Notation and Jargon

θ: parameter(s)
X: data
$p(\theta|X)$: probability of θ given X (i.e., posterior probability density)
$p(\theta)$: prior distribution for θ
$p(X|\theta)$: likelihood

$$
\begin{aligned}
p(\theta|X) &= \frac{p(\theta, X)}{p(X)} \\
&= \frac{p(X|\theta)p(\theta)}{\int_\theta p(X, \theta)d\theta} \\
&= \frac{p(X|\theta)p(\theta)}{\int_\theta p(X|\theta)p(\theta)d\theta}
\end{aligned}
$$

$E(\theta|X)$: expected value of θ given X (i.e., posterior mean)

Motivation:

Bayesian inference centers on the posterior probability distribution $p(\theta|X)$. For example, Bayesians will typically want to report the posterior mean $E(\theta|X)$ as a point estimate of θ. If parameter(s) θ is(are) discrete, then the posterior mean is

$$E(\theta|X) = \sum_{\theta} \theta p(\theta|X). \tag{1}$$

If parameter(s) θ is(are) continuous, then the posterior mean is

$$E(\theta|X) = \int_{\theta} \theta p(\theta|X) d\theta. \tag{2}$$

Often, the above sums or the above integrals are not easy to evaluate. It is often the situation for continuous parameter(s) θ that the density $p(\theta|X)$ can be evaluated for each value of θ even though the exact solution of the above integral for the posterior mean is not easy to obtain. In such a situation, it may be of interest to characterize the posterior distribution $p(\theta|X)$ by randomly sampling values of θ from this posterior density.

Random Samples

Let $\theta_1, \theta_2, \ldots, \theta_N$ be N independent samples from the posterior probability distribution $p(\theta|X)$. If N is large, a good estimate of the posterior mean for θ is the sample mean for θ,

$$E(\theta|X) \doteq \frac{1}{N} \sum_{i=1}^{N} \theta_i. \tag{3}$$

A good estimate of the posterior mean for θ^2 might be the sample mean for θ^2,

$$E(\theta^2|X) \doteq \frac{1}{N} \sum_{i=1}^{N} {\theta_i}^2. \tag{4}$$

In general, a good estimate of the posterior mean for some function $f(\theta)$ of θ might be the sample mean for $f(\theta)$,

$$E(f(\theta)|X) \doteq \frac{1}{N} \sum_{i=1}^{N} f(\theta_i). \tag{5}$$

Summary: Because Bayesian inference revolves around posterior dist. $p(\theta|X)$ and what can be computed from it, there is much value to obtaining random sample $\theta_1, \theta_2, \ldots, \theta_N$ from $p(\theta|X)$.

Rejection Sampling

Sometimes, posterior distribution has simple form and it is easy to directly sample $\theta_1, \theta_2, \ldots, \theta_N$ from $p(\theta|X)$.

For cases where direct sampling is not easy, rejection sampling might work. With rejection sampling, we choose a probability distribution $g(\theta)$ for θ that is straightforward to sample from.

We assume that we know or can find some positive number c such that $cg(\theta) \geq p(\theta|X)$ for all values of θ. Notice that one implication of this requirement is that we cannot have $g(\theta) = 0$ for some value of θ where $p(\theta|X)$ is not zero.

Rejection sampling algorithm:

```
For $i = 1$ to $N$   {
1. Sample $\theta^*$ from $g(\theta)$
2. Set $r = \frac{p(\theta^*|X)}{cg(\theta^*)}$
3. With probability $r$, set $\theta_i = \theta^*$. Otherwise, “reject” $\theta^*$ and go to Step 1 above   }
```

Note 1: We can carry out Step 3 above by sampling a random number U from a uniform distribution between 0 and 1. Because the probability that $U < r$ is equal to r, this means that we set $\theta_i = \theta^*$ if $U < r$ and we reject θ^* if $U \geq r$.

