Large deviation principle for Metropolis-Hastings based Markov chains

Federica Milinanni

Algorithms & Computationally Intensive Inference seminars

Warwick, May 7, 2024

Joint work with
Pierre Nyquist, Chalmers & Gothenburg University

Outline

Large deviation principle for Metropolis-Hastings chains

LDP on non-compact state spaces

Alternative representation of the LDP rate function

Algorithm tuning

The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm

Algorithm used to generate a Markov chain of **samples** $\{\theta_0, \theta_1, \theta_2, \dots\}$ to approximate a **target** probability measure π

The Metropolis-Hastings algorithm

Algorithm used to generate a Markov chain of **samples** $\{\theta_0, \theta_1, \theta_2, \dots\}$ to approximate a **target** probability measure π

Algorithm

1. generate a proposal $\Theta^* \sim J(\cdot|\theta_k)$ from a **proposal distribution** $J(\cdot|\theta_k)$

The Metropolis-Hastings algorithm

Algorithm used to generate a Markov chain of **samples** $\{\theta_0, \theta_1, \theta_2, \dots\}$ to approximate a **target** probability measure π

Algorithm

1. generate a proposal $\Theta^* \sim J(\cdot|\theta_k)$ from a **proposal distribution** $J(\cdot|\theta_k)$
2. set

$$\theta_{k+1} = \begin{cases} \theta^* & \text{with probability} \quad \alpha(\theta_k, \theta^*) \\ \theta_k & \text{with probability} \quad 1 - \alpha(\theta_k, \theta^*) \end{cases}$$

   with Metropolis-Hasting **acceptance probability**

$$\alpha(\theta_k, \theta^*) := \min\left\{1, \frac{\pi(\theta^*)J(\theta_k|\theta^*)}{\pi(\theta_k)J(\theta^*|\theta_k)}\right\}$$

The Metropolis-Hastings algorithm

- $\{\theta_0, \theta_1, \theta_2, \dots\}$ MH samples
- The **empirical measure** of the MH Markov chain converges to the **target** π w.p.1.

$$L^n(dx) := \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\theta_i}(dx) \xrightarrow[n\to\infty]{} \pi(dx), \quad \text{w.p.1.}$$

Tools for convergence analysis

- Spectral gap
- Asymptotic variance
- Mixing times
- Poincaré inequalities
- Logarithmic Sobolev inequalities

The Metropolis-Hastings algorithm

Tools for convergence analysis

- Spectral gap
- Asymptotic variance
- Mixing times
- Poincaré inequalities
- Logarithmic Sobolev inequalities

Our contribution: **Large deviation principle** for the MH algorithm as a **complementary** convergence analysis method

Tools for convergence analysis

- Spectral gap
- Asymptotic variance
- Mixing times
- Poincaré inequalities
- Logarithmic Sobolev inequalities

Our contribution: **Large deviation principle** for the MH algorithm as a **complementary** convergence analysis method

Advantage: Study the convergence of the **empirical measure** generated by the MH samples

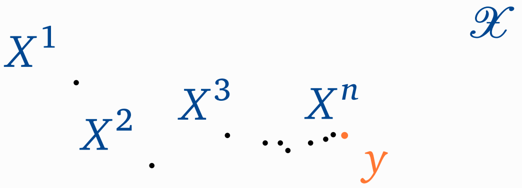
Large deviation principle

Large deviation principle

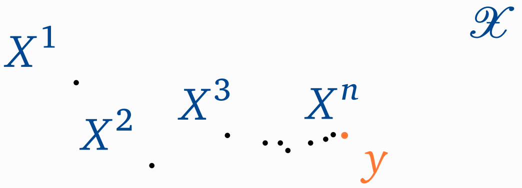
Assume $X^n \xrightarrow[n\to\infty]{} x \neq y$ in probability.

Large deviation principle

Assume $X^n \xrightarrow[n\to\infty]{} x \neq y$ in probability.

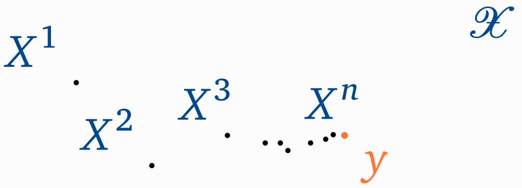


Assume $X^n \xrightarrow[n\to\infty]{} x \neq y$ in probability.



- $\mathbb{P}(X^n \approx y) \approx 0$ (Rare event)

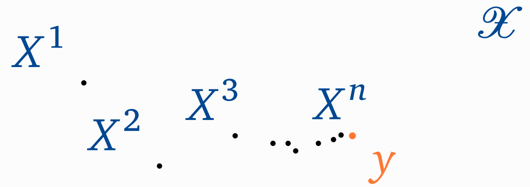
Assume $X^n \xrightarrow[n\to\infty]{} x \neq y$ in probability.



- $\mathbb{P}(X^n \approx y) \approx 0$ (Rare event)
- $\mathbb{P}(X^n \approx y) \xrightarrow[n\to\infty]{} 0$ exponentially fast

Large deviation principle

Assume $X^n \xrightarrow[n\to\infty]{} x \neq y$ in probability.



