Electronic Journal of Statistics
Vol. 14 (2020) 857–889
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1680

On a Metropolis–Hastings importance sampling estimator

Daniel Rudolf[*]

Institute for Mathematical Stochastics
University of Goettingen
Goldschmidtstr. 7
37077 Göttingen
Germany
e-mail: daniel.rudolf@uni-goettingen.de

and

Björn Sprungk[†]

Faculty of Mathematics and Computer Science
Technische Universität Bergakademie Freiberg
09596 Freiberg
Germany
e-mail: bjoern.sprungk@math.tu-freiberg.de

Abstract: A classical approach for approximating expectations of functions w.r.t. partially known distributions is to compute the average of function values along a trajectory of a Metropolis–Hastings (MH) Markov chain. A key part in the MH algorithm is a suitable acceptance/rejection of a proposed state, which ensures the correct stationary distribution of the resulting Markov chain. However, the rejection of proposals causes highly correlated samples. In particular, when a state is rejected it is not taken any further into account. In contrast to that we consider a MH importance sampling estimator which explicitly incorporates all proposed states generated by the MH algorithm. The estimator satisfies a strong law of large numbers as well as a central limit theorem, and, in addition to that, we provide an explicit mean squared error bound. Remarkably, the asymptotic variance of the MH importance sampling estimator does not involve any correlation term in contrast to its classical counterpart. Moreover, although the analyzed estimator uses the same amount of information as the classical MH estimator, it can outperform the latter in scenarios of moderate dimensions as indicated by numerical experiments.

MSC 2010 subject classifications: Primary 62-04, 60J05, 60J22; secondary 60F05, 62F15.

Keywords and phrases: Metropolis–Hastings algorithm, importance sampling, Markov chains, variance reduction, central limit theorem.

Received October 2019.

[*]Supported by the Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences (Volkswagen Foundation) and the Campus laboratory AIMS.
[†]Supported by the DFG within the project 389483880.

1. Introduction

Motivation. A fundamental task in computational science and statistics is the computation of expectations w.r.t. a partially unknown probability measure μ on a measurable space $(G, \mathcal{G})$ determined by

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(x) = \frac{\rho(x)}{Z}, \quad x \in G, \tag{1.1}$$

where μ_0 denotes a σ-finite reference measure on G and where the normalizing constant $Z = \int_G \rho(x)\,\mu_0(\mathrm{d}x) \in (0, \infty)$ is typically unknown. Thus, given a function $f\colon G \to \mathbb{R}$ the goal is to compute $\mathbb{E}_\mu(f) = \int_G f(x)\mu(\mathrm{d}x)$ only by using evaluations of f and ρ. Here, a plain Monte Carlo estimator for the approximation of $\mathbb{E}_\mu(f)$ based on independent μ-distributed random variables is, in general, infeasible due to the unknown normalizing constant Z and the fact that we only have access to function evaluations of ρ. However, a possible and very common approach is the construction of a Markov chain for approximate sampling w.r.t. μ. In particular, the well-known Metropolis–Hastings (MH) algorithm provides a general scheme for simulating a Markov chain $(X_n)_{n\in\mathbb{N}}$ with stationary distribution μ. Under appropriate assumptions the distribution of X_n of such a *MH Markov chain* converges to μ and the *classical MCMC estimator* for $\mathbb{E}_\mu(f)$ is then given by the sample average

$$S_n(f) = \frac{1}{n}\sum_{k=1}^{n} f(X_k). \tag{1.2}$$

The statistical efficiency of $S_n(f)$ highly depends on the autocorrelation of the time series $(f(X_n))_{n\in\mathbb{N}}$. In particular, a large autocorrelation diminishes the efficiency of $S_n(f)$. An essential part in the MH algorithm is the acceptance/rejection step: Given $X_n = x$, a sample y of $Y_{n+1} \sim P(x, \cdot)$ is drawn, where P denotes a *proposal transition kernel*. But only with a certain probability this y is accepted as the next state, that is $X_{n+1} := y$, and otherwise it is rejected, such that $X_{n+1} := x$. This indicates that a potential reason for a high autocorrelation is the rejection of proposed states. Hence, the question arises whether it is possible to derive a more efficient estimator for $\mathbb{E}_\mu(f)$ based on the potentially less correlated time series $(f(Y_n))_{n\in\mathbb{N}}$ determined by the sample of proposals Y_n.

Main result. In this paper we consider and analyze a modification of the classical estimator from (1.2) of the form

$$A_n(f) = \frac{\sum_{k=1}^{n} w(X_k, Y_k) f(Y_k)}{\sum_{k=1}^{n} w(X_k, Y_k)},$$

which we call *MH importance sampling estimator*. The (importance) weight w is chosen in such a way that we obtain a consistent estimator. More detailed, we set $w(x, y) := \frac{\mathrm{d}\mu_0}{\mathrm{d}P(x,\cdot)}(y) \cdot \rho(y)$ assuming the existence of the density $\frac{\mathrm{d}\mu_0}{\mathrm{d}P(x,\cdot)}$ for each $x \in G$. The appeal of the modified estimator is that it is still based

on the MH algorithm and needs no additional function evaluations of ρ and f, while, after appropriate tuning in scenarios of moderate dimensions, it can outperform the classical estimator as we illustrate in a few numerical examples in Section 4. Moreover, it can be seen and studied as an importance sampling corrected MCMC estimator, or as an importance sampling estimator using an underlying MH Markov chain for providing the importance distributions. In this paper we have chosen the first point of view and exploit the fact that the *augmented MH Markov chain* $(X_n, Y_n)_{n\in\mathbb{N}}$ inherits several desirable[1] properties of the original MH Markov chain $(X_n)_{n\in\mathbb{N}}$ such as Harris recurrence, see Lemma 3.1. By using those properties we prove the following results for the estimator A_n:

- Theorem 3.1: A strong law of large numbers (SLLN), i.e., for functions $f \in L^1(\mu)$ we have almost surely $A_n(f) \to \mathbb{E}_\mu(f)$ as $n \to \infty$;
- Theorem 3.2: A central limit theorem (CLT), that is, for any $f \in L^2(\mu)$ the scaled error $\sqrt{n}(A_n(f) - \mathbb{E}_\mu(f))$ converges in distribution to a mean-zero normal distribution $\mathcal{N}(0, \sigma_A^2(f))$ with *asymptotic variance* $\sigma_A^2(f)$ given by

$$\sigma_A^2(f) := \int_G \int_G (f(y) - \mathbb{E}_\mu(f))^2 \frac{\mathrm{d}\mu}{\mathrm{d}P(x,\cdot)}(y)\mu(\mathrm{d}y)\mu(\mathrm{d}x);$$

- Theorem 3.3: An estimate of the mean squared error $\mathbb{E}\,|A_n(f) - \mathbb{E}_\mu(f)|^2$ for bounded functions $f\colon G \to \mathbb{R}$.

Here, we denote by $L^p(\mu)$, $p \in [1, \infty)$ the Lebesgue space of functions $f\colon G \to \mathbb{R}$ which are p-integrable w.r.t. μ. It is remarkable that in the asymptotic variance $\sigma_A^2(f)$ of the CLT there is no covariance or correlation term. However, there appears the density of μ w.r.t. $P(x,\cdot)$ which quantifies the difference of the employed importance distribution given by the proposal transition kernel $P(x,\cdot)$ and the desired distribution μ.

Related literature. Importance sampling is a well-established technique for approximating expectations, see [4, 27] for textbook introductions, which has recently attracted considerable attention in terms of theory and application, see for example [1, 6, 15, 34]. In particular, its combination with Markov chain Monte Carlo methods is exploited by several authors. For example, Botev et al. [3] use the MH algorithm in order to approximately sample from the minimum variance importance distribution. Vihola et al. [39] consider general importance sampling estimators based on an underlying Markov chain and Martino et al. [22] propose a hierarchical approach where a mixture importance distribution close to μ is constructed based on the (accepted) samples X_k in the MH algorithm. Schuster and Klebanov [35] follow a similar idea to the latter, but rather use the proposals Y_k of the MH algorithm and their asymptotic distribution as the importance distribution. Indeed, the idea of using all proposed states generated in the MH algorithm for estimating expectations such as $\mathbb{E}_\mu(f)$ is not new. For instance, Frenkel suggests in [11, 12] an approximation scheme which

[1] Surprisingly, the augmented MH Markov chain is in general not reversible but still has a stationary distribution.

recycles the rejected states in a MH algorithm. In the work of Delmas and Jourdain [8] this method is used in a control variate variance reduction approach and it is analyzed in a general framework. It turns out that for the Barker-algorithm the method is indeed beneficial, whereas for the MH-algorithm this is not necessarily the case. In particular, an estimator similar to $A_n(f)$ as above but for sampling from normalized densities was already introduced by Casella and Robert [5]. However, besides some numerical examples it was not further studied in [5] whereas their main focus, variance reduction of sampling methods by Rao-Blackwellization, got extended by [2, 9]. In particular, the theoretical results of Douc and Robert [9] provide variance reduction guarantees for their MH based estimator while keeping the additional computation cost under control. In contrast to that, using the estimator A_n does not increase the number of function evaluations, but we also do not provide a guarantee of improvement.

Outline. First, we provide some basic preliminaries on Markov chains and the corresponding classical MCMC estimator S_n. In Section 3 we introduce the MH importance sampling estimator, study properties of the aforementioned augmented MH Markov chain $(X_n, Y_n)_{n\in\mathbb{N}}$ and state the main results. In Section 4 we compare the classical MCMC estimator S_n with A_n numerically in two representative examples and draw some conclusions in Section 5.

2. Preliminaries on Markov chain Monte Carlo

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The random variables considered throughout the paper (mainly) map from this probability space to a measurable space $(G, \mathcal{G})$. A *(time-homogeneous) Markov chain* is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ which satisfy for any $A \in \mathcal{G}$ and any $n \in \mathbb{N}$ that $\mathbb{P}$-almost surely

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

where $K\colon G\times\mathcal{G} \to [0,1]$ denotes a *transition kernel*, i.e., $K(x,\cdot)$ is a probability measure for any $x \in G$ and the mapping $x \mapsto K(x,A)$ is measurable for any $A \in \mathcal{G}$. Our focus is on Markov chains designed for approximate sampling of the distribution μ. Such Markov chains typically have μ as their *stationary distribution*, i.e., their transition kernels K satisfy $\mu K = \mu$, where $\mu K(A) := \int_G K(x,A)\, \mu(\mathrm{d}x)$ for any $A \in \mathcal{G}$.

2.1. The Metropolis–Hastings algorithm

Let $P\colon G \times \mathcal{G} \to [0,1]$ be a *proposal transition kernel* satisfying the following structural assumption.

Assumption 2.1. For any $x \in G$ the proposal $P(x,\cdot)$ possesses a density $p(x,\cdot)$ w.r.t. μ_0 and for any $y \in G$ assume

$$\rho(y) > 0 \quad \Longrightarrow \quad p(x,y) > 0 \quad \forall x \in G.$$

This condition has some useful implications, see Proposition 2.1. Moreover, for example for $G \subseteq \mathbb{R}^d$, $\mathcal{G} = \mathcal{B}(G)$ and μ_0 being the Lebesgue measure, any Gaussian proposal, such as a Gaussian- or Langevin-random walk, satisfies it. Assumption 2.1 allows us to define the finite "acceptance ratio" $r(x, y)$ for the MH algorithm for any $x, y \in G$ according to [38, Section 2] by

$$r(x,y) := \begin{cases} \frac{\rho(y)p(y,x)}{\rho(x)p(x,y)} & \rho(x)p(x,y) > 0, \\ 1 & \text{otherwise.} \end{cases}$$

Then, the *MH algorithm*, which provides a realization of a Markov chain $(X_n)_{n\in\mathbb{N}}$, works as follows:

Algorithm 2.1. Assume that $X_n = x$, then the next state X_{n+1} is generated by the following steps:

1. Draw $Y_n \sim P(x, \cdot)$ and $U \sim \text{Unif}[0, 1]$ independently, call the result y and u, respectively.
2. Set $\alpha(x, y) := \min\{1, r(x, y)\}$.
3. Accept y with probability $\alpha(x, y)$, that is, if $u < \alpha(x, y)$, then set $X_{n+1} = y$, otherwise set $X_{n+1} = x$.

The Markov chain generated by the MH algorithm is called *MH Markov chain*, and its transition kernel, which we also call *MH (transition) kernel*, is given by

$$K(x, A) := \int_A \alpha(x, y)P(x, \mathrm{d}y) + \mathbf{1}_A(x) \int_G \alpha^c(x, y)P(x, \mathrm{d}y), \quad A \in \mathcal{G}, \tag{2.1}$$

where $\alpha^c(x, y) := 1 - \alpha(x, y)$. It is well-known that the transition kernel K in (2.1) is reversible w.r.t. μ, that is, $K(x, \mathrm{d}y)\mu(\mathrm{d}x) = K(y, \mathrm{d}x)\mu(\mathrm{d}y)$. In particular, this implies that μ is a stationary distribution of K.

2.2. Strong law of large numbers, central limit theorem and mean squared error bound

For convergence, in particular the strong law of large numbers, we need the concepts of ϕ-irreducibility and Harris recurrence: Given a σ-finite measure ϕ on $(G, \mathcal{G})$, a Markov chain $(X_n)_{n\in\mathbb{N}}$ is *ϕ-irreducible* if for each $A \in \mathcal{G}$ with $\phi(A) > 0$ and each $x \in G$ there exists an $n = n(x, A) \in \mathbb{N}$ such that $\mathbb{P}(X_n \in A \mid X_1 = x) > 0$. Furthermore, a Markov chain $(X_n)_{n\in\mathbb{N}}$ is *Harris recurrent* if it is ϕ-irreducible and satisfies for each $A \in \mathcal{G}$ with $\phi(A) > 0$ that for any $x \in G$

$$\mathbb{P}(X_n \in A \text{ infinitely often} \mid X_1 = x) = 1.$$

It is proven in [37, Corollary 2] that μ-irreducibility of a MH Markov chain $(X_n)_{n\in\mathbb{N}}$ implies Harris recurrence. Moreover, it is known that Assumption 2.1 ensures μ-irreducibility and, thus, Harris recurrence:

Proposition 2.1 ([25, Lemma 1.1]). *Given Assumption 2.1 the Markov chain $(X_n)_{n\in\mathbb{N}}$ realized by the MH algorithm is μ-irreducible.*

We recall the SLLN of the classical MCMC estimator $S_n(f)$ given in (1.2) based on the concept of Harris recurrence.

