Survey of Hamiltonian Monte Calro

竹田航太

2021 年 2 月 6 日

目次

概要

Betancourt の survey 論文 [1] をもとにしている．本稿はハミルトニアンモンテカルロ (Hamiltonian Monte Calro: HMC) の理論的背景を数学的な詳細には立ち入らず，物理的なイメージを交えつつ直感的に説明する．まず，ランダムサンプリングの基本である Malkov Chain Monte Calro(MCMC) について説明する．この方法は直感的であり実装も非常に単純だが，計算効率が悪く得られる結果の有効性の保証も弱い．

HMC は生まれた当初は Hybrid Monte Carlo という呼び名ついていたように，gradient の情報を利用した deterministic な遷移と stochastic な遷移を合わせてサンプリングを行う．パラメータ空間を位相空間に拡張して Hamiltonian を保存するように deterministic に動き，その後 stochastic に別の等 Hamilton 面に移動するという 2step を繰り返す．このような HMC のサンプリングは効率がよく，得られる結果の有効性も理論的に広く保証されている．

積分時間やエネルギー遷移など HMC 実装上の自由度が残されておりそのチューニングや複雑化に改善の可能性がある．

History of Hamiltonian Monte Calro

1980 年代末期に Hybrid Monte Carlo と言う名前で Lattice Quantum Chromodynamics の計算のために用いられた．[2, Hybrid Monte Carlo]

数年後，Radford Neal がこの手法のポテンシャルに気づき Bayesian neural networks の開拓時に利用した．[3, An Improved Acceptance Procedure for the Hybrid Monte Carlo Algorithm]

次の 10 年でこの手法が教科書に載りはじめる．[4, Infomation Theory] で Hamiltonian Monte Carlo という言葉が初めて使われた．その他の教科書は [5, bishop:2006:PRML]

Neal’s influential review (Neal, R. M. (2011). MCMC Using Hamiltonian Dynamics. In Handbook of Markov Chain Monte Carlo (S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng, eds.) CRC Press, New York.) で実際の統計的数値計算に用いる方法を紹介.

C++ のライブラリ Stan にも使用されている．

1 Computing Expactation

興味のある確率分布から情報を得る際に何らかの関数の期待値を計算することは重要．以下のような設定の問題を考える．

設定: $D \in \mathbb{N}$, 標本空間 $Q \subset \mathbb{R}^D$, target distribution density π, 目的関数 f on Q に対して期待値 $\mathbb{E}_\pi[f]$ を求める．

$\mathbb{E}_\pi[f]$ は積分で表せる．

$$\mathbb{E}_\pi[f] = \int_Q f(q)\pi(q)dq$$

変数変換

$$= \int_{Q'} f(q')\pi(q')dq'$$

積分は一般には解析的に計算できるとは限らないので数値的に近似する必要がある．ただし，ここで $|E_\pi[f]| = \infty$ であるような場合でも数値的には収束しているように見える場合があるので注意が必要である．[6, chapter 9]

1.1 計算量の削減

keys

- density が最大になるような mode の近傍で積分を近似
- density と volume のバランス

数値積分の非効率性について

積分への寄与が小さい標本空間上の領域を sample すると効率が悪いので，目的の分布と関数を見て積分への寄与が大きい領域を判断する必要がある．(パラメータの取り方を変える)

実用上は 1 つの density に対して複数の目的関数についての期待値を計算することがある．このとき複数の目的関数について一様に効率の良いパラメータの取り方を見つける必要がある．

パラメータの取り方の問題について少なくともの 1 次元の積分の場合には効率の良い結果が得られている．'Keep in mind, however, that if only a single expectation is in fact of interest then exploiting the structure of that function can provide significant improvements ' in [7]

積分を近似評価する際に重要なのは density が大きくなる mode の近傍である．この直感は MLE や Laplace approximation など多くの統計的手法でも見られる．しかしながら，高次元になるとこの直感に反することが起きる．density が大きい領域周辺の volume が非常に小さくなる場合があり結果として積分への寄与が小さくなってしまう．このため，density と volume の

