entropy

MDPI

Article

Convergence Rates for the Constrained Sampling via Langevin Monte Carlo

Yuanzheng Zhu

School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China; zhuyz0626@smail.swufe.edu.cn

Abstract: Sampling from constrained distributions has posed significant challenges in terms of algorithmic design and non-asymptotic analysis, which are frequently encountered in statistical and machine-learning models. In this study, we propose three sampling algorithms based on Langevin Monte Carlo with the Metropolis–Hastings steps to handle the distribution constrained within some convex body. We present a rigorous analysis of the corresponding Markov chains and derive non-asymptotic upper bounds on the convergence rates of these algorithms in total variation distance. Our results demonstrate that the sampling algorithm, enhanced with the Metropolis–Hastings steps, offers an effective solution for tackling some constrained sampling problems. The numerical experiments are conducted to compare our methods with several competing algorithms without the Metropolis–Hastings steps, and the results further support our theoretical findings.

Keywords: Bayesian computation; constrained sampling; convex support; Langevin Monte Carlo; MCMC; mixing time bound

1. Introduction

Sampling from distributions with some constraints has extensive applications in statistics, machine-learning, and operations research, among other areas. Some distributions have bounded support, such as the simple but versatile uniform distribution which serves as the foundation for a series of Monte Carlo methods, as discussed in [1]. Furthermore, many statistical inference problems involve estimating parameters subject to constraints on the parameter space, which defines a posterior distribution with bounded support in a Bayesian setting. Examples include Latent Dirichlet Allocation [2], truncated data problems in failure and survival time studies [3], ordinal data models [4], constrained lasso and ridge regressions [5], and non-negative matrix factorization [6]. In Bayesian learning, sampling from posterior distributions is a fundamental primitive, used for exploring posterior distributions, identifying the unknown parameters [7,8], obtaining credible intervals, and solving inverse problems [7,8]. Finally, constrained sampling has great potential in solving constrained optimization problems [9,10].

Many Markov Chain Monte Carlo (MCMC) algorithms have been extensively studied for sampling from probability distributions with convex support or more generally with constrained parameters, mainly in the fields of Bayesian statistics and theoretical computer science. Early work includes, among others, [1,11–14]. Firstly, based on MCMC algorithms, a direct solution involves discarding samples that violate the constraints, thereby exclusively retaining samples that satisfy the constraints; see, for example, [1,15,16]. However, these rejection-type approaches may encounter an excessive number of rejections or an extremely large acceptance rate within some local subspace that satisfies the constraints, which leads to poor mixing and computational inefficiency, especially for complicated constraints and the high dimensional distributions [17,18]. Secondly, some literature draws inspiration from penalty functions in optimization problems and considers the construction of barriers along the boundaries of the constrained domain, effectively constraining the sampling process within the constrained area. These approaches encounter a major challenge

check for updates

Citation: Zhu, Y. Convergence Rates for the Constrained Sampling via Langevin Monte Carlo. *Entropy* **2023**, *25*, 1234. https://doi.org/10.3390/e25081234

Academic Editor: Antonio M. Scarfone

Received: 26 June 2023
Revised: 8 August 2023
Accepted: 15 August 2023
Published: 18 August 2023

when the samples reach the boundaries of the constraints, necessitating the implementation of a mechanism based on reflection to redirect them back into the constrained region. To address this issue, Ref. [19] extended the Hamiltonian Monte Carlo (HMC) method by setting the potential energy outside the constraint region to infinity, restricting the states to the desired domain. Ref. [20] extended the HMC method to sample from truncated multivariate Gaussian distributions, and Ref. [21] proposed an approach that involves mapping the constrained domain onto a sphere in an augmented space. Thirdly, motivated by the constrained optimization methods, the constrained sampling problem can be reformulated as an unconstrained sampling problem via suitable transformations. Following this idea, Ref. [22] proposed a family of novel algorithms based on HMC through the introduction of Lagrange multipliers that address a broader range of constrained sampling problems. More recently, Ref. [23] tackled the constrained sampling problem via the mirror-Langevin algorithm. In spite of the widespread adoption of these MCMC methods, most of them have primarily focused on the algorithm design and lack the rigorous theoretical analysis of convergence rates.

Among all the MCMC algorithms, a class of algorithms based on the Langevin dynamics has garnered significant attention in both practical applications and theoretical analyses [24–27]. It has recently witnessed a notable increase in non-asymptotic analyses of these algorithms, initiated by the seminal work of [28]. In the setting of unconstrained sampling, Ref. [29] extended the theoretical analysis of convergence rates by studying with decreasing step size, and Refs. [30,31] derived corresponding convergence results based on alternative distances. These theoretical analyses focus on the Langevin algorithm without the Metropolis–Hastings step. More recently, Refs. [32,33] have shown that incorporating the Metropolis–Hastings step can significantly improve the convergence rate of the associated Langevin algorithm. In the setting of constrained sampling, Ref. [34] suggested a Euclidean projection step in the Langevin algorithms for the constrained case (PLMC) and derived the convergence rate of the associated Markov chain. Ref. [35] presented a detailed theoretical analysis for a proximal version of the Langevin algorithm that incorporates the Moreau-Yosida envelope of the indicator function (MYULA) to handle the distributions that are restricted to a convex body. Ref. [36] constructed the mirrored Langevin algorithm (MLD) using a mirror map to constrain the domain, which achieves the same convergence rate as its unconstrained counterpart [28]. However, these constrained sampling algorithms are all developed based on the Langevin algorithm without incorporating the Metropolis–Hastings steps, thus not leveraging the fast mixing advantages of them.

In this paper, we considered the constrained Langevin Monte Carlo with the Metropolis–Hastings step for sampling from the distributions restricted to some convex support. Firstly, for certain constraints, we re-examine the simple and intuitive rejection-type methods for sampling from constrained distributions, and reach a surprising discovery that the corresponding algorithm still retained the advantage of rapid convergence by carefully selecting the step size parameter. Subsequently, for the more generally constrained domain, we build upon the framework proposed in [35], incorporating the Metropolis–Hastings step for further refinement, and analyze the convergence rate of the corresponding Markov chain. We present detailed non-asymptotic analysis for these constrained algorithms and achieve notably enhanced convergence rates in the total variation distance. Compared with the best rate in [36], our results show that adopting the Metropolis–Hastings step in some constrained MCMC algorithms can also lead to an exponentially improved dependence on the error tolerance.

The rest of the paper is organized as follows. In Section 2, we introduce the preliminaries and the problem set-up of our study. Then, we propose the constrained sampling algorithms tailored to different types of constraint regions in Section 3. Section 4 provides the non-asymptotic theoretical results of the proposed algorithms. The numerical experiments and comparisons are presented in Section 5. Some Markov chain basics are provided in Appendix A and all the technical proofs are deferred to Appendix B.

Notation: Let $\lceil a \rceil$ represent the smallest integer not less than $a \in \mathbb{R}$. For a vector $x \in \mathbb{R}^d$, we use $|x|_2$ to denote its Euclidean norm. For a $q \times q$ symmetric matrix A, denote by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ the smallest and largest eigenvalues of A, respectively, and let A^{T} be its transpose. For two square matrices A and B, we write $A \preceq B$ if $(B - A)$ is a positive semi-definite matrix. Denote by $I(\cdot)$ the indicator function. For $r > 0$, let $\mathbb{B}(x, r) = \{y \in \mathbb{R}^d : |y - x|_2 \leq r\}$ denote a closed Euclidean ball with center x and radius r. For two real-valued sequences a_n and b_n, we say $a_n = O(b_n)$ if there exists a universal constant c such that $a_n \leq cb_n$, and $a_n = \tilde{O}(b_n)$ if $a_n \leq c_n b_n$ where the sequence c_n grows at most poly-logarithmically with n. For any two probability measures μ and ν, denote by $\|\mu - \nu\|_{\mathrm{TV}}$ the total variation distance between μ and ν.

2. Preliminaries and Problem Set-Up

In this section, we introduce the MCMC sampling methods with its mixing analysis, the traditional unconstrained Metropolis-Adjusted Langevin Algorithm (MALA), and our problem set-up for this paper.

2.1. Markov Chain Monte Carlo and Mixing

Consider a distribution Π equipped with a density $\pi : \mathbb{R}^d \mapsto \mathbb{R}_+$ such that

$$\pi(x) \propto e^{-U(x)} \tag{1}$$

for some potential function $U : \mathbb{R}^d \mapsto \mathbb{R}$. In certain scenarios, it is necessary to perform sampling from this distribution. For example, many statistical applications involve estimating the expectation of a function $g(X)$ for $X \sim \pi$, where analytical and numerical computation is infeasible. Monte Carlo approximation provides a solution by generating samples from Π and using sample mean to estimate the population expectation. Hence, the key point is to access samples from Π.

MCMC represents a class of popular sampling algorithms, which construct an appropriate Markov chain whose stationary distribution is Π or close to Π in certain metrics. The class of the Metropolis–Hastings algorithms refers to a type of MCMC method that ensures the corresponding Markov chain converges to the target distribution by incorporating the Metropolis–Hastings step. The Metropolis–Hastings algorithms usually take two steps to generate a Markov chain: a proposal step and a reject-accept step. At each iteration, a sample is generated from the proposal distribution in the proposal step, and it is updated as a new state of the Markov chain with probability determined by the Metropolis–Hastings correction in the reject-accept step.

Given an error tolerance $\varepsilon \in (0, 1)$, in order to obtain an ε-accurate sample with respect to some metric, one simulates the Markov chain for a certain number of steps k, as determined by a mixing time analysis. Specifically, we are concerned about how many steps the chain needs to take such that the current distribution of the chain is ε-close to the target distribution Π. Based on this, we define the ε-mixing time with respect to the target distribution Π as

$$\tau(\varepsilon; \mathbb{P}^0, \Pi) = \min\{k \in \mathbb{N} : \|\mathcal{T}^k(\mathbb{P}^0) - \Pi\|_{\mathrm{TV}} \leq \varepsilon\} \tag{2}$$

for the error tolerance $\varepsilon \in (0, 1)$, where $\mathcal{T}$ is the transition operator of the Markov chain and $\mathcal{T}^k(\mathbb{P}^0)$ is the distribution of the Markov chain at k-th step from an initial distribution $\mathbb{P}^0$.

2.2. Metropolis-Adjusted Langevin Algorithm

Consider the problem of sampling from the distribution with density defined as (1). MALA [26,27] adopts the Gaussian distribution $\mathcal{N}\{x_k - h\nabla U(x_k), 2hI_p\}$ as the proposal distribution for the k-th step, where x_k is the current state and $h > 0$ is a proper step size, and performs a Metropolis–Hastings accept-reject step. MALA is the standard Metropolis–Hastings algorithm applied to the Langevin dynamics, and the associated Langevin-type algorithms belong to a family of gradient-based MCMC sampling algo-

rithms [37]. The Langevin-type algorithms can be understood as the Euler discretization of the Langevin dynamics:

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t\,,$$

where $W_t (t \geq 0)$ is the standard Brownian motion on $\mathbb{R}^d$.

Algorithm 1 provides the unconstrained MALA for sampling from the distribution supported on $\mathbb{R}^d$, where $\phi_h(\cdot \mid x)$ denotes the probability density function of $\mathcal{N}\{x - h\nabla U(x), 2hI_d\}$.

Algorithm 1 Metropolis-adjusted Langevin algorithm

Input: a sample $x^0 \in \mathbb{R}^d$ from an initial distribution $\mathbb{P}^0$, the step size h
  for $k = 0, 1, 2, \ldots, K-1$ **do**
    Proposal step: $y^{k+1} \leftarrow x^k - h\nabla U(x^k) + \xi$, where $\xi \sim \mathcal{N}(0, 2hI_p)$
    Accept-reject step:
    compute $\alpha^{k+1} = \min\left\{1, \dfrac{\phi_h(x^k \mid y^{k+1})\pi(y^{k+1})}{\phi_h(y^{k+1} \mid x^k)\pi(x^k)}\right\}$
    sample u^{k+1} from the uniform distribution on $[0, 1]$
    if $\alpha^{k+1} \geq u^{k+1}$, **then** $x^{k+1} \leftarrow y^{k+1}$
    else $x^{k+1} \leftarrow x^k$
    end if
  end for
Output: $x^1, x^2, \ldots, x^K$

2.3. Problem Set-Up

In this part, we consider the problem of sampling from a target distribution or posterior Π^* supported on a compact set $\mathcal{X} \subset \mathbb{R}^d$ equipped with a density π^*. It can be written in the form

$$\pi^*(x) = \frac{\exp\{-U(x)\}I(x \in \mathcal{X})}{\int_{\mathcal{X}} \exp\{-U(y)\}\,\mathrm{d}y} \tag{3}$$

for some potential function $U : \mathbb{R}^d \mapsto \mathbb{R}$. Assume that the function $U(\cdot)$ and the set $\mathcal{X}$ satisfy the following assumptions:

Assumption 1. *$U(\cdot)$ is a twice continuously differentiable, L-smooth and m-strongly convex function on $\mathbb{R}^d$. That is, there exist universal constants $L \geq m > 0$ such that*

$$\frac{m}{2}|y - x|_2^2 \leq U(y) - U(x) - \{\nabla U(x)\}^{\mathrm{T}}(y - x) \leq \frac{L}{2}|y - x|_2^2$$

for any $x, y \in \mathbb{R}^d$.

Assumption 2. *$\mathcal{X} \subset \mathbb{R}^d$ is a compact and convex set satisfying*

$$\mathbb{B}(x^*, r) \subset \mathcal{X} \subset \mathbb{B}(x^*, R)$$

for some universal constants $0 < r \leq R$ and $x^* \in \mathcal{X}$.

Hereafter, we assume that the above two assumptions hold, which is frequently used in the literature for the analysis of constrained sampling algorithms [34–36]. We will modify the MALA in Algorithm 1 to adapt to sampling from the above constrained distribution, and analyse its non-asymptotic theoretical properties and derive the mixing time bound in terms of the problem dimension d and the error tolerance ε.

3. The Constrained Langevin Algorithms

In this section, we present three sampling algorithms based on MALA to handle the distribution constrained within some convex body $\mathcal{X}$. As discussed in [34], the inherent

challenges in constrained sampling problems arise from the complex properties on the boundary of the constraint region, and the lack of the curvature in the potential function. To tackle these challenges, Ref. [34] initially studied constrained sampling from the uniform distribution on $\mathcal{X}$, and then extended the exploration to more general distributions. Similarly, we begin our investigation by examining some simple constrained regions and progressively extend our analysis to more complex constraint scenarios.

