Introduction to Marko Chain Monte Carlo

Meaningful Modeling of Epidemiologic Data, 2016
AIMS, Muizenberg, South Africa

Steve Bellan, PhD, MPH
Postdoctoral Fellow
The University of Texas at Austin

Bayes Theorem

∩ denotes "and"

| denotes “given”

Bayesian Statistics

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})\, P(\text{model})}{P(\text{data})}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{P(\text{data})}$$

By “model”, we often mean a specific model parameterization

Bayesian Statistics

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters})\, P(\text{parameters})}{P(\text{data})}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{P(\text{data})}$$

Bayesian Statistics

$$P(\theta \mid y) = \frac{P(y \mid \theta)\, P(\theta)}{P(y)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{P(y)}$$

What is $P(y)$?

$$P(y) = \int P(y \mid \theta) P(\theta) d\theta$$

Probability of observing *y* marginal over all possible values of *θ*.

Bayesian Statistics

$$P(\theta \mid y) = \frac{P(y \mid \theta)\, P(\theta)}{P(y)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{P(y)}$$

What is $P(y)$?

$$P(y) = \int \text{likelihood} \times \text{prior}\; d\theta$$

Probability of observing y marginal over all possible values of θ.

Bayesian Statistics

$$P(\theta \mid y) = \frac{P(y \mid \theta)\, P(\theta)}{P(y)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\int \text{likelihood} \times \text{prior}\, d\theta}$$

Denominator is a scalar number, but rarely has an analytical solution.

Very difficult to calculate in practice!

How can we calculate posterior without denominator?

Bayesian Statistics

$$P(\theta \mid y) \propto P(y \mid \theta)\, P(\theta)$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

The posterior is proportional to the numerator up to some scalar constant.

We also know that the posterior integrates to 1 because it is a PDF.

$$\int P(\theta \mid y)\, d\theta = 1$$

Bayesian Statistics

$$P(\theta \mid y) \propto P(y \mid \theta)\, P(\theta)$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

For a specified model world,

the posterior distribution of θ is based on

observations y & our prior beliefs about θ.

Motivating MCMC

Want to calculate the posterior but can't do it directly because of the troublesome denominator integral.

However, we do know the following:

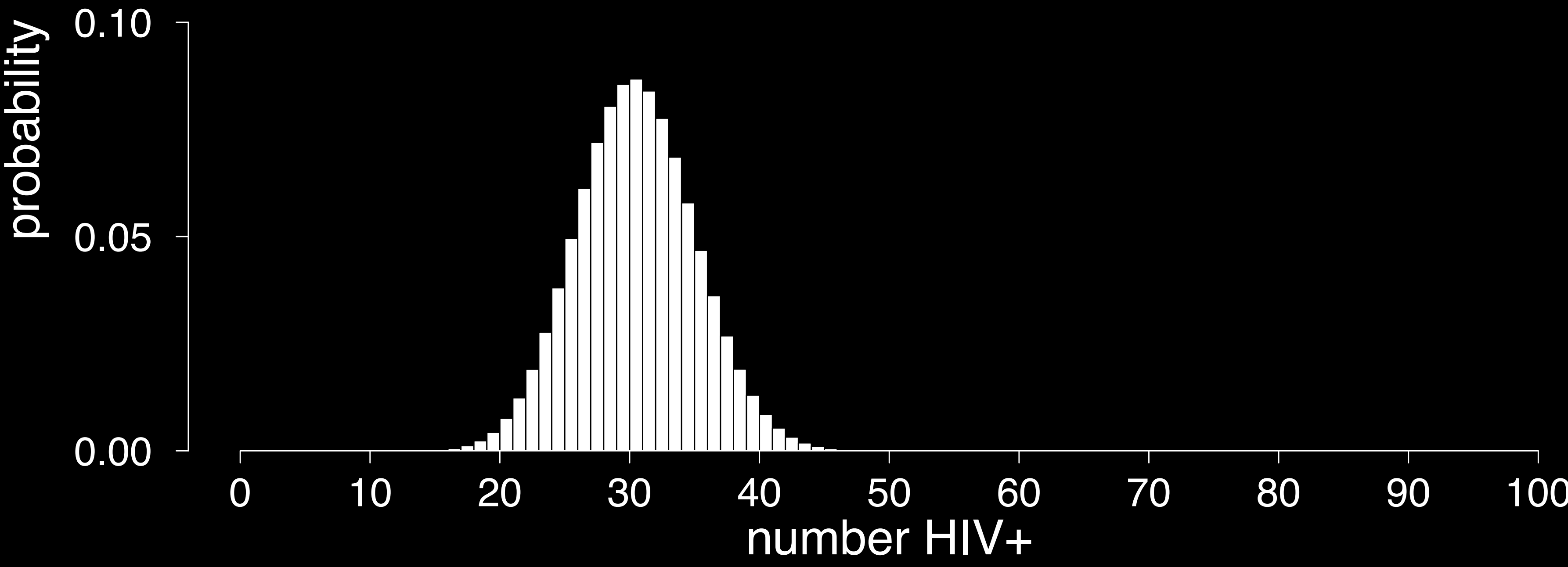
$$P(\theta \mid y) \propto P(y \mid \theta)\, P(\theta)$$

$$\int P(\theta \mid y)\, d\theta = 1$$

MCMC takes advantage of these two items to numerically approximate the posterior

In a population of 1,000,000 people with a true prevalence of 30%, the probability distribution of number of positive individuals if 100 are sampled:

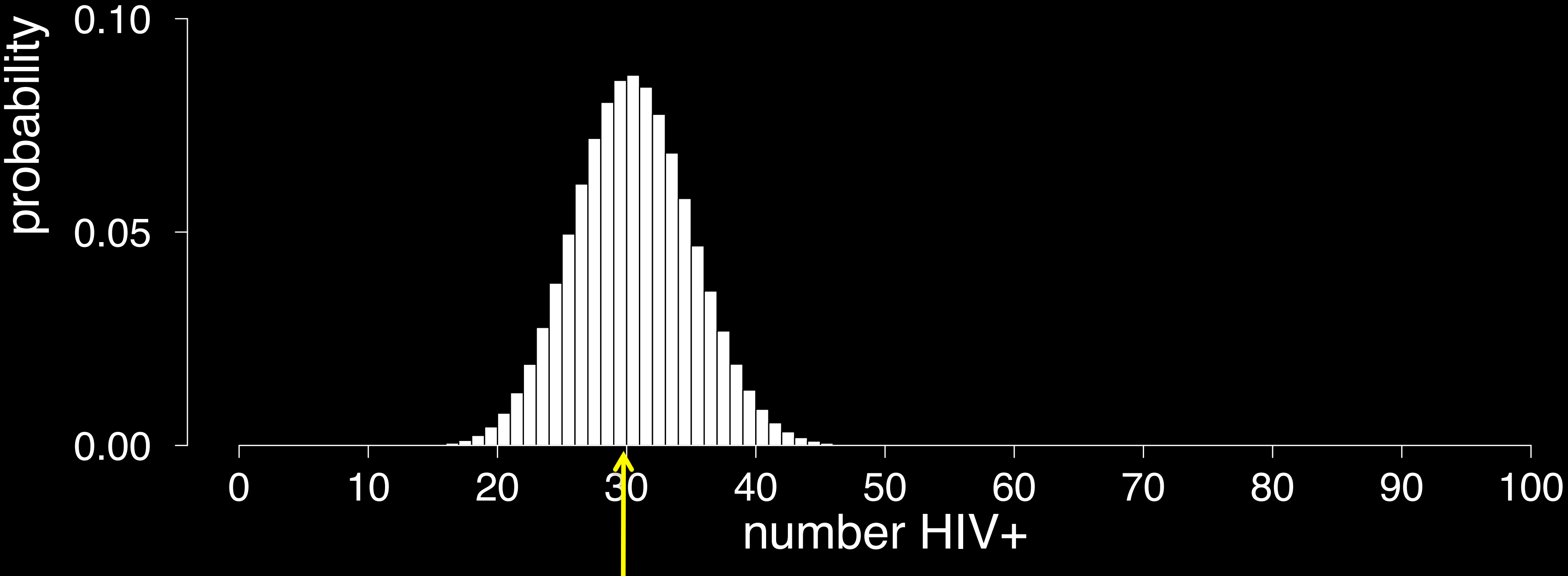
$$f(x) = \binom{100}{x}(0.3)^x(0.7)^{100-x}$$




```
barplot(dbinom(x = 0:100, size = 100, prob = .3), names.arg = 0:size)
```

In a population of 1,000,000 people with a true prevalence of 30%, the probability distribution of number of positive individuals if 100 are sampled:

$$f(x) = \binom{100}{x}(0.3)^x(0.7)^{100-x}$$



We sample 100 people once and 28 are positive:

```
> rbinom(n = 1, size = 100, prob = .3)
[1] 28
```

Defining Likelihood

- L(parameter | data) = p(data | parameter)
- Not a probability distribution.
- Probabilities taken from many different distributions.

function of x

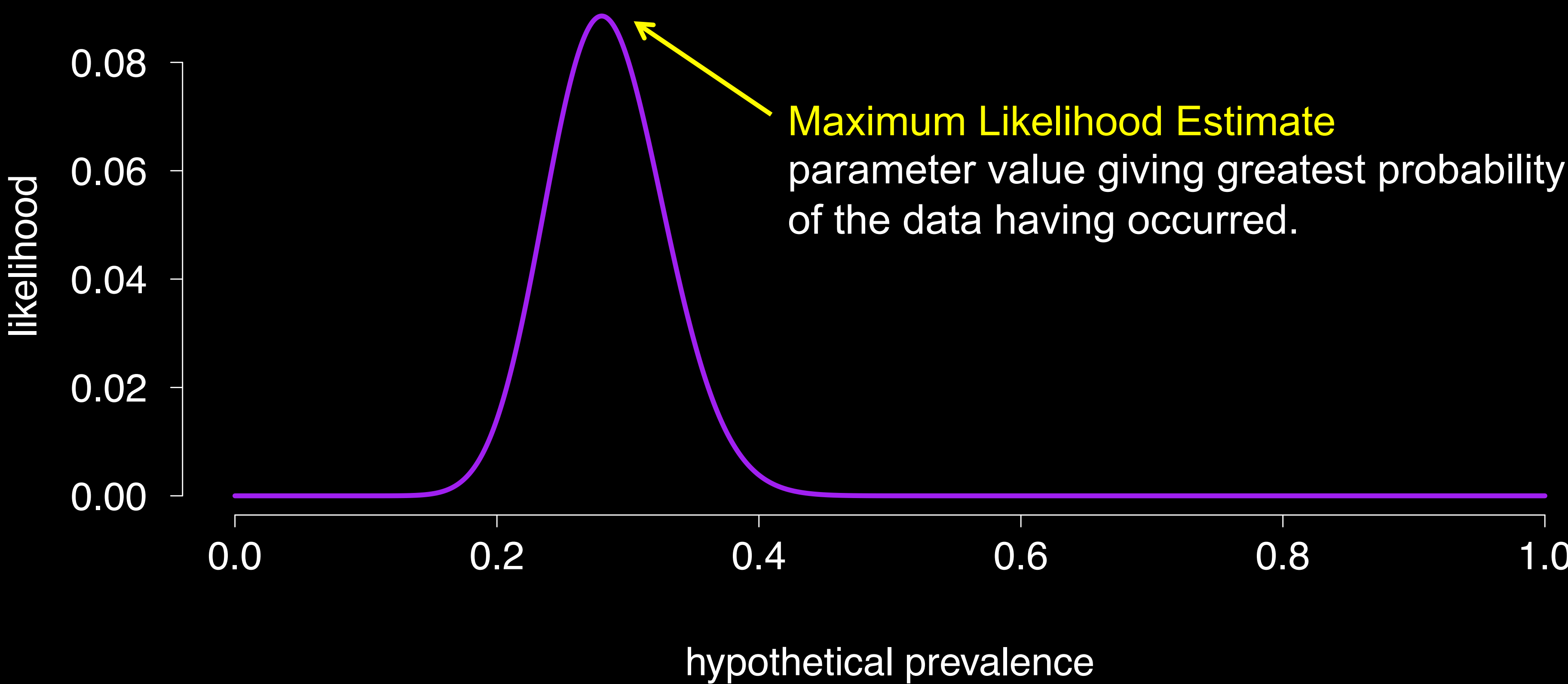
PDF: $f(x|p) = \binom{n}{x}(p)^x(1-p)^{n-x}$

LIKELIHOOD: $L(p|x) = \binom{n}{x}(p)^x(1-p)^{n-x}$

function of p

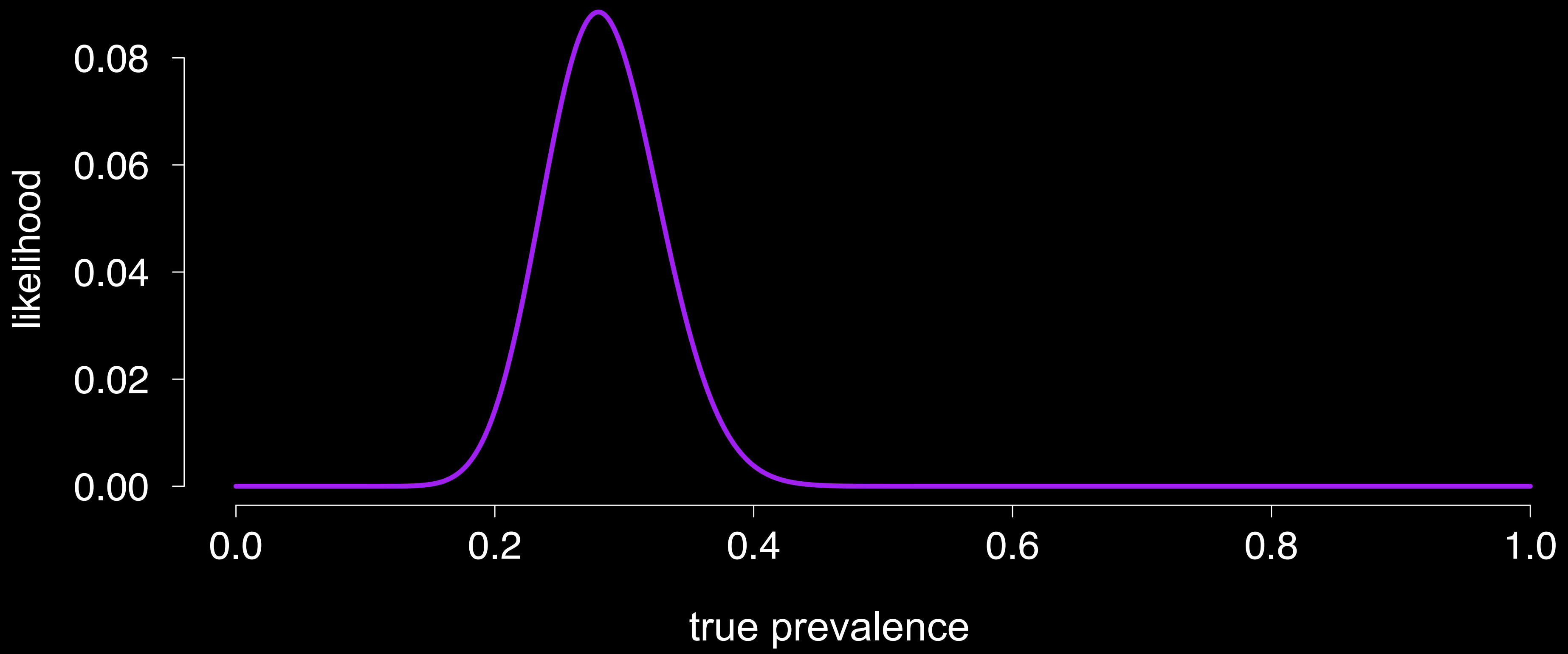
Frequentist ML: Estimate the MLE and 95% confidence bounds.