Note 2: We hope $p(\theta^*|X)$ and $cg(\theta^*)$ tend to not be too different from each other. Otherwise, we will often reject θ^* in Step 3 above. This may make it computationally infeasible to obtain desired random sample.

Importance Sampling

To emphasize that posterior mean $E(\theta|X)$ is expected value of θ according to posterior distribution $p(\theta|X)$ we use $E_{p(\theta|X)}(\theta)$ as another way to refer to the posterior mean.

In general, expected value of some function $f(\theta)$ of θ with respect to some probability distribution $g(\theta)$ will be written $E_{g(\theta)}(f(\theta))$.

If θ is discrete,

$$E_{g(\theta)}(f(\theta)) = \sum_{\theta} f(\theta)g(\theta). \tag{6}$$

If θ is continuous,

$$E_{g(\theta)}(f(\theta)) = \int_{\theta} f(\theta)g(\theta)d\theta. \tag{7}$$

Now, we assume $\theta_1, \theta_2, \ldots, \theta_N$ is sample from $g(\theta)$ rather than from the posterior density.

Importance sampling is way of adjusting a sample $\theta_1, \theta_2, \ldots, \theta_N$ from $g(\theta)$ so that it can be used to estimate quantities related to other probability distributions (e.g., the posterior probability distribution $p(\theta|X)$).

For a parameter θ that is defined on continuous values (the same ideas would apply if it was a discrete-valued parameter), we can approximate the posterior mean of some function $f(\theta)$ because

$$\begin{aligned} E(f(\theta)|X) = E_{p(\theta|X)}(f(\theta)) &= \int_\theta f(\theta)p(\theta|X)d\theta \\ &= \int_\theta f(\theta)\frac{g(\theta)}{g(\theta)}p(\theta|X)d\theta \\ &= \int_\theta f(\theta)\frac{p(\theta|X)}{g(\theta)}g(\theta)d\theta \\ &= E_{g(\theta)}(f(\theta)\frac{p(\theta|X)}{g(\theta)}). \qquad (8) \end{aligned}$$

Because $\theta_1, \theta_2, \ldots, \theta_N$ is a sample from $g(\theta)$, we could therefore use the approximation,

$$E_{p(\theta|X)}(f(\theta)) = E_{g(\theta)}(f(\theta)\frac{p(\theta|X)}{g(\theta)}) \doteq \frac{1}{N}\sum_{i=1}^{N} f(\theta_i)w_i, \tag{9}$$

where w_i is referred to as an importance weight and is

$$w_i = \frac{p(\theta_i|X)}{g(\theta_i)}. \tag{10}$$

However, a (usually) better approximation is

$$E_{p(\theta|X)}(f(\theta)) \doteq \frac{\Sigma_{i=1}^{N} f(\theta_i)w_i}{\Sigma_{i=1}^{N} w_i}. \tag{11}$$

We will not justify this latter approximation, but notice that

$$E_{g(\theta)}(w_i) = 1$$

Note 1: An advantage of Equation 11 is that the same approximation would result if all weights were multiplied by some number. The implication is that we only need to be able to estimate the weights up to some constant of proportionality that is shared among the weights. This is often useful for Bayesian applications because

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}. \tag{12}$$

Numerator of above ratio is likelihood multiplied by prior density. Denominator is term that is typically difficult to calculate but it fortunately is not a function of θ. This means the $p(X)$ in the numerator and denominator terms of Equation 11 can cancel each other out and we need only to calculate numerator of ratio in Equation 12.

Note 2: Approximation of Equation 11 is likely to be good if importance weights do not vary much and is likely to be bad if importance weights substantially vary. One implication is that importance sampling tends to be most successful when $g(\theta)$ and $p(\theta|X)$ are quite similar.

Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods have revolutionized statistics.

