Analysis of a Class of Multilevel Markov Chain Monte Carlo Algorithms Based on Independent Metropolis–Hastings

Juan P. Madrigal-Cianci†, Fabio Nobile‡, and Raúl Tempone§

Abstract. In this work, we present, analyze, and implement a class of multilevel Markov chain Monte Carlo (ML-MCMC) algorithms based on independent Metropolis–Hastings proposals for Bayesian inverse problems. In this context, the likelihood function involves solving a complex differential model, which is then approximated on a sequence of increasingly accurate discretizations. The key point of this algorithm is to construct highly coupled Markov chains together with the standard multilevel Monte Carlo argument to obtain a better cost-tolerance complexity than a single-level MCMC algorithm. Our method extends the ideas of Dodwell et al., [SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 1075–1108] to a wider range of proposal distributions. We present a thorough convergence analysis of the ML-MCMC method proposed, and show, in particular, that (i) under some mild conditions on the (independent) proposals and the family of posteriors, there exists a unique invariant probability measure for the coupled chains generated by our method, and (ii) that such coupled chains are uniformly ergodic. We also generalize the cost-tolerance theorem of Dodwell et al. to our wider class of ML-MCMC algorithms. Finally, we propose a self-tuning continuation-type ML-MCMC algorithm. The presented method is tested on an array of academic examples, where some of our theoretical results are numerically verified. These numerical experiments evidence how our extended ML-MCMC method is robust when targeting some pathological posteriors, for which some of the previously proposed ML-MCMC algorithms fail.

Key words. Bayesian inversion, multilevel Monte Carlo, Markov chain Monte Carlo, uncertainty quantification

MSC codes. 35R30, 62F15, 62P30, 65M32, 60J05, 60J22

DOI. 10.1137/21M1420927

1. Introduction. By now, multilevel Monte Carlo (MLMC) methods are well-established computational techniques [13] to compute expectations that arise in stochastic simulations in cases in which the stochastic model can not be simulated exactly, rather approximated, at different levels of accuracy and, as such, at different computational costs. Despite their

---

Received by the editors May 18, 2021; accepted for publication (in revised form) August 22, 2022; published electronically March 3, 2023.

Funding: This project was partially funded by the King Abdullah University of Science and Technology through grant OSR-2015-CRG4-2584-01: “Advanced Multi-Level sampling techniques for Bayesian Inverse Problems with applications to subsurface.” The first and second authors also acknowledge support from the Swiss Data Science Center (SDSC) through grant P18-09. The third author acknowledges support from the Alexander von Humboldt foundation.

Protocol Labs (juan.madrigalcianci@protocol.ai).

SB-MATH-CSQI, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland (fabio.nobile@epfl.ch).

Computer, Electrical and Mathematical Sciences and Engineering, KAUST, Thuwal, Saudi Arabia, and Alexander von Humboldt professor in Mathematics of Uncertainty Quantification, RWTH Aachen University, Aachen 52062, Germany (raul.tempone@kaust.edu.sa, tempone@uq.rwth-aachen.de).
widespread applicability, extending these MLMC ideas to MLMC Monte Carlo (ML-MCMC) methods to compute expectations with respect to (w.r.t.) a complex target distribution from which independent (whether exact or approximate) sampling is not accessible, has only recently been attempted, with only a handful of works dedicated to this task. This situation arises, for instance, in Bayesian inverse problems (BIPs) where one wishes to compute the expectation $\mathbb{E}_{\mu^\gamma}[\mathcal{D}]$ of some output quantity of interest $\mathcal{D}$ with respect to the posterior measure $\mu^\gamma$ of some parameters $\Theta$ given some indirect noise measurements $y = \mathcal{F}(\Theta) + \eta$, where $\eta$ is the additive noise and $\mathcal{F}$ is the forward operator, which may involve the solution of a differential equation. At their core, ML-MCMC methods for BIP introduce a hierarchy of discretization levels $\ell = 0, 1, \ldots, L$ of the underlying forward operator, which in turn induce a family of posterior probability measures $\mu^\gamma_{\ell}$, approximating $\mu^\gamma$ with increasing levels of accuracy as $\ell \to \infty$. Given some $\mu^\gamma$-integrable quantity of interest $\mathcal{D}$, one can approximate the expectation of $\mathcal{D}$ over $\mu^\gamma$ by the usual telescoping sum argument of MLMC, namely,

\[
\mathbb{E}_{\mu^\gamma}[\mathcal{D}] \simeq \mathbb{E}_{\mu^\gamma}[\mathcal{D}_L] = \mathbb{E}_{\mu^\gamma}[\mathcal{D}_0] + \sum_{\ell=1}^{L} \left( \mathbb{E}_{\mu^\gamma}[\mathcal{D}_\ell] - \mathbb{E}_{\mu^\gamma_{\ell-1}}[\mathcal{D}_{\ell-1}] \right)
\]

with $\Delta \mathcal{E}_\ell := \mathbb{E}_{\mu^\gamma_{\ell}}[\mathcal{D}_\ell] - \mathbb{E}_{\mu^\gamma_{\ell-1}}[\mathcal{D}_{\ell-1}]$, $\Delta \mathcal{E}_0 = \mathbb{E}_{\mu^\gamma}[\mathcal{D}_0]$, and where, for $\ell = 0, 1, \ldots, L$, $\mathcal{D}_\ell$ is a $\mu^\gamma_{\ell}$-integrable, level-$\ell$ approximation of the quantity of interest $\mathcal{D}$. This telescoping sum presents the basis for various types of multilevel techniques for BIPs. The work [18], for example, approximates the expectation (1.1) by splitting each $\Delta \mathcal{E}_\ell$ into three different terms, which are then computed using a mixture of importance-sampling and MCMC techniques. A multi-index generalization of such a method is presented in [20]. In addition, similar multilevel ideas have also been attempted in the context of multilevel sequential Monte Carlo (MLSMC) in the works [2, 21, 26].

In this work, we rather follow the approach proposed in [9], which is probably the first proposition of multilevel ideas for BIP and which consists in approximating $\mathbb{E}_{\mu^\gamma}[\mathcal{D}_L]$ by the following ergodic estimator:

\[
\mathbb{E}_{\mu^\gamma}[\mathcal{D}_L] \approx \frac{1}{N_0} \sum_{n=0}^{N_0} \mathcal{D}_0(\Theta^n_{0,0}) + \sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{n=0}^{N_\ell} \mathcal{D}_\ell(\Theta^n_{\ell,\ell}) - \mathcal{D}_{\ell-1}(\Theta^n_{\ell-1,\ell-1}),
\]

where $\{\Theta^n_{\ell,\ell}\}_{n=0}^{N_\ell}$ is a (truncated) Markov chain with invariant distribution $\mu^\gamma_{\ell}$. In this work, by ML-MCMC we refer to methods relying upon an expansion of the form (1.2). The key idea here is to couple $\{\Theta^n_{\ell-1}\}_{n=0}^{N_{\ell-1}}$, $\{\Theta^n_{\ell,\ell}\}_{n=0}^{N_\ell}$, i.e., the marginals of the truncated joint chain $\{\Theta^n_{\ell-1,\ell}\}_{n=0}^{N_\ell}$ so that they are highly correlated and the variance of the ergodic estimator $\mathbb{E}_{\mu^\gamma}[\mathcal{D}_L]$ is smaller and smaller as $\ell$ increases. By carefully choosing $N_\ell$, this method
can achieve a much better sampling complexity (in terms of cost versus tolerance) than its single level counterparts (see [9]).

Most of the existing literature on ML-MCMC has focused on the construction of these types of couplings [7, 9]. In [9], the authors use (an approximation of) the posterior distribution at the previous discretization level \( \ell - 1 \) as an independent proposal for level \( \ell \). This is practically implemented by subsampling from the samples obtained at the previous level, i.e., \( \{\Theta_{n,\ell-1}^{n-1}\}_{n=0}^{N_{\ell-1}} \). Such an idea has been recently expanded in [7], where the subsampling idea is combined with the so-called dimension independent likelihood informed (DILI) MCMC method of [8] to generate proposed samples at level 0 in their ML-MCMC algorithm. In addition, the recent work [30] improves upon the ML-MCMC ideas of [9] by rewriting them in an adaptive, delayed-acceptance Metropolis framework, as in [28].

An alternative idea to couple the chains at two consecutive discretization levels has been presented in [22]. There, instead of using independent samplers, the authors use rejection-free Markov transitions kernels, such as the Gibbs sampler, driven by the same sequence of random variables, something that is also known as common random numbers (c.f. [1, Chapter 5, section 6]), to couple the two chains.

We mention that the topic of coupling Markov chains has also received attention in recent years with the goal of constructing so-called unbiased ergodic estimators, where the bias here is associated with the finite length of the chain and not to the discretization of the underlying mathematical model. This is achieved by generating two copies of the Markov chain of interest using maximal coupling techniques [19, 27], together with the telescoping idea presented in the seminal work of [14]. We remark that these ideas also take advantage of common random numbers (as defined in the previous paragraph) and will be utilized in our context in Algorithm 3.1. These methods have also been used to construct variance reduction techniques [1], such as antithetic variables and control variates, for ergodic estimators obtained from Markov chains [36]. A combination of these unbiased estimation ideas with a multilevel construction has been recently proposed in [17].

However, investigating more theoretical aspects of ML-MCMC algorithms, such as the existence of an invariant measure for the coupled chains and the type of convergence to such measure (provided it exists), have been widely overlooked, and one of the aims of this paper is to fill this gap.

This work presents several novel contributions. First, we present an ML-MCMC algorithm where chains are coupled using independent Metropolis–Hastings-type proposals as in [9], however, allowing for a wider class of admissible proposals. In particular, we show that the subsampling approach in [9] can be replaced by a properly chosen independent Metropolis–Hastings (IMH) proposal (that is, a proposal for which the proposed state is independent of the current state of the chain), which proposes the same state to the two chains \( \{\Theta_{n,\ell-1}^{n-1}, \Theta_{n,\ell}^{n}\}_{n=0}^{N_{\ell}} \) targeting (asymptotically in \( n \)) \( \mu_{\ell-1}^{y}, \mu_{\ell}^{y} \), respectively, which is then accepted by the usual Metropolis–Hastings criterion. This ensures the coupling of the chains. Such a proposal can be, for example, the prior, a Laplace approximation, or even a kernel density approximation of the posterior at the previous level. Obviously, the choice of proposal has a direct impact on the joint invariant distribution \( \nu_{\ell} \) of the coupled chain \( \{\Theta_{n,\ell-1}^{n}, \Theta_{n,\ell}^{n}\}_{n=0}^{N_{\ell}} \) (provided it exists) and hence on the variance of the ergodic estimator \(\sum_{n}T_{n}^{n}\).
The main contribution of this work is an in-depth convergence analysis of our extended ML-MCMC method. More precisely, we provide sufficient conditions on the (marginal) level $\ell$ posterior and on the proposal probability measure $Q_{\ell}$ so that there exists a unique joint invariant probability measure for the coupled chain. Such a contribution is presented in Theorem 4.3, where it is shown that, under some mild conditions on $Q_{\ell}, \mu_{\ell}, \mu_{\ell-1}$, the ML-MCMC algorithm presented here (i) has a unique invariant probability measure for the joint chain at level $\ell$ and (ii) is uniformly ergodic. Following [40], we provide also computable, quantitative, nonasymptotic error estimators for the ergodic estimator (1.1). This allows us, on the one hand, to generalize the cost-tolerance result of [9] to our extended ML-MCMC method and, on the other hand, to propose an adaptive ML-MCMC algorithm in which the number of levels $L$ and chain lengths $N_{\ell}$ are determined on the fly, in the spirit of the continuation MLMC method of [6].

The rest of the paper is organized as follows. In section 2 we introduce some notation and standard results in the theory of Markov chains and BIPs. We present our ML-MCMC method in section 3 and then proceed to analyze its convergence in section 4. Section 5 is dedicated to the generalization to our case of the cost-tolerance analysis result of [9]. In section 6 we discuss the continuation-type algorithm and implementation details. In section 7 we numerically verify some of our results in two academic examples, and, lastly, we present some conclusions and finalizing remarks in section 8.

2. Background.

2.1. Background on Markov chains. We recall some definitions and standard results from the theory of Markov chains. Let $(X, \| \cdot \|_X)$ be a separable Banach space with Borel $\sigma$-algebra $\mathcal{B}(X)$, and denote by $\mathcal{M}(X)$ the set of probability measures on $(X, \mathcal{B}(X))$. For some $\mu \in \mathcal{M}(X)$ and any $q \in [1, \infty]$, we define the Banach spaces

$$L_q(X, \mu) := \left\{ f : X \to \mathbb{R}_{\mu} \text{ measurable s.t. } \int_X |f(\theta)|^q \mu(d\theta) < \infty \right\}$$

if $q < \infty$,

$$L_\infty(X, \mu) := \left\{ f : X \to \mathbb{R}_{\mu} \text{ measurable s.t. } \text{ess sup}_{\theta \in X} |f(\theta)| < \infty \right\}$$

if $q = \infty$,

endowed with the norms

$$\|f\|_{L_q(X, \mu)} := \left( \int_X |f(\theta)|^q \mu(d\theta) \right)^{1/q}$$

if $q < \infty$,

$$\|f\|_{L_\infty(X, \mu)} := \text{ess sup}_{\theta \in X} |f(\theta)|$$

if $q = \infty$.

In the particular case where $q = 2$, $L_2(X, \mu)$ is a Hilbert space equipped with the inner product

$$\langle f, g \rangle_\mu := \int_X f(\theta) g(\theta) \mu(d\theta), \ f, g \in L_2(X, \mu).$$

For any $q \in [1, \infty]$, we define the subspace $L_q^0(X, \mu)$ of $L_q(X, \mu)$ as

$$L_q^0(X, \mu) := \left\{ f \in L_q(X, \mu) \text{ s.t. } \int_X f(\theta) \mu(d\theta) = 0 \right\}.$$
Recall that a Markov transition kernel is a function $p : X \times \mathcal{B}(X) \to [0, 1]$ such that

1. For each $A \in \mathcal{B}(X)$, the mapping $X \ni \theta \mapsto p(\theta, A)$, is a $\mathcal{B}(X)$-measurable real-valued function.
2. For each $\theta$ in $X$, the mapping $\mathcal{B}(X) \ni A \mapsto p(\theta, A)$, is a probability measure on $(X, \mathcal{B}(X))$.

Given a Markov transition kernel $p$, we denote by $P$ its associated Markov transition operator which acts to the left on measures, $\mu \mapsto \mu P \in \mathcal{M}(X)$, and to the right on functions, $f \mapsto Pf$, measurable on $(X, \mathcal{B}(X))$, such that

$$
(\mu P)(A) = \int_X p(\theta, A) \mu(d\theta), \quad \forall A \in \mathcal{B}(X),
$$

$$
(Pf)(\theta) = \int_X f(\theta)p(\theta, d\theta), \quad \forall \theta \in X.
$$

We say that a Markov operator $P$ is $\mu$-invariant if $\mu P = \mu$. We say that a Markov operator $P$ is $\mu$-reversible if

$$
\int_B p(\theta, A) \mu(d\theta) = \int_A p(\theta, B) \mu(d\theta), \quad A, B \in \mathcal{B}(X).
$$

It is a well-known fact that if $P$ is $\mu$-reversible, then $P$ is $\mu$-invariant. The reverse is in general not true. The main idea behind using MCMC methods to sample a measure of interest $\mu$ on $(X, \mathcal{B}(X))$, is to create a Markov chain with initial state $\Theta^0 \sim \lambda^0$, for some $\lambda^0 \in \mathcal{M}(X)$, using a $\mu$-invariant Markov transition operator $P$. The Markov chain $\{\Theta^n\}_{n \in \mathbb{N}}$ is then generated by sampling $\Theta^n \sim p(\Theta^{n-1}, \cdot) \forall n \in \mathbb{N}$. Throughout this work we will use $\theta \in X$ to denote the realization of the random variable $\Theta$, with the notable exception of the data $y$ in the formulation of the BIP. Under some suitable conditions, it can be shown that $\lambda^0 P^n \to \mu$ as $n \to \infty$, where the convergence is with respect to a suitable distance between probability measures.

### 2.2. BIPs and MCMC

Let $(X, \| \cdot \|_X)$ and $(Y, \| \cdot \|_Y)$ be separable Banach spaces with associated Borel $\sigma$-algebras $\mathcal{B}(X)$, $\mathcal{B}(Y)$, and let us define the forward operator $\mathcal{F} : X \to Y$. In inverse problems, we use some data $y \in Y$, usually polluted by some random noise $\varepsilon \sim \mu_{\text{noise}}$, to determine a possible state $\theta \in X$ that may have generated the data. Assuming an additive noise model, the relationship between $\theta$ and $y$ is given by

$$
y = \mathcal{F}(\theta) + \varepsilon, \quad \varepsilon \sim \mu_{\text{noise}}.
$$

Here, $\theta$ can be a set of parameters of a possibly nonlinear partial differential equation (PDE), and the data correspond to some observable (output) quantities related to the solution of the PDE. The forward operator $\mathcal{F}$ describes then the parameter-to-outputs map, and its evaluation at any $\theta \in X$ implies the solution of a PDE. As such, it is, in general, not possible to evaluate $\mathcal{F}$ exactly, and one has rather to use numerical methods. We denote by $\mathcal{F}_L$ the numerical approximation of the forward map at an accuracy level $L$. In a Bayesian setting, we consider the parameter $\theta$ to be uncertain and model it as a random variable $\Theta$ with a
given prior probability measure $\mu_{pr}$ on $(X,\mathcal{B}(X))$. Such a prior models the knowledge we have on the uncertainty in $\Theta$, before observing the data $y$. For simplicity, we will assume that $Y = \mathbb{R}^n$ for some $n \geq 1$, and that $\mu_{\text{noise}}$ has a density $\pi_{\text{noise}}$, with respect to the Lebesgue measure, and that such a measure is strictly positive and bounded, i.e., $\exists A > 0$ such that $0 < \pi_{\text{noise}}(z) \leq A \forall z \in Y$. Furthermore, if we assume that the noise $\varepsilon$ and $\Theta$ are statistically independent (when seen as random variables on their respective spaces), then we have that $\mathbb{P}(y - \mathcal{F}_L(\Theta) \in \cdot | \Theta = \theta) = \mathbb{P}(\varepsilon \in \cdot)$; i.e., $y - \mathcal{F}_L(\Theta)$ conditioned on $\Theta = \theta$ has the same distribution as $\varepsilon$. This allows us to define the likelihood function

$$\pi_L(y|\theta) := \pi_{\text{noise}}(y - \mathcal{F}_L(\theta))$$

and write it in terms of a potential function $\Phi_L(\theta; y) : X \times Y \to [0, \infty)$:

$$\Phi_L(\theta; y) = -\log [\pi_{\text{noise}}(y - \mathcal{F}_L(\theta))/A].$$

The scaling factor $A^{-1}$ in the definition of the potential is introduced to make it nonnegative.

The function $\Phi_L(\theta; y)$ is a measure of the misfit between the recorded data $y$ and the predicted value (at accuracy level $L$) $\mathcal{F}_L(\theta)$ and often depends only on $||y - \mathcal{F}_L(\theta)||_Y$. From the Bayes theorem (see, e.g., [10, section 10.2]), it follows that the BIP consists in determining (or approximating) the conditional probability measure $\mu_L^y(\cdot) = \mathbb{P}(\Theta \in \cdot | y)$ of $\Theta$ given the observed data $y$, which we write in terms of its Radon–Nikodym derivative with respect to $\mu_{pr}$:

$$\pi_L^y(\theta) = \frac{d\mu_L^y(\theta)}{d\mu_{pr}(\theta)} = \frac{\pi_{\text{noise}}(y - \mathcal{F}_L(\theta))}{\int \pi_{\text{noise}}(y - \mathcal{F}_L(\theta))\mu_{pr}(d\theta)} = \frac{1}{Z_L} \exp (-\Phi_L(\theta; y)),$$

with $Z_L := \int_X \exp(-\Phi_L(\theta; y))\mu_{pr}(d\theta)$.

