, there exists a random variable W_x with $\mathbb{E}\|W_x\| \le K$ for a constant $K < \infty$ independent of x and*

$$\|Y_x\| \le \|x\| + \|W_x\| \qquad \textit{almost surely.} \tag{9}$$

Additionally, assume that there exists a radius $R > 0$, constants $\ell \in (0,1]$ and $\widetilde{\ell} \in [0,1)$ such that for all $x \in B_R(0)^c := \{x \in \mathbb{R}^d \colon \|x\| > R\}$ there is a set $D_x \in \mathcal{B}(\mathbb{R}^d)$ satisfying *(a)* $P(x, D_x) \geq \ell$,

(b) $\sup_{y \in D_x} \|y\| \leq \widetilde{\ell} \|x\|$.

Then $V(x) := \|x\|$ *is a Lyapunov function for* P *with* $L := R + K < \infty$ *and* $\delta := 1 - (1 - \widetilde{\ell})\ell < 1$.

Proof. We distinguish whether $x \in B_R(0)$ or $x \in B_R(0)^c$. Consider the case $x \in B_R(0)$: By assumption we have a.s. $V(Y_x) \leq V(x) + V(W_x)$, such that

$$\begin{aligned} PV(x) = \mathbb{E}V(Y_x) \leq V(x) + \mathbb{E}\|W_x\| \quad &\leq R + K \\ &\leq \delta V(x) + R + K. \end{aligned}$$

Consider the case $x \in B_R(0)^c$: We have

$$PV(x) = \int_{D_x} V(y)\, P(x, \mathrm{d}y) + \int_{D_x^c} V(y)\, P(x, \mathrm{d}y).$$

For the first term we obtain

$$\int_{D_x} V(y)\, P(x, \mathrm{d}y) \underset{\text{(b)}}{\leq} \widetilde{\ell}\, V(x)\, P(x, D_x).$$

To bound the second term observe that

$$\begin{aligned} \int_{D_x^c} V(y)\, P(x, \mathrm{d}y) &= \mathbb{E}(\mathbf{1}_{D_x^c}(Y_x)\, V(Y_x)) \\ &\underset{(9)}{\leq} V(x)\mathbb{P}(Y_x \in D_x^c) + \mathbb{E}\|W_x\|. \end{aligned}$$

We have

$$\mathbb{P}(Y_x \in D_x^c) = 1 - \mathbb{P}(Y_x \in D_x) = 1 - P(x, D_x)$$

and combining both estimates above yields

$$\begin{aligned} PV(x) &\leq \widetilde{\ell}\, V(x)\, P(x, D_x) + (1 - P(x, D_x))\, V(x) + L \\ &= [1 - P(x, D_x) + \widetilde{\ell}\, P(x, D_x)]\, V(x) + L. \end{aligned}$$

By the fact that $1 - (1 - \widetilde{\ell})P(x, D_x) \leq \delta$ the assertion is proven. □

We apply this proposition in the context of elliptical slice sampling and obtain the following result.

Lemma 3.3. *Assume that there exists an* $\alpha \in (0, 1/\sqrt{2}]$ *and* $R > 0$, *such that* $B_{\alpha\|x\|}(0) \subseteq G_{\varrho(x)}$ *for all* $x \in B_R(0)^c$. *Then, the function* $V(x) := \|x\|$ *is a Lyapunov function for* E *with some* $\delta \in [0, 1)$ *and* $L \in [0, \infty)$.

Proof. From (2) we have for all $x, w \in \mathbb{R}^d$ and any $\theta \in [0, 2\pi]$ that

$$\|p(x, w, \theta)\| \leq \|x\| + \|w\|. \tag{10}$$

Thus, condition (9) is satisfied for the transition kernel E with $W_x \sim \mu_0$ being the random variable W in line 1 of Algorithm 1. Next, we show that for any $x \in B_R(0)^c$ and $D_x := B_{\alpha\|x\|}(0)$ the assumptions (a) and (b) of Proposition 3.2 are satisfied for an $\ell \in (0, 1]$ and an $\widetilde{\ell} \in [0, 1)$. Obviously, $\sup_{y \in D_x} \|y\| \leq \widetilde{\ell}\|x\|$ for $\widetilde{\ell} = \frac{1}{\sqrt{2}} < 1$ even for all $x \in \mathbb{R}^d$. Thus, it is sufficient to find a number $\ell \in (0, 1]$ such that

$$E(x, D_x) \geq \ell, \qquad \forall x \in B_R(0)^c.$$

For this notice that the probability to move to a set $A \in \mathcal{B}(\mathbb{R}^d)$ after all trials described in the lines 6–13 of Algorithm 1 is larger than the probability to move to A after exactly one iteration of the loop. Thus, for any $x, w \in \mathbb{R}^d$, $t \in [0, \varrho(x)]$ and $A \in \mathcal{B}(\mathbb{R}^d)$ we have

$$E_{x,w,t}(A) \geq \frac{1}{2\pi} \int_0^{2\pi} \mathbb{1}_{A \cap G_t}(p(x, w, \theta))\, \mathrm{d}\theta, \tag{11}$$

with $E_{x,w,t}$ as given in (3). Further, notice that for any $x \in B_R(0)^c$ and any $t \in [0, \varrho(x)]$ we have

$$D_x = B_{\alpha\|x\|}(0) \subseteq G_{\varrho(x)} \subseteq G_t.$$

Defining $\widetilde{\Theta}$ to be a $[0, 2\pi]$-uniformly distributed random variable and using (11) we have for any $x \in B_R(0)^c, w \in \mathbb{R}^d$ and $t \in [0, \varrho(x)]$ that

$$\begin{aligned} E_{x,w,t}(D_x) &\geq \frac{1}{2\pi} \int_0^{2\pi} \mathbb{1}_{D_x}(p(x, w, \theta))\mathrm{d}\theta \\ &= \mathbb{P}\left(p(x, w, \widetilde{\Theta}) \in D_x\right). \end{aligned}$$

Additionally, let $W \sim \mu_0$ be independent of $\widetilde{\Theta}$. Then we have for all $x \in B_R(0)^c$

$$E(x, D_x) \geq \mathbb{P}\left(p(x, W, \widetilde{\Theta}) \in D_x\right). \tag{12}$$

Hence, we need to study the event $\{p(x, W, \widetilde{\Theta}) \in D_x\}$ in more detail. We have

$$p(x, W, \widetilde{\Theta}) \in D_x \iff \|p(x, W, \widetilde{\Theta})\| \leq \alpha\|x\|,$$

which is equivalent to

$$\begin{aligned} \|p(x, W, \widetilde{\Theta})\|^2 = \|x\|^2 \cos^2(\widetilde{\Theta}) &+ \|W\|^2 \sin^2(\widetilde{\Theta}) \\ &+ 2\langle x, W\rangle \sin(\widetilde{\Theta}) \cos(\widetilde{\Theta}) \\ &\leq \alpha^2 \|x\|^2, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^d$. Defining

$$\begin{aligned} A_W &:= \|x\|^2 - \|W\|^2, \\ B_W &:= 2\langle x, W\rangle, \\ C_W &:= (2\alpha^2 - 1)\|x\|^2 - \|W\|^2, \end{aligned}$$

and using the trigonometric identities

$$\begin{aligned}\cos(2\theta) &= 2\cos^2(\theta) - 1 = 1 - 2\sin^2(\theta),\\ \sin(2\theta) &= 2\cos(\theta)\sin(\theta),\end{aligned}$$

we have that $p(x, W, \widetilde{\Theta}) \in D_x$ is equivalent to

$$A_W \cos(2\widetilde{\Theta}) + B_W \sin(2\widetilde{\Theta}) \le C_W.$$

Letting $\varphi_W \in [0, 2\pi)$ be an angle satisfying

$$\cos(\varphi_W) = \frac{A_W}{\sqrt{A_W^2 + B_W^2}}, \quad \sin(\varphi_W) = \frac{B_W}{\sqrt{A_W^2 + B_W^2}},$$

and using the cosine of sum identity we get

$$\cos(2\widetilde{\Theta} - \varphi_W) \le \frac{C_W}{\sqrt{A_W^2 + B_W^2}}.$$

At this point we have

$$\left\{p(x, W, \widetilde{\Theta}) \in D_x\right\} = \left\{\cos(2\widetilde{\Theta} - \varphi_W) \le \frac{C_W}{\sqrt{A_W^2 + B_W^2}}\right\}. \quad (13)$$

