

NONLINEAR LEAST SQUARES, MAXIMUM LIKELIHOOD ESTIMATION AND BAYESIAN INFERENCE

Alistair Forbes
National Physical Laboratory

23rd June 2008, AMCTM VIII Paris
(c) Crown Copyright

Overview

Summary of uncertainty evaluation associated with nonlinear least squares model fitting: forward and inverse approaches

Behaviour of Markov chain Monte Carlo algorithms

Simple example

Model fitting

$$y_i = \phi(\mathbf{x}_i, \boldsymbol{\alpha}) + \epsilon_i,$$

y_i measured response, $\phi(\mathbf{x}, \boldsymbol{\alpha})$ modelled response depending on variables \mathbf{x} and parameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^\top$, e.g., $\phi(x, \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 x$, ϵ_i a random effect.

Data vector $\mathbf{y} = (y_1, \dots, y_m)^\top \in \mathcal{R}^m$ corresponding to variable values \mathbf{x}_i , \mathbf{x}_i known accurately.

Given $\{\mathbf{x}_i\}_{i=1}^m$, $\boldsymbol{\alpha} \mapsto \boldsymbol{\phi}(\boldsymbol{\alpha}) = (\phi(\mathbf{x}_1, \boldsymbol{\alpha}), \dots, \phi(\mathbf{x}_m, \boldsymbol{\alpha}))^\top$ describes an n -dimensional surface in \mathcal{R}^m .

Least squares estimate $\hat{\boldsymbol{\alpha}}$ minimises $d(\mathbf{y}, \boldsymbol{\alpha}) = \|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\alpha})\|$.

If $\mathbf{y} \in N(\boldsymbol{\phi}(\boldsymbol{\alpha}), \sigma^2 I)$, then $\hat{\boldsymbol{\alpha}}$ maximises $p(\mathbf{y}|\boldsymbol{\alpha}) \propto \exp\left\{-\frac{1}{2\sigma^2}d^2(\mathbf{y}, \boldsymbol{\alpha})\right\}$.

Uncertainty associated with the least squares estimate (forward)

With α fixed, $\mathbf{y} \sim N(\phi(\alpha), \sigma^2 I)$, the LS estimate is a function $\mathbf{a} = \mathcal{A}(\mathbf{y})$ of \mathbf{y} .

Use linearisation about \mathbf{a} to propagate the uncertainties associated with \mathbf{y} :

$$V_{\mathbf{a}} = \sigma^2 (J^T J)^{-1}, \quad J_{ij} = \frac{\partial f_i}{\partial \alpha_j}, \quad f_i(\alpha) = y_i - \phi(\mathbf{x}_i, \alpha).$$

Alternatively, use MC to estimate $\mathbf{a}|\alpha$: $\mathbf{y}_q \in N(\phi(\alpha), \sigma^2 I)$, $\mathbf{a}_q = \mathcal{A}(\mathbf{y}_q)$.

Inferences about α , given data vector \mathbf{y} (inverse)

Bayes' theorem:

$$p(\alpha|\mathbf{y}) \propto p(\mathbf{y}|\alpha)p(\alpha).$$

If $\mathbf{y} \in N(\phi(\alpha), \sigma^2 I)$, $p(\alpha) \propto 1$, then

$$p(\alpha|\mathbf{y}) \propto \exp \left\{ -\frac{1}{2\sigma^2} d^2(\mathbf{y}, \alpha) \right\}.$$

Shape of the distribution is determined by $\|\mathbf{y} - \phi(\alpha)\|$ as a function of α . The LS solution maximises $p(\alpha|\mathbf{y})$.

Using a linear or quadratic approximation to $\phi(\alpha)$ about \mathbf{a} , $p(\alpha|\mathbf{y})$ is approximated by

$$N(\mathbf{a}, V_{\mathbf{a}}) \quad \text{or} \quad N(\mathbf{a}, V_{\mathbf{a}}^Q), \quad V_{\mathbf{a}}^Q + \sigma^2 \left(J^T J + \sum_i f_i G_i \right)^{-1}, \quad G_{i,jk} = \frac{\partial^2 f_i}{\partial \alpha_j \partial \alpha_k}.$$

Note that $V_{\mathbf{a}}^Q$ depends on \mathbf{y} through $f_i = y_i - \phi(\mathbf{x}_i, \alpha)$.

Use Markov chain MC to estimate $p(\alpha|\mathbf{y})$ more accurately.

Forward and inverse uncertainty evaluation

Forward: $p(\mathbf{a}|\boldsymbol{\alpha})$ evaluated for $\boldsymbol{\alpha} = \mathbf{a}$, approximated by $N(\mathbf{a}, V_{\mathbf{a}})$, or MC.

Inverse: $p(\boldsymbol{\alpha}|\mathbf{y})$, approximated by $N(\mathbf{a}, V_{\mathbf{a}})$, $N(\mathbf{a}, V_{\mathbf{a}}^Q)$, or MCMC.

$p(\mathbf{a}|\boldsymbol{\alpha})$ characterises the behaviour of the measurement and estimation method, [Rossi et al., AMCTM VII].

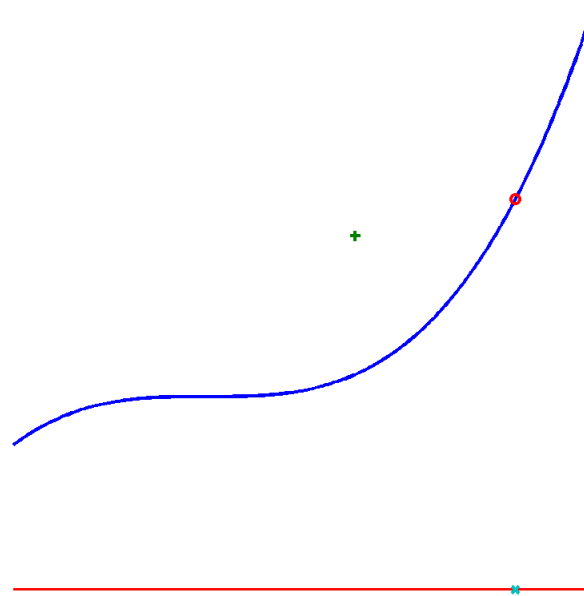
$p(\boldsymbol{\alpha}|\mathbf{a})$ is used to make inferences about $\boldsymbol{\alpha}$, given a measurement result.

For nonlinear models $p(\boldsymbol{\alpha}|\mathbf{y})$ and $p(\mathbf{a}|\boldsymbol{\alpha})$ are generally different distributions.

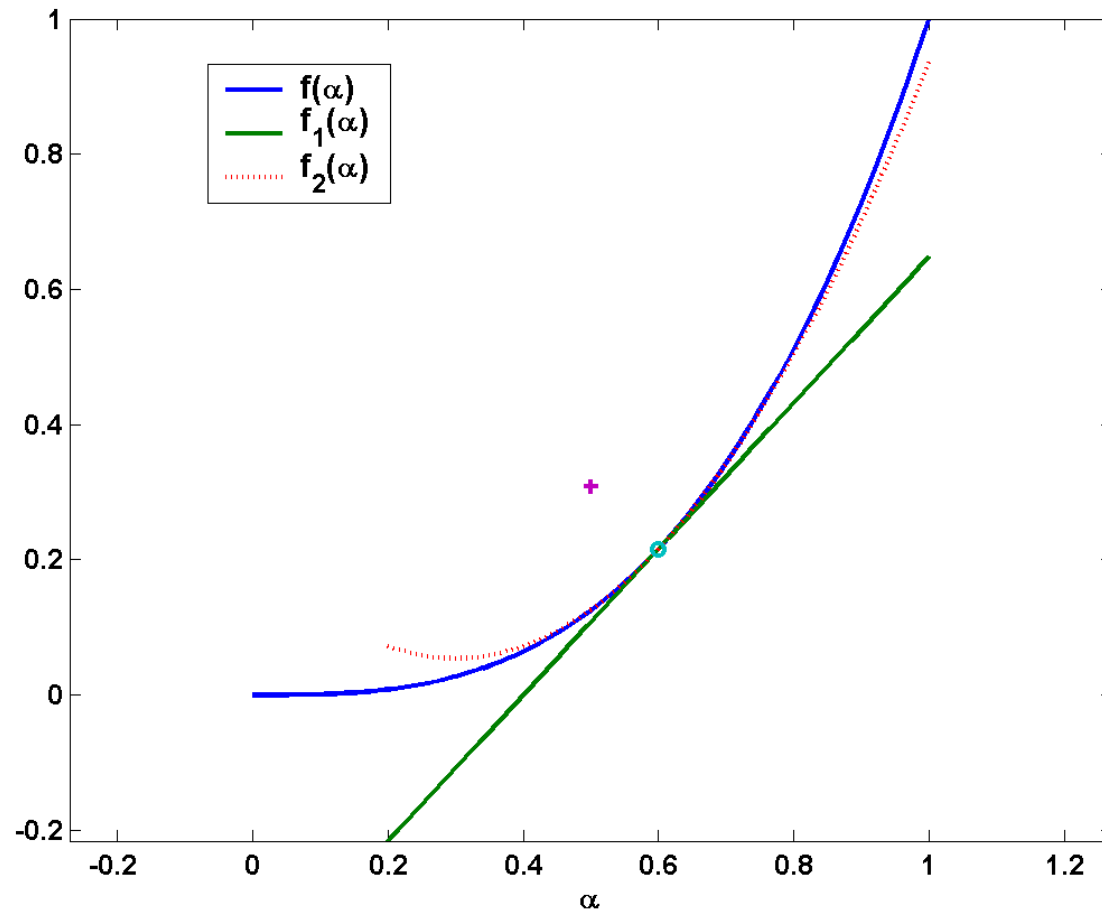
[For linear models they are the same.]

