Sampling high-dimensional probability distributions & Bayesian learning

Gabriel STOLTZ

gabriel.stoltz@enpc.fr

(CERMICS, Ecole des Ponts & MATHERIALS team, INRIA Paris)

UM6P research school, November 2019

Outline

- **Examples of high-dimensional probability measures**
  - Statistical physics
  - Bayesian inference
- **Markov chain methods**
  - Metropolis–Hastings algorithm
  - Hybrid Monte Carlo and its variants
- **Methods based on stochastic differential equations**
  - An introduction to SDEs (generators, invariant measure, discretization, etc)
  - Langevin-like dynamics
- **Variance reduction techniques**
- **Large scale Bayesian inference**
  - Mini-batching
  - Adaptive Langevin dynamics

General references (1)

- Computational Statistical Physics
  - D. Frenkel and B. Smit, *Understanding Molecular Simulation, From Algorithms to Applications* (Academic Press, 2002)
  - M. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation* (Oxford, 2010)
  - M. P. Allen and D. J. Tildesley, *Computer simulation of liquids* (Oxford University Press, 1987)
  - D. C. Rapaport, *The Art of Molecular Dynamics Simulations* (Cambridge University Press, 1995)
  - T. Schlick, *Molecular Modeling and Simulation* (Springer, 2002)
- Computational Statistics [my personal references... many more out there!]
  - J. Liu, *Monte Carlo strategies in scientific computing*, Springer, 2008
  - W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov chain Monte Carlo in practice* (Chapman & Hall, 1996)
- Machine learning and sampling
  - C. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006)

General references (2)

- Sampling the canonical measure
  - L. Rey-Bellet, Ergodic properties of Markov processes, *Lecture Notes in Mathematics*, **1881** 1–39 (2006)
  - E. Cancès, F. Legoll and G. Stoltz, Theoretical and numerical comparison of some sampling methods, *Math. Model. Numer. Anal.* **41**(2) (2007) 351-390
  - T. Lelièvre, M. Rousset and G. Stoltz, *Free Energy Computations: A Mathematical Perspective* (Imperial College Press, 2010)
  - B. Leimkuhler and C. Matthews, *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods* (Springer, 2015).
  - T. Lelièvre and G. Stoltz, Partial differential equations and stochastic methods in molecular dynamics, *Acta Numerica* **25**, 681-880 (2016)

- Convergence of Markov chains
  - S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability* (Cambridge University Press, 2009)
  - R. Douc, E. Moulines, P. Priouret and P. Soulier, *Markov chains* (Springer, 2018)

Sampling high-dimensional probability measures

Statistical physics (1)

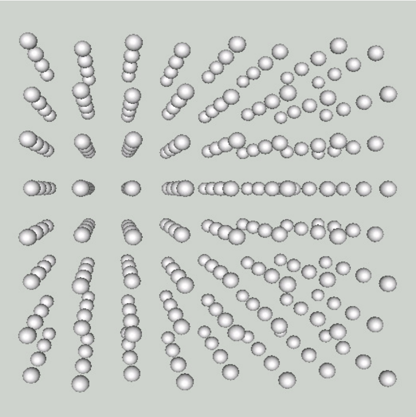
- **Aims of computational statistical physics**
  - numerical microscope
  - computation of average properties, static or dynamic

- **Orders of magnitude**
  - distances $\sim 1\ \mathring{A} = 10^{-10}$ m
  - energy per particle $\sim k_{\mathrm{B}}T \sim 4 \times 10^{-21}$ J at room temperature
  - atomic masses $\sim 10^{-26}$ kg
  - time $\sim 10^{-15}$ s
  - number of particles $\sim \mathcal{N}_A = 6.02 \times 10^{23}$

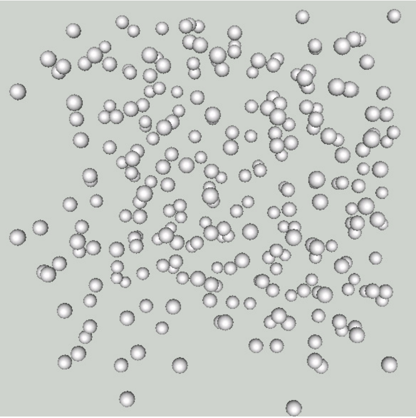
- **"Standard" simulations**
  - 10^6 particles ["world records": around 10^9 particles]
  - integration time: (fraction of) ns ["world records": (fraction of) μs]

Statistical physics (2)

What is the melting temperature of argon?



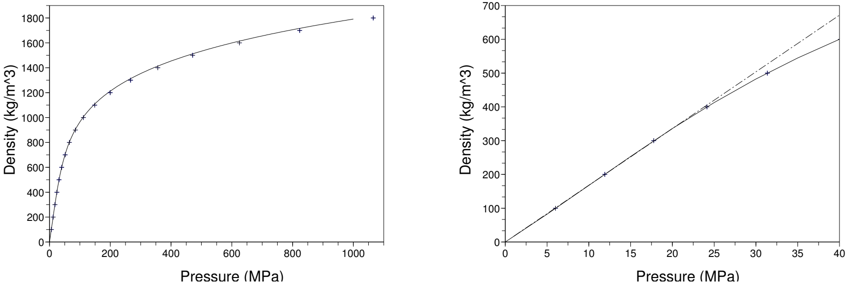
(a) Solid argon (low temperature)



(b) Liquid argon (high temperature)

Statistical physics (3)

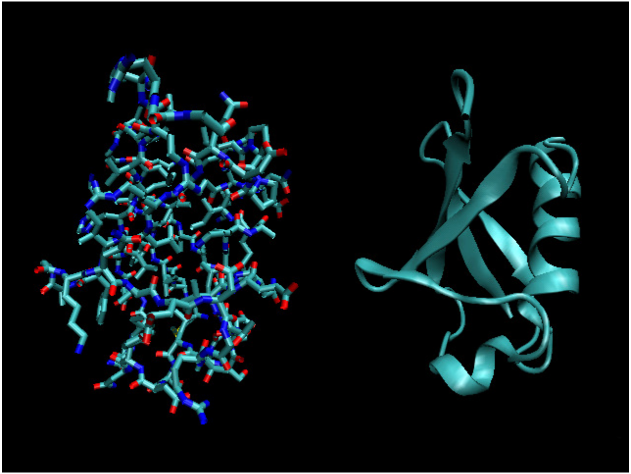
"Given the structure and the laws of interaction of the particles, what are the macroscopic properties of the matter composed of these particles?"



Equation of state (pressure/density diagram) for argon at $T = 300$ K

Statistical physics (4)

What is the structure of the protein? What are its typical conformations, and what are the transition pathways from one conformation to another?



Statistical physics (5)

• Microstate of a classical system of N particles:

$$(q,p) = (q_1,\ldots,q_N,\ p_1,\ldots,p_N) \in \mathcal{E}$$

Positions q (configuration), momenta p (to be thought of as $M\dot{q}$)

• In the simplest cases, $\mathcal{E} = \mathcal{D} \times \mathbb{R}^{3N}$ with $\mathcal{D} = \mathbb{R}^{3N}$ or $\mathbb{T}^{3N}$

• More complicated situations can be considered: molecular constraints defining submanifolds of the phase space

• Hamiltonian $H(q,p) = E_{\mathrm{kin}}(p) + V(q)$, where the kinetic energy is

$$E_{\mathrm{kin}}(p) = \frac{1}{2}\, p^T M^{-1} p, \qquad M = \begin{pmatrix} m_1 \,\mathrm{Id}_3 & & 0 \\ & \ddots & \\ 0 & & m_N \,\mathrm{Id}_3 \end{pmatrix}.$$

Statistical physics (6)

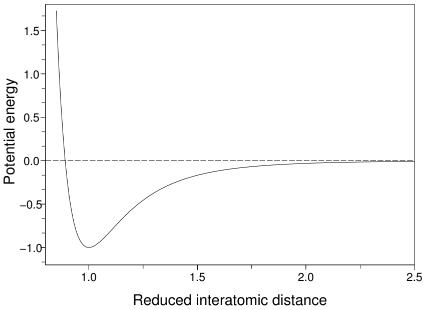
- All the physics is contained in V
  - ideally derived from quantum mechanical computations
  - in practice, empirical potentials for large scale calculations