Note that A_W, B_W, C_W, φ_W are all random variables which depend on W, but are independent of $\widetilde{\Theta}$. We aim to condition on the event $\|W\|^2 \le \frac{\alpha^2 R^2}{2-\alpha^2}$. In this case $C_W < 0$ and $A_W > 0$, such that

$$0 > \frac{C_W}{\sqrt{A_W^2 + B_W^2}} \ge \frac{C_W}{A_W} = \frac{(2\alpha^2 - 1)\|x\|^2 - \|W\|^2}{\|x\|^2 - \|W\|^2}.$$

The last fraction can be rewritten as

$$\frac{(\alpha^2 - 1)(\|x\|^2 - \|W\|^2) + \alpha^2\|x\|^2 + (\alpha^2 - 2)\|W\|^2}{\|x\|^2 - \|W\|^2}$$

or equivalently as

$$(\alpha^2 - 1) + \frac{\alpha^2}{\|x\|^2 - \|W\|^2}\left(\|x\|^2 + \frac{\alpha^2 - 2}{\alpha^2}\|W\|^2\right).$$

The second term is non-negative, therefore, we have

$$\frac{C_W}{\sqrt{A_W^2 + B_W^2}} \ge \alpha^2 - 1 > -1. \quad (14)$$

With $\ell_R := \mathbb{P}\left(\|W\|^2 \le \frac{\alpha^2 R^2}{2-\alpha^2}\right) > 0$ we have

$$\mathbb{P}\left(p(x, W, \widetilde{\Theta}) \in D_x\right) \ge \mathbb{P}\left(p(x, W, \widetilde{\Theta}) \in D_x \middle| \|W\|^2 \le \frac{\alpha^2 R^2}{2 - \alpha^2}\right)\ell_R.$$

Now using (13) and (14) we have that

$$\mathbb{P}\left(p(x, W, \widetilde{\Theta}) \in D_x\right) \ge \mathbb{P}\left(\cos(2\widetilde{\Theta} - \varphi_W) \le \alpha^2 - 1 \middle| \|W\|^2 \le \frac{\alpha^2 R^2}{2 - \alpha^2}\right)\ell_R.$$

For any random variable ξ independent of $\widetilde{\Theta}$ we have that the distribution of $\cos(2\widetilde{\Theta} - \xi)$ coincides with the distribution of $\cos(2\widetilde{\Theta})$, since $\widetilde{\Theta}$ is uniformly distributed on $[0, 2\pi]$. Recall that φ_W is independent of $\widetilde{\Theta}$. Therefore, with

$$\varepsilon_\alpha := \mathbb{P}\left(\cos(2\widetilde{\Theta}) \le \alpha^2 - 1\right) > 0$$

we have

$$\mathbb{P}\left(\cos(2\widetilde{\Theta} - \varphi_W) \le \alpha^2 - 1 \middle| \|W\|^2 \le \frac{\alpha^2 R^2}{2 - \alpha^2}\right) = \varepsilon_\alpha.$$

Putting everything together, we conclude that

$$E(x, D_x) \overset{(12)}{\ge} \mathbb{P}\left(p(x, W, \widetilde{\Theta}) \in D_x\right) \ge \varepsilon_\alpha \ell_R > 0$$

and all assumptions of Proposition 3.2 are then satisfied with $\ell := \varepsilon_\alpha \ell_R$. □

3.3. Small Set

In this section we show that under suitable assumptions any compact set is small w.r.t. the transition kernel E of elliptical slice sampling.

Lemma 3.4. *Assume that ϱ is bounded away from 0 and ∞ on any compact set. Then any compact set $G \subset \mathbb{R}^d$ is small w.r.t. E and the measure $\varepsilon \cdot \lambda_G$, where $\varepsilon > 0$ is some constant and λ_G denotes the d-dimensional Lebesgue measure restricted to G.*

Proof. Let $x \in G$ be arbitrary and recall that for any $A \in \mathcal{B}(\mathbb{R}^d)$ we have

$$E(x, A) = \frac{1}{\varrho(x)} \int_0^{\varrho(x)} \int_{\mathbb{R}^d} E_{x,w,t}(A)\mu_0(\mathrm{d}w)\mathrm{d}t,$$

where we argued in (11) that

$$E_{x,w,t}(A) \ge \frac{1}{2\pi} \int_0^{2\pi} \mathbb{1}_{A \cap G_t}(p(x, w, \theta))\,\mathrm{d}\theta,$$

for any $w \in \mathbb{R}^d$ and $t \in [0, \varrho(x)]$. Therefore, we obtain

$$E(x, A) \ge \frac{1}{2\pi\varrho(x)} \int\limits_0^{\varrho(x)} \int\limits_{\mathbb{R}^d} \int\limits_0^{2\pi} \mathbb{1}_{A \cap G_t}(p(x, w, \theta))\mathrm{d}\theta\mu_0(\mathrm{d}w)\mathrm{d}t.$$

Changing the order of integration yields

$$E(x,A) \geq \frac{1}{2\pi\varrho(x)} \int_0^{\varrho(x)} \int_0^{2\pi} \mathbb{E}\big(\mathbb{1}_{A\cap G_t}(p(x,W,\theta))\big) \mathrm{d}\theta \mathrm{d}t$$

for some random vector $W \sim \mathcal{N}(0,C)$. Define the auxiliary random vector $Z_{x,\theta} := p(x,W,\theta)$ with corresponding distribution $\nu_{x,\theta} := \mathcal{N}(x\cos(\theta), \sin^2(\theta)C)$. Then

$$\begin{aligned} E(x,A) &\geq \frac{1}{2\pi\varrho(x)} \int_0^{\varrho(x)} \int_0^{2\pi} \mathbb{E}\big(\mathbb{1}_{A\cap G_t}(Z_{x,\theta})\big) \mathrm{d}\theta \mathrm{d}t \\ &= \frac{1}{2\pi\varrho(x)} \int_0^{\varrho(x)} \int_0^{2\pi} \int_A \mathbb{1}_{G_t}(z) \nu_{x,\theta}(\mathrm{d}z) \mathrm{d}\theta \mathrm{d}t. \end{aligned}$$

Using the fact that $\mathbb{1}_{G_t}(z) = \mathbb{1}_{[0,\varrho(z)]}(t)$ we have

$$E(x,A) \geq \int_0^{2\pi} \int_A \frac{1}{2\pi\varrho(x)} \int_0^{\varrho(x)} \mathbb{1}_{[0,\varrho(z)]}(t) \mathrm{d}t \, \nu_{x,\theta}(\mathrm{d}z) \mathrm{d}\theta.$$