Usually, MCMC methods are applied in Bayesian frameworks.

$$\begin{aligned} p(\theta|X) &= \frac{p(\theta, X)}{p(X)} \\ &= \frac{p(X|\theta)p(\theta)}{\int_\theta p(X, \theta)d\theta} \\ &= \frac{p(X|\theta)p(\theta)}{\int_\theta p(X|\theta)p(\theta)d\theta} \end{aligned}$$

In many situations, determining the exact value of the integral in denominator is difficult.

The MCMC idea is to approximate $\Pr(\theta \mid X)$ by sampling a large number of θ values from $\Pr(\theta \mid X)$.

So, θ values with a higher posterior probability are more likely to be sampled than θ values with a low posterior probability.

Question: How is this sampling achieved?

Answer: A Markov chain is constructed and simulated. The states of this chain represent values of θ. The stationary distribution of this chain is $\Pr(\theta \mid X)$.

In other words, we start the chain at some initial value of θ. After running the chain for a long enough time, the probability of the chain being at some particular state will be approximately equal to the posterior probability of the state.

Let $\theta^{(t)}$ be the value of θ after t steps of the Markov chain where $\theta^{(0)}$ is the initial value.

Each step of the Markov chain involves randomly proposing a new value of θ based on the current value of θ. Call the proposed value θ^*.

We decide with some probability to either accept θ^* as our new state or to reject the proposed θ^* and remain at our current state.

The Hastings (Hastings 1970) algorithm is a way to make this decision and force the stationary distribution of the chain to be $p(\theta|X)$.

According to the Hastings algorithm, what state should we adopt at step $t+1$ if $\theta^{(t)}$ is the current state and θ^* is the proposed state?

Let $J(\theta^*|\theta^{(t)})$ be the "jumping" distribution, i.e. the probability of proposing θ^* given that the current state is $\theta^{(t)}$.

Define r as

$$r = \frac{p(\theta^*|X)J(\theta^{(t)}|\theta^*)}{p(\theta^{(t)}|X)J(\theta^*|\theta^{(t)})} = \frac{p(X|\theta^*)p(\theta^*)J(\theta^{(t)}|\theta^*)}{p(X|\theta^{(t)})p(\theta^{(t)})J(\theta^*|\theta^{(t)})}$$

With probability equal to the minimum of r and 1, we set

$$\theta^{(t+1)} = \theta^*.$$

Otherwise, we set

$$\theta^{(t+1)} = \theta^{(t)}.$$

For the Hastings algorithm to yield the stationary distribution $p(\theta|X)$, there are a few required conditions.

The most important condition is that it must be possible to reach each state from any other in a finite number of steps.

Also, the Markov chain can't be periodic.

MCMC implementation details:

The Markov chain should be run as long as possible.

We may have T total samples after running our Markov chain. They would be $\theta^{(1)}$, $\theta^{(2)}$, . . ., $\theta^{(T)}$.

The first B $(1 \leq B < T)$ of these samples are often discarded (i.e. not used to approximate the posterior).

The period before the chain has gotten these B samples that will be discarded is referred to as the "burn–in" period.

The reason for discarding these samples is that the early samples typically are largely dependent on the initial state of the Markov chain and often the initial state of the chain is (either intentionally or unintentionally) atypical with respect to the posterior distribution.

The remaining samples $\theta^{(B+1)}$, $\theta^{(B+2)}$, $\ldots$, $\theta^{(T)}$ are used to approximate the posterior distribution. For example, the average among the sampled values for a parameter might be a good estimate of its posterior mean.

Question: How do we know that we have run our Markov chain long enough to get a good approximation of the posterior distribution?

Annoying Answer: We don't.