両方を考慮する必要がある．

1.2 The Geometry of High-Dimensional Spaces

keys

- marginal volume

高次元空間では「外側の体積」がとても大きくなる．

例えば，$[0,1]$ 区間を 3 分割することを考える．すると中心 ($[\frac{1}{3}, \frac{2}{3}]$) の長さは全体の $\frac{1}{3}$．同様に 2 次元で考えると $[0,1] \times [0,1]$ を 9 分割することができるが中心の面積は全体の $\frac{1}{9}$．3 次元では $\frac{1}{27}$，D 次元では $\frac{1}{3^D}$ となり中心の体積の割合が非常に小さい．

もう少し厳密には，中心からの距離が δr 変化したときの D 次元球体の体積の変化の割合は

$$\delta V \propto r^{D-1} \delta r$$

と半径 r の冪乗に比例し r が大きいほど (中心から遠ざかるほど) 半径の摂動に対する体積の変化は大きくなる．また，次元 D が大きくなるほどこの効果は強くなる．

1.3 The Geometry of High-Dimensional Probability Distributions

keys

- typical set
- concentration of measure

typical set: 前のセクションにある通り高次元空間では体積の増加が中心より周辺の方が大きくなる．このため積分への寄与は「density が大きくなる mode の近傍」のより「density が小さい周辺」の方が大きくなってしまうこともある．このように平均 (積分) への寄与を考える際には density と volume の両方を考慮する必要がある．これらを考慮した上で積分への寄与の大きい領域のことを typical set と呼ぶ.[*1]

concentration of measure: 次元が上がるにつれて density と volume が (支配的な寄与をする程度に) 共に大きいような領域は狭くなっていき，singular になっていく．この領域以外の積分への寄与は小さくなり，無視できるようになる．このため typical set 上での積分を計算することで効率よく良い精度で積分を評価できる．

[*1] typical set とは元々情報エントロピーを近似できる点の集合と定義されたようである．

2 Malkov Chain Monte Calro

ランダムな移動により typical set を探索する．Markov chain は十分な時間があれば typical set に到達する．

2.1 Estimating Expectations with Markov Chains

keys

- Markov transitions
- 条件付き確率 (conditional probability)
- 詳細釣り合い (detailed balance)

目的の分布の density に従う点列を $(q_1, \cdots, q_n)$ を構成しそれらによって期待値を推定する．以下を満たすような transition を定義してそれに従って sample 空間を探索する．これを満たすような π は $\mathbb{T}$ に対して stational という．

$$\pi(q) = \int_Q dq' \pi(q') \mathbb{T}(q|q')$$

このとき Markov transition は以下の詳細釣り合い条件を満たしていることが望ましい．

$$\pi(q)\mathbb{T}(q'|q) = \pi(q')\mathbb{T}(q|q') \tag{2.1}$$

Markov transition$\mathbb{T}(q|q')$ に従って Markov Chain$\{q_1, q_2, \cdots, q_N\}$ を生成しそれによって平均を推定する．

$$\hat{f}_N = \frac{1}{N} \sum_{n=0}^{N} f(q_n) \to \mathbb{E}_\pi[f] \ (N \to \infty)$$

2.2 ideal behavior

keys

- CLT
- effctive sample size
- sample bias

Markov Chain に期待される漸近挙動として CLT がある．

$$\hat{f}_N \sim \mathcal{N}(\mathbb{E}_\pi[f], \sigma_{mcmc})$$

ただし，σ_{mcmc}(Markov Chain Monte Carlo Standard Error) は以下で与えられ,

$$\sigma_{mcmc} \coloneqq \sqrt{\frac{\mathrm{Var}_\pi[f]}{N_{ess}}}$$

有効サンプルサイズ (effective sample size) は以下で定義される.

$$N_{ess} \coloneqq \frac{N}{1 + 2\sum_{l=1}^{\infty} \rho_l}$$
$$\rho_l = \sum_{n=l}^{N} f(q_n) f(q_{n-l})$$

2.3 Pathological Behavoir

keys

- pathological curvature
- geometric ergordicity
- split $\hat{\mathrm{R}}$ static

typical set に sigular な曲率を持つ領域が存在する場合，Markov Chain がその付近に留まって振動してしまう．目的の分布が Markov Chain に対して pathological な挙動を示すかどうかは Markov transition に依存する.

