

# Optimal tuning of Hybrid Monte Carlo

J. M. Sanz-Serna

Universidad de Valladolid, Spain

Joint work with A. Beskos, N. S. Pillai,  
G. O. Roberts & A. M. Stuart

## I. THE SET-UP

## MARKOV CHAIN MONTE CARLO (MCMC) METHODS

- Wish to sample from a density  $\Pi$

$$\Pi(Q) = \exp(-\mathcal{V}(Q)), \quad \mathcal{V} : \mathbb{R}^N \rightarrow \mathbb{R}.$$

- MCMC methods construct a Chain  $Q^0 \mapsto Q^1 \mapsto Q^2 \mapsto \dots$  that has  $\Pi$  as an invariant density.
- Hope that after ‘burn in’ steps,  $Q^n$  distributed according to  $\Pi$ .
- Examples of methods follow.

(i) Random-Walk Metropolis (RWM) (Metropolis *et al.* 1953)

- From current location  $Q$  chain makes a ( $\Pi$ -independent) proposal

$$Q' = Q + \sqrt{h}Z,$$

where  $h > 0$  is a parameter and  $Z$  is drawn from a symmetric distribution (typically  $Z \sim N(0, I)$ ).

- Proposal accepted with probability (Metropolis[-Hastings] rule)

$$a(Q, Q') = 1 \wedge \frac{\Pi(Q')}{\Pi(Q)}.$$

- If accepted,  $Q'$  is next state of chain. Else chain stays at  $Q$ .
- Markov chain reversible wrt invariant target density  $\Pi$ .
- If  $h$  is large, proposals are typically not accepted.
- If  $h$  is small, samples are highly correlated.
- If dimensionality  $N$  is high, poor mixing, slow moves and high correlation between samples.

(ii) Metropolis-adjusted Langevin algorithm (MALA) (Roberts & Tweedie 1996)

- Proposal searches for ( $\Pi$ -dependent) high-probability locations

$$Q' = Q + \frac{h}{2} \nabla \log \Pi(Q) + \sqrt{h} Z, \quad Z \sim N(0, I)$$

(Euler discretization with stepsize  $h$  of Langevin equation

$$dQ_t = \frac{1}{2} \nabla \log \Pi(Q_t) dt + dW_t.)$$

- Metropolis-Hastings recipe for acceptance produces Markov chain reversible wrt invariant target density.

(iii) Hybrid Monte Carlo Method (HMC) (Duane *et al.* 1987)

- Introduce Hamiltonian fnctn ( $\mathcal{M}$  sym. pos. def. mass matrix)

$$\mathcal{H}(Q, P) = \frac{1}{2} \langle P, \mathcal{M}^{-1} P \rangle + \mathcal{V}(Q),$$

with equations of motion

$$\frac{dQ}{dt} = \mathcal{M}^{-1} P, \quad \frac{dP}{dt} = -\nabla \mathcal{V}(Q).$$

- The corresponding solution flow

$$(Q(t), P(t)) = \Phi_t(Q(0), P(0))$$

(i) *conserves energy*:  $\mathcal{H} \circ \Phi_t = \mathcal{H}$ .

(ii) *conserves the volume element*  $dQdP$ .