Notice that

$$\begin{aligned} \frac{1}{\varrho(x)} \int_0^{\varrho(x)} \mathbb{1}_{[0,\varrho(z)]}(t) \mathrm{d}t &= \frac{1}{\varrho(x)} \min\{\varrho(x), \varrho(z)\} \\ &= \min\left\{1, \frac{\varrho(z)}{\varrho(x)}\right\}. \end{aligned}$$

Moreover, for all $x,z \in G$ by the boundedness assumption on ϱ we have

$$\min\left\{1, \frac{\varrho(z)}{\varrho(x)}\right\} \geq \min\left\{1, \frac{\inf_{a\in G} \varrho(a)}{\sup_{a\in G} \varrho(a)}\right\} =: \beta > 0.$$

Thus,

$$\begin{aligned} E(x,A) &\geq \frac{\beta}{2\pi} \int_0^{2\pi} \nu_{x,\theta}(A\cap G) \mathrm{d}\theta \\ &\geq \frac{\beta}{2\pi} \int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \nu_{x,\theta}(A\cap G) \mathrm{d}\theta. \end{aligned}$$

Since G is a compact set, there exists a finite constant $\kappa > 0$, such that

$$(z - x\cos\theta)^T C^{-1} (z - x\cos\theta) \leq \kappa, \qquad \forall x,z \in G.$$

Moreover, for all $\theta \in \left[\frac{\pi}{4}, \frac{\pi}{2}\right]$ we have that $\frac{1}{2} \leq \sin^2(\theta) \leq 1$. Therefore, the factors of the density of the Gaussian distribution $\nu_{x,\theta}$ satisfy

$$\exp\left(-\frac{(z - x\cos\theta)^T C^{-1} (z - x\cos\theta)}{2\sin^2(\theta)}\right) \geq \exp(-\kappa),$$

and

$$\left(2\pi\sin^2(\theta)\right)^{-\frac{d}{2}} \det(C)^{-\frac{1}{2}} \geq (2\pi)^{-\frac{d}{2}} \det(C)^{-\frac{1}{2}}.$$

Hence,

$$\nu_{x,\theta}(A\cap G) \geq \frac{\exp(-\kappa)}{(2\pi)^{\frac{d}{2}} \det(C)^{\frac{1}{2}}} \lambda_G(A),$$

such that finally with $\varepsilon := \frac{\beta}{8} \frac{\exp(-\kappa)}{(2\pi)^{\frac{d}{2}} \det(C)^{\frac{1}{2}}}$ we have

$$E(x,A) \geq \varepsilon \cdot \lambda_G(A),$$

which finishes the proof. □

Remark 3.5. For a compact set $G \subset \mathbb{R}^d$ suppose that $\varrho\colon G \to (0,\infty)$ with $0 < \inf_{x\in G} \varrho(x)$ and $\sup_{x\in G} \varrho(x) < \infty$. In this setting the same arguments as in the proof of Lemma 3.4 can be used to verify that the whole state space G is small w.r.t. elliptical slice sampling. This leads to the fact that elliptical slice sampling is uniformly ergodic in this scenario, see for example Theorem 15.3.1 in (Douc et al., 2018). For a summary of different ergodicity properties and their relations to each other we refer to Section 3.1 in (Rudolf, 2012).

3.4. Proof of Theorem 2.2

We apply Proposition 3.1. First, recall that in (Murray et al., 2010) it is verified that elliptical slice sampling is reversible w.r.t. μ and therefore μ is a stationary distribution of E. Hence, it is sufficient to provide a Lyapunov function and to check the smallness of S_R. By Assumption 2.1 part 2. the requirements for Lemma 3.3 are satisfied, such that $V(x) := \|x\|$ is a Lyapunov function with $\delta \in [0,1)$ and $L \in [0,\infty)$. By Assumption 2.1 part 1. using Lemma 3.4 we obtain that for any $R > 2L/(1-\delta)$ the set $S_R = B_R(0)$ is compact and therefore small w.r.t. transition kernel E and some non-trivial finite measure. Therefore, all requirements of Proposition 3.1 are satisfied and the statement of Theorem 2.2 follows.

4. Illustrative Examples

In this section we verify in toy scenarios as well as more demanding settings the conditions of Assumption 2.1 to illustrate the applicability of our result. In the supplementary we provide a discussion in terms of the exponential family.

4.1. Gaussian Posterior

In (Murray et al., 2010) Gaussian regression is considered as test scenario for elliptical slice sampling, since there the posterior distribution is again Gaussian. We see covering that setting as a minimal requirement for our theory: Here, for some $x_0 \in \mathbb{R}^d$ we have

$$\varrho(x) = \exp\left(-\frac{1}{2}(x-x_0)^T \Sigma^{-1} (x-x_0)\right), \quad x \in \mathbb{R}^d, \tag{15}$$

that is, ϱ is proportional to a Gaussian density with non-degenerate covariance matrix Σ. Thus, the matrix $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric, positive-definite, and we denote its eigenvalues by $\lambda_1, \dots, \lambda_d$. Notice that all eigenvalues are strictly positive and define $\lambda_{\min} := \min_{i=1,\dots,d} \lambda_i$, $\lambda_{\max} := \max_{i=1,\dots,d} \lambda_i$. The covariance matrix induces a norm $\|\cdot\|_{\Sigma^{-1}}$ on $\mathbb{R}^d$ by

$$\|x\|^2_{\Sigma^{-1}} = x^T \Sigma^{-1} x.$$

It is well-known that the Euclidean and the Σ^{-1}-norm are equivalent. One has

$$\lambda_{\max}^{-1}\|x\|^2 \leq \|x\|^2_{\Sigma^{-1}} \leq \lambda_{\min}^{-1}\|x\|^2, \quad \forall x \in \mathbb{R}^d. \tag{16}$$

Now we are able to formulate and prove the following proposition guaranteeing the applicability of Theorem 2.2.

Proposition 4.1. *For ϱ defined in (15) Assumption 2.1 is satisfied with $R = 4\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\|x_0\|$ and $\alpha = \frac{1}{2}\sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}$.*

Proof. Observe that ϱ is continuous, bounded by 1 and strictly larger than 0 everywhere, such that part 1. of Assumption 2.1 is true. By exploiting both inequalities in (16) we show part 2. of Assumption 2.1, that is, we verify for all $x \in B_{4\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\|x_0\|}(0)^c$ holds $B_{\frac{1}{2}\sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}\|x\|}(0) \subseteq G_{\varrho(x)}$. For this fix $x \in B_{4\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\|x_0\|}(0)^c$. Therefore, we have

$$\|x\| \geq 4\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\,\|x_0\|. \tag{17}$$

Now let $y \in B_{\frac{1}{2}\sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}\|x\|}(0)$. Therefore, we have

$$\|y\| \leq \frac{1}{2}\sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}\|x\|, \tag{18}$$

and one might observe that

$$G_{\varrho(x)} = \{y \in \mathbb{R}^d \colon \|y - x_0\|_{\Sigma^{-1}} \leq \|x - x_0\|_{\Sigma^{-1}}\}.$$