Geometrical interpretation of $p(\alpha|y)$

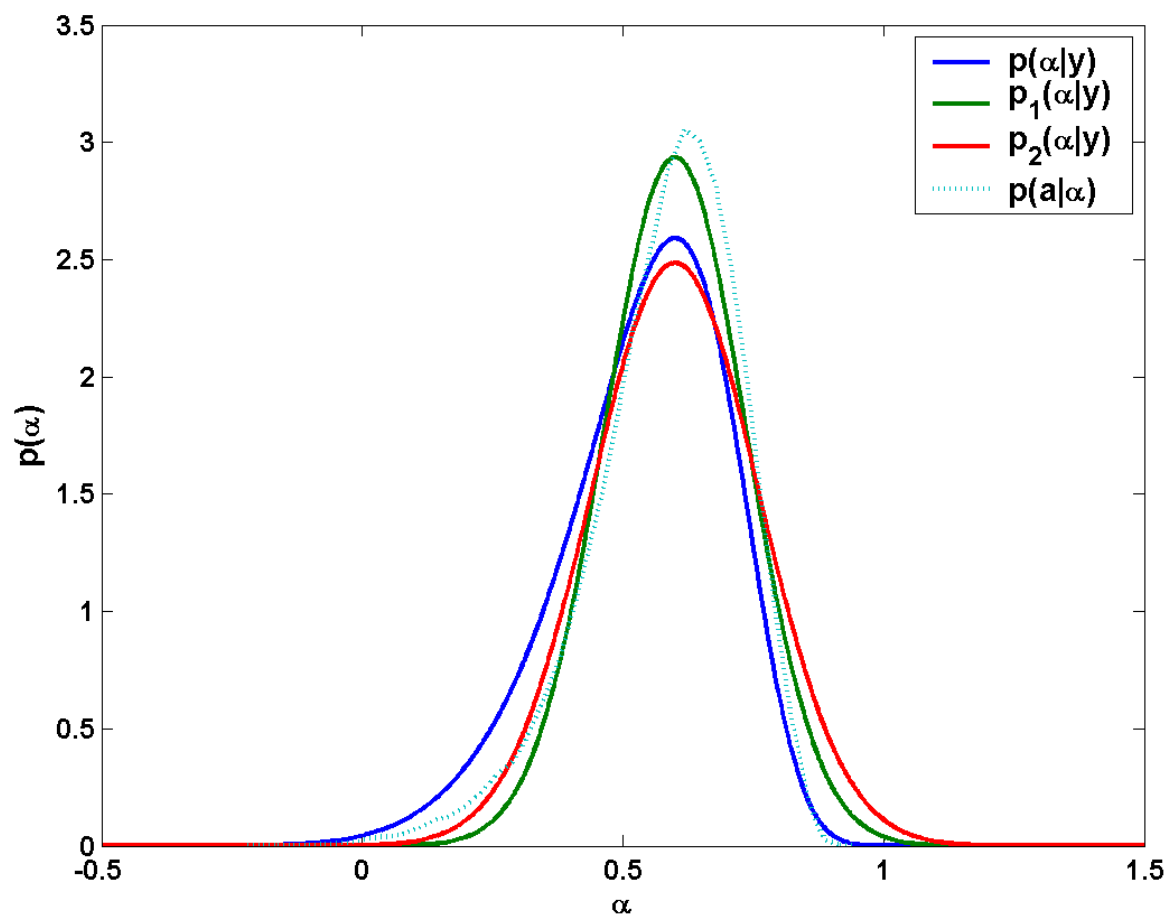
$$y_1 \in N(\alpha, \sigma_1^2), \quad y_2 \in N(\alpha^3, \sigma_2^2), \quad \alpha \mapsto (\alpha, \alpha^3)^\top$$