$$P(x = 28, n = 100 \mid \theta)$$



Bayesian Inference: Calculate the posterior probability distribution of every possible parameter value

$$P(x = 28, n = 100 \mid \theta)$$



$$P(\theta \mid x = 28, n = 100)$$

MCMC

What is the posterior probability distribution for the prevalence θ

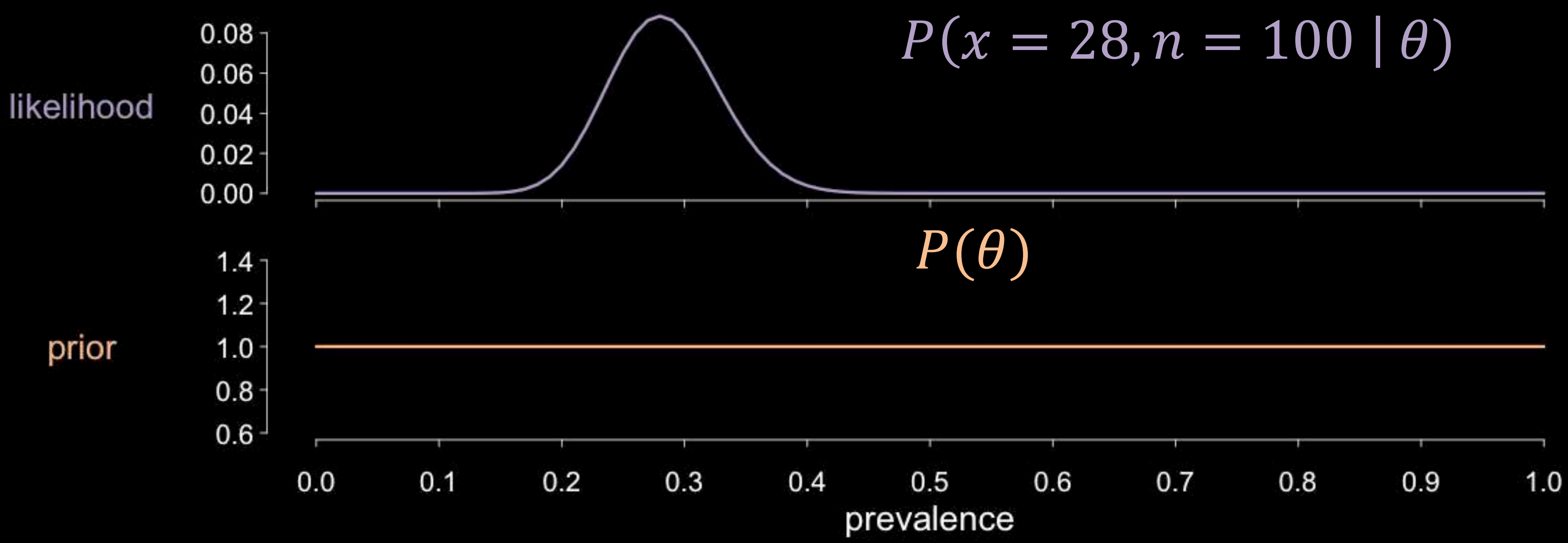
- when we observe 28/100 positive
- & if we have no prior beliefs about plausible prevalence

$$P(\theta \mid x = 28, n = 100) \propto P(x = 28, n = 100 \mid \theta)\, P(\theta)$$

$$\int P(\theta \mid x = 28, n = 100)\, d\theta = 1$$

Let’s use the Metropolis-Hastings MCMC algorithm to estimate

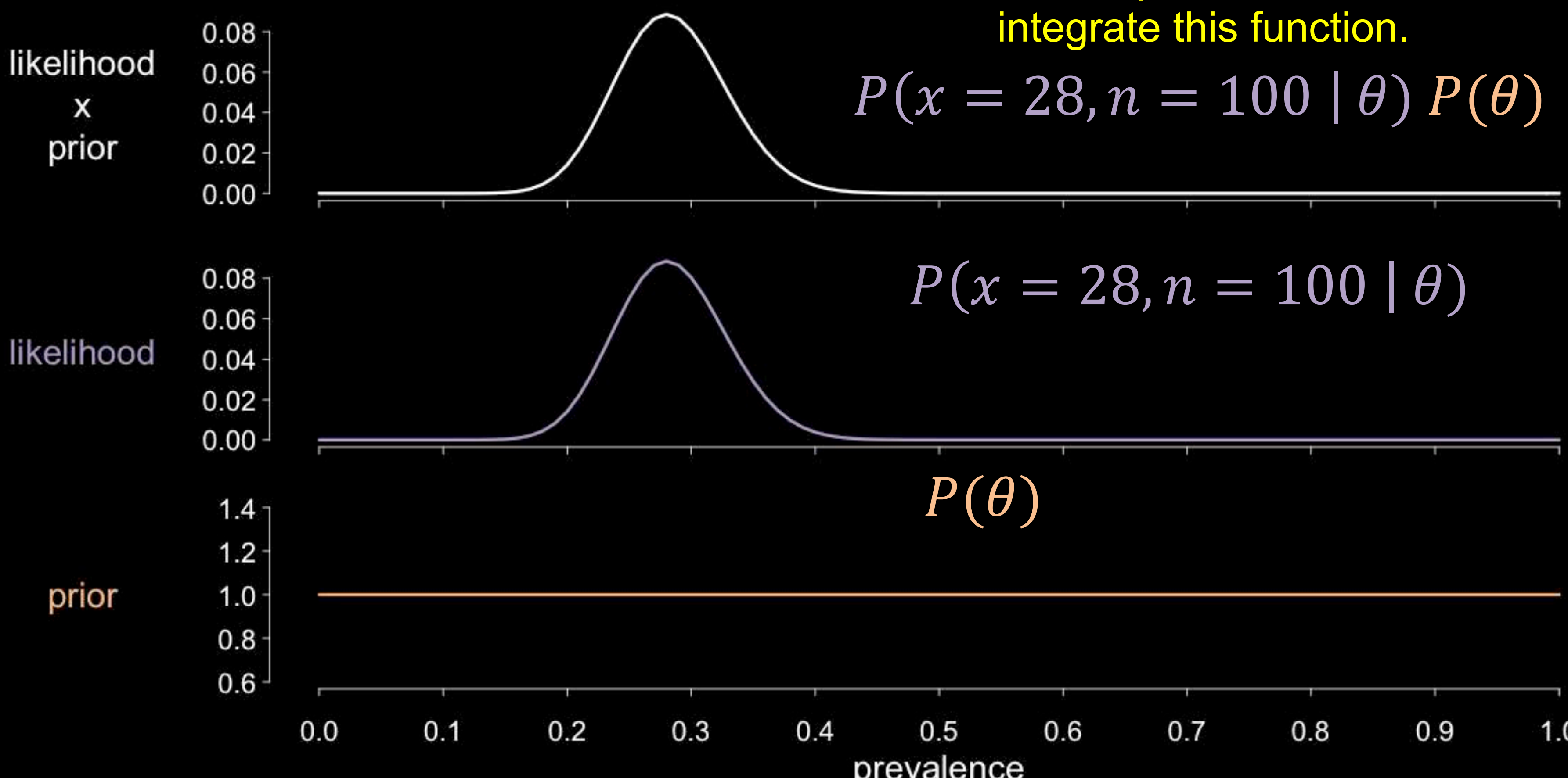
$$P(\theta \mid x = 28, n = 100)$$



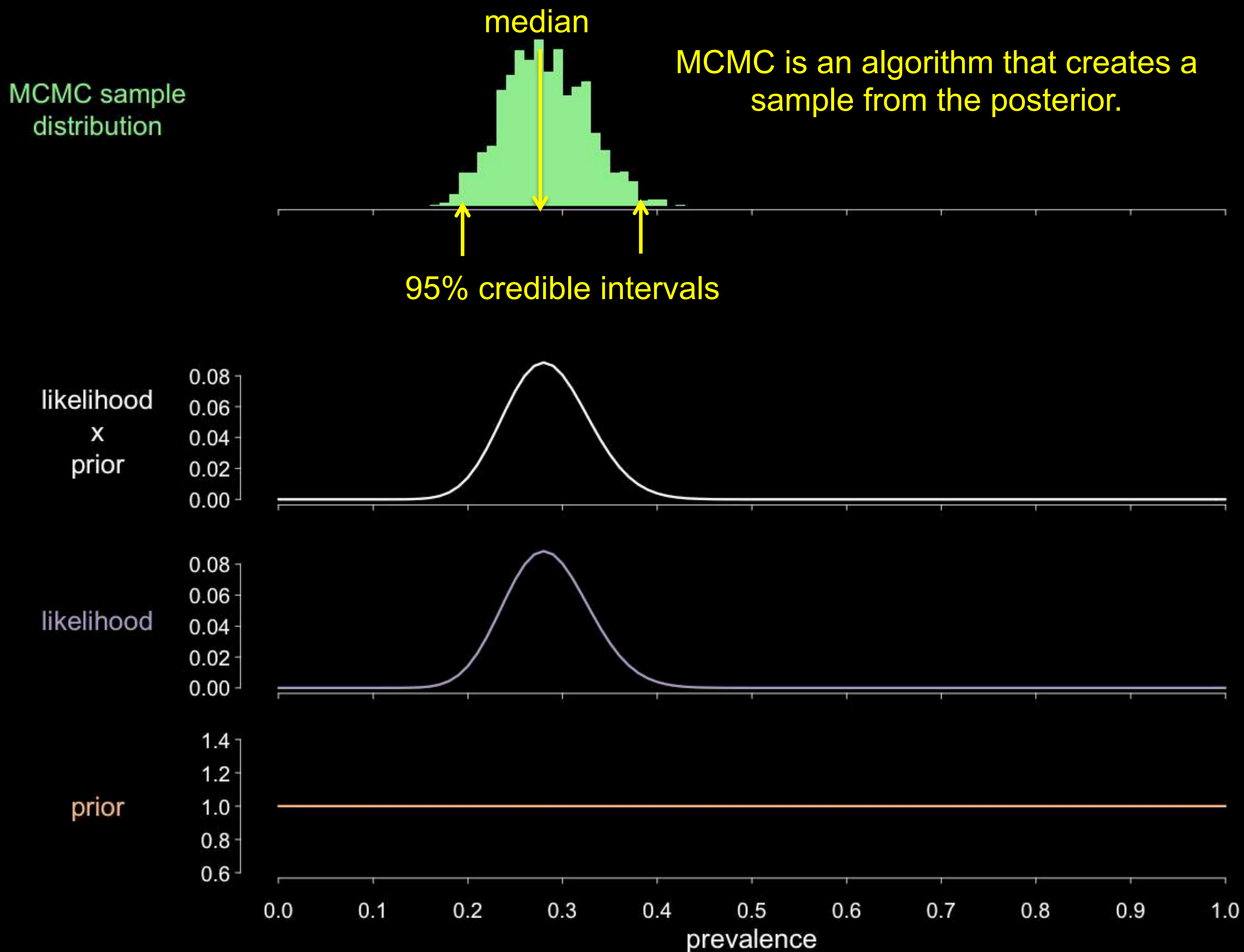
$P(x = 28, n = 100 \mid \theta)$
likelihood
0.08
0.06
0.04
0.02
0.00
$P(\theta)$
prior
1.4
1.2
1.0
0.8
0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
prevalence

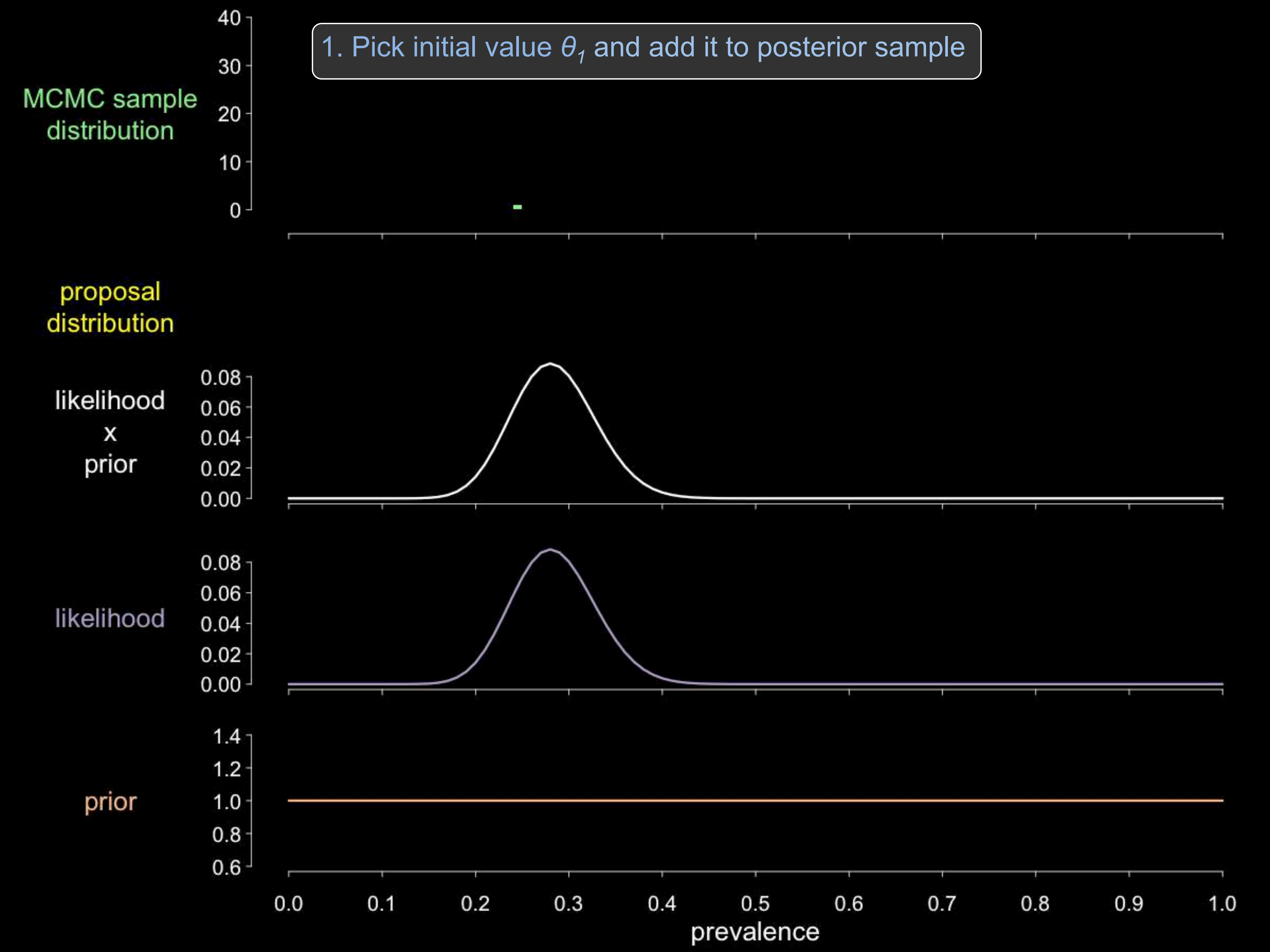
$$P(\theta \mid x = 28, n = 100) = \frac{P(x = 28, n = 100 \mid \theta)\, P(\theta)}{\int P(x = 28, n = 100 \mid \theta)\, P(\theta) d\theta}$$

For nontrivial problems, we cannot integrate this function.

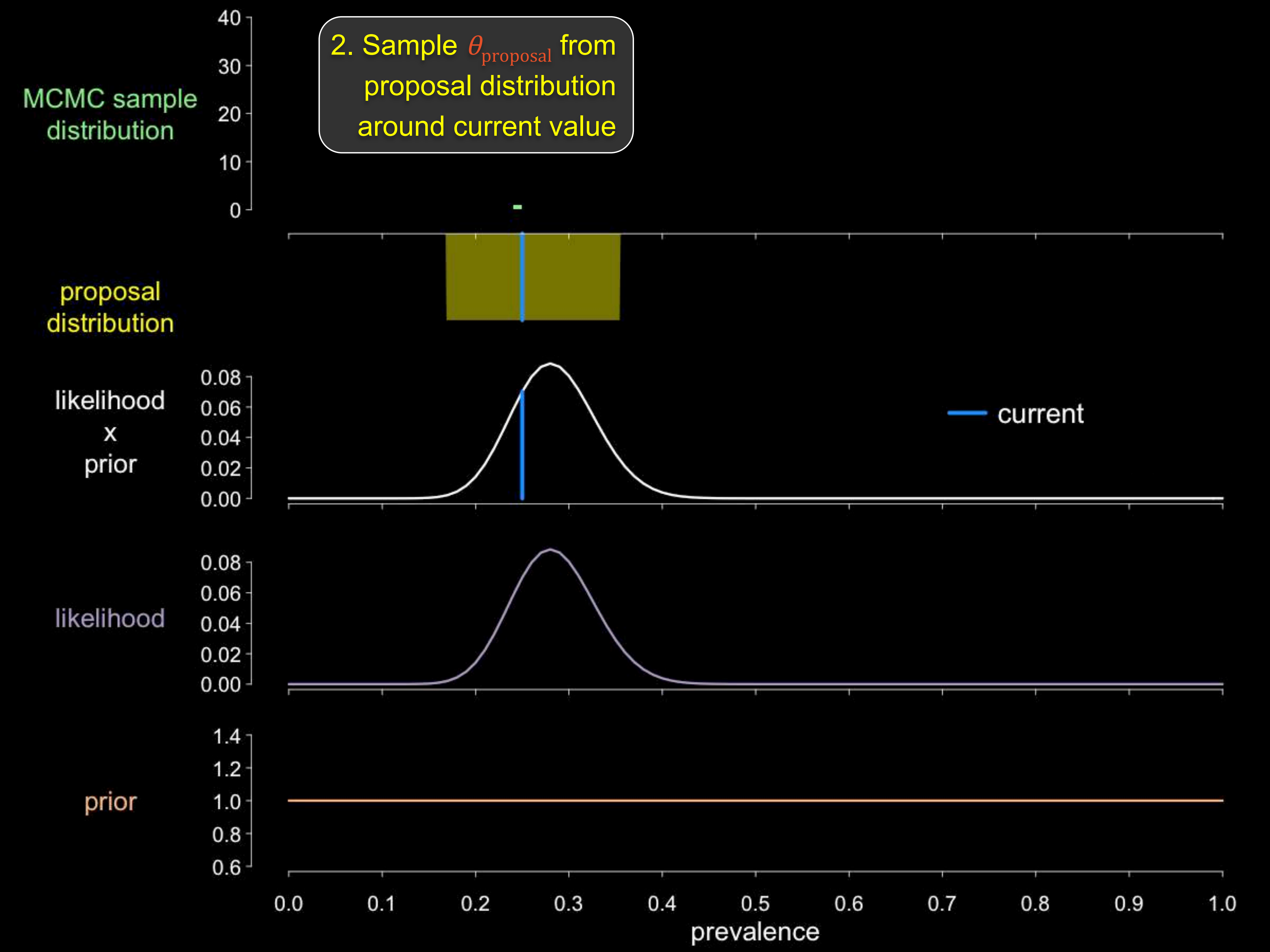


MCMC is an algorithm that creates a sample from the posterior.