With this we obtain

$$\begin{aligned}
&\|y - x_0\|_{\Sigma^{-1}} \leq \|y\|_{\Sigma^{-1}} + \|x_0\|_{\Sigma^{-1}} \underset{(16)}{\leq} \frac{\|y\|}{\lambda_{\min}^{1/2}} + \|x_0\|_{\Sigma^{-1}} \\
&\underset{(18)}{\leq} \frac{\|x\|}{2\lambda_{\max}^{1/2}} + \|x_0\|_{\Sigma^{-1}} \underset{(16)}{\leq} \frac{\|x\|_{\Sigma^{-1}}}{2} + \|x_0\|_{\Sigma^{-1}} \\
&= \|x\|_{\Sigma^{-1}} - \|x_0\|_{\Sigma^{-1}} - \frac{\|x\|_{\Sigma^{-1}}}{2} + 2\|x_0\|_{\Sigma^{-1}} \\
&\underset{(16)}{\leq} \|x - x_0\|_{\Sigma^{-1}} - \frac{\|x\|}{2\lambda_{\max}^{1/2}} + 2\|x_0\|_{\Sigma^{-1}} \\
&\underset{(17)}{\leq} \|x - x_0\|_{\Sigma^{-1}} - \frac{2\|x_0\|}{\lambda_{\min}^{1/2}} + 2\|x_0\|_{\Sigma^{-1}} \underset{(16)}{\leq} \|x - x_0\|_{\Sigma^{-1}}
\end{aligned}$$

which provides the desired result. □

In Gaussian process regression as well as Bayesian inverse problems with linear forward maps the resulting posterior distribution has again a Gaussian density ϱ with respect to the Gaussian prior μ_0. However, in these applications the corresponding covariance matrix of ϱ is typically positive semi-definite, and we have to replace Σ^{-1} in (15) by its pseudo-inverse $\Sigma^{\dagger}$. We emphasize that also in this more general situation Assumption 2.1 is satisfied, since ϱ is then simply constant on the null space of Σ and on its orthogonal complement we can apply Proposition 4.1.

4.2. Multi-modality

In the previous section we considered the setting of a Gaussian posterior distribution μ. In particular, ϱ had just a single peak. It seems that such a requirement is not necessary to verify the crucial Assumption 2.1. Here we introduce a class of density functions which might behave almost arbitrarily in their "center" (the central part of the state space) and exhibit a certain tail behavior. For formulating the result, let $|\cdot|$ be a norm on $\mathbb{R}^d$ which is equivalent to the Euclidean norm $\|\cdot\|$, that is, there exist constants $c_1, c_2 \in (0, \infty)$ such that

$$c_1\|x\| \leq |x| \leq c_2\|x\|, \quad \forall x \in \mathbb{R}^d. \tag{19}$$

Proposition 4.2. *For some $R' > 0$ and some $x_0 \in \mathbb{R}^d$ let $\varrho_{R'} \colon B_{R'}(x_0) \to (0,\infty)$ be continuous and let $r \colon [R', \infty) \to (0, \infty)$ be decreasing. Furthermore, suppose that*

$$\inf_{z \in B_{R'}(x_0)} \varrho_{R'}(z) \geq \sup_{t \geq R'} r(t). \tag{20}$$

Then, the function

$$\varrho(x) := \begin{cases} \varrho_{R'}(x) & x \in B_{R'}(x_0) \\ r(|x - x_0|) & x \in B_{R'}(x_0)^c, \end{cases}$$

satisfies Assumption 2.1 with $R = \max\{R', 4\frac{c_2}{c_1}\|x_0\|\}$ and $\alpha = \frac{c_1}{2c_2}$.

Proof. By the continuity of $\varrho_{R'}$ and the fact that r is strictly positive as well as decreasing part 1. of Assumption 2.1 is satisfied. For part 2. let $x \in B_R(0)^c$, i.e., $\|x\| > R'$ and $\|x\| > 4\frac{c_2}{c_1}\|x_0\|$. Hence, we have by (20) and the decreasing property of r that

$$G_{\varrho(x)} = B_{R'}(x_0) \cup \{y \in B_{R'}(x_0)^c \colon |y - x_0| \leq |x - x_0|\}.$$

Now let $y \in B_{\alpha\|x\|}(0)$ and distinguish two cases:

1. For $y \in B_{R'}(x_0)$ we immediately have $y \in G_{\varrho(x)}$, and we are done.

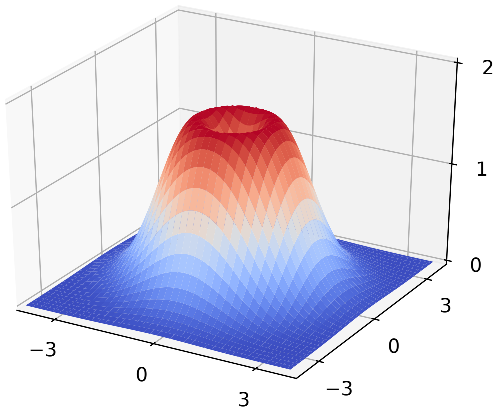

Figure 1. Plot of the function $x \mapsto \exp(\|x\| - \frac{1}{2}\|x\|^2)$ for $d = 2$.

2. For $y \in B_{R'}(x_0)^c \cap B_{\alpha\|x\|}(0)$ we obtain due to $\|y\| \leq \frac{c_1}{2c_2}\|x\|$ that

$$|y - x_0| \leq |y| + |x_0| \underset{(19)}{\leq} \frac{1}{2}|x| + |x_0|$$

and, furthermore, by exploiting $\|x\| > 4\frac{c_2}{c_1}\|x_0\|$ that

$$\begin{aligned}\frac{1}{2}|x| + |x_0| &= |x| - |x_0| - \frac{1}{2}|x| + 2|x_0| \\ &\underset{(19)}{\leq} |x - x_0| - 2|x_0| + 2|x_0| = |x - x_0|,\end{aligned}$$

which leads again to $y \in G_{\varrho(x)}$.

Both cases combined yield the statement. □

To state an example which satisfies the assumption of Proposition 4.2 we consider the following "volcano density".

Example 4.3. Set $\varrho(x) := \exp(\|x\| - \frac{1}{2}\|x\|^2)$. Let $|\cdot| = \|\cdot\|$, $x_0 = 0$, $R' = 2$, $r(t) := \exp(t - t^2/2)$ and $\varrho_{R'}$ be the restriction of ϱ to $B_2(0)$. It is easily checked that for this choice of parameters all required properties are satisfied. One can argue that the function ϱ is highly multimodal, since its maximum is attained on a $d-1$-dimensional manifold (a sphere). For illustration, it is plotted in Figure 1.

4.3. Volcano Density and Limitations of the Result

In the last section we showed the applicability of Theorem 2.2 for a "volcano density". Here we use this density differently. Namely, $\mu(\mathrm{d}x) \propto \exp\left(\|x\| - \frac{1}{2}\|x^2\|\right)\mathrm{d}x$, that is, the Lebesgue density of μ is proportional to the function plotted in Figure 1. Setting $\mu_0 = \mathcal{N}(0, I)$ with identity matrix I, we obtain

$$\varrho(x) = \exp(\|x\|), \quad x \in \mathbb{R}^d. \tag{21}$$

Observe that in this setting for any $x \in \mathbb{R}^d$ we have

$$G_{\varrho(x)} = \{y \in \mathbb{R}^d \colon \|y\| \geq \|x\|\} = B_{\|x\|}(0)^c,$$

such that $G_{\varrho(x)}$ never completely contains a ball around the origin and Assumption 2.1 cannot be satisfied. For this scenario we conduct numerical experiments in various dimensions, namely, $d = 10, 30, 100, 300, 1000$. Although, our sufficient Assumption 2.1 is not satisfied[2], we still observe a good performance of the elliptical slice sampler. In particular, its statistical efficiency in terms of the *effective sample size (ESS)* seems to be independent of the dimension, see Figure 2. To check whether this "dimension-independent" behavior is inherently due to the particular setting or not, we also consider other Markov chain based sampling algorithms.

