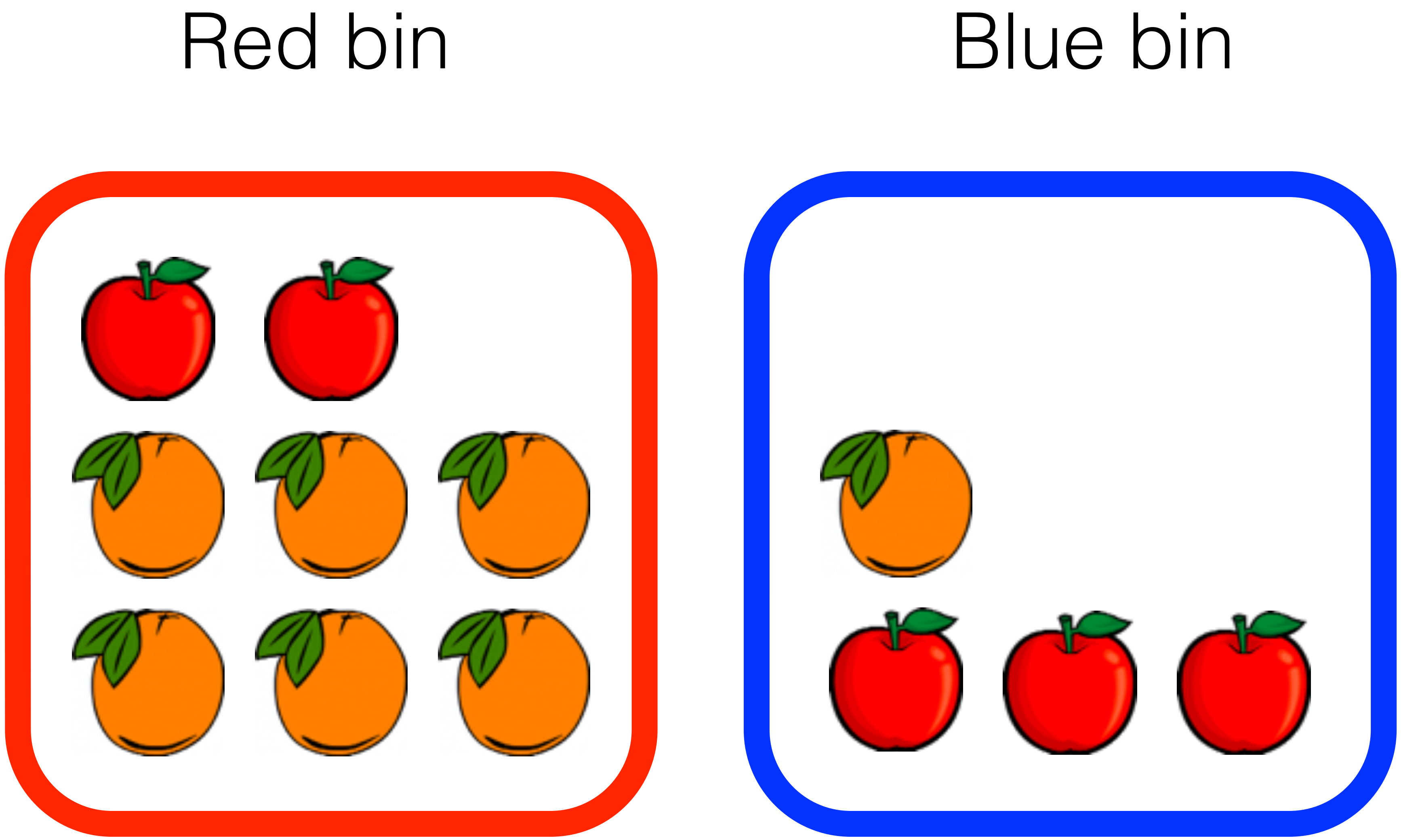
Introduction to Bayesian Inference

Brooks Paige

Goals of this lecture

- Understand joint, marginal, and conditional probability distributions
- Understand expectations of functions of a random variable
- Understand how Monte Carlo methods allow us to approximate expectations
- Goal for the subsequent exercise: understand how to implement basic Monte Carlo inference methods

Simple example: discrete probability

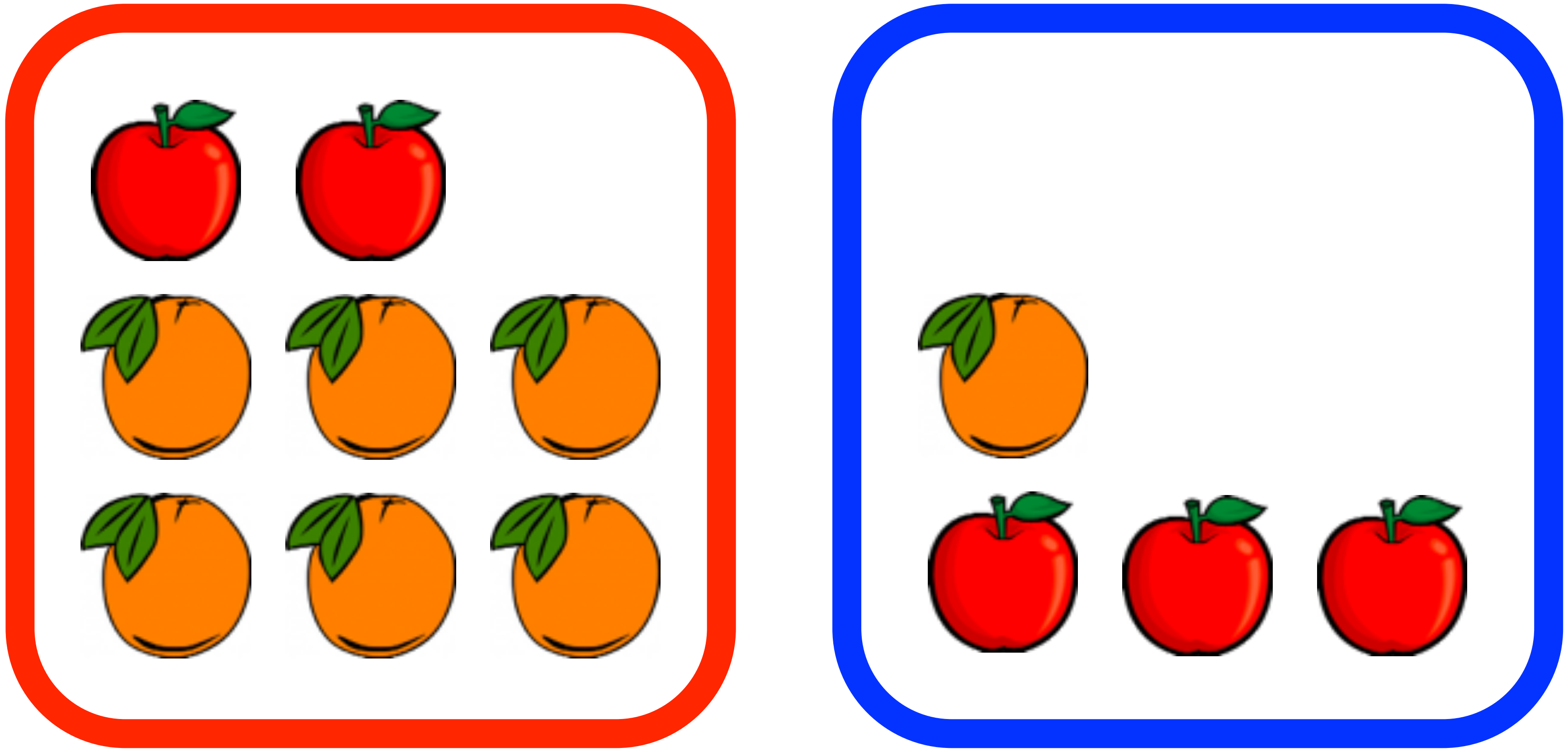


Simple example: discrete probability

"First I pick a bin, then I pick a single fruit from the bin"

p(red bin) = 2/5 p(blue bin) = 3/5

p(apple|red) = 2/8 p(apple|blue) = 3/4



Simple example: discrete probability

"First I pick a bin, then I pick a single fruit from the bin"

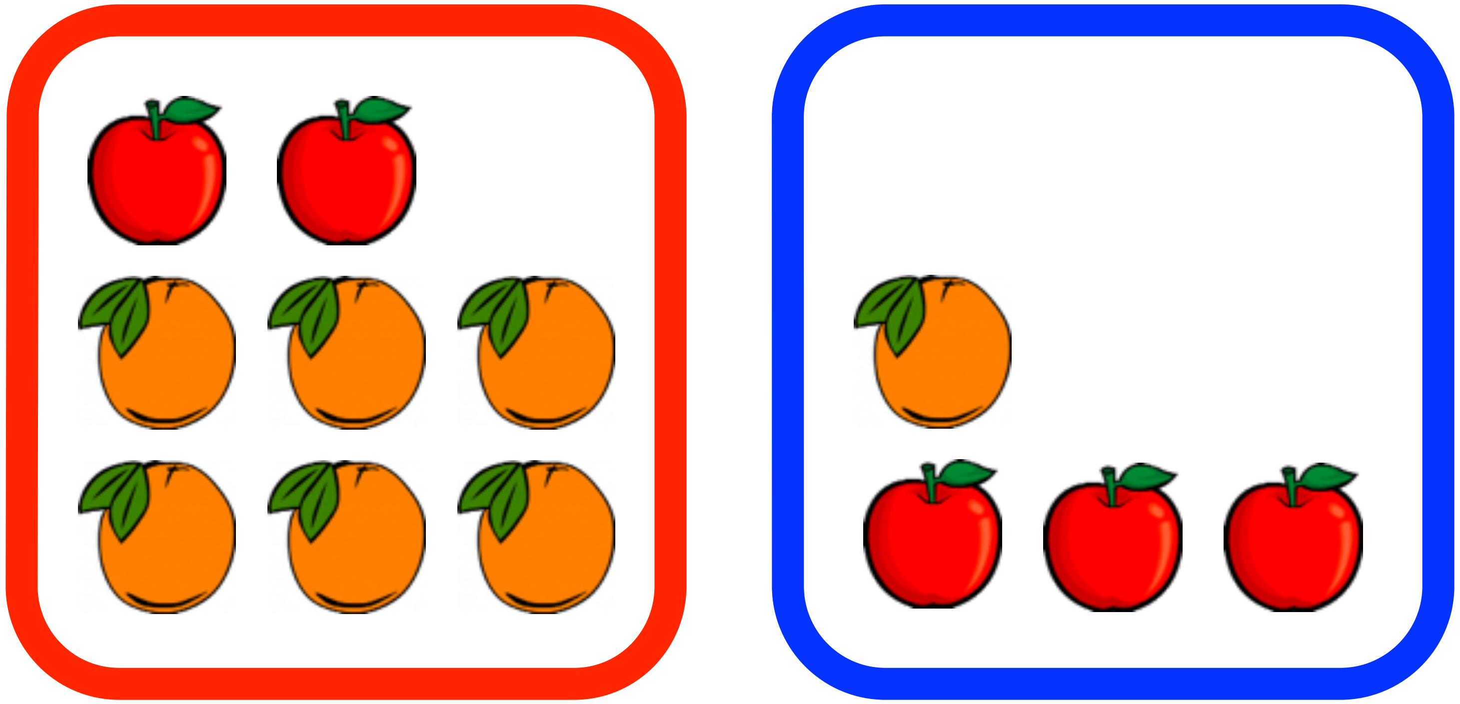
Easy question: what is the probability I pick the red bin?

