A large deviation principle for the empirical measures of Metropolis–Hastings chains

Citation for the original published paper (version of record):
Milinanni, F., Nyquist, P. (2024). A large deviation principle for the empirical measures of Metropolis–Hastings chains. Stochastic Processes and their Applications, 170.
http://dx.doi.org/10.1016/j.spa.2023.104293

Stochastic Processes and their Applications 170 (2024) 104293

Contents lists available at ScienceDirect

Stochastic Processes and their Applications

journal homepage: www.elsevier.com/locate/spa

A large deviation principle for the empirical measures of Metropolis–Hastings chains

Federica Milinanni [a,*], Pierre Nyquist [b]

[a] KTH Royal Institute of Technology, Department of Mathematics, Stockholm, 100 44, Sweden
[b] Chalmers University of Technology and University of Gothenburg, Department of Mathematical Sciences, Gothenburg, 412 96, Sweden

ARTICLE INFO

MSC:
primary 60F10
60C05
secondary 60G57
60J05

Keywords:
Large deviations
Empirical measure
Markov chain Monte Carlo
Metropolis–Hastings

ABSTRACT

To sample from a given target distribution, Markov chain Monte Carlo (MCMC) sampling relies on constructing an ergodic Markov chain with the target distribution as its invariant measure. For any MCMC method, an important question is how to evaluate its efficiency. One approach is to consider the associated empirical measure and how fast it converges to the stationary distribution of the underlying Markov process. Recently, this question has been considered from the perspective of large deviation theory, for different types of MCMC methods, including, e.g., non-reversible Metropolis–Hastings on a finite state space, non-reversible Langevin samplers, the zig-zag sampler, and parallel tempering. This approach, based on large deviations, has proven successful in analysing existing methods and designing new, efficient ones. However, for the Metropolis–Hastings algorithm on more general state spaces, the workhorse of MCMC sampling, the same techniques have not been available for analysing performance, as the underlying Markov chain dynamics violate the conditions used to prove existing large deviation results for empirical measures of a Markov chain. This also extends to methods built on the same idea as Metropolis–Hastings, such as the Metropolis-Adjusted Langevin Method or ABC-MCMC. In this paper, we take the first steps towards such a large-deviations based analysis of Metropolis–Hastings-like methods, by proving a large deviation principle for the empirical measures of Metropolis–Hastings chains. In addition, we also characterize the rate function and its properties in terms of the acceptance- and rejection-part of the Metropolis–Hastings dynamics.

1. Introduction

Sampling from a given probability distribution is an essential problem in a range of areas, for example biology, physics, epidemiology and ecology, and statistics. The most common approach is Markov chain Monte Carlo (MCMC), which allows the user to sample from a target probability distribution π, by generating an ergodic Markov chain $\{X_i\}_{i\geq 0}$ with π as stationary distribution. These sampling techniques are particularly helpful when it is not possible to use methods that simulate directly from π, for example for computing posterior distributions in a Bayesian setting, or more generally when π is only known up to a normalizing constant. Because of this, MCMC methods are now widely used across scientific disciplines, and are integral tools in areas such as computational chemistry and physics, statistics and machine learning [1,4,41].

Because of their prevalence in a range of fields, the performance of MCMC algorithms has become an important topic within applied probability and computational statistics. In principle, even the standard Metropolis–Hastings algorithm [29,35] can be used to sample from essentially any target distribution π. However, when the underlying problem, and thus the distribution π, becomes more and more complex, convergence speed or the cost per iteration becomes an issue. Analysing and improving the convergence

* Corresponding author.
E-mail addresses: fedmil@kth.se (F. Milinanni), pnyquist@chalmers.se (P. Nyquist).

https://doi.org/10.1016/j.spa.2023.104293
Received 16 May 2023; Received in revised form 20 November 2023; Accepted 31 December 2023
Available online 3 January 2024

speed of a given class of algorithms, as well as comparing the performance of different types of algorithms, is therefore not only interesting from a theoretical perspective, it is also of central importance for applications, where fast and accurate methods are needed for increasingly complex problems.

When analysing performance of MCMC methods, the rate of convergence of time averages is a central quantity for comparing different methods, and for choosing hyperparameters. The fundamental idea underlying MCMC is that for an observable $f \in L^1(\pi)$, for an ergodic Markov chain $\{X_i\}_{i\in\mathbb{N}}$ with invariant distribution π, the n-step average $\frac{1}{n}\sum_{i=0}^{n-1} f(X_i)$ can be used to approximate the expectation $\mathbb{E}_\pi[f(X)]$. This average can be viewed as the integral of f with respect to the empirical measure of the Markov process. The rate of convergence of the empirical measure is therefore directly linked to the performance of a given MCMC method.

Because of the role the empirical measure plays in MCMC, and for Monte Carlo methods in general, in the past decade there has been an increasing interest in using the theory of large deviations for empirical measures to study the performance of MCMC methods [8,9,16,23,37–40]. However, surprisingly, existing large deviation results do not cover the empirical measure arising from the Metropolis–Hastings algorithm [29,35] on a general state space. Thus, in order to use a large deviation approach to analyse this foundational algorithm, or more advanced MCMC methods built on the same ideas as Metropolis–Hastings—such as the Metropolis-Adjusted Langevin Method (MALA) [7,43,45] and methods based on Approximate Bayesian Computation (ABC) (see [5,33] for an overview and further references)—the relevant large deviation results must first be established. This is the main contribution of this paper: we prove the large deviation principle for the empirical measures associated with Markov chains arising from the Metropolis–Hastings algorithm. This sets the stage for future work proving similar results for Markov chains with dynamics that resemble those of Metropolis–Hastings, and for analysing the corresponding MCMC methods.

The theory of large deviations has become a cornerstone in modern probability theory, with a wide range of applications. In the context of Monte Carlo methods, it has been known for a long time that for rare-event simulation, sample-path large deviations results are integral to analysing and designing efficient algorithms; see [4,11,12] and the references therein. In the MCMC setting, the theory remains much less explored for analysing performance and designing new, efficient methods. Instead, standard tools for convergence analysis of sampling methods based on ergodic Markov processes include: the spectral gap of the associated dynamics, mixing times of the process, asymptotic variance and functional inequalities (Poincaré, log-Sobolev) [6,15,26,27,30,47]. However, these tools mainly provide information about convergence of the associated n-step transition operator or the law of the process, neither of which are directly linked to the convergence of the empirical measure. Empirical measure large deviations are instead concerned precisely with the convergence of the empirical measure. This is in turn linked to the transient behaviour of the underlying Markov chain, which is of central importance for the performance of MCMC methods.

To the best of our knowledge, the first works on using large deviation theory to study the convergence of the empirical measures arising from MCMC sampling are [23,37]. Therein, the authors analyse the performance of parallel tempering, one of the most frequently applied MCMC methods in computational chemistry and physics, from the perspective of large deviations, leading to the construction of a new type of method known as infinite swapping. In the subsequent work [16], empirical measure large deviations and associated stochastic control problems are used to analyse the convergence properties of parallel tempering and infinite swapping. In [24] the authors study methods like parallel tempering and infinite swapping in the low-temperature regime, and use empirical measure large deviations to solve the long-standing open problem of optimal temperature selection. Similarly, in [8,38–40] a large deviation approach is used to analyse certain irreversible samplers. In [9], large deviations for the empirical measures of certain piecewise deterministic Markov processes, including the zig-zag sampler, are obtained, and the associated rate function is used to address a key question concerning the optimal choice of the so-called switching rate of the zig-zag process. The results therein also highlight the differences in considering convergence of empirical averages, and in studying the convergence to equilibrium with, e.g., the spectral gap; see also [47,49].

In this paper we focus on the Metropolis–Hastings algorithm [35] (described in Section 2.3), the most classical MCMC method and the main building block for many more advanced methods [1,4,41,48]. Because of its importance in the area of Monte Carlo sampling, the method is well-studied and classical results on convergence properties and performance include [13,28,34,42,44,46]; see also [20,36] and the references therein for the general theory of Markov chains. However, despite significant efforts over long time, there are still gaps in our understanding of the theoretical properties of this fundamental class of algorithms. As an example, in a recent tour de force [2,3] the authors develop a functional analytical framework, aimed at analysing Markov chains arising in sampling algorithms, and obtain the first explicit convergence bounds for the Metropolis algorithm. In [8] a non-reversible version of Metropolis–Hastings is introduced and studied. One of the methods used for analysing performance is large deviations for the associated empirical measure. Because the setting is a finite state space S, the classical results [17,19], due to Donsker and Varadhan, give the large deviation principle. To the best of our knowledge, this is the only work that studies large deviations for Markov chains arising from algorithms of Metropolis–Hastings-type. In [8] the focus is on the effects of non-reversibility, and there is thus no attempt at extending the large deviation results to the setting where the state space S is instead a (uncountable) subset of $\mathbb{R}^d$.

The pioneering work by Donsker and Varadhan [17–19] is often the starting point for empirical measure large deviations for Markov processes, and their results have been extended in numerous directions; see [12,14,25] and the references therein. However, it is pointed out in [22] (see also Section 2.2) that even for fairly simple continuous-time pure-jump processes, the results by Donsker and Varadhan, or more general versions of them such as in, e.g., [32], do not hold. This is because all such large deviation results rely on the transition probability function of the Markov process to have a density with respect to some reference measure. In [22] the authors show how this condition can be replaced by a more general transitivity condition (Condition 2.1 in the current paper) to ensure that a large class of processes are covered. However, for the Metropolis–Hastings chains, neither of these conditions hold due to the rejection part of the dynamics. The purpose of this paper is to show that, despite this violation of the standard transitivity conditions, the empirical measures of the Metropolis–Hastings chain do satisfy a large deviation principle. In fact, as discussed

in Remark 3.4, the result holds for a more general class of discrete time Markov processes whose transition kernel $K(x, dy)$ is a mixture of a part with density with respect to some reference measure, and a Dirac mass in x. The proof is based on the weak convergence approach [12,21], which is described in some more detail in Sections 2.2 and 4. With the large deviation results established, our future work is aimed at (i) analysing the performance and comparing various Metropolis–Hastings algorithms using the rate function, and comparing the conclusion to, e.g., the recent results [2]; (ii) investigate whether optimal scaling results, similar to the celebrated results in [28,44], can be obtained from a large deviation perspective; (iii) extend the results to cover more advanced MCMC algorithms, such as ABC-MCMC. These topics are all significant undertakings in their own right and we leave them to be investigated separately in future work.

The remainder of the paper is organized as follows. In Section 2 we provide the preliminaries needed for the paper: notation and definitions, a brief overview of large deviations for empirical measures, and a description of the Metropolis–Hastings algorithm. Next, in Section 3 we present the assumptions used for the Metropolis–Hastings chain. The main result is stated in Theorem 4.1 in Section 4. In this section we also show some properties of the associated rate function. The proof of Theorem 4.1 is divided into two parts, in Sections 5 and 6 we prove the Laplace upper and lower bound, respectively, which combined prove Theorem 4.1.

2. Preliminaries

2.1. Notation and definitions

Throughout the paper we work with some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We use a.s. and w.p. 1 as shorthand for almost sure, or almost surely, and with probability 1, respectively.

For a Polish space S, with a translation invariant metric d_S, $\mathcal{B}(S)$ is the Borel $\sigma-$algebra on S, and $C(S)$ and $C_b(S)$ denote the spaces of functions $f : S \to \mathbb{R}$ that are continuous, and bounded and continuous, respectively. For any $r \in \mathbb{R}_+$ and $x \in S$, $B_r(x)$ is the open ball of radius r with centre in x:

$$B_r(x) = \{y \in S : d_S(x, y) < r\}.$$

When $S \subseteq \mathbb{R}^d$, for some $d \geq 1$, we take λ to denote Lebesgue measure on $\mathbb{R}^d$. We abuse notation a bit in that λ is generically taken to represent Lebesgue measure, regardless of the underlying dimension d. For integration with respect to λ we use the standard notation dx for $\lambda(dx)$.

For a measure η on S, and a measurable function f on S, we denote the integral of f with respect to η by $\eta(f) = \int_S f(x)\eta(dx)$. When f is the indicator of a set A, we write $\eta(A) = \int_A \eta(dx)$.

The space of probability measures on S is denoted by $\mathcal{P}(S)$. Given $\gamma \in \mathcal{P}(S^2)$, denote by $[\gamma]_1$ and $[\gamma]_2$ the first and second marginals of γ, respectively. For $\mu \in \mathcal{P}(S)$, define

$$A(\mu) = \{\gamma \in \mathcal{P}(S^2) : [\gamma]_1 = [\gamma]_2 = \mu\}. \tag{1}$$

We consider the topology of weak convergence on $\mathcal{P}(S)$: $\nu_n \to \nu$ in this topology if, for all $f \in C_b(S)$,

$$\nu_n(f) = \int_S f(x)\nu_n(dx) \to \int_S f(x)\nu(dx) = \nu(f), \ n \to \infty.$$

We use $\nu_n \Rightarrow \nu$ as shorthand notation for $\{\nu_n\} \subset \mathcal{P}(S)$ converging weakly to $\nu \in \mathcal{P}(S)$. Unless otherwise stated, we equip $\mathcal{P}(S)$ with the *Lévy–Prohorov metric*, denoted d_{LP}: for $\nu, \mu \in \mathcal{P}(S)$,

$$d_{LP}(\nu, \mu) = \inf\{\epsilon > 0 : \ \nu(A) \leq \mu(A^\epsilon) + \epsilon, \text{ for all closed subsets } A \subset S\},$$

where $A^\epsilon = \{x \in S : d_S(x, A) < \epsilon\}$. This metric is compatible with the topology of weak convergence (see, e.g., [12], Theorem A.1), and turns $\mathcal{P}(S)$ into a Polish space. For any signed measure η on S, the *total variation norm* of η, $\|\eta\|_{TV}$, is defined as

$$\|\eta\|_{TV} = \sup_f |\eta(f)|,$$

where the supremum is taken over all measurable functions bounded by 1. For $\nu, \mu \in \mathcal{P}(S)$, the total variation norm provides an upper bound on d_{LP}:

$$d_{LP}(\nu, \mu) \leq \|\nu - \mu\|_{TV}.$$

For a measurable space $(\mathcal{Y}, \mathcal{A})$, let $q(y, dx)$ be a collection of probability measures on S parameterized by $y \in Y$. Then q is called a *stochastic kernel* on S given $\mathcal{Y}$ if, for every $A \in \mathcal{B}(S)$, the map $y \mapsto q(y, A) \in [0, 1]$ is measurable.