For estimating the ESS we use an empirical proxy of the autocorrelation function

$$\gamma_f(k) := \mathrm{Corr}(f(X_{n_0}), f(X_{n_0+k})),$$

of the underlying Markov chain $(X_k)_{k\in\mathbb{N}}$ for a chosen quantity of interest $f\colon \mathbb{R}^d \to \mathbb{R}$ where n_0 denotes a burn-in parameter. Since the ESS takes the form

$$\mathrm{ESS}(n, f, (X_k)_{k\in\mathbb{N}}) = n\left(1 + 2\sum_{k=0}^{\infty}\gamma_f(k)\right)^{-1},$$

where $n \in \mathbb{N}$ denotes the chosen sample size, we approximate it by using the empirical proxy of $\gamma_f(k)$ and truncating the summation at $k = 10^4$.

In Figure 2 we display estimates of the ESS for four different Markov chain Monte Carlo algorithms. Namely, the random walk Metropolis algorithm (RWM), the preconditioned Crank-Nicolson Metropolis (pCN), the simple slice sampler and the elliptical one. For each algorithm we set the initial state to be $0 \in \mathbb{R}^d$ and compute the ESS for $f(x) := \log(1 + \|x\|)$, $n_0 := 10^5$ and $n := 10^6$. Both Metropolis algorithms (the RWM and the pCN Metropolis) were tuned to an averaged acceptance probability of approximately 0.25. We clearly see in Figure 2 the dimension-dependence of the ESS for the simple slice sampler[3] and the RWM. In contrast to that, the results for the elliptical slice sampler and the pCN Metropolis indicate a dimension-independent efficiency. Let us remark that elliptical slice sampling does not need to be tuned in comparison to the pCN Metropolis, which performs similarly. However, the price for this is the requirement of evaluating the function ϱ

[2] In the supplementary material we provide further discussions how Assumption 2.1 can be satisfied in this scenario by taking a modification into account.

[3] In the light of (Natarovskii et al., 2021) the dimension-dependent behavior for simple slice sampling is not surprising. There, for a certain class of ϱ a spectral gap of size $1/d$ is proven.

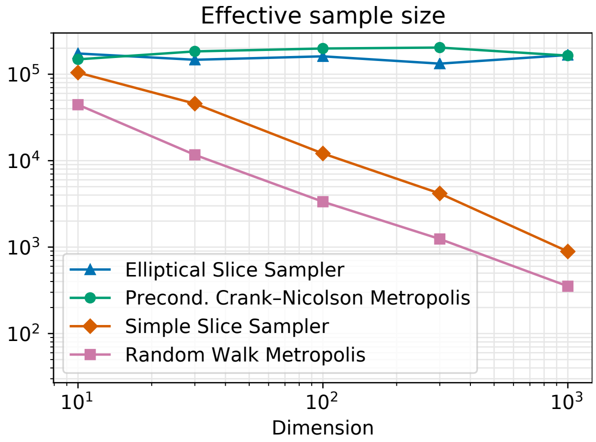

Figure 2. Proxies for ESS for different MCMC algorithms depending on the dimension of the space.

more often within a single transition. Here the function ϱ was evaluated on average 1.5 times in each iteration of the elliptical slice sampler. Intuitively, the example of this section is not covered by our theorem, since the tail behavior of ϱ is "bad". Namely, for $\|x\| \to \infty$ we have $\varrho(x) \to \infty$. It seems that for convergence only the tail behavior of likelihood times prior considered as Lebesgue density matters.

Let us briefly comment on different approaches how to verify the numerically observed dimension-independence. Similarly to the strategy employed in (Hairer et al., 2014; Rudolf & Sprungk, 2018) for the pCN Metropolis one might be able to extend elliptical slice sampling on infinite-dimensional Hilbert spaces. If one proves the existence of an absolute spectral gap of the correspondent transition kernel, then this directly gives bounds of the total variation distance of the nth step distribution to the stationary one. Due to the infinite-dimensional setting one might argue that the estimate must be independent of the dimension. Another approach is to prove dimension-free Wasserstein contraction rates, as, for example, has been done in (Eberle, 2016; Eberle et al., 2019; De Bortoli & Durmus, 2019) for diffusion processes.

4.4. Logistic Regression

Suppose data $(\xi_i, y_i)_{i=1,\dots,N}$ with $\xi_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$ for $i = 1,\dots,N$ is given. For logistic regression the function $\varrho\colon \mathbb{R}^d \to (0,\infty)$ takes the form

$$\varrho(x) = \prod_{i=1}^{N} \frac{1}{1+\exp(-y_i x^T \xi_i)}, \quad x \in \mathbb{R}^d. \qquad (22)$$

Moreover, assume we have a Gaussian prior distribution μ_0 on $\mathbb{R}^d$ with $\mu_0 = \mathcal{N}(0,I)$. Thus, the distribution of interest, i.e., the posterior distribution μ is determined by $\mu(\mathrm{d}x) \propto \varrho(x)\,\mu_0(\mathrm{d}x)$. The function ϱ does not satisfy Assumption 2.1, since it has no vanishing tails. For example for $d = N = \xi_1 = y_1 = 1$ we have $\varrho(x) = (1+\exp(-x))^{-1}$, which is increasing with $G_{\varrho(x)} = [x,\infty)$ for all $\forall x \in \mathbb{R}$. Thus, ϱ cannot satisfy Assumption 2.1. In the general setting the phenomena is the same and the arguments are similar.

Therefore, our theory for elliptical slice sampling seems not to be applicable. However, with a "tail-shift" modification we can satisfy Assumption 2.1. The idea is to take a "small" part of the Gaussian prior and shift it to the likelihood function, such that it gets sufficiently nice tail behavior.

For arbitrary $\varepsilon \in (0,1)$ set $\widetilde{\mu}_0 := \mathcal{N}(0,(1-\varepsilon)^{-1}I)$ and

$$\widetilde{\varrho}(x) := \varrho(x)\exp\left(-\varepsilon\|x\|^2/2\right). \qquad (23)$$

Observe that $\widetilde{\varrho}$ has, in contrast to ϱ, exponential tails. Moreover, note that $\mu_0(\mathrm{d}x) \propto \exp(-\varepsilon\|x\|^2/2)\widetilde{\mu}_0(\mathrm{d}x)$ and therefore

$$\mu(\mathrm{d}x) \propto \varrho(x)\mu_0(\mathrm{d}x) \propto \widetilde{\varrho}(x)\widetilde{\mu}_0(\mathrm{d}x).$$

Now considering μ as given through $\widetilde{\varrho}$ and $\widetilde{\mu}_0$ our main theorem is applicable. In the supplementary material we prove the following result and provide a discussion of the "tail-shift" modification.

Proposition 4.4. *For $\varepsilon \in (0,1)$ the function $\widetilde{\varrho}$ given in (23) satisfies Assumption 2.1 for $\alpha = \varepsilon/2$ and $R = 4N\min_{i=1,\dots,N}\|\xi_i\|/\varepsilon$.*

Finally, note that for having the guarantee of geometric ergodicity of elliptical slice sampling one can choose $\varepsilon \in (0,1)$ arbitrarily small, whereas for $\varepsilon = 0$ our theory does not apply.

5. Conclusion

In this paper we provide a mild sufficient condition for the geometric ergodicity of the elliptical slice sampler in finite dimensions. In particular, it is satisfied if the density of the target measure with respect to a Gaussian measure μ_0 is continuous, strictly positive and has a sufficiently nice tail behavior. Besides that our numerical results indicate that (a) our condition is not necessary and (b) the elliptical slice sampler shows a dimension-independent efficiency. Both issues will be addressed in future research.

Acknowledgements

We thank the anonymous referees for their valuable remarks, in particular, for bringing the "tail-shift" modification to our attention. VN thanks the DFG Research Training Group 2088 for their support. BS acknowledges support of the DFG within project 389483880. DR gratefully acknowledges support of the DFG within project 432680300 – SFB 1456 (subproject B02).