Notice that $\mu_L^y$ (resp., $\pi_L^y$) is a numerical approximation of a usually unattainable posterior measure (resp., density) $\mu^y$ (resp., $\pi^y$). The goal of Bayesian analysis is often to explore the posterior measure or compute posterior expectations of some $\mu^y$-integrable output quantity of interest $\mathcal{Q} : X \to \mathbb{R}$ by drawing samples from the posterior measure. A common method for performing such a task is to use MCMC methods, the most celebrated of which is, perhaps, the Metropolis–Hastings algorithm [16, 32], which we describe briefly. Let $Q : X \times X \to \mathbb{R}_+$ be a continuous and strictly positive function such that, for any $\theta \in X$, $\int_X Q(\theta, z)\mu_{pr}(dz) = 1$. Thus, for any fixed $\theta \in X$, $Q(\theta, \cdot)$ induces a probability measure $Q(\theta, \cdot)$ having Radon–Nikodym derivative with respect to the prior given by $\frac{dQ(\theta, \cdot)}{d\mu_{pr}}(z) = Q(\theta, z)$. Notice that, since $Q(\theta, z)$ is assumed to be strictly positive for any $\theta, z \in X$, $\mu_{pr}$ and $Q(\theta, \cdot)$ are equivalent probability measures for any $\theta \in X$. Furthermore, define the following three probability measures on $(X \times X, \mathcal{B}(X \times X))$:

$$\eta_L(d\theta, dz) := \mu_L^y(d\theta)Q(\theta, dz),$$

$$\eta_L^T(d\theta, dz) := \mu_L^y(dz)Q(z, d\theta),$$

$$\eta_{pr}(d\theta, dz) := \mu_{pr}(d\theta)\mu_{pr}(dz).$$
It is known (see, e.g., [42]) that whenever \( \eta^T \) is absolutely continuous w.r.t \( \eta_L (\eta^T_L \ll \eta_L) \), one can construct a well-defined Metropolis–Hastings algorithm targeting \( \mu^y_L \). Clearly, since both \( \pi^y_L \) and \( Q \) are strictly positive, both \( \eta_L \) and \( \eta^T_L \) are equivalent with respect to \( \eta_{pr} \), with Radon–Nikodym derivatives given by

\[
\frac{d\eta^T}{d\eta_{pr}}(\theta, z) = \frac{d\mu^y_L}{d\mu_{pr}}(\theta) \frac{d\hat{Q}(\theta, \cdot)}{d\mu_{pr}}(z) = \pi^y_L(\theta) Q(\theta, z),
\]

\[
\frac{d\eta_L}{d\eta_{pr}}(\theta, z) = \frac{d\mu^y_L}{d\mu_{pr}}(z) \frac{d\hat{Q}(\cdot, z)}{d\mu_{pr}}(\theta) = \pi^y_L(z) Q(\theta, z).
\]

As such, one can define the so-called (level L) Metropolis acceptance ratio

\[
\frac{d\eta^T}{d\eta_L}(\theta, z) = \left( \frac{d\eta^T}{d\eta_{pr}} \right)(\theta, z) = \pi^y_L(z) Q(\theta, z) \frac{\pi^y_L(z) Q(\theta, z)}{\pi^y_L(\theta) Q(\theta, z)},
\]

and, similarly, one can define the (level L) Metropolis–Hastings acceptance probability \( \alpha_L : X \times X \to [0, 1] \) by

\[
\alpha_L(\theta, z) = \min \left\{ 1, \frac{d\eta^T}{d\eta_L}(\theta, z) \right\}.
\]

The Metropolis–Hastings algorithm constructs a Markov chain \( \{\Theta^n\}_{n \in \mathbb{N}} \) by iteratively proposing at each step \( n \) a candidate state \( Z \sim Q(\theta^n, \cdot) \) given \( \Theta^n = \theta^n \), and setting \( \Theta^{n+1} = Z \) with probability \( \alpha_L(\theta^n, Z) \), and otherwise setting \( \Theta^{n+1} = \theta^n \). Iterating this procedure produces a truncated Markov chain \( \{\Theta^n\}_{n=1}^N \) with \( \Theta^n \) being asymptotically (in \( n \)) distributed as \( \mu^y_L \), provided the chain is geometrically ergodic (c.f. section 4). The procedure is depicted in Algorithm 2.1.

Once samples \( \{\Theta^n\}_{n=1}^N \), drawn approximately from \( \mu^y_L \) have been obtained by some MCMC algorithm, the posterior expectation \( E_{\mu^y} [2] \) can be first approximated by \( E_{\mu^y} [2_L] \), which in

---

**Algorithm 2.1 Metropolis–Hastings.**

1: procedure Metropolis–Hastings\((\pi^y_L, Q, N, \lambda^0)\).
2: Sample \( \Theta^0 \sim \lambda^0 \).
3: for \( n = 0, \ldots, N - 1 \) do
4: \hspace{1em} Given \( \Theta^n = \theta^n \), sample \( Z \sim Q(\theta^n, \cdot) \).
5: \hspace{1em} Set \( \Theta^{n+1} = Z \) with acceptance probability
6: \hspace{1em} \hspace{1em} \alpha_L(\theta^n, Z) = \min \left[ 1, \frac{\pi^y_L(Z) Q(Z, \theta^n)}{\pi^y_L(\theta^n) Q(\theta^n, Z)} \right].
7: \hspace{1em} end for
8: Output \( \{\Theta^n\}_{n=0}^N \).
9: end procedure

---
turn gets approximated by the following ergodic estimator

\begin{equation}
\hat{\mathcal{D}}_L := \frac{1}{N} \sum_{n=1}^{N} \mathcal{D}_L(\Theta^n),
\end{equation}

where \( \mathcal{D}_L : X \rightarrow \mathbb{R} \) is a \( \mu^\theta \)-integrable numerical approximation of \( \mathcal{D} \).

The Metropolis–Hastings algorithm induces the following Markov transition kernel \( p : X \times \mathcal{B}(X) \rightarrow [0,1] \)

\[ p(\theta, A) = \int_A \alpha_L(\theta, z)Q(\theta, z)\mu_{pr}(dz) + \mathbb{I}_{\{\theta \in A\}} \int_X (1 - \alpha_L(\theta, z))Q(\theta, z)\mu_{pr}(dz) \]

for every \( \theta \in X \) and \( A \in \mathcal{B}(X) \). Here, \( \mathbb{I}_{\{\theta \in A\}} \) is the characteristic function of the set \( A \); i.e., \( \mathbb{I}_{\{\theta \in A\}} = 1 \) if \( \theta \in A \) and \( \mathbb{I}_{\{\theta \in A\}} = 0 \) otherwise. Thus, such an algorithm can be understood as iteratively applying the Markov transition operator \( P \) associated to \( p \) to the initial probability measure \( \lambda^0 \).

### 3. ML-MCMC.

In general, the Metropolis–Hastings algorithm requires a large number of samples to converge—it is not uncommon for \( N \) to be of the order of tens of thousands. Furthermore, as it can be seen from Algorithm 2.1, every time \( \alpha_L(\theta^n, z) \) is evaluated, one needs to evaluate the posterior density at each new proposed state. In PDE-based inverse problems, where evaluating \( \pi^\theta(z) \) implies solving a possibly nonlinear and time-dependent PDE on a sufficiently fine mesh (i.e., with high accuracy), the cost associated to the Metropolis–Hastings (MH) algorithm can rapidly become prohibitive. One way to alleviate this issue is by introducing multilevel techniques. To that end, let \( \{M^\ell\}_{\ell=0}^L \) be a hierarchy of discretization parameters of the underlying mathematical model \( \mathcal{F}(\cdot) \) in (2.1), which could, for instance, represent the number of degrees of freedom used in the discretization of the underlying PDE. In what follows, we consider only geometric sequences for \( \{M^\ell\}_{\ell=0}^L \) with \( M^\ell = sM^{\ell-1} \) for some \( M_0 > 0 \) and \( s > 1 \). We denote the corresponding discretized forward models by \( \mathcal{F}^\ell(\cdot) \) and the corresponding approximate quantity of interest by \( \mathcal{D}_\ell \). We assume that the accuracy of the discretization, as well as the cost of evaluating the discretized model, increases as \( \ell \) (and hence \( M^\ell \)) increases. This hierarchy of discretizations, in turn, induces a hierarchy of posterior probability measures \( \{\mu^\ell\}_{\ell=0}^L \) approximating \( \mu^\theta \) with increasing accuracy and cost. Notice that we can write the posterior expectation \( \mathbb{E}_{\mu^\theta}[\mathcal{D}] \), approximated on the finest available discretization level \( L \), in terms of the following telescoping sum:

\[ \mathbb{E}_{\mu^\theta}[\mathcal{D}] \approx \mathbb{E}_{\mu_0}[\mathcal{D}_L] = \mathbb{E}_{\mu_0}[\mathcal{D}_0] + \sum_{\ell=1}^{L} \left( \mathbb{E}_{\mu_\ell}[\mathcal{D}_\ell] - \mathbb{E}_{\mu_{\ell-1}}[\mathcal{D}_{\ell-1}] \right) . \]

This motivates introducing the following ML-MCMC version of the ergodic estimator (2.3):

\begin{equation}
\hat{\mathcal{D}}_L\{N_\ell\}_{\ell=0}^L := \frac{1}{N_0} \sum_{n=0}^{N_0} [\mathcal{D}_0(\Theta^0_n)] + \sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{n=0}^{N_\ell} \left( \mathcal{D}_\ell(\Theta^\ell_n) - \mathcal{D}_{\ell-1}(\Theta^{\ell-1}_n) \right) .
\end{equation}

Here we have introduced a notation where, for every \( \Theta^\ell_n \), the first subindex \( \ell, \ell = 1, \ldots, L \), represents the current level in the telescoping sum (3.1), whilst the second subindex \( j = \ell - 1, \ell \),
represents the discretization level of the posterior measure associated to \( \Theta_{n,j} \), i.e., \( \Theta_{n,j} \sim \mu_j^n \) asymptotically in \( n \). Notice that the differences \( Y_{\ell,n}^\ell \) are small, in general, if the components of the tuple \( (\Theta_{n-1}^\ell, \Theta_{n,j}^\ell) \) are close to each other in a suitable sense (e.g., \( \mathbb{P}(\Theta_{n-1}^\ell \neq \Theta_{n,j}^\ell) \to 0 \) as \( \ell \to 0 \)). The key of the method is then to design a coupled Markov chain \( \{\Theta_{\ell-1}^\ell, \Theta_{\ell,j}^\ell\}_{n \in \mathbb{N}} \) for which \( \Theta_{\ell-1}^\ell, \Theta_{\ell,j}^\ell \) stay highly correlated and close to each other for every \( n \), while, at the same time, keeping the right marginal invariant distributions \( \mu_{\ell-1}^\ell, \mu_{\ell,j}^\ell \), respectively. This is necessary for the terms in (3.1) to telescope in the mean. Constructing a coupled Markov chain (with marginal target measures \( \mu_{\ell-1}^\ell, \mu_{\ell,j}^\ell \) for which \( \Theta_{\ell-1}^\ell - \Theta_{\ell,j}^\ell \to 0 \) in a suitable sense, as \( \ell \to \infty \), will in turn result in \( \mathbb{V}_n[Y] \to 0 \) as \( \ell \to \infty \), where \( \nu_{\ell} \in \mathcal{M}(X) \) is the invariant measure of the coupled Markov chain (provided it exists). Hence, by using an adequate proposal distribution and properly choosing \( L \) and \( \{N_{\ell}\}_{\ell=0} \) one can obtain a significantly better complexity than that of a single-level MCMC estimator (see [9] for a general complexity result of the ML-MCMC approach). To achieve this, following [9], we will use what we call an IMH coupling of \( \Theta_{\ell-1}^\ell, \Theta_{\ell,j}^\ell \). The main idea of such a coupling is to create two simultaneous Markov chains \( \{\Theta_{\ell-1}^\ell\}_{n \in \mathbb{N}}, \{\Theta_{\ell,j}^\ell\}_{n \in \mathbb{N}} \) at two adjacent discretization levels, using as a proposal a probability measure \( Q_\ell \) (where \( \mu_j^\ell \ll Q_\ell \) \( j = \ell - 1, \ell \)), having a (strictly positive) \( \mu_{\text{pr}} \)-density \( Q_\ell \), in such a way that \( Q_\ell \) generates proposed states \( Z \) independently of the current state of either chain, and that at every iteration, the same candidate state \( Z \) is proposed as the new state of both chains, which then accept or reject it using the standard MH accept-reject step with the same uniform random variable \( U \sim \mathcal{U}(0,1) \). This will in turn guarantee that, marginally \( \Theta_{\ell-1}^\ell, \Theta_{\ell,j}^\ell \) asymptotically in \( n \) for both \( j = \ell - 1, \ell \) (that is, the marginal chains follow the right distribution), and that \( \{\Theta_{\ell,j}^\ell\}_{n \in \mathbb{N}}, \{\Theta_{\ell,j}^\ell\}_{n \in \mathbb{N}} \) are highly correlated provided the acceptance rate is sufficiently high. A depiction of one step of such a coupling procedure is shown in Algorithm 3.1. The full ML-MCMC procedure is presented in Algorithm 3.2. Notice that at each level \( \ell = 1, 2, \ldots, L \), the coupled, truncated, chains \( \{\Theta_{\ell,j}^\ell\}_{n \in \mathbb{N}} \) generated by Algorithm 3.2 start from the same state; i.e., we set \( \Theta_{0,j}^0 \) for the same state; i.e., we set \( \Theta_{0,j}^0 = \Theta_{0,j}^0 \).

**Algorithm 3.1 One-step IMH coupling.**

1: procedure IMH-Coupling(\{\pi_{\ell-1}^\ell, \pi_{\ell,j}^\ell\}, \{\Theta_{\ell,j}^\ell\}_{n \in \mathbb{N}}, Q_\ell)
2: Sample \( Z \sim Q_\ell \).
3: Sample \( U \sim \mathcal{U}(0,1) \).
4: for \( j = \ell - 1, \ell \) do
5: \hspace{1cm} Given \( \Theta_{\ell,j} = \theta_{\ell,j}^n \), set \( \Theta_{\ell,j}^{n+1} = Z \) if \( U < \alpha_j(\theta_{\ell,j}, Z) \), where
6: \hspace{1cm} \[ \alpha_j(\theta_{\ell,j}, Z) := \min \left[ 1, \frac{\pi_{\ell}^y(Z)Q_\ell(\theta_{\ell,j}^n)}{\pi_{\ell,j}^y(\theta_{\ell,j}^n)Q_\ell(Z)} \right] \].
7: end for
8: Output \( \{\Theta_{\ell,j}^{n+1}\} \).
9: end procedure

Copyright © by SIAM and ASA. Unauthorized reproduction of this article is prohibited.
Algorithm 3.2 ML-MCMC.

1: procedure ML-MCMC(\{\pi^y_L\}_{L=0}, Q, \{N_0\}_{L=0}, \lambda^0)
2: if \ell = 0 then
3: \{\Theta^n_{0,0}\}_{n=0}^{N_0} = \text{Metropolis–Hastings} (\pi^y_0, Q, N_0, \lambda^0)
4: Set \chi_{0,0} = \{\Theta^n_{0,0}\}_{n=0}^{N_0}.
5: end if
6: for \ell = 1, \ldots, L do
7: “Construct” \(Q_\ell\).
8: Sample \(\Theta^n_{0,\ell-1}\) from, e.g., \(\chi_{\ell-1,\ell-1}\), and set \(\Theta^0_{0,\ell-1} = \Theta^0_{0,\ell-1}\)
9: for \(n = 0, \ldots, N_\ell - 1\) do
10: \# Create a coupled chain using IMH coupling
11: \{\Theta^n_{\ell-1,\ell}, \Theta^n_{\ell+1}\} = \text{IMH Coupling}(\{\pi^y_{\ell-1,\ell}, \pi^y_{\ell,\ell}\}, \{\Theta^n_{0,\ell-1}, \Theta^n_{0,\ell}\}, Q_\ell)
12: end for
13: Set \(\chi_{\ell,j} = \{\Theta^n_{\ell,j}\}_{n=0}^{N_\ell}, j = \ell - 1, \ell\).
14: end for
15: Output \(\chi_{0,0} \cup \{\chi_{\ell-1,\ell}, \chi_{\ell,\ell}\}_{\ell=1}^L \) and \(\mathcal{S}_{L,L}^{\{N_0\}_{L=0}^\ell}\).
16: end procedure

Algorithm 3.1 is, effectively, a type of independent sampler Metropolis [1] on the marginal chains. As such, the sampling efficiency of such an algorithm will critically depend on how well the proposal \(Q_\ell\) approximates \(\mu^y_\ell\) and \(\mu^y_{\ell-1}\); choosing a proposal \(Q_\ell\) that closely resembles \(\mu^y_\ell\) or \(\mu^y_{\ell-1}\) will reduce the amount of rejection steps, hence enhancing the mixing of the chains (see [1, 4] for a more in-depth discussion on this). In principle, \(Q_\ell\) can be chosen to be, e.g., the prior, an empirical version of the posterior based on the samples \(\{\Theta^n_{\ell-1,\ell-1}\}_{n=0}^{N_\ell}\) collected at the previous level, as originally proposed in [9], or any reasonable approximation of \(\mu^y_\ell\) such as, e.g., a Laplace approximation or a kernel density estimator (KDE), again based on the samples \(\{\Theta^n_{\ell-1,\ell-1}\}_{n=0}^{N_\ell}\) collected at the previous level. We remark that, in many applications of interest (e.g., MCMC for an infinite-dimensional BIP), one can construct reasonably accurate Laplace approximations of the posteriors under some mild assumptions on the problem setting. On a parallel, ongoing work, we are focusing on implementing our ML-MCMC algorithm using (potentially infinite-dimensional) Laplace approximations of the posterior. From an implementation point of view, such approximations can be efficiently constructed using software libraries such as hIPPYlib [43].

Notice that each step of Algorithm 3.1, given \(Z = z\) and \((\Theta^n_{\ell-1,\ell}, \Theta^n_{\ell,\ell}) = (\theta^n_{\ell-1,\ell}, \theta^n_{\ell,\ell})\), produces one out of four possible configurations \(S_1, S_2, S_3, S_4\), namely,

\[
\begin{align*}
S_1 & : (\Theta^n_{\ell-1,\ell}, \Theta^n_{\ell,\ell}) = (z, z) \quad \text{(both chains accept the proposed state)}, \\
S_2 & : (\Theta^n_{\ell-1,\ell}, \Theta^n_{\ell,\ell}) = (z, \theta^n_{\ell,\ell}) \quad \text{(chain at level } \ell - 1 \text{ accepts and chain at level } \ell \text{ rejects)}, \\
S_3 & : (\Theta^n_{\ell-1,\ell}, \Theta^n_{\ell,\ell}) = (\theta^n_{\ell-1,\ell}, z) \quad \text{(chain at level } \ell - 1 \text{ rejects and chain at level } \ell \text{ accepts)}, \\
S_4 & : (\Theta^n_{\ell-1,\ell}, \Theta^n_{\ell,\ell}) = (\theta^n_{\ell-1,\ell}, \theta^n_{\ell,\ell}) \quad \text{(both chains reject the proposed state)}. 
\end{align*}
\]
This is illustrated in Figure 1. More formally, set $\mathbb{X}^2 \ni \theta^n_\ell := (\theta^n_{\ell,\ell-1}, \theta^n_{\ell,\ell})$. Then, for any $A \in \mathcal{B}(\mathbb{X}^2)$, Algorithm 3.1 induces the multilevel Markov transition kernel $p_\ell : \mathbb{X}^2 \times \mathcal{B}(\mathbb{X}^2) \mapsto [0,1]$ given by

\begin{align}
(3.2) \quad p_\ell (\theta^n_\ell, A) := & \int_{\mathbb{X}} \min\{\alpha_{\ell-1}(\theta^n_{\ell,\ell-1}, z), \alpha_\ell(\theta^n_{\ell,\ell}, z)\} Q_\ell(z) \mathbb{I}_{\{(z,z) \in A\}} \mu_{\text{pr}}(dz) \\
& + \int_{\mathbb{X}} (\alpha_{\ell-1}(\theta^n_{\ell,\ell-1}, z) - \alpha_\ell(\theta^n_{\ell,\ell}, z))^+ Q_\ell(z) \mathbb{I}_{\{(z,\theta^n_{\ell,\ell}) \in A\}} \mu_{\text{pr}}(dz) \\
& + \int_{\mathbb{X}} (\alpha_\ell(\theta^n_{\ell,\ell}, z) - \alpha_{\ell-1}(\theta^n_{\ell,\ell-1}, z))^+ Q_\ell(z) \mathbb{I}_{\{\theta^n_{\ell,\ell-1}, z \in A\}} \mu_{\text{pr}}(dz) \\
& + \mathbb{I}_{\{(\theta^n_{\ell,\ell-1}, \theta^n_{\ell,\ell}) \in A\}} \left( 1 - \int_{\mathbb{X}} \max\{\alpha_{\ell-1}(\theta^n_{\ell,\ell-1}, z), \alpha_\ell(\theta^n_{\ell,\ell}, z)\} Q_\ell(z) \mu_{\text{pr}}(dz) \right),
\end{align}

where $(x)^+ := \frac{x+|x|}{2}$, $x \in \mathbb{R}$. Notice that each line on the right-hand side of (3.2) corresponds to the transition kernel proposing to move from the state $\theta^n_\ell$ to one of the four possible configurations $S_i$, $i = 1, 2, 3, 4$. Although it is clear that $p_\ell$ targets the right marginals, the properties related to the convergence of the chain generated by $p_\ell$, such as irreducibility, existence of an invariant (joint) measure $\nu_\ell$, or geometric ergodicity, are not obvious. We will investigate these convergence properties on the following section.

4. Convergence analysis of the ML-MCMC algorithm. We now proceed to analyze the convergence of the levelwise coupled chains generated by Algorithm 3.1. The main result of this section is stated in Theorem 4.3. Loosely speaking, this theorem (i) gives conditions for the existence and uniqueness of a joint invariant measure of the multilevel Markov transition kernel (3.2), and (ii) indicates that such a kernel generates a uniformly ergodic chain under certain conditions (i.e., a chain that converges exponentially fast to its invariant distribution independently of the initial state of the chain, c.f. term $V(\theta^0)$ in (4.3)).
Notice that, at each level, $\ell$, Algorithm 3.1 creates two coupled chains using the same proposal $Q_{\ell}$. This in turn induces two Markov transition kernels, each generating a marginal chain. We formalize this in the following definition.