Theorem 2.1 (SLLN for S_n, [26, Theorem 17.0.1]). *Let $(X_n)_{n\in\mathbb{N}}$ be a Harris recurrent Markov chain with stationary distribution μ on G and let $f \in L^1(\mu)$. Then,*

$$S_n(f) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_\mu(f),$$

for any initial distribution, *i.e., any distribution of X_1.*

This theorem justifies that the classical MCMC method based on the MH algorithm yields a consistent estimator. Moreover, for $S_n(f)$ also a central limit theorem can be shown. Deriving a CLT is an important issue in studying MCMC and a lot of conditions which imply a CLT are known, for an overview we refer to the survey paper [16] and the references therein. We require some further terminology. Let $\mathrm{K}\colon L^2(\mu) \to L^2(\mu)$ be the *transition operator* associated to the transition kernel K of a Markov chain $(X_n)_{n\in\mathbb{N}}$ given by

$$(\mathrm{K}f)(x) := \int_G f(y)K(x,\mathrm{d}y), \quad f \in L^2(\mu).$$

For $n \geq 2$ and $f \in L^2(\mu)$ we have

$$\mathrm{K}^n f(x) = \int_G f(y)K^n(x,\mathrm{d}y),$$

where K^n is the n-step transition kernel, which is recursively defined by

$$K^n(x,A) := \int_G K(y,A)K^{n-1}(x,\mathrm{d}y), \quad A \in \mathcal{G}.$$

Note that the transition operator recovers the transition kernel, namely, for $n \geq 1$ we have

$$(\mathrm{K}^n \mathbf{1}_A)(x) = K^n(x,A), \qquad x \in G,\ A \in \mathcal{G}.$$

We also need the concept of the asymptotic variance: Let $(X_n^*)_{n\in\mathbb{N}}$ denote a Markov chain with transition kernel K starting at stationarity, i.e., the stationary distribution μ is also the initial one. Then, for $f \in L^2(\mu)$ the *asymptotic variance* of the classical MCMC estimator $S_n(f)$ for $\mathbb{E}_\mu(f)$ is given by

$$\sigma_S^2(f) := \lim_{n\to\infty} n \cdot \mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^n f(X_k^*)\right)$$

whenever the limit exists. One can easily see that the asymptotic variance admits the following representation in terms of the autocorrelation of the time series $(f(X_n^*))_{n\in\mathbb{N}}$. Namely,

$$\sigma_S^2(f) = \mathrm{Var}_\mu(f)\left(1 + 2\sum_{k=1}^\infty \mathrm{Corr}(f(X_1^*), f(X_{1+k}^*))\right), \tag{2.2}$$

where $\mathrm{Var}_\mu(f) := \mathbb{E}_\mu(f - \mathbb{E}_\mu(f))^2$ denotes the variance of f w.r.t. μ and $\mathrm{Corr}(\cdot,\cdot)$ the correlation between random variables.

Theorem 2.2 (CLT for S_n)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Harris recurrent Markov chain with transition kernel K and stationary distribution μ. For $f \in L^2(\mu)$, if either*

1. $\sum_{k=1}^{\infty} k^{-3/2} \left(\mathbb{E}_\mu[\sum_{j=0}^{k-1} \mathrm{K}^j(f - \mathbb{E}_\mu(f))]^2\right)^{1/2} < \infty$ *or*
2. *K is reversible w.r.t. μ and $\sigma_S^2(f) < \infty$,*

then we have for any initial distribution

$$\sqrt{n}(S_n(f) - \mathbb{E}_\mu(f)) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma_S^2(f))$$

with $\sigma_S^2(f)$ as in (2.2).

The theorem is justified by the following arguments. First, by [25, Proposition 17.1.6] it is sufficient to have a CLT when the initial distribution is a stationary one. In that case the Markov chain is an ergodic stationary process. Under condition 1., where no reversibility is necessary, the statement follows then by arguments derived in the introduction of [24]. Under condition 2. the statement follows based on [18, Corollary 1.5]. Although MH Markov chains are μ-reversible by construction, we encounter in the following a non-reversible Markov chain and derive a CLT by verifying condition 1.

The SLLN and the CLT only contain asymptotic statements, but one might be interested in explicit error bounds. For $f \in L^2(\mu)$ the *mean squared error* of the classical MCMC estimator $S_n(f)$ is given by $\mathbb{E}\,|S_n(f) - \mathbb{E}_\mu(f)|^2$. Depending on different convergence properties of the underlying Markov chain different error bounds are known, see for example [17, 20, 21, 31, 32, 33]. In particular, there is a relation between the asymptotic variance $\sigma_S^2(f)$ and the mean squared error of S_n: If $X_1 \sim \mu$, then

$$\lim_{n\to\infty} n \cdot \mathbb{E}\,|S_n(f) - \mathbb{E}_\mu(f)|^2 = \sigma_S^2(f),$$

and some of the error bounds have the same asymptotic behavior, see [20] and also [33].

3. The MH importance sampling estimator

The CLT for the MCMC estimator $S_n(f)$ shows that its statistical efficiency determined by the asymptotic variance $\sigma_S^2(f)$ is diminished by a large autocorrelation of $(f(X_n^*))_{n\in\mathbb{N}}$ or $(f(X_n))_{n\in\mathbb{N}}$, respectively. A reason for a large autocorrelation is the rejection of proposed states. In particular, the sequence of proposed states $(f(Y_n))_{n\in\mathbb{N}}$ is potentially less correlated than the MH Markov chain itself, since no rejection is involved. For example, if a proposal kernel P on $G = \mathbb{R}^d$ is absolutely continuous w.r.t. the Lebesgue measure, and $X_n \sim \mu$, then we have

$$0 = \mathbb{P}(Y_{n+1} = Y_n) \le \mathbb{P}(X_{n+1} = X_n) = \int_G \alpha^c(x,y)\, P(x, \mathrm{d}y)\ \mu(\mathrm{d}x).$$

Thus, one may ask whether it is beneficial, in terms of a higher statistical efficiency, to consider an estimator based on $(f(Y_n))_{n\in\mathbb{N}}$ rather than $(f(X_n))_{n\in\mathbb{N}}$. Such an estimator might be of the form

$$A_n(f) = \frac{\sum_{k=1}^n w_k f(Y_k)}{\sum_{k=1}^n w_k}$$

with suitable weights w_k. The reason for the latter is the fact that $Y_n \sim P(X_n,\cdot)$ does not follow the distribution μ. In fact, even if $X_n \sim \mu$, then $Y_n \sim \mu P$, hence, we need to apply an importance sampling correction in order to obtain a consistent estimator $A_n(f)$. To this end, Assumption 2.1 ensures the existence of:

$$\bar{\rho}(x,y) := Z\,\frac{\mathrm{d}\mu}{\mathrm{d}P(x,\cdot)}(y) \qquad \forall x,y\in G. \tag{3.1}$$

Indeed, by the fact that $p(x,y)=0$ implies $\rho(y)=0$ (Assumption 2.1) we have

$$\bar{\rho}(x,y) = \begin{cases} \rho(y)/p(x,y), & \rho(y)>0,\\ 0, & \rho(y)=0. \end{cases}$$

Moreover, the acceptance ratio $r(x,y)$ can be expressed only in terms of $\bar{\rho}$:

$$r(x,y) = \begin{cases} \frac{\bar{\rho}(y,x)}{\bar{\rho}(x,y)} & \bar{\rho}(x,y)>0,\\ 1 & \text{otherwise.} \end{cases}$$

As it turns out, $\bar{\rho}$ provides the correct weights w_k for an estimator $A_n(f)$, as indicated by the next result.

Proposition 3.1. *Let Assumption 2.1 be satisfied. Then, for any $f\in L^1(\mu)$, we have*

$$\mathbb{E}_\mu(f) = \frac{\int_G\int_G f(y)\bar{\rho}(x,y)P(x,\mathrm{d}y)\mu(\mathrm{d}x)}{\int_G\int_G \bar{\rho}(x,y)P(x,\mathrm{d}y)\mu(\mathrm{d}x)}$$

with $\bar{\rho}$ as in (3.1).

Proof. We have

$$\begin{aligned}\mathbb{E}_\mu(f) &= \int_G f(y)\frac{\bar{\rho}(x,y)}{Z}P(x,\mathrm{d}y) = \int_G\int_G f(y)\frac{\bar{\rho}(x,y)}{Z}P(x,\mathrm{d}y)\mu(\mathrm{d}x)\\ &= \frac{\int_G\int_G f(y)\bar{\rho}(x,y)P(x,\mathrm{d}y)\mu(\mathrm{d}x)}{\int_G\int_G \bar{\rho}(x,y)P(x,\mathrm{d}y)\mu(\mathrm{d}x)},\end{aligned}$$

where the last equality follows from

$$\begin{aligned}Z &= \int_G \rho(y)\mu_0(\mathrm{d}y) = \int_G \frac{\mathrm{d}\mu_0}{\mathrm{d}P(x,\cdot)}(y)\rho(y)P(x,\mathrm{d}y)\\ &= \int_G\int_G \frac{\mathrm{d}\mu_0}{\mathrm{d}P(x,\cdot)}(y)\rho(y)P(x,\mathrm{d}y)\mu(\mathrm{d}x)\\ &= \int_G\int_G \bar{\rho}(x,y)P(x,\mathrm{d}y)\mu(\mathrm{d}x).\end{aligned}$$

□

Proposition 3.1 motivates the following estimator.

Definition 3.1. Let Assumption 2.1 be satisfied and let $(X_n)_{n\in\mathbb{N}}$ be a MH Markov chain, where $(Y_n)_{n\in\mathbb{N}}$ denotes the corresponding proposal sequence. Then, given $f \in L^1(\mu)$, the *MH importance sampling estimator* for $\mathbb{E}_\mu(f)$ is

$$A_n(f) := \frac{\sum_{k=1}^n \bar{\rho}(X_k, Y_k) f(Y_k)}{\sum_{k=1}^n \bar{\rho}(X_k, Y_k)} \tag{3.2}$$

with $\bar{\rho}$ defined in (3.1).

Remark 3.1. *The dependence on ρ in A_n is explicitly given within $\bar{\rho}$, whereas the dependence on ρ of the classical estimator S_n realized with the MH algorithm is rather implicit. Namely, it appears only in the acceptance probability of the MH algorithm. However, in many situations the computational cost for function evaluations of ρ are much larger than for function evaluations of f, such that it seems counterintuitive to use the information of the value of ρ at the proposed state, which was expensive to compute, not any further.*

Remark 3.2. *The estimator $A_n(f)$ is related to self-normalizing importance sampling estimators for $\mathbb{E}_\mu(f)$ of the form*

$$\frac{\sum_{k=1}^n w_k f(\xi_k)}{\sum_{k=1}^n w_k},$$

where $(\xi_k)_{k\in\mathbb{N}}$ is an arbitrary sequence of random variables $\xi_k \sim \phi_k$ and where $w_k = \frac{\mathrm{d}\mu_0}{\mathrm{d}\phi_k}(\xi_k)\rho(\xi_k)$ are the corresponding importance weights. For $(\xi_k)_{k\in\mathbb{N}} = (Y_k)_{k\in\mathbb{N}}$ being the proposal sequence in the MH algorithm for realizing a μ-reversible Markov chain $(X_k)_{k\in\mathbb{N}}$, we recover $A_n(f)$ with $\phi_k = P(X_k, \cdot)$. In other words, $A_n(f)$ can be viewed as an importance sampling estimator where the importance distributions ϕ_k are determined by a MH Markov chain.

Remark 3.3. *Related to the previous remark we highlight a recent approach similar but slightly different to ours. Namely, the authors of [35] propose and study a self-normalizing importance sampler where the importance distribution is $\phi_k = \mu P$, i.e., the stationary distribution of the proposal sequence in the MH algorithm. Moreover, we remark that the particular form of the estimator $A_n(f)$ in the case of already normalized weights appeared in [5, Section 5], but without any further analysis. Since self-normalizing is rather inevitable in practice, we continue studying $A_n(f)$ as given in (3.2).*

3.1. The augmented MH Markov chain and its properties

In order to analyze the MH importance sampling estimator A_n we consider the *augmented MH Markov chain* $(X_n, Y_n)_{n\in\mathbb{N}}$ on $G \times G$ consisting of the original MH Markov chain $(X_n)_{n\in\mathbb{N}}$ and the associated sequence of proposals $(Y_n)_{n\in\mathbb{N}}$. The transition kernel K_{aug} of the augmented MH Markov chain is given by

$$K_{\mathrm{aug}}\left((x,y), \mathrm{d}u\mathrm{d}v\right) := \delta_y(\mathrm{d}u)P(y,\mathrm{d}v)\alpha(x,y) + \delta_x(\mathrm{d}u)P(x,\mathrm{d}v)\alpha^c(x,y)$$

for $x, y \in G$, where δ_z denotes the Dirac-measure at $z \in G$. Now we derive a useful representation of K_{aug} and the MH kernel K, which simplify several arguments. To this end, we define the probability measure

$$\nu(\mathrm{d}x\mathrm{d}y) := P(x, \mathrm{d}y)\mu(\mathrm{d}x) \tag{3.3}$$

on $(G \times G, \mathcal{G} \otimes \mathcal{G})$ and let $L^2(\nu)$ be the space of functions $g\colon G \times G \to \mathbb{R}$ which satisfy

$$\|g\|_\nu := \left(\int_{G\times G} |g(x,y)|^2 \nu(\mathrm{d}x\mathrm{d}y) \right)^{1/2} < \infty.$$

By K_{aug} the transition operator $\mathrm{K}_{\mathrm{aug}}\colon L^2(\nu) \to L^2(\nu)$ is induced. Furthermore, for a given proposal transition kernel P we define a linear operator $\widehat{\mathrm{P}}\colon L^2(\nu) \to L^2(\mu)$ by

$$(\widehat{\mathrm{P}}g)(x) := \int_G g(x,y) P(x, \mathrm{d}y).$$

It is easily seen that its adjoint operator $\widehat{\mathrm{P}}^*\colon L^2(\mu) \to L^2(\nu)$ is given by

$$(\widehat{\mathrm{P}}^* f)(x,y) = f(x),$$

i.e., $\langle \widehat{\mathrm{P}}g, f\rangle_\mu = \langle g, \widehat{\mathrm{P}}^* f\rangle_\nu$, where $\langle\cdot,\cdot\rangle_\mu$ and $\langle\cdot,\cdot\rangle_\nu$ denote the inner products in $L^2(\mu)$ and $L^2(\nu)$, respectively. Let H be the transition kernel on $G \times G$ given by

$$H((x,y), \mathrm{d}u\mathrm{d}v) := \alpha(x,y)\delta_{(y,x)}(\mathrm{d}u\mathrm{d}v) + \alpha^c(x,y)\delta_{(x,y)}(\mathrm{d}u\mathrm{d}v)$$

and let $\mathrm{H}\colon L^2(\nu) \to L^2(\nu)$ denote the associated transition operator. The following properties are useful for the subsequent analysis.