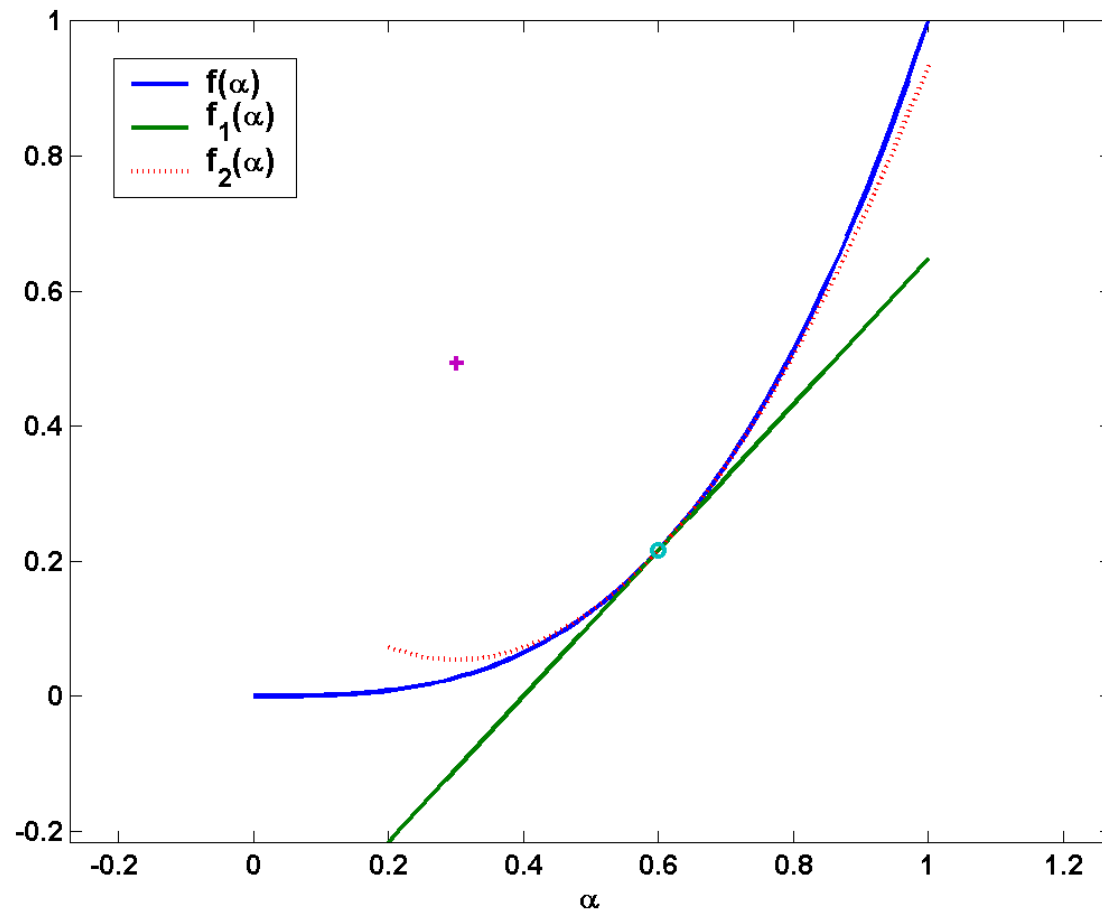
Nonlinear example I



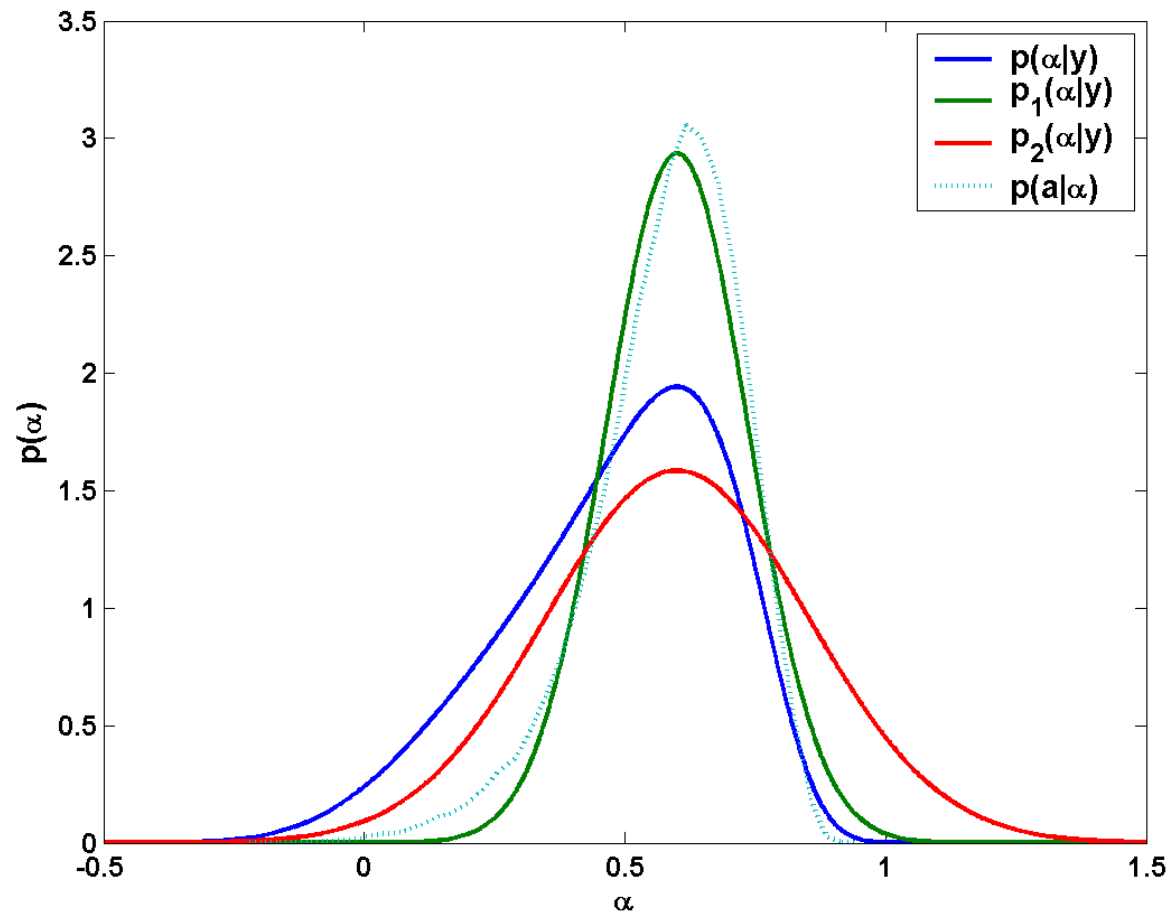
Posterior distributions I



Nonlinear example II



Posterior distributions II



Markov chain Monte Carlo

Simulation method to sample from any distribution $p(\alpha)$.

One example of the Metropolis-Hastings algorithm:

Given \mathbf{a}_q , $p(\mathbf{a}_q)$ and $q(\mathbf{a}_q)$, draw $\mathbf{a}^* \in q(\alpha)$, an approximating distribution.

Evaluate $p(\mathbf{a}^*)$ and $q(\mathbf{a}^*)$ and $r_q = \min\{1, p(\mathbf{a}^*)q(\mathbf{a}_q)/p(\mathbf{a}_q)q(\mathbf{a}^*)\}$.

Draw $u_q \in R[0, 1]$.

If $u_q < r_q$ accept \mathbf{a}^* : set $\mathbf{a}_{q+1} = \mathbf{a}^*$, otherwise set $\mathbf{a}_{q+1} = \mathbf{a}_q$.

Eventually \mathbf{a}_q are samples from $p(\alpha)$.

Convergence of Markov chains

Example: random walk

$$a_q = ae_q + ba_{q-1}, \quad e_q \in N(0, 1), \quad a^2 + b^2 = 1;$$

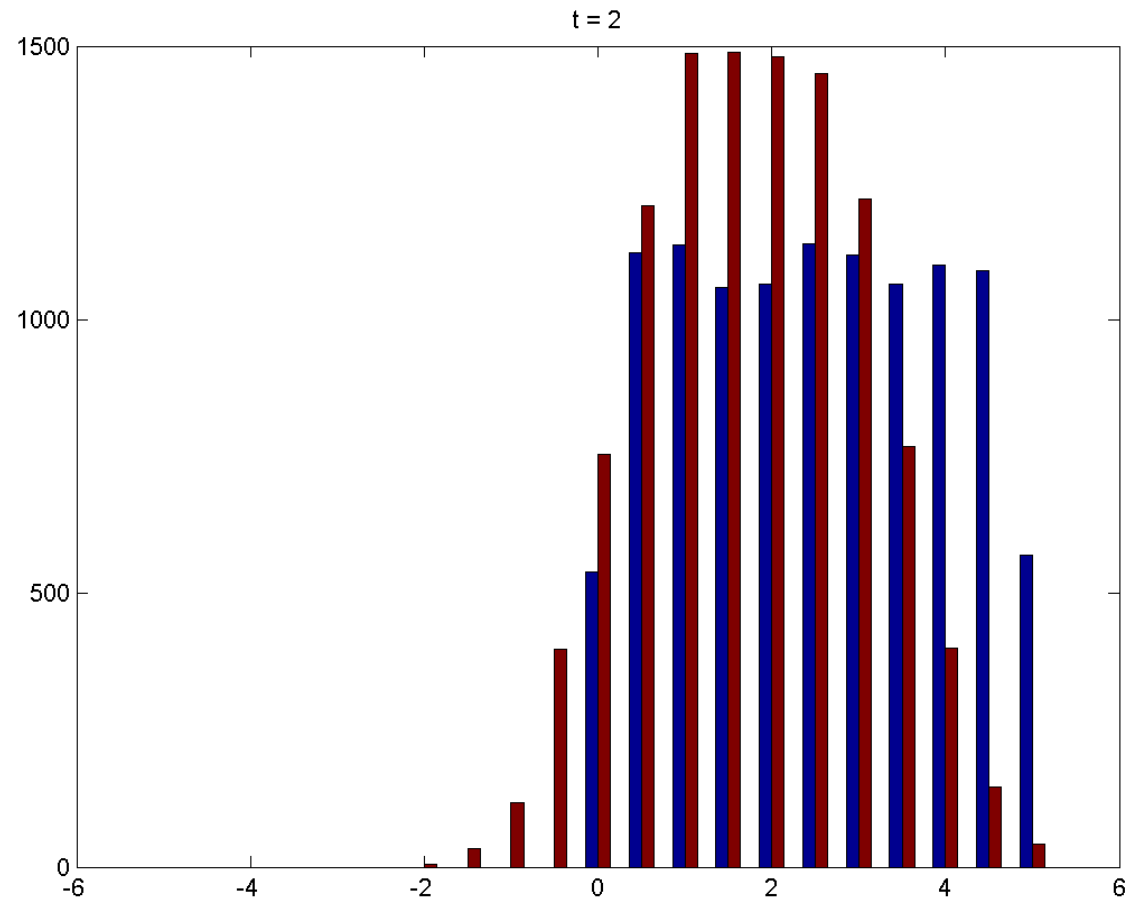
Let a_1 be a random sample from *any* distribution and generate a_q as above. Eventually a_q will be a sample from $N(0, 1)$.

Note that if a_{q-1} is a sample from $N(0, 1)$ then a_q is also: $N(0, 1)$ is the limit distribution of the chain.

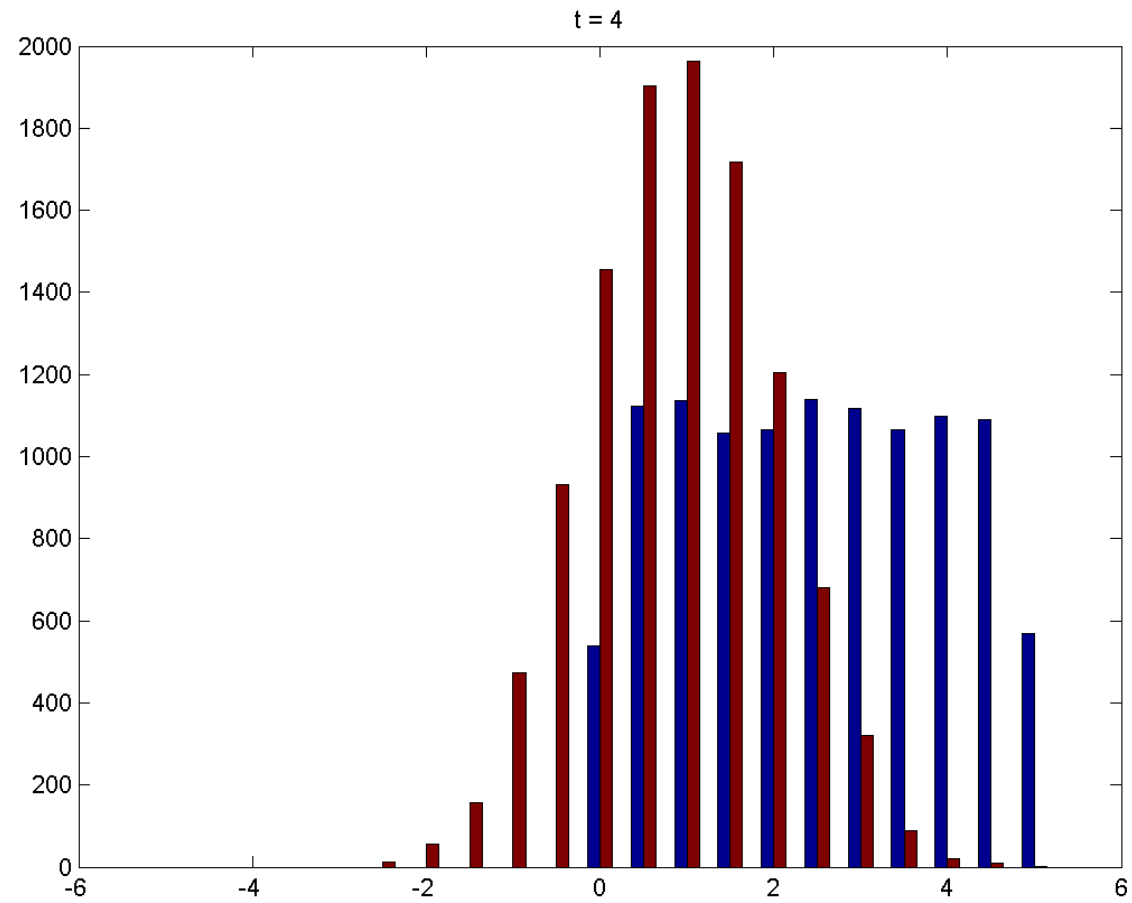
For discrete distributions, the chain is specified by a transition matrix with 1 as a maximum eigenvalue and the limit distribution as the corresponding eigenvector.

