Bernoulli **24**(3), 2018, 1726–1786
https://doi.org/10.3150/16-BEJ911

Unbiased Monte Carlo: Posterior estimation for intractable/infinite-dimensional models

SERGIOS AGAPIOU[1], GARETH O. ROBERTS[2] and
SEBASTIAN J. VOLLMER[3]

[1]*Department of Mathematics and Statistics, University of Cyprus. E-mail:* Agapiou.Sergios@ucy.ac.cy
[2]*Department of Statistics, University of Warwick. E-mail:* Gareth.O.Roberts@warwick.ac.uk
[3]*Department of Statistics/Mathematical Institute, University of Warwick and Alan Turing Institute. E-mail:* svollmer@turing.ac.uk

We provide a general methodology for unbiased estimation for intractable stochastic models. We consider situations where the target distribution can be written as an appropriate limit of distributions, and where conventional approaches require truncation of such a representation leading to a systematic bias. For example, the target distribution might be representable as the L^2-limit of a basis expansion in a suitable Hilbert space; or alternatively the distribution of interest might be representable as the weak limit of a sequence of random variables, as in MCMC. Our main motivation comes from infinite-dimensional models which can be parameterised in terms of a series expansion of basis functions (such as that given by a Karhunen–Loeve expansion). We introduce and analyse schemes for direct unbiased estimation along such an expansion. However, a substantial component of our paper is devoted to the study of MCMC schemes which, due to their infinite dimensionality, cannot be directly implemented, but which can be effectively estimated unbiasedly. For all our methods we give theory to justify the numerical stability for robust Monte Carlo implementation, and in some cases we illustrate using simulations. Interestingly the computational efficiency of our methods is usually comparable to simpler methods which are biased. Crucial to the effectiveness of our proposed methodology is the construction of appropriate couplings, many of which resonate strongly with the Monte Carlo constructions used in the coupling from the past algorithm.

Keywords: Bayesian inverse problems; coupling; Markov chain Monte Carlo in infinite dimensions; unbiased estimation

1. Introduction

Bayesian analyses of complex models often lead to posterior distributions which are only available indirectly as an appropriate limit of a sequence of probability measures. A classical example of this is Markov Chain Monte Carlo (MCMC), which constructs an algorithm to access the posterior distribution which involves creating Markov chains with the required limiting distribution. Rather different examples come from infinite-dimensional models, for example arising in inference for continuous-time stochastic processes, and in inverse problems where the quantity to be inferred is naturally expressed as a function in an appropriate Hilbert space. In these examples, the exact representation of the posterior distribution is via an infinite sum (perhaps representing a basis expansion) or the limit of a sequence of approximations perhaps derived from time discretisations. Thus, in these contexts we have an indirect representation of the posterior distribution.

The conventional approach to such an indirect representation is to truncate:

1350-7265

- to run the MCMC for long enough;
- to choose a fixed fine time-discretisation;
- or to take sufficiently many terms in the series expansion.

The main problem with this general approach is that the accuracy of the approximation produced is highly application-specific and very difficult to analyse.

It is a common misconception that *exact* methods, which avoid truncation approximations entirely, are either impossible or prohibitively computationally expensive (see [23] for some examples involving simulation of SDEs). Although stochastic simulation directly from the posterior distribution is generally not feasible, it turns out to be very commonly feasible and practical to obtain unbiased estimates for any arbitrary posterior expected functional of interest. This is the focus of the present paper, which builds on the contributions of [26].

Fundamental to the success of these methods is the construction of suitable couplings to ensure our estimators have finite variances. Much of these constructions resonate with the huge body of literature inspired by the coupling from the past algorithm of Propp and Wilson [25], although crucially our methods are substantially more general as we do not require the strong *coalescent* couplings needed for coupling from the past.

Although we shall state most of our results quite generally, our main applications will be in the area of Bayesian inverse problems. We construct unbiased estimators in four settings:

- for chains with exactly computable transitions, which posses a simulatable contracting coupling between runs started at different positions (Section 2 – bias due to using finite time distributions);
- for linear Gaussian conjugate infinite dimensional Bayesian inverse problems (Section 3 – bias due to discretisation);
- for non-linear infinite dimensional inverse problems with uniform series priors using the independence sampler (Section 4 – bias due to discretisation of the states and finite time);
- for measures with log-Lipschitz densities with respect to infinite dimensional Gaussians using the pCN algorithm (Section 5 – bias due to discretisation of the states and finite time).

There are many operational choices in our procedures, and we have only just begun exploring all the options. Optimisation of our procedures is therefore an important and difficult question which leads on from our work here. From the examples, we have considered here, we have however been surprised by the apparent efficiency of essentially ad-hoc choices for algorithm parameters. Thus, our methods seem very promising as practical and general approaches which circumvent the systematic error of existing approaches.

Although our work is significantly more technical in nature than [26], we see our main contributions here as methodological rather than mathematical, and in this light have tried to keep technicalities to a minimum, particularly in the main body of the paper. For instance, we refrain from expressing or proving our results for the most general Hilbert space-valued functions, even though a generalisation to this context is completely straightforward.

1.1. Overview of existing results

We now briefly outline the recent results by Chang-han Rhee and Peter Glynn [26,27], which we extend in the following sections (see also work by Don McLeish [20]). The objective is to efficiently simulate an unbiased estimator of the expectation of a real valued random variable Y. We consider settings in which the exact simulation of Y is impossible due to the infinite cost associated with generating an exact sample, thus in order to perform a Monte Carlo simulation one needs to use approximations Y_i of Y. This introduces a bias in the Monte Carlo estimator of the expectation, which in turn results in suboptimal rates of convergence with respect to the computational budget c. In particular, instead of the optimal $\mathcal{O}(c^{-1/2})$ rate of convergence, we get slower rates. In the aforementioned works, the goal is twofold: first to construct unbiased estimators of the expectation of interest using an appropriate combination of biased ones, and second to find conditions which secure that the variance and the computational cost of the constructed estimator are such that the optimal rate of convergence with respect to the budget c is achieved.

The starting point is a neat randomisation idea for unbiased estimation of infinite sums, which traces back to John von Neumann and Stanislaw Ulam in the context of matrix inversion [12,33]. The idea was more recently employed by Peter Glynn in the setting of time integral estimation [14]. Assume that the approximations Y_i satisfy $\mathbb{E}(Y_i) \to \mathbb{E}(Y)$ as $i \to \infty$. Then one can express the expectation of Y as a telescoping sum

$$\mathbb{E}(Y) = \sum_{i=0}^{\infty} \mathbb{E}(Y_i - Y_{i-1}),$$

where $Y_{-1} = 0$ by convention. Provided the approximations are good enough so that Fubini's theorem applies, this suggests that an unbiased estimator for $\mathbb{E}(Y)$ is the sum $\sum_{i=0}^{\infty}(Y_i - Y_{i-1})$. However, this estimator cannot be generated in finite time, so the idea is to use a random truncation point N and correct for the introduced bias. Indeed, let N be an integer-valued random variable which is independent of the random approximations Y_i and is such that $P(N \geq i) > 0$ for all $i \in \mathbb{N}$. Then, letting $\Delta_i = Y_i - Y_{i-1}$, and again assuming that the approximations are good enough so that we can interchange expectation with summation, we have that

$$\mathbb{E}\left[\sum_{i=0}^{N} \frac{\Delta_i}{P(N \geq i)}\right] = \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{1_{\{N \geq i\}}\Delta_i}{P(N \geq i)}\right] = \sum_{i=0}^{\infty} \mathbb{E}(\Delta_i) = \mathbb{E}(Y),$$

so that the estimator

$$Z := \sum_{i=0}^{N} \frac{\Delta_i}{P(N \geq i)} \tag{1.1}$$

is unbiased.

In order for the estimator Z to be practical, we need to also have that its variance, $\operatorname{var}(Z)$, as well as the expected work required to generate a copy of it, $\mathbb{E}(\tau)$, are finite. Letting t_i be the expected incremental effort required to calculate Y_i, we have that

$$\mathbb{E}(\tau) = \mathbb{E}\left(\sum_{i=0}^{N} t_i\right) = \sum_{i=0}^{\infty} t_i P(N \geq i), \tag{1.2}$$

while in [26,27] it is shown that

$$\mathrm{var}(Z) = \sum_{i=0}^{\infty} \frac{\beta_i}{P(N \geq i)},$$

where $\beta_i = \mathcal{O}(\mathbb{E}[(Y - Y_i)^2])$. It is hence apparent that there is a competition between $P(N \geq i)$ decaying fast enough so that the expected work required to generate Z is finite, but not too fast so that $\mathrm{var}(Z)$ is also finite. In order to obtain that both the expected work and the variance of the estimator are finite, the rate of convergence of $\mathbb{E}[(Y - Y_i)^2]$ needs to be faster than the rate at which the expected incremental effort t_i goes to ∞.

The following proposition is proved in [26] and is very useful for verifying the unbiasedness and finite variance of the proposed estimator. Here and elsewhere, we use the notation $\|h\|_2 := (\mathbb{E}[h^2])^{\frac{1}{2}}$.

Proposition 1.1 (Proposition 6, [26]). *Suppose that* $(\Delta_i : i \geq 0)$ *is a sequence of real-valued random variables and let* N *be an integer-valued random variable which is independent of the* Δ_i*'s and satisfies* $\mathbb{P}(N \geq i) > 0$ *for all* $i \geq 0$. *Assume that*

$$\sum_{i \leq l} \frac{\|\Delta_i\|_2 \|\Delta_l\|_2}{\mathbb{P}(N \geq i)} < \infty.$$

Then $Y_n := \sum_{i=0}^{n} \Delta_i$ *converges in* L^2 *to a limit* $Y := \sum_{i=0}^{\infty} \Delta_i$ *as* $n \to \infty$. *Let* $\alpha = \mathbb{E}Y (= \lim_{n\to\infty} \mathbb{E}Y_n)$ *and suppose that for all* i, $\tilde{\Delta}_i$ *is a copy of* Δ_i *such that* $\{\tilde{\Delta}_i\}$ *are mutually independent. Then* $\tilde{Z} := \sum_{i=0}^{N} \frac{\tilde{\Delta}_i}{\mathbb{P}(N \geq i)}$ *is an unbiased estimator for* α *with finite second moment*

$$\mathbb{E}\tilde{Z}^2 = \sum_{i=0}^{\infty} \frac{\tilde{\nu}_i}{\mathbb{P}(N \geq i)},$$

where $\tilde{\nu}_i = \mathrm{var}(\Delta_i) + (\alpha - \mathbb{E}Y_{i-1})^2 - (\alpha - \mathbb{E}Y_i)^2$.

Remark 1.2. In Proposition C.1 in the Supplementary Material [3], we generalise Proposition 1.1 to cover estimation of expectations of Hilbert space-valued random variables Y. Nevertheless, in order to avoid overcomplicating our presentation, we state and prove our results for real-valued random variables Y and only comment on their applicability in the more general Hilbert space setting.

Under the assumption that both $\mathrm{var}(Z)$ and $\mathbb{E}(\tau)$ are finite, Glynn and Whitt's results on general estimators imply that a central limit theorem holds

$$c^{1/2}\big(\hat{\alpha}(c) - \mathbb{E}(Y)\big) \Rightarrow \sqrt{\mathbb{E}(\tau)\,\mathrm{var}(Z)}N(0, 1), \tag{1.3}$$

where $\hat{\alpha}(c)$ is the Monte Carlo estimator produced from independent replicates of Z that can be generated after c units of computer time [15]. This immediately gives that the estimator converges at the optimal square root rate. Furthermore, the above central limit theorem supports theoretically the intuition that the product of the variance and the expected work is a good measure of efficiency of the estimator and consequently suggests that the choice of distribution for N can be optimised by minimising this product.

In the work of Rhee and Glynn [26,27], this programme has been developed and carried out in two general settings. The first setting is simulation of SDEs, in which these ideas are directly applicable to many of the available discretisation schemes. An important observation in this setting is that for lower order schemes, like the Euler–Maruyama discretisation, this methodology does not work since the convergence of $E[(Y - Y_i)^2]$ is not quick enough compared to the increase in the cost of producing Y_i. On the other hand, with respect to the bias aspect of the problem, there is no need to use discretisation schemes of particularly high order, since for example, the Millstein scheme is already enough to secure the optimal square root convergence rate of the Monte Carlo estimator. The second setting is the study of ergodic Markov chains, where the aim is to estimate expectations with respect to the invariant measure and the finite-time distributions are used to define the approximations Y_i. In this setting, the theory is not immediately applicable, since although the finite-time distributions converge to the invariant measure, in general the random variables Y_n defined through the outcome of the Markov chain, may not converge in the L^2 sense. For this reason, one needs to construct an appropriate coupling to enable the sequence of approximations to converge in L^2 and thus to permit the application of Proposition 1.1. In [26], such couplings are constructed for uniformly ergodic, contracting and Harris chains (see Section 1.2 below).

In infinite-dimensional contexts (such as those arising in Bayesian inverse problems) it is usually impossible to implement the infinite-dimensional MCMC algorithms required to sample from the target distribution (though see [5], for an example where it can be done).

A rather different application of the ideas of unbiasing by taking random differences, has been introduced by [9,16,19], which build on the Multilevel Monte Carlo (MLMC) method of Mike Giles [13]. This method makes substantial progress in the construction of algorithms which unbiasedly estimate chosen finite-dimensional summaries from infinite-dimensional MCMC methods. However, these methods do not avoid bias due to Markov chain burn-in. In the present paper, we will provide practical unbiasing methods which circumvent bias, either from the need for finite-dimensional approximation, or from Markov chain burn in.

1.2. Glynn and Rhee's results for exact estimation in the context of ergodic Markov chains

Before moving on with the presentation of our results, we briefly recall the methodology of [26] for constructing an appropriate coupling in the setting of uniformly ergodic Markov chains; we will build our extension to the MCMC in function space setting on this methodology. Let $X = \{X_n\}_{n\in\mathbb{N}}$ be a Markov chain in a state space $\mathcal{X}$, with transition probabilities $P(x, A)$ and invariant distribution π. A *uniformly recurrent* Markov chain is one for which there exists a probability measure ν on $\mathcal{X}$, a constant $\lambda > 0$ and an integer $m \geq 1$, such that

$$P^m(x, B) \geq \lambda\nu(B),$$

for any $x \in \mathcal{X}$ and any measurable $B \subset \mathcal{X}$. In other words, this condition says that the whole space $\mathcal{X}$ is "small", see [22], Section 5.2, which by [22], Theorem 16.0.2, is one of the conditions that are equivalent to the uniform ergodicity of the Markov chain. In particular, the Markov chain converges to its invariant distribution, however this does not guarantee that X converges in L^2. In order to find a coupling of X_n and X_{n+1} such that they come closer in L^2 as n increases, the authors of [26] define the random functions

$$\varphi_n(\cdot) := I_n \xi_n + (1 - I_n)\phi_n(\cdot),$$

where I_n are independent and identically distributed Bernoulli random variables with success probability λ, ξ_n are independent random variables drawn according to ν, and ϕ_n are random functions representing the transition $Q(x, B) = \frac{P(x,B)-\lambda\nu(B)}{1-\lambda}$, that is, $P(\phi_n(x) \in B) = Q(x, B)$. They then recursively express the chain X_n as $X_{n+1} = \varphi_n(X_n)$, where $X_0 = x$. Since $\varphi_n(x)$ are independent and identically distributed according to $P(x, \cdot)$, one can then define $\tilde{X}_n$ to be the backwards process

$$\begin{aligned} \tilde{X}_{n+1} &:= \varphi_0 \circ \cdots \circ \varphi_n(x) \\ &\overset{\mathcal{L}}{=} \varphi_n \circ \cdots \circ \varphi_0(x) \\ &= X_{n+1}. \end{aligned} \tag{1.4}$$

Note that φ_n is constant with positive probability λ, so that with probability $1 - (1 - \lambda)^n$, at least one of the φ_k, $k \in \{1, \dots, n\}$, is a constant (random) function. The advantage of working with the backwards process is that contrary to the forward process, if φ_n is a constant function then all $\tilde{X}_k$ for $k > n$ are equal to the same constant. We hence have that as n increases, with probability which goes to 1, $\tilde{X}_n = \tilde{X}_{n+1}$.

This is particularly useful for estimating the expectation of a bounded function $f : \mathcal{X} \to \mathbb{R}$ with respect to the equilibrium distribution π, $\mathbb{E}_\pi[f]$. An obvious choice of approximating sequence in this setting is the sequence of the images under f of the chain after a finite number of steps, hence we let $Y_i = f(X_i)$. Then

$$\begin{aligned} \mathbb{E}\big[(Y_i - Y_{i-1})^2\big] &= \mathbb{E}\big[\big(f(\tilde{X}_i) - f(\tilde{X}_{i-1})\big)^2\big] \\ &\le \|f\|_\infty^2 P(\varphi_j \text{ is not constant for all } j \le i) \\ &\le \|f\|_\infty^2 (1 - \lambda)^i. \end{aligned}$$

We thus have that the Y_i converge in L^2 and the unbiasing programme described in the previous subsection can be applied.

1.3. Implementation of the backwards construction

At a high level, Rhee and Glynn's general approach is to represent the chain using random functions $\varphi(x, W)$, where W represents all the randomness needed to simulate the transition. Then the evolution of the chain is written as $X_n = \varphi_n \circ \cdots \circ \varphi_0(x)$, where $\varphi_i(\cdot) = \varphi(\cdot, W_i)$ for some independent identically distributed sequence W_i. As described above, the backwards technique consists in considering the chain $\tilde{X}_n = \varphi_0 \circ \cdots \circ \varphi_n(x) \overset{\mathcal{L}}{=} X_n$. Under appropriate assumptions (contraction or uniform ergodicity) this technique turns the weak convergence of the chain X_n to its equilibrium distribution, to almost sure convergence of $\tilde{X}_n$ to a limiting random variable $\tilde{X}$. The chain $\tilde{X}$ is then used to obtain the approximations $Y_i = f(\tilde{X}_i)$, and hence the differences $\Delta_i = Y_i - Y_{i-1} = f(\tilde{X}_i) - f(\tilde{X}_{i-1})$ which are used for generating the unbiased estimator Z. It is important to observe, that completely independent copies of Δ_i at different levels i are used both for the algorithm and the analysis, see Proposition 1.1.

We remark that the above described coupling is also used as the fundamental idea in the *coupling from the past* algorithm for sampling perfectly from the invariant distribution of a Markov chain [25]. Furthermore, note that the backwards technique in the above described form, has the disadvantage that in order to pass from $\tilde{X}_n$ to $\tilde{X}_{n+1}$ we need to recompute the whole chain. This means that in order to compute Δ_i, we first need to produce Y_{i-1} and then start from scratch to produce Y_i (this discussion does not apply for producing Δ_i and Δ_{i+1} since they are assumed to be independent). For the benefit of the reader and since no implementation details are given in [26], we now describe a reasonable implementation. This implementation is easier than the coupling from the past algorithm, however the probabilistic construction is very similar. We will later generalise this construction to cover sampling from infinite dimensional target measures, using the finite-time distributions of a hierarchy of Markov chains with state spaces of increasing dimension (see Sections 4 and 5).

We start by noticing that it is not necessary to construct Y_i's that have the correct distribution, but rather it suffices to generate Δ_i's which have the correct expectation (this is silently observed in [26] but the authors do not seem to exploit it). We present this in a more general setting, and in particular we consider approximation levels i that correspond to a_i time steps, where $\{a_i\}_{i\in\mathbb{N}}$ is a strictly increasing sequence of positive integer numbers. We will show later in Section 6 that the choice of a_i has a huge impact on the efficiency of the estimator.

The random variables $\tilde{X}_{a_i}$ and $\tilde{X}_{a_{i-1}}$ needed to generate Δ_i when using the backwards technique, are given as

$$\begin{aligned}\tilde{X}_{a_i} &= \varphi\bigl(\varphi\bigl(\dots\varphi\bigl(\varphi(x_0, W_{a_i}), W_{a_i-1}\bigr)\dots, W_2\bigr), W_1\bigr),\\ \tilde{X}_{a_{i-1}} &= \varphi\bigl(\varphi\bigl(\dots\varphi\bigl(\varphi(x_0, W_{a_{i-1}}), W_{a_{i-1}-1}\bigr)\dots, W_2\bigr), W_1\bigr).\end{aligned}$$

The same set of random variables, can be generated sequentially as the algorithm progresses. To do this, we introduce the chains $\mathcal{T}^i, \mathcal{B}^i$ corresponding to the "top" and "bottom" approximation levels, respectively, which appear in the definition of Δ_i. We set

$$\mathcal{T}^i_{-a_i} = x_0,$$

and to get $\mathcal{T}^i_{-a_{i-1}}$ we simulate until $-a_{i-1}$, that is we set

$$\mathcal{T}^i_{-a_{i-1}} = \varphi\bigl(\dots\varphi\bigl(\varphi(x_0, W_{a_i}), W_{a_i-1}\bigr)\dots, W_{a_{i-1}+1}\bigr). \tag{1.5}$$

We then set $\mathcal{B}^i_{-a_{i-1}} = x_0$, and simulate $\mathcal{T}^i$ and $\mathcal{B}^i$ jointly up to time 0, hence obtaining

$$\begin{aligned} \mathcal{T}^i_0 &= \varphi\big(\varphi\big(\dots\varphi\big(\varphi\big(\mathcal{T}^i_{-a_{i-1}}, W_{a_{i-1}}\big), W_{a_{i-1}-1}\big)\dots, W_2\big), W_1\big), \\ \mathcal{B}^i_0 &= \varphi\big(\varphi\big(\dots\varphi\big(\varphi(x_0, W_{a_{i-1}}), W_{a_{i-1}-1}\big)\dots, W_2\big), W_1\big). \end{aligned} \tag{1.6}$$

Thus we have $\mathcal{B}^i_0 = \tilde{X}_{a_{i-1}}$ and $\mathcal{T}^i_0 = \tilde{X}_{a_i}$, and can define $\Delta_i = f(\mathcal{T}^i_0) - f(\mathcal{B}^i_0)$. Furthermore, observe that the direction of enumeration of the W's does not matter in this construction, since the a_i's are a priori fixed and can hence be generated as the algorithm progresses.

Alternatively, one can think of this construction in terms of couplings. Let

$$P(x, \cdot) = \mathcal{L}\big(\varphi(x, W)\big)$$

be the transition kernel of X_i. Moreover,

$$K\big((x, y), \cdot\big) := \mathcal{L}\big(\varphi(x, W), \varphi(y, W)\big)$$

is a coupling of $P(x, \cdot)$ and $P(y, \cdot)$. This coupling allows us to write (1.5) and (1.6) as

1. $\mathcal{T}^i_{-a_i} = x_0$, then simulate according to P up to $\mathcal{T}^i_{-a_{i-1}}$;
2. set $\mathcal{B}^i_{-a_{i-1}} = x_0$, then simulate $(\mathcal{T}^i, \mathcal{B}^i)$ jointly according to K up to $(\mathcal{T}^i_0, \mathcal{B}^i_0)$.

Under the assumption that

$$\sum_{i \le l} \frac{\|\Delta_i\|_2 \|\Delta_l\|_2}{\mathbb{P}(N \ge i)} < \infty, \tag{1.7}$$

which has to be verified for different classes of Markov chains, we can define retrospectively the approximations $Y_i := \sum_{k=0}^{i} \Delta_k$, and apply Proposition 1.1 and more generally the programme developed by Rhee and Glynn, to get an unbiased estimator of $\mathbb{E}_\pi[f]$ with optimal cost.

1.4. Notation

We always denote the state space by $\mathcal{X}$, although we work under assumptions on the state space which differ between sections. As stated earlier, we use the notation $\|h\|_2 = (\mathbb{E}[h^2])^{\frac{1}{2}}$. We use f to denote the function whose expectations we want to estimate and denote by $\mathbb{E}_\pi[f]$ the expectation of f under a probability measure π. For two sequences k_j and h_j of positive real numbers, $k_j \asymp h_j$ means that $\frac{k_j}{h_j}$ is bounded away from zero and infinity as $j \to \infty$, $k_j \lesssim h_j$ means that $\frac{k_j}{h_j}$ is bounded as $j \to \infty$, and $k_j \sim h_j$ means that $\frac{k_j}{h_j} \to 1$ as $j \to \infty$.

1.5. Organisation of the paper

In Section 2, we consider unbiased estimation of expectations with respect to the limiting distribution of an ergodic Markov chain in a state space in which we assume that transitions can be computed exactly. The source of the bias is then only the use of finite time distributions to approximate the limiting distribution (burn-in). Compared to the contracting chain setting of [26], we work under the weaker assumption that there exists a simulatable contracting coupling between runs of the chain started at different states (see Remark 2.9).

In Section 3, we consider unbiased estimation of posterior expectations in Gaussian-conjugate Bayesian linear inverse problems in a separable Hilbert space. Since in this setting the posterior is also Gaussian, the source of the bias is only the discretisation.

In Section 4, we consider estimation of posterior expectations in a nonlinear Bayesian inverse problem setting in function space, with uniform series priors and under assumptions which ensure the uniform ergodicity of the independence sampler at any fixed discretisation level of the state space. In this case the bias is both due to discretisation of the state and burn-in. We achieve unbiased estimation by constructing a hierarchy of coupled independence samplers in state spaces of increasing dimension.

In Section 5, we consider target measures which are absolutely continuous with respect to a Hilbert space Gaussian reference measure, under assumptions on the log-density which secure the existence of a simulatable contracting coupling of the pCN algorithm at any fixed discretisation level of the state space. In this case, the bias is again due to both discretisation and burn-in. We achieve unbiased estimation by constructing a hierarchy of coupled pCN algorithms in state spaces of increasing dimension.