$p(\text{red bin}) = 2/5$

$p(\text{apple}|\text{red}) = 2/8$

$p(\text{blue bin}) = 3/5$

$p(\text{apple}|\text{blue}) = 3/4$



Simple example: discrete probability

"First I pick a bin, then I pick a single fruit from the bin"

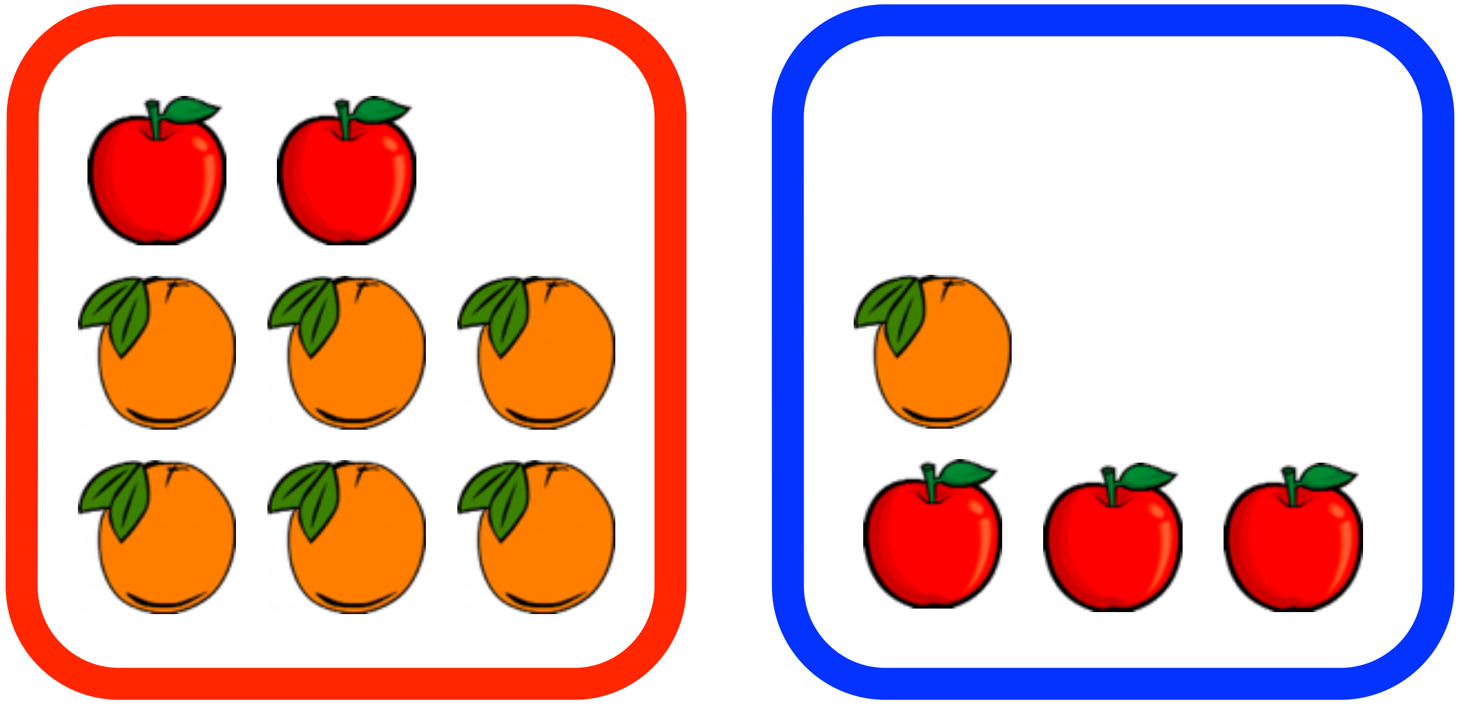
Easy question: If I first pick the red bin, what is the probability I pick an orange?

$p(\text{red bin}) = 2/5$

$p(\text{apple}|\text{red}) = 2/8$

$p(\text{blue bin}) = 3/5$

$p(\text{apple}|\text{blue}) = 3/4$



Simple example: discrete probability

"First I pick a bin, then I pick a single fruit from the bin"

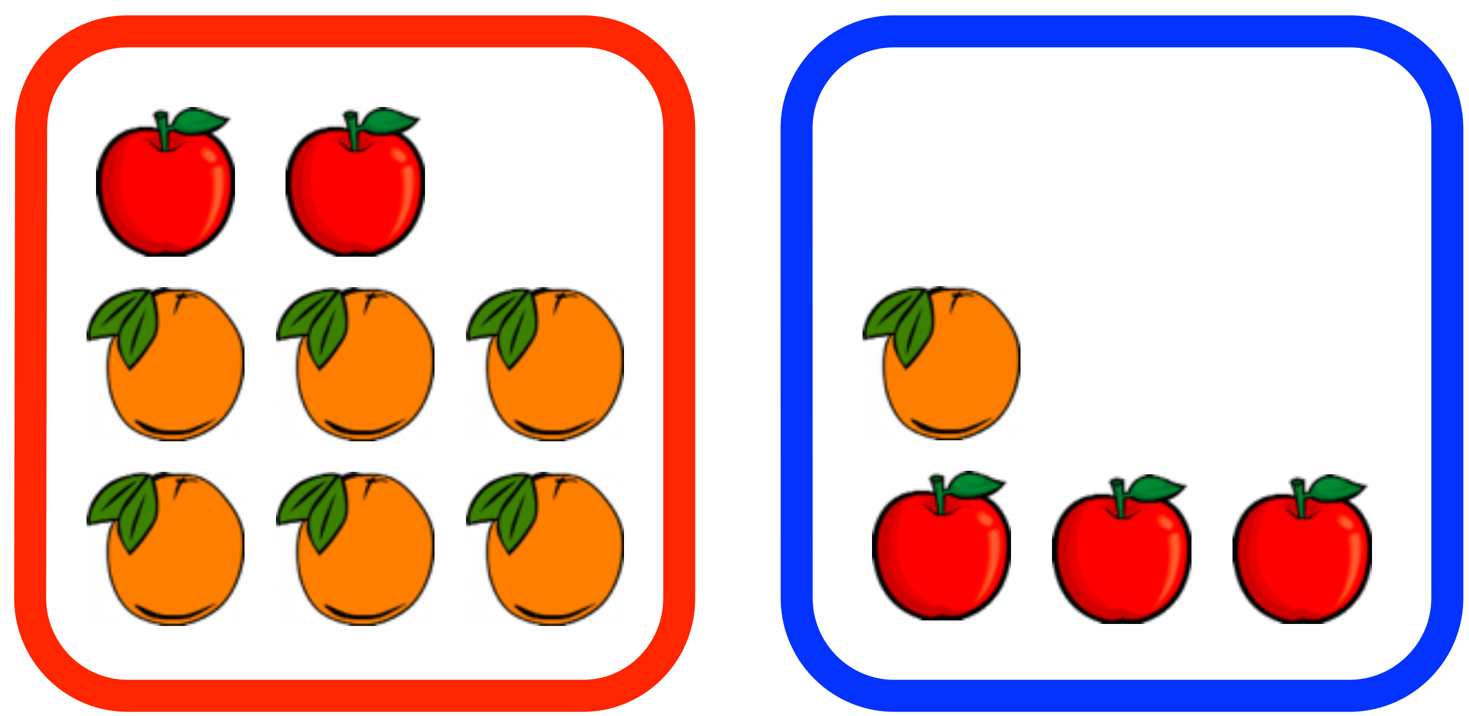
Less easy question: What is the overall probability of picking an apple?

p(red bin) = 2/5

p(apple|red) = 2/8

p(blue bin) = 3/5

p(apple|blue) = 3/4



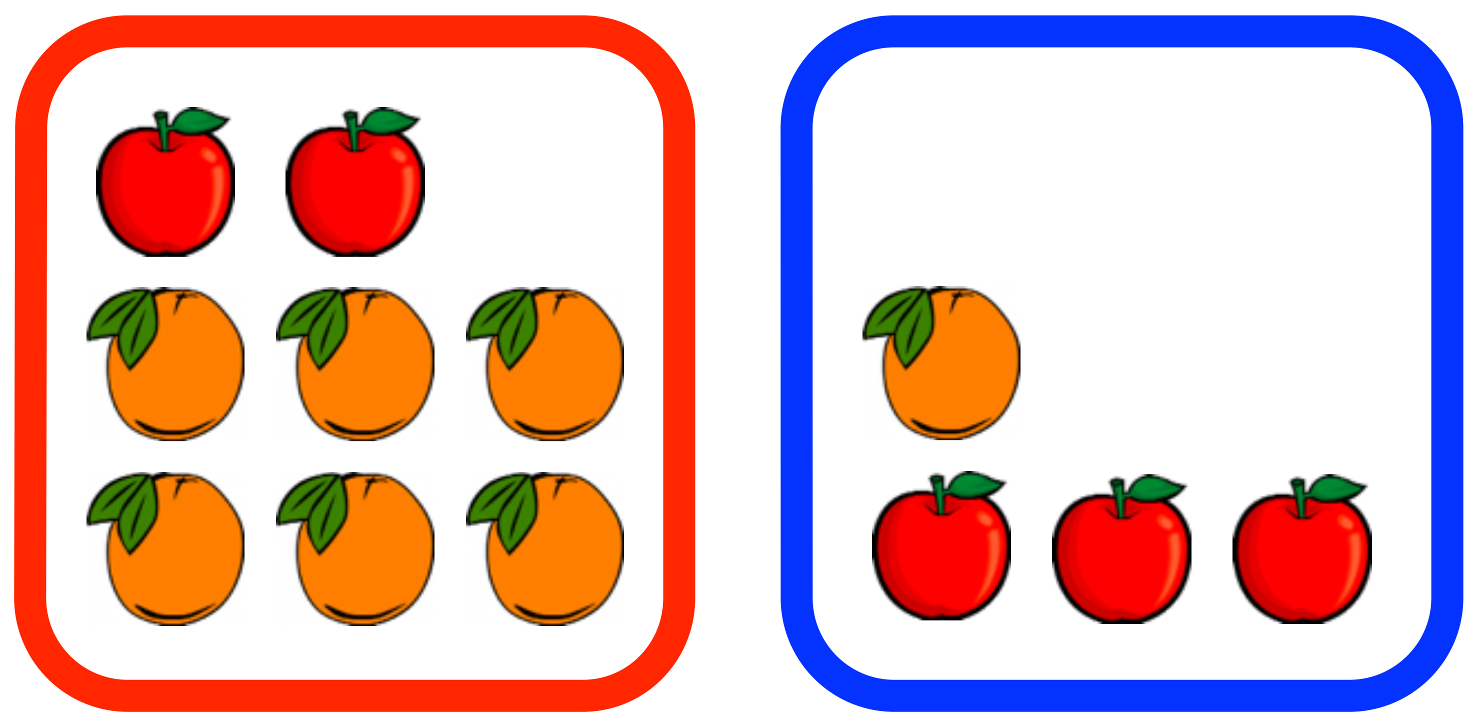
Simple example: discrete probability

"First I pick a bin, then I pick a single fruit from the bin"

Hard question: If I pick an orange, what is the probability that I picked the blue bin?

p(red bin) = 2/5
p(apple|red) = 2/8

p(blue bin) = 3/5
p(apple|blue) = 3/4



What is inference?

- The “hard question” requires reasoning backwards in our generative model

- Our generative model specifies these probabilities explicitly:
  - A “marginal” probability *p(bin)*
  - A “conditional” probability $p(fruit \mid bin)$
  - A “joint” probability *p(fruit, bin)*

- How can we answer questions about different conditional or marginal probabilities?
  - *p(fruit)*: “what is the overall probability of picking an orange?”
  - $p(bin|fruit)$: “what is the probability I picked the blue bin, given I picked an orange?”

Rules of probability

We just need two basic rules of probability.

- **Sum rule:**

$$p(y) = \sum_x p(y, x) \qquad p(x) = \sum_y p(y, x)$$

- **Product rule:**

$$p(y, x) = p(y \mid x) p(x) = p(x \mid y) p(y)$$

- These rules define the relationship between *marginal*, *joint*, and *conditional* distributions.

Bayes' Rule

Bayes' rule relates two conditional probabilities:

$$\underbrace{p(x \mid y)}_{\text{Posterior}} = \underbrace{p(y \mid x)}_{\text{Likelihood}}\,\underbrace{p(x)}_{\text{Prior}}/p(y)$$

Mini–exercise

$$\sum_{x} p(x|y) = \ ???$$

Use the sum and product rules!

Simple example: discrete probability

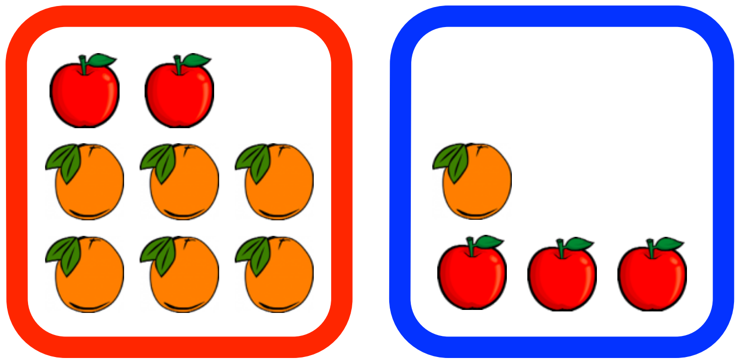
"First I pick a bin, then I pick a single fruit from the bin"

USE THE SUM RULE: What is the overall probability of picking an apple?

$$p(\text{apple}) = p(\text{apple}|\text{red})p(\text{red}) + p(\text{apple}|\text{blue})p(\text{blue})$$

$$= \quad 2/8 \quad \text{x} \quad 2/5 \quad + \quad 3/4 \quad \text{x} \quad 3/5$$

$$= 0.55$$



Simple example: discrete probability

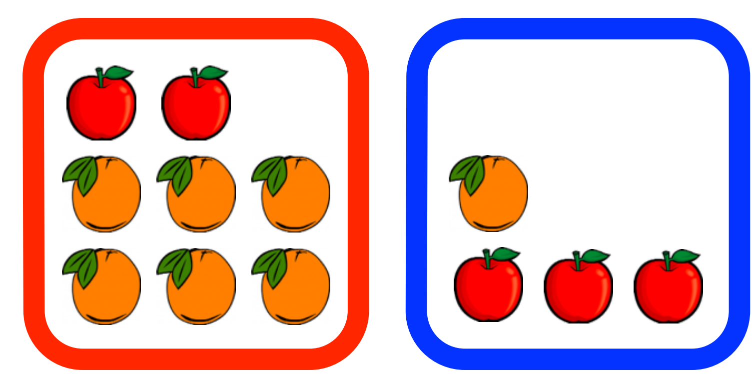
"First I pick a bin, then I pick a single fruit from the bin"

