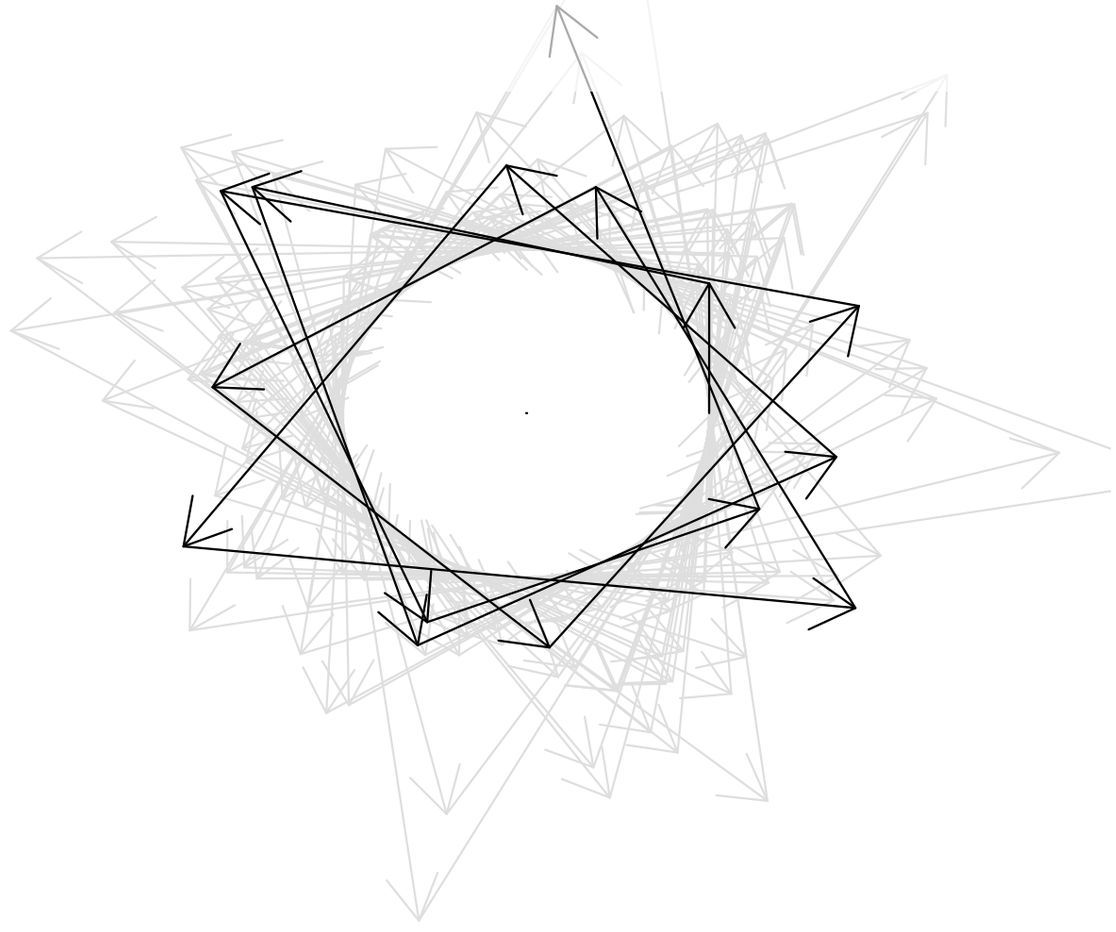
Monte Carlo methods

Alexandre Bouchard-Côté



Project

- Due: April 26 (send code + pdf by email)
- By ~~Monday (March 26)~~ Tuesday April 3 send informal plan for project by email
- See Syllabus page on website for more info
- Encouraged to combine with your research and/or other classes

Stopping criteria

- Many diagnostic exist
  - All have limitations
  - Some are dubious
- Best approach is CLT (with same caveats as IS): for a 95% confidence interval, use

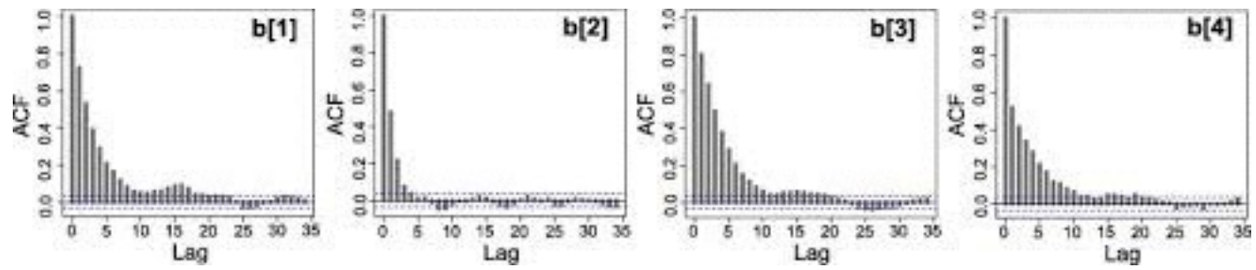
$$I_n \pm 1.96\sqrt{\sigma^2_{\text{asympt}}/n}$$

- The asymptotic variance is:

$$\sigma^2(\phi) = \mathbb{V}_\pi[\phi(X_1)] + 2\sum_{k=2}^{\infty} \mathbb{C}ov_\pi[\phi(X_1), \phi(X_k)].$$

Estimation of the asymptotic variance

- Direct method: estimate the auto-correlations (ACF)



- Can be done quickly with FFT
- But estimator has infinite variance! Need to truncate. Still unstable in many practical scenarios.