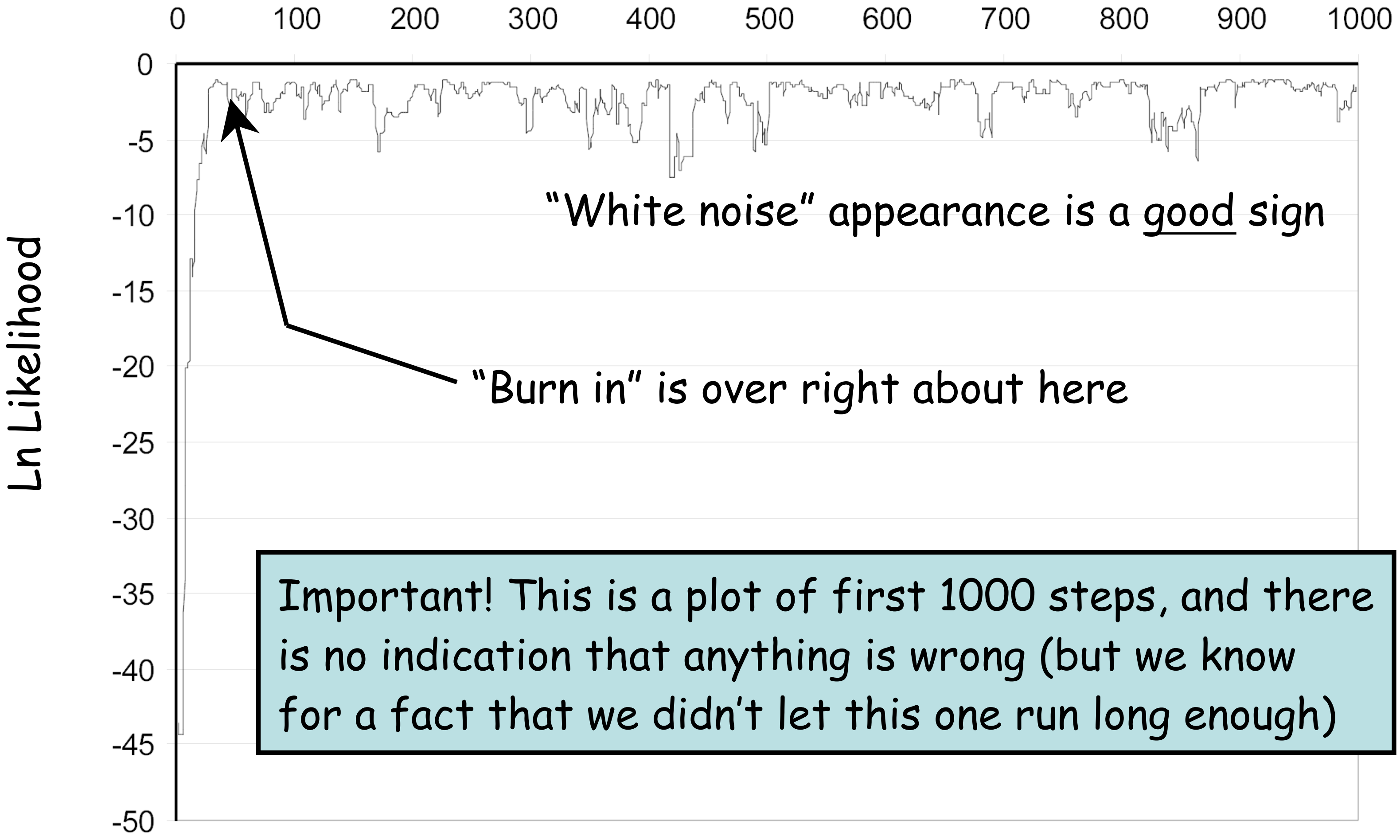
Diagnostics (not guaranteed to detect that chain is too short but often work in practice):

1. Run multiple independent chains from different initial states. See if these chains give approximately the same approximation of posterior.

2. Look at **lots** of plots, especially "trace plots" where parameter value or some statistic is on y-axis and where step number is on x-axis. Want to see "white noise" pattern in trace plot.