The convergence is governed by the second largest eigenvalue.

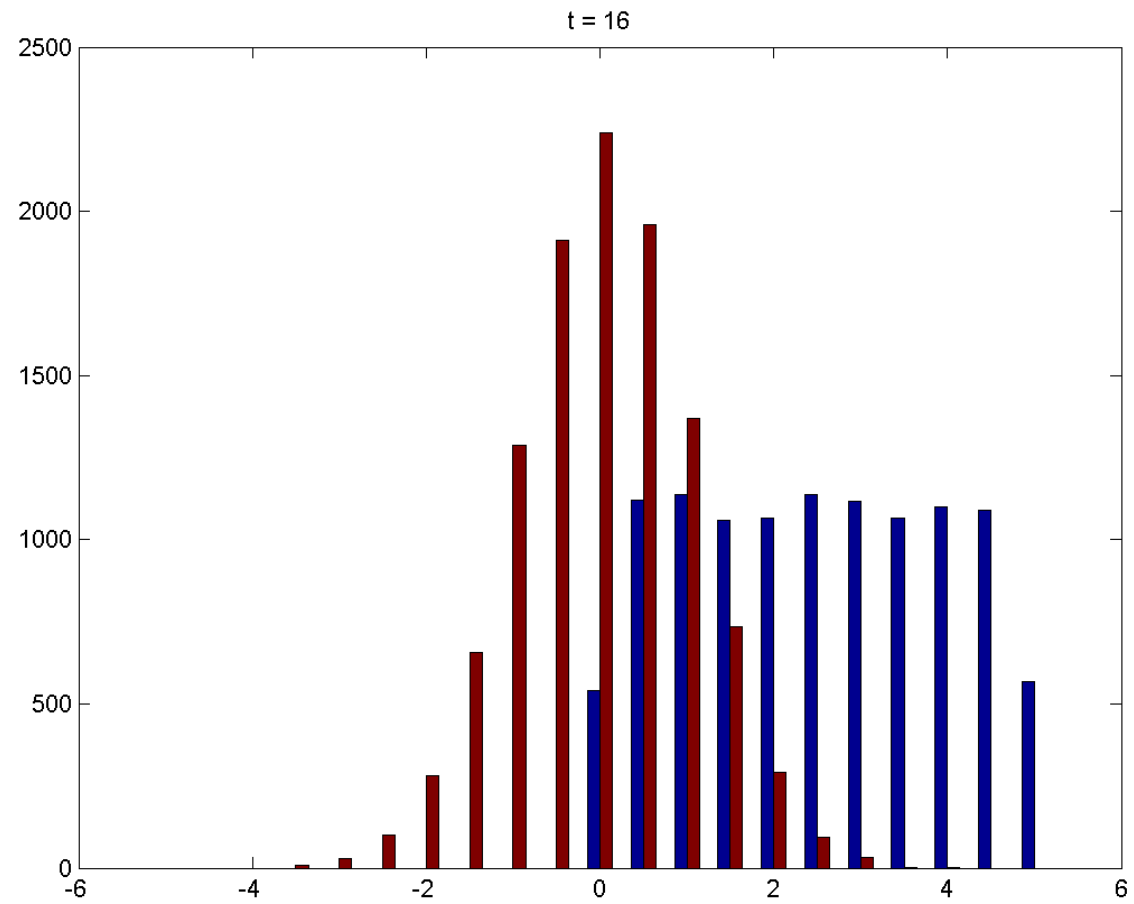
Random walk: $a = b = \sqrt{2}$



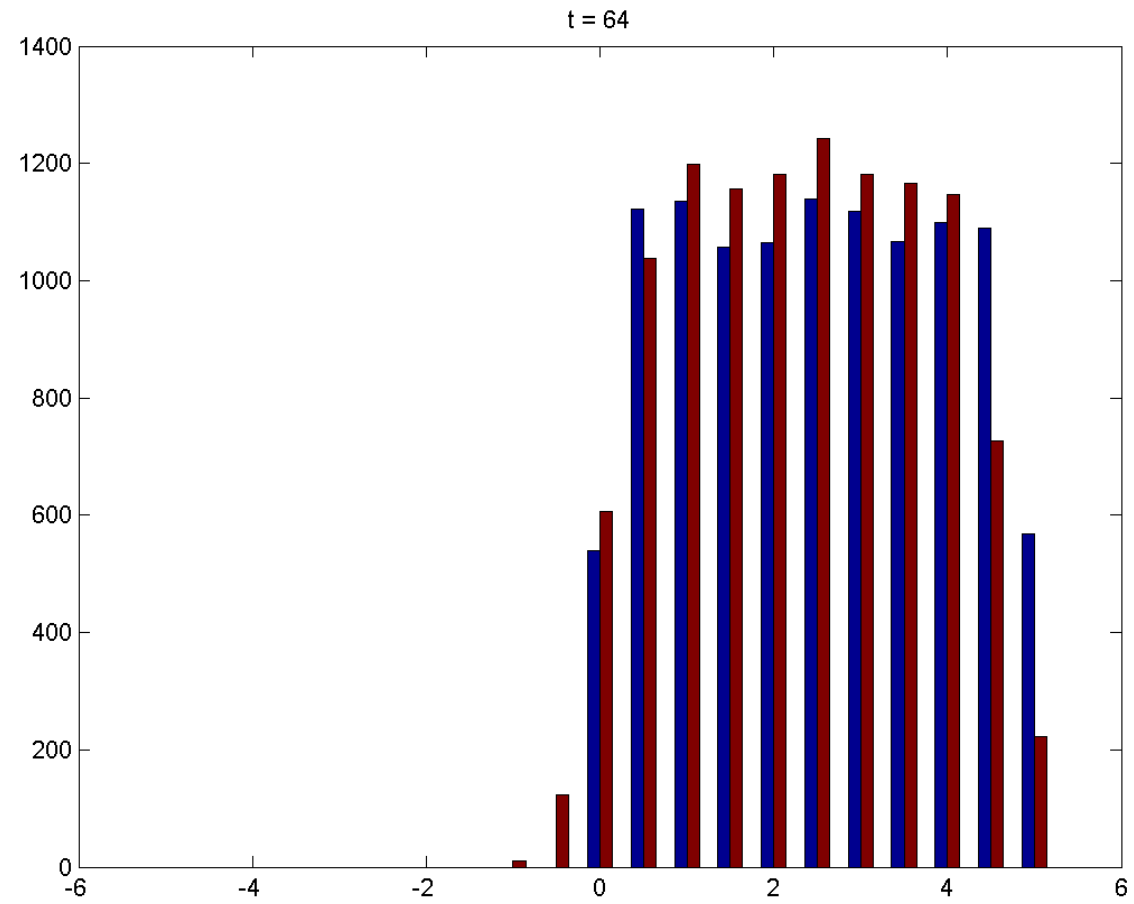
Random walk: $a = b = \sqrt{2}$



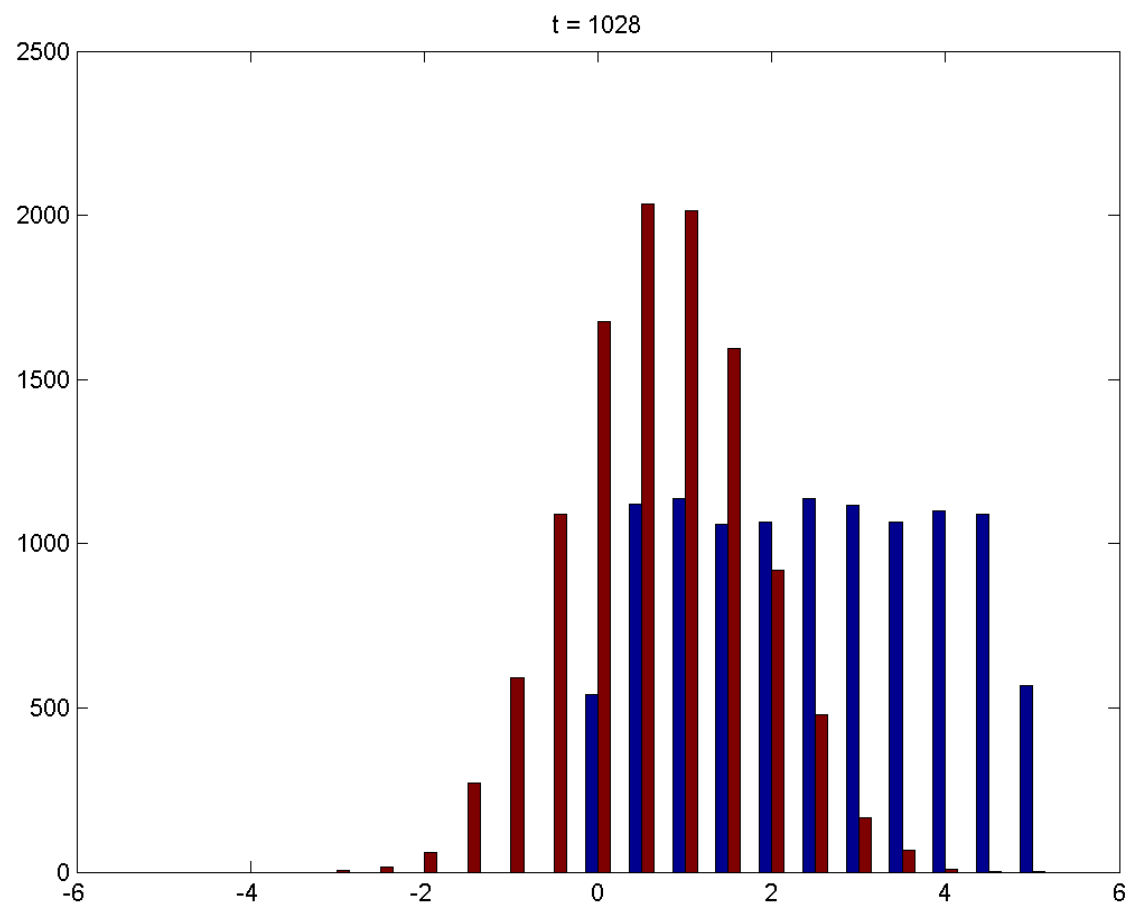
Random walk: $a = b = \sqrt{2}$



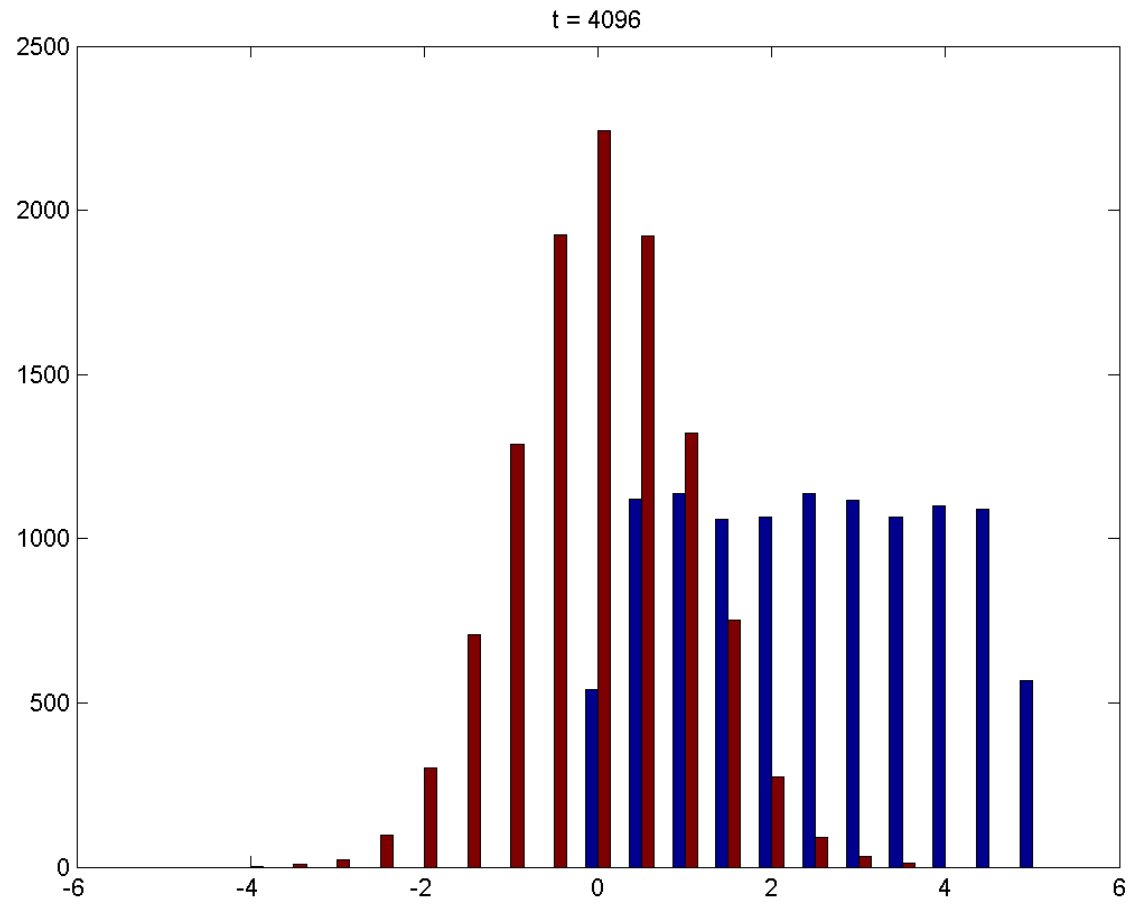
Random walk: $a = 0.05$