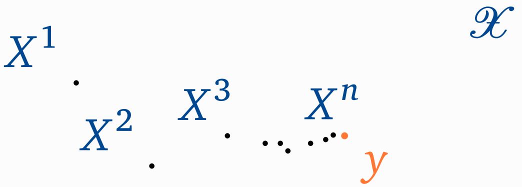
- $\mathbb{P}(X^n \approx y) \approx 0$ (Rare event)
- $\mathbb{P}(X^n \approx y) \xrightarrow[n\to\infty]{} 0$ exponentially fast
- Large deviation theory studies the exponential **decay rate**

$$\mathbb{P}(X^n \approx y) \approx e^{-n \cdot I(y)}$$

rate function

Large deviation principle

Assume $X^n \xrightarrow[n\to\infty]{} x \neq y$ in probability.



- $\mathbb{P}(X^n \approx y) \approx 0$ (Rare event)
- $\mathbb{P}(X^n \approx y) \xrightarrow[n\to\infty]{} 0$ exponentially fast
- Large deviation theory studies the exponential **decay rate**

$$\mathbb{P}(X^n \approx y) \approx e^{-n \cdot I(y)}$$

rate function

Higher rate function $\Rightarrow$ faster convergence

Large deviation principle

$\{X^n\}$ satisfies a **large deviation principle** on $\mathscr{X}$ with **speed** n and **rate function** I if

- $I : \mathscr{X} \to [0, \infty]$ has compact level sets
- $\forall$ measurable $A \subset \mathscr{X}$,

$$\begin{aligned} -\inf_{x \in A^\circ} I(x) &\leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X^n \in A^\circ) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X^n \in \bar{A}) \leq -\inf_{x \in \bar{A}} I(x) \end{aligned}$$

Idea:

$$\mathbb{P}(X^n \in A) \approx e^{-n \inf_{x \in A} I(x)}$$

LDP for sampling algorithms

LDP for sampling algorithms

- Target probability measure $\pi \in \mathscr{P}(S)$, (e.g. $S \subseteq \mathbb{R}^d$)

LDP for sampling algorithms

- Target probability measure $\pi \in \mathscr{P}(S)$, (e.g. $S \subseteq \mathbb{R}^d$)
- Algorithm samples $\{\theta^n\} \subset S$

LDP for sampling algorithms

- Target probability measure $\pi \in \mathscr{P}(S)$, (e.g. $S \subseteq \mathbb{R}^d$)
- Algorithm samples $\{\theta^n\} \subset S$
- Empirical measure $\{L^n\} \subset \mathscr{P}(S)$

$$L^n(dx) := \frac{1}{n} \sum_{i=0}^{n-1} \delta_{\theta_i}(dx)$$

LDP for sampling algorithms

- Target probability measure $\pi \in \mathscr{P}(S)$, (e.g. $S \subseteq \mathbb{R}^d$)
- Algorithm samples $\{\theta^n\} \subset S$
- Empirical measure $\{L^n\} \subset \mathscr{P}(S)$

$$L^n(dx) := \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\theta_i}(dx) \xrightarrow[n\to\infty]{} \pi(dx), \quad \text{w.p.1.}$$

LDP for sampling algorithms

- Target probability measure $\pi \in \mathscr{P}(S)$, (e.g. $S \subseteq \mathbb{R}^d$)
- Algorithm samples $\{\theta^n\} \subset S$
- Empirical measure $\{L^n\} \subset \mathscr{P}(S)$

$$L^n(dx) := \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\theta_i}(dx) \xRightarrow[n\to\infty]{} \pi(dx), \quad \text{w.p.1.}$$

$$\mathbb{P}(L^n \approx \pi) \approx 1$$

LDP for sampling algorithms

- Target probability measure $\pi \in \mathscr{P}(S)$, (e.g. $S \subseteq \mathbb{R}^d$)
- Algorithm samples $\{\theta^n\} \subset S$
- Empirical measure $\{L^n\} \subset \mathscr{P}(S)$

$$L^n(dx) := \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\theta_i}(dx) \xRightarrow[n\to\infty]{} \pi(dx), \quad \text{w.p.1.}$$

$$\mathbb{P}(L^n \approx \pi) \approx 1$$

$$\mathbb{P}(L^n \approx \mu) \approx 0, \quad \mu \neq \pi$$

LDP for sampling algorithms

- Target probability measure $\pi \in \mathscr{P}(S)$, (e.g. $S \subseteq \mathbb{R}^d$)
- Algorithm samples $\{\theta^n\} \subset S$
- Empirical measure $\{L^n\} \subset \mathscr{P}(S)$

$$L^n(dx) := \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\theta_i}(dx) \xRightarrow[n\to\infty]{} \pi(dx), \quad \text{w.p.1.}$$

$$\mathbb{P}(L^n \approx \pi) \approx 1$$

$$\mathbb{P}(L^n \approx \mu) \approx 0, \quad \mu \neq \pi$$

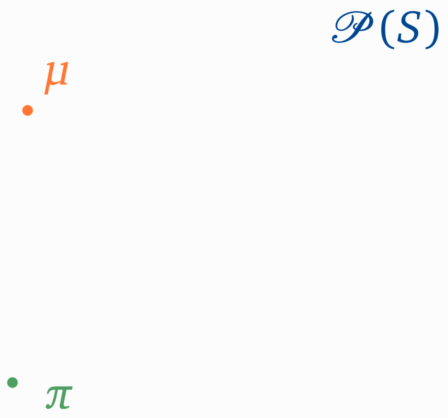
- $\{L^n\}$ satisfies a **large deviation principle** on $\mathscr{P}(S)$ with rate function I if