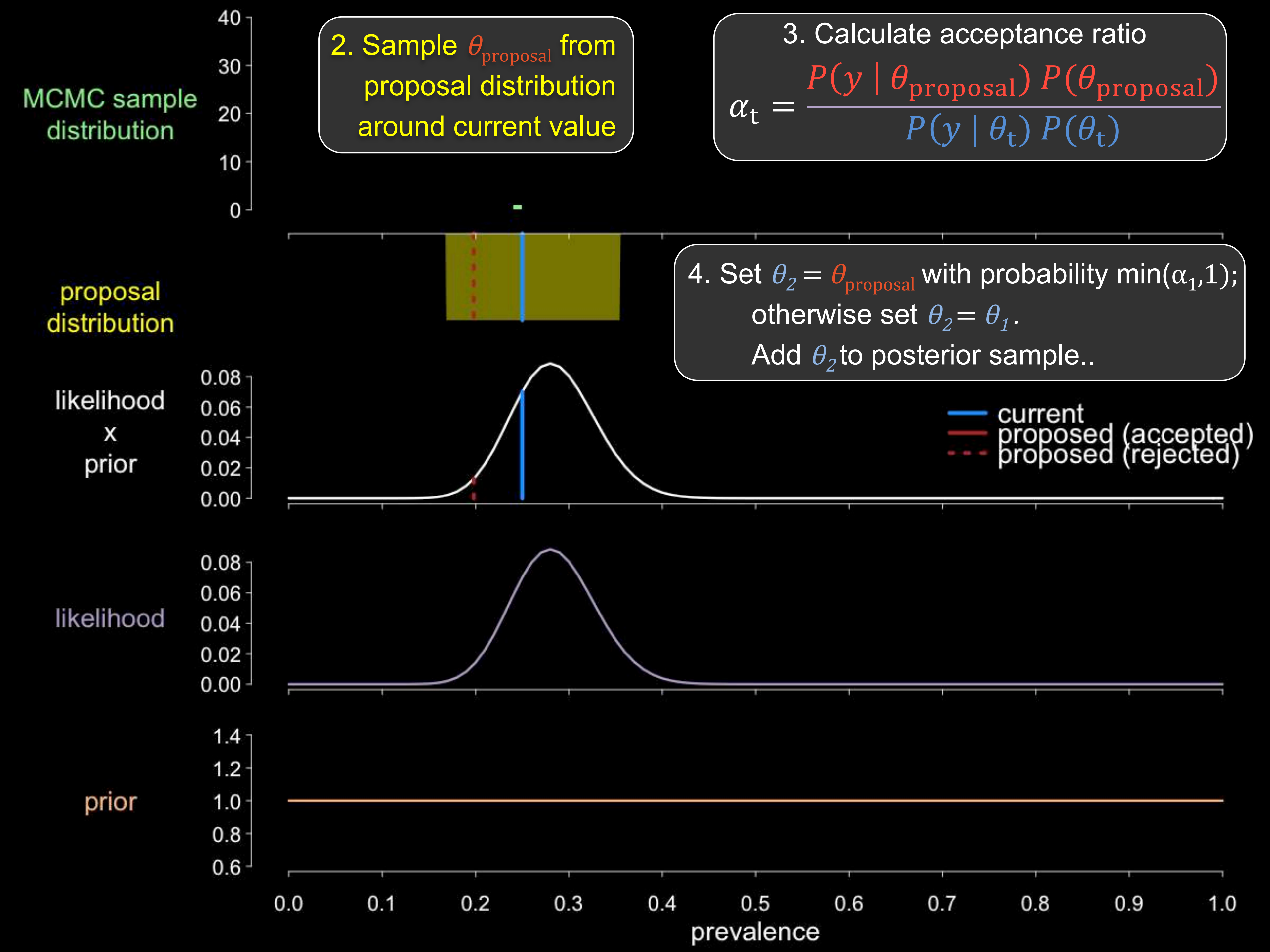




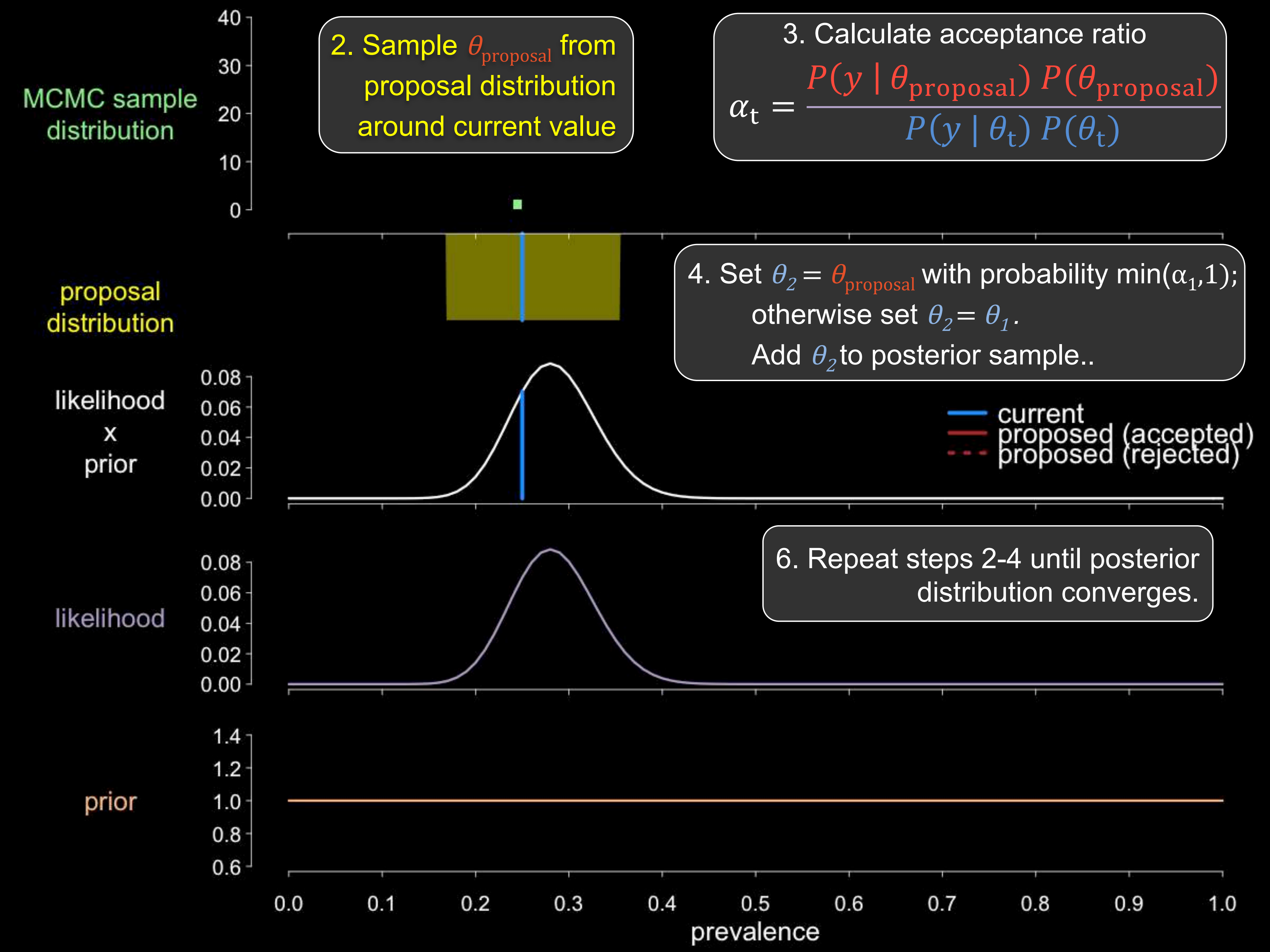
1. Pick initial value θ_1 and add it to posterior sample
MCMC sample distribution
40
30
20
10
0
proposal distribution
likelihood x prior
0.08
0.06
0.04
0.02
0.00
likelihood
0.08
0.06
0.04
0.02
0.00
prior
1.4
1.2
1.0
0.8
0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
prevalence

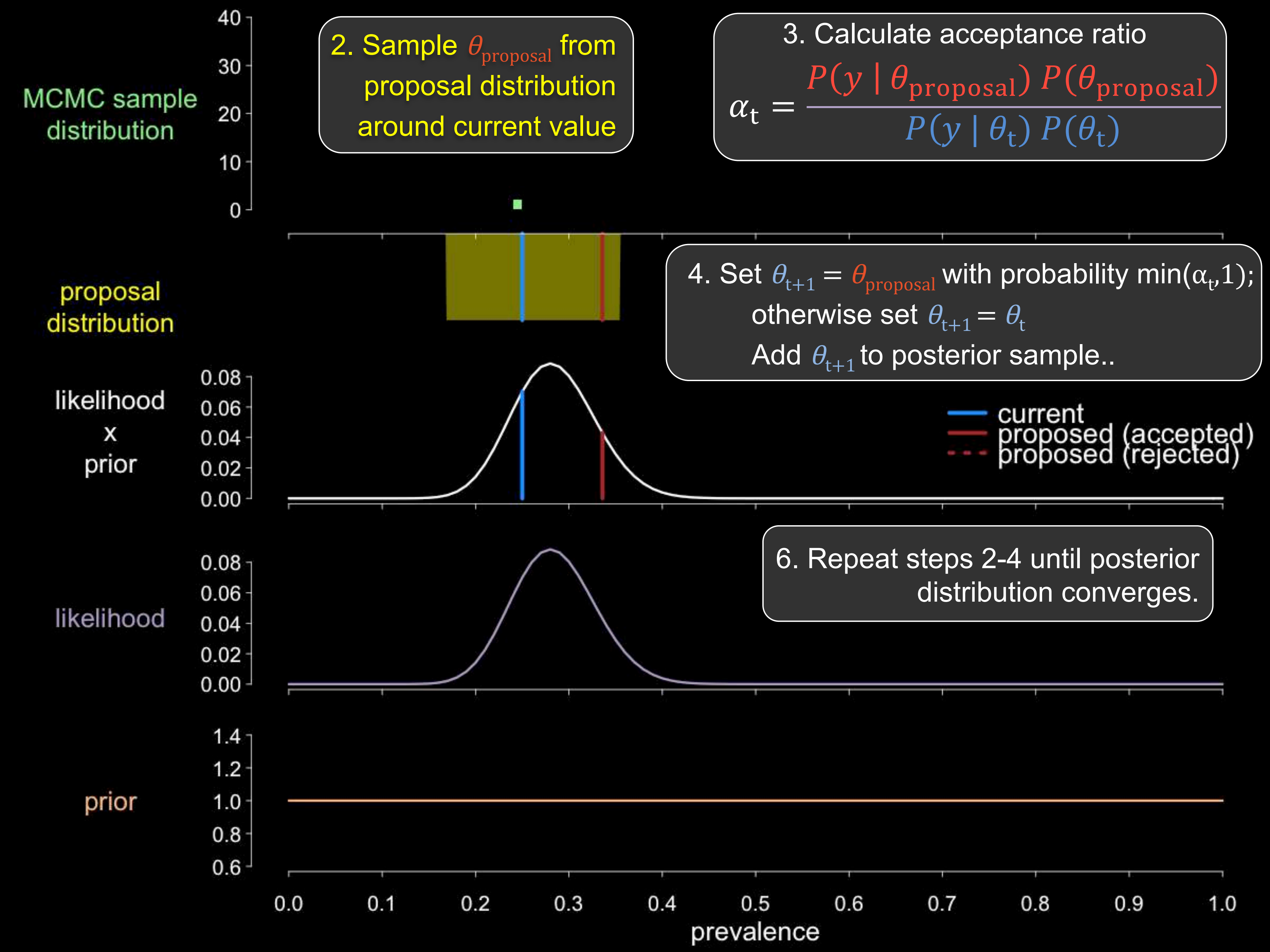


2. Sample θ_{proposal} from proposal distribution around current value
MCMC sample distribution
40
30
20
10
0
proposal distribution
likelihood x prior
0.08
0.06
0.04
0.02
0.00
current
likelihood
0.08
0.06
0.04
0.02
0.00
prior
1.4
1.2
1.0
0.8
0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
prevalence

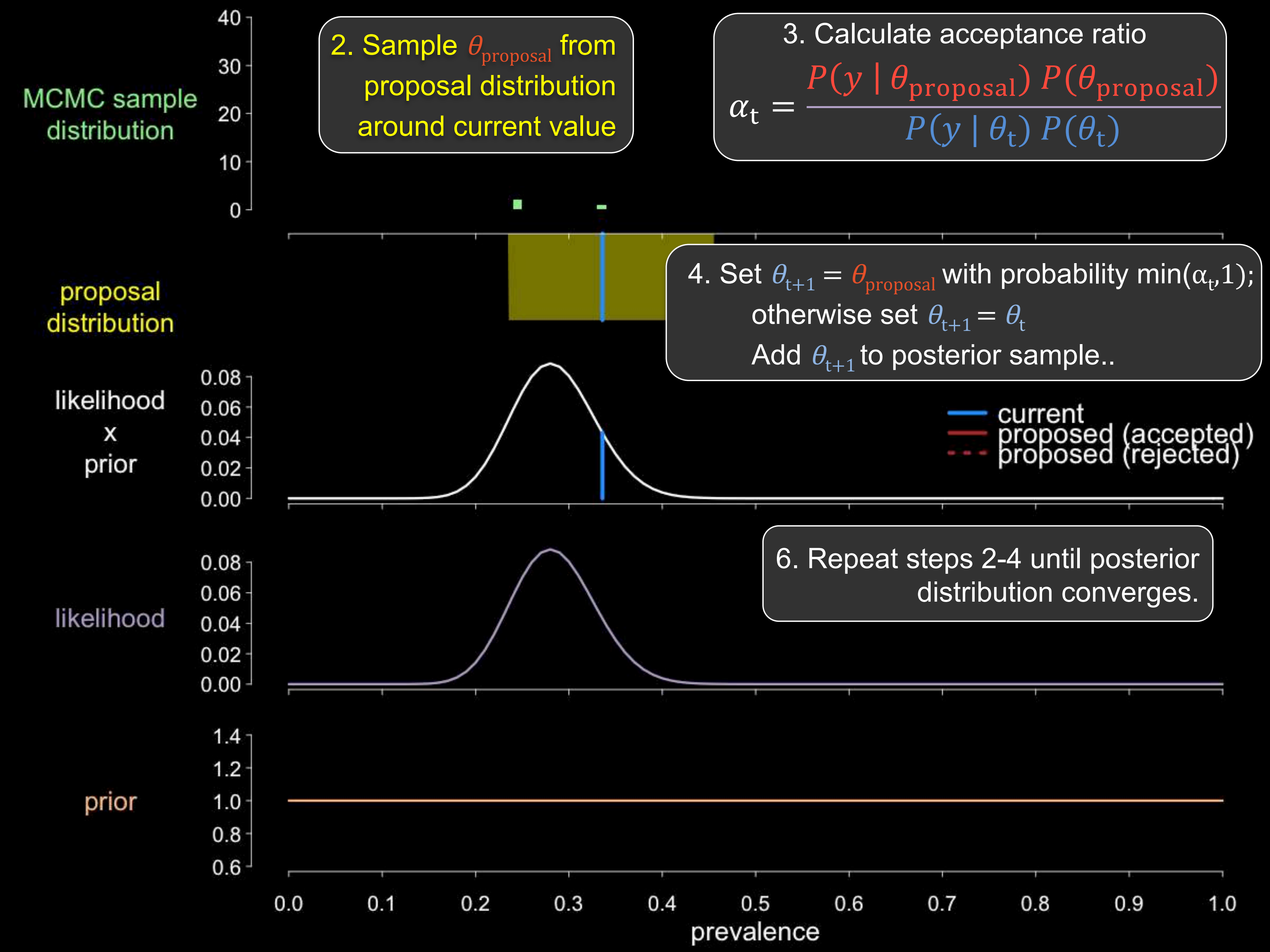


MCMC sample distribution
2. Sample θ_{proposal} from proposal distribution around current value
3. Calculate acceptance ratio
$\alpha_t = \frac{P(y \mid \theta_{\text{proposal}})\, P(\theta_{\text{proposal}})}{P(y \mid \theta_t)\, P(\theta_t)}$
proposal distribution
4. Set $\theta_2 = \theta_{\text{proposal}}$ with probability $\min(\alpha_1, 1)$; otherwise set $\theta_2 = \theta_1$. Add θ_2 to posterior sample..
likelihood x prior
current
proposed (accepted)
proposed (rejected)
likelihood
prior
prevalence
40
30
20
10
0
0.08
0.06
0.04
0.02
0.00
1.4
1.2
1.0
0.8
0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

