Tutorial on Markov Chain Monte Carlo

Kenneth M. Hanson

Los Alamos National Laboratory

This presentation available under http://www.lanl.gov/home/kmh/

Acknowledgements

- MCMC experts
  - Julian Besag, Jim Guberantus, John Skilling, Malvin Kalos
- General discussions
  - Greg Cunningham, Richard Silver

Problem statement

- Parameter space of n dimensions represented by vector $\mathbf{x}$
- Given an "arbitrary" **target** probability density function (pdf), $q(\mathbf{x})$, draw a set of samples $\{\mathbf{x}_k\}$ from it
- Only requirement typically is that, given $\mathbf{x}$, one be able to evaluate $Cq(\mathbf{x})$, where C is an unknown constant
  - MCMC algorithms do not typically require knowledge of the normalization constant of the target pdf; from now on the multiplicative constant C will not be made explicit
- Although focus here is on continuous variables, MCMC can be applied to discrete variables as well

Uses of MCMC

- Permits evaluation of the expectation values of functions of **x**, e.g.,

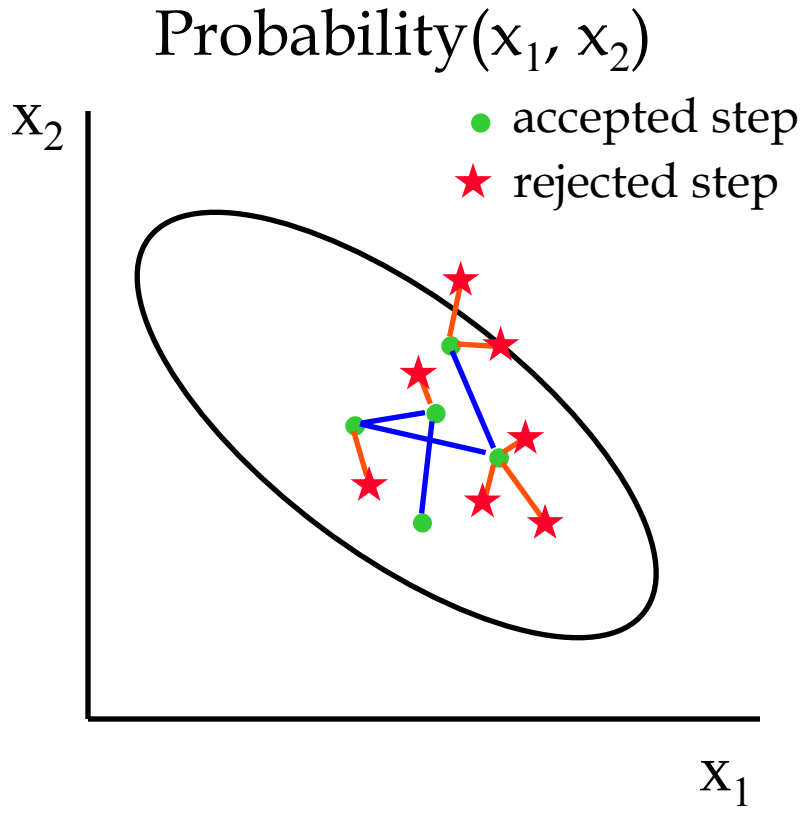
$$\langle f(\mathbf{x})\rangle = \int f(\mathbf{x})\, q(\mathbf{x})\, d\mathbf{x} \cong (1/\mathrm{K})\, \Sigma_k\, f(\mathbf{x}_k)$$

  - typical use is to calculate mean $\langle \mathbf{x} \rangle$ and variance $\langle (\mathbf{x} - \langle \mathbf{x} \rangle)^2 \rangle$
- Also useful for evaluating integrals, such as the partition function for properly normalizing the pdf
- Dynamic display of sequences provides visualization of uncertainties in model and range of model variations
- Automatic marginalization; when considering any subset of parameters of an MCMC sequence, the remaining parameters are marginalized over

Markov Chain Monte Carlo

Generates sequence of random samples from an arbitrary probability density function

- Metropolis algorithm:
  - draw trial step from symmetric pdf, i.e., $t(\Delta \mathbf{x}) = t(-\Delta \mathbf{x})$
  - accept or reject trial step
  - simple and generally applicable
  - relies only on calculation of target pdf for any **x**



Metropolis algorithm

- Select initial parameter vector $\mathbf{x}_0$
- Iterate as follows: at iteration number k

  (1) create new trial position $\mathbf{x}^* = \mathbf{x}_k + \Delta\mathbf{x}$,
  where $\Delta\mathbf{x}$ is randomly chosen from $t(\Delta\mathbf{x})$

  (2) calculate ratio $r = q(\mathbf{x}^*)/q(\mathbf{x}_k)$

  (3) accept trial position, i.e. set $\mathbf{x}_{k+1} = \mathbf{x}^*$
  if $r \geq 1$ or with probability r, if $r < 1$
  otherwise stay put, $\mathbf{x}_{k+1} = \mathbf{x}_k$