ESS for MC

- Idea is similar as for IS, but still tied to a test function:

$$\frac{\mathrm{ESS(N)}}{N} \rightarrow \frac{\sigma^2_{\mathrm{iid}}}{\sigma^2_{\mathrm{asymptotic}}}$$

- Can estimate using ACF as in last slide
- Better method: batch estimators.
  - Segment the MCMC trace into chunks of length $\sqrt{n}$
  - Assume sampler is good enough so that behaviour across blocks is nearly iid
- Standard metric in MCMC literature to compare samplers: ESS per second or ESS per operation

Asymptotic variance and ESS for MC

Def 43

- References:
  - Honest exploration of intractable probability distributions (2001). Jones and Hobert.
    - Monte Carlo standard errors for MCMC (2008). Flegal.
- Multivariate confidence version:
  - Multivariate output analysis for Markov chain Monte Carlo (2015). Vats et al.

Ingenious MCMC constructions

Terminology

Collapsed sampler: analytically marginalize some of the variables, and run MCMC on the *reduced* state space

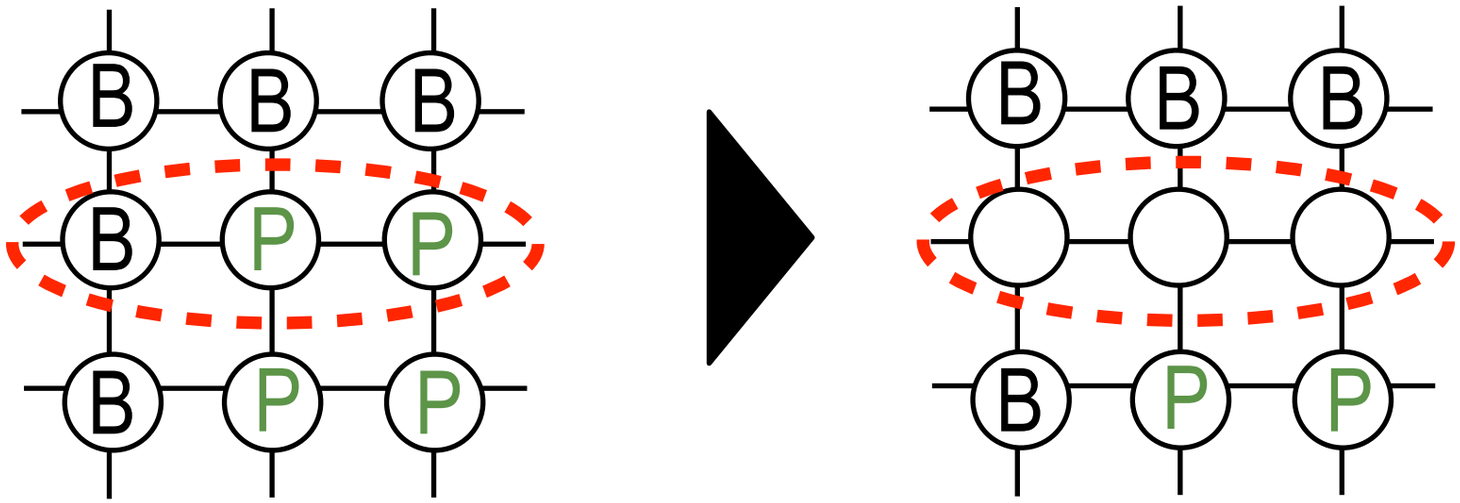
Example: *HMM global parameter inference while summing over latent dynamic states*