Random walk: $a = 0.05$



Random walk: $a = 0.05$



Ergodicity

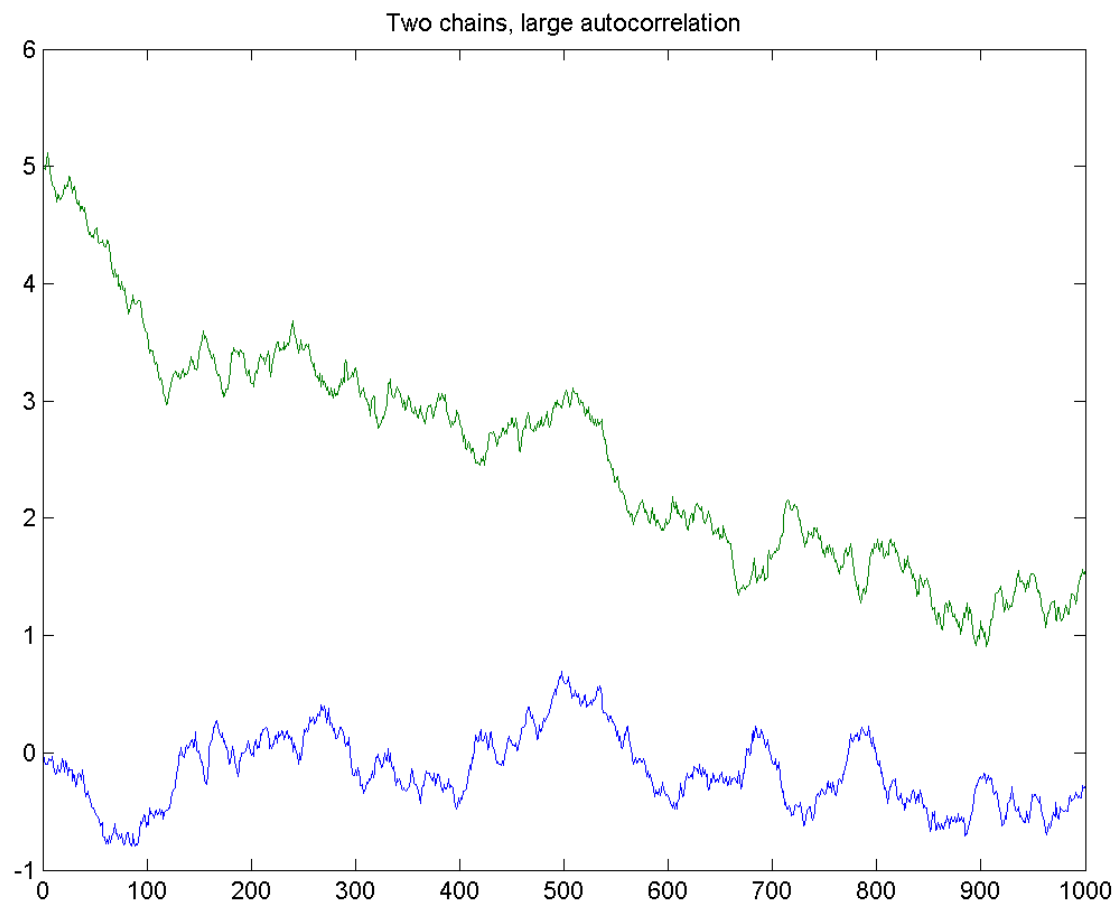
Once a chain has converged, a_q , $q = N, N + 1, \dots$, is a sample from the limit distribution.

With multiple chains starting at \mathbf{a}_1 , and $A = [\mathbf{a}_1 \ \mathbf{a}_2 \ \dots]$, both the columns and rows will represent samples from the limit distribution.