References

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

Bierkens, J., Grazzi, S., Kamatani, K., and Roberts, G. The boomerang sampler. In *International Conference on Machine Learning*, pp. 908–918. PMLR, 2020.

Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, pp. 424–446, 2013.

De Bortoli, V. and Durmus, A. Convergence of diffusions and their discretizations: from continuous to discrete processes and back. *arXiv preprint arXiv:1904.09808*, 2019.

Douc, R., Moulines, E., Priouret, P., and Soulier, P. *Markov chains*. Springer, 2018.

Eberle, A. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3): 851–886, 2016.

Eberle, A., Majka, M. B., et al. Quantitative contraction rates for Markov chains on general state spaces. *Electronic Journal of Probability*, 24, 2019.

Fagan, F., Bhandari, J., and Cunningham, J. P. Elliptical slice sampling with expectation propagation. In *UAI*, 2016.

Hahn, P. R., He, J., and Lopes, H. F. Efficient sampling for Gaussian linear regression with arbitrary priors. *Journal of Computational and Graphical Statistics*, 28(1):142–154, 2019.

Hairer, M. and Mattingly, J. C. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pp. 109–117. Springer, 2011.

Hairer, M., Stuart, A. M., and Vollmer, S. J. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24(6):2455–2490, 2014.

Hosseini, B. and Johndrow, J. E. Spectral gaps and error estimates for infinite-dimensional Metropolis-Hastings with non-Gaussian priors. *arXiv preprint arXiv:1810.00297*, 2018.

Medina-Aguayo, F., Rudolf, D., and Schweizer, N. Perturbation bounds for Monte Carlo within Metropolis via restricted approximations. *Stochastic processes and their applications*, 130(4):2200–2227, 2020.

Meyn, S. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2009.

Murray, I. and Graham, M. Pseudo-marginal slice sampling. In *Artificial Intelligence and Statistics*, pp. 911–919, 2016.

Murray, I., Adams, R., and MacKay, D. Elliptical slice sampling. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 541–548, 2010.

Natarovskii, V., Rudolf, D., and Sprungk, B. Quantitative spectral gap estimate and Wasserstein contraction of simple slice sampling. *The Annals of Applied Probability*, 31 (2):806–825, 2021.

Neal, R. M. *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, Ontario, Canada, 1993.

Neal, R. M. Regression and classification using Gaussian process priors. *J. M. Bernardo et al., editors, Bayesian Statistics*, 6:475–501, 1999.

Neal, R. M. Slice sampling. *Annals of statistics*, pp. 705–741, 2003.

Nishihara, R., Murray, I., and Adams, R. P. Parallel MCMC with generalized elliptical slice sampling. *The Journal of Machine Learning Research*, 15(1):2087–2112, 2014.

Roberts, G. O. and Tweedie, R. L. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.

Rudolf, D. Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Mathematicae*, 485:1–93, 2012.

Rudolf, D. and Schweizer, N. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24 (4A):2610–2639, 2018.

Rudolf, D. and Sprungk, B. On a generalization of the preconditioned Crank–Nicolson Metropolis algorithm. *Foundations of Computational Mathematics*, 18(2):309–343, 2018.

B.2 Supplementary material

Supplementary Material to "Geometric Convergence of Elliptical Slice Sampling"

Viacheslav Natarovskii [1] Daniel Rudolf [1] Björn Sprungk [2]

1. Derivation of Proposition 3.1

We comment on deriving Proposition 3.1 (formulated in the article) from the results in (Hairer & Mattingly, 2011). For stating the Harris ergodic theorem shown in (Hairer & Mattingly, 2011) we need to introduce the following weighted supremum norm. For a chosen weight function $V\colon \mathbb{R}^d \to [0,\infty)$ and for $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ define

$$\|\varphi\|_V := \sup_{x\in\mathbb{R}^d} \frac{|\varphi(x)|}{1+V(x)}.$$

One may think of V as the Lyapunov function of a generic transition kernel P. Now we state Theorem 1.2 from (Hairer & Mattingly, 2011) on $\mathbb{R}^d$.

Theorem 1.1. *Let P be a transition kernel on $\mathbb{R}^d$. Assume that $V\colon \mathbb{R}^d \to [0,\infty)$ is a Lyapunov function of P with $\delta \in [0,1)$ and $L \in [0,1)$. Additionally, for some constant $R > 2L/(1-\delta)$ let*

$$S_R := \{x \in \mathbb{R}^d \colon V(x) \le R\}$$

be a small set w.r.t. P and a non-zero measure ν on $\mathbb{R}^d$. Then, there is a unique stationary distribution $\mu_\star$ of P on $\mathbb{R}^d$ and there exist constants $\gamma \in (0,1)$ as well as $C < \infty$ such that

$$\|P^n\varphi - \mu_\star(\varphi)\|_V \le C\gamma^n \|\varphi - \mu_\star(\varphi)\|_V, \qquad (1)$$

where $P^n\varphi(x) := \int_{\mathbb{R}^d} \varphi(y) P^n(x,\mathrm{d}y)$ and $\mu_\star(\varphi) := \int_{\mathbb{R}^d} \varphi(y)\mu_\star(\mathrm{d}y)$ for any $x \in \mathbb{R}^d$ as well as any $n \in \mathbb{N}$.

Let us assume that all requirements of the previous theorem are satisfied. Then, for any $x \in \mathbb{R}^d$ we have

$$\begin{aligned}
&\|P^n(x,\cdot) - \mu_\star\|_{\mathrm{tv}} = \sup_{\|f\|_\infty \le 1} |P^n f(x) - \mu_\star(f)| \\
&\le \sup_{\|\varphi\|_V \le 1} |P^n\varphi(x) - \mu_\star(\varphi)| \\
&= (1+V(x)) \sup_{\|\varphi\|_V \le 1} \frac{|P^n\varphi(x) - \mu_\star(\varphi)|}{1+V(x)} \\
&\le (1+V(x)) \sup_{\|\varphi\|_V \le 1} \|P^n\varphi - \mu_\star(\varphi)\|_V \\
&\le (1+V(x))C\gamma^n \sup_{\|\varphi\|_V \le 1} \|\varphi - \mu_\star(\varphi)\|_V \\
&\le 2(1+V(x))C\gamma^n,
\end{aligned}$$

which shows that the statement of Proposition 3.1 is a consequence of Theorem 1.1.

2. Further Example from the Exponential Family

We formulate a consequence of Proposition 4.2 (stated in the article) in terms of properties of the exponential family and provide examples which eventually satisfy our regularity condition. For the convenience of the reader we repeat the assumption which guarantees the applicability of the main theorem.

Assumption 2.1. *The function $\varrho\colon \mathbb{R}^d \to (0,\infty)$ satisfies the following properties:*

1. *It is bounded away from 0 and ∞ on any compact set.*
2. *There exists an $\alpha > 0$ and $R > 0$, such that*

$$B_{\alpha\|x\|}(0) \subseteq G_{\varrho(x)} \qquad \textit{for } \|x\| > R.$$

It is clear that regularity properties for members of the exponential family are required, since already by part 1. of the former assumption we need that ϱ has full support. For example, ϱ coming from the exponential distribution does not work, since then it is not bounded away from 0 on any compact set where ϱ is equal to 0.