3. Numerical measure such as "Gelman-Rubin diagnostic" (google it) or "effective sample size" (see next pages).

History plots



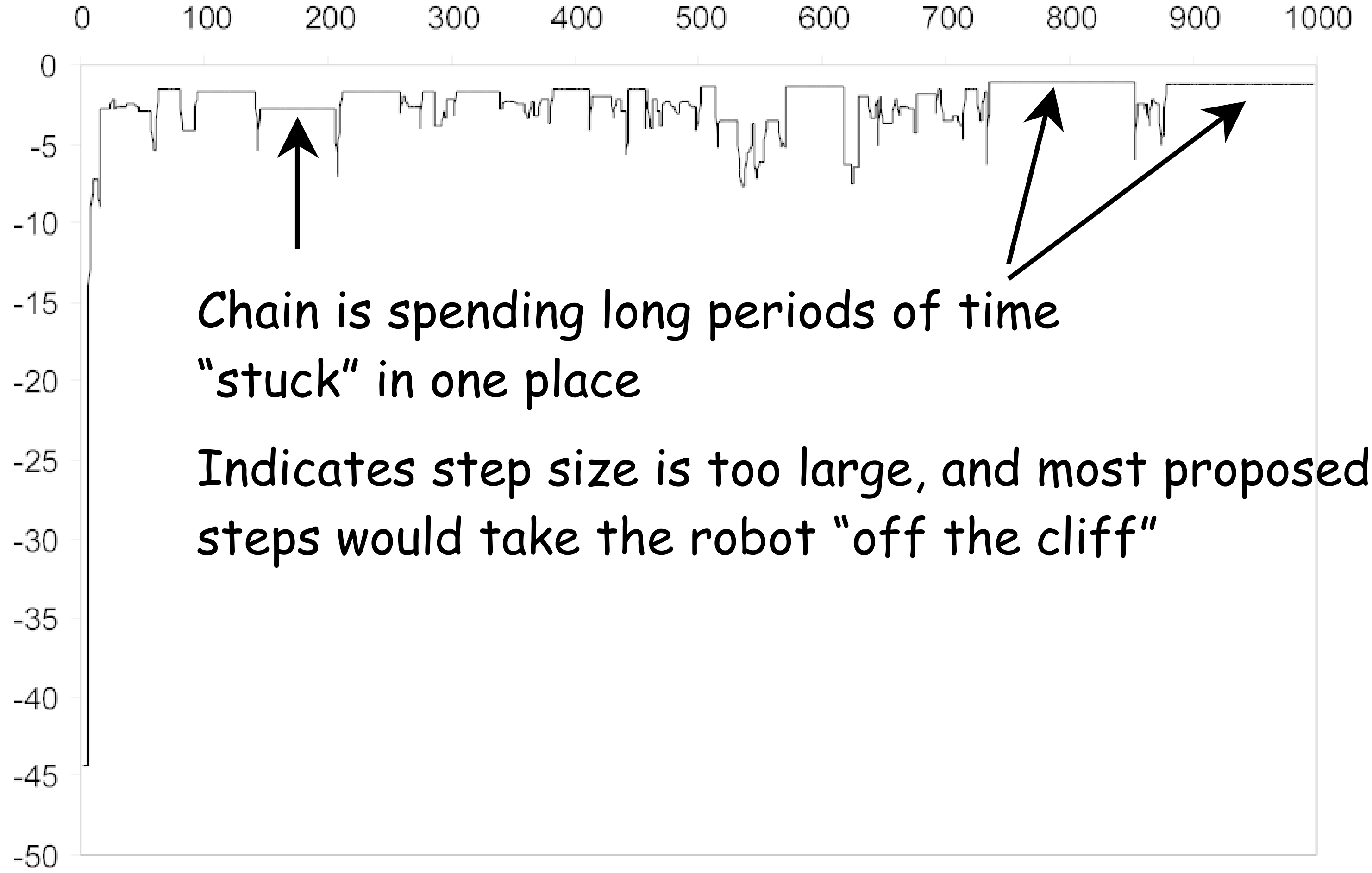
"White noise" appearance is a good sign

"Burn in" is over right about here

Important! This is a plot of first 1000 steps, and there is no indication that anything is wrong (but we know for a fact that we didn't let this one run long enough)