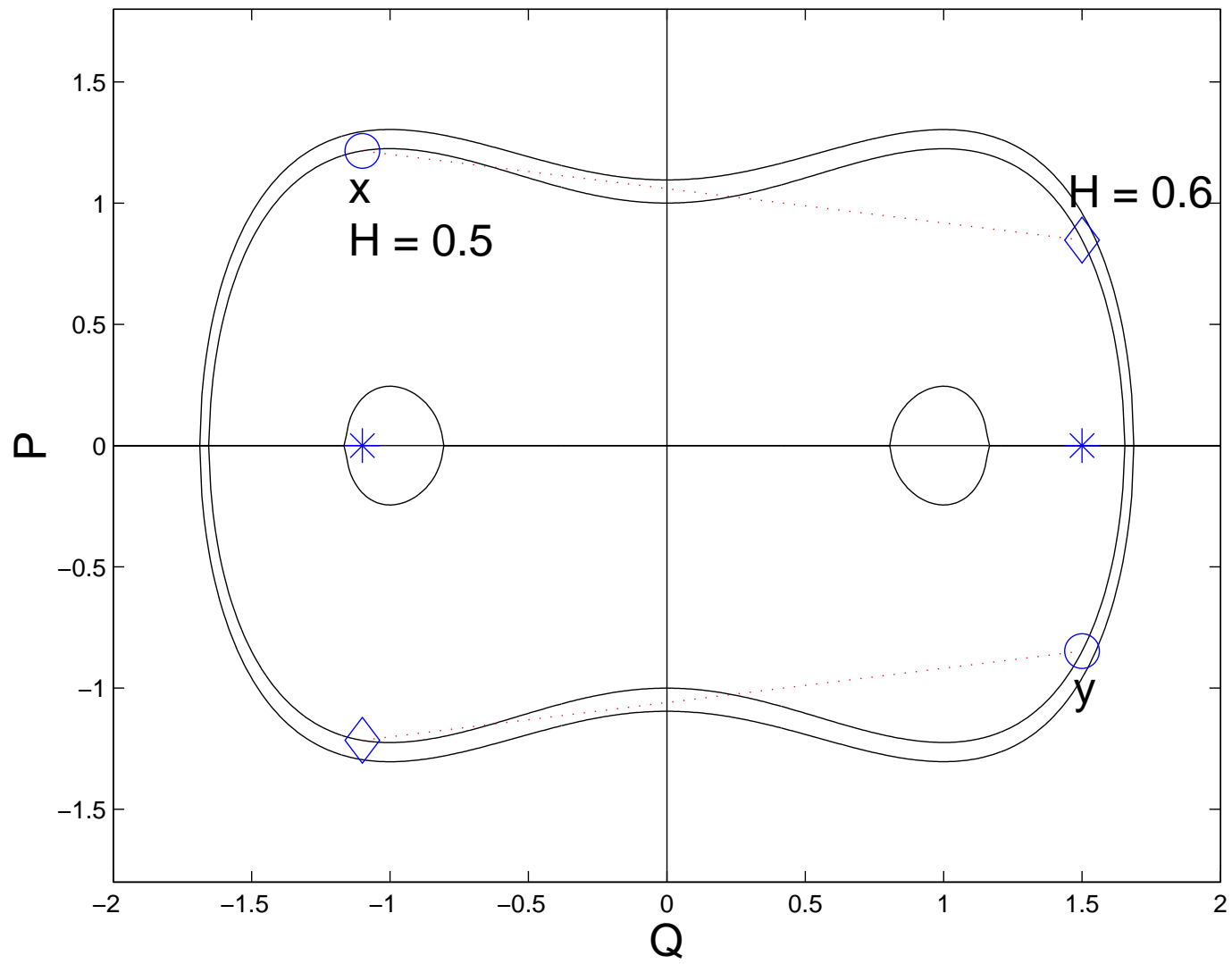
(iii) *is reversible*: if  $\mathcal{S}(Q, P) = (Q, -P)$ , then  $(\Phi_t)^{-1} \circ \mathcal{S} = \mathcal{S} \circ \Phi_t$ .

- From these properties it follows that  $\Phi_t$  preserves any density depending only on  $\mathcal{H}$ , in particular it preserves

$$\exp(-\mathcal{H}(Q, P)) = \exp((1/2)\langle P, \mathcal{M}^{-1}P \rangle) \exp(-\mathcal{V}(Q)),$$

and its marginal, ie our target  $\Pi(Q) = \exp(-\mathcal{V}(Q))$ . ( $P$  is Gaussian with covariance matrix  $\mathcal{M}$ .)

- Therefore  $\Phi_T$ , for any fixed  $T$ , may be used to build Markov chain for  $Q$ .
- ‘Global’ moves offer potential for faster exploration of state space.



- $\Phi_t$  not known and must resort to numerical simulation. Verlet algorithm is typical choice. A step of length  $h$  from  $(Q_0, P_0)$  is

$$\begin{aligned} P_{h/2} &= P_0 - \frac{h}{2} \nabla \mathcal{V}(Q_0), \\ Q_h &= Q_0 + h \mathcal{M}^{-1} P_{h/2}, \\ P_h &= P_{h/2} - \frac{h}{2} \nabla \mathcal{V}(Q_h). \end{aligned}$$

- The scheme gives rise to a map:

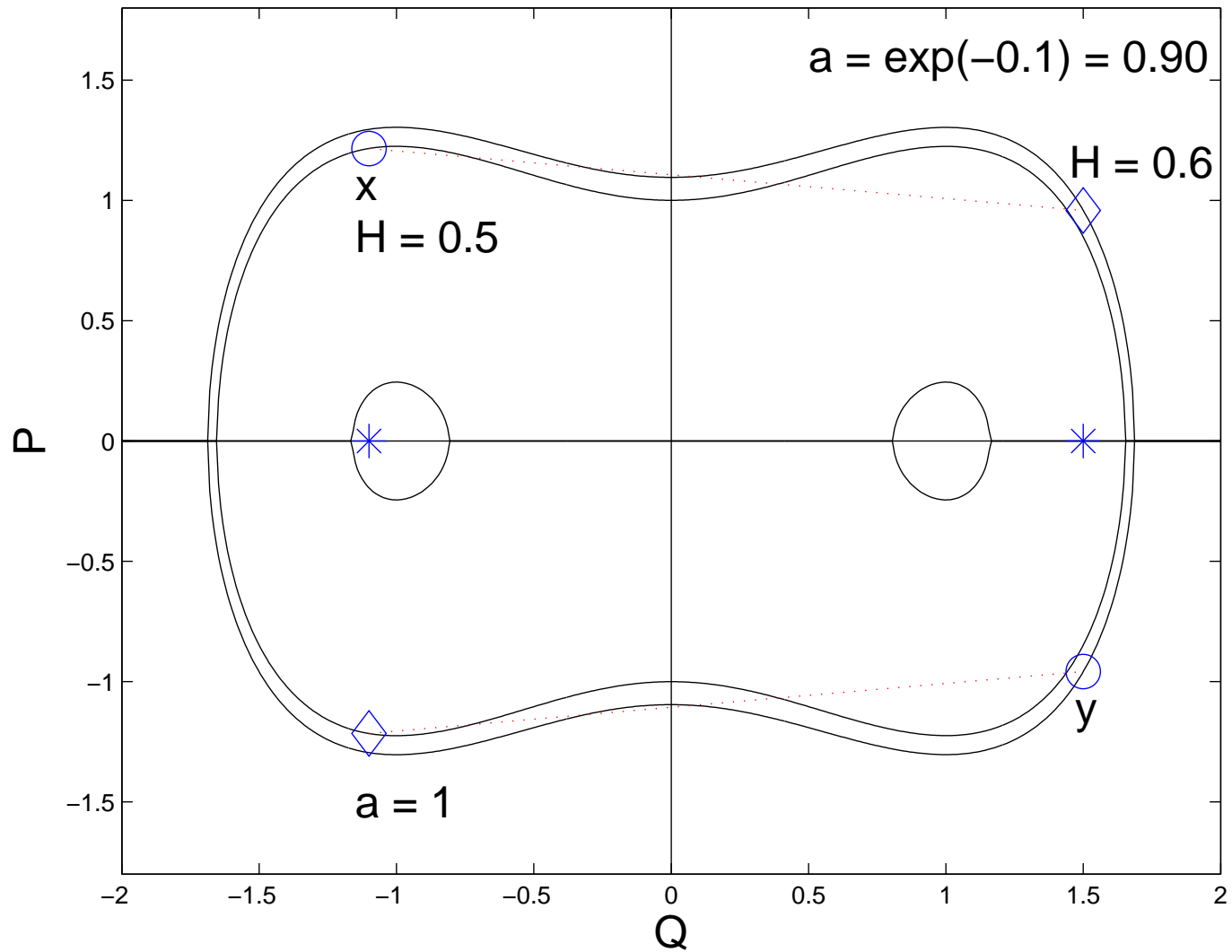
$$\Psi_h: (Q_0, P_0) \mapsto (Q_h, P_h)$$

which approximates the flow  $\Phi_h$ .

- Solution at time  $T$  is approximated by taking  $\lfloor \frac{T}{h} \rfloor$  Verlet steps:

$$(Q(T), P(T)) = \Phi_T((Q(0), P(0))) \approx \Psi_h^{\lfloor \frac{T}{h} \rfloor}((Q(0), P(0))).$$

- $\Psi_h^{(T)} = \Psi_h^{\lfloor \frac{T}{h} \rfloor}$  is *volume preserving and reversible* but does not conserve energy exactly and an accept/reject mechanism has to be introduced.
- HMC reads (for chosen  $h, T, \mathcal{M}$ ):
  - (i) Sample a momentum  $P \sim N(0, \mathcal{M})$ .
  - (ii) Accept proposed update  $Q'$  defined via  $(Q', P') = \Psi_h^{(T)}(Q, P)$  with probability:
 
$$a((Q, P), (Q', P')) := 1 \wedge \exp\{\mathcal{H}(Q, P) - \mathcal{H}(Q', P')\}.$$
- $\Pi$  is invariant density for chain.