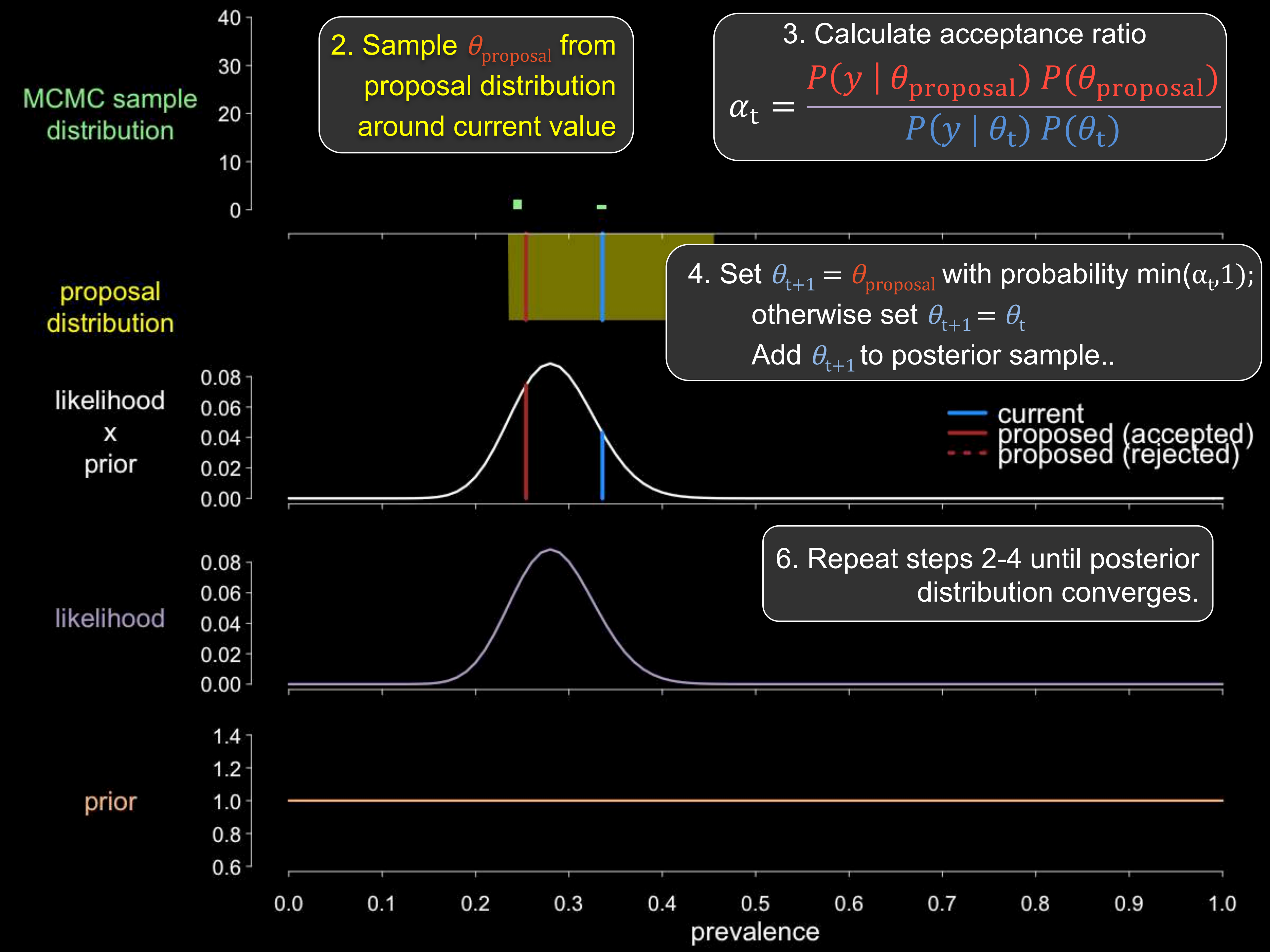

MCMC sample distribution
40
30
20
10
0
2. Sample θ_{proposal} from proposal distribution around current value
3. Calculate acceptance ratio
$\alpha_t = \frac{P(y \mid \theta_{\text{proposal}})\, P(\theta_{\text{proposal}})}{P(y \mid \theta_t)\, P(\theta_t)}$
proposal distribution
4. Set $\theta_2 = \theta_{\text{proposal}}$ with probability min$(\alpha_1,1)$;
otherwise set $\theta_2 = \theta_1$.
Add θ_2 to posterior sample..
likelihood x prior
0.08
0.06
0.04
0.02
0.00
current
proposed (accepted)
proposed (rejected)
6. Repeat steps 2-4 until posterior distribution converges.
likelihood
0.08
0.06
0.04
0.02
0.00
prior
1.4
1.2
1.0
0.8
0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
prevalence



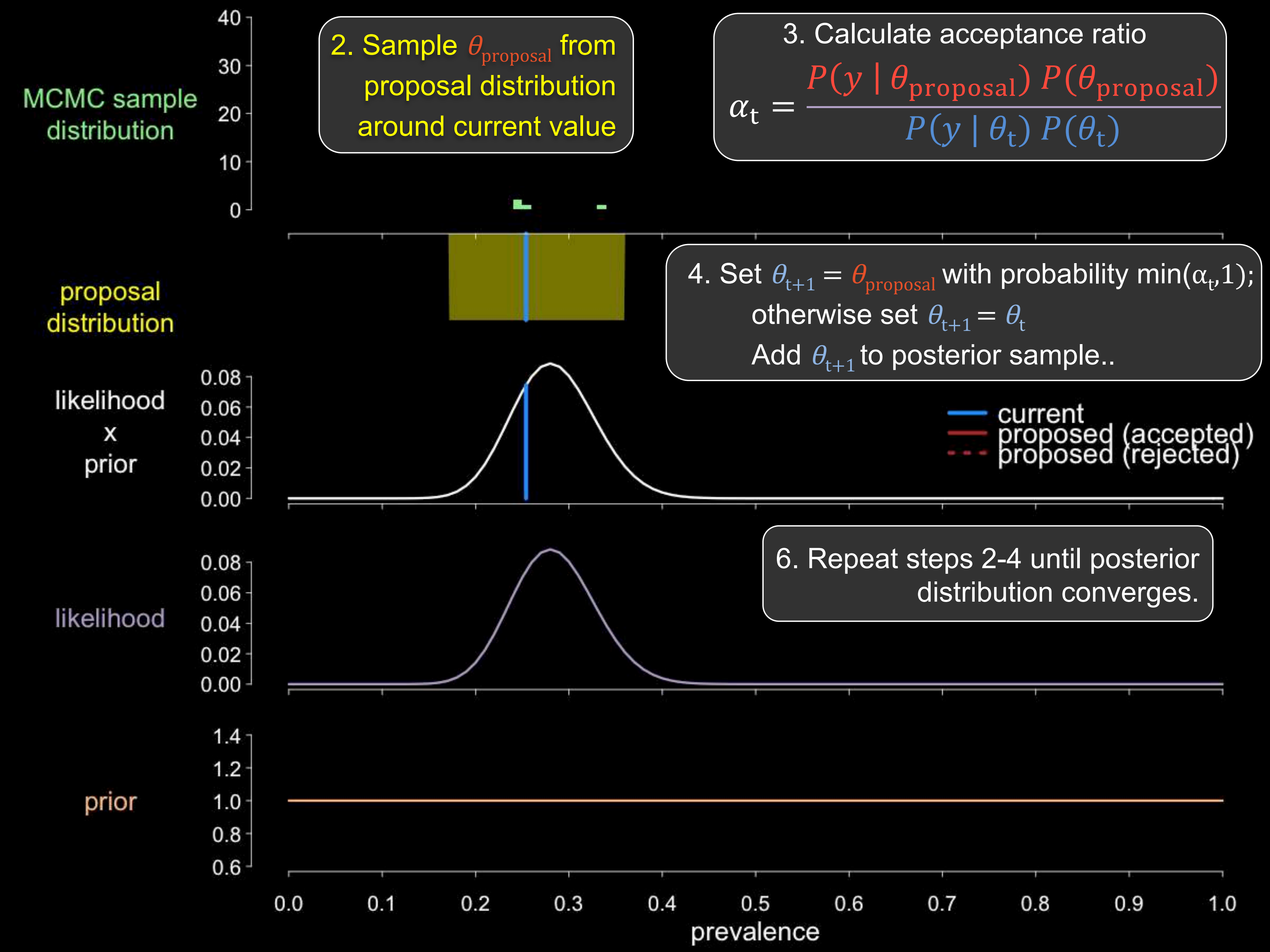
2. Sample $\theta_{proposal}$ from proposal distribution around current value
3. Calculate acceptance ratio
$\alpha_t = \frac{P(y \mid \theta_{proposal})\, P(\theta_{proposal})}{P(y \mid \theta_t)\, P(\theta_t)}$
4. Set $\theta_{t+1} = \theta_{proposal}$ with probability $\min(\alpha_t, 1)$;
otherwise set $\theta_{t+1} = \theta_t$
Add θ_{t+1} to posterior sample..
6. Repeat steps 2-4 until posterior distribution converges.
MCMC sample distribution
proposal distribution
likelihood x prior
likelihood
prior
current
proposed (accepted)
proposed (rejected)
prevalence
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0



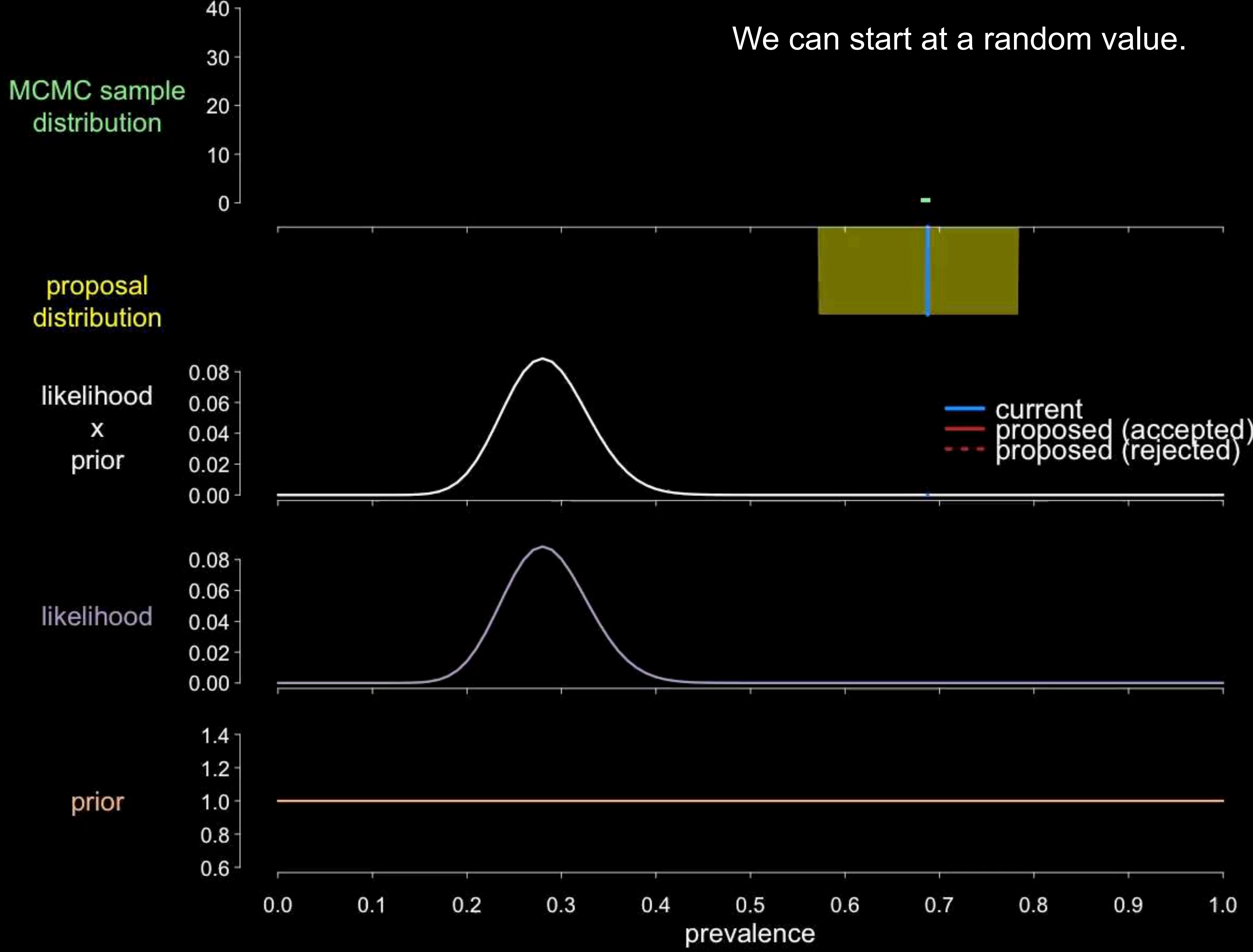
2. Sample θ_{proposal} from proposal distribution around current value
3. Calculate acceptance ratio
$\alpha_t = \frac{P(y \mid \theta_{\text{proposal}})\, P(\theta_{\text{proposal}})}{P(y \mid \theta_t)\, P(\theta_t)}$
4. Set $\theta_{t+1} = \theta_{\text{proposal}}$ with probability $\min(\alpha_t, 1)$; otherwise set $\theta_{t+1} = \theta_t$
Add θ_{t+1} to posterior sample..
6. Repeat steps 2-4 until posterior distribution converges.
MCMC sample distribution
proposal distribution
likelihood x prior
likelihood
prior
current
proposed (accepted)
proposed (rejected)
prevalence



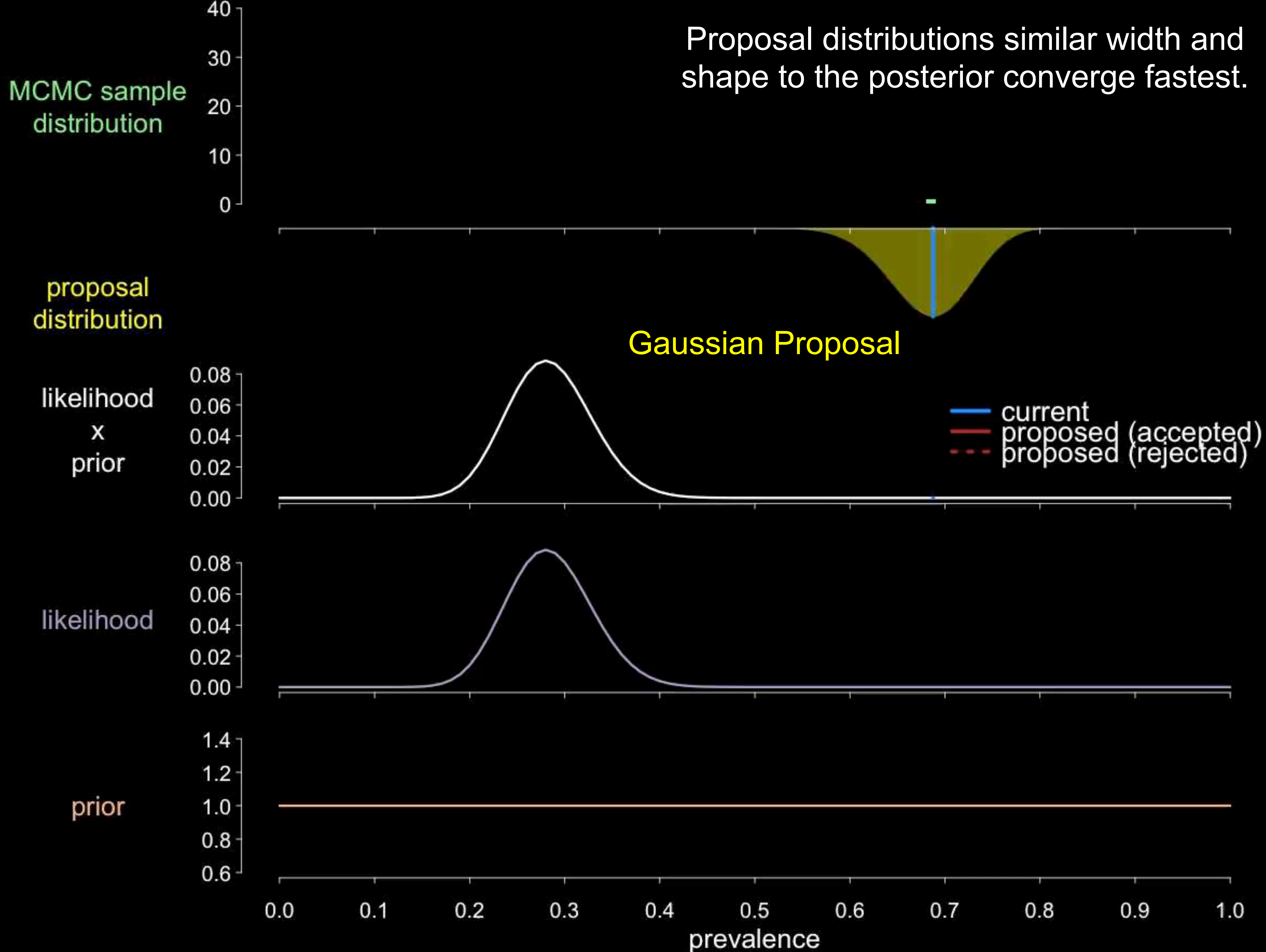
2. Sample $\theta_{proposal}$ from proposal distribution around current value
3. Calculate acceptance ratio
$\alpha_t = \frac{P(y \mid \theta_{proposal})\, P(\theta_{proposal})}{P(y \mid \theta_t)\, P(\theta_t)}$
4. Set $\theta_{t+1} = \theta_{proposal}$ with probability $\min(\alpha_t, 1)$;
otherwise set $\theta_{t+1} = \theta_t$
Add θ_{t+1} to posterior sample..
6. Repeat steps 2-4 until posterior distribution converges.
MCMC sample distribution
proposal distribution
likelihood x prior
likelihood
prior
current
proposed (accepted)
proposed (rejected)
prevalence