In Section 6, we present a comparison between the performance of the ergodic average of an MCMC run and the performance of the Monte Carlo estimator constructed using the unbiasing procedure. This is first done in a 1-dimensional Gaussian autoregression setting and then for a Bayesian logistic regression model.

The main body of the paper ends with concluding remarks in Section 7. The proofs of the results in Sections 2, 4 and 5, as well as the statements and proofs of some necessary intermediate results are contained in Section 8. The proofs of the results in Section 3 are straightforward calculations and are included in the Supplementary Material [3]. Also in the Supplementary Material we provide some further considerations in the context of Section 3, the generalisation of Proposition 1.1 to Hilbert space-valued random variables, and an elliptic inverse problem example which satisfies the assumptions of Section 4.

2. Wasserstein convergence of Markov chains and unbiased estimators of equilibrium expectations

In this section, we consider the problem of constructing unbiased estimators for expectations with respect to limiting distributions of Markov chains. As discussed in Section 1.1, the techniques developed in [27], have been applied in [26] in this setting and in particular for uniformly recurrent, contracting and Harris chains. The approximation is achieved by considering the finite-time distributions, and then the challenge is to construct a coupling which guarantees that the chain comes close in L^2 as time increases. In general, the approach taken in [26], is to use intelligent techniques that turn convergence in distribution to almost sure convergence (for example, the backwards process technique described in Section 1.1). We now show that this is not necessary, but instead a simulatable coupling between chains started at different positions is sufficient, provided this coupling drives the two chains towards each other quickly enough in expectations.

Let $\mathcal{X}$ be a general state space. Throughout this section, d denotes a distance-like function, that is a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ which is symmetric, lower semi-continuous and which vanishes when the two arguments are equal. Let $X = \{X_n\}_{n\in\mathbb{N}}$ be a Markov chain with transition probabilities $P(x, .)$ and invariant distribution π. Our aim is to find an unbiased estimate for the expectation $\mathbb{E}Y := \mathbb{E}_\pi[f]$, where $f : \mathcal{X} \to \mathbb{R}$ is an s-Hölder continuous function with respect to d for some $s \in (0, 1]$, that is

$$\|f\|_s := \sup_{x\neq y} \frac{|f(x) - f(x)|}{d^s(x, y)} < \infty.$$

Assumption 2.1. *We work under the following assumptions on the chain X in terms of the distance-like function d:*

i. *there exists a simulatable coupling $K((x, y), (dx', dy'))$ between the transition probabilities $P(x, dx')$ and $P(y, dy')$, which satisfies*

$$K^n d^{2s} \le cr^n d^{2s} \qquad \textit{for some } r < 1; \tag{2.1}$$

ii. *there exists a point $x_0 \in \mathcal{X}$ such that*

$$\sup_n P^n d(x_0, \cdot) < \infty. \tag{2.2}$$

Remark 2.2. We comment the following about Assumptions 2.1.

1. Assumption 2.1.i. is more general than the contracting chains case considered in Chapter 3.4 in [26], since the coupling is allowed to depend on both x and y; for more details, see Remark 2.9 below.
2. Assumption 2.1.i. is related to the s-Wasserstein distance-like function d_s associated with d, which is given by

$$d_s(\nu_1, \nu_2) = \left(\inf_{\pi\in\Gamma(\nu_1,\nu_2)} \int_{\mathcal{X}\times\mathcal{X}} d^s(x, y)\pi(dx, dy) \right)^{\frac{1}{s}},$$

with $\Gamma(\nu_1, \nu_2)$ being the set of couplings of ν_1 and ν_2 (all measures on $\mathcal{X} \times \mathcal{X}$ with marginals ν_1 and ν_2). Since K constitutes a particular coupling, it follows that

$$d_{2s}^{2s}\big(P^n(x, \cdot), P^n(y, \cdot)\big) \le K^n d^{2s}.$$

That is, our assumption is stronger than the corresponding assumption on the transition probabilities in terms of d_s because we need K to be simulatable.

3. Finally, observe that Assumption 2.1.ii. can be established by picking a distance d that is bounded or compatible with a Lyapunov function of the underlying Markov chain.

As discussed in Section 1.3, we generate the differences Δ_i directly and through them define the approximations Y_i. Let $\{a_i\}$ be an increasing sequence of integers. We generate Δ_i as specified

Algorithm 1 Coupled contraction for unbiased estimation

```
Fix starting point $x_0 \in \mathcal{X}$ once and for all. For $i = 0$
    • set $\mathcal{T}^0_{-a_0} = x_0$ and run the chain until $\mathcal{T}^0_0$;
    • set $\Delta_0 = f(\mathcal{T}^0_0)$.
For $i \geq 1$ do
    • set $\mathcal{T}^i_{-a_i} = x_0$ and run the chain until $\mathcal{T}^i_{-a_{i-1}}$;
    • set $\mathcal{B}^i_{-a_{i-1}} = x_0$;
    • evolve $\mathcal{B}^i_k$ and $\mathcal{T}^i_k$ jointly according to $K$ up to time 0;
    • set $\Delta_i = f(\mathcal{T}^i_0) - f(\mathcal{B}^i_0)$.
```

in Algorithm 1 and where x_0 is defined in Assumption 2.1.ii. We denote by $\mathcal{T}^i_k$ and $\mathcal{B}^i_k$ the chains coupled through the kernel K for $k = -a_{i-1} + 1, \dots, 0$.

In order to follow the general idea of the unbiasing technique as outlined in Section 1.1, we now make an assumption about the computing time of generating Δ_i.

Assumption 2.3. *The expected computing time t_i of generating Δ_i satisfies*

$$t_i \lesssim a_i.$$

This seems a reasonable assumption as Δ_i can be produced using a_i steps following K. We have the following result on the estimator Z defined in equation (1.1).

Theorem 2.4. *Suppose Assumption 2.1 (existence of contracting coupling) and Assumption 2.3 are satisfied, and that $f : \mathcal{X} \to \mathbb{R}$ is s-Hölder continuous with respect to d, for some $s \in (0, 1]$. Then there exist choices for a_i and $\mathbb{P}(N \geq i)$, such that*

$$Z = \sum_{i=0}^{N} \frac{\Delta_i}{\mathbb{P}(N \geq i)}$$

is an unbiased estimator of $\mathbb{E}_\pi[f]$ with finite variance and finite expected computing time. In particular, an example of such choices is $a_i \lesssim r^{(2\epsilon-1)i}$ and $\mathbb{P}(N \geq i) \propto r^{(1-\epsilon)i}$ for any $0 < \epsilon < \frac{1}{2}$.

Note that the exponential convergence in Assumption 2.1.i. makes the calculations easier, however the same argument works for sufficiently fast sub-exponential convergence.

Assumption 2.5. *There exists a simulatable coupling $K((x, y), (dx', dy'))$ between the transition probabilities $P(x, dx')$ and $P(y, dy')$, which satisfies*

$$K^n d^{2s} \leq C n^{-2r} d^{2s}, \tag{2.3}$$

where $r > \frac{1}{2}$.

Theorem 2.6. *Suppose Assumption 2.5, Assumption 2.1.ii. and Assumption 2.3 are satisfied, and that $f : \mathcal{X} \to \mathbb{R}$ is s-Hölder continuous with respect to d, for some $s \in (0, 1]$. Then there exist choices for a_i and $\mathbb{P}(N \geq i)$, such that*

$$Z = \sum_{i=0}^{N} \frac{\Delta_i}{\mathbb{P}(N \geq i)}$$

is an unbiased estimator of $\mathbb{E}_\pi[f]$ with finite variance and finite expected computing time. In particular, an example of such choices is $a_i \asymp i^k$ and $\mathbb{P}(N \geq i) \propto i^{-2rk+2+\epsilon}$ for $k > \frac{3}{2s-1}$ and any $0 < \epsilon < -3 - (1 - 2s)k$.

Remark 2.7. The Assumption 2.5 can be verified using drift conditions and coupling sets which are provided in the article [10]. Note that in this case it is not even clear that the ergodic average of the underlying Markov chain satisfies a central limit theorem, while the construction above remains valid. For $r \leq \frac{1}{2}$, the decay of $\|\Delta_i\|_2$ is not fast enough to allow for Z to have both finite variance and finite expected computing time.

Remark 2.8. Using Proposition C.1 in the Supplementary Material [3] which generalises Proposition 1.1, it is straightforward to check that Theorems 2.4 and 2.6 can be extended to hold for estimating expectations with respect to π of functions $f : \mathcal{X} \to H$ which are s-Hölder continuous with respect to d, where $(H, \langle\cdot,\cdot\rangle_H, \|\cdot\|_H)$ is a Hilbert space.

Remark 2.9. This section is a genuine generalisation of Section 3.4 of [26]. In this reference, the authors consider Markov chains that can be represented through iterated random functions which satisfy

$$X_{n+1} = \varphi_n(X_n) = \varphi(X_n, \xi_n),$$

with ξ_n independent and identically distributed, without loss of generality, U[0, 1] random variables. Under the assumption that

$$\sup_{x \neq y} \mathbb{E}\left(\frac{d(\varphi(x), \varphi(y))}{d(x, y)}\right)^{2\gamma} = r < 1, \tag{2.4}$$

for some $r < 1$ the general procedure can be applied to $\Delta_i = f(\tilde{X}_i) - f(\tilde{X}_{i-1})$ where $\tilde{X}_i$ is the backwards chain discussed in Section 1.2. In the language of the present section, the coupling in [26], Section 3.4, is specified through the random function, that is we can use

$$K\big((x, y), (d\tilde{x}, d\tilde{y})\big) = \mathcal{L}\big(\varphi(x, \xi), \varphi(y, \xi)\big)$$

which turns (2.4) into (2.1). We now show an example of a coupling of a Markov chain that leads to a faster contraction in (2.1) than any representation of the Markov chain through a random function. In particular, the coupling cannot be represented by a random function. Consequently, this section indeed genuinely generalises the results of [26].

Example 2.10. Consider the Markov chain given by

$$X_{n+1} \sim (X_n + U) \mod 2\pi, \tag{2.5}$$

where $U \sim \mathrm{U}[-2, 2]$. We denote the corresponding transition kernel by $P(x, \cdot)$ and note that it is of the form $P(x, \cdot) = U(A_x)$ where $A_x \subset [0, 2\pi]$. It is easy to check that $|A_x \cap A_{\tilde{x}}| \geq 8 - 2\pi$ for any $x, \tilde{x} \in [0, 2\pi]$, so that we have coupling probability for the 1-step maximal coupling at least $\frac{8-2\pi}{4} = 2 - \frac{\pi}{2} > \frac{1}{3}$. More precisely, this maximal coupling can be written as

$$Q(x_1, x_2) = \mathcal{L}\big((Y, Y)\mathbb{1}_{[0, \frac{|A_{x_1} \cap A_{x_2}|}{4}]}(U) + (Y_1, Y_2)\big(\mathbb{1}_{(\frac{|A_{x_1} \cap A_{x_2}|}{4}, 1]}(U)\big)\big),$$

where $Y_i \sim U(A_{x_i} \setminus A_{x_1} \cap A_{x_2})$, $Y \sim U(A_{x_1} \cap A_{x_2})$ and $U \sim \mathrm{U}[0, 1]$ are independent random variables. Note that this coupling clearly satisfies Assumption 2.1 with

$$Qd \leq \left(1 - \frac{8 - 2\pi}{4}\right)d,$$

for d the discrete metric.

In contrast, suppose there is a random function $\varphi(\cdot, \xi)$ such that $P(x, \cdot) = \mathcal{L}(\varphi(x, \xi))$ and

$$\mathbb{E}d\big(\varphi(x, \xi), \varphi(y, \xi)\big) < \frac{2}{3}d(x, y), \tag{2.6}$$

for every $x, y \in [0, 2\pi]$. Then consider the three points: $x_1 = 0$, $x_2 = \frac{2\pi}{3}$, $x_3 = \frac{4\pi}{3}$. It is easy to check that the minorisation measures between $P(x_1, \cdot)$ and $P(x_2, \cdot)$, $P(x_2, \cdot)$ and $P(x_3, \cdot)$, and $P(x_1, \cdot)$ and $P(x_3, \cdot)$, necessarily lie in the intervals $[0, \frac{2\pi}{3}]$, $[\frac{2\pi}{3}, \frac{4\pi}{3}]$ and $[\frac{4\pi}{3}, 2\pi]$, respectively (that is, $A_{x_1} \cap A_{x_2} \subset [0, \frac{2\pi}{3}]$, $A_{x_2} \cap A_{x_3} \subset [\frac{2\pi}{3}, \frac{4\pi}{3}]$ and $A_{x_1} \cap A_{x_3} \subset [\frac{4\pi}{3}, 2\pi]$). This observation implies that the sets

$$\begin{aligned} &\big\{\xi \mid \varphi(x_1, \xi) = \varphi(x_2, \xi)\big\}, \\ &\big\{\xi \mid \varphi(x_1, \xi) = \varphi(x_3, \xi)\big\} \quad \text{and} \\ &\big\{\xi \mid \varphi(x_2, \xi) = \varphi(x_3, \xi)\big\} \end{aligned}$$

are pairwise disjoint. Since d is the discrete metric, for (2.6) to hold each of the above sets needs to have probability exceeding $\frac{1}{3}$. This is a contradiction.

3. Unbiased estimation for Bayesian linear inverse problems

In this section, we consider the problem of estimating expectations with respect to the posterior distribution arising in Bayesian linear inverse problems in function space. We assume Gaussian prior and noise distributions, hence the posterior is available analytically and the only source of bias is the discretisation. We show that Glynn and Rhee's programme can directly be adapted in this setting to perform unbiased estimation of posterior expectations.

3.1. Setup

We work in a separable Hilbert space $(\mathcal{X}, \langle\cdot,\cdot\rangle, \|\cdot\|)$ and consider the inverse problem of finding an unknown function $u \in \mathcal{X}$ from a blurred, noisy observation y. In particular, we consider the data model

$$y = Ku + \xi,$$

where $\xi \sim \mathcal{N}(0, I)$ is additive Gaussian white noise and K is the forward operator which is assumed to be linear and bounded. We put a Gaussian prior $\mu_0 = \mathcal{N}(0, \mathcal{C}_0)$ on the unknown u, where $\mathcal{C}_0$ is a positive definite, selfadjoint and trace class linear operator. We make the following assumption on the operators K and $\mathcal{C}_0$.

Assumption 3.1. *The linear operators K and $\mathcal{C}_0$ commute with each other and K^*K and $\mathcal{C}_0$ are mutually diagonalizable with common complete orthonormal basis $\{e_\ell\}_{\ell\in\mathbb{N}}$ in $\mathcal{X}$. In particular, there exist $p \geq 0, a > \frac{1}{2}$ such that the eigenvalues of K^*K and $\mathcal{C}_0$ decay as ℓ^{-4p} and ℓ^{-2a}, respectively.*

In this diagonal setting, it is straightforward to check that the posterior, denoted by μ^y, is also Gaussian almost surely with respect to the joint distribution of (u, y), [2]. We hence have $\mu^y = \mathcal{N}(m, \mathcal{C})$, where the mean and precision operator (inverse covariance) are given by

$$\mathcal{C}^{-1} = \mathcal{C}_0^{-1} + K^*K, \tag{3.1}$$

$$\mathcal{C}^{-1}m = K^*y. \tag{3.2}$$

We make the following assumption concerning the observed data.

Assumption 3.2. *We have a fixed realisation of the data, y, which has the regularity of the noise, that is, there exist $c_-, c_+ > 0$ such that for all $\ell \in \mathbb{N}$, $y_\ell = \langle y, e_\ell\rangle \in (c_-, c_+)$.*

This assumption is reasonable, since given that $u \in \mathcal{X}$, in order to have that the inverse problem is ill-posed and hence worthy of consideration, the noise needs to be outside of the range of K. This means that the noise needs to be the roughest part of the data.

Gaussianity suggests that we can in theory draw exactly from $u|y \sim \mathcal{N}(m, \mathcal{C})$, however in practice this is impossible to achieve in finite time due to the infinite-dimensionality of the posterior. In the present setting, the approximation is achieved by considering truncations of the Karhunen–Loeve expansion of $\mathcal{N}(m, \mathcal{C})$. Let $x \sim \mathcal{N}(\mu, \Sigma)$ be a Gaussian random variable in $\mathcal{X}$, where $\mu \in \mathcal{X}$ and $\Sigma : \mathcal{X} \to \mathcal{X}$ is a selfadjoint, positive definite and trace class linear operator in $\mathcal{X}$. Then the operator Σ possesses a set of eigenvalue-eigenfuction pairs $\{\sigma_\ell, \psi_\ell\}_{\ell\in\mathbb{N}}$, where $\sigma_\ell > 0,\ \ell \in \mathbb{N}$ are summable and $\{\psi_\ell\}_{\ell\in\mathbb{N}}$ forms a complete orthonormal basis in $\mathcal{X}$. We can then write $x = \sum_{\ell=1}^{\infty}(\mu_\ell + \sqrt{\sigma_\ell}\gamma_\ell)\psi_\ell$, where $\mu_\ell = \langle\mu, \psi_\ell\rangle$ and $\{\gamma_\ell\}_{\ell\in\mathbb{N}}$ are independent and identically distributed standard Gaussian random variables in $\mathbb{R}$; this is the Karhunen–Loeve expansion of x, [1].

In particular, the Gaussian random variable $u|y \sim \mathcal{N}(m, \mathcal{C})$ can be written as

$$u|y = \sum_{\ell=1}^{\infty} (m_\ell + \sqrt{c_\ell}\zeta_\ell) e_\ell,$$

where c_ℓ are the eigenvalues of $\mathcal{C}$ (which is also diagonalizable in the basis $\{e_\ell\}_{\ell\in\mathbb{N}}$), ζ_ℓ are independent and identically distributed standard normal random variables and $m^\ell = \langle m, e_\ell\rangle$. One can then define the approximations $u^i|y^i$ of $u|y$ at level $i \in \mathbb{N}$, by truncating its Karhunen–Loeve expansion to the first j_i terms,

$$u^i|y^i := \sum_{\ell=1}^{j_i} (m_\ell + \sqrt{c_\ell}\zeta_\ell) e_\ell,$$

where $\{j_i\}_{i\in\mathbb{N}}$ is an increasing sequence of positive integers. Using equations (3.2) and (3.1), together with Assumption 3.1, we get that

$$u^i|y^i = \sum_{\ell=1}^{j_i} \frac{\ell^{-2p} y_\ell}{\ell^{2a} + \ell^{-4p}} e_\ell + \sum_{\ell=1}^{j_i} \frac{\zeta_\ell}{(\ell^{2a} + \ell^{-4p})^{\frac{1}{2}}} e_\ell. \tag{3.3}$$

Approximating expectations with respect to the posterior μ^y, by expectations with respect to the laws μ_j^y of the truncated Karhunen–Loeve expansion, introduces a bias. In the next subsection, we demonstrate how Glynn and Rhee's unbiased estimation programme for SDE's (see Section 1.1), can be applied directly in the setting of linear inverse problems to obtain unbiased estimates of expectations with respect to the posterior μ^y.

3.2. Main result

Suppose that we want to estimate $\mathbb{E}_{\mu^y}[f] = \mathbb{E}[Y]$, where $Y := f(u|y)$ and $f : \mathcal{X} \to \mathbb{R}$ is s-Hölder continuous for some $s \in (0, 1]$. We define the approximations $Y_i = f(u^i|y^i)$ for $i \in \mathbb{N}$ and as in Section 1.1 the differences $\Delta_i = Y_i - Y_{i-1}$, where $Y_{-1} := 0$. We make the following assumption which will be needed for controlling the expected computing time of the proposed estimator.

Assumption 3.3. *The expected computing time t_i for generating Δ_i satisfies*

$$t_i \lesssim j_i.$$

This is a reasonable assumption, since we require j_i Gaussian draws to produce $u^i|y^i$. We have the following result on the estimator Z defined in equation (1.1), which holds under Assumptions 3.1, 3.2, 3.3:

Theorem 3.4. *Let $f : \mathcal{X} \to \mathbb{R}$ be s-Hölder continuous for some $s \in (0, 1]$ and assume $a > \frac{1+s}{2s}$, that is, that the eigenvalues of the prior covariance decay sufficiently fast. Then, there exist* *choices of* j_i *and* $\mathbb{P}(N \geq i)$, *such that*

$$\tilde{Z} = \sum_{i=0}^{N} \frac{\tilde{\Delta}_i}{\mathbb{P}(N \geq i)}$$

is an unbiased estimator of $\mathbb{E}_{\mu^y}[f]$ *with finite variance and finite expected computing time. Here, as in Proposition* 1.1, *each* $\tilde{\Delta}_i$ *is an independent copy of* Δ_i *as defined above. In particular, two examples of such choices are*:

(i) $j_i = 2^i$ *and* $\mathbb{P}(N \geq i) \propto 2^{\frac{(2-\epsilon)s(1-2a)i}{2}}$, *for any* $\epsilon \in (0, \frac{2+2s-4as}{s(1-2a)})$;

(ii) $j_i \lesssim i^q$, *and* $\mathbb{P}(N \geq i) \propto i^{s(q-1-2aq)+2+\epsilon}$, *for* $q > \frac{s-3}{1+s-2as}$ *and for any* $\epsilon \in (0, s-3-q(1+s-2as))$.

The assumption on the regularity of the prior, $a > \frac{1+s}{2s}$, in Theorem 3.4, is more severe than the usual $a > \frac{1}{2}$ which is required for the formulation of the Bayesian linear inverse problem. In Section A in the Supplementary Material [3], we show how to modify the estimator Z in order to relax this assumption, in the case that f is a linear functional hence Lipschitz continuous.

Remark 3.5. Using Proposition C.1 in the Supplementary Material [3] which generalises Proposition 1.1, it is straightforward to check that Theorem 3.4 can be extended to hold for estimating posterior expectations of functions $f : \mathcal{X} \to H$ which are s-Hölder continuous where $(H, \langle\cdot,\cdot\rangle_H, \|\cdot\|_H)$ is a Hilbert space. In particular, the theorem holds for unbiased estimation of the posterior mean.

4. Unbiased estimation for Bayesian inverse problems with uniform priors, using the independence sampler

In this section, we consider infinite dimensional state spaces and extend the considerations of Glynn and Rhee on unbiased estimation of expectations with respect to the limiting distribution of a Markov chain, to remove not only the bias introduced due to the use of the finite-time distributions as approximations of the target distribution, but also the bias introduced due to the necessity to discretise. For expository reasons, we do this in an idealised nonlinear Bayesian inverse problem setting, and present an unbiased version of the independence sampler to approximate expectations with respect to the posterior. Later on in Section 5, we extend our results to more elaborate settings and present an unbiased version of the preconditioned Crank–Nicholson algorithm.

4.1. Setup

We consider the inverse problem of finding an unknown function u from noisy indirect observations $y \in \mathbb{R}^d$. We assume the data model

$$y = G(u) + \eta,$$

where $G : \mathcal{X} \to \mathbb{R}^d$ is the observation operator and $\eta \sim N(0, I)$ is the observational noise. A typical example in the inverse problems literature, is the situation that G maps the diffusion coefficient u of an elliptic partial differential equation, to the solution evaluated at a set of finite points [8], Section 3.4. Henceforward, we identify the function u with a sequence $u = \{u_k\}_{k\in\mathbb{N}} \in \mathbb{R}^\infty$, which represents the coefficients of the unknown function in some series expansion.

Let $u_k^\star \downarrow 0$ and consider the sequence of j-dimensional state spaces

$$\mathcal{X}_j = \prod_{k=1}^{j} [-u_k^\star, u_k^\star],$$

assumed to be embedded in the infinite dimensional state space $\mathcal{X} := \prod_{k=1}^\infty [-u_k^\star, u_k^\star]$. We denote by Π_j the projection onto $\mathcal{X}_j \subset \mathcal{X}$, $\Pi_j : \mathcal{X} \to \mathcal{X}_j$, $\Pi_j u = (u_1, \dots, u_j, 0, \dots)$. The reader should think of an element $u \in \mathcal{X}$ as the collection of coefficients (for example Fourier) of a function which decay at a prescribed rate. Depending on the particular expansion used, the decay of the coefficients translates to smoothness of the corresponding function. We put a uniform prior on $u \in \mathcal{X}$,

$$\mu_0 = \bigotimes_{k=1}^{\infty} (\lambda \mid_{[-u_k^*, u_k^*]}), \tag{4.1}$$

where λ denotes the Lebesque measure, treating all components as uniformly distributed over the range and independent of all the other components. Such priors have been used in the inverse problem setting in [30]; see again [8], Section 3.4, for a less technical version. In particular, in these references it is shown that under certain conditions on the basis used in the series expansion and on the continuity and boundedness of the forward operator G, the posterior distribution μ^y of $u|y$ is well defined and given by

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp\bigl(-\|y - G(u)\|_{\mathbb{R}^d}^2\bigr).$$

However, in general μ^y is not available in closed form and on the contrary it can be a very complicated infinite dimensional probability measure. In order to probe the posterior, one needs to discretise and sample. We discuss how to do this naively using an independence sampler in Section 4.2, while in Section 4.3 we modify the independence sampler to achieve unbiased estimation of expectations with respect to μ^y.

4.2. Approximations to the forward problem and a naive independence sampler

In the assumed inverse problem setting, it is natural to discretise u in $\mathcal{X}_j$ and to approximate the observation operator G by $G_j : \mathcal{X} \to \mathbb{R}^d$ which depends on u only through the projection $\Pi_j u$, that is

$$G_j(u) = G_j(\Pi_j u). \tag{4.2}$$

We use the notation $G_\infty = G$ and work under the following assumption.