Lemma 3.1. *With the above notation we have that*

1. H *is self-adjoint and* $\|\mathrm{H}\|_{L^2(\nu)\to L^2(\nu)} = 1$*;*
2. $\widehat{\mathrm{P}}^*\widehat{\mathrm{P}}\colon L^2(\nu) \to L^2(\nu)$ *is a projection and*

$$\|\widehat{\mathrm{P}}\|_{L^2(\nu)\to L^2(\mu)} = \|\widehat{\mathrm{P}}^*\|_{L^2(\mu)\to L^2(\nu)} = 1;$$

3. $\mathrm{K} = \widehat{\mathrm{P}}\mathrm{H}\widehat{\mathrm{P}}^*$ *and* $\mathrm{K}_{aug} = \mathrm{H}\widehat{\mathrm{P}}^*\widehat{\mathrm{P}}$*;*
4. ν *given in* (3.3) *is a stationary distribution of* K_{aug}*;*
5. $\mathrm{K}_{aug}^n = \mathrm{H}\widehat{\mathrm{P}}^*\mathrm{K}^{n-1}\widehat{\mathrm{P}}$ *and* $\mathrm{K}^n = \widehat{\mathrm{P}}\mathrm{K}_{aug}^{n-1}\mathrm{H}\widehat{\mathrm{P}}^*$ *for* $n \geq 2$*.*

Proof. **To 1.:** Let $g_1, g_2 \in L^2(\nu)$. Then, by the choice of $\alpha(x,y)$ we have $\alpha(x,y)\nu(\mathrm{d}x\mathrm{d}y) = \alpha(y,x)\nu(\mathrm{d}y\mathrm{d}x)$, and self-adjointness follows from

$$\begin{aligned}
\langle \mathrm{H}g_1, g_2\rangle_\nu &= \int_{G\times G} (\alpha(x,y)g_1(y,x) + \alpha^c(x,y)g_1(x,y))g_2(x,y)\nu(\mathrm{d}x\mathrm{d}y) \\
&= \int_{G\times G} g_1(y,x)g_2(x,y)\alpha(x,y)\nu(\mathrm{d}x\mathrm{d}y) \\
&\quad + \int_{G\times G} \alpha^c(x,y)g_1(x,y)g_2(x,y)\nu(\mathrm{d}x\mathrm{d}y)
\end{aligned}$$

$$
\begin{aligned}
&= \int_{G\times G} g_1(x,y)g_2(y,x)\alpha(y,x)\nu(\mathrm{d}x\mathrm{d}y) \\
&\quad + \int_{G\times G} \alpha^c(x,y)g_1(x,y)g_2(x,y)\nu(\mathrm{d}x\mathrm{d}y) \\
&= \langle g_1, \mathrm{H}g_2\rangle_\nu.
\end{aligned}
$$

Since H is induced by the transition kernel H the operator norm is one.

To 2.: It is easily seen that $\widehat{\mathrm{P}}^*\widehat{\mathrm{P}}$ is a projection. Moreover, it is well-known that the norm of an operator and its adjoint coincide, which yields the statement in combination with

$$
1 = \left\|\widehat{\mathrm{P}}^*\widehat{\mathrm{P}}\right\|_{L^2(\nu)\to L^2(\nu)} = \|\widehat{\mathrm{P}}\|_{L^2(\nu)\to L^2(\mu)}.
$$

To 3.: The representations can be verified by a straightforward calculation.

To 4.: For any $A, B \in \mathcal{G}$ we have

$$
\begin{aligned}
\nu K_{\mathrm{aug}}(A\times B) &= \int_{G^2} (\mathrm{H}\widehat{\mathrm{P}}^*\widehat{\mathrm{P}}\mathbf{1}_{A\times B})(x,y)P(x,\mathrm{d}y)\mu(\mathrm{d}x) \\
&= \int_G (\widehat{\mathrm{P}}\mathrm{H}\widehat{\mathrm{P}}^*\widehat{\mathrm{P}}\mathbf{1}_{A\times B})(x)\mu(\mathrm{d}x) \\
&= \int_G (\mathrm{K}\widehat{\mathrm{P}}\mathbf{1}_{A\times B})(x)\mu(\mathrm{d}x) \\
&= \int_G (\widehat{\mathrm{P}}\mathbf{1}_{A\times B})(x)\mu(\mathrm{d}x) = \nu(A\times B),
\end{aligned}
$$

where the last-but-one equality follows from the fact that μ is a stationary distribution of K. Since the Cartesian products $A \times B$ provide a generating system of $\mathcal{G}\otimes\mathcal{G}$ the result follows by the uniqueness theorem of probability measures.

To 5.: These representations are a direct consequence of 3. □

Note that statement 5 of Lemma 3.1 yields for $n \geq 1$ and $g \in L^2(\nu)$ that

$$
\begin{aligned}
(K_{\mathrm{aug}}^n\, g)(x,y) = \alpha(x,y) &\int_{G^2} g(u,v)P(u,\mathrm{d}v)K^{n-1}(y,\mathrm{d}u) \\
&+ \alpha^c(x,y)\int_{G^2} g(u,v)P(u,\mathrm{d}v)K^{n-1}(x,\mathrm{d}u).
\end{aligned} \tag{3.4}
$$

Remark 3.4. *In general, the transition kernel K_{aug} is not reversible w.r.t. ν. Since reversibility is equivalent to self-adjointness of the Markov operator this can be seen by the fact that $K_{\mathrm{aug}}^* = \widehat{P}^*\widehat{P}H$, which does not necessarily coincides with K_{aug}. For convenience of the reader we also provide a simple example which illustrates the non-reversibility. Consider a finite state space $G = \{1,2\}$ equipped with the counting measure μ_0 with $\rho(i) = 1/2$ and $P(i,j) = 1/2$ for all $i,j \in G$ such that $\alpha(i,j) = 1$. Then the transition matrix K_{aug} is given by*

$$
K_{\mathrm{aug}}((i,j),(k,\ell)) = \frac{\delta_j(\{k\})}{2}
$$

for any $i,j,k,\ell \in G$. Here reversibility is equivalent to $K_{aug}((i,j),(k,\ell)) = K_{aug}((k,\ell),(i,j))$ for all $i,j,k,\ell \in G$, which is not satisfied for $i=j=\ell=1$ and $k=2$.

Now, using Lemma 3.1 we show that stability properties of the MH kernel K pass over to K_{aug}. The proof of the following result is adapted from [39, Lemma 24].

Lemma 3.2. *Assume that ϕ is a σ-finite measure on $(G,\mathcal{G})$ and let K denote the MH kernel as in (2.1).*

- *If K is ϕ-irreducible, then K_{aug} is ϕ_P-irreducible on $G\times G$, where the σ-finite measure ϕ_P is given by $\phi_P(\mathrm{d}x\mathrm{d}y) := P(x,\mathrm{d}y)\phi(\mathrm{d}x)$.*
- *If K is Harris recurrent (w.r.t. ϕ), then K_{aug} is also Harris recurrent (w.r.t. ϕ_P).*

Proof. For $A \in \mathcal{G}\otimes\mathcal{G}$ and $x\in G$ define

$$A_2(x) := \{y\in G : (x,y)\in A\} \in \mathcal{G},$$
$$A_1 := \{x\in G : A_2(x)\neq \emptyset\} \in \mathcal{G},$$

so that $A_2(x)$ is the slice of A for fixed first component x and A_1 is the "projection" of the set A on the first component space. For $\varepsilon>0$ let

$$A_1(\varepsilon) := \{x\in G : P(x,A_2(x)) > \varepsilon\}.$$

By the use of (3.4) we prove the irreducibility statement: Assume that $A\in\mathcal{G}\otimes\mathcal{G}$ with $\phi_P(A)>0$. Then, $\phi(A_1)>0$, since otherwise

$$\phi_P(A) = \int_A P(x,\mathrm{d}y)\phi(\mathrm{d}x) = \int_{A_1} P(x,A_2(x))\,\phi(\mathrm{d}x)$$

is zero. By the same argument, one obtains that there exists an $\varepsilon>0$ such that $\phi(A_1(\varepsilon))>0$, since otherwise

$$\phi_P(A) = \int_{\bigcup_{\varepsilon>0} A_1(\varepsilon)} P(x,A(x))\,\phi(\mathrm{d}x)$$

is zero. Because of the ϕ-irreducibility of K, we have for $x,y\in G$ that there exist $n_x,n_y\in\mathbb{N}$ such that $K^{n_x}(x,A_1(\varepsilon))>0$ and $K^{n_y}(y,A_1(\varepsilon))>0$. Hence, if $\alpha(x,y)>0$, then

$$\begin{aligned} K_{\mathrm{aug}}^{n_y+1}((x,y),A) &\overset{(3.4)}{\geq} \alpha(x,y)\int_A P(u,\mathrm{d}v)K^{n_y}(y,\mathrm{d}u)\\ &= \alpha(x,y)\int_{A_1} P(u,A_2(u))K^{n_y}(y,\mathrm{d}u)\\ &\geq \alpha(x,y)\int_{A_1(\varepsilon)} P(u,A_2(u))K^{n_y}(y,\mathrm{d}u)\\ &\geq \alpha(x,y)\,\varepsilon\,K^{n_y}(y,A_1(\varepsilon)) > 0. \end{aligned}$$

Otherwise, if $\alpha^c(x,y) = 1$, we obtain analogously

$$K_{\text{aug}}^{n_x+1}((x,y),A) \geq \alpha^c(x,y) \int_A P(u,\mathrm{d}v) K^{n_x}(x,\mathrm{d}u) \geq \varepsilon\, K^{n_x}(x, A_1(\varepsilon)) > 0.$$

In other words, for $(x,y) \in G \times G$ we find an $n \in \mathbb{N}$ (depending on $\alpha(x,y)$) such that $K_{\text{aug}}^n((x,y),A) > 0$, which proves the ϕ_P-irreducibility.

We turn to the Harris recurrence: Let K be Harris recurrent w.r.t. ϕ and let $\phi_P(A) > 0$. As above, we can conclude that there exists an $\varepsilon > 0$ such that $\phi(A_1(\varepsilon)) > 0$. Furthermore, for the augmented Markov chain $(X_n, Y_n)_{n\in\mathbb{N}}$ with transition kernel K_{aug} we have

$$\mathbb{P}\left((X_n,Y_n) \in A\right) = \mathbb{P}\left(Y_n \in A_2(X_n)\right) = P(X_n, A_2(X_n)).$$

By $\phi(A_1(\varepsilon)) > 0$ and the fact that $(X_n)_{n\in\mathbb{N}}$ is Harris recurrent w.r.t. ϕ, with probability one there are infinitely many distinct times $(\tau_k)_{k\in\mathbb{N}}$, such that $X_{\tau_k} \in A_1(\varepsilon)$ for any $k \in \mathbb{N}$. Hence

$$\begin{aligned}\mathbb{P}\left(\sum_{n=1}^{\infty} \mathbf{1}_A(X_n,Y_n) = \infty\right) &= \mathbb{P}\left(\sum_{n=1}^{\infty} \mathbf{1}_{A_2(X_n)}(Y_n) = \infty\right) \\ &\geq \mathbb{P}\left(\sum_{k=1}^{\infty} \mathbf{1}_{A_2(X_{\tau_k})}(Y_{\tau_k}) = \infty\right).\end{aligned}$$

Note that by construction $\mathbf{1}_{A_2(X_{\tau_k})}(Y_{\tau_k})$ are Bernoulli random variables with success probability of at least ε. Moreover, they are conditionally independent given $(X_{\tau_k})_{k\in\mathbb{N}}$. Hence,

$$\mathbb{P}\left(\sum_{k=1}^{\infty} \mathbf{1}_{A_2(X_{\tau_k})}(Y_{\tau_k}) = \infty \,\middle|\, (X_{\tau_k})_{k\in\mathbb{N}}\right) = 1 \qquad \mathbb{P}\text{-a.s.}$$

yields

$$\mathbb{P}\left(\sum_{k=1}^{\infty} \mathbf{1}_{A_2(X_{\tau_k})}(Y_{\tau_k}) = \infty\right) = \mathbb{E}\left[\mathbb{P}\left(\sum_{k=1}^{\infty} \mathbf{1}_{A_2(X_{\tau_k})}(Y_{\tau_k}) = \infty \,\middle|\, (X_{\tau_k})_{k\in\mathbb{N}}\right)\right] = 1,$$

which shows that the augmented MH Markov chain is Harris recurrent. □

Remark 3.5. *Another consequence of Lemma 3.1 interesting on its own is that also geometric ergodicity is inherited by the augmented MH Markov chain. However, since this fact is not relevant for the remainder of the paper, we postpone the discussion of geometric ergodicity and its inheritance to Appendix A.*

3.2. Strong law of large numbers and central limit theorem

A consistency statement in form of a SLLN of the MH importance sampling estimator defined in (3.2) is stated and proven in the following. A key argument in the proofs is the inheritance of Harris recurrence of $(X_n)_{n\in\mathbb{N}}$ to the augmented MH Markov chain $(X_n, Y_n)_{n\in\mathbb{N}}$.