For a Markov chain $\{X_i\}_{i\in\mathbb{N}}$ taking values in S, for a given $x_0 \in S$, we denote by $\mathbb{P}_{x_0}$ the distribution of $\{X_i\}_{i\in\mathbb{N}}$ starting at x_0. The associated expectation operator is denoted by $\mathbb{E}_{x_0}$. The *transition probability function*, or *transition kernel*, of a Markov chain is a stochastic kernel q, such that the distribution of X_i given X_{i-1} is given by $q(X_{i-1}, \cdot)$. We say that a transition probability function $q(x, dy)$ on $S \times \mathcal{P}(A)$ satisfies the *Feller property* if, for any sequence $\{x_n\}_{n\in\mathbb{N}}$ such that $x_n \to x \in S$ as $n \to \infty$, $q(x_n, \cdot) \Rightarrow q(x, \cdot)$.

Given a measure $\mu \in \mathcal{P}(S)$ and a transition kernel $q(x, dy)$, we say that μ is *invariant* for q, or for the corresponding Markov chain, if for all $A \in \mathcal{B}(S)$,

$$\mu(A) = \int_S q(x, A)\mu(dx).$$

For $\nu \in \mathcal{P}(S)$, $R(\cdot \parallel \nu) : \mathcal{P}(S) \to [0, \infty]$ is the *relative entropy* (with respect to ν), defined by

$$R(\mu \parallel \nu) = \begin{cases} \int_S \log\left(\frac{d\mu}{d\nu}\right) d\mu, & \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

We recall the following properties of relative entropy (see Lemmas 1.4.1 and 1.4.3 in [21]): $R(\cdot \parallel \cdot)$ is jointly convex and jointly lower semi-continuous with respect to the weak topology on $\mathcal{P}(S)^2$, and $R(\mu \parallel \nu) = 0$ if and only if $\mu = \nu$. Another useful property follows from the chain rule for relative entropy (see Theorem 2.6 and Corollary 2.7 in [12]): given two transition kernels p, q, for any $\mu \in \mathcal{P}(S)$,

$$R(\mu \otimes p \parallel \mu \otimes q) = \int_S R(p(x, \cdot) \parallel q(x, \cdot))\mu(dx). \tag{2}$$

Lastly, for a set A, A° and $\bar{A}$ denote the *interior* and *closure* of the set, respectively, and $x \mapsto I\{x \in A\}$ is the *indicator function* of the set A. When the set is a singleton, $A = \{y\}$, we write $I\{x = y\}$. We also use δ_y to denote this case.

2.2. Large deviations for empirical measures of a Markov chain

Consider a Markov chain $\{X_i\}_{i \geq 0}$ with state space S and transition probability function p. The *empirical measure*, L^n, associated with the chain $\{X_i\}$ is defined as

$$L^n(\cdot) = \frac{1}{n} \sum_{i=0}^{n-1} \delta_{X_i}(\cdot). \tag{3}$$

For each n, this is a random element of $\mathcal{P}(S)$. We can also view $\{L^n\}_{n \geq 0}$ as a stochastic process in $\mathcal{P}(S)$.

In the context of MCMC methods, empirical measures are essential objects as they are used for forming approximations for any observable: for a given observable $f \in C_b(S)$, we have

$$L^n(f) = \frac{1}{n} \sum_{i=0}^{n-1} f(X_i).$$

If the Markov chain X has an invariant distribution $\pi \in \mathcal{P}(S)$ and is ergodic, we have $L^n(f) \to \pi(f)$, a.s. as $n \to \infty$. Thus, there is a direct link between the convergence properties of the empirical measure L^n and the performance of Monte Carlo methods based on time averages for approximating observables.

Classical methods for studying performance of MCMC methods are often mixing properties or asymptotic variance, which are not directly linked to the empirical measure L^n of the underlying Markov chain. The theory of large deviations on the other hand, is concerned precisely with deviations of L^n from π as the number of steps n grows. It therefore serves as a useful complement to the more traditional methods for analysing performance of a given MCMC method, as well as for designing new algorithms.

At the heart of the theory of large deviations is the *large deviation principle* (LDP): the sequence $\{L^n\}$ is said to satisfy an LDP with speed n and *rate function* $I : S \to [0, \infty]$, if I is lower semi-continuous, has compact sub-level sets and for any measurable $A \subset \mathcal{P}(S)$,

$$-\inf_{\nu \in A^\circ} I(\nu) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(L^n \in A^\circ) \leq \limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(L^n \in \bar{A}) \leq -\inf_{\nu \in \bar{A}} I(\nu).$$

The gist of these inequalities is that, if $\{L^n\}$ satisfies an LDP with speed n and rate function I, then for any $\nu \in \mathcal{P}(S)$ and n large,

$$\mathbb{P}(L^n \approx \nu) \simeq \exp\{-nI(\nu)\}.$$

The definition of an LDP makes this statement rigorous in the limit $n \to \infty$.

For any metric space, an equivalent formulation of the LDP is the *Laplace principle* (see e.g., Theorems 1.5 and 1.8 in [12]). In the setting of the empirical measures $\{L^n\}$, we have that this sequence satisfies a Laplace principle, with speed n and rate function I (same as in the LDP), if for any $F \in C_b(\mathcal{P}(S))$,

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}\left[e^{-nF(L^n)}\right] = -\inf_{\nu \in \mathcal{P}(S)} \{F(\nu) + I(\nu)\}. \tag{4}$$

The starting point for large deviations of empirical measures of Markov processes is the pioneering work of Donsker and Varadhan [17,19]. A central assumption in those works is that the transition probability function p has a density with respect to some reference measure. This is a reasonable transitivity assumption for processes that involve something that, in some sense, resembles a diffusive term. However, in [22] the authors show that it is a rather restrictive condition and as an example construct a simple continuous-time pure-jump process for which it does not hold. The following alternative condition on p was used in [22] to establish an LDP for the empirical measures of a Markov process.

Condition 2.1 (*Condition 6.3 in [12]*)**.** The transition kernel p of the Markov chain X is such that there exist positive integers l_0 and n_0, such that for all x and ζ in S,

$$\sum_{i \geq l_0} 2^{-i} p^{(i)}(x, dy) \ll \sum_{j \geq n_0} 2^{-j} p^{(j)}(\zeta, dy), \tag{5}$$

where $p^{(k)}$ denotes the $k-$step transition probability.

This condition is general enough to cover a large class of Markov processes, both in discrete and continuous time; see e.g., [12,22] and the references therein. However, it does not cover the case when X comes from a Metropolis–Hastings scheme, as we show with a simple counterexample in Section 4. Condition 2.1, or variations of it, is a key ingredient in existing work on large deviations for Markov chains. Because it is not satisfied for Metropolis–Hastings, in order to use large deviations to analyse the performance of such algorithms, and with an outlook towards more advanced MCMC methods that build on the Metropolis–Hastings algorithm—e.g., MALA and ABC-MCMC—we must first establish the relevant LDP. This is the main contribution of this paper.

2.3. Metropolis–Hastings algorithm

We now give a brief description of the Metropolis–Hastings (MH) algorithm for constructing a Markov chain $\{X_i\}_{i\geq 0}$ with the target measure π as invariant distribution. For simplicity we restrict ourselves to the setting where $S \subseteq \mathbb{R}^d$ and π is equivalent to Lebesgue measure. More abstract settings are possible as well, see for example [48]. However this would require different assumptions and modifications of the proof of the large deviation principle in Section 2.2.

The main ingredient of the MH algorithm is the *proposal distribution* $J(\cdot|x) \in \mathcal{P}(S)$, defined for all $x \in S$. If the chain after n steps is in some state $X_n = x_n$, a proposal Y_{n+1} for the next state X_{n+1} is generated from $J(\cdot|x_n)$. This is followed by an acceptance–rejection step, which is defined in terms of the *Hastings ratio*,

$$\varpi(x, y) = \min\left\{1, \frac{\pi(y)J(x|y)}{\pi(x)J(y|x)}\right\};$$

where if $\pi(x)J(y|x) = 0$, we set $\varpi(x, y) = 1$. The proposed move from $X_n = x_n$ to $X_{n+1} = Y_{n+1}$ is accepted with probability $\varpi(x_n, Y_{n+1})$, and rejected with probability $1 - \varpi(x_n, Y_{n+1})$. In the latter case, we set $X_{n+1} = x_n$. The pseudocode for the update step in the MH algorithm is presented in Algorithm 2.1.

Algorithm 2.1 Metropolis–Hastings algorithm

Given $X_n = x_n$,

1: Generate a proposal $Y_{n+1} \sim J(\cdot|x_n)$

2: Set

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{with probability } \varpi(x_n, Y_{n+1}) \\ x_n & \text{with probability } 1 - \varpi(x_n, Y_{n+1}) \end{cases}$$

Define a transition kernel $a(x, dy)$ and a function $r : S \to [0, 1]$ by

$$a(x, dy) = \min\left\{1, \frac{\pi(y)J(x|y)}{\pi(x)J(y|x)}\right\} J(dy|x), \tag{6}$$

and

$$r(x) = 1 - a(x, S) = 1 - \int_S a(x, dy). \tag{7}$$

The kernel a corresponds to the acceptance-part of the MH algorithm, i.e., it corresponds to transitions to proposed states that are accepted in the MH algorithm. Similarly, r corresponds to the rejection part: it represents the probability of rejecting a proposed state, and thus remaining at the current state of the chain. With these definitions, the dynamics of the MH algorithm corresponds to generating a Markov chain $\{X_i\}_{i\geq 0}$, the MH chain, with transition kernel

$$K(x, dy) = a(x, dy) + r(x)\delta_x(dy). \tag{8}$$

For a more in-depth look at the MH algorithm and its various properties, see for example [41] and the references therein. A key observation is that due to the form of the Hastings ratio, and the corresponding kernel K, under reasonable assumptions on the proposal distribution J, the MH chain $\{X_i\}_{i\geq 0}$ generated according to the above has π as its unique invariant measure.

3. Assumptions

In this section, we state the assumptions we make on the MH chain defined in Section 2.3. Rather than aiming to make them as general as possible, we have aimed for assumptions, primarily on the proposal distribution J, that are tangible from the perspective of MCMC methods. One alternative, commonly used when studying this type of Markov chain, is to assume the existence of some Lyapunov function [31,32,34,45]. Although this ensures the convergence of the empirical measures, for the large deviation results additional assumptions are still needed; see e.g., the Donsker–Varadhan-like assumption on the transition kernel in [32].

As mentioned in Section 2.3, we make the assumption that $S \subseteq \mathbb{R}^d$, for some $d \geq 1$. We make a slight abuse of notation, in that we let $\pi(\cdot)$, $J(\cdot|x)$, and $a(x, \cdot)$ denote both the measures and the corresponding density functions. In order to establish the LDP, we make the following additional assumptions.

(A.1) S is an open subset of $\mathbb{R}^d$ and the target probability measure π is equivalent to λ on S (i.e., $\pi \ll \lambda$ and $\lambda \ll \pi$). The probability density $\pi(x)$ is a continuous function.

(A.2) The proposal distribution $J(\cdot|x)$ is absolutely continuous with respect to the target measure π (i.e., $J(\cdot|x) \ll \pi$), for all $x \in S$. The probability density $J(y|x)$ is a continuous and bounded function of x and y, and it satisfies

$$J(y|x) > 0, \quad \forall (x, y) \in S^2. \tag{9}$$

(A.3) There exists a Lyapunov function $U : S \to [0, \infty)$ such that the following properties hold:

(a) $\inf_{x\in S} \left[U(x) - \log \int_S e^{U(y)} K(x, dy)\right] > -\infty$

(b) For each $M < \infty$, the set

$$\left\{ x \in S \ : \ U(x) - \log \int_S e^{U(y)} K(x, dy) \le M \right\}$$

is a relatively compact subset of S.

(c) For every compact set $K \subset S$ there exists $C_K < \infty$ such that

$$\sup_{x\in K} U(x) \le C_K.$$

Because π and λ are equivalent measures, the support of π is all of S. However, it is not necessarily the case that $\pi(x) > 0$ for all $x \in S$, as there may exist a (nonempty) set $E \subset S$, such that $\lambda(E) = 0$ and $\pi(x) = 0$ for $x \in E$. Therefore, define the set S_+ as

$$S_+ = \{y \in S \ : \ \pi(y) > 0\}. \tag{10}$$

Observe that S_+ is an open subset of S, being the density function $\pi(x)$ continuous.

Assumptions (A.1)–(A.2) are used to show that the MH transition kernel K, and thus the MH chain $\{X_i\}_{i\ge 0}$, has certain properties needed for the LDP, including that π is the unique invariant distribution. Assumption (A.3) replaces a compactness-assumption on S for proving the LDP. In the case of a compact state space S, this assumption is not needed.

Remark 3.1. We start by showing that the combination of (A.1) and (A.2) ensures continuity and boundedness of the components a (acceptance part) and r (rejection part) of the MH transition kernel K.

Consider $x \in S_+$. Assumption (A.1) and Assumption (A.2) imply that $J(\cdot|x) \ll \lambda$. Therefore, the acceptance part (6) of $K(x, \cdot)$ is absolutely continuous with respect to the Lebesgue measure (i.e., $a(x, \cdot) \ll \lambda$), and its density is given by

$$a(x, y) = \min\left\{1, \frac{\pi(y)J(x|y)}{\pi(x)J(y|x)}\right\} J(y|x). \tag{11}$$

Since $\pi(x)$ is continuous for all $x \in S$ and $J(x|y)$ is continuous and bounded for all $(x, y) \in S^2$, we have $a(x, y) \in C_b(S_+ \times S)$.

