THE HAMILTONIAN MONTE CARLO METHOD AND GEOMETRIC INTEGRATION

J. M. Sanz-Serna
Universidad Carlos III de Madrid

I. HANDLING PROBABILITY DISTRIBUTIONS IN PREHISTORIC TIMES

• Working with a probability distribution function $\rho : \mathbb{R} \to \mathbb{R}$ requires the knowledge of quantities such as its mean and variance

$$\mu = \int_{\mathbb{R}} x\rho(x)\,dx, \qquad \sigma^2 = \int_{\mathbb{R}} x^2\rho(x)\,dx - \mu^2,$$

or, for real a, the probability

$$\mathbb{P}\Big((-\infty, a)\Big) = \int_{-\infty}^{a} \rho(x)\,dx = \int_{\mathbb{R}} 1_{(-\infty,a)}(x)\,\rho(x)\,dx.$$

In short, the knowledge of integrals/expectations

$$\int_{\mathbb{R}} F(x)\rho(x)\,dx$$

for relevant real-valued functions F.

• Before computers were widely available, these integrals were calculated once and for all (in general by numerical quadrature). . .

• . . . and then compiled in tables such as . . .

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

• The use of tables has clear limitations. Among others:

+ It makes you to restrict your attention to a small bunch of distributions (standard normal, Student, χ^2, ...)

Such a small bunch may be sufficient to solve many typical statistical problems in a frequentist framework, but not in a Bayesian framework. In Bayesian statistics the posterior distribution changes with the available data. Bayesian statistics only took off once the 'tabulating' approach was superseded.

+ The class of integrands F is also restricted (in the example above to $1_{(-\infty,a)}$ for 11 values of a).

• In the multivariate case where $x \in \mathbb{R}^d$, $d > 1$ the interest is computing quantities

$$\int_{\mathbb{R}^d} F(x)\rho(x)\,dx,$$

(for instance

$$\mathbb{P}(\Omega) = \int_{\mathbb{R}^d} 1_\Omega(x)\,\rho(x)\,dx\ ).$$

These integrals/expectations cannot really be tabulated.

• Worse than that: unless d is small, those integrals cannot even be computed accurately by conventional cubature rules. [A (tensor product) rule with three nodes in each variate requires almost six thousand functions evaluations in $\mathbb{R}^{10}$ and in excess of three billion function evaluations in $\mathbb{R}^{20}$.] [The curse of dimensionality (Bellman 1957).]

II. AN ALTERNATIVE USING COMPUTERS: MONTE CARLO

• Under fairly general assumptions, the law of large numbers shows that the integral we wish to compute, i.e.

$$\int_{\mathbb{R}^d} F(x)\rho(x)\,dx,$$

is the (almost sure) limit of the random sequence

$$\frac{1}{N}\Big(F(X_1)+\cdots+F(X_N)\Big),$$

where the X_n are independent random variables each with pdf ρ.

• This suggests the (naive) Monte Carlo quadrature rule

$$\int_{\mathbb{R}^d} F(x)\rho(x)\,dx \simeq \frac{1}{N}\Big(F(x_1)+\cdots+F(x_N)\Big),$$

where x_n are independent draws of a random variable with pdf ρ. [Note equal weights.]

• Archetypal example: ρ uniform in $[0,1] \times [0,1]$, $F(x) = 1_\Omega(x)$, $\Omega \subset [0,1] \times [0,1]$,

$$\text{Area}(\Omega) \simeq \frac{1}{N} \#\{x_n \in \Omega\}.$$

• Error bounds for the naive rule are $\mathcal{O}(1/\sqrt{N})$ (computers needed!).

• The good news: bounds are independent of dimension d and regularity of F.

• The fly in the ointment: for most probability distributions, generating independent draws is not feasible.

• An alternative. The law of large numbers that underpins the formula

$$\int_{\mathbb{R}^d} F(x)\rho(x)\,dx \simeq \frac{1}{N}\Big(F(x_1)+\cdots+F(x_N)\Big),$$

also holds if the random variables X_n are not assumed to be independent, but form a Markov chain for which the pdf ρ is invariant. [This roughly means that X_{n+1} depends on X_n but in such a way that, if X_n has pdf ρ, so does X_{n+1}.]

• Metropolis, Rosenbluth, Rosenbluth, Teller and Teller showed in 1953 that it is always possible to construct a suitable Markov chain for which realizations x_n of the random variables X_n may be easily generated in a computer.

The Random Walk Metropolis algorithm. Choose a value $h > 0$. Once X_n has been defined $(n = 0, 1, \dots)$:

- Define the proposal $X^\star_{n+1} = X_n + hZ_n$ where Z_n is standard normal (and independent from past). [Hence the name random walk.]

- [Accept/reject mechanism.] Define $U_n \sim \mathcal{U}[0, 1]$ and then
  - If $a(X_n) = \rho(X^\star_{n+1})/\rho(X_n) \geq U_n$, set $X_{n+1} = X^\star_{n+1}$ (the proposal has been accepted). [a is the acceptance probability.]
  - Else, set $X_{n+1} = X_n$ (the proposal has been rejected).