- An example: Lennard-Jones pair interactions to describe noble gases

$$V(q_1,\dots,q_N) = \sum_{1\leqslant i<j\leqslant N} v(|q_j - q_i|)$$

$$v(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

$$\text{Argon:}\quad \begin{cases} \sigma = 3.405\times 10^{-10}\ \mathrm{m} \\ \varepsilon/k_{\mathrm{B}} = 119.8\ \ \mathrm{K} \end{cases}$$



Statistical physics (7)

- Macrostate of the system described by a probability measure

Equilibrium thermodynamic properties (pressure,...)

$$\langle \varphi \rangle_\mu = \mathbb{E}_\mu(\varphi) = \int_{\mathcal{E}} \varphi(q,p)\, \mu(dq\, dp)$$

- Choice of thermodynamic ensemble
  - least biased measure compatible with the observed macroscopic data
  - Volume, energy, number of particles, ... fixed exactly or in average
  - Equivalence of ensembles (as $N \to +\infty$)

- Canonical ensemble = measure on (q,p), average energy fixed H

$$\mu_{\mathrm{NVT}}(dq\, dp) = Z_{\mathrm{NVT}}^{-1}\, \mathrm{e}^{-\beta H(q,p)}\, dq\, dp$$

with $\beta = \dfrac{1}{k_{\mathrm{B}}T}$ the Lagrange multiplier of the constraint $\displaystyle\int_{\mathcal{E}} H\, \rho\, dq\, dp = E_0$

Bayesian inference (1)

- Data set $\{y_i\}_{i=1,\dots,N_{\text{data}}}$
- Elementary likelihood $P(y|q)$, with q parameters of probability measure
- A priori distribution of the parameters p_{prior} (usually not so informative)

Aim

Find the values of the parameters q describing correctly the data: sample

$$\nu(q) \propto p_{\text{prior}}(q) \prod_{i=1}^{N_{\text{data}}} P(y_i|q)$$

- Example of Gaussian mixture model

Bayesian inference (2)

- Elementary likelihood approximated by mixture of K Gaussians

$$P(y\,|\,\theta) = \sum_{k=1}^{K} a_k \sqrt{\frac{\lambda_k}{2\pi}} \exp\left(-\frac{\lambda_k}{2}(y-\mu_k)^2\right)$$

- Parameters $\theta = (a_1, \dots, a_{K-1}, \mu_1, \dots, \mu_K, \lambda_1, \dots, \lambda_K)$ with

$$\mu_k \in \mathbb{R}, \quad \lambda_k \geqslant 0, \quad 0 \leqslant a_k \leqslant 1, \quad a_1 + \cdots + a_K = 1$$

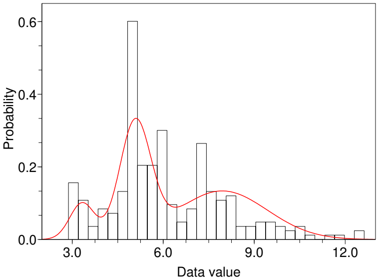
- Prior distribution: Random beta model: additional variable
  - uniform distribution of the weights a_k
  - $\mu_k \sim \mathcal{N}\left(M, R^2/4\right)$ with M = mean of data, $R = \max y_i - \min y_i$
  - $\lambda_k \sim \Gamma(\alpha, \beta)$ with $\beta \sim \Gamma(g, h)$, $g = 0.2$ and $h = 100g/\alpha R^2$

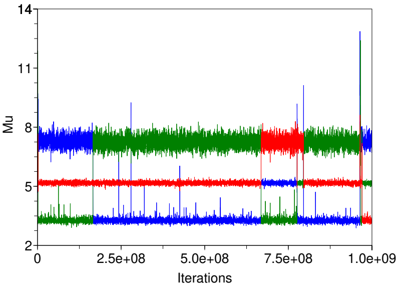
Aim

Find the values of the parameters (namely θ, and possibly K as well) describing correctly the data

[RG97] S. Richardson and P. J. Green. *J. Roy. Stat. Soc. B*, 1997.
[JHS05] A. Jasra, C. Holmes and D. Stephens, Statist. Science, 2005

Bayesian inference (3)



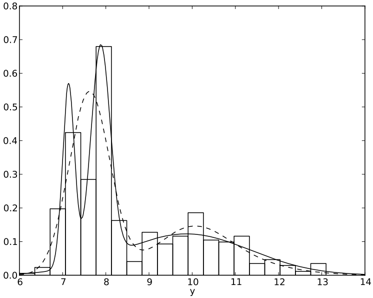


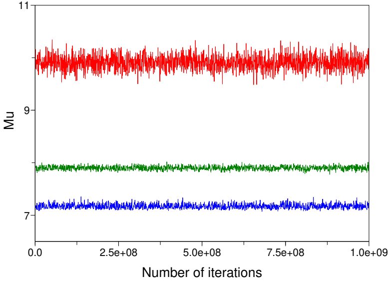
Left: Lengths of snappers ($N_{\mathrm{data}} = 256$), and a possible fit for $K = 3$ using the last configuration from the trajectory plotted in the right picture.

Right: Typical sampling trajectory, Metropolis/Gaussian random walk with $(\sigma_q, \sigma_\mu, \sigma_v, \sigma_\beta) = (0.0005, 0.025, 0.05, 0.005)$.

[IS88] A. J. Izenman and C. J. Sommer, *J. Am. Stat. Assoc.*, 1988.
[BMY97] K. Basford *et al.*, *J. Appl. Stat.*, 1997

Bayesian inference (4)



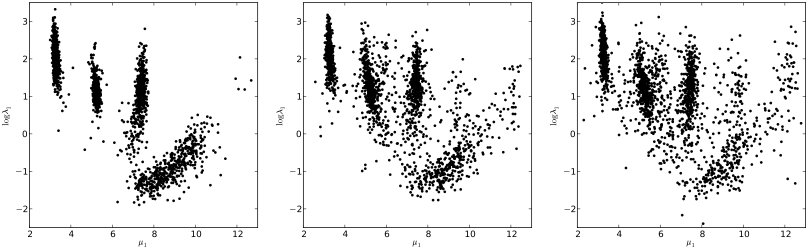


Left: Thickness of Mexican stamps ("Hidalgo stamp data", $N_{\rm data} = 485$), and two possible fits for $K = 3$ ("genuine multimodality", solid line: dominant mode).

Right: Typical sampling trajectory

[TSM86] D. Titterington *et al.*, *Statistical Analysis of Finite Mixture Distributions*, 1986.
[FS06] S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models*, 2006.

Bayesian inference (5)



Scatter plot of the marginal distribution of $(\mu_1, \log \lambda_1)$ for the Fish data, for various values of $K = 4, 5, 6$

Standard techniques to sample probability measures (1)

- The basis is the generation of numbers uniformly distributed in $[0,1]$
- Deterministic sequences which look like they are random...
  - Early methods: linear congruential generators ("chaotic" sequences)

$$x_{n+1} = ax_n + b \mod c, \qquad u_n = \frac{x_n}{c-1}$$

  - Known defects: short periods, point alignments, etc, which can be (partially) patched by cleverly combining several generators

- More recent algorithms: shift registers, such as Mersenne-Twister

$\rightarrow$ defaut choice in *e.g.* Scilab, available in the GNU Scientific Library

- Randomness tests: various flavors

Standard techniques to sample probability measures (2)

- Classical distributions are obtained from the uniform distribution by...
  - inversion of the cumulative function $F(x) = \int_{-\infty}^{x} f(y)\,dy$ (which is an increasing function from $\mathbb{R}$ to $[0,1]$)

$$X = F^{-1}(U) \sim f(x)\,dx$$

    Proof: $\mathbb{P}\{a < X \leqslant b\} = \mathbb{P}\{a < F^{-1}(X) \leqslant b\} = \mathbb{P}\{F(a) < U \leqslant F(b)\} = F(b) - F(a) = \int_a^b f(x)\,dx$