Auxiliary variable: *augment* the state space to facilitate sampling

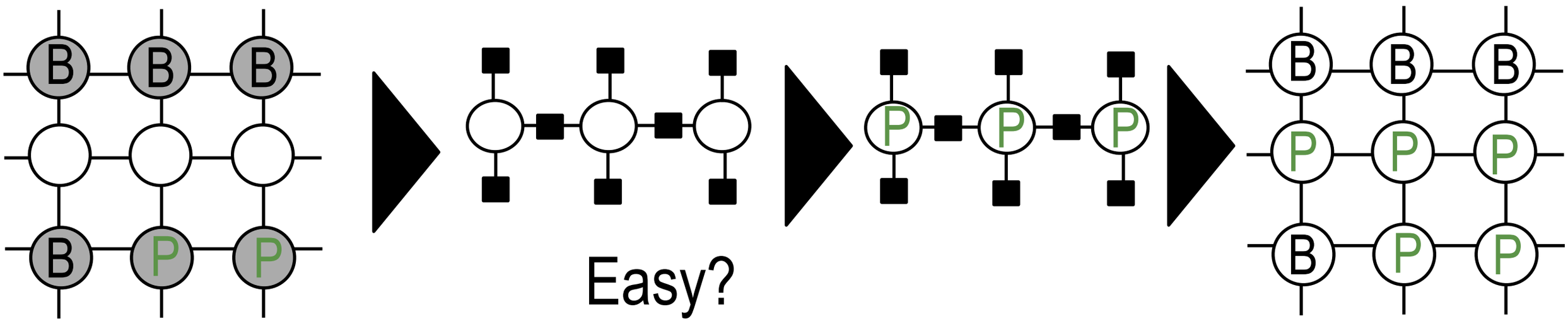
Example: *slice sampling*

Another collapsed example: Collapsed Gibbs samplers

Loop: pick a subset of nodes N at random, erase the contents of the guessed values in N, freeze the value of the nodes not in N



Then: resample a value for the nodes in N conditioning on all the others, and write this to the current state at N



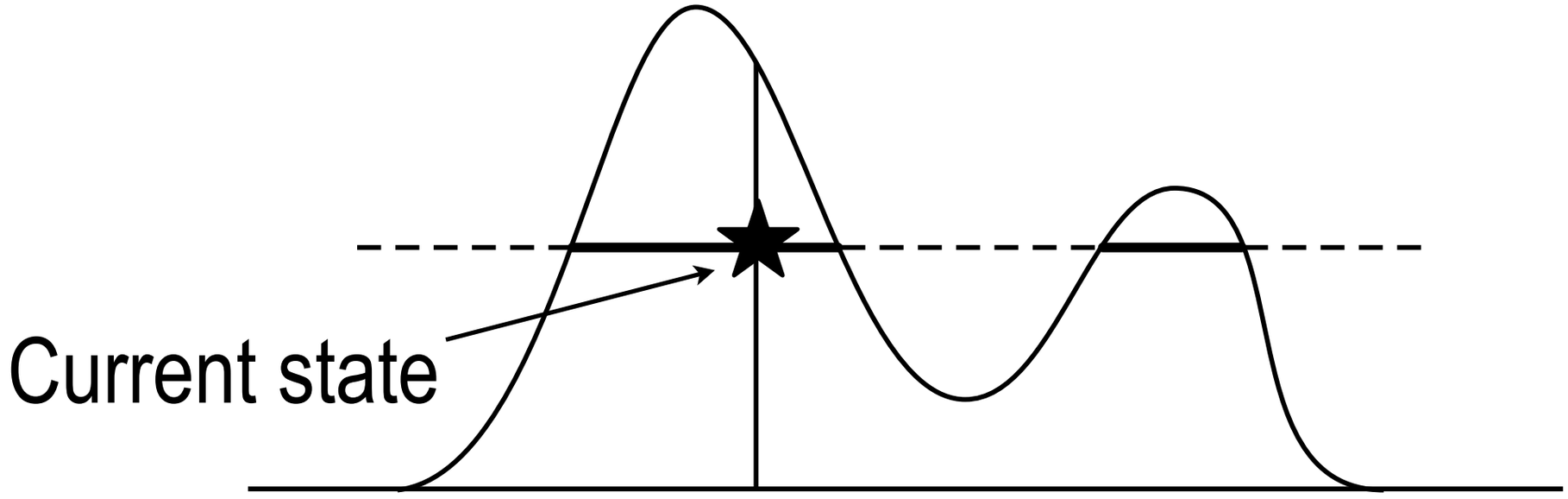
Easy?

Slice sampling

Goal: sampling from a r.v. X with density $f(x)/Z$, where Z is difficult to compute

Intuition: use a MCMC defined on the 2D space defined as the graph of the density