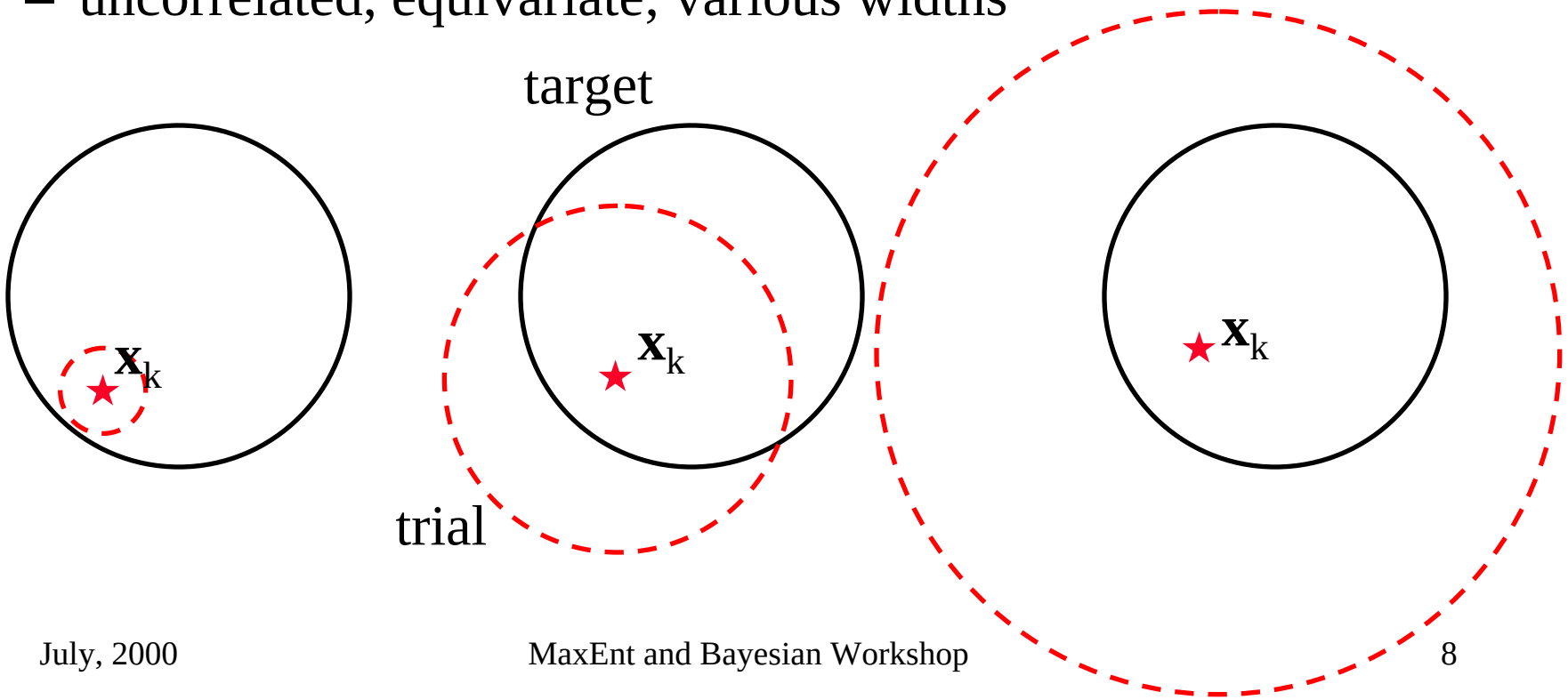
- Requires only computation of $q(\mathbf{x})$
- Creates Markov chain since $\mathbf{x}_{k+1}$ depends only on $\mathbf{x}_k$

Choice of trial distribution

- Loose requirements on trial distribution *t*()
  - stationary; independent of position
- Often used functions include
  - *n*-D Gaussian, isotropic and uncorrelated
  - *n*-D Cauchy, isotropic and uncorrelated
- Choose width to “optimize” MCMC efficiency
  - rule of thumb: aim for acceptance fraction of about 25%