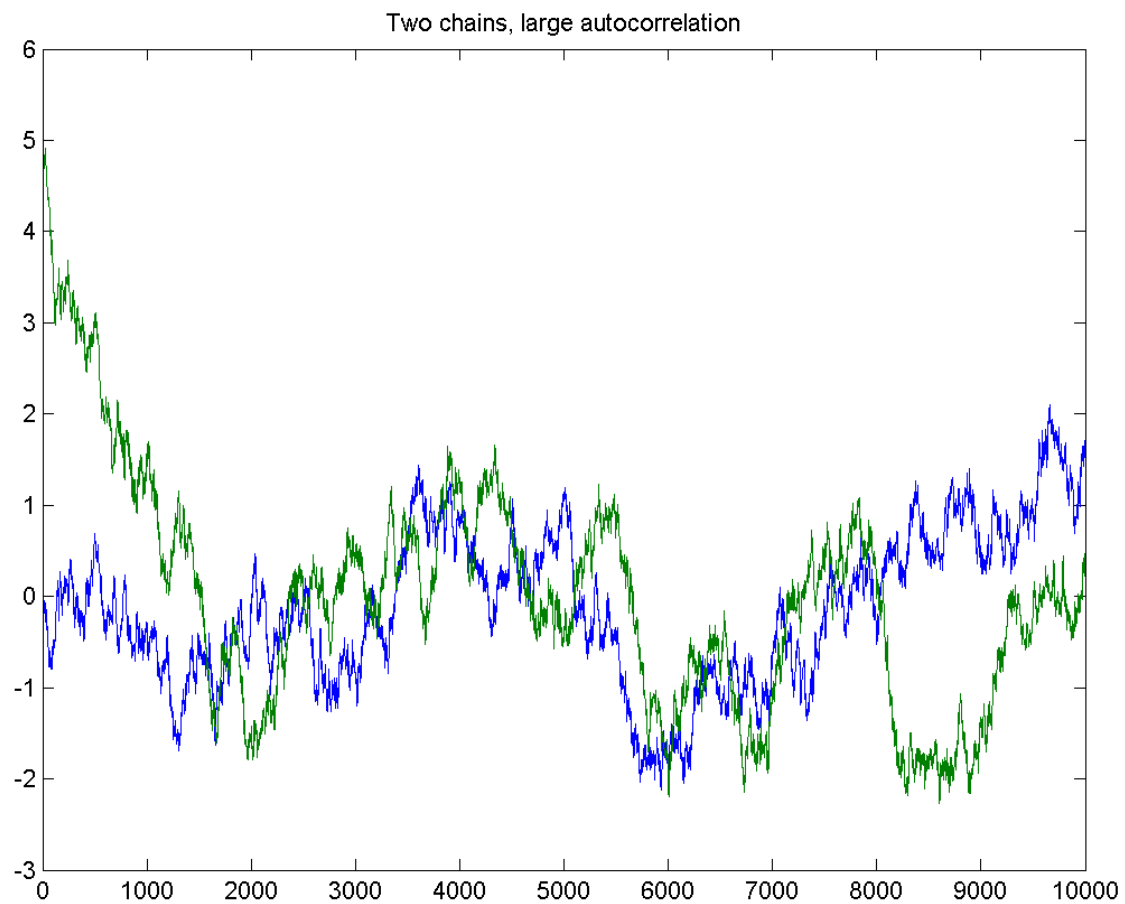
But we might need a large sample to generate accurate statistics.

For Metropolis-Hastings algorithms, need to choose the proposal distribution $q(\alpha)$ carefully. The scheme above is designed to minimise autocorrelation in the chains.