From the continuity of $a(x, y)$ on $S_+ \times S$, we obtain that $r(x) = 1 - a(x, S)$ is also continuous for all $x \in S_+$. This continuity extends to all of S. First, if $x \notin S_+$, so that $\pi(x) = 0$, then

$$r(x) = 1 - a(x, S) = 1 - \int_S J(y|x)dy = 0,$$

since $\pi(y) > 0$ for λ-almost all $y \in S$. Take $x \notin S_+$ and a sequence $\{x_n\} \subset S$ that converges to x. From the continuity of the target density function π, $\pi(x_n) \to \pi(x) = 0$. Moreover, for a fixed y such that $\pi(y) > 0$, we have $a(x_n, y) \to J(y|x)$ as $n \to \infty$. To see this, note that since π and λ are equivalent, $\pi(y) > 0$ for $\lambda-$almost all $y \in S$. It follows that $\lim_{n\to\infty} a(x_n, y) = J(y|x)$ for $\lambda-$almost all $y \in S$. Recalling that J is bounded, by dominated convergence we have

$$\lim_{n\to\infty} a(x_n, S) = \lim_{n\to\infty} \int_S a(x_n, y)dy = \int_S \lim_{n\to\infty} a(x_n, y)dy = \int_S J(y|x)dy = 1.$$

This in turn implies that

$$\lim_{n\to\infty} r(x_n) = 1 - \lim_{n\to\infty} a(x_n, S) = 0.$$

Since $r(x) = 0$, this shows that r is continuous on S.

Remark 3.2. Next, we show that (A.1)–(A.2) ensure that K has the target measure π as its unique invariant distribution, and the MH chain $\{X_i\}_{i\ge 0}$ is ergodic.

Let $x \in S_+$ as defined in (10). Since $\lambda \ll \pi$ by (A.1), $\pi(y) > 0$ for $\lambda-$almost every $y \in S$. Moreover, by Assumption (A.2), $J(x|y) > 0$ for all $(x, y) \in S^2$. It follows that $a(x, y) > 0$ for $\lambda-$a.e. $y \in S$. This in turn implies that $\lambda \ll a(x, \cdot)$, and λ and $a(x, \cdot)$ are equivalent measures for all $x \in S_+$. By transitivity, $a(x, \cdot)$ and $a(y, \cdot)$ are equivalent for all $x, y, \in S_+$. We now show that from this it follows that the MH transition kernel K is indecomposable, i.e. there are no disjoint Borel sets $A_1, A_2 \in \mathcal{B}(S)$ such that

$$K(x, A_1) = 1 \ \forall x \in A_1 \quad \text{and} \quad K(y, A_2) = 1 \ \forall y \in A_2.$$

We argue by contradiction. Assume that two such sets exist. Then,

$$1 = K(x, A_1) = a(x, A_1) + r(x)\delta_x(A_1). \tag{12}$$

Since $\lambda \ll a(x,\cdot)$, we have $a(x,S) > 0$, and thus $r(x) = 1 - a(x,S) < 1$ for all $x \in S$. Combined with (12), this shows $a(x,A_1) > 0$. It follows from $a(x,\cdot)$ and $a(y,\cdot)$ being equivalent measures that $a(y,A_1) > 0$, which contradicts the assumption. Hence, $K(x,dy)$ is indecomposable. By Theorem 7.16 in [10], π is the unique invariant distribution for the MH transition kernel $K(x,dy)$ and the Markov chain associated with π and $K(x,dy)$ is ergodic.

Remark 3.3. The existence of a Lyapunov function U satisfying Assumption (A.3) can be shown for different instances of the MH algorithm. For example, in forthcoming work we give precise results on conditions on the tail decays of π and J for Assumption (A.3) to hold for independent MH and MALA. These results are in line with those of [34,45] on uniform and geometric ergodicity.

For the case $S = \mathbb{R}^d$, Section 8.2 in [21] describes a class of models for which a Lyapunov function U that satisfies (A.3) exists. Here we present their example adapted to the MH kernel K. For specific choices of J and/or π, this assumption can be made more explicit (or verified).

Let $b : \mathbb{R}^d \to \mathbb{R}^d$ be measurable. Denote by $\langle\cdot,\cdot\rangle$ the scalar product in $\mathbb{R}^d$ and for $\alpha \in \mathbb{R}^d$ define

$$H_b(x,\alpha) = \log\left[\int_{\mathbb{R}^d} e^{\langle\alpha, y-x-b(x)\rangle} a(x,dy) + r(x)e^{-\langle\alpha,b(x)\rangle}\right].$$

Consider the following assumptions.

(a) b is bounded on all compact sets in $\mathbb{R}^d$

(b) there exists $r > 0$ such that

$$\sup_{x\in\mathbb{R}^d} H_b(x,\alpha) < \infty,$$

for all $\alpha \in \mathbb{R}^d$ that satisfy $\|\alpha\| \le r$

(c) there exists a Lipschitz continuous function $U : \mathbb{R}^d \to [0,\infty)$ for which

$$\lim_{\|x\|\to\infty} [U(x+b(x)) - U(x)] = -\infty.$$

If (a), (b) and (c) hold, then U is a Lyapunov function as required by Assumption (A.3).

A natural choice for b is

$$b(x) = \int_{\mathbb{R}^d} y \cdot a(x,dy) - (1 - r(x)) \cdot x,$$

and the corresponding H_b is

$$H_b(x,\alpha) = -\langle\alpha, \int_{\mathbb{R}^d} y \cdot a(x,dy) + r(x)\cdot x\rangle + \log\left[\int_{\mathbb{R}^d} e^{\langle\alpha,y\rangle} a(x,dy) + e^{\langle\alpha,x\rangle} r(x)\right].$$

Note that if the space S is compact, then Assumption (A.3) is automatically satisfied (for example, take $U(x) \equiv 0$).

Remark 3.4. The large deviation principle that we prove in the present paper holds for a broader class of Markov chains than those of MH type. In particular, Theorem 4.1 remains valid if Assumptions (A.1) and (A.3) are satisfied, in conjunction with the following assumptions, which generalize Assumption (A.2).

(B.1) The Markov chain transition kernel can be decomposed as

$$K(x,dy) = a(x,dy) + r(x)\delta_x(dy),$$

where $a(x,\cdot)$ is a measure on S for all $x \in S$, the map $x \mapsto a(x,A)$ is measurable for every $A \in \mathcal{B}(S)$, and $r(x) = 1 - \int_S a(x,dy)$.

(B.2) The measure $a(x,\cdot)$ is absolutely continuous with respect to the target measure π (i.e., $a(x,\cdot) \ll \pi$), for all $x \in S$. The probability density $a(x,y)$ is a continuous and bounded function of x and y, and it satisfies

$$a(x,x) > 0, \quad \forall x \in S_+,$$

and

$$\int_S a(x,y)dy = 1, \quad \forall x \notin S_+,$$

i.e. if $\pi(x) = 0$, then $r(x) = 0$.

(B.3) The Markov chain associated with π and $K(x,dy)$ is ergodic.

4. Large deviations for empirical measures of Metropolis–Hastings chains

We are now ready to state our main result, an LDP for the sequence $\{L^n\}$ of empirical measures of the MH chain $\{X_i\}_{i\ge0}$ with invariant distribution π (see Section 2.3 for the definition).

Theorem 4.1. *Let $\{X_i\}_{i\geq 0}$ be the Metropolis–Hastings chain from Section 2.3 and $K(x,dy)$ the associated transition kernel. Let $\{L^n\}_{n\geq 0} \subset \mathcal{P}(S)$ be the corresponding sequence of empirical measures, defined in (3). Under Assumptions (A.1)–(A.3), with $A(\mu)$ as in (1), $\{L^n\}_{n\geq 0}$ satisfies an LDP with speed n and rate function*

$$I(\mu) = \inf_{\gamma\in A(\mu)} R(\gamma \parallel \mu \otimes K). \tag{13}$$

As mentioned in Section 2.1, we consider $\mathcal{P}(S)$ as a metric space (equipped with, e.g., the Lévy–Prohorov metric). Therefore, the LDP is equivalent to the Laplace principle, and we will use the latter to prove Theorem 4.1. More specifically, the proof is split up into proving the *Laplace principle upper bound*,

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{E}_{x_n}\left[e^{-nF(L^n)}\right] \geq \inf_{\mu\in\mathcal{P}(S)} \{F(\mu)+I(\mu)\}, \tag{14}$$

and the *Laplace principle lower bound*,

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{E}_{x}\left[e^{-nF(L^n)}\right] \leq \inf_{\mu\in\mathcal{P}(S)} \{F(\mu)+I(\mu)\}, \tag{15}$$

for every $F \in C_b(\mathcal{P}(S))$, every sequence $\{x_n\} \subset S$ and $x \in S$. The respective proofs are given in Sections 5 and 6. The starting point for both bounds is the following representation formula (Proposition 6.1 in [12]): for every bounded, measurable $F : \mathcal{P}(S) \to \mathbb{R}$,

$$-\frac{1}{n}\log \mathbb{E}\left[e^{-nF(L^n)}\right] = \inf_{\{\bar{\mu}_i^n\}} \mathbb{E}\left[F(\bar{L}^n) + \frac{1}{n}\sum_{i=1}^{n} R(\bar{\mu}_i^n \parallel K(\bar{X}_{i-1}^n,\cdot))\right], \tag{16}$$

where $\bar{L}^n$ is the controlled empirical measure, $\bar{L}^n = \frac{1}{n}\sum_{i=0}^{n-1}\delta_{\bar{X}^n}$, and the conditional distribution of $\bar{X}_i^n$ given $\sigma(\bar{X}_1^n,\dots,\bar{X}_{i-1}^n)$ is $\bar{\mu}_i^n$. The infimum is over all such controls, i.e., random probability measures, $\bar{\mu}_i^n$, such that $\bar{\mu}_i^n$ is measurable with respect to $\mathcal{F}_{i-1}^n = \sigma\left(\bar{X}_1^n,\dots,\bar{X}_{i-1}^n\right)$, with $\mathcal{F}_0^n = \{\emptyset,\Omega\}$; see [12,21] for more details.

For the upper bound, under Assumptions (A.1)–(A.3), the proof of Proposition 6.13 in [12], with the additional arguments in Section 6.10 therein to account for a non-compact state space, can be applied in our setting as well. The only thing that needs to be verified is the Feller property of the MH transition kernel K (see Lemma 5.2).

The work for proving Theorem 4.1 lies in proving the lower bound (15). Existing results rely on some variation of Condition 2.1. However, such a condition is not applicable in our setting, as the following simple example shows: Take an $x \in S$ such that $r(x) > 0$ (i.e., when in x, there is a positive probability of rejecting a proposed move and stay in x) and consider the Borel set $A = \{x\} \in \mathcal{B}(S)$. If $x \neq \zeta$, then $K^{(j)}(\zeta, A) = K^{(j)}(\zeta,\{x\}) = 0, \forall j \geq 0$. However, $K^{(i)}(x,\{x\}) > 0, \forall i \geq 0$, since $r(x) > 0$. This shows that (5) does not hold for all $x \in S$, and Condition 2.1 does not hold for the MH kernel K, nor for kernels of similar type, such as those arising in ABC-MCMC or MALA.

In Section 6 we show how the Laplace principle lower bound can be shown for the MH chain without relying on a transitivity assumption like Condition 2.1. The main point is that due to the specific structure of the MH kernel, under Assumptions (A.1)–(A.2) the chain retains the properties that are important for proving the LDP (and typically guaranteed by something like Condition 2.1 combined with other assumptions).

The main difficulty in the proof arises from the fact that, contrary to the setting in [12], for $\nu \in \mathcal{P}(S)$, $I(\nu) < \infty$ does not imply that $\nu \ll \pi$. In [12] this implication is used in defining near-optimal controls in the representation (16), which in turn can be used to prove the lower bound.

Before proceeding with the proofs of the upper and lower bounds, in the following section we give some different characterizations and properties of the rate function I in (13). Note that although here the rate function is phrased in terms of relative entropy, other (equivalent) formulations are also possible, similar to, e.g., the early work by Donsker and Varadhan [17]. Establishing such alternative formulations, and their use in analysing MCMC methods, is the topic of forthcoming work by the authors.

4.1. Characterization and properties of the rate function

We first express the rate function (13) in a more convenient form. By Lemma 6.8(a) in [12], the probability measures in the set $A(\nu)$ are of the form

$$\gamma(dx \times dy) = \nu(dx)\, q(x,dy),$$

for a transition kernel $q(x,dy)$ such that ν is invariant for q. Therefore, using (2), the chain rule for relative entropy, we can rewrite (13) as

$$I(\nu) = \inf_{q\in\mathcal{Q}} \int_S R(q(x,\cdot) \parallel K(x,\cdot))\nu(dx), \tag{17}$$

where $\mathcal{Q}$ denotes the set of all the transition kernels $q(x,dy)$ on S such that ν is an invariant distribution for q. Lemma 6.8(b) in [12] guarantees the existence of a minimizing q in the definition of $I(\nu)$, under the assumption $I(\nu) < \infty$. That is, there exists a transition kernel q with stationary distribution ν such that

$$I(\nu) = \int_S R(q(x,\cdot) \parallel K(x,\cdot))\nu(dx). \tag{18}$$

The representation (18) of the rate function allows us to characterize the minimizers q, based on the form of the MH transition kernel K (8), as the following result shows.

Lemma 4.2. *If $I(\nu) < \infty$, then the transition kernel $q(x, dy)$ in (18) is $q(x,\cdot) \ll K(x,\cdot)$ ν-a.s. In particular, it is of the form*

$$q(x,\cdot) = \alpha(x,\cdot) + \rho(x)\delta_x(\cdot), \quad \nu\text{-a.s.}, \tag{19}$$

with $\alpha(x,\cdot) \ll a(x,\cdot)$ ν-a.s. and $\rho(x)$ is a measurable function.