MCMC sample distribution
40
30
20
10
0
2. Sample θ_{proposal} from proposal distribution around current value
3. Calculate acceptance ratio
$\alpha_t = \frac{P(y \mid \theta_{\text{proposal}})\, P(\theta_{\text{proposal}})}{P(y \mid \theta_t)\, P(\theta_t)}$
proposal distribution
4. Set $\theta_{t+1} = \theta_{\text{proposal}}$ with probability $\min(\alpha_t, 1)$;
otherwise set $\theta_{t+1} = \theta_t$
Add θ_{t+1} to posterior sample..
likelihood x prior
0.08
0.06
0.04
0.02
0.00
current
proposed (accepted)
proposed (rejected)
6. Repeat steps 2-4 until posterior distribution converges.
likelihood
0.08
0.06
0.04
0.02
0.00
prior
1.4
1.2
1.0
0.8
0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
prevalence

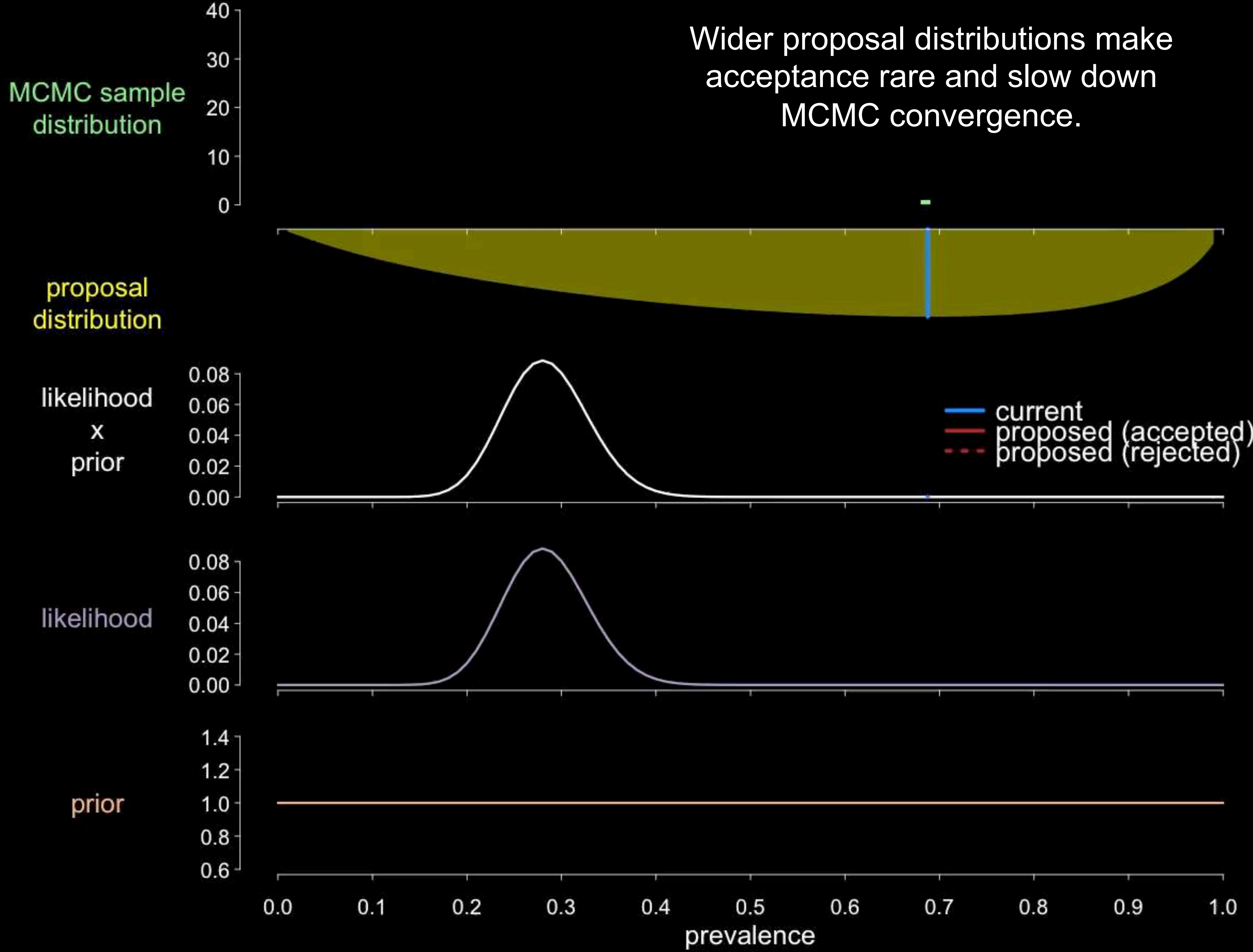
We can start at a random value.



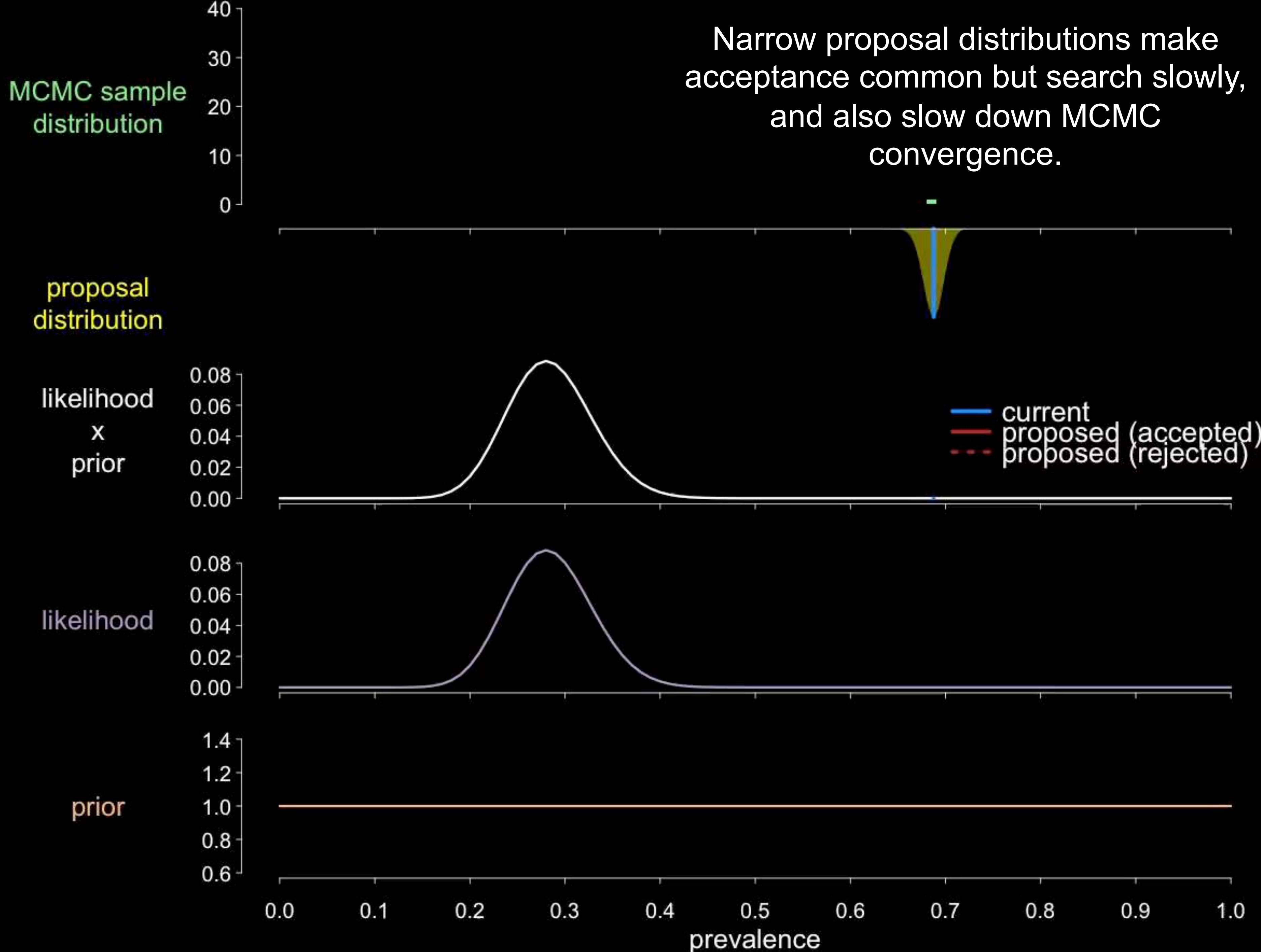
Proposal distributions similar width and shape to the posterior converge fastest.

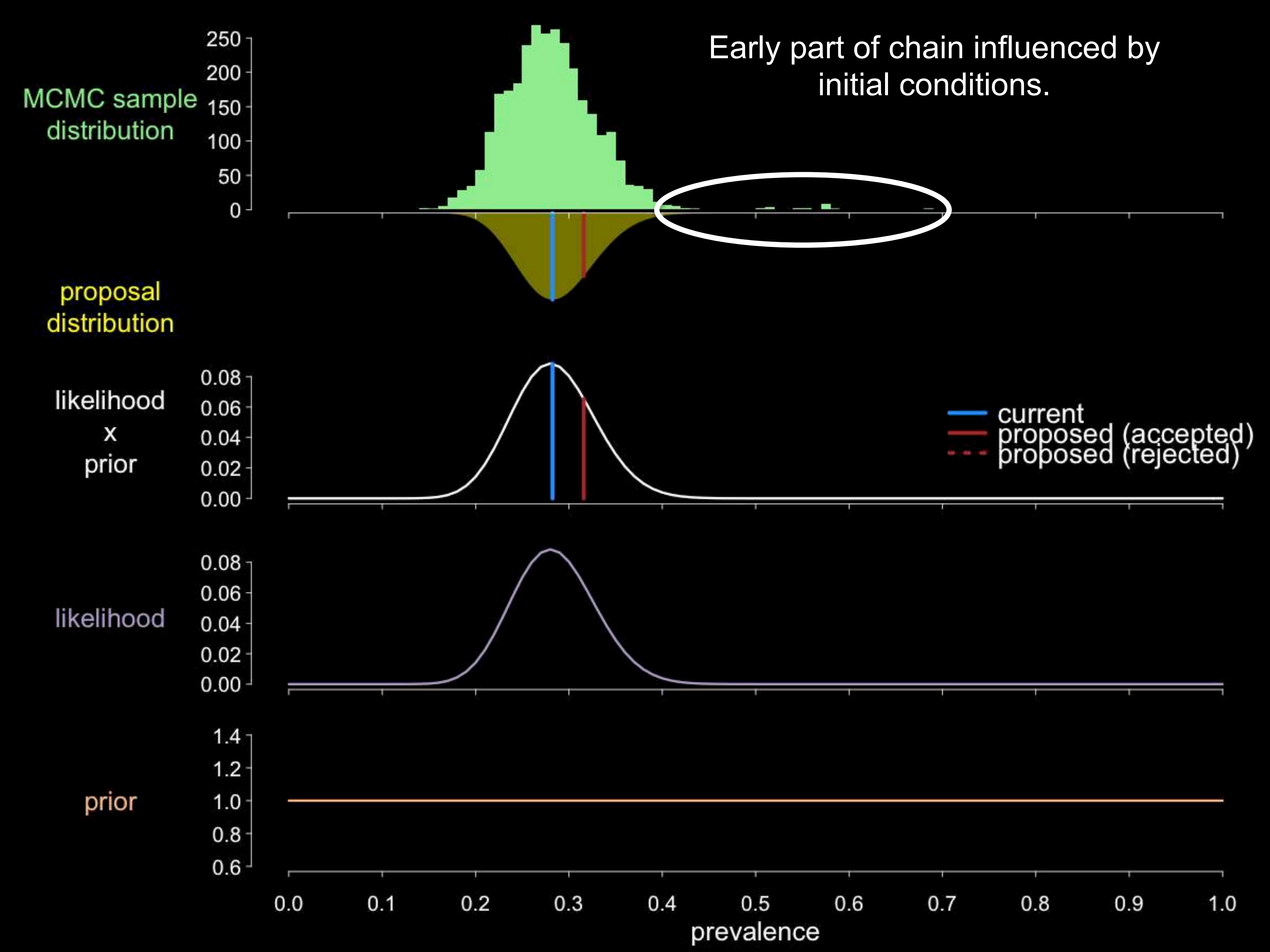


Wider proposal distributions make acceptance rare and slow down MCMC convergence.



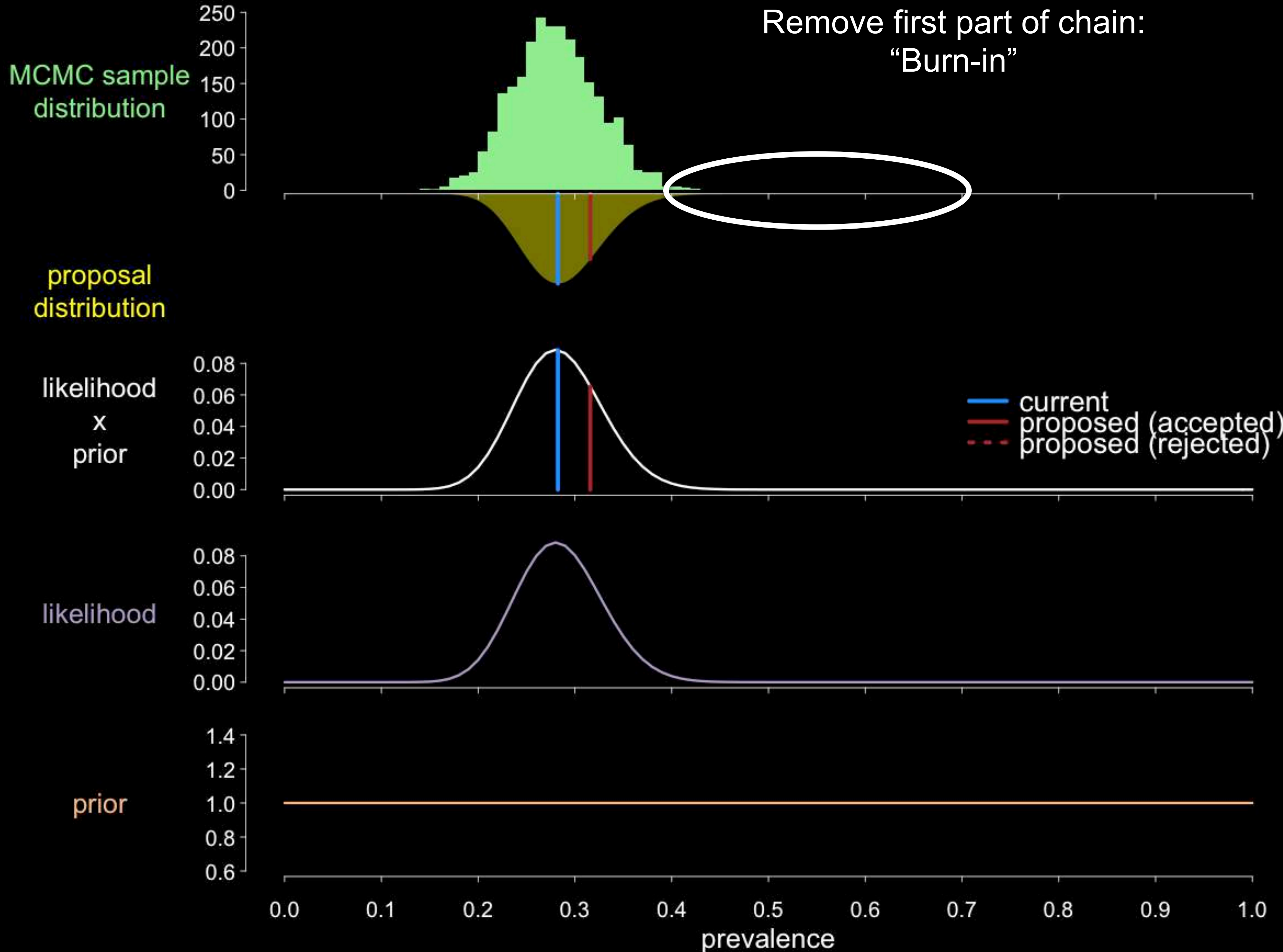
Narrow proposal distributions make acceptance common but search slowly, and also slow down MCMC convergence.