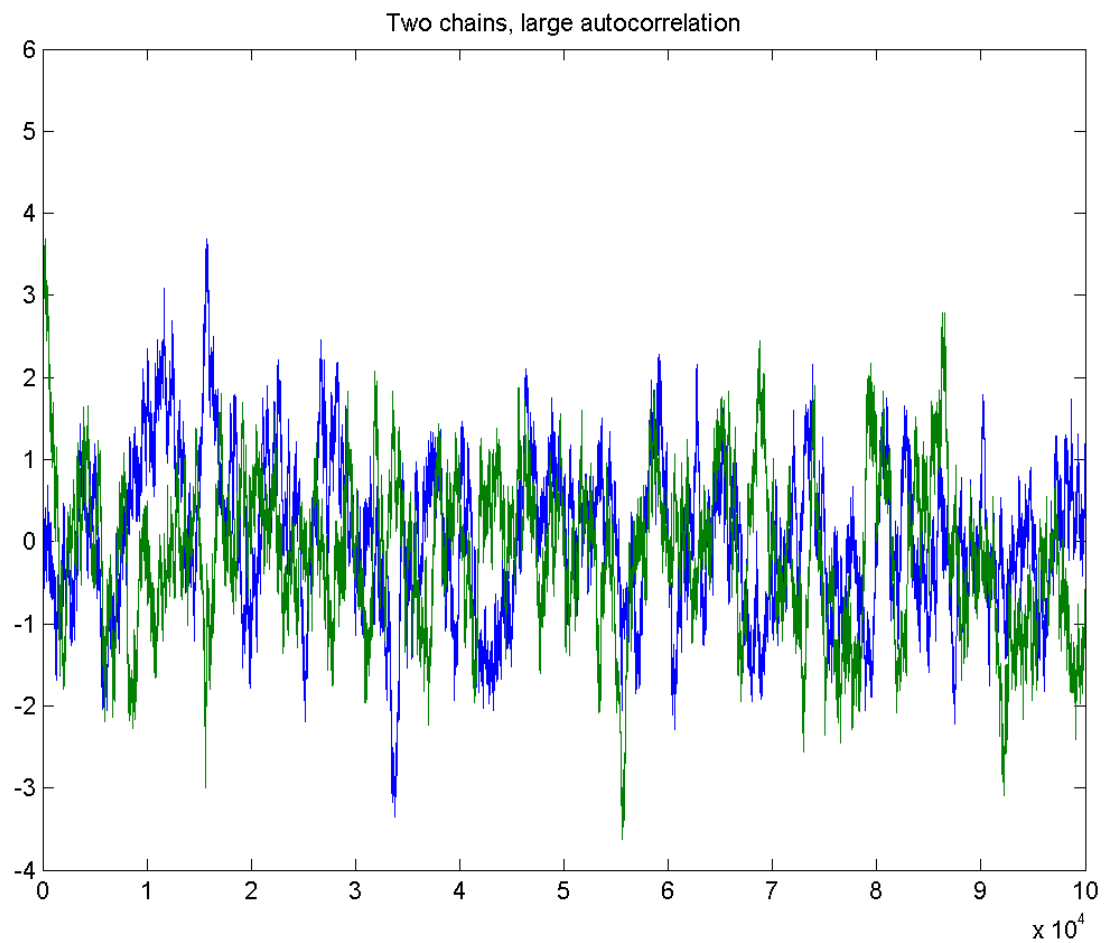
Random walk: $a = 0.05$



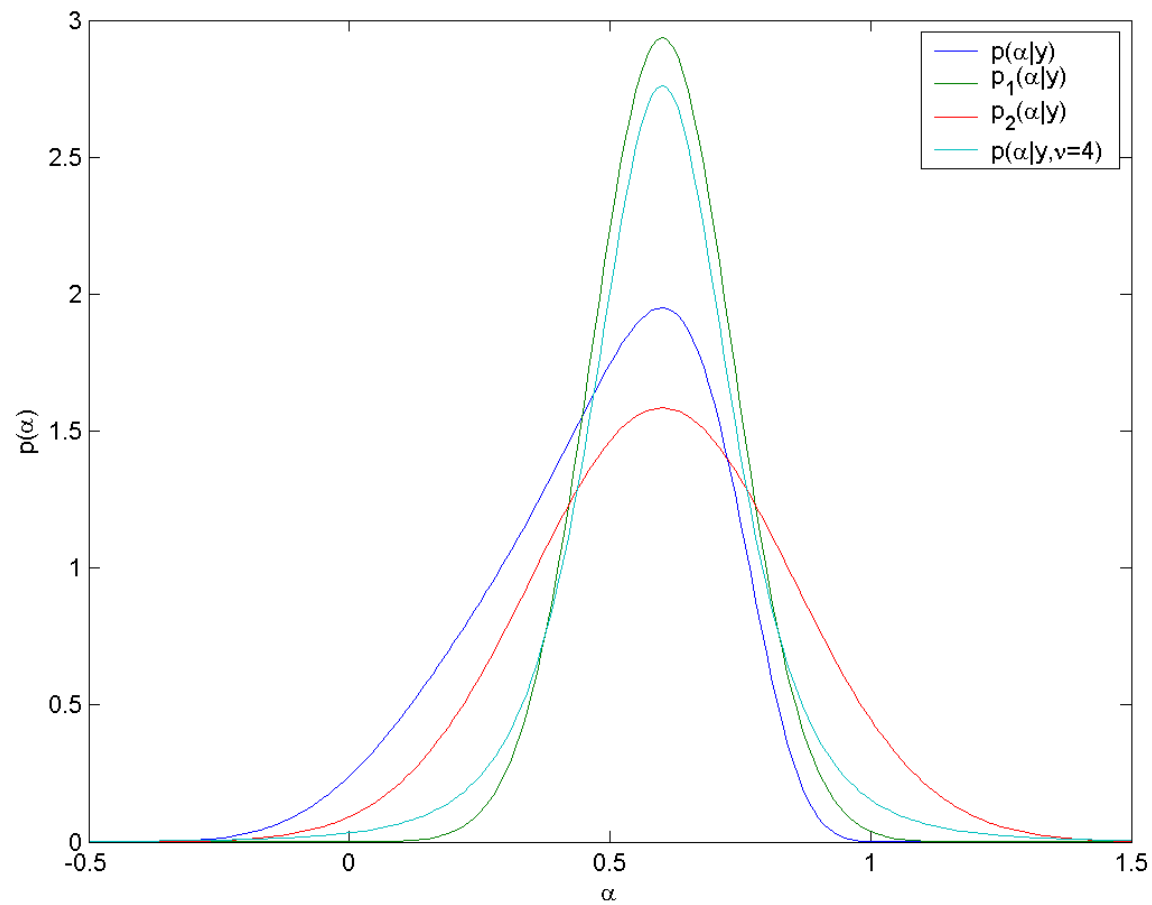
Random walk: $a = 0.05$



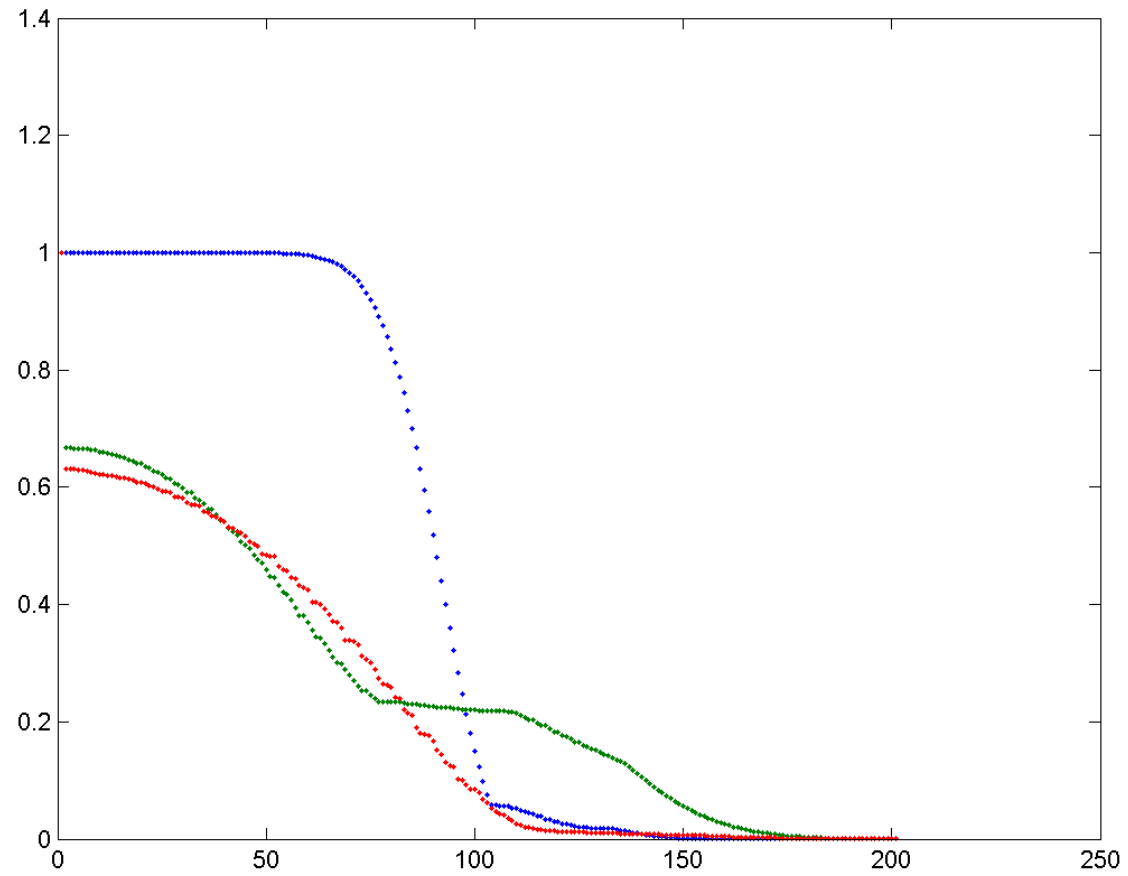
Random walk: $a = 0.05$



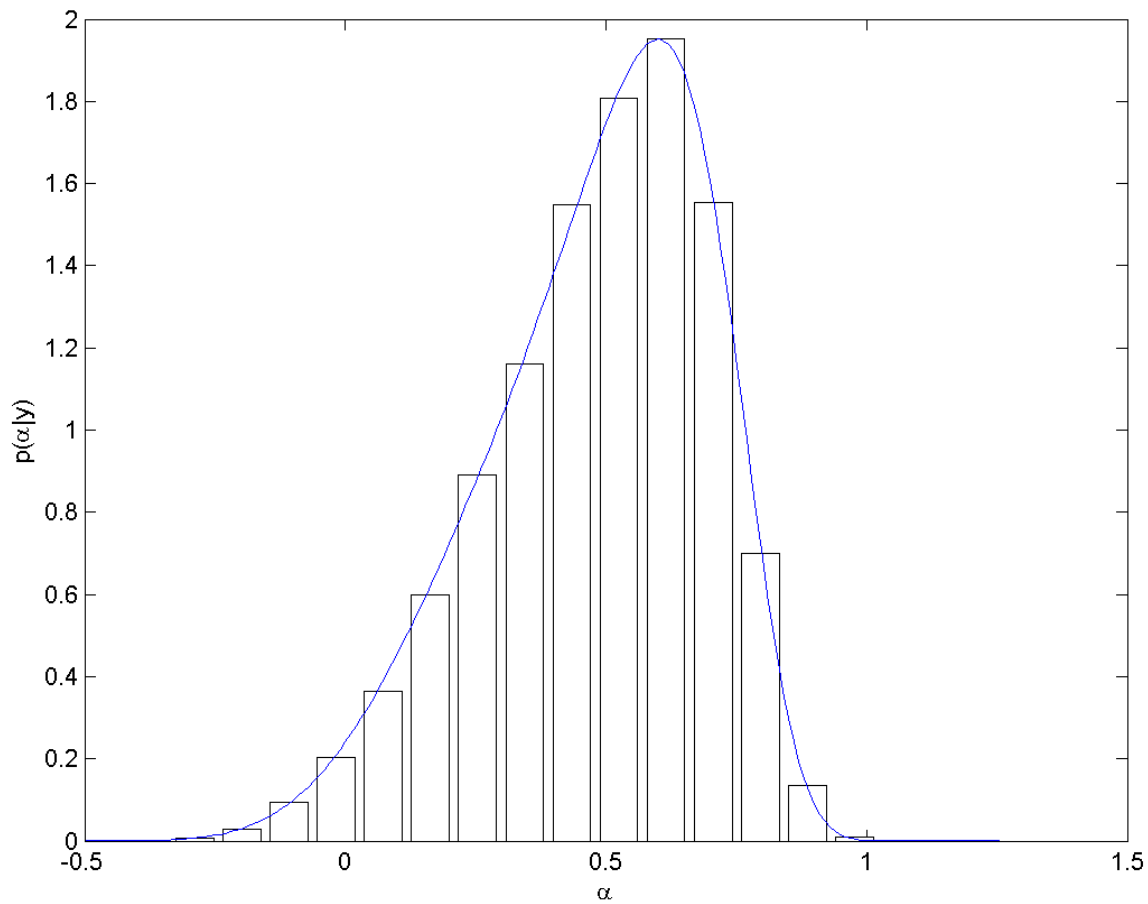
Three approximating distributions for $p(\alpha|y)$



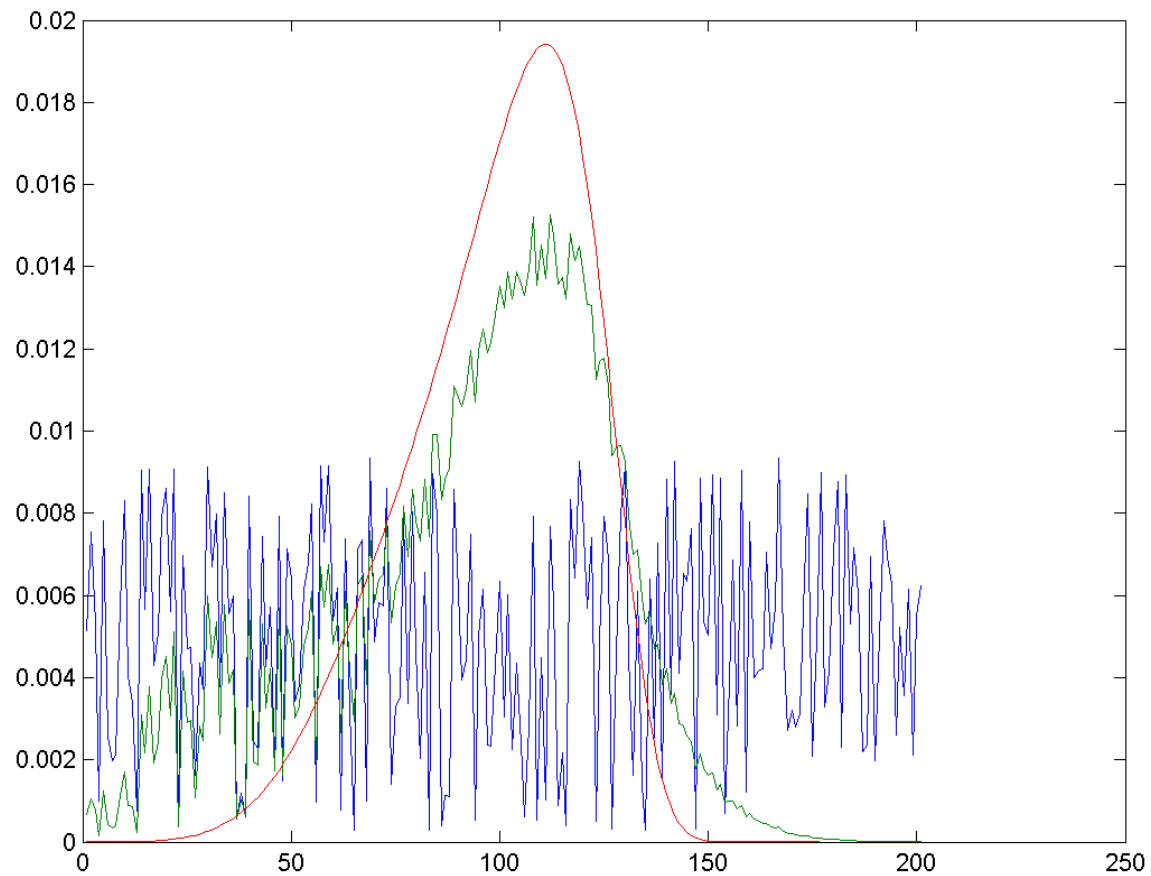
Eigenvalues of the corresponding transition matrices