Assumption 4.1. *There exists some* $\beta > 1$, *such that the observation operator and its approximations satisfy*

$$\sup_{u \in \mathcal{X}} \| G_j(u) - G(u) \|_{\mathbb{R}^d} \lesssim j^{-\beta},$$

$$\sup_{u \in \mathcal{X}} \| G(u) \|_{\mathbb{R}^d} < \infty.$$

Notice that Assumption 4.1 implies

$$\sup_{u \in \mathcal{X}} \| G_j(u) - G_{\tilde{j}}(u) \|_{\mathbb{R}^d} \lesssim (j \wedge \tilde{j})^{-\beta}. \tag{4.3}$$

For a concrete example of G and the relevant discretisations, which satisfies Assumption 4.1 see Section D in the Supplementary Material [3].

We define the projected priors $\mu_{0,j} = \Pi_{j\star}\mu_0$ on $\mathcal{X}_j$, which combined with the approximation of the observation operator give rise to the approximate posteriors

$$\frac{d\mu_j^y}{d\mu_{0,j}}(u) \propto \exp\left(-\frac{1}{2} \| y - G_j(u) \|_{\mathbb{R}^d}^2 \right).$$

Approximating an expectation with respect to μ^y by an expectation with respect to μ_j^y, results in a discretisation error which is quantified in [7,19] and [8]. Moreover, the expectations with respect to μ_j^y are not available analytically but they are amenable to approximation using Markov chain Monte Carlo algorithms. Again for illustration, we consider the (regular) independence sampler, the Metropolis–Hastings algorithm arising from the state-independent proposal $\mu_{0,j}$, see Algorithm 2. We denote the resulting Markov chain by $X_\cdot^j$ and its transition kernel by P_j.

Algorithm 2 Independence sampler

Generate X_0^j. Iterate the following steps for $k = 1, \ldots, K_{\max}$:

1. $\xi^j \sim \mu_{0,j}$
2. set $X_{k+1}^j = \xi^j$ with probability

$$\alpha_j\left(X_k^j, \xi^j\right) = 1 \wedge \exp\left(\frac{1}{2} \| y - G_j\left(X_k^j\right) \|_{\mathbb{R}^d}^2 - \frac{1}{2} \| y - G_j\left(\xi^j\right) \|_{\mathbb{R}^d}^2 \right) \tag{4.4}$$

and $X_{k+1}^j = X_k^j$ otherwise.

It is shown in [32], that the boundedness of G_j, implies a deterministic lower bound on the acceptance probability

$$\alpha_j \geq \alpha_\star > 0. \tag{4.5}$$

In this case, the Monte Carlo error can be controlled explicitly because the Markov chain $X_\cdot^j$ is uniformly ergodic due to (4.5), [19]. The overall error in the approximation

$$\mathbb{E}_{\mu^y}[f] \approx \mathbb{E}_{\mu_j^y}[f] \approx \frac{1}{K_{\max}} \sum_{k=1}^{K_{\max}} f\bigl(X_k^j\bigr)$$

has two contributions, the Monte Carlo error and the discretisation error. In particular, the discretisation error is chosen at the beginning of the MCMC computation and can only be reduced by restarting the computations from scratch. In the next subsection, we formulate the modified independence sampler which leads to unbiased estimation of posterior expectations.

4.3. Unbiased estimation using the independence sampler

We now present a version of the independence sampler which leads to the removal of both the bias due to the use of the finite-time distributions and the bias due to the discretisation of the posterior.

We use the unbiasing programme of Glynn and Rhee as introduced in Sections 1.1 and 1.3, in order to construct an unbiased estimator Z of the posterior expectation $\mathbb{E}_{\mu^y}[f]$, for some function $f : \mathcal{X} \to \mathbb{R}$. For two increasing sequences of integers a_i and j_i, representing the time-step and the discretisation level respectively, we would like to set $\Delta_i = f(X_{a_i}^{j_i}) - f(X_{a_{i-1}}^{j_{i-1}})$ in the definition of Z in Proposition 1.1, where the chains $X_\cdot^{j_i}$ and $X_\cdot^{j_{i-1}}$ are the (regular) independence sampler chains introduced in the previous subsection following the transition kernels P_{j_i} and $P_{j_{i-1}}$, respectively. For the unbiasing technique to work, we need to construct an appropriate coupling between the two chains, so that $\|\Delta_i\|_2$ decays sufficiently quickly for Proposition 1.1 to apply, and the expected computing time is finite. In order to achieve this, we generate Δ_i using a "top" level chain in $\mathcal{X}_{j_i}$ and a "bottom" level chain in $\mathcal{X}_{j_{i-1}}$, which we denote by $\mathcal{T}_\cdot^i$ and $\mathcal{B}_\cdot^i$, and which perform a_i and a_{i-1} steps, respectively. According to Proposition 1.1, we need Δ_i to be independent for different i, hence the two chains $\mathcal{T}_\cdot^i$ and $\mathcal{B}_\cdot^{i+1}$ both following the transition kernel P_{j_i} in $\mathcal{X}_{j_i}$, are constructed independently. Nevertheless, the chains at different levels are coupled as follows:

1. $\mathcal{T}_\cdot^i$ is coupled to $\mathcal{B}_\cdot^i$ which follows the transition kernel $P_{j_{i-1}}$ on $\mathcal{X}_{j_{i-1}}$;
2. $\mathcal{B}_\cdot^{i+1}$ is coupled to $\mathcal{T}_\cdot^{i+1}$ which follows the transition kernel $P_{j_{i+1}}$ on $\mathcal{X}_{j_{i+1}}$.

The following diagram illustrates the construction of the Δ_i:

$$\begin{array}{rcccccl}
 & & & & x_0 = \mathcal{T}^0_{-a_0} & \dots & \mathcal{T}^0_0 \;\}\Delta_0 = f(\mathcal{T}^0_0) \\
 & & & & x_0 = \mathcal{B}^1_{-a_0} & \dots & \mathcal{B}^1_0 \\
 & & & & | & | & | \;\;\}\Delta_1 = f(\mathcal{T}^1_0) - f(\mathcal{B}^1_0) \\
 & & x_0 = \mathcal{T}^1_{-a_1} & \dots & \mathcal{T}^1_{-a_0} & \dots & \mathcal{T}^1_0 \\
 & & x_0 = \mathcal{B}^2_{-a_1} & \dots & \mathcal{B}^2_{-a_0} & \dots & \mathcal{B}^2_0 \\
 & & | & | & | & | & | \;\;\}\Delta_2 = f(\mathcal{T}^2_0) - f(\mathcal{B}^2_0) \\
x_0 = \mathcal{T}^2_{-a_2} & \dots & \mathcal{T}^2_{-a_1} & \dots & & \dots & \mathcal{T}^2_0 \\
x_0 = \mathcal{B}^3_{-a_2} & \dots & \mathcal{B}^3_{-a_1} & \dots & & \dots & \mathcal{B}^3_0
\end{array}$$

Here | indicates coupling between two chains. We would like to point out a connection to Multilevel Markov Chain Monte Carlo (MLMCMC) [9,16,19]. Both the present method and MLMCMC couple Markov chains in different dimensions. However, the method presented in this section can be seen as taking a diagonal approach between the unbiasing approach of [26] and the MLMCMC idea; this also applies for our method of coupling pCN algorithms presented in the next section. More precisely, we couple Markov chains in different dimensions performing a different number of steps. In this way, we remove the bias due to both discretisation and the finite number of iterations. In contrast in MLMCMC both contributions to the bias remain, however it achieves an efficient distribution of computations between discretisation levels, which reduces the cost of producing estimators with a certain error level compared to standard MCMC.

The couplings above arise form the minorisation due to the lower bound on the acceptance probability. They can be represented using the random functions $\varphi^i_{\mathcal{T}}$ and $\varphi^i_{\mathcal{B}}$, defined as:

$$\begin{aligned}
\varphi^i_{\mathcal{T}}(x, W^i) &= \mathbb{1}_{[0,\alpha_\star]}(U^i_1)\xi^i_1 \\
&\quad + \mathbb{1}_{(\alpha_\star,1]}(U^i_1)\Big(\mathbb{1}_{[0,\frac{\alpha_{j_i}(x,\xi^i_2)-\alpha_\star}{1-\alpha_\star}]}(U^i_2)\xi^i_2 + \mathbb{1}_{(\frac{\alpha_{j_i}(x,\xi^i_2)-\alpha_\star}{1-\alpha_\star},1]}(U^i_2)x\Big), \\
\varphi^i_{\mathcal{B}}(x, W^i) &= \mathbb{1}_{[0,\alpha_\star]}(U^i_1)\Pi_{j_i-1}\xi^i_1 \\
&\quad + \mathbb{1}_{(\alpha_\star,1]}(U^i_1)\Big(\mathbb{1}_{[0,\frac{\alpha_{j_i-1}(x,\Pi_{j_i-1}\xi^i_2)-\alpha_\star}{1-\alpha_\star}]}(U^i_2)\Pi_{j_i-1}\xi^i_2 + \mathbb{1}_{(\frac{\alpha_{j_i-1}(x,\Pi_{j_i-1}\xi^i_2)-\alpha_\star}{1-\alpha_\star},1]}(U^i_2)x\Big),
\end{aligned} \tag{4.6}$$

where $W^i = (U^i_1, U^i_2, \xi^i_1, \xi^i_2)$, for $U^i_l \sim \mathrm{U}[0, 1]$ and $\xi^i_l \sim \mu^{j_i}_0$, $l = 1, 2$, which are all independent of each other.

The functions $\varphi^i_{\mathcal{T}}$ and $\varphi^i_{\mathcal{B}}$ are constructed by minorising the transition kernels P_{j_i} and $P_{j_{i-1}}$, using the proposal distributions $\mu^{j_i}_0$ and $\mu^{j_{i-1}}_0$, respectively. The uniform random variable U^i_1 is used to construct the "coin" for switching between the minorising measure and the residual kernel. The residual kernel is still a Metropolis–Hastings kernel with a corrected acceptance probability and U^i_2 is used for acceptance and rejection. The coupling between the "top" and "bottom" chains used to construct Δ_i, will be achieved through the use of the same random seeds in the random functions $\varphi^i_{\mathcal{T}}$ and $\varphi^i_{\mathcal{B}}$.

The construction of Δ_i is given in detail in Algorithm 3. In Lemma 8.1, we derive bounds on the decay of $\|\Delta_i\|_2$ which are sufficient for the unbiasing programme to work. In order to achieve

Algorithm 3 Coupled independence samplers for unbiased estimation

```
Fix a starting point $x_0 \in \mathcal{X}_{j_0}$ once and for all. For $i = 0$, generate $\Delta_0$ as follows:
    1. set $\mathcal{T}^0_{-a_0} = x_0$ on $\mathcal{X}_{j_0}$ and simulate according to Algorithm 2 up to $\mathcal{T}^0_0$;
    2. set $\Delta_0 = f(\mathcal{T}^0_0)$.
For $i \geq 1$, generate $\Delta_i$ as follows:
    1. set $\mathcal{T}^i_{-a_i} = x_0$ and simulate according to Algorithm 2 upto $\mathcal{T}^i_{-a_{i-1}}$ in dimension $j_i$;
    2. set $\mathcal{B}^i_{-a_{i-1}} = x_0$;
    3. for $k = -a_{i-1} + 1, \dots, 0$ simulate $\mathcal{T}^i_k$ and $\mathcal{B}^i_k$ as coupled independence samplers as described below:
        (a) draw $U^i_l \overset{\text{i.i.d.}}{\sim} \mathrm{U}[0,1]$ and $\xi^i_l \overset{\text{i.i.d.}}{\sim} \mu^{j_i}_0$ for $l = 1, 2$ independently from everything else and
            set $W^i = (U^i_1, U^i_2, \xi^i_1, \xi^i_2)$ as the collection of all random input to do the $k$th step;
        (b) set
                $\mathcal{T}^i_k = \varphi^i_{\mathcal{T}}(\mathcal{T}^i_{k-1}, W^i),$
                $\mathcal{B}^i_k = \varphi^i_{\mathcal{B}}(\mathcal{B}^i_{k-1}, W^i);$                    (4.7)
    4. set $\Delta_i = f(\mathcal{T}^i_0) - f(\mathcal{B}^i_0)$.
```

this, we use the decomposition

$$\begin{aligned} \|\Delta_i\|_2^2 &\leq 2\big\|f(\mathcal{B}^i_0) - f(\Pi_{j_{i-1}}\mathcal{T}^i_0)\big\|_2^2 + 2\big\|f(\mathcal{T}^i_0) - f(\Pi_{j_{i-1}}\mathcal{T}^i_0)\big\|_2^2 \\ &:= 2(E_1 + E_2). \end{aligned} \tag{4.8}$$

The first term E_1 measures the difference in the lower level of the two coupled chains $\mathcal{T}^i_\cdot$ and $\mathcal{B}^i_\cdot$ used to generate Δ_i, while E_2 has to do with the dependence of the function f on higher modes. By the definition of the couplings, see (4.6), it is clear that in order to control E_1 it suffices to make sure that the two chains have the same acceptance behaviour with high probability; we use Assumption 4.1 and the implied uniform ergodicity to show this. On the other hand, to control E_2 we make the following assumption on f:

Assumption 4.2. *We assume that* $f : \mathcal{X} \to \mathbb{R}$ *satisfies*

$$\sup_{x \in \mathcal{X}} \big| f(\Pi_j x) - f(\Pi_{\tilde{j}} x)\big| \lesssim (j \wedge \tilde{j})^{-\frac{\kappa}{2}},$$

for some $\kappa > 1$.

Note that for specific examples of the function f, the last assumption is essentially an assumption on the decay of the sequence defining the space $\mathcal{X}$, $u^\star_k$. Under our assumptions, in Lemma 8.1, we derive bounds on the decay of $\|\Delta_i\|_2$ which are sufficient for the unbiasing procedure to work.

In order to control the expected computing time of the estimator Z, we make the following assumption on the cost of generating Δ_i.

Assumption 4.3. *Let $r := \beta \wedge \kappa$, where $\beta, \kappa > 1$ are defined in Assumptions 4.1 and 4.2, respectively. We assume that the computational cost of one step of the chain at level j_i is*

$$s_i \lesssim j_i^{\theta},$$

with $\theta < r$. Therefore, since we need a_i steps of the chain to generate Δ_i, the expected computing time t_i of Δ_i satisfies

$$t_i \lesssim a_i j_i^{\theta}.$$

Remark 4.4. The simultaneous validity of Assumptions 4.1, 4.2 and 4.3 depends on a relationship between the properties of G, the regularity of f and, most importantly, the smoothness of the space $\mathcal{X}$ as expressed by the decay of the sequence $u_k^{\star}$. Making this explicit in full generality is beyond the scope of this paper, however we do provide an example in Section D of the Supplementary Material [3].

We have the following result on the estimator Z defined in equation (1.1):

Theorem 4.5. *Suppose that the forward model satisfies Assumption 4.1 with $\beta > 1$ and the observable $f : \mathcal{X} \to \mathbb{R}$ satisfies Assumption 4.2 with $\kappa > 1$ and let $r := \beta \wedge \kappa > 1$. Furthermore, assume that the computational cost of one step of the chain satisfies Assumption 4.3 with $\theta < r$. Then there is a choice of a_i, j_i and $\mathbb{P}(N \geq i)$, such that*

$$Z = \sum_{i=0}^{N} \frac{\Delta_i}{\mathbb{P}(N \geq i)}$$

is an unbiased estimator of $\mathbb{E}_{\mu^y}[f]$ with finite variance and finite expected computing time. For example this works for the choice $j_i = i^q$, $a_i \sim \frac{q\beta}{c_\star} \log(i)$, for $c_\star = -\log(1-\alpha_\star)$ and $\mathbb{P}(N \geq i) \propto i^{-t}$ where $q > \frac{3}{r-\theta}$ and $t \in (1+\theta q, rq-2)$. Note that under our assumptions the choices of q and t are simultaneously admissible.

Remark 4.6. Using Proposition C.1 in the Supplementary Material [3] which generalises Proposition 1.1, it is straightforward to check that Theorem 4.5 can be extended to hold for estimating expectations with respect to μ^y of functions $f : \mathcal{X} \to H$ which satisfy an assumption of the type of Assumption 4.2.

5. Unbiased estimation for Gaussian-based target measures, using coupled pCN algorithms

In Section 4, we showed that it is possible to couple the independence sampler in order to achieve unbiased estimation for an idealised Bayesian inverse problem setting in function space. Our coupling construction relied on assumptions on the inverse problem, which secured that the independence sampler is uniformly ergodic. However, for many measures of interest the independence sampler is not uniformly ergodic; in fact, if there exist areas of positive target measure, in which the density of the proposal with respect to the target vanishes, the independence sampler is not even geometrically ergodic [21].

In this section, we extend the methodology of the last section, and couple the preconditioned Crank–Nicholson (pCN) algorithm in order to perform unbiased estimation in more difficult situations. The pCN algorithm first appeared in [6] as the PIA algorithm, and recently has received a lot of interest from the Bayesian inverse problem community due to the fact that it is well defined in the function space setting. In particular, it was shown in [18] that pCN achieves a dimension-independent geometric rate of convergence for Gaussian-based target measures, that is, measures that have density with respect to a Gaussian measure. Below we also consider Gaussian-based target measures, and although we do not have uniform ergodicity of the pCN algorithm, we show that it is possible to perform unbiased estimation, by extending known contraction results for the pCN algorithm and using a combination of the techniques applied in Sections 2 and 4.

5.1. Setup

We work in a separable Hilbert space $(\mathcal{X}, \langle\cdot,\cdot\rangle, \|\cdot\|)$, and consider target mesaures μ which can be expressed as log-Lipschitz changes of measure from a Gaussian reference measure μ_0. In particular, let μ_0 be a Gaussian measure in $\mathcal{X}$ with Karhunen–Loeve expansion (see Section 3) of the form

$$\mu_0 = \mathcal{L}\left(\sum_{\ell=0}^{\infty} \sqrt{\lambda_\ell}\gamma_\ell e_\ell\right), \quad \gamma_\ell \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),\ \lambda_\ell \lesssim \ell^{-2a}, \tag{5.1}$$

where $\{e_\ell\}_{\ell\in\mathbb{N}}$ is a complete orthonormal basis in $\mathcal{X}$ and $a > \frac{1}{2}$ is a regularity parameter. We consider the target measure μ, given as

$$\frac{d\mu}{d\mu_0}(x) \propto \exp\bigl(-g(x)\bigr), \tag{5.2}$$

where $g : \mathcal{X} \to \mathbb{R}$ is Lipschitz continuous.

We define the approximate reference measures $\mu_{0,j}$ through the truncated Karhunen–Loeve expansion

$$\mu_{0,j} = \mathcal{L}\left(\sum_{l=0}^{j} \sqrt{\lambda_\ell}\gamma_\ell e_\ell\right).$$

The measures $\mu_{0,j}$ are then supported on the j-dimensional space $\mathcal{X}_j := \operatorname{span}\{e_1, \dots, e_j\} \subset \mathcal{X}$. In the following, we identify the spaces $\mathcal{X}_j$ with the corresponding subsets of $\mathcal{X}$ and denote by Π_j the projection onto $\mathcal{X}_j$. We consider the sequence of truncated target measures μ_j defined through

$$\frac{d\mu_j}{d\mu_{0,j}}(x) \propto \exp\bigl(-g(x)\bigr). \tag{5.3}$$

Algorithm 4 pCN algorithm

```
Fix $\rho \in (0,1)$. Generate $X_0^j$. Iterate the following steps for $k = 1, \ldots, K_{\max}$:
    1. $\xi^j \sim \mu_{0,j}$;
    2. set $\hat{X}_{k+1}^j = \rho X_k^j + \sqrt{1-\rho^2}\xi^j$;
    3. set $X_{k+1}^j = \hat{X}_{k+1}^j$ with probability
           $\alpha\big(X_k^j, \xi^j\big) = 1 \wedge \exp\big(g\big(X_k^j\big) - g\big(\hat{X}_{k+1}^j\big)\big)$
       and $X_{k+1}^j = X_k^j$ otherwise.
```

Approximating an expectation with respect to μ by an expectation with respect to μ_j, results in a discretisation error which is quantified in [7,19] and [8]. Furthermore, the expectations with respect to μ_j are in general not available analytically but they are amenable to approximation using Markov chain Monte Carlo algorithms. In particular, we consider the Markov chains corresponding to the pCN algorithms applied to μ^j, see Algorithm 4. We denote the resulting Markov chain by $X_{\cdot}^j$ and the corresponding Metropolis–Hastings Markov kernel by P_j. In a similar way to Section 4, in the next subsection we use appropriately coupled pCN algorithms to achieve the removal of both the discretisation bias as well as the bias introduced by the use of finite time distributions.

5.2. Unbiased estimation using the pCN algorithm

Our aim is to obtain an unbiased estimator of $\mathbb{E}_\mu[f]$, for some function $f : \mathcal{X} \to \mathbb{R}$. As in Section 4, for two increasing sequences of integers a_i and j_i we would like to set $\Delta_i = f(X_{a_i}^{j_i}) - f(X_{a_{i-1}}^{j_{i-1}})$ in the definition of Z in Proposition 1.1, where the chains $X_{\cdot}^{j_i}$ and $X_{\cdot}^{j_{i-1}}$ are the (regular) pCN chains introduced in the previous subsection following the transition kernels P_{j_i} and $P_{j_{i-1}}$, respectively. For the unbiasing technique to work, we need to construct an appropriate coupling between the two chains, so that $\|\Delta_i\|_2$ decays sufficiently quickly for Proposition 1.1 to apply, and the expected computing time is finite. In order to achieve this, we again generate Δ_i using a "top" level chain in $\mathcal{X}_{j_i}$ and a "bottom" level chain in $\mathcal{X}_{j_{i-1}}$, which we denote by $\mathcal{T}_{\cdot}^i$ and $\mathcal{B}_{\cdot}^i$, and which perform a_i and a_{i-1} steps, respectively. According to Proposition 1.1, we need Δ_i to be independent for different i, hence the two chains $\mathcal{T}_{\cdot}^i$ and $\mathcal{B}_{\cdot}^{i+1}$ both following the transition kernel P_{j_i} in $\mathcal{X}_{j_i}$, are constructed independently. Nevertheless, the chains at different levels are coupled as follows:

1. $\mathcal{T}_{\cdot}^i$ is coupled to $\mathcal{B}_{\cdot}^i$ which follows the transition kernel $P_{j_{i-1}}$ on $\mathcal{X}_{j_{i-1}}$;
2. $\mathcal{B}_{\cdot}^{i+1}$ is coupled to $\mathcal{T}_{\cdot}^{i+1}$ which follows the transition kernel $P_{j_{i+1}}$ on $\mathcal{X}_{j_{i+1}}$.

The following diagram illustrates the construction of the Δ_i:

$$
\begin{array}{rcrcrcll}
 & & & & x_0 = \mathcal{T}^0_{-a_0} & \dots & \mathcal{T}^0_0 & \}\Delta_0 = f(\mathcal{T}^0_0) \\
 & & & & x_0 = \mathcal{B}^1_{-a_0} & \dots & \mathcal{B}^1_0 & \\
 & & & & | & | & | & \}\Delta_1 = f(\mathcal{T}^1_0) - f(\mathcal{B}^1_0) \\
 & & x_0 = \mathcal{T}^1_{-a_1} & \dots & \mathcal{T}^1_{-a_0} & \dots & \mathcal{T}^1_0 & \\
 & & x_0 = \mathcal{B}^2_{-a_1} & \dots & \mathcal{B}^2_{-a_0} & \dots & \mathcal{B}^2_0 & \\
 & & | & | & | & | & | & \}\Delta_2 = f(\mathcal{T}^2_0) - f(\mathcal{B}^2_0) \\
x_0 = \mathcal{T}^2_{-a_2} & \dots & \mathcal{T}^2_{-a_1} & \dots & & \dots & \mathcal{T}^2_0 & \\
x_0 = \mathcal{B}^3_{-a_2} & \dots & \mathcal{B}^3_{-a_1} & \dots & & \dots & \mathcal{B}^3_0 &
\end{array}
$$

where | indicates coupling between two chains. The coupling is achieved by using the same random seed $\xi^i \sim \mu_{0,i}$ in the pCN proposal for $\mathcal{T}^i_\cdot$ and $\mathcal{B}^i_\cdot$, as well as the same uniform variable U^i for acceptance or rejection. The random variables ξ^i and U^i are taken to be independent of each other, as well as independent of ξ^j and U^j for $j \neq i$. We use the random functions $\varphi^i_{\hat{\mathcal{T}}}$, $\varphi^i_{\hat{\mathcal{B}}}$ to denote the pCN proposals $\hat{\mathcal{T}}^i_\cdot$ and $\hat{\mathcal{B}}^i_\cdot$ for the chains $\mathcal{T}^i_\cdot$ and $\mathcal{B}^i_\cdot$, respectively, where

$$\varphi^i_{\hat{\mathcal{T}}}(x, \xi^i) := \rho x + (1 - \rho^2)^{\frac{1}{2}} \xi^i,$$

$$\varphi^i_{\hat{\mathcal{B}}}(x, \xi^i) := \rho x + (1 - \rho^2)^{\frac{1}{2}} \Pi_{j_{i-1}} \xi^i.$$

Furthermore, we use the random functions $\varphi^i_{\mathcal{T}}$, $\varphi^i_{\mathcal{B}}$ to represent the chains $\mathcal{T}^i_\cdot$ and $\mathcal{B}^i_\cdot$, respectively, where

$$\varphi^i_{\mathcal{T}}(\mathcal{T}^i_{k-1}, W^i_k) := \mathbb{1}_{[0,\alpha(\mathcal{T}^i_{k-1}, \hat{\mathcal{T}}^i_k)]}(U^i_k)\hat{\mathcal{T}}^i_k + \mathbb{1}_{(\alpha(\mathcal{T}^i_{k-1}, \hat{\mathcal{T}}^i_k), 1]}(U^i_k)\mathcal{T}^i_{k-1},$$

$$\varphi^i_{\mathcal{B}}(\mathcal{B}^i_{k-1}, W^i_k) := \mathbb{1}_{[0,\alpha(\mathcal{B}^i_{k-1}, \hat{\mathcal{B}}^i_k)]}(U^i_k)\hat{\mathcal{B}}^i_k + \mathbb{1}_{(\alpha(\mathcal{B}^i_{k-1}, \hat{\mathcal{B}}^i_k), 1]}(U^i_k)\mathcal{B}^i_{k-1}.$$

The construction of Δ_i is given in detail in Algorithm 5.