$$\mathbb{P}(L^n \approx \mu) \approx e^{-n\cdot I(\mu)}$$

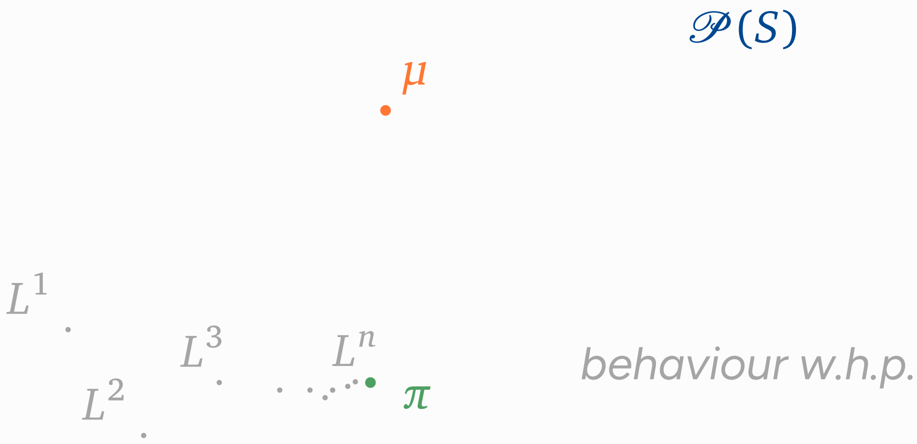
LDP for sampling algorithms

$\mathcal{P}(S)$

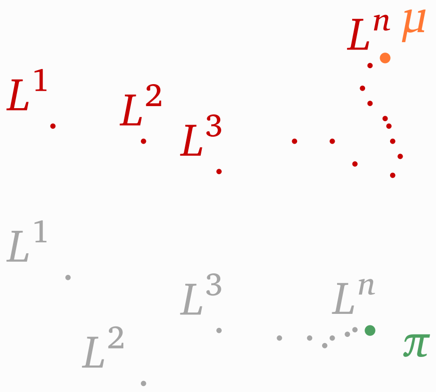
π


$\mathscr{P}(S)$
μ
π

LDP for sampling algorithms



LDP for sampling algorithms



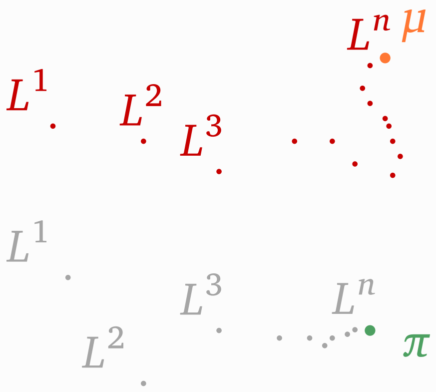
$\mathscr{P}(S)$

rare event

behaviour w.h.p.

LDP for sampling algorithms

$\mathscr{P}(S)$



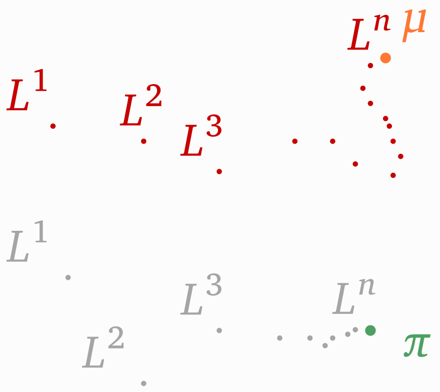
rare event

behaviour w.h.p.

$$\mathbb{P}(L^n \approx \mu) \approx e^{-n \cdot I(\mu)}$$

LDP for sampling algorithms

$\mathscr{P}(S)$



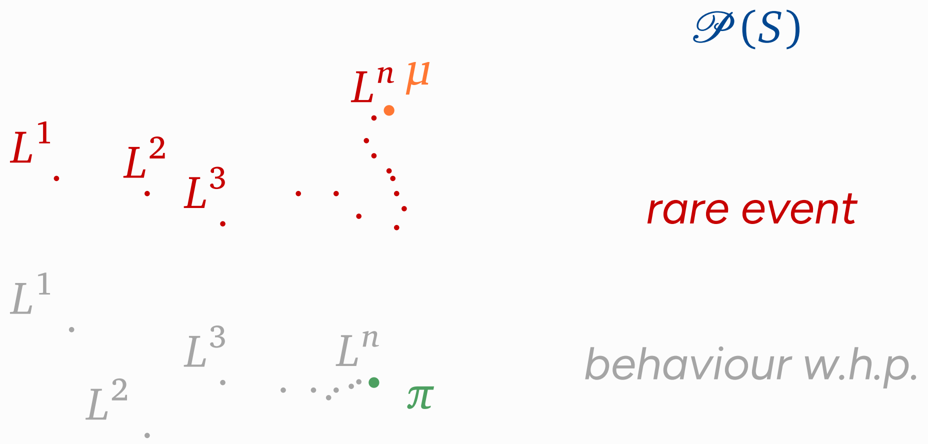
rare event

behaviour w.h.p.

