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

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

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>.

[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

$$p(x, w, \theta) := \cos(\theta)\, x + \sin(\theta)\, w, \qquad (2)$$

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

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

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

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

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.,

---

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 θ;
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 θ;
13: **end while**
14: $y \leftarrow p(x, w, \theta)$

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

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

where E^n is iteratively defined as

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

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

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*

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

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

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

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

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

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

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 $P: \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, 1]$ be a generic transition kernel.

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

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

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

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

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 $V: \mathbb{R}^d \to [0, \infty)$ with $\delta \in [0, 1)$ and $L \in [0, \infty)$ ((7) is satisfied). Additionally, for some constant $R > 2L/(1 - \delta)$ let*

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

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

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

(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 $\mathbb{E}\|W_x\| \leq K$ for a constant $K < \infty$ independent of x and*

$$\|Y_x\| \leq \|x\| + \|W_x\| \qquad \text{almost surely.} \tag{9}$$

Additionally, assume that there exists a radius $R > 0$, constants $\ell \in (0, 1]$ and $\widetilde{\ell} \in [0, 1)$ such that for all $x \in B_R(0)^c := \{x \in \mathbb{R}^d : \|x\| > R\}$ 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

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

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

$$\int_{D_x^c} V(y)\, P(x, \mathrm{d}y) = \mathbb{E}(\mathbf{1}_{D_x^c}(Y_x)\, V(Y_x))$$

$$\underset{\text{(9)}}{\leq} V(x)\mathbb{P}(Y_x \in D_x^c) + \mathbb{E}\|W_x\|.$$

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

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

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

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

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

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

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

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 $\left\{p(x, W, \widetilde{\Theta}) \in D_x\right\}$ 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

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

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

$$A_W := \|x\|^2 - \|W\|^2,$$
$$B_W := 2\langle x, W\rangle,$$
$$C_W := (2\alpha^2 - 1)\|x\|^2 - \|W\|^2,$$

and using the trigonometric identities

$$\cos(2\theta) = 2\cos^2(\theta) - 1 = 1 - 2\sin^2(\theta),$$
$$\sin(2\theta) = 2\cos(\theta)\sin(\theta),$$

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

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

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

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

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

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

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 $\ell := \varepsilon_\alpha \ell_R$. $\qquad \square$

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

$$E(x, A) \ge$$
$$\frac{1}{2\pi\varrho(x)} \int_0^{\varrho(x)} \int_{\mathbb{R}^d} \int_0^{2\pi} \mathbb{1}_{A \cap G_t}(p(x, w, \theta)) \mathrm{d}\theta \mu_0(\mathrm{d}w) \mathrm{d}t.$$

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

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

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

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

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,

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

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

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

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

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

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

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

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

$$\|x\| \geq 4\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \|x_0\|. \quad (17)$$

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

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

and one might observe that

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

With this we obtain

$$\|y - x_0\|_{\Sigma^{-1}} \leq \|y\|_{\Sigma^{-1}} + \|x_0\|_{\Sigma^{-1}} \underset{(16)}{\leq} \frac{\|y\|}{\lambda_{\min}^{1/2}} + \|x_0\|_{\Sigma^{-1}}$$

$$\underset{(18)}{\leq} \frac{\|x\|}{2\lambda_{\max}^{1/2}} + \|x_0\|_{\Sigma^{-1}} \underset{(16)}{\leq} \frac{\|x\|_{\Sigma^{-1}}}{2} + \|x_0\|_{\Sigma^{-1}}$$

$$= \|x\|_{\Sigma^{-1}} - \|x_0\|_{\Sigma^{-1}} - \frac{\|x\|_{\Sigma^{-1}}}{2} + 2\|x_0\|_{\Sigma^{-1}}$$

$$\underset{(16)}{\leq} \|x - x_0\|_{\Sigma^{-1}} - \frac{\|x\|}{2\lambda_{\max}^{1/2}} + 2\|x_0\|_{\Sigma^{-1}}$$

$$\underset{(17)}{\leq} \|x - x_0\|_{\Sigma^{-1}} - \frac{2\|x_0\|}{\lambda_{\min}^{1/2}} + 2\|x_0\|_{\Sigma^{-1}} \underset{(16)}{\leq} \|x - x_0\|_{\Sigma^{-1}}$$

which provides the desired result. $\square$

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

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

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*

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

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

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

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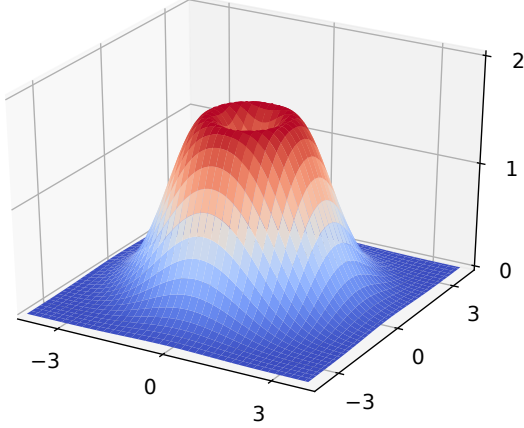
Figure 1. Plot of the function $x \mapsto \exp(\|x\| - \frac{1}{2}\|x\|^2)$ for $d = 2$.

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

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

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

Both cases combined yield the statement. $\qquad\square$

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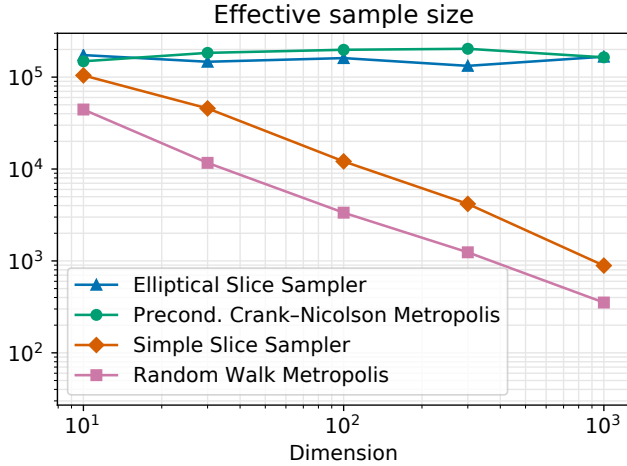
Figure 2. Proxies for ESS for different MCMC algorithms depending on the dimension of the space.

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

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

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 As-

sumption 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

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

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

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. Markov Chains and Stochastic Stability 2 edn (Cambridge: Cambridge University)(with a prologue by PW Glynn). 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.