Definition 4.1 (marginal kernel). For a given level $\ell$, $\ell = 1, 2, \ldots, L$ and proposal $Q_{\ell}$, we define the $\mu^y_j$-invariant marginal Markov transition kernel $p_{\ell,j} : X \times \mathcal{B}(X) \rightarrow \{0, 1\}$, with $j = \ell - 1, \ell$, as

\[
p_{\ell,j}(\theta_{\ell,j}, A) := \int_A \alpha_j(\theta_{\ell,j}, z)Q_{\ell}(z)\mu_{\text{pr}}(dz) + \mathbb{I}_{\{\theta_{\ell,j} \in A\}} \int_X (1 - \alpha_j(\theta_{\ell,j}, z))Q_{\ell}(z)\mu_{\text{pr}}(dz)
\]

for any $\theta_{\ell,j} \in X$, and $A \in \mathcal{B}(X)$. Similarly, we denote by $P_{\ell,j}$ its corresponding marginal Markov transition operator.

Clearly, the marginal chain $\{\Theta^y_{\ell,j}\}_{n \in \mathbb{N}}, \ j = \ell - 1, \ell$ generated by (4.1) is indeed a Markov chain. Furthermore, notice that, by construction, $P_{\ell,j}$ is $\mu^y_j$-invariant, i.e., $\mu^y_j P_{\ell,j} = \mu^y_j$.

We make the following assumptions on the proposal and the (marginal) posterior densities.

Assumption A (assumptions on proposal and posterior densities). The following conditions hold for all $\ell = 1, \ldots, L$:

A.1. There exists a positive constant $c \in (0, 1)$, independent of $\ell$, such that

$$\text{ess inf}_{z \in X} \left\{Q_{\ell}(z)/\pi^y_j(z)\right\} \geq c > 0, \quad j = \ell - 1, \ell.$$

A.2. The proposal density $Q_{\ell}$ is continuous. Furthermore, for any fixed $y \in Y$, the potential function $\Phi_{\ell}(:; y) : X \rightarrow \mathbb{R}_+$ is continuous and strictly positive.

A.3. There exist positive constants $r > 1$, and $C_r$, independent of $\ell$, such that $\int_X Q_{\ell}(\theta)\mu_{\text{pr}}(d\theta) \leq C_r$, for any $\ell$.

Assumption A.1 implies that the tails of the proposal $Q_{\ell}$ must decay more slowly than those of $\mu^y_{\ell}$, $\mu^y_{\ell-1}$ at infinity, (i.e., $Q_{\ell}$ has heavier tails than $\mu^y_j$, $j = \ell - 1, \ell$). It is the opinion of the authors that Assumption A.1 is a restrictive assumption; however, it is a crucial one for the convergence of the coupled chains generated by the IMH algorithm. Assumption A.2 requires the potential to be strictly positive in $X$ (for some fixed $y \in Y$). In our setting, this assumption is easily satisfied by a suitable rescaling as in (2.2). This assumption will be used in the next section (c.f. Lemmata 5.6 and 5.9). Lastly, Assumption A.3 is an integrability condition on $Q_{\ell}$ with respect to the prior and will also be used in the next section.

4.1. Convergence of the levelwise coupled chain. In most MCMC methods, one typically designs a Markov chain with a given invariant probability measure, which automatically ensures the existence of (at least) one invariant probability measure. However, this is not the case for ML-MCMC algorithms (including the one presented here), and as such, we now proceed to demonstrate that such an invariant measure uniquely exists. We begin this section by recalling some definitions and standard results from the theory of Markov chains.
We say that a Markov kernel $p : X \times B(X) \to [0, 1]$ is $\psi$-irreducible if there exists a nonzero, $\sigma$-finite measure $\psi$ on $(X, B(X))$ such that for all measurable sets $A \in B(X)$ with $\psi(A) > 0$, and for all $\theta \in X$, there exists a positive integer $n$, possibly depending on $\theta$ and $A$, such that $p^n(\theta, A) > 0$. We say that a set $S \in B(X)$ is $\nu_m$-small if there exists $m \in \mathbb{N}$ and a nontrivial, positive measure $\nu$ on $(X, B(X))$ such that

$$p^n(\theta, A) \geq \nu(A) \quad \forall \theta \in S, \ A \in B(X).$$

We call a set $S \in B(X)$ small if it is $\nu_m$-small with $m = 1$. Given a set $A \in B(X)$ and a Markov chain $\{\Theta^n\}_{n \in \mathbb{N}}$, we define the occupation time of $\{\Theta^n\}_{n \in \mathbb{N}}$ in $A$ as $T_A(\{\Theta^n\}_{n \in \mathbb{N}}) := \sum_{n \in \mathbb{N}} \{\Theta^n \in A\}$. Notice that $T_A(\{\Theta^n\}_{n \in \mathbb{N}})$ is a random variable. Furthermore, we say that a set $A$ is recurrent if $\mathbb{E}[T_A(\{\Theta^n\}_{n \in \mathbb{N}})] = +\infty$. We say that a Markov chain is Harris recurrent if it is $\psi$-irreducible and every set $A \in B(X)$ with $\psi(A) > 0$ is recurrent. Similarly, we say that a Markov operator $P : L_r(X, \mu) \to L_r(X, \mu)$, $r \in [1, \infty]$, is Harris recurrent if it induces a Harris recurrent chain. Given a $\mu$-invariant and irreducible Markov operator $P$ and a probability measure $\mu^0 \in \mathcal{M}_r(X, \mu)$, we say that the Markov chain $\{\Theta^n\}_{n \in \mathbb{N}}$ generated by $P$ with $\Theta^0 \sim \mu^0$ is geometrically ergodic if there exists $\rho \in (0, 1)$ and a function $V : X \to \mathbb{R}_+$, such that, for any $\mu$-integrable function $f$ it holds that

$$\sup_{\|f\|_{L_\infty(X, \mu)} \leq V} \left| \int_X f(\theta)p^n(\theta, d\theta') - \int_X f(\theta)\mu(d\theta) \right| \leq V(\theta^0)\rho^n \quad \forall \theta^0 \in X \ \forall n \geq 1.$$

We say that a chain is uniformly ergodic if the previous bound can be obtained independently of the initial state of the chain; i.e., if there exists a constant $V_{\text{max}} < +\infty$ such that $V(\theta) \leq V_{\text{max}} \ \forall \theta \in X$. The following theorem is a collection of standard results in the theory of Markov chains (see, e.g., [33]).

**Theorem 4.2.** Let $P$ be a Markov operator satisfying a minorization condition of the form (4.2) with $S = X$, $n = 1$, and some measure $\nu$. Then, the Markov chain generated by $P$ is Harris recurrent, is $\nu$-irreducible, has a unique invariant probability measure $\mu$, and converges to it with uniform ergodicity, with a rate given by $\rho = 1 - \nu(X)$; i.e., there exists a bounded function $V : X \to [1, V_{\text{max}}]$ with $V_{\text{max}} < +\infty$ such that

$$\sup_{\|f\|_{L_\infty(X, \mu)} \leq V} \left| \int_X f(\theta)p^n(\theta^0, d\theta') - \int_X f(\theta)\mu(d\theta) \right| \leq V_{\text{max}}(1 - \nu(X))^n \ \forall \theta^0 \in X, \ n \geq 1,$$  

with $V_{\text{max}}$ independent of $\theta^0$.

**Proof.** Since $X$ is small, it follows from the minorization condition (4.2) that $p(\theta, A) \geq \nu(A)$ for any $\theta \in X$ and $A \in B(X)$ with $\nu(A) > 0$. Once $\nu$-irreducibility has been established, Harris recurrence follows from [33, Theorem 8.3.6], and the existence of a unique invariant measure $\mu$ follows from [33, Theorem 10.0.1]. From this, it follows from [33, Theorem 16.0.2] that the Markov chain converges to $\mu$ with uniform ergodicity and rate $\rho = 1 - \nu(X)$.

The main result of this subsection is given below.

**Theorem 4.3** (uniform ergodicity of the coupled chain). Suppose that Assumption A holds. Then, for any level $\ell = 0, 1, 2, \ldots, L$, there exists a unique invariant probability measure $\nu_\ell$ on

Copyright © by SIAM and ASA. Unauthorized reproduction of this article is prohibited.
\((X^2, \mathcal{B}(X^2))\) for the Markov transition operator \(P_\ell\). Furthermore, the Markov chain induced by such an operator is uniformly ergodic, i.e.,

\[
\sup_{\|f\|_{L_\infty(X^2, \nu_\ell)} \leq 1} \left| \int_{X^2} f(\theta_\ell') p_\ell^n(\theta_\ell, d\theta_\ell') - \int_{X^2} f(\theta_\ell) \nu_\ell(d\theta_\ell) \right| \leq 2(1 - \rho_\ell)^n \quad \forall \theta_\ell \in X^2, n \geq 1,
\]

with \(\rho_\ell := c \int_X \min\{\pi_\ell^y(z), \pi_\ell^y(1-z)\} \mu_\ell(dz)\), and \(c \in (0, 1)\) as in Assumption A.

**Proof.** We begin by showing that the whole space \(X^2\) is a small set. Indeed, notice that for any \((\theta_\ell, \ell-1, \theta_\ell, \ell) = \theta_\ell \in X^2\) and \(A \in \mathcal{B}(X^2)\), it follows from (3.2) that

\[
p_\ell(\theta_\ell, A) \geq \int_X \min\{\alpha_{\ell-1}(\theta_\ell, \ell-1, z), \alpha_{\ell}(\theta_\ell, \ell, z)\} Q_\ell(z) \mathbb{1}_{\{(z,z) \in A\}} \mu_\ell(dz)
\]

\[
= \int_X \min\left\{1, \frac{\pi_{\ell-1}^y(z)}{Q_\ell(z)} \pi_\ell^y(\theta_\ell, \ell-1), \frac{\pi_\ell^y(z) Q_\ell(\theta_\ell, \ell)}{Q_\ell(z)} \pi_\ell^y(\theta_\ell, \ell)\right\} Q_\ell(z) \mathbb{1}_{\{(z,z) \in A\}} \mu_\ell(dz)
\]

\[
\geq \int_X \min\left\{1, \frac{\pi_{\ell-1}^y(z)}{Q_\ell(z)}, \frac{\pi_\ell^y(z) c}{Q_\ell(z)}\right\} Q_\ell(z) \mathbb{1}_{\{(z,z) \in A\}} \mu_\ell(dz) \quad \text{(By Assumption A)}
\]

\[
= c \int_X \min\left\{\pi_{\ell-1}^y(z), \pi_\ell^y(z)\right\} Q_\ell(z) \mathbb{1}_{\{(z,z) \in A\}} \mu_\ell(dz)
\]

where we have set

\[
\tilde{\nu}_\ell(A) := \int_X \min\left\{\pi_\ell^y(z), \pi_{\ell-1}^y(z)\right\} \mathbb{1}_{\{(z,z) \in A\}} \mu_\ell(dz)
\]

Notice that \(\tilde{\nu}_\ell\) defines then a measure on \(X^2\). Thus, since such a minorization condition holds for the whole space, \(X^2\) is a small set, and the chain is \(\tilde{\nu}_\ell\)-irreducible. Setting \(\rho_\ell := c \tilde{\nu}_\ell(X^2)\), it then follows from Theorem 4.2 that the Markov chain generated by \(P_\ell\) is Harris recurrent, it admits a unique invariant probability measure \(\nu_\ell\), and that the chain is uniformly ergodic with

\[
\sup_{\|f\|_{L_\infty(X^2, \nu_\ell)} \leq 1} \left| \int_{X^2} f(\theta_\ell') p_\ell^n(\theta_\ell, d\theta_\ell') - \int_{X^2} f(\theta_\ell) \nu_\ell(d\theta_\ell) \right| \leq 2(1 - \rho_\ell)^n \quad \forall \theta_\ell \in X^2, n \geq 1.
\]

We have demonstrated that the joint chain generated by the multilevel algorithm with independent proposals (i) has an invariant measure and (ii) is uniformly ergodic. Moreover, we will show in section 5 that \(\rho_\ell \to c\) as \(\ell \to \infty\) under some mild assumptions (c.f. Assumption C).

Notice that the previous theorem is closely related to the following standard result in the theory of Markov chains (see, e.g., [31]), and which we recall here for convenience.

**Theorem 4.4 (uniform ergodicity of IMH).** For any \(\ell = 1, 2, \ldots, \ell-1, \ell\), let \(p_{\ell,j} : X \times \mathcal{B}(X) \to [0, 1]\) denote the \(\mu_\ell^y\)-reversible Markov transition kernel associated with an IMH algorithm with proposal \(Q_\ell\). If \(Q_\ell\) and \(\mu_\ell^y\) are such that \(\text{ess inf}_{z \in X}\{Q_\ell(z)/\pi_\ell^y(z)\} > 0\), then \(p_{\ell,j}\) is uniformly ergodic. Conversely, if \(\text{ess inf}_{z \in X}\{Q_\ell(z)/\pi_\ell^y(z)\} = 0\), then \(p_{\ell,j}\) fails to be geometrically ergodic.

**Proof.** See [31, Theorem 2.1].
Thus, from Theorem 4.4, Assumption A.1 also implies uniform ergodicity of the marginal chains of the ML-MCMC algorithm. We remark, however, that such a result cannot directly be used instead of our Theorem 4.3, since it requires a priori the existence of an invariant probability measure for the chain.

The choice of \( Q_\ell \) is delicate for the ML-MCMC algorithm to work. For instance, consider the case \( L = 1 \), \( \mu_0^y = N(1,1) \) and \( \mu_1^y = N(\frac{1}{2},1) \). What might initially appear to be a good proposal for the coupled chain at level \((\ell - 1, \ell) = (0,1)\) is to take \( Q_1 = \mu_0^y \), i.e., the (exact) posterior at the previous level. (Notice that in practical applications, exact sampling from \( \mu_0^y \) will not be possible.) However, this proposal choice does not lead to a geometrically ergodic chain given Theorem 4.4, because \( Q_1(z)/\pi_1^y(z) \) has essential supremum 0. The idea of proposing from the previous level is somehow what is advocated in [9], which could work only if \( \exists c_1, c_2 \in \mathbb{R}_+ \) such that \( c_1 \leq \pi_{\ell-1}^y(z)/\pi_\ell^y(z) \leq c_2 \) \( \forall z \in X \) and \( \forall \ell \).

Notice that, by construction, the ML-MCMC Algorithm 3.2 starts from a measure \( \hat{\lambda}^0(A) := \lambda^0(A_{\Delta}) \), \( \lambda^0 \ll \mu_{pr} \), where, for any set \( A \in \mathcal{B}(X^2) \), we define \( A_{\Delta} := \{ z \in X : (z, z) \in A \} \). We now show that, for any level \( \ell = 1, 2, \ldots, L \), \( \hat{\lambda}^0 \ll \nu_{\ell} \). Before proceeding to do so, we recall a result regarding the convergence of probability measures on general Banach spaces. By definition, the Markov transition kernel \( p(\cdot, \cdot) \) induces a probability measure \( p(\theta, \cdot) \) on \((X, \mathcal{B}(X))\) for any \( \theta \in X \). Such a measure \( p(\theta, \cdot) \) is tight if for any \( \epsilon > 0 \), there exists a compact set \( K_\epsilon \in \mathcal{B}(X) \) such that \( p(\theta, K_\epsilon) > 1 - \epsilon \). A measure \( p(\theta, \cdot) \) is tight if and only if \( p(\theta, A) \) is the suprema of \( p(\theta, K) \) over the compact subsets \( K \) of \( A \) [3]. In addition, we present the following result.

**Theorem 4.5.** If \( X \) is a separable Banach space, then, every probability measure on \((X, \mathcal{B}(X))\) is tight.

**Proof.** See [3, Theorem 1.3].

We can now show that the starting measure \( \hat{\lambda}^0 \) is absolutely continuous with respect to \( \nu_{\ell} \).

**Theorem 4.6 (absolute continuity of initial measure).** Suppose Assumption A holds. Then, for any level \( \ell = 1, 2, \ldots, L \), \( \hat{\lambda}^0 \ll \nu_{\ell} \).

**Proof.** Let \( A \in \mathcal{B}(X^2) \) be a compact set such that \( \nu_{\ell}(A) = 0 \) (the case of noncompact sets will be shown later). Furthermore, from the tightness of \( \nu_{\ell} \), we have that, given some \( \epsilon > 0 \), there exists a compact \( K_\epsilon \in \mathcal{B}(X^2) \) such that \( \nu_{\ell}(K_\epsilon) \geq 1 - \epsilon \). We then have that

\[
0 = \nu_{\ell}(A) = \int_{X^2} p_{\ell}(\theta_{\ell}, A) \nu_{\ell}(d\theta_{\ell}) \\
\geq \int_{X^2} \int_{A_{\Delta}} \min_j \left\{ \min \left\{ Q_{\ell}(z), \frac{\pi_j^y(z)Q_{\ell}(\theta_{\ell,j})}{\pi_j^y(\theta_{\ell,j})} \right\} \right\} \mu_{pr}(dz) \nu_{\ell}(d\theta_{\ell}) \\
\geq \int_{K_\epsilon} \int_{A_{\Delta}} \min_j \left\{ \min \left\{ Q_{\ell}(z), \frac{\pi_j^y(z)Q_{\ell}(\theta_{\ell,j})}{\pi_j^y(\theta_{\ell,j})} \right\} \right\} \mu_{pr}(dz) \nu_{\ell}(d\theta_{\ell}).
\]

(4.4)

By Assumption A and the compactness of \( K_\epsilon \) and \( A \), we have that there exists a \( c' > 0 \) such that \( c' \leq \min_j \left\{ \min \left\{ Q_{\ell}(z), \frac{\pi_j^y(z)Q_{\ell}(\theta_{\ell,j})}{\pi_j^y(\theta_{\ell,j})} \right\} \right\} \forall \theta_{\ell} \in K_\epsilon \forall z \in A_{\Delta} \). Then, we get that

\[
(4.4) \geq c'(1 - \epsilon) \mu_{pr}(A_{\Delta}).
\]
which implies that $\mu_{pr}(A_{\Delta}) = 0$. Moreover, since $\lambda^0 \ll \mu_{pr}$, we have $\tilde{\lambda}^0(A) = \lambda^0(A_{\Delta}) = 0$, and as such, $\tilde{\lambda}^0 \ll \nu_{c}$. Now, suppose $A$ is not compact. Since $\lambda^0$ is a tight probability measure (c.f. Theorem 4.5), then it follows that

$$\tilde{\lambda}^0(A) = \sup_{K \subseteq A} \tilde{\lambda}^0(K) = 0,$$

and we can conclude as in the previous case. \hfill \blacksquare

### 4.2. Nonasymptotic bounds on the levelwise ergodic estimator.

Recall that if the Markov operator $P$ defines a map $P : L_q(X, \mu) \to L_q(X, \mu)$, for some $q \in [1, \infty]$, we can define the norm of $P$ as

$$\|P\|_{L_q(X, \mu)} := \sup_{\|f\|_{L_q(X, \mu)} = 1} \|Pf\|_{L_q(X, \mu)}.$$

It is well known that $P$ is always a weak contraction in $L_q(X, \mu)$, i.e., $\|P\|_{L_q(X, \mu)} \leq 1$. We define the $L_q(X, \mu)$-spectral gap of $P : L_q(X, \mu) \to L_q(X, \mu)$ as

$$\gamma_q[P] := 1 - \|P\|_{L_q(X, \mu)}.$$

Whenever $\gamma_q[P] > 0$, it can be shown that $\nu^0 P^n$ converges to $\mu$ for any $\nu^0 \in \mathcal{M}(X)$ in the $L_q'$ distance between probability measures, where $q^{-1} + q'^{-1} = 1$ (see, e.g., [25, 40]).

We now turn to analyze the levelwise contribution to the ML-MCMC ergodic estimator (3.1), which we write hereafter in more general terms, including a burn-in phase.

For $\ell = 1, 2, \ldots, L$, let $\mathcal{D}_{\ell} - \mathcal{D}_{\ell-1} = : Y_{\ell} : X^2 \to \mathbb{R}$ be a $\nu_{\ell}$ square-integrable function, and $\nu^0$ be a probability measure on $(X^2, \mathcal{B}(X^2))$ such that $\nu^0 \ll \nu_{\ell}$. We remark that $\nu_{\ell}$ is the joint invariant measure induced by $P_{\ell}$. In addition, for any $\nu_{\ell}$-integrable function $h$, we write $\nu_{\ell}(h) := \mathbb{E}_{\nu_{\ell}}[h]$, and denote by $\mathbb{E}_{\nu_{\ell}}[P_{\ell}[H]]$ the expectation of a quantity $H$ which depends on a Markov chain $\{\Theta_{n,\ell}\}_{n \geq 0}$ generated by $P_{\ell}$, starting from an initial probability measure $\nu^0$.