2.3.1 CLT for Markov Chain

Markov chain により近似された平均が CLT に従って真の平均に近づくことを保証する条件がいくつかある．詳細は [8] によるが以下のいくつかを確認する必要がある.

(1) 可逆性 (reversibblity): これは詳細釣り合い条件のこと
(2) ϕ-既約制 (ϕ-irreducibility)
(3) 非周期性 (aperiodicity)
(4) 一様エルゴード性 (uniform ergodicity)
(5) 幾何的エルゴード性 (geometric ergodicity)

しかし，これらは理論的な条件であり単純な問題でなければ確認するのは困難であるので代わりに数値的な判定法に頼りたい．それが [9] の split $\hat{\mathrm{R}}$ static である.

2.4 The Metropolis-Hastings Algorithm

keys

- Metropolis-Hastings Algorithm
- proposal density
- intuitive and simple implementation
- poor parfomance with high-dimension and complex target distributions

2.4.1 Markov Chain の構成

目的の分布 (特に density が与えられている) から Markov Chain を生成するには適切な Markov transition を構成する必要があるがこれは non-trivial な問題である．それを解決するのが Metropolis-Hasting Algorithm である．2step で簡単に可逆な (詳細釣り合い条件を満たす)Markov Chain を構成できる．

提案分布を導入しその density を $Q(q'|q)$ とする．各時刻で次の Chain の候補を提案し，以下の確率で採択する．

$$a(q'|q) = \min\left(1, \frac{Q(q|q')\pi(q')}{Q(q'|q)\pi(q)}\right) \tag{2.2}$$

2.4.2 実用

提案分布としてよく使われるのは Gaussian で $Q(q'|q) = \mathcal{N}(q'|q, \sigma)$ この場合 Random Walk Metropolis と呼ばれる．$Q(q'|q)$ が対象になるので (2.2) は簡単にかけて

$$a(q'|q) \min\left(1, \frac{\pi(q')}{\pi(q)}\right)$$

Random Walk Metropolis は実装が簡単でありかつ直感的なアルゴリズムなので広く用いられてきた．しかし，単純さゆえ高次元サンプル空間や複雑な目的分布に対してのパフォーマンスはよくない．提案分布の分散が大きいと候補点は typical set の外に出てしまい reject される．一方，分散が小さいと候補点が typical set に入り accept されるが移動距離が小さく探索効率が悪い．結果的に MCMC は biased で大きな自己相関を持ってしまう．

ただし，これを解決するため Anealing という方法がありさらにそれを並行して行うレプリカ交換法というのもあるらしい．

2.4.3 まとめ

MCMC の収束は step 数の平方根のオーダー程度が限界である．これはパラメータ空間の次元には依存しないが遅い．

3 Hamiltonian Monte Calro の基礎

提案と採択といった MCMC の戦略は高次元空間ではうまくいかない．指数的に多い ($3^D - 1$ のような) 選択肢の中か特異的に少ない typical set への方向を選ぶ必要がある．typical set(あるいは target distribution) の幾何的情報を利用し，より deterministic な探索が求められる．