    Example: exponential law of density $\lambda \mathrm{e}^{-\lambda x}\mathbf{1}_{\{x \geqslant 0\}}$, $F(x) = \mathbf{1}_{\{x \geqslant 0\}}(1 - \mathrm{e}^{-\lambda x})$, so that $X = -\frac{1}{\lambda}\ln U$

  - change of variables: standard Gaussian $G = \sqrt{-2\ln U_1}\cos(2\pi U_2)$

    Proof: $\mathbb{E}(f(X,Y)) = \frac{1}{2\pi}\int_{\mathbb{R}^2} f(x,y)\,\mathrm{e}^{-(x^2+y^2)/2}\,dx\,dy = \int_0^{+\infty} f\left(\sqrt{r}\cos\theta, \sqrt{r}\sin\theta\right)\frac{1}{2}\mathrm{e}^{-r/2}\,dr\,\frac{d\theta}{2\pi}$

  - using the rejection method

    Find a probability density g and a constant $c \geqslant 1$ such that $0 \leqslant f(x) \leqslant cg(x)$. Generate i.i.d. variables $(X^n, U^n) \sim g(x)\,dx \otimes \mathcal{U}[0,1]$, compute $r^n = \frac{f(X^n)}{cg(X^n)}$, and accept X^n if $r^n \geqslant U^n$

Standard techniques to sample probability measures (3)

- The previous methods work only
  - for low-dimensional probability measures
  - when the normalization constants of the probability density are known

- In more complex cases, one needs to resort to trajectory averages

Ergodic methods

$$\frac{1}{N_{\text{iter}}} \sum_{n=1}^{N_{\text{iter}}} \varphi(x^n) \xrightarrow[N_{\text{iter}} \to +\infty]{} \int \varphi \, d\mu$$

- **Find methods for which**
  - the convergence is guaranteed? (and in which sense?)
  - error estimates are available? (typically with Central Limit Theorem)

Standard techniques to sample probability measures (4)

- Assume that $x^n \sim \pi$ are idependently and identically distributed (i.i.d.)

Law of Large Numbers for $\varphi \in L^1(\pi)$

$$S_{N_{\rm iter}} = \frac{1}{N_{\rm iter}} \sum_{n=1}^{N_{\rm iter}} \varphi(x^n) \xrightarrow[N_{\rm iter}\to+\infty]{} \mathbb{E}_\pi(\varphi) = \int_{\mathcal{X}} \varphi \, d\pi \quad \text{almost surely}$$

Central Limit Theorem for $\varphi \in L^2(\pi)$

$$\sqrt{N_{\rm iter}} \left(S_{N_{\rm iter}} - \int \varphi \, d\pi \right) \xrightarrow[N_{\rm iter}\to+\infty]{\rm law} \mathcal{N}(0,\sigma_\varphi^2), \;\; \sigma_\varphi^2 = \int_{\mathcal{X}} \left[\varphi - \mathbb{E}_\pi(\varphi)\right]^2 \, d\pi$$

- This should be thought of in practice as $S_{N_{\rm iter}} \simeq \mathbb{E}_\pi(\varphi) + \dfrac{\sigma_\varphi}{\sqrt{N_{\rm iter}}}\mathcal{G}$

Outline

- **Examples of high-dimensional probability measures**
  - Statistical physics
  - Bayesian inference
- **Markov chain methods**
  - Metropolis–Hastings algorithm
  - Hybrid Monte Carlo and its variants
- **Methods based on stochastic differential equations**
  - An introduction to SDEs (generators, invariant measure, discretization, etc)
  - Langevin-like dynamics
- **Variance reduction techniques**
- **Large scale Bayesian inference**
  - Mini-batching
  - Adaptive Langevin dynamics

Metropolis–Hastings algorithms

Metropolis-Hastings algorithm (1)

- Markov chain method[1,2], on position space

  - Given q^n, propose $\tilde{q}^{n+1}$ according to transition probability $T(q^n, \tilde{q})$
  - Accept the proposition with probability $\min\left(1, r(q^n, \widetilde{q}^{n+1}\right)$ where

$$r(q, q') = \frac{T(q', q)\,\nu(q')}{T(q, q')\,\nu(q)}, \qquad \nu(dq) \propto \mathrm{e}^{-\beta V(q)}.$$

  If acception, set $q^{n+1} = \tilde{q}^{n+1}$; otherwise, set $q^{n+1} = q^n$.

- Example of proposals
  - Gaussian displacement $\tilde{q}^{n+1} = q^n + \sigma\, G^n$ with $G^n \sim \mathcal{N}(0, \mathrm{Id})$
  - Biased random walk[3,4] $\tilde{q}^{n+1} = q^n - \alpha \nabla V(q^n) + \sqrt{\frac{2\alpha}{\beta}}\, G^n$

[1] Metropolis, Rosenbluth ($\times 2$), Teller ($\times 2$), *J. Chem. Phys.* (1953)
[2] W. K. Hastings, *Biometrika* (1970)
[3] G. Roberts and R.L. Tweedie, *Bernoulli* (1996)
[4] P.J. Rossky, J.D. Doll and H.L. Friedman, *J. Chem. Phys.* (1978)

Metropolis-Hastings algorithm (2)

- The normalization constant in the canonical measure needs not be known
- Transition kernel: accepted moves + rejection

$$P(q,dq') = \min\Big(1, r(q,q')\Big) T(q,q')\, dq' + \Big(1-\alpha(q)\Big)\delta_q(dq'),$$

where $\alpha(q) \in [0,1]$ is the probability to accept a move starting from q:

$$\alpha(q) = \int_{\mathcal{D}} \min\Big(1, r(q,q')\Big) T(q,q')\, dq'.$$

- Rejection rate $1-\alpha(q) \sim \sqrt{\sigma}$ for RWMH, and $\alpha^{3/2}$ for MALA
- The canonical measure is reversible with respect to ν

$$P(q,dq')\nu(dq) = P(q',dq)\nu(dq')$$

This implies invariance: $\displaystyle\int_{\mathcal{D}}\int_{\mathcal{D}} \varphi(q')P(q,dq')\,\nu(dq) = \int_{\mathcal{D}} \varphi(q)\,\nu(dq)$

Metropolis-Hastings algorithm (3)

- Proof: Detailed balance on the absolutely continuous parts

$$\begin{aligned}\min\left(1, r(q,q')\right) T(q,dq')\nu(dq) &= \min\left(1, r(q',q)\right) r(q,q') T(q,dq')\nu(dq) \\ &= \min\left(1, r(q',q)\right) T(q',dq)\nu(dq')\end{aligned}$$

using successively $\min(1,r) = r \min\left(1, \frac{1}{r}\right)$ and $r(q,q') = \dfrac{1}{r(q',q)}$

- Equality on the singular parts $(1-\alpha(q))\,\delta_q(dq')\nu(dq) = (1-\alpha(q'))\delta_{q'}(dq)\nu(dq')$

$$\begin{aligned}\int_{\mathcal{D}}\int_{\mathcal{D}} \phi(q,q')\,(1-\alpha(q))\,\delta_q(dq')\nu(dq) &= \int_{\mathcal{D}} \phi(q,q)(1-\alpha(q))\nu(dq) \\ &= \int_{\mathcal{D}}\int_{\mathcal{D}} \phi(q,q')(1-\alpha(q'))\delta_{q'}(dq)\nu(dq')\end{aligned}$$

- Note: other acceptance ratios $R(r)$ possible as long as $R(r) = rR(1/r)$, but the Metropolis ratio $R(r) = \min(1,r)$ is optimal in terms of asymptotic variance[5]

[5] P. Peskun, *Biometrika* (1973)

Metropolis-Hastings algorithm (4)

- Irreducibility: for almost all q_0 and any set $\mathcal{S}$ of positive measure, there exists n such that

$$P^n(q_0, \mathcal{S}) = \int_{x \in \mathcal{D}} P(q_0, dx)\, P^{n-1}(x, \mathcal{S}) > 0$$

- Assume also aperiodicity (comes from rejections)

- Pathwise ergodicity[6] $\displaystyle \lim_{N_{\mathrm{iter}} \to +\infty} \frac{1}{N_{\mathrm{iter}}} \sum_{n=1}^{N_{\mathrm{iter}}} \varphi(q^n) = \int_{\mathcal{D}} \varphi(q)\, \nu(dq)$