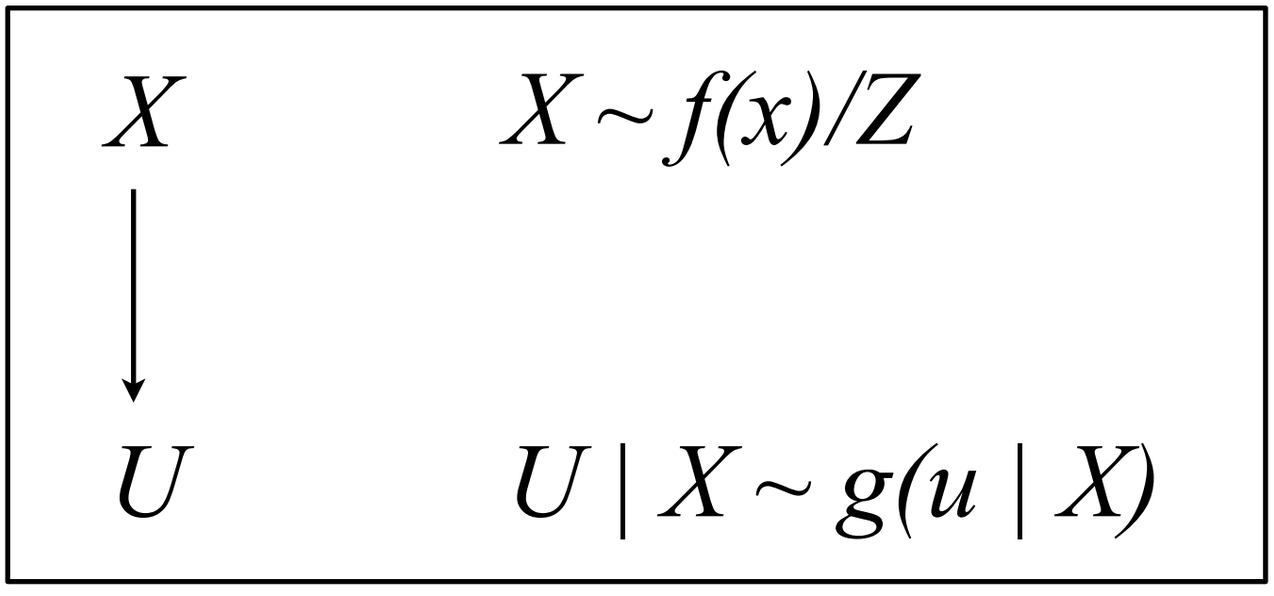
Moves: sample uniformly vertically or horizontally



Slice sampling

Goal: sampling from a r.v. X with density $f(x)/Z$, where Z is difficult to compute

General auxiliary variable construction: adding a new random variable U with the following graphical model does not change the marginal distribution of X, no matter what is the conditional density g of $U \mid X$

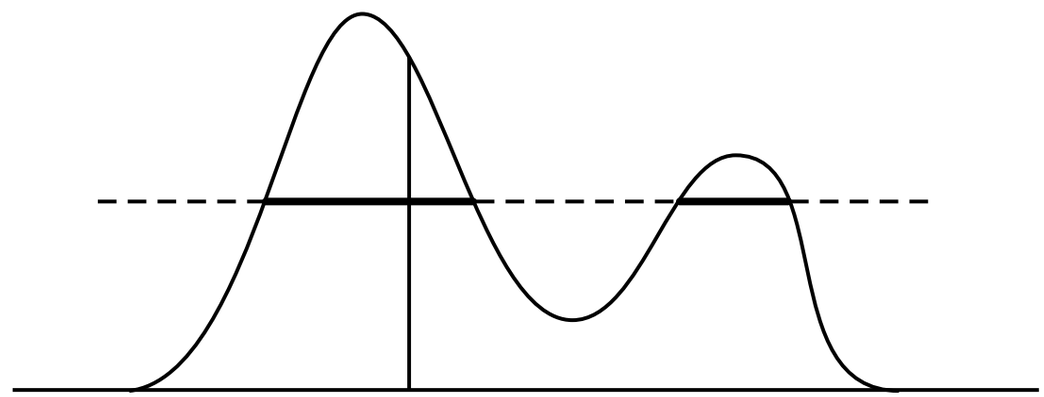


Slice sampler

$X \qquad X \sim f(x)/Z$

$\downarrow$

$U \qquad U \mid X \sim \text{Uni}[0, f(X)]$



Vertical move: $U \mid X \sim \text{Uni}[0, f(X)]$

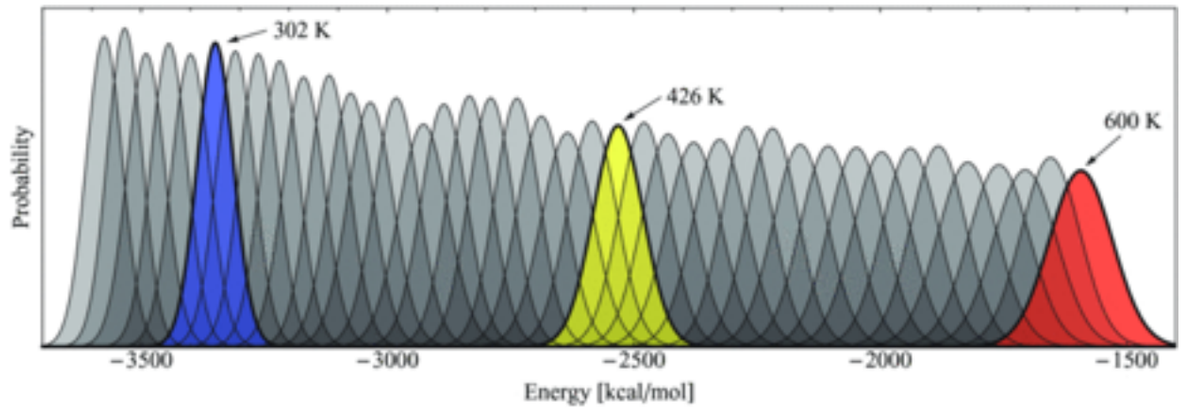
Horizontal move: $X \mid U \sim \text{Uni}\{x : f(x) \geq U\}$