For $W \sim \mu_{0,j_i} \otimes \mathrm{U}[0, 1]$, we define

$$K^{j_i}_{j_{i-1}}\bigl((x_1, x_2), \cdot\bigr) := \mathcal{L}\bigl(\varphi^i_{\mathcal{B}}(x_2, W), \varphi^i_{\mathcal{T}}(x_1, W)\bigr), \tag{5.4}$$

$$K^{j_i}_{j_i}\bigl((x_1, x_2), \cdot\bigr) := \mathcal{L}\bigl(\varphi^i_{\mathcal{T}}(x_1, W), \varphi^i_{\mathcal{T}}(x_2, W)\bigr), \tag{5.5}$$

that is, $K^{j_i}_{j_{i-1}}$ and $K^{j_i}_{j_i}$ are the couplings between $P_{j_{i-1}}(x_1, \cdot)$ and $P_{j_i}(x_2, \cdot)$ and $P_{j_i}(x_1, \cdot)$ and $P_{j_i}(x_2, \cdot)$, respectively. In order to simplify the notation we will write K_i for $K^{j_i}_{j_i}$.

In contrast to Section 4, in the present setting we do not have uniform ergodicity, thus we can only rely on the contracting property of the pCN algorithm in some distance (or distance-like function) d, in order to get the required decay of $\|\Delta_i\|_2$ for the unbiasing programme to work. We stress here, that the readily available results in the literature concern the contraction of the pCN algorithm at a fixed dimension j_i, and in particular the contraction of the coupling K_i in certain distances; see the results of Durmus and Moulines [10], Durmus et al. [11] and Hairer et al. [18].

Algorithm 5 Coupled pCN algorithms for unbiased estimation

```
Fix a starting point $x_0 \in \mathcal{X}_{j_0}$ once and for all. For $i = 0$, generate $\Delta_0$ as follows:
  1. set $\mathcal{T}^0_{-a_0} = x_0$ on $\mathcal{X}_{j_0}$ and simulate according to $P_{j_0}$ up to $\mathcal{T}^0_0$;
  2. set $\Delta_0 = f(\mathcal{T}^0_0)$.
For, $i \geq 1$, generate $\Delta_i$ as follows: We generate $\Delta_i$ for $i \geq 1$ as follows:
  1. set $\mathcal{T}^i_{-a_i} = x_0$ and run the chain until $\mathcal{T}^i_{-a_{i-1}}$ according to $P_{j_i}$;
  2. set $\mathcal{B}^i_{-a_{i-1}} = x_0$;
  3. for $k = -a_{i-1} + 1, \dots, 0$ run $\mathcal{T}^i_\cdot$ and $\mathcal{B}^i_\cdot$ as coupled pCN algorithms, as described below:
    (a) draw $\xi^i_k \sim \mu_{0,j_i}$ and $U^i_k \sim \mathrm{U}[0,1]$ independently from everything else and set $W^i_k = (\xi^i_k, U^i_k)$ as the collection of all random inputs for the $k$th step;
    (b) propose
          $\hat{\mathcal{T}}^i_k = \varphi^i_{\hat{\mathcal{T}}}(\mathcal{T}^i_{k-1}, \xi^i_k) = \rho\mathcal{T}^i_{k-1} + (1-\rho^2)^{\frac{1}{2}}\xi^i_k,$
          $\hat{\mathcal{B}}^i_k = \varphi^i_{\hat{\mathcal{B}}}(\mathcal{B}^i_{k-1}, \xi^i_k) = \rho\mathcal{B}^i_{k-1} + (1-\rho^2)^{\frac{1}{2}}\Pi_{j_{i-1}}\xi^i_k;$
    (c) set
          $\mathcal{T}^i_k = \varphi^i_{\mathcal{T}}(\mathcal{T}^i_{k-1}, W^i_k) = \mathbb{1}_{[0,\alpha(\mathcal{T}^i_{k-1},\hat{\mathcal{T}}^i_k)]}(U^i_k)\hat{\mathcal{T}}^i_k + \mathbb{1}_{(\alpha(\mathcal{T}^i_{k-1},\hat{\mathcal{T}}^i_k),1]}(U^i_k)\mathcal{T}^i_{k-1},$
          $\mathcal{B}^i_k = \varphi^i_{\mathcal{B}}(\mathcal{B}^i_{k-1}, W^i_k) = \mathbb{1}_{[0,\alpha(\mathcal{B}^i_{k-1},\hat{\mathcal{B}}^i_k)]}(U^i_k)\hat{\mathcal{B}}^i_k + \mathbb{1}_{(\alpha(\mathcal{B}^i_{k-1},\hat{\mathcal{B}}^i_k),1]}(U^i_k)\mathcal{B}^i_{k-1};$
  4. Set $\Delta_i = f(\mathcal{T}^i_0) - f(\mathcal{B}^i_0)$.
```

Instead, we need to work harder in order to show a form of contraction of the transdimensional coupling $K^{j_i}_{j_{i-1}}$ in the same distances, which happens asymptotically as $i \to \infty$. We achieve this by using the triangle inequality to combine the existing contraction results at a fixed level i, with estimates on the large i behaviour of the transdimensional coupling when the two chains are started from the same initial condition. Once we get the appropriate behaviour of $K^{j_i}_{j_{i-1}}$ in d, it is straightforward to obtain estimates of the decay of $\|\Delta_i\|_2$ in a similar way to Section 2, which hold for functions f having sufficient Hölder regularity in the same distance d.

We work under the following assumption on the log-change of measure g, which ensures the contraction of the pCN algorithm in a fixed state space (see Section 8.3.2 for details).

Assumption 5.1. *The function $g : \mathcal{X} \to \mathbb{R}$ is globally Lipschitz and there exist positive constants C, R_1, R_2, such that for $x \in \mathcal{X}$ with $\|x\| \geq R_1$*

$$\inf_{z \in B(\rho x, R_2)} \exp\bigl(g(x) - g(z)\bigr) > C, \tag{5.6}$$

where ρ is as in the definition of the pCN algorithm.

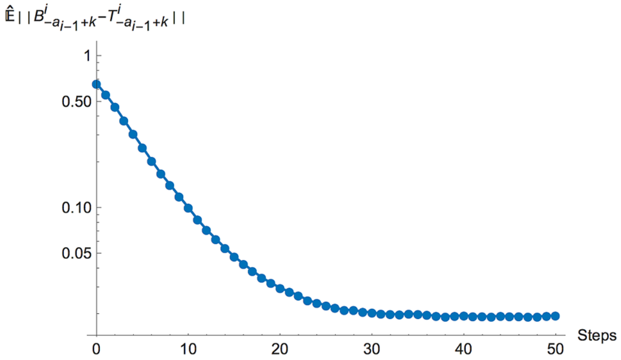

Figure 1. Estimates of $\mathbb{E}\|\mathcal{T}^i_{-a_{i-1}+k} - \mathcal{B}^i_{-a_{i-1}+k}\|$ based on 10 000 runs, plotted against k. Here $g(x) = \|x\|$, $\rho = 0.7$, $j_{i-1} = 15$, $j_i = 17$ and $a_i - a_{i-1} = 8$.

We first consider the distance $d_\tau = 1 \wedge \frac{\|x-y\|}{\tau}$. In Lemma 8.2, we derive results of the form

$$\mathbb{E}d_\tau\left(\mathcal{T}^i_0, \mathcal{B}^i_0\right) \leq Cr^{a_{i-1}} + C_{j_{i-1},j_i}, \tag{5.7}$$

where C_{j_{i-1},j_i} is a constant only depending on $j_{i-1} < j_i$, such that

$$C_{j_{i-1},j_i} \to 0 \qquad \text{as } i \to \infty.$$

Explicit bounds on C and r can be obtained as outlined in Section 8.3.2. We note here, that the bound in (5.7) agrees with the qualitative behaviour that we observe in simulations, see Figure 1.

We consider unbiased estimation of $\mathbb{E}_\mu[f]$, where f is s-Hölder for $s \in [\frac{1}{2}, 1]$ with respect to the distance d_τ. We note that this class of functions f does not depend on the choice of $\tau > 0$. For such a function f, the boundedness of the distance d_τ implies the bound

$$\|\Delta_i\|_2^2 \leq \|f\|_s^2 \mathbb{E}d_\tau\left(\mathcal{T}^i_0, \mathcal{B}^i_0\right)^{2s} \leq \|f\|_s^2 \mathbb{E}d_\tau\left(\mathcal{T}^i_0, \mathcal{B}^i_0\right). \tag{5.8}$$

Balancing the two terms on the right-hand side of (5.7), gives rise to sufficiently sharp bounds on $\|\Delta_i\|_2$, see Lemma 8.2 again.

In order to follow the unbiasing programme, we pose the following assumption on the expected computing time.

Assumption 5.2. *The expected computing time to simulate $K^{j_i}_{j_{i-1}}$ satisfies*

$$s_i \lesssim j_i^\theta$$

with $\theta \geq 1$. Therefore, since we need a_i steps of the chain to generate Δ_i, the expected computing time t_i of Δ_i satisfies

$$t_i \lesssim a_i j_i^\theta.$$

We have the following result on the estimator Z defined in (1.1).

Theorem 5.3. *Assume that the target measure μ is given as in (5.1), where g satisfies Assumption 5.1. Suppose that Assumption 5.2 is satisfied for $\theta \geq 1$ and let $f : \mathcal{X} \to \mathbb{R}$ be s-Hölder continuous with respect to d_τ, for some $s \in [\frac{1}{2}, 1]$. Assume that $a > \theta + \frac{1}{2}$, where a represents the regularity of the reference measure, see (5.1). Then there are choices of a_i, j_i and $\mathbb{P}(N \geq i)$, such that*

$$Z = \sum_{i=0}^{N} \frac{\Delta_i}{\mathbb{P}(N \geq i)}$$

is an unbiased estimate of $\mathbb{E}_\mu[f]$ with finite variance and finite expected computing time. For example, for any $m \in \mathbb{N}$, this works if we choose $a_i = mi$, $j_i \sim r^{\frac{2mi}{1-2a}}$ and $\mathbb{P}(N \geq i) \propto r^{(m-\epsilon)i}$, where $\epsilon \in (0, \frac{2\theta m}{1-2a} + m)$.

Note that the choice of m does not affect the finiteness of the variance or the expected computing time of Z. However, our intuition from the numerical experiments presented in Section 6 for problems of fixed dimension, suggests that a good choice of m has a large impact on the efficiency of the algorithm (see Figure 3). We expect this to be the case in the transdimensional setting too, and for this reason choose to allow this flexibility in the formulation of the theorem.

The last result shows that the unbiasing procedure can be applied for estimating posterior expectations with respect to functions that are Hölder continuous with respect to the bounded distance d_τ. In particular, f needs to be bounded which does not allow the estimation of the mean or the second moment. We now show that it is possible to obtain unbiased estimates for unbounded functions, under a stronger assumption on the regularity of the reference measure μ_0. This is achieved by considering the distance-like function $\tilde{d}(x, y) := \sqrt{d_\tau(x, y)(1 + V(x) + V(y))}$ with $V(x) = \exp(\|x\|)$.

Indeed, in Lemma 8.3 we obtain bounds of the form

$$\mathbb{E}\tilde{d}\big(\mathcal{T}_0^i, \mathcal{B}_0^i\big) \lesssim r^{a_{i-1}} + C_{j_{i-1}, j_i}^{\frac{1}{2}}, \tag{5.9}$$

where $C_{j_{i-1}}^{j_i}$ is the same constant as in (5.7) and $r \in (0, 1)$. Since $\tilde{d}$ is unbounded, a bound of the type of (5.8) is not possible for general $s \geq \frac{1}{2}$, and so we need to restrict ourselves to the estimation of $\mathbb{E}_\mu[f]$ where f is $\frac{1}{2}$-Hölder continuous in $\tilde{d}$. In this case, we immediately have

$$\|\Delta_i\|_2^2 \leq \|f\|_{\frac{1}{2}}^2 \mathbb{E}\tilde{d}\big(\mathcal{T}_0^i, \mathcal{B}_0^i\big), \tag{5.10}$$

and as before, we can balance the two terms on the right-hand side of (5.9) to get sufficiently sharp bounds on $\|\Delta_i\|_2$, see Lemma 8.3. Note that the square root on $C_{j_{i-1}}^{j_i}$ is the source of the stronger assumption on the regularity of the reference measure μ_0. We get the following result.

Theorem 5.4. *Assume that the target measure μ is given as in (5.1), where g satisfies Assumption 5.1. Suppose that Assumption 5.2 is satisfied for $\theta \geq 1$ and let $f : \mathcal{X} \to \mathbb{R}$ be $\frac{1}{2}$-Hölder* *continuous with respect to $\tilde{d}$. Assume that $a > 2\theta + \frac{1}{2}$, where a represents the regularity of the reference measure, see* (5.1). *Then there are choices of a_i, j_i and $\mathbb{P}(N \geq i)$, such that*

$$Z = \sum_{i=0}^{N} \frac{\Delta_i}{\mathbb{P}(N \geq i)}$$

is an unbiased estimate of $\mathbb{E}_\mu[f]$ with finite variance and finite expected computing time. For example, for any $m \in \mathbb{N}$, this works if we choose $a_i = mi$, $j_i \sim r^{\frac{4mi}{1-2a}}$ and $\mathbb{P}(N \geq i) \propto r^{(m-\epsilon)i}$, where $\epsilon \in (0, \frac{4\theta m}{1-2a} + m)$.

Remark 5.5. Let $(H, \langle\cdot,\cdot\rangle_H, \|\cdot\|_H)$ be another Hilbert space. Using Proposition C.1 in the Supplementary Material [3] which generalises Proposition 1.1, it is straightforward to check that Theorems 5.3 and 5.4 can be extended to the estimation of expectations of functions $f : \mathcal{X} \to H$ which are Hölder continuous. In particular, using Theorem 5.4, we can perform unbiased estimation of all moments of μ.

Indeed, observe that all functions $f : \mathcal{X} \to H$ satisfying $\|f(x) - f(y)\|_H \leq C\|x - y\|^{\frac{1}{4}} \exp(\frac{1}{8}(\|x\| \vee \|y\|))$ are $\frac{1}{2}$-Hölder continuous with respect to $\tilde{d}$; this follows by separate inspection of the cases $\frac{\|x-y\|}{\tau} \leq 1$ and $\frac{\|x-y\|}{\tau} > 1$. In the former

$$\|f(x) - f(y)\|_H \leq \frac{C\tau^{\frac{1}{4}}\|x-y\|^{\frac{1}{4}}}{\tau^{\frac{1}{4}}}\left(\exp\left(\frac{1}{2}\left(\|x\| \vee \|y\|\right)\right)\right)^{\frac{1}{4}} \leq C\tau^{\frac{1}{4}}\tilde{d}(x,y)^{\frac{1}{2}},$$

while in the latter

$$\begin{aligned}\|f(x) - f(y)\|_H &\leq C\left(\|x\|^{\frac{1}{4}} \vee \|y\|^{\frac{1}{4}}\right)\exp\left(\frac{1}{8}\left(\|x\| \vee \|y\|\right)\right) \\ &\leq \tilde{C}\exp\left(\frac{1}{4}\left(\|x\| \vee \|y\|\right)\right) \leq \left(\exp\left(\|x\| \vee \|y\|\right)\right)^{\frac{1}{4}} \\ &\leq \tilde{C}\tilde{d}(x,y)^{\frac{1}{2}}.\end{aligned}$$

Using this observation, it is straightforward to check that we can apply the unbiasing procedure to $f(x) = x$ and $f(x) = x \otimes x$ (or to the finite dimensional approximations $f(x) = \Pi_j x$ and $f(x) = \Pi_j x(\Pi_j x)^t$) to obtain unbiased estimates of the mean and the second moment, respectively.

Remark 5.6. In this section, we focused on the discretisation of the input of g, x. However, in most practical scenarios like those arising in Bayesian inverse problems, g is based on a solution operator to a Partial Differential Equation and hence g itself needs to be discretised, say by g^l. We provide an example of how it is possible to do this in the setting for uniformly ergodic Markov chains in Section D of the Supplementary Material [3]. In order to make possible the unbiased estimation using the pCN algorithm in practical problems, the analysis in this section needs to be adapted accordingly. This is beyond the scope of the present paper, but it will be the topic of follow-up work.

6. Comparison of the unbiasing procedure and the ergodic average

In Section 2, we have shown how the unbiasing procedure can be applied to the estimation of expectations with respect to the invariant distribution π of a Markov chain that exhibits a simulatable contracting coupling. The existence of such a coupling implies that the Markov chain is ergodic, thus, the ergodic average constitutes a consistent estimator of $\mathbb{E}_\pi[f]$, for sufficiently nice functions f. In this section, we investigate how estimators constructed by averaging over independent runs of the unbiasing procedure perform compared to the ergodic average.

We compare the two methods using the Mean Square Error – work product (MSE-work product)

$$\text{MSE} \times \mathbb{E}(\text{computing time}), \tag{6.1}$$

which has also been used as a performance measure in [27], in the setting of unbiased estimation of expectations with respect to diffusions. For estimators constructed by averaging over unbiased estimators, the MSE-work product has the attractive property that it does not depend on the number of instances L that are averaged over. The reason for this is that the variance is scaled by $\frac{1}{L}$ whereas the expected computing time is multiplied by L. Using Proposition 1.1 and the expression (1.2), we see that the MSE-work product for the unbiasing procedure studied in the present paper is

$$\left(\sum_i \frac{\nu_i}{\bar{F}_i} - \left(\mathbb{E}_\pi[f]\right)^2\right)\left(\sum_i \bar{F}_i t_i\right). \tag{6.2}$$

Here t_i denotes the expected computing time to generate Δ_i, $\bar{F}_i = \mathbb{P}(N \geq i)$ and

$$\nu_i = \|\Delta_i\|_2^2 + 2\mathbb{E}\Delta_i(\mathbb{E}Y - \mathbb{E}Y_i) = \text{Var}(\Delta_i) + (\mathbb{E}Y - \mathbb{E}Y_{i-1})^2 - (\mathbb{E}Y - \mathbb{E}Y_i)^2, \tag{6.3}$$

where $Y \sim f_\star\pi$ and $Y_i = \sum_{k=0}^{i} \Delta_k$.

There are (uncountably) many choices of the number of time steps a_i used to construct Δ_i in Algorithm 1, and the probabilities $\bar{F}_i$, that yield unbiased estimators with finite variance and finite expected computing time. For a fair comparison with the ergodic average, we need to optimise the MSE-work product with respect to a_i and $\bar{F}_i$. Since this is difficult in general, we consider the example of 1-dimensional contracting normals in Section 6.1. We note that this example is also covered by the theory in [26], however we use it to

- compare the performance of the ergodic average of the Markov chain with the average of unbiased estimators of the type presented in Section 2;
- show that the added flexibility of choosing a_i, is crucial for optimizing the performance of the unbiased estimator (note that in [26] a_i is restricted to be equal to i);
- illustrate that we do not need sharp bounds on the properties of the coupling in order to tune the unbiased estimator;
- show numerical results suggesting that in a parallel setting the unbiasing procedure can be superior.

In Section 6.2, we consider posterior inference for a Bayesian logistic regression model and get the same findings as for contracting normals. Even though we cannot verify the contracting assumption of Section 2, we demonstrate that even a naive implementation of the unbiasing procedure leads to a competitive algorithm.

6.1. Contracting normals

We consider the example of 1-dimensional contracting normals, that is, the Markov chain defined by

$$X_{n+1} = \rho X_n + \sqrt{1-\rho^2}\xi_{n+1}, \tag{6.4}$$

for $\rho \in (0, 1)$ and $\xi_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. This Markov chain is ergodic with the standard normal distribution as invariant distribution, that is $\pi = \mathcal{N}(0, 1)$. The construction of the unbiased estimator Z follows from Section 2, by considering the coupling

$$K\bigl((x, y), (dx', dy')\bigr) = \mathcal{L}\bigl(\rho x + \sqrt{1-\rho^2}\xi, \rho y + \sqrt{1-\rho^2}\xi\bigr),$$

where $\xi \sim \mathcal{N}(0, 1)$. It is straightforward to check that this coupling satisfies Assumption 2.1.i. with geometric rate of contraction $r = \rho$, for the distance $d(x, y) = |x - y|$. The corresponding "top" and "bottom" chains have the form

$$\mathcal{T}^i_{k+1} = \rho \mathcal{T}^i_k + \sqrt{1-\rho^2}\xi^i_{k+1},$$
$$\mathcal{B}^i_{k+1} = \rho \mathcal{B}^i_k + \sqrt{1-\rho^2}\xi^i_{k+1},$$

where $\xi^i_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. The expected computing time is $t_i = T_{\text{step}} \times a_i$, where T_{step} is the expected computing time to simulate one step of the chain, while the ν_i can be bounded using the bounds on $\|\Delta_i\|_2$.

For this chain, there are analytic expressions for ν_i if we consider the estimation of $\mathbb{E}_\pi[f]$ for f being a polynomial. In the following we consider the simple function $f(x) = x$, which is trivially Lipschitz in d so that Theorem 2.4 applies. In this case, we simply have that $\Delta_i = \mathcal{T}^i_0 - \mathcal{B}^i_0$.

In Section 6.1.1, we find an explicit asymptotic expression for the MSE-work product for the ergodic average. We discuss the problem of finding good choices of a_i and $\bar{F}_i$ for the unbiasing procedure in Section 6.1.2. Even though we are not able to give a satisfying answer to the optimisation problem, we show in Section 6.1.3 that informed choices of a_i and $\bar{F}_i$ lead to a competitive performance of the unbiased estimator compared to the ergodic average, as measured by the MSE-work product. Such informed choices require precise knowledge of ν_i, which in practice is not available. In Section 6.1.4, we investigate the effect on the optimisation over $\bar{F}_i$ for fixed a_i, of using the exact values ν_i for $i \le i_0$ and only upper bounds for $i > i_0$. We demonstrate that this already leads to a considerable improvement over using upper bounds for all i. Finally, in Section 6.1.5, we present a comparison of the unbiasing procedure and the ergodic average in terms of computing time in the parallel computing setting. This comparison is not exhaustive but suggests future investigation.

6.1.1. *The MSE-work product for the ergodic average*

The MSE-work product of the ergodic average for $f(x) = x$ for contracting normals can be calculated explicitly. Indeed, we first iterate (6.4) to obtain

$$\sum_{i=0}^{n} X_i = \frac{1-\rho^{n+1}}{1-\rho} X_0 + \sum_{i=1}^{n} \xi_i \sum_{j=0}^{n-i} \rho^j \sqrt{1-\rho^2}.$$

Using this formula, we obtain an expression for the MSE as follows

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}\sum_{i=0}^{n} X_i - 0\right)^2 &= \frac{\mathbb{E}X_0^2}{n^2}\left(\frac{1-\rho^{n+1}}{1-\rho}\right)^2 + \frac{(1-\rho^2)}{n^2}\sum_{i=1}^{n}\left(\frac{1-\rho^{n-i+1}}{1-\rho}\right)^2 \\ &= \frac{1}{n^2}\left(\frac{1-\rho^{n+1}}{1-\rho}\right)^2 \mathbb{E}X_0^2 + \frac{1}{n}\frac{1+\rho}{1-\rho}\frac{1}{n}\sum_{i=1}^{n}\left(1-\rho^{n-i+1}\right)^2.\end{aligned}$$

This allows us to calculate the asymptotic performance as $n \to \infty$

$$\lim_{n\to\infty} \mathrm{MSE} \times \mathbb{E}\,(\text{computing time}) = \frac{1+\rho}{1-\rho} T_{\text{step}}. \tag{6.5}$$

It is important to note that non-asymptotic effects such as burn-in lead to a worse MSE-work product for finite n.

For general Markov chains the expression in (6.5) generalises to

$$\mathrm{Var}_\pi(f)\frac{1+\rho}{1-\rho} T_{\text{step}}$$

which is an asymptotic upper bound on the MSE-work product if ρ denotes the L^2-spectral gap. This result can be found in [29].