- Central limit theorem for Markov chains under additional assumptions:

$$\sqrt{N_{\mathrm{iter}}} \left| \frac{1}{N_{\mathrm{iter}}} \sum_{n=1}^{N_{\mathrm{iter}}} \varphi(q^n) - \int_{\mathcal{D}} \varphi(q)\, \nu(dq) \right| \xrightarrow[N_{\mathrm{iter}} \to +\infty]{\mathrm{law}} \mathcal{N}(0, \sigma_\varphi^2)$$

[6] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability* (1993)

Metropolis-Hastings algorithm (5)

- The asymptotic variance σ_φ^2 takes into account the correlations:

$$\sigma_\varphi^2 = \mathrm{Var}_\nu(\varphi) + 2\sum_{n=1}^{+\infty} \mathbb{E}_\nu\Big[\big(\varphi(q^0) - \mathbb{E}_\nu(\varphi)\big)\big(\varphi(q^n) - \mathbb{E}_\nu(\varphi)\big)\Big]$$

Proof: Consider $\widetilde{\varphi} = \varphi - \mathbb{E}_\nu(\varphi)$ and the average $\widetilde{\Phi}_{N_{\mathrm{iter}}} = \dfrac{1}{N_{\mathrm{iter}}}\displaystyle\sum_{n=1}^{N_{\mathrm{iter}}} \widetilde{\varphi}(q^n)$

Compute $N_{\mathrm{iter}}\mathbb{E}_\nu\left(\widetilde{\Phi}_{N_{\mathrm{iter}}}^2\right) = \dfrac{1}{N_{\mathrm{iter}}}\displaystyle\sum_{n,m=0}^{N_{\mathrm{iter}}} \mathbb{E}_\nu\Big(\widetilde{\varphi}(q^n)\widetilde{\varphi}(q^m)\Big)$

Stationarity $\mathbb{E}_\nu\Big(\widetilde{\varphi}(q^n)\widetilde{\varphi}(q^m)\Big) = \mathbb{E}_\nu\Big(\widetilde{\varphi}(q^{n-m})\widetilde{\varphi}(q^0)\Big)$ for $n \geqslant m$, implies

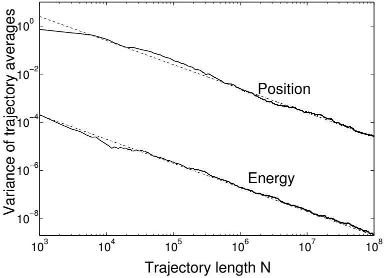
$$N_{\mathrm{iter}}\mathbb{E}_\nu\left(\widetilde{\Phi}_{N_{\mathrm{iter}}}^2\right) = \mathbb{E}_\nu\Big(\widetilde{\varphi}\left(q^0\right)^2\Big) + 2\sum_{n=1}^{N_{\mathrm{iter}}}\left(1 - \frac{n}{N_{\mathrm{iter}}}\right)\mathbb{E}_\nu\Big(\widetilde{\varphi}(q^n)\widetilde{\varphi}(q^0)\Big)$$

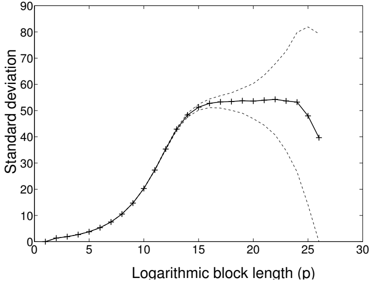
Metropolis-Hastings algorithm (6)

- Estimation of σ_φ^2 by block averaging (batch means)

$$\sigma_\varphi^2 = \lim_{N,M\to+\infty} \frac{N}{M}\sum_{k=1}^{M}\left(\Phi_N^k - \Phi_{NM}^1\right)^2, \qquad \Phi_N^k = \frac{1}{N}\sum_{n=(k-1)N+1}^{kN} \varphi(q^n)$$

Expected $\Phi_N^k \sim \displaystyle\int_{\mathcal{X}} \varphi\, d\nu + \frac{\sigma_\varphi}{\sqrt{N}}\mathscr{G}^k$, with $\mathscr{G}^k$ i.i.d.





Metropolis-Hastings algorithm (7)

• Useful rewriting: number of correlated steps $\sigma_\varphi^2 = N_{\rm corr} \mathrm{Var}_\nu(\varphi)$

• Numerical efficiency: trade-off between acceptance and sufficiently large moves in space to reduce autocorrelation (rejection rate around 0.5)[7]

• Refined Monte Carlo moves such as

- "non physical" moves
- parallel tempering
- replica exchanges
- Hybrid Monte-Carlo

• A way to stabilize discretization schemes for SDEs

[7] Roberts/Gelman/Gilks (1997), ..., Jourdain/Lelièvre/Miasojedow (2012)

Hybrid Monte–Carlo

The Hamiltonian dynamics (1)

Hamiltonian dynamics

$$\left\{ \begin{aligned} \frac{dq(t)}{dt} &= \nabla_p H(q(t),p(t)) = M^{-1}p(t) \\ \frac{dp(t)}{dt} &= -\nabla_q H(q(t),p(t)) = -\nabla V(q(t)) \end{aligned} \right.$$

Assumed to be well-posed (*e.g.* when the energy is a Lyapunov function)

- Flow: $\phi_t(q_0,p_0)$ solution at time t starting from initial condition (q_0,p_0)

- Why Hamiltonian formalism? (instead of working with velocities?)
  - Note that the vector field is divergence-free

$$\mathrm{div}_q\Big(\nabla_p H(q(t),p(t))\Big) + \mathrm{div}_p\Big(-\nabla_q H(q(t),p(t))\Big) = 0$$

  - Volume preservation $\displaystyle\int_{\phi_t(B)} dq\,dp = \int_B dq\,dp$

The Hamiltonian dynamics (2)

- Other properties
  - Preservation of energy $H \circ \phi_t = H$

$$\frac{d}{dt}\Big[H\big(q(t),p(t)\big)\Big] = \nabla_q H(q(t),p(t))\cdot\frac{dq(t)}{dt} + \nabla_p H(q(t),p(t))\cdot\frac{dp(t)}{dt} = 0$$

  - Time-reversibility $\phi_{-t} = S \circ \phi_t \circ S$ where $S(q,p) = (q,-p)$

    Proof: use $S^2 = \mathrm{Id}$ and note that

$$S \circ \phi_{-t}(q_0,p_0) = \big(q(-t), -p(-t)\big)$$

    is a solution of the Hamiltonian dynamics starting from $(q_0, -p_0)$, as is $\phi_t \circ S(q_0,p_0)$. Conclude by uniqueness of solution.

  - Symmetry $\phi_{-t} = \phi_t^{-1}$ (in general, $\phi_{t+s} = \phi_t \circ \phi_s$)

The Hamiltonian dynamics (3)

- Numerical integration: usually Verlet scheme[8] (Strang splitting)

Störmer-Verlet scheme

$$\begin{cases} p^{n+1/2} = p^n - \dfrac{\Delta t}{2} \nabla V(q^n) \\ q^{n+1} = q^n + \Delta t \; M^{-1} p^{n+1/2} \\ p^{n+1} = p^{n+1/2} - \dfrac{\Delta t}{2} \nabla V(q^{n+1}) \end{cases}$$

- Properties:
  - Symplectic, symmetric, time-reversible
  - One force evaluation per time-step, linear stability condition $\omega \Delta t < 2$
  - In fact, $M \dfrac{q^{n+1} - 2q^n + q^{n-1}}{\Delta t^2} = -\nabla V(q^n)$

[8] L. Verlet, *Phys. Rev.* **159**(1) (1967) 98-105

Hybrid Monte Carlo (1)

- Measure $\mu(dq\,dp) = \mathrm{e}^{-\beta H(q,p)}\,dq\,dp$ with marginal $\nu(dq) = \mathrm{e}^{-\beta V(q)}\,dq$

- Markov chain in the configuration space[9,10]: parameters τ and Δt
  - generate momenta p^n according to $Z_p^{-1}\,\mathrm{e}^{-\beta p^T M^{-1} p/2}\,dp$
  - compute an approximation of the flow $\Phi_\tau(q^n, p^n) = (\tilde{q}^{n+1}, \tilde{p}^{n+1})$ of the Hamiltonian dynamics (i.e. Verlet scheme with $\tau/\Delta t$ timesteps)
  - set $q^{n+1} = \tilde{q}^{n+1}$ with probability $\min\Big(1, \mathrm{e}^{-\beta(H(\tilde{q}^{n+1},\tilde{p}^{n+1}) - H(q^n,p^n))}\Big)$; otherwise set $q^{n+1} = q^n$.