Proof. If $I(\nu) < \infty$, then (18) implies $R(q(x,\cdot) \parallel K(x,\cdot)) < \infty$ ν−a.s. By the definition of relative entropy, this means that $q(x,\cdot) \ll K(x,\cdot)$ ν−a.s. Recall that

$$K(x, dy) = a(x, y)dy + r(x)\delta_x(dy),$$

i.e. $K(x,\cdot)$ is a mixture of a transition kernel $a(x,\cdot) \ll \lambda$, and a point mass in x. Therefore, for the transition kernel $q(x,\cdot)$ to be $q(x,\cdot) \ll K(x,\cdot)$ ν−a.s., it must be of the form

$$q(x, y) = \alpha(x, y)dy + \rho(x)\delta_x(dy),$$

where $\alpha(x,\cdot) \ll a(x,\cdot) \ll \lambda$, and $\rho(x) = 0$ if $r(x) = 0$. In particular, $\rho(x)$ must be a measurable function in order to make $x \mapsto q(x, A)$ a measurable function for every $A \in \mathcal{B}(S)$, and therefore q a stochastic kernel. □

With the characterization of q from Lemma 4.2, we can write the rate function (18) in a more explicit way.

Proposition 4.3. *If $I(\nu) < \infty$, then the rate function can be expressed as*

$$I(\nu) = \int_S \int_S \log\left(\frac{\alpha(x,y)}{a(x,y)}\right) \alpha(x,y)\, dy\, \nu(dx) + \int_S \log\left(\frac{\rho(x)}{r(x)}\right) \rho(x)\, \nu(dx), \tag{20}$$

with $\alpha(x,y)$ and $\rho(x)$ as in Lemma 4.2.

Proof. Applying the definition of relative entropy in (18), the rate function becomes

$$I(\nu) = \int_S \int_S \log\big(f_x(y)\big)\, q(x, dy)\nu(dx), \tag{21}$$

where f_x denotes the Radon–Nikodym derivative of the transition kernel $q(x,\cdot)$ with respect to $K(x,\cdot)$ for a fixed $x \in S$. By Lemma 4.2 f_x exists ν-a.s. and, by combining (8) and (19),

$$f_x(y) = \frac{\alpha(x,y)}{a(x,y)} I\{y \neq x\} + \frac{\rho(x)}{r(x)} I\{y = x\}. \tag{22}$$

Indeed, let $A \in \mathcal{B}(S)$ and recall that $a(x,\cdot) \ll \lambda$. Then, it holds

$$\begin{aligned}\int_A f_x(y)K(x, dy) &= \int_A \left(\frac{\alpha(x,y)}{a(x,y)} I\{y \neq x\} + \frac{\rho(x)}{r(x)} I\{y = x\}\right)\big(a(x,y)dy + r(x)\delta_x(dy)\big)\\ &= \int_A \alpha(x,y)dy + \rho(x)\delta_x(A) = q(x, A),\end{aligned}$$

for ν-almost all $x \in S$. This proves that $f_x(y)$ in (22) is the Radon–Nikodym derivative of $q(x,\cdot)$ with respect to $K(x,\cdot)$ for ν-almost all x in S.

Replacing $f_x(y)$ in (21) with (22) gives

$$\begin{aligned}I(\nu) &= \int_S \int_S \log\left(\frac{\alpha(x,y)}{a(x,y)} I\{y \neq x\} + \frac{\rho(x)}{r(x)} I\{y = x\}\right)\big(\alpha(x,y)dy + \rho(x)\delta_x(dy)\big)\, \nu(dx)\\ &= \int_S \left(\int_S \log\left(\frac{\alpha(x,y)}{a(x,y)}\right) \alpha(x,y)dy + \log\left(\frac{\rho(x)}{r(x)}\right) \rho(x)\right) \nu(dx),\end{aligned}$$

which leads to (20). □

We end this section with an alternative characterization of the rate function, that highlights the fact that measures $\nu \in \mathcal{P}(S)$ for which $I(\nu) < \infty$ need not be absolutely continuous with respect to π.

For any $\nu \in \mathcal{P}(S)$, by the Lebesgue decomposition theorem, we have

$$\nu = (1 - p) \cdot \nu_\lambda + p \cdot \nu_s, \tag{23}$$

where $p \in [0, 1]$, $\nu_\lambda, \nu_s \in \mathcal{P}(S)$, with $\nu_\lambda \ll \lambda$ and $\nu_s \perp \lambda$. Note that p is specific to ν, which we suppress in the notation. Associated with the decomposition (23), we also define the partition $S = S_\lambda \cup S_s$, with $S_\lambda \cap S_s = \emptyset$, $\nu_s(S_\lambda) = 0$ and $\lambda(S_s) = 0$. The following Lemma shows that $I(\nu)$ is split into two parts, one corresponding to ν_λ and one corresponding to ν_s.

Lemma 4.4. *Let $\nu \in \mathcal{P}(S)$ with $I(\nu) < \infty$ and consider its decomposition as in (23). Let $q(x, dy)$ be a transition kernel on S with invariant distribution ν, that satisfies*

$$I(\nu) = \int_S R(q(x,\cdot) \parallel K(x,\cdot))\nu(dx).$$

Define $\mathcal{Q}_\lambda$ and $\mathcal{Q}_s$ as the set of transitions kernels that ν_λ and ν_s are invariant for, respectively. The following holds:

(a) $q \in \mathcal{Q}_\lambda \cap \mathcal{Q}_s$, *i.e. both* ν_λ *and* ν_s *are invariant for* q,
(b) the rate function satisfies

$$I(\nu) = (1-p)I(\nu_\lambda) + pI(\nu_s). \tag{24}$$

Proof. (a) By Lemma 4.2, we can write

$$q(x,\cdot) = \alpha(x,\cdot) + \rho(x)\delta_x(\cdot), \quad \nu\text{-a.s.},$$

where $\alpha(x,\cdot) \ll \lambda$. By invariance of ν for q, for all $A \in \mathcal{B}(S)$,

$$\nu(A) = \int_S q(x,A)\nu(dx) = \int_S \alpha(x,A)\nu(dx) + \int_A \rho(x)\nu(dx).$$

If we consider $A = S_s$, for which $\lambda(S_s) = 0$, then $\alpha(x,S_s) = 0$ for ν-almost all $x \in S$ (because of $\alpha(x,\cdot) \ll \lambda$), and thus $\nu(S_s) = \int_{S_s} \rho(x)\nu(dx)$. On the other hand, $\nu(S_s) = \int_{S_s} \nu(dx)$. This implies that for all $x \in S_s$ ν-a.s., we have that $\rho(x) = 1$ a.s., and therefore $q(x,dy) = \delta_x(dy)$.

With the form of q on S_s established, for $A \in \mathcal{B}(S)$, we have

$$\int_S q(x,A)\nu_s(dx) = \int_{S_s} q(x,A)\nu_s(dx) = \int_{S_s} \delta_x(A)\nu_s(dx) = \nu_s(A).$$

This proves that ν_s is invariant for q, which means that $q \in \mathcal{Q}_s$.

We now show that ν_λ is also invariant for q. The decomposition (23) combined with the invariance of ν for q, and given that $q(x,\cdot) = \delta_x(dy)$, ν_s-a.s., gives, for $A \in \mathcal{B}(S)$,

$$\begin{aligned}(1-p)\cdot\nu_\lambda(A) + p\cdot\nu_s(A) = \nu(A) &= \int q(x,A)\nu(dx)\\ &= (1-p)\cdot\int q(x,A)\nu_\lambda(dx) + p\cdot\int q(x,A)\nu_s(dx)\\ &= (1-p)\cdot\int q(x,A)\nu_\lambda(dx) + p\cdot\nu_s(A).\end{aligned}$$

It follows that $\nu_\lambda(A) = \int q(x,A)\nu_\lambda(dx)$. Since $A \in \mathcal{B}(S)$ was chosen arbitrarily, ν_λ is invariant for q, i.e., $q \in \mathcal{Q}_\lambda$.

To prove (b), by convexity of I (see Lemma 6.10(a) in [12]),

$$I(\nu) = I\left((1-p)\cdot\nu_\lambda + p\cdot\nu_s\right) \le (1-p)\cdot I(\nu_\lambda) + p\cdot I(\nu_s). \tag{25}$$

On the other hand, by the decomposition (23),

$$I(\nu) = \int_S R(q(x,\cdot)\parallel K(x,\cdot))\nu(dx) = (1-p)\cdot\int_S R(q(x,\cdot)\parallel K(x,\cdot))\nu_\lambda(dx) + p\cdot\int_S R(q(x,\cdot)\parallel K(x,\cdot))\nu_s(dx). \tag{26}$$

From part (a), q is an element of both $\mathcal{Q}_\lambda$ and $\mathcal{Q}_s$. Therefore,

$$\int_S R(q(x,\cdot)\parallel K(x,\cdot))\nu_\lambda(dx) \ge \inf_{\tilde{q}\in\mathcal{Q}_\lambda}\int_S R(\tilde{q}(x,\cdot)\parallel K(x,\cdot))\nu_\lambda(dx).$$

The right-hand side of the previous display is precisely $I(\nu_\lambda)$. Similarly,

$$\int_S R(q(x,\cdot)\parallel K(x,\cdot))\nu_s(dx) \ge \inf_{\tilde{q}\in\mathcal{Q}_s}\int_S R(\tilde{q}(x,\cdot)\parallel K(x,\cdot))\nu_s(dx),$$

and the right-hand side of this inequality is now $I(\nu_s)$. The two inequalities together with (26) imply

$$I(\nu) \ge (1-p)\cdot I(\nu_\lambda) + p\cdot I(\nu_s).$$

Combined with the opposite inequality (25), this proves the desired equality (24). □

5. Laplace principle upper bound

In this section we prove the Laplace principle upper bound (14).

Proposition 5.1. *Let* $\{L^n\}_{n\ge0}$ *be the empirical measures defined in* (3) *and* $\{x_n\}_{n\ge0}$ *any sequence in* S. *Take* $F \in C_b(\mathcal{P}(S))$ *and define* $I : \mathcal{P}(S) \to [0,\infty]$ *as in* (13). *Assume* (A.1), (A.2) *and* (A.3). *Then,*

$$\liminf_{n\to\infty} -\frac{1}{n}\log\mathbb{E}_{x_n}\left[e^{-nF(L^n)}\right] \ge \inf_{\nu\in\mathcal{P}(S)}[F(\nu) + I(\nu)].$$

As mentioned in Section 4, under (A.1)–(A.3), the arguments from [12] can be used. We include the main steps here for self-containment and convenience of the reader; we emphasize that once the Feller property of $K(x,dy)$ has been established, this part of the proof goes precisely as in [12].

Lemma 5.2. *Under Assumptions* (A.1)–(A.2), *the Metropolis–Hastings transition kernel* $K(x,\cdot)$ *satisfies the Feller property.*

Proof. Recall the form (8) for K, with $a(x,y)$ in (11) corresponding to the probability density of the acceptance part and r corresponding to the rejection part. The assumptions ensure that both a and r are continuous and bounded functions of x (see Remark 3.1). Consider now a function $f \in C_b(S)$, and a sequence $\{x_n\}_{n\in\mathbb{N}} \subset S$ such that $x_n \to x \in S$. By dominated convergence, we have

$$\begin{aligned}\int_S f(y)K(x_n,dy) &= \int_S f(y)a(x_n,y)dy + f(x_n)r(x_n)\\ &\to \int_S f(y)a(x,y)dy + f(x)r(x) = \int_S f(y)K(x,dy).\end{aligned}$$

An application of the Portmanteau theorem then completes the proof. □

Proof of Proposition 5.1. In (16), take a control sequence $\{\bar{\mu}_i^n\}$ such that

$$\mathbb{E}\left[F(\bar{L}^n) + \frac{1}{n}\sum_{i=1}^n R(\bar{\mu}_i^n \parallel K(\bar{X}_{i-1}^n,\cdot))\right] \le \inf_{\{\hat{\mu}_i^n\}} \mathbb{E}\left[F(\hat{L}^n) + \frac{1}{n}\sum_{i=1}^n R(\hat{\mu}_i^n \parallel K(\hat{X}_{i-1}^n,\cdot))\right] + \frac{1}{n},$$

where $\bar{L}^n$ is the controlled empirical measure associated with $\{\bar{\mu}_i^n\}$. Let

$$\lambda^n(dx\times dy) = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{X}_{i-1}^n}(dx)\bar{\mu}_i^n(dy).$$

By Assumption (A.3), $\{(\bar{L}^n,\lambda^n)\}$ is tight; see Section 6.10 in [12]. Thus, there is a subsequence, also denoted by n, such that $\{(\bar{L}^n,\lambda^n)\}$ converges along that subsequence, to some limit $(\bar{L},\lambda)$, and it is enough to prove the upper bound (14) for this subsequence. In fact, taking $n\to\infty$, we have

$$\begin{aligned}\liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{E}_{x_n}\left[e^{-nF(\bar{L}^n)}\right] &\ge \mathbb{E}\left[F(\bar{L}) + R(\lambda \parallel \bar{L}\otimes K)\right]\\ &\ge \inf_{\nu\in\mathcal{P}(S)}\left[F(\nu) + \inf_{\gamma\in A(\nu)} R(\gamma \parallel \nu\otimes K)\right]\\ &= \inf_{\nu\in\mathcal{P}(S)}\left[F(\nu) + I(\nu)\right]. \quad \square\end{aligned}$$

6. Laplace principle lower bound

We now proceed to prove the Laplace principle lower bound (15).