Theorem 3.1. *Let Assumption 2.1 be satisfied. Then, for any initial distribution and any $f \in L^1(\mu)$ we have*

$$A_n(f) = \frac{\frac{1}{n}\sum_{k=1}^n \bar{\rho}(X_k, Y_k) f(Y_k)}{\frac{1}{n}\sum_{k=1}^n \bar{\rho}(X_k, Y_k)} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_\mu(f). \tag{3.5}$$

Proof. Assumption 2.1 implies μ-irreducibility and Harris recurrence of the MH Markov chain $(X_n)_{n\in\mathbb{N}}$ due to Proposition 2.1. This yields, due to Lemma 3.2, that also the transition kernel K_{aug} is Harris recurrent. Hence, by Theorem 2.1 we have for each $h \in L^1(\nu)$ that

$$\frac{1}{n}\sum_{k=1}^n h(X_k, Y_k) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}_\nu(h).$$

Define $h_1(x,y) := \bar{\rho}(x,y)f(y)$ and $h_2(x,y) := \bar{\rho}(x,y)$. Since $\mathbb{E}_\nu(h_2) = Z < \infty$ and $\mathbb{E}_\nu(h_1) = \mathbb{E}_\mu(f) \cdot Z < \infty$, we have $h_1, h_2 \in L^1(\nu)$ and, thus, the numerator and denominator on the left-hand side of (3.5) converge a.s. to $\mathbb{E}_\nu(h_1)$ and $\mathbb{E}_\nu(h_2)$. The assertion follows then by the continuous mapping theorem and $\mathbb{E}_\nu(h_1)/\mathbb{E}_\nu(h_2) = \mathbb{E}_\mu(f)$. □

The next goal is to derive a CLT, which provides a way to quantify the asymptotic behavior of A_n. Since the augmented Markov chain $(X_n, Y_n)_{n\in\mathbb{N}}$ is, in general, not reversible w.r.t. ν, we aim to use condition 1 of Theorem 2.2.

Theorem 3.2. *Let Assumption 2.1 be satisfied and assume for $f \in L^1(\mu)$ that*

$$\sigma_A^2(f) := \int_G \int_G (f(y) - \mathbb{E}_\mu(f))^2 \frac{\mathrm{d}\mu}{\mathrm{d}P(x,\cdot)}(y)\, \mu(\mathrm{d}y)\mu(\mathrm{d}x)$$

is finite. Then, for any initial distribution, we have

$$\sqrt{n}(A_n(f) - \mathbb{E}_\mu(f)) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma_A^2(f)).$$

Proof. We frequently use the identity

$$\int_G g(x,y)\bar{\rho}(x,y)P(x,\mathrm{d}y) = Z\int_G g(x,y)\mu(\mathrm{d}y), \tag{3.6}$$

for any $x \in G$ and any $g\colon G^2 \to \mathbb{R}$ for which one of the two integrals exist. Define the centered version of f by $f_c(y) := f(y) - \mathbb{E}_\mu(f)$ and set $h_3(x,y) := \bar{\rho}(x,y)f_c(y)$ for $x, y \in G$. Note that $\mathbb{E}_\nu(h_3) = 0$ and $h_3 \in L^2(\nu)$, since

$$\begin{aligned}
\mathbb{E}_\nu(h_3^2) &= \int_G \int_G f_c(y)^2 \bar{\rho}(x,y)^2 P(x,\mathrm{d}y)\mu(\mathrm{d}x) \\
&\overset{(3.6)}{=} Z\int_G \int_G f_c(y)^2 \bar{\rho}(x,y)\mu(\mathrm{d}y)\mu(\mathrm{d}x) \\
&= Z^2 \int_G \int_G f_c(y)^2 \frac{\mathrm{d}\mu}{\mathrm{d}P(x,\cdot)}(y)\mu(\mathrm{d}y)\mu(\mathrm{d}x) = Z^2\sigma_A^2(f) < \infty.
\end{aligned}$$

With the representation (3.4) one obtains for any $k \geq 2$ that

$$\begin{aligned}
\mathrm{K}_{\mathrm{aug}}^k h_3(x,y) &= \int_{G\times G} \bar{\rho}(u,v) f_c(v) K_{\mathrm{aug}}^k(x,y,\mathrm{d}u\,\mathrm{d}v) \\
&= \alpha(x,y) \int_G \int_G \bar{\rho}(u,v) f_c(v) P(u,\mathrm{d}v) K^{k-1}(y,\mathrm{d}u) \\
&\quad + \alpha^c(x,y) \int_G \int_G \bar{\rho}(u,v) f_c(v) P(u,\mathrm{d}v) K^{k-1}(x,\mathrm{d}u) \\
&= 0,
\end{aligned}$$

where the last equality follows from

$$\int_G f_c(v)\bar{\rho}(u,v) P(u,\mathrm{d}v) \overset{(3.6)}{=} Z\, \mathbb{E}_\mu(f_c) = 0 \qquad \forall u \in G.$$

By the same argument we obtain $\mathrm{K}_{\mathrm{aug}} h_3 = 0$. Hence, for the augmented MH Markov chain $(X_n, Y_n)_{n\in\mathbb{N}}$ condition 1. of Theorem 2.2 is satisfied for the function h_3 and by the inheritance of the Harris recurrence from K to K_{aug}, see Lemma 3.2, we get

$$\frac{1}{\sqrt{n}} \sum_{k=1}^n h_3(X_k, Y_k) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma_S^2(h_3)).$$

Here

$$\sigma_S^2(h_3) = \mathrm{Var}(h_3(X_1,Y_1)) + 2\sum_{k=1}^\infty \mathrm{Cov}(h_3(X_1,Y_1), h_3(X_{k+1},Y_{k+1})).$$

By exploiting again the fact that $\mathrm{K}_{\mathrm{aug}}^k h_3 = 0$ for $k \geq 1$ we obtain

$$\mathrm{Cov}(h_3(X_1,Y_1), h_3(X_{k+1},Y_{k+1})) = \int_{G\times G} (\mathrm{K}_{\mathrm{aug}}^k h_3)(x,y) h_3(x,y) \nu(\mathrm{d}x\mathrm{d}y) = 0,$$

such that

$$\sigma_S^2(h_3) = \mathrm{Var}(h_3(X_1,Y_1)) = Z^2 \sigma_A^2(f).$$

Further,

$$\sqrt{n}(A_n(f) - \mathbb{E}_\mu(f)) = \frac{n^{-1/2}\sum_{j=1}^n h_3(X_j,Y_j)}{\frac{1}{n}\sum_{j=1}^n \bar{\rho}(X_j,Y_j)}.$$

The denominator converges by Theorem 2.1 to Z as well as

$$n^{-1/2} \sum_{k=1}^n h_3(X_k,Y_k) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, Z^2\sigma_A^2(f)),$$

such that by Slutsky's Theorem the assertion is proven. □

Remark 3.6. *It is remarkable that the asymptotic variance $\sigma_A^2(f)$ of $A_n(f)$ coincides with the asymptotic variance of the importance sampling estimator*

$$\frac{\sum_{k=1}^n \bar{\rho}(X_k,Y_k)f(Y_k)}{\sum_{k=1}^n \bar{\rho}(X_k,Y_k)}$$

given independent *random variables $(X_k,Y_k) \sim \nu$ for $k \in \mathbb{N}$, see [1, Section 2.3.1] or [27, Section 9.2]. Here, ν denotes the stationary measure of the augmented MH Markov chain given in (3.3). Hence, the fact that $A_n(f)$ is based on the, in general, dependent sequence $(X_k,Y_k)_{k\in\mathbb{N}}$ of the augmented MH Markov chain, does surprisingly not effect its asymptotic variance.*

Remark 3.7. *Often it is of interest to estimate the asymptotic variance appearing in a CLT. For a given $f \in L^1(\mu)$ the corresponding quantity, given by Theorem 3.2, can be rewritten as*

$$\sigma_A^2(f) = \frac{\int_{G\times G}(f(y)-\mathbb{E}_\mu(f))^2\bar{\rho}(x,y)^2P(x,\mathrm{d}y)\mu(\mathrm{d}x)}{\left(\int_{G\times G}\bar{\rho}(x,y)P(x,\mathrm{d}y)\mu(\mathrm{d}x)\right)^2}.$$

Given this representation of $\sigma_A^2(f)$ we suggest estimating it by

$$\frac{n\cdot\sum_{k=1}^n\left[f(Y_k)-\frac{1}{n}\sum_{j=1}^n f(X_j)\right]^2\bar{\rho}(X_k,Y_k)^2}{\left(\sum_{k=1}^n\bar{\rho}(X_k,Y_k)\right)^2}$$

where $\frac{1}{n}\sum_{j=1}^n f(X_j)$ can also be replaced by $A_n(f)$.

Now we turn to a non-asymptotic analysis, where the error criterion is the mean squared error.

3.3. Mean squared error bound

In this section we provide explicit bounds for the mean squared error of A_n. Those estimates are an immediate consequence of the following two lemmas, which are similar to the arguments in [23, Theorem 2] and [1, Theorem 2.1].

Lemma 3.3. *Let $(X_n,Y_n)_{n\in\mathbb{N}}$ denote an augmented MH Markov chain. For $f\colon G\to\mathbb{R}$ define*

$$D(f) := \int_{G\times G} f(y)\bar{\rho}(x,y)P(x,\mathrm{d}y)\mu(\mathrm{d}x),$$
$$D_n(f) := \frac{1}{n}\sum_{j=1}^n \bar{\rho}(X_j,Y_j)f(Y_j).$$

Then, for bounded f, i.e., $\|f\|_\infty := \sup_{x\in G}|f(x)| < \infty$, we have

$$\mathbb{E}\,|A_n(f)-\mathbb{E}_\mu(f)|^2 \le \frac{2}{D(1)^2}\left(\|f\|_\infty^2\,\mathbb{E}\,|D(1)-D_n(1)|^2 + \mathbb{E}\,|D_n(f)-D(f)|^2\right).$$

Proof. Observe that $D(1) = Z$. Further

$$\begin{aligned}\mathbb{E}\left|A_n(f) - \mathbb{E}_\mu(f)\right|^2 &= \mathbb{E}\left|\frac{D_n(f)}{D_n(1)} - \frac{D(f)}{Z}\right|^2 \\ &= \mathbb{E}\left|\frac{D_n(f)}{D_n(1)} - \frac{D_n(f)}{Z} + \frac{D_n(f)}{Z} - \frac{D(f)}{Z}\right|^2.\end{aligned}$$

Using the fact that $(a+b)^2 \le 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$ gives

$$\begin{aligned}\mathbb{E}\left|A_n(f) - \mathbb{E}_\mu(f)\right|^2 &\le 2\mathbb{E}\left|\frac{D_n(f)}{D_n(1)} - \frac{D_n(f)}{Z}\right|^2 + 2\mathbb{E}\left|\frac{D_n(f)}{Z} - \frac{D(f)}{Z}\right|^2 \\ &= \frac{2}{Z^2}\mathbb{E}\left|\frac{D_n(f)}{D_n(1)}\left(D_n(1) - Z\right)\right|^2 + \frac{2\mathbb{E}\left|D_n(f) - D(f)\right|^2}{Z^2} \\ &\le \frac{2}{Z^2}\left(\|f\|_\infty^2\,\mathbb{E}\left|D_n(1) - Z\right|^2 + \mathbb{E}\left|D_n(f) - D(f)\right|^2\right).\end{aligned}$$

□

Lemma 3.4. *Assume that the initial distribution is the stationary one, that is, $X_1 \sim \mu$. Then, with the notation from Lemma 3.3, we have*

$$n \cdot \mathbb{E}\left|D_n(f) - D(f)\right|^2 = \int_{G^2} f(y)^2 \bar{\rho}(x,y)^2 P(x, \mathrm{d}y)\mu(\mathrm{d}x) - Z^2 \mathbb{E}_\mu(f)^2.$$

Proof. Observe that

$$D(f) = \int_G \int_G f(y)\bar{\rho}(x,y)P(x,\mathrm{d}y)\mu(\mathrm{d}x) \overset{(3.6)}{=} Z \cdot \mathbb{E}_\mu(f).$$

Define the centered function $g_c(x,y) := \bar{\rho}(x,y)f(y) - Z \cdot \mathbb{E}_\mu(f)$ for any $x, y \in G$. We have

$$\begin{aligned}\mathbb{E}\left|D_n(f) - D(f)\right|^2 &= \frac{1}{n^2}\sum_{j=1}^{n} \mathbb{E}\left[g_c(X_j, Y_j)^2\right] \\ &\quad + \frac{2}{n^2}\sum_{j=1}^{n-1}\sum_{i=j+1}^{n} \mathbb{E}\left[g_c(X_i, Y_i)g_c(X_j, Y_j)\right].\end{aligned}$$

Exploiting the fact that the initial distribution is the stationary one we obtain for $i \ge j$ that

$$\mathbb{E}\left[g_c(X_i, Y_i)g_c(X_j, Y_j)\right] = \int_{G\times G} g_c(x,y)\left(K_{\mathrm{aug}}^{i-j} g_c\right)(x,y)P(x,\mathrm{d}y)\mu(\mathrm{d}x).$$

In the case $k := i - j > 1$ we have by representation (3.4) that

$$\begin{aligned}\mathrm{K}_{\mathrm{aug}}^k g_c(x,y) &= \alpha(x,y)\int_G\int_G g_c(u,v)P(u,\mathrm{d}v)K^{k-1}(y,\mathrm{d}u) \\ &\quad + \alpha^c(x,y)\int_G\int_G g_c(u,v)P(u,\mathrm{d}v)K^{k-1}(x,\mathrm{d}u)\end{aligned}$$

and

$$\int_G g_c(u,v)P(u,\mathrm{d}v) = \int_G f(v)\bar{\rho}(u,v)P(u,\mathrm{d}v) - Z\cdot\mathbb{E}_\mu(f) \overset{(3.6)}{=} 0$$

leads to $\mathrm{K}_{\mathrm{aug}}^k g_c(x,y) = 0$. By similar arguments we obtain $\mathrm{K}_{\mathrm{aug}} g_c(x,y) = 0$. Hence,

$$\begin{aligned}\mathbb{E}\left|D_n(f) - D(f)\right|^2 &= \frac{1}{n}\mathbb{E}\left[g_c(X_1,Y_1)^2\right] \\ &= \frac{1}{n}\left(\int_{G^2} f(y)^2\bar{\rho}(x,y)^2 P(x,\mathrm{d}y)\mu(\mathrm{d}x) - Z^2\mathbb{E}_\mu(f)^2\right).\end{aligned}$$

□

By the combination of both lemmas we derive the following theorem.