Consider the following ergodic estimator:

$$\hat{Y}_{\ell,N_{\ell},n_{\ell},\ell} := \frac{1}{N_{\ell}} \sum_{n = 1}^{N_{\ell}} Y_{\ell}(\Theta_{n,\ell}^{n+n_{\ell},\ell}), \quad \Theta_{\ell}^{n} \sim P_{\ell}(\Theta_{n,\ell}^{n-1}, \cdot),$$

where $n_{\ell}, \ell \in \mathbb{N}$ is the usual burn-in period. The aim of this section is to provide error bounds on the nonasymptotic statistical mean square error (MSE) of (4.5), namely,

$$\text{MSE}(\hat{Y}_{\ell,N_{\ell},n_{\ell},\ell}; \nu^0) := \mathbb{E}_{\nu_{\ell}}[P_{\ell} \left( \hat{Y}_{\ell,N_{\ell},n_{\ell},\ell} - \nu_{\ell}(Y_{\ell}) \right)^2].$$

In particular, we aim at obtaining a bound of the form

$$\text{MSE}(\hat{Y}_{\ell,N_{\ell},n_{\ell},\ell}; \nu^0) \leq C_{\text{mse},\ell} \frac{\mathbb{V}_{\nu_{\ell}}[Y_{\ell}]}{N_{\ell}},$$

with $\mathbb{V}_{\nu_{\ell}}[Y_{\ell}]$ understood as $\mathbb{V}_{\Theta_{\ell} \sim \nu_{\ell}}[Y_{\ell}(\Theta_{\ell})]$, for some level-dependent, positive constant $C_{\text{mse},\ell}$. Such a bound is presented in Theorem 4.7, the main result of this subsection. We emphasize...
that for the purposes of this work, we will require bounds on the MSE of the estimator $\hat{Y}_\ell$ to be given as in (4.6), i.e., in terms of the variance of $Y_\ell$ with respect to $\nu_\ell$ (equivalently, in terms of the $L_2(X^2, \nu_\ell)$-norm). As it will become evident in section 5, this is done in order to satisfy an assumption required by a result bounding the computational complexity of our ML-MCMC algorithm (c.f. Result T3 in Theorem 5.2), as it will become more evident in section 5. While we are aware of several other results presenting nonasymptotic bounds on the MSE, such as the one in [23] as well as [40, Theorems 3.34 and 3.41], we remark that these results do not fit our setting, either because they rely upon reversibility (which does not hold for our case) [40, Theorem 3.34], they rely upon assumptions that are impractical to show in our case [40, Theorem 3.41], or because they present bounds in terms of norms that do not fit our purposes [23] and [40, Theorem 3.41]. We will discuss this in further detail towards the end of this section, in section 4.2.1. Instead, inspired by the error analysis of [40] and the so-called pseudo-spectral approach of [35], we construct a bound of the form (4.6) for general (i.e., not necessarily multilevel) nonreversible, discrete-time Markov chains. To the best of the authors’ knowledge, this result is new.

For any $q, q' \in [1, \infty]$ with $\frac{1}{q} + \frac{1}{q'} = 1$, we denote the adjoint operator of $P_\ell : L_q(X^2, \nu_\ell) \to L_q(X^2, \nu_\ell)$ by $P^*_\ell : L_{q'}(X^2, \nu_\ell) \to L_{q'}(X^2, \nu_\ell)$. It can be shown [40] that a Markov operator acting on $L_2(X^2, \nu_\ell)$ is self-adjoint if and only if it is $\nu_\ell$-reversible. Notice that, even though $P_\ell$ is not reversible, the multiplicative reversibilization $R_{\ell,k} = (P^*_\ell)^k P_\ell^k$ is for any $k \geq 1$.

Moreover, we define the so-called pseudo-spectral gap of the Markov operator $P_\ell : L_2(X^2, \nu_\ell) \to L_2(X^2, \nu_\ell)$ as

\begin{equation}
\gamma_{ps}[P_\ell] := \max_{k \geq 1} \left\{ \gamma_2((P^*_\ell)^k P_\ell^k/k), \quad k \in \mathbb{N}. \right\}
\end{equation}

As mentioned in [35], the pseudo-spectral gap can be understood as a generalization of the $L_2$-spectral gap of the multiplicative reversibilization $R_{\ell,k} = (P^*_\ell)^k P_\ell^k$. It is shown in [35, Proposition 3.4] that for a uniformly ergodic chain with Markov kernel, $P_\ell$, \exists $\tau_\ell < \infty$ such that $\gamma_{ps}[P_\ell] \geq \frac{1}{\tau_\ell} > 0$, where $\tau_\ell$ is the so-called mixing time.

**Theorem 4.7 (nonasymptotic bound on the MSE).** Suppose Assumption A holds. Furthermore, for any $\ell = 1, 2, \ldots, L$, let $Y_\ell \in L_2(X^2, \nu_\ell)$, and write $g_\ell(\theta_\ell) = Y_\ell(\theta_\ell) - \int_{X^2} Y_\ell(\theta_\ell) \nu_\ell(d\theta_\ell)$, and assume the Markov chain generated by $P_\ell$ is started from a measure $\nu^0$ with $\nu^0 \ll \nu_\ell$, and \frac{dw_\ell}{d\nu_\ell} \in L_{\infty}(X^2, \nu_\ell).$ Then,

\begin{equation}
\text{MSE}(\hat{Y}_{\ell,N_\ell,n_\ell}, \nu^0) := \mathbb{E}_{\nu^0,P_\ell} \left| \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} g_\ell(\Theta_\ell^{n,n_\ell}) \right|^2 \leq C_{\text{mse},\ell} \frac{\nu_\ell \left[ Y_\ell \right]}{N_\ell},
\end{equation}

where $C_{\text{mse},\ell} = C_{\text{inv},\ell} + C_{\text{nse},\ell}$, with

\begin{equation}
C_{\text{inv},\ell} = \left(1 + \frac{4}{\gamma_{ps}[P_\ell]} \right), \quad C_{\text{nse},\ell} = \left(2 \left\| \frac{d \nu^0}{d \nu_\ell} - 1 \right\|_{L_{\infty}(X^2, \nu_\ell)} \left(1 + \frac{4}{\gamma_{ps}[P_\ell]} \right) \right),
\end{equation}

and where $\gamma_{ps}[P_\ell]$ is the pseudo-spectral gap of $P_\ell$, defined in (4.7).
The proof of this theorem is technical and presented in the appendix for a general uniformly ergodic (nonreversible) Markov chain. Notice that the Assumption \( \nu^0 \ll \nu_\ell \) holds in our setting by Theorem 4.6 for \( \nu^0(A) = \lambda^0(A_\Delta) \).

We will work under the additional assumption that \( C_{\text{mse}, \ell} \not\to \infty \) as \( \ell \to \infty \).

**Assumption B.** There exists a level-independent constant \( C_{\text{mse}} < +\infty \) such that, for any \( \ell = 0, 1, \ldots, \) it holds that \( C_{\text{mse}, \ell} < C_{\text{mse}} \).

Such an assumption is required for the proof of complexity results such as that in Theorem 5.4 (c.f. [9, Theorem 3.4]). Notice that from (4.9), one has that \( C_{\text{mse}, \ell} \to \infty \) as \( \ell \to \infty \) if either \( \gamma_{\text{ps}}[P_\ell] \to 0 \) (Case I) or if \( \left\| \frac{d\nu_\ell}{d\nu_{\ell-1}} - 1 \right\|_{L_\infty(X^2, \mu_{\ell})} \to \infty \) as \( \ell \to \infty \) (Case II). Regarding Case I, loosely speaking, \( \gamma_{\text{ps}}[P_\ell] \in (0, 1) \) quantifies how fast the chain is mixing; in the sense that a torpidly mixing chain will result in small values of \( \gamma_{\text{ps}}[P_\ell] \), and conversely, a rapidly mixing chain will have values of \( \gamma_{\text{ps}}[P_\ell] \) close to one. Thus, (and once again, loosely speaking) Assumption B can be seen as a way of preventing the case where the proposals become increasingly worse as one goes deeper on the levels, to the point that \( Q_\ell \) and \( \mu_j \), \( j = \ell - 1, \ell \), become mutually singular as \( \ell \to \infty \). Thus, we expect such an assumption to be satisfied provided that the proposals do not degenerate (i.e., become mutually orthogonal with respect to either posterior measure) with \( \ell \). As for Case II, notice that since we are using an IMH sampler, under the assumption that \( \mu_\ell \) and \( \mu_{\ell-1} \) are not mutually singular probability measures, and that \( Q_\ell \) is a reasonably good proposal the joint measures \( \nu_\ell \) have mass in the diagonal (since it can happen that both chains at levels \( \ell - 1 \) and \( \ell \) accept the same proposed state in Algorithm 3.1). Furthermore, since the joint chain is started from the diagonal, and the mass of \( \nu_\ell \) gets increasingly concentrated in the diagonal as \( \ell \to \infty \) (intuitively, this happens because \( \mu_\ell \), \( \mu_{\ell-1} \to \mu^0 \) as \( \ell \to \infty \), which makes the probability of only one chain accepting the proposal go to zero asymptotically in \( \ell \); c.f. Lemma 5.10 for a rigorous statement on this), we then expect the term \( \left\| \frac{d\nu_\ell}{d\nu_{\ell-1}} - 1 \right\|_{L_\infty(X^2, \nu_{\ell})} \) to stay bounded (or to even decrease) in \( \ell \).

**4.2.1. Discussion of other bounds on the nonasymptotic MSE.** We now discuss some other works presenting bounds on the nonasymptotic MSE of an estimator. The work [24] presents such a type of bound provided some assumptions on the convergence of the chain are satisfied. In particular, such a bound is given in terms of the \( V \)-norm of \( f : X \to \mathbb{R} \)

\[
|f|_V := \sup_{\theta \in X} \frac{|f(\theta)|}{V(\theta)}, \quad f\text{ measurable},
\]

with \( V : X \to [1, \infty] \) understood to be a Lyapunov function associated to the Markov operator \( P \) (see, e.g., [33] for more details). However, we decide against using such a result for two reasons. On the one hand, as we will see in the following section, we aim at constructing such a bound in terms of the asymptotic variance (c.f. Result T3 in Theorem 5.2), i.e., in terms of the \( L_2 \)-norm, to be able to use the complexity result from [9]. On the other hand, because of the uniform ergodicity of the chain, the Lyapunov function associated with \( P \) is constant and therefore the bound applies only to bounded functions; this is in contrast to our result which applies to \( L_2 \) functions.
Two other bounds are presented in [40, Theorems 3.34 and 3.41]. The bound in [40, Theorem 3.34] is perhaps the most similar to our work, which we now describe in our setting. Such a result, relies upon the assumptions that

1. the chain generated by \( P_\ell \) is reversible, and
2. that \( P_\ell \) is \( L_1 \)-exponentially convergent as defined in [40]; i.e., there exist positive constants \( \hat{c}_\ell < +\infty \) and \( a_\ell \in (0, 1) \) such that

\[
\| P_\ell^n - \tilde{\nu}_\ell \|_{L_1(X^2, \nu_\ell) \to L_1(X^2, \nu_\ell)} \leq \hat{c}_\ell a_\ell^n,
\]

where \( \tilde{\nu}_\ell f := \nu_\ell(f) = \int_{X^2} f(\theta_\ell) \nu_\ell(d\theta_\ell) \), is the averaging operator and \( f : X^2 \to \mathbb{R} \) is some \( \nu_\ell \)-integrable function.

Under these assumptions, together with the assumption that the Markov chain generated by \( P_\ell \) is started from a measure \( \nu^0 \) with \( \nu^0 \ll \nu_\ell \), and \( \frac{d\nu^0}{d\nu_\ell} \in L_\infty(X^2, \nu_\ell) \), such a work presents the following bound:

\[
\mathbb{E}_{\nu^0, P_\ell} \left[ \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} g(\Theta_\ell^{n+\nu_\ell, t}) \right]^2 \leq \frac{\mathbb{V}_{\nu_\ell}[Y_\ell]}{N_\ell} \left( \frac{2}{1 - b_\ell} + \frac{2\hat{c}_\ell \| \frac{d\nu^0}{d\nu_\ell} - 1 \|_{L_\infty(X^2, \nu_\ell)} a_\ell^{n_\ell, t}}{N_\ell(1 - a_\ell)^2} \right),
\]

with \( b_\ell < a_\ell \), where \( 1 - b_\ell \) is the \( L_2(X^2, \nu_\ell) \)-spectral gap of \( P_\ell \) (known to exist given the \( L_1 \)-exponential convergence of \( P_\ell \) and the reversibility of the chain; see [40, Proposition 3.24]). Notice that the first term inside the parenthesis in the right-hand side of (4.10) is associated to the variance contribution to the MSE, while the second term corresponds to the statistical squared bias and is of higher order in \( N_\ell \). While the bound in (4.10) is indeed sharper than the one we present in Theorem 4.7, it can not be used for our purposes, as it relies upon the assumption of reversibility of the joint chain, which is not satisfied in our context, as well as on the additional requirement of \( L_1 \)-exponential convergence, which is impractical to verify in this setting. Another similar bound is presented in [40, Theorem 3.41]. Such a bound does not require reversibility of the chain; however, it requires the existence of a positive \( L_2(X^2, \nu_\ell) \)-spectral gap (as opposed to ours, which requires a positive pseudo-spectral gap). Furthermore, said bound is given in terms of an \( L_p(X^2, \nu_\ell) \)-norm with \( p > 2 \), which is, once again, outside of our setting.

5. Cost analysis of the ML-MCMC algorithm. For \( \ell = 0, 1, \ldots, L \), let \( \mathcal{L}_\ell : X \mapsto \mathbb{R} \) be a \( \mu^0_\ell \)-integrable quantity of interest, denote by \( \mathbb{E}_{\mu_\ell} \) (resp., \( \mathbb{V}_{\mu_\ell} \)) the expectation (resp., variance) with respect to the whole ML-MCMC sampling procedure, and denote by \( E = \mathbb{E}_{\nu^0, P_\ell} \) (resp., \( V = \mathbb{V}_{\nu^0, P_\ell} \)) the expectation with respect to the law of a Markov chain generated by the \( \nu_\ell \)-invariant Markov operator \( P_\ell \) started form an initial measure \( \nu^0 \). The total MSE of the multilevel estimator (3.1) is given by the following:

\[
\hat{c}_{\mu_\ell}(\mathcal{L}_{L, \{N_\ell\}_{\ell=0}}^{\nu^0}) := \mathbb{E}_{\mu_\ell} \left[ \left( \mathcal{L}_{L, \{N_\ell\}_{\ell=0}}^{\nu^0} - \mathbb{E}_{\mu^0}[\mathcal{L}] \right)^2 \right].
\]

Notice that the estimator \( \mathcal{L}_{L, \{N_\ell\}_{\ell=0}}^{\nu^0} \) also depends on \( \{P_\ell\}_{\ell=1}^L \), the burn-in, and initial measure for each level; however, for the sake of readability, we opted not to write these
dependencies explicitly throughout this section. The previous term can be upper bounded by

$$
\hat{e}_{ML}(\widehat{\mathcal{L}}_{\{N_i\}}_{i=0}^\infty) = \mathbb{V}_{ML}[\widehat{\mathcal{L}}_{\{N_i\}}_{i=0}^\infty] + \mathbb{E}_{ML} \left[ \mathcal{L}_{\{N_i\}}_{i=0}^\infty \right] - \mathbb{E}_{\mu^\ell}[\mathcal{L}]^2
\leq \mathbb{V}_{ML}[\widehat{\mathcal{L}}_{\{N_i\}}_{i=0}^\infty] + 2 \left[ \mathbb{E}_{ML} \left[ \mathcal{L}_{\{N_i\}}_{i=0}^\infty \right] - \mathbb{E}_{\mu^\ell}[\mathcal{L}]^2 \right]^2 + 2 \left[ \mathbb{E}_{\mu^\ell}[\mathcal{L}] - \mathbb{E}_{\mu^\ell}[\mathcal{L}] \right]^2.
$$

Notice that

$$\mathbb{V}_{ML}[\widehat{\mathcal{L}}_{\{N_i\}}_{i=0}^\infty] = \sum_{\ell=0}^L \mathbb{V}[\hat{Y}_\ell] + 2 \sum_{0 \leq \ell \leq \ell' \leq L} \text{Cov}(\hat{Y}_\ell, \hat{Y}_{\ell'}) \leq 2(L + 1) \sum_{\ell=0}^L \mathbb{V}[\hat{Y}_\ell].$$

Furthermore, we have that

$$2 \left[ \mathbb{E}_{ML} \left( \mathcal{L}_{\{N_i\}}_{i=0}^\infty \right) - \mathbb{E}_{\mu^\ell}[\mathcal{L}] \right]^2 \leq 2(L + 1) \sum_{\ell=0}^L \left( \mathbb{E}[\hat{Y}_\ell] - \mathbb{E}_{\nu^\ell}[\hat{Y}_\ell] \right)^2.$$

Recognizing the levelwise (statistical) MSE of $\hat{Y}_\ell$ as

$$\text{MSE}(\hat{Y}_\ell) = \mathbb{V}[\hat{Y}_\ell] + \left( \mathbb{E}[\hat{Y}_\ell] - \mathbb{E}_{\nu^\ell}[\hat{Y}_\ell] \right)^2,$$

we then have that

$$\hat{e}_{ML}(\widehat{\mathcal{L}}_{\{N_i\}}_{i=0}^\infty) \leq 2(L + 1) \sum_{\ell=0}^L \text{MSE}(\hat{Y}_\ell) + 2 \left[ \mathbb{E}_{\mu^\ell}[\mathcal{L}] - \mathbb{E}_{\mu^\ell}[\mathcal{L}] \right]^2 =: e_{ML}(\widehat{\mathcal{L}}_{\{N_i\}}_{i=0}^\infty).$$

For some tolerance $\text{tol} > 0$, we denote the minimal computational cost required to obtain $e_{ML}(\widehat{\mathcal{L}}_{\{N_i\}}_{i=0}^\infty) \leq \text{tol}^2$ by $C(e_{ML}(\widehat{\mathcal{L}}_{\{N_i\}}_{i=0}^\infty), \text{tol}^2)$. The focus of this section is to provide upper bounds on this computational cost, while quantifying the computational advantage of the ML-MCMC method over its single-level counter part (at level $L$). In particular, our result can be thought of as an extension of [9, Theorem 3.4]. The main result of this section is presented in Theorem 5.2. To establish a cost-tolerance relation, we must first make assumptions on the decay of the discretization error and the corresponding increase in computational cost for the evaluation of $\mathcal{F}_\ell$ as a function of the discretization parameter $M_\ell = s^\ell M_0$.

**Assumption C.** For any $\ell \geq 0$, the following hold:

- **C.1.** There exist positive functions $C_F, C_\Phi : \mathcal{X} \to \mathbb{R}_+$ independent of $\ell$, and positive constants $C_\epsilon, \alpha$ independent of $\theta$ and $\ell$ such that
  
  (a) $\|F_\ell(\theta) - F(\theta)\|_Y \leq C_F(\theta) s^{-\alpha \ell} \forall \theta \in \mathcal{X}$,
  
  (b) $|\Phi(\theta; y) - \Phi(\theta; y)| \leq C_\Phi(\theta) \|F_\ell(\theta) - F(\theta)\|_Y \forall \theta \in \mathcal{X}$,
  
  (c) $\int_\mathcal{X} \exp(C_F(\theta) C_F(\theta)) \mu_\theta(d\theta) \leq C_\epsilon < \infty$. 

Copyright © by SIAM and ASA. Unauthorized reproduction of this article is prohibited.
C.2. Given a $\mu^\ell$-integrable quantity of interest $\mathcal{D}_\ell$, there exists a function $C_q : X \to \mathbb{R}_+$ independent of $\ell$ and positive constants $\hat{C}_q, \alpha_q, \ell_m$, and $m > 2$, independent of $\theta$ and $\ell$ such that

(a) $|\mathcal{D}_\ell(\theta) - \mathcal{D}(\theta)| \leq C_q(\theta)s^{-\alpha_q}\ell \forall \theta \in X.$

(b) \(\int_X C^2_q(\theta)\mu_{pr}(d\theta) \leq C^2_q < \infty\).

(c) \((\int_X |\mathcal{D}_\ell(\theta)|^m\mu_{pr}(d\theta))^{1/m} \leq \ell_m < \infty\).

C.3. There exist positive constants $\gamma$ and $C_\gamma$, such that, for each discretization level $\ell$, the computational cost of obtaining one sample from a $\mu^\ell$-integrable quantity of interest $\mathcal{D}_\ell(\theta_{\ell,\ell})$, $\theta_{\ell,\ell} \sim \mu^\ell$, with $\theta_{\ell,\ell}$ generated by Algorithm 3.1, denoted by $C_\ell(\mathcal{D}_\ell)$, scales as

$$C_\ell(\mathcal{D}_\ell) \leq C_\gamma s^{\gamma \ell}.$$  

Remark 5.1. Notice that, with a slight abuse of notation, we have used the symbol $\alpha$ to denote the (strong) rate in C.1, and $\alpha\ell(\cdot, \cdot)$ to denote acceptance probability at level $\ell$. We hope this does not create any confusion.

For all $\ell = 1, 2, \ldots, L$, and for a $\mu^\ell$-integrable quantity of interest $\mathcal{D}_j$, $j = \ell - 1, \ell$, we write

$$Y_\ell(\Theta_\ell) := \mathcal{D}_\ell(\Theta_\ell) - \mathcal{D}_{\ell-1}(\Theta_{\ell-1,\ell}), \ (\Theta_{\ell-1,\ell}, \Theta_{\ell,\ell}) := \Theta_\ell \sim \nu_\ell,$$

$$\hat{Y}_{\ell,N_\ell} := \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} Y_\ell(\Theta_\ell^{n+n_\ell,\ell}), \ \Theta_\ell^{n+n_\ell,\ell} \sim \rho_\ell(\Theta_\ell^{n-1,\ell}).$$

We state the main result of this section.