Hamiltonian Monte Calro(以下 HMC) を理解する第 1 歩として，確率的システムを物理的に捉え直すことにある．確率密度 (density)$\pi(q)$ を以下のように書き直す．

$$\pi(q) \propto e^{-V(q)}$$

ただし，$V(q) = \log(\pi(q))+\{$ 任意の q の加法的関数 $\}$．ここで $V(q)$ はポテンシャルに相当する．

3.1 効率の良い Markov transition と物理的イメージ

keys

- gradient
- independent on parametrization
- auxiliary momentum

効率よく typical set を探索するため typical set に沿ったベクトル場に従って次の点をとることを考える．このベクトル場を目的分布のみから構成できれば良い．

目的関数から得られる自然な情報として density の gradient(勾配) がある．しかし gradient に沿って動くと mode(peak) の近くに吸い寄せられてしまう．(パラメータの取り方で目的関数さらには gradient も変わるが gradient は mode の近傍でパラメータに sensitive) gradient の情報を生かす以下の考え方がある．

パラメータ普遍性 (parametrization-invariant): gradient の情報をより扱いやすくするには typical set 方向のパラメータの取り方に依存しないような不変量を見つける必要がある．このことの詳しい理解には微分幾何学の知識が必要である．

物理的イメージ: 基本的に任意の確率的な系にはそれに対応する (直感的にわかりやすい) 物理系が存在する．今の場合だと，mode，gradient, typical set は惑星，重力，衛星軌道に対応する．軌道上の衛星が惑星に落ちていかないのは momentum(運動量) があるからである．ただし運動量が大きすぎても小さすぎても衛星は軌道から外れてしまう．衛星に適切な運動量を与えることで衛星は軌道に留まり，系は保存系になる．

gradient の情報と適切な運動量を与えることで効率の良い typical set の探索が実現する．

3.2 相空間と Hamilton 方程式

keys

- Hamilton 方程式
- 保存系

3.2.1 物理的知識

物理における保存系を考える場合は体積が保存される必要がある．Hamilton 方程式を満たす系では体積が保存するという Liouville の定理が知られている．付録 A

Definition 3.1 (Hamilton 方程式)**.** N 粒子系の相空間 $\mathbb{R}^{2N}$ を考える．$(\mathbf{q}, \mathbf{p}) \in R^{2N}$ が *Hamilton* 方程式を満たすとはある $H(q,p)$ があって，$i = 1, 2, \cdots, N$ で

$$\frac{\partial q_i}{\partial t} = \frac{\partial H}{\partial p_i} \tag{3.1}$$

$$\frac{\partial p_i}{\partial t} = -\frac{\partial H}{\partial q_i} \tag{3.2}$$

が成り立つこと．

H のことを *Hamiltonian* という．

3.2.2 保存系の構成

保存系を考えるためにせん断応力 (sear stress) のみが働くような空間を想定する．パラメータ q の空間での変化を運動量 p の空間で補完して (q,p) の空間全体で体積の保存を実現する．このため次元は 2 倍になる．$(q_n \to (q_n, p_n))$ 目的分布も $\pi(q,p)$ は条件付確率を用いて以下のように拡張する．

$$\pi(q,p) = \pi(p|q)\pi(q)$$

さらに，$\pi(q,p)$ はパラメータの取り方に依存しないので Hamiltonian $H(q,p)$ を用いて

$$\pi(q,p) = e^{-H(q,p)} \tag{3.3}$$

とかく．$H(q,p)$ を以下のように運動エネルギーとポテンシャルエネルギーに分解できる．

$$\begin{aligned} H(q,p) &= -\log(\pi(q,p)) \\ &= -\log(\pi(p|q)) - \log(\pi(q)) \\ &= K(q,p) + V(q) \end{aligned}$$

3.3 Hamiltonnian Markov 遷移

元のパラメータ空間での $q \to q'$ の遷移を考える.

(1) $q \to \tilde{p}$: 位置 q に対して $\tilde{p} \sim \pi(\tilde{p}|q)$ で運動量 $\tilde{p}$ をとる.
(2) $(q, \tilde{p}) \to \phi_t(q, \tilde{p})$: Hamilton 方程式に従って時間 t だけ進める.
(3) $\phi_t(q, \tilde{p}) \to q'$: $\phi_t(q, \tilde{p})$ を射影して q' を得る.