- Rejection rate of order Δt^2 when $\tau = \mathrm{O}(1)$, and Δt^3 for $\tau = \Delta t$

- **Various extensions**, including correlated momenta, random times τ, constraints, ...

- Ergodicity is an issue (quadratic potential with $\tau =$ period)

---

[9] S. Duane, A. Kennedy, B. Pendleton and D. Roweth, *Phys. Lett. B* (1987)

[10] Ch. Schütte, *Habilitation Thesis* (1999)

(Generalized) Hybrid Monte Carlo (1)

- Transformation $S = S^{-1}$ leaving $\mu(dx)$ invariant, *e.g.* $S(q,p) = (q,-p)$

- Assume that $r(x,x') = \dfrac{T(S(x'),S(dx))\,\pi(dx')}{T(x,dx')\,\pi(dx)}$ is defined and positive

Generalized Hybrid Monte Carlo

- given x^n, propose a new state $\tilde{x}^{n+1}$ from x^n according to $T(x^n,\cdot)$;
- accept the move with probability $\min\left(1, r(x^n,\tilde{x}^{n+1})\right)$, and set in this case $x^{n+1} = \tilde{x}^{n+1}$; otherwise, set $x^{n+1} = S(x^n)$.

- Reversibility up to S, *i.e.* $P(x,dx')\,\mu(dx) = P(S(x'),S(dx))\,\mu(dx')$

- Standard HMC: $T(q,dq') = \delta_{\Phi_\tau(q)}(dq')$, momentum reversal upon rejection (not important since momenta are resampled, but is important when momenta are partially resampled)

(Generalized) Hybrid Monte Carlo (2)

Complete algorithm ($M = \mathrm{Id}$, $\beta = 1$): starting from (q^n, p^n),

- Partially resample momenta as $p^{n+1/2} = \alpha p^n + \sqrt{1-\alpha^2}\, G^n$
- Perform one Verlet step as $(\widetilde{q}^{n+1}, \widetilde{p}^{n+1}) = \Phi_{\Delta t}(q^n, p^n)$
- Compute the acceptance probability $a^n = \mathrm{e}^{H(q^n,p^n) - H(\widetilde{q}^{n+1}, \widetilde{p}^{n+1})}$
- Sample $U^n \sim \mathcal{U}[0,1]$
- If $U^n \leqslant a^n$, set $(q^{n+1}, p^{n+1}) = (\widetilde{q}^{n+1}, \widetilde{p}^{n+1})$
  otherwise set $(q^{n+1}, p^{n+1}) = (q^n, -p^{n+1/2})$

• Ergodicity no longer is an issue (irreducibility much easier to prove than for standard HMC)

Outline

- **Examples of high-dimensional probability measures**
  - Statistical physics
  - Bayesian inference
- **Markov chain methods**
  - Metropolis–Hastings algorithm
  - Hybrid Monte Carlo and its variants
- **Methods based on stochastic differential equations**
  - An introduction to SDEs (generators, invariant measure, discretization, etc)
  - Langevin-like dynamics
- **Variance reduction techniques**
- **Large scale Bayesian inference**
  - Mini-batching
  - Adaptive Langevin dynamics

Langevin dynamics

- Stochastic perturbation of the Hamiltonian dynamics : friction $\gamma > 0$

$$\begin{cases} dq_t = M^{-1} p_t \, dt \\ dp_t = -\nabla V(q_t)\, dt - \gamma M^{-1} p_t \, dt + \sqrt{\dfrac{2\gamma}{\beta}} \, dW_t \end{cases}$$

- **Motivations**
  - Ergodicity can be proved and is indeed observed in practice
  - Many useful extensions

- **Aims**
  - Understand the meaning of this equation
  - Understand why it samples the canonical ensemble
  - Implement appropriate discretization schemes
  - Estimate the errors (systematic biases vs. statistical uncertainty)

A (practical) introduction to SDEs

An intuitive view of the Brownian motion (1)

- Independant Gaussian increments whose variance is proportional to time

$$\forall\, 0 < t_0 \leqslant t_1 \leqslant \cdots \leqslant t_n, \qquad W_{t_{i+1}} - W_{t_i} \sim \mathcal{N}(0, t_{i+1} - t_i)$$

where the increments $W_{t_{i+1}} - W_{t_i}$ are independent

- $G \sim \mathcal{N}(m, \sigma^2)$ distributed according to the probability density

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

- The solution of $dq_t = \sigma dW_t$ can be thought of as the limit $\Delta t \to 0$

$$q^{n+1} = q^n + \sigma\sqrt{\Delta t}\, G^n, \qquad G^n \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

where q^n is an approximation of $q_{n\Delta t}$

- Note that $q^n \sim \mathcal{N}(q^0, \sigma^2 n \Delta t)$

- Multidimensional case: $W_t = (W_{1,t}, \ldots, W_{d,t})$ where W_i are independent

An intuitive view of the Brownian motion (2)

- Analytical study of the process: law $\psi(t,q)$ of the process at time t

$\rightarrow$ distribution of all possible realizations of q_t for

  - a given initial distribution $\psi(0,q)$, *e.g.* δ_{q^0}
  - and all realizations of the Brownian motion

Averages at time t

$$\mathbb{E}\Big(A(q_t)\Big) = \int_{\mathcal{D}} A(q)\,\psi(t,q)\,dq$$

- Partial differential equation governing the evolution of the law

Fokker-Planck equation

$$\partial_t \psi = \frac{\sigma^2}{2}\Delta\psi$$

Here, simple heat equation $\rightarrow$ "diffusive behavior"

An intuitive view of the Brownian motion (3)

- Proof: Taylor expansion, beware random terms of order $\sqrt{\Delta t}$

$$
\begin{aligned}
A\left(q^{n+1}\right) &= A\left(q^n + \sigma\sqrt{\Delta t}\, G^n\right) \\
&= A\left(q^n\right) + \sigma\sqrt{\Delta t} G^n \cdot \nabla A\left(q^n\right) + \frac{\sigma^2 \Delta t}{2}\left(G^n\right)^T \left(\nabla^2 A\left(q^n\right)\right) G^n + \mathrm{O}\Big(\Delta t^{3/2}\Big)
\end{aligned}
$$

Taking expectations (Gaussian increments G^n independent from the current position q^n)

$$
\mathbb{E}\left[A\left(q^{n+1}\right)\right] = \mathbb{E}\left[A\left(q^n\right) + \frac{\sigma^2 \Delta t}{2} \Delta A\left(q^n\right)\right] + \mathrm{O}\Big(\Delta t^{3/2}\Big)
$$

Therefore, $\mathbb{E}\left[\dfrac{A\left(q^{n+1}\right) - A\left(q^n\right)}{\Delta t} - \dfrac{\sigma^2}{2}\Delta A\left(q^n\right)\right] \to 0$. On the other hand,

$$
\mathbb{E}\left[\frac{A\left(q^{n+1}\right) - A\left(q^n\right)}{\Delta t}\right] \to \partial_t\Big(\mathbb{E}\left[A(q_t)\right]\Big) = \int_{\mathcal{D}} A(q)\partial_t \psi(t,q)\, dq.
$$

This leads to

$$
0 = \int_{\mathcal{D}} A(q)\partial_t\psi(t,q)\, dq - \frac{\sigma^2}{2}\int_{\mathcal{D}} \Delta A(q)\, \psi(t,q)\, dq = \int_{\mathcal{D}} A(q)\left(\partial_t\psi(t,q) - \frac{\sigma^2}{2}\Delta\psi(t,q)\right) dq
$$

This equality holds for all observables A.