• The algorithm will compute expectations for (almost) arbitrary $F(x)$ and $\rho(x)$ in any number of dimensions d.

• If the correlation between the random variables X_n increases, the number of samples N to achieve a target accuracy of the quadrature rule has to be increased.

• A large value of h typically leads to many rejections (particularly so in high dimensions) and therefore to large correlations because, when the proposal is rejected $X_{n+1} = X_n$.

• A small value of h results in the proposal $X^{\star}_{n+1} = X_n + hZ_n$ being near X_n, thus increasing the probability of acceptance, but then the correlations in the chain are also high.

• Roberts, Gelman and Gilks 1997: when the target is a product of d identically distributed components. . .

+ h has to be chosen proportional to $1/d$.

+ Algorithm needs $\mathcal{O}(d^2)$ work to make $\mathcal{O}(1)$ moves in the state space $\mathbb{R}^d$.

+ The exploration of state space is optimal when the acceptance rate is $0.234\ldots$, regardless of the specific distribution being sampled.

• Idea to improve the algorithm: use proposals that avail themselves of information on the target pdf ρ.

• The Hybrid Monte Carlo (HMC) method is one (among many others) algorithm based on that idea.

• Introduced in the Physics literature by Duane, Kenney, Pendleton and Roweth 1987.

• Neal made it known to the Statistics community, where the acronym HMC is now read as Hamiltonian Monte Carlo and the algorithm and its variants are extremely popular.

• Ideally HMC offers the possibility of proposals that are far away from the current state and yet are accepted with high probability.

III. A FIRST SMALL DETOUR: STATISTICAL PHYSICS

• For a conservative mechanical system, Newton's second law reads

$$M\ddot{q} = -\nabla V(q),$$

($q \in \mathbb{R}^d$ collects the positions, d is the number of degrees of freedom, the matrix M contains the masses and V is the potential energy).

• As t varies, the total energy $(1/2)\dot{q}(t)^T M\dot{q}(t) + V(q(t))$ is conserved.

• Now assume that the system, rather than being isolated from the environment, is inside a heat bath at constant (absolute) temperature $1/\beta$. (Think of a protein inside the human body.) Molecules of the heat bath hit the system and interchange energy with it.

• Keeping track of all interchanges is impossible and a statistical description is needed. (Maxwell, Boltzmann, Gibbs,. . .)

• Statistical mechanics uses the Hamiltonian formulation of mechanics. This introduces a new independent variable $p = M\dot{q}$ (momentum). The space $\mathbb{R}^d \times \mathbb{R}^d$ of pairs (q, p) is the phase space.

Newton's law is rewritten as the first-order system

$$\dot{q} = M^{-1}p, \qquad \dot{p} = -\nabla V(q)$$

i.e. in the symmetric form

$$\dot{q} = \frac{\partial H}{\partial p}, \qquad \dot{p} = -\frac{\partial H}{\partial q},$$

where $H(q, p) = (1/2)p^T M^{-1}p + V(q)$ is the total energy of the system expressed as a function of q and p.

• In a heat bath $q(t), p(t)$ evolve so as to preserve the canonical probability measure in phase space: $d\mu = (1/Z)\exp(-\beta H(q, p))\, dqdp$, where Z is the normalizing constant $\int_{\mathbb{R}^d \times \mathbb{R}^d} \exp(-\beta H) dqdp$.

• In view of the product structure

$$\exp(-\beta H(q,p)) = \exp\Big(-\beta(1/2)p^T M^{-1} p\Big) \times \exp\Big(-\beta V(q)\Big),$$

q and p are stochastically independent.

• The momenta have a Gaussian density (proportional to)

$$\exp(-\beta(1/2)p^T M^{-1} p)$$

(Maxwell's distribution). From here it follows that the average kinetic energy is $1/(2\beta) \times d$: the absolute temperature $1/\beta$ is twice the average kinetic energy per degree of freedom.

• The positions q have the Boltzmann density $\propto \exp(-\beta V(q))$: minima of the potential energy are modes of the probability. As the temperature diminishes those minima carry more and more probability.

IIII. A SECOND SMALL DETOUR: SYMMETRIES OF THE HAMILTONIAN DYNAMICS.

• For the Hamiltonian system

$$\dot{q} = \frac{\partial H}{\partial p}, \qquad \dot{p} = -\frac{\partial H}{\partial q},$$

with arbitrary H, denote by $\varphi_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ the solution flow, i.e. $\varphi_t(q,p)$ is the value at time t of the solution with initial values (q,p) at the initial time $t = 0$. The flow has important geometric properties.

• For each t the flow preserves volume in phase space (Liouville): $\forall \Omega \subset \mathbb{R}^d \times \mathbb{R}^d$, $\varphi_t(\Omega)$ has the same $2d$-dimensional measure as Ω. [In fact, the flow has a stronger property: *symplecticness* (Poincaré).]