3.4 まとめ

物理的な知見から効率の良い Markov 遷移を構成する方針を得ることができた.

4 Effcient Hamiltonian Monte Carlo

keys

- 運動エネルギーの選択
- 積分時間のチューニング

HMC の効率に関わる自由度は上記の 2 つ.

4.1 相空間の幾何

keys

- HMC の再解釈
- Hamiltonian level set
- microcanonical distribution

Hamilton 方程式の解軌道は Hamilonian の等値線 (level set) となるので, エネルギー E を用いて特徴付けられる.

$$H^{-1}(E) = \{(q, p) \in \mathbb{R}^{2D}; H(q, p) = E\} \subset \mathbb{R}^{2D-1}$$

この表現で相空間上の density を以下のように書き換えられる.

$$\pi(q, p) = \pi(\theta_E|E)\pi(E)$$

相空間上の点をエネルギー E と level set 上の角度 θ_E で一意的に表せる.

HMC を以下 2 つの step で再解釈できる.

(d) deterministic な同一 level set 内の移動

(s) stochastic な異なる level set 間の移動.

この分解の各 step を解析することで HMC の効率に関わるポイントが見えてくる．(d) では積分時間のチューニングが必要であり，level set の幾何情報に依存する．(s) での探索効率は選択する運動エネルギー遷移の分布 $\pi(E|q)$(運動エネルギー $K(q,p)$ の形に依存) が $\pi(E)$ にどれくらい近いかに依存する．$\pi(E)$ がより heavy-tailed だと探索効率が悪くなる．

4.2 最適な運動エネルギー選択

HMC における level set 内の deterministic な移動の効率に影響を与える運動エネルギーの選び方を考える．運動エネルギーを選ぶ上での要請は level set $H^{-1}(E)$ が uniform であることと運動エネルギー遷移の分布 $\pi(E|q)$ が $\pi(E)$ と「近い」こと．

4.2.1 Euclidean Gaussian

まず Gaussian の運動エネルギーを考える．

相空間の位置の方向に重み $M \in \mathbb{R}^{D\times D}$ で距離を定める．

$$d(q,q') = (q-q')^\top M(q-q')$$

duality から運動量方向に誘導される距離は

$$d(p,p') = (p-p')^\top M^{-1}(p-p')$$

この距離から分布を構成できる．例えば

$$\pi(p|q) = \mathcal{N}(p;0,M)$$

これは以下の運動エネルギーに対応する．

$$K(q,p) = \frac{1}{2}p^\top Mp + \log(|M|) + \text{const}$$

次に最適な重み M を考える．パラメータ空間の変換によって M^{-1} が目的分布の Covariance に近づくほど相関が小さくなり Hamiltonian level set が「一様」になる．このことから最適な M は

$$M^{-1} = \mathrm{Cov}_\pi[q] = \mathbb{E}_\pi[(q-\mu)^\top(q-\mu)] \qquad (\mu = \mathbb{E}_\pi[q])$$

4.2.2 Riemanian Gaussian

目的分布が Gaussian であっても global$M$ によって「一様」な level set を達成できないこともある．これに対応するために Euclid 幾何を一般化して Riemann 幾何を考える．つまり距離

M を位置 q に依存して変化する $\Sigma(q)$ に拡張する．Euclid の場合と同様に Gaussian を考えると

$$\pi(p|q) = \mathcal{N}(p; 0, \Sigma(q))$$
$$K(q,p) = \frac{1}{2}p^\top \Sigma(q)^{-1}p + \log(|\Sigma(q)|) + \text{const}$$

実装は [10]