Theorem 3.3. *Assume that the initial distribution of an augmented MH Markov chain $(X_n,Y_n)_{n\in\mathbb{N}}$ is the stationary one, i.e., $X_1 \sim \mu$. Then, for bounded $f\colon G\to\mathbb{R}$ we obtain*

$$\mathbb{E}\left|A_n(f) - \mathbb{E}_\mu(f)\right|^2 \le \frac{4}{n}\left\|f\right\|_\infty^2 \int_{G\times G} \frac{\mathrm{d}\mu}{\mathrm{d}P(x,\cdot)}(y)\mu(\mathrm{d}y)\mu(\mathrm{d}x).$$

Remark 3.8. *Let us mention here two things: First, we assumed that the initial distribution is the stationary one. This assumption is certainly restrictive, we refer to [20, 33] for techniques to derive explicit error bounds for more general initial distribution. Second, the factor*

$$4\left\|f\right\|_\infty^2 \int_{G\times G} \frac{\mathrm{d}\mu}{\mathrm{d}P(x,\cdot)}(y)\mu(\mathrm{d}y)\mu(\mathrm{d}x)$$

in the estimate is an upper bound of the asymptotic variance $\sigma_A^2(f)$ derived in Theorem 3.2. We conjecture that the estimate actually holds with $\sigma_A^2(f)$ instead of this upper bound.

3.4. Optimal calibration of proposals

Given the explicit expression for the asymptotic variance $\sigma_A^2(f)$ involving the proposal kernel P, we can ask for an optimal choice of the kernel $P\colon G\times\mathcal{G}\to[0,1]$ in order to minimize $\sigma_A^2(f)$. However, finding an optimal kernel among all admissible kernels is, in general, an infeasible task. In practice, one often considers common types of proposal kernels $P = P_s$ with a tunable step size parameter $s>0$ and ask for the optimal value of s. For example, given a measure μ on $G\subseteq\mathbb{R}^d$, we can use the *random walk* proposal

$$P_s(x,\cdot) = \mathcal{N}(x, s^2C), \qquad s>0, \tag{3.7}$$

where $x\in\mathbb{R}^d$ and $C\in\mathbb{R}^{d\times d}$ denotes a covariance matrix, within a MH algorithm. For this proposal and the classical path average estimator $S_n(f)$ it is

widely known that a good step size $s_S^\star$ is chosen in such a way that the average acceptance rate is

$$\int_G \alpha(x,y) P_{s_S^\star}(x,\mathrm{d}y)\mu(\mathrm{d}x) \approx 0.234.$$

For a justification and further details we refer to [29]. For the MH importance sampling estimator $A_n(f)$ we look for an optimal step size $s_A^\star$. Optimal in the sense that it minimizes the asymptotic variance of $A_n(f)$, thus, we ask for

$$s_A^\star := \operatorname*{argmin}_{s>0} V(s), \quad V(s) := \int_G \int_G (f(y) - \mathbb{E}_\mu(f))^2 \frac{\rho(y)}{p_s(x,y)}\, \mu(\mathrm{d}y)\mu(\mathrm{d}x),$$

where $p_s(x,\cdot)$ denotes the density of $P_s(x,\cdot)$ w.r.t. the reference measure μ_0. If we assume that the mapping $s \mapsto p_s(x,y)$ is differentiable for each $(x,y) \in G\times G$ with derivative $\frac{\mathrm{d}}{\mathrm{d}s}p_s(x,y)$, then any s minimizing $V(s)$ satisfies

$$0 = \frac{\mathrm{d}}{\mathrm{d}s}V(s) = \int_G \int_G (f(y) - \mathbb{E}_\mu(f))^2 \rho(y) \frac{\frac{\mathrm{d}}{\mathrm{d}s}p_s(x,y)}{p_s^2(x,y)}\, \mu(\mathrm{d}y)\mu(\mathrm{d}x). \tag{3.8}$$

By the fact that $\mu(\mathrm{d}y)\mu(\mathrm{d}x) \propto \bar\rho_s(x,y)\, P_s(x,\mathrm{d}y)\mu(\mathrm{d}x)$, where $\bar\rho_s(x,y) = \frac{\rho(y)}{p_s(x,y)}$, we can rewrite (3.8) and approximate $\frac{\mathrm{d}}{\mathrm{d}s}V(s)$ by using (X_k,Y_k), $k=1,\dots,n$, from the augmented Markov chain. Thus

$$\begin{aligned} 0 &= \int_G \int_G (f(y) - \mathbb{E}_\mu(f))^2\, \bar\rho_s^2(x,y)\, \frac{\frac{\mathrm{d}}{\mathrm{d}s}p_s(x,y)}{p_s(x,y)}\, P_s(x,\mathrm{d}y)\mu(\mathrm{d}x) \\ &\approx \frac{1}{n}\sum_{k=1}^n \left(f(Y_k) - \frac{1}{n}\sum_{j=1}^n f(X_j) \right)^2 \bar\rho_s^2(X_k,Y_k) \frac{\frac{\mathrm{d}}{\mathrm{d}s}p_s(X_k,Y_k)}{p_s(X_k,Y_k)}. \end{aligned}$$

In practice we can calibrate s such that the empirical average on the right-hand side is close to zero. We demonstrate the feasibility of this approach for two common proposals.

Example 3.1 (Optimal calibration of the random walk-MH)**.** *We consider μ_0 as the Lebesgue measure on $G \subseteq \mathbb{R}^d$ and P_s as in (3.7). Thus,*

$$p_s(x,y) = \frac{1}{s^d\sqrt{\det(2\pi C)}} \exp\left(-\frac{\|y-x\|_C^2}{2s^2} \right)$$

where $\|y-x\|_C^2 := (y-x)^\top C^{-1}(y-x)$, and

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s}p_s(x,y) &= \left(-ds^{-d-1} + s^{-d-3}\|y-x\|_C^2\right) \frac{\exp\left(-\frac{\|y-x\|_C^2}{2s^2}\right)}{\sqrt{\det(2\pi C)}} \\ &= \left(-ds^{-1} + s^{-3}\|y-x\|_C^2\right) p_s(x,y). \end{aligned}$$

Hence, the necessary condition (3.8) *boils down to*

$$0 = \int_G \int_G (f(y) - \mathbb{E}_\mu(f))^2 \ \bar{\rho}_s^2(x,y) \ \left(-ds^{-1} + s^{-3}\|y-x\|_C^2\right) \ P_s(x,\mathrm{d}y)\mu(\mathrm{d}x),$$

which can be rewritten as

$$s^2 = \frac{\int_G \int_G (f(y) - \mathbb{E}_\mu(f))^2 \ \bar{\rho}_s^2(x,y) \ \|y-x\|_C^2 \, P_s(x,\mathrm{d}y)\mu(\mathrm{d}x)}{d \int_G \int_G (f(y) - \mathbb{E}_\mu(f))^2 \ \bar{\rho}_s^2(x,y) \, P_s(x,\mathrm{d}y)\mu(\mathrm{d}x)}.$$

In practice, we then can seek an $s_\star > 0$ *such that for* n *states* (X_k, Y_k) *of the augmented MH Markov chain generated by the proposal* $P_{s_\star}(x,\cdot) = \mathcal{N}(x, s_\star^2 C)$ *we have*

$$s_\star^2 \approx \frac{\sum_{k=1}^n \left(f(Y_k) - \frac{1}{n}\sum_{j=1}^n f(X_j)\right)^2 \bar{\rho}_{s_\star}^2(X_k,Y_k) \ \|Y_k - X_k\|_C^2}{d \sum_{k=1}^n \left(f(Y_k) - \frac{1}{n}\sum_{j=1}^n f(X_j)\right)^2 \bar{\rho}_{s_\star}^2(X_k,Y_k)}. \tag{3.9}$$

Example 3.2 (Optimal calibration of the MALA)**.** *Another common proposal on* $G = \mathbb{R}^d$ *is the one of the* Metropolis-adjusted Langevin algorithm (MALA)*, given by*

$$P_s(x,\cdot) = \mathcal{N}\left(x + \frac{s^2}{2}\nabla \log \rho(x), s^2 I_d\right), \tag{3.10}$$

where we assume that $\log \rho \colon G \to \mathbb{R}$ *is differentiable and* I_d *denotes the identity matrix in* $\mathbb{R}^d$*. The resulting proposal density is*

$$p_s(x,y) = \frac{1}{s^d (2\pi)^{d/2}} \exp\left(-\frac{\|y - m_s(x)\|^2}{2s^2}\right),$$

with $m_s(x) := x + \frac{s^2}{2}\nabla \log \rho(x)$*. In order to compute the derivative* $\frac{\mathrm{d}}{\mathrm{d}s} p_s(x,y)$ *we first obtain*

$$\frac{\mathrm{d}}{\mathrm{d}s}\|y - m_s(x)\|^2 = -2s(y - m_s(x))^\top \nabla \log \rho(x),$$

which then yields

$$\frac{\mathrm{d}}{\mathrm{d}s} p_s(x,y) = \left(-ds^{-1} + s^{-3}\|y - m_s(x)\|^2 + s^{-1}(y - m_s(x))^\top \nabla \log \rho(x)\right) \ p_s(x,y).$$

Thus, in the case of MALA the necessary condition (3.8) *is equivalent to*

$$s^2 = \frac{\int_G \int_G (f(y) - \mathbb{E}_\mu(f))^2 \ \bar{\rho}_s^2(x,y) \ \|y - m_s(x)\|^2 \, P_s(x,\mathrm{d}y)\mu(\mathrm{d}x)}{\int_G \int_G (f(y) - \mathbb{E}_\mu(f))^2 \, \bar{\rho}_s^2(x,y) \, [d - (y - m_s(x))^\top \nabla \log \rho(x)] \, P_s(x,\mathrm{d}y)\mu(\mathrm{d}x)}.$$

Again, in practice we seek for an $s_\star > 0$ *such that given* n *states* (X_k, Y_k) *of the augmented MH Markov chain generated by the MALA proposal* $P_{s_\star}$ *in* (3.10) *we have* $s_\star^2$ *close to*

$$\frac{\sum_{k=1}^n \left(f(Y_k) - \frac{1}{n}\sum_{j=1}^n f(X_j)\right)^2 \bar{\rho}_{s_\star}^2(X_k,Y_k) \ \|Y_k - m_{s_\star}(X_k)\|^2}{\sum_{k=1}^n \left(f(Y_k) - \frac{1}{n}\sum_{j=1}^n f(X_j)\right)^2 \bar{\rho}_{s_\star}^2(X_k,Y_k) \, [d - (Y_k - m_{s_\star}(X_k))^\top \nabla \log \rho(X_k)]}. \tag{3.11}$$

4. Numerical examples

We want to illustrate the benefits as well as the limitations of the MH importance sampling estimator $A_n(f)$ at two simple but representative examples. To this end, we compare the considered $A_n(f)$ to the classical path average estimator $S_n(f)$ as well as to two other established estimators using also the proposed states Y_k generated in the MH algorithm. Namely

- the *waste-recycling Monte Carlo estimator*, for further details we refer to [11, 12, 8], given by

$$WR_n(f) := \sum_{k=1}^{n} (1 - \alpha(X_k, Y_k))\, f(X_k) + \alpha(X_k, Y_k) f(Y_k);$$

- another Markov chain importance sampling estimator also based on the proposed states, see [35], given by

$$B_n(f) := \frac{\sum_{k=1}^{n} \widetilde{w}_n(X_{1:n}, Y_k) f(Y_k)}{\sum_{k=1}^{n} \widetilde{w}_n(X_{1:n}, Y_k)}, \qquad \widetilde{w}_n(X_{1:n}, Y_k) := \frac{\rho(Y_k)}{\sum_{j=1}^{n} p(X_j, Y_k)}.$$

The notation $X_{1:n}$ within $B_n(f)$ stands for $X_1, \dots, X_n$. In the following we provide two comments w.r.t. B_n and the other estimators.

Remark 4.1. *For the convenience of the reader we justify heuristically that $B_n(f)$ approximates $\mathbb{E}_\mu(f)$. For this let ν_Y be the marginal distribution of the stationary probability measure ν on $G \times G$ of the augmented Markov chain $(X_k, Y_k)_{k\in\mathbb{N}}$, that is,*

$$\nu_Y(\mathrm{d}y) := \int_G \nu(\mathrm{d}x\mathrm{d}y) = \int_G P(x, \mathrm{d}y)\mu(\mathrm{d}x).$$

Intuitively, ν_Y can be considered as the asymptotic distribution of the proposed states. The empirically computed weights $\widetilde{w}_n(X_{1:n}, Y_k)$ in $B_n(f)$ approximate importance sampling weights $\widetilde{w}(Y_k) \propto \frac{\mathrm{d}\mu}{\mathrm{d}\nu_Y}(Y_k)$ resulting from the asymptotic distribution ν_Y of the proposed states Y_k. Now if we substitute $\widetilde{w}_n(X_{1:n}, Y_k)$ within $B_n(f)$ by $\widetilde{w}(Y_k)$ we have an importance sampling estimator based on the proposed states which approximates $\mathbb{E}_\mu(f)$.