Let $|\cdot|$ be a norm on $\mathbb{R}^d$, which is equivalent to the Euclidean norm $\|\cdot\|$, that is, there exist constants $c_1, c_2 \in (0,\infty)$ such

[1] Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Göttingen, Germany [2] Faculty of Mathematics and Computer Science, Technische Universität Bergakademie Freiberg, Germany. Correspondence to: Viacheslav Natarovskii <vnataro@uni-goettingen.de>, Daniel Rudolf <daniel.rudolf@uni-goettingen.de>, Björn Sprungk <bjoern.sprungk@math.tu-freiberg.de>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

that

$$c_1\|x\| \leq |x| \leq c_2\|x\|, \qquad \forall x \in \mathbb{R}^d. \tag{2}$$

We obtain the following result:

Corollary 2.2. *Let ϱ be proportional to the mapping*

$$x \mapsto \exp(\eta(x)^T\mu - A(x)), \qquad x \in \mathbb{R}^d,$$

for some $\eta : \mathbb{R}^d \to \mathbb{R}^k$, $\mu \in \mathbb{R}^k$ and $A : \mathbb{R}^d \to \mathbb{R}$ with $k \in \mathbb{N}$. Assume that there exists an increasing function $\varphi : [0,\infty) \to \mathbb{R}$ as well as a point $x_0 \in \mathbb{R}^d$, such that

$$\eta(x)^T\mu - A(x) = -\varphi(|x - x_0|), \qquad \forall x \in \mathbb{R}^d,$$

or equivalently, such that ϱ is proportional to the mapping

$$x \mapsto \exp(-\varphi(|x - x_0|)), \qquad x \in \mathbb{R}^d.$$

Then ϱ satisfies Assumption 2.1 with $R = 4\frac{c_2}{c_1}\|x_0\|$ and $\alpha = \frac{c_1}{2c_2}$.

Proof. Apply Proposition 4.2 from the article with arbitrary $R' > 0$, function $r(t) := \exp(-\varphi(t))$ and $\varrho_{R'}(x) = \exp(-\varphi(|x - x_0|))$ defined on $B_{R'}(x_0)$. □

Now we illustrate how to use the former corollary.

2.1. Gaussian density

Despite having the Gaussian setting already covered in Section 4.1 of the article, we show that this canonical member of the exponential family can also be treated with Corollary 2.2.

For any $x_0 \in \mathbb{R}^d$ and any symmetric, positive-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$ the classical Gaussian setting, where

$$\varrho(x) = \exp\left(-\frac{1}{2}(x - x_0)^T\Sigma^{-1}(x - x_0)\right), \quad x \in \mathbb{R}^d,$$

corresponds to a member of the exponential family with $k = 1$, $\mu = -1$, $A(x) = 0$ and

$$\eta(x) = \frac{1}{2}(x - x_0)^T\Sigma^{-1}(x - x_0).$$

It can be rewritten as

$$\varrho(x) = \exp(-\varphi(\|x - x_0\|_{\Sigma^{-1}})), \quad x \in \mathbb{R}^d,$$

with the continuous increasing function $\varphi(t) = t$ and a norm $|\cdot| = \|\cdot\|_{\Sigma^{-1}}$, defined by

$$\|x\|_{\Sigma^{-1}} := x^T\Sigma^{-1}x. \tag{3}$$

Note that the norm is equivalent to the Euclidean one since

$$\lambda_{\text{max}}^{-1}\|x\|^2 \leq \|x\|_{\Sigma^{-1}}^2 \leq \lambda_{\text{min}}^{-1}\|x\|^2, \quad \forall x \in \mathbb{R}^d, \tag{4}$$

where λ_{min} is the smallest and λ_{max} is the largest eigenvalue of the symmetric, positive-definite matrix Σ. Thus, all requirements of Corollary 2.2 are satisfied and therefore Assumption 2.1 is fulfilled.

2.2. Multivariate t-distribution

For any $\nu > 1, x_0 \in \mathbb{R}^d$ and any symmetric, positive-definite matrix Σ we have

$$\varrho(x) = \left(1 + \frac{1}{\nu}(x - x_0)^T\Sigma^{-1}(x - x_0)\right)^{-(\nu+d)/2},$$

for $x \in \mathbb{R}^d$. This corresponds to a member of the exponential family with $k = 1$, $\mu = -1$, $A(x) = 0$ and

$$\eta(x) = \frac{\nu + d}{2}\log\left(1 + \frac{1}{\nu}(x - x_0)^T\Sigma^{-1}(x - x_0)\right).$$

Using $|\cdot| = \|\cdot\|_{\Sigma^{-1}}$ as defined in (3) and the fact that $\varphi \colon [0,\infty) \to \mathbb{R}$, given by

$$\varphi(t) := \frac{\nu + d}{2}\log\left(1 + \frac{1}{\nu}t\right), \quad t \geq 0,$$

is increasing we can apply Corollary 2.2 and therefore Assumption 2.1 is satisfied.

3. "Tail-Shift" Modification

If $\varrho \colon \mathbb{R}^d \to (0,\infty)$ has "poor" tail behavior and therefore does not satisfy Assumption 2.1, as e.g. in the scenario of the "volcano density" or logistic regression considered in the article, then a "tail-shift" modification might help. The idea is to take a small part of the Gaussian prior and shift it to ϱ to get sufficiently "nice" tails.

Assume that the distribution of interest μ is determined by $\varrho \colon \mathbb{R}^d \to (0,\infty)$ and prior distribution $\mu_0 = \mathcal{N}(0, C)$, that is,

$$\mu(\mathrm{d}x) \propto \varrho(x)\mu_0(\mathrm{d}x).$$

For arbitrary $\varepsilon \in (0,1)$ set

$$f(x) := \exp\left(-\frac{\varepsilon}{2}x^TC^{-1}x\right), \qquad x \in \mathbb{R}^d,$$

and $\widetilde{\mu}_0 := \mathcal{N}(0, (1-\varepsilon)^{-1}C))$. Note that

$$\mu_0(\mathrm{d}x) \propto f(x)\widetilde{\mu}_0(\mathrm{d}x). \tag{5}$$

The function f represents the part of μ_0 which we shift from the prior to ϱ. For doing this rigorously we define

$$\widetilde{\varrho}(x) := \varrho(x)f(x), \quad x \in \mathbb{R}^d, \tag{6}$$

and obtain an alternative representation of μ. Namely,

$$\mu(\mathrm{d}x) \propto \varrho(x)\mu_0(\mathrm{d}x) \underset{(5)}{\propto} \varrho(x)f(x)\widetilde{\mu}_0(\mathrm{d}x) \underset{(6)}{=} \widetilde{\varrho}(x)\widetilde{\mu}_0(\mathrm{d}x).$$

Using the representation of μ in terms of $\widetilde{\varrho}$ and $\widetilde{\mu}_0$ it might be possible to satisfy Assumption 2.1 for $\widetilde{\varrho}$ as the following example shows.

Example 3.1. We apply the "tail-shift" modification to the "volcano density" considered in Section 4.3 in the article. Recall that

$$\varrho(x) = \exp(\|x\|), \quad x \in \mathbb{R}^d,$$

and $\mu_0 = \mathcal{N}(0, I)$. For $\varepsilon \in (0,1)$ after setting

$$f(x) := \exp\left(-\frac{\varepsilon}{2}\|x\|^2\right),$$

we obtain $\mu_0(\mathrm{d}x) \propto f(x)\widetilde{\mu}_0(\mathrm{d}x)$ with $\widetilde{\mu}_0 = \mathcal{N}(0, (1-\varepsilon)^{-1}I)$ and

$$\widetilde{\varrho}(x) = \exp\left(\|x\| - \frac{\varepsilon}{2}\|x\|^2\right).$$

By applying Proposition 4.2 from the article with $|\cdot| = \|\cdot\|$, $x_0 = 0$, $R' = 2\varepsilon^{-1}$ and $r(t) := \exp(t - \varepsilon t^2/2)$ as well as $\varrho_{R'}$ being the restriction of $\widetilde{\varrho}$ to $B_{R'}(0)$ we get that Assumption 2.1 is satisfied.