4.2.3 Non Gaussian

理論上は運動エネルギーは non Gaussian も考えられるが理論的なサポートはまだない．また計算上のパフォーマンスは良くない．高次元パラメータ空間では比較的弱い条件でエネルギーの周辺分布 $\pi(E)$ は中心極限定理に従い Gaussian に収束し，その場合 non Gaussian の運動エネルギーのパフォーマンスは良くない．

4.3 積分時間

keys

- dyanamic ergodicity
- dynamic tuning
- No-U-Turn Sampler

運動エネルギーを決めると Hamiltonian level set が決まる．あとは level set に合わせて最適な積分時間 $T(q,p)$ を求める．

4.3.1 理論

ここでの目標は相空間上の分布を Hamilton orbit[*2]に制限した分布からうまくサンプルすることである．(orbit が目的の分布をよく近似しているかどうかは別の問題であり，運動エネルギー選択に依存する．)

ここで dynamic ergodicity という概念を導入する．一般にエルゴード性と呼ばれるものである．dynamic ergodicity は積分時間を長くすれば HMC の trajectory からの一様なサンプルが orbit 上の目的分布に近づくことを保証する．詳細は付録 B

例えば，目的分布が $\pi_\beta \propto e^{-|q|^\beta}$, 運動エネルギーが $\pi(p|q) = \mathcal{N}(0,1)$ である場合，最適な積分時間は $T_{opti}(q,p) \propto (E)^{\frac{2-\beta}{2\beta}}$ となる．特に $\beta < 2$ の場合は $E = H(q,p)$ が大きくなると最適積分時間 T_{opti} も大きくなる．heavy-tailed であるほど tail の探索に時間がかかる．

[*2] HMC の chain が取りうる全ての点の集合を orbit とよぶ.

4.3.2 実用

実用上は Hmilton trajectory から動的に T_{opti} を同定する必要がある．現在使われている方法は以下

(1) No-U-Turn(NUTS) termination creterion: 初期の点から時間前後方向に Hamiltona trajectory を伸ばしていき両端が近づいたら止めて trajectory からサンプルする．[11]
(2) Exhaustive termination: Hamiltonian から十分な積分時間計算する．NUTS より robust.

積分時間のチューニングは Open problem である．

5 Inplementing Hamiltonian Monte Carlo in Practice

keys

- Sympletic 法

一般に Hamilton 方程式の解析解を求めるのは困難なので数値的に trajectory を計算する．ここでは Hamilton 方程式の数値解法の最適化について議論を行う．

5.1 Sympletic 法

Hamilton 方程式を数値的に解く一般的な方法が sympletic 法である．この方法は「エネルギー」と「体積」を保つ数値スキームである．Hamiltonian に近い modified Hamiltonian を保存し厳密解の近くに留まる軌道を描く．数値誤差が時間発達しないので長時間積分が可能.

5.1.1 アルゴリズム

よく使われる 2 次の Sympletic 法は leap-frog 法により時間 step を ϵ として,

$$
\begin{aligned}
p_{n+\frac{1}{2}} &= p_n - \frac{\epsilon}{2}\frac{dV}{dq}(q_n) \\
q_{n+1} &= q_n + \epsilon p_{n+\frac{1}{2}} \\
p_{n+1} &= p_{n+\frac{1}{2}} - \frac{\epsilon}{2}\frac{dV}{dq}(q_{n+1})
\end{aligned}
$$

5.1.2 Sympletic 法の問題

Sympletic 法では小さな誤差が時間発達しないものの常に残る.

5.2 積分誤差の補正

keys

- acceptance probability
- Hamiltonian による監視

積分の誤差からくる bias を補正する方法の１つとして, Hamilton 遷移を相空間上の Metropolis-Hasting の proposal として扱う方法がある.