Remark 4.2. *By the fact that within the estimators $S_n(f)$, $A_n(f)$, and $WR_n(f)$ only one or two separate sums over $k = 1, \dots, n$ appear, the number of arithmetic operations and therefore the complexity is $\mathcal{O}(n)$. In contrast to that, within the alternative Markov chain importance sampling estimator $B_n(f)$ an additional summation of complexity $\mathcal{O}(n)$ is required for the computation of each weight $\widetilde{w}_n(X_{1:n}, Y_k)$, $k = 1, \dots, n$, such that the overall number of arithmetic operations is $\mathcal{O}(n^2)$ for $B_n(f)$. To take this into account, we often compare the former three estimators to $B_{\sqrt{n}}(f)$. Besides that an optimal tuning of the proposal step size of the estimator $B_n(f)$ is left open in [35], however, the authors suggest to simply use the usual calibration rule for the classical path average estimator $S_n(f)$ from [29].*

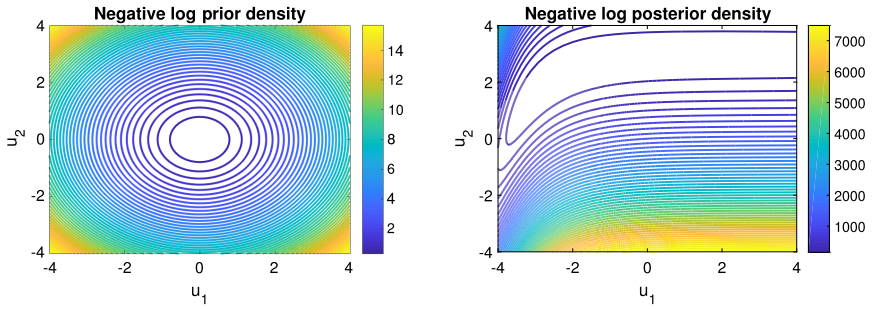

FIG 1. *Contour plot of the normal prior and the resulting posterior density for the example of Section 4.1.*

4.1. Bayesian inference for a differential equation

We consider a boundary value problem in one spatial dimension $x \in [0,1]$ which serves as a simple model for, e.g., stationary groundwater flow:

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(\exp(u_1)\,\frac{\mathrm{d}}{\mathrm{d}x}p(x)\right) = 1, \quad p(0) = 0,\ p(1) = u_2. \tag{4.1}$$

Here, the unknown parameters $u = (u_1, u_2)$ involving the log-diffusion coefficient u_1 and the Dirichlet data u_2 at the righthand boundary $x = 1$ shall be inferred given noisy observations $y \in \mathbb{R}^2$ of the solution p at $x_1 = 0.25$ and $x_2 = 0.75$. This inference setting has been already applied as a test case for sampling and filtering methods in [10, 13, 14]. We place a Gaussian prior on $u = (u_1, u_2)$, namely, $\mu_0 \sim N(0, I_2)$ where I_2 denotes the identity matrix in $\mathbb{R}^2$. The observation vector is given by $y = (27.5, 79.7)$ and we assume an additive measurement noise $\varepsilon \sim N(0, 0.01 I_2)$, i.e., the likelihood $L(y|u)$ of observing y given a fixed value $u \in \mathbb{R}^2$ is

$$L(y|u) := \frac{100}{2\pi}\exp\left(-\frac{100}{2}\|y - F(u)\|^2\right)$$

where $\|\cdot\|$ denotes the Euclidean norm and $F\colon \mathbb{R}^2 \to \mathbb{R}^2$ the mapping $(u_1, u_2) \mapsto (p(x_1), p(x_2))$ with

$$p(x) = u_2 x + \frac{\exp(-u_1)}{2}(x - x^2), \qquad x \in [0,1].$$

The resulting posterior measure for u given the observation y follows then the form (1.1) with $\rho(u) := L(y|u)$. The negative log prior and posterior density are presented in Figure 1.

For approximate sampling of the posterior μ we apply now the random walk-MH algorithm and MALA, see Section 3.4, with various values of the step size s. We let the Markov chains run for $n = 10^4$ iterations after a burn-in of $n_0 = 10^3$ iterations. Then, we use the generated path of the (augmented) MH Markov

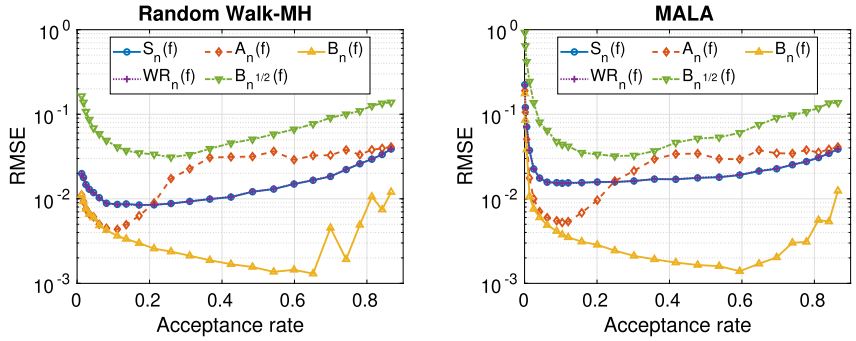

FIG 2. *RMSE for computing the posterior mean w.r.t. average acceptance rate for the example of Section 4.1.*

chain in order to estimate the posterior mean $\mathbb{E}_\mu(\mathbf{f})$ where $\mathbf{f}(u) := (f_1(u), f_2(u))$ with $f_i(u) := u_i$. We approximate $\mathbb{E}_\mu(\mathbf{f})$ by the various estimators $E_n(\mathbf{f})$ discussed at the begining of this section, that is,

$$E_n(\mathbf{f}) \in \{S_n(\mathbf{f}), A_n(\mathbf{f}), B_n(\mathbf{f}), B_{\sqrt{n}}(\mathbf{f}), WR_n(\mathbf{f})\}.$$

The true value $\mathbb{E}_\mu(\mathbf{f})$ of the posterior mean is computed by Gauss quadrature employing 1500 Gauss–Hermite nodes in each dimension which ensures a quadrature error smaller than 10^{-4}. For each choice of the step size s we run $M = 1,200$ independent Markov chains and, thus, compute M realizations of the estimators $S_n(\mathbf{f})$, $A_n(\mathbf{f})$, $B_n(\mathbf{f})$, $B_{\sqrt{n}}(\mathbf{f})$, and $WR_n(\mathbf{f})$, respectively. We use these M realizations in order to empirically estimate the root mean squared error (RMSE)

$$\mathrm{RMSE}_{E_n}(\mathbf{f}) := \left(\mathbb{E}\,\|E_n(\mathbf{f}) - \mathbb{E}_\mu(\mathbf{f})\|^2\right)^{1/2},$$

of the various estimators $E_n(\mathbf{f}) \in \{S_n(\mathbf{f}), A_n(\mathbf{f}), B_n(\mathbf{f}), B_{\sqrt{n}}(\mathbf{f}), WR_n(\mathbf{f})\}$ for each chosen step size s of the proposal kernels. The results are displayed in Figure 2 and Figure 3, respectively.[2]

Comparison of $A_n(\mathbf{f})$ to $S_n(\mathbf{f})$: In Figure 2 we observe that for a certain range of s, the MH importance sampling estimator $A_n(\mathbf{f})$ provides a significant error reduction for both proposal kernels, the random-walk and MALA. In particular, the global minimum for the error of $A_n(\mathbf{f})$ is smaller than for $S_n(\mathbf{f})$. In fact, it is roughly half the size for both proposals. Hence, given the optimal step size s the MH importance sampling method can indeed outperform the classical path average estimator. In this example, we could reduce the RMSE by 50% without

[2]We note that in each setting the squared norm of the bias of the estimators $E_n(f_i)$ is roughly the same size as their variance, i.e., the magnitude of the displayed RMSE coincides basically with $\sqrt{2}$ times the standard deviation of the corresponding estimator. For a larger sample size n the percentage of the bias in the RMSE would have decreased, however, the computation of $B_n(\mathbf{f})$ would have become unfeasible.

a significant additional cost. We comment below on how to find this optimal step size for $A_n(\mathbf{f})$.

Comparison of $A_n(\mathbf{f})$ to other estimators: We observe in Figure 2 that the waste recycling estimator $WR_n(\mathbf{f})$ basically coincides with the classical path average estimator $S_n(\mathbf{f})$ for both proposals and all chosen step sizes s, i.e., it yields no improvement and is outperformed by $A_n(\mathbf{f})$. Concerning the Markov chain importance sampling estimators $B_n(\mathbf{f})$ we obtain a further improvement on $A_n(\mathbf{f})$ and nearly can reduce the RMSE by an order of magnitude compared to $S_n(\mathbf{f})$. However, this performance comes at the price of a signicant larger complexity. If we consider the Markov chain importance sampling estimators $B_{\sqrt{n}}(\mathbf{f})$ with the same complexity as the other estimators $S_n(\mathbf{f})$, $A_n(\mathbf{f})$, and $WR_n(\mathbf{f})$, we in fact observe a worse performance to the other estimators for all chosen step sizes. Thus, in the error-vs-complexity sense the estimator $A_n(\mathbf{f})$ performs best among all considered estimators if calibrated correctly.

Optimal calibration of $A_n(\mathbf{f})$: Concerning the optimal step size for $A_n(\mathbf{f})$ we present in Figure 3 a verification of the approach outlined in Section 3.4. For both MH algorithms, the random walk-MH and MALA, we display in the top row the RMSE of $S_n(\mathbf{f})$ and $A_n(\mathbf{f})$ w.r.t. the chosen step sizes. In the bottom row we display for each step size value s the relation of s^2 to the empirical functionals

$$J_{\mathbf{f}}(s) := \frac{\sum_{k=1}^n \left\| \mathbf{f}(Y_k) - \frac{1}{n}\sum_{j=1}^n \mathbf{f}(X_j) \right\|^2 \bar{\rho}_s^2(X_k, Y_k)\ \|Y_k - X_k\|_C^2}{d \sum_{k=1}^n \left\| \mathbf{f}(Y_k) - \frac{1}{n}\sum_{j=1}^n \mathbf{f}(X_j) \right\|^2 \bar{\rho}_s^2(X_k, Y_k)}$$

and

$$J(s) := \frac{\sum_{k=1}^n \bar{\rho}_s^2(X_k, Y_k)\ \|Y_k - X_k\|_C^2}{d \sum_{k=1}^n \bar{\rho}_s^2(X_k, Y_k)}$$

for the random walk-MH and the corresponding $J_{\mathbf{f}}(s)$ and $J(s)$ for MALA based on (3.11). In Section 3.4 we derived as a necessary condition for the optimal step size $s_\star$ that $s_\star^2 \approx J_{\mathbf{f}}(s_\star)$ for both kind of proposals. Here, we can indeed verify this condition: the optimal $s_\star$, which was calibrated by hand following the rule $s_\star^2 \approx J_{\mathbf{f}}(s_\star)$, shows indeed also the smallest RMSE in the top row. The optimal $s_\star$, $J_{\mathbf{f}}(s_\star)$, and its RMSE are highlighted by a green marker in Figure 3. Besides that, choosing the rather "objective" functional $J(s)$, which is independent of the particular quantity of interest $\mathbf{f}$, and apply the alternative calibration rule $s_\star^2 \approx J(s_\star)$ does not yield to a step size with minimal RMSE for $A_n(\mathbf{f})$ — although the alternatively calibrated step size and the resulting RMSE are not that far off from the true optimum. In summary, Figure 3 verifies that the approach in Section 3.4 can indeed be applied in practice for finding the optimal step size for the MH importance sampling estimator $A_n(\mathbf{f})$.

4.2. Bayesian inference for probit regression (PIMA data)

The second example is a test problem for logistic regression, see, e.g., [7] for a discussion. Here, nine predictors $x_i \in \mathbb{R}^9$ such as diastolic blood pressure, body

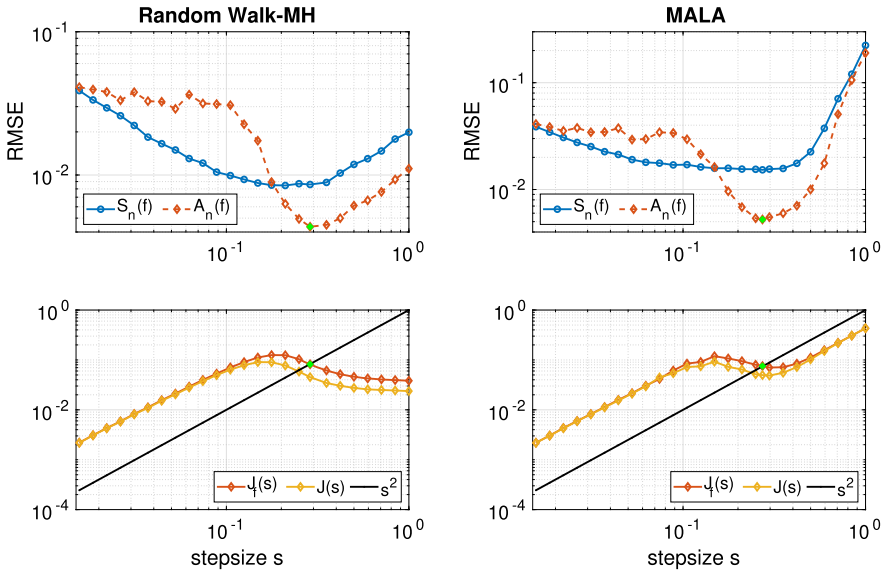

FIG 3. *RMSE for mean w.r.t. step size s for the example of Section 4.1.*

mass index, or age are fitted to the binary outcome $y_i \in \{-1, 1\}$ for diagnosing diabetes for $N = 768$ members $i = 1, \dots, N$, of the Pima Indian tribe. For more details about the data we refer to [36]. Following [7] the likelihood $L(y|\beta)$ for the outcome $y \in \{-1, 1\}^N$ of the diagnosis is modeled by

$$L(y|\beta) := \prod_{i=1}^{N} \Phi(y_i \beta^\top x_i),$$

where Φ denotes the cumulative distribution function of a univariate standard normal distribution and $\beta \in \mathbb{R}^9$ the unknown regression coefficients (including the intercept). Moreover, we take independent Gaussian priors for each component of β as suggested in [7], i.e., the prior is $\mu_0 = \mathcal{N}(0, \Lambda)$ where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_9)$ with $\lambda_1 = 20$ and $\lambda_i = 5$ for $i \geq 2$. Given the data set $(x_i, y_i)_{i=1}^N \in \mathbb{R}^{10 \times N}$ the resulting posterior for β is of the form (1.1) with $\mu_0 = \mathcal{N}(0, \Lambda)$ and

$$\rho(\beta) := \prod_{i=1}^{N} \Phi(y_i \beta^\top x_i).$$

For this example we test the performance of the MH importance sampling estimator in several dimensions $d = 2, \dots, 9$. To this end, we modify the regression model for each d by setting $\beta = (\beta_1, \dots, \beta_d, 0, \dots, 0) \in \mathbb{R}^9$ and only infer the values of the components β_i for $i = 1, \dots, d$. Hence, the posterior from which we would like to sample is a measure on $\mathbb{R}^d$, $d = 2, \dots, 9$. For each $d = 2, \dots, 9$ we perform the same simulations as in the first example, i.e.,

