

Monte Carlo Markov Sampling

Elena Sellentin

Université de Genève
Département de Physique Théorique

MACSS Summer school
8th-12th May 2017

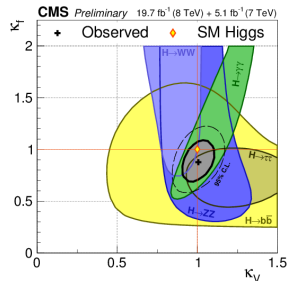
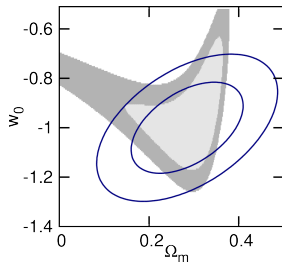
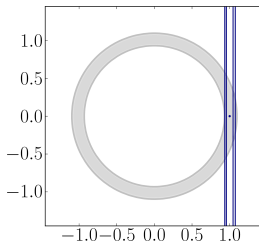
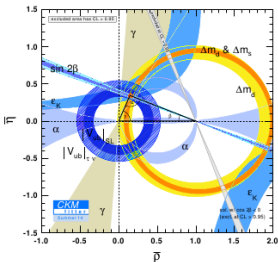
Bayesian Inference

$$\mathcal{P}(\theta, \mathcal{M}|x) = \frac{L(x|\theta, \mathcal{M})\pi(\theta)}{\epsilon(x|\mathcal{M})} \quad (1)$$

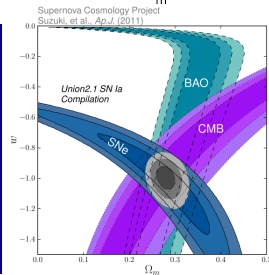
- $\mathcal{P}(\theta, \mathcal{M}|x)$: the posterior.
- $L(x|\theta, \mathcal{M})$: the likelihood.
- $\pi(\theta)$: the priors.
- $\epsilon(x|\mathcal{M})$: the evidence ('marginal likelihood').

→ Even for a known likelihood, the posterior can be difficult to obtain. Need sampling techniques (MCMC, Gibbs...) & DALI.

Examples



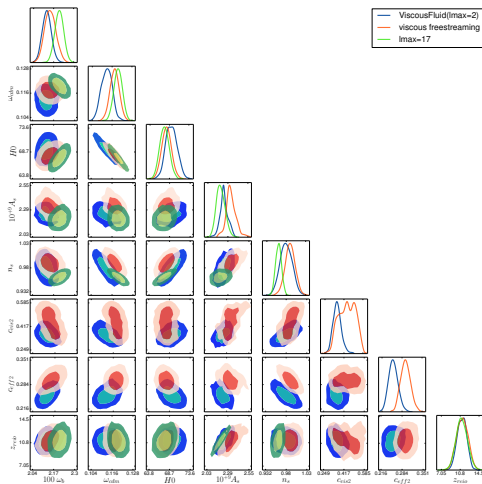
$$L = \frac{1}{\sqrt{|2\pi F^{-1}|}} \exp\left(-\frac{1}{2}[\Delta p]^T F[\Delta p]\right) + ?$$



2 Options:

MCMC (today)
or
DALI & Fisher matrix (tomorrow)

Triangle plots



Sellentin & Durrer (2015)



The Metropolis-Hastings Algorithm

Detailed Balance

Equilibrium between the occupation of two states \mathcal{P}_i is reached if

$$r_{i \rightarrow j} \mathcal{P}_i = r_{j \rightarrow i} \mathcal{P}_j, \quad (2)$$

and the transition probability from \mathcal{P}_i to state \mathcal{P}_j has rate $r_{i \rightarrow j}$.

⇒ Same principle as e.g. in photon emission/absorption from electronic shells in atoms [search for 'Einstein coefficients']

→ Let's turn the transition between two states into a running chain.