We want to emphasize here that different representations of μ lead, eventually, to different algorithms. Observe that one can choose $\varepsilon \in (0,1)$ arbitrarily small and the requirements for the main theorem are satisfied, whereas for $\varepsilon = 0$ our theory does not apply. Unfortunately it is not always easy to verify Assumption 2.1 in the modified setting.

In the following, we provide another tool for showing Assumption 2.1. Independent of the "tail-shift" modification it can be used to prove that for certain $\varrho\colon \mathbb{R}^d \to (0,\infty)$ the main theorem is applicable.

Proposition 3.2. *For $\varrho\colon \mathbb{R}^d \to (0,\infty)$ and some $R > 0$ suppose that there are continuous functions $\varrho_\ell : \mathbb{R}^d \to (0,\infty)$ and $\varrho_u : \mathbb{R}^d \to (0,\infty)$, such that*

$$\varrho_\ell(x) \le \varrho(x) \le \varrho_u(x), \qquad \forall x \in \mathbb{R}^d. \tag{7}$$

Furthermore, assume that for some $\alpha > 0$ we have

$$A_x := \{y \in \mathbb{R}^d : \varrho_\ell(y) \ge \varrho_u(x)\} \supseteq B_{\alpha\|x\|}(0) \tag{8}$$

for any $x \in B_R(0)^c$. Then ϱ satisfies Assumption 2.1 with constants R and α.

Proof. Obviously, ϱ is bounded away from 0 and ∞ on any compact set, since ϱ_ℓ and ϱ_u are strictly positive and continuous. Therefore, part 1. of Assumption 2.1 is satisfied. For part 2. notice that for all $x \in B_R^c(0)$ holds $G_{\varrho(x)} \supseteq A_x$, since, if $y \in A_x$, then $\varrho_\ell(y) \ge \varrho_u(x)$ and therefore

$$\varrho(y) \underset{(7)}{\ge} \varrho_\ell(y) \ge \varrho_u(x) \underset{(7)}{\ge} \varrho(x).$$

Thus,

$$G_{\varrho(x)} \supseteq A_x \supseteq B_{\alpha\|x\|}(0), \qquad \forall x \in B_R(0)^c,$$

which finishes the proof. □

We apply the former proposition to the logistic regression example and therefore prove Proposition 4.4 from the article.

3.1. Logistic Regression

For some data $(\xi_i, y_i)_{i=1,\dots,N}$ with $\xi_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$ for $i = 1,\dots,N$ let

$$\varrho(x) = \prod_{i=1}^{N} \frac{1}{1+\exp(-y_i x^T \xi_i)}, \quad x \in \mathbb{R}^d. \tag{9}$$

In this case ϱ does not satisfy Assumption 2.1, see Section 4.4 in the main article. Using the "tail-shift" modification changes the picture.

Let $\mu_0 = \mathcal{N}(0, I)$ and note that for arbitrary $\varepsilon \in (0,1)$, with

$$f(x) := \exp(-\varepsilon\|x\|^2/2), \qquad \theta \in \mathbb{R}^d,$$

the measure μ_0 can be expressed as

$$\mu_0(\mathrm{d}x) \propto f(x)\widetilde{\mu}_0(\mathrm{d}x)$$

with $\widetilde{\mu}_0 := \mathcal{N}(0, (1-\varepsilon)^{-1}I)$. Therefore, $\widetilde{\varrho}$ from (6) takes the form

$$\widetilde{\varrho}(x) = \exp(-\varepsilon\|x\|^2/2) \prod_{i=1}^{N} \frac{1}{1+\exp(-y_i x^T \xi_i)}.$$

Observe that $\widetilde{\varrho}$ has, in contrast to ϱ, exponential tails. To apply Proposition 3.2 to $\widetilde{\varrho}$ we need to find suitable lower and upper bounds which satisfy the conditions formulated in (7) and (8). For any $x \in \mathbb{R}^d$ we have by applying the Cauchy-Schwarz inequality that

$$\exp(-\beta\|x\|) \le \varrho(x) \le 1,$$

where $\beta := 2N \min_{i=1,\dots,N} \|\xi_i\|$. Taking this into account, with

$$\varrho_\ell(x) := \exp(-\varepsilon\|x\|^2/2)\exp(-\beta\|x\|),$$
$$\varrho_u(x) := \exp(-\varepsilon\|x\|^2/2),$$

we have the desired lower and upper bound for $\widetilde{\varrho}$. For A_x defined in (8) (based on ϱ_ℓ and ϱ_u) we show that

$$A_x \supseteq \left\{z \in \mathbb{R}^d\colon \|z\| \le \frac{\varepsilon}{2}\|x\|\right\} \tag{10}$$

for all $x \in \mathbb{R}^d$ with $\|x\| \ge 2\beta/\varepsilon$. For this notice that

$$\begin{aligned} A_x &= \left\{z \in \mathbb{R}^d\colon -\beta\|z\| - \varepsilon\|z\|^2/2 \ge -\varepsilon\|x\|^2/2\right\} \\ &= \left\{z \in \mathbb{R}^d : \varepsilon\|z\|^2 + 2\beta\|z\| - \varepsilon\|x\|^2 \le 0\right\} \\ &= \left\{z \in \mathbb{R}^d : \|z\| \le -\beta + \sqrt{\beta^2 + \varepsilon^2\|x\|^2}\right\} \\ &\supseteq \left\{z \in \mathbb{R}^d : \|z\| \le \varepsilon\|x\| - \beta\right\}, \end{aligned}$$

where the inclusion is due to the fact that $\sqrt{\beta^2 + \varepsilon^2\|x\|^2} \ge \varepsilon\|x\|$. We conclude that for any $x \in \mathbb{R}^d$ with $\|x\| \ge 2\beta/\varepsilon$, or equivalently, $\beta \le \varepsilon\|x\|/2$, condition (10) holds true. Thus, all requirements of Proposition 3.2 are fulfilled for

$\alpha = \varepsilon/2$ and $R = 2\beta/\varepsilon$ and therefore $\widetilde{\varrho}$ satisfies Assumption 2.1.

We summarize that the application of the main theorem, which gives geometric ergodicity of elliptical slice sampling, depends on the representation of μ. As pointed out for

$$\mu(\mathrm{d}x) \propto \varrho(x)\mu_0(\mathrm{d}x),$$

with $\varrho\colon \mathbb{R}^d \to (0,\infty)$ and $\mu_0 = \mathcal{N}(0,C)$, it might be possible that Assumption 2.1 is not satisfied. Therefore, for elliptical slice sampling with this representation of μ we do not provide any ergodicity guarantee. However, by using the "tail-shift" modification it is likely that one can find $\widetilde{\varrho}\colon \mathbb{R}^d \to (0,\infty)$ and a Gaussian measure $\widetilde{\mu}_0$ with

$$\mu(\mathrm{d}x) \propto \widetilde{\varrho}(x)\widetilde{\mu}_0(\mathrm{d}x),$$

such that for $\widetilde{\varrho}$ Assumption 2.1 is satisfied and the geometric ergodicity theorem for elliptical slice sampling is applicable for $\widetilde{\varrho}$ and $\widetilde{\mu}_0$.

References

Hairer, M. and Mattingly, J. C. Yet another look at Harris ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pp. 109–117. Springer, 2011.