USE BAYES' RULE: If I pick an orange, what is the probability that I picked the blue bin?

$$p(\text{blue}|\text{orange}) = \frac{p(\text{orange}|\text{blue})p(\text{blue})}{p(\text{orange})}$$

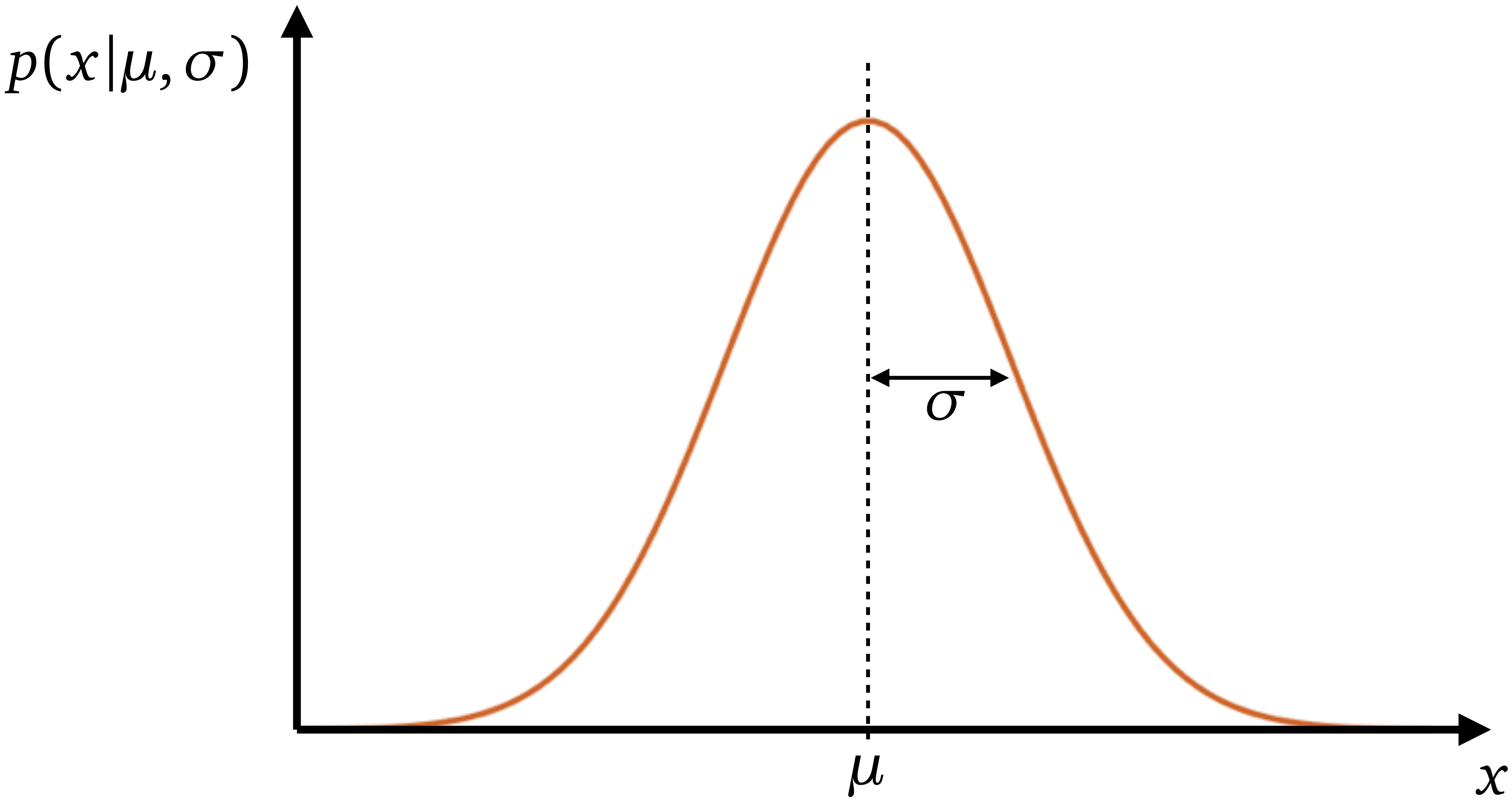
$$= \frac{1/4 \quad \times \quad 3/5}{6/8 \times 2/5 \ + \ 1/4 \times 3/5}$$

$$= 1/3$$



Continuous probability

The normal distribution



$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$

A simple continuous example

- Measure the temperature of some water using an inexact thermometer

- The actual water temperature *x* is somewhere near room temperature of 22°; we record an estimate *y*.

$$x \sim \mathrm{Normal}(22, 10)$$
$$y|x \sim \mathrm{Normal}(x, 1)$$

Easy question: what is $p(y \mid x = 25)$?

Hard question: what is $p(x \mid y = 25)$?

Rules of probability: continuous

- For real-valued *x*, the sum rule becomes an *integral*:

$$p(y) = \int p(y, x)\mathrm{d}x$$

- Bayes' rule:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y, x)\mathrm{d}x}$$

Integration is harder than addition!

Bayes' rule:

$$p(x|y=25) = \frac{p(x)p(y=25|x)}{p(y=25)}$$

Sum rule, in the denominator:

$$p(y=25) = \int p(x)p(y=25|x)\mathrm{d}x$$

In general this integral is intractable, and we can only evaluate up to a normalizing constant

Monte Carlo inference

General problem:

$$p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{y} \mid \mathbf{x})p(\mathbf{x})/p(\mathbf{y})$$

Posterior Likelihood Prior

- Our *data* is given by *y*
- Our generative model specifies the prior and likelihood
- We are interested in answering questions about the *posterior* distribution of *p(x | y)*

General problem:

$$\underbrace{p(x \mid y)}_{\text{Posterior}} = \underbrace{p(y \mid x)}_{\text{Likelihood}}\underbrace{p(x)}_{\text{Prior}}/p(y)$$

- Typically we are not trying to compute a probability density function for $p(x \mid y)$ as our end goal

- Instead, we want to compute *expected values* of some function $f(x)$ under the posterior distribution

Expectation

- Discrete and continuous:

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x.$$

- Conditional on another random variable:

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Key Monte Carlo identity

- We can approximate expectations using *samples* drawn from a distribution *p.* If we want to compute

$$\mathbb{E}[f] = \int p(x) f(x)\,\mathrm{d}x.$$

we can approximate it with a finite set of points sampled from *p(x)* using

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

which becomes exact as *N* approaches infinity.

How do we draw samples?

- Simple, well-known distributions: samplers exist (for the moment take as given)
- We will look at:

1. Build samplers for complicated distributions out of samplers for simple distributions compositionally
2. Rejection sampling
3. Likelihood weighting
4. Markov chain Monte Carlo

Ancestral sampling from a model

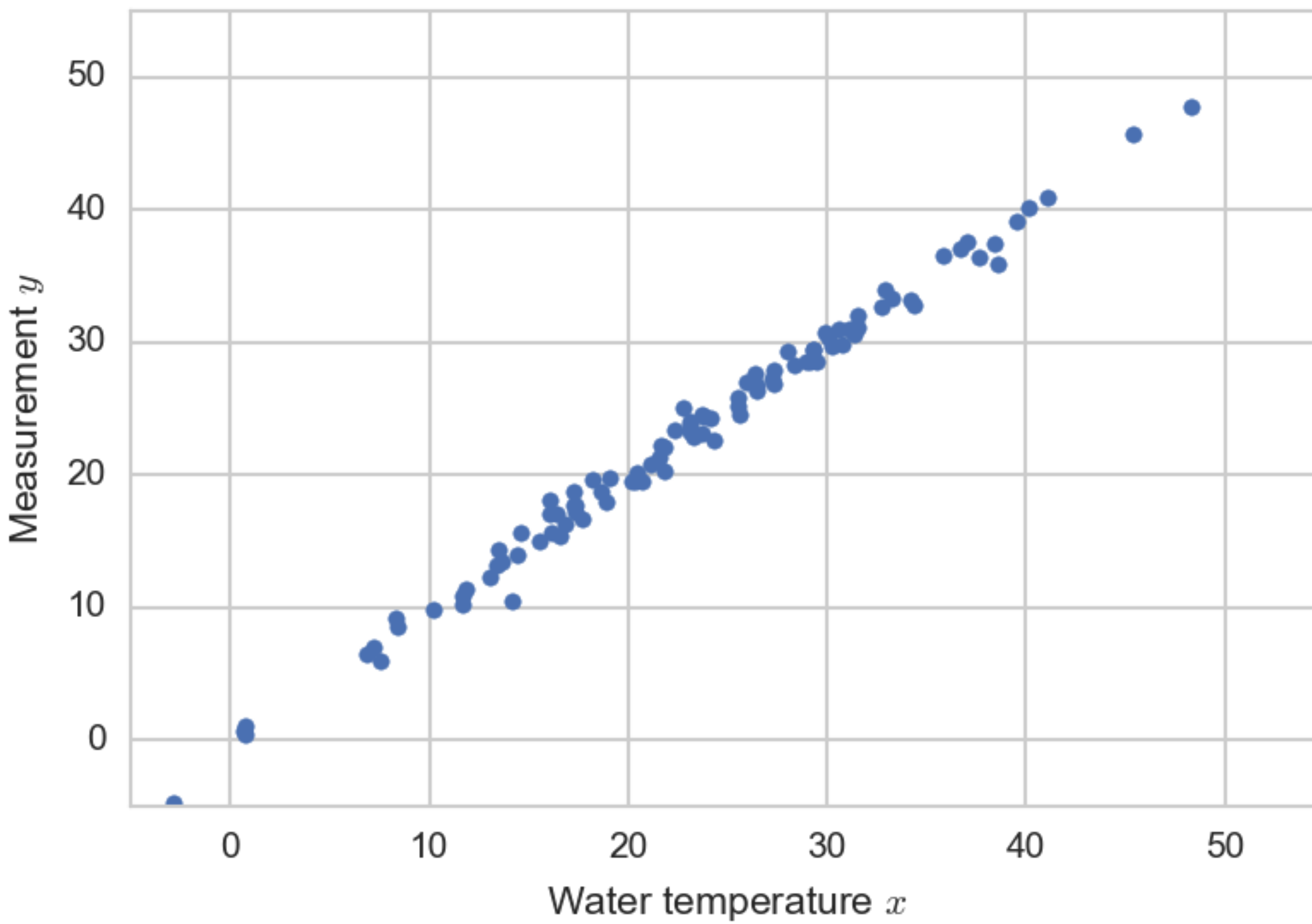
- In our example with estimating the water temperature, suppose we already know how to sample from a normal distribution.

$$x \sim \mathrm{Normal}(22, 10)$$
$$y|x \sim \mathrm{Normal}(x, 1)$$

  We can sample *y* by literally simulating from the generative process: we first sample a “true” temperature *x*, and then we sample the observed *y*.

- This draws a sample from the **joint** distribution *p(x, y)*.

Samples from the joint distribution



Conditioning via rejection

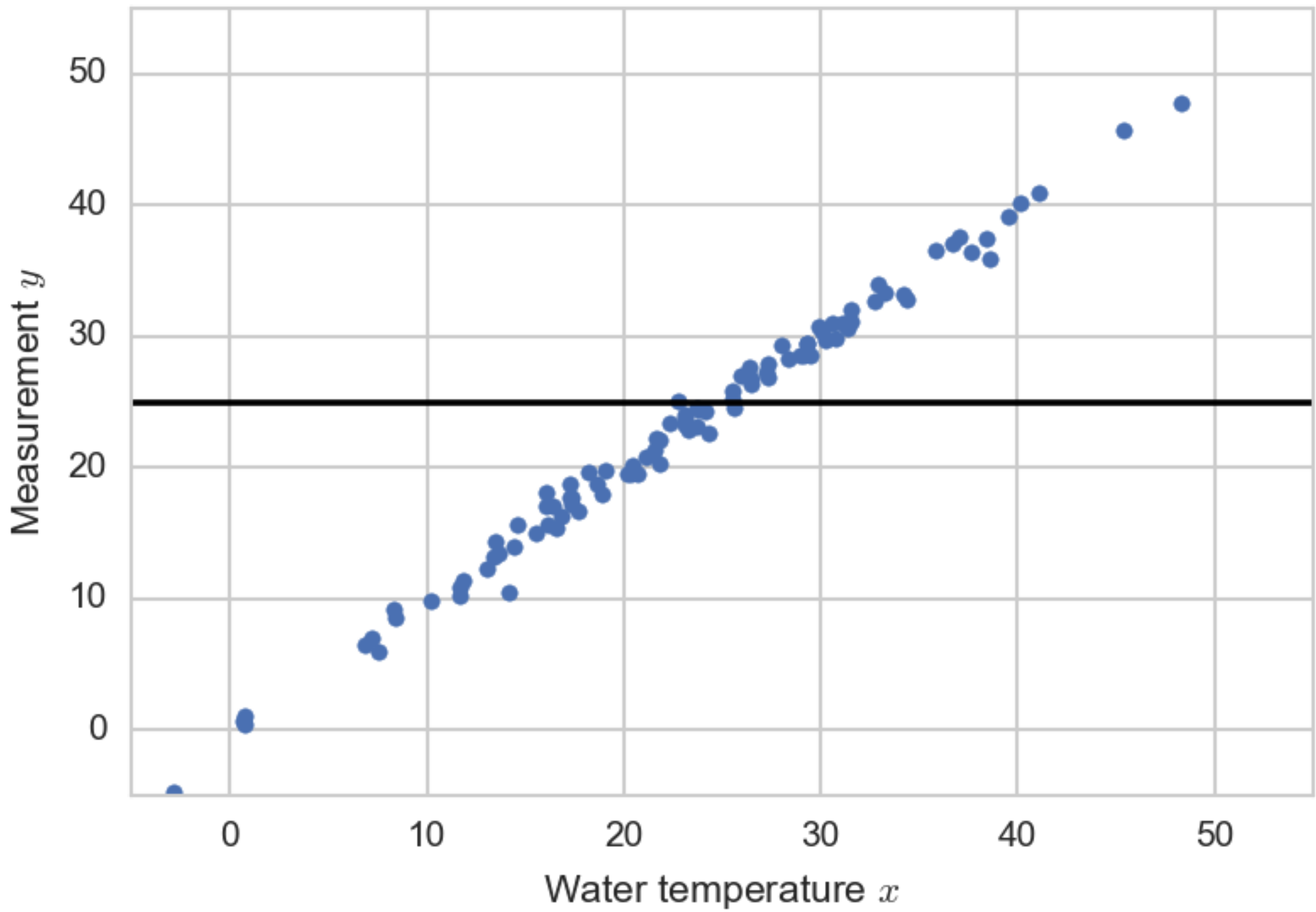
- What if we want to sample from a conditional distribution? The simplest form is via rejection.