Metropolis Hastings Algorithm

- 1 Provide a guess for a Gaussian approximation $\mathcal{G}_P(\theta)$ to the posterior (Fisher matrix/previous sample covariance matrix)
- 2 FOR $i = 0$ TO N_{MCMC}
if $i = 0$ evaluate the posterior P at point θ_0 in parameter space.
Else use the current θ_i of the chain.
- 3 Draw a random step $\Delta\theta_i \sim \mathcal{G}_P(\theta)$ and calculate $R = \frac{P(\theta_i + \Delta\theta_i)}{P(\theta_i)}$.
- 4 IF $R > 1$, then the posterior probability at the new point $\theta_i + \Delta\theta_i$ is larger than the old probability; the new point is then accepted as $\theta_{i+1} = \theta_i + \Delta\theta_i$.
- 5 IF $R < 1$, then draw $\alpha \sim \text{Uniform}[0, 1]$.
IF $\alpha > R$, then $\theta_{i+1} = \theta_i$, i.e. the point $\theta_i + \Delta\theta_i$ is rejected because it has too low a probability.
IF, however, $\alpha < R$, then $\theta_{i+1} = \theta_i + \Delta\theta_i$, i.e. the trial point is accepted because statistical equilibrium demands a population of the low- \mathcal{L} states as well.
- 6 Store all points θ_i in a chain.

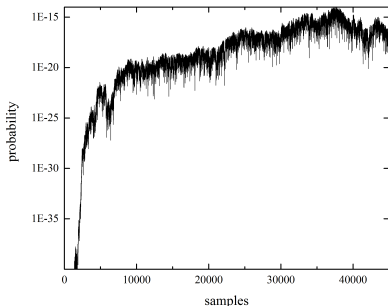
Convergence

- If the chain ran for long enough ('has converged'), then $n(\boldsymbol{\theta}) \propto \mathcal{P}(\boldsymbol{\theta})$.
- Detailed balance: reacts to $\mathcal{P}(\boldsymbol{\theta}_i)/\mathcal{P}(\boldsymbol{\theta}_j)$.
- $\Rightarrow \mathcal{N}$ is missing (but not needed for parameter inference).

Burn-in period

Reaching equilibrium needs interactions.

- Thermodynamics: put cold object into warmer environment
- Radiation physics: put a phosphorescent object into the dark
- MCMC: burn-in period (searching for the peak)

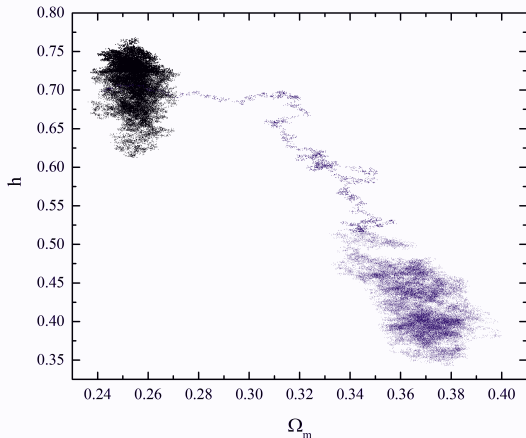


Plot: K. Wolz

Burn-in

Cut away the unrepresentative burn-in period:

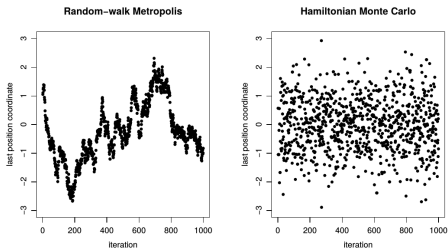
- Rat tails in likelihood plots
- log-likelihood steadily increases



Plot: K. Wolz

Monitor convergence

- Trade-off: acceptance rate \leftrightarrow correlation between samples
- measure the correlation length of chains
- thin highly correlated chains
- if you have multiple peaks, sample all of them



Neil 2012, arXiv: 1206.1901

Convergence Diagnostics

- Plot the samples without gridding.
- Find the best-fitting values of all your chains and compare.
- Is half your chain still representative of the posterior?
- Does binning still influence the credibility intervals?
- Is the chain squeezed against any prior-boundaries?
- Compute the auto-correlation length of your chain for all parameters & potentially thin until sequential points are uncorrelated ('potentially' depends on posterior-shape)

$$C(T) \approx \frac{1}{M-T} \sum_{m=1}^{M-T} [f(T+m) - \bar{f}][f(m) - \bar{f}]$$

Convergence Diagnostics

Gelman-Rubin Test: Intra-Chain variance vs. Inter-chain variance.