Theorem 5.2 (decay of errors). For any $\ell = 0, 1, \ldots, L$, let $\mathcal{D}_\ell$ be an $L_1(X, \mu^\ell)$-integrable quantity of interest, and suppose Assumptions A, B, and C hold. Then, there exist positive constants $C_w, C_v, C_{mse}$, independent of $\ell$ such that

T1. (Weak convergence) $|E_{\mu^\ell}[\mathcal{D}_\ell] - E_{\mu^\ell}[\mathcal{D}]| \leq C_w s^{-\alpha_w}\ell$,

T2. (Strong convergence) $\nu_{v_\ell}[Y_\ell] \leq C_v s^{-\beta}\ell$.

T3. (MSE bound) $\text{MSE}(\hat{Y}_{\ell,N_\ell}) \leq N^{-1}_{\ell}C_{mse}\nu_{v_\ell}[Y_\ell]$. Here, $\alpha_w = \min\{\alpha_q, \alpha\}$ and $\beta = \min\{2\alpha_q, \alpha(1 - 2/m)\}$, with $\alpha, \alpha_q$, and $m$ as in Assumption C.

The proof of Theorem 5.2 is presented in section 5.1. It has been shown in [9, Theorem 3.4] that, if a ML-MCMC algorithm satisfies conditions T1–T3, then it has a complexity (cost-tolerance relation) analogous to a standard MLMC algorithm to compute expectations (when independent sampling from the underlying probability measure is possible) up to logarithmic terms. This result is stated in Theorem 5.4 below. The purpose of Theorem 5.2 is to show that our class of ML-MCMC algorithms does actually fulfill conditions T1–T3.

Remark 5.3. Throughout this work, we have the tacit assumption that the chain at level $0$, i.e., the one that does not require an IMH sampler, is geometrically ergodic with respect to $\mu^0_0$.

Theorem 5.4 (see [9, Theorem 3.4]). Under the same assumptions as in Theorem 5.2, with $\alpha_w \geq \frac{1}{2}\min\{\gamma, \beta\}$ for any $\text{tol} > 0$ there exist a number of levels $L = L(\text{tol})$, a decreasing
sequence of integers \( \{N_{\ell}(\text{tol})\}_{\ell=0}^{L} \), and a positive constant \( C_{\text{ML}} \) independent of \( \text{tol} \), such that the MSE bound of the multilevel estimator, \( e_{\text{ML}}(\hat{{\varphi}}_{L,N_{\ell}})_{\ell=0}^{L} \), satisfies
\[
e_{\text{ML}}(\hat{{\varphi}}_{L,N_{\ell}})_{\ell=0}^{L} \leq \text{tol}^{2},
\]
whereas, the corresponding total ML-MCMC cost is bounded by
\[
C \left( e_{\text{ML}}(\hat{{\varphi}}_{L,N_{\ell}})_{\ell=0}^{L}, \text{tol}^{2} \right) \leq C_{\text{ML}} \begin{cases} \text{tol}^{-2} |\log \text{tol}| & \text{if } \beta > \gamma, \\ \text{tol}^{-2} |\log \text{tol}|^{3} & \text{if } \beta = \gamma, \\ \text{tol}^{-2+\left(\gamma-\beta\right)/\alpha_{w}} |\log \text{tol}| & \text{if } \beta < \gamma. \end{cases}
\]

**Proof.** Just as in [9], the proof of this theorem follows from (5.1) and the proof of [5, Theorem 1] \( \square \)

We remark that the rates in Theorem 5.4 are independent of the dimension of \( X \). It is shown also in [9] that the cost of obtaining an equivalent single-level (at level \( L \)) MSE of an estimator \( \hat{{\varphi}}_{N} \) based on a single-level MCMC algorithm (e.g., standard MH) (denoted by \( e_{\text{SL}} \)), generated by a reversible and geometrically ergodic Markov kernel is given by
\[
C \left( e_{\text{SL}}(\hat{{\varphi}}_{N}), \text{tol}^{2} \right) \leq C_{\text{SL}} \text{tol}^{-2-\gamma/\alpha_{w}}, \quad C_{\text{SL}} \in \mathbb{R}_{+},
\]
where \( \alpha_{w} \) and \( \gamma \) are the same constants as in Theorem 5.2, and \( C_{\text{SL}} \) is a positive constant independent of \( \text{tol} \).

**5.1. Proof of Theorem 5.2.** We will decompose the proof of Theorem 5.2 in a series of auxiliary results. Notice that T3 is obtained from Theorem 4.7 with a level dependent constant and we postulated in Assumption B that this constant can be bounded by a finite, level-independent constant \( C_{\text{mse}} \), and as such, we can use it in T3. Thus, we just need to prove that T1 and T2 hold. This is done in Lemmas 5.7 and 5.11. We begin by proving some auxiliary results needed to prove implication T1.

**Lemma 5.5.** Suppose Assumption C holds. Then, for \( \ell = 1, 2, \ldots, L \) it holds for \( Z_{\ell} = \int_{X} e^{-\Phi_{\ell}(\theta; y)} \mu_{\text{pr}}(d\theta) \) that
\[
c_{l} \leq Z_{\ell} \leq C_{e},
\]
where \( c_{l} = \int_{X} \exp(-\Phi(\theta; y) - C_{F}(\theta)C_{\Phi}(\theta)) \mu_{\text{pr}}(d\theta) \) and \( C_{e} \) as in Assumption C.

**Proof.** From Assumption C.1 one has that, \( \forall \ell \geq 0, \) and \( \theta \in X, \)
\[
\Phi(\theta; y) - C_{\Phi}(\theta)C_{F}(\theta) \leq \Phi_{\ell}(\theta; y) \leq \Phi(\theta; y) + C_{\Phi}(\theta)C_{F}(\theta).
\]
Hence,
\[
Z_{\ell} = \int_{X} \exp(-\Phi_{\ell}(\theta; y)) \mu_{\text{pr}}(d\theta) \leq \int_{X} \exp(-\Phi(\theta; y) - C_{\Phi}(\theta)C_{F}(\theta)) \mu_{\text{pr}}(d\theta) \\
\leq \int_{X} \exp(C_{\Phi}(\theta)C_{F}(\theta)) \mu_{\text{pr}}(d\theta) = C_{e},
\]
Copyright © by SIAM and ASA. Unauthorized reproduction of this article is prohibited.
where the last step follows from the assumption of nonnegativity of $\Phi(\theta; y)$ (c.f. (2.2)). Similarly, $Z_\ell \geq \int_X \exp(-\Phi(\theta; y) - C_\theta(\theta) C_F(\theta)) \mu_{\text{pr}}(d\theta) = c_\ell$, independently of $\ell$.

Lemma 5.6. Suppose Assumption C holds. Then, for any $\ell \geq 1$, there exist positive functions $C_{\pi,\ell}(\theta) : X \to \mathbb{R}_+$, $\tilde{C}_{\pi,\ell}(\theta) : X \to \mathbb{R}_+$, such that

\begin{align}
|\pi_{\ell}^y(\theta) - \pi_{\ell-1}^y(\theta)| & \leq C_{\pi,\ell}(\theta)s^{-\alpha \ell} \quad \forall \theta \in X, \\
|\pi_{\ell}^y(\theta) - \pi_{\ell}^y(\theta)| & \leq \tilde{C}_{\pi,\ell}(\theta)s^{-\alpha \ell} \quad \forall \theta \in X.
\end{align}

Moreover, $C_{\pi,\ell}(\theta) = (\pi_{\ell}^y(\theta) + \pi_{\ell-1}^y(\theta))K_{\pi,\ell}(\theta)$, $\tilde{C}_{\pi,\ell}(\theta) = (\pi_{\ell}^y(\theta) + \pi_{\ell}^y(\theta))\tilde{K}_{\pi,\ell}(\theta)$, with

\begin{align*}
\tilde{K}_{\pi,\ell}(\theta) & = C_\theta(\theta)C_F(\theta) + c_\ell^{-1}C_e, \\
K_{\pi,\ell}(\theta) & = (1 + s^\alpha)\tilde{K}_{\pi,\ell}(\theta).
\end{align*}

Furthermore, we have that for any $p \in [1, +\infty]$

\begin{align*}
K_p & := \left( \int_X |K_{\pi,\ell}(\theta)|^p \mu_{\text{pr}}(d\theta) \right)^{1/p} < +\infty, \\
\tilde{K}_p & := \left( \int_X |\tilde{K}_{\pi,\ell}(\theta)|^p \mu_{\text{pr}}(d\theta) \right)^{1/p} < +\infty.
\end{align*}

**Proof.** We begin with the proof of (5.2). We consider first the case $\Phi_{\ell}(\theta; y) \leq \Phi_{\ell-1}(\theta; y)$:

\begin{align*}
|\pi_{\ell}^y(\theta) - \pi_{\ell-1}^y(\theta)| & = \left| \frac{e^{-\Phi_{\ell}(\theta; y)}}{Z_\ell} - \frac{e^{-\Phi_{\ell-1}(\theta; y)}}{Z_{\ell-1}} \right| \\
& \leq \left| \frac{e^{-\Phi_{\ell}(\theta; y)}}{Z_\ell} - \frac{e^{-\Phi_{\ell-1}(\theta; y)}}{Z_\ell} \right| + \left| \frac{e^{-\Phi_{\ell-1}(\theta; y)}}{Z_\ell} - \frac{e^{-\Phi_{\ell-1}(\theta; y)}}{Z_{\ell-1}} \right|
\end{align*}

We first focus on $I$. A straightforward application of the mean value theorem gives

\begin{align}
|e^{-\Phi_{\ell}(\theta; y)} - e^{-\Phi_{\ell-1}(\theta; y)}| & \leq e^{-\Phi_{\ell}(\theta; y)}|\Phi_{\ell}(\theta; y) - \Phi_{\ell-1}(\theta; y)|.
\end{align}

Thus, we have from (5.4), together with Assumptions C.1 that

\begin{align}
I & = Z_{\ell-1}^{-\ell} \left| e^{-\Phi_{\ell}(\theta; y)} - e^{-\Phi_{\ell-1}(\theta; y)} \right| \leq \pi_{\ell}^y(\theta)|\Phi_{\ell}(\theta; y) - \Phi_{\ell-1}(\theta; y)| \\
& \leq \pi_{\ell}^y(\theta) (|\Phi_{\ell}(\theta; y) - \Phi(\theta; y)| + |\Phi_{\ell-1}(\theta; y) - \Phi(\theta; y)|) \\
& \leq \pi_{\ell}^y(\theta)C_\theta(\theta) (|\Phi_{\ell}(\theta; y) - \Phi(\theta; y)| + |\Phi_{\ell-1}(\theta; y) - \Phi(\theta; y)|) \\
& \leq \pi_{\ell}^y(\theta)C_\theta(\theta)C_F(\theta)(1 + s^\alpha)s^{-\alpha \ell}.
\end{align}

We now shift our attention to $II$. Following a similar procedure as for $I$, we have that

\begin{align*}
II & \leq \frac{\pi_{\ell-1}^y(\theta)}{Z_\ell} \int_X \left| e^{-\Phi_{\ell}(\theta; y)} - e^{-\Phi_{\ell-1}(\theta; y)} \right| \mu_{\text{pr}}(d\theta) \\
& \leq \frac{\pi_{\ell-1}^y(\theta)}{Z_\ell} \int_X e^{-\min\{\Phi_{\ell}(\theta; y), \Phi_{\ell-1}(\theta; y)\} |\Phi_{\ell}(\theta; y) - \Phi_{\ell-1}(\theta; y)|} \mu_{\text{pr}}(d\theta) \\
& \leq \pi_{\ell-1}^y(\theta) (1 + s^\alpha)s^{-\alpha \ell} e_\ell^{-1} \int_X C_\theta(\theta)C_F(\theta) e^{-\min\{\Phi_{\ell}(\theta; y), \Phi_{\ell-1}(\theta; y)\}} \mu_{\text{pr}}(d\theta)
\end{align*}
\[
\pi^y_{\ell-1}(\theta)(1 + s^\alpha) s^{-\alpha} \ell^{-1} \int_\mathcal{X} C_\Phi(z) C_{\bar{\mathcal{F}}}(z) \mu_{pr}(dz) \leq \pi^y_{\ell-1}(\theta)(1 + s^\alpha) s^{-\alpha} \ell^{-1} C_c,
\]

where in the last step we have used that

\[
\int_\mathcal{X} C_\Phi(\theta) C_{\bar{\mathcal{F}}}(\theta) \mu_{pr}(d\theta) \leq \int_\mathcal{X} \exp(C_\Phi(\theta) C_{\bar{\mathcal{F}}}(\theta)) \mu_{pr}(d\theta) \leq C_c.
\]

Adding (5.5) and (5.6) gives the desired result with

\[
C'_{\pi,\ell}(\theta) = (\pi^y_{\ell-1}(\theta) C_\Phi(\theta) C_{\bar{\mathcal{F}}}(\theta) + \pi^y_{\ell-1}(\theta) c^{-1}_I C_c) (1 + s^\alpha).
\]

The case \( \Phi_\ell(\theta; y) > \Phi_{\ell-1}(\theta; y) \) can be treated analogously by considering the alternative splitting

\[
|\pi^y_\ell(\theta) - \pi^y_{\ell-1}(\theta)| \leq \left( \left| \frac{e^{-\Phi_\ell(\theta; y)}}{\mathcal{Z}_\ell} - \frac{e^{-\Phi_{\ell-1}(\theta; y)}}{\mathcal{Z}_{\ell-1}} \right| + \left| \frac{e^{-\Phi_\ell(\theta; y)}}{\mathcal{Z}_\ell} - \frac{e^{-\Phi_{\ell-1}(\theta; y)}}{\mathcal{Z}_{\ell-1}} \right| \right) \cdot
\]

which yields the constant \( C'_{\pi,\ell}(\theta) = (\pi^y_{\ell-1}(\theta) C_\Phi(\theta) C_{\bar{\mathcal{F}}}(\theta) + \pi^y_{\ell-1}(\theta) c^{-1}_I C_c)(1 + s^\alpha) \). Thus, one can obtain the desired bound \(|\pi^y_\ell(\theta) - \pi^y_{\ell-1}(\theta)| \leq C_{\pi,\ell}(\theta) s^{-\alpha} \) with

\[
C_{\pi,\ell}(\theta) = (\pi^y_{\ell-1}(\theta) + \pi^y_\ell(\theta)) K_{\pi,\ell}(\theta),
\]

\[
K_{\pi,\ell}(\theta) = (C_\Phi(\theta) C_{\bar{\mathcal{F}}}(\theta) + c^{-1}_I C_c)(1 + s^\alpha).
\]

A similar procedure shows that the bound (5.3) holds with

\[
\bar{C}_{\pi,\ell}(\theta) = (\pi^y_\ell(\theta) + \pi^y_{\ell-1}(\theta)) \bar{K}_{\pi,\ell}(\theta),
\]

\[
\bar{K}_{\pi,\ell}(\theta) = C_\Phi(\theta) C_{\bar{\mathcal{F}}}(\theta) + c^{-1}_I C_c.
\]

We finally remark that

\[
K_p := \left( \int_\mathcal{X} |K_{\pi,\ell}(\theta)|^p \mu_{pr}(d\theta) \right)^{1/p} = (1 + s^\alpha) \left( \int_\mathcal{X} (C_\Phi(\theta) C_{\bar{\mathcal{F}}}(\theta) + c^{-1}_I C_c)^p \mu_{pr}(d\theta) \right)^{1/p} \leq (1 + s^\alpha) \left( \frac{p}{c} \right) \left( \int_\mathcal{X} \exp \left\{ C_\Phi(\theta) C_{\bar{\mathcal{F}}}(\theta) + c^{-1}_I C_c \right\} \mu_{pr}(d\theta) \right)^{1/p} \leq (1 + s^\alpha) \left( \frac{p}{c} \right) \left( C_e \exp \left\{ c^{-1}_I C_c \right\} \right)^{1/p} \leq +\infty.
\]

A similar calculation for \( \bar{K}_{\pi,\ell} \) leads to

\[
\bar{K}_p = \left( \int_\mathcal{X} |\bar{K}_{\pi,\ell}(\theta)|^p \mu_{pr}(d\theta) \right)^{1/p} \leq \left( \frac{p}{c} \right) \left( C_e \exp \left\{ c^{-1}_I C_c \right\} \right)^{1/p} \leq +\infty.
\]

We can now show implication T1.
Lemma 5.7. Suppose Assumption C holds. Then, for any $\ell = 0, 1, \ldots, L$, there exists a positive constant $C_w \in \mathbb{R}_+$, independent of $\ell$, such that

$$|E_{\Theta_{t,\sim \mu}} [\mathcal{D}_{t}(\Theta_{t,\ell})] - E_{\Theta_{t,\sim \mu^n}} [\mathcal{D}(\Theta)]| \leq C_w s^{-\alpha_w \ell},$$

with $\alpha_w = \min \{ \alpha_q, \alpha \}$ and $\alpha_q, \alpha$ as in Assumption C.

Proof. To shorten notation, we will write $E_{\Theta_{t,\sim \mu}} [\mathcal{D}_{t}(\Theta_{t,\ell})]$ as $E_{\mu} [\mathcal{D}]$ (and similarly for $E_{\Theta_{t,\sim \mu^n}} [\mathcal{D}(\Theta)]$). We follow an approach similar to that of [9]:

$$|E_{\mu} [\mathcal{D}] - E_{\mu^n} [\mathcal{D}]| \leq |E_{\mu} [\mathcal{D}] - E_{\mu} [\mathcal{D}]| + |E_{\mu} [\mathcal{D}] - E_{\mu^n} [\mathcal{D}]|.$$

For the first term, we have that

$$|E_{\mu} [\mathcal{D}] - E_{\mu^n} [\mathcal{D}]| \leq E_{\mu} [||\mathcal{D} - \mathcal{D}||] \leq \left( \int X C_q(\theta) \mu^n_y (d\theta) \right) s^{-\alpha_{q,\ell}} \leq \frac{s^{-\alpha_{q,\ell}}}{L_\ell} \int X C_q(\theta) \mu_{pr} (d\theta) \leq c_I^{-1} C_q s^{-\alpha_{q,\ell}}.$$

For the second term, we have that

$$|E_{\mu} [\mathcal{D}] - E_{\mu^n} [\mathcal{D}]| = \left| \int X \mathcal{D}(\theta) [\pi_y^n(\theta) - \pi_y^n(\theta)] \mu_{pr} (d\theta) \right| \leq \int X |\mathcal{D}(\theta)| (\pi_y^n(\theta) + \pi_y^n(\theta)) \tilde{K}_{\pi,\ell}(\theta) \mu_{pr} (d\theta) s^{-\alpha_{\ell}}.$$

Working on the first term of the previous integral, we obtain from Hölder’s inequality that

$$\left| \int X \mathcal{D}(\theta) \pi_y^n(\theta) \tilde{K}_{\pi,\ell}(\theta) \mu_{pr} (d\theta) \right| \leq \left( \int X |\mathcal{D}(\theta)| \mu_{pr} (d\theta) \right)^{1/m} \left( \int X \pi_y^n(\theta) |\tilde{K}_{\pi,\ell}(\theta)|^{m'} \mu_{pr} (d\theta) \right)^{1/m'} \leq C_m c_I^{-1} \tilde{K}_{m',\ell},$$

where we have taken $m$ as in Assumption C, $m' = 1 - 1/m$ and $\tilde{K}_{m',\ell}$ as in Lemma 5.6. A similar bound holds for the second term in (5.8), thus leading to

$$|E_{\mu} [\mathcal{D}] - E_{\mu^n} [\mathcal{D}]| \leq 2c_I^{-1} C_m \tilde{K}_{m',\ell} s^{-\alpha_{\ell}}.$$

The desired result follows from (5.7) and (5.9), with $\alpha_w = \min \{ \alpha_q, \alpha \}$, and a level independent constant $C_w = c_I^{-1} (2C_m \tilde{K}_{m',\ell} + C_q)$. \hfill \Box

We now shift our attention to implication T2. We first prove several auxiliary results. For any given level $\ell = 0, 1, \ldots, L$, we say that the joint chains created by Algorithm 3.1 are synchronized at step $n$ if $\Theta_{t,\ell}^n = \Theta_{t,\ell-1}^n$. Conversely, we say they are unsynchronized at step $n$ if $\Theta_{t,\ell}^n \neq \Theta_{t,\ell-1}^n$. Notice that if the chains are synchronized at a state $\Theta_{t,\ell}^n = \Theta_{t,\ell-1}^n = \theta$, and the new proposed state at the $(n+1)$th iteration of the algorithm is $Z$, they de-synchronize.
at the next step with probability $|\alpha_\ell(\theta, z) - \alpha_{\ell-1}(\theta, z)|$ (c.f. Figure 1). Intuitively, one would expect that such a probability goes to 0 as $\ell \to \infty$. We formalize this intuition below.

**Lemma 5.8.** Suppose Assumption C.1 holds. Then, the following bound holds

$$|\alpha_\ell(\theta, z) - \alpha_{\ell-1}(\theta, z)| \leq h_\ell(\theta, z)s^{-\alpha_\ell}, \quad \theta, z \in X,$$

with

$$h_\ell(\theta, z) := \frac{Q_\ell(\theta)}{Q_\ell(z)} \frac{1}{\pi^y_\ell(\theta)\pi^y_{\ell-1}(\theta)} \left| \pi^y_\ell(z)C_{\pi,\ell}(\theta) + \pi^y_\ell(\theta)C_{\pi,\ell}(z) \right|$$

and $C_{\pi,\ell}(\cdot)$ as in Lemma 5.6.