Proposition 6.1. *Let $\{L^n\}_{n\ge 0}$ be the empirical measures defined in (3) and define $I : \mathcal{P}(S) \to [0,\infty]$ as in (13). Assume (A.1)–(A.2). Then, for $x \in S$,*

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{E}_x\left[e^{-nF(L^n)}\right] \le \inf_{\nu\in\mathcal{P}(S)}[F(\nu)+I(\nu)]. \tag{27}$$

As described in Section 4, in proving Theorem 4.1, the lower bound is where the lack of Condition 2.1 for the MH kernel K plays a role. To see why the lack of this transitivity condition becomes an issue, one of the consequences of the condition is that if $\nu \in \mathcal{P}(S)$ is such that $I(\nu) < \infty$, then $\nu \ll \pi$. This property plays an important role in the proof of the LDP for empirical measures of a Markov chain in [12]—it is implicitly used to define a sequence of near-optimal controls in the representation (16). Here, because of the rejection part of the MH kernel, which is the reason Condition 2.1 does not hold, the implication is not true in general. As a counterexample, consider an $x_0 \in S$ such that $r(x_0) > 0$ and take $\nu = \delta_{x_0} \in \mathcal{P}(S)$. Then ν is not absolutely continuous with respect to λ, and thus not with respect to π. Consider the transition kernel $\tilde{q}(x,\cdot) = \delta_x$. Then ν is invariant for $\tilde{q}$ and from (17),

$$I(\nu) \le \int_S R(\delta_x(\cdot) \parallel K(x,\cdot))\nu(dx) = R(\delta_{x_0}(\cdot) \parallel K(x_0,\cdot)).$$

From (22), the Radon–Nikodym derivative of $\delta_{x_0}(\cdot)$ with respect to $K(x,\cdot)$, for $x = x_0$, is given by $f_{x_0}(y) = \frac{1}{r(x_0)} I\{y = x_0\}$. It follows that the rate function is finite, since

$$I(\nu) \le R(\delta_{x_0}(\cdot) \parallel K(x_0,\cdot)) \le \log\frac{1}{r(x_0)} < \infty.$$

We circumvent the problem of not having Condition 2.1 by showing that if $\nu \in \mathcal{P}(S)$ is such that $I(\nu) < \infty$, then there exists another probability measure $\nu^* \in \mathcal{P}(S)$ that is arbitrarily close to ν, and satisfies $\nu^\star \ll \pi$ and $I(\nu^*) \le I(\nu) + \varepsilon$.

To prove the existence of such a measure, recall that the decomposition (23) allows us to separate ν into two parts: one part, ν_λ, with a density with respect to λ (and thus with respect to π) and one, ν_s, that is singular with respect to λ. The idea is to approximate the latter with measures that are absolutely continuous with respect to λ. This allows us to construct near-optimal controls in the representation formula, which in turn are used to prove Proposition 6.1.

The following is a brief outline of the argument.

In Lemma 6.2, we characterize the transition kernels q that achieve the infimum in (17) for $\nu_s \in \mathcal{P}(S)$ such that $\nu_s \perp \lambda$ and $I(\nu_s) < \infty$. Next, in Lemma 6.3, we construct a sequence of random measures $\{\nu_s^n\} \subset \mathcal{P}(S)$ that are absolutely continuous with

respect to λ, $\nu_s^n \Rightarrow \nu_s$ as $n \to \infty$, and $\lim_{n\to\infty} I(\nu_s^n) \leq I(\nu_s)$. This construction allows us to show (Lemma 6.4) that for any $\nu \in \mathcal{P}(S)$ such that $I(\nu) < \infty$, for any $\varepsilon > 0$, there exists a $\nu^\dagger \in \mathcal{P}(S)$ that is arbitrarily close to ν, $\nu^\dagger \ll \lambda$ and $I(\nu^\dagger) \leq I(\nu) + \varepsilon$. The existence of such a probability is then used in Lemma 6.5 to prove the existence of a $\nu^* \in \mathcal{P}(S)$ with the desired properties. From there, the proof of Proposition 6.1 follows largely that of [12].

Lemma 6.2. *Let $\nu_s \in \mathcal{P}(S)$ be such that $\nu_s \perp \lambda$ and $I(\nu_s) < \infty$. Then, $q(x,\cdot) = \delta_x(\cdot)$ ν_s−a.s. is the only transition kernel that satisfies (18), i.e.,*

$$I(\nu_s) = \int_S R(\delta_x(\cdot) \parallel K(x,\cdot))\nu_s(dx) = \int_S \log \frac{1}{r(x)} \nu_s(dx)$$

Proof. By Lemma 4.2, if $I(\nu_s) < \infty$, then the kernels $q(x,\cdot)$ that satisfy (18) are of the form $\alpha(x,\cdot) + \rho(x)\delta_x(\cdot)$, ν_s−a.s. with $\alpha(x,\cdot) \ll a(x,\cdot)$, ν_s−a.s. Moreover, $a(x,\cdot)$ is in turn absolutely continuous with respect to λ. Observe that since the set S_s satisfies $\lambda(S_s) = 0$, then $a(x,S_s) = 0$ and therefore $\alpha(x,S_s) = 0$. On the other hand, $\nu_s(S_s) = 1$ by definition, and by invariance the following must hold:

$$\begin{aligned}
1 &= \nu_s(S_s) \\
&= \int_S q(x,S_s)\nu_s(dx) \\
&= \int_S \left(\alpha(x,S_s) + \rho(x)\delta_x(S_s)\right)\nu_s(dx) \\
&= \int_S \left(0 + \rho(x)\delta_x(S_s)\right)\nu_s(dx) \\
&= \int_{S_s} \rho(x)\nu_s(dx).
\end{aligned}$$

Given that $\rho(x) \leq 1$ $\forall x \in S$, $\int_{S_s} \rho(x)\nu_s(dx) = 1$ can only hold if $\rho(x) \equiv 1$ ν_s−a.s. We conclude that the singular measure ν_s admits only $q(x,\cdot) = \delta_x(\cdot)$ ν_s−a.s. as invariant kernel. This implies that

$$I(\nu_s) = \int_S R(\delta_x(\cdot) \parallel K(x,\cdot))\nu_s(dx).$$

Furthermore, by Proposition 4.3, we have

$$\int_S R(\delta_x(\cdot) \parallel K(x,\cdot))\nu_s(dx) = \int_S \log \frac{1}{r(x)} \nu_s(dx).$$

This completes the proof. □

We now move to the construction of a sequence of random measures $\{\nu_s^n\} \subset \mathcal{P}(S)$ that can be used to approximate ν_s arbitrarily well and satisfy $\lim_{n\to\infty} I(\nu_s^n) \leq I(\nu)$ a.s., while maintaining absolute continuity with respect to λ. To facilitate this, we define, for $\varepsilon > 0$ and $x \in S_+$,

$$\begin{aligned}
\Delta_\varepsilon(x) = \sup\{t \ :\ &|\log a(x,x) - \log a(y,z)| < \varepsilon \text{ and} \\
&|\log r(x) - \log r(y)| < \varepsilon, \qquad \forall y,z \in B_t(x)\}.
\end{aligned} \tag{28}$$

Lemma 6.3. *Take $\nu_s \in \mathcal{P}(S)$ such that $\nu_s \perp \lambda$ and $I(\nu_s) < \infty$. Let $\{Y_i\}_{i=1}^\infty$ be independent and identically distributed according to ν_s. For $n \in \mathbb{N}$, define*

$$\varrho^n = \min\left\{\frac{1}{n}, \min_{1\leq i\leq n} \Delta_{\frac{1}{n}}(Y_i), \frac{1}{2}\min_{Y_i\neq Y_j} d_S(Y_i,Y_j), \frac{1}{2}\min_{1\leq i\leq n} d_S(\partial S, Y_i), \min_{1\leq i\leq n} a(Y_i,Y_i)\right\}. \tag{29}$$

Let $V_n = \lambda(B_{\varrho^n}(0))$, the (Lebesgue) volume of the balls of radius ϱ^n, and define the sequence of random measures $\{\nu_s^n\}_{n\in\mathbb{N}} \subset \mathcal{P}(S)$ by

$$\nu_s^n(dx) := \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n I\{x \in B_{\varrho^n}(Y_i)\}\lambda(dx). \tag{30}$$

This sequence satisfies the following properties:

(a) $\nu_s^n \ll \lambda$ *for all* $n \in \mathbb{N}$,
(b) $\nu_s^n \Rightarrow \nu_s$ *a.s.,*
(c) There is an $n_0 \in \mathbb{N}$ *such that, for all* $n > n_0$, $I(\nu_s^n) < \infty$ *a.s.,*
(d) $\lim_{n\to\infty} I(\nu_s^n) \leq I(\nu_s)$ *a.s.*

Before we embark on the proof, some comments on the construction. First, because we consider ν_s such that $I(\nu_s) < \infty$, ν_s can only put mass on points in S_+: if $\nu_s(x) > 0$ for some x such that $\pi(x) = 0$, then $r(x) = 0$ (see Remark 3.1). By Lemma 4.2, the corresponding transition kernel is of the form $q(x,\cdot) = \alpha(x,\cdot)$, where $\alpha(x,\cdot) \ll a(x,\cdot)$. This is not compatible with ν_s being singular with respect to λ; see also Lemma 6.2. Thus, the Y_is used in the construction are in S_+ ν_s-a.s.

Next, we verify that for any fixed $n \in \mathbb{N}$, the radius ϱ^n of the $B_{\varrho^n}(Y_i)$-balls is well-defined, i.e., $\varrho^n > 0$ ν_s-a.s. Note that if $\nu_s = \delta_x$ for some $x \in S_+$, then ϱ^n becomes

$$\varrho^n = \min\left\{\frac{1}{n}, \Delta_{\frac{1}{n}}(x), \frac{1}{2}d_S(\partial S, x), a(x,x)\right\}.$$

For a generic ν_s such that $I(\nu_s) < \infty$, we have for $Y_i \sim \nu_s$,

$$\mathbb{E}\left[\log\frac{1}{r(Y_i)}\right] = \int_S \log\frac{1}{r(x)}\nu_s(dx) = I(\nu_s) < \infty.$$

It follows that $r(Y_i) > 0$ w.p. 1. From Assumption (A.2) we have $a(Y_i, Y_i) = J(Y_i|Y_i) > 0$. Since the support of ν_s is in S_+ (an open subset of S; see Assumption (A.1)), and $a(Y_i, Y_i)$ and $r(Y_i)$ are both strictly positive ν_s-a.s., the continuity of $r(\cdot)$ and $a(\cdot,\cdot)$ on S and $S_+ \times S$, respectively, ensures that $\Delta_{\frac{1}{n}}(Y_i) > 0$, $i = 1, \dots, n$. Moreover, $d_S(Y_i, Y_j) > 0$ for $Y_i \neq Y_j$ by definition, and $d_S(\partial S, Y_i) > 0$ ν_s-a.s. since the support of ν_s is a subset of S_+, which is an open subset of S. Combined, these show that $\varrho^n > 0$ ν_s-a.s.

Proof of Lemma 6.3. Part (a) follows directly from the definition (30) of ν_s^n. In particular, since λ and π are equivalent measures (Assumption (A.1)), then $\nu_s^n \ll \pi$.

To prove (b), i.e. that the sequence $\{\nu_s^n\}$ converges weakly to ν_s a.s., we show that for any bounded and Lipschitz continuous function f it holds that $\int_S f d\nu_s^n \to \int_S f d\nu_s$ a.s. An application of the Portmanteau theorem then gives the claim.

To this end, let $f \in C_b(S)$ be Lipschitz continuous and denote its Lipschitz constant by $L_f < \infty$. For $n \in \mathbb{N}$, we have

$$\int_S f(x)\nu_s^n(dx) = \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n \int_{B_{\varrho^n}(Y_i)} f(x)\lambda(dx). \tag{31}$$

The Lipschitz continuity of f implies that for all $x \in B_{\varrho^n}(Y_i)$ and for all $i \in \{1, \dots, n\}$,

$$f(Y_i) - L_f \cdot \varrho^n \le f(x) \le f(Y_i) + L_f \cdot \varrho^n.$$

By integrating over $B_{\varrho^n}(Y_i)$ and dividing by V_n, it follows that

$$f(Y_i) - L_f \cdot \varrho^n \le \frac{1}{V_n}\int_{B_{\varrho^n}(Y_i)} f(x)\lambda(dx) \le f(Y_i) + L_f \cdot \varrho^n, \quad i = 1, \dots, n.$$

This implies the following bounds on the integral (31):

$$\frac{1}{n}\sum_{i=1}^n f(Y_i) - L_f \cdot \varrho^n \le \int_S f(x)\nu_s^n(dx) \le \frac{1}{n}\sum_{i=1}^n f(Y_i) + L_f \cdot \varrho^n.$$

By the strong law of large numbers, $\frac{1}{n}\sum_{i=1}^n \delta_{Y_i}(\cdot) \Rightarrow \nu_s(\cdot)$ a.s., and it follows that $\frac{1}{n}\sum_{i=1}^n f(Y_i) \to \int_S f d\nu_s$ a.s. Moreover, by construction $\varrho^n \to 0$ as $n \to \infty$, which implies $L_f \cdot \varrho^n \to 0$. The squeeze theorem now yields the desired result.

We now move to part (c). To show that $I(\nu_s^n)$ is finite for large enough $n \in \mathbb{N}$, we first note that by construction, $V_n \to 0$ as $n \to \infty$. Therefore, there is an $n_0 \in \mathbb{N}$ such that $V_n < 1$ for all $n > n_0$. Henceforth, we only consider such n.

Recall the characterization (17) of the rate function,

$$I(\nu_s^n) = \inf_q \int_S R(q(x,\cdot) \parallel K(x,\cdot))\nu_s^n(dx),$$

where the infimum is taken over all the transition kernels $q(x, dy)$ that are ν_s^n−invariant. We will now construct such a transition kernel $q^n(x, dy)$, for which it also holds that

$$\int_S R(q^n(x,\cdot) \parallel K(x,\cdot))\nu_s^n(dx) < \infty.$$

This in turn implies that $I(\nu_s^n) < \infty$. The collection of transition kernels $\{q^n\}$ will also be used to show part (d).

We begin by defining $N^n(x)$ as the number of $B_{\varrho^n}(Y_i)$, $i = 1, \dots, n$, that $x \in S$ belongs to,

$$N^n(x) = \sum_{i=1}^n I\{x \in B_{\varrho^n}(Y_i)\}.$$

Next, we define q^n by

$$q^n(x, dy) = \frac{1}{N^n(x)}\sum_{i=1}^n I\{x \in B_{\varrho^n}(Y_i)\}I\{y \in B_{\varrho^n}(Y_i)\}dy + \left(1 - V_n\right)\delta_x(dy),$$

for x such that $N^n(x) \ge 1$, and otherwise $q^n(x, dy) = \delta_x(dy)$. Then, for all $x \in S$, $q^n(x,\cdot)$ is a transition probability: if $N^n(x) \ge 1$,

$$q^n(x, S) = \frac{1}{N^n(x)}\sum_{i=1}^n I\{x \in B_{\varrho^n}(Y_i)\}V_n + \left(1 - V_n\right)\delta_x(S) = 1,$$

and, for $N^n(x) = 0$, it holds immediately that $q^n(x, S) = 1$. Moreover, due to the choice of $n > n_0$ $q^n(x, A) \in [0, 1]$, for every $A \in \mathcal{B}(S)$.