3.1. Constrained Langevin Algorithm via Rejection

We initially discuss the case where the constraint region $\mathcal{X}$ is an Euclidean ball on $\mathbb{R}^d$, where the boundary can be characterized by a curve equation. If $\mathcal{X} = \mathbb{B}(x^*, R)$ for some universal constant $R > 0$ and $x^* \in \mathbb{R}^d$, we consider the simple and intuitive rejection-type methods via the Metropolis–Hastings accept-reject step for sampling from the distribution with density defined as (3). The constrained MALA for $\mathcal{X} = \mathbb{B}(x^*, R)$ outlined in Algorithm 2 as follows, where $\phi_h(\cdot \mid x)$ denotes the probability density function of the Gaussian distribution $\mathcal{N}\{x - h\nabla U(x), 2hI_d\}$.

Algorithm 2 The MALA for Euclidean ball constrained domain

Input: a sample $x^0 \in \mathcal{X}$ from an initial distribution $\mathbb{P}^0$, the step size h
  for $k = 0, 1, 2, \ldots, K-1$ **do**
    Proposal step: $y^{k+1} \leftarrow x^k - h\nabla U(x^k) + \xi$, where $\xi \sim \mathcal{N}(0, 2hI_p)$
    Accept-reject step:
    if $y^{k+1} \in \mathcal{X}$ **then**
      compute $\alpha^{k+1} = \min\left\{1, \dfrac{\phi_h(x^k \mid y^{k+1})\pi^*(y^{k+1})}{\phi_h(y^{k+1} \mid x^k)\pi^*(x^k)}\right\}$
      sample u^{k+1} from the uniform distribution on $[0, 1]$
      if $\alpha^{k+1} \geq u^{k+1}$, **then** $x^{k+1} \leftarrow y^{k+1}$
      else $x^{k+1} \leftarrow x^k$
      end if
    else $x^{k+1} \leftarrow x^k$
    end if
  end for
Output: $x^1, x^2, \ldots, x^K$

Compared with Algorithm 1, this modified algorithm forces the Markov chain to stay at the current state when it jumps out of the limited state space $\mathcal{X} = \mathbb{B}(x^*, R)$, which is a quite natural extension of the unconstrained MALA. This idea is not completely novel. Ref. [34] suggested a projection step in unadjusted Langevin algorithm for sampling from a log-concave distribution with compact support. Ref. [10] proposed an MALA for constrained optimization, where they used a similar step to constrain the Markov chain to stay at a given state space. Due to the favorable properties on the boundary of constrained domain $\mathcal{X} = \mathbb{B}(x^*, R)$, we can establish the theoretical results of Algorithm 2; see Lemma A1 in Appendix B for details.

3.2. Norm-Constrained Domain

Regularization is a technique commonly used in machine-learning and statistical modeling. As discussed in [38], some models with regularization can be reformulated as the distributions with norm-constraint on the parameters. Notice that the L_p-norm for the vector $x = (x_1, x_2, \ldots, x_d)^\intercal \in \mathbb{R}^d$ is defined as

$$|x|_p = \begin{cases} \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}, & p \in (0, \infty) \\ \max_{1 \leq i \leq d} |x_i|, & p = \infty. \end{cases}$$

For the norm-constrained domain $\mathcal{X} = \{x \in \mathbb{R}^d : |x|_p \leq C\}$ with some universal constant $C > 0$, we can transform it into the Euclidean ball $\mathbb{B}(0,1)$ via a vector-valued function $f : \mathcal{X} \mapsto \mathbb{B}(0,1)$. Specifically, for any $x = (x_1, x_2, \ldots, x_d)^{\mathrm{T}} \in \mathcal{X}$, we have $y = f(x) =: \{f_1(x), f_2(x), \ldots, f_d(x)\}^{\mathrm{T}}$ with

$$f_i(x) = \begin{cases} C^{-p/2}\mathrm{sgn}(x_i)|x_i|^{p/2}, & p \in (0,\infty) \\ x_i \dfrac{|x|_\infty}{C|x|_2}, & p = \infty \end{cases}, \quad 1 \leq i \leq d$$

such that $y \in \mathbb{B}(0,1)$. Due to the bijective nature of the function $f : \mathcal{X} \mapsto \mathbb{B}(0,1)$, its inverse function $f^{-1} =: g : \mathbb{B}(0,1) \mapsto \mathcal{X}$ can be defined accordingly. Similarly, for any $y = (y_1, y_2, \ldots, y_d)^{\mathrm{T}} \in \mathbb{B}(0,1)$, we have $x = g(y) =: \{g_1(y), g_2(y), \ldots, g_d(y)\}^{\mathrm{T}}$ with

$$g_i(y) = \begin{cases} C\mathrm{sgn}(y_i)|y_i|^{2/p}, & p \in (0,\infty) \\ Cy_i \dfrac{|y|_2}{|y|_\infty}, & p = \infty \end{cases}, \quad 1 \leq i \leq d$$

such that $x \in \mathcal{X}$. By utilizing the vector-valued functions $f(\cdot)$ and $g(\cdot)$ defined above, we can employ the Euclidean ball constrained sampling algorithm, as described in Section 3.1, to tackle the norm-constrained domain $\mathcal{X} = \{x \in \mathbb{R}^d : |x|_p \leq C\}$. The computational process is outlined in Algorithm 3, where

$$\pi^{\mathbb{B}(0,1)}(x) = \frac{\exp\{-U(x)\}I\{x \in \mathbb{B}(0,1)\}}{\int_{\mathbb{B}(0,1)} \exp\{-U(y)\}\,\mathrm{d}y}$$

with the potential function $U(\cdot)$.

Algorithm 3 The MALA for norm-constrained domain

Input: a sample $x^0 \in \mathcal{X}$ from an initial distribution $\mathbb{P}^0$, the step size h
  for $k = 0, 1, 2, \ldots, K-1$ **do**
    Transformation step: $y^k \leftarrow f(x^k)$
    Proposal step: $z^{k+1} \leftarrow y^k - h\nabla U(y^k) + \xi$, where $\xi \sim \mathcal{N}(0, 2hI_p)$
    Accept-reject step:
    if $z^{k+1} \in \mathbb{B}(0,1)$ **then**
      compute $\alpha^{k+1} = \min\left\{1, \dfrac{\phi_h(y^k \mid z^{k+1})\pi^{\mathbb{B}(0,1)}(z^{k+1})}{\phi_h(z^{k+1} \mid y^k)\pi^{\mathbb{B}(0,1)}(y^k)}\right\}$
      sample u^{k+1} from the uniform distribution on $[0,1]$
      if $\alpha^{k+1} \geq u^{k+1}$, **then** $y^{k+1} \leftarrow z^{k+1}$
      else $y^{k+1} \leftarrow y^k$
      end if
    else $y^{k+1} \leftarrow y^k$
    end if
    Transformation step: $x^{k+1} \leftarrow g(y^{k+1})$
  end for
Output: $x^1, x^2, \ldots, x^K$

Compared with Algorithm 2, the Algorithm 3 achieves the $\mathcal{X} \to \mathbb{B}(0,1) \to \mathcal{X}$ transformation by incorporating two transformation steps, thereby addressing the norm-constrained sampling problems. The main purpose of this approach is to facilitate theoretical analysis by leveraging the well-understood properties of the boundary of the Euclidean ball compared to the boundary of the norm-constrained domain; see Appendix B.7 for details.

3.3. Constrained Langevin Algorithm via an Approximation of the Indicator Function

We proceed to discuss the constrained sampling for more general constraint regions. Given $\mathcal{X} \in \mathbb{R}^d$, define

$$\iota_{\mathcal{X}}(x) =: -\log\{I(x \in \mathcal{X})\} = \begin{cases} 0, & \text{If } x \in \mathcal{X} \\ \infty, & \text{If } x \notin \mathcal{X} \end{cases} \tag{4}$$

for any $x \in \mathbb{R}^d$. Then, the target distribution Π^* with density defined as (3) can be reformulated as

$$\pi^*(x) = \frac{\exp\{-V_{\mathcal{X}}(x)\}}{\int_{\mathcal{X}} \exp\{-V(y)\}\,\mathrm{d}y} \tag{5}$$

with the potential function $V_{\mathcal{X}} : \mathbb{R}^d \mapsto \mathbb{R}$ satisfying

$$V_{\mathcal{X}}(\cdot) = U(\cdot) + \iota_{\mathcal{X}}(\cdot), \tag{6}$$

where $\iota_{\mathcal{X}}(\cdot)$ is defined in (4). Notice that $\iota_{\mathcal{X}}(\cdot)$ is a convex function on $\mathbb{R}^d$. Under Assumption 1, we then know that the potential function $V_{\mathcal{X}}(\cdot)$ is smooth and strongly convex on $\mathbb{R}^d$. By this transformation, the problem of constrained sampling is apparently converted into an unconstrained counterpart. However, the non-differentiability of the function $V_{\mathcal{X}}(\cdot)$ on the boundary of $\mathcal{X}$ poses a challenge when applying the gradient-based unconstrained sampling algorithms. To address this issue, we can approximate the function $\iota_{\mathcal{X}}(\cdot)$ by a differentiable function such as the Moreau-Yosida (MY) envelope [35]. The MY envelope of $\iota_{\mathcal{X}}(\cdot)$ is defined as

$$\iota_{\mathcal{X}}^{\lambda}(x) = \inf_{y \in \mathbb{R}^d} \{\iota_{\mathcal{X}}(x) + (2\lambda)^{-1}|x - y|_2^2\} = (2\lambda)^{-1}|x - \mathrm{Pro}_{\mathcal{X}}(x)|_2^2 \tag{7}$$

for any $x \in \mathbb{R}^d$, where $\lambda > 0$ is a regularization parameter and $\mathrm{Pro}_{\mathcal{X}}(\cdot)$ is the projection function onto $\mathcal{X}$. By [35], the function $\iota_{\mathcal{X}}^{\lambda}(\cdot)$ is convex and continuously differentiable with the gradient

$$\nabla \iota_{\mathcal{X}}^{\lambda}(x) = \lambda^{-1}\{x - \mathrm{Pro}_{\mathcal{X}}(x)\} \tag{8}$$

for any $x \in \mathbb{R}^d$, and it holds that

$$|\nabla \iota_{\mathcal{X}}^{\lambda}(x) - \nabla \iota_{\mathcal{X}}^{\lambda}(y)|_2 \leq \lambda^{-1}|x - y|_2 \tag{9}$$

for any $x, y \in \mathbb{R}^d$. Then the approximation of $V_{\mathcal{X}}(\cdot)$ defined as (6) can be given by

$$V_{\mathcal{X}}^{\lambda}(\cdot) = U(\cdot) + \iota_{\mathcal{X}}^{\lambda}(\cdot), \tag{10}$$

which is continuously differentiable, smooth and strongly convex on $\mathbb{R}^d$ if $U(\cdot)$ satisfying Assumption 1. Define the distribution $\Pi^{*,\lambda}$ with density

$$\pi^{*,\lambda}(x) = \frac{\exp\{-V_{\mathcal{X}}^{\lambda}(x)\}}{\int_{\mathbb{R}^d} \exp\{-V^{\lambda}(y)\}\,\mathrm{d}y}. \tag{11}$$

Recall that the target distribution Π^* with the reformulated density defined as (5). As discussed in [35], under some mild conditions including Assumptions 1 and 2, the approximation error between Π^* and $\Pi^{*,\lambda}$ in total variation distance can be made arbitrarily small by adjusting the regularization parameter λ. Therefore, we can utilize the gradient-based unconstrained sampling algorithms, such as the MALA presented in Algorithm 1, for constructing an appropriate Markov chain whose stationary distribution is close to Π^*; see Algorithm 4 for details, where $\phi_h^{\lambda}(\cdot \mid x)$ denotes probability density function of the Gaussian distribution $\mathcal{N}\{x - h\{\nabla U(x) + \nabla \iota_{\mathcal{X}}^{\lambda}(x)\}, 2hI_d\}$ with $\nabla \iota_{\mathcal{X}}^{\lambda}(\cdot)$ defined as (8).

Algorithm 4 The MALA for convex constrained domain

Input: a sample $x^0 \in \mathbb{R}^d$ from an initial distribution $\mathbb{P}^0$, the step size h
for $k = 0, 1, 2, \ldots, K-1$ **do**
Proposal step: $y^{k+1} \leftarrow x^k - h\{\nabla U(x^k) + \nabla \iota_{\mathcal{X}}^{\lambda}(x^k)\} + \xi$, where $\xi \sim \mathcal{N}(0, 2hI_p)$
Accept-reject step:
compute $\alpha^{k+1} = \min\left\{1, \dfrac{\phi_h^{\lambda}(x^k \mid y^{k+1})\pi^{*,\lambda}(y^{k+1})}{\phi_h^{\lambda}(y^{k+1} \mid x^k)\pi^{*,\lambda}(x^k)}\right\}$
sample u^{k+1} from the uniform distribution on $[0, 1]$
if $\alpha^{k+1} \geq u^{k+1}$, **then** $x^{k+1} \leftarrow y^{k+1}$
else $x^{k+1} \leftarrow x^k$
end if
end for
Output: $x^1, x^2, \ldots, x^K$

4. Theoretical Results

In this section, we first analyze the properties of the Markov chains determined by the three constrained sampling algorithms presented in Section 3, and then establish the mixing time bounds of these Markov chains.

4.1. Properties of the Markov Chains

The outcomes $\{x^1, \ldots, x^K\}$ from each algorithm presented in Section 3 form a Markov chain, whose properties are established in Propositions 1, 2, and 3, respectively, as below.

Proposition 1. *For $\mathcal{X} = \mathbb{B}(x^*, R)$ with some universal constant $R > 0$ and $x^* \in \mathbb{R}^d$, the Markov chain determined by Algorithm 2 is Π^*-irreducible, smooth, and reversible with respect to the stationary distribution Π^* with density π^* defined as (3) (The definition of the Π^*-irreducible, reversible, and smooth Markov chain is deferred to Appendix A).*

Remark 1. *Proposition 1 shows that the Markov chain determined by Algorithm 2 enjoys a series of nice properties as the unconstrained MALA, which form the basis for the study of the mixing time bounds of such Markov chain.*

The similar properties hold for the Markov chains determined by Algorithms 3 and 4 as well.