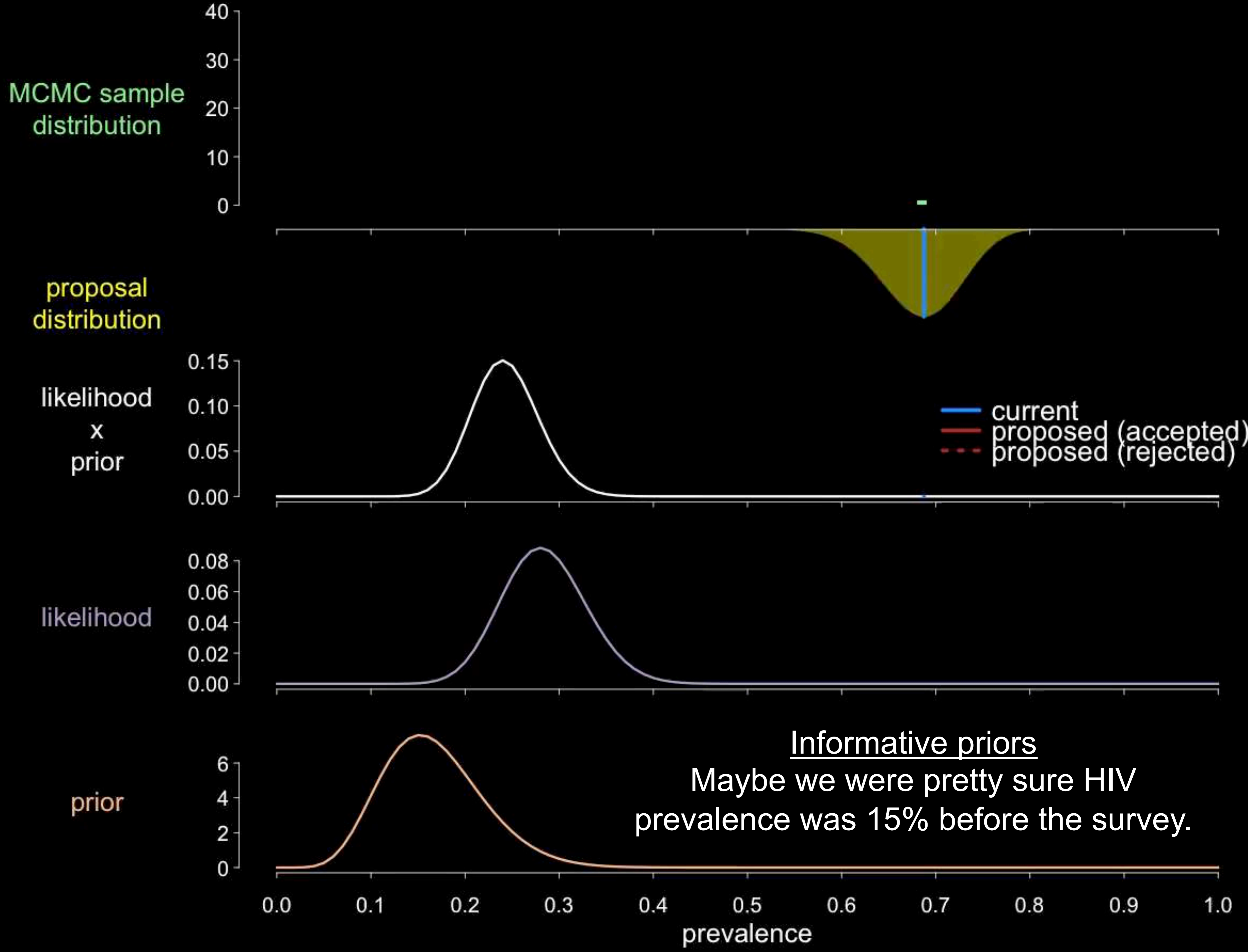




Early part of chain influenced by initial conditions.
MCMC sample distribution
250
200
150
100
50
0
proposal distribution
likelihood x prior
0.08
0.06
0.04
0.02
0.00
current
proposed (accepted)
proposed (rejected)
likelihood
0.08
0.06
0.04
0.02
0.00
prior
1.4
1.2
1.0
0.8
0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
prevalence

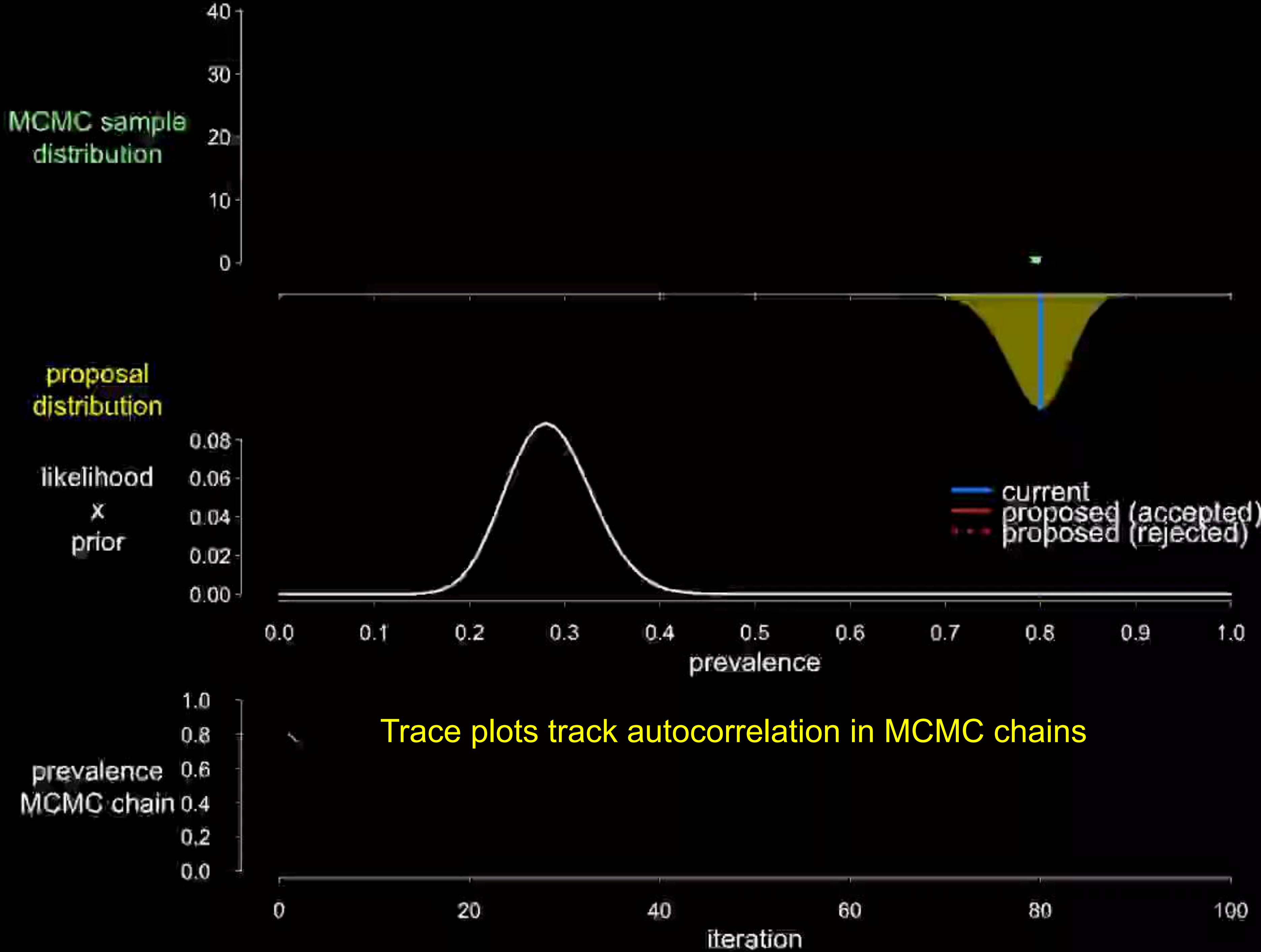
Remove first part of chain:
“Burn-in”

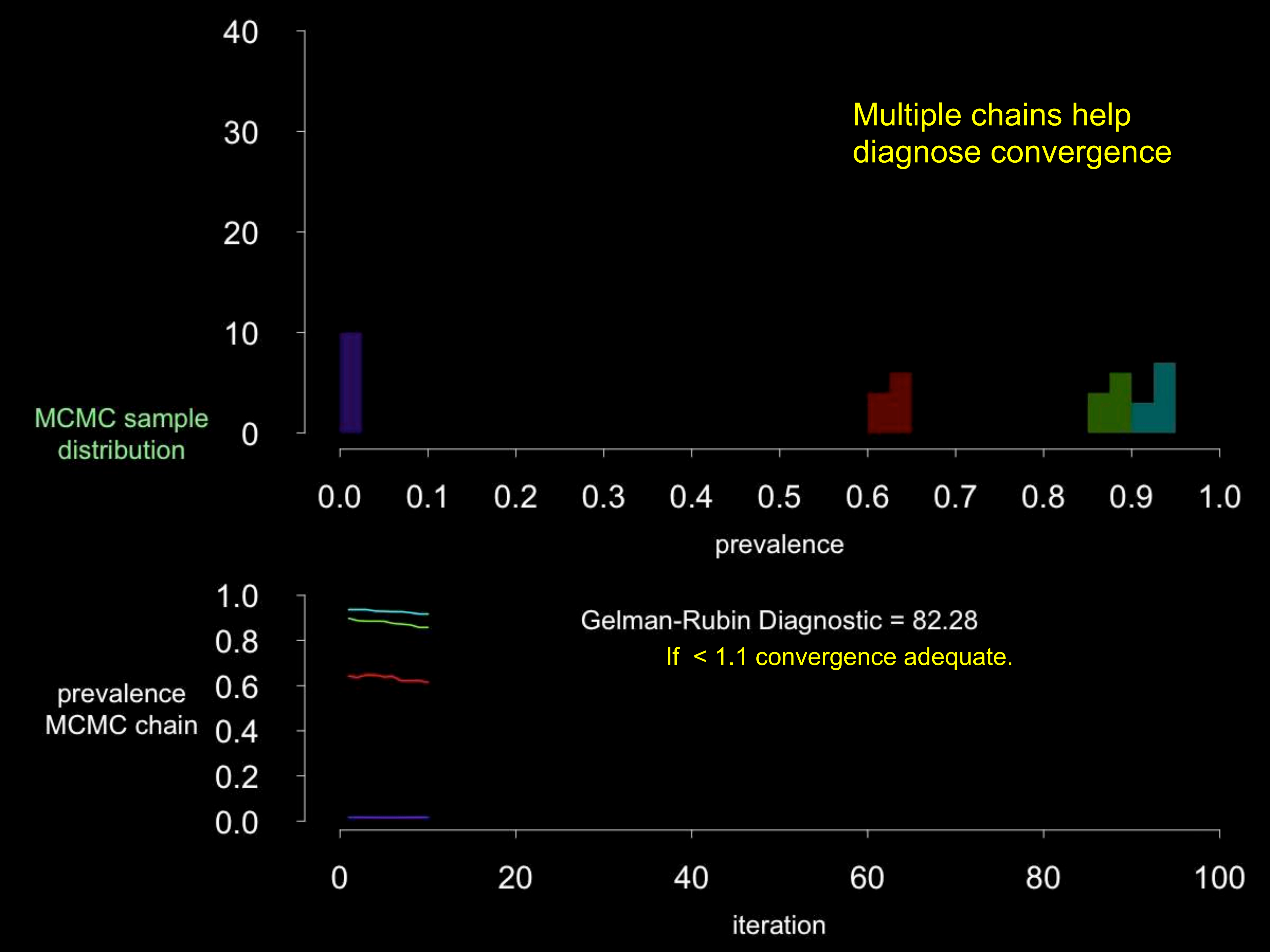




<u>Informative priors</u>

Maybe we were pretty sure HIV prevalence was 15% before the survey.


MCMC sample distribution
proposal distribution
likelihood x prior
current
proposed (accepted)
proposed (rejected)
prevalence
prevalence MCMC chain
iteration
Trace plots track autocorrelation in MCMC chains



Multiple chains help
diagnose convergence
MCMC sample
distribution
40
30
20
10
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
prevalence
prevalence
MCMC chain
1.0
0.8
0.6
0.4
0.2
0.0
Gelman-Rubin Diagnostic = 82.28
If < 1.1 convergence adequate.
0
20
40
60
80
100
iteration

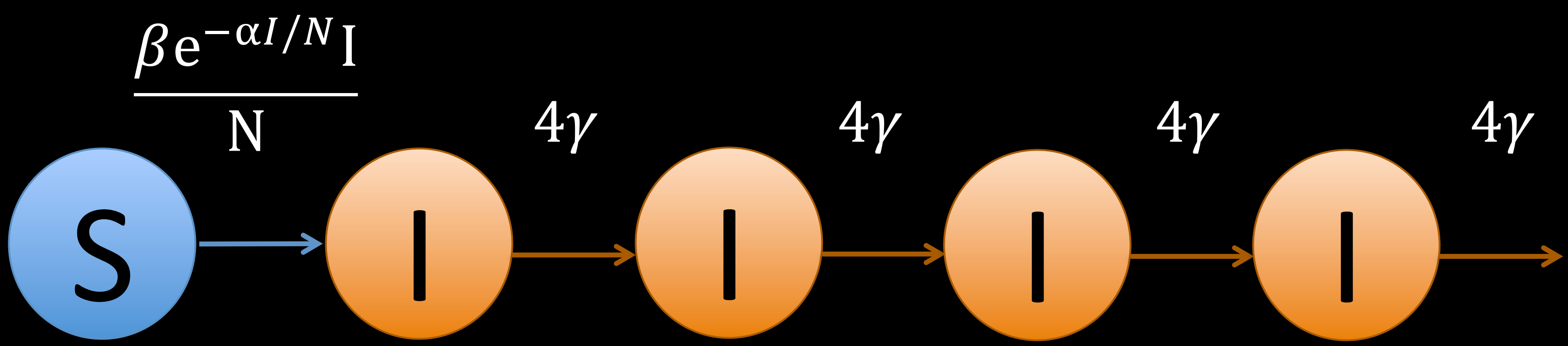
Defining MCMC

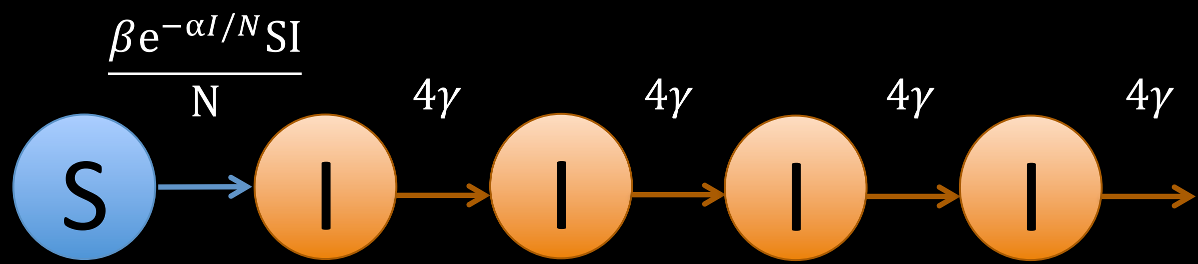
- Markov Chain = each state is a function of the last state
- Monte Carlo = stochastic Markov Chains
- MCMC = Markov Chains that eventually converge to a desired probability distribution.
- Samples are correlated!
- Convergence guaranteed, but only in long run.
- Convergence diagnostics are important

Multivariate MCMC

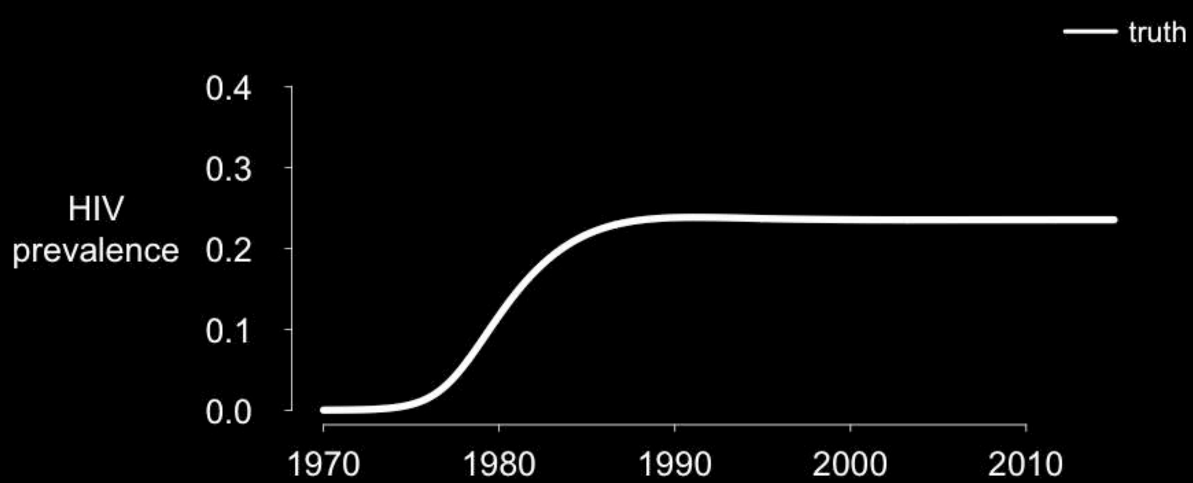
- Proposing parameters in N-dimensional space.
- Propose parameters one at a time (sequential sampling)