Note: Efficient alternatives to the horizontal move exist (state-of-the-art: doubling+shrinking procedure, Neal, 2003)
See https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/main/java/blang/mcmc/RealSliceSampler.java

Annealing and tempering

- Key idea: using sequences of distributions
  - Denoted, for $t = 0 .. 1$, π_t
- The case $t = 0$ should be easy ('heated')
  - ideally, such that we can get exact idd samples in poly-time for that temperature
- The case $t = 1$ should coincide with the target of interest ('room temperature')

Sequences of π



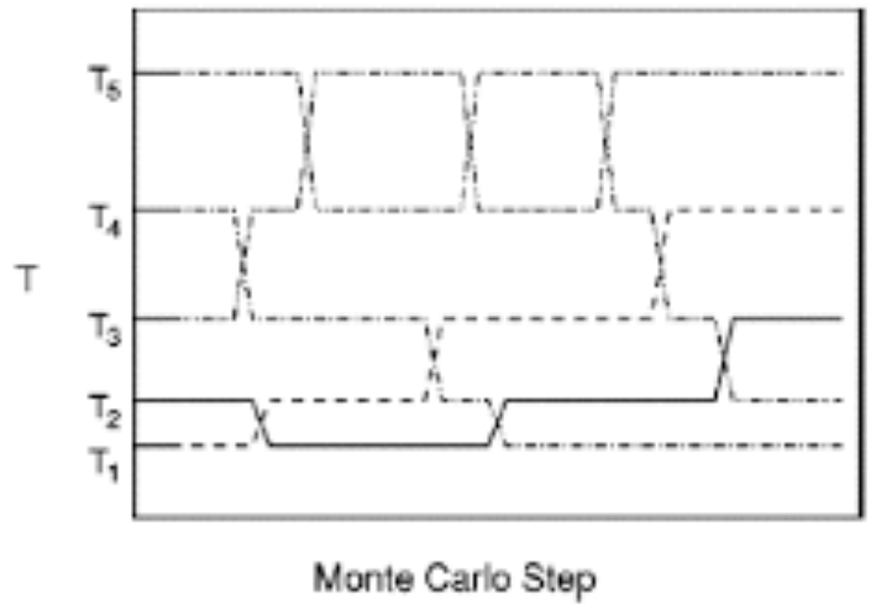
- Examples
  - Naive: exponentiate the whole target
    - Problem: we don't want non-normalizable targets
    - Solution: Exponentiate only likelihood
  - Other issues
    - hard constraints/restrictions in the likelihood
    - computation: interpolate number of datapoints [Project]
  - Automatic creation of sequences of distributions in Blang: https://www.stat.ubc.ca/~bouchard/blang/Inference_and_runtime.html
- Sparsity considerations (changing *t* should be O(1))

Annealing

- Make temperature random
- Extract subset of samples where $t = 1$
- Exact simulation version exists:
  - Moller and Nicholls 1999
- Problem?

Parallel tempering

- Have all the temperature exist at same time
  - I.e. state space is product instead of union
  - Swap temperatures