- Use the ancestral sampling procedure to simulate from the generative process, draw a sample of x and a sample of y. These are drawn together from the joint distribution $p(x, y)$.

- To estimate the posterior $p(x \mid y = 25)$, we say that x is a sample from the posterior if its corresponding value $y = 25$.

- **Question:** is this a good idea?

Conditioning via rejection



Black bar shows measurement at *y = 25*.

How many of these samples from the joint have *y = 25*?

Conditioning via importance sampling

- One option is to sidestep sampling from the posterior $p(x \mid y = 3)$ entirely, and draw from some proposal distribution $q(x)$ instead.

- Instead of computing an expectation with respect to $p(x|y)$, we compute an expectation with respect to $q(x)$:

$$\begin{aligned}
\mathbb{E}_{p(x|y)}[f(x)] &= \int f(x)p(x|y)\mathrm{d}x \\
&= \int f(x)p(x|y)\frac{q(x)}{q(x)}\mathrm{d}x \\
&= \mathbb{E}_{q(x)}\left[f(x)\frac{p(x|y)}{q(x)}\right]
\end{aligned}$$

Conditioning via importance sampling

- Define an "importance weight" $W(x) = \dfrac{p(x|y)}{q(x)}$
- Then, with $x_i \sim q(x)$

$$\mathbb{E}_{p(x|y)}[f(x)] = \mathbb{E}_{q(x)}\left[f(x)W(x)\right] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)W(x_i)$$

- Expectations now computed using *weighted* samples from *q(x)*, instead of unweighted samples from *p(x|y)*

Conditioning via importance sampling

- Typically, can only evaluate *W(x)* up to a constant (but this is not a problem):

$$W(x_i) = \frac{p(x_i|y)}{q(x_i)} \qquad\qquad w(x_i) = \frac{p(x_i, y)}{q(x_i)}$$

- Approximation:

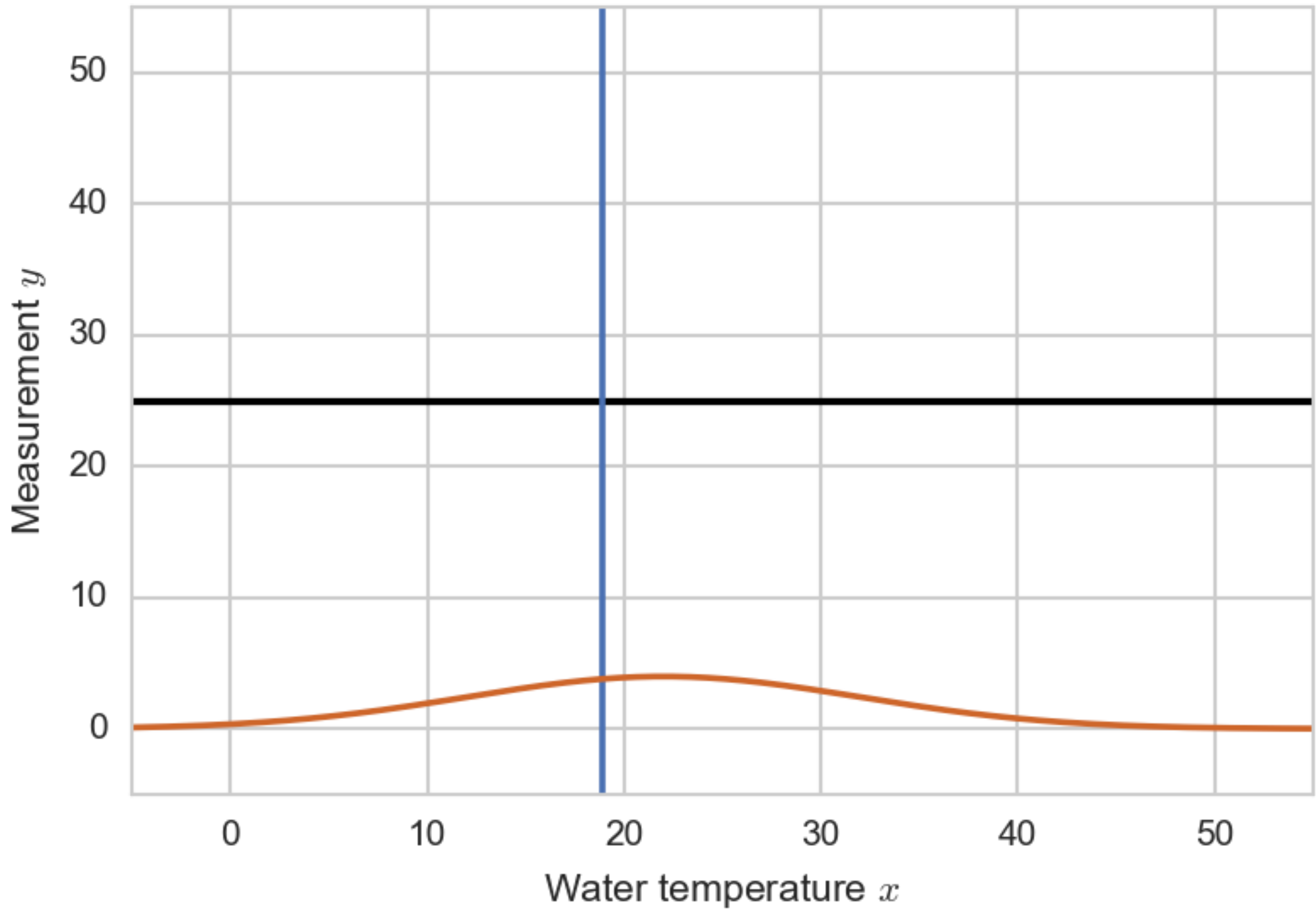
$$W(x_i) \approx \frac{w(x_i)}{\sum_{j=1}^{N} w(x_j)}$$

$$\mathbb{E}_{p(x|y)}[f(x)] \approx \sum_{i=1}^{N} \frac{w(x_i)}{\sum_{j=1}^{N} w(x_j)} f(x_i)$$

Conditioning via importance sampling

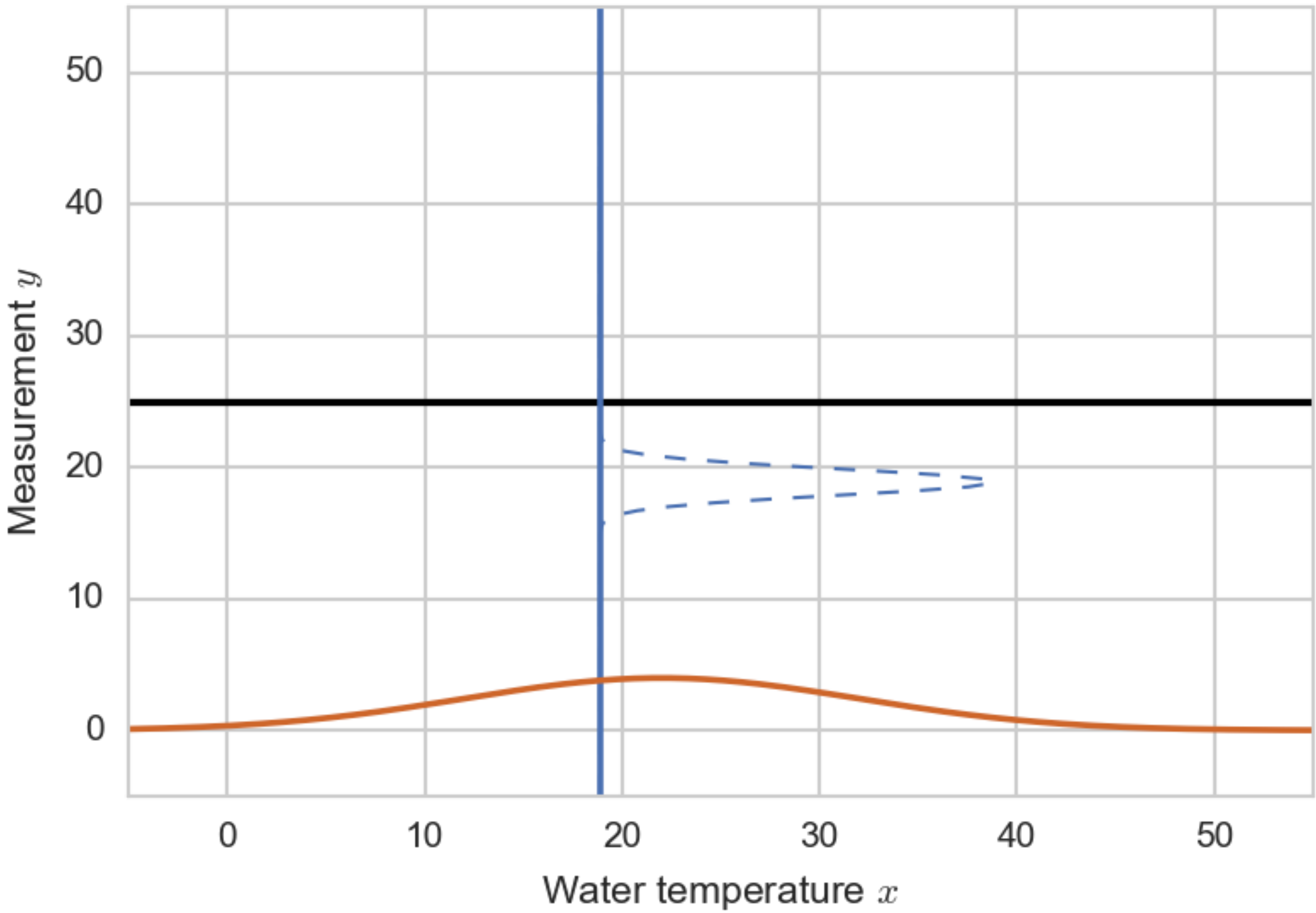
- We already have very simple proposal distribution we know how to sample from: the prior *p(x)*.
- The algorithm then resembles the rejection sampling algorithm, except instead of sampling both the latent variables and the observed variables, we only sample the latent variables
- Then, instead of a “hard” rejection step, we use the values of the latent variables and the data to assign “soft” weights to the sampled values.