**Proof.** From the definition of $\alpha_\ell$, and the fact that $\psi(x) := \min\{1, x\}$ is Lipschitz continuous with constant 1, it can be seen that

$$|\alpha_\ell(\theta, z) - \alpha_{\ell-1}(\theta, z)| \leq \frac{Q_\ell(\theta)}{Q_\ell(z)} \frac{1}{\pi^y_\ell(\theta)\pi^y_{\ell-1}(\theta)} \left| \pi^y_\ell(z)(\pi^y_{\ell-1}(\theta)) + \pi^y_\ell(\theta)(\pi^y_{\ell-1}(\theta)) \right|
\leq \frac{Q_\ell(\theta)}{Q_\ell(z)} \frac{1}{\pi^y_\ell(\theta)\pi^y_{\ell-1}(\theta)} \left| \pi^y_\ell(z)C_{\pi,\ell}(\theta) + \pi^y_\ell(\theta)C_{\pi,\ell}(z) \right| s^{-\alpha_\ell}.$$

\[\blacksquare\]

**Lemma 5.9.** Suppose Assumptions A and C hold, and denote the diagonal set of $X^2$ as $\Delta := \{(\theta, z) \in X^2 \text{ s.t. } \theta = z\}$. The transition probability to $\Delta^c$ for the coupled chain of Algorithm 3.1 is such that

$$p_\ell(\theta_\ell, \Delta^c) \leq R_\ell(\theta) s^{-\alpha_\ell} \quad \forall \theta_\ell = (\theta, \theta) \in \Delta,$$

with

$$R_\ell(\theta) = \frac{Q_\ell(\theta)}{\pi^y_\ell(\theta)\pi^y_{\ell-1}(\theta)} \left( C_{\pi,\ell}(\theta) + \pi^y_\ell(\theta)K_1 \right),$$

and $C_{\pi,\ell}(\cdot)$ and $K_1$ as in Lemma 5.6. Moreover, whenever $\theta_\ell \in \Delta^c$,

$$p_\ell(\theta_\ell, \Delta^c) \leq 1 - c \int_X \min\{\pi^y_\ell(\theta), \pi^y_{\ell-1}(\theta)\} \mu_{pr}(d\theta),$$

where $c$ is the same constant as in Assumption A.1. Furthermore, $\exists \delta > 0$ independent of $\ell$ such that

$$\inf_{\ell \in \mathbb{N}} \int_X \min\{\pi^y_\ell(\theta), \pi^y_{\ell-1}(\theta)\} \mu_{pr}(d\theta) > \delta > 0.$$
Proof. We begin with the first inequality. For $\theta_{\ell} \in \Delta$ and from the definition of $p_{\ell}$ we obtain

$$p_{\ell}(\theta_{\ell}, \Delta^c) = \int_{X} (\alpha_{\ell-1}(\theta_{\ell-1}, z) - \alpha_{\ell}(\theta_{\ell}, z))^+ Q_{\ell}(z) \mathbb{1}_{\{(z, \theta_{\ell}, z) \in \Delta^c\}} \mu_{pr}(dz) \geq 0,$$

where the first and last term in (3.2) are both 0. Writing $\theta_{\ell, i} = \theta_{\ell, i-1} = \theta$, it then follows from Lemma 5.8 that

$$p_{\ell}(\theta_{\ell}, \Delta^c) \leq \frac{Q_{\ell}(\theta)}{\pi_{\ell}^y(\theta) \pi_{\ell-1}^y(\theta)} \int_{X} |\pi_{\ell}^y(z)C_{\pi, \ell}(\theta) + \pi_{\ell}^y(\theta)C_{\pi, \ell}(z)| \mu_{pr}(dz) \leq \frac{Q_{\ell}(\theta)}{\pi_{\ell}^y(\theta) \pi_{\ell-1}^y(\theta)} \left( C_{\pi, \ell}(\theta) + \pi_{\ell}^y(\theta) \int_{X} C_{\pi, \ell}(z) \mu_{pr}(dz) \right) \leq \frac{Q_{\ell}(\theta)}{\pi_{\ell}^y(\theta) \pi_{\ell-1}^y(\theta)} \left( C_{\pi, \ell}(\theta) + 2c_1^{-1} K_1 \pi_{\ell}^y(\theta) \right).$$

Thus, one has that $\forall \theta_{\ell} \in \Delta$,

$$p_{\ell}(\theta_{\ell}, \Delta^c) \leq R_{\ell}(\theta)s_{-\alpha \ell},$$

with

$$R_{\ell}(\theta) = \frac{Q_{\ell}(\theta)}{\pi_{\ell}^y(\theta) \pi_{\ell-1}^y(\theta)} \left( C_{\pi, \ell}(\theta) + 2c_1^{-1} K_1 \pi_{\ell}^y(\theta) \right).$$

We now focus on the second inequality which holds for $\theta_{\ell} \in \Delta^c$. Thus, from the fact that $\max\{a, b\} - |a - b| = \min\{a, b\} \forall a, b \in \mathbb{R}$ and using Assumption A.1, we obtain

$$p_{\ell}(\theta_{\ell}, \Delta^c) \leq \int_{X} \left( 1 - \min_{j=\ell-1, \ell} \left\{ \alpha_j(\theta_{\ell, j}, u) \right\} \right) Q_{\ell}(u) \mu_{pr}(du) \leq 1 - \int_{X} \min_{j=\ell-1, \ell} \left\{ \frac{Q_{\ell}(u)}{c \pi_{\ell}^y(\theta)} \right\} \mu_{pr}(du) = 1 - \int_{X} \min_{j=\ell-1, \ell} \left\{ \frac{Q_{\ell}(u)}{\pi_{\ell}^y(\theta)} \right\} \mu_{pr}(du) \leq 1 - c \int_{X} \min_{j=\ell-1, \ell} \left\{ \pi_{\ell}^y(u) \right\} \mu_{pr}(du),$$

where $c$ is the same constant as in Assumption A.1 (notice that $c < 1$).

Finally, we show that the integral term in the previous equation is lower bounded by a strictly positive constant independent of $\ell$. First notice that

$$\lim_{\ell \to \infty} \int_{X} \min_{j=\ell-1, \ell} \left\{ \pi_{\ell}^y(\theta) \right\} \mu_{pr}(d\theta) = 1 - \lim_{\ell \to \infty} \frac{1}{2} \int_{X} |\pi_{\ell}^y(\theta) - \pi_{\ell-1}^y(\theta)| \mu_{pr}(d\theta) \geq \lim_{\ell \to \infty} (1 - K_1 s^{-\alpha \ell}) = 1,$$
and, by definition,
\[ \int_X \min_{j=1, \ell} \left\{ \pi_j^n(\theta) \right\} \mu_{\text{pr}}(d\theta) \leq 1 \quad \forall \ell \in \mathbb{N}. \]

Thus, the sequence \( \{ \int_X \min_{j=1, \ell} \{ \pi_j^n(\theta) \} \mu_{\text{pr}}(d\theta) \}_{\ell \in \mathbb{N}} \) has 1 as an accumulation point, as \( \ell \to \infty \), and there exists \( \delta' > 0 \) and \( \ell' \geq 0 \) such that, for any \( \ell \geq \ell' \), \( \int_X \min_{j=1, \ell} \{ \pi_j^n(\theta) \} \mu_{\text{pr}}(d\theta) \geq \delta' \). Lastly, recall that by Assumption A.2, \( \pi^n_{\ell} \) and \( \pi^n_{\ell-1} \) are continuous and strictly positive. Thus, for any compact set \( A \subset X \) with \( \mu_{\text{pr}}(A) > 0 \) and for any \( \ell = \{0, 1, \ldots, \ell'\} \), we have
\[ \int_X \min_{j=1, \ell} \{ \pi_j^n(\theta) \} \mu_{\text{pr}}(d\theta) \geq \int_{A} \min_{j=1, \ell} \{ \pi_j^n(\theta) \} \mu_{\text{pr}}(d\theta) =: \delta > 0. \]

Thus setting \( \delta = \min_{0 \leq \ell \leq \ell'} \{ \delta_{\ell} \} \), and \( \delta = \min \{ \delta, \delta' \} \) we obtain that, for any \( \ell \geq 0 \),
\[ \int_X \min_{j=1, \ell} \{ \pi_j^n(\theta) \} \mu_{\text{pr}}(d\theta) \geq \delta > 0. \]

Lemma 5.10. Let \( (\Theta_{\ell, \ell-1}, \Theta_{\ell}) \sim \nu_\ell \), and suppose Assumptions A and C hold. Then, \( \forall \ell = 1, 2, \ldots, L \), it holds that
\[ \mathbb{P}_{\nu_\ell}(\Theta_{\ell, \ell} \neq \Theta_{\ell, \ell-1}) \leq C_{r, \ell} s^{-\alpha \ell}, \]

where
\[ C_{r, \ell} = \frac{2C_r K_{r, \ell} (c_r^{-1} + 1)}{c \int_X \min_{j=1, \ell} \{ \pi_j^n(\theta), \pi_j^n(\theta) \} \mu_{\text{pr}}(d\theta) > 0,} \]

with \( c \) as in Assumption A.1 and \( r \) as in Assumption A.3. Furthermore, \( C_{r, \ell} \to C^* > 0 \) as \( \ell \to \infty \).

Proof. Let \( \{ \Theta^n_\ell \}_{n \in \mathbb{N}} \) be a Markov chain with transition kernel \( p_\ell \), so that at stationarity \( \Theta^n_\ell \sim \nu_\ell \forall n \in \mathbb{N} \). For notational simplicity, for the remainder of this proof we will write \( P_n := \mathbb{P}_{\nu_n}(\Theta^n_{\ell, \ell-1} \neq \Theta^n_{\ell, \ell}) \), \( n \in \mathbb{N} \). Let \( Z_{\Delta, \ell} := \int_{\Delta} \nu_\ell(d\theta_\ell) = (1 - P_n) \). From Lemma 5.9 we obtain, for any \( n \in \mathbb{N} \),
\[ \mathbb{P}_{\nu_\ell}(\Theta^{n+1}_{\ell, \ell} \neq \Theta_{n+1, \ell-1} \mid \Theta^n_\ell \in \Delta) = Z_{\Delta, \ell}^{-1} \int_{\Delta} p_\ell(\theta_\ell, \Delta^c) \nu_\ell(d\theta_\ell) \]
\[ \leq \frac{s_{n+1} - \alpha \ell}{Z_{\Delta, \ell}} \int_{\Delta} R_\ell(\theta) \nu_\ell(d\theta_\ell) \quad (\text{with } \theta_\ell = (\theta, \theta) \text{ on } \Delta) \]
\[ \leq \frac{s_{n+1} - \alpha \ell}{Z_{\Delta, \ell}} \int_{\Delta} \frac{Q_\ell(\theta_\ell)}{\pi^{-1}_n(\theta)} (K_{\ell, \ell}(\theta_\ell) + 2c^{-1}_r K_1) \nu_\ell(d\theta_\ell) + \frac{s_{n+1} - \alpha \ell}{Z_{\Delta, \ell}} \int_{\Delta} \frac{Q_\ell(\theta_\ell)}{\pi^{-1}_n(\theta)} K_{\ell, \ell}(\theta_\ell) \nu_\ell(d\theta_\ell). \]

We begin with integral I:
\[ I = \int_{X^2} \frac{Q_\ell(\theta_\ell, \ell-1)}{\pi^{-1}_n(\theta_\ell, \ell-1)} (K_{\ell, \ell}(\theta_\ell, \ell-1) + 2c^{-1}_r K_1) \mathbf{1}_{(\theta_\ell, \ell-1, \theta_\ell, \ell) \in \Delta} \nu_\ell(d\theta_\ell) \]
\[ \leq \int_X \frac{Q_\ell(\theta_\ell, \ell-1)}{\pi^{-1}_n(\theta_\ell, \ell-1)} (K_{\ell, \ell}(\theta_\ell, \ell-1) + 2c^{-1}_r K_1) \nu_\ell(d\theta_\ell, d\theta_\ell). \]
\[ = \int_X Q_\ell(\theta_{\ell,\ell-1}) (K_\pi,\ell(\theta_{\ell,\ell-1}) + 2c_r^{-1}K_1) \mu_{pr}(d\theta_{\ell,\ell-1}) \]
\[ \leq \left( \int_X |Q_\ell(\theta_{\ell,\ell-1})|^{1/\tau} \mu_{pr}(d\theta_{\ell,\ell-1}) \right)^{1/\tau} \left( 2c_r^{-1}K_1 + \left( \int_X |K_\pi,\ell(\theta_{\ell,\ell-1})|^{1/\tau} \mu_{pr}(d\theta_{\ell,\ell-1}) \right)^{1/\tau} \right) \]
\[ = C_r(2c_r^{-1} + 1)K_{\ell,\ell}. \]

Similarly, for II, we get
\[ II = \int_X \frac{Q_\ell(\theta_{\ell,\ell})}{\pi(\theta_{\ell,\ell})} K_\pi,\ell(\theta_{\ell,\ell}) \mathbb{I}_{\{\theta_{\ell,\ell-1,\ell} \in \Delta\}} \nu_\ell(d\theta_{\ell,\ell-1}, d\theta_{\ell,\ell}) \]
\[ \leq \int_X Q_\ell(\theta_{\ell,\ell}) K_\pi,\ell(\theta_{\ell,\ell}) \mu_{pr}(d\theta_{\ell,\ell}) \leq C_rK_{\ell,\ell}. \]

Setting \( \hat{C} = 2C_rK_{r}(c_r^{-1} + 1) \), one then has
\[ P_{\nu_\ell}(\Theta_{\ell,\ell}^{n+1} \neq \Theta_{\ell,\ell-1}^{n+1}|\Theta_{\ell,\ell}^n = \Theta_{\ell,\ell-1}^n) \leq \hat{C} \Delta_{\ell,\ell}^{-1} s^{-\alpha \ell} := Z_{\Delta_{\ell,\ell}}^{-1} s, \]
where we have set \( s = \hat{C} s^{-\alpha \ell} \). Similarly, letting \( Z_{\Delta_{\ell,\ell}} := \int_{\Delta_{\ell,\ell}} \nu_\ell(d\theta_{\ell}) \), one has
\[ P_{\nu_\ell}(\Theta_{\ell,\ell}^{n+1} \neq \Theta_{\ell,\ell-1}^{n+1}|\Theta_{\ell,\ell}^n \in \Delta_{\ell,\ell}) = Z_{\Delta_{\ell,\ell}}^{-1} \int_{\Delta_{\ell,\ell}} \nu_\ell(\theta_{\ell}) \] \[ \leq 1 - c \int_X \min\{\pi_\ell^n(\theta), \pi_{\ell-1}^n(\theta)\} \mu_{pr}(d\theta) =: \tilde{c}_\ell \] (From Lemma 5.9).

We can then write the desynchronization probability at the \((n+1)\)th step as
\[ P_{n+1} = P_{\nu_\ell}(\Theta_{\ell,\ell}^{n+1} \neq \Theta_{\ell,\ell-1}^{n+1}) = P_{\nu_\ell}(\Theta_{\ell,\ell}^{n+1} \neq \Theta_{\ell,\ell-1}^{n+1}|\Theta_{\ell,\ell}^n = \Theta_{\ell,\ell-1}^n) P_{\nu_\ell}(\Theta_{\ell,\ell}^n = \Theta_{\ell,\ell-1}^n) \]
\[ + P_{\nu_\ell}(\Theta_{\ell,\ell}^{n+1} \neq \Theta_{\ell,\ell-1}^{n+1}|\Theta_{\ell,\ell}^n \neq \Theta_{\ell,\ell-1}^n) P_{\nu_\ell}(\Theta_{\ell,\ell}^n \neq \Theta_{\ell,\ell-1}^n) \]
\[ \leq Z_{\Delta_{\ell,\ell}}^{-1} s (1 - P_n) + \tilde{c}_\ell P_n \]
\[ \leq s + \tilde{c}_\ell P_n. \]

However, by stationarity, we have that \( P_{n+1} = P_n = P_{\nu_\ell}(\Theta_{\ell,\ell} \neq \Theta_{\ell,\ell-1}) \). Thus, from (5.11) we get that
\[ P_{\nu_\ell}(\Theta_{\ell,\ell-1} \neq \Theta_{\ell,\ell}) \leq \frac{\hat{C} s^{-\alpha \ell}}{1 - \tilde{c}_\ell} = \frac{\hat{C}}{c \int_X \min\{\pi_\ell^n(\theta), \pi_{\ell-1}^n(\theta)\} \mu_{pr}(d\theta)} s^{-\alpha \ell}. \]

Notice that by (5.10) the integral term in the denominator is lower bounded by a constant \( \delta \) independent of the level. Furthermore, this integral converges to 1 as \( \ell \to \infty \).

We are now ready to prove implication T2.

**Lemma 5.11.** Suppose Assumptions C and A.2 hold. Then, for any \( \ell \geq 1 \), there exists a positive constant \( C_v \) such that
\[ \mathbb{V}_{\nu_\ell}[Y_{\ell}] \leq C_v s^{-\beta \ell}. \]
where $\beta = \min\{2\alpha_q, \alpha(1 - 2/m)\}$, $\alpha$, $\alpha_q$, $m$ as in Assumption C, and where we have once again used the notation $V_{\nu_l}[Y_l] = V_{\Theta_l \sim \nu_l}[Y_l(\Theta_l)]$.

Proof. Write $\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell}) - \mathcal{L}_{l-1}(\Theta_{l,\ell-1}))^2]$ as $\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell}) - \mathcal{L}_{l-1}(\Theta_{l,\ell-1}))^2]$. We follow a similar argument to that of [9, Lemma 4.8]. From Young’s inequality we have

$$\mathbb{E}_{\nu_l}[V_{\nu_l}[Y_l]] \leq 2\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell}) - \mathcal{L}_{l-1}(\Theta_{l,\ell-1}))^2] + 2\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell-1}) - \mathcal{L}_{l-1}(\Theta_{l,\ell-1}))^2].$$

Notice that in the case where $\mathcal{L}(\cdot)$ and $\mathcal{L}_{l-1}(\cdot)$ are the same (which could happen when the quantity of interest, seen as a functional, is mesh-independent), the second term vanishes. Otherwise, we have, using Assumption C.2, that

$$\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell-1}) - \mathcal{L}_{l-1}(\Theta_{l,\ell-1}))^2] \leq 2\tilde{C}_q^2(1 + s^{2\alpha_q})s^{-2\alpha_q\ell}.$$

As for the first term, notice that it is only nonzero whenever $\Theta_{l,\ell} \neq \Theta_{l,\ell-1}$. Thus, it can be rewritten as

$$2\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell}) - \mathcal{L}(\Theta_{l,\ell-1}))^2] = 2\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell}) - \mathcal{L}(\Theta_{l,\ell-1}))^21_{\{\Theta_{l,\ell} \neq \Theta_{l,\ell-1}\}}].$$

In turn, applying Hölder’s inequality, we can further write the above expression as

$$2\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell}) - \mathcal{L}(\Theta_{l,\ell-1}))^21_{\{\Theta_{l,\ell} \neq \Theta_{l,\ell-1}\}}] \leq 2\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell}) - \mathcal{L}(\Theta_{l,\ell-1}))^2]^{2/m} \mathbb{E}_{\nu_l}[1_{\{\Theta_{l,\ell} \neq \Theta_{l,\ell-1}\}}]^{1/m'} \quad (\text{with } m' = m/(m-2))$$

$$= 2\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell}) - \mathcal{L}(\Theta_{l,\ell-1}))^2]^{2/m} \mathbb{P}_{\nu_l}(\Theta_{l,\ell} \neq \Theta_{l,\ell-1})^{1/m'}.$$  

From Assumption C.2(c), it follows that one can bound the first term in (5.12) by

$$\mathbb{E}_{\nu_l}[(\mathcal{L}(\Theta_{l,\ell}) - \mathcal{L}(\Theta_{l,\ell-1}))^2]^{2/m} \leq \left(\mathbb{E}_{\nu_l}[\mathcal{L}(\Theta_{l,\ell})^2]^{2/m} + \mathbb{E}_{\nu_l}[\mathcal{L}(\Theta_{l,\ell-1})^2]^{2/m}\right)^2 \leq 4c_l^{-2/m}C_m^{-2}.$$  

Moreover, from Lemma 5.10, we have that $\mathbb{P}_{\nu_l}(\Theta_{l,\ell} \neq \Theta_{l,\ell-1}) \leq C_{r,\ell}s^{-\alpha\ell}$. Thus,

$$\mathbb{E}_{\nu_l}[Y_l] \leq C_{e}s^{-\beta\ell},$$

where $C_{e} = 8c_l^{-2/m}C_m^{-2}\max\{C_{r,\ell}\} + 4\tilde{C}_q^2(1 + s^{2\alpha_q}).$

6. Implementation. We begin with a discussion on how to choose the optimal number of samples $N_{\ell}$. Recall that, for $\ell = 0, \ldots, L$, we denote by $C_{\ell}$ the total cost of evaluating
ANALYSIS OF ML-MCMC ALGORITHMS BASED ON IMH

(\mathcal{D}_{\ell-1}, \mathcal{D}_\ell) at level \ell using Algorithm 3.2. The total cost of the ML-MCMC estimator is then given by

$$C \left( \widehat{\mathcal{D}}_{L \{N_L\}} \right) = \sum_{\ell=0}^{L} C_\ell N_\ell. \quad (6.1)$$