Proposition 2. *For $\mathcal{X} = \{x \in \mathbb{R}^d : |x|_p \leq C\}$ with some universal constant $C > 0$, the Markov chain determined by Algorithm 3 is Π^*-irreducible, smooth, and reversible with respect to the stationary distribution Π^* with density π^* defined as (3).*

Proposition 3. *Under Assumption 2, the Markov chain determined by Algorithm 4 is $\Pi^{*,\lambda}$-irreducible, smooth, and reversible with respect to the distribution $\Pi^{*,\lambda}$ with density $\pi^{*,\lambda}$ defined as (11).*

4.2. Mixing Time Bounds of the Markov Chains

For a distribution Π supported on $\mathcal{X} \subset \mathbb{R}^d$ with the density π, recall that the ε-mixing time with respect to Π is defined as (2). A β-warm initial distribution $\mathbb{P}^0$ with density p^0 with respect to the distribution Π is commonly used for the mixing time analysis, which satisfies

$$\sup_{x \in \mathcal{X}} \frac{p^0(x)}{\pi(x)} \leq \beta$$

for some finite constant $\beta > 0$. We say that the Markov chain is ς-lazy if at each iteration the chain is forced to stay at the previous state with probability at least ς. It is a convenient assumption for theoretical analysis of the convergence rate, but not likely to be used in

practice since the lazy steps slow down the mixing rate of Markov chain. Given the definitions above and some Markov chain basics in Appendix A, we can obtain the following results for some well-behaved Markov chains defined on $\{\mathcal{X}, \mathscr{B}(\mathcal{X})\}$.

Lemma 1. *Consider a reversible, Π-irreducible, ς-lazy, and smooth Markov chain defined on $\{\mathcal{X}, \mathscr{B}(\mathcal{X})\}$ with stationary distribution Π supported on $\mathcal{X}$. For any error tolerance $\varepsilon \in (0,1)$ and β-warm initial distribution $\mathbb{P}^0$, the ε-mixing time with respect to Π satisfying*

$$\tau(\varepsilon; \mathbb{P}^0, \Pi) \leq \left\lceil \frac{4}{\varsigma} \int_{4\beta^{-1}}^{\varepsilon^{-2}} \frac{\mathrm{d}v}{v\tilde{\Omega}^2(v)} \right\rceil,$$

where $\tau(\varepsilon; \mathbb{P}^0, \Pi)$ and $\tilde{\Omega}(\cdot)$ are defined, respectively, in (2) and (A4).

Remark 2. *Lemma 1 provides a control on the mixing time of a Markov chain on $\mathcal{X}$ in terms of $\tilde{\Omega}(\cdot)$. This result can be seen as an extension of Lemma 3 in [33] to the case where a Markov chain defined on $\{\mathcal{X}, \mathscr{B}(\mathcal{X})\}$. We then can readily derive the mixing time bound if a lower bound for $\tilde{\Omega}(\cdot)$ is known.*

The following lemma gives a lower bound for $\Omega(\cdot)$.

Lemma 2. *Assume that the distribution Π supported on $\mathcal{X}$ with the density π satisfy the log-isoperimetry inequality defined as (A1) for some constant $\hat{c} > 0$. If a reversible Markov chain with stationary distribution Π satisfies $\sup_{x,y\in\mathcal{X}:\,|x-y|_2\leq\Delta} \|\mathcal{T}_x - \mathcal{T}_y\|_{\mathrm{TV}} \leq 1-\delta$ for some $\delta \in (0,1)$ and $\Delta > 0$, it then holds that*

$$\Omega(v) \geq \frac{\delta}{4} \min\left\{1, \frac{\Delta}{4\hat{c}} \log^{1/2}\left(1 + \frac{1}{v}\right)\right\}$$

for any $v \in (0, 1/2]$, where $\mathcal{T}_x$ is the one-step transition distribution of this Markov chain at $x \in \mathcal{X}$ and $\Omega(\cdot)$ is the conductance profile of this Markov chain defined in (A3).

Remark 3. *Lemma 2 states a lower bound for the conductance profile of a Markov chain on $\mathcal{X}$. Similar results can be found in the [33,39,40]. Lemma 2, together with Lemma 1, provides a general framework for obtaining mixing time bound of a well-behaved Markov chain on $\mathcal{X}$.*

Based on Lemmas 1 and 2, we can drive the upper bounds for each ε-mixing time of the Markov chains determined by the three constrained sampling algorithms presented in Section 3.

Theorem 1. *For $\mathcal{X} = \mathbb{B}(x^*, R)$ with some universal constant $R > 0$ and $x^* \in \mathbb{R}^d$, let Assumption 1 hold with $L^{3/8}R^{3/4} \geq 16/\sqrt{d} + 8$ and $L^{-15/8}m^2R^{1/4} \geq 12d$. Given a β-warm initial distribution $\mathbb{P}^0$ and an error tolerance $\varepsilon \in (0,1)$, the Markov chain determined by Algorithm 2 satisfies*

$$\tau(\varepsilon; \mathbb{P}^0, \Pi^*) = O\left(\frac{L^{7/4}R^{3/2}d}{m} \log\frac{\log\beta}{\varepsilon}\right)$$

for any step size h satisfying

$$\frac{1}{L^{7/4}R^{3/2}d} \leq h \leq \min\left[\frac{R^2(1-\tilde{c})^2}{4\{\log^{1/2}(16/u) + \sqrt{d}\}^2}, \frac{\sqrt{u}}{4\sqrt{3}L^{3/2}R}, \frac{u}{128L\{\log^{1/2}(16/u) + \sqrt{d}\}^2}\right]$$

with $\tilde{c} = \{1 + (L^{-7/2}R^{-3}d^{-2} - L^{-11/4}R^{-3/2}d^{-1})m^2\}^{1/2}$ and some constant $u \in (1/2, 1)$, where Π^* with density π^* defined as (3).

Remark 4. *Theorem 1 presents a sharp mixing time bound for Algorithm 2 with a β-warm initial distribution as $\tilde{O}\{d\log(1/\varepsilon)\}$ up to β and L, m, R which are specified in Assumptions 1 and 2.*

This result improves upon the previously known mixing time bounds for constrained sampling algorithms in [34–36]*; see Table 1 for details.*

Table 1. Convergence rates for sampling from log-concave distributions with bounded support.

Assumptions	$\|\cdot\|_{\mathrm{TV}}$ Rate	Algorithms
$0I_d \preceq \nabla^2 U(x) \preceq LI_d$	$\tilde{O}(d^{12}\varepsilon^{-12})$	PLMC in [34]
$mI_d \preceq \nabla^2 U(x) \preceq LI_d$	$\tilde{O}(d^5\varepsilon^{-6})$	MYULA in [35]
$mI_d \preceq \nabla^2 U(x)$	$\tilde{O}(d\varepsilon^{-2})$	MLD in [36]
$mI_d \preceq \nabla^2 U(x) \preceq LI_d$	$\tilde{O}\{d\log(1/\varepsilon)\}$	Algorithms 2 and 3 in this paper
$mI_d \preceq \nabla^2 U(x) \preceq LI_d$	$\tilde{O}(d^3\varepsilon^{-2})$	Algorithm 4 in this paper

For sampling from the norm-constrained domain $\mathcal{X} = \{x \in \mathbb{R}^d : |x|_p \le C\}$ with some universal constant $C > 0$, we transform it into the sampling from Euclidean ball $\mathbb{B}(0,1)$ as shown in Algorithm 3; then, the similar result holds for the Markov chain determined by Algorithm 3 as well.

Corollary 1. *For $\mathcal{X} = \{x \in \mathbb{R}^d : |x|_p \le C\}$ with some universal constant $C > 0$, let Assumption 1 hold with $L^{3/8} \ge 16/\sqrt{d} + 8$ and $L^{-15/8}m^2 \ge 12d$. Given a β-warm initial distribution $\mathbb{P}^0$ and an error tolerance $\varepsilon \in (0,1)$, the Markov chain determined by Algorithm 3 satisfies*

$$\tau(\varepsilon; \mathbb{P}^0, \Pi^*) = O\left(\frac{L^{7/4}d}{m}\log\frac{\log\beta}{\varepsilon}\right)$$

for any step size h satisfying

$$\frac{1}{L^{7/4}d} \le h \le \min\left[\frac{(1-\bar{c})^2}{4\{\log^{1/2}(16/u)+\sqrt{d}\}^2}, \frac{\sqrt{u}}{4\sqrt{3}L^{3/2}}, \frac{u}{128L\{\log^{1/2}(16/u)+\sqrt{d}\}^2}\right]$$

with $\bar{c} = \{1 + (L^{-7/2}d^{-2} - L^{-11/4}d^{-1})m^2\}^{1/2}$ and some constant $u \in (1/2, 1)$, where Π^* with density π^* defined as (3)*.*

For the Markov chain determined by Algorithm 4, we can also derive a sharp mixing time bound by the mixing time analysis for sampling from log-concave distribution without constraints in [33] and the approximation error between Π^* and $\Pi^{*,\lambda}$ in [35].

Theorem 2. *Let Assumptions 1 and 2 hold, and assume that there exists a universal constant $\tilde{C} > 0$ such that $\exp\{\inf_{x\in\mathcal{X}^c} U(x) - \sup_{x\in\mathcal{X}} U(x)\} \ge \tilde{C}$. Given the initial distribution $\mathbb{P}^0 = \mathcal{N}\{x^\star, (L+\lambda^{\star-1})^{-1}I_d\}$ with $x^\star = \arg\min_{x\in\mathbb{R}^d} V_{\mathcal{X}}^{\lambda^\star}(x)$ and an error tolerance $\varepsilon \in (0,1)$, the Markov chain determined by Algorithm 4 satisfies*

$$\tau(\varepsilon; \mathbb{P}^0, \Pi^*) = O\left[\frac{(L+\lambda^{\star-1})d}{m}\log\frac{d}{\varepsilon}\cdot\max\left\{1, \sqrt{\frac{L+\lambda^{\star-1}}{dm}}\right\}\right]$$

for the step size h satisfying

$$h = c\frac{1}{(L+\lambda^{\star-1})d\cdot\max\left\{1, \sqrt{\frac{L+\lambda^{\star-1}}{dm}}\right\}}$$

with some universal constant $c > 0$, where $V_{\mathcal{X}}^{\lambda^\star}(\cdot)$ is defined as in (10) *with $\lambda^\star := 8\pi^{-1}\varepsilon^2 r^2 d^{-2}\tilde{C}^2$, and Π^* with density π^* defined as* (3)*.*

Remark 5. *Theorem 2 presents a mixing time bound for Algorithm 4 with a feasible initial distribution as* $O\{d^3\varepsilon^{-2}\log(d/\varepsilon)\}$ *up to* L, m, r *which are specified in Assumptions 1 and 2 if we choose the regularization parameter* $\lambda = \lambda^\star$. *This result improves upon the mixing time bound for constrained sampling algorithm without incorporating the Metropolis–Hastings step in* [35]; *see Table 1 for details.*

5. Numerical Experiments

In this section, we conduct numerical experiments to validate the theoretical properties derived in Section 4 and compare the constrained sampling algorithms presented in Section 3 with three competing MCMC algorithms for sampling from constrained log-concave distributions listed in Table 1 under various simulation settings. The implementation of these algorithms involves the selection of a step size. For Algorithms 2 and 3, we follow Theorem 1 and Corollary 3, respectively, to select the step size. For Algorithm 4, we choose the step size as that in [32] for the MALA for sampling from log-concave distribution without constraints. The step size choice of the other three MCMC algorithms follows the recommendation in the associated papers; see Table 2 for details.

Table 2. Step sizes for sampling from log-concave distributions with bounded support.

Algorithms	Step Size
PLMC in [34]	$L^{-1}d^{-2}$
MYULA in [35]	$\{d\max(d,L)\}^{-1}$
MLD in [36]	the grid search
Algorithm 2 in this paper	$L^{-7/4}R^{-3/2}d^{-1}$
Algorithm 3 in this paper	$L^{-7/4}d^{-1}$
Algorithm 4 in this paper	$\{(L+\lambda^{\star-1})\max[d,\{m^{-1}d(L+\lambda^{\star-1})\}^{1/2}]\}^{-1}$

5.1. Sampling from the Euclidean Ball Constrained Domain

We consider the problem of sampling from a truncated multivariate Gaussian distribution on $\mathcal{X}$, which admits the density

$$\pi^*(x) \propto \exp\left\{-\frac{(x-\mu)^\intercal\Sigma^{-1}(x-\mu)}{2}\right\}I(x\in\mathcal{X}),$$

where the mean $\mu = 0$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is a diagonal matrix with $\lambda_{\max}(\Sigma) = 10$ and $\lambda_{\min}(\Sigma) = 1$. For this target distribution, the potential function $U(\cdot)$ and its derivatives are given as $U(x) = 2^{-1}x^\intercal\Sigma^{-1}x$, $\nabla U(x) = \Sigma^{-1}x$, and $\nabla^2 U(x) = \Sigma^{-1}$. Therefore, the function $U(\cdot)$ is smooth with parameter $L = \lambda_{\min}^{-1}(\Sigma)$ and strongly convex with parameter $m = \lambda_{\max}^{-1}(\Sigma)$ on $\mathbb{R}^d$. We select $\mathcal{X} = \mathbb{B}(\mathbf{0}, R)$ with $R = 5$, the initial distribution $\mathbb{P}^0 = \mathcal{N}_{\mathcal{X}}\{\mathbf{0}, (2L)^{-1}I_d\}$, and use the inverse transformation algorithm [14] to generate an initial point from $\mathbb{P}^0$. We compare Algorithm 2 with the three sampling algorithms in literature given in Table 2, and follow the recommendation in the associated papers to choose the initial points of the three sampling algorithms.

5.1.1. The Trace Graphs of Sampling Algorithms

To initiate a preliminary assessment of the convergence properties of these algorithms, we commence with simple sample trace plots. Write $x = (x_1,\ldots,x_d)^\intercal \in \mathbb{R}^d$ and $\mu = (\mu_1,\ldots,\mu_d)^\intercal \in \mathbb{R}^d$. Figure 1 depicts the traces of x_1 of the Markov chains determined by the four sampling algorithms under dimension $d = 10$. Evidently, in comparison to the other three algorithms, Algorithm 2 exhibits a notably faster mixing time, as evidenced by the trace consistently remaining around its mean $\mu_1 = 0$. Conversely, the traces of the other three sampling algorithms exhibit greater fluctuations and deviate more from $\mu_1 = 0$.