$$\mathbb{P}(L^n \approx \mu) \approx e^{-n \cdot I(\mu)}$$

Note: $I(\pi) = 0$

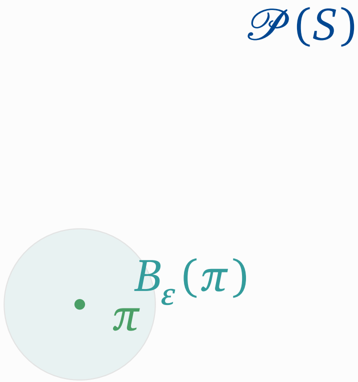
LDP for sampling algorithms



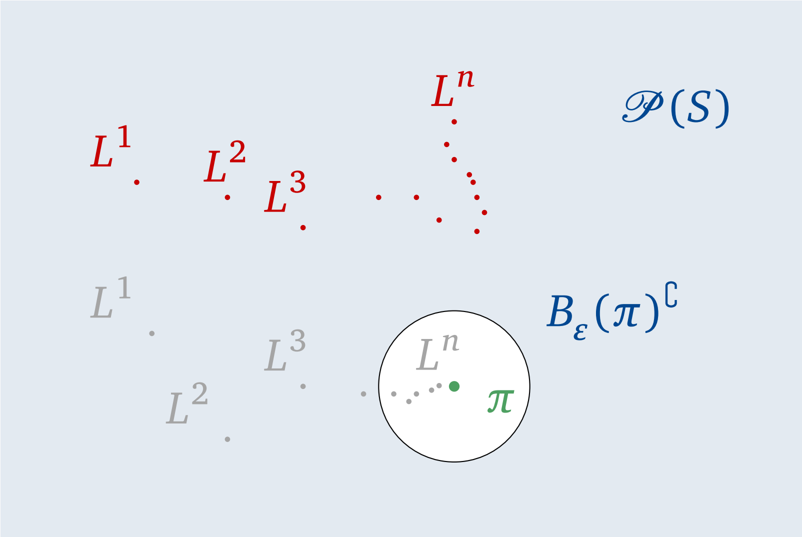
$$\mathbb{P}(L^n \approx \mu) \approx e^{-n \cdot I(\mu)}$$

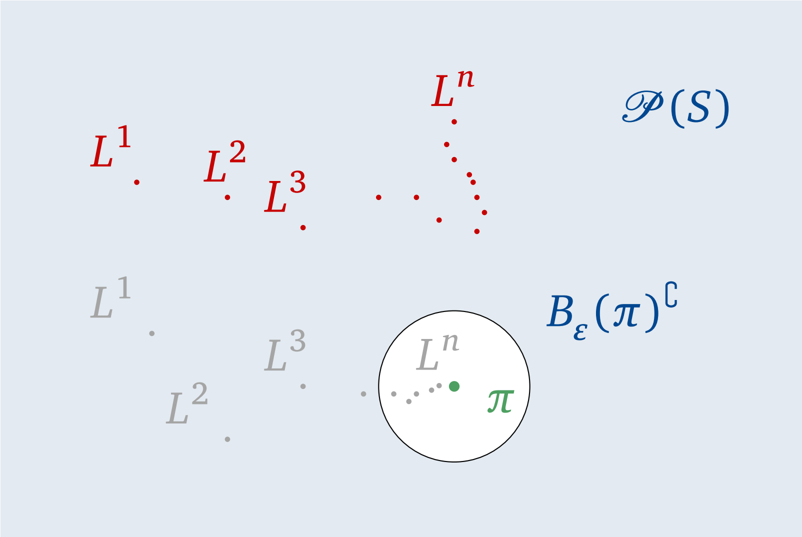
Note: $\quad I(\pi) = 0 \quad \Rightarrow \quad \mathbb{P}(L^n \approx \pi) \approx 1$

LDP for sampling algorithms



LDP for sampling algorithms





$$\mathbb{P}\left(L^n \in B_\varepsilon(\pi)^\complement\right) \approx \exp\left\{-n \cdot \inf_{\mu \in B_\varepsilon(\pi)^\complement} I(\mu)\right\}$$

Higher $\inf_{\mu \in B_\varepsilon(\pi)^\complement} I(\mu)$ $\Rightarrow$ faster convergence to π

Large Deviation Principle for Observables

If $\{L^n\}$ satisfies a LDP with rate function I,

$$\mathbb{P}(L^n \approx \mu) \approx e^{-n \cdot I(\mu)}$$

Large Deviation Principle for Observables

If $\{L^n\}$ satisfies a LDP with rate function I,

$$\mathbb{P}(L^n \approx \mu) \approx e^{-n \cdot I(\mu)}$$

Given an observable $f \in C(S)$, by the **contraction principle** we can define a rate function $\tilde{I}_f$ for which

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \approx m\right) \approx e^{-n \cdot \tilde{I}_f(m)}$$

Large Deviation Principle for Observables

If $\{L^n\}$ satisfies a LDP with rate function I,

$$\mathbb{P}(L^n \approx \mu) \approx e^{-n \cdot I(\mu)}$$

Given an observable $f \in C(S)$, by the **contraction principle** we can define a rate function $\tilde{I}_f$ for which

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \approx m\right) \approx e^{-n \cdot \tilde{I}_f(m)}$$

More formally,

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \in (m-\varepsilon, m+\varepsilon)\right) = \tilde{I}_f(m)$$

LDP for sampling algorithms

There has been growing interest in exploring large deviations as a tool to analyze the speed of convergence of sampling algorithms:

- Parallel Tempering

  [Dupuis P., Liu Y., Plattner N., and Doll J. D., 2012]

  [Doll J., Dupuis P. and Nyquist P., 2016]

  [Dupuis P. and Wu G.-J., 2022]

- Irreversible Langevine Samplers

  [Rey-Bellet L., Spiliopoulos K., 2015, 2016]

- Zig-Zag Sampler

  [Bierkens J., Nyquist P., Schlottke M. C., 2021]

LDP for sampling algorithms

A result of LDP for the **Metropolis-Hastings** algorithm is already available

[Bierkens J., 2016]

LDP for sampling algorithms

A result of LDP for the **Metropolis-Hastings** algorithm is already available

[Bierkens J., 2016]

It considers

- a **non-reversible** version of the algorithm
- a **countable** space S

LDP for sampling algorithms

A result of LDP for the **Metropolis-Hastings** algorithm is already available

[Bierkens J., 2016]

It considers

- a **non-reversible** version of the algorithm
- a **countable** space S

We stated a **large deviation principle** for the **Metropolis-Hastings** algorithm on $S \subseteq \mathbb{R}^d$

LDP for Metropolis-Hastings chains

Assumptions

(A.1) Target probability measure $\pi \ll$ Lebesgue measure with continuous density

(A.2) Proposal distribution $J(\cdot \mid x) \ll$ Lebesgue measure with continuous and bounded density

(A.3) There exists a Lyapunov function $U : S \to [0, \infty)$ satisfying certain properties...