In order to bound the statistical contribution of the total error bound, we will require, from (4.8) and (5.1), that

$$2(L + 1) \sum_{\ell=0}^{L} C_{\text{mse}} \frac{\mathbb{V}_{\nu} [Y_\ell]}{N_\ell} \leq \frac{\text{tol}^2}{2},$$

where \text{tol} is some user-prescribed tolerance. However, it is in general not a simple task to compute or estimate the constant \( C_{\text{mse}} \). We will ignore it hereafter and aim at bounding the quantity

$$2(L + 1) \sum_{\ell=0}^{L} \frac{\mathbb{V}_{\nu} [\hat{Y}_\ell]}{N_\ell} \leq \frac{\text{tol}^2}{2}. \quad (6.2)$$

To that end we will use the so-called batched means estimator of \( \mathbb{V}_{\nu} [Y_\ell] \) denoted by \( \hat{\sigma}_\ell^2 \) (see [12] for further details). In this case, treating \( N_\ell \) as a real number and minimizing (6.1) subject to (6.2), gives the optimal samples sizes

$$N_\ell = \left[ 2 \text{tol}^{-2} \sqrt{\hat{\sigma}_\ell^2 / C_\ell} \left( \sum_{j=0}^{L} \sqrt{\hat{\sigma}_j^2 / C_j} \right) \right], \quad (6.3)$$

where \([ \cdot ]\) is the ceiling function. Lastly, we also need to ensure that the second contribution to the total error, i.e., the discretization bias at level L, is such that

$$\sqrt{2} \left| \mathbb{E}_{\mu_L^*}[\mathcal{D}_L] - \mathbb{E}_{\mu^*}[\mathcal{D}] \right| \leq \frac{\text{tol}}{\sqrt{2}}.$$

Notice that from T1 it follows

$$\left| \mathbb{E}_{\mu_L^*}[\mathcal{D}_L] - \mathbb{E}_{\mu^*}[\mathcal{D}] \right| = \sum_{j=L+1}^{\infty} \left( \mathbb{E}_{\mu_j^*}[\mathcal{D}_j] - \mathbb{E}_{\mu_{j-1}}[\mathcal{D}_{j-1}] \right) \approx \frac{[\widehat{\mathcal{D}}_L - \widehat{\mathcal{D}}_{L-1}]}{1 - \alpha_{\omega}}.$$

Thus, to achieve a total (estimated) MSE of the ML-MCMC estimator less than \( \text{tol}^2 \), we need to check that

$$2(L + 1) \left( \sum_{\ell=0}^{L} \frac{\hat{\sigma}_\ell^2}{N_\ell} \right) + 2 \left( \frac{[\widehat{\mathcal{D}}_L - \widehat{\mathcal{D}}_{L-1}]}{1 - \alpha_{\omega}} \right)^2 \leq \text{tol}^2.$$
In practice, the set of parameters $\mathcal{P} := \{C_w, \alpha_w, \{\sigma^2_{\ell}\}_{\ell=0}^L, C_\sigma, \beta, C_\gamma, \gamma\}$ need to be estimated with a preliminary run over $L_0$ levels, using $N_\ell$, $\ell = 0, 1, \ldots, L_0$ samples per level. However, the main disadvantage of this procedure is that for computationally expensive problems, this screening phase can be quite inefficient. In particular, if $L_0$ is chosen too large, then the screening phase might turn out to be more expensive than the overall ML-MCMC simulation on the optimal hierarchy $\{0, 1, \ldots, L\}$. On the other hand, if $L_0$ (or $N_\ell$) is chosen too small, the extrapolation (or estimation) of the values of $\mathcal{P}$ might be quite unreliable, particularly at higher levels. In the MLMC literature, one way of overcoming these issues is with the so-called continuation MLMC method [6]. We will present a continuation-type ML-MCMC (C-ML-MCMC) algorithm in the following subsection, based on the works [6, 37].

**6.1. A C-ML-MCMC.** The key idea behind this method is to iteratively implement a ML-MCMC algorithm with a sequence of decreasing tolerances while, at the same time, progressively improving the estimation of the problem dependent parameters $\mathcal{P}$. As presented before, these parameters directly control the number of levels and sample sizes. Following [6], we introduce the family of tolerances $\text{tol}_i$, $i = 0, 1, \ldots$, given by

$$\text{tol}_i = \begin{cases} r_1^{i-1}r_2^{-1}\text{tol}, & i < i_E, \\ r_2^{i-1}\text{tol}, & i \geq i_E, \end{cases}$$

where $r_1 \geq r_2 > 1$, so that $\text{tol}_{i_E-1} \geq \text{tol} > \text{tol}_{i_E}$, with

$$i_E := \left\lceil -\frac{\log(\text{tol}) + \log(r_2) + \log(\text{tol}_0)}{\log(r_1)} \right\rceil. \tag{6.4}$$

The idea is then to iteratively run the ML-MCMC algorithm for each of the tolerances $\text{tol}_i$, $i = 0, 1, \ldots$ until the algorithm achieves convergence, based on the criterion defined in the previous subsection. Iterations for which $i < i_E$ are used to obtain increasingly more accurate estimates of $\mathcal{P}$. Notice that when $i = i_E$, the problem is solved with a slightly smaller tolerance $r_2^{-1}\text{tol}$ for some carefully chosen $r_2$. Solving at this slightly smaller tolerance is done in order to prevent any extra unnecessary iterations due to the statistical nature of the estimated quantities. Furthermore, if the algorithm has not converged at the $i_E$th iteration, it keeps running for even smaller tolerances $\text{tol}_i$, $i > i_E$, to account for cases where the estimates of $\mathcal{P}$ are unstable. Thus, at the $i$th iteration of the C-ML-MCMC algorithm, we run the Algorithm 3.2 with an iteration-dependent number of levels $L_i$, where $L_i$ is obtained by solving the following discrete constrained optimization problem:

$$\begin{align*}
\arg\min_{L_{i-1} \leq L \leq L_{\text{max}}} & \left\{ 2\text{tol}_i^{-2}2(L + 1) \left( \sum_{j=0}^{L_i} \sqrt{C_{\beta}s^{-\alpha_{\beta}}C_j} \right)^2 \right\} \\
\text{s.t.} & \quad C_w s^{-\alpha_w L_i} \leq \frac{\text{tol}_i}{\sqrt{2}}. \tag{6.5}
\end{align*}$$

Here, $L_{-1} = L_0$ is a given minimum number of levels, $L_{\text{max}}$ is chosen as the maximum number of levels given a computational budget (which could be dictated, for example, by the minimum mesh size imposed by memory or computational restrictions). Furthermore, notice that (6.5) is easily solved by exhaustive search.

We now have everything needed to implement the C-ML-MCMC algorithm, which we present in Algorithm 6.1.
Algorithm 6.1 C-ML-MCMC.

1: procedure C-ML-MCMC(\{\pi_y^{\ell}\}_{\ell=0}^L, Q, N, L_0, L_{\max}, \{\nu_y^{\ell}\}_{\ell=0}^L, tol_0, tol_1, r_1, r_2)
2:     # Preliminary run
3:     Compute \(i_E\) according to (6.4). Set \(N_\ell = \tilde{N}_\ell, \ell = 0, 1, \ldots, L_0\).
4:     \{\{\Theta_n^{\ell}\}_{n=0}^{\tilde{N}_\ell} \sim \mathcal{N}(\theta_0, 1)\}_{\ell=0}^{L_0} = \text{ML-MCMC (\{\pi_y^{\ell}\}_{\ell=0}^L, Q, \{N_\ell\}_{\ell=0}^L, \lambda_0)}.
5:     Compute estimates for the parameters \(\mathcal{P}\) using least squares fit
6:     set \(i = 1\) and \(t_E = \infty\).
7:     # Starts continuation algorithm
8:     while \(i < i_E\) or \(t_E > tol\) do
9:         Update tolerance \(tol = tol_{i-1}/r_k\), where \(k = 1\) if \(i < i_E\) and \(k = 2\) otherwise.
10:        Compute \(L_i = L_{i-1}(L_{\max}, tol_i, \mathcal{P})\) using (6.5)
11:        Compute \(N_\ell = N_\ell(L_i, tol_i, \mathcal{P})\) for \(\ell = 0, 1, \ldots, L_i\), using (6.3).
12:        Q can be constructed using samples from previous iterations
13:        \{\{\Theta_n^{\ell}\}_{n=0}^{N_\ell} \sim \mathcal{N}(\theta_0, 1)\}_{\ell=0}^{L_i} = \text{ML-MCMC (\{\pi_y^{\ell}\}_{\ell=0}^L, Q, \{N_\ell\}_{\ell=0}^L, \lambda_0)}
14:        Update estimates for \(\mathcal{P}\) using least squares fit
15:        Update total error \(t_E = 2(L + 1)(\sum_{\ell=0}^{L_i} \tilde{\sigma}^2_\ell/N_\ell) + 2(C_w s_{-a_{\ell}})^2\)
16:        \(i = i + 1\)
17:     end while
18:     Return \(\hat{\mathcal{S}}_{L_i}(N_\ell)_{\ell=0}^{L_i}\) computed with (3.1).
19: end procedure

7. Numerical experiments. We present the first two “sanity check” experiments aimed at numerically verifying the theory presented in previous sections.

In the following two experiments we will compare our proposed ML-MCMC algorithm to that of [9], which, by construction, does not satisfy our Assumption A.1. The aim of these experiments is to verify the theoretical results of the previous sections, as well as to provide a setting for which our methods might be better suited than the subsampling approach of [9]. For ease of exposition, we will consider as a quantity of interest \(D(\Theta) = \Theta - \Theta_{\mu}\), and we will assume that the cost of evaluating the posterior density at each level grows as \(2^\ell\), with \(\gamma = 1\). For both experiments, we implement the subsampling ML-MCMC algorithm of [9] with a level-dependent subsampling rate \(t_\ell := \min(1 + 2\sum_{k=0}^{N_\ell} \hat{\theta}_k, 5)\), where \(\hat{\theta}_k\) is the so-called log-k auto-correlation time and \(1 + 2\sum_{k=0}^{N_\ell} \hat{\theta}_k\) is the so-called integrated auto-correlation time [4].

7.1 Nested Gaussians. We begin with a scenario for which both our ML-MCMC method and that of [9] can be applied. In this case we aim at sampling from the family of posteriors \(\mu_y^{\ell} = \mathcal{N}(1, 1 + 2^{-\ell}), \ell = 0, 1, 2, \ldots, \) which approximate \(\mu^y = \mathcal{N}(1, 1)\) as \(\ell \to \infty\). For the ML-MCMC method proposed in the current work, we will use a fixed proposal across all levels given by \(Q_\ell = Q = \mathcal{N}(1, 3)\). Such proposal is chosen to guarantee that Assumption A.1 is fulfilled. The family of posteriors and the proposal \(Q\) used in our ML-MCMC algorithm are depicted in Figure 2. For both algorithms, the proposal distribution at level \(\ell = 0\) is a random walk Metropolis (RWM) proposal \(Q_0(\theta_0^{\ell}, \cdot) = \mathcal{N}(\theta_0^{\ell}, 1)\), provided that at the nth iteration the Markov chain is at a state \(\Theta^y = \theta^n\). This proposal is chosen to guarantee an acceptance rate of about 40%, the value deemed close to optimum for MCMC in one dimension [4].
Figure 2. Family of posteriors $\mu^\nu_\ell$ and fixed proposal distribution $Q$ for the nested Gaussians example.

Figure 3. True posterior $\mu^\nu_\ell$ for different levels $\ell = 0, 3, 6$ and histogram of the samples of $\theta_\ell \sim \mu^\nu_\ell$ obtained with the ML-MCMC algorithm described herein with $Q_\ell = \mathcal{N}(1, 3)$ (top row) and the subsampling ML-MCMC algorithm (bottom row). As we can see, both methods are able to obtain samples from the right posterior distribution.

As a sanity check, we begin by verifying that both algorithms target the right marginal distribution at different levels. This can be seen in Figure 3, where the histograms of samples obtained with a simple ML-MCMC algorithm with proposal $Q$ and prescribed number of levels $L = 7$ and number of samples $N_\ell = 50000$ for $\ell = 0, 1, \ldots, L$ (top row) and the algorithm of [9] (bottom row) are shown for levels $\ell = 0, 3, 6$. The true posterior at level $\ell$ is shown in red. As it can be seen, both methods can sample from the right marginal distribution for the family of posteriors considered herein.

We now aim at verifying the rates presented in Theorem 5.2. To that end, we run the ML-MCMC algorithm 100 independent times. For each independent run, we obtained 50,000 samples on each level and investigate the behavior of $|\mathbb{E}_{\nu_\ell}[Y_\ell]|$ (Figure 4(left)) and $\mathbb{V}_{\nu_\ell}[Y_\ell]$ (Figure 4(right)) with respect to the level $\ell$. As it can be seen from Figure 4, both $|\mathbb{E}_{\nu_\ell}[Y_\ell]|$ and $\mathbb{V}_{\nu_\ell}[Y_\ell]$ decay with respect to $\ell$ with nearly the same estimated rate $\approx -1.34$ for the ML-MCMC algorithm discussed in this work, close to the predicted one in Theorem 5.2. It can be
ANALYSIS OF ML-MCMC ALGORITHMS BASED ON IMH

Figure 4. (Left) $E_{\nu_\ell}[Y_\ell]$ versus level. (Right) $\nu_\ell[Y_\ell]$ vs. level. In both figures, the rates were estimated over 100 independent runs, with 50,000 samples per level, on each run. Solid lines indicate the average value, dashed lines indicate 95% confidence intervals.

Figure 5. (Left) Number of samples versus level for both algorithms. (Right) Synchronization rate versus level for both algorithms.

seen, however, from Figure 4(right), that the variance decay of the subsampling algorithm is slightly better than the one obtained by the method presented herein. This, in turn, results in a smaller overall sample size at each level for a given particular error tolerance, as it can be seen in Figure 5(left). We believe that this difference in rate is due to (i) the slightly higher synchronization rate of the subsampling ML-MCMC algorithm (Figure 5(right)) and (ii) the fact that the convergence rate of the marginal chain in the subsampling algorithm also increases with level, which is not necessarily the case for our method. These results suggest that, for this particular case, it is more cost-efficient to use the subsampling ML-MCMC algorithm.
We plot sample size vs level (Figure 5(left)) and synchronization rate versus level (Figure 5(right)). Both figures were obtained from 100 independent runs: solid lines indicate the average value, and dashed lines indicate 95% confidence intervals. The computation of $N_{\ell}$ for each level $\ell = 0, 1, \ldots, L$ was done by estimating $\hat{\sigma}_{\ell}$ with 50,000 samples per level and a tolerance $\text{tol} = 0.07$. It can be seen from Figure 5(left) that the subsampling algorithm requires a smaller number of samples per level. From Figure 5(right) we can see that both algorithms tend to have a synchronization rate of 1, as expected. It can be seen that the subsampling algorithm provides a slightly higher synchronization rate for the problem at hand.

Lastly, we perform some robustness experiments for our C-ML-MCMC algorithm. To that end, we run Algorithm 6.1 using the same level independent proposals $Q_{\ell} = Q = \mathcal{N}(2, 3)$ for three different prescribed tolerances $\text{tol} = \{0.025, 0.05, 0.1\}$. The algorithm is run for a total of 100 independent times. At each run $k$, we compute the total squared error of the ML estimator obtained from the $k$th run of the C-ML-MCMC algorithm given by

\[
er^2_k := \left( \hat{\mathcal{Y}}(\mathcal{L}_{\ell=0}^{\infty} - \mu^y(\mathcal{Q})) \right)^2
\]

and plot it in Figure 6. As we can see, we obtain estimators whose MSE is less than the prescribed tolerance, as desired. This evidences the robustness of Algorithm 6.1 when computing quantities of interest for a given tolerance.

**7.2. Shifting Gaussians.** We now move to a slightly more challenging problem, which is better suited for our proposed method. In this case, we aim at sampling from the family of posteriors $\mu^y_{\ell} = \mathcal{N}(2^{-\ell+2}, 1)$, $\ell = 0, 1, 2, \ldots, L$, which approximate $\mu^y = \mathcal{N}(0, 1)$ as $\ell \to \infty$. Once again, for the ML-MCMC method proposed in the current work, we will use a fixed proposal across all levels given by $Q_{\ell} = Q = \mathcal{N}(2, 3)$. Such a proposal is chosen to guarantee that Assumption A.1 is fulfilled. The posterior and proposal densities are shown in Figure 7. Just
as in the experiment presented in section 7.1, the proposal distribution at level $\ell = 0$ for both algorithms is a RWM proposal $Q_0(\theta_0^n, \cdot) = \mathcal{N}(\theta_0^n, 1)$. This proposal is chosen to guarantee an acceptance rate of about 40%. Once again, we begin by investigating the correctness of the corresponding marginals. To that end we run both algorithms for $L = 6$, obtaining 50,000 samples per level, and plot the resulting histograms of $\mu y_{\ell}$ for levels $\ell = 0, 3, 6$. Such results are presented in Figure 8. As it can be seen, the ML-MCMC presented herein (Figure 8(top row)) can sample from the correct marginals. On the contrary, the subsampling ML-MCMC algorithm is not able to produce samples from the correct distributions, at least for the number of samples considered, as it can be seen in Figure 8(bottom row). We believe that this is due to Assumption A.1 not being satisfied since there is a very small overlap between the posterior at level 0 and the posteriors at higher levels. Sampling from the wrong marginal distribution will in turn result in biased estimators when using the subsampling method [9].

We now proceed to verify the converge rates stated in Theorem 5.2. Notice that for this particular setting we have $|E_{\nu^0}[\mathcal{L}_\ell] - E_{\nu^\ell}[\mathcal{L}]| = 2^{-\ell+1}$. We run Algorithm 3.2 100
Figure 9. (Top left) Estimated expected value of $\mathbb{E}_{\nu_{\ell}}[Q_{\ell}]$ for both ML-MCMC algorithms and the true mean of $\mathbb{E}_{\nu_{\ell}}[Q_{\ell}]$ for different values of $\ell$. (Top right) Expected value of $Y_{\ell} = Q_{\ell} - Q_{\ell-1}$ obtained with both algorithms for different values of $\ell$. (Bottom left) Variance of $Y_{\ell}$ obtained with both algorithms for different values of $\ell$. (Bottom right) Number of samples per level for each method with $\text{tol} = 0.07$. On all plots, dashed lines represent a 95% confidence interval estimated over 100 independent runs of each algorithm.

independent times, obtaining 50,000 samples on each level for every run. The accuracy of the theoretical rates in Theorem 5.2 is numerically verified in Figure 9. However, as it can be seen in Figure 9(top left) the sample mean of $\mathbb{E}_{\nu_{\ell}}[Q_{\ell}]$ obtained with the subsampling algorithm does not decay as $2^{-\ell}$, which confirms the bias of the subsampling ML-MCMC algorithm. The decay rates $\alpha_{w}$ and $\beta$, corresponding to the decay in weak and strong error, respectively, for the ML-MCMC algorithm with fixed proposals, are verified to be 1, as theoretically expected, in Figure 9(top right and bottom left). The optimal number of samples per level is presented in Figure 9(bottom left). Once again, the subsampling ML-MCMC provides a smaller number of samples and variances than the method presented herein, however, at the price of a biased estimator. Furthermore, it can be seen from Figure 10 that the synchronization rate of both methods tends to 1 with $\ell$, as expected.
Lastly, we once again perform some robustness experiments for our C-ML-MCMC algorithm. To that end, we run Algorithm 6.1 using the same level independent proposals $Q_\ell = Q = \mathcal{N}(2, 3)$ for three different prescribed tolerances $\text{tol} = \{0.1, 0.07, 0.06\}$ for a total of 100 independent runs. Similar as in the previous example, for each independent run $k$ of the C-ML-MCMC algorithm, we compute $\text{er}_k^2$ as in (7.1) and plot it in Figure 11. Once again, we obtain estimators whose MSE is close to the prescribed tolerance, as desired. This further evidences the robustness of Algorithm 6.1.
7.3. Subsurface flow. We consider a slightly more challenging problem for which we try to recover the probability distribution of the stochastic permeability field in Darcy’s subsurface flow equation (7.2), based on some noise-polluted measured data. In particular, let $D = [0, 1]^2$, $X = \mathbb{R}^4$, $(x_1, x_2) =: x \in D$, $\partial D = \Gamma_N \cup \Gamma_D$, with $\Gamma_N \cap \Gamma_D = \emptyset$, where $\Gamma_D := \{(x_1, x_2) \in \partial D, \text{s.t. } x_1 = \{0, 1\}\}$, and $\Gamma_N = \partial D \setminus \Gamma_D$. Darcy’s subsurface equation is given by

$$
\begin{aligned}
-\nabla_x \cdot (\kappa(x, \theta) \nabla_x u(x, \theta)) = 1, & \quad x \in D, \theta \in X, \\
u(x, \theta) = 0, & \quad x \in \Gamma_D, \theta \in X, \\
\partial_n u(x, \theta) = 0, & \quad x \in \Gamma_N, \theta \in X,
\end{aligned}
$$

(7.2)

where $u$ represents the pressure (or hydraulic head), and we model the stochastic permeability $\kappa(x, \theta)$ for $(\theta_1, \theta_2, \theta_3, \theta_4) =: \theta \in X$ as

$$
\kappa(x, \theta) = \exp \left( \theta_1 \cos(\pi x) + \frac{\theta_2}{2} \sin(\pi x) + \frac{\theta_3}{3} \cos(2\pi x) + \frac{\theta_4}{4} \sin(2\pi x) \right),
$$

with $\theta_i \sim \mathcal{N}(0, 1)$, $i = 1, 2, 3, 4$. Data $y$ is modeled by the solution of (7.2) observed at a grid of $9 \times 9$ equally spaced points in $D$ (hence $Y = \mathbb{R}^{9 \times 9}$) and polluted by a normally distributed noise $\eta \sim \mathcal{N}(0, \sigma_{\text{noise}}^2 I_{81 \times 81})$, with $\sigma_{\text{noise}} = 0.004$, which corresponds to approximately 1% noise and $I_{81 \times 81}$ is the 81-dimensional identity matrix. At each discretization level $\ell \geq 0$, the solution to (7.2) is numerically approximated using the finite element method on a triangular mesh of $2^\ell \times 16 \times 2^\ell \times 16$ elements, which is computationally implemented using the FEniCS library [29]. Such a library includes optimal solvers for the forward model, for which $\gamma$ can be reasonably taken equal to 1. Thus, the map $\theta \mapsto \mathcal{F}_\ell(\theta)$ is to be understood as the numerical solution of (7.2) at a discretization level $\ell$, observed at a grid of $9 \times 9$ equally spaced points, for a particular value of $\theta \in X$. This, in turn, induces a level dependent potential

$$
\Phi_\ell(\theta; y) := \frac{1}{2\sigma_{\text{noise}}^2} \| y - \mathcal{F}_\ell(\theta) \|^2
$$

and prior $\mu_{\text{pr}} = \mathcal{N}(0, I_{4 \times 4})$. In the above expressions, $\| \cdot \|$ denotes the Frobenius norm on $\mathbb{R}^{9 \times 9}$. Given that we are on a finite-dimensional setting, $\mu_{\text{pr}}$ has a density with respect to the Lebesgue measure, and as such, we can define the unnormalized posterior density $\tilde{\pi}_\ell^y : X \rightarrow \mathbb{R}_+$ w.r.t the Lebesgue measure given by

$$
\tilde{\pi}_\ell^y(\theta) = \exp \left( -\Phi_\ell(\theta; y) - \frac{1}{2} \theta^T \theta \right).
$$

As a quantity of interest we consider the average pressure over the physical domain, that is, $\mathcal{Q}(\theta) = \int_D u(x, \theta) \, dx$. We implement our ML-MCMC algorithm to approximate $\mathbb{E}_{\mu_{\text{pr}}}[\mathcal{Q}]$. In particular, we use RWM at level 0 with Gaussian proposals $\mathcal{N}(0, \sigma_{\text{rwm}}^2 I_{4 \times 4})$ with step-size $\sigma_{\text{rwm}} = 0.05$, which produces an acceptance rate of about 24%. For the proposal $Q_\ell$ at higher levels $\ell \geq 1$, we use a mixture between the prior and a kernel density approximation (KDE) obtained from the samples obtained at the previous level $\ell - 1$. This choice of the mixture is made so that Assumption A holds.