Likelihood weighting schematic



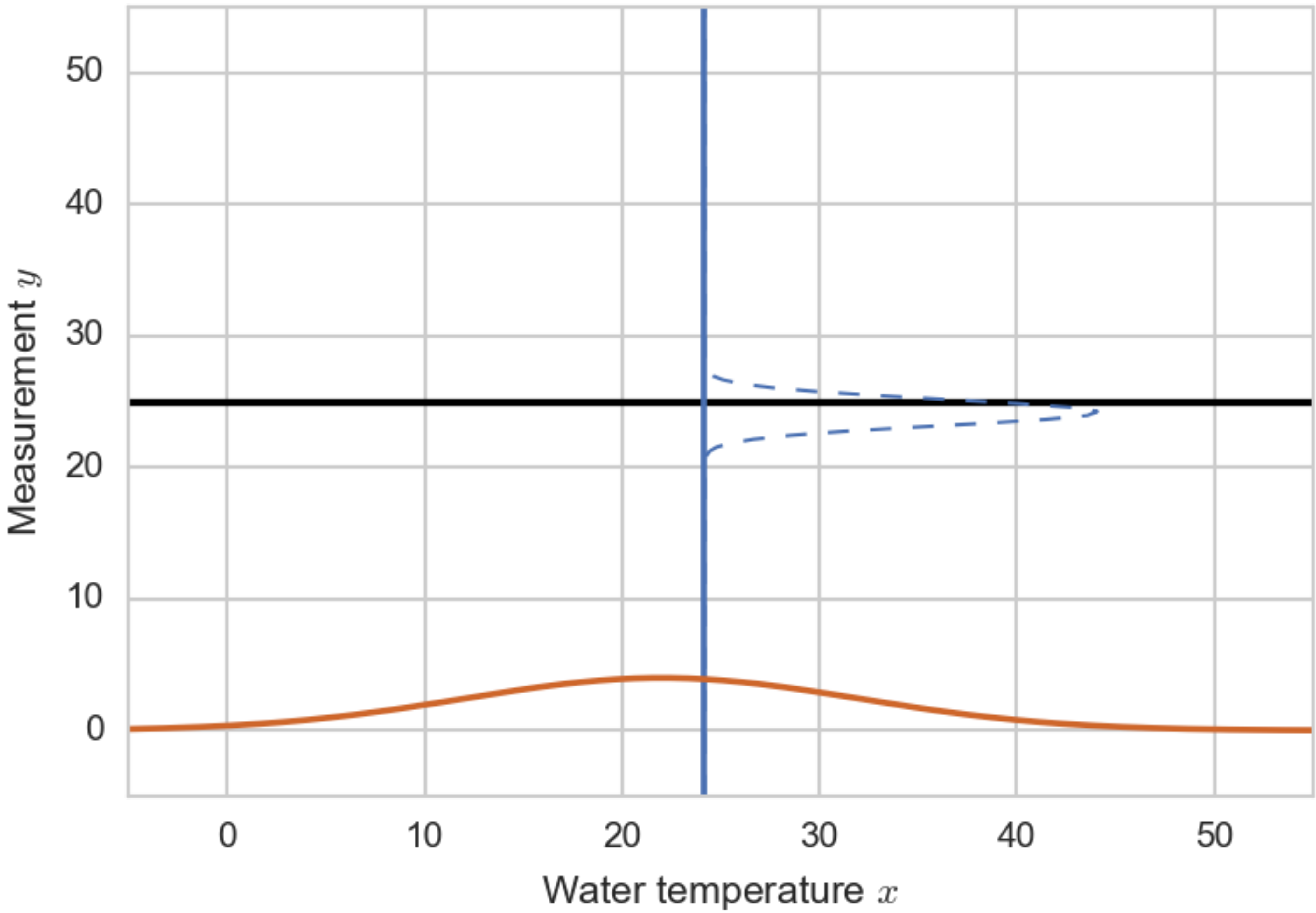
Draw a sample of *x* from the prior

Likelihood weighting schematic



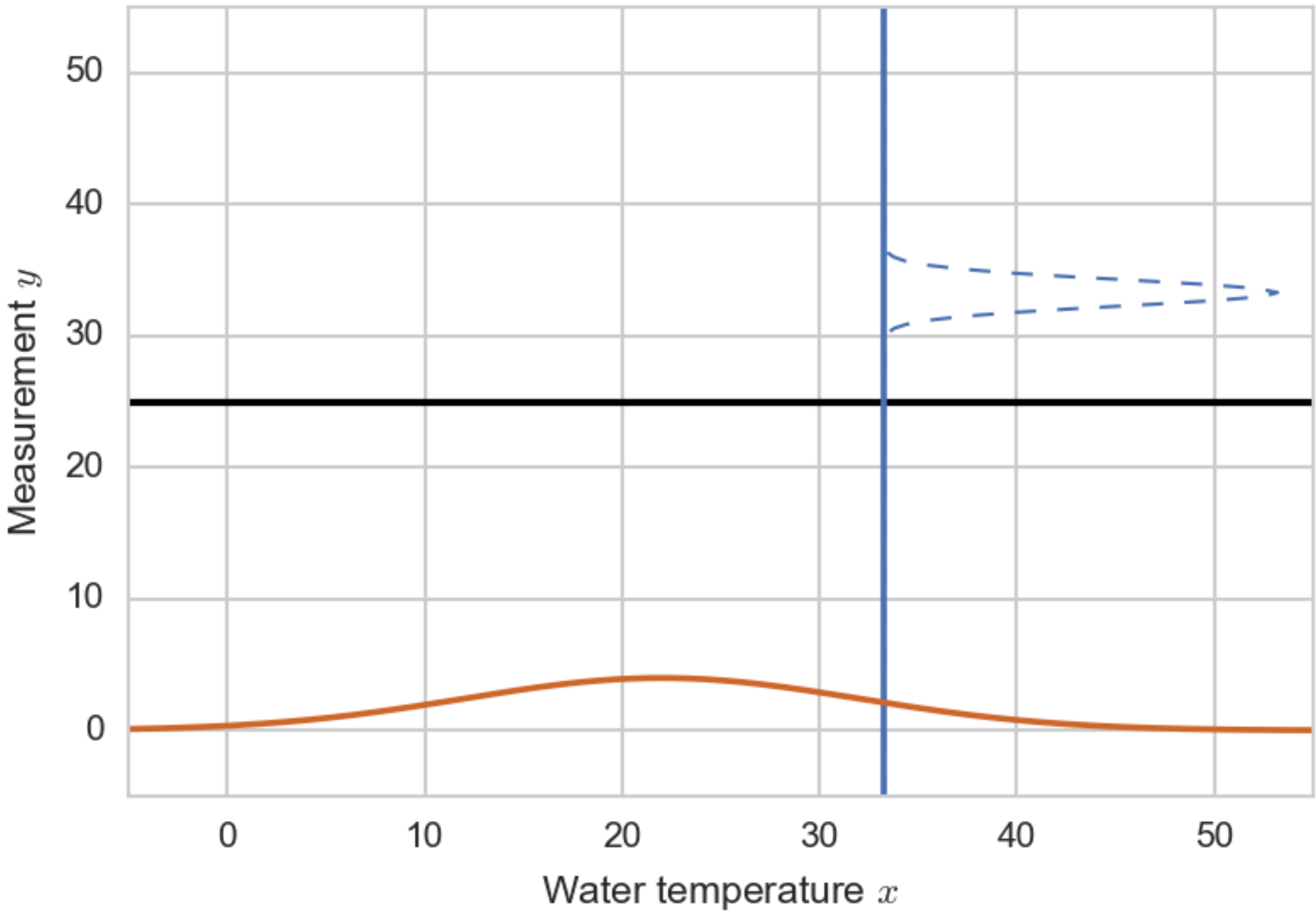
What does *p(y|x)* look like for this sampled *x* ?

Likelihood weighting schematic



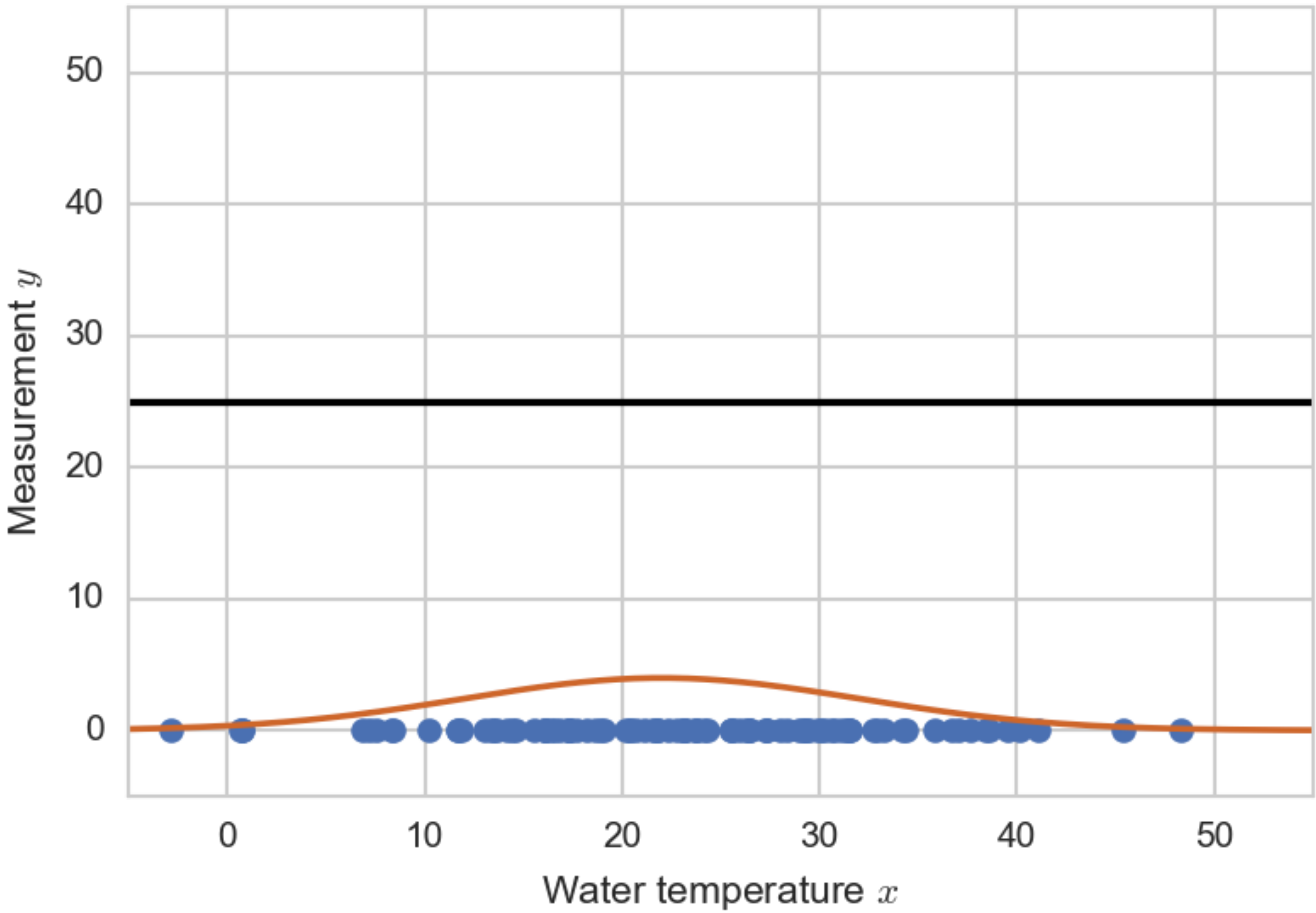
What does $p(y|x)$ look like for this sampled x?

Likelihood weighting schematic



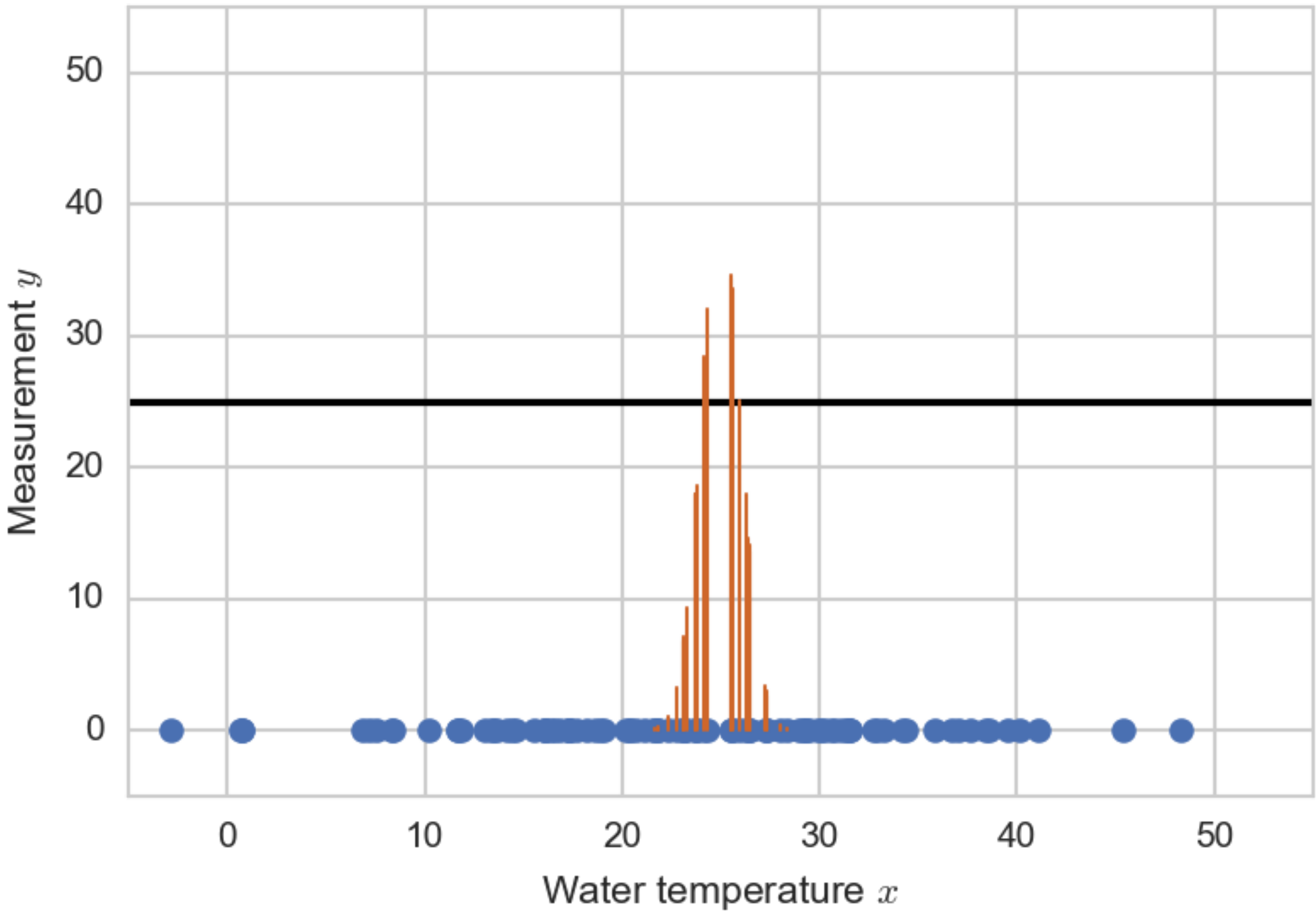
What does $p(y|x)$ look like for this sampled x ?