LDP for Metropolis-Hastings chains

Assumptions

(A.1) Target probability measure $\pi \ll$ Lebesgue measure with continuous density

(A.2) Proposal distribution $J(\cdot|x) \ll$ Lebesgue measure with continuous and bounded density

(A.3) There exists a Lyapunov function $U : S \to [0, \infty)$ satisfying certain properties...

Metropolis-Hastings transition kernel

$$K(x, dy) = \min\left\{1, \frac{\pi(y)J(x|y)}{\pi(x)J(y|x)}\right\} J(y|x)dy + r(x)\delta_x(dy)$$

where $r(x) = 1 - \int_S \min\left\{1, \frac{\pi(y)J(x|y)}{\pi(x)J(y|x)}\right\} J(y|x)dy$.

Marginal distributions

Given $\gamma \in \mathscr{P}(S \times S)$,
$[\gamma]_1,\ [\gamma]_2 \in \mathscr{P}(S)$ denote the first and second marginal of γ:

$$[\gamma]_1(A) = \gamma(A, S) \qquad [\gamma]_2(A) = \gamma(S, A)$$

with $A \in \mathscr{B}(S)$

Marginal distributions

Given $\gamma \in \mathscr{P}(S \times S)$,
$[\gamma]_1,\ [\gamma]_2 \in \mathscr{P}(S)$ denote the first and second marginal of γ:

$$[\gamma]_1(A) = \gamma(A, S) \qquad [\gamma]_2(A) = \gamma(S, A)$$

with $A \in \mathscr{B}(S)$

Relative Entropy (= Kullback-Leibler Divergence)

$\mu, \nu \in \mathscr{P}(S)$,

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

LDP for Metropolis-Hastings chains

Theorem (M., Nyquist 2024a)

The sequence of MH empirical measures $\{L^n\}$ satisfies a large deviation principle with speed n and rate function $I : \mathscr{P}(S) \to [0, +\infty]$

$$\mu \mapsto I(\mu) = \inf_{\substack{\gamma \in \mathscr{P}(S \times S) \\ [\gamma]_1 = [\gamma]_2 = \mu}} R(\gamma \parallel \mu \otimes K)$$

Theorem (M., Nyquist 2024a)

The sequence of MH empirical measures $\{L^n\}$ satisfies a large deviation principle with speed n and rate function $I : \mathscr{P}(S) \to [0, +\infty]$

$$\mu \mapsto I(\mu) = \inf_{\substack{\gamma \in \mathscr{P}(S \times S) \\ [\gamma]_1 = [\gamma]_2 = \mu}} R(\gamma \parallel \mu \otimes K)$$

$$= \inf_{\substack{\gamma \in \mathscr{P}(S \times S) \\ [\gamma]_1 = [\gamma]_2 = \mu}} R(\gamma(dx, dy) \parallel \mu(dx) K(x, dy))$$

LDP for Metropolis-Hastings chains

Theorem (M., Nyquist 2024a)

The sequence of MH empirical measures $\{L^n\}$ satisfies a large deviation principle with speed n and rate function $I : \mathscr{P}(S) \to [0, +\infty]$

$$\mu \mapsto I(\mu) = \inf_{\substack{\gamma \in \mathscr{P}(S \times S) \\ [\gamma]_1 = [\gamma]_2 = \mu}} R(\gamma \parallel \mu \otimes K)$$

$$= \inf_{\substack{\gamma \in \mathscr{P}(S \times S) \\ [\gamma]_1 = [\gamma]_2 = \mu}} R(\gamma(dx, dy) \parallel \mu(dx) K(x, dy))$$

$$= \inf_{\substack{q(x, dy) \\ \mu \text{ invariant for } q}} R(\mu(dx) q(x, dy) \parallel \mu(dx) K(x, dy)).$$

LDP on non-compact state spaces

LDP on non-compact state spaces

To guarantee the LDP for MH on a non-compact state space we assume the existence of a **Lyapunov function** $U : S \to [0, \infty)$ with the following properties:

LDP on non-compact state spaces

To guarantee the LDP for MH on a non-compact state space we assume the existence of a **Lyapunov function** $U : S \to [0, \infty)$ with the following properties: let

$$F_U(x) = U(x) - \log \int_S e^{U(y)} K(x, dy);$$

(a) $\inf_{x \in S} F_U(x) > -\infty$,

(b) $F_U(x)$ has relatively compact sublevel sets,

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

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

LDP on non-compact state spaces

When S is compact, $U \equiv 0$ satisfies (a)-(c).

When S is non-compact, verifying the existence of U is not immediate.