Slide Courtesy of Dr. Paul Lewis, University of Connecticut

Slow mixing

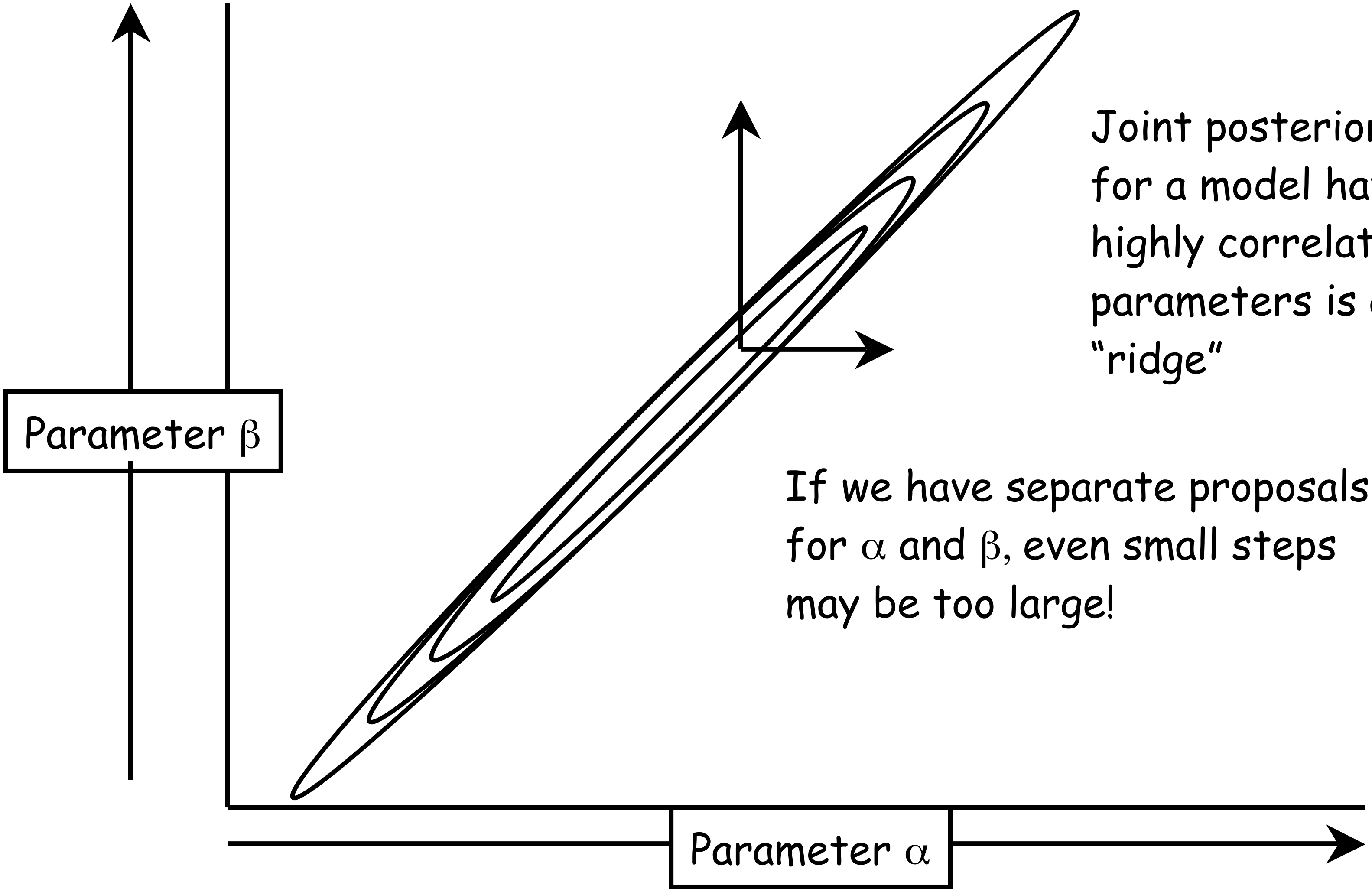


Chain is spending long periods of time "stuck" in one place

Indicates step size is too large, and most proposed steps would take the robot "off the cliff"