We begin by numerically verifying the converge rates stated in Theorem 5.2. To that end, we run Algorithm 3.2 20 independent times, obtaining 10,000 samples per run at each level.
\(\ell = 0, 1, 2, 3\). We plot the obtained rates in Figure 12. As we can see, we numerically verify that \(\alpha_w = \beta = 2.0\), as predicted by our theory.

Lastly, we once again perform some robustness experiments for our C-ML-MCMC algorithm, with \(L_{\text{max}} = 3\). To that end, we first estimate \(\mathbb{E}_{\nu}[\mathcal{L}] \approx \mu_4^y(\mathcal{D}_4)\) by performing 50 independent runs of a single-level MCMC algorithm at a discretization level \(\ell = 4\), obtaining 2000 samples on each simulation. In particular, each independent run implements a RWM sampler, using proposals given by \(\mathcal{N}(0, \sigma_{\text{rwm}}^2 I_{4 \times 4})\) with step-size \(\sigma_{\text{rwm}} = 0.05\), which produces an acceptance rate of about 21\%. We run Algorithm 6.1 using the same mixture of independent proposals as before for different tolerance levels \(\text{tol} = \{1.1 \times 10^{-4}, 2.0 \times 10^{-4}, 3.0 \times 10^{-4}\}\). The C-ML-MCMC algorithm is run 20 independent times for each tolerance \(\text{tol}_i\). For each independent run \(k = 1, 2, \ldots, 20\), let \(\hat{\mathcal{L}}^{(k)}_{\text{tol}_i} = \{N_{\ell}\}_{\ell=0}^{L(\text{tol}_i)}\), with \(L(\text{tol}_i) \leq 3\), denote the ML estimator obtained from the \(k\)th run at tolerance \(\text{tol}_i\). We compute the (approximate) total error squared \(\tilde{e}_{i,k}^2\) at the \(k\)th run with a tolerance \(\text{tol}_i\) as

\[
\tilde{e}_{i,k}^2 = \left( \hat{\mu}_4^y(\mathcal{D}_4) - \mu_4^y(\mathcal{D}_4) \right)^2
\]

and plot it versus a given tolerance in Figure 13. As expected, the MSE of the obtained estimators is less than the prescribed tolerance. This further evidences the robustness of Algorithm 6.1.

8. Conclusions and future outlook. In the current work we have presented a general ML-MCMC framework based on independent MH proposals, which can be understood as an extension of currently existing methods [9]. Furthermore, we have provided what is, to the best of the authors knowledge, the first thorough study on the theory of ML-MCMC algorithms based on independent proposals; we have investigated necessary conditions for the existence

\[\ell = 0, 1, 2, 3\]. We plot the obtained rates in Figure 12. As we can see, we numerically verify that \(\alpha_w = \beta = 2.0\), as predicted by our theory.

Lastly, we once again perform some robustness experiments for our C-ML-MCMC algorithm, with \(L_{\text{max}} = 3\). To that end, we first estimate \(\mathbb{E}_{\nu}[\mathcal{L}] \approx \mu_4^y(\mathcal{D}_4)\) by performing 50 independent runs of a single-level MCMC algorithm at a discretization level \(\ell = 4\), obtaining 2000 samples on each simulation. In particular, each independent run implements a RWM sampler, using proposals given by \(\mathcal{N}(0, \sigma_{\text{rwm}}^2 I_{4 \times 4})\) with step-size \(\sigma_{\text{rwm}} = 0.05\), which produces an acceptance rate of about 21\%. We run Algorithm 6.1 using the same mixture of independent proposals as before for different tolerance levels \(\text{tol} = \{1.1 \times 10^{-4}, 2.0 \times 10^{-4}, 3.0 \times 10^{-4}\}\). The C-ML-MCMC algorithm is run 20 independent times for each tolerance \(\text{tol}_i\). For each independent run \(k = 1, 2, \ldots, 20\), let \(\hat{\mathcal{L}}^{(k)}_{\text{tol}_i} = \{N_{\ell}\}_{\ell=0}^{L(\text{tol}_i)}\), with \(L(\text{tol}_i) \leq 3\), denote the ML estimator obtained from the \(k\)th run at tolerance \(\text{tol}_i\). We compute the (approximate) total error squared \(\tilde{e}_{i,k}^2\) at the \(k\)th run with a tolerance \(\text{tol}_i\) as

\[
\tilde{e}_{i,k}^2 = \left( \hat{\mu}_4^y(\mathcal{D}_4) - \mu_4^y(\mathcal{D}_4) \right)^2
\]

and plot it versus a given tolerance in Figure 13. As expected, the MSE of the obtained estimators is less than the prescribed tolerance. This further evidences the robustness of Algorithm 6.1.

8. Conclusions and future outlook. In the current work we have presented a general ML-MCMC framework based on independent MH proposals, which can be understood as an extension of currently existing methods [9]. Furthermore, we have provided what is, to the best of the authors knowledge, the first thorough study on the theory of ML-MCMC algorithms based on independent proposals; we have investigated necessary conditions for the existence
and convergence to a unique invariant probability measure for the methods discussed herein, as well as presented sufficient conditions under which the cost-tolerance result presented in [9] can be generalized to our method. Furthermore, we have verified our theoretical results on three academic examples. Lastly, we have presented a nonasymptotic bound on the statistical MSE in the spirit of [40] under minimal assumptions on the Markov transition operator. We remark that such a bound can be applied outside the context of ML-MCMC.

We intend to carry out a number of future extensions of the work presented herein, both from a theoretical and computational perspective. On the one hand, from a theoretical perspective, there are still some unanswered questions regarding the class of methods at hand, and a further analysis of the spectral properties of $P_\ell$ would be an interesting contribution. In addition, given the unattainable nature of the invariant joint probability measure of the ML-MCMC algorithm, it is also unclear at this point how such a measure depends on the choice of proposal.

On the other hand, from a computational perspective, one future research direction stems from the construction of the level $\ell$ proposal $Q_\ell$. Intuitively, such a proposal should resemble the posterior $\mu_{\ell-1}^x$ or $\mu_{\ell}^y$ (or a convex combination of both) as much as possible, while at the same time satisfying Assumption A.1. Some preliminary experiments suggest that density approximation techniques, such as KDE [41] or, more recently, flow-based generative models [34], can be used to construct efficient ML-MCMC proposals. Furthermore we would also be interested in extending these results to a multi-index setting as in [15].

Furthermore, we argue that our IMH setting can help overcome issues associated with multimodality. Indeed, provided that one has a sufficiently good proposal distribution (i.e., an accurate approximation of the posteriors at levels $\ell, \ell - 1$, satisfying Assumption A.1), multimodality is less of an issue in the IMH case than in other MH algorithms, since (a) the proposed state of either chain does not depend on their current state (contrary to diffusion-based proposals, such as RWM), and (b) the proposed state gets accepted by both chains with some nonnegligible probability, thus synchronizing the chains in a few steps. Of course,
Assumption A.1 implies that the proposal has nonnegligible mass around each of the modes; hence, in a sense, it should respect the multimodality of the posterior. On a similar line of thought, parallel tempering methods (see, e.g., [25]) which are highly used to sample multimodal distributions, can be understood as applying an IMH proposal in their swapping step. This is in contrast to other coupling techniques, such as reflection maximal coupling [19], where, even for non-Metropolized algorithms, multimodality might become catastrophic for the convergence of the method if the dynamics associated with the Markov chain are chaotic or noncontractive (see also [11]). Naturally, one needs to be careful with the construction of the proposal, and we aim at exploring this further by combining our ML-MCMC approach with (G)PT in future work.

Lastly, an additional research direction would be to dig deeper in the combination of ML-MCMC with the unbiased estimators, as in [19]. Some work in this direction has been recently presented in [17], with the multilevel part being carried out as in [14], which is a slightly different approach from ours.

Appendix A. Bounding the statistical MSE. We now present a nonasymptotic bound of the MSE of the ergodic estimator \( \hat{f}_{N,n} = \frac{1}{N} \sum_{i=1}^{N} f(\Theta^n \circ \pi_0) \) obtained with finitely many samples \( \{\Theta^n\}_{n=1}^N \) from a nonreversible Markov chain \( \{\Theta^n\}_{n \in \mathbb{N}} \). We remark that the bound presented in this work has utility beyond the multilevel construction. Throughout this section we will let \((X, \| \cdot \|_X)\) be a separable Banach space with associated Borel \( \sigma \)-algebra \( \mathcal{B}(X) \). We will also let \( P : L_2(X, \nu) \rightarrow L_2(X, \nu) \) be a \( \nu \)-invariant Markov operator for some probability measure \( \nu \) on \((X, \mathcal{B}(X))\). Lastly, we define the averaging operator as \( \bar{\nu} f \coloneqq \nu(f) = \int_X f(\theta) \nu(d\theta) \).

Theorem A.1 (nonasymptotic bound on the MSE). Let \( f \in L_2(X, \nu) \), be a \( \nu \)-square integrable function, and write \( g(\theta) = f(\theta) - \int_X f(\theta) \nu(d\theta) \). Suppose the Markov operator \( P \) is uniformly ergodic and nonnecessarily reversible. In addition, assume the chain generated by \( P \) starts from an initial probability measure \( \nu^0 \ll \nu \), with \( \frac{d\nu^0}{d\nu} \in L_\infty(X, \nu) \). Then,

\[
\text{MSE}(\hat{f}_{N,n} ; \nu^0) = E_{\nu^0,P} \left| \frac{1}{N} \sum_{n=1}^{N} g(\Theta^n \circ \pi_0) \right|^2 \leq \frac{\nu(f)}{N} \left( C_{\text{inv}} + C_{\text{ns}} \right),
\]

where \( C_{\text{inv}} = 1 + \frac{4}{\gamma_{\text{ps}}[P]} \), \( C_{\text{ns}} = \left( 2 \left\| \frac{d\nu^0}{d\nu} - 1 \right\|_{L_\infty} \left( 1 + \frac{4}{\gamma_{\text{ps}}[P]} \right) \right) \), and \( \gamma_{\text{ps}}[P] \) is the pseudospectral gap of \( P \), defined in (4.7).

The proof of Theorem A.1 is decomposed into a series of auxiliary results. We now present a first bound of the form (A.1) for chains which are started at stationarity, i.e., whenever \( \nu^0 = \nu \). Although this is usually not the case, the following lemma will be useful in the proof of Theorem A.1.

Lemma A.2 (MSE bound starting at stationarity). Under the same assumptions as in Theorem A.1 and with \( \nu^0 = \nu \), it holds

\[
\text{MSE}(\hat{f}_{N,n} = 0; \nu) := E_{\nu,P} \left| \frac{1}{N} \sum_{n=1}^{N} g(\Theta^n \circ \pi_0) \right|^2 \leq \frac{\nu(f)}{N} \left( 1 + \frac{4}{\gamma_{\text{ps}}[P]} \right).\]
Proof. We follow a similar approach to those presented in [35, Theorem 3.2] and [39, section 3]. In order to ease notation, for the remainder of this proof we write \( L_q = L_q(X, \nu) \), \( q \in [1, \infty] \). We can write the MSE of a Markov chain generated by \( P \) starting at \( \nu \) as

\[
(A.2) \quad \mathbb{E}_{\nu, P} \left[ \frac{1}{N} \sum_{n=1}^{N} g(\Theta^n) \right]^2 = \frac{1}{N^2} \sum_{n=1}^{N} \mathbb{E}_{\nu, P}[g(\Theta^n)^2] + \frac{2}{N^2} \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \mathbb{E}_{\nu, P}[g(\Theta^i)g(\Theta^j)].
\]

Working on the expectation of the second term on the right-hand side we get from the Cauchy–Schwarz inequality that

\[
\mathbb{E}_{\nu, P}[g(\Theta^i)g(\Theta^j)] = \langle g, P^{i-j} g \rangle_{\nu} = \langle g, (P - \widehat{P})^{i-j} g \rangle_{\nu} \leq \|g\|_{L_2}^2 \|(P - \widehat{P})^{i-j}\|_{L_2 \to L_2}.
\]

Notice that for any \( k \geq 1 \), we have

\[
\|(P - \widehat{P})^{i-j}\|_{L_2 \to L_2} \leq \|(P - \widehat{P})^k\|_{L_2 \to L_2}^{\lfloor \frac{i-j}{k} \rfloor}
\]

where \( \lfloor \cdot \rfloor \) is the floor function. Now, let \( k_{ps} \) be the smallest integer such that

\[
(A.3) \quad k_{ps}\gamma_{ps}[P] = \gamma_2[(P^*)^{k_{ps}} P^{k_{ps}}] = 1 - \|(P^* - \widehat{P})^{k_{ps}} (P - \widehat{P})^{k_{ps}}\|_{L_2 \to L_2},
\]

which is strictly positive for uniformly ergodic chains (see [38, section 3.3]). Then, from (A.2), (A.3), and (A.4), we obtain

\[
\frac{2}{N^2} \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \mathbb{E}_{\nu, P}[g(\Theta^i)g(\Theta^j)] \leq \frac{2}{N^2} \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \|g\|_{L_2}^2 (1 - k_{ps}\gamma_{ps}[P])^{\lfloor \frac{i-j}{k_{ps}} \rfloor}.
\]

For notational simplicity we write \( \varrho = (1 - k_{ps}\gamma_{ps}[P]) \). We then have that

\[
\frac{2}{N^2} \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \|g\|_{L_2}^2 \varrho^{\lfloor \frac{i-j}{k_{ps}} \rfloor} \leq \frac{2}{N^2} \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \|g\|_{L_2}^2 \sum_{m=0}^{\infty} \varrho^{\lfloor \frac{m}{k_{ps}} \rfloor} \leq \frac{2}{N^2} \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \|g\|_{L_2}^2 \sum_{m=0}^{\infty} \varrho^{\frac{m}{k_{ps}}} 
\]

\[
= \frac{2}{N} \|g\|_{L_2}^2 \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \frac{k_{ps}}{1 - \varrho^{1/2}} = \frac{2}{N} \|g\|_{L_2}^2 \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \frac{k_{ps}}{1 - \varrho^{1/2}} = \frac{2}{N} \|g\|_{L_2}^2 \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \frac{k_{ps}}{N} \frac{1 + \varrho^{1/2}}{1 - \varrho} 
\]

where the second inequality comes from the definition of the floor function \( \lfloor \cdot \rfloor \). We now shift our attention to the first term in (A.2). Using Hölder’s inequality with \( q = \infty, q' = 1 \), together with the fact that \( P \) is a weak contraction in \( L_q(X, \nu) \), for any \( q \in [1, \infty] \), we obtain

\[
\frac{1}{N^2} \sum_{n=1}^{N} \mathbb{E}_{\nu, P}[g(\Theta^n)^2] = \frac{1}{N^2} \sum_{n=1}^{N} \langle 1, P^n g^2 \rangle_{\nu} \leq \frac{1}{N^2} \sum_{n=1}^{N} \| P^n g^2 \|_{L_1} \leq \frac{\|g\|_{L_1}^2}{N}.
\]
Lastly, notice that
\[
\|g^2\|_{L^1} = \int_X |g^2(\theta)| \nu(d\theta) = \int_X g^2(\theta) \nu(d\theta) = \|g\|_{L^2}^2.
\]

Hence, we obtain the bound
\[
(A.5) \quad \frac{1}{N^2} \sum_{n=1}^{N} E_{\nu,P}[g(\Theta^n)^2] \leq \frac{\|g\|_{L^2}^2}{N}.
\]

Thus, from (A.2) and (A.5), together with the observation that \(\|g\|_{L^2}^2 = \int_X (f(\theta) - \widehat{\nu}(f))^2 \nu(d\theta) = \mathbb{V}_\nu[f]\), we finally obtain
\[
E_{\nu,P} \left| \frac{1}{N} \sum_{n=1}^{N} g(\Theta^n) \right|^2 \leq \frac{\|g\|_{L^2}^2}{N} \left( 1 + \frac{4}{\gamma_{ps}[P]} \right) + \frac{2}{N^2} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} \mathcal{H}^{j+n_k}(g^2) + \frac{2}{N^2} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} \mathcal{H}^{j+n_k}(gP^{k-j} g),
\]
\[
(A.6)
\]

where
\[
\mathcal{H}^i(h) = \left\langle (P^i - \widehat{\nu}) h, \left( \frac{d\nu^0}{d\nu} - 1 \right) \right\rangle_\nu, \quad i \in \mathbb{N}, \ h \in L^2(X, \nu).
\]

**Proof.** See [40, Proposition 3.29].

We can now prove Theorem A.1.

**Proof of Theorem A.1.** Once again, for the remainder of this proof we write \(L_q = L_q(X, \nu), \ q \in [1, \infty]\). From Lemma A.3 we get
\[
(H.7) \quad \mathcal{H}^{j+n_k}(g^2) = \left\langle (P^{j+n_k} - \widehat{\nu}) g^2, \left( \frac{d\nu^0}{d\nu} - 1 \right) \right\rangle_\nu,
\]
Lemma A.3. Lastly, combining these results with Lemma A.2 and once again observing that (A.10)

Using Hölder’s inequality with \( q' = \infty, q = 1 \) on the right-hand side of (A.7) gives

where the last inequality comes from the definition of operator norm. Moreover, since the Markov operators are weak contractions, we have that \( \| (P_{j+n_0} - \tilde{\nu}) \|_{L_1 \to L_1} \leq 2\gamma j \in \mathbb{N} \), which gives the bound

Summing over \( j \) gives

Following similar procedure for (A.8) we obtain

Furthermore, from Hölder’s inequality (with \( q' = q = 2 \)) and the fact that \( \tilde{\nu}(g) = 0 \) we get

where the last inequality follows from the same pseudo-spectral gap argument used in the proof of Lemma A.2. Adding over \( j \) and \( k \) gives

Notice then that (A.9) and (A.10) provide a bound on the second and third term in Lemma A.3. Lastly, combining these results with Lemma A.2 and once again observing that \( \| g \|_{L_2}^2 = \mathcal{V}_\nu[f] \) gives the desired result

\[
(A.6) \leq \frac{\mathcal{V}_\nu[f]}{N} \left( 1 + \frac{2}{\gamma_{ps}[P]} \right) + \frac{\mathcal{V}_\nu[f]}{N} \left( 2 \left\| \frac{d\nu^0}{d\nu} - 1 \right\|_{L_\infty} \left( 1 + \frac{4}{\gamma_{ps}[P]} \right) \right).
\]
Acknowledgments. The authors would like to extend their gratitude to the anonymous reviewers whose comments significantly improved the quality of this manuscript. The authors would also like to thank Dr. Sebastian Krumscheid, Dr. Panagiotis Tsilifis, and Dr. Anamika Pandey for fruitful discussions during the very early stages of this work.

REFERENCES