## II. THE PROBLEM

- Wish to investigate behavior of MCMC when **dimension  $N$  is large** as in

1. Molecular dynamics —origin of HMC—.
  2. Bayesian estimation of a *function* (say the initial condition in a fluid dynamics problem). Discretization replaces the function by a vector in high dimensions. (A Stuart's talk.)
- To make headway consider the iid simplified scenario

$$\Pi(Q) = \exp\left\{-\sum_{i=1}^d V(q_i)\right\}, \quad V : \mathbb{R}^m \rightarrow \mathbb{R}, \quad N = m \times d.$$

RWM & MALA (Roberts, Gelman & Gilks (1997), Roberts & Rosenthal (1998)) ( $d = 1$ ,  $N = d$ .)

- RWM: under the scaling  $h = \ell/d$  the acceptance probability approaches, as  $d \rightarrow \infty$ , a nontrivial value

$$\mathbb{E}[a(Q, Q')] \rightarrow a_{RWM}(\ell) \in (0, 1).$$

- Requires  $\mathcal{O}(d)$  steps to make  $\mathcal{O}(1)$  moves in state space.
- The speed of exploration is optimal when  $a_{RWM}(\ell) = 0.234$ , regardless of the target density.

- MALA: under the scaling  $h = \ell/d^{1/3}$  the acceptance probability approaches, as  $d \rightarrow \infty$ , a nontrivial value

$$\mathbb{E}[a(Q, Q')] \rightarrow a_{MALA}(\ell) \in (0, 1).$$

- Requires  $\mathcal{O}(d^{1/3})$  steps to to make  $\mathcal{O}(1)$  moves in state space.
- The speed of exploration is optimal when  $a_{MALA}(\ell) = 0.574$ , regardless of the target density.

---

WHAT IS THE SITUATION FOR HMM?

### III. THE RESULT

Will prove here (in agreement with experiments/heuristics Gupta *et al.* (1990), Chen *et al.* (2000)):

- HMC: under the scaling  $h = \ell/d^{1/4}$ , as  $d \rightarrow \infty$ , the acceptance probability approaches a nontrivial value

$$\mathbb{E}[a(Q, Q')] \rightarrow a_{HMC}(\ell) \in (0, 1).$$

- Requires a  $\mathcal{O}(d^{1/4})$  Verlet steps to make  $\mathcal{O}(1)$  moves in state space.
- The speed of exploration is optimal when  $a_{HMC}(\ell) = 0.651$ , regardless of the target density.

### III. THE MATHS

## HMC in the iid scenario

- Write  $Q = (q_i)_{i=1}^d$   $P = (p_i)_{i=1}^d$ , with  $q_i, p_i$   $m$ -dimensional. Write

$$X = (x_i)_{i=1}^d; \quad x_i := (q_i, p_i) \in \mathbb{R}^{2m}.$$

(In general: small case letters for individual particles, large for aggregate.)

- The Hamiltonian is

$$\mathcal{H}(Q, P) = \sum_{i=1}^d H(q_i, p_i); \quad H(q, p) := \frac{1}{2} \langle p, M^{-1} p \rangle + V(q),$$

- The equations of motion for each of the  $d$  particles are

$$\frac{dq}{dt} = M^{-1} p, \quad \frac{dp}{dt} = -\nabla V(q).$$

- Particles evolve without coupling but are connected thru acceptance probability ( $T$  is henceforth fixed)

$$a(X, Y) = 1 \wedge \exp \left( \sum_{i=1}^d [H(x_i) - H(\psi_h^{(T)}(x_i))] \right)$$

( $\psi_h^{(T)}$  denotes Verlet solution operator for single particle).

- Need to estimate (in sense of analysis) the energy increment for single particle

$$\Delta(x, h) = H(\psi_h^{(T)}(x)) - H(\phi_T(x)) = H(\psi_h^{(T)}(x)) - H(x).$$

Smaller  $\Delta$  means higher  $a$ . (For true flow,  $\Delta \equiv 0$  and  $a = 1$ .)