- Run M different chains with different starting points, let $m \in [1, M]$.
- m th chain: $\theta_1^m, \theta_2^m, \theta_3^m, \dots, \theta_{N_m}^m$.
- Discard the burnins.

- Calculate for each parameter θ , the posterior mean

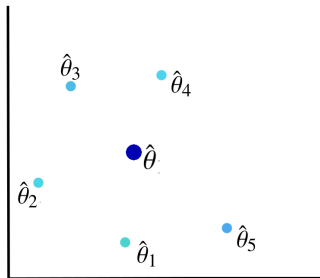
$$\hat{\theta}_m = \frac{1}{N_m} \sum_i^{N_m} \theta_i^m,$$

- ...and the intra-chain variance

$$\sigma_m^2 = \frac{1}{N_m - 1} \sum_i^{N_m} (\theta_i^m - \hat{\theta}_m)^2.$$

- Calculate $\hat{\theta}$, the mean of all chains

$$\hat{\theta} = \frac{1}{M} \sum_m^M \hat{\theta}_m.$$



Gelman-Rubin cntd.

- Compute how the individual means vary around the joint mean

$$B = \frac{N}{M-1} \sum_{m=1}^M (\hat{\theta}_m - \hat{\theta})^2$$

- Compute the averaged variances of the chains

$$W = \frac{1}{M} \sum_{m=1}^M \sigma_m^2$$

- Define $\hat{V} = \frac{N-1}{N} W + \frac{M+1}{MN} B$; under convergence, this is an unbiased estimator of the true variance. But if the chains have converged, then W is *also* an unbiased estimate of the true variance. Hence...

- ...test whether $R = \sqrt{\hat{V}/W} \approx 1$.

If it is not, convergence has not been reached.

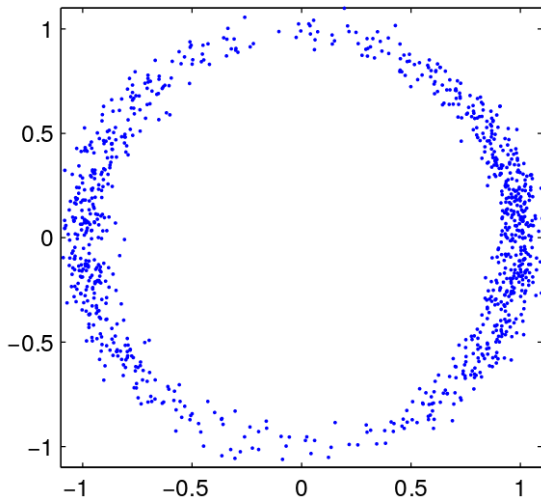
- Various refinements exist, see Gelman & Rubin (1992), Brooks & Gelman (1997).



Hamilton Monte Carlo

A clever way of distributing MH-samples

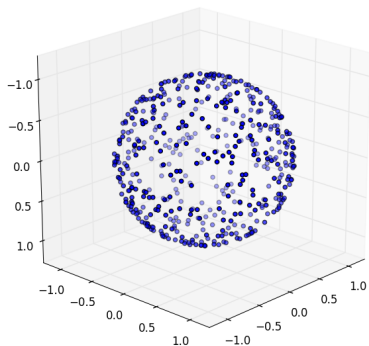
Difficult posterior shapes



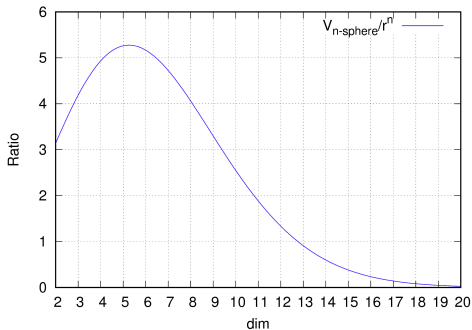
Hajian 2006

The Curse of Dimensionality

With increasing d , have ever more possibilities to go 'wrong' with a d -dimensional random step.