2-dimensional example

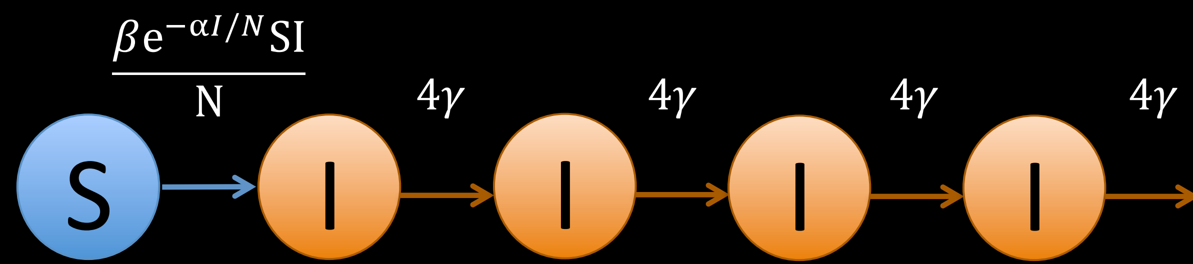


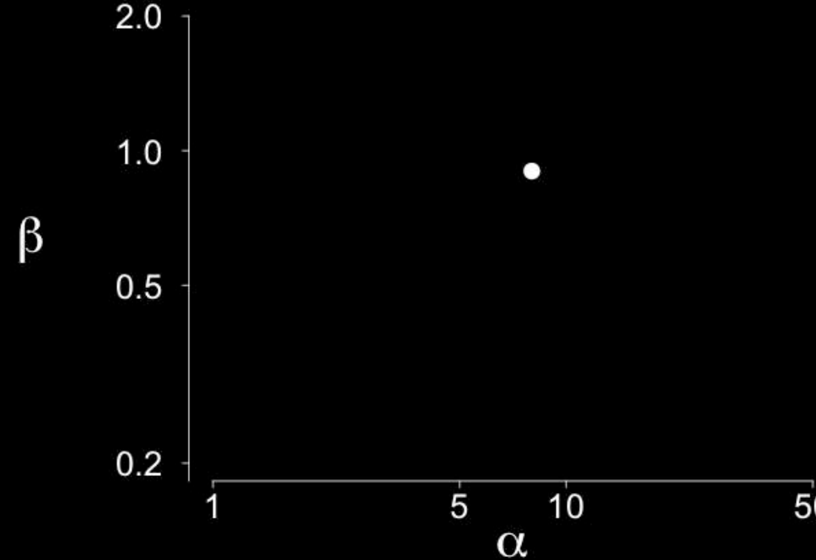


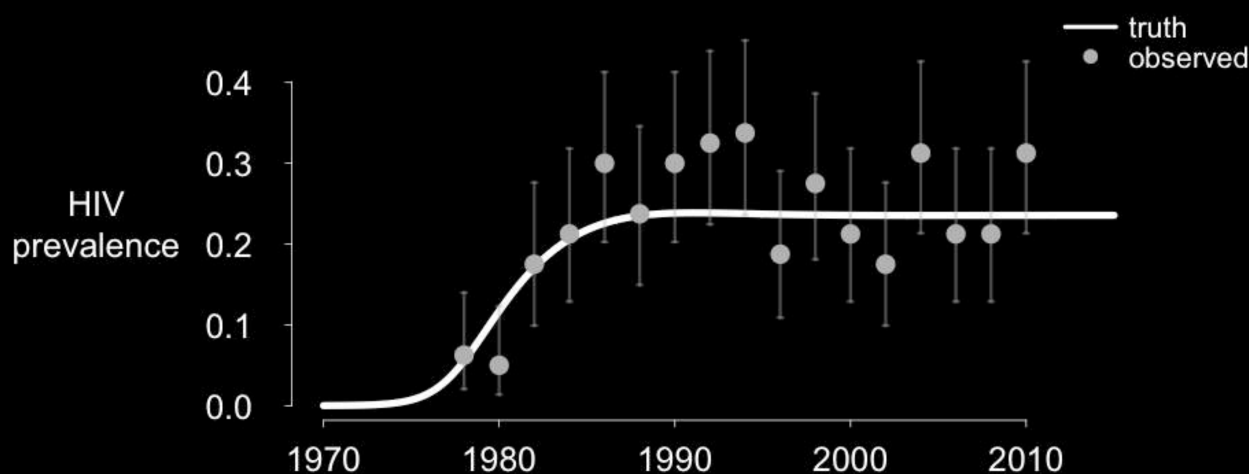
$\frac{\beta e^{-\alpha I/N} SI}{N}$
S
I
I
I
I
4γ
4γ
4γ
4γ



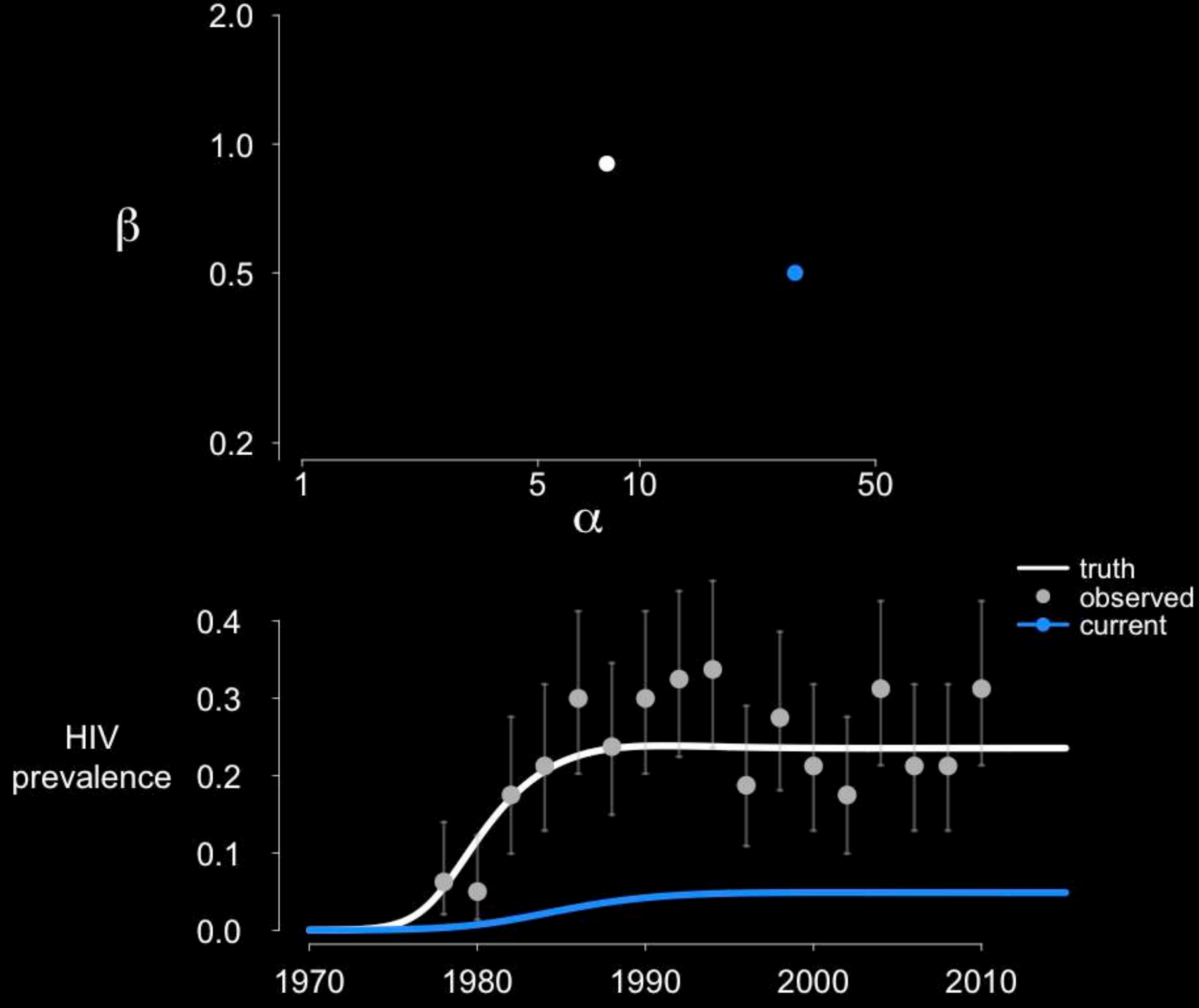
truth
HIV prevalence
0.4
0.3
0.2
0.1
0.0
1970
1980
1990
2000
2010


$\frac{\beta e^{-\alpha I/N} SI}{N}$
S
I
I
I
I
4γ
4γ
4γ
4γ

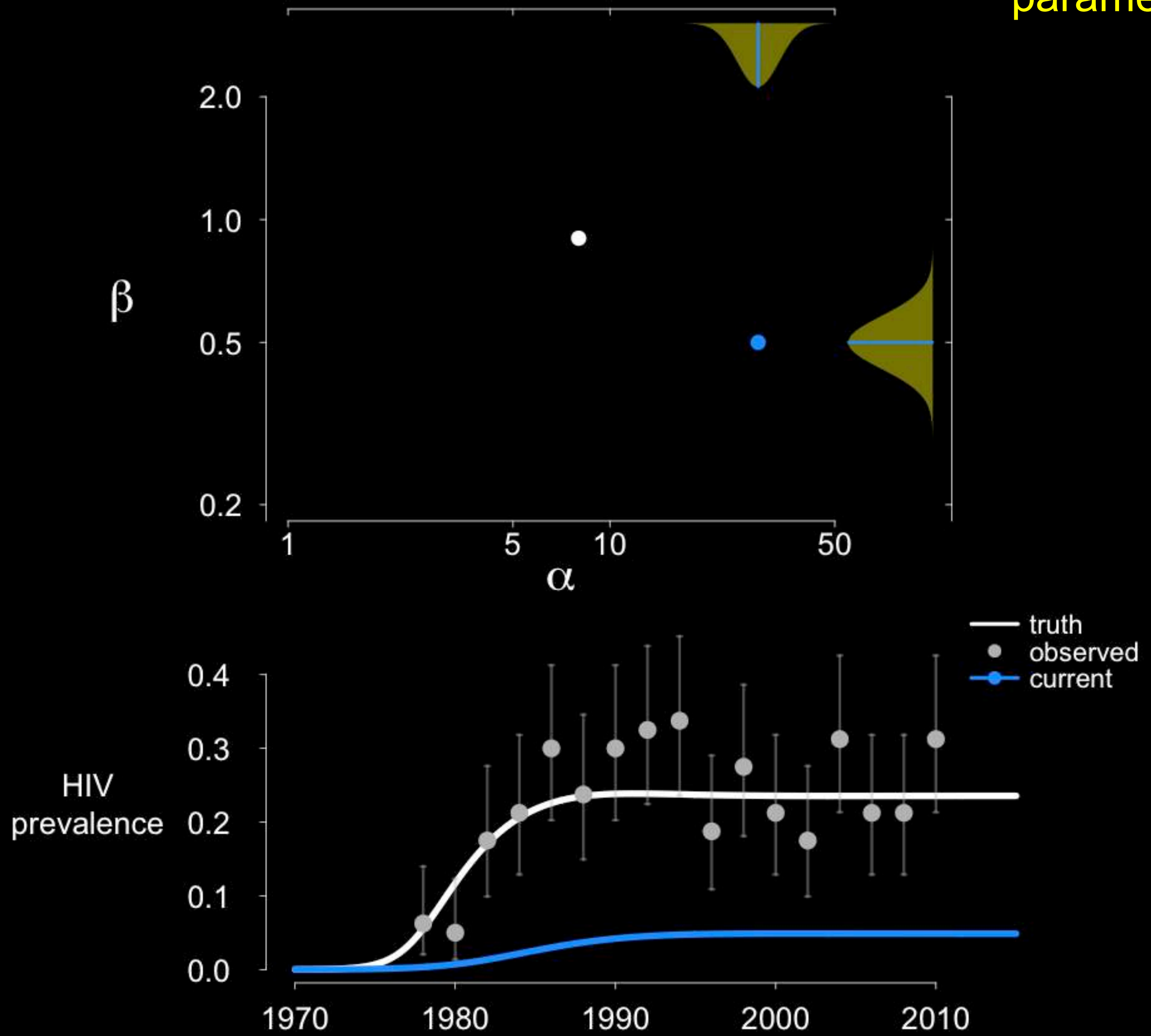

β
2.0
1.0
0.5
0.2
1
5
10
50
α


truth
observed
HIV prevalence
0.4
0.3
0.2
0.1
0.0
1970
1980
1990
2000
2010

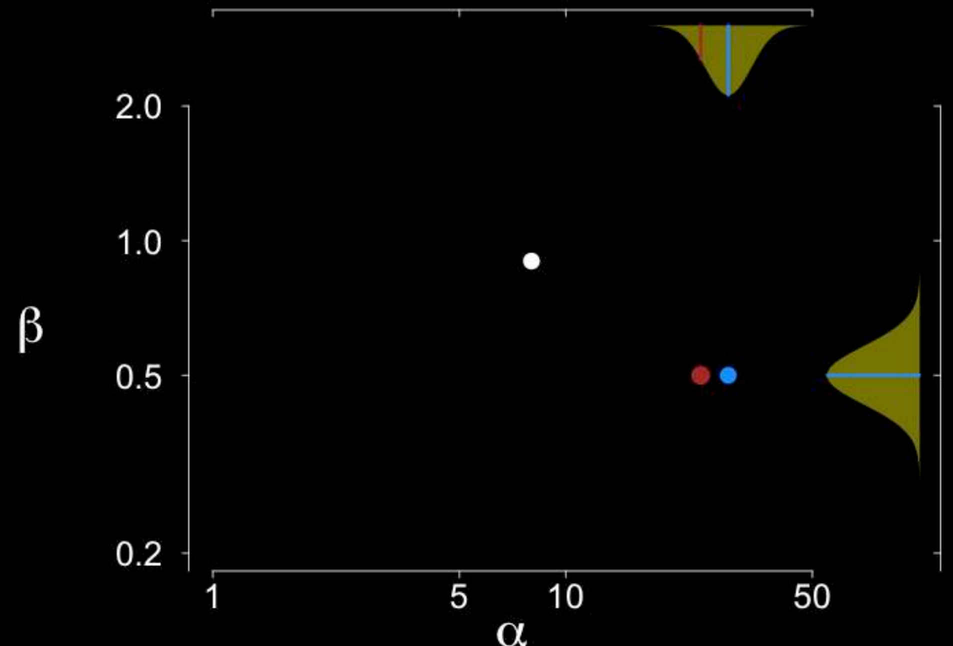
Start with an initial guess for both parameters.