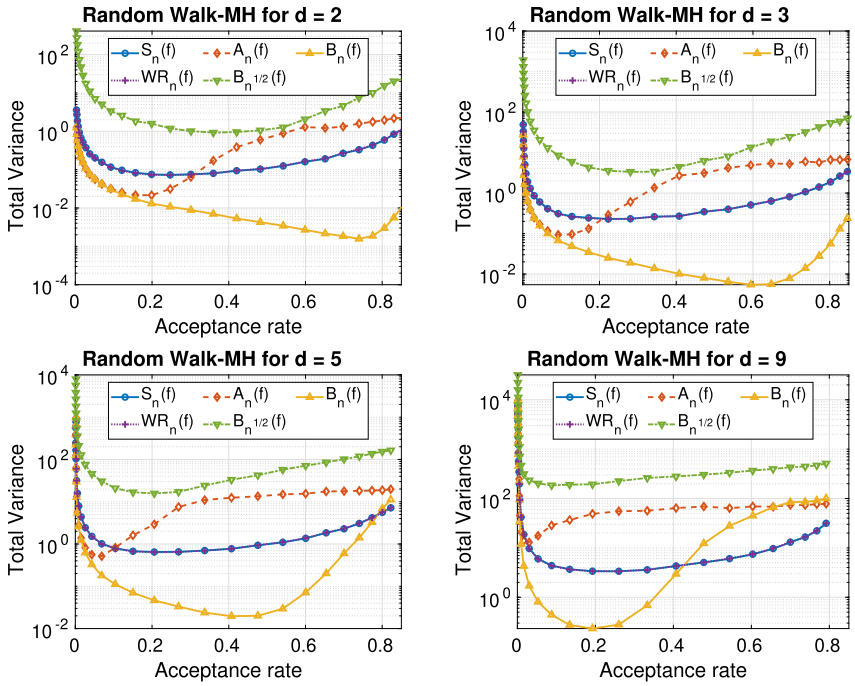

FIG 4. *Total variances of estimators w.r.t. average acceptance rate in various dimensions for the example from Section 4.2.*

we generate Markov chains by the MH algorithm using the Gaussian random walk proposal from Section 4.1 with varying step size parameter s. Then, we compute the estimates $E_n(\mathbf{f}) := (E_n(f_i))_{i=1,\dots,d}$ for $\mathbf{f}(\beta) = (f_i(\beta))_{i=1,\dots,d}$ with $f_i(\beta) = \beta_i$ where $E_n(\mathbf{f})$ is again a placeholder for the particular estimator at choice, i.e., $E_n(\mathbf{f}) \in \{S_n(\mathbf{f}), A_n(\mathbf{f}), B_n(\mathbf{f}), WR_n(\mathbf{f}), B_{\sqrt{n}}(\mathbf{f})\}$ as in Section 4.1. For each choice of the step size s we repeat this procedure $M = 1,200$ times and use the results to compute empirical estimates for the total variance of the estimators $E_n(\mathbf{f})$ given by

$$\mathrm{Var}(E_n(\mathbf{f})) := \sum_{i=1}^{d} \mathrm{Var}(E_n(f_i)).$$

The number of iterations of the MH Markov chain as well as the burn-in length are the same as in Section 4.1. In Figure 4 we present for several choices of d the resulting plots for the total variance of the estimators w.r.t. average acceptance rate in the MH algorithm, similar to Figure 2 in the previous section. Furthermore, Table 1 displays the ratios of the total variances $\mathrm{Var}(E_n(\mathbf{f}))/\mathrm{Var}(S_n(\mathbf{f}))$ for various estimators $E_n(\mathbf{f}))$ (each one with its optimal step size) in dimensions $d = 2, \dots, 9$. Our results can be summarized as follows.

Performance of $A_n(\mathbf{f})$: For small dimensions, like $d = 2, \dots, 5$, we observe

TABLE 1
Ratio of the minimal total variances $\mathrm{Var}(E_n(\mathbf{f}))/\mathrm{Var}(S_n(\mathbf{f}))$ *for various estimators* E_n *and dimensions* d *for the example of Section 4.2.*

Dimension d	$\frac{\mathrm{Var}(A_n(\mathbf{f}))}{\mathrm{Var}(S_n(\mathbf{f}))}$	$\frac{\mathrm{Var}(B_n(\mathbf{f}))}{\mathrm{Var}(S_n(\mathbf{f}))}$	$\frac{\mathrm{Var}(WR_n(\mathbf{f}))}{\mathrm{Var}(S_n(\mathbf{f}))}$	$\frac{\mathrm{Var}(B_{\sqrt{n}}(\mathbf{f}))}{\mathrm{Var}(S_n(\mathbf{f}))}$
2	0.30	0.02	0.98	12.68
3	0.41	0.02	0.98	14.56
4	0.59	0.03	0.99	18.63
5	0.79	0.03	0.99	24.19
6	1.40	0.03	0.99	31.98
7	2.11	0.04	0.99	43.59
8	2.98	0.06	0.99	48.84
9	3.86	0.07	1.00	55.10

that the minimal total variance of $A_n(\mathbf{f})$ is smaller or at most as large as the minimal total variance of $S_n(\mathbf{f})$, see also Table 1. However, for dimensions $d \geq 6$ the MH importance sampling estimator $A_n(\mathbf{f})$ shows a higher total variance than the classical path average estimator $S_n(\mathbf{f})$. In particular, we observe in Table 1 that the performance of $A_n(\mathbf{f})$ compared to $S_n(\mathbf{f})$ seems to decline more and more for increasing dimension.

Performance of other estimators: As in Section 4.1 the waste recycling estimator $WR_n(\mathbf{f})$ basically coincides with the path average estimator $S_n(\mathbf{f})$ for any considered dimension d. Also for the Markov chain importance sampling estimators $B_n(\mathbf{f})$ and $B_{\sqrt{n}}(\mathbf{f})$ the performance compared to $S_n(\mathbf{f})$ and $A_n(\mathbf{f})$ is similar to Section 4.1, i.e., $B_n(\mathbf{f})$ outperformes all other estimators — but at a higher cost — whereas its cost-equivalent version $B_{\sqrt{n}}(\mathbf{f})$ performes worse than any other estimator. However, also $B_n(\mathbf{f})$ and $B_{\sqrt{n}}(\mathbf{f})$ seem to suffer from higher dimensions of the state space as indicated in Table 1, i.e., their total variance relative to the total variance of $S_n(\mathbf{f})$ becomes larger as d increases.

Optimal calibration: We observe that for $S_n(\mathbf{f})$ the total variance becomes minimal for average acceptance rates between 0.2 and 0.25. This is in accordance with the well-known asymptotic result on optimal a-priori step size choices, see [29]. The same optimal calibration holds true for the waste recycling estimator $WR_n(\mathbf{f})$. For MH importance sampling estimator $A_n(\mathbf{f})$ the minimal total variance is obtained for smaller and smaller average acceptance rates as the dimension d increases. In fact, the numerical results indicate that the optimal proposal step size s for $A_n(\mathbf{f})$ remains constant w.r.t. the dimension d. This is in contrast to the classical MCMC estimator where the optimal asymptotic a-priori step size s behaves for a product density ρ like d^{-1} for the Gaussian random walk proposal, see [29]. Moreover, for each dimension d the minimal total variances of $A_n(\mathbf{f})$ where obtained for step sizes satisfying the optimal calibration rules outlined in Section 3.4. Concerning the estimators $B_n(\mathbf{f})$ and $B_{\sqrt{n}}(\mathbf{f})$ we also observe that the optimal performance occurs for decreasing acceptance rates as the dimension d increases. Here, the numerical results suggest that the optimal proposal step size s even increases mildly with the dimensions d.

5. Conclusion

In this work we studied a MH importance sampling estimator A_n for which we showed a SLLN, a CLT, and an explicit estimate of the mean squared error. A remarkable property of this estimator is that its asymptotic variance does not contain any autocorrelation term, in fact

$$\operatorname{Corr}(\bar{\rho}(X_k, Y_k)f(Y_k), \bar{\rho}(X_m, Y_m)f(Y_m)) = \delta_k(\{m\}).$$

This is in sharp contrast to the asymptotic variance of the classical MCMC estimator S_n, see (2.2). Additionally, we performed numerical experiments which indicate that the MH importance sampling estimator can outperform the classical one. This requires the correct tuning of the underlying MH Markov chain in terms of the proposal step size where the estimator A_n seems to benefit from rather small average acceptance rates in contrast to optimal scaling results for the MCMC estimator. However, we exhibit a decreasing efficiency of the MH importance sampling estimator for increasing dimension in the numerical experiments. Indeed, the classical MCMC estimator performs better for larger dimensions. This is very likely related to the well-known degeneration of efficiency for importance sampling in high dimensions, see for example the discussion [1, Section 2.5.4].

Appendix A: Inheritance of geometric ergodicity

A transition kernel $K\colon G \times \mathcal{G} \to [0,1]$ with stationary distribution μ is $L^2(\mu)$-*geometrically ergodic* if there exists a constant $r \in [0,1)$ such that for all probability measures η on G with $\frac{\mathrm{d}\eta}{\mathrm{d}\mu} \in L^2(\mu)$ there is $C_\eta \in [0,\infty)$ satisfying

$$d_{\mathrm{TV}}(\mu, \eta K^n) \le C_\eta\, r^n \qquad \forall n \in \mathbb{N}, \tag{A.1}$$

where d_{TV} denotes the total variation distance. Note that if $\frac{\mathrm{d}\eta}{\mathrm{d}\mu}$ exists, then

$$d_{\mathrm{TV}}(\mu,\eta) := \sup_{A\in\mathcal{G}} |\mu(A) - \eta(A)| = \frac{1}{2}\int_G \left|\frac{\mathrm{d}\eta}{\mathrm{d}\mu}(x) - 1\right| \mu(\mathrm{d}x).$$

In addition to the exponential convergence, $L^2(\mu)$-geometric ergodicity also yields advantages concerning the CLT for the classical MCMC estimator $S_n(f)$ for $\mathbb{E}_\mu(f)$.

Proposition A.1 ([28, Corollary 2.1])**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with μ-reversible, $L^2(\mu)$-geometrically ergodic transition kernel. Then, for $f \in L^2(\mu)$ we have $\sigma_S^2(f) < \infty$ and $\sqrt{n}(S_n(f) - \mathbb{E}_\mu(f)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_S^2(f))$ as $n \to \infty$.*

A further important aspect, see e.g. [28], is the relation between $L^2(\mu)$-geometric ergodicity of a μ-reversible transition kernel K and spectral properties of the associated self-adjoint transition operator. To this end, we introduce $L_0^2(\mu)$ as the space of all $g \in L^2(\mu)$ satisfying $\mathbb{E}_\mu(g) = 0$.

Proposition A.2 ([28, Theorem 2.1], [19, Proposition 1.5])**.** *Let the transition kernel* $K\colon G\times\mathcal{G}\to[0,1]$ *be* μ*-reversible. Then,* K *is* $L^2(\mu)$*-geometrically ergodic if and only if*

$$\|\mathrm{K}\|_{L_0^2(\mu)\to L_0^2(\mu)}<1. \tag{A.2}$$

The condition (A.2) is often referred to as the existence of a positive $L^2(\mu)$-*spectral gap* of K:

$$\mathrm{gap}_\mu(\mathrm{K}):=1-\|\mathrm{K}\|_{L_0^2(\mu)\to L_0^2(\mu)}>0.$$

By Lemma 3.1 we obtain easily the following relation between the norms of $\mathrm{K}\colon L_0^2(\mu)\to L_0^2(\mu)$ and the corresponding operator $\mathrm{K}_{\mathrm{aug}}\colon L_0^2(\nu)\to L_0^2(\nu)$ of the augmented MH Markov chain.