• The flow preserves energy: $H(\varphi_t(q,p)) = H(q,p)$.

• For the special form $H(q,p) = (1/2)p^T M^{-1} p + V(q)$ we found above, the flow is reversible: if $\varphi_t(q,p) = (q^\star, p^\star)$, then $\varphi_t(q^\star, -p^\star) = (q, -p)$.

• As a consequence, the flow preserves the canonical probability measure [$d\mu \propto \exp(-\beta H(q,p))\, dqdp$]: i.e. $\forall \Omega \subset \mathbb{R}^d \times \mathbb{R}^d$, $\varphi_t(\Omega)$ carries the same probability as Ω. [But note that the Hamiltonian dynamics does not describe the motions of the system in the heat bath.]

• We are now ready to leave the detours and go back to our task: given a target probability distribution with density $\rho(x)$ in $\mathbb{R}^d$ construct a Markov chain that has it as an invariant distribution.

V. THE IDEA BEHIND HAMILTONIAN MONTE CARLO SAMPLING

A Markov chain that does the job. [$T > 0$ is a parameter.]

- Write q instead of x and $\rho(q) = \exp(-V(q))$.

- In the phase space of the variable (q, p) consider the Hamiltonian system associated with $H = (1/2)p^T p + V(q)$ and the solution flow φ_T.

- If Q_n is an element of the chain, then Q_{n+1} is defined as follows.

  + Generate P_n from pdf $\propto \exp(-(1/2)p^T p))$, independent from Q_n (and from past).

  + Define $(Q_{n+1}, \widetilde{P}_{n+1}) = \varphi_T(Q_n, P_n)$ ($\widetilde{P}_{n+1}$ is discarded).

- Proof: the Hamiltonian flow φ_T preserves canonical probability measure $d\mu \propto \exp(-(1/2)p^T p - V(q))dqdp$.

• Good news: by suitably choosing T, Q_{n+1} may be far away from Q_n (implications: low correlation, chain explores quickly $\mathbb{R}^d$).

• Bad news: φ_T only known in trivial cases.

• Good idea: use a numerical approximation Ψ to φ_T, i.e. at each step of the Markov chain, integrate numerically the Hamiltonian dynamics with step-length h in the interval $0 \leq t \leq T$.

• Additional bad news: No numerical integrator simultaneously preserves volume and energy (Ge and Mardsden 1988). Thus no Ψ preserves the canonical distribution μ.

The construction of integrators that preserve symmetries of the system being integrated is the aim of Geometric Integration (SS 1997).

There exist explicit integrators that preserve volume and are reversible.

• Additional good idea: Introduce an accept/reject mechanism so as to enforce exact conservation of μ.

We have finally arrived at HMC:

VI. HAMILTONIAN MONTE CARLO SAMPLING

- If Q_n is an element of the Markov chain, then Q_{n+1} is defined as follows.

  + Generate P_n from pdf $\propto \exp(-(1/2)p^T p))$, independent from Q_n (and from past).

  + Find $(Q^\star_{n+1}, \widetilde{P}_{n+1})$ by integrating numerically, from the initial condition (Q_n, P_n), the Hamiltonian dynamics. The integrator must be volume-preserving and reversible.

  + Accept the proposal $Q^\star_{n+1}$ with probability

$$\min\Big(1, \exp\big(-[H(Q^\star_{n+1}, \widetilde{P}_{n+1}) - H(Q_n, P_n)]\big)\Big).$$

  (Upon acceptance $Q_{n+1} = Q^\star_{n+1}$, upon rejection $Q_{n+1} = Q_n$; on both cases $\widetilde{P}_{n+1}$ is discarded.)

• Since $H(\varphi_T(Q_n, P_n)) = H(Q_n, P_n)$,

$$H(Q^\star_{n+1}, \widetilde{P}_{n+1}) - H(Q_n, P_n) = H(Q^\star_{n+1}, \widetilde{P}_{n+1}) - H(\varphi_t(Q_n, P_n))$$

is the energy error in the integration. Hence, as the time-step approaches 0 with T fixed, the acceptance probability approaches 100%. (But the integration becomes more expensive.)

• The algorithm thus provides the possibility of generating proposals away from the current state (by suitably choosing T) that may be accepted with high probability (by suitably reducing the step-size).

• The Störmer/Verlet/leapfrog algorithm is the used in practice, but better alternatives exist (particularly for large problems) (joint work with Blanes, Casas, Akhmatskaya, etc.).

• In the product scenario, Beskos, Pillai, Roberts, SS and Stuart 2013 show (for integrators that are second order accurate):

+ h has to be chosen proportional to $1/d^{1/4}$.

+ Algorithm needs $\mathcal{O}(d^{5/4})$ work to make $\mathcal{O}(1)$ moves in the state space $\mathbb{R}^d$.

+ The exploration of state space is optimal when the acceptance rate is $0.651\ldots$, regardless of the specific distribution being sampled.