まず, 提案分布 (proposal density) による提案 $(q,p) \to (q',p')$ の確率を $Q(q',p'|q,p)$ とかく. 今，点 (q_0,p_0) から L step だけ Sympletic 法で進めた点を (q_L,p_L) とすると，deterministic な移動なので $Q(q',p'|q_0,p_0) = \delta(q'-q_L)\delta(p'-p_L)$ となる．しかし，これでは可逆でなくなる. (i.e. $Q(q_0,p_0|q_L,p_L) = 0$ なので Metropolis-Hasting の採択確率が以下のように 0 となってしまう.

$$a(q_L,p_L|q_0,p_0) = \min\left\{1, \frac{Q(q_0,p_0|q_L,p_L)}{Q(q_L,p_L|q_0,p_0)}\frac{\pi(q_L,p_L)}{\pi(p_0,q_0)}\right\} = \min\left\{1, \frac{0}{1}\frac{\pi(q_L,p_L)}{\pi(p_0,q_0)}\right\} = 0$$

)

これを回避するために移動後に momentaum を flip する．$(q_0,p_0) \to (q_L,p_L) \overset{flip}{\to} (q_L,-p_L)$ これにより

$$Q(q',p'|q_0,p_0) = \delta(q'-q_L)\delta(p'+p_L)$$

となり $Q(q_0,p_0|q_L,-p_L) = Q(q_L,-p_L|q_0,p_0) = 1$ なので

$$\begin{aligned}
a(q_L,-p_L|q_0,p_0) &= \min\left\{1, \frac{Q(q_0,p_0|q_L,-p_L)}{Q(q_L,-p_L|q_0,p_0)}\frac{\pi(q_L,-p_L)}{\pi(p_0,q_0)}\right\} \\
&= \min\left\{1, \frac{1}{1}\frac{\pi(q_L,-p_L)}{\pi(p_0,q_0)}\right\} \\
&= \min\left\{1, \frac{e^{-H(q_L,-p_L)}}{e^{-H(p_0,q_0)}}\right\} \\
&= \min\left\{1, \exp(-H(q_L,-p_L)+H(p_0,q_0))\right\}
\end{aligned}$$

と採択確率を構成することができる.

5.3 Sympletic 法のパラメータ

keys

- 時間 step ϵ
- 階数 K
- modified Hamiltonian

Sympletic 法実装の自由度として時間 step ϵ と階数 K がある．ϵ を小さくし，K を大きくすると精度が上がるがその分時間がかかる．

Sympletic 法は Hamiltonian の代わりに ϵ^K 程度 modify された Hamiltonian $\tilde{H}$ を保存するのでその起動は ϵ^Kmodified Hamiltonian level set となる．modified Hamiltonian level set の安定性などを調べることによって ϵ や K を決める．

2 次の Sympletic 法の単純な実装で modified Hamiltonian level set が well-behaved な場合には H と $\tilde{H}$ の差を計算コストと平均採択率の関係で評価できる．[12] これから ϵ のチューニングを行う．

6 Robustness of Hamiltonian Monte Carlo

keys

- 理論的保証
- 計算の監視

いくら高速なアルゴリズムでも typical set をうまくサンプルしていなければ意味がない．この問題に対する現状の結果としては単純な HMC でも広いクラスの目的分布について geometric ergordicity が成り立つことが知られている．[13] このクラスは Random Walk Metropolis のような non-gradient mdethod よりもずっと広いクラスである．

6.1 Diagnosing Porrly-Chosen Kinetic Energy

運動エネルギーの選択が適切かどうかを Hamiltonian trajectory から診断する方法がある．Bayesian fraction of missing information を利用する．

$$\text{E-BFMI} := \frac{\mathbb{E}_\pi[\text{Var}_{\pi_{E|q}}[E|q]]}{\text{Var}_{\pi_E}[E]} \approx \text{E-B}\hat{\text{F}}\text{MI} := \frac{\sum_{n=1}^{N}(E_n - E_{n-1})^2}{\sum_{n=0}^{N}(E_n - \bar{E})^2}$$

これが 0.3 以下だと望ましくない．

6.2 Diagnosing Regions of High Curvature

High Curvature な領域が typical set に存在するとその近傍では MCMC や HMC ではうまくサンプルできず bias が生まれる．特に hierarchical model などはこのような性質を持つ．