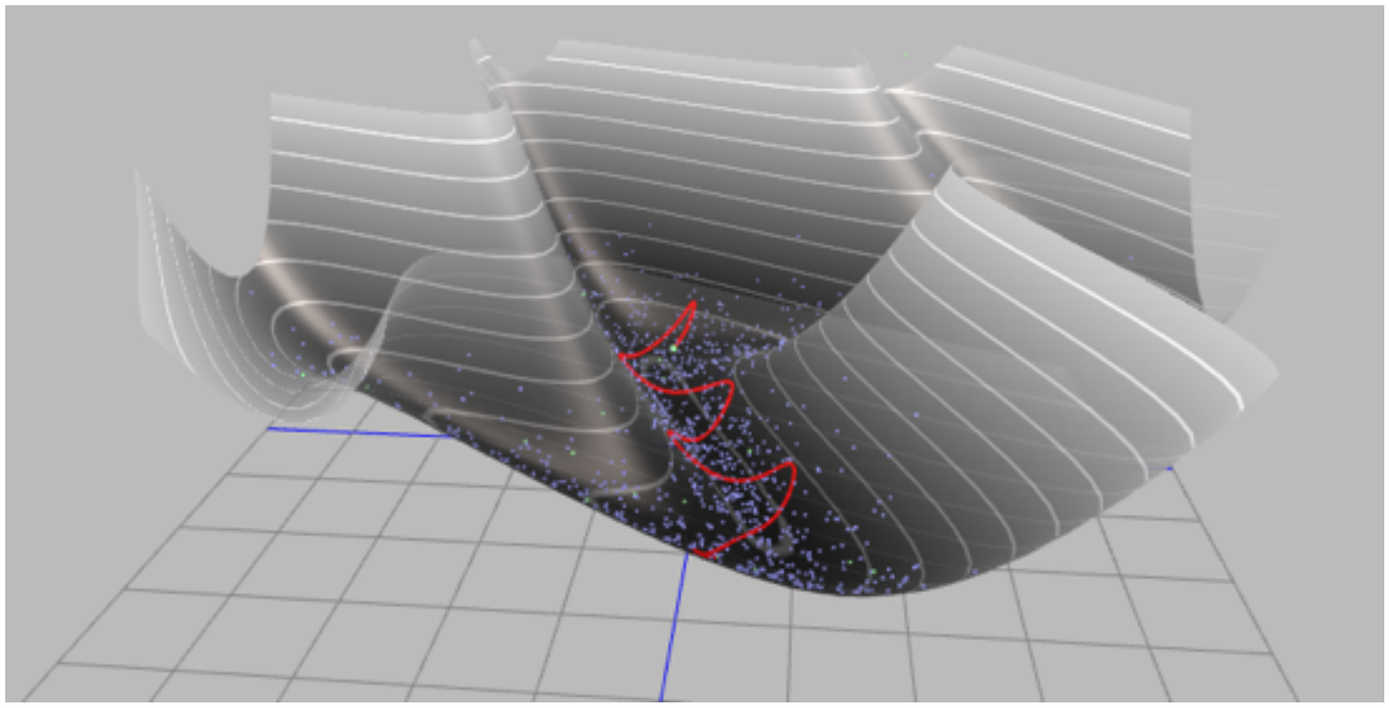


- Normalization constants cancel out now!
- Parallel implementation attractive
- See https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/main/java/blang/engines/ParallelTempering.java

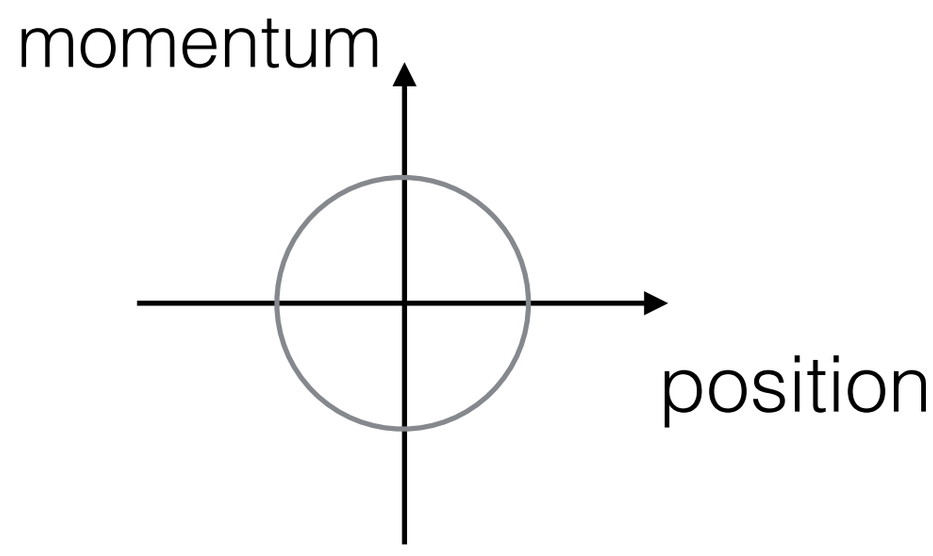
Gradient-based methods

Hamiltonian Monte Carlo: intuition

- Physical ball rolling on the *energy*
  - E(x) = -log(p(x))
- Motion described by the *Hamiltonian flow*



- Phase space on a Gaussian target:



HMC: auxiliary variables

- Physics' notation: z = (q, p)
  - position *q*
  - Augment the state with a momentum random variable *p*
- Put an auxiliary distribution on p, with f(p) = exp(-K(p)) and s.t. K(p) = K(-p), e.g. normal.

$$H(q,p) = U(q) + K(p), \quad U(q) = q^2/2, \quad K(p) = p^2/2$$

- Can think of p as a velocity (when the mass matrix, i.e. covariance of f(x) is identity).
- Statistical notation would be then z = (x, v)

Exact HMC

- MCMC kernel is a non-reversible
- Given by a Dirac delta: $k(z, dz') = \delta_{\Phi(z)}(dz')$
- Φ is the Hamiltonian flow, i.e. solutions of the differential equations