Slide Courtesy of Dr. Paul Lewis, University of Connecticut

The problem of co-linearity

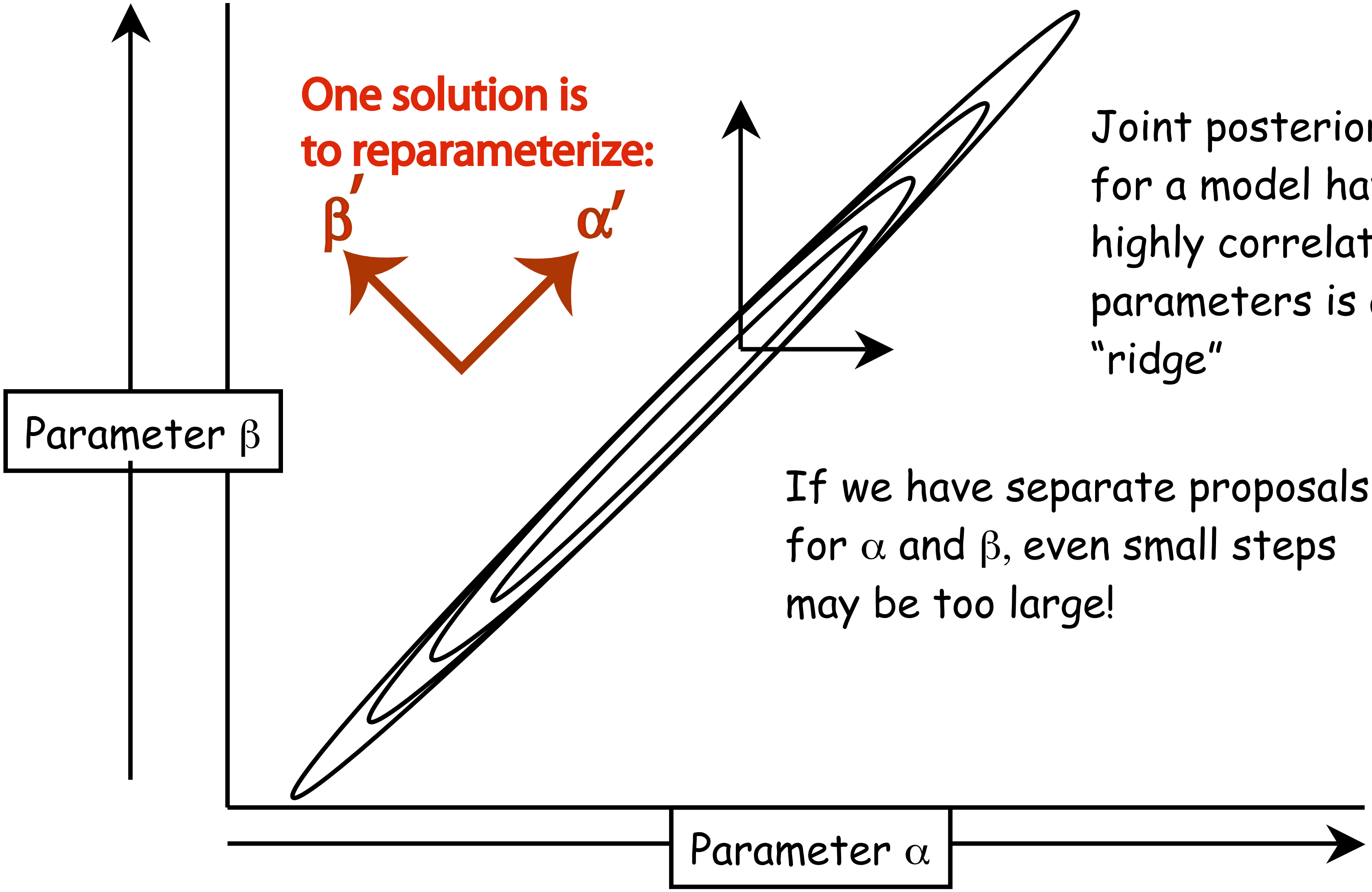


Joint posterior density for a model having two highly correlated parameters is a narrow "ridge"

If we have separate proposals for α and β, even small steps may be too large!

Slide Courtesy of Dr. Paul Lewis, University of Connecticut

The problem of co-linearity



One solution is to reparameterize:

Joint posterior density for a model having two highly correlated parameters is a narrow "ridge"

If we have separate proposals for α and β, even small steps may be too large!

Slide Courtesy of Dr. Paul Lewis, University of Connecticut

Imagine $\theta_1, \theta_2, \ldots, \theta_N$ are independent samples from posterior probability distribution with mean $\mu = E[\theta|X]$ and variance σ^2.

Sample mean $\overline{\theta}$ estimates μ . . .

$$\overline{\theta} = \frac{1}{N} \sum_{i=1}^{N} \theta_i$$

$$\mathrm{Var}(\overline{\theta}) = \frac{\sigma^2}{N}$$

If samples $\theta_1, \theta_2, \ldots, \theta_N$ are **autocorrelated** (i.e., are **not** independent) samples from posterior distribution, then define N_{eff} such that

$$\mathrm{Var}(\overline{\theta}) = \frac{\sigma^2}{N_{\text{eff}}}.$$

N_{eff} is the "effective sample size" for θ.

$$N_{\text{eff}} = \frac{N}{1 + 2\,\Sigma_{k=1}^{\infty}\,\rho_k(\theta)}$$

where $\rho_k(\theta)$ is known as autocorrelation of lag k (i.e., correlation between value and value k steps later).

Problems with MCMC approaches:

1. They are difficult to implement. Implementations may need to be clever to be computationally tractable.