General SDEs (1)

- State of the system $X \in \mathbb{R}^d$, m-dimensional Brownian motion, diffusion matrix $\sigma \in \mathbb{R}^{d\times m}$

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t$$

to be thought of as the limit as $\Delta t \to 0$ of (X^n approximation of $X_{n\Delta t}$)

$$X^{n+1} = X^n + \Delta t\, b\left(X^n\right) + \sqrt{\Delta t}\,\sigma(X^n)G^n, \qquad G^n \sim \mathcal{N}\left(0, \mathrm{Id}_m\right)$$

- Generator

$$\mathcal{L} = b(x)\cdot\nabla + \frac{1}{2}\sigma\sigma^T(x) : \nabla^2 = \sum_{i=1}^d b_i(x)\partial_{x_i} + \frac{1}{2}\sum_{i,j=1}^d \left[\sigma\sigma^T(x)\right]_{i,j}\partial_{x_i}\partial_{x_j}$$

- Proceeding as before, it can be shown that

$$\partial_t\Big(\mathbb{E}\left[A(X_t)\right]\Big) = \int_{\mathcal{X}} A\,\partial_t\psi = \mathbb{E}\Big[\left(\mathcal{L}A\right)(X_t)\Big] = \int_{\mathcal{X}}\left(\mathcal{L}A\right)\psi$$

General SDEs (2)

Fokker-Planck equation

$$\partial_t \psi = \mathcal{L}^* \psi$$

where $\mathcal{L}^*$ is the adjoint of $\mathcal{L}$

$$\int_{\mathcal{X}} (\mathcal{L}A)(x)\, B(x)\, dx = \int_{\mathcal{X}} A(x)\, (\mathcal{L}^* B)(x)\, dx$$

- Invariant measures are stationary solutions of the Fokker-Planck equation

Invariant probability measure $\psi_\infty(x)\, dx$

$$\mathcal{L}^* \psi_\infty = 0, \qquad \int_{\mathcal{X}} \psi_\infty(x)\, dx = 1, \qquad \psi_\infty \geqslant 0$$

- When $\mathcal{L}$ is elliptic (*i.e.* $\sigma\sigma^T$ has full rank: the noise is sufficiently rich), the process can be shown to be irreducible = accessibility property

$$P_t(x, \mathcal{S}) = \mathbb{P}(X_t \in \mathcal{S} \mid X_0 = x) > 0$$

General SDEs (3)

- Sufficient conditions for ergodicity
  - irreducibility
  - existence of an invariant probability measure $\psi_\infty(x)\,dx$

Then the invariant measure is unique and

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \varphi(X_t)\,dt = \int_{\mathcal{X}} \varphi(x)\,\psi_\infty(x)\,dx \qquad \text{a.s.}$$

- Rate of convergence given by Central Limit Theorem: $\widetilde{\varphi} = \varphi - \int \varphi\,\psi_\infty$

$$\sqrt{T}\left(\frac{1}{T}\int_0^T \varphi(X_t)\,dt - \int \varphi\,\psi_\infty\right) \xrightarrow[T\to+\infty]{\text{law}} \mathcal{N}(0,\sigma_\varphi^2)$$

with $\sigma_\varphi^2 = 2\,\mathbb{E}\left[\int_0^{+\infty} \widetilde{\varphi}(X_t)\widetilde{\varphi}(X_0)dt\right]$ (proof: later, discrete time setting)

SDEs: numerics (1)

- Numerical discretization: various schemes (Markov chains in all cases)

- Example: Euler-Maruyama

$$X^{n+1} = X^n + \Delta t\, b(X^n) + \sqrt{\Delta t}\, \sigma(X^n)\, G^n, \qquad G^n \sim \mathcal{N}(0, \mathrm{Id}_d)$$

- Standard notions of error: fixed integration time $T < +\infty$
  - Strong error $\displaystyle\sup_{0 \leqslant n \leqslant T/\Delta t} \mathbb{E}|X^n - X_{n\Delta t}| \leqslant C\Delta t^p$
  - Weak error: $\displaystyle\sup_{0 \leqslant n \leqslant T/\Delta t} \Big|\mathbb{E}\left[\varphi\left(X^n\right)\right] - \mathbb{E}\left[\varphi\left(X_{n\Delta t}\right)\right]\Big| \leqslant C\Delta t^p$ (for any φ)
  - "mean error" *vs.* "error of the mean"

- Example: for Euler-Maruyama, weak order 1, strong order $1/2$ (1 when σ constant)

SDEs: numerics (2)

- Trajectorial averages: estimator $\Phi_{N_{\rm iter}} = \dfrac{1}{N_{\rm iter}} \displaystyle\sum_{n=1}^{N_{\rm iter}} \varphi(X^n)$

- Numerical scheme ergodic for the probability measure $\psi_{\infty,\Delta t}$

- Two types of errors to compute averages w.r.t. invariant measure
  - Statistical error, quantified using a Central Limit Theorem

$$\Phi_{N_{\rm iter}} = \int \varphi\, \psi_{\infty,\Delta t} + \frac{\sigma_{\Delta t,\varphi}}{\sqrt{N_{\rm iter}}} \mathscr{G}_{N_{\rm iter}}, \qquad \mathscr{G}_{N_{\rm iter}} \sim \mathcal{N}(0,1)$$

  - Systematic errors
    - perfect sampling bias, related to the finiteness of Δt

$$\left| \int_{\mathcal{X}} \varphi\, \psi_{\infty,\Delta t} - \int_{\mathcal{X}} \varphi\, \psi_\infty \right| \leqslant C_\varphi\, \Delta t^p$$

    - finite sampling bias, related to the finiteness of $N_{\rm iter}$

SDEs: numerics (3)

Expression of the asymptotic variance: correlations matter!

$$\sigma^2_{\Delta t,\varphi} = \mathrm{Var}(\varphi) + 2\sum_{n=1}^{+\infty} \mathbb{E}\Big(\widetilde{\varphi}(X^n)\widetilde{\varphi}(X^0)\Big), \qquad \widetilde{\varphi} = \varphi - \int \varphi\,\psi_{\infty,\Delta t}$$

where $\mathrm{Var}(\varphi) = \displaystyle\int_{\mathcal{X}} \widetilde{\varphi}^2 \psi_{\infty,\Delta t} = \int_{\mathcal{X}} \varphi^2 \psi_{\infty,\Delta t} - \left(\int_{\mathcal{X}} \varphi\psi_{\infty,\Delta t}\right)^2$

- Note also that $\sigma^2_{\Delta t,\varphi} \sim \dfrac{2}{\Delta t}\mathbb{E}\left[\displaystyle\int_0^{+\infty} \widetilde{\varphi}(X_t)\widetilde{\varphi}(X_0)\,dt\right]$

- Estimation with block averaging for instance, or approximation of integrated autocorrelation

Langevin-like dynamics

Overdamped Langevin dynamics

- SDE on the configurational part only (momenta trivial to sample)

$$dq_t = -\nabla V(q_t)\, dt + \sqrt{\frac{2}{\beta}} dW_t$$

- Invariance of the canonical measure $\nu(dq) = \psi_0(q)\, dq$

$$\psi_0(q) = Z^{-1}\, \mathrm{e}^{-\beta V(q)}, \qquad Z = \int_{\mathcal{D}} \mathrm{e}^{-\beta V(q)}\, dq$$

- Generator $\mathcal{L} = -\nabla V(q) \cdot \nabla_q + \dfrac{1}{\beta}\Delta_q$
  - invariance of ψ_0: adjoint $\mathcal{L}^* \varphi = \mathrm{div}_q \left((\nabla V)\varphi + \dfrac{1}{\beta}\nabla_q \varphi \right)$
  - elliptic generator hence irreducibility and ergodicity
- Discretization $q^{n+1} = q^n - \Delta t\, \nabla V(q^n) + \sqrt{\dfrac{2\Delta t}{\beta}}\, G^n$ (+ Metropolization)

Langevin dynamics (1)

- Stochastic perturbation of the Hamiltonian dynamics

$$\begin{cases} dq_t = M^{-1} p_t \, dt \\ dp_t = -\nabla V(q_t)\, dt - \gamma M^{-1} p_t \, dt + \sigma \, dW_t \end{cases}$$

- γ,σ may be matrices, and may depend on q

- Generator $\mathcal{L} = \mathcal{L}_{\mathrm{ham}} + \mathcal{L}_{\mathrm{thm}}$

$$\mathcal{L}_{\mathrm{ham}} = p^T M^{-1} \nabla_q - \nabla V(q)^T \nabla_p = \sum_{i=1}^{dN} \frac{p_i}{m_i} \partial_{q_i} - \partial_{q_i} V(q) \partial_{p_i}$$

$$\mathcal{L}_{\mathrm{thm}} = -p^T M^{-1} \gamma^T \nabla_p + \frac{1}{2} \left(\sigma \sigma^T \right) : \nabla_p^2 \qquad \left(= \frac{\sigma^2}{2} \Delta_p \text{ for scalar } \sigma \right)$$

- Irreducibility can be proved (control argument)

Langevin dynamics (2)

- Invariance of the canonical measure to conclude to ergodicity?