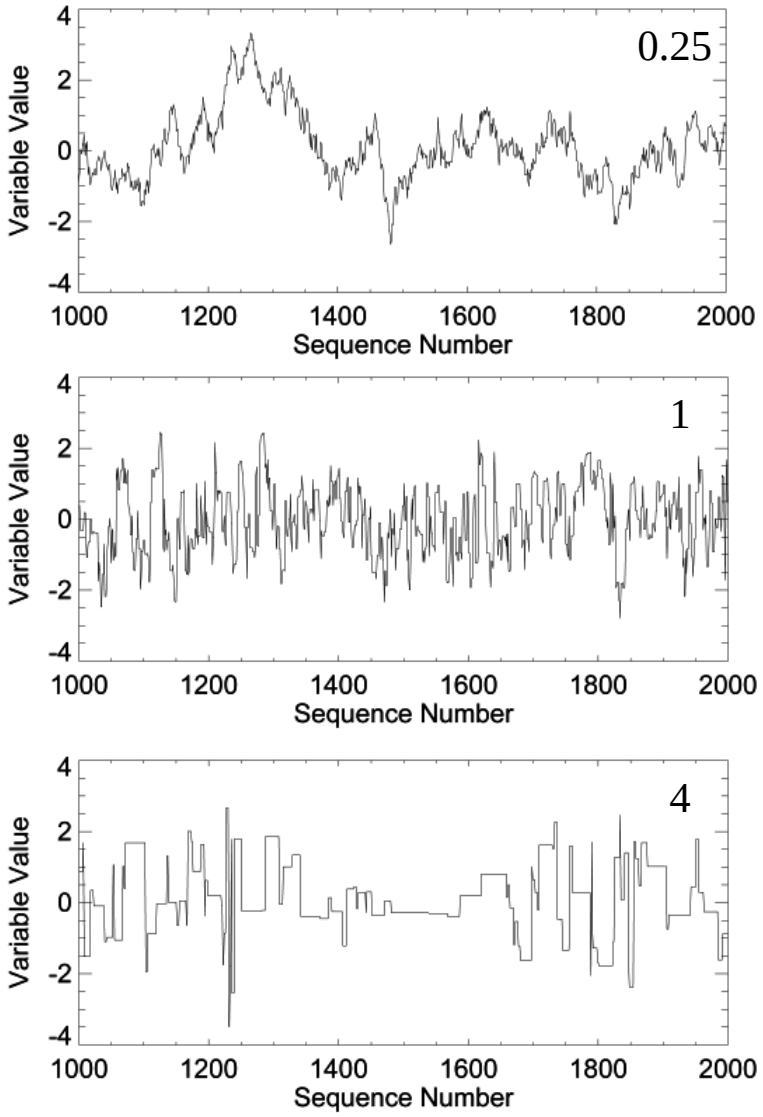
Experiments with the Metropolis algorithm

- Target distribution $q(\mathbf{x})$ is n dimensional Gaussian
  - uncorrelated, univariate (isotropic with unit variance)
  - most generic case
- Trial distribution $t(\Delta\mathbf{x})$ is n dimensional Gaussian
  - uncorrelated, equivariate; various widths



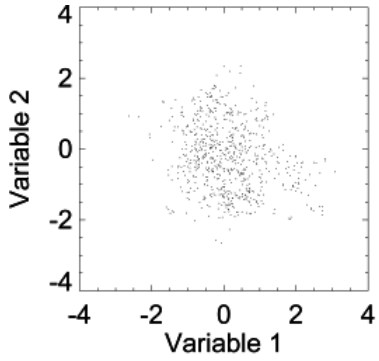
MCMC sequences for 2D Gaussian

- results of running Metropolis with ratios of width of trial to target of 0.25, 1, and 4
- when trial pdf is much smaller than target pdf, movement across target pdf is slow
- when trial width same as target, samples seem to sample target pdf better
- when trial much larger than target, trials stay put for long periods, but jumps are large
  - This example from Hanson and Cunningham (SPIE, 1998)

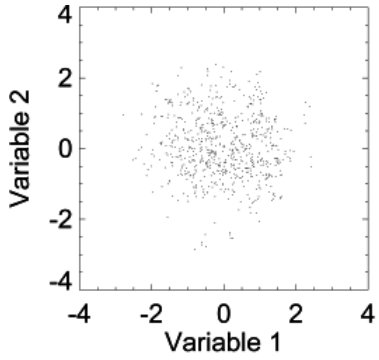


MCMC sequences for 2D Gaussian

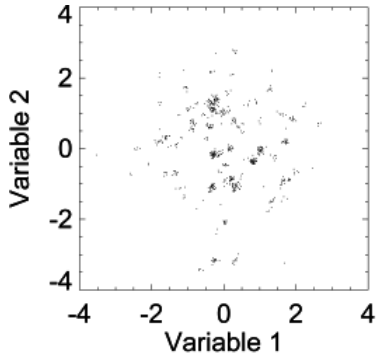
- results of running Metropolis with ratios of width of trial to target of 0.25, 1, and 4
- display accumulated 2D distribution for 1000 trials
- viewed this way, it is difficult to see difference between top two images
- when trial pdf much larger than target, fewer splats, but further apart



0.25



1



4

MCMC - autocorrelation and efficiency

- In MCMC sequence, subsequent parameter values are usually correlated
- Degree of correlation quantified by autocorrelation function:

$$\rho(l) = \frac{1}{N}\sum_{i=1}^{N} y(i)\,y(i-l)$$

where *y(x)* is the sequence and *l* is lag

- For Markov chain, expect exponential

$$\rho(l) = \exp\left[-\left|\frac{l}{\lambda}\right|\right]$$

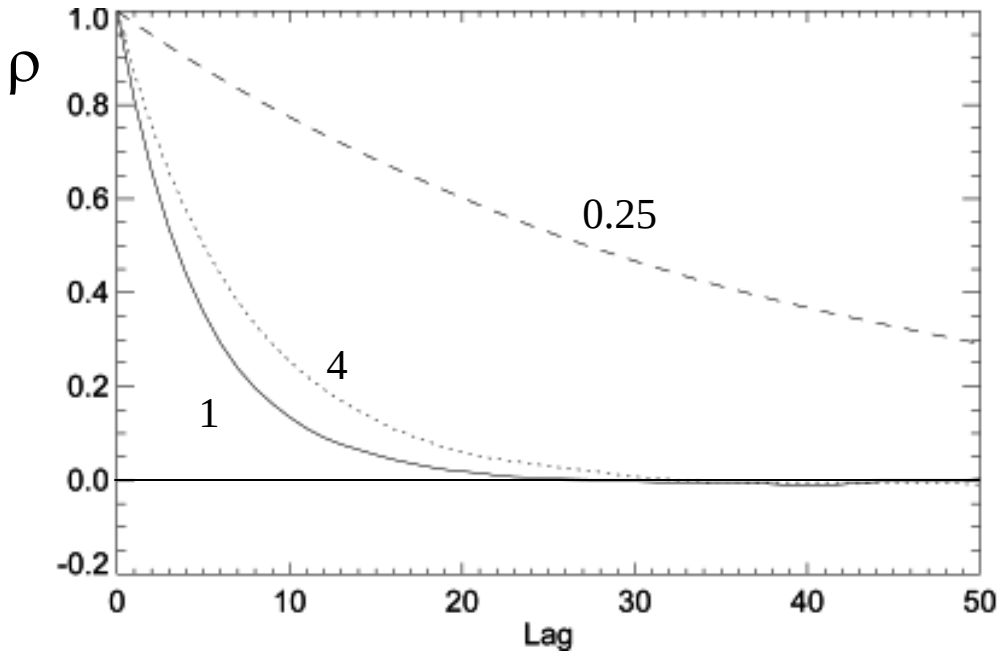
- Sampling efficiency is

$$\eta = [1 + 2\sum_{l=1}^{\infty} \rho(l)]^{-1} = \frac{1}{1+2\lambda}$$

- In other words, η^{-1} iterates required to achieve one statistically independent sample

Autocorrelation for 2D Gaussian

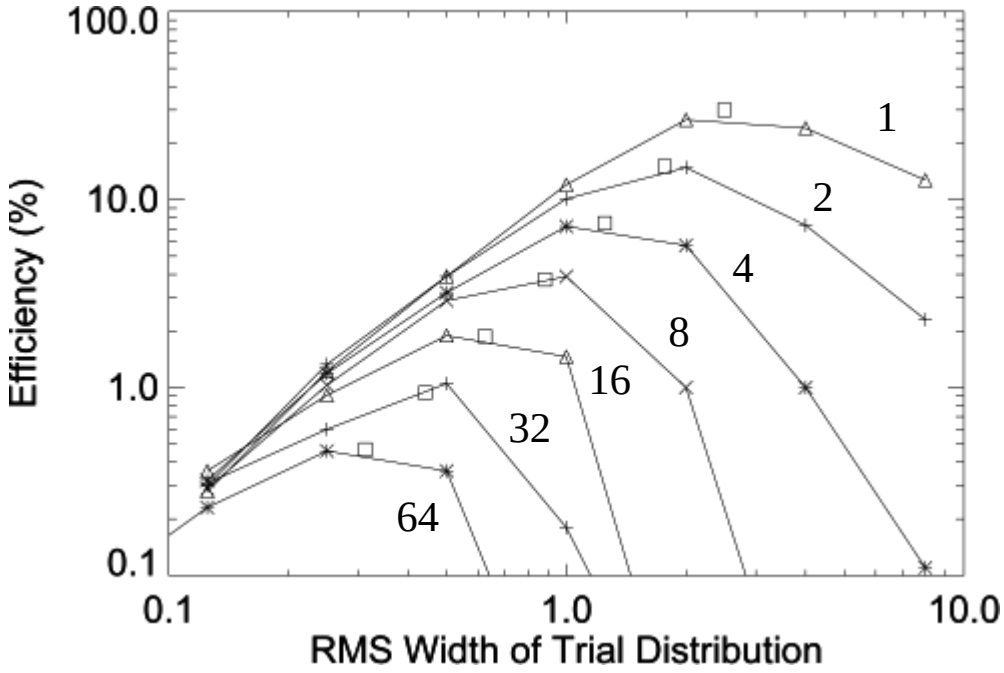
- plot confirms that the autocorrelation drops slowly when the trial width is much smaller than the target width; MCMC efficiency is poor
- best efficiency is when trial about same size as target (for 2D)



Normalized autocovariance for various widths of trial pdf relative to target: 0.25, 1, and 4