Choose proposal distributions for both parameters



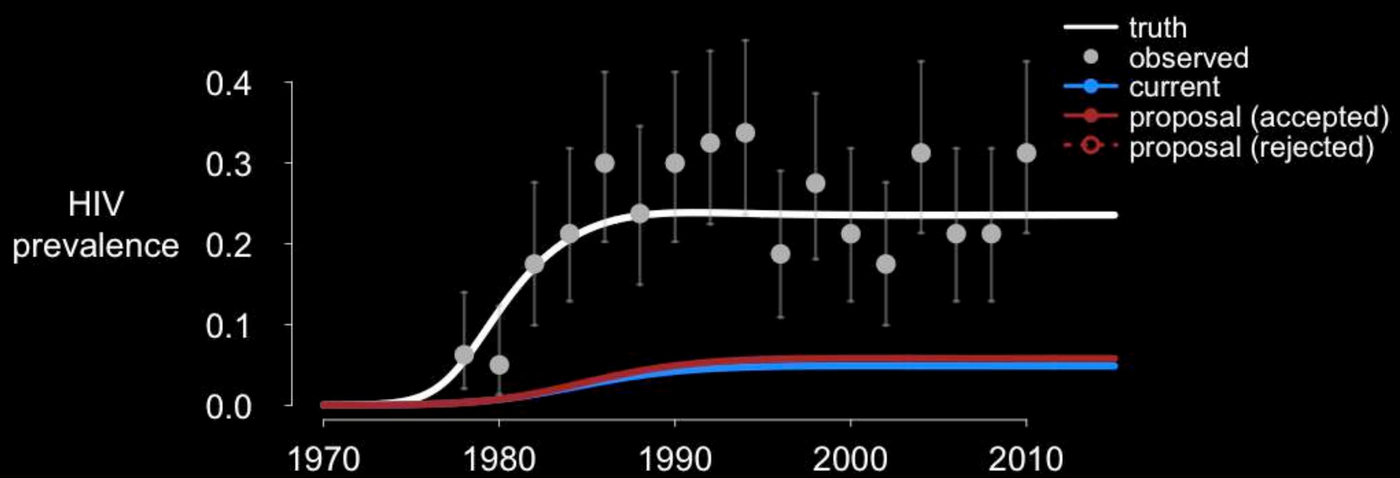
$$\alpha_t = \frac{P(y \mid \theta_{\text{proposal}})\, P(\theta_{\text{proposal}})}{P(y \mid \theta_t)\, P(\theta_t)}$$



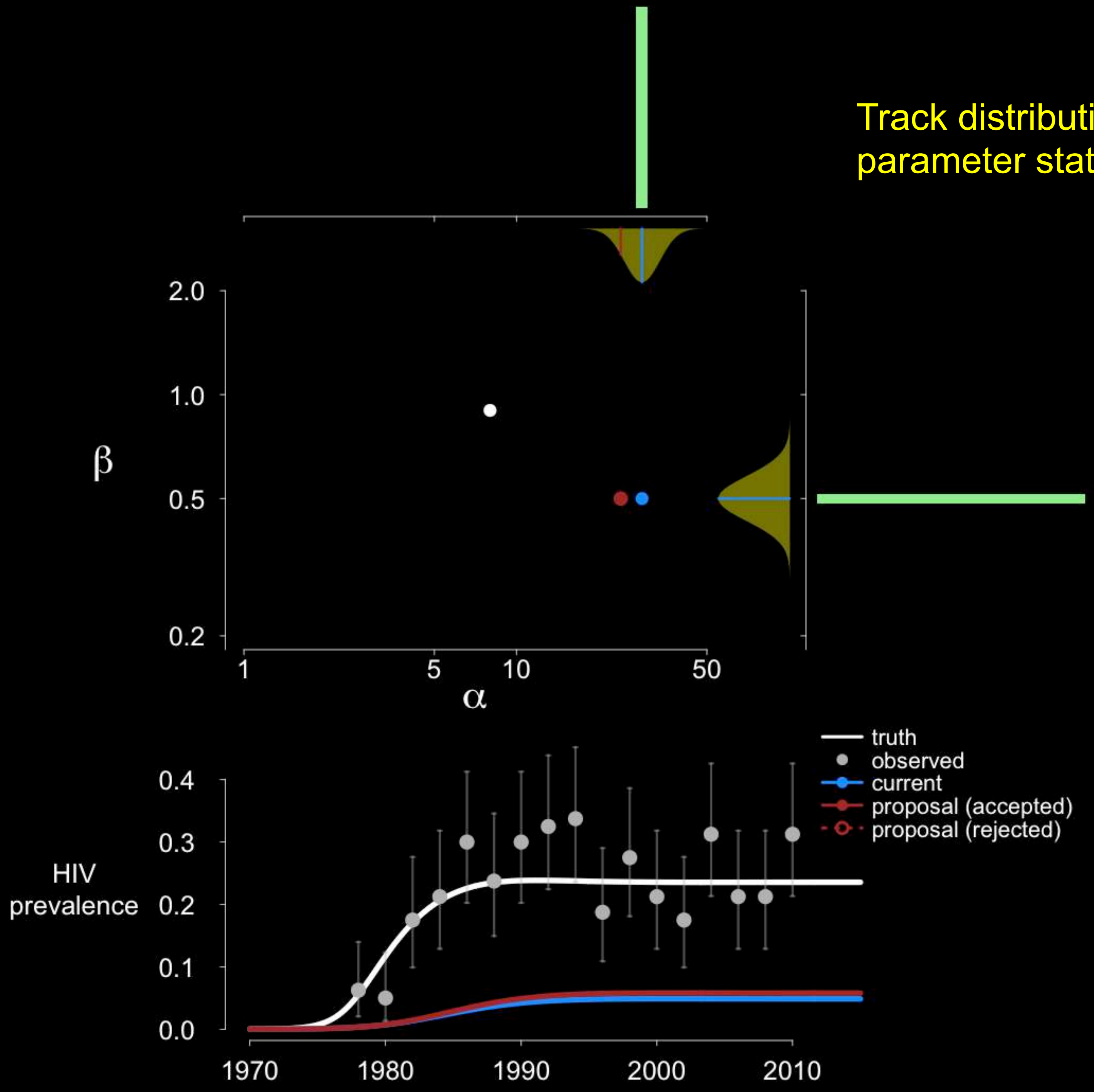
Propose new value for parameter at a time.

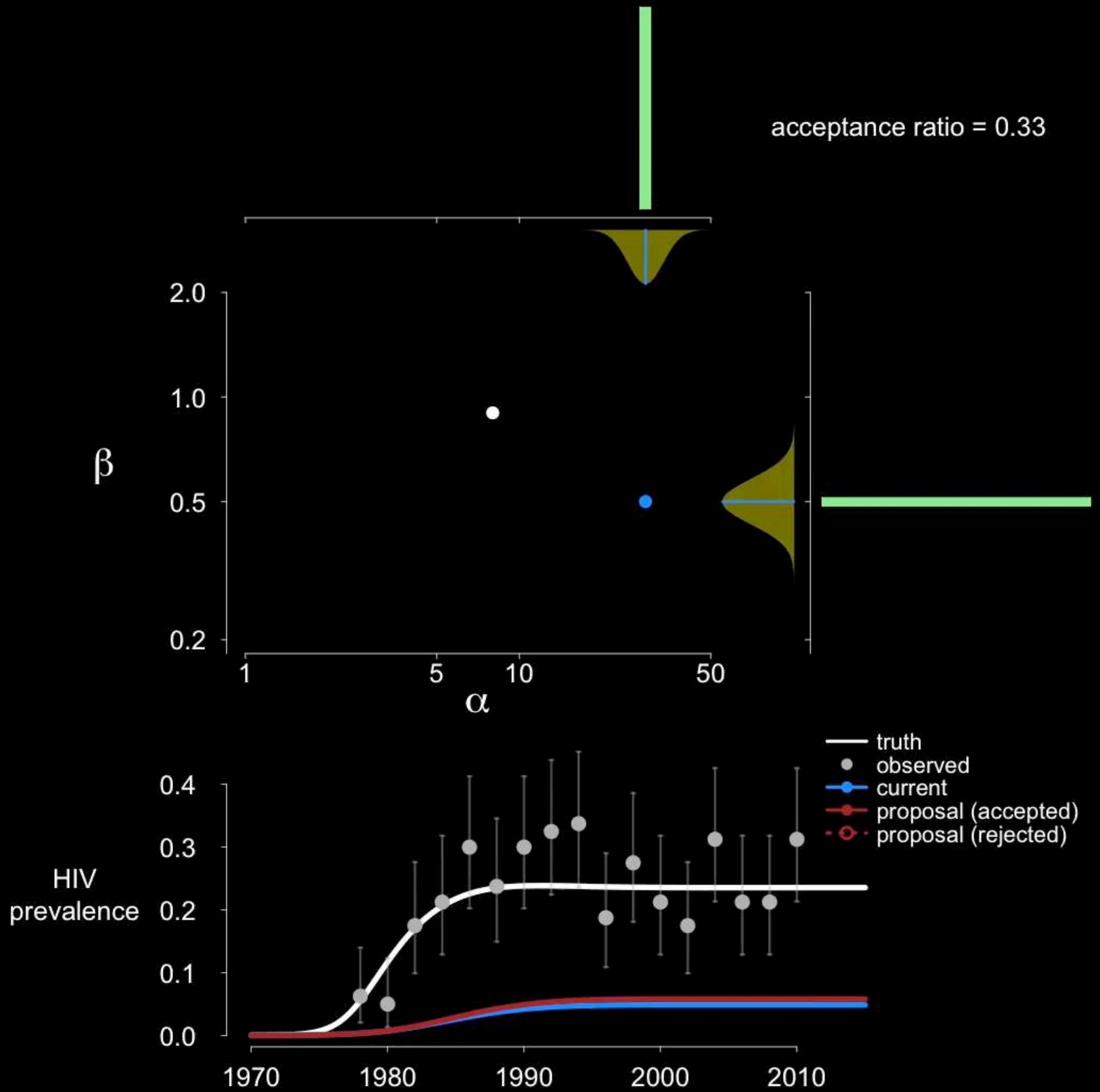
Accept new proposal with probability $\min(\alpha_t, 1)$.

Otherwise stay at same state.

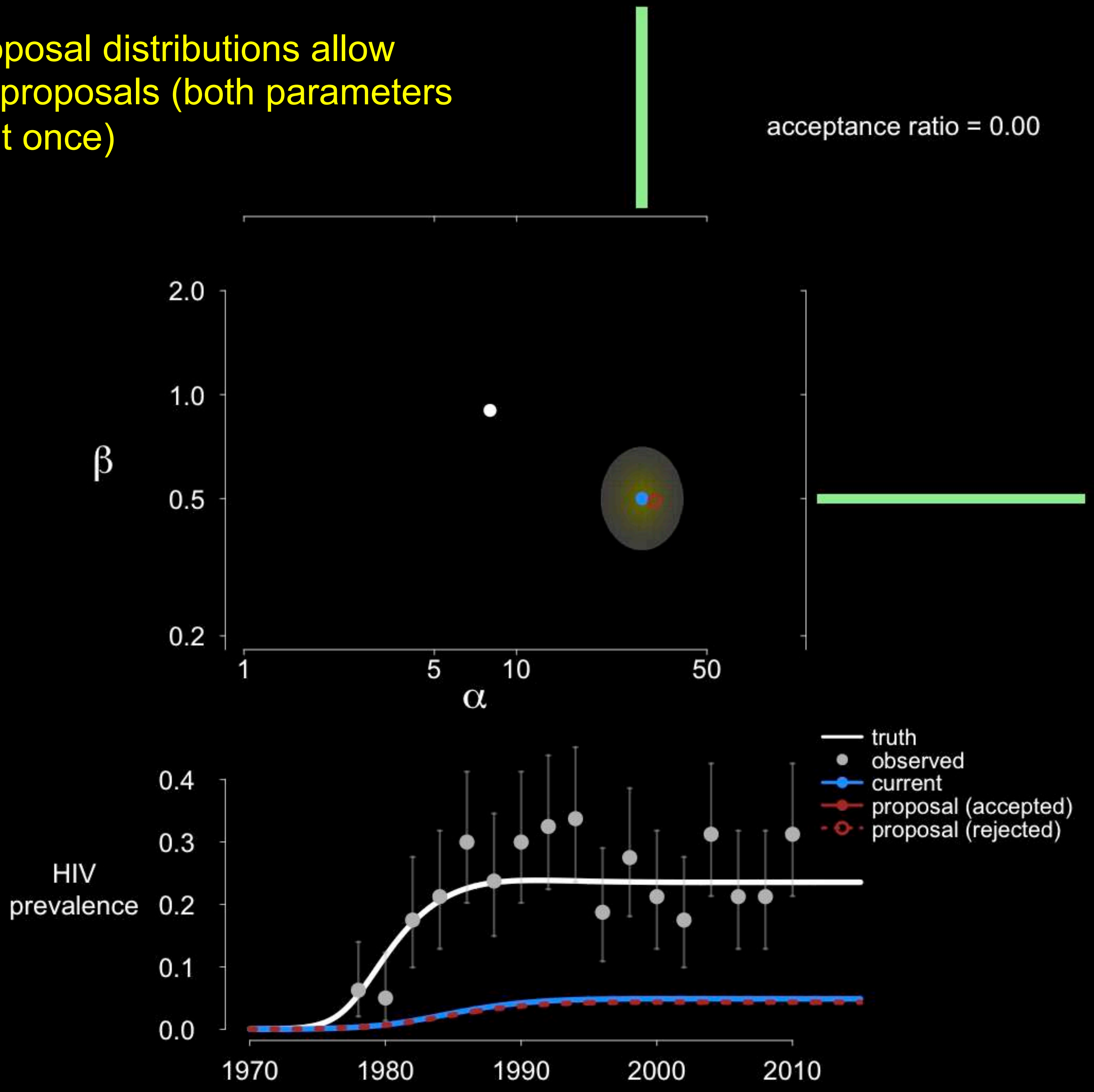


Track distribution of parameter states.

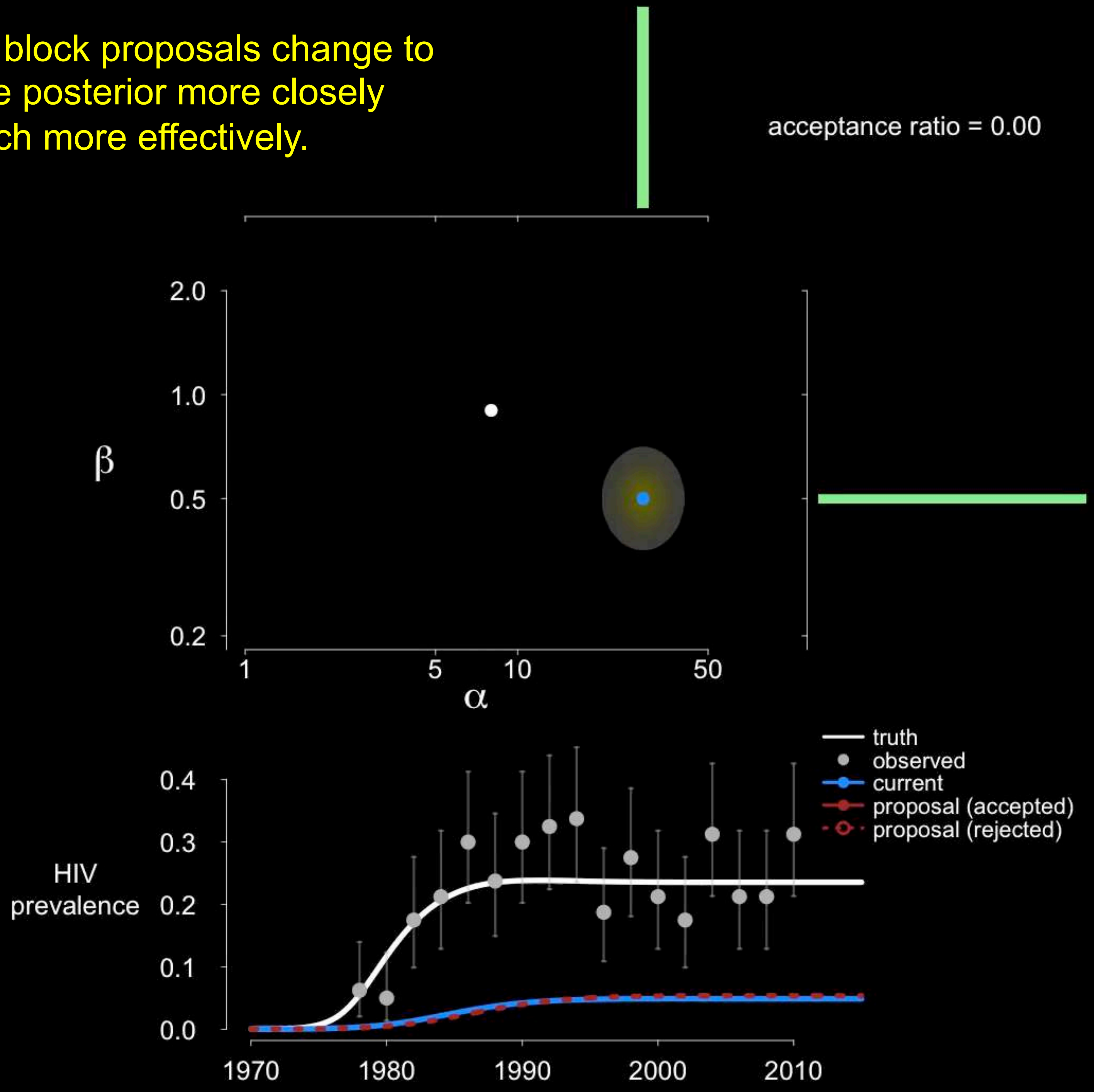



acceptance ratio = 0.33
2.0
1.0
β
0.5
0.2
1
5
10
50
α
truth
observed
current
proposal (accepted)
proposal (rejected)
0.4
0.3
HIV prevalence
0.2
0.1
0.0
1970
1980
1990
2000
2010

Block proposal distributions allow diagonal proposals (both parameters change at once)



Adaptive block proposals change to match the posterior more closely and search more effectively.



Multivariate MCMC

- Gelman-Rubin diagnostic assesses univariate & multivariate convergence
- Assess trace plots of all parameters
- Block sampling of collinear parameters increases efficiency
- Finding a "good" first guess more challenging for greater # parameters fit

Acceptance Ratio

- Ideal rate is 50% for 1-dimensional fitting
- Approaches 23% for N-dimensional fitting

MCMC Algorithms

- Metropolis-Hastings
- Gibbs Sampler
- Hamiltonian MCMC
- No U-turn Sampler

- Block Sampling
- Adaptive MCMC

More info at

http://www.bayesian-inference.com/mcmc

Moving Past MCMC

- Approximate Bayesian Computation

  Match data characteristics, rather than explicit likelihoods

- Particle Filters

  Fit process noise and observation noise

- Particle MCMC

  Combine both approaches