Likelihood weighting schematic



Compute $p(y|x)$ for *all* of our x drawn from the prior

Likelihood weighting schematic



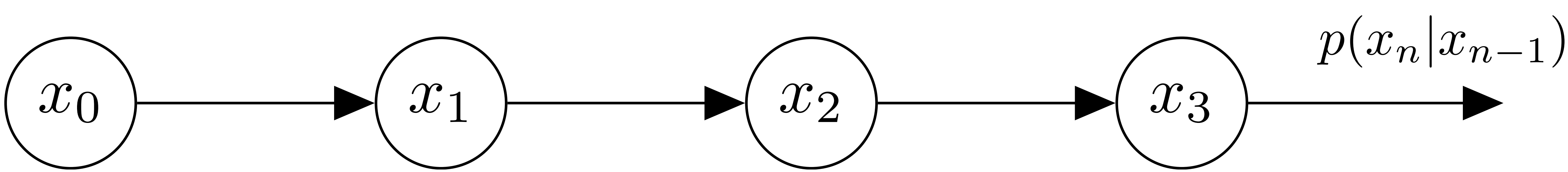
Assign weights (vertical bars) to samples for a representation of the posterior

Conditioning via MCMC

- **Problem**: Likelihood weighting degrades poorly as the dimension of the latent variables increases, unless we have a very well-chosen proposal distribution *q(x)*.

- **An alternative**: Markov chain Monte Carlo (MCMC) methods draw samples from a target distribution by performing a biased random walk over the space of the latent variables *x*.

- Idea: create a Markov chain such that the sequence of states $x_0, x_1, x_2, \ldots$ are samples from $p(x \mid y)$



Conditioning via MCMC

- MCMC also uses a proposal distribution, but this proposal distribution makes **local** changes to the latent variables *x*. The proposal $q(x' \mid x)$ defines a conditional distribution over x' given a current value x.

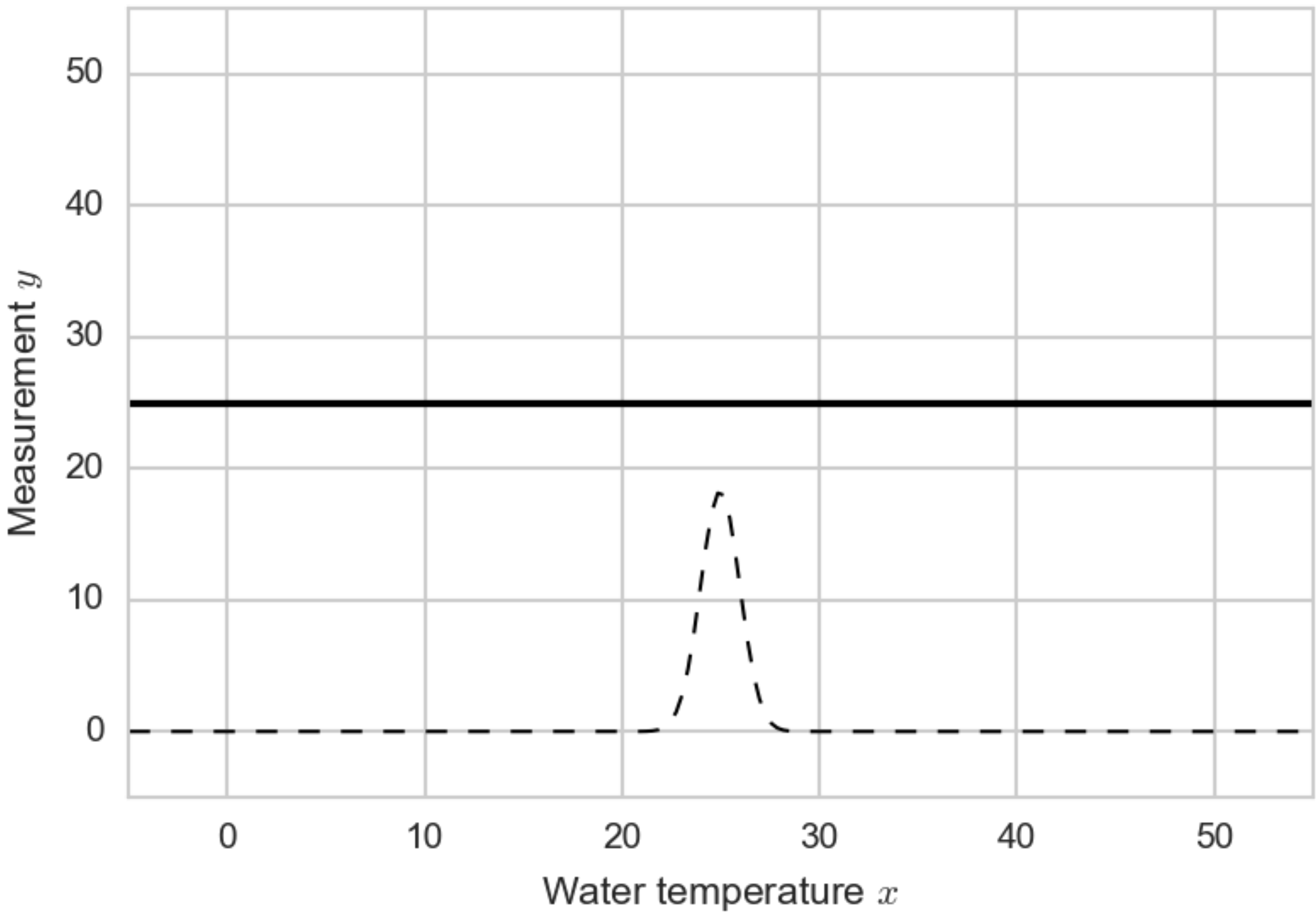
  - Typical choice: add small amount of Gaussian noise

- We use the proposal and the joint density to define an "acceptance ratio"

$$A(x \to x') = \min\left(1,\ \frac{p(x', y)q(x|x')}{p(x, y)q(x'|x)}\right)$$

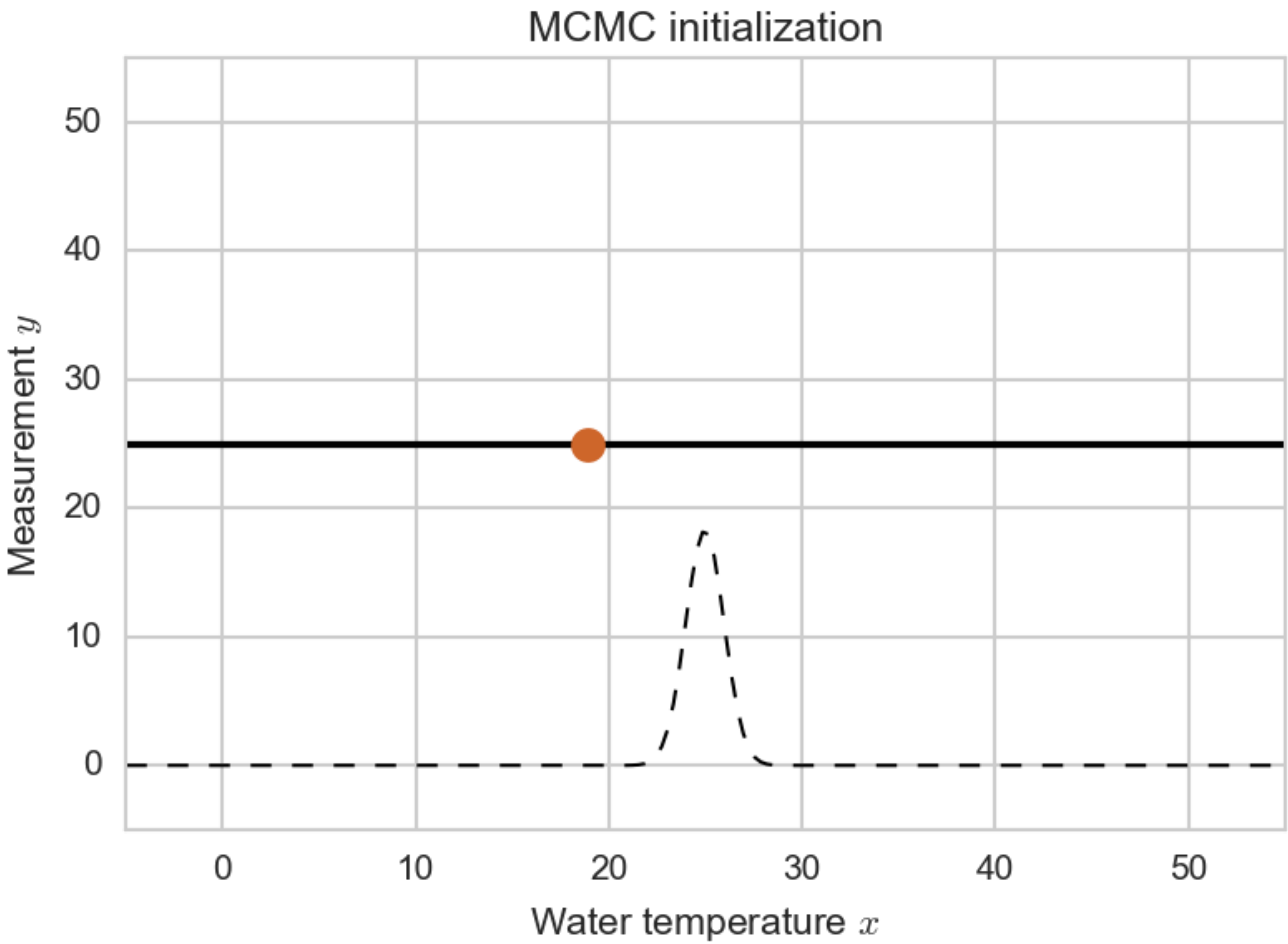
- With probability A we "move" state with the new value x', otherwise we stay at x.