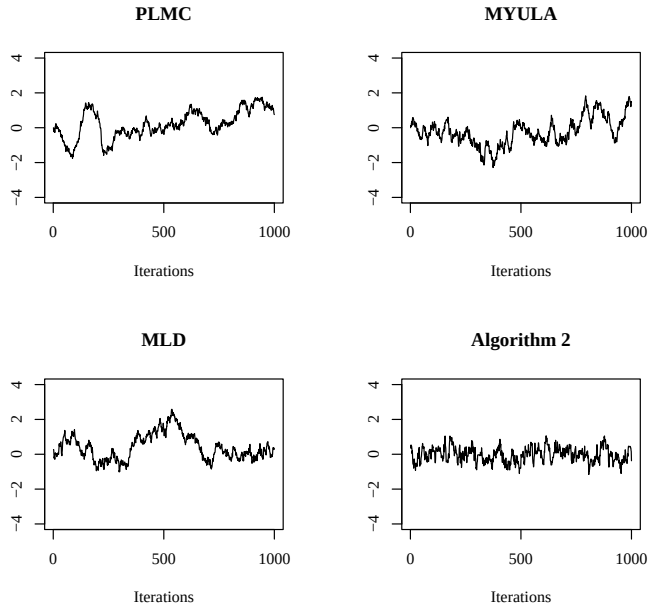


Figure 1. The trace graphs of x_1 of the Markov chain determined by the four sampling algorithms.

Figure 2 illustrates the histograms and densities corresponding to these traces of x_1. Similarly, it is evident that Algorithm 2 achieves sample means closer to $\mu_1 = 0$, along with the least variance. Conversely, the sample means obtained from the other three sampling algorithms exhibit a certain degree of deviation from $\mu_1 = 0$, accompanied by heavier tails.

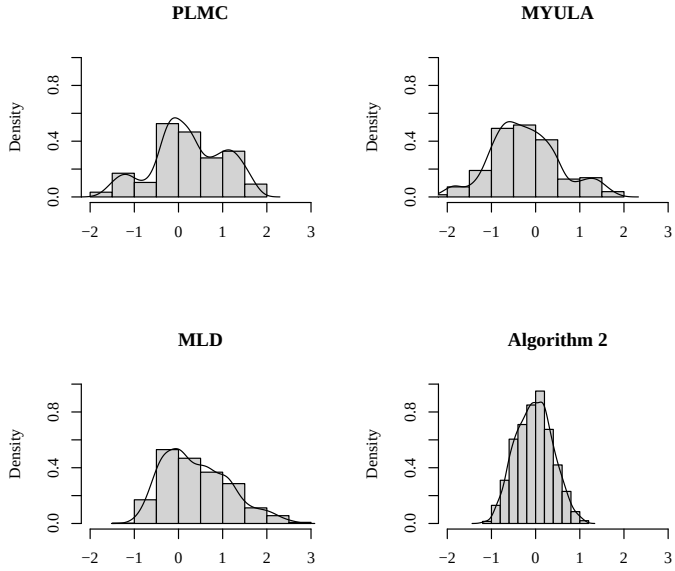


Figure 2. The densities of x_1 of the Markov chain determined by the four sampling algorithms.

5.1.2. Dimension and Error Dependence of Algorithm 2

The goal of this simulation is to demonstrate that the dimension and error tolerance dependence of the mixing time bound for Algorithm 2 both conform to the theoretical results shown in Theorem 1.

Since the total variation distance between continuous measures is hard to estimate, we use the error in quantiles along some direction for convergence diagnostics in the experiments. In the spirit of [33], we measure the error in the 95% quantile of the sample distribution and the true distribution in the direction along the eigenvector of Σ corresponding to $\lambda_{\min}(\Sigma)$. The approximate mixing time $\hat{k}_{\text{mix}}(\varepsilon)$ is then defined as the smallest iteration k when such error between the distribution of the Markov chain at iteration k

and the target distribution falls below the error tolerance ε. We simulate 20 independent runs of the Markov chain of the algorithms with $N = 20{,}000$ samples at each run to determine the approximate mixing time $\hat{k}_{\text{mix}}(\varepsilon)$. Then the final $\hat{k}_{\text{mix}}(\varepsilon)$ is the average of these 20 independent runs.

Figure 3a shows the dependence of the approximate mixing time $\hat{k}_{\text{mix}}(0.2)$ as a function of dimension d for Algorithm 2. By the linear regression for $\hat{k}_{\text{mix}}(0.2)$ with respect to d, we conclude that the mixing time of Algorithm 2 is linear in d with slope 4.137 and R-squared 0.991. Figure 3b presents the dependence of the approximate mixing time $\hat{k}_{\text{mix}}(\varepsilon)$ on the inverse of the error tolerance ε^{-1} for Algorithm 2 under $d = 4$. The linear regression for the approximate mixing time $\hat{k}_{\text{mix}}(\varepsilon)$ with respect to ε^{-1} suggests that the mixing time of Algorithm 2 is linear in $\log(\varepsilon^{-1})$ with slope 15.854 and R-squared 0.994, which is consistent with the theoretical results given in Theorem 1.

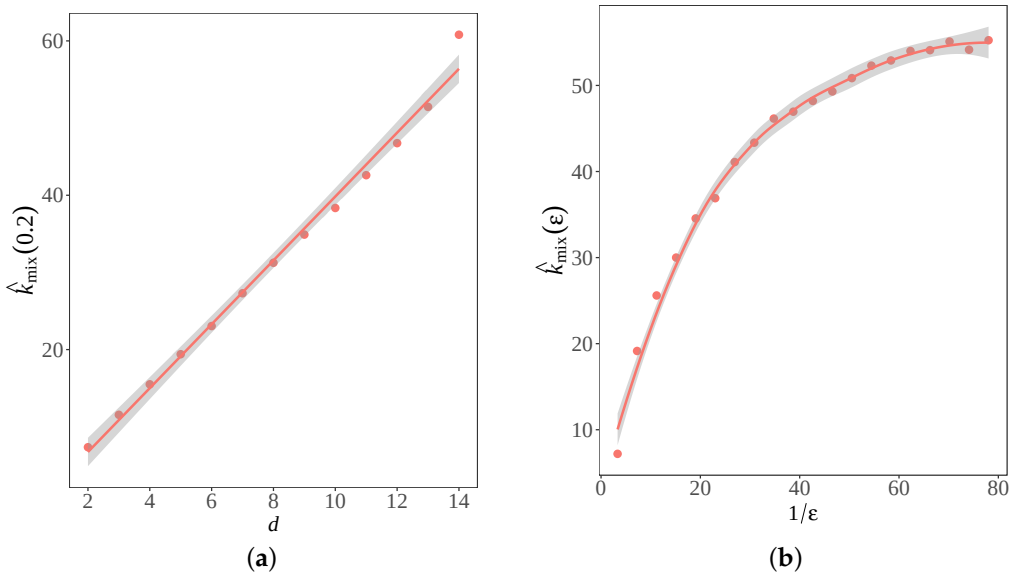


Figure 3. Approximate mixing time with respect to dimension and error tolerance of Algorithm 2. (**a**) Dimension dependence for fixed error tolerance. (**b**) Error tolerance dependence for fixed dimension.

5.1.3. Comparison with Competitive Algorithms

Figure 4a shows the dependence of the approximate mixing time $\hat{k}_{\text{mix}}(0.2)$ on the problem dimension d for the four sampling algorithms. Compared with the other three algorithms, the approximate mixing time of Algorithm 2 seems more robust to dimension. When d is small, the approximate mixing time of the four algorithms is comparatively close. However, as the dimension d increases, the approximate mixing time of PLMC and MYULA increases rapidly, showing a polynomial order with respect to d. Moreover, the dimension dependence of MLD and Algorithm 2 both indicate a linear growth trend, and MLD needs a few more steps than Algorithm 2 to reach the same error tolerance.

Figure 4b presents the dependence of the approximate mixing time $\hat{k}_{\text{mix}}(\varepsilon)$ on the inverse of the error tolerance ε^{-1} for the four sampling algorithms under $d = 4$. The regression analysis shows that the approximate mixing time $\hat{k}_{\text{mix}}(\varepsilon)$ of PLMC and MYULA increases in polynomial order of ε^{-1}. When ε^{-1} is relatively small, MLD and Algorithm 2 have similar approximate mixing time. With the increase in ε^{-1}, the strength of Algorithm 2 gets more significant. For MLD, the linear regression for the approximate mixing time $\hat{k}_{\text{mix}}(\varepsilon)$ with respect to ε^{-2} yields a slope of 1.934 and R-squared 0.984, suggesting the error tolerance dependence of order ε^{-2}.

It is noteworthy that the above analysis not only suggests significantly better dimension and error tolerance dependence of the constrained MALA but also partly verifies the theoretical convergence rates of the three methods for comparison.

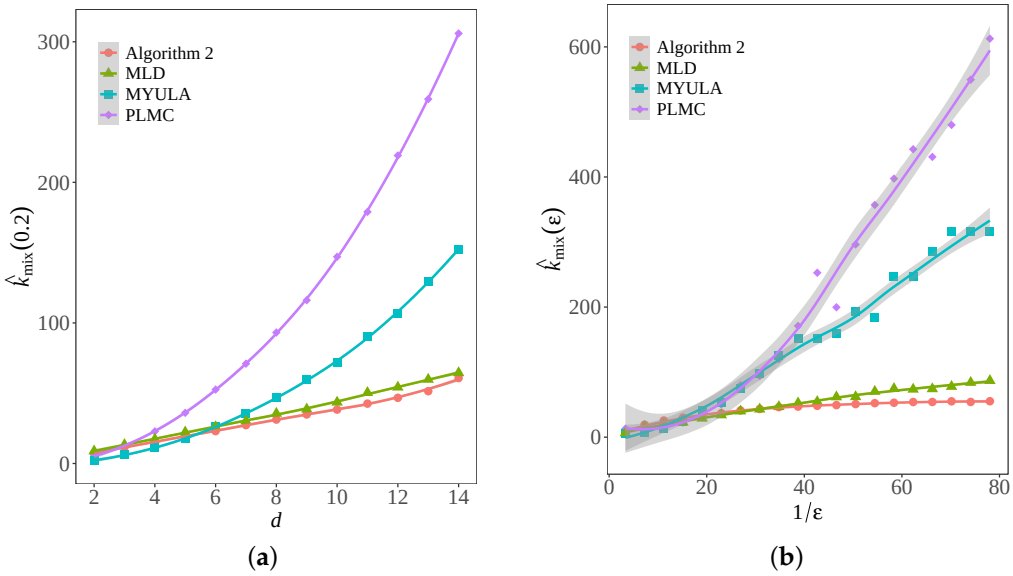


Figure 4. Approximate mixing time with respect to dimension and error tolerance dependence of the four sampling algorithms. (**a**) Dimension dependence for fixed error tolerance. (**b**) Error tolerance dependence for fixed dimension.

5.2. Bayesian Regularized Regression

The regularized regression involves adding a penalty term on the objective function of the regression model, which helps to control the complexity of the model and prevent it from fitting the noise in the data. In this section, we validate the effectiveness of Algorithm 3 for constrained sampling involving the Bayesian regularized regression.

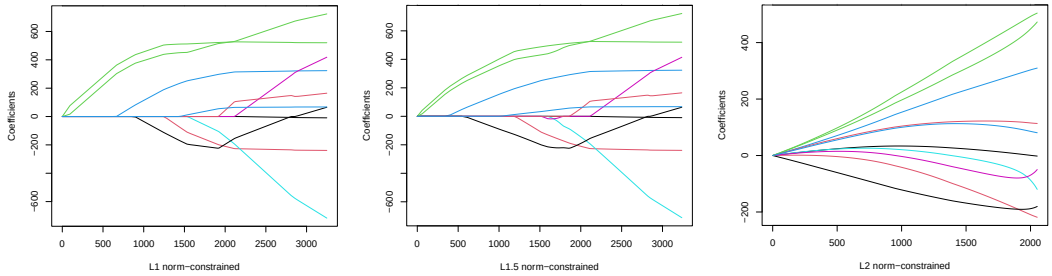
Given the independent and identically observations $y = (y_1, y_2, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ which follow from the Gaussian distribution with mean $X\beta$ and covariance matrix $\sigma^2 I_n$, we consider the regression models where the parameter are obtain by minimizing the square of Euclidean norm of the residual subject to a norm-constraint on the regression parameter as follows:

$$\min_{\beta \in \mathbb{R}^d} |y - X\beta|_2^2 \quad \text{subject to} \quad |\beta|_p \leq C$$

for some universal constant $C > 0$, where $X \in \mathbb{R}^{n \times d}$ is the design matrix, $\beta \in \mathbb{R}^d$ is the regression parameter, and $|\beta|_p$ is the L_p-norm of β. In Bayesian setting, many regularization techniques correspond to imposing certain prior distributions on model parameters. We then consider sampling from the distribution with density

$$\pi^*(x) \propto \exp\left\{ -\frac{|y - X\beta|_2^2}{2\sigma^2} \right\} I(x \in \mathcal{X}),$$

and obtaining the parameter estimates $\hat{\beta}$ via the maximum a posteriori probability (MAP) estimate, where $\mathcal{X} = \{x \in \mathbb{R}^d : |x|_p \leq C\}$. We use the diabetes data studied in [41], and set the burn-in period to be 10^3 iterations and $\sigma^2 = 1$. Figure 5 presents the paths of the parameter estimates under different norm constraints, which demonstrate that Algorithm 3 can effectively handle the norm-constrained sampling problems.



(**a**) (**b**) (**c**)

Figure 5. Bayesian regularized regression via Algorithm 3, where distinct colors represent various trajectories of parameter estimates for distinct variables. (**a**) L_1—norm-constraint. (**b**) $L_{1.5}$—norm-constraint. (**c**) L_2—norm-constraint.

5.3. Truncated Multivariate Gaussian Distribution

The final comparison was made by examining the sampling performance of MYULA in [35] and Algorithm 4 in the setting of a more general truncated multivariate Gaussian distribution. We consider the same setup as in [35]. Specifically, the density of the target distribution is defined as follows:

$$\pi^*(x) \propto \exp\left\{-\frac{(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}{2}\right\} I(x \in \mathcal{X}),$$

where $\mathcal{X}$ is a convex set and the origin 0 is on its boundary. Let $\mu = 0$, the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ with (i,j)-th element given by $(\Sigma)_{i,j} = 1/(1+|i-j|)$, and $\mathcal{X} = [0,5]\times[0,1]$. We generate 10^6 samples for Algorithm 4, and set the burn-in period to be the initial 10% iterations.

Table 3 presents the mean and covariance estimation results of the target distribution based on the samples generated by MYULA and Algorithm 4. For comparison purposes, the results of MYULA align with those reported in [35]. With the same number of iterations, Algorithm 4 outperforms MYULA in terms of the estimation results. This indicates that incorporating the Metropolis–Hastings step in Algorithm 4 leads to improvements in the mixing time.