We study the existence of U in some instances of

- Independent Metropolis-Hastings (IMH)
- Metropolis-adjusted Langevin algorithm (MALA)
- Random Walk Metropolis (RWM)

Independent Metropolis-Hastings

- target density $\pi(x) \propto e^{-\eta|x|^\alpha}$
- independent proposal $f(y) \propto e^{-\gamma|y|^\beta}$

A Lyapunov function $U : S \to [0, \infty)$ satisfying properties (a)-(c) exists if and only if

1. $\alpha = \beta$ and $\eta > \gamma$,
2. or $\alpha > \beta$.

LDP on non-compact state spaces

Independent Metropolis-Hastings

- target density $\pi(x) \propto e^{-\eta|x|^{\alpha}}$
- independent proposal $f(y) \propto e^{-\gamma|y|^{\beta}}$

A Lyapunov function $U : S \to [0, \infty)$ satisfying properties (a)-(c) exists if and only if

1. $\alpha = \beta$ and $\eta > \gamma$,
2. or $\alpha > \beta$.

Remark

If 1. or 2. is satisfied $\Rightarrow$ the MH Markov chain is **uniformly ergodic**
If neither 1. or 2. is satisfied $\Rightarrow$ the MH Markov chain is **not** even **geometrically ergodic**

[Mengersen K. L., Tweedie R. L., 1996]

LDP on non-compact state spaces

Theorem (M., Nyquist 24b)

Consider the target density $\pi(x) \propto e^{-\eta|x|^{\alpha}}$ *and the independent proposal density* $f(y) \propto e^{-\gamma|y|^{\beta}}$ *in the* ***Independent Metropolis-Hastings*** *algorithm. Suppose that either of the following holds:*

i) $\alpha = \beta$ and $\eta > \gamma$,

ii) $\alpha > \beta$.

Then, the empirical measures of the associated Metropolis-Hastings chain satisfies an LDP with speed n *and rate function* I.

Metropolis-adjusted Langevin algorithm

- target density $\pi(x) \propto e^{-\gamma|x|^\beta}$
- MALA proposal density $J(y|x) \propto \exp\left\{-\frac{1}{2\varepsilon}\left|y - x + \frac{\varepsilon\gamma\beta}{2}|x|^{\beta-2}x\right|^2\right\}$

A Lyapunov function $U : S \to [0, \infty)$ satisfying properties (a)-(c) exists if and only if $\beta = 2$ and $\varepsilon\gamma < 2$, or $1 < \beta < 2$.

Metropolis-adjusted Langevin algorithm

- target density $\pi(x) \propto e^{-\gamma|x|^{\beta}}$
- MALA proposal density $J(y|x) \propto \exp\left\{-\frac{1}{2\varepsilon}\left|y - x + \frac{\varepsilon\gamma\beta}{2}|x|^{\beta-2}x\right|^2\right\}$

A Lyapunov function $U : S \to [0, \infty)$ satisfying properties (a)-(c) exists if and only if $\beta = 2$ and $\varepsilon\gamma < 2$, or $1 < \beta < 2$.

Remark

In $d = 1$,

- if $\beta = 2$ and $\varepsilon\gamma < 2$, or $1 < \beta < 2 \Rightarrow$ the MALA chain is **geometrically ergodic**
- if $0 < \beta < 1, \beta > 2$,
  or $\beta = 2$ and $\varepsilon\gamma \geq 2 \Rightarrow$ the MALA chain is **not geometrically ergodic**
- if $\beta = 1 \Rightarrow$ the MALA chain is **geometrically ergodic** on the positive real line

[Roberts G. O., Tweedie R. L., 1996]

LDP on non-compact state spaces

Theorem (M., Nyquist 24b)

Consider a target density $\pi(x) \propto e^{-\gamma|x|^\beta}$ *and let* $J(y|x)$ *be the corresponding* ***MALA*** *proposal density with discretization step* ε*,*

$$J(y|x) \propto \exp\left\{-\frac{1}{2\varepsilon}\left|y - x + \frac{\varepsilon\gamma\beta}{2}|x|^{\beta-2}x\right|^2\right\}.$$

Suppose that either of the following holds:

i) $\beta = 2$ and $\varepsilon\gamma < 2$,

ii) $1 < \beta < 2$.

Then, the empirical measures of the associated Metropolis-Hastings chain satisfy an LDP with speed n *and rate function* I*.*

Random Walk Metropolis

- random walk proposal $J(y\,|\,x) = \hat{J}(y - x) = \hat{J}(x - y)$

There exists no function $U : S \to [0, \infty)$ that satisfies the properties (a)-(c) on the Lyapunov function.

LDP on non-compact state spaces

Random Walk Metropolis

- random walk proposal $J(y\,|\,x) = \hat{J}(y-x) = \hat{J}(x-y)$

There exists no function $U : S \to [0, \infty)$ that satisfies the properties (a)-(c) on the Lyapunov function.

Remark

The RWM Markov chain is **not uniformly ergodic** for any target π.
However, with additional assumptions on the tail decays, the RWM Markov chain is **gemetrically ergodic**.

[Mengersen K. L., Tweedie R. L., 1996]

LDP on non-compact state spaces

		Assumption (A.3)	geometric ergodicity
IMH	$\alpha = \beta$ and $\eta > \gamma$, or $\alpha > \beta$	✓	✓
	otherwise	✗	✗
MALA	$\beta = 2$ and $\varepsilon\gamma < 2$, or $1 < \beta < 2$	✓	✓
	$\beta = 1$	✗	✓
	otherwise	✗	✗
RWM	tail decays as in [MT96]	✗	✓
	otherwise	✗	✗

LDP on non-compact state spaces

		Assumption (A.3)	geometric ergodicity
IMH	$\alpha = \beta$ and $\eta > \gamma$, or $\alpha > \beta$	✓	✓
	otherwise	✗	✗
MALA	$\beta = 2$ and $\varepsilon\gamma < 2$, or $1 < \beta < 2$	✓	✓
	$\beta = 1$	✗	✓
	otherwise	✗	✗
RWM	tail decays as in [MT96]	✗	✓
	otherwise	✗	✗

We hypothesise that Assumption (A.3) is too strict and we pose the following

Open problem

Assume that the MH chain $\{X_i\}$ is **geometrically ergodic**.
Does the corresponding empirical measure $\{L^n\}$ satisfy an **LDP**?