Fluctuation/dissipation relation

$$\sigma\sigma^T = \frac{2}{\beta}\gamma \qquad \text{implies} \qquad \mathcal{L}^*\left(\mathrm{e}^{-\beta H}\right) = 0$$

- Proof for scalar γ, σ: a simple computation shows that

$$\mathcal{L}_{\mathrm{ham}}^* = -\mathcal{L}_{\mathrm{ham}}, \qquad \mathcal{L}_{\mathrm{ham}} H = 0$$

- Overdamped Langevin analogy $\mathcal{L}_{\mathrm{thm}} = \gamma\left(-p^T M^{-1}\nabla_p + \frac{1}{\beta}\Delta_p\right)$

$\to$ Replace q by p and $\nabla V(q)$ by $M^{-1}p$

$$\mathcal{L}_{\mathrm{thm}}^*\left[\exp\left(-\beta\frac{p^T M^{-1} p}{2}\right)\right] = 0$$

- Conclusion: $\mathcal{L}_{\mathrm{ham}}^*$ and $\mathcal{L}_{\mathrm{thm}}^*$ both preserve $\mathrm{e}^{-\beta H(q,p)}\, dq\, dp$

Langevin dynamics (3)

- Exponential convergence of semigroup $\mathrm{e}^{t\mathcal{L}}$ on Banach spaces $E \cap L_0^2(\mu)$
  - Lyapunov techniques[11] on $L_W^\infty(\mathcal{E}) = \left\{\varphi \, \text{measurable}, \left\|\frac{\varphi}{W}\right\|_{L^\infty} < +\infty\right\}$
  - Hypocoercive[12] setup $H^1(\mu)$, with hypoelliptic regularization[13], or directly[14] $L^2(\mu)$
  - Coupling techniques[15]

- Allows to define the asymptotic variance (with $\Pi\varphi = \varphi - \mathbb{E}_\mu(\varphi)$)

$$\sigma_\varphi^2 = 2\int_0^{+\infty} \int \left(\mathrm{e}^{t\mathcal{L}}\Pi\varphi\right) \Pi\varphi \, d\mu \, dt = 2\int (-\mathcal{L}^{-1}\Pi\varphi)\Pi\varphi \, d\mu$$

[11] L. Rey-Bellet, *Lecture Notes in Mathematics* (2006), Hairer/Mattingly (2011)
[12] Villani (2009) and before Talay (2002), Eckmann/Hairer (2003), Hérau/Nier (2004)
[13] F. Hérau, *J. Funct. Anal.* **244**(1), 95-118 (2007)
[14] Dolbeault, Mouhot and Schmeiser (2009, 2015); Armstrong and Mourrat (2019)
[15] Eberle, Guillin and Zimmer (2019)

Numerical integration of the Langevin dynamics (1)

- Splitting strategy: Hamiltonian part + fluctuation/dissipation

$$\left\{\begin{aligned} dq_t &= \ \ M^{-1}\,p_t\,dt \\ dp_t &= -\nabla V(q_t)\,dt \end{aligned}\right. \qquad \oplus \qquad \left\{\begin{aligned} dq_t &= 0 \\ dp_t &= -\gamma M^{-1} p_t\,dt + \sqrt{\frac{2\gamma}{\beta}}\,dW_t \end{aligned}\right.$$

- Hamiltonian part integrated using a Verlet scheme

- Analytical integration of the fluctuation/dissipation part

$$d\left(\mathrm{e}^{\gamma M^{-1} t} p_t\right) = \mathrm{e}^{\gamma M^{-1} t}\left(dp_t + \gamma M^{-1} p_t\,dt\right) = \sqrt{\frac{2\gamma}{\beta}}\mathrm{e}^{\gamma M^{-1} t}\,dW_t$$

so that

$$p_t = \mathrm{e}^{-\gamma M^{-1} t}\,p_0 + \sqrt{\frac{2\gamma}{\beta}}\int_0^t \mathrm{e}^{-\gamma M^{-1}(t-s)}\,dW_s$$

It can be shown that $\displaystyle\int_0^t f(s)\,dW_s \sim \mathcal{N}\left(0, \int_0^t f(s)^2 ds\right)$

Numerical integration of the Langevin dynamics (2)

- Trotter splitting (define $\alpha_{\Delta t} = \mathrm{e}^{-\gamma M^{-1}\Delta t}$, choose $\gamma M^{-1}\Delta t \sim 0.01 - 1$)

$$\begin{cases} p^{n+1/2} = p^n - \dfrac{\Delta t}{2}\nabla V(q^n), \\ q^{n+1} = q^n + \Delta t\, M^{-1} p^{n+1/2}, \\ \widetilde{p}^{n+1} = p^{n+1/2} - \dfrac{\Delta t}{2}\nabla V(q^{n+1}), \\ p^{n+1} = \alpha_{\Delta t}\widetilde{p}^{n+1} + \sqrt{\dfrac{1-\alpha_{2\Delta t}}{\beta}M}\, G^n, \end{cases}$$

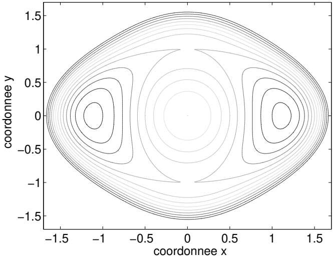
Error estimate on the invariant measure $\mu_{\Delta t}$ of the numerical scheme

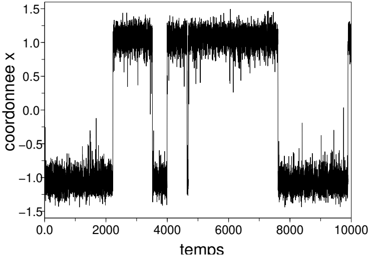
There exist a function f such that, for any smooth observable ψ,

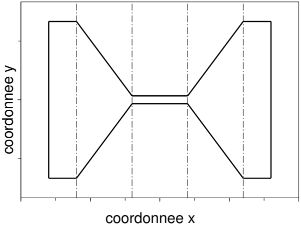
$$\int_{\mathcal{E}} \psi\, d\mu_{\Delta t} = \int_{\mathcal{E}} \psi\, d\mu + \Delta t^2 \int_{\mathcal{E}} \psi\, f\, d\mu + \mathrm{O}(\Delta t^3)$$

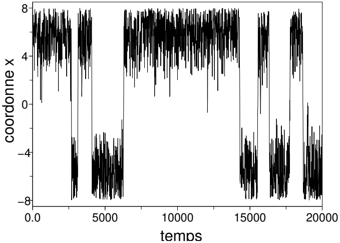
- Strang splitting more expensive and not more accurate

Metastability: large variances...









Need for variance reduction techniques!