Table 3. The mean and covariance estimation results obtained by MYULA and Algorithm 4.

Assumptions	Mean	Covariance
The truth	$\begin{pmatrix}0.790\\0.488\end{pmatrix}$	$\begin{pmatrix}0.326 & 0.017\\0.017 & 0.080\end{pmatrix}$
MYULA	$\begin{pmatrix}0.758\pm0.052\\0.484\pm0.016\end{pmatrix}$	$\begin{pmatrix}0.309\pm0.038 & 0.017\pm0.009\\0.017\pm0.009 & 0.088\pm0.002\end{pmatrix}$
Algorithm 4	$\begin{pmatrix}0.781\pm0.034\\0.491\pm0.009\end{pmatrix}$	$\begin{pmatrix}0.317\pm0.012 & 0.017\pm0.004\\0.017\pm0.004 & 0.082\pm0.003\end{pmatrix}$

6. Discussion and Conclusions

In this article, we propose three sampling algorithms based on Langevin Monte Carlo with the Metropolis–Hastings steps to handle the distribution constrained within some convex body, and establish the mixing time bounds of these algorithms for sampling from strongly log-concave distributions. Under certain conditions, these bounds are sharper than existing algorithms in the literature. Furthermore, in comparison to existing algorithms, the suggested constrained sampling algorithms are simpler, more intuitive, and easier to operate in some cases.

Our results demonstrate that the sampling algorithm, enhanced with the Metropolis–Hastings step, offers an effective solution for tackling some constrained sampling problems. Numerical experiments fully illustrate the advantages of the proposed algorithms. Although we focus on the strongly log-concave distributions in the theoretical analysis, the

proposed algorithm can be readily applied to weakly log-concave distributions or non-convex potential functions. Simultaneously, we recognize that there are various aspects of the sampling algorithms that can be further improved. For instance, potential enhancements could involve the multiple importance sampling methods or adaptive techniques. We leave the investigation of its theoretical properties under such scenarios for future work.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are included within the article.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Some Markov Chain Basics

Consider the time-homogeneous (We say that a Markov chain is time-homogeneous in which the probability of any state transition is independent of time.) Markov chains defined on a measurable state space $\{\mathcal{X}, \mathscr{B}(\mathcal{X})\}$ with a transition probability $\Psi : \mathcal{X} \times \mathscr{B}(\mathcal{X}) \mapsto [0,1]$. The transition probability satisfies

$$\Psi(x, \mathrm{d}y) \geq 0 \ \ \forall\, x \in \mathcal{X}\,, \text{ and } \int_{y \in \mathcal{X}} \Psi(x, \mathrm{d}y) = 1\,.$$

The k-th step transition probability defined recursively as

$$\Psi^k(x, \mathrm{d}y) = \int_{z \in \mathcal{X}} \Psi^{k-1}(x, \mathrm{d}z)\Psi(z, \mathrm{d}y)\,.$$

For a distribution Π on $\mathcal{X}$, a Markov chain defined on $\{\mathcal{X}, \mathscr{B}(\mathcal{X})\}$ is called Π-irreducible if for each $A \in \mathscr{B}(\mathcal{X})$ with $\Pi(A) > 0$ and each $x \in \mathcal{X}$, there exists $k \in \mathbb{N}$ such that $\Psi^k(x, A) > 0$. A Markov chain defined on $\{\mathcal{X}, \mathscr{B}(\mathcal{X})\}$ with transition probability Ψ: $\mathcal{X} \times \mathscr{B}(\mathcal{X}) \mapsto [0,1]$ and stationary distribution Π is called reversible if it satisfies the detailed balance condition $\Pi(\mathrm{d}x)\Psi(x, \mathrm{d}y) = \Pi(\mathrm{d}y)\Psi(y, \mathrm{d}x)$ for any $x, y \in \mathcal{X}$.

Smooth chain assumption. We say that the Markov chain satisfies the smooth chain condition if its transition probability $\Psi : \mathcal{X} \times \mathscr{B}(\mathcal{X}) \mapsto [0,1]$ can be expressed in the form

$$\Psi(x, \mathrm{d}y) = \psi(x, y)\,\mathrm{d}y + \iota_x \delta_x(\mathrm{d}y)$$

for any $x, y \in \mathcal{X}$, where $\psi(\cdot\,, \cdot)$ is the transition kernel satisfying $\psi(x, y) \geq 0$ for any $x, y \in \mathcal{X}$, ι_x denotes the one-step probability of the chain to stay at its current state x, and $\delta_x(\cdot)$ is the Dirac-delta function at x.

Log-isoperimetric inequality. A distribution Π supported on $\mathcal{X}$ with density π is said to satisfy the log-isoperimetry inequality with some constant $\hat{c} > 0$ if

$$\Pi(S_3) \geq \frac{d(S_1, S_2)}{2\hat{c}} \min\{\Pi(S_1), \Pi(S_2)\} \log^{1/2}\left[1 + \frac{1}{\min\{\Pi(S_1), \Pi(S_2)\}}\right] \tag{A1}$$

for any partition $(S_1,\ S_2,\ S_3)$ of $\mathcal{X}$, where $\Pi(S_i) = \int_{S_i} \pi(x)\,\mathrm{d}x$ and $d(S_1, S_2) = \inf_{x \in S_1, y \in S_2} |x - y|_2$.

Conductance profile. Given a Markov chain with transition probability $\Psi : \mathcal{X} \times \mathscr{B}(\mathcal{X}) \mapsto [0,1]$ and stationary distribution Π with density π, its stationary flow $\omega(\cdot) : \mathscr{B}(\mathcal{X}) \mapsto \mathbb{R}$ is defined as

$$\omega(S) = \int_S \Psi(x, S^{\mathrm{c}})\pi(x)\,\mathrm{d}x \tag{A2}$$

for any $S \in \mathscr{B}(\mathcal{X})$. For any $v \in (0, 1/2]$, the conductance profile is given by

$$\Omega(v) = \inf_{S:\Pi(S)\in(0,v]} \frac{\omega(S)}{\Pi(S)}. \tag{A3}$$

Furthermore, the extended conductance profile is defined as

$$\tilde{\Omega}(v) = \begin{cases} \Omega(v), & v \in (0, 1/2], \\ \Omega(1/2), & v \in (1/2, \infty). \end{cases} \tag{A4}$$

Appendix B. Proofs

Appendix B.1. Proof of Proposition 1

Proof of Proposition 1. Denote by $\Psi(x, \cdot)$ the transition probability of the Markov chain at $x \in \mathcal{X}$ determined by Algorithm 2. For any $x \in \mathcal{X}$, let $\mathcal{P}_{x,h} = \mathcal{N}\{x - h\nabla U(x), 2hI_d\}$ with the step size h. Write the density of $\mathcal{P}_{x,h}$ as $\phi_h(\cdot \mid x)$. For any $x \in \mathcal{X}$, denote by $\alpha_x(y) = \min\{1, R_x(y)\}$ the acceptance probability for any $y \in \mathbb{R}^d$, where

$$R_x(y) = \frac{\pi^*(y)\phi_h(x \mid y)}{\pi^*(x)\phi_h(y \mid x)} I(y \in \mathcal{X}).$$

Then, the transition probability of the associated Markov chain at $x \in \mathcal{X}$ has a probability mass $\psi_x = 1 - \int_{\mathcal{X}} \phi_h(y \mid x)\alpha_x(y)\,\mathrm{d}y$. Define the transition kernel

$$\psi(x, y) = \phi_h(y \mid x)\alpha_x(y) I(y \in \mathcal{X} \setminus \{x\})$$

for $x \in \mathcal{X}$. Then, the transition probability $\Psi : \mathcal{X} \times \mathscr{B}(\mathcal{X}) \mapsto [0, 1]$ satisfies

$$\Psi(x, \mathrm{d}y) = \psi_x \delta_x(\mathrm{d}y) + \psi(x, y)\,\mathrm{d}y, \tag{A5}$$

where $\delta_x(\cdot)$ is the Dirac-delta function at x. By the smooth chain condition given in Appendix A, we know the Markov chain with the transition probability $\Psi(\cdot, \cdot)$ is smooth.

Recall that Π^* is the distribution on $\mathcal{X}$ with the density π^* defined as (3). Since

$$\alpha_x(y)\pi^*(x)\phi_h(y \mid x) = \alpha_y(x)\pi^*(y)\phi_h(x \mid y)$$

for any $x, y \in \mathcal{X}$, then $\pi^*(x)\psi(x, y) = \pi^*(y)\psi(y, x)$ for any $x, y \in \mathcal{X}$. Together with (A5), for any $A, B \in \mathscr{B}(\mathcal{X})$, it holds that

$$\begin{aligned} \int_A \pi^*(x)\Psi(x, B)\,\mathrm{d}x &= \int_{A\cap B} \pi^*(x)\psi_x\,\mathrm{d}x + \int_{(x,y)\in A\times B} \pi^*(x)\psi(x, y)\,\mathrm{d}x\mathrm{d}y \\ &= \int_B \pi^*(x)\psi_x\delta_x(A)\,\mathrm{d}x + \int_{(x,y)\in A\times B} \pi^*(y)\psi(y, x)\,\mathrm{d}x\mathrm{d}y \\ &= \int_B \pi^*(x)\Psi(x, A)\,\mathrm{d}x \end{aligned}$$

with $\delta_x(A) = I(x \in A)$, which implies $\Pi^*(A) = \int_A \pi^*(x)\Psi(x, \mathcal{X})\,\mathrm{d}x = \int_{\mathcal{X}} \pi^*(x)\Psi(x, A)\,\mathrm{d}x$ for any $A \in \mathscr{B}(\mathcal{X})$. Thus, Π^* is the stationary distribution of the Markov chain with the transition probability $\Psi(\cdot, \cdot)$. Hence, such Markov chain is reversible.

Furthermore, by (A5), we have

$$\Psi(x, A) = \psi_x\delta_x(A) + \int_A \psi(x, y)\,\mathrm{d}y$$

for any $x \in \mathcal{X}$ and $A \in \mathscr{B}(\mathcal{X})$. For any $A \in \mathscr{B}(\mathcal{X})$ with $\Pi^*(A) > 0$, due to $\Pi^*(A) = \int_A \pi^*(x)\,\mathrm{d}x$, we know the Lebesgue measure of A is nonzero. Since $\alpha_x(y) \le 1$ and $\mathcal{X} = \mathbb{B}(x^*, R)$ for some universal constant $R > 0$ and $x^* \in \mathbb{R}^d$, we know $\psi_x \ge 1 - \int_{\mathcal{X}} \phi_h(y \mid x)\,\mathrm{d}y > 0$ for any $x \in \mathcal{X}$. If $A = \{x\}$, $\Psi(x, A) \ge \psi_x > 0$. If $A \ne \{x\}$, we know the

Lebesgue measure of $A \setminus \{x\}$ is also nonzero, which implies $\Psi(x, A) \geq \int_{A\setminus\{x\}} \psi(x, y)\,\mathrm{d}y > 0$. Thus, the Markov chain with the transition probability $\Psi(\cdot\,, \cdot)$ is Π^*-irreducible. We complete the proof of Proposition 1. □

Appendix B.2. Proof of Proposition 2

Proof of Proposition 2. Recall $\mathcal{X} = \{x \in \mathbb{R}^d : |x|_p \leq C\}$ for some universal constant $C > 0$. Notice that the additional two steps are introduced in Algorithm 3 only for the purpose of establishing a one-to-one mapping between $\{x \in \mathbb{R}^d : |x|_p \leq C\}$ and $\mathbb{B}(0, 1)$, and they do not affect the properties of the Markov chain. Using the same arguments in the proof of Proposition 1, we can obtain the results of Proposition 2. □

Appendix B.3. Proof of Proposition 3

Proof of Proposition 3. The proof is almost identical to that of Proposition 1. Recall the distribution $\Pi^{*,\lambda}$ with density

$$\pi^{*,\lambda}(x) = \frac{\exp\{-V^\lambda_{\mathcal{X}}(x)\}}{\int_{\mathcal{X}} \exp\{-V^\lambda(y)\}\,\mathrm{d}y},$$

where $V^\lambda_{\mathcal{X}}(\cdot) = U(\cdot) + \iota^\lambda_{\mathcal{X}}(\cdot)$ with $\nabla\iota^\lambda_{\mathcal{X}}(\cdot)$ defined as (8). Let $\phi^\lambda_h(\cdot \,|\, x)$ be the probability density function of the Gaussian distribution $\mathcal{N}\{x - h\{\nabla U(x) + \nabla\iota^\lambda_{\mathcal{X}}(x)\}, 2hI_d\}$. We only need to replace $\{\Pi^*, \pi^*, \phi_h(\cdot \,|\, x)\}$ which appeared in the proof of Proposition 1 by $\{\Pi^{*,\lambda}, \pi^{*,\lambda}, \phi^\lambda_h(\cdot \,|\, x)\}$ and all the arguments still hold. □

Appendix B.4. Proof of Lemma 1

Proof of Lemma 1. We introduce some notation first. Denote by π the density function of Π, and $L_2(\pi)$ the space of square integrable functions defined on $\mathcal{X}$ under the density π, that is,

$$\int_{\mathcal{X}} g^2(x)\pi(x)\,\mathrm{d}x < \infty$$

for any $g \in L_2(\pi)$. The Dirichlet form $\mathcal{E}_\Psi : L_2(\pi) \times L_2(\pi) \mapsto \mathbb{R}$ associated with the transition probability $\Psi(\cdot\,, \cdot)$ is defined as follows:

$$\mathcal{E}_\Psi(g, h) = \frac{1}{2}\int_{(x,y)\in\mathcal{X}^2} \{g(x) - h(y)\}^2\Psi(x, \mathrm{d}y)\pi(x)\,\mathrm{d}x. \tag{A6}$$

For any $g \in L_2(\pi)$, let

$$\mathbb{E}_\pi(g) = \int_{\mathcal{X}} g(x)\pi(x)\,\mathrm{d}x \quad \text{and} \quad \mathrm{Var}_\pi(g) = \int_{\mathcal{X}} \{g(x) - \mathbb{E}_\pi(g)\}^2\pi(x)\,\mathrm{d}x.$$