6.1.2. *The MSE-work product for estimators based on the unbiasing procedure*

For contracting normals the expressions for ν_i can be derived analytically using (6.3). For simplicity we consider $X_0 = 0$ (so that in Algorithm 1, we set $x_0 = 0$) for which we obtain

$$\begin{aligned}\nu_0 &= \left(1-\rho^{2a_0}\right), \\ \nu_i &= \rho^{2a_{i-1}}\left(1-\rho^{2(a_i-a_{i-1})}\right).\end{aligned} \tag{6.6}$$

Thus, the optimisation of the MSE-work product is similar to the one encountered in [27] for unbiased estimation of expectations based on diffusions. More precisely, the authors of [27] consider the optimisation problem

$$\begin{aligned} &\min_{\bar{F}} \left(\sum_n \frac{\nu_n}{\bar{F}_n}\right)\left(\sum_n \bar{F}_n t_n\right) \\ &\text{subject to} \qquad \bar{F}_i \geq \bar{F}_{i+1} \\ &\qquad\qquad\qquad\ \bar{F}_i > 0 \\ &\qquad\qquad\qquad\ \bar{F}_0 = 1. \end{aligned} \tag{6.7}$$

They show using the Cauchy–Schwarz inequality, that the choice

$$\mathbb{P}(N \geq i) = \bar{F}_i = \frac{\sqrt{\frac{\nu_i}{t_i}}}{\sqrt{\frac{\nu_0}{t_0}}}, \tag{6.8}$$

gives rise to the lower bound

$$\left(\sum_n \sqrt{\nu_n t_n}\right)^2. \tag{6.9}$$

Therefore the minimum is attained by this choice of $\bar{F}_i$ provided that it is feasible, that is, provided ν_i/t_i is decreasing.

In the setting of (6.6), we have the following explicit optimisation problem

$$\begin{aligned} &\min\left(\sum_{i=1}^{\infty} \frac{\rho^{2a_{i-1}}(1-\rho^{2(a_i-a_{i-1})})}{\bar{F}_i} + 1 - \rho^{2a_0}\right)\sum_{i=0}^{\infty} \bar{F}_i a_i \\ &\text{subject to} \qquad \bar{F}_0 = 1 \geq \bar{F}_1 \geq \bar{F}_2 \geq \cdots \\ &\qquad\qquad\qquad\ \bar{F}_i > 0 \\ &\qquad\qquad\qquad\ 0 < a_0 < a_1 < \cdots \\ &\qquad\qquad\qquad\ a_i \in \mathbb{N}. \end{aligned} \tag{6.10}$$

In contrast to [27], we want to optimise the MSE-work product with respect to both $\bar{F}_i$ and a_i. However, even in this simple case we do not know the solution, but instead present a comparison based on informed choices of a_i and $\bar{F}_i$ in the next subsection.

6.1.3. *Initial results based on informed parameter choices*

The minimisation over both a_i and $\bar{F}_i$ could be achieved by first minimising over $\bar{F}_i$ for fixed a_i and then minimising the resulting expression over a_i. If for a_i the choice of $\bar{F}_i$ given in (6.8) is feasible, then the minimum is given by (6.9). If it is not feasible the minimisation over $\bar{F}_i$ is not clear.

Even though we cannot optimise explicitly over all choices of a_i, we do so over the sub class $a_i = mi + m$ for $m \in \mathbb{N}$. The expected computing time of Δ_i, $t_i = T_{\text{step}} a_i$, is monotonically increasing. Moreover, it is straight forward to check for this choice of a_i, that v_i is decreasing such that the choice of $\bar{F}_i$ in (6.8) is feasible. As a result, this choice of $\bar{F}_i$ gives rise to the optimal MSE-work product of the unbiased estimator for any fixed m, and the corresponding (optimal) MSE-work product can be obtained using (6.9) as follows

$$\begin{aligned}
&\left(\sum_{i=1}^{\infty} \sqrt{\rho^{2mi}\left(1-\rho^{2m}\right) T_{\text{step}}\, m(i+1)} + \sqrt{T_{\text{step}} m\left(1-\rho^{2m}\right)}\right)^2 \\
&\quad = T_{\text{step}} \left(\sqrt{m\left(1-\rho^{2m}\right)} \sum_{j=1}^{\infty} \rho^{m(j-1)} \sqrt{j}\right)^2 \\
&\quad = T_{\text{step}} \left(\rho^{-m} \sqrt{m\left(1-\rho^{2m}\right)} \mathrm{Li}_{-\frac{1}{2}} \rho^m\right)^2,
\end{aligned}$$

where Li denotes the polylogarithm function. Subsequently, we assume that $T_{\text{step}} = 1$ since it is only a multiplicative constant of the minimum and it does not change the optimal choice of $\bar{F}_i$ in (6.8).

We compare the MSE-work product of the ergodic average, given in (6.5), to the optimal MSE-work product of the unbiased estimator for a fixed m. Again, we would like to stress that this comparison is advantageous for the ergodic average because we disregard non-asymptotic effects such as burn-in. In Figure 2, we plot the MSE-work product of the ergodic average, the

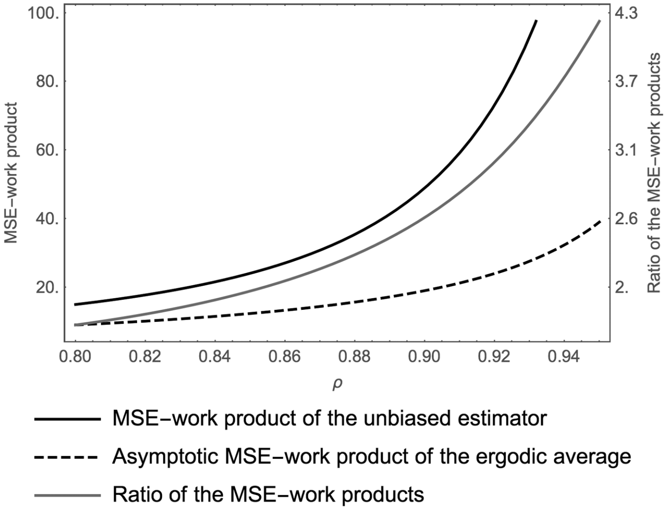

Figure 2. MSE-work products for the ergodic average and the unbiasing procedure for $a_i = 4i + 4$ and $\bar{F}_i$ chosen optimally, plotted against ρ, in the contracting normals example. In different scale we plot the ratio of the MSE-work product of the unbiased estimator over the MSE-work product of the ergodic average.

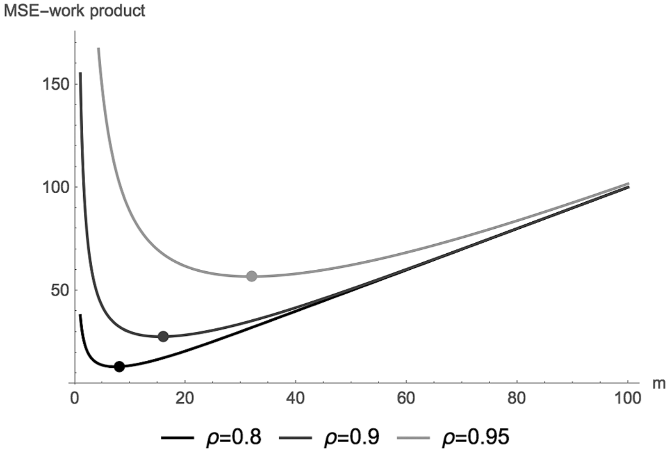

Figure 3. MSE-work products for the unbiasing procedure with $a_i = mi + m$ and $\bar{F}_i$ chosen optimally, for $\rho = 0.8$, 0.9 and 0.95, plotted against m, in the contracting normals example. The markers correspond to the choice $m = \lceil \frac{w}{\log \rho} \rceil$ with $w = -1.632$.

optimal MSE-work product of the unbiased estimator for $m = 4$, and their ratio, as functions of ρ. We observe that as ρ increases towards 1 the ratio of the MSE-work product of the unbiasing procedure over the one of the ergodic average explodes.

In an effort to improve the performance of the unbiased estimator, we allow m to depend on ρ. In order to illustrate the impact of m, we plot the MSE-work product as a function of m for different values of ρ in Figure 3. We observe that for small values of m the MSE-work product is very large, however for the optimal choice of m the value of MSE-work product is relatively small. As $\rho \to 1$, the optimal value of m increases.

We next try to roughly find the optimal value of m for a given ρ, and to do this we make the ansatz that m should be of the form $\lceil \frac{w}{\log \rho} \rceil$ for $w < 0$. The reason for this choice is that it at least keeps the values of ν_i roughly at the same magnitude as $\rho \to 1$, even though the value of t_i increases. This choice will be justified further subsequently. Let's suppose for the moment that m is a continuous variable and we set it to $\frac{w}{\log \rho}$. In this case, we consider the ratio of the MSE-work product of the unbiasing procedure over the one of the ergodic average, given by

$$\begin{aligned}\text{rMSE-work} &= \frac{(\rho^{-m}\sqrt{m(1-\rho^{2m})}\mathrm{Li}_{-\frac{1}{2}}\rho^m)^2}{\frac{1+\rho}{1-\rho}}\\ &= \left(e^{-w}\sqrt{1-e^{2w}}\mathrm{Li}_{-\frac{1}{2}}\left(e^w\right)\sqrt{w}\right)^2\left(\frac{1-\rho}{(1+\rho)\log(\rho)}\right).\end{aligned}$$

It is clear, that minimisation of this ratio over w does not depend on ρ. Optimisation of the first parenthesis gives that it attains its minimum at $w = -1.632$. This choice of w gives rise to the circular markers in Figure 3 which are clearly close to the optimal values of m for all the plotted values of ρ.

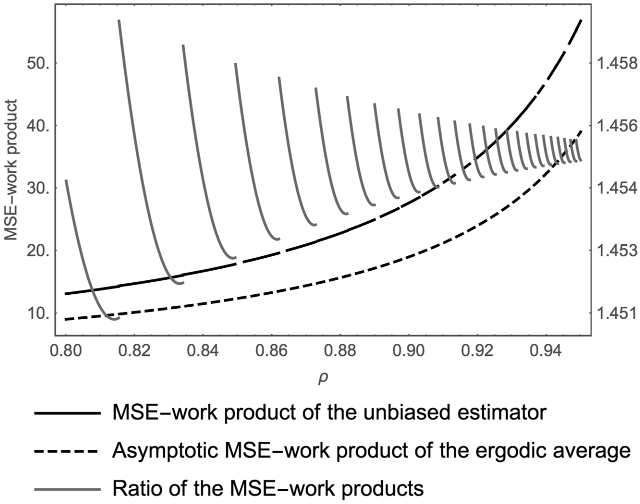

Figure 4. MSE-work products for the ergodic average and the unbiasing procedure for $a_i = \lceil \frac{w}{\log \rho} \rceil i + \lceil \frac{w}{\log \rho} \rceil$ with $w = -1.632$ and $\bar{F}_i$ chosen optimally, plotted against ρ, in the contracting normals example. In different scale we plot the ratio of the MSE-work product of the unbiased estimator over the MSE-work product of the ergodic average.

In Figure 4, we plot again the MSE-work product of the ergodic average, the (optimised over $\bar{F}_i$) MSE-work product of the unbiased estimator for $m = \lceil \frac{w}{\log \rho} \rceil$ and $w = -1.632$, and their ratio. In this case, we observe that the ratio stays bounded above by 1.5 as $\rho \to 1$, that is even as the convergence of the underlying chain deteriorates. Notice that the oscillation of the ratio comes from the use of the ceiling function.

6.1.4. *Tuning*

At first sight, it seems necessary to have a very precise knowledge of the coupling, in terms of for example tight bounds on ν_i, in order to tune the unbiased estimator. In this subsection, we show that if we only have good estimates $\nu_i \approx \hat{\nu}_i$ for $i \leq i_0$ and use a crude bound on ν_i for $i > i_0$, then the performance of the unbiased estimator remains close to the optimal behaviour. More precisely, instead of the optimisation problem (6.7), we consider

$$\begin{aligned} &\min \left(\sum_{i=0}^{i_0} \frac{\hat{\nu}_i}{\bar{F}_i} + \sum_{i=i_0+1}^{\infty} \frac{\nu_i^\star}{\bar{F}_i} \right) \left(\sum_{i=0}^{\infty} a_i \bar{F}_i \right) \\ &\text{subject to} \qquad \bar{F}_0 = 1 \geq \bar{F}_1 \geq \cdots \\ &\qquad\qquad\qquad \bar{F}_i > 0. \end{aligned} \tag{6.11}$$

In order to illustrate this, we again consider the behaviour of the unbiased estimator with the fixed choice $a_i = 4i + 4$. We fix $\rho = 0.5$ and suppose that $\nu_i^\star = \nu_i(\tilde{\rho})$ for $i > i_0 = 3$ are our upper bounds on $\nu_i(\rho)$ for some $\tilde{\rho} \geq 0.5$. Moreover, we use the exact value of ν_i for $i \leq 3$. We then optimise $\bar{F}_i^\star$ in the parametric family $C\tilde{\rho}^{a_{i-1}}$ for $i > i_0$ leading to the following optimisation problem:

$$\begin{aligned}
&\min\left(\sum_{i=0}^{i_0}\frac{\hat{\nu}_i}{\bar{F}_i} + \sum_{i=i_0+1}^{\hat{i}}\frac{\nu_i^\star}{\bar{F}_i^\star(C)}\right)\left(\sum_{i=0}^{i_0}a_i\bar{F}_i + \sum_{i=0}^{i_0}a_i\bar{F}_i^\star(C)\right)\\
&\text{with respect to} \qquad \bar{F}_1,\dots,\bar{F}_{i_0},C\\
&\text{subject to} \qquad \bar{F}_0 = 1 \geq \bar{F}_1 \geq \cdots \geq \bar{F}_{i_0} \geq \bar{F}_i^\star(C)\\
&\qquad\qquad\qquad\quad \bar{F}_i > 0.
\end{aligned} \tag{6.12}$$

A numerical solution to this optimization problem using Mathematica results in a significant improvement in the performance of the unbiased estimator as shown in Figure 5. We see that having good estimates of ν_i even for just the first three levels and using crude bounds for the higher levels, greatly improves the performance of the unbiasing procedure. Naturally, as the bounds for the higher levels get worse (that is, as $\tilde{\rho}$ increases), the performance deteriorates.

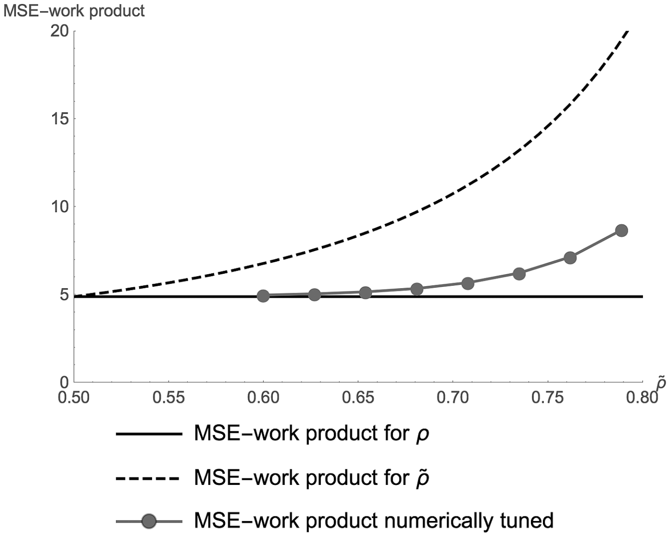

Figure 5. MSE-work products in the contracting normals example, plotted against ρ, for unbiased procedures based on knowledge of true ν_i for $\rho = 0.5$ (black), only an upper bound on ν_i using $\nu_i(\tilde{\rho})$ for $\tilde{\rho} \geq 0.5$ (dashed) and numerical optimisation of $\bar{F}_i$ using the exact values of ν_i for $i = 1, 2, 3$ and upper bounds $\nu_i(\tilde{\rho})$ for $i > 3$ (grey). More precisely, we optimise $\bar{F}_i$ for $i = 1, \dots, 3$ and C in $\bar{F}_i^\star(C) = C\tilde{\rho}^{a_{i-1}}$ subject to the constraint that $\bar{F}_3 > \bar{F}_4$.

6.1.5. *Comparison in the parallel setting*

We compare the ergodic average to the unbiasing procedure by measuring CPU time. We consider $\rho = 0.8$. To make the comparison fair after each step of the Markov chain the algorithm sleeps for 1 millisecond. In this way, the generation of N has negligible effect on the comparison as it should do for most large scale inference procedures and the computing time is determined by the distribution of N and the number of steps performed. Subsequently, we describe the procedure both for the ergodic average and the unbiasing procedure in a 10 core parallel setting.

1. For the ergodic average, we draw a random number M between 10 and 10 000. Each core performs M steps and we measure the time it takes to do these steps. We average over the chains and the steps of each chain. We plot the squared error versus the time, which gives rise to one black dot in the left panel of Figure 6.
2. We draw a random time uniformly distributed on a log-scale between 0.1 and 10 seconds and let each core produce unbiased estimates. When the time is up we average over the obtained unbiased estimates and plot the squared error of the resulting unbiased estimator against time, giving rise to the grey dots in the left panel Figure 6.

In the right panel of Figure 6, we smooth the results of the above simulation procedure and produce 95% confidence tubes for the MSE for the ergodic average (black) and the unbiased estimator (white). In this particular setting, it seems that the unbiasing method is competitive. Whereas this result is in no way conclusive, it suggests further investigation.

6.2. Logistic regression

We apply the findings of Section 2 on unbiased estimators based on contracting couplings to posterior inference for a Bayesian logistic regression model. Even though we cannot verify the assumption of Section 2 and we cannot tune the unbiasing procedure, we demonstrate in this section that a hands-on application of the unbiasing procedure leads to a competitive algorithm.

We assume the data $y_i \in \{-1, 1\}$ for $i = 1, \dots, M$ is modelled by

$$p(y_i|T_i, \beta) = h\big(y_i \beta^t T_i\big), \tag{6.13}$$

where $h(z) = \frac{1}{1+\exp(-z)} \in [0, 1]$. We put a Gaussian prior $\mathcal{N}(0, I)$ on the regression coefficient $\beta \in \mathbb{R}^d$ and consider a fixed design matrix $T \in \mathbb{R}^{M\times d}$ which we specify later on. By Bayes' rule, the posterior π satisfies

$$\pi(\beta) \propto \exp\left(-\frac{1}{2}\|\beta\|^2\right) \prod_{i=1}^{N} h\big(y_i \beta^t T_i\big).$$

Thus, the target measure has a density with respect to a centred Gaussian distribution, which is such that the pCN algorithm satisfies the Assumption 2.1 of Section 2 as shown in [18] and [10]. We provide a brief summary of the relevant results to the contraction of the pCN algorithm in Section 8.3.2.

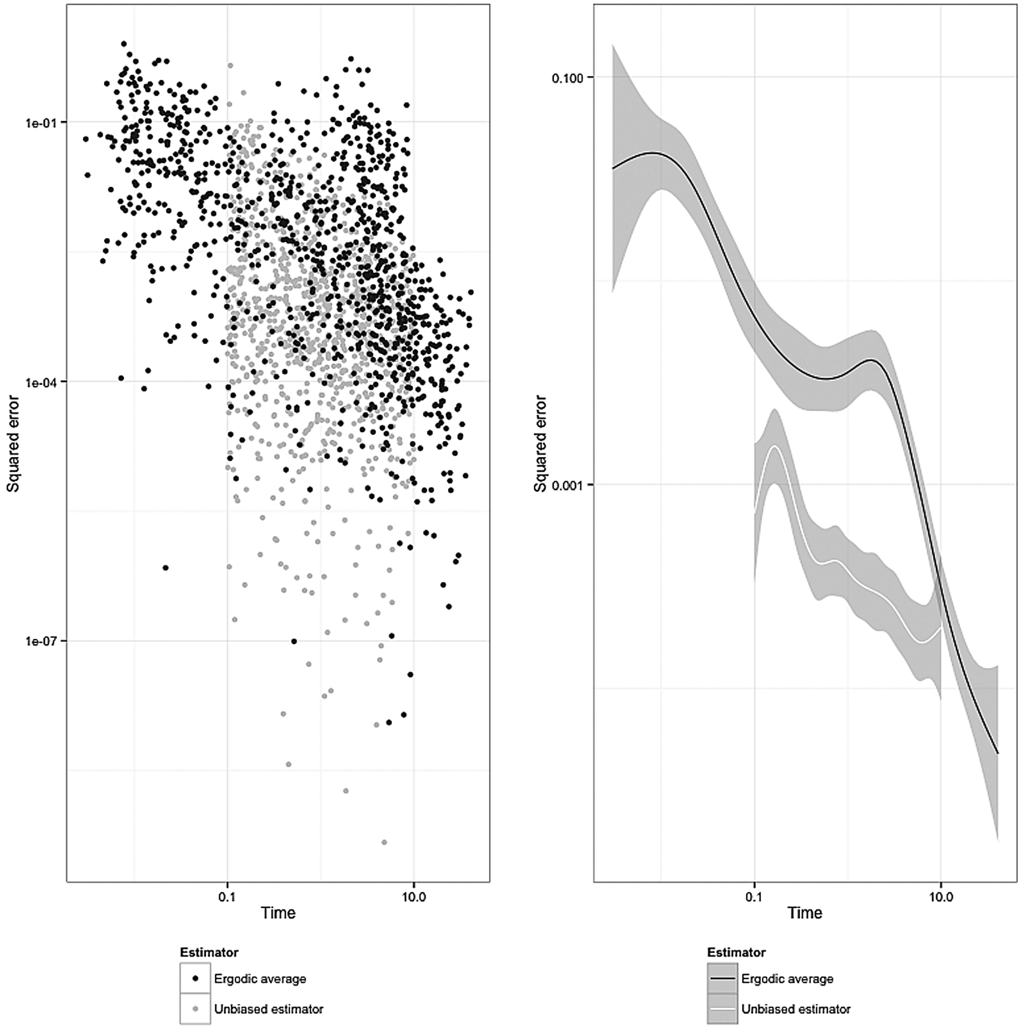

Figure 6. (a) Measurements of squared errors and running times for the ergodic average and unbiasing procedure for independent runs, in a 10-core parallel setting (left panel) and (b) 95% confidence tube for the MSE based on the data on the left panel using a generalised additive model (right panel), in the contracting normals example with $\rho = 0.8$.

For the problem at hand, the prior mean is 0 and the posterior mean is typically far from 0. The proposal of the pCN algorithm only takes into account the prior and pushes towards 0. Furthermore, the covariance matrix changes from prior to posterior as well. This has to be corrected by the rejection step of the Metropolis–Hastings algorithm. The result is that the coupling of the corresponding pCN algorithm with the same random input and different initial states has a contraction rate close to 1. For this reason, the unbiasing procedure is difficult to apply for this coupling.

A solution to this difficulty is to modify the pCN algorithm as in [24], and in particular to consider the Metropolis–Hastings algorithm with proposal given by

$$\beta' = c + \rho(\beta - c) + \sqrt{1-\rho^2}\xi, \qquad \text{with } \xi \sim \mathcal{N}(0, C). \tag{6.14}$$

The resulting Markov chain preserves the non-centered Gaussian distribution $\mathcal{N}(c, C)$. Reasonable choices for c and C are:

1. posterior mean and posterior covariance as estimated by a MCMC run;
2. Laplace approximation based on a maximum a posteriori estimator;
3. the minimiser of the Kullback–Leibler divergence of π from $\nu = \mathcal{N}(c, C)$

$$D_{KL}(\nu\|\pi) = \int \log\left(\frac{d\nu}{d\pi}\right) d\nu,$$

as suggested in [24].

For simplicity, we take the first approach using 10^6 steps of the random walk Metropolis (RWM) algorithm to estimate the values of the posterior mean and covariance. We consider $d = 3$ and $N = 100$ data points and choose the design matrix to be

$$T = \begin{pmatrix} T_{1,1} & T_{1,2} & 1 \\ T_{2,1} & T_{2,2} & 1 \\ \vdots & \vdots & \vdots \\ T_{100,1} & T_{100,2} & 1 \end{pmatrix},$$

for a fixed sample of $T_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ for $i = 1, \dots, 100$ and $j = 1, 2$.

We now consider the estimation of the mean of the third component under the posterior. We apply the unbiasing procedure to the coupling arising from using the same $\xi \sim \mathcal{N}(0, C)$ in the proposal (6.14), and the same uniform random variable for the accept and reject step of the corresponding Metropolis–Hastings algorithm. The contraction property of this coupling has not been established, however, we estimate the contraction factor by fitting a line with slope $s \approx 0.75$ to the log-plot of the averaged distance, see Figure 7. This suggests that Assumption 2.1 is satisfied with $r = \exp(s)$. We take a more conservative approach and set $r := \exp(\frac{1}{2}s)$. We choose $a_i = mi + m$ with $m = \lceil \frac{-1.632}{\log r} \rceil$ and $\bar{F}_i = r^{m \cdot i}$ which closely resembles our "optimised" choice for the contracting normals chain with $\rho = r$ in Section 6.1.

In the following, we compare the MSE-work product for

1. the ergodic average of the modified pCN algorithm over 10 000 steps started at c;
2. the average of 100 independent realisations of the unbiased estimator, as described in Section 2 and for $x_0 = c$.

For both algorithms, we record the squared error and the CPU time it took to generate the estimator. Because we are using CPU-time it actually matters how many unbiased estimators, we average over. This is in contrast to the idealised properties of the MSE-work error described at the beginning of Section 6. This is the reason for averaging over 100 independent realisations of the unbiased estimator, rather than just taking one sample as we did in Section 6.