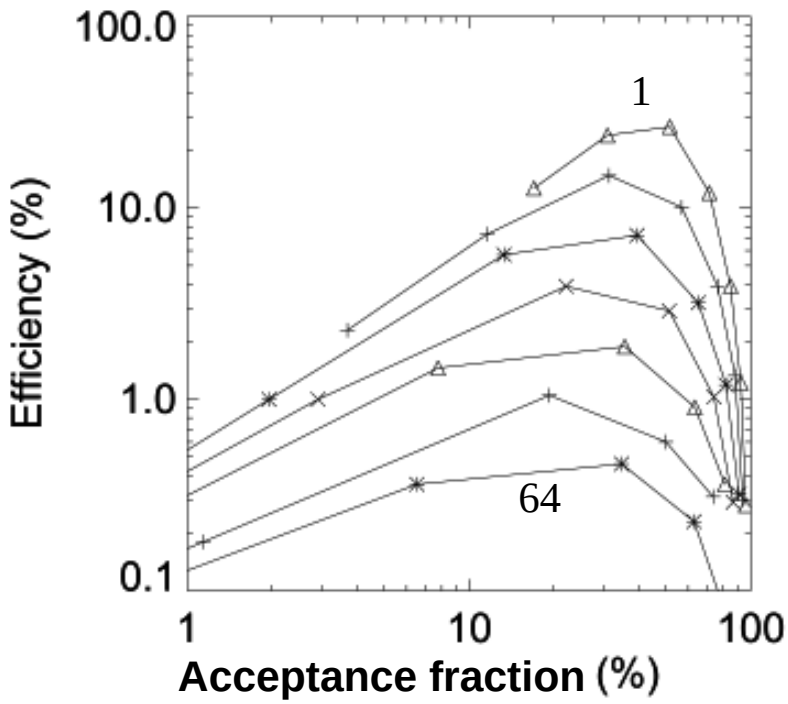
Efficiency as function of width of trial pdf

- for univariate Gaussians, with 1 to 64 dimensions
- efficiency as function of width of trial distributions
- boxes are predictions of optimal efficiency from diffusion theory [A. Gelman, et al., 1996]
- efficiency drops reciprocally with number of dimensions



Efficiency as function of acceptance fraction

- for univariate Gaussians, with 1 to 64 dimensions
- efficiency as function of acceptance fraction
- best efficiency is achieved when about 25% of trials are accepted for a moderate number of dimensions

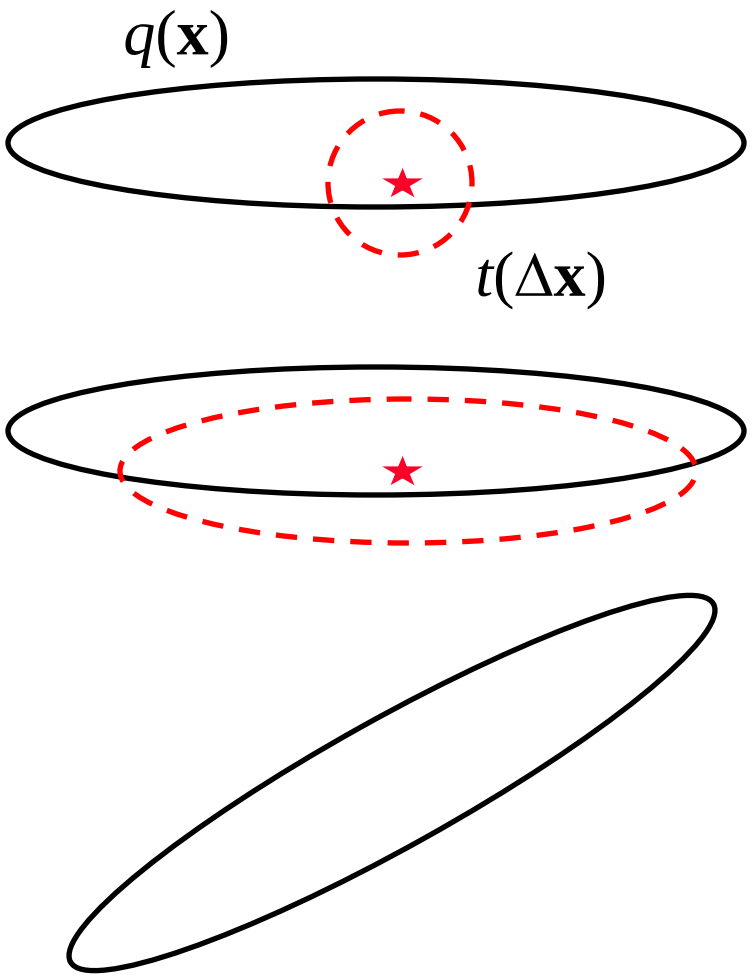


Efficiency of Metropolis algorithm

- Results of experimental study agree with predictions from diffusion theory (A. Gelman et al., 1996)
- Optimum choice for width of Gaussian trial distribution occurs for acceptance fraction of about 25% (but is a weak function of number of dimensions)
- Optimal statistical efficiency: $\eta \sim 0.3/n$
  - for simplest case of uncorrelated, equivariate Gaussian
  - correlation and variable variance generally decreases efficiency