For a measurable non-empty subset $S \subset \mathcal{X}$, the spectral gap is defined as

$$\lambda(S) = \inf_{g\in c_0^+(S)} \frac{\mathcal{E}_\Psi(g, g)}{\mathrm{Var}_\pi(g)},$$

where $c_0^+(S) = \{g \in L_2(\pi) : \mathrm{supp}(g) \subset S, g \geq 0, \mathrm{Var}_\pi(g) > 0\}$. Define the spectral profile $\Lambda(\cdot)$ as

$$\Lambda(v) = \inf_{S:\,\Pi(S)\in(0,v]} \lambda(S) \tag{A7}$$

for any $v \in (0, \infty)$. If the current state of a Markov chain admits the distribution $\mathbb{P}$ with density p, we write $\mathcal{T}(p)$ as the distribution of its next state. The proof of Lemma 1 includes two steps. The first step is to show

$$\tau(\varepsilon; \mathbb{P}^0, \Pi) \leq \frac{1}{\varsigma} \int_{4\beta^{-1}}^{\varepsilon^{-2}} \frac{\mathrm{d}v}{v\Lambda(v)}\,.$$

The second step is to show that the spectral profile and the conductance profile defined in (A3) are related as

$$\Lambda(v) \geq \begin{cases} \dfrac{\Omega^2(v)}{2}\,, & v \in (0, 1/2]\,, \\ \dfrac{\Omega^2(1/2)}{4}\,, & v \in (1/2, \infty)\,. \end{cases}$$

Notice that $\Pi(\mathcal{X}) = 1$. Replacing the restricted conductance profile and restricted spectral gap in the proof of Lemma 1 in [33] by the conductance profile and spectral gap, respectively, and using the similar arguments in the proof of Lemma 1 in [33], we can obtain the results of the two steps. Then, Lemma 1 can be constructed immediately. □

Appendix B.5. Proof of Lemma 2

Proof of Lemma 2. Denote by π the density function of the distribution Π. For any measurable non-empty subset $A_1 \subset \mathcal{X}$ such that $0 < \Pi(A_1) \leq 1/2$, we have $\Pi(A_2) \geq 1/2 \geq \Pi(A_1)$, where $A_2 = \mathcal{X} \setminus A_1$. Given $\delta > 0$, we define the following sets

$$A_1' = \{x \in A_1 : \Psi(x, A_2) < \delta/2\}\,, \quad A_2' = \{x \in A_2 : \Psi(x, A_1) < \delta/2\}$$

and $A_3' = \mathcal{X} \setminus (A_1' \bigcup A_2')$, where $\Psi : \mathcal{X} \times \mathscr{B}(\mathcal{X}) \mapsto [0, 1]$ is the transition probability of the considered Markov chain.

On the one hand, if $\Pi(A_1') \leq \Pi(A_1)/2$, then $\Pi(A_1 \setminus A_1') \geq \Pi(A_1)/2$. Thus,

$$\int_{A_1} \Psi(x, A_2)\pi(x)\,\mathrm{d}x \geq \int_{A_1 \setminus A_1'} \Psi(x, A_2)\pi(x)\,\mathrm{d}x \geq \frac{\delta}{2} \int_{A_1 \setminus A_1'} \pi(x)\,\mathrm{d}x \geq \frac{\delta}{4}\Pi(A_1)\,.$$

Similarly, if $\Pi(A_2') \leq \Pi(A_2)/2$, we have $\int_{A_2} \Psi(x, A_1)\pi(x)\,\mathrm{d}x \geq \delta\Pi(A_2)/4$. By the detailed balance condition and the Fubini's theorem, it holds that

$$\begin{aligned} \int_{A_1} \Psi(x, A_2)\pi(x)\,\mathrm{d}x &= \int_{x \in A_1} \int_{y \in A_2} \Psi(x, \mathrm{d}y)\pi(x)\,\mathrm{d}x \\ &= \int_{x \in A_1} \int_{y \in A_2} \Psi(y, \mathrm{d}x)\pi(y)\,\mathrm{d}y \\ &= \int_{A_2} \Psi(y, A_1)\pi(y)\,\mathrm{d}y = \int_{A_2} \Psi(x, A_1)\pi(x)\,\mathrm{d}x\,. \end{aligned} \tag{A8}$$

Therefore, if $\Pi(A_1') \leq \Pi(A_1)/2$ or $\Pi(A_2') \leq \Pi(A_2)/2$, we have

$$\int_{A_1} \Psi(x, A_2)\pi(x)\,\mathrm{d}x \geq \frac{\delta}{4} \min\{\Pi(A_1), \Pi(A_2)\} = \frac{\delta}{4}\Pi(A_1)\,.$$

On the other hand, we consider the case with $\Pi(A_1') > \Pi(A_1)/2$ and $\Pi(A_2') > \Pi(A_2)/2$. Notice that $\mathcal{T}_x(\cdot) = \Psi(x, \cdot)$. By the definition of the total variation distance, for any $x \in A_1'$ and $y \in A_2'$, we have

$$\|\mathcal{T}_x - \mathcal{T}_y\|_{\mathrm{TV}} \geq \Psi(x, A_1) - \Psi(y, A_1) = 1 - \Psi(x, A_2) - \Psi(y, A_1) > 1 - \delta\,.$$

Since $\sup_{x,y\in\mathcal{X}:|x-y|_2\le\Delta}\|\mathcal{T}_x-\mathcal{T}_y\|_{\mathrm{TV}}\le 1-\delta$, we know $|x-y|_2>\Delta$, which implies $d(A_1',A_2'):=\inf_{x\in A_1',y\in A_2'}|x-y|_2\ge\Delta$. Recall $A_3'=\mathcal{X}\setminus(A_1'\bigcup A_2')$. By (A8),

$$\begin{aligned}\int_{A_1}\Psi(x,A_2)\pi(x)\,\mathrm{d}x&=\frac{1}{2}\int_{A_1}\Psi(x,A_2)\pi(x)\,\mathrm{d}x+\frac{1}{2}\int_{A_2}\Psi(x,A_1)\pi(x)\,\mathrm{d}x\\&\ge\frac{1}{2}\int_{A_1\setminus A_1'}\Psi(x,A_2)\pi(x)\,\mathrm{d}x+\frac{1}{2}\int_{A_2\setminus A_2'}\Psi(x,A_1)\pi(x)\,\mathrm{d}x\\&\ge\frac{\delta}{4}\Pi(A_3')\,.\end{aligned}\tag{A9}$$

Since $\Pi(A_1')>\Pi(A_1)/2$, $\Pi(A_2')>\Pi(A_2)/2$ and the sets (A_1',A_2',A_3') partition $\mathcal{X}$, by the log-isoperimetry inequality given in (A1), it holds that

$$\begin{aligned}\Pi(A_3')&\ge\frac{d(A_1',A_2')}{2\hat{c}}\min\{\Pi(A_1'),\Pi(A_2')\}\log^{1/2}\left[1+\frac{1}{\min\{\Pi(A_1'),\Pi(A_2')\}}\right]\\&\ge\frac{\Delta}{4\hat{c}}\min\{\Pi(A_1),\Pi(A_2)\}\log^{1/2}\left[1+\frac{2}{\min\{\Pi(A_1),\Pi(A_2)\}}\right]\\&\ge\frac{\Delta}{4\hat{c}}\Pi(A_1)\log^{1/2}\left\{1+\frac{1}{\Pi(A_1)}\right\},\end{aligned}\tag{A10}$$

where the second inequality follows from the fact that $x\log^{1/2}(1+x^{-1})$ is non-decreasing in $x>0$. By (A9) and (A10), we have

$$\int_{A_1}\Psi(x,A_2)\pi(x)\,\mathrm{d}x\ge\frac{\delta\Delta}{16\hat{c}}\Pi(A_1)\log^{1/2}\left\{1+\frac{1}{\Pi(A_1)}\right\}.$$

Putting the two cases together, it holds that

$$\omega(A_1)=\int_{A_1}\Psi(x,A_2)\pi(x)\,\mathrm{d}x\ge\frac{\delta}{4}\Pi(A_1)\min\left[1,\frac{\Delta}{4\hat{c}}\log^{1/2}\left\{1+\frac{1}{\Pi(A_1)}\right\}\right]$$

for any measurable non-empty subset $A_1\subset\mathcal{X}$ with $0<\Pi(A_1)\le 1/2$. Due to $\inf_{x\in(0,v]}\log^{1/2}(1+x^{-1})=\log^{1/2}(1+v^{-1})$, by the definition of the conductance profile given in (A9), we have

$$\Omega(v)\ge\frac{\delta}{4}\min\left\{1,\frac{\Delta}{4\hat{c}}\log^{1/2}\left(1+\frac{1}{v}\right)\right\}$$

for any $v\in(0,1/2]$. We complete the proof of Lemma 2. □

Appendix B.6. Proof of Theorem 1

For any $x\in\mathcal{X}$, let $\mathcal{P}_{x,h}=\mathcal{N}\{x-h\nabla U(x),2hI_d\}$ with the step size h. For $\mathcal{X}=\mathbb{B}(x^*,R)$ with some universal constant $R>0$ and $x^*\in\mathbb{R}^d$, without loss of generality, we set $x^*=\arg\min_{x\in\mathbb{R}^d}U(x)$. Under Assumption 1, we know $\nabla U(x^*)=0$.

Lemma A1. *Let $\mathcal{X}=\mathbb{B}(x^*,R)$ for some universal constant $R>0$ and $x^*=\arg\min_{x\in\mathbb{R}^d}U(x)$, and Assumption 1 hold. For any step size $h\in(0,2L^{-1}]$ with L specified in Assumption 1, it holds that*

$$\|\mathcal{P}_{x,h}-\mathcal{P}_{x,h}\|_{\mathrm{TV}}\le\frac{|x-y|_2}{\sqrt{2h}}\tag{A11}$$

for any $x,y\in\mathcal{X}$. Furthermore, if $L^{3/8}R^{3/4}\ge 16d^{-1/2}+8$ and $L^{-15/8}m^2R^{1/4}\ge 12d$, for any $u\in(1/2,1)$, it holds that

$$\sup_{x\in\mathcal{X}}\|\mathcal{P}_{x,h}-\mathcal{T}_x\|_{\mathrm{TV}}\le\frac{u}{4}\tag{A12}$$

for any step size h satisfying

$$\frac{1}{L^{7/4}R^{3/2}d} \le h \le \min\left[\frac{R^2(1-\tilde{c})^2}{4\{\log^{1/2}(16u^{-1})+\sqrt{d}\}^2}, \frac{\sqrt{u}}{4\sqrt{3}L^{3/2}R}, \frac{u}{128L\{\log^{1/2}(16u^{-1})+\sqrt{d}\}^2}\right]$$

with $\tilde{c} = \{1+(L^{-7/2}R^{-3}d^{-2} - L^{-11/4}R^{-3/2}d^{-1})m^2\}^{1/2}$, where m is specified in Assumption 1, and $\mathcal{T}_x$ is the one-step transition distribution of the associated Markov chain involved in Algorithm 2 at $x \in \mathcal{X}$.

Proof of Lemma A1. Firstly, we prove the first claim (A11) of this lemma. Recall $\mathcal{P}_{x,h} = \mathcal{N}\{x - h\nabla U(x), 2hI_d\}$ with the step size h. For any $x, y \in \mathcal{X}$, by the Pinsker's inequality, we have

$$\|\mathcal{P}_{x,h} - \mathcal{P}_{y,h}\|_{\mathrm{TV}} \le \sqrt{2\mathrm{KL}(\mathcal{P}_{x,h} \,\|\, \mathcal{P}_{y,h})} = (2h)^{-1/2}|\{x - h\nabla U(x)\} - \{y - h\nabla U(y)\}|_2\,,$$

where $\mathrm{KL}(\mathcal{P}_{x,h} \,\|\, \mathcal{P}_{y,h})$ is the Kullback–Leibler divergence between $\mathcal{P}_{x,h}$ and $\mathcal{P}_{y,h}$. Under Assumption 1, by the Taylor expansion, it holds that

$$\begin{aligned}|\{x - h\nabla U(x)\} - \{y - h\nabla U(y)\}|_2 &= |\{I_d - h\nabla^2 U(z)\}(x-y)|_2 \\ &\le \|I_d - h\nabla^2 U(z)\|_2 |x-y|_2\end{aligned}$$

for some z lying on the jointing line between x and y. Since $\mathcal{X} = \mathbb{B}(x^*, R)$ for some universal constant $R > 0$ and $U(\cdot)$ is L-smooth and m-strongly convex on $\mathcal{X}$, by Theorems 2.1.6 and 2.1.11 of [42], we have $mI_d \preceq \nabla^2 U(z) \preceq LI_d$. Due to $h \in (0, 2L^{-1}]$, then

$$\lambda_{\max}\{I_d - h\nabla^2 U(z)\} \le \lambda_{\max}(I_d) + \lambda_{\max}\{-h\nabla^2 U(z)\} \le 1 - mh \le 1\,,$$

and

$$\lambda_{\min}\{I_d - h\nabla^2 U(z)\} \ge \lambda_{\min}(I_d) + \lambda_{\min}\{-h\nabla^2 U(z)\} \ge 1 - Lh \ge -1$$

for all $z \in \mathcal{X}$. Therefore, we can obtain $\sup_{z\in\mathcal{X}} \|I_d - h\nabla^2 U(z)\}\|_2 \le 1$, which implies that

$$\|\mathcal{P}_{x,h} - \mathcal{P}_{y,h}\|_{\mathrm{TV}} \le \frac{|x-y|_2}{\sqrt{2h}}$$

for any $x, y \in \mathcal{X}$. It yields the claim (A11).

Next, we will prove the second claim (A12) of this lemma. Write the density of $\mathcal{P}_{x,h}$ as $\phi_h(\cdot \,|\, x)$. Notice that the one-step transition distribution of the associated Markov chain at $x \in \mathcal{X}$ has a probability mass

$$\mathcal{T}_x(\{x\}) = 1 - \int_{\mathcal{X}} \phi_h(z \,|\, x)\alpha_x(z)\,\mathrm{d}z\,,$$

and admits a transition kernel $\phi_h(z \,|\, x)\alpha_x(z)I(z \in \mathcal{X} \setminus \{x\})$, where

$$\alpha_x(z) = \min\left\{1, \frac{\pi^*(z)\phi_h(x \,|\, z)}{\pi^*(x)\phi_h(z \,|\, x)} I(z \in \mathcal{X})\right\}.$$

By the definition of the total variation distance, we have

$$\begin{aligned}\|\mathcal{P}_{x,h} - \mathcal{T}_x\|_{\mathrm{TV}} &= \frac{1}{2}\mathcal{T}_x(\{x\}) + \frac{1}{2}\int_{\mathbb{R}^d} |\phi_h(z \,|\, x) - \phi_h(z \,|\, x)\alpha_x(z)I(z \in \mathcal{X}\setminus\{x\})|\,\mathrm{d}z \\ &= 1 - \int_{\mathcal{X}} \phi_h(z \,|\, x)\alpha_x(z)\,\mathrm{d}z \\ &= 1 - \mathbb{E}_{z\sim\mathcal{P}_{x,h}}\alpha_x(z)\end{aligned}$$

for any $x \in \mathcal{X}$. By the Markov's inequality, it holds that

$$\mathbb{E}_{z\sim\mathcal{P}_{x,h}}\alpha_x(z) \geq C\mathbb{P}_{z\sim\mathcal{P}_{x,h}}\left\{\frac{\pi^*(z)\phi_h(x\,|\,z)I(z\in\mathcal{X})}{\pi^*(x)\phi_h(z\,|\,x)} \geq C\right\} \tag{A13}$$

for any $C \in (0,1]$. In the sequel, we will derive a lower bound for this tail probability.