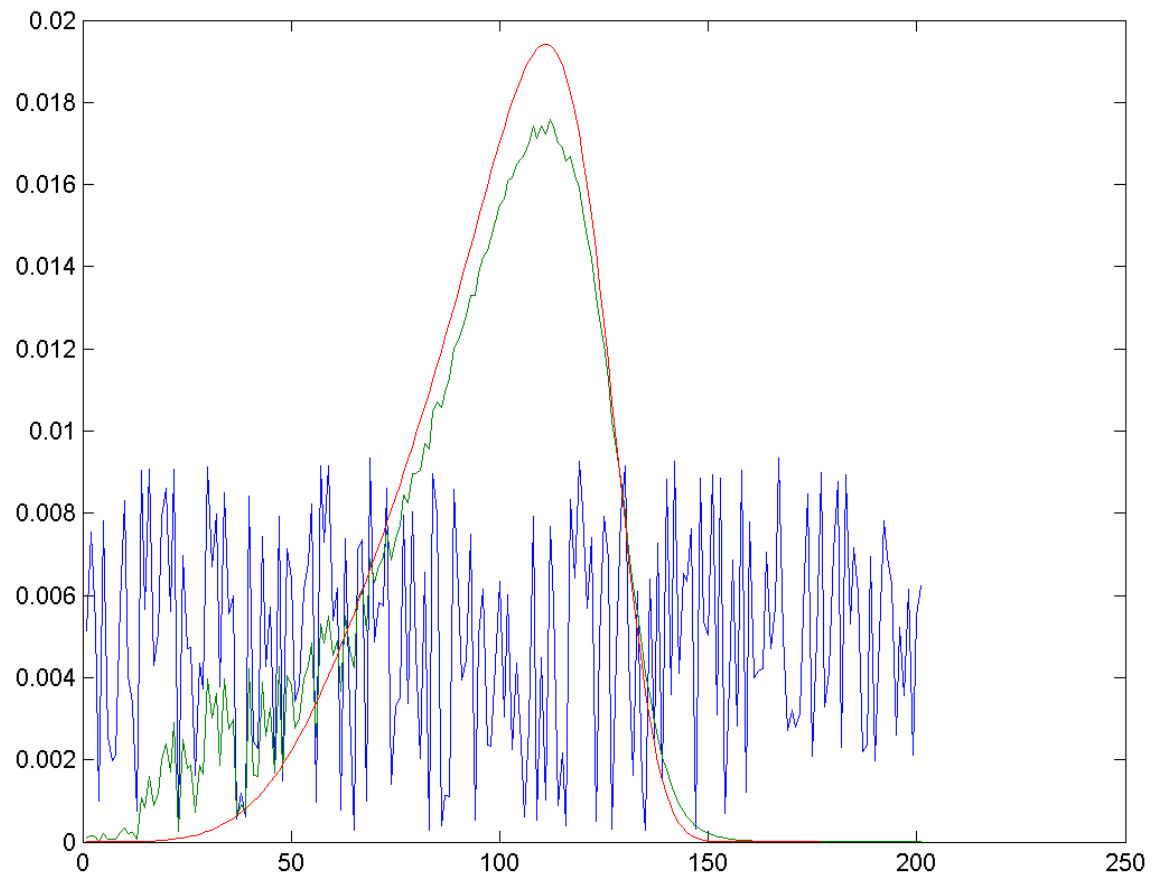
$p(\alpha|y)$ estimated by MCMC



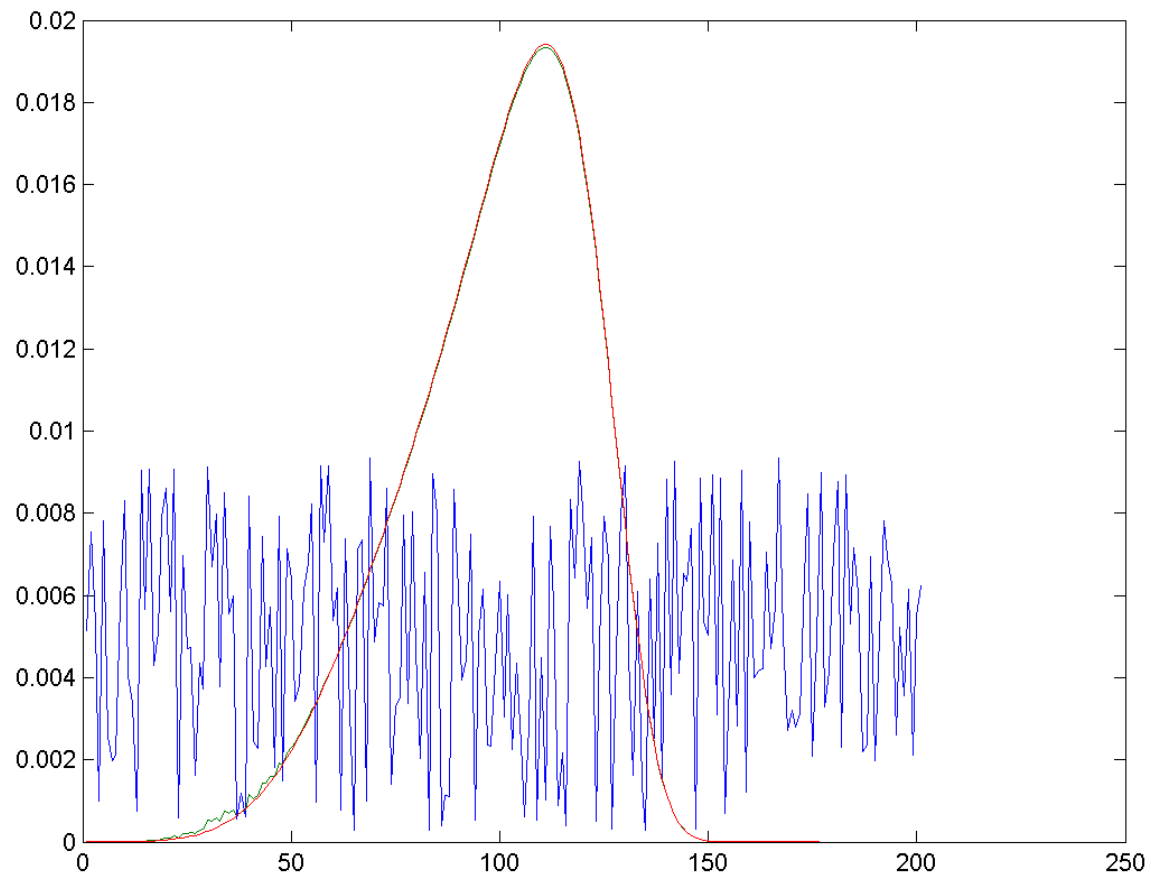
Convergence to $p(\alpha|y)$ from an arbitrary distribution: 1 step



Convergence to $p(\alpha|y)$ from an arbitrary distribution: 2 steps



Convergence to $p(\alpha|y)$ from an arbitrary distribution: 10 steps



Summary

Forward: $p(\mathbf{a}|\boldsymbol{\alpha})$, $N(\mathbf{a}, V_{\mathbf{a}})$, MC, system characterisation

Inverse: $p(\boldsymbol{\alpha}|\mathbf{y})$, $N(\mathbf{a}, V_{\mathbf{a}})$, $N(\mathbf{a}, V_{\mathbf{a}}^Q)$, MCMC, inferences about $\boldsymbol{\alpha}$, given \mathbf{y} .

MCMC algorithms $A = [\mathbf{a}_1, \mathbf{a}_2, \dots]$. More chains if autocorrelation is a problem.