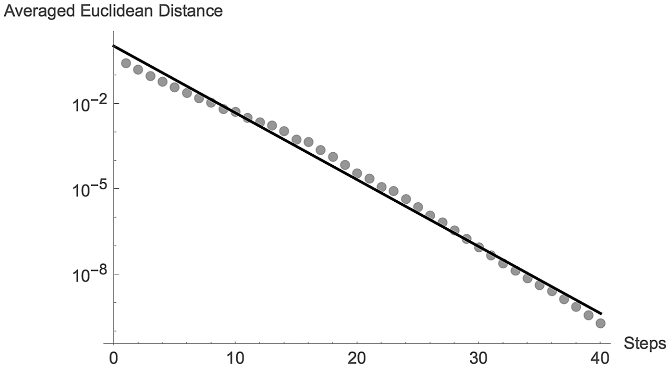

Figure 7. Empirical contraction property of the modified pCN algorithm based on simulations of the coupling and averaging over independent runs, in the logistic regression example. The contraction factor is estimated using a least squares fit.

We repeat this 10 000 times and visualise the results using box plots in Figure 8. Notice that the distribution of the squared error for the unbiased estimator is much more heavy-tailed compared to that of the modified pCN algorithm. This becomes more apparent in the histograms in Figure 9, where we can see that there exist outliers with large squared error for the unbiasing procedure. We use this data to estimate the ratio of MSE work products to be 4.15 and obtain a 95%-confidence interval $(3.86, 4.55)$ for the ratio using the pivotal bootstrap method.

In conclusion, we again see that the unbiasing procedure has competitive performance compared to the ergodic average, even with a crude choice of parameters and without using parallelisation.

7. Conclusion and future directions

We considered unbiased estimation in intractable and/or infinite dimensional settings. In particular, we showed how to unbiasedly estimate expectations with respect to the limiting distributions of Markov chains in possibly infinite dimensional state spaces. To do this, we generalised the methodology developed in [26] for removing the bias due to the burn-in time of the Markov chain, to cover the case that only a simulatable contracting coupling between runs of the chain started at different states is available (see Section 2). We then used a hierarchy of coupled Markov chains in state spaces of increasing dimension, to remove the bias due to the discretisation of the infinite-dimensional state space (see Sections 4 and 5).

Our focus has been on the methodological aspect, to show what it is possible to achieve, rather than to produce fully optimised results. It is crucial for the performance of the unbiasing procedure to have good couplings between runs of the chain started at different states. There is a great body of literature on couplings which can be potentially exploited in order to on the one hand improve the results presented in the present paper and on the other hand extend the application of the unbiasing procedure to other algorithms.

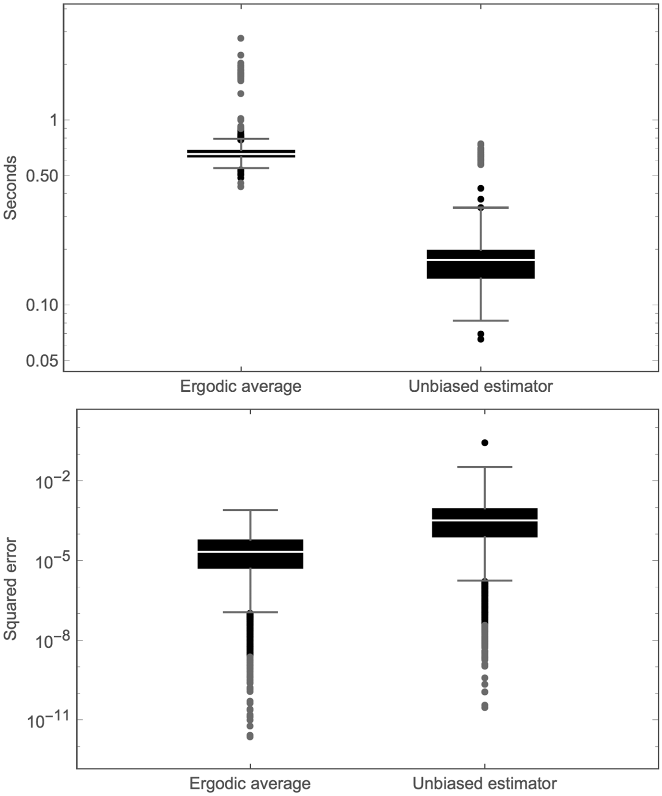

Figure 8. Estimation of the mean of the third component under the posterior in the logistic regression example. Box plots based on 10 000 independent simulations of the simulation time and the squared error for the ergodic average (10 000 steps) and the unbiasing procedure (average over 100 instances).

Furthermore, as we demonstrated in Section 6, the tuning of the parameters appearing in the unbiasing procedure, namely the distribution $\bar{F}_i$ of the random truncation point N, the number of steps performed at each approximation level a_i and the dimension of each approximation level j_i, has a huge impact on the performance. It is thus very important to develop an efficient algorithm that adapts the choice of these parameters and improves the simulation on the fly. This is particularly crucial for the transdimensional framework, since the cost of producing samples in high dimensions rapidly increases and hence the best possible management of the available resources is crucial.

One of the big advantages of the unbiasing methodology, is that it is very easily parallelisable. On the one hand, we can use multiple cores to produce multiple copies of the unbiased estimator Z, while on the other hand the generation of the differences Δ_i is also readily parallelisable since we assume that they are mutually independent across different levels. Moreover, it is straightforward to manage heterogeneous computer architectures, by generating Δ_i's at low levels using slower CPU's and GPU's and reserving the faster processors for higher levels.

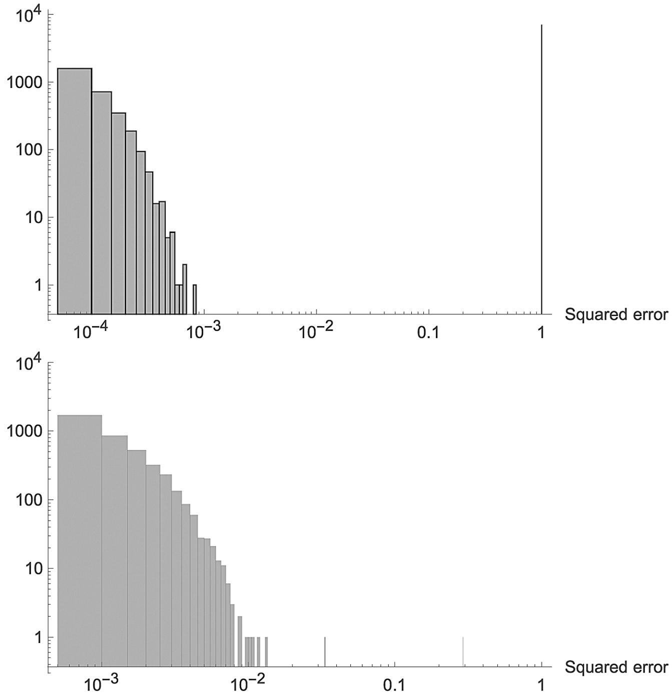

Figure 9. Histograms of the squared error for the ergodic average (top) and the unbiasing procedure (bottom), in the logistic regression example. The outliers for the unbiasing procedure illustrate the heavy tails of the distribution of the squared error.

Since the submission of this paper, [31] has appeared which on the one hand uses stratification to construct unbiased estimators with reduced variance, and on the other hand compares the performance of various unbiasing schemes to MLMC. Numerical simulations are provided in the context of estimating expectations with respect to solutions of SDE's, which suggest that the unbiasing schemes asymptotically match the performance of MLMC while avoiding the concern of bias.

In the context of estimation of expectations with respect to the limiting distribution of a Markov chain, we provided a range of initial results on the performance of the unbiasing procedure against the ergodic average (see Section 6). It is clear from these results, which are not optimally tuned and do not make full use of parallel computing, that the unbiasing method is competitive. We are hence very much looking forward to further simulations and comparisons in problems of higher computational complexity which are closer to real-life applications.

8. Proofs

We now present the proofs of the results contained in Sections 2, 4 and 5. The proofs of the results in Section 3 are provided in Section B of the Supplementary Material [3].

8.1. Proofs of the results in Section 2

Proof of Theorem 2.4. We first use Proposition 1.1 which gives conditions securing unbiasedness and finite variance of Z and then make sure that these conditions are compatible with a finite expected computing time. Let $\mathcal{F}_k = \sigma(\{\mathcal{T}_\ell^i, \mathcal{B}_\ell^i | \ell \le k\})$. We bound

$$\begin{aligned}
\|\Delta_i\|_2^2 &\le \|f\|_s^2 \mathbb{E} d^{2s}(\mathcal{T}_0^i, \mathcal{B}_0^i) \\
&\le \|f\|_s^2 \mathbb{E}\mathbb{E}\big(d^{2s}(\mathcal{T}_0^i, \mathcal{B}_0^i) | \mathcal{F}_{-a_{i-1}}\big) \\
&\le \|f\|_s^2 \mathbb{E}\big(K^{a_{i-1}} d^{2s}(\mathcal{T}_{-a_{i-1}}^i, x_0)\big) \\
&\le c\|f\|_s^2 r^{a_{i-1}} \mathbb{E} d^{2s}(\mathcal{T}_{-a_{i-1}}^i, x_0) \\
&\le c r^{a_{i-1}},
\end{aligned} \tag{8.1}$$

where the last step follows from (2.2) and where we use c as a positive constant which maybe different from occurrence to occurrence.

By the considerations at the end of Section 1.3, it suffices to verify (1.7) to get the unbiasedness and finite variance of Z. Using (8.1), we have

$$\begin{aligned}
\sum_{i \le l} \frac{\|\Delta_i\|_2 \|\Delta_l\|_2}{\mathbb{P}(N \ge i)} &= \sum_{i=0}^{\infty} \frac{\|\Delta_i\|_2}{\mathbb{P}(N \ge i)} \sum_{l=i}^{\infty} \|\Delta_l\|_2 \le c \sum_{i=0}^{\infty} \frac{\|\Delta_i\|_2}{\mathbb{P}(N \ge i)} \frac{r^{\frac{1}{2}i}}{1 - r^{\frac{1}{2}}} \\
&\le c \sum_{i=0}^{\infty} \frac{r^i}{\mathbb{P}(N \ge i)},
\end{aligned} \tag{8.2}$$

where we used that $r < 1$ and $a_i \ge i$. It is hence sufficient to choose the distribution of N such that $\sum_i \frac{r^i}{\mathbb{P}(N \ge i)} < \infty$ in order to have finite variance of the estimator Z; a valid choice is for example $\mathbb{P}(N \ge i) \propto r^{(1-\epsilon)i}$ for $\epsilon > 0$ which can be arbitrarily small.

Regarding the expected computing time of Z, we have that it is equal to $\sum_{i=0}^{\infty} t_i \mathbb{P}(N \ge i)$, where t_i is the expected time to generate Δ_i. By Assumption 2.3, we have a mild condition on the growth of a_i. For example, $a_i \lesssim r^{(2\epsilon - 1)i}$ works provided $\epsilon < \frac{1}{2}$ so that we have $a_i \ge i$ as required. □

Proof of Theorem 2.6. The proof is very similar to the proof of Theorem 2.4, where the estimate (8.1) is replaced by

$$\|\Delta_i\|_2^2 \le c a_{i-1}^{-2r}. \tag{8.3}$$

For example choose $a_i = i^k$. Then we have that

$$\sum_{i \le l} \frac{\|\Delta_i\|_2 \|\Delta_l\|_2}{\mathbb{P}(N \ge i)} \le c \sum_{i=0}^{\infty} \frac{i^{-2rk+1}}{\mathbb{P}(N \ge i)},$$

where c a positive constant which may change between different occurrences. The right-hand side is finite, if for example we choose $\mathbb{P}(N \geq i) \propto i^{-2rk+2+\epsilon}$ for $\epsilon > 0$ which can be arbitrarily small. Under this choice, the expected computing time

$$\mathbb{E}(\tau) = \sum_{i=0}^{\infty} i^k \mathbb{P}(N \geq i) = c \sum_{i=0}^{\infty} i^{k-2rk+2+\epsilon}.$$

For the last sum to be finite, we need to choose $0 < \epsilon < -3 - (1-2r)k$ and this choice is possible only if $k > \frac{3}{2s-1}$. □

8.2. Proofs of the results in Section 4

We first state and prove a crucial lemma which establishes that the Δ_i decay sufficiently quickly for the unbiasing programme to work. Then we provide the proof of Theorem 4.5.

Lemma 8.1. *Suppose Assumptions 4.1 and 4.2 are satisfied for some $\beta, \kappa > 1$, respectively. Then for $a_i \sim \frac{\beta}{c_\star} \log(j_i)$, where $c_\star = -\log(1-\alpha_\star)$, we have*

$$\|\Delta_i\|_2^2 \lesssim j_{i-1}^{-(\beta \wedge \kappa)}.$$

Proof. In order to bound E_1, we need to be able to control the different behaviour of the independence sampler in dimension j_{i-1} and j_i if driven by the same underlying randomness W. For this reason, we introduce the random functions $b^i_{\mathcal{T}}$, $b^i_{\mathcal{B}}$ below, taking values in $\{1, 2, 3\}$,

$$\begin{aligned}
b^i_{\mathcal{T}}(x, W) &= 1 \cdot \mathbb{1}_{[0,\alpha_\star]}\bigl(U^i_1\bigr) + 2 \cdot \mathbb{1}_{(\alpha_\star,1]}\bigl(U^i_1\bigr)\mathbb{1}_{[0,\frac{\alpha_{j_i}(x,\xi^i_2)-\alpha_\star}{1-\alpha_\star}]}\bigl(U^i_2\bigr) \\
&\quad + 3 \cdot \mathbb{1}_{(\alpha_\star,1]}\bigl(U^i_1\bigr)\mathbb{1}_{(\frac{\alpha_{j_i}(x,\xi^i_2)-\alpha_\star}{1-\alpha_\star},1]}\bigl(U^i_2\bigr), \\
b^i_{\mathcal{B}}(x, W) &= 1 \cdot \mathbb{1}_{[0,\alpha_\star]}\bigl(U^i_1\bigr) + 2 \cdot \mathbb{1}_{(\alpha_\star,1]}\bigl(U^i_1\bigr)\mathbb{1}_{[0,\frac{\alpha_{j_{i-1}}(x,\Pi_{j_{i-1}}\xi^i_2)-\alpha_\star}{1-\alpha_\star}]}\bigl(U^i_2\bigr) \\
&\quad + 3 \cdot \mathbb{1}_{(\alpha_\star,1]}\bigl(U^i_1\bigr)\mathbb{1}_{(\frac{\alpha_{j_{i-1}}(x,\Pi_{j_{i-1}}\xi^i_2)-\alpha_\star}{1-\alpha_\star},1]}\bigl(U^i_2\bigr),
\end{aligned}$$

such that $b^i_{\mathcal{T}}(\mathcal{T}^i_{k-1}, W^i_k)$ and $b^i_{\mathcal{B}}(\mathcal{B}^i_{k-1}, W^i_k)$ denotes the branch of the random functions $\varphi^i_{\mathcal{T}}$ and $\varphi^i_{\mathcal{B}}$ that was taken to go from $\mathcal{T}^i_{k-1}$ to $\mathcal{T}^i_k$ and from $\mathcal{B}^i_{k-1}$ to $\mathcal{B}^i_k$, respectively. More precisely,

1. if $b^i_{\mathcal{T}}(\mathcal{T}^i_{k-1}, W^i_k) = 1$ then $\mathcal{T}^i_k = \xi^i_{1,k}$;
2. if $b^i_{\mathcal{T}}(\mathcal{T}^i_{k-1}, W^i_k) = 2$ then $\mathcal{T}^i_k = \xi^i_{2,k}$;
3. if $b^i_{\mathcal{T}}(\mathcal{T}^i_{k-1}, W^i_k) = 3$ then $\mathcal{T}^i_k = \mathcal{T}^i_{k-1}$

and the analogous statement for $b^i_{\mathcal{B}}(\mathcal{B}^i_{k-1}, W^i_k)$ and $\mathcal{B}^i_k$. For economy of notation, we define $b^i_{\mathcal{T},k} := b^i_{\mathcal{T}}(\mathcal{T}^i_{k-1}, W^i_k)$ and similarly for $b^i_{\mathcal{B},k}$. Note that $b^i_{\mathcal{T},k} = 1$ if and only if $b^i_{\mathcal{B},k} = 1$ such that this leads to synchronisation because in this case $\mathcal{B}^i_k = \Pi_{j_{i-1}}\xi^i_{1,k} = \Pi_{j_{i-1}}\mathcal{T}^i_k$. Notice that this synchronisation property is preserved to $l > k$ as long as $b^i_{\mathcal{B},\tilde{k}} = b^i_{\mathcal{T},\tilde{k}}$ for $\tilde{k} = k+1,\dots,l$.

This notation allows us to bound E_1 as follows:

$$E_1 \leq \|f\|^2_\infty \mathbb{P}\bigl(\neg\bigl\{\exists k \leq 0 \text{ s.t. } b^i_{\mathcal{T},k} = 1 \text{ and } \forall l > k \; b^i_{\mathcal{T},l} = b^i_{\mathcal{B},l}\bigr\}\bigr),$$

because $f(\Pi_{j_{i-1}}\mathcal{T}^i_0) - f(\mathcal{B}^i_0) = 0$ on the event $\{\exists k \leq 0 \; b^i_{\mathcal{T},k} = 1 \text{ and } \forall l > k \; b^i_{\mathcal{T},l} = b^i_{\mathcal{B},l}\}$. In order to bound the above we introduce the filtration $\mathcal{F}_k = \sigma(\{W^i_m\}_{m \leq k})$ and notice that $\tau = \inf\{k : b^i_{\mathcal{T},k} = 1\}$ is a stopping time with respect to $\{\mathcal{F}_k\}$. We have

$$\begin{aligned}
E_1 &\leq \|f\|^2_\infty \mathbb{P}\bigl(\{\tau > 0\} \cup \bigl\{\tau \leq 0 \text{ and } \exists 0 \geq l > \tau : \; b^i_{\mathcal{T},l} \neq b^i_{\mathcal{B},l}\bigr\}\bigr)\\
&\leq \|f\|^2_\infty \bigl(\mathbb{P}\bigl(\{\tau > 0\}\bigr) + \mathbb{P}\bigl(\bigl\{\tau \leq 0 \text{ and } \exists 0 \geq l > \tau : \; b^i_{\mathcal{T},l} \neq b^i_{\mathcal{B},l}\bigr\}\bigr)\bigr)\\
&\leq \|f\|^2_\infty \bigl((1-\alpha_\star)^{a_{i-1}} + \mathbb{P}\bigl(\bigl\{\tau \leq 0 \text{ and } \exists 0 \geq l > \tau : \; b^i_{\mathcal{T},l} \neq b^i_{\mathcal{B},l}\bigr\}\bigr)\bigr).
\end{aligned}$$

Notice that for $A = \{\tau \leq 0 \text{ and } \exists 0 \geq l > \tau : b^i_{\mathcal{T},l} \neq b^i_{\mathcal{B},l}\}$, we have

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{E}(\mathbb{1}_{\tau \leq 0}\mathbb{1}_{\{\exists 0 \geq l > \tau \; b^i_{\mathcal{T},l} \neq b^i_{\mathcal{B},l}\}})\\
&= \mathbb{E}\bigl(\mathbb{1}_{\tau \leq 0}\mathbb{E}(\mathbb{1}_{\{\exists 0 \geq l > \tau \; b^i_{\mathcal{T},l} \neq b^i_{\mathcal{B},l}\}} \mid \tau)\bigr)\\
&= \mathbb{E}\bigl(\mathbb{1}_{\tau \leq 0}\mathbb{P}_{(\mathcal{T}^i_\tau,\mathcal{B}^i_\tau)}\bigl(\exists -\tau \geq l > 0 : b^i_{\mathcal{T},l} \neq b^i_{\mathcal{B},l}\bigr)\bigr)\\
&\leq \mathbb{P}(\tau \leq 0) \sup_{\Pi_j(\mathcal{T}^i_0) = \mathcal{B}^i_0} \mathbb{P}_{(\mathcal{T}^i_0,\mathcal{B}^i_0)}\bigl(\exists -\tau \geq l > 0 : b^i_{\mathcal{T},l} \neq b^i_{\mathcal{B},l}\bigr),
\end{aligned}$$

where in the third identity we made use of the strong Markov property. In order to bound the supremum on the right-hand side above, we introduce

$$\delta_i := \sup_{\Pi_j(x_\mathcal{T}) = x_\mathcal{B}} \mathbb{P}\bigl(b_{\mathcal{T}^i}(x_\mathcal{T}, W) \neq b^i_\mathcal{B}(x_\mathcal{B}, W)\bigr),$$

where $x_\mathcal{T}$ and $x_\mathcal{B}$ live in $\mathcal{X}_{j_i}$ and $\mathcal{X}_{j_{i-1}}$, respectively. We have

$$\begin{aligned}
\delta_i &\leq \sup_{\Pi_j(x_\mathcal{T}) = x_\mathcal{B}} \mathbb{E}\bigl|\alpha_{j_i}\bigl(x_\mathcal{T}, \xi^i_2\bigr) - \alpha_{j_{i-1}}\bigl(x_\mathcal{B}, \Pi_{j_{i-1}}\xi^i_2\bigr)\bigr|\\
&= \sup_{\Pi_j(x_\mathcal{T}) = x_\mathcal{B}} \mathbb{E}\biggl|1 \wedge \exp\biggl(\frac{1}{2}\bigl\|y - G_{j_i}(x_\mathcal{T})\bigr\|^2_{\mathbb{R}^d} - \frac{1}{2}\bigl\|y - G_{j_i}\bigl(\xi^i_2\bigr)\bigr\|^2_{\mathbb{R}^d}\biggr)\\
&\quad - 1 \wedge \exp\biggl(\frac{1}{2}\bigl\|y - G_{j_{i-1}}(x_\mathcal{B})\bigr\|^2_{\mathbb{R}^d} - \frac{1}{2}\bigl\|y - G_{j_{i-1}}\bigl(\Pi_{j_{i-1}}\xi^i_2\bigr)\bigr\|^2_{\mathbb{R}^d}\biggr)\biggr|\\
&\lesssim \sup_{\Pi_j(x_\mathcal{T}) = x_\mathcal{B}} \bigl|\bigl\|y - G_{j_i}(x_\mathcal{T})\bigr\|^2_{\mathbb{R}^d} - \bigl\|y - G_{j_{i-1}}(x_\mathcal{B})\bigr\|^2_{\mathbb{R}^d}\bigr|\\
&\quad + \mathbb{E}\bigl|\bigl\|y - G_{j_i}\bigl(\xi^i_2\bigr)\bigr\|^2_{\mathbb{R}^d} - \bigl\|y - G_{j_{i-1}}\bigl(\Pi_{j_{i-1}}\xi^i_2\bigr)\bigr\|^2_{\mathbb{R}^d}\bigr|.
\end{aligned}$$

Using Assumption 4.1, we get

$$\begin{aligned}
&\bigl|\bigl\|y - G_{j_i}(x_{\mathcal{T}})\bigr\|^2_{\mathbb{R}^d} - \bigl\|y - G_{j_{i-1}}(x_{\mathcal{B}})\bigr\|^2_{\mathbb{R}^d}\bigr| \\
&\quad = \bigl|\bigl\langle 2y - G_{j_i}(x_{\mathcal{T}}) - G_{j_{i-1}}(x_{\mathcal{B}}), G_{j_i}(x_{\mathcal{T}}) - G_{j_{i-1}}(x_{\mathcal{B}})\bigr\rangle\bigr| \\
&\quad \le \bigl\|2y - G_{j_i}(x_{\mathcal{T}}) - G_{j_{i-1}}(x_{\mathcal{B}})\bigr\|_{\mathbb{R}^d}\bigl\|G_{j_i}(x_{\mathcal{T}}) - G_{j_{i-1}}(x_{\mathcal{B}})\bigr\|_{\mathbb{R}^d} \\
&\quad \lesssim j_{i-1}^{-\beta}
\end{aligned}$$

and similarly

$$\bigl|\bigl\|y - G_{j_i}\bigl(\xi_2^i\bigr)\bigr\|^2_{\mathbb{R}^d} - \bigl\|y - G_{j_{i-1}}\bigl(\Pi_{j_{i-1}}\xi_2^i\bigr)\bigr\|^2_{\mathbb{R}^d}\bigr| \lesssim j_{i-1}^{-\beta}.$$

Combining, we obtain

$$\delta_i \lesssim j_{i-1}^{-\beta}.$$

We next introduce the σ-algebra

$$\mathcal{S}_l = \sigma\bigl(\bigl\{b^l_{\mathcal{T},k} = b^i_{\mathcal{B},k} \mid k = 1, \dots, l\bigr\}\bigr)$$

with the convention that $S_0 = \{0, \Omega\}$ and let $\Pi_{j_{i-1}}\mathcal{T}_0^i = \mathcal{B}_0^i$ be arbitrary. Then we calculate

$$\begin{aligned}
&\mathbb{P}_{(\mathcal{T}_0^i,\mathcal{B}_0^i)}\bigl(\exists -\tau \ge l > 0 : b^i_{\mathcal{T},l} \ne b^i_{\mathcal{B},l}\bigr) \\
&\quad = 1 - \mathbb{P}_{(\mathcal{T}_0^i,\mathcal{B}_0^i)}\bigl(\forall l < -\tau b^i_{\mathcal{T},l} = b^i_{\mathcal{B},l}\bigr) \\
&\quad = 1 - \prod_{l=1}^{-\tau}\mathbb{P}_{(\mathcal{T}_0^i,\mathcal{B}_0^i)}\bigl(b^i_{\mathcal{T},l} = b^i_{\mathcal{B},l}|S_{l-1}\bigr) \\
&\quad = 1 - \prod_{l=1}^{-\tau}\bigl(1 - \mathbb{P}_{(\mathcal{T}_0^i,\mathcal{B}_0^i)}\bigl(b^i_{\mathcal{T},l} \ne b^i_{\mathcal{B},l}|S_{l-1}\bigr)\bigr) \\
&\quad = 1 - \prod_{l=1}^{-\tau}\bigl(1 - \mathbb{E}\bigl(\mathbb{P}_{(\mathcal{T}_0^i,\mathcal{B}_0^i)}\bigl(b^i_{\mathcal{T},l} \ne b^i_{\mathcal{B},l}|\mathcal{F}_{l-1} \vee S_{l-1}\bigr)|S_{l-1}\bigr)\bigr),
\end{aligned}$$

where the last step follows by the tower property of conditional expectation. The Markov property and Bernulli's inequality yield

$$\begin{aligned}
&\mathbb{P}_{(\mathcal{T}_0^i,\mathcal{B}_0^i)}\bigl(\exists -\tau \ge l > 0 : b^i_{\mathcal{T},l} \ne b^i_{\mathcal{B},l}\bigr) \\
&\quad = 1 - \prod_{l=1}^{-\tau}\bigl(1 - \mathbb{E}\bigl(\mathbb{P}_{(\mathcal{T}_{l-1}^i,\mathcal{B}_{l-1}^i)}\bigl(b^i_{\mathcal{T},1} \ne b^i_{\mathcal{B},1}\bigr)|S_{l-1}\bigr)\bigr) \\
&\quad \le 1 - \prod_{l=1}^{-\tau}(1 - \delta_i) = 1 - (1 - \delta_i)^{-\tau} \le 1 - (1 - \delta_i)^{a_{i-1}} \le a_{i-1}\delta_{j_{i-1}j_i}.
\end{aligned}$$

Hence, we have

$$\mathbb{P}(A) \le a_{i-1}\delta_i,$$

and putting things together, we obtain that

$$E_1 \le \|f\|_\infty^2\big((1-\alpha_\star)^{a_{i-1}} + a_{i-1}\delta_i\big).$$

We thus have

$$\mathbb{E}_1 \le \|f\|_\infty^2\big(\exp(-c_\star a_{i-1}) + a_{i-1} j_{i-1}^{-\beta}\big),$$

where $c_\star = -\log(1-\alpha_\star)$. In order to optimise the right-hand side and since the first term decreases while the second term increases with a_i, we need to balance the two terms by choosing a_i as an appropriate function of j_i. Indeed, using [4], Lemma 4.5, we have that for $a_i \sim \frac{\beta}{c_\star}\log(j_i)$ the two terms are asymptotically balanced as $i \to \infty$ so that for this choice $E_1 \lesssim j_{i-1}^{-\beta}$ (check that this is true).