Alternative representation of the LDP rate function

Rate function decomposition

- By the Lebesgue decomposition theorem, for any $\mu \in \mathscr{P}(S)$,

$$\mu = (1-p) \cdot \mu_\lambda + p \cdot \mu_s,$$

where $p \in [0,1]$, $\mu_\lambda, \mu_s \in \mathscr{P}(S)$, with $\mu_\lambda \ll \lambda$ and $\mu_s \perp \lambda$.

Rate function decomposition

- By the Lebesgue decomposition theorem, for any $\mu \in \mathscr{P}(S)$,

$$\mu = (1-p)\cdot\mu_\lambda + p\cdot\mu_s,$$

  where $p \in [0,1]$, $\mu_\lambda, \mu_s \in \mathscr{P}(S)$, with $\mu_\lambda \ll \lambda$ and $\mu_s \perp \lambda$.
- The rate function satisfies

$$I(\mu) = (1-p)\cdot I(\mu_\lambda) + p\cdot I(\mu_s)$$

Rate function decomposition

- By the Lebesgue decomposition theorem, for any $\mu \in \mathscr{P}(S)$,

$$\mu = (1-p) \cdot \mu_\lambda + p \cdot \mu_s,$$

where $p \in [0,1]$, $\mu_\lambda, \mu_s \in \mathscr{P}(S)$, with $\mu_\lambda \ll \lambda$ and $\mu_s \perp \lambda$.

- The rate function satisfies

$$I(\mu) = (1-p) \cdot I(\mu_\lambda) + p \cdot I(\mu_s)$$

- The rate function $I(\mu_s)$ is

$$I(\mu_s) = -\int_S \log r(x) \mu_s(dx)$$

Alternative representation of the LDP rate function

[Work in progress]

If $\mu_\lambda \ll \lambda$, the rate function I admits an **alternative representation** in the more classical Donsker Varadhan flavour:

$$I(\mu_\lambda) = -\inf_{u \in \mathcal{U}} \int_S \log\left(\frac{Ku}{u}\right)(x)\mu_\lambda(dx),$$

where $\mathcal{U} = \{u \in C(S), u > 0\}$, and

$$Ku(x) = \int_S u(y)K(x,dy).$$

Rate function lower and upper bounds

This alternative representation of the LDP rate function allows us to find the following bounds.

- $\mu \in \mathscr{P}(S)$,

$$I(\mu) \leq -\int_S \log r(x)\mu(dx) \leq +\infty$$

Rate function lower and upper bounds

This alternative representation of the LDP rate function allows us to find the following bounds.

- $\mu \in \mathscr{P}(S)$,

$$I(\mu) \leq -\int_S \log r(x)\mu(dx) \leq +\infty$$

- $\mu_\lambda \in \mathscr{P}(S)$, $\mu_\lambda \ll \lambda$, $\theta = \frac{d\mu_\lambda}{d\lambda}$

$$I(\mu_\lambda) \geq -\log \iint \sqrt{\theta(x)\theta(y)}K(x, dy)\pi(dx)$$

Variational formula for the lower bound

$$-\log \iint e^{-k} d(\pi \otimes K) = \inf_{\gamma \in \mathscr{P}(S \times S)} R(\gamma \parallel \pi \otimes K) - \iint k d\gamma$$

The $\inf$ is achieved by $\gamma_0 \in \mathscr{P}(S \times S)$ that satisfies

$$\frac{d\gamma_0}{d(\pi \otimes K)}(x, y) = \frac{e^{-k(x,y)}}{\iint e^{-k(x,y)} (\pi \otimes K)(dx, dy)}$$

Variational formula for the lower bound

$$-\log \iint e^{-k} d(\pi \otimes K) = \inf_{\gamma \in \mathscr{P}(S\times S)} R(\gamma \parallel \pi \otimes K) - \iint k d\gamma$$

The $\inf$ is achieved by $\gamma_0 \in \mathscr{P}(S \times S)$ that satisfies

$$\frac{d\gamma_0}{d(\pi \otimes K)}(x, y) = \frac{e^{-k(x,y)}}{\iint e^{-k(x,y)} (\pi \otimes K)(dx, dy)}$$

Let $\theta(x) = \frac{d\mu}{d\pi}$ and $k(x, y) = -\log \sqrt{\theta(x)\theta(y)}$, we obtain (after few intermediate steps)

$$-\log \iint \sqrt{\theta(x)\theta(y)} d(\pi \otimes K) \leq \inf_{\substack{\gamma \in \mathscr{P}(S\times S) \\ [\gamma]_1 = [\gamma]_2 = \mu}} R(\gamma \parallel \pi \otimes K) - \int \log \theta d\mu = I(\mu)$$