Further considerations

- When target distribution $q(\mathbf{x})$ not isotropic
  - difficult to accommodate with isotropic $t(\Delta\mathbf{x})$
  - each parameter can have different efficiency
  - desirable to vary width of different $t(\mathbf{x})$ to approximately match $q(\mathbf{x})$
  - recovers efficiency of univariate case
- When $q(\mathbf{x})$ has correlations
  - $t(\mathbf{x})$ should match shape of $q(\mathbf{x})$

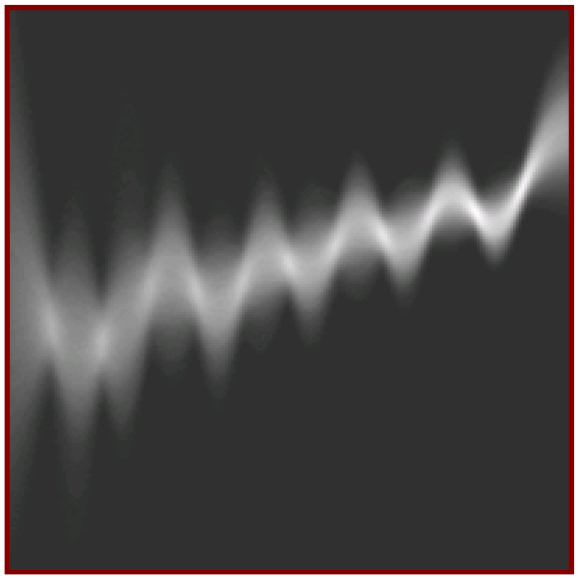


MCMC - Issues

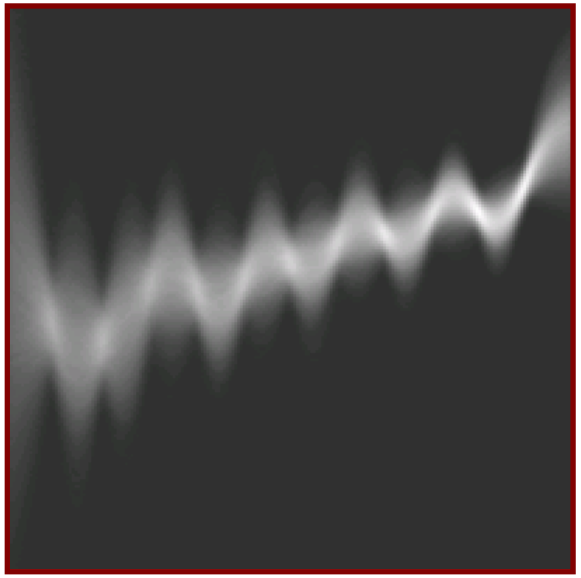
- Identification of convergence to target pdf
  - is sequence in thermodynamic equilibrium with target pdf?
  - validity of estimated properties of parameters (covariance)
- Burn in
  - at beginning of sequence, may need to run MCMC for awhile to achieve convergence to target pdf
- Use of multiple sequences
  - different starting values can help confirm convergence
  - natural choice when using computers with multiple CPUs
- Accuracy of estimated properties of parameters
  - related to efficiency, described above
- Optimization of efficiency of MCMC

MCMC - Multiple runs

– Multiple runs starting with different random number seed confirm MCMC sequences have converged to the target pdf



First MCMC sequence



Second, independent MCMC sequence

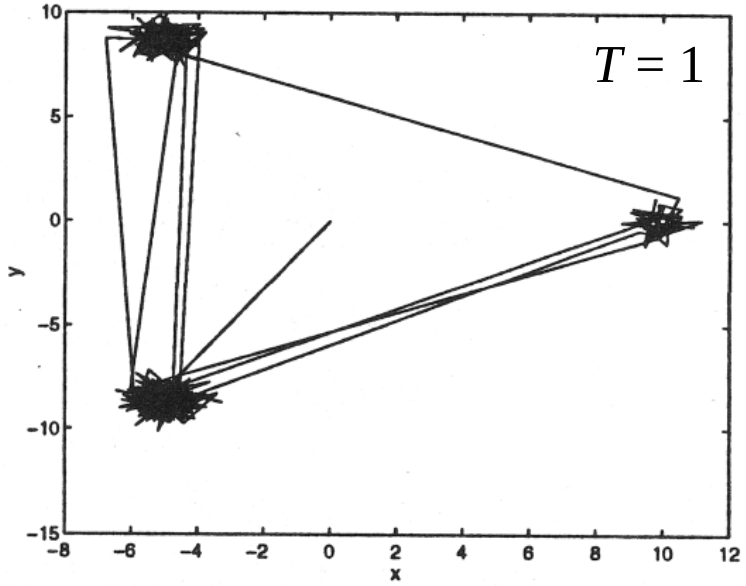
Examples of multiple MCMC runs from my talk on the analysis of Rossi data (http://public.lanl.gov/kmh/talks/maxent00a.pdf)