Now we treat the second term E_2, which by Assumption 4.2 satisfies

$$E_2 = \mathbb{E}\big(f\big(\mathcal{T}_{a_i}^{j_i}\big) - f\big(\Pi_{j_{i-1}}\mathcal{T}_{a_i}^{j_i}\big)\big)^2 \lesssim j_{i-1}^{-\kappa}.$$

Finally, combining the bounds for E_1 and E_2 yields

$$\|\Delta_i\|_2^2 \le 2(E_1+E_2) \lesssim j_{i-1}^{-(\beta\wedge\kappa)}.$$ □

Proof of Theorem 4.5. In order for the unbiasing procedure to work, we need to have both finite computing time and finite variance of the estimator Z. By the considerations at the end of Section 1.3, it suffices to verify (1.7) to get the unbiasedness and finite variance of Z. Below, we use c as a generic positive constant which may be change between occurrences.

Using Lemma 8.1 and according to the stated choices of the relevant parameters, we have

$$\sum_{i\le l}\frac{\|\Delta_i\|_2\|\Delta_l\|_2}{\mathbb{P}(N\ge i)} = \sum_{i=0}^{\infty}\frac{\|\Delta_i\|_2}{\mathbb{P}(N\ge i)}\sum_{l=i}^{\infty}\|\Delta_l\|_2$$

$$\le c\sum_{i=0}^{\infty} i^{-\frac{rq}{2}+t}\sum_{l=i}^{\infty} l^{-\frac{rq}{2}} \le c\sum_{i=0}^{\infty} i^{1-rq+t},$$

provided $q > \frac{2}{r}$ which holds since $q > \frac{3}{r-\theta}$. The right-hand side is finite provided $t < rq-2$.

Regarding the expected computing time of Z, by Assumption 4.3, we have

$$\mathbb{E}[\tau] = \sum_{i}^{\infty} t_i\mathbb{P}(N\ge i) \lesssim \sum_{i}^{\infty} i^{\theta q - t}\log(i),$$

which is finite provided $t > 1+\theta q$.

Concatenating, we have that for the unbiased procedure to work we need to choose $t \in (1+\theta q, rq-2)$, which is possible since $1+\theta q < rq-2$ under the assumption $q > \frac{3}{r-\theta}$. □

8.3. Proofs of the results in Section 5

In this section we present the proofs of Theorems 5.3 and 5.4. The crucial step for both proofs is to derive bounds on $\mathbb{E}d(\mathcal{T}_0^i, \mathcal{B}_0^i)$ for the appropriate distance d, which in turn give bounds on the decay of $\|\Delta_i\|_2$; the method of generating $\mathcal{T}_0^i, \mathcal{B}_0^i$ and Δ_i is summarised in Algorithm 5. The main idea used to obtain such bounds is explained in Section 8.3.1 under artificial conditions. The rigorous bounds used for Theorems 5.3 and 5.4 are contained in Lemma 8.2 and Lemma 8.3 in Section 8.3.3, respectively. Obtaining these bounds is based on results known for coupling of the pCN algorithms on the same state space which are summarised in Section 8.3.2. Finally, in Section 8.3.4 we put things together and prove Theorems 5.3 and 5.4.

8.3.1. *Main idea*

In the following, we show how to obtain bounds on the contraction of the transdimensional coupling $K_{j_{i-1}}^{j_i}$ defined in (5.4), under the following artificial assumption on the fixed state space coupling K_i defined in (5.5):

$$\mathbb{E}_W d_\tau\bigl(\varphi_{\mathcal{T}}^i(x_1, W), \varphi_{\mathcal{T}}^i(x_2, W)\bigr) \le r d_\tau(x_1, x_2). \tag{8.4}$$

This assumption does not hold for the pCN algorithm, but allows us to present the strategy of our proofs while avoiding technicalities and overloaded notation. The fact that d_τ satisfies the triangle inequality is crucial for our analysis. In particular, it allows us to introduce intermediate steps $\mathcal{I}_k^i = \varphi_{\mathcal{I}}^i(\mathcal{B}_{k-1}^i, W_k^i)$ by performing a transition from a state of the lower level chain $\mathcal{B}_k^i$, according to the transition kernel P_{j_i} of the high level chain $\mathcal{T}_k^i$. This enables us to use (8.4) to control the distance between $\mathcal{I}_k^i$ and $\mathcal{T}_k^i$, while at the same time $\mathcal{I}_k^i$ is with high probability close to $\mathcal{B}_k^i = \varphi_{\mathcal{B}}^i(\mathcal{B}_{k-1}^i, W_k^i)$, since they have the same starting point. We show that this intuition is accurate below.

The intermediate step $\mathcal{I}_k^i$ is constructed as follows:

$$\begin{aligned}
\hat{\mathcal{I}}_k^i &= \varphi_{\hat{\mathcal{T}}}^i\bigl(\mathcal{B}_{k-1}^i, \xi_k^i\bigr) = \rho\mathcal{B}_{k-1}^i + \bigl(1-\rho^2\bigr)^{\frac{1}{2}}\xi_k^i, \\
\mathcal{I}_k^i &= \varphi_{\mathcal{T}}^i\bigl(\mathcal{B}_{k-1}^i, W_k^i\bigr) = \mathbb{1}_{[0,\alpha(\mathcal{B}_{k-1}^i,\hat{\mathcal{I}}_k^i)]}\bigl(U_k^i\bigr)\hat{\mathcal{I}}_k^i + \mathbb{1}_{(\alpha(\mathcal{B}_{k-1}^i,\hat{\mathcal{I}}_k^i),1]}\bigl(U_k^i\bigr)\mathcal{B}_{k-1}^i.
\end{aligned}$$

Using the triangle inequality, we get the bound

$$\begin{aligned}
\mathbb{E}d_\tau\bigl(\mathcal{T}_k^i, \mathcal{B}_k^i\bigr) &\le \mathbb{E}\bigl[d_\tau\bigl(\mathcal{T}_k^i, \mathcal{I}_k^i\bigr) + d_\tau\bigl(\mathcal{I}_k^i, \mathcal{B}_k^i\bigr)\bigr] \\
&= \mathbb{E}\bigl[\mathbb{E}\bigl(d_\tau\bigl(\mathcal{T}_k^i, \mathcal{I}_k^i\bigr) + d_\tau\bigl(\mathcal{I}_k^i, \mathcal{B}_k^i\bigr)|\mathcal{F}_{k-1}\bigr)\bigr],
\end{aligned} \tag{8.5}$$

where $\mathcal{F}_k = \sigma(\{\mathcal{T}_l^i, \mathcal{B}_l^i\} \mid l \le k)$. We use (8.4) together with the Markov property in order to get

$$\mathbb{E}\bigl[d_\tau\bigl(\mathcal{T}_k^i, \mathcal{I}_k^i\bigr)|\mathcal{F}_{k-1}\bigr] = \mathbb{E}\bigl[(K_i d_\tau)\bigl(\mathcal{T}_{k-1}^i, \mathcal{B}_{k-1}^i\bigr)\bigr] \le \mathbb{E}\bigl[r d_\tau\bigl(\mathcal{T}_{k-1}^i, \mathcal{B}_{k-1}^i\bigr)\bigr]. \tag{8.6}$$

Therefore it is left to consider $\mathbb{E}(d_\tau(\mathcal{I}_0^i, \mathcal{B}_0^i)|\mathcal{F}_{-1})$. Since $d_\tau \leq 1$, we have the bound

$$\begin{aligned}\mathbb{E}\big(d_\tau\big(\mathcal{I}_k^i, \mathcal{B}_k^i\big) \mid \mathcal{F}_{k-1}\big) \leq {} & \mathbb{E}(0 \cdot \mathbb{1}_{\text{both reject}} \mid \mathcal{F}_{k-1}) \\ & + \mathbb{E}\left(\frac{(1-\rho^2)^{\frac{1}{2}} \|\xi_k^i - \Pi_{j_{i-1}} \xi_k^i\|}{\tau} \cdot \mathbb{1}_{\text{both accept}} \mid \mathcal{F}_{k-1}\right) \\ & + \mathbb{E}(1 \cdot \mathbb{1}_{\text{one accepts}} \mid \mathcal{F}_{k-1}),\end{aligned} \tag{8.7}$$

where $\mathbb{1}_{\text{one accepts}} = \mathbb{1}_{\{\mathcal{I}_k^i = \hat{\mathcal{I}}_k^i \text{ xor } \mathcal{B}_k^i = \hat{\mathcal{B}}_k^i\}}$. The probability that only one of the chains accepts can be bounded using the Markov property as follows:

$$\begin{aligned}&\mathbb{P}(\text{one accepts} \mid \mathcal{F}_{k-1}) \\ &\quad = \mathbb{E}_{\xi_k^i, U_k^i}\big(\mathbb{1}_{[\alpha(\mathcal{B}_{k-1}^i, \hat{\mathcal{I}}_k^i), \alpha(\mathcal{B}_{k-1}^i, \hat{\mathcal{B}}_k^i)]}\big(U_k^i\big) + \mathbb{1}_{[\alpha(\hat{\mathcal{B}}_{k-1}^i, \mathcal{B}_k^i), \alpha(\mathcal{B}_{k-1}^i, \hat{\mathcal{I}}_k^i)]}\big(U_k^i\big)\big) \\ &\quad = \mathbb{E}_{\xi_k^i}\big|\alpha\big(\mathcal{B}_{k-1}^i, \hat{\mathcal{I}}_k^i\big) - \alpha\big(\mathcal{B}_{k-1}^i, \hat{\mathcal{B}}_k^i\big)\big| \\ &\quad = \mathbb{E}_{\xi_k^i}\big|1 \wedge \exp\big(g\big(\mathcal{B}_{k-1}^i\big) - g\big(\hat{\mathcal{I}}_k^i\big)\big) - 1 \wedge \exp\big(g\big(\mathcal{B}_{k-1}^i\big) - g\big(\hat{\mathcal{B}}_k^i\big)\big)\big| \\ &\quad \leq C_g \mathbb{E}_{\xi_k^i}\big\|\hat{\mathcal{I}}_k^i - \hat{\mathcal{B}}_k^i\big\| \\ &\quad \leq C_g\big(1-\rho^2\big)^{\frac{1}{2}} \mathbb{E}\big\|\xi_k^i - \Pi_{j_{i-1}} \xi_k^i\big\|,\end{aligned} \tag{8.8}$$

where C_g depends on the Lipschitz constant of the log-change of measure g. The second term on the right-hand side of (8.7) is of similar form, so that we get the overall bound

$$\begin{aligned}\mathbb{E} d_\tau\big(\mathcal{B}_k^i, \mathcal{I}_k^i\big) & \leq \left(\frac{1}{\tau} + C_g\right)\big(1-\rho^2\big)^{\frac{1}{2}} \mathbb{E}\big\|\xi_k^i - \Pi_{j_{i-1}} \xi_k^i\big\| \\ & \leq \left(\frac{1}{\tau} + C_g\right)\big(1-\rho^2\big)^{\frac{1}{2}} K \sqrt{\sum_{k=j_{i-1}+1}^{j_i} \lambda_k} \\ & =: C_{j_{i-1}, j_i},\end{aligned} \tag{8.9}$$

where we used Cauchy–Schwarz inequality in the last step. Repeated use of the Markov property and the bounds (8.5), (8.6) and (8.9), yields that

$$\begin{aligned}\mathbb{E} d_\tau\big(\mathcal{T}_0^i, \mathcal{B}_0^i\big) & \leq \mathbb{E}\big[\mathbb{E}\big(d_\tau\big(\mathcal{I}_0^i, \mathcal{B}_0^i\big) + d_\tau\big(\mathcal{T}_0^i, \mathcal{I}_0^i\big)|\mathcal{F}_{-1}\big)\big] \\ & \leq \mathbb{E}\big[r d_\tau\big(\mathcal{T}_{-1}^i, \mathcal{B}_{-1}^i\big) + C_{j_{i-1}, j_i}\big] \\ & \leq r\big(r \mathbb{E} d_\tau\big(\mathcal{T}_{-2}^i, \mathcal{B}_{-2}^i\big) + C_{j_{i-1}, j_i}\big) + C_{j_{i-1}, j_i} \\ & \quad \vdots\end{aligned} \tag{8.10}$$

$$\begin{aligned}
&\leq r\big(\dots\big(r\mathbb{E}d_\tau\big(\mathcal{T}^i_{-a_{i-1}},\mathcal{B}^i_{-a_{i-1}}\big)+C_{j_{i-1},j_i}\big)\dots\big)+C_{j_{i-1},j_i}\\
&\leq r^{a_{i-1}}\mathbb{E}d_\tau\big(\mathcal{T}^i_{-a_{i-1}},\mathcal{B}^i_{-a_{i-1}}\big)+C_{j_{i-1},j_i}\frac{1-r^{a_{i-1}}}{1-r}\\
&\leq r^{a_{i-1}}+C_{j_{i-1},j_i}\frac{1-r^{a_{i-1}}}{1-r},
\end{aligned}$$

where in the last step we used that $d_\tau \leq 1$. Our strategy thus indeed gives a bound on the contraction of the transdimensional coupling $K^{j_i}_{j_{i-1}}$ under the artificial assumption (8.4) on the contraction of K_i. We use the same strategy to get the required contraction bounds in the more realistic settings considered in Lemmas 8.2 and 8.3.

8.3.2. *Overview of the coupling bounds*

We next describe how the existing literature yields that the fixed state space coupling in (5.5) leads to contraction with respect to

$$d_\tau(x,y)=1\wedge\frac{\|x-y\|}{\tau}\quad\text{and}\quad\tilde{d}(x,y):=\sqrt{d_\tau(x,y)\big(1+V(x)+V(y)\big)}.$$

For simplicity, below we assume that $j_i=i$. This particular coupling is called *the basic coupling*, [18]. Recall that the contraction bound for a particular coupling is always an upper bound for the Wasserstein distance of the transition kernel, see Remark 2.2.2. In the following, we summarise the relevant results and make connections to geometric ergodicity.

Verifying that a particular coupling contracts is often difficult, but [10,17] and [11] give verifiable conditions which resemble the well-known conditions for geometric and polynomial ergodicity. Geometric ergodicity is usually established using the Harris theorem by verifying the existence of a Lyapunov function, also called geometric drift condition. That is, it suffices to show the existence of a function V, $0<\lambda<1$ and $b>0$ such that

$$PV\leq\lambda V+b \tag{8.11}$$

and showing that an appropriate small set exists, see [28], Section 3.4. The problem is that the resulting error bounds on the ergodic average deteriorate with dimension because it is difficult to find *good* small sets.

This problem is alleviated when considering Wasserstein convergence. In particular, the article [17] establishes a weak Harris theorem. It shows exponential convergence with respect to the Wasserstein distance based on $\tilde{d}(x,y)=\sqrt{d(x,y)(1+V(x)+V(y))}$, for $d(x,y)\leq 1$ a distance-like function. Below we use the letter d to also denote the Wasserstein distance and hope that this does not cause confusion. The small set condition of the Harris theorem is replaced by the requirements that:

1. a sub-level set S of V is d-small, that is, for all x and y in S

$$d\big(P(x,\cdot),P(y,\cdot)\big)\leq s<1; \tag{8.12}$$

2. the transition kernel P is d-contracting, that is, there is a $0 < c < 1$ such that for $d(x, y) < 1$

$$d\big(P(x, \cdot), P(y, \cdot)\big) \le cd(x, y). \tag{8.13}$$

For a summary of the weak Harris theorem we refer the reader to Section 2.2.1 of [18]. Equations (8.12) and (8.13) are typically established using the fact that the Wasserstein distance can be bounded using a particular coupling. That is, it suffices to establish the existence of couplings $K^{(1)}$ and $K^{(2)}$ such that

$$\big(K^{(1)}d\big)(x, y) \le s < 1, \tag{8.14}$$

$$\big(K^{(2)}d\big)(x, y) \le cd(x, y) \qquad \text{for } d(x, y) < 1. \tag{8.15}$$

An inspection of the proof of the weak Harris theorem in [17], shows that in fact the contraction property is established for the coupling arising from

- if $d(x, y) < 1$ use coupling $K^{(2)}$;
- else if $x, y \in S$ use coupling $K^{(1)}$;
- else use any coupling,

rather than directly for the Wasserstein distance which takes the infimum over all couplings. Thus, the same is true for Theorems 2.14 and 2.17 of [18], that derive a non-explicit but dimension-independent contraction rate for the basic coupling K_i of the pCN algorithm, for target measures which are changes of measure from a Gaussian distribution with log-density satisfying Assumption 5.1. More precisely, the proof shows that:

- (8.11) is satisfied for $V(x) = \exp(\|x\|)$ and with b and λ which are dimension independent;
- there exists a dimension-independent τ, such that the basic coupling K_i satisfies both (8.14) and (8.15) for the distance d_τ, for s and c which are also dimension-independent.

In particular, this shows that for any $0 < r < 1$, there exists $n_0 = n_0(r) \in \mathbb{N}$ such that

$$\big((K_i)^{n_0(r)}\tilde{d}\big)(x, y) \le r\tilde{d}(x, y) \qquad \text{for any } x, y \in \mathcal{X}_i, i \in \mathbb{N}, \tag{8.16}$$

for

$$\tilde{d}(x, y) := \sqrt{d_\tau(x, y)\big(1 + V(x) + V(y)\big)}$$

where $V(x) = \exp(\|x\|)$.

The work of [17] has been extended

- in [10] to cover polynomial ergodicity using more complicated drift conditions;
- in [11] to obtain more explicit bounds in the geometric case.

The article [10] explicitly considers the pCN algorithm. In particular, equation (68) in [10] establishes that

$$K_i d_\tau \le d_\tau. \tag{8.17}$$

Combining Proposition 12 of [10], Lemma 3.2 of [18] and Theorem 1 of [11], we get that the basic coupling decays exponentially

$$(K_i)^n d_\tau(x,y) \le Cr^n\bigl(V(x)+V(y)\bigr), \tag{8.18}$$

where C is dimension independent and V as above.

In the next subsection we will employ the above contraction results for the fixed state-space basic coupling K_i of the pCN algorithm, to show the decay of the transdimensional coupling $K^{j_i}_{j_{i-1}}$ of the pCN algorithm, defined in (5.4).

8.3.3. *Coupling bounds between $\mathcal{T}_0^i$ and $\mathcal{B}_0^i$*

In the following we use the results reviewed in the previous subsection to obtain coupling bounds between $\mathcal{T}_0^i$ and $\mathcal{B}_0^i$ in terms of $d_\tau(x,y) = 1 \wedge \frac{\|x-y\|_{\mathcal{X}}}{\tau}$ (Lemma 8.2) and $\tilde{d}(x,y) := \sqrt{d(x,y)(1+V(x)+V(y))}$ (Lemma 8.3). These bounds in turn imply bounds on the decay of $\|\Delta_i\|_2$ which are crucial for the proofs of Theorem 5.3 and 5.4.

Lemma 8.2. *Under Assumption 5.1, there exist $\tau, C > 0$ and $r \in (0,1)$, such that*

$$\mathbb{E} d_\tau\bigl(\mathcal{T}_0^i, \mathcal{B}_0^i\bigr) \le Cr^{a_{i-1}} + a_{i-1}C_{j_{i-1},j_i}, \tag{8.19}$$

with $C_{j_{i-1},j_i} := \sqrt{\sum_{k=j_{i-1}+1}^{j_i} \lambda_k}$. In particular if $f : \mathcal{X} \to \mathbb{R}$ is s-Hölder continuous with respect to d_τ for some $s \in [\frac{1}{2}, 1]$, for the choice $j_i \sim r^{-a_i \frac{2}{2\alpha-1}}$ we get the estimate

$$\|\Delta_i\|_2^2 \lesssim r^{a_{i-1}}. \tag{8.20}$$

Proof. From Section 8.3.2, we know that there are τ, C and $r \in (0,1)$ independent of i, such that the fixed state space coupling K_i satisfies

$$\begin{aligned} (K_i)^n d_\tau(x,y) &\le Cr^n\bigl(V(x)+V(y)\bigr) \\ K_i d_\tau &\le d_\tau, \end{aligned} \tag{8.21}$$

for any $n \in \mathbb{N}$ and for the Lyapunov function $V(x) = \exp(\|x\|)$. This statement is much weaker than the artificial assumption (8.4) in Section 8.3.1 because of the multiplicative constant and the Lyapunov function. As a result we cannot simply recurse as in (8.10). In the following, the constant C may change from occurrence to occurrence, but C it is always independent of n and i.