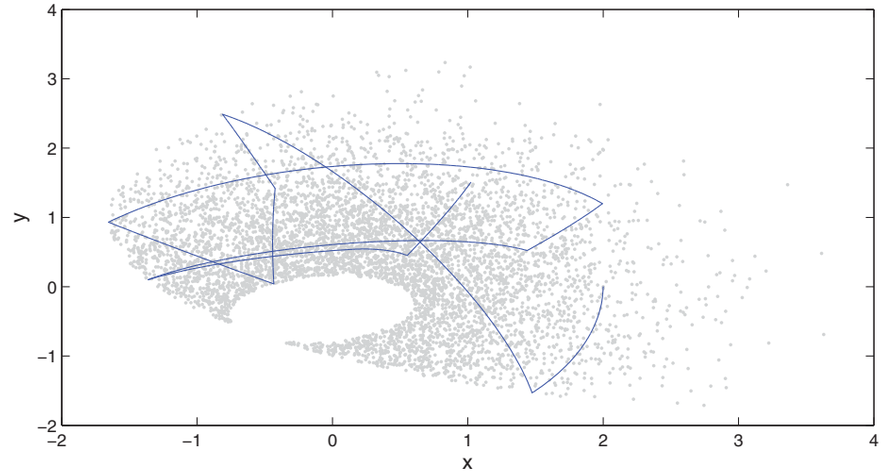
$$\begin{aligned} \frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i} \end{aligned} \quad \Longrightarrow \quad \begin{aligned} \frac{dq_i}{dt} &= [M^{-1}p]_i \\ \frac{dp_i}{dt} &= -\frac{\partial U}{\partial q_i} \end{aligned}$$

- Exact HMC: Analytic solution only in special cases, e.g. for (truncated) normal target we get:

$$q(t) = r\cos(a+t), \quad p(t) = -r\sin(a+t)$$

Application: truncated normal distributions

- See Pakman and Paninski (2014)
- Truncated normal arise in many practical contexts:
  - Probit and tobit models
  - Bayesian splines for positive functions
  - Bayesian lasso



$$y_i = \text{sign}(w_i)$$

$$w_i = -\mathbf{z}_i \cdot \boldsymbol{\beta} + \varepsilon_i$$

$$\varepsilon_i \sim \mathcal{N}(0, 1)$$

Exact HMC: invariance

- MCMC kernel is a non-reversible
- Given by a Dirac delta: $k(z, dz') = \delta_{\Phi(z)}(dz')$
- Invariance equivalent to:
  - given $Z \sim$ extended target π'
    $\pi'(x, v) = \pi(x) \times \text{normal}(v)$
  - Define $Y = \Phi(Z)$
  - Do we have $Y \sim \pi'$?

Exact HMC: invariance

- By change of variable formula, break into two factors:

$$f_Y(y) = f_Z(\Phi^{-1}(y)) \, |\det J_{\Phi^{-1}}(y)|$$

hence ingredient to show Y ~ π' are:

- Φ invertible (yes, set v ⟵ -v)
- *Conservation of Hamiltonian*: first factor is constant
- *Volume preservation*: second factor is constant

Conservation of Hamiltonian

- Want $f(z) = f(\Phi(z))$
- Enough: no infinitesimal Hamiltonian changes, $H' = 0$
- Use total derivative identity

$$\frac{dH}{dt} = \sum_{i=1}^{d} \left[\frac{dq_i}{dt} \frac{\partial H}{\partial q_i} + \frac{dp_i}{dt} \frac{\partial H}{\partial p_i} \right]$$

- Then substitute our choice of the differential equation:

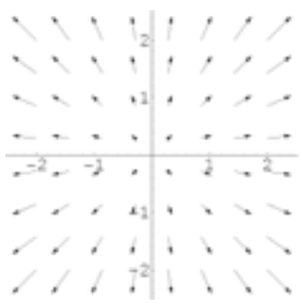
$$\begin{aligned} \frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i} \end{aligned} \implies \sum_{i=1}^{d} \left[\frac{dq_i}{dt} \frac{\partial H}{\partial q_i} + \frac{dp_i}{dt} \frac{\partial H}{\partial p_i} \right] = \sum_{i=1}^{d} \left[\frac{\partial H}{\partial p_i} \frac{\partial H}{\partial q_i} - \frac{\partial H}{\partial q_i} \frac{\partial H}{\partial p_i} \right] = 0$$

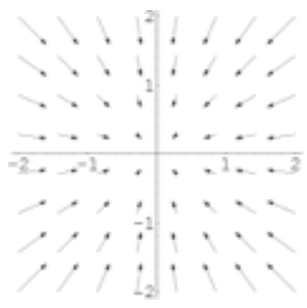
Volume preservation

The preservation of volume by Hamiltonian dynamics can be proved in several ways. One is to note that the divergence of the vector field defined by equations (2.1) and (2.2) is zero, which can be seen as follows:

$$\sum_{i=1}^{d}\left[\frac{\partial}{\partial q_i}\frac{dq_i}{dt}+\frac{\partial}{\partial p_i}\frac{dp_i}{dt}\right]=\sum_{i=1}^{d}\left[\frac{\partial}{\partial q_i}\frac{\partial H}{\partial p_i}-\frac{\partial}{\partial p_i}\frac{\partial H}{\partial q_i}\right]=\sum_{i=1}^{d}\left[\frac{\partial^2 H}{\partial q_i \partial p_i}-\frac{\partial^2 H}{\partial p_i \partial q_i}\right]=0 \quad (2.13)$$

A vector field with zero divergence can be shown to preserve volume (Arnold, 1989).