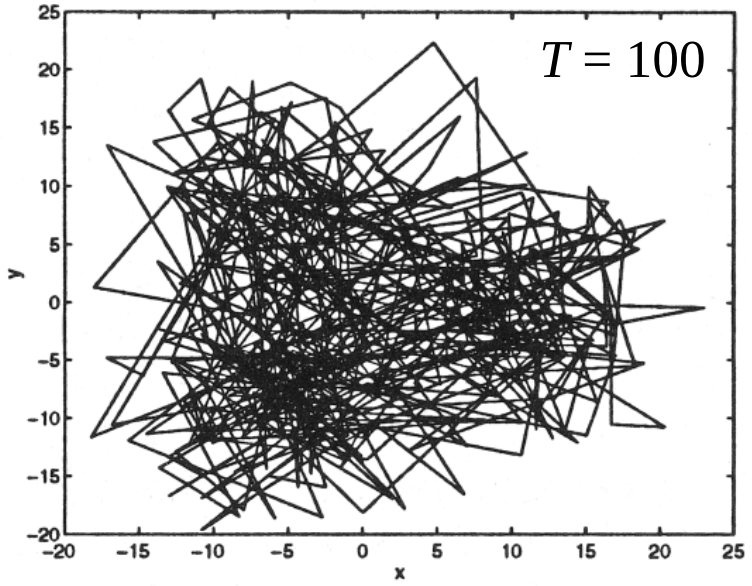
Annealing

- Introduction of fictitious temperature
  - define functional $\varphi(\mathbf{x})$ as minus-logarithm of target probability
    $$\varphi(\mathbf{x}) = - \log(q(\mathbf{x}))$$
  - scale φ by an inverse "temperature" to form new pdf
    $$q'(\mathbf{x}, T) = \exp[- T^{-1} \varphi(\mathbf{x})]$$
  - $q'(\mathbf{x}, T)$ is flatter than $q(\mathbf{x})$ for $T > 1$ (called annealing)
- Uses of annealing (also called tempering)
  - allows MCMC to move between multiple peaks in $q(\mathbf{x})$
  - simulated annealing optimization algorithm (takes lim $T \rightarrow 0$)

Annealing to handle multiple peaks

- Example - target distribution is three narrow, well separated peaks
- For original distribution (T = 1), an MCMC run of 10000 steps rarely moves between peaks
- At temperature T = 100 (right), MCMC moves easily between peaks and through surrounding regions
- from M-D Wu and W. J. Fitzgerald, *Maximum Entropy and Bayesian Methods* (1996)





Other MCMC algorithms

- Gibbs
  - vary only one component of **x** at a time
  - draw new value of x_j from conditional $q(x_j | x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots)$
- Metropolis-Hastings
  - allows use of nonsymmetric trial functions, $t(\Delta\mathbf{x}; \mathbf{x}_k)$, suitably chosen to improve efficiency
  - use $r = [t(\Delta\mathbf{x}; \mathbf{x}_k)\, q(\mathbf{x}^*)] / [t(-\Delta\mathbf{x}; \mathbf{x}^*)\, q(\mathbf{x}_k)]$
- Langevin technique
  - uses gradient* of minus-log-prob to shift trial function towards regions of higher probability
  - uses Metropolis-Hastings