Algorithm tuning

Algorithm tuning

The MH proposal probability $J(\cdot|x)$ depends on parameters p

$$J(\cdot|x) = J(\cdot|x; p)$$

Algorithm tuning

The MH proposal probability $J(\cdot|x)$ depends on parameters p

$$J(\cdot|x) = J(\cdot|x;p)$$

Thus, the MH transition kernel depends on parameters p

$$K(x,dy) = K(x,dy;p).$$

Algorithm tuning

The MH proposal probability $J(\cdot|x)$ depends on parameters p

$$J(\cdot|x) = J(\cdot|x; p)$$

Thus, the MH transition kernel depends on parameters p

$$K(x, dy) = K(x, dy; p).$$

Therefore, the rate function depends on the parameters p

$$I(\mu) = \inf R(\gamma \parallel \mu(dx) \otimes K(x, dy; p)) = I(\mu; p).$$

Algorithm tuning

The MH proposal probability $J(\cdot|x)$ depends on parameters p

$$J(\cdot|x) = J(\cdot|x; p)$$

Thus, the MH transition kernel depends on parameters p

$$K(x, dy) = K(x, dy; p).$$

Therefore, the rate function depends on the parameters p

$$I(\mu) = \inf R(\gamma \parallel \mu(dx) \otimes K(x, dy; p)) = I(\mu; p).$$

Higher rate function $\Rightarrow$ faster convergence

Algorithm tuning

The MH proposal probability $J(\cdot|x)$ depends on parameters p

$$J(\cdot|x) = J(\cdot|x; p)$$

Thus, the MH transition kernel depends on parameters p

$$K(x, dy) = K(x, dy; p).$$

Therefore, the rate function depends on the parameters p

$$I(\mu) = \inf R(\gamma \parallel \mu(dx) \otimes K(x, dy; p)) = I(\mu; p).$$

Higher rate function $\Rightarrow$ faster convergence

Choose p so that $I(\mu; p)$ is maximized

Example: Independent Metropolis-Hastings

- Target $\pi \sim \mathcal{N}(0, 1)$
- Independent proposal $f \sim \mathcal{N}(m, s^2)$

Example: Independent Metropolis-Hastings

- Target $\pi \sim \mathcal{N}(0,1)$
- Independent proposal $f \sim \mathcal{N}(m, s^2)$
- Rate function lower bound

$$I(\mu) \geq -\log\left(1 - \frac{1}{2}\iint \min\left\{\frac{f(x)}{\pi(x)}, \frac{f(y)}{\pi(y)}\right\}\left(\sqrt{\mu(x)\pi(y)} - \sqrt{\mu(y)\pi(x)}\right)^2 dx dy\right)$$

Example: Independent Metropolis-Hastings

- Target $\pi \sim \mathcal{N}(0,1)$
- Independent proposal $f \sim \mathcal{N}(m, s^2)$
- Rate function lower bound

$$I(\mu) \geq -\log\left(1 - \frac{1}{2}\iint \min\left\{\frac{f(x)}{\pi(x)}, \frac{f(y)}{\pi(y)}\right\}\left(\sqrt{\mu(x)\pi(y)} - \sqrt{\mu(y)\pi(x)}\right)^2 dx\,dy\right)$$

- Maximized when $(m, s) = (0, 1)$

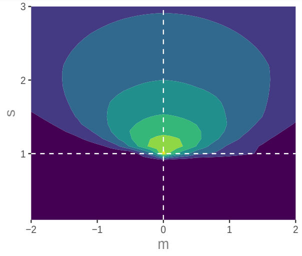
Example: Independent Metropolis-Hastings

- Target $\pi \sim \mathcal{N}(0,1)$
- Independent proposal $f \sim \mathcal{N}(m, s^2)$
- Rate function lower bound

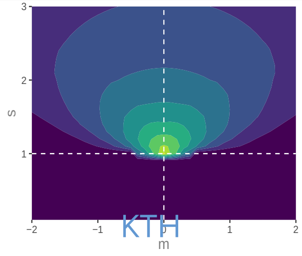
$$I(\mu) \geq -\log\left(1 - \frac{1}{2}\iint \min\left\{\frac{f(x)}{\pi(x)}, \frac{f(y)}{\pi(y)}\right\}\left(\sqrt{\mu(x)\pi(y)} - \sqrt{\mu(y)\pi(x)}\right)^2 dx\,dy\right)$$

- Maximized when $(m, s) = (0, 1)$

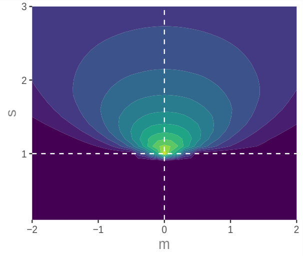
$\mu \sim \mathcal{N}(1,2)$



$\mu \sim Gamma(3,5)$



$\mu \sim Unif(0,1)$



Future directions

Future directions

- LDP assumptions:
  - Generalise assumptions to the LDP to account for methods such as ABC-MCMC
  - Open problem: geometric ergodicity $\Rightarrow$ LDP?
- Rate function:
  - Find a more explicit formula for the rate function I
  - Design an algorithm to approximate the rate function or find the argmax
- Use the LDP:
  - Compare MCMC algorithms
  - Compare LDP with other convergence analysis tools
  - Design algorithms to precalibrate MH-based MCMC methods

References

- Milinanni, F., & Nyquist, P. (2024). *A large deviation principle for the empirical measures of Metropolis–Hastings chains.* Stochastic Processes and their Applications, 170, 104293.
- Milinanni, F., & Nyquist, P. (2024). *On the large deviation principle for Metropolis-Hastings Markov Chains: the Lyapunov function condition and examples.* arXiv preprint arXiv:2403.08691.

KTH
VETENSKAP
OCH KONST