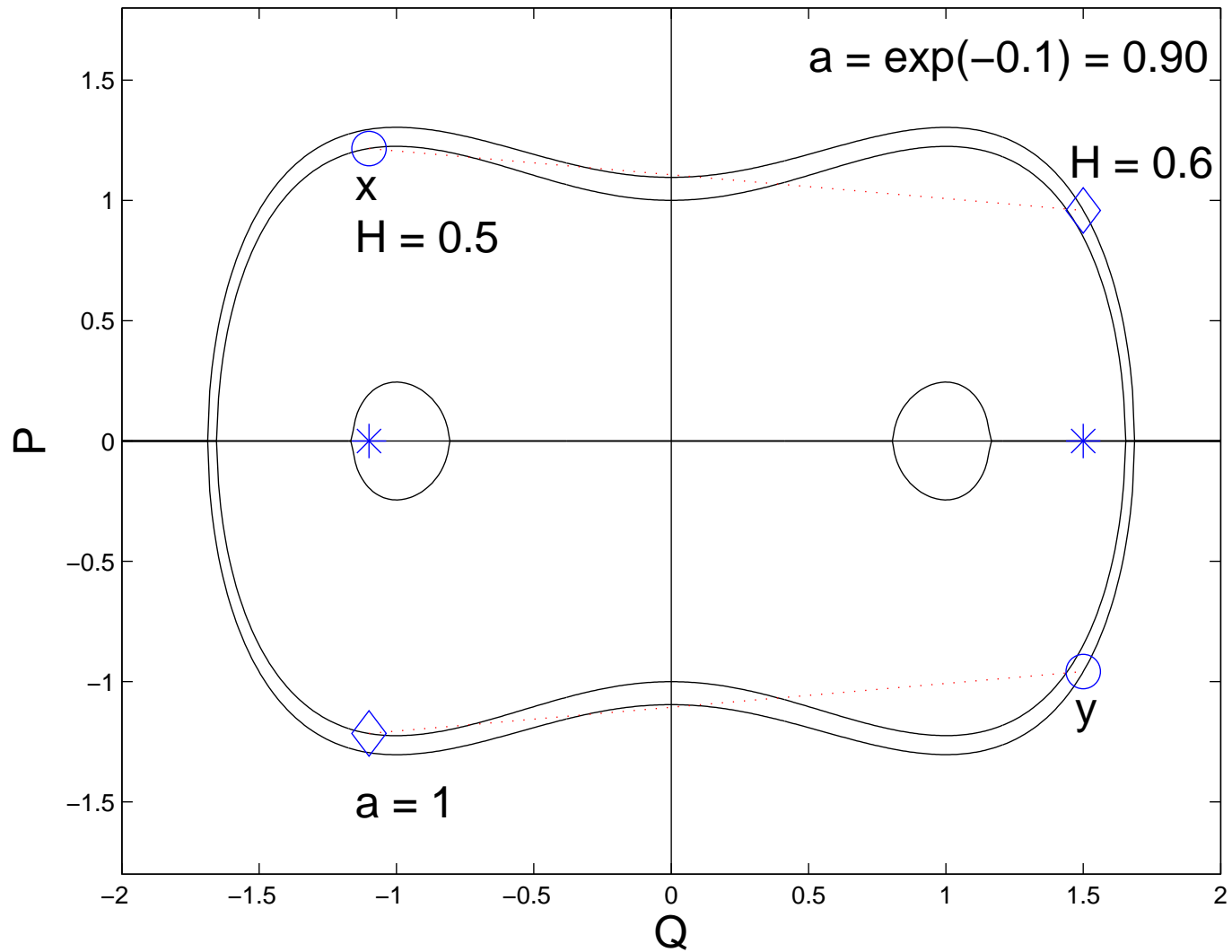
## Energy increments: pointwise

- For fixed  $x$ ,  $\Delta(x, h)$  is expected to be  $\mathcal{O}(h^2)$ . More precisely, under suitable hypotheses on the potential  $V$  (see technicalities later), the following condition holds:

C(1) There exist functions  $\alpha(x)$ ,  $\rho(x, h)$  such that

$$\Delta(x, h) = h^2\alpha(x) + h^2\rho(x, h)$$

with  $\lim_{h \rightarrow 0} \rho(x, h) = 0$ .