Pathological な領域に入ると計算が発散するのですぐわかる．Pathological Curvature への対応は ϵ を小さくするもしくは強い事前分布で正則化すること．

6.3 Limitations of Diagnostics

上記の診断方法は必要条件に過ぎないので，split $\hat{R}$ statics などと合わせて考えるべき．

7 Conclusion

MCMC と比べて HMC は計算効率だけでなく，得られたサンプル結果の有効性もより強く保証する．non-Gaussian や非ユークリッド運動エネルギーの利用．

また，geometric method を拡張した Adiabatic Monte Carlo(パラメータ β を追加) なども考えられている．

HMC は非常に理論的なサポートがしっかりしているが実用上はデータ駆動判定する必要がある部分が多い．

付録 A Liouville の定理

一般に物理学では以下の質量保存則が仮定される．

Theorem 付録 A.1 (質量保存則・連続の式)**.**

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \tag{付録 A.1}$$

Hamilton 方程式と質量保存則を合わせると Liouville の定理が導かれる．

Theorem 付録 A.2. N 粒子系の相空間 $\mathbb{R}^{2N}$ を考える．$(\mathbf{q}, \mathbf{p}) \in R^{2N}$ が *Hamilton* 方程式を満たすとする．ρ を粒子の密度，$\mathbf{u}$ を粒子の流れとすると，以下が成り立つ．

$$\frac{d\rho}{dt} = 0 \tag{付録 A.2}$$

つまり，*Hamilton* 方程式に沿った運動で体積は保存する．

付録 B　dynamic ergodicity

ここで dynamic ergodicity という概念を導入する．一般にエルゴード性と呼ばれる物である．HMC の chain が取りうる全ての点の集合を orbit とよび ϕ とかく．また，$z \in \mathbb{R}^{D\times D}$ を出発して時間 t 進めた orbit 上の点を $\phi_t^H(z)$ とかく．

Definition 付録 B.1 (dynamic ergodicity)**.** *HamiltonianH* に対して，その軌道が *dynamic ergodicity* を満たすとは任意の $f : \mathbb{R}^{D\times D} \to \mathbb{R}$ に対して以下が成り立つこと．

$$\lim_{T\to\infty} \langle f \rangle_{\phi^H}(z, T) \coloneqq \lim_{T\to\infty} \frac{1}{T} \int_0^T f \circ \phi_t^H(z) dt = \mathbb{E}_{\pi_{H^{-1}(E)}}[f] \tag{付録 B.1}$$

参考文献

[1] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo, 2018.

[2] Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics Letters B*, 195(2):216 – 222, 1987.

[3] Radford M. Neal. An improved acceptance procedure for the hybrid monte carlo algorithm. *Journal of Computational Physics*, 111(1):194 – 203, 1994.

[4] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] Timothy John Sullivan. *Introduction to uncertainty quantification*, volume 63. Springer, 2015.

[7] Antonietta Mira, Reza Solgi, and Daniele Imparato. Zero variance markov chain monte carlo for bayesian estimators. *Statistics and Computing*, 23(5):653–662, Jul 2012.

[8] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space markov chains and mcmc algorithms. *Probability Surveys*, 1(0):20–71, 2004.

[9] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis, 3rd Ed.* 01 2014.

[10] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[11] Matthew D. Homan and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, January 2014.

[12] M. J. Betancourt, Simon Byrne, and Mark Girolami. Optimizing the integrator step size for hamiltonian monte carlo, 2015.

[13] Samuel Livingstone, Michael Betancourt, Simon Byrne, and Mark Girolami. On the geometric ergodicity of hamiltonian monte carlo, 2018.

[14] Michael Betancourt. Identifying the optimal integration time in hamiltonian monte carlo, 2016.

[15] Stan user guide reference manual. `https://stan-ja.github.io/gh-pages-html`.