Credit (left): Wikipedia



⇒ we need some guidance

Hamiltonian: (governs the evolution of trajectories in phase-space)

$$H(\boldsymbol{\theta}, \mathbf{u}) = U(\boldsymbol{\theta}) + K(\mathbf{u}). \quad (3)$$

Introducing a potential $U(\boldsymbol{\theta})$

$$U(\boldsymbol{\theta}) = -\log P(\boldsymbol{\theta}). \quad (4)$$

Introduce kinetic energy

$$K(\mathbf{u}) = \mathbf{u}^T \mathbf{u} / 2, \quad \mathbf{u} \sim \mathcal{G}(\mathbf{0}, \mathbf{I}), \quad (5)$$

Know Hamiltonian equations of motion:

$$\dot{\boldsymbol{\theta}} = \mathbf{u}, \quad u_i = -\frac{\partial H}{\partial \theta_i}. \quad (6)$$

Solving Hamiltonian equations: deterministic.

Need a source of randomness \Rightarrow randomize initial velocities.

Initial $\mathbf{u} \sim \mathcal{G}$.

$$\exp(-H(\boldsymbol{\theta}, \mathbf{u})) = P(\boldsymbol{\theta})\mathcal{G}(0, \mathcal{I}) \quad (7)$$

So if the auxiliary velocities \mathbf{u} are marginalized over, which is equivalent to not protocolling them in the chain, then sampling $\exp(-H)$ samples $P(\boldsymbol{\theta})$.

Hamilton Monte Carlo Algorithm

- 1 For $i = 0$ TO N_{MCMC}
if $i = 0$, choose a starting point θ_0 ,
else use the current θ_i of the chain.
- 2 Draw a random velocity $\mathbf{u}_i \sim \mathcal{G}(0, \mathcal{I})$.
Leapfrog loop
 - 1 Use θ_i and \mathbf{u}_i as initial conditions for the Hamiltonian equations of motions.
 - 2 For $j = 0$ to N_L
make leapfrog steps that update $(\theta_j, \mathbf{u}_j) \rightarrow (\theta_{j+1}, \mathbf{u}_{j+1})$
- 3 Having arrived at $(\theta_{N_L}, \mathbf{u}_{N_L})$, calculate
 $R = \exp[-H(\theta_i, \mathbf{u}_i) + H(\theta_{N_L}, \mathbf{u}_{N_L})]$.
- 4 If $R > 1$, the new point is accepted, $\theta_{i+1} = \theta_i$.
- 5 If $R < 1$, draw $\alpha \sim \text{Uniform}[0, 1]$.
If $\alpha > R$, then $\theta_{i+1} = \theta_i$, i.e. the trial point θ_{N_L} is rejected.
If $\alpha < R$, then $\theta_{i+1} = \theta_{N_L}$, i.e. the trial point is accepted.

HMC vs. MH

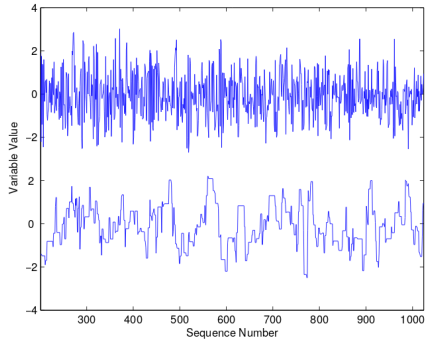


FIG. 1: Samples drawn from an isotropic six-dimensional Gaussian distribution using the HMC (top) and the Metropolis algorithm with optimal step-size (bottom).

Hajian 2006