To show that $q^n(x,\cdot)$ is also invariant for ν_s^n, consider a set $A \in \mathcal{B}(S)$. We have

$$\nu_s^n(A) = \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n \int_A I\{x \in B_{\varrho^n}(Y_i)\}\lambda(dx) = \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n \lambda(A \cap B_{\varrho^n}(Y_i)).$$

Take $x \in S$. If $N^n(x) \geq 1$,

$$\begin{aligned} q^n(x,A) &= \frac{1}{N^n(x)}\sum_{i=1}^n I\{x \in B_{\varrho^n}(Y_i)\}\int_A I\{y \in B_{\varrho^n}(Y_i)\}\lambda(dy) + \left(1 - V_n\right)\delta_x(A) \\ &= \frac{1}{N^n(x)}\sum_{i=1}^n I\{x \in B_{\varrho^n}(Y_i)\}\lambda(A \cap B_{\varrho^n}(Y_i)) + \left(1 - V_n\right)\delta_x(A). \end{aligned}$$

From this it follows that

$$\begin{aligned} \int_S q^n(x,A)d\nu_s^n(dx) &= \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n \int_{B_{\varrho^n}(Y_i)}\left(\frac{1}{N^n(x)}\sum_{j=1}^n I\{x \in B_{\varrho^n}(Y_j)\}\lambda(A \cap B_{\varrho^n}(Y_j)) + \left(1 - V_n\right)\delta_x(A)\right)\lambda(dx) \\ &= \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n \left(V_n\lambda(A \cap B_{\varrho^n}(Y_i)) + (1 - V_n)\lambda(A \cap B_{\varrho^n}(Y_i))\right) \\ &= \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n \lambda(A \cap B_{\varrho^n}(Y_i)) = \nu_s^n(A), \end{aligned}$$

where in the second equality we use that, due to the definition of ϱ^n, there are no overlaps between the $B_{\varrho^n}(Y_i)$-balls. If instead $N^n(x) = 0$, then $q^n(x,A) = \delta_x(A)$, and we have

$$\int_S \delta_x(A)\nu_s^n(dx) = \int_A \nu_s^n(dx) = \nu_s^n(A).$$

Combined with the computation for $N^n(x) \geq 1$, this proves the invariance.

From (17), $I(\nu_s^n)$ is defined in terms of the infimum over the set of ν_s^n-invariant kernels (17). Therefore,

$$\begin{aligned} I(\nu_s^n) &\leq \int_S R(q^n(x,\cdot) \parallel K(x,\cdot))\nu_s^n(dx) \\ &= \int_{\{x:N^n(x)=0\}} R(q^n(x,\cdot) \parallel K(x,\cdot))\nu_s^n(dx) + \int_{\{x:N^n(x)\geq 1\}} R(q^n(x,\cdot) \parallel K(x,\cdot))\nu_s^n(dx). \end{aligned}$$

For the first integral in the last display, since ν_s^n has no mass on $\{x \in S : N^n(x) = 0\}$, this integral is zero. For the second integral, we have

$$\begin{aligned} &\int_{\{x:N^n(x)\geq 1\}} R(q^n(x,\cdot) \parallel K(x,\cdot))\nu_s^n(dx) \\ &= \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n \int_{B_{\varrho^n}(Y_i)}\left(\int_{B_{\varrho^n}(Y_i)} \log\frac{1}{a(x,y)}\lambda(dy) + (1 - V_n)\cdot\log\frac{1 - V_n}{r(x)}\right)\lambda(dx) \\ &= \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n \left(\iint_{(B_{\varrho^n}(Y_i))^2} \log\frac{1}{a(x,y)}\lambda(dy)\lambda(dx) + (1 - V_n)\int_{B_{\varrho^n}(Y_i)} \log\frac{1 - V_n}{r(x)}\lambda(dx)\right) \end{aligned}$$

Recalling that we only consider $n > n_0$, so that $V_n < 1$, we obtain the upper bound

$$\begin{aligned} &\int_{\{x:N^n(x)\geq 1\}} R(q^n(x,\cdot) \parallel K(x,\cdot))\nu_s^n(dx) \\ &\leq \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n \left(\iint_{(B_{\varrho^n}(Y_i))^2} \log\frac{1}{a(x,y)}\lambda(dy)\lambda(dx) + \int_{B_{\varrho^n}(Y_i)} \log\frac{1}{r(x)}\lambda(dx)\right). \end{aligned} \tag{32}$$

From the definition of ϱ^n (see (29)), it holds that $\varrho^n \leq a(Y_i,Y_i)$ and $\varrho^n \leq \Delta_{\frac{1}{n}}(Y_i)$ for all $i = 1,\dots,n$. Moreover, the definition of $\Delta_{\frac{1}{n}}$ implies that, for a fixed $i = 1,\dots,n$ and $(x,y) \in \left(B_{\varrho^n}(Y_i)\right)^2$,

$$\log\frac{1}{a(x,y)} = -\log a(x,y) + \log a(Y_i,Y_i) - \log a(Y_i,Y_i) < -\log a(Y_i,Y_i) + \frac{1}{n} \leq -\log\varrho^n + \frac{1}{n} = -\log\left(C_d V_n^{\frac{1}{d}}\right) + \frac{1}{n}, \tag{33}$$

for some constant C_d that depends on the dimension d of the space $S \subseteq \mathbb{R}^d$. Similarly, for a fixed $i = 1,\dots,n$ and $x \in B_{\varrho^n}(Y_i)$,

$$\log\frac{1}{r(x)} = -\log r(x) \leq -\log r(Y_i) + \frac{1}{n}. \tag{34}$$

Using the inequalities (33) and (34) in (32) gives the upper bound

$$I(\nu_s^n) \leq \frac{1}{n}\frac{1}{V_n}\sum_{i=1}^n \left(V_n^2\left(-\log\left(C_d V_n^{\frac{1}{d}}\right) + \frac{1}{n}\right) + V_n\left(-\log r(Y_i) + \frac{1}{n}\right)\right)$$

$$= -V_n \log\left(C_d V_n^{\frac{1}{d}}\right) + \frac{V_n}{n} + \frac{1}{n}\sum_{i=1}^{n} \log\frac{1}{r(Y_i)} + \frac{1}{n},$$

whenever $n > n_0$. Since $V_n \to 0$ by construction, we conclude that

$$\lim_{n\to\infty} I(\nu_s^n) \le \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log\frac{1}{r(Y_i)} = \int_S \log\frac{1}{r(x)}\nu_s(dx) = I(\nu_s) \quad \text{a.s.},$$

where the second-to-last equality follows from the strong law of large numbers, and the last equality is motivated by Lemma 6.2. □

Lemma 6.4. *Let $\nu \in \mathcal{P}(S)$ be such that $I(\nu) < \infty$. Take $\varepsilon > 0$ and $\delta > 0$. There exists a probability measure $\nu^\dagger \in \mathcal{P}(S)$ absolutely continuous with respect to the Lebesgue measure and such that*

$$d_{LP}(\nu^\dagger, \nu) < \frac{\delta}{2} \quad \text{and} \quad I(\nu^\dagger) < I(\nu) + \varepsilon.$$

Proof. First, if $\nu \ll \lambda$ there is nothing to prove. Therefore, suppose this does not hold and the decomposition (23) is non-trivial.

Sample $\{Y_i\}_{i=1}^{\infty}$ i.i.d. ν_s and define the sequence of random probability measures $\{\nu_s^n\}_{n\in\mathbb{N}}$ as in the construction in Lemma 6.3. Motivated by the decomposition (23) for ν, we define a new sequence of random probability measures $\{\nu^n\}_{n\in\mathbb{N}}$ by

$$\nu^n = (1-p)\cdot\nu_\lambda + p\cdot\nu_s^n,$$

where $p \in [0,1]$ is the same as in (23), again suppressing in the notation that p depends on ν. By part (a) of Lemma 6.3, $\nu_s^n \ll \lambda$ for all n. It follows that $\nu^n \ll \lambda$. Moreover, from part (b) of the same Lemma, ν^n converges weakly to ν ν_s-a.s. Therefore, for any $\omega \in \Omega$ outside of a ν_s-null set, there is an $N_\delta = N_\delta(\omega) \in \mathbb{N}$ such that

$$d_{LP}(\nu^n(\omega), \nu) < \frac{\delta}{2}, \quad \forall n \ge N_\delta(\omega).$$

Consider now $I(\nu^n)$. By convexity,

$$I(\nu^n) \le (1-p)\cdot I(\nu_\lambda) + p\cdot I(\nu_s^n),$$

for which the right-hand-side is finite w.p. 1 whenever $n \ge n_0$. Combined with part (d) of Lemma 6.3, this yields that, ν_s-a.s.,

$$\lim_{n\to\infty} I(\nu^n) \le (1-p)I(\nu_\lambda) + p\cdot I(\nu_s) = I(\nu).$$

Similar to before, this implies that for any $\omega \in \Omega$ outside of a ν_s-null set, there is a $N_\varepsilon = N_\varepsilon(\omega) \in \mathbb{N}$, such that

$$I(\nu^n(\omega)) < I(\nu) + \varepsilon, \quad \forall n \ge N_\varepsilon(\omega).$$

As a consequence, for any $\omega \in \Omega$ outside of a null set, we can define

$$N(\omega) = \max\{N_\delta(\omega), N_\varepsilon(\omega), n_0\}.$$

Then, for $n \ge N(\omega)$, $\nu^n(\omega) \ll \lambda$, $d_{LP}(\nu^n(\omega), \nu) < \delta/2$, and $I(\nu^n(\omega)) < I(\nu) + \varepsilon$. Since this is outside a ν_s-null set, it has positive probability also under ν. This proves the existence of a measure $\nu^\dagger$ with the claimed properties. □

We emphasize that the randomness of the sequence $\{\nu^n\}$ is entirely due to the sequence of random variables $\{Y_i\}_{i=1}^{\infty}$. Thus, the set of outcomes of $\{Y_i\}$ that lead to a measure ν^n with the desired properties is a set with strictly positive probability. This guarantees the existence of a measure $\nu^\dagger$ with the claimed properties. The following result is a version of Lemma 6.17 in [12].

Lemma 6.5. *Let $\nu \in \mathcal{P}(S)$ satisfy $I(\nu) < \infty$. Under (A.1)–(A.2), for given $\varepsilon > 0$ and $\delta > 0$, there exists $\nu^* \in \mathcal{P}(S)$ with the following properties:*

- *(a) $d_{LP}(\nu^*, \nu) < \delta$;*
- *(b) $\nu^* \ll \pi$ and $\pi \ll \nu^*$;*
- *(c) there exists a transition probability function $q^*(x, dy)$ on S such that ν^* is an invariant measure of $q^*(x, dy)$, the associated Markov chain is ergodic, and*

$$I(\nu^*) < I(\nu) + \varepsilon. \tag{35}$$

Proof. To prove (a), by Lemma 6.4, there exists a measure $\nu^\dagger$ that satisfies

$$d_{LP}(\nu^\dagger, \nu) < \frac{\delta}{2} \quad \text{and} \quad I(\nu^\dagger) < I(\nu) + \varepsilon \tag{36}$$

Define ν^* by

$$\nu^* = \left(1 - \frac{\delta}{4}\right)\nu^\dagger + \frac{\delta}{4}\pi. \tag{37}$$

Then,

$$d_{LP}(\nu^*, \nu^\dagger) \le \|\nu^* - \nu^\dagger\|_{TV} = \left\|\left(1-\frac{\delta}{4}\right)\nu^\dagger + \frac{\delta}{4}\pi - \nu^\dagger\right\|_{TV} = \frac{\delta}{4}\|\pi - \nu^\dagger\|_{TV} \le \frac{\delta}{2}.$$

Combining this with (36) and the triangle inequality now yields the desired upper bound on $d_{LP}(\nu^*, \nu)$,

$$d_{LP}(\nu^*, \nu) \le d_{LP}(\nu^\dagger, \nu) + d_{LP}(\nu^*, \nu^\dagger) < \delta.$$

(b). The first part of follows from $\nu^\dagger \ll \lambda$ (see Lemma 6.4) and the fact that, by Assumption (A.1), $\lambda \ll \pi$. For the second part, for any $A \in \mathcal{B}(S)$, $\nu^*(A) \ge \frac{\delta}{4}\pi(A)$ by construction. Thus, $\pi \ll \nu^*$.