Lemma A.1. *With the same notation introduced in Section 3.1 we have that*

1. $\|\mathrm{K}^n\|_{L_0^2(\mu)\to L_0^2(\mu)}\le\left\|\mathrm{K}_{aug}^{n-1}\right\|_{L_0^2(\nu)\to L_0^2(\nu)}$ *and*

$$\left\|\mathrm{K}_{aug}^{n}\right\|_{L_0^2(\nu)\to L_0^2(\nu)}\le\left\|\mathrm{K}^{n-1}\right\|_{L_0^2(\mu)\to L_0^2(\mu)}$$

   for $n\ge 2$*;*
2. $\|\mathrm{K}\|_{L_0^2(\mu)\to L_0^2(\mu)}\le\|\mathrm{K}_{aug}\|_{L_0^2(\nu)\to L_0^2(\nu)}$ *and the spectrum of* K_{aug} *is non-negative and real as well as the spectral radius* $r(\mathrm{K}_{aug}\mid L_0^2(\nu))$ *of* K_{aug} *on* $L_0^2(\nu)$ *satisfies*

$$r(\mathrm{K}_{aug}\mid L_0^2(\nu))\le\|\mathrm{K}\|_{L_0^2(\mu)\to L_0^2(\mu)}.$$

Proof. **To 1.:** Note that $\widehat{\mathrm{P}}^*f\in L_0^2(\nu)$, $\mathrm{H}g\in L^2(\nu)$ and $\widehat{\mathrm{P}}g\in L_0^2(\mu)$ for any $f\in L_0^2(\mu)$ and $g\in L_0^2(\nu)$. By applying Lemma 3.1 we have

$$\|\mathrm{K}^n\|_{L_0^2(\mu)\to L_0^2(\mu)}=\left\|\widehat{\mathrm{P}}\mathrm{K}_{\mathrm{aug}}^{n-1}\mathrm{H}\widehat{\mathrm{P}}^*\right\|_{L_0^2(\mu)\to L_0^2(\mu)}\le\left\|\mathrm{K}_{\mathrm{aug}}^{n-1}\right\|_{L_0^2(\nu)\to L_0^2(\nu)},$$

since

$$\left\|\widehat{\mathrm{P}}\right\|_{L_0^2(\nu)\to L_0^2(\mu)}\le\left\|\widehat{\mathrm{P}}\right\|_{L^2(\nu)\to L^2(\mu)}=1\quad\text{and}\quad\left\|\mathrm{H}\widehat{\mathrm{P}}^*\right\|_{L_0^2(\mu)\to L_0^2(\nu)}\le 1.$$

Similarly

$$\left\|\mathrm{K}_{\mathrm{aug}}^{n}\right\|_{L_0^2(\nu)\to L_0^2(\nu)}=\left\|\mathrm{H}\widehat{\mathrm{P}}^*\mathrm{K}^{n-1}\widehat{\mathrm{P}}\right\|_{L_0^2(\nu)\to L_0^2(\nu)}\le\left\|\mathrm{K}^{n-1}\right\|_{L_0^2(\mu)\to L_0^2(\mu)}.$$

To 2.: By the fact that $\mathrm{K}\colon L_0^2(\mu)\to L_0^2(\mu)$ is self-adjoint, properties of the spectral radius formula for self-adjoint operators and statement 1 we have

$$\begin{aligned}\|\mathrm{K}\|_{L_0^2(\mu)\to L_0^2(\mu)}&=\lim_{n\to\infty}(\|\mathrm{K}^n\|_{L_0^2(\mu)\to L_0^2(\mu)})^{1/n}\le\lim_{n\to\infty}(\left\|\mathrm{K}_{\mathrm{aug}}^{n-1}\right\|_{L_0^2(\nu)\to L_0^2(\nu)})^{1/n}\\&=\|\mathrm{K}_{\mathrm{aug}}\|_{L_0^2(\nu)\to L_0^2(\nu)}.\end{aligned}$$

Unfortunately, K_{aug} is in general not reversible, see Remark 3.4, such that $\mathrm{K}_{\mathrm{aug}}$ is not self-adjoint. Thus, we can only estimate the spectral radius of $\mathrm{K}_{\mathrm{aug}}\colon L_0^2(\nu) \to L_0^2(\nu)$, but not the operator norm. The same argument yields to

$$r(\mathrm{K}_{\mathrm{aug}} \mid L_0^2(\nu)) \leq \|\mathrm{K}\|_{L_0^2(\mu)\to L_0^2(\mu)}\,.$$

Finally, since $\mathrm{K}_{\mathrm{aug}}$ is a product of two self-adjoint operators and, additionally, the projection $\widehat{\mathrm{P}}^*\widehat{\mathrm{P}}$ is positive, we obtain by [30, Proposition 4.1] that the spectrum of $\mathrm{K}_{\mathrm{aug}}\colon L_0^2(\nu) \to L_0^2(\nu)$ is real and non-negative. □

Since K_{aug} is not reversible, we can not argue that a positive $L^2(\nu)$-spectral gap of $\mathrm{K}_{\mathrm{aug}}$, which due to statement 2 of Lemma A.1 is implied by a positive $L^2(\mu)$-spectral gap of K, yields the $L^2(\nu)$-geometric ergodicity of the augmented MH Markov chain. However, by using also statement 1 of Lemma A.1 we indeed obtain the inheritance of geometric ergodicity.

Corollary A.1. *Assume that the MH transition kernel K with stationary distribution μ on G is $L^2(\mu)$-geometrically ergodic. Then, the augmented MH transition kernel K_{aug} is $L^2(\nu)$-geometrically ergodic with ν as in (3.3).*

Proof. By Proposition A.2 and the μ-reversibility of K we have that $r := \|\mathrm{K}\|_{L_0^2(\mu)\to L_0^2(\mu)} < 1$. Let η be a probability distribution on $G \times G$ such that $\frac{\mathrm{d}\eta}{\mathrm{d}\nu} \in L^2(\nu)$. With the notation of the adjoint operator we use (for details we refer to [33, Lemma 3.9]) that

$$\frac{\mathrm{d}(\eta K_{\mathrm{aug}}^n)}{\mathrm{d}\nu}(x,y) = (\mathrm{K}_{\mathrm{aug}}^n)^* \left[\frac{\mathrm{d}\eta}{\mathrm{d}\nu}\right](x,y), \quad \nu\text{-a.e.}$$

as well as

$$\left\|(\mathrm{K}_{\mathrm{aug}}^n)^*\right\|_{L_0^2(\nu)\to L_0^2(\nu)} = \left\|\mathrm{K}_{\mathrm{aug}}^n\right\|_{L_0^2(\nu)\to L_0^2(\nu)}\,.$$

Then, for $n \geq 2$ we have

$$\begin{aligned}
2d_{\mathrm{TV}}(\nu, \eta K_{\mathrm{aug}}^n) &= \int_{G\times G} \left|\frac{\mathrm{d}(\eta K_{\mathrm{aug}}^n)}{\mathrm{d}\nu}(x,y) - 1\right| \nu(\mathrm{d}x\mathrm{d}y) \\
&= \int_{G\times G} \left|(\mathrm{K}_{\mathrm{aug}}^n)^* \left[\frac{\mathrm{d}\eta}{\mathrm{d}\nu}\right](x,y) - 1\right| \nu(\mathrm{d}x\mathrm{d}y) \\
&= \int_{G\times G} \left|(\mathrm{K}_{\mathrm{aug}}^n)^* \left[\frac{\mathrm{d}\eta}{\mathrm{d}\nu}(x,y) - 1\right]\right| \nu(\mathrm{d}x\mathrm{d}y) \\
&\leq \left\|(\mathrm{K}_{\mathrm{aug}}^n)^* \left[\frac{\mathrm{d}\eta}{\mathrm{d}\nu} - 1\right]\right\|_\nu \\
&\leq \left\|\mathrm{K}_{\mathrm{aug}}^n\right\|_{L_0^2(\nu)\to L_0^2(\nu)} \left\|\frac{\mathrm{d}\eta}{\mathrm{d}\nu} - 1\right\|_\nu \\
&\leq \|\mathrm{K}\|^{n-1}_{L_0^2(\mu)\to L_0^2(\mu)} \left\|\frac{\mathrm{d}\eta}{\mathrm{d}\nu} - 1\right\|_\nu \leq C_\eta\, r^n
\end{aligned}$$

with $C_\eta := \frac{1}{r}\left\|\frac{\mathrm{d}\eta}{\mathrm{d}\nu} - 1\right\|_\nu$, where we used the fact that $(\frac{\mathrm{d}\eta}{\mathrm{d}\nu} - 1) \in L_0^2(\nu)$ as well as statement 1 of Lemma A.1. □

Acknowledgements

We thank I. Klebanov, K. Łatuszyński, A. Lee, I. Schuster, and M. Vihola for helpful and inspiring discussions.

References

[1] Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science 32*, 3 (2017), 405–431. MR3696003

[2] Atchadé, Y. F., and Perron, F. Improving on the independent Metropolis-Hastings algorithm. *Statistica Sinica 15*, 1 (2005), 3–18. MR2125717

[3] Botev, Z. I., L'Ecuyer, P., and Truffin, B. Markov chain importance sampling with applications to rare event probability estimation. *Statistics and Computing 23* (2013), 271–285. MR3016944

[4] Cappé, O., Moulines, E., and Ryden, T. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. MR2159833

[5] Casella, G., and Robert, C. Rao-Blackwellisation of sampling schemes. *Biometrika 83* (1996), 81–94. MR1399157

[6] Chatterjee, S., and Diaconis, P. The sample size required in importance sampling. *Ann. Appl. Probab. 28*, 2 (2018), 1099–1135. MR3784496

[7] Chopin, N., and Ridgway, J. Leave Pima indians alone: Binary regression as a benchmark for Bayesian computation. *Statistical Science 32*, 1 (02 2017), 64–87. MR3634307

[8] Delmas, J., and Jourdain, B. Does waste recycling really improve the multi-proposal Metropolis–Hastings algorithm? An analysis based on control variates. *Journal of Applied Probability 46*, 4 (12 2009), 938–959. MR2582699

[9] Douc, R., and Robert, C. A vanilla Rao–Blackwellization of Metropolis–Hastings algorithms. *Ann. Statist. 39*, 1 (2011), 261–277. MR2797846

[10] Ernst, O., Sprungk, B., and Starkloff, H.-J. Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems. *SIAM/ASA J. Uncertainty Quantification 3*, 1 (2015), 823–851. MR3400030

[11] Frenkel, D. Speed-up of Monte Carlo simulations by sampling of rejected states. *Proceedings of the National Academy of Sciences 101* (2004), 17571–17575.

[12] Frenkel, D. Waste-recycling Monte Carlo. *In Computer Simulations in Condensed Matter: From Materials to Chemical Biology (Lecture Notes Phys. 703), Springer, Berlin* (2006), 127–137.

[13] Garbuno-Inigo, A., Hoffman, F., Li, W., and Stuart, A. M. Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. Preprint arXiv:1903.08866v3, 2019. MR4059375

[14] Herty, M., and Visconti, G. Kinetic methods for inverse problems. *Kinetic & Related Models 12*, 5 (2019), 1109–1130. MR4027079
[15] Hinrichs, A. Optimal importance sampling for the approximation of integrals. *Journal of Complexity 26*, 2 (2010), 125–134. MR2607728
[16] Jones, G. On the Markov chain central limit theorem. *Probability Surveys 1* (2004), 299–320. MR2068475
[17] Joulin, A., and Ollivier, Y. Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab. 38*, 6 (2010), 2418–2442. MR2683634
[18] Kipnis, C., and Varadhan, S. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics 104*, 1 (1986), 1–19. MR0834478
[19] Kontoyiannis, I., and Meyn, S. Geometric ergodicity and the spectral gap of non-reversible markov chains. *Probab. Theory Relat. Fields 154*, 1 (2012), 327–339. MR2981426
[20] Łatuszynski, K., Miasojedow, B., and Niemiro, W. Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli 19*, 5A (2013), 20133–2066. MR3129043
[21] Łatuszyński, K., and Niemiro, W. Rigorous confidence bounds for MCMC under a geometric drift condition. *Journal of Complexity 27*, 1 (2011), 23–38. MR2745298
[22] Martino, L., Elvira, V., Luengo, D., and Corander, J. Layered adaptive importance sampling. *Statistics and Computing 27*, 3 (2016), 599–623. MR3613588
[23] Mathé, P., and Novak, E. Simple Monte Carlo and the Metropolis algorithm. *Journal of Complexity 23*, 4-6 (2007), 673–696. MR2372022
[24] Maxwell, M., and Woodroofe, M. Central limit theorems for additive functionals of Markov chains. *Ann. Probab. 28*, 2 (2000), 713–724. MR1782272
[25] Mengersen, K., and Tweedie, R. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist. 24*, 1 (1996), 101–121. MR1389882
[26] Meyn, S., and Tweedie, R. *Markov chains and stochastic stability*, first ed. Cambridge University Press, 1993. MR2509253
[27] Owen, A. *Monte Carlo theory, methods and examples.* 2013.
[28] Roberts, G., and Rosenthal, J. Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab. 2* (1997), no. 2, 13–25. MR1448322
[29] Roberts, G., and Rosenthal, J. Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science 16*, 4 (2001), 351–367. MR1888450
[30] Rosenthal, J., and Rosenthal, P. Spectral bounds for certain two-factor non-reversible mcmc algorithms. *Electron. Commun. Probab. 20* (2015), 10 pp. MR3434208
[31] Rudolf, D. Explicit error bounds for lazy reversible Markov chain Monte Carlo. *Journal of Complexity 25*, 1 (2009), 11–24. MR2475305
[32] Rudolf, D. Error bounds of computing the expectation by Markov chain

Monte Carlo. *Monte Carlo Methods and Applications 16* (2010), 323–342. MR2747819

[33] Rudolf, D. Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Mathematicae 485* (2012), 93 pp. MR2977521

[34] Schuster, I. Gradient Importance Sampling. Preprint arXiv: 1507.05781v1, 2015.

[35] Schuster, I., and Klebanov, I. Markov chain importance samping – a highly efficient estimator for MCMC. Preprint arXiv:1805.07179v3, 2020.

[36] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., and Johannes, R. S. Using the ADAP learning algorithm to forecast the onset of Diabetes Mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* (1988), pp. 261–265.

[37] Tierney, L. Markov chains for exploring posterior distributions. *Ann. Stat. 22*, 4 (1994), 1701–1762. MR1329166

[38] Tierney, L. A note on the Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab. 8* (1998), 1–9. MR1620401

[39] Vihola, M., Helske, J., and Franks, J. Importance sampling type estimators based on approximate marginal MCMC. Preprint arXiv:1510.02577v5, 2018.