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*

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

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 38th International Conference on Machine Learning, PMLR 139, 2021.

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

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

where $\lambda_{\min}$ is the smallest and $\lambda_{\max}$ 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

$$\begin{aligned}
\varrho_\ell(x) &:= \exp(-\varepsilon\|x\|^2/2)\exp(-\beta\|x\|),\\
\varrho_u(x) &:= \exp(-\varepsilon\|x\|^2/2),
\end{aligned}$$

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.