* adjoint differentiation affords efficient gradient calculation

Hamiltonian hybrid algorithm

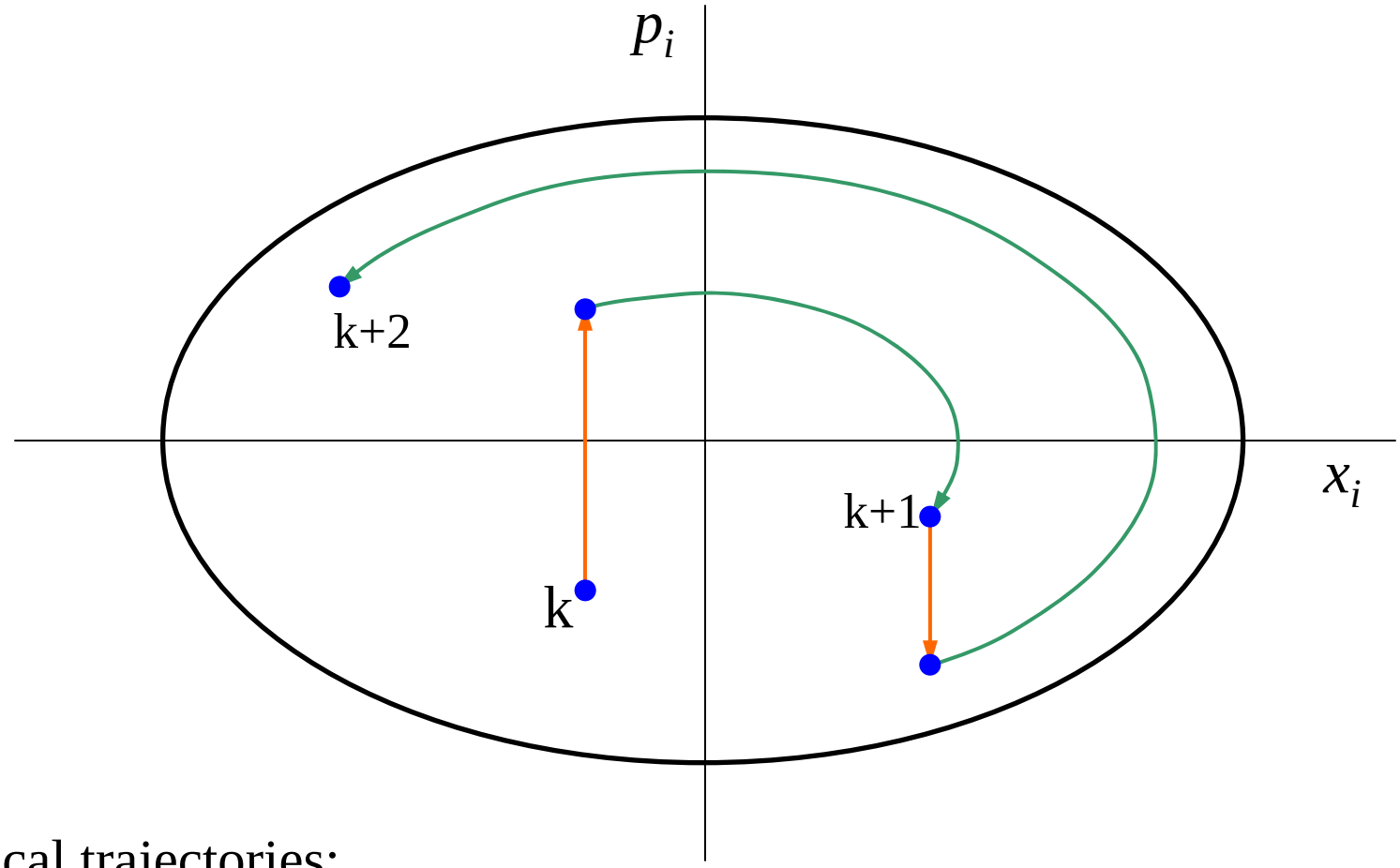
- Hamiltonian hybrid algorithm
  - called hybrid because it alternates Gibbs & Metropolis steps
  - associate with each parameter x_i a momentum p_i
  - define a Hamiltonian

    $H = \varphi(\mathbf{x}) + \Sigma\, p_i^2/(2\, m_i)$; where $\varphi = -\log(q(\mathbf{x}))$
  - new pdf:

    $$q'(\mathbf{x}, \mathbf{p}) = \exp(-H(\mathbf{x}, \mathbf{p})) = q(\mathbf{x}) \exp(-\Sigma\, p_i^2/(2\, m_i))$$
  - can easily move long distances in ($\mathbf{x}$, $\mathbf{p}$) space at constant H using Hamiltonian dynamics, so Metropolis step is very efficient
  - uses gradient* of φ (minus-log-prob)
  - Gibbs step in $\mathbf{p}$ for constant $\mathbf{x}$ is easy
  - efficiency may be better than Metropolis for large dimensions

* adjoint differentiation affords efficient gradient calculation

Hamiltonian hybrid algorithm



Typical trajectories:

red path - Gibbs sample from momentum distribution

green path - trajectory with constant H, follow by Metropolis

Conclusions

- MCMC provides good tool for exploring the posterior and hence for drawing inferences about models and parameters
- For valid results, care must be taken to
  - verify convergence of the sequence
  - exclude early part of sequence, before convergence reached
  - be wary of multiple peaks that need to be sampled
- For good efficiency, care must be taken to
  - adjust the size and shape of the trial distribution; rule of thumb is to aim for 25% trial acceptance for $5 < n < 100$
- A lot of research is happening - don't worry, be patient

Short Bibliography

- "Posterior sampling with improved efficiency," K. M. Hanson and G. S. Cunningham, *Proc. SPIE* **3338**, 371-382 (1998); includes introduction to MCMC
- *Markov Chain Monte Carlo in Practice*, W. R. Gilks et al., (Chapman and Hall, 1996); excellent general-purpose book
- "Efficient Metropolis jumping rules," A. Gelman et al, in *Bayesian Statistics* **5**, J. M. Bernardo et al., (Oxford Univ., 1996); diffusion theory
- "Bayesian computation and stochastic systems," J. Besag et al., *Stat. Sci.* **10**, 3-66 (1995); MCMC applied to image analysis
- *Bayesian Learning for Neural Networks*, R. M. Neal, (Springer, 1996); Hamiltonian hybrid MCMC
- "Bayesian multinodal evidence computation by adaptive tempered MCMC," M-D Wu and W. J. Fitzgerald, in *Maximum Entropy and Bayesian Methods*, K. M. Hanson and R. N. Silver, eds., (Kluwer Academic, 1996); annealing
- "Inversion based on complex simulations," K. M. Hanson, *Maximum Entropy and Bayesian Methods*, G. J. Erickson et al., eds., (Kluwer Academic, 1998); describes adjoint differentiation and its usefulness

More articles and slides under http://www.lanl.gov/home/kmh/