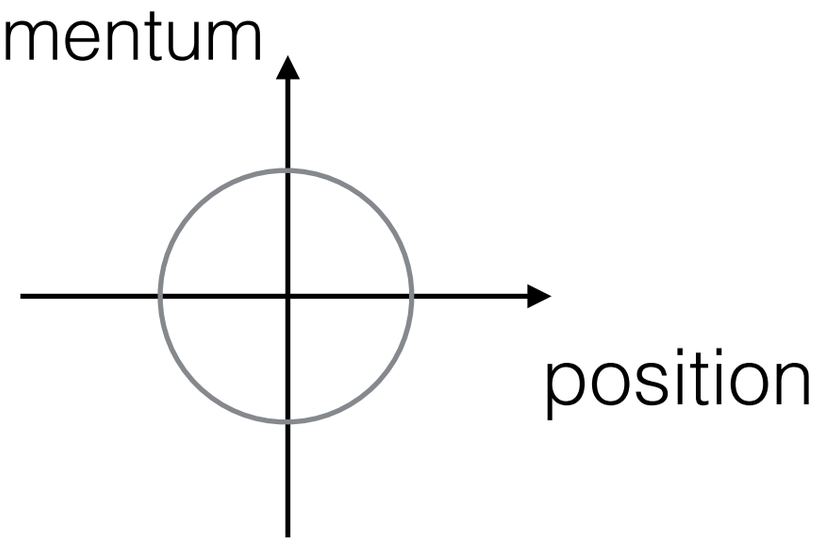




- See Neal (2012). MCMC using Hamiltonian dynamics for another, more direct argument

Exact HMC: irreducibility

- Easy to see non irreducible in phase space



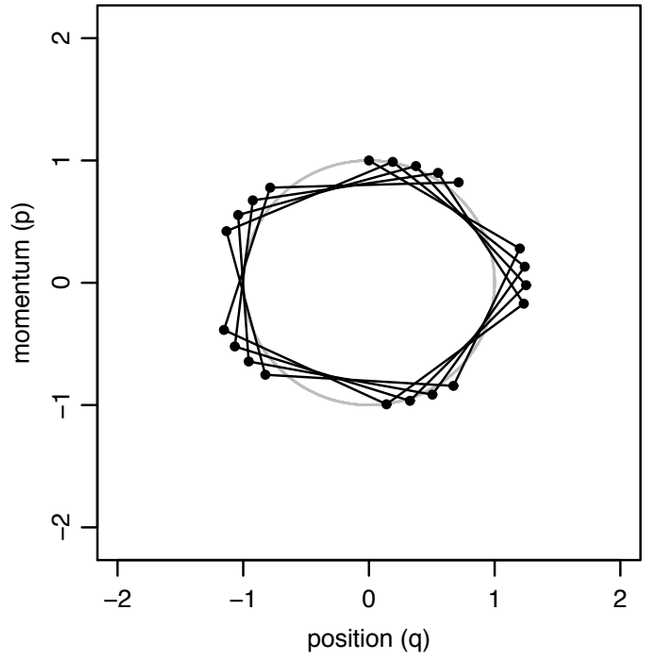
- Solution: refresh momentum

Leap-frog HMC

- We can't simulate the exact Hamiltonian flow for most targets of interest.
- Idea:
  - solve the differential equation using numerical methods and initial condition given by current point
    - can be done so that volume still preserved (e.g. with leap-frog integrator)
  - Hamiltonian no longer exactly preserved, so use MH to accept-reject

Leap-frog HMC

- Numerical solution example:



- Algorithm:

$$p_i(t+\varepsilon/2) \;=\; p_i(t) \;-\; (\varepsilon/2)\,\frac{\partial U}{\partial q_i}(q(t))$$

$$q_i(t+\varepsilon) \;=\; q_i(t) \;+\; \varepsilon\,\frac{p_i(t+\varepsilon/2)}{m_i}$$

$$p_i(t+\varepsilon) \;=\; p_i(t+\varepsilon/2) \;-\; (\varepsilon/2)\,\frac{\partial U}{\partial q_i}(q(t+\varepsilon))$$

- Properties:
  - reversibility
  - symplecticness