## Energy increments: averaged over phase space

- Due to the reversibility of the Verlet algorithm, to each point  $x$  in phase space, there corresponds a  $y$  such that  $\Delta(x, h) = -\Delta(y, h)$ . Hence  $\Delta$  averages to zero wrt Lebesgue measure.
- But here averages have to be taken wrt  $\exp(-H)$  and cancellation is only partial: If the expectation

$$\mu(h) = \mathbb{E}[\Delta(x, h)] = \int_{\mathbb{R}^{2m}} \Delta(x, h) e^{-H(x)} dx,$$

exists, then

$$2\mu(h) = \int_{\mathbb{R}^{2m}} \Delta(x, h) \left[ e^{-H(x)} - e^{-H(\psi_h^{(T)}(x))} \right] dx,$$

ie

$$2\mu(h) = \int_{\mathbb{R}^{2m}} \Delta(x, h) \left[ 1 - \exp(-\Delta(x, h)) \right] e^{-H(x)} dx.$$

- Since  $1 - \exp(-\Delta) \approx \Delta$  the integral should approximate the second moment of  $\Delta$ , which should behave as  $O(h^4)$ . To make this precise, consider:

C(2) There exists a function  $D : \mathbb{R}^{2m} \rightarrow \mathbb{R}$  such that

$$\sup_{0 \leq h \leq 1} \frac{|\Delta(x, h)|^2}{h^4} \leq D(x); \quad \int_{\mathbb{R}^{2m}} D(x) e^{-H(x)} dx < \infty.$$

Then . . .

THM: Under (C1)–(C2), the expectation and variance of  $\Delta$  satisfy

$$\lim_{h \rightarrow 0} \frac{\mu(h)}{h^4} = \mu, \quad \lim_{h \rightarrow 0} \frac{\sigma^2(h)}{h^4} = \Sigma,$$

for the constants (depending on target and numerical method):

$$\Sigma = \int_{\mathbb{R}^{2m}} \alpha^2(x) e^{-H(x)} dx, \quad \mu = \Sigma/2.$$

- To compute the acceptance probability  $a$  we have to add  $d$  independent variables with same distribution. Hence scale  $h \sim d^{-1/4}$  to get a nontrivial distributional limit via CLT.

## Expected acceptance probability

**THM** Assume (C1)–(C2) and  $h = \ell \cdot d^{-1/4}$ , for a constant  $\ell > 0$ . Then at stationarity, ie for  $X \sim \exp\{-\mathcal{H}\}$ ,

$$\lim_{d \rightarrow \infty} \mathbb{E}[a(X, Y)] = 2 \Phi(-\ell^2 \sqrt{\Sigma}/2) = a(\ell) \in (0, 1),$$

where  $\Sigma$  is the potential-dependent constant defined before. ( $\Phi$  standard normal distribution function.)

**Remark.** Proofs only use that Verlet is volume preserving and reversible and second-order accurate. So result holds for any other method with those properties. For volume preserving, reversible methods of order  $2\nu$  a similar result holds with  $h = \ell \cdot d^{-1/(2\nu)}$ .

## Technicalities

**THM:** Assume that the potential  $V$  is bounded from below and four times continuously differentiable, with bounded derivatives of orders 2, 3 and 4. If either

1.  $\nabla V$  is bounded and  $\int_{\mathbb{R}^m} |V(q)|^8 e^{-V(q)} dq < \infty$ .
2. There exist constants  $K_1, K_2 > 0$  and  $0 < \gamma \leq 1$  such that for all  $|q| \geq K_2$ ,  $V(q) \geq K_1 |q|^\gamma$ .

Then Conditions (C1)–(C2) hold for Verlet.

## Optimizing the algorithm

- $\text{eff} = \ell a(\ell)$  is a reasonable measure of the efficiency of the algorithm:
- **Proof in terms of computational work.** Idea: Work to compute a proposal  $\propto (T/h)d = Td^{5/4}/\ell$ . But expected number of proposals until (and including) the first accepted is  $\propto 1/a(\ell)$ . So expected work before acceptance  $\propto Td^{5/4}/[\ell a(\ell)]$ .
- **Alternative proof in terms of jumping distance.** (See written version of this talk.)

## Maximizing the efficiency

- Have to maximize  $\text{eff} = \ell a(\ell) = \ell^2 \Phi(-\ell^2 \sqrt{\Sigma}/2)$ , where  $\Sigma$  is unknown to user and depends on target distribution,  $h$ ,  $T$  &  $\mathcal{M}$ .
- Use  $a$  as independent variable:

$$\text{eff}(a) = \frac{\sqrt{2}}{\Sigma^{1/4}} \cdot a \cdot \left( \Phi^{-1}\left(1 - \frac{a}{2}\right) \right)^2.$$

Hence  $\Sigma^{1/4} \text{eff}(a)$  is distribution independent!

- The function  $a(\Phi^{-1}(1 - a/2))^{1/2}$ ,  $a \in [0, 1]$  is concave and vanishes at  $a = 0, 1$ . Its (unique) maximum reached at  $a \approx 0.651$
- Parameters tuned to reach acceptance rates close to that value.