MCMC schematic



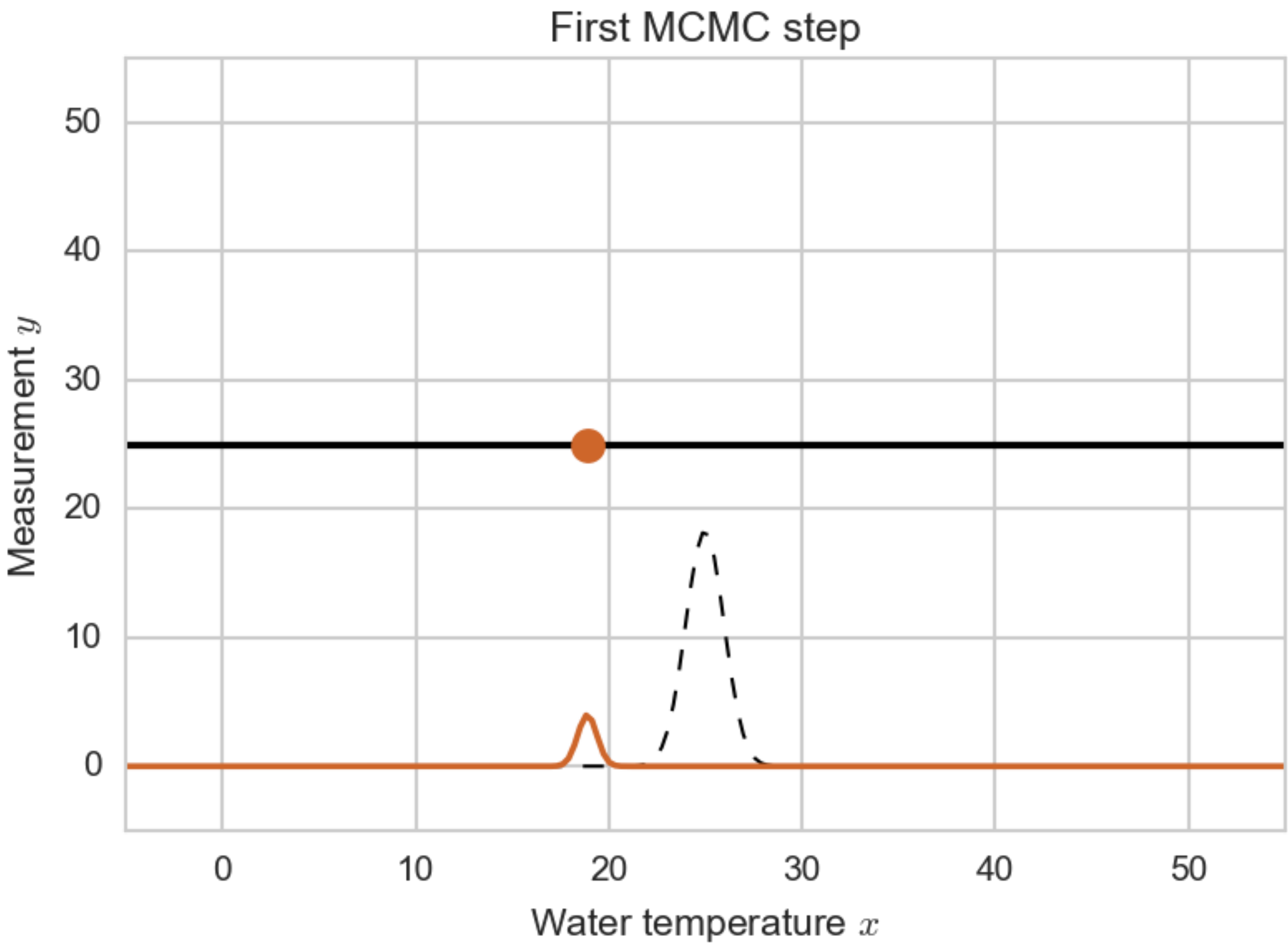
The (unnormalized) joint distribution *p(x,y)* is shown as a dashed line

MCMC schematic



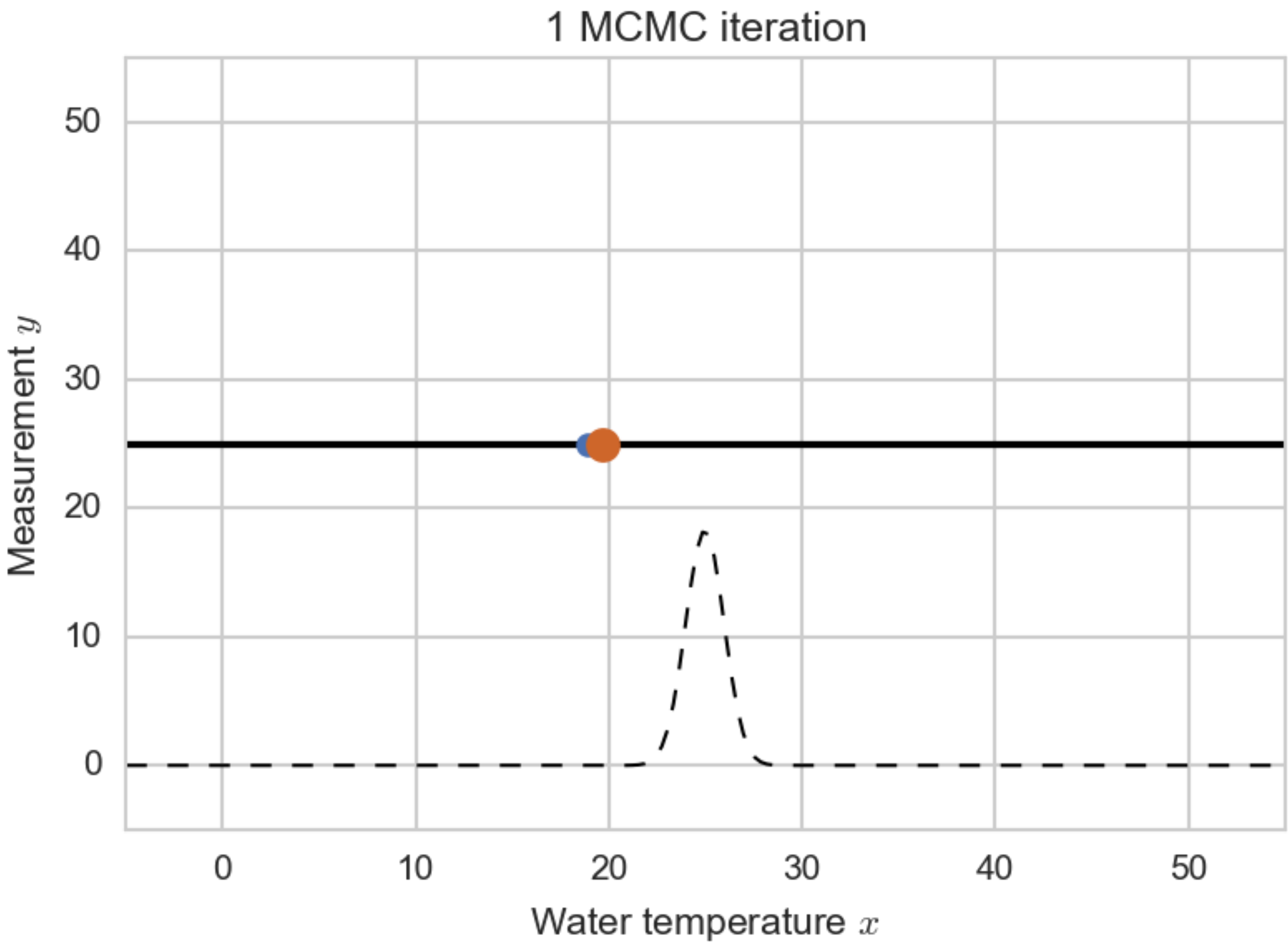
Initialize arbitrarily (e.g. with a sample from the prior)

MCMC schematic



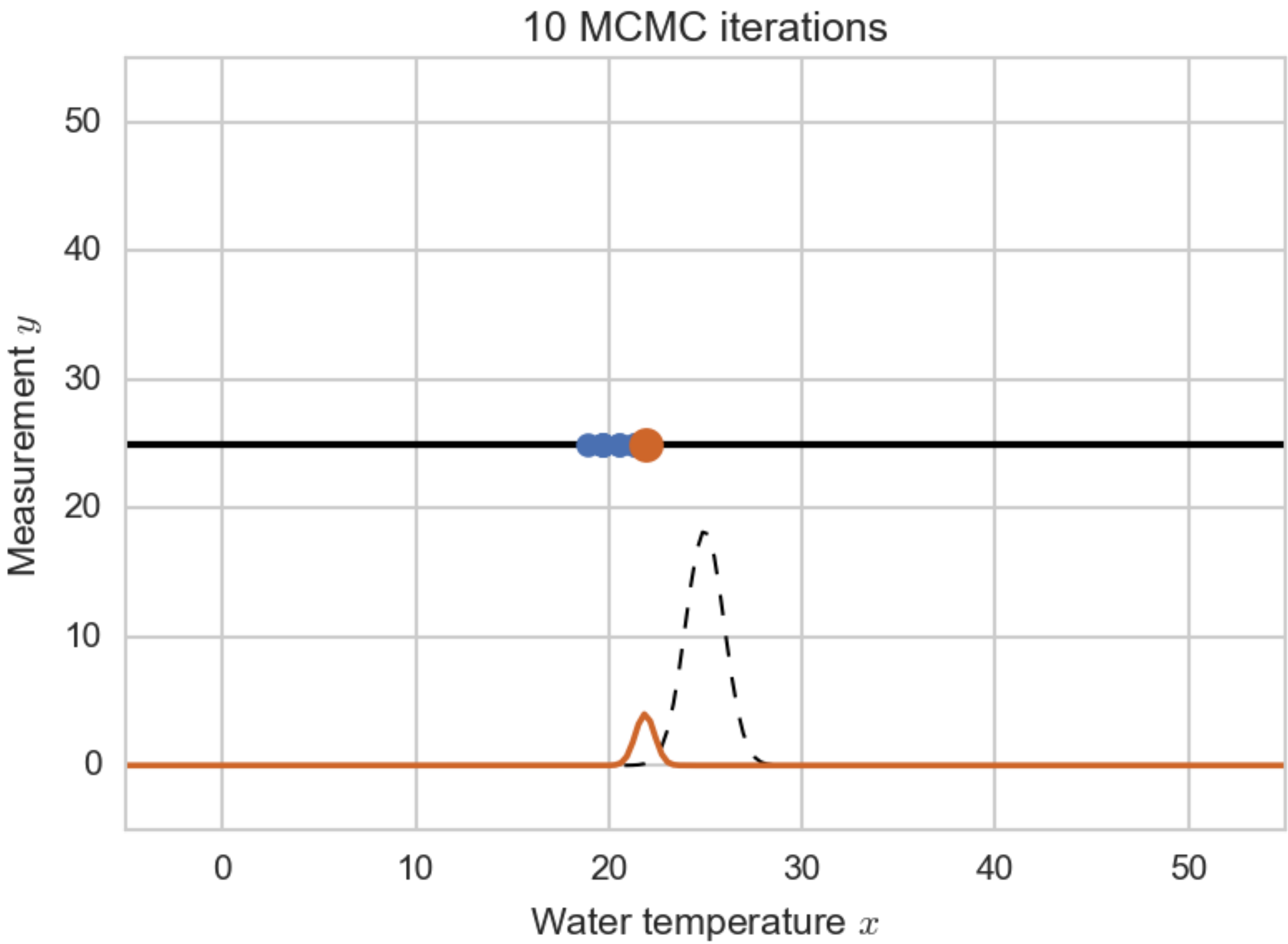
Propose a local move on *x* from a transition distribution

MCMC schematic



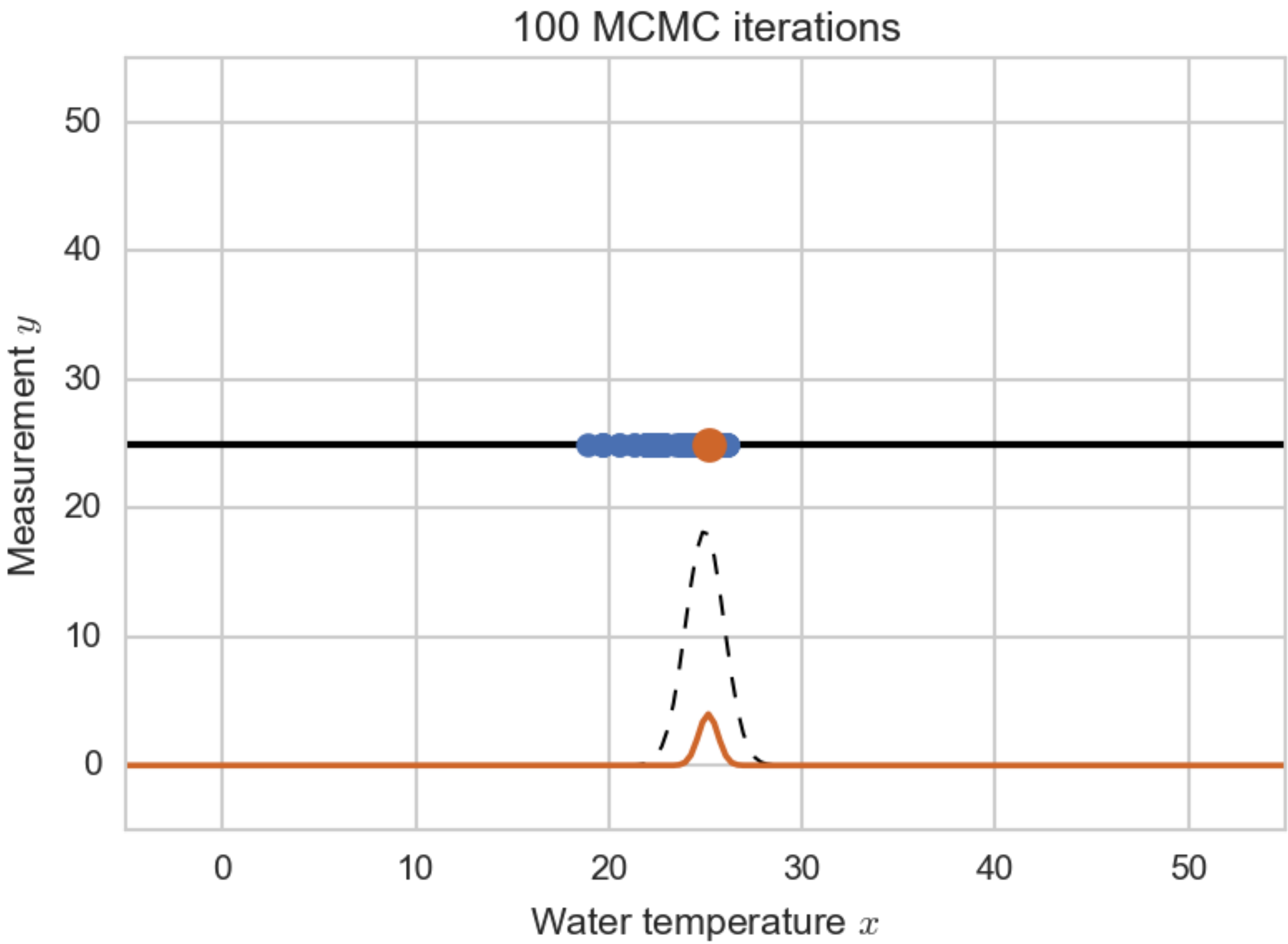
Here, we proposed a point in a region of higher probability density, and accepted