Outline

- **Examples of high-dimensional probability measures**
  - Statistical physics
  - Bayesian inference
- **Markov chain methods**
  - Metropolis–Hastings algorithm
  - Hybrid Monte Carlo and its variants
- **Methods based on stochastic differential equations**
  - An introduction to SDEs (generators, invariant measure, discretization, etc)
  - Langevin-like dynamics
- **Variance reduction techniques**
- **Large scale Bayesian inference**
  - Mini-batching
  - Adaptive Langevin dynamics

Main strategies for variance reduction

- **Example:** computation of the integral $\displaystyle\int_{[-1/2,1/2]^d} f$
  - Estimation with i.i.d. variables $X^i \sim \mathcal{U}([-1/2,1/2]^d)$ as $S_{N_{\rm iter}} = N_{\rm iter}^{-1}\left(f(X^1) + \dots + f(X_{N_{\rm iter}})\right)$
  - Asymptotic variance $\sigma_f^2 = \mathrm{Var}(f) \to$ reduce it?
- **Various methods** (i.i.d. context, but can be extended to MCMC)
  - Antithetic variables $I_{N_{\rm iter}} = \frac{1}{2N_{\rm iter}} \sum_{i=1}^{N_{\rm iter}} \Big(f\left(X^i\right) + f\left(-X^i\right) \Big)$
  - Control variates with $\sigma_{f-g}^2 \ll \sigma_f^2$ and g analytically integrable

$$I_{N_{\rm iter}} = \frac{1}{N_{\rm iter}} \sum_{i=1}^{N_{\rm iter}} (f-g)\left(X^i\right) + \int_{[-1/2,1/2]^d} g$$

  - Stratification: partition domain, sample subdomains, aggregate
  - Importance sampling

Importance sampling

- **Importance sampling function $\widetilde{V}$**
  - Target measure $\pi_0(dx) = Z_0^{-1} \mathrm{e}^{-V(x)}\, dx$
  - Sample a modified target measure $\pi_{\widetilde{V}}(dx) = Z_{\widetilde{V}}^{-1} \mathrm{e}^{-(V+\widetilde{V})(x)}\, dx$
  - Reweight sample points $x^n \sim \pi_{\widetilde{V}}$ by $\mathrm{e}^{\widetilde{V}}$

$$\widehat{\varphi}_{N_{\mathrm{iter}},\widetilde{V}} = \frac{\displaystyle\sum_{n=1}^{N_{\mathrm{iter}}} \varphi(x^n) \mathrm{e}^{\widetilde{V}(x^n)}}{\displaystyle\sum_{n=1}^{N_{\mathrm{iter}}} \mathrm{e}^{\widetilde{V}(x^n)}} \xrightarrow[N_{\mathrm{iter}}\to+\infty]{\text{a.s.}} \frac{\displaystyle\int \varphi\, \mathrm{e}^{\widetilde{V}}\, d\pi_{\widetilde{V}}}{\displaystyle\int \mathrm{e}^{\widetilde{V}}\, d\pi_{\widetilde{V}}} = \int \varphi\, d\pi_0$$

- In practice, replace $-\nabla V$ with $-\nabla V - \nabla \widetilde{V}$ (in Langevin, MALA, etc)

- A good choice of the importance sampling function can improve the performance of the estimator... but a bad choice can degrade it!

High dimensional importance sampling

- **General strategy:**
  - find some low-dimensional (nonlinear) function $\xi(x)$ which encodes the metastability of the sampling method
  - bias by the associated free energy: $\widetilde{V}(x) = F(\xi(x))$ with

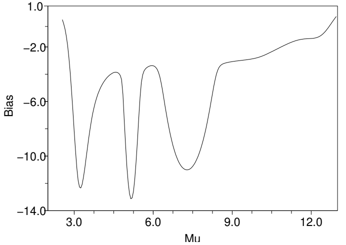
$$\mathrm{e}^{-F(z)} = \int \mathrm{e}^{-V(x)} \, \delta_{\xi(x)-z}(dx)$$

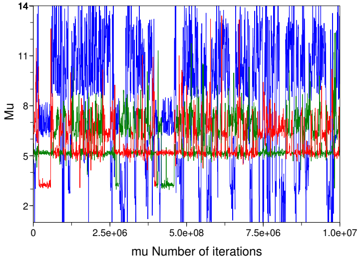
  - Simple case: $\xi(x) = x_1$, in which case

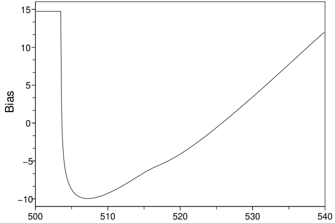
$$F(z) = -\ln \left(\int \mathrm{e}^{-V(z,x_2,\dots,x_d)} \, dx_2 \dots dx_d \right)$$

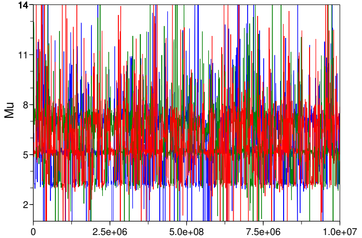
- **Various methods to compute the free energy**: thermodynamic integration, umbrella sampling, adaptive methods, ...

Free energy biasing for Bayesian inference









Choices $\xi(x) = \mu_1$ and $\xi(x) = V(x)$

[CLS12] N. Chopin, T. Lelièvre and G. Stoltz, *Statist. Comput.*, 2012

Outline

- **Examples of high-dimensional probability measures**
  - Statistical physics
  - Bayesian inference
- **Markov chain methods**
  - Metropolis–Hastings algorithm
  - Hybrid Monte Carlo and its variants
- **Methods based on stochastic differential equations**
  - An introduction to SDEs (generators, invariant measure, discretization, etc)
  - Langevin-like dynamics
- **Variance reduction techniques**
- **Large scale Bayesian inference**
  - Mini-batching
  - Adaptive Langevin dynamics

Bayesian inference in the large data context

- **Data $\{y_i\}_{i=1,\dots,N_{\rm data}}$ to be explained by a statistical model**
  - Sample q from $\nu(dq) = \mathrm{e}^{-V(q)}\,dq = Z_\nu^{-1} p_{\rm prior}(q) \prod_{i=1}^{N_{\rm data}} P(y_i|q)\,dq$
  - For usual MCMC methods, each step costs $\mathrm{O}(N_{\rm data})$

- **Mini-batching**: Stochastic gradient Langevin dynamics[16]
  - Assumption: for $1 \ll \mathcal{N} \ll N_{\rm data}$ and $J_{\mathcal{N}} \in \{1,\dots,N\}^{\mathcal{N}}$,

$$\nabla(\ln\rho)(q) + \frac{N_{\rm data}}{\mathcal{N}} \sum_{j \in J_{\mathcal{N}}} \nabla(\ln P(y_j|q)) = -\nabla V(q) + \mathcal{G}, \quad \mathcal{G} \sim \mathcal{N}(0, \Sigma(q))$$

  - Amounts to introducing an additional Brownian motion of unknown magnitude $\to$ bias
  - Assume that $\Sigma(q)$ is constant [Work of Inass Sekkat...]

[16] Welling/Teh, *ICML* (2011)

Removing the mini-batching bias

- Phase-space extension: momenta p and variable friction ζ

Adaptive Langevin dynamics[13]: unknown σ (scalar, for simplicity)

$$\begin{aligned} dq_t &= M^{-1} p_t \, dt, \\ dp_t &= \left(-\nabla V(q_t) - \zeta_t M^{-1} p_t\right) dt + \sigma \, dW_t, \\ d\zeta_t &= \frac{1}{m} \left(p_t^T M^{-2} p_t - \beta^{-1} \mathrm{Tr}\left(M^{-1}\right)\right) dt \end{aligned}$$

- Invariant measure with marginal in q is always ν (whatever σ)

$$\exp\left(-\beta\left[\frac{p^T M^{-1} p}{2} + V(q) + \frac{m}{2}\left(\zeta - \frac{\beta\sigma^2}{2}\right)^2\right]\right) dq \, dp \, d\zeta$$

- Convergence/CLT for time averages[17]

[17] B. Leimkuhler, M. Sachs and G. Stoltz, Hypocoercivity properties of adaptive Langevin dynamics, *arXiv preprint* **1908.09363**

[13] A. Jones and B. Leimkuhler, *J. Chem. Phys.* (2011); Ding et al., *NIPS* (2014); B. Leimkuhler and X. Shang, *SIAM J. Sci. Comput.* (2015)