Notice that

$$\frac{\pi^*(z)\phi_h(x\,|\,z)}{\pi^*(x)\phi_h(z\,|\,x)} = \exp\left[\frac{4h\{U(x)-U(z)\}+|z-x+h\nabla U(x)|_2^2-|x-z+h\nabla U(z)|_2^2}{4h}\right].$$

For the numerator of this exponent, we have

$$\begin{aligned}
&4h\{U(x)-U(z)\}+|z-x+h\nabla U(x)|_2^2-|x-z+h\nabla U(z)|_2^2\\
=&4h\{U(x)-U(z)\}+|z-x|_2^2+|h\nabla U(x)|_2^2+2h(z-x)^{\mathrm{T}}\nabla U(x)\\
&-|x-z|_2^2-|h\nabla U(z)|_2^2-2h(x-z)^{\mathrm{T}}\nabla U(z)\\
=&2h\{U(x)-U(z)-(x-z)^{\mathrm{T}}\nabla U(x)\}+2h\{U(x)-U(z)-(x-z)^{\mathrm{T}}\nabla U(z)\}\\
&+h^2\{|\nabla U(x)|_2^2-|\nabla U(z)|_2^2\}.
\end{aligned}$$

Since $U(\cdot)$ is L-smooth and m-strongly convex on $\mathcal{X}$, it holds that

$$U(x)-U(z)-(x-z)^{\mathrm{T}}\nabla U(x) \geq -\frac{L}{2}|x-z|_2^2,\ U(x)-U(z)-(x-z)^{\mathrm{T}}\nabla U(z) \geq \frac{m}{2}|x-z|_2^2$$

for any $x,z \in \mathcal{X}$. By the Cauchy–Schwarz's inequality, triangle inequality, and Theorem 2.1.5 of [42], we know

$$\begin{aligned}
|\nabla U(x)|_2^2-|\nabla U(z)|_2^2 &= \{\nabla U(x)+\nabla U(z)\}^{\mathrm{T}}\{\nabla U(x)-\nabla U(z)\}\\
&\geq -|\nabla U(x)+\nabla U(z)|_2|\nabla U(x)-\nabla U(z)|_2\\
&\geq -|\nabla U(x)+\nabla U(z)-\nabla U(x)+\nabla U(x)|_2 L|x-z|_2\\
&\geq -\{2|\nabla U(x)|_2+L|x-z|_2\}L|x-z|_2
\end{aligned}$$

for any $x,z \in \mathcal{X}$. Since $\mathcal{X} = \mathbb{B}(x^*,R)$ for some universal constant $R > 0$ and $x^* = \arg\min_{x\in\mathbb{R}^d} U(x)$, by Assumption 1, it holds that

$$|\nabla U(x)|_2 = |\nabla U(x)-\nabla U(x^*)|_2 \leq L|x-x^*|_2 \leq LR$$

for any $x \in \mathcal{X}$. Thus,

$$\frac{\pi^*(z)\phi_h(x\,|\,z)}{\pi^*(x)\phi_h(z\,|\,x)} \geq \exp\Bigg\{\underbrace{-\frac{L-m}{4}|x-z|_2^2-\frac{hL^2R}{2}|x-z|_2-\frac{hL^2}{4}|x-z|_2^2}_{T}\Bigg\} \tag{A14}$$

for any $x,z \in \mathcal{X}$. Since $z \sim \mathcal{P}_{x,h} = \mathcal{N}\{x-h\nabla U(x), 2hI_d\}$ and $\nabla U(x^*) = 0$, we have

$$|x-z|_2 = |h\nabla U(x)-(2h)^{1/2}\xi|_2 \leq h|\nabla U(x)|_2+(2h)^{1/2}|\xi|_2 \leq hLR+(2h)^{1/2}|\xi|_2$$

and $|x-z|_2^2 \leq 2h^2L^2R^2+4h|\xi|_2^2$ for some $\xi \sim \mathcal{N}(0,I_d)$, which implies

$$T \geq -\frac{3}{2}h^2L^3R^2-2hL|\xi|_2^2-\frac{1}{\sqrt{2}}h^{3/2}L^2R|\xi|_2$$

if $hL \leq 1$. Recall $\mathcal{X} = \mathbb{B}(x^*, R)$. Under Assumption 1, by Theorems 2.1.5, 2.1.9 and 2.1.10 of [42], it holds that

$$\begin{aligned}|x - h\nabla U(x) - x^*|_2^2 &= |x - x^*|_2^2 - 2h(x - x^*)^{\mathrm{T}}\nabla U(x) + h^2|\nabla U(x)|_2^2 \\ &\leq |x - x^*|_2^2 + (h^2 - hL^{-1})|\nabla U(x)|_2^2 \\ &\leq \{1 + (h^2 - hL^{-1})m^2\}R^2 \leq R^2\end{aligned}$$

for any $x \in \mathcal{X}$ if $hL \leq 1$. Recall $z = x - h\nabla U(x) + (2h)^{1/2}\xi$. Select $\tilde{c} \in (0,1)$ satisfying $\tilde{c}^2 = 1 + (L^{-7/2}R^{-3}d^{-2} - L^{-11/4}R^{-3/2}d^{-1})m^2$, which can be guaranteed by $L \geq m$ and $L^{3/8}R^{3/4} \geq 16d^{-1/2} + 8$. Then

$$|z - x^*|_2 \leq R\tilde{c} + (2h)^{1/2}|\xi|_2$$

for any $h \in [L^{-7/4}R^{-3/2}d^{-1}, L^{-1} - L^{-7/4}R^{-3/2}d^{-1}]$. For such selected h, we have

$$\{|\xi|_2 \leq (2h)^{-1/2}R(1 - \tilde{c})\} \subset \{z \in \mathcal{X}\}.$$

Since $L^{3/8}R^{3/4} \geq 16d^{-1/2} + 8$ and $L^{-15/8}m^2R^{1/4} \geq 12d$, by Lemma 1 of [43], for any given $u \in (1/2, 1)$, we have

$$\begin{aligned}\mathbb{P}_{z\sim\mathcal{P}_{x,h}}\left(T \geq -\frac{u}{8}, z \in \mathcal{X}\right) &\geq \mathbb{P}\left\{T \geq -\frac{u}{8}, |\xi|_2 \leq \frac{R(1-\tilde{c})}{\sqrt{2h}}\right\} \\ &\geq \mathbb{P}\left\{|\xi|_2^2 \leq \frac{R^2(1-\tilde{c})^2}{2h}\right\} - \mathbb{P}\left\{\left(\sqrt{\frac{3}{2}}hL^{3/2}R + \sqrt{2hL}|\xi|_2\right)^2 \geq \frac{u}{8}\right\} \\ &\geq \mathbb{P}\left[|\xi|_2^2 \leq 2\left\{\log^{1/2}\left(\frac{16}{u}\right) + \sqrt{d}\right\}^2\right] - \mathbb{P}\left(|\xi|_2^2 \geq \frac{u}{64hL}\right) \\ &\geq 1 - \frac{u}{8}\end{aligned}$$

for any step size h satisfying

$$\frac{1}{L^{7/4}R^{3/2}d} \leq h \leq \min\left[\frac{R^2(1-\tilde{c})^2}{4\{\log^{1/2}(16u^{-1}) + \sqrt{d}\}^2}, \frac{\sqrt{u}}{4\sqrt{3}L^{3/2}R}, \frac{u}{128L\{\log^{1/2}(16u^{-1}) + \sqrt{d}\}^2}\right].$$

Together with (A14), it holds that

$$\mathbb{P}_{z\sim\mathcal{P}_{x,h}}\left\{\frac{\pi^*(z)\phi_h(x\,|\,z)I(z \in \mathcal{X})}{\pi^*(x)\phi_h(z\,|\,x)} \geq \exp\left(-\frac{u}{8}\right)\right\} \geq 1 - \frac{u}{8}$$

for any $x \in \mathcal{X}$. Select $C = \exp(-u/8)$ in (A13). Due to $\exp(-u/8) \geq 1 - u/8$, we have

$$\mathbb{E}_{z\sim\mathcal{P}_{x,h}}\alpha_x(z) \geq \left(1 - \frac{u}{8}\right)^2 \geq 1 - \frac{u}{4},$$

which implies $\|\mathcal{P}_{x,h} - \mathcal{T}_x\|_{\mathrm{TV}} \leq u/4$ for any $x \in \mathcal{X}$. Therefore, we have the result (A12). We complete the proof of Lemma A1. □

Lemma A2. *Let $\mathcal{X} = \mathbb{B}(x^*, R)$ for some universal constant $R > 0$ and $x^* \in \mathbb{R}^d$, and Assumption 1 hold. The target distribution Π^* with density π^* defined as (3) satisfies the log-isoperimetry inequality given in (A1) with constant $\hat{c} = m^{-1/2}$, where m is specified in Assumption 1.*

Proof of Lemma A2. Let p denote the density of the Gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$, and let Π be a distribution with density $\pi = q \cdot p$, where q is a log-concave function supported on $\mathcal{X}$. From Lemma 16 in [33], it holds that

$$\Pi(S_3) \geq \frac{d(S_1, S_2)}{2\sigma} \min\{\Pi(S_1), \Pi(S_2)\} \log^{1/2}\left[1 + \frac{1}{\min\{\Pi(S_1), \Pi(S_2)\}}\right] \tag{A15}$$

for any partition S_1, S_2, S_3 of $\mathcal{X}$.

We now prove that the target distribution Π^* with density π^* defined as (3) satisfies the log-isoperimetry inequality defined as (A1). Notice that

$$\pi^*(x) = \left(\frac{2\pi}{m}\right)^{d/2} \frac{\exp\{-U(x) + m|x|_2^2/2\}}{\int_{\mathcal{X}} \exp\{-U(y)\}\, \mathrm{d}y} I(x \in \mathcal{X}) \cdot \frac{\exp(-m|x|_2^2/2)}{(2\pi/m)^{d/2}},$$

where $U(\cdot)$ is m-strongly convex on $\mathcal{X}$. By Theorem 2.1.11 of [42], we know $U(\cdot) - m|\cdot|_2^2/2$ is convex on $\mathcal{X}$. Since the indicator function $I(\cdot \in \mathcal{X})$ is log-concave on $\mathcal{X}$ and the class of log-concave functions is closed under multiplication, then π^* can be expressed as the product of a log-concave function and the density of the normal distribution $\mathcal{N}(0, m^{-1} I_d)$. By (A15), the distribution Π^* satisfies the log-isoperimetry inequality defined as (A1) with constant $\hat{c} = m^{-1/2}$. We complete the proof of Lemma A2. □

Proof of Theorem 1. Let $\mathcal{T}_x^{\mathrm{L}}$ be the one-step transition distribution of the Markov chain determined by the 1/2-lazy version of Algorithm 2, at $x \in \mathcal{X}$. Then we have

$$\mathcal{T}_x^{\mathrm{L}}(A) = \frac{1}{2}\delta_x(A) + \frac{1}{2}\mathcal{T}_x(A)$$

for any $A \in \mathscr{B}(\mathcal{X})$, where $\delta_x(\cdot)$ is the Dirac-delta function at $x \in \mathcal{X}$ and $\mathcal{T}_x$ is the one-step transition distribution of the associated Markov chain determined by Algorithm 2, at $x \in \mathcal{X}$. By the definition of lazy chain and Proposition 1, we know that the Markov chain with transition distribution $\mathcal{T}_x^{\mathrm{L}}$ is 1/2-lazy, Π^*-irreducible, smooth, and reversible with respect to the distribution Π^* with density π^* defined as (3).