We define recursively for $l \in \mathbb{N}$ and for a fixed $r \in \mathbb{Z}$, the l-step random functions

$$\begin{aligned} \varphi_{\mathcal{B}}^{i,l}\bigl(x, \{W_s\}_{s=1+r}^{l+r}\bigr) &:= \varphi_{\mathcal{B}}^{i}\bigl(\varphi_{\mathcal{B}}^{i,l-1}\bigl(x, \{W_s\}_{s=r+1}^{l+r-1}\bigr), W_{l+r}\bigr) \\ \varphi_{\mathcal{T}}^{i,l}\bigl(x, \{W_s\}_{s=1+r}^{l+r}\bigr) &:= \varphi_{\mathcal{T}}^{i}\bigl(\varphi_{\mathcal{T}}^{i,l-1}\bigl(x, \{W_s\}_{s=r+1}^{r+l-1}\bigr), W_{l+r}\bigr) \end{aligned}$$

with the convention that $\varphi_{\mathcal{T}}^{i,0} = \varphi_{\mathcal{B}}^{i,0} = Id$. Using the triangle inequality and following the strategy described at the beginning of Section 8.3.1, we can bound

$$\begin{aligned}
\mathbb{E}d_\tau\big(\mathcal{T}_0^i, \mathcal{B}_0^i\big) &= \mathbb{E}d_\tau\big(\varphi_{\mathcal{T}}^{i,a_{i-1}}\big(\varphi_{\mathcal{T}}^{i,a_i-a_{i-1}}\big(x_0, \{W_s\}_{s=-a_i+1}^{-a_{i-1}}\big), \{W_s\}_{s=-a_{i-1}+1}^{0}\big), \\
&\quad \varphi_{\mathcal{B}}^{i,a_{i-1}}\big(x_0, \{W_s\}_{s=-a_{i-1}+1}^{0}\big)\big) \\
&\le \mathbb{E}d_\tau\big(\varphi_{\mathcal{T}}^{i,a_{i-1}}\big(\mathcal{T}_{-a_{i-1}}^i, \{W_s\}_{s=-a_{i-1}+1}^{0}\big), \varphi_{\mathcal{T}}^{i,a_{i-1}}\big(x_0, \{W_s\}_{s=-a_{i-1}+1}^{0}\big)\big) \qquad (8.22) \\
&\quad + \mathbb{E}d_\tau\big(\varphi_{\mathcal{T}}^{i,a_{i-1}}\big(x_0, \{W_s\}_{s=-a_{i-1}+1}^{0}\big), \varphi_{\mathcal{B}}^{i,a_{i-1}}\big(x_0, \{W_s\}_{s=-a_{i-1}+1}^{0}\big)\big) \\
&=: R_1 + R_2.
\end{aligned}$$

Using (8.21) and the Markov property, we have that

$$R_1 \le C r^{a_{i-1}} \mathbb{E}\big(V(x_0) + V\big(\mathcal{T}_{a_{i-1}}^i\big)\big) \le C r^{a_{i-1}}, \qquad (8.23)$$

where the second inequality follows from the fact that $\sup_{j,n} P_j^n V(x_0) < \infty$. This can be seen by induction on (8.11) which as discussed in the previous subsection for the pCN algorithm is satisfied with dimension independent parameters, which gives

$$P^n V = \lambda^n V + \frac{1-\lambda^n}{1-\lambda} b,$$

implying

$$\sup_n P^n V \le V + \frac{1}{1-\lambda} b.$$

By repeatedly using the triangle inequality in order to introduce intermediate steps which differ from each other in the evolution at only one time-step, we can estimate R_2 as below:

$$\begin{aligned}
R_2 &\le \mathbb{E}\big[d_\tau\big(\varphi_{\mathcal{T}}^{i,a_{i-1}-1}\big(\varphi_{\mathcal{T}}^i(x_0, W_{-a_{i-1}+1}), \{W_s\}_{s=-a_{i-1}+2}^{0}\big), \\
&\quad \varphi_{\mathcal{T}}^{i,a_{i-1}-1}\big(\varphi_{\mathcal{B}}^i(x_0, W_{-a_{i-1}+1}), \{W_s\}_{s=-a_{i-1}+2}^{0}\big)\big) \\
&\quad + d_\tau\big(\varphi_{\mathcal{T}}^{i,a_{i-1}-1}\big(\varphi_{\mathcal{B}}^i(x_0, W_{-a_{i-1}+1}), \{W_s\}_{s=-a_{i-1}+2}^{0}\big), \\
&\quad \varphi_{\mathcal{B}}^{i,a_{i-1}-1}\big(\varphi_{\mathcal{B}}^i(x_0, W_{a_{i-1}+1}), \{W_s\}_{s=-a_{i-1}+2}^{0}\big)\big)\big] \\
&\quad \vdots \\
&\le \mathbb{E} \sum_{k=0}^{a_{i-1}-1} d_\tau\big(\varphi_{\mathcal{T}}^{i,k}\big(\varphi_{\mathcal{T}}^i\big(\varphi_{\mathcal{B}}^{i,a_i-1-k}\big(x_0, \{W_s\}_{s=-a_{i-1}+1}^{-k-1}\big), W_{-k}\big), \{W_s\}_{s=-k+1}^{0}\big), \\
&\quad \varphi_{\mathcal{T}}^{i,k}\big(\varphi_{\mathcal{B}}^i\big(\varphi_{\mathcal{B}}^{i,a_i-1-k}\big(x_0, \{W_s\}_{s=-a_{i-1}+1}^{-k-1}\big), W_{-k}\big), \{W_s\}_{s=-k+1}^{0}\big)\big).
\end{aligned}$$

Since $K_i d_\tau \le d_\tau$ by (8.21), we hence have

$$R_2 \le \mathbb{E} \sum_{k=0}^{a_{i-1}-1} d_\tau \big(\varphi_{\mathcal{T}}^i \big(\varphi_{\mathcal{B}}^{i, a_i - 1 - k} \big(x_0, \{W_s\}_{s=-a_{i-1}+1}^{-k-1} \big), W_{-k} \big), \\ \varphi_{\mathcal{B}}^i \big(\varphi_{\mathcal{B}}^{i, a_i - 1 - k} \big(x_0, \{W_s\}_{s=-a_{i-1}+1}^{-k-1} \big), W_{-k} \big) \big).$$

Similarly to (8.9), we see that for each k the summands are bounded by C_{j_{i-1}, j_i}, hence we get that

$$R_2 \le a_{i-1} C_{j_{i-1}, j_i} \tag{8.24}$$

with $C_{j_{i-1}, j_i} := \sqrt{\sum_{k=j_{i-1}+1}^{j_i} \lambda_k}$. Combining (8.23) and (8.24) with (8.22), we get the desired bound (8.19).

To (roughly) optimise the right-hand side of (8.19), we require that the two terms have bounds of the same order. Since $\lambda_\ell \lesssim \ell^{-2\alpha}$, we have that

$$C_{j_{i-1}, j_i} = \left(\sum_{k=j_{i-1}+1}^{j_i} \lambda_k \right)^{\frac{1}{2}} \lesssim j_{i-1}^{\frac{1-2a}{2}}.$$

We hence require $a_{i-1} j_{i-1}^{\frac{1-2a}{2}} = r^{a_{i-1}}$ and using [4], Lemma 4.5, we get that the choice $j_i \sim r^{a_i \frac{2}{1-2a}}$ as $i \to \infty$ yields that

$$\mathbb{E} d_\tau \big(\mathcal{T}_0^i, \mathcal{B}_0^i \big) \lesssim r^{a_{i-1}}.$$

The estimate (8.20) then follows from (5.8). □

Lemma 8.3. *Under Assumption 5.1, there exist τ and $r \in (0, 1)$, such that*

$$\mathbb{E} \tilde{d} \big(\mathcal{T}_0^i, \mathcal{B}_0^i \big) \lesssim r^{a_{i-1}} + C_{j_{i-1}, j_i}^{\frac{1}{2}}, \tag{8.25}$$

with $C_{j_{i-1}, j_i} := \sqrt{\sum_{k=j_{i-1}+1}^{j_i} \lambda_k}$. In particular if $f : \mathcal{X} \to \mathbb{R}$ is $\frac{1}{2}$-Hölder continuous with respect to $\tilde{d}$, for the choice $j_i \sim C r^{-a_i \frac{4}{2\alpha - 1}}$ we have the estimate

$$\| \Delta_i \|_2^2 \lesssim r^{a_{i-1}}. \tag{8.26}$$

Proof. Combining (8.17) with (8.11), which as discussed in the previous subsection are both satisfied for the pCN algorithm with dimension independent constants $b > 0$, $0 < \lambda < 1$ and

$\tau > 0$ and for $V(x) = \exp(\|x\|)$, we get the following bound

$$\begin{aligned}\big((K_i)^n \tilde{d}\big)(x, y) &\le \big((K_i)^n d_\tau(x, y)\big)^{\frac{1}{2}} \big(1 + \big((K_i)^n V\big)(x) + \big((K_i)^n V\big)(y)\big)^{\frac{1}{2}} \\ &\le \big(d_\tau(x, y)\big)^{\frac{1}{2}} \left(1 + V(x) + V(y) + \frac{2b}{1-\lambda}\right)^{\frac{1}{2}} \qquad (8.27) \\ &\le \sqrt{\frac{2b}{1-\lambda} + 1} \cdot \tilde{d}(x, y) := \tilde{b}\tilde{d}(x, y).\end{aligned}$$

In the first inequality we used Cauchy–Schwarz and in the second, we used that (8.11) implies

$$P^n V \le \lambda^n V + \frac{b}{1-\lambda}. \qquad (8.28)$$

A major problem is that $\tilde{d}$ usually does not satisfy the triangle inequality, so that the method described in Section 8.3.1 cannot be applied directly. However, this can be circumvented using the technique of Section 4.1.1 of [18]. More precisely, we define

$$\hat{d}(x, y) := \sqrt{\inf_{n, x=z_1, \dots, y=z_n} \sum_{j=1}^{n-1} d_0(z_j, z_{j+1})}$$

with $d_0 = d_\tau(1 + V(x) + V(y))$; $\hat{d}$ satisfies the triangle inequality by construction. Following the proof of Lemma 4.1.1 in [18], it is possible to show that there exists a constant $C_L \le 1$ such that

$$C_L \tilde{d} \le \hat{d} \le \tilde{d}.$$

Using (8.16), we choose $n_0 = n_0(\frac{C_L}{2})$ such that

$$\big((K_j)^{n_0} \tilde{d}\big)(x, y) \le \frac{C_L}{2} \tilde{d}(x, y) \qquad \text{for any } j. \qquad (8.29)$$

Using the triangle inequality for $\hat{d}$ and $\tilde{d} \le \frac{1}{C_L} \hat{d}$, we get

$$\begin{aligned}\mathbb{E}\tilde{d}\big(\mathcal{T}_0^i, \mathcal{B}_0^i\big) \le &\frac{1}{C_L} \mathbb{E}\tilde{d}\big(\varphi_{\mathcal{T}}^{i,n_0}\big(\mathcal{T}_{-n_0}^i, \{W_s\}_{s=-n_0+1}^0\big), \varphi_{\mathcal{T}}^{i,n_0}\big(\mathcal{B}_{-n_0}^i, \{W_s\}_{s=-n_0+1}^0\big)\big) \\ &+ \frac{1}{C_L} \mathbb{E}\tilde{d}\big(\varphi_{\mathcal{B}}^{i,n_0}\big(\mathcal{B}_{-n_0}^i, \{W_s\}_{s=-n_0+1}^0\big), \varphi_{\mathcal{T}}^{i,n_0}\big(\mathcal{B}_{-n_0}^i, \{W_s\}_{s=-n_0+1}^0\big)\big) \\ =: &R_1 + R_2,\end{aligned}$$

where the l-step random functions $\varphi_{\mathcal{T}}^{i,l}, \varphi_{\mathcal{B}}^{i,l}$ are defined as in the proof of Lemma 8.2. Using (8.29), we get that

$$R_1 \le \frac{1}{2} \tilde{d}\big(\mathcal{T}_{-n_0}^i, \mathcal{B}_{-n_0}^i\big).$$

Using the triangle inequality for $\hat{d}$, $\hat{d} \leq \tilde{d}$ and (8.27), we have

$$\begin{aligned} R_2 \leq{} & \frac{1}{C_L}\mathbb{E}\sum_{k=0}^{n_0-1} \hat{d}\big\{\varphi_{\mathcal{T}}^{i,k}\big(\varphi_{\mathcal{T}}^{i}\big(\varphi_{\mathcal{B}}^{i,n_0-1-k}\big(\mathcal{B}^i_{-n_0}, \{W_s\}_{s=-n_0+1}^{-k-1}\big), W_{-k}\big), \{W_s\}_{s=-k+1}^{0}\big), \\ & \varphi_{\mathcal{T}}^{i,k}\big(\varphi_{\mathcal{B}}^{i}\big(\varphi_{\mathcal{B}}^{i,n_0-1-k}\big(\mathcal{B}^i_{-n_0}, \{W_s\}_{s=-n_0+1}^{-k-1}\big), W_{-k}\big), \{W_s\}_{s=-k+1}^{0}\big)\big\} \\ \leq{} & \frac{\tilde{b}}{C_L}\mathbb{E}\sum_{k=0}^{n_0-1} \tilde{d}\big\{\varphi_{\mathcal{T}}^{i}\big(\varphi_{\mathcal{B}}^{i,n_0-1-k}\big(\mathcal{B}^i_{-n_0}, \{W_s\}_{s=-n_0+1}^{-k-1}\big), W_{-k}\big), \\ & \varphi_{\mathcal{B}}^{i}\big(\varphi_{\mathcal{B}}^{i,n_0-1-k}\big(\mathcal{B}^i_{-n_0}, \{W_s\}_{s=-n_0+1}^{-k-1}\big), W_{-k}\big)\big\}. \end{aligned} \tag{8.30}$$

The next step is to derive a bound on the one-step difference between $\varphi^i_{\mathcal{B}}(x, W)$ and $\varphi^i_{\mathcal{T}}(x, W)$. We obtain the following bound using the Cauchy–Schwarz inequality, (8.9) and (8.11)

$$\begin{aligned} \mathbb{E}\tilde{d}\big(\varphi^i_{\mathcal{B}}(x, W), \varphi^i_{\mathcal{T}}(x, W)\big) &\leq \big(\mathbb{E}d_\tau\big(\varphi^i_{\mathcal{B}}(x, W), \varphi^i_{\mathcal{T}}(x, W)\big)\big)^{\frac{1}{2}} \\ &\quad \cdot \big(1 + \mathbb{E}V\big(\varphi_{\mathcal{B}}(x, W)\big) + \mathbb{E}V\big(\varphi_{\mathcal{T}}(x, W)\big)\big)^{\frac{1}{2}} \\ &\lesssim C^{\frac{1}{2}}_{j_{i-1},j_i}\big(1 + 2\lambda V(x) + 2b\big)^{\frac{1}{2}}. \end{aligned}$$

Notice that the application of Cauchy–Schwarz here leads to $C^{\frac{1}{2}}_{j_{i-1},j_i}$ instead of C_{j_{i-1},j_i} in Section 8.3.3. This is the reason for the stronger condition $a > 2\theta + \frac{1}{2}$ in Theorem 5.4 compared to $a > \theta + \frac{1}{2}$ in Theorem 5.3. Using this bound on the right-hand side of (8.30), yields

$$R_2 \leq \frac{\tilde{b}}{C_L}\mathbb{E}\sum_{k=0}^{n_0-1} C^{\frac{1}{2}}_{j_{i-1},j_i}\big(1 + 2\lambda V\big(\mathcal{B}^i_{-k-1}\big) + 2b\big)^{\frac{1}{2}}.$$

Using the Cauchy–Schwarz inequality together with (8.28) we thus get that

$$\begin{aligned} R_2 &\leq \frac{\tilde{b}}{C_L}C^{\frac{1}{2}}_{j_{i-1},j_i}\sum_{k=0}^{n_0-1}\left(\mathbb{E}1 + 2\lambda\lambda^{a_{i-1}-k-1}V(x_0) + 2\lambda\frac{2b}{1-\lambda} + 2b\right)^{\frac{1}{2}} \\ &\leq M\frac{\tilde{b}}{C_L}C^{\frac{1}{2}}_{j_{i-1},j_i}n_0\left(1 + 2V(x_0) + \frac{8b}{1-\lambda}\right)^{\frac{1}{2}}, \end{aligned}$$

where the constant M only depends on λ and n_0 and we used that $\lambda < 1$ and a_i is increasing. We abuse notation and write $M = M\frac{\tilde{b}}{C_L}n_0$. Combining the bounds for R_1 and R_2, we obtain that

$$\mathbb{E}\tilde{d}\big(\mathcal{T}^i_0, \mathcal{B}^i_0\big) \leq \frac{1}{2}\mathbb{E}\tilde{d}\big(\mathcal{T}^i_{-n_0}, \mathcal{B}^i_{-n_0}\big) + MC^{\frac{1}{2}}_{j_{i-1},j_i}\left(1 + 2V(x_0) + \frac{8b}{1-\lambda}\right)^{\frac{1}{2}}.$$

Finally, using the Markov property we can iterate the above bound $k = \lfloor \frac{a_{i-1}}{n_0} \rfloor$times to obtain

$$\begin{aligned}\mathbb{E}\tilde{d}(\mathcal{T}_0^i, \mathcal{B}_0^i) &\le \left(\frac{1}{2}\right)^k \mathbb{E}\tilde{d}(\mathcal{T}^i_{-kn_0}, \mathcal{B}^i_{-kn_0}) + 2MC^{\frac{1}{2}}_{j_{i-1},j_i}\left(1 + 2V(x_0) + \frac{8b}{1-\lambda}\right)^{\frac{1}{2}} \\ &\le \left(\frac{1}{2}\right)^k \mathbb{E}\tilde{b}\tilde{d}(\mathcal{T}^i_{-a_{i-1}}, x_0) + 2MC^{\frac{1}{2}}_{j_{i-1},j_i}\left(1 + 2V(x_0) + \frac{8b}{1-\lambda}\right)^{\frac{1}{2}} \\ &\le \left(\frac{1}{2}\right)^k \tilde{b}\mathbb{E}\sqrt{1 + V(x_0) + V(\mathcal{T}^i_{-a_{i-1}})} + CC^{\frac{1}{2}}_{j_{i-1},j_i} \\ &\le \left(\frac{1}{2}\right)^k \tilde{b}\sqrt{1 + 2V(x_0) + \frac{b}{1-\lambda}} + CC^{\frac{1}{2}}_{j_{i-1},j_i} \lesssim r^{a_{i-1}} + C^{\frac{1}{2}}_{j_{i-1},j_i}\end{aligned}$$

where to get the first inequality we summed-up a geometric series, in the second inequality we used (8.27) and where $r = (\frac{1}{2})^{n_0(\frac{C_L}{2})^{-1}}$.

The rest of the proof is very similar to the last part of the proof of Lemma 8.2 (using (5.10) instead of (5.8)), and is hence omitted. □

8.3.4. *Proofs of Theorems* 5.3 *and* 5.4

In order to prove Theorems 5.3 and 5.4, we need to first use Proposition 1.1 which gives conditions securing unbiasedness and finite variance of Z and then make sure that these conditions are compatible with a finite expected computing time.

Proof of Theorem 5.3. By the considerations at the end of Section 1.3, in order to get the unbiasedness and finite variance of Z it suffices to verify (1.7). Using (8.20), we have that for the stated choices of a_i and j_i, it holds

$$\begin{aligned}\sum_{i\le l} \frac{\|\Delta_i\|_2\|\Delta_l\|_2}{\mathbb{P}(N \ge i)} = \sum_{i=0}^{\infty} \frac{\|\Delta_i\|_2}{\mathbb{P}(N \ge i)} \sum_{l=i}^{\infty} \|\Delta_l\|_2 \lesssim \sum_{i=0}^{\infty} \frac{\|\Delta_i\|_2}{\mathbb{P}(N \ge i)} \frac{r^{\frac{m}{2}i}}{1 - r^{\frac{1}{2}}} \\ \lesssim \sum_{i=0}^{\infty} \frac{r^{mi}}{\mathbb{P}(N \ge i)},\end{aligned} \tag{8.31}$$

where $r < 1$ is defined in Lemma 8.2. It is hence sufficient to choose the distribution of N such that $\sum_i \frac{r^{mi}}{\mathbb{P}(N \ge i)} < \infty$. A valid choice is for example, $\mathbb{P}(N \ge i) \propto r^{(m-\epsilon)i}$ for $\epsilon > 0$ which can be arbitrarily small.

Regarding the expected computing time of Z, we have that it is equal to $\sum_{i=0}^{\infty} t_i \mathbb{P}(N \ge i)$, where t_i is the expected time to generate Δ_i. By Assumption 5.2, we have $t_i \lesssim a_i j_i^{\theta}$, hence

$$\sum_{i=0}^{\infty} t_i \mathbb{P}(N \ge i) \lesssim \sum_{i=0}^{\infty} i r^{(\frac{2\theta m}{1-2a} + m - \epsilon)i}.$$

To get that the right-hand side is finite, we need to have $\epsilon < \frac{2\theta m}{1-2a} + m$ and such a choice is possible since $a > \theta + \frac{1}{2}$. □

Proof of Theorem 5.4. The proof is almost identical to the proof of Theorem 5.3, and is hence omitted. □

Remark 8.4. In Theorem 4.5, in Section 4, we give an example of parameter choices for which the unbiasing procedure works, which is such that a_i grow logarithmically and j_i polynomially in i. A simple calculation shows that we could have made the same choices here and would have ended up with the same condition on the regularity, α, of the reference measure μ_0. The present choice implies that the random variable N has moments of all orders. On the other hand the dimensionality j_N increases exponentially in N. Thus, the comparison of both approaches depends on the concrete choices.

Acknowledgements

The majority of the work of Sergios Agapiou was completed at the University of Warwick, supported by ERC under Grant 226488 and by EPSRC under the EQUIP Project EP/K034154/1. The work of Gareth Roberts was supported by EPSRC under Grants EP/K034154/1, EP/K014463/1, and EP/D002060/1. Sebastian Vollmer was supported by EPSRC under Grants EP/N000188/1 and EP/K009850/1. The authors are thankful to the anonymous referee and the editor Eric Moulines for their useful comments which helped improve the presentation of the paper.

Supplementary Material

Supplement to "Unbiased Monte Carlo: posterior estimation for intractable/infinite-dimensional models" (DOI: 10.3150/16-BEJ911SUPP; .pdf). We provide detailed proofs, further consideration on unbiased estimators for Bayesian linear inverse problems and an elliptic inverse problem as detailed example.

References

[1] Adler, R.J. (1990). *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*. *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, 12. Hayward, CA: IMS. MR1088478

[2] Agapiou, S., Larsson, S. and Stuart, A.M. (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process. Appl.* **123** 3828–3860. MR3084161

[3] Agapiou, S., Roberts, G.O. and Vollmer, S.J. Supplement to "Unbiased Monte Carlo: Posterior estimation for intractable/infinite-dimensional models." DOI:10.3150/16-BEJ911SUPP.

[4] Agapiou, S., Stuart, A.M. and Zhang, Y.-X. (2014). Bayesian posterior contraction rates for linear severely ill-posed inverse problems. *J. Inverse Ill-Posed Probl.* **22** 297–321. MR3215928

[5] Beskos, A., Papaspiliopoulos, O., Roberts, G.O. and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 333–382. MR2278331

[6] Beskos, A., Roberts, G., Stuart, A. and Voss, J. (2008). MCMC methods for diffusion bridges. *Stoch. Dyn.* **8** 319–350. MR2444507

[7] Cotter, S.L., Dashti, M. and Stuart, A.M. (2010). Approximation of Bayesian inverse problems for PDEs. *SIAM J. Numer. Anal.* **48** 322–345. MR2608372

[8] Dashti, M. and Stuart, A.M. The Bayesian approach to inverse problems. In *Handbook of Uncertainty Quantification* (R. Ghanem, D. Higdon and H. Owhadi, eds.). Springer.

[9] Dodwell, T.J., Ketelsen, C., Scheichl, R. and Teckentrup, A.L. (2015). A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. *SIAM/ASA J. Uncertain. Quantificat.* **3** 1075–1108. MR3418231

[10] Durmus, A., Fort, G. and Moulines, É. (2016). Subgeometric rates of convergence in Wasserstein distance for Markov chains. *Ann. Inst. H. Poincaré Probab. Statist.* **52** 1799–1822.

[11] Durmus, A. and Moulines, É. (2015). Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis adjusted Langevin algorithm. *Stat. Comput.* **25** 5–19. MR3304898

[12] Forsythe, G.E. and Leibler, R.A. (1950). Matrix inversion by a Monte Carlo method. *Math. Tables Other Aids Comput.* **4** 127–129. MR0038138

[13] Giles, M.B. (2008). Multilevel Monte Carlo path simulation. *Oper. Res.* **56** 607–617. MR2436856

[14] Glynn, P.W. (1983). Randomized estimators for time integrals. Tech. rep., Mathematics Research Center, University of Wisconsin, Madison.

[15] Glynn, P.W. and Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Oper. Res.* **40** 505–520. MR1180030

[16] Gruhlke, D. (2014). Convergence of multilevel MCMC methods on path spaces Ph.D. thesis, Universitäts-und Landesbibliothek, Bonn.

[17] Hairer, M., Mattingly, J.C. and Scheutzow, M. (2011). Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations. *Probab. Theory Related Fields* **149** 223–259. MR2773030

[18] Hairer, M., Stuart, A.M. and Vollmer, S.J. (2014). Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* **24** 2455–2490. MR3262508

[19] Hoang, V.H., Schwab, C. and Stuart, A.M. (2013). Complexity analysis of accelerated MCMC methods for Bayesian inversion. *Inverse Probl.* **29** 085010, 37. MR3084684

[20] McLeish, D. (2011). A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods Appl.* **17** 301–315. MR2890424

[21] Mengersen, K.L. and Tweedie, R.L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24** 101–121. MR1389882

[22] Meyn, S. and Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge: Cambridge Univ. Press. MR2509253

[23] Peluchetti, S. and Roberts, G.O. (2012). A study on the efficiency of exact methods for diffusion simulation. In *Monte Carlo and Quasi-Monte Carlo Methods* 2010. *Springer Proc. Math. Stat.* **23** 161–187. Springer, Heidelberg. MR3173833

[24] Pinski, F.J., Simpson, G., Stuart, A.M. and Weber, H. (2015). Algorithms for Kullback–Leibler approximation of probability measures in infinite dimensions. *SIAM J. Sci. Comput.* **37** A2733–A2757. MR3424069

[25] Propp, J.G. and Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms* (*Atlanta*, *GA*, 1995) **9** 223–252. MR1611693

[26] Rhee, C.H. (2013). Unbiased estimation with biased samples Ph.D. thesis, Stanford University.

[27] Rhee, C.-H. and Glynn, P.W. (2015). Unbiased estimation with square root convergence for SDE models. *Oper. Res.* **63** 1026–1043. MR3422533

[28] Roberts, G.O. and Rosenthal, J.S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.* **1** 20–71. MR2095565

[29] Rudolf, D. (2012). Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Math.* (*Rozprawy Mat.*) **485** 1–93. MR2977521

[30] Schwab, C. and Stuart, A.M. (2012). Sparse deterministic approximation of Bayesian inverse problems. *Inverse Probl.* **28** 045003, 32. MR2903278

[31] Vihola, M. (2015). Unbiased estimators and multilevel Monte Carlo. arXiv preprint, arXiv:1512.01022.

[32] Vollmer, S.J. (2015). Dimension-independent MCMC sampling for inverse problems with non-Gaussian priors. *SIAM/ASA J. Uncertain. Quantificat.* **3** 535–561. MR3370004

[33] Wasow, W.R. (1952). A note on the inversion of matrices by random walks. *Math. Tables Other Aids Comput.* **6** 78–81. MR0055033

Received December 2014 and revised July 2016