MCMC schematic



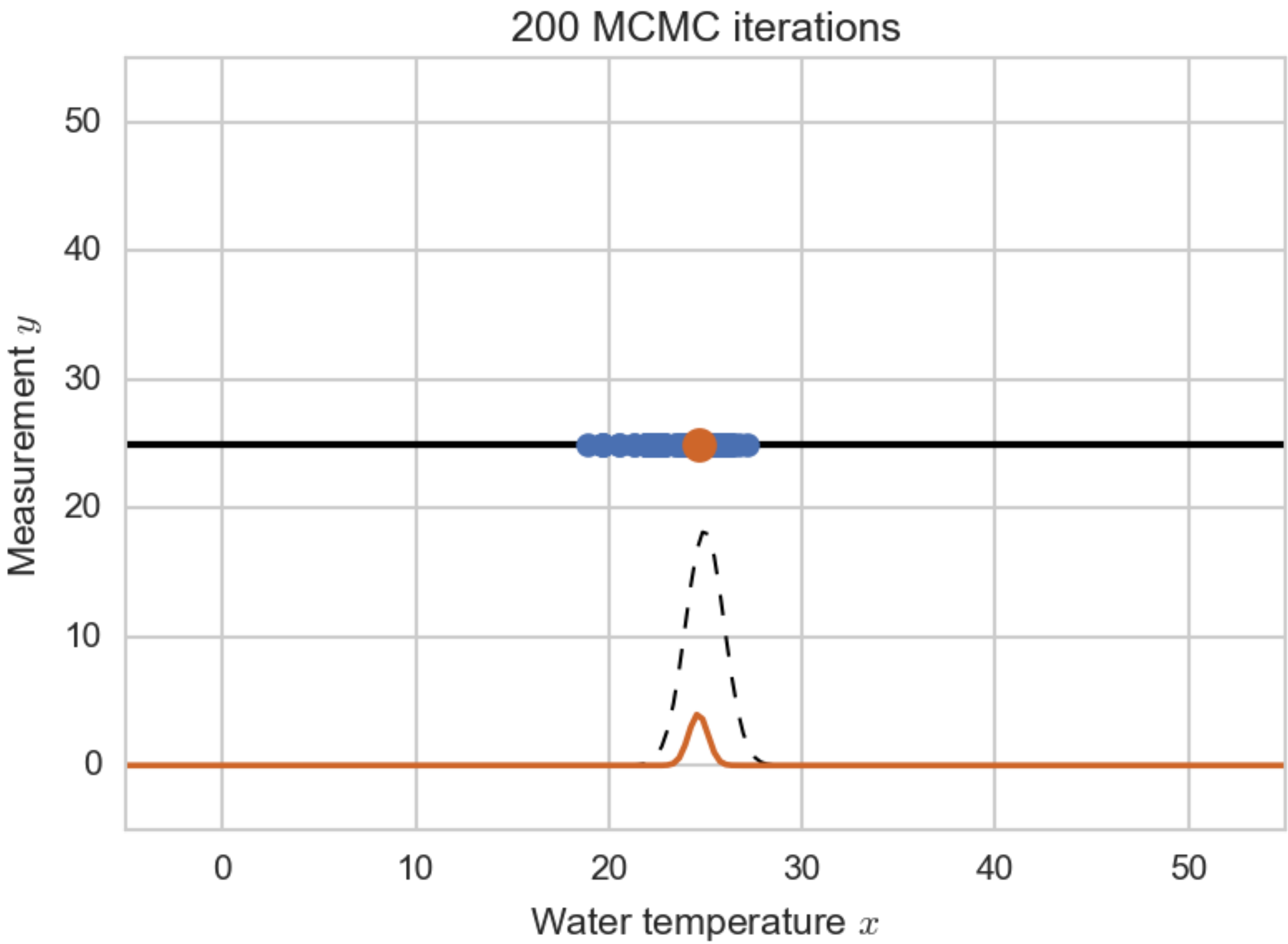
Continue: propose a local move, and accept or reject.
At first, this will look like a stochastic search algorithm!

MCMC schematic



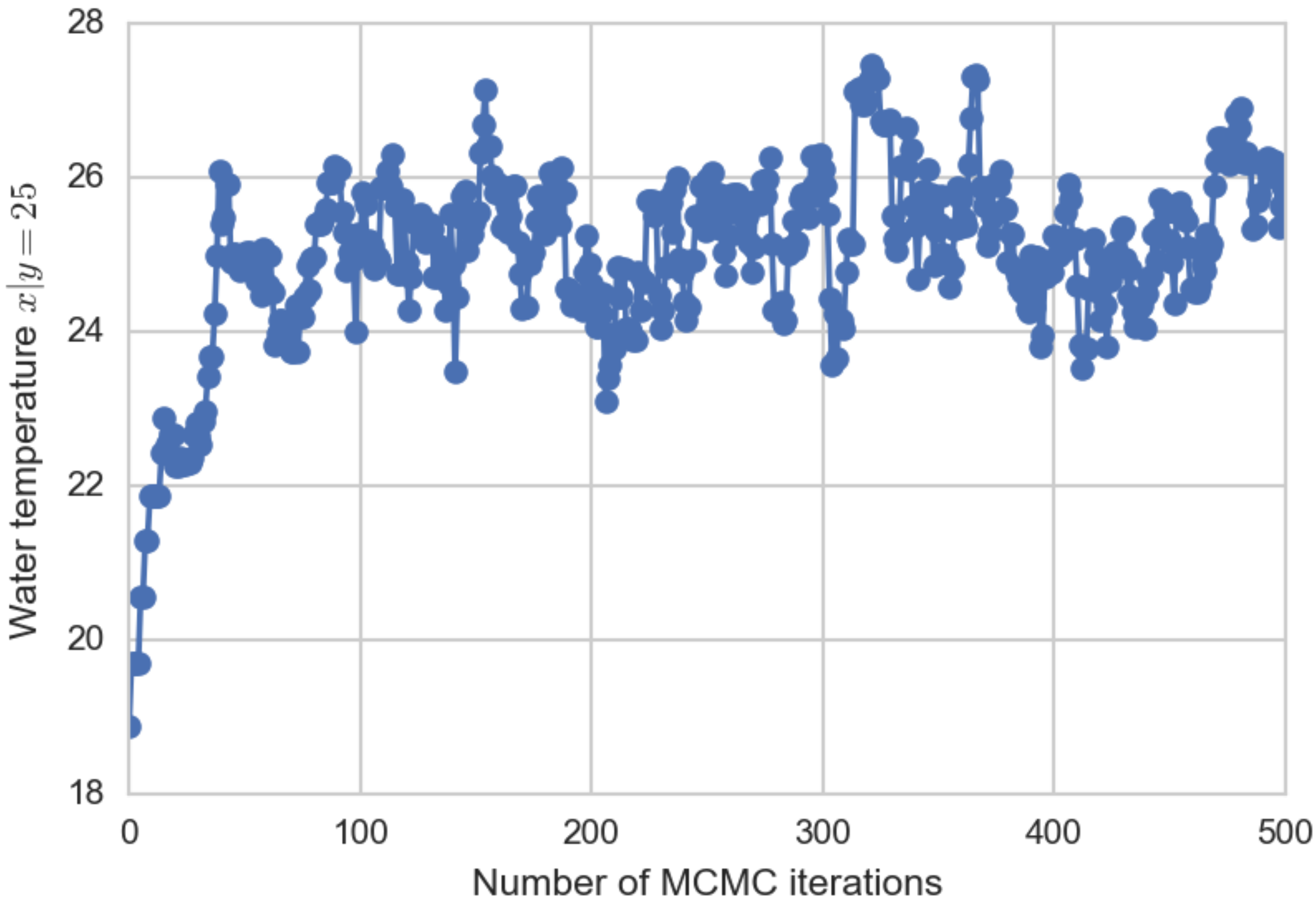
Once in a high-density region, it will explore the space

MCMC schematic



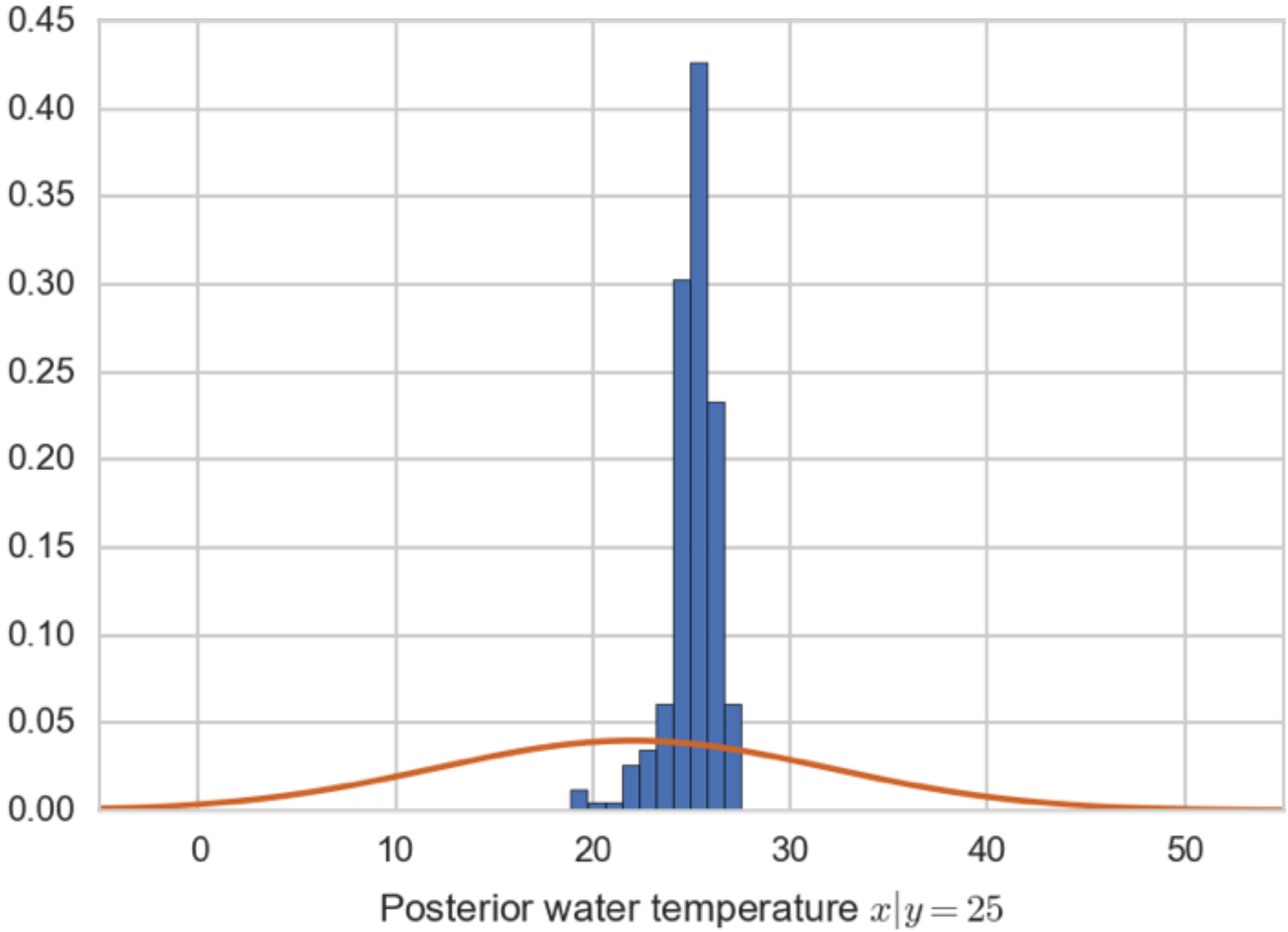
Once in a high-density region, it will explore the space

MCMC schematic



Helpful diagnostic: a “trace plot” of the path of the sampled values, as the number of MCMC iterations increases

MCMC schematic



Histogram of trace plot, overlaid on prior probability density

Now: exercises

- **Part one:** a model much like the model we just looked at, Gaussian data with a latent Gaussian distributed mean

  A. implement likelihood weighting for this model

  B. this is one of the *very few* continuous models where exact inference is possible. Do the math, and check if your sampler is correct!

- **Part two:** seven scientists are performing an experiment to estimate the value of a particular physical constant. Most of them find similar results, but a few differ by surprisingly much. Do I trust all these scientists equally? What is the "real" value? Write an MCMC sampler to find out!