Recall $\mathcal{P}_{x,h}$ is the proposal distribution involved in Algorithm 2 and the 1/2-lazy version of Algorithm 2. For any $x, y \in \mathcal{X}$ such that $|x - y|_2 \leq (2^{-1}h)^{1/2}u$ for some $u \in (1/2, 1)$ and the step size h satisfying $h \geq L^{-7/4}R^{-3/2}d^{-1}$ and

$$h \leq \min\left[\frac{R^2(1-\tilde{c})^2}{4\{\log^{1/2}(16u^{-1}) + \sqrt{d}\}^2}, \frac{\sqrt{u}}{4\sqrt{3}L^{3/2}R}, \frac{u}{128L\{\log^{1/2}(16u^{-1}) + \sqrt{d}\}^2}\right]$$

with $\tilde{c} = \{1 + (L^{-7/2}R^{-3}d^{-2} - L^{-11/4}R^{-3/2}d^{-1})m^2\}^{1/2}$, by the triangle inequality and Lemma A1, it holds that

$$\begin{aligned}
\|\mathcal{T}_x^{\mathrm{L}} - \mathcal{T}_y^{\mathrm{L}}\|_{\mathrm{TV}} &\leq \frac{1}{2} + \frac{1}{2}\|\mathcal{T}_x - \mathcal{T}_y\|_{\mathrm{TV}} \\
&\leq \frac{1}{2} + \frac{1}{2}(\|\mathcal{T}_x - \mathcal{P}_{x,h}\|_{\mathrm{TV}} + \|\mathcal{P}_{x,h} - \mathcal{P}_{y,h}\|_{\mathrm{TV}} + \|\mathcal{P}_{y,h} - \mathcal{T}_y\|_{\mathrm{TV}}) \\
&\leq \frac{1+u}{2}.
\end{aligned}$$

Recall $\mathcal{X} = \mathbb{B}(x^*, R)$ for some universal constant $R > 0$ and $x^* \in \mathbb{R}^d$. Under Assumption 1, Lemma A2 implies that the distribution Π^* with density π^* satisfies the log-isoperimetry inequality given in (A1) with constant $\hat{c} = m^{-1/2}$. Using Lemma 2 with $\delta = 2^{-1}(1-u)$ and $\Delta = (2h)^{1/2}u$, we have

$$\Omega(v) \geq \frac{1-u}{8} \min\left\{1, \frac{u\sqrt{hm}}{4\sqrt{2}} \log^{1/2}\left(1 + \frac{1}{v}\right)\right\}$$

for any $v \in (0, 1/2]$, where $\Omega(\cdot)$ is the conductance profile defined in (A3) for Markov chain with transition distribution $\mathcal{T}_x^{\mathrm{L}}$. For the above selected u and h, define the function

$$\Upsilon(v) = \begin{cases} \dfrac{1-u}{8} \min\left\{1, \dfrac{u\sqrt{hm}}{4\sqrt{2}} \log^{1/2}\left(\dfrac{1}{v}\right)\right\}, & v \in (0, 1/2] \\ \dfrac{1-u}{8} \min\left\{1, \dfrac{u\sqrt{hm}}{4\sqrt{2}} (\log 2)^{1/2}\right\}, & v \in (1/2, \infty) \end{cases}$$

for any $v > 0$. Recall that

$$\tau(\varepsilon; \mathbb{P}^0, \Pi^*) = \min\{k \in \mathbb{N} : \|\mathcal{T}^k(\mathbb{P}^0) - \Pi^*\|_{\mathrm{TV}} \leq \varepsilon\}$$

for an error tolerance $\varepsilon \in (0, 1)$, where $\mathcal{T}^k(\mathbb{P}^0)$ is the distribution of the Markov chain with transition distribution $\mathcal{T}_x^{\mathrm{L}}$ at the k-th step. Let

$$\tilde{\Omega}(v) = \begin{cases} \Omega(v), & v \in (0, 1/2], \\ \Omega(1/2), & v \in (1/2, \infty). \end{cases}$$

be the extended conductance profile of such Markov chain. By Lemma 1, it holds that

$$\tau(\varepsilon; \mathbb{P}^0, \Pi^*) \leq \left\lceil 8 \int_{4\beta^{-1}}^{\varepsilon^{-2}} \frac{\mathrm{d}v}{v\tilde{\Omega}^2(v)} \right\rceil \leq \left\lceil 8 \int_{4\beta^{-1}}^{\varepsilon^{-2}} \frac{\mathrm{d}v}{v\Upsilon^2(v)} \right\rceil.$$

If $\beta > 8$ and $h \leq 32u^{-2}\{m \log(\beta/4)\}^{-1}$, it then holds that

$$\frac{u\sqrt{hm}}{4\sqrt{2}} (\log 2)^{1/2} < \frac{u\sqrt{hm}}{4\sqrt{2}} \log^{1/2}\left(\frac{\beta}{4}\right) \leq 1,$$

which implies

$$\tau(\varepsilon; \mathbb{P}^0, \Pi^*) = O\left(\frac{1}{hm} \log \frac{\log \beta}{\varepsilon}\right).$$

Together with $h \geq L^{-7/4} R^{-3/2} d^{-1}$, we complete the proof of Theorem 1. □

Appendix B.7. Proof of Corollary 1

Proof of Corollary 1. Recall $\mathcal{X} = \{x \in \mathbb{R}^d : |x|_p \leq C\}$ for some universal constant $C > 0$. Since the additional two steps are introduced in Algorithm 3 only transforming the sampling from the norm-constrained region $\{x \in \mathbb{R}^d : |x|_p \leq C\}$ to the Euclidean ball $\mathbb{B}(0, 1)$, the convergence rate of the two processes remains consistent. Using the same arguments in the proof of Theorem 1 with $R = 1$ and $x^* = 0$, we can obtain the results of Corollary 1. □

Appendix B.8. Proof of Theorem 2

Proof of Theorem 2. Recall that the distribution $\Pi^{*,\lambda}$ with density

$$\pi^{*,\lambda}(x) = \frac{\exp\{-V_{\mathcal{X}}^{\lambda}(x)\}}{\int_{\mathbb{R}^d} \exp\{-V^{\lambda}(y)\}\, \mathrm{d}y}$$

for a regularization parameter $\lambda > 0$, where $V_{\mathcal{X}}^{\lambda}(\cdot)$ is defined as in (10), and the target distribution Π^* with density

$$\pi^*(x) = \frac{\exp\{-U(x)\} I(x \in \mathcal{X})}{\int_{\mathcal{X}} \exp\{-U(y)\}\, \mathrm{d}y}$$

for some potential function $U : \mathbb{R}^d \mapsto \mathbb{R}$. Under Assumptions 1 and 2, if there exists a universal constant $\tilde{C} > 0$ such that $\exp\{\inf_{x\in\mathcal{X}^c} U(x) - \sup_{x\in\mathcal{X}} U(x)\} \geq \tilde{C}$, by Proposition 4 in [35], we have

$$\|\Pi^{*,\lambda} - \Pi^*\|_{\mathrm{TV}} \leq \varepsilon \tag{A16}$$

for $\lambda = 8\pi^{-1}\varepsilon^2 r^2 d^{-2}\tilde{C}^2$ with the error tolerance $\varepsilon \in (0,1)$, where $r > 0$ is specified in Assumption 2.

Notice that $V_{\mathcal{X}}^{\lambda}(\cdot) = U(\cdot) + \iota_{\mathcal{X}}^{\lambda}(\cdot)$ with $\iota_{\mathcal{X}}^{\lambda}(\cdot)$ defined as (7). Under Assumption 1, by (9) and Theorem 2.1.5 in [42], we know that the function $V_{\mathcal{X}}^{\lambda}(\cdot)$ is twice continuously differentiable, $(L+\lambda^{-1})$-smooth and m-strongly convex on $\mathbb{R}^d$. Given the initial distribution $\mathbb{P}^0 = \mathcal{N}\{x^\star, (L+\lambda^{-1})^{-1} I_d\}$ with $x^\star = \arg\min_{x\in\mathbb{R}^d} V_{\mathcal{X}}^{\lambda}(x)$ and an error tolerance $\varepsilon \in (0,1)$, by Theorem 5 of [33], the Markov chain determined by Algorithm 4 satisfies

$$\tau(\varepsilon; \mathbb{P}^0, \Pi^{*,\lambda}) = O\left[\frac{(L+\lambda^{-1})d}{m} \log\frac{d}{\varepsilon} \cdot \max\left\{1, \sqrt{\frac{L+\lambda^{-1}}{dm}}\right\}\right]$$

with the step size

$$h = c\frac{1}{(L+\lambda^{-1})d \cdot \max\left\{1, \sqrt{\frac{L+\lambda^{-1}}{dm}}\right\}},$$

where $c > 0$ is a universal constant. Together with (A16), by the definition of ε-mixing time and the triangle inequality, we have

$$\tau(\varepsilon; \mathbb{P}^0, \Pi^{*}) = O\left[\frac{(L+\lambda^{-1})d}{m} \log\frac{d}{\varepsilon} \max\left\{1, \sqrt{\frac{L+\lambda^{-1}}{dm}}\right\}\right]$$

with $\lambda = 8\pi^{-1}\varepsilon^2 r^2 d^{-2}\tilde{C}^2$. Hence, we complete the proof of Theorem 2. □

References

1. Gelfand, A.E.; Smith, A.F.; Lee, T.M. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Am. Stat. Assoc.* **1992**, *87*, 523–532. [CrossRef]
2. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
3. Klein, J.P.; Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated Data*; Springer: New York, NY, USA, 2005; pp. 5–18.
4. Johnson, V.E.; Albert, J.H. *Ordinal Data Modeling*; Springer: New York, NY, USA, 2006; pp. 126–157.
5. Celeux, G.; El Anbari, M.; Marin, J.M.; Robert, C.P. Regularization in regression: Comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Anal.* **2012**, *7*, 477–502. [CrossRef]
6. Paisley, J.W.; Blei, D.M.; Jordan, M.I. Bayesian nonnegative matrix factorization with stochastic variational inference. In *Handbook of Mixed Membership Models and Their Applications*; Airoldi, E.M., Blei, D.M., Erosheva, E.A., Fienberg, S.E., Eds.; CRC Press: Boca Raton, FL, USA, 2014; pp. 205–224.
7. Khodadadian, A.; Parvizi, M.; Teshnehlab, M.; Heitzinger, C. Rational design of field-effect sensors using partial differential equations, Bayesian inversion, and artificial neural networks. *Sensors* **2022**, *22*, 4785. [CrossRef] [PubMed]
8. Noii, N.; Khodadadian, A.; Ulloa, J.; Aldakheel, F.;Wick, T.; François, S.;Wriggers, P. Bayesian inversion with open-source codes for various one-dimensional model problems in computational mechanics. *Arch. Comput. Methods Eng.* **2022**, *29*, 4285–4318. [CrossRef]
9. Ma, Y.A.; Chen Y.; Jin C.; Flammarion N.; Jordan M.I. Sampling can be faster than optimization. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 20881–20885. [CrossRef] [PubMed]
10. Mangoubi, O.; Vishnoi, N.K. Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. In Proceedings of the 32nd Conference on Learning Theory, Phoenix, AZ, USA, 25–28 June 2019; pp. 2259–2293.
11. Dyer, M.; Frieze, A. Computing the volume of convex bodies: A case where randomness provably helps. *Probabilistic Comb. Its Appl.* **1991**, *44*, 123–170.
12. Rodriguez-Yam, G.; Davis, R.A.; Scharf, L.L. Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression. In *Technical Report*; Unpublished Manuscript; Colorado State University: Fort Collins, CO, USA, 2004.
13. Lovász, L.; Vempala, S. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms* **2007**, *30*, 307–358. [CrossRef]

14. Chen, M.H.; Shao, Q.M.; Ibrahim, J.G. *Monte Carlo Methods in Bayesian Computation*; Springer: New York, NY, USA, 2012; pp. 191–212.
15. Dyer, M.; Frieze, A.; Kannan, R. A random polynomial-time algorithm for approximating the volume of convex bodies. *J. ACM* **1991**, *38*, 1–17. [CrossRef]
16. Lang, L.; Chen, W.S.; Bakshi, B.R.; Goel, P.K.; Ungarala, S. Bayesian estimation via sequential Monte Carlo sampling—Constrained dynamic systems. *Automatica* **2007**, *43*, 1615–1622. [CrossRef]
17. Chaudhry, S.; Lautzenheiser, D.; Ghosh, K. An efficient scheme for sampling in constrained domains. *arXiv* **2021**, arXiv:2110.10840.
18. Lan, S.; Kang, L. Sampling constrained Continuous probability distributions: A review. *arXiv* **2021**, arXiv:2209.12403.
19. Neal, R.M. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*; Brooks, S., Gelman, A., Jones, G., Meng, X.L., Eds.; CRC Press: Boca Raton, FL, USA, 2011; pp. 113–162.
20. Pakman, A.; Paninski, L. Exact hamiltonian Monte Carlo for truncated multivariate gaussians. *J. Comput. Graph. Stat.* **2014**, *23*, 518–542. [CrossRef]
21. Lan, S.; Shahbaba, B. Sampling constrained probability distributions using spherical augmentation. In *Algorithmic Advances in Riemannian Geometry and Applications*; Minh, H.Q., Murino, V., Eds.; Springer: New York, NY, USA, 2016; pp. 25–71.
22. Brubaker, M.; Salzmann, M.; Urtasun, R. A family of MCMC methods on implicitly defined manifolds. In Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, La Palma, Canary Islands, Spain, 21–23 April 2012; pp. 161–172.
23. Ahn, K.; Chewi, S. Efficient constrained sampling via the mirror-Langevin algorithm. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; pp. 28405–28418.
24. Parisi, G. Correlation functions and computer simulations. *Nucl. Phys. B* **1981**, *180*, 378–384. [CrossRef]
25. Grenander, U.; Miller, M.I. Representations of knowledge in complex systems. *J. R. Stat. Soc. Ser. B (Methodol.)* **1994**, *56*, 549–581. [CrossRef]
26. Roberts, G.O.; Tweedie, R.L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **1996**, *2*, 341–363. [CrossRef]
27. Roberts, G.O.; Stramer, O. Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.* **2002**, *4*, 337–357. [CrossRef]
28. Dalalyan, A.S. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B (Methodol.)* **2017**, *79*, 651–676. [CrossRef]
29. Durmus, A.; Moulines, E. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Bernoulli* **2017**, *27*, 1551–1587. [CrossRef]
30. Cheng, X.; Bartlett, P. Convergence of Langevin MCMC in KL-divergence. In Proceedings of the Machine Learning Research, Lanzarote, Spain, 7–9 April 2018; pp. 186–211.
31. Durmus, A.; Moulines, E. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **2019**, *25*, 2854–2882. [CrossRef]
32. Dwivedi, R.; Chen, Y.; Wainwright, M.J.; Yu, B. Log-concave sampling: Metropolis-Hastings algorithms are fast. *J. Mach. Learn. Res.* **2019**, *20*, 1–42.
33. Chen, Y.; Dwivedi, R.; Wainwright, M.J.; Yu, B. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *J. Mach. Learn. Res.* **2020**, *21*, 3647–3717.
34. Bubeck, S.; Eldan, R.; Lehec, J. Finite-time analysis of projected Langevin Monte Carlo. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 1243–1251.
35. Brosse, N.; Durmus, A.; Moulines, É.; Pereyra, M. Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In Proceedings of the 2017 Conference on Learning Theory, Amsterdam, The Netherlands, 7–10 July 2017; pp. 319–342.
36. Hsieh, Y.P.; Kavis, A.; Rolland, P.; Cevher, V. Mirrored langevin dynamics. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 1–10.
37. Roberts, G.O.; Rosenthal, J.S. General state space Markov chains and MCMC algorithms. *Probab. Surv.* **2004**, *1*, 20–71. [CrossRef]
38. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
39. Kannan, R.; Lovász, L.; Montenegro, R. Blocking conductance and mixing in random walks. *Comb. Probab. Comput.* **2006**, *15*, 541–570. [CrossRef]
40. Lee, Y.T.; Vempala, S.S. Stochastic localization + Stieltjes barrier = tight bound for log-Sobolev. In Proceedings of the Annual ACM SIGACT Symposium on Theory of Computing, Los Angeles, CA, USA, 25–29 June 2018; pp. 1122–1129.
41. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499. [CrossRef]
42. Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*; Springer: New York, NY, USA, 2003; pp. 51–101.
43. Laurent, B.; Massart, P. Adaptive estimation of a quadratic functional by model selection. *Ann. Stat.* **2000**, *28*, 1302–1338. [CrossRef]