We now prove part (c), following the steps in [12, Lemma 6.17]. Since $I(\nu^\dagger) < \infty$, by Lemma 6.8(b) in [12], we can choose a transition kernel $q(x, dy)$ with invariant measure $\nu^\dagger$ and

$$\int_S R(q(x,\cdot) \parallel K(x,\cdot))\nu^\dagger(dx) = I(\nu^\dagger).$$

Define the $\gamma^\dagger$, θ and γ^* in $\mathcal{P}(S^2)$ by,

$$\gamma^\dagger = \nu^\dagger \otimes q,$$
$$\theta = \pi \otimes K,$$

and

$$\gamma^* = \left(1 - \frac{\delta}{4}\right)\gamma^\dagger + \frac{\delta}{4}\theta.$$

Both marginals of $\gamma^\dagger$ equal $\nu^\dagger$. Similarly, both marginals of θ equal π. From (37) it then follows that both marginals of γ^* equal ν^*. From Lemma 6.8(a) in [12], there exists a transition kernel $q^*(x, dy)$ that has ν^* as invariant probability distribution and such that $\gamma^* = \nu^* \otimes q^*$. Using the convexity of relative entropy, the property $R(\alpha \parallel \alpha) = 0$ and (36), we obtain the upper bound (35):

$$\begin{aligned}
I(\nu^*) &\le \int_S R(q^*(x,\cdot) \parallel K(x,\cdot))\nu^*(dx) \\
&= R(\gamma^* \parallel \nu^* \otimes K) \\
&= R\left(\left(1 - \frac{\delta}{4}\right)\nu^\dagger \otimes q + \frac{\delta}{4}\pi \otimes K \middle\| \left(1 - \frac{\delta}{4}\right)\nu^\dagger \otimes K + \frac{\delta}{4}\pi \otimes K\right) \\
&= \left(1 - \frac{\delta}{4}\right) R(\nu^\dagger \otimes q \parallel \nu^\dagger \otimes K) + \frac{\delta}{4} R(\pi \otimes K \parallel \pi \otimes K) \\
&= \left(1 - \frac{\delta}{4}\right) I(\nu^\dagger) \\
&< I(\nu) + \varepsilon.
\end{aligned}$$

It remains to show that the Markov process associated with q^* is ergodic. Let $f = \frac{d\nu^*}{d\pi}$ be the Radon–Nikodym derivative of ν^* with respect to π, which is well-defined by part (b). Since $\nu^*(A) \ge \frac{\delta}{4}\pi(A)$ for all $A \in \mathcal{B}(S)$, for all $x \in S$, $f(x) \ge \frac{\delta}{4}$. We observe that for any $A, B \in \mathcal{B}(S)$,

$$\gamma^*(A \times B) = \int_A q^*(x, B)\nu^*(dx) = \int_A q^*(x, B) f(x)\pi(dx),$$

and, from the definition of γ^*,

$$\gamma^*(A \times B) \ge \frac{\delta}{4}\theta(A \times B) = \frac{\delta}{4}\int_A K(x, B)\pi(dx).$$

It follows that

$$q^*(x, B) \ge \frac{\delta}{4f(x)} K(x, B), \quad \forall x,\ \pi - \text{a.s.},$$

for all $B \in \mathcal{B}(S)$. Thus, π-a.s. for $x \in S$, $K(x,\cdot) \ll q^*(x,\cdot)$. To show absolute continuity in the reverse direction, note that from

$$\int_S R(q^*(x,\cdot) \parallel K(x,\cdot))\nu^*(dx) < \infty,$$

it follows that $R(q^*(x,\cdot) \parallel K(x,\cdot)) < \infty$. Thus, $q^*(x,\cdot) \ll K(x,\cdot)$, ν^*-a.s. As ν^* and π are equivalent measures, we obtain that $q^*(x,\cdot)$ and $K(x,\cdot)$ are equivalent π-a.s. This means that there exists a Borel set $C \in \mathcal{B}(S)$ such that $\pi(C) = 0 = \nu^*(C)$, and $q^*(x,\cdot)$ and $K(x,\cdot)$ are equivalent for all x in the complement of C. If we redefine $q^*(x,\cdot) = K(x,\cdot)$ for all $x \in C$, we obtain the equivalence between $q^*(x, dy)$ and $K(x,\cdot)$ for all $x \in S$. Besides, being $\nu^*(C) = 0$, the newly defined $q^*(x,\cdot)$ still has ν^* as invariant measure. To show that $q^*(x,\cdot)$ is ergodic, recall that in Remark 3.2 we proved that there are no disjoint Borel sets $A_1, A_2 \in \mathcal{B}(S)$ such that

$$K(x, A_1) = 1\ \forall x \in A_1 \quad \text{and} \quad K(y, A_2) = 1\ \forall y \in A_2.$$

Because $q^*(x,\cdot)$ and $K(x,\cdot)$ are equivalent for all $x \in S$, it follows that also $q^*(x,\cdot)$ satisfies the property that there do not exist disjoint $A_1, A_2 \in \mathcal{B}(S)$ for which

$$q^*(x, A_1) = 1\ \forall x \in A_1 \quad \text{and} \quad q^*(y, A_2) = 1\ \forall y \in A_2,$$

meaning that $q^*(x,\cdot)$ is indecomposable. Therefore, by Theorem 7.16 in [10], ν^* is the unique invariant distribution for $q^*(x, dy)$ and the Markov chain associated with ν^* and $q^*(x, dy)$ is ergodic. □

We are ready to prove Proposition 6.1, the Laplace principle lower bound. The following proof is mostly based on the proof of Proposition 6.15 in [12], with minor changes due to the lack of Condition 2.1. The main work has been done in Lemmas 6.2–6.5, and most of the proof from [12] now goes through, with some minor modifications to rely on those results rather than Condition 2.1. We include the full argument for self-containment and convenience for the reader.

Proof of Proposition 6.1. To prove the Laplace lower bound (27), it is sufficient to consider only bounded Lipschitz continuous functions F (see Corollary 1.10 in [12]). Since we have endowed $\mathcal{P}(S)$ with the Lévy–Prohorov metric d_{LP}, a function $F \in C_b(\mathcal{P}(S))$ is Lipschitz if

$$\sup_{\nu_1 \neq \nu_2} \frac{|F(\nu_1) - F(\nu_2)|}{d_{LP}(\nu_1, \nu_2)} < \infty.$$

Recall that $\{X_i\}_{i\geq 0}$ denotes the Metropolis–Hastings chain, as described in Section 2.3, and $\{L^n\}_n$ the associated sequence of empirical measures. We now construct a nearly optimal sequence of controls in the variational representation (16),

$$-\frac{1}{n}\log \mathbb{E}_x\left[e^{-nF(L^n)}\right] = \inf_{\{\bar{\mu}_i^n\}} \mathbb{E}_x\left[F(\bar{L}^n) + \frac{1}{n}\sum_{i=1}^{n} R(\bar{\mu}_i^n \parallel K(\bar{X}_{i-1}^n, \cdot))\right]. \tag{38}$$

Let $\varepsilon > 0$ be given and choose $\nu \in \mathcal{P}(S)$ such that

$$F(\nu) + I(\nu) \leq \inf_{\mu \in \mathcal{P}(S)} [F(\mu) + I(\mu)] + \varepsilon < \infty \tag{39}$$

Since F is continuous, there exists $\delta > 0$ such that $d_{LP}(\mu, \nu) < \delta$ implies $|F(\mu) - F(\nu)| < \varepsilon$. In Lemma 6.5 it is shown that, for any such pair δ, ε, there exists a probability measure $\nu^* \in \mathcal{P}(S)$ and a transition probability $q^*(x, dy)$ such that ν^* is an invariant measure for $q^*(x, dy)$, the Markov chain with initial distribution ν^* and transition probability $q^*(x, dy)$ is ergodic, and

$$I(\nu^*) \leq \int_S R(q^*(x, \cdot) \parallel K(x, \cdot))\nu^*(dx) < I(\nu) + \varepsilon < \infty. \tag{40}$$

Moreover, Part (a) of the Lemma ensures $d_{LP}(\nu^*, \nu) < \delta$, which then implies

$$F(\nu^*) < F(\nu) + \varepsilon. \tag{41}$$

Thus, ν^* is such that

$$F(\nu^*) + I(\nu^*) < F(\nu) + I(\nu) + 2\varepsilon.$$

The transition probability function q^* associated with ν^* is now used to define the controls,

$$\bar{\mu}_i^n(dy) = q^*(\bar{X}_{i-1}^n, dy), \quad i = 1, \dots, n. \tag{42}$$

With the inequalities (40)–(41) established, and the choice (42) for the controls, we can proceed with the same arguments as in the proof of Proposition 6.15 in [12].

With the choice (42), the running costs for the controlled chain $\bar{X}^n$ are

$$\frac{1}{n}\sum_{i=1}^{n} R(\bar{\mu}_i^n(\cdot) \parallel K(\bar{X}_{i-1}^n, \cdot)) = \frac{1}{n}\sum_{i=0}^{n-1} R(q^*(\bar{X}_i^n, \cdot) \parallel K(\bar{X}_i^n, \cdot)).$$

The $\bar{\mu}_i^n$s only give the conditional distributions for $\bar{X}_i^n$ for $i = 1, \dots, n$. For the distribution of the initial point $\bar{X}_0^n$, consider two choices: δ_x and ν^*. Let $\mathbb{P}_x$ and $\mathbb{P}^*$ denote the corresponding probability measures and let $\mathbb{E}_x$ and $\mathbb{E}^*$ be the associated expectation, respectively. Define D^n and D_x^n as the expected difference between the empirical average of the relative entropy between q^* and K, and its mean, under $\mathbb{P}^*$ and $\mathbb{P}_x$, respectively,

$$D^n = \mathbb{E}^*\left[\left|\frac{1}{n}\sum_{i=0}^{n-1} R(q^*(\bar{X}_i^n, \cdot) \parallel K(\bar{X}_i^n, \cdot)) - \int_S R(q^*(\xi, \cdot) \parallel K(\xi, \cdot))\nu^*(d\xi)\right|\right],$$

and

$$D_x^n = \mathbb{E}_x\left[\left|\frac{1}{n}\sum_{i=0}^{n-1} R(q^*(\bar{X}_i^n, \cdot) \parallel K(\bar{X}_i^n, \cdot)) - \int_S R(q^*(\xi, \cdot) \parallel K(\xi, \cdot))\nu^*(d\xi)\right|\right].$$

From the definition of the controls (42), and since ν^* is an invariant measure of $q^*(x, dy)$, all terms of the controlled process $\{\bar{X}_i^n\}_{i=0}^n$ with ν^* as initial distribution are distributed according to ν^*. By the non-negativity of the relative entropy and $R(\cdot \parallel \cdot)$ and (40), we obtain

$$\mathbb{E}^*\left[\left|R(q^*(\bar{X}_i^n, \cdot) \parallel K(\bar{X}_i^n, \cdot))\right|\right] = \int_S R(q^*(\xi, \cdot) \parallel K(\xi, \cdot))\nu^*(d\xi) < I(\nu) + \varepsilon < \infty.$$

The L^1-ergodic theorem [10, Corollary 6.25] then gives

$$\lim_{n\to\infty} D^n = 0.$$

Moreover, note that $D^n = \int_S D_x^n \nu^*(dx)$. Therefore, the convergence of D^n also implies that

$$\lim_{n\to\infty} \int_S D_x^n \nu^*(dx) = 0.$$

Convergence in probability of D_x^n to 0 now follows from Chebyshev's inequality: for any $c > 0$,

$$\lim_{n\to\infty} \nu^*\{x \in S \ : \ D_x^n \geq c\} \leq \lim_{n\to\infty} \frac{1}{c} \int_{\{x : D_x^n \geq c\}} D_x^n \nu^*(dx) \leq \frac{1}{c} \lim_{n\to\infty} D^n = 0.$$

From this convergence in probability, for every subsequence of $\{n\}$ there is a further subsequence, which we also denote by $\{n\}$, such that the convergence is w.p. 1. That is, there is a Borel set Φ_1 with $\nu^*(\Phi_1) = 1$, such that along such (sub)subsequences and for all $x \in \Phi_1$,

$$\lim_{n\to\infty} \mathbb{E}_x \left| \frac{1}{n} \sum_{i=0}^{n-1} R(q^*(\bar{X}_i^n, \cdot) \parallel K(\bar{X}_i^n, \cdot)) - \int_S R(q^*(\xi, \cdot) \parallel K(\xi, \cdot)) \nu^*(d\xi) \right| = 0. \tag{43}$$

Abusing notation, we now fix such a subsubsequence $\{n\}$. The previous argument shows the a.s. convergence of the running costs and we now consider the corresponding sequence of controlled empirical measures $\{\bar{L}^n\}$. Because $S \subset \mathbb{R}^d$, there is a countable convergence-determining class $\Xi \subset C_b(S)$ (see e.g. Appendix A in [12]). For each $g \in \Xi$, we define the set

$$\mathcal{A}(g) = \left\{ \omega \in \Omega \ : \ \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} g(\bar{X}_i^n(\omega)) = \int_S g(x) \nu^*(dx) \right\}.$$

By the pointwise ergodic theorem [10, Sect. 6.5],

$$\mathbb{P}^* \{\mathcal{A}(g)\} = 1.$$

Observing that $\mathbb{P}^* \{\mathcal{A}(g)\} = \int_S \mathbb{P}_x \{\mathcal{A}(g)\} \nu^*(dx)$, we obtain

$$\int_S \mathbb{P}_x \{\mathcal{A}(g)\} \nu^*(dx) = 1.$$

This implies that $\mathbb{P}_x \{\mathcal{A}(g)\} = 1$ a.s., i.e., there exists a Borel set $\Phi_2(g) \in \mathcal{B}(S)$ with $\nu^*(\Phi_2(g)) = 1$ and such that $\mathbb{P}_x \{\mathcal{A}(g)\} = 1$ for $x \in \Phi_2(g)$.

To establish the convergence of $\bar{L}^n$, we define $\Phi_2 = \cap_{g\in\Xi} \Phi_2(g)$. Since Ξ is countable, Φ_2 satisfies $\nu^*(\Phi_2) = 1$. Then, for all initial points $\bar{X}_0^n = x \in \Phi_2$,

$$\lim_{n\to\infty} \int_S g \, d\bar{L}^n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} g(\bar{X}_i^n) = \int_S g \, d\nu^*,$$

$\mathbb{P}_x$−a.s. for all $g \in \Xi$. Because Ξ a convergence determining class, it follows that $\bar{L}^n \Rightarrow \nu^*$ $\mathbb{P}_x$−a.s. for all $x \in \Phi_2$. From the continuity of F on $\mathcal{P}(S)$, we then have

$$\lim_{n\to\infty} F(\bar{L}^n) = F(\nu^*), \tag{44}$$

for all $x \in \Phi_2$.

We now combine the arguments for the running costs and the controlled empirical measures to show the Laplace principle lower bound on a set of ν^*-measure 1. Define the set $\Phi = \Phi_1 \cap \Phi_2 \subset S$. Since $\nu^*(\Phi) = \nu^*(\Phi_2) = 1$, we have $\nu^*(\Phi) = 1$. For all $x \in \Phi$, both (43) and (44) hold, and

$$\begin{aligned}
\limsup_{n\to\infty} -\frac{1}{n} \log \mathbb{E}_x e^{-nF(L^n)} &\leq \lim_{n\to\infty} \mathbb{E}_x \left[F(\bar{L}^n) + \frac{1}{n} \sum_{i=0}^{n-1} R(q^*(\bar{X}_i^n, \cdot) \parallel K(\bar{X}_i^n, \cdot)) \right] \\
&= F(\nu^*) + \int_S R(q^*(\xi, \cdot) \parallel K(\xi, \cdot)) \nu^*(d\xi) \\
&\leq F(\nu) + I(\nu) + 2\varepsilon \\
&\leq \inf_{\mu\in\mathcal{P}(S)} [F(\mu) + I(\mu)] + 3\varepsilon,
\end{aligned}$$

where the inequality on the third line comes from (40) and (41), while the inequality on the last line follows from (39). By taking the limit $\varepsilon \to 0$ we obtain the upper bound (27) for all $x \in \Phi$.