2. For the kinds of complicated situations that biologists face, it may be very difficult to know how fast the Markov chain converges to the desired posterior distribution.

There are diagnostics for evaluating whether a chain has converged to the posterior distribution but often the diagnostics do not provide a guarantee of convergence.

Example diagnostics include randomly choosing the initial state of the Markov chain and then determining whether different MCMC runs yield about the same estimated posterior distribution.

Also, people will often look at how the posterior probability (or something proportional to it such as $\Pr\left(\theta^{(t)}\right)\Pr\left(X \mid \theta^{(t)}\right)$) changes as t changes. If the chain has some obvious pattern or trend in terms of a plot of the posterior probability versus t, then the chain has not converged.

Problems with Bayesian approaches in general:

1. Disagreements over priors.

2. Heavy Computational Requirements

(problem 2 is rapidly becoming less noteworthy)

Markov Chain Monte Carlo and Relatives

CARLIN, B.P., and T.A. LOUIS. 1996. Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall, London.

GELMAN, A., J.B. CARLIN, H.S. STERN, and D.B. RUBIN. 1995. Bayesian Data Analysis. Chapman and Hall, London.

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

METROPOLIS, N., A.W. ROSENBLUTH, M.N. ROSENBLUTH, A.H. TELLER, and E. TELLER. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. **21**: 1087–1092.

The MCMCRobot software by Dr. Paul Lewis is an excellent software program for illustrating the Metropolis-Hastings algorithm. It is freely available at:

http://phylogeny.uconn.edu/software/