We conclude the proof by extending this result from Φ to the whole space S. Whereas [12] relies on the transitivity condition (5) for this extension, we instead rely on the properties of the MH kernel; this requires only minor changes in the argument.

By Lemma 6.5, ν^* and π are equivalent, thus $\nu^*(\Phi) = 1$ implies $\pi(\Phi) = 1$. Moreover, π and λ are equivalent measures by Assumption (A.1), and we have

$$a(x, \Phi) = \int_\Phi a(x, y) dy = \int_S a(x, y) dy = a(x, S),$$

for all $x \in S$. As a consequence, $K(x, \Phi) \geq a(x, \Phi) = a(x, S)$, which is strictly positive for all $x \in S$ (see Remark 3.2). It follows that

$$K(x, \Phi) > 0, \qquad \forall x \in S. \tag{45}$$

Define $\tilde{L}^n$ as the empirical measure of $X_1, \dots, X_n$,

$$\tilde{L}^n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

Since L^n and $\tilde{L}^n$ only differ in the first and last summands,

$$\|\tilde{L}^n - L^n\|_{TV} \leq \frac{2}{n}.$$

Let $L_F < \infty$ denote the Lipschitz constant of F with respect to the Lévy–Prohorov metric. For all $\omega \in \Omega$,

$$F(L^n) \leq F(\tilde{L}^n) + L_F \cdot d_{LP}(L^n, \tilde{L}^n) \leq F(\tilde{L}^n) + L_F \|L^n - \tilde{L}^N\|_{TV} \leq F(\tilde{L}^n) + \frac{2L_F}{n}.$$

Take any $x \in S$ and $n \in \mathbb{N}$. Since the X_i^ns evolve according to K, using the previous inequality we have,

$$\begin{aligned}
\mathbb{E}_x\left[e^{-nF(L^n)}\right] &\geq e^{-2L_F}\mathbb{E}_x\left[e^{-nF(\tilde{L}^n)}\right] \\
&= e^{-2L_F}\int_S \mathbb{E}\left[e^{-nF(\tilde{L}^n)} \mid X_1 = y\right] K(x, dy) \\
&= e^{-2L_F}\int_S \mathbb{E}_y\left[e^{-nF(L^n)}\right] K(x, dy) \\
&\geq e^{-2L_F}\int_{\Phi} \mathbb{E}_y\left[e^{-nF(\tilde{L}^n)}\right] K(x, dy),
\end{aligned} \tag{46}$$

where the equality on the third line is due to the Markov property. With this lower bound established, from here we can again follow the proof of Proposition 6.15 in [12]. Let $\varepsilon > 0$ be fixed. We have that (27) holds for all $y \in \Phi$, why for each such y there exists an $N(y, \varepsilon) \in \mathbb{N}$ such that

$$-\frac{1}{n}\log \mathbb{E}_y\left[e^{-nF(L^n)}\right] \leq \inf_{\mu \in \mathcal{P}(S)} [F(\mu) + I(\mu)] + \varepsilon \tag{47}$$

for all $n \geq N(y, \varepsilon)$. Without loss of generality, take $N(y, \varepsilon)$ as the smallest integer with this property. Then, the function $S \to \mathbb{N}$ that maps y into $N(y, \varepsilon)$ is measurable, the sets

$$\Phi^{(i)} = \{y \in \Phi \ : \ N(y, \varepsilon) = i\} \subset S$$

are disjoint Borel sets, and $\Phi = \cup_{i=1}^{\infty} \Phi^{(i)}$.

Because $K(x, \Phi) > 0$ for all $x \in S$ (see (45)), we have that for all $x \in S$ there exists an $i_0 \in \mathbb{N}$ such that $K(x, \Phi^{(i_0)}) > 0$. Combined with the bounds in (46), and (47), this implies that for all $n \geq i_0$,

$$\begin{aligned}
\mathbb{E}_x\left[e^{-nF(L^n)}\right] &\geq e^{-2L_F}\int_{\Phi} \mathbb{E}_y\left[e^{-nF(\tilde{L}^n)}\right] K(x, dy) \\
&\geq e^{-2L_F}\int_{\Phi^{(i_0)}} \mathbb{E}_y\left[e^{-nF(\tilde{L}^n)}\right] K(x, dy) \\
&\geq e^{-2L_F} \exp\left\{-n\left(\inf_{\mu \in \mathcal{P}(S)} [F(\mu) + I(\mu)] + \varepsilon\right)\right\} K(x, \Phi^{(i_0)}).
\end{aligned}$$

It follows that

$$\begin{aligned}
\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{E}_x\left[e^{-nF(L^n)}\right] &\leq \inf_{\mu \in \mathcal{P}(S)} [F(\mu) + I(\mu)] + \varepsilon + \lim_{n\to\infty} \frac{2L_F - \log K(x, \Phi^{(i_0)})}{n} \\
&= \inf_{\mu \in \mathcal{P}(S)} [F(\mu) + I(\mu)] + \varepsilon.
\end{aligned}$$

In the limit $\varepsilon \to 0$, we have for all $x \in S$,

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{E}_x\left[e^{-nF(L^n)}\right] \leq \inf_{\mu \in \mathcal{P}(S)} [F(\mu) + I(\mu)] + \varepsilon.$$

This concludes the proof of the Laplace principle lower bound. □

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Professors I. Kontoyiannis and S. P. Meyn for comments on the first version of the paper, and for pointing out their previous work [31,32], and Prof. A. Wang for insightful comments that lead to a refinement of Assumption (A.2). We also thank the anonymous referee for constructive comments and questions that have helped to improve the clarity of this paper.

The research of FM and PN was supported by the Swedish e-Science Research Centre (SeRC). PN was also supported by Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, Sweden, and by the Swedish Research Council (VR-2018-07050, VR-2023-03484).

References

[1] C. Andrieu, N. de Freitas, A. Doucet, M.I. Jordan, An introduction to MCMC for machine learning, Mach. Learn. 50 (1) (2003) 5–43.
[2] C. Andrieu, A. Lee, S. Power, A.Q. Wang, Explicit convergence bounds for Metropolis Markov chains: isoperimetry, spectral gaps and profiles, 2022, Preprint; arXiv:2211.08959.
[3] C. Andrieu, A. Lee, S. Power, A.Q. Wang, Poincaré inequalities for Markov chains: a meeting with cheeger, Lyapunov and Metropolis, 2022, Preprint; arXiv:2208.05239.
[4] S. Asmussen, P.W. Glynn, Stochastic Simulation : Algorithms and Analysis, in: Stochastic Modelling and Applied Probability, Springer, New York, 2007.
[5] M.A. Beaumont, Approximate Bayesian computation, Annu. Rev. Stat. Appl. 6 (2019) 379–403.
[6] M. Bédard, J.S. Rosenthal, Optimal scaling of Metropolis algorithms: heading toward general target distributions, Can. J. Statist. 36 (4) (2008) 483–503.
[7] J. Besag, Comments on representations of knowledge in complex systems by U. Grenander and M. I. Miller, J. R. Stat. Soc. Ser. B Stat. Methodol. 56 (591–592) (1994) 4.
[8] J. Bierkens, Non-reversible Metropolis-Hastings, Stat. Comput. 26 (6) (2016) 1213–1228.
[9] J. Bierkens, P. Nyquist, M.C. Schlottke, Large deviations for the empirical measure of the zig-zag process, Ann. Appl. Probab. 31 (6) (2021) 2811–2843.
[10] L. Breiman, Probability, Society for Industrial and Applied Mathematics, 1992.
[11] J.A. Bucklew, Introduction to Rare Event Simulation, in: Springer Series in Statistics, Springer-Verlag, New York, 2004.
[12] A. Budhiraja, P. Dupuis, Analysis and Approximation of Rare Events: Representations and Weak Convergence Methods, in: Probability Theory and Stochastic Modelling Ser., vol. 94, Springer, New York, NY, 2019.
[13] O.F. Christensen, G.O. Roberts, J.S. Rosenthal, Scaling limits for the transient phase of local metropolis–hastings algorithms, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2) (2005) 253–268.
[14] A. Dembo, O. Zeitouni, Large deviations techniques and applications, Appl. Math. (NY) 38 (1994).
[15] P. Diaconis, S. Holmes, R.M. Neal, Analysis of a nonreversible Markov chain sampler, Ann. Appl. Probab. 10 (3) (2000) 726–752.
[16] J.D. Doll, P. Dupuis, P. Nyquist, A large deviation analysis of certain qualitative properties of parallel tempering and infinite swapping algorithms, Appl. Math. Optim. 78 (1) (2018) 103–144.
[17] M.D. Donsker, S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time, I, Comm. Pure Appl. Math. 28 (1) (1975a) 1–47.
[18] M.D. Donsker, S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time, II, Comm. Pure Appl. Math. 28 (2) (1975b) 279–301.
[19] M.D. Donsker, S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time, III, Comm. Pure Appl. Math. 29 (4) (1976) 389–461.
[20] R. Douc, E. Moulines, P. Priouret, P. Soulier, Markov Chains, Springer, 2018.
[21] P. Dupuis, R.S. Ellis, A Weak Convergence Approach to the Theory of Large Deviations, in: Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1997.
[22] P. Dupuis, Y. Liu, On the large deviation rate function for the empirical measures of reversible jump Markov processes, Ann. Probab. 43 (3) (2015) 1121–1156.
[23] P. Dupuis, Y. Liu, N. Plattner, J.D. Doll, On the infinite swapping limit for parallel tempering, Multiscale Model. Simul. 10 (3) (2012) 986–1022.
[24] P. Dupuis, G.-J. Wu, Analysis and optimization of certain parallel monte carlo methods in the low temperature limit, Multiscale Model. Simul. 20 (1) (2022) 220–249.
[25] J. Feng, T.G. Kurtz, Large Deviations for Stochastic Processes, in: Mathematical Surveys and Monographs, American Mathematical Society, 2006.
[26] B. Franke, C.-R. Hwang, H.-M. Pai, S.-J. Sheu, The behavior of the spectral gap under growing drift, Trans. Amer. Math. Soc. 362 (3) (2010) 1325–1350.
[27] A. Frigessi, P. di Stefano, C.-R. Hwang, S.J. Sheu, Convergence rates of the gibbs sampler, the Metropolis algorithm and other single-site updating dynamics, J. R. Stat. Soc. Ser. B Stat. Methodol. 55 (1) (1993) 205–219.
[28] A. Gelman, W.R. Gilks, G.O. Roberts, Weak convergence and optimal scaling of random walk metropolis algorithms, Ann. Appl. Probab. 7 (1) (1997) 110–120.
[29] W. Hastings, Monte Carlo sampling methods using Markov chains and their application, Biometrika 57 (1970) 97–109.
[30] C.-R. Hwang, S.-Y. Hwang-Ma, S.-J. Sheu, Accelerating diffusions, Ann. Appl. Probab. 15 (2) (2005) 1433–1444.
[31] I. Kontoyiannis, S.P. Meyn, Spectral theory and limit theorems for geometrically ergodic Markov processes, Ann. Appl. Probab. 13 (1) (2003) 304–362.
[32] I. Kontoyiannis, S.P. Meyn, Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes, Electron. J. Probab. 10 (2005) 61–123.
[33] P. Marjoram, J. Molitor, V. Plagnol, S. Tavaré, Markov chain Monte Carlo without likelihoods, Proc. Natl. Acad. Sci. USA 100 (26) (2003) 15324–15328.
[34] K.L. Mengersen, R.L. Tweedie, Rates of convergence of the Hastings and Metropolis algorithms, Ann. Statist. 24 (1) (1996) 101–121.
[35] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (6) (1953) 1087.
[36] S.P. Meyn, R.L. Tweedie, Markov Chains and Stochastic Stability, Springer Science & Business Media, 2012.
[37] N. Plattner, J.D. Doll, P. Dupuis, H. Wang, Y. Liu, J.E. Gubernatis, An infinite swapping approach to the rare-event sampling problem, J. Chem. Phys. 135 (13) (2011) 134111.
[38] L. Rey-Bellet, K. Spiliopoulos, Irreversible langevin samplers and variance reduction: a large deviations approach, Nonlinearity 28 (7) (2015) 2081.
[39] L. Rey-Bellet, K. Spiliopoulos, Variance reduction for irreversible langevin samplers and diffusion on graphs, Electron. Commun. Probab. 20 (2015).
[40] L. Rey-Bellet, K. Spiliopoulos, Improving the convergence of reversible samplers, J. Stat. Phys. 164 (3) (2016) 472–494.
[41] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, second ed., Springer, New York, New York, NY, 2004.
[42] G.O. Roberts, J.S. Rosenthal, Geometric ergodicity and hybrid Markov chains, Electron. Commun. Probab. 2 (1997) 13–25.
[43] G.O. Roberts, J.S. Rosenthal, Optimal scaling of discrete approximations to langevin diffusions, J. R. Stat. Soc. Ser. B Stat. Methodol. 60 (1) (1998) 255–268.
[44] G.O. Roberts, J.S. Rosenthal, Optimal scaling for various Metropolis–Hastings algorithms, Stat. Sci. 16 (4) (2001) 351–367.
[45] G.O. Roberts, R.L. Tweedie, Exponential convergence of langevin distributions and their discrete approximations, Bernoulli 2 (4) (1996) 341–363.
[46] G.O. Roberts, R.L. Tweedie, Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms, Biometrika 83 (1) (1996) 95–110.
[47] J.S. Rosenthal, Asymptotic variance and convergence rates of nearly-periodic Markov chain Monte Carlo algorithms, J. Amer. Statist. Assoc. 98 (461) (2003) 169–177.
[48] L. Tierney, A note on Metropolis–Hastings kernels for general state spaces, Ann. Appl. Probab. 8 (1) (1998) 1–9.
[49] M. Vialaret, F. Maire, On the convergence time of some non-reversible Markov chain Monte Carlo methods, Methodol. Comput. Appl. Probab. 22 (2020) 1349–1387.