More Monte Carlo Methods

Brooks Paige

COMP0171

Ancestral sampling

We can build samplers for complicated distributions out of samplers for simple distributions compositionally.

This is particularly useful for sampling from joint distributions defined by generative models, i.e. by a sequence of conditional distributions.

Ancestral sampling

We can build samplers for complicated distributions out of samplers for simple distributions compositionally.

This is particularly useful for sampling from joint distributions defined by generative models, i.e. by a sequence of conditional distributions.

We've already seen a generative model for N coin flips:

$$\begin{aligned} x &\sim \mathcal{U}([0,1]) \\ y_i|x &\sim \text{Bernoulli}(x), \qquad\qquad i = 1, \ldots, N. \end{aligned}$$

We can simulate from the joint distribution over $(x, y_1, \ldots, y_N)$ by sampling from the conditional distributions in turn.

Ancestral sampling

$x \sim \mathcal{U}([0,1])$ $\qquad$ $y_1, y_2, y_3 | x \sim \text{Bernoulli}(x)$

x=0.214; y=[0 0 0]	x=0.164; y=[0 0 0]	x=0.584; y=[1 0 1]
x=0.794; y=[1 0 1]	x=0.674; y=[1 1 0]	x=0.297; y=[1 0 1]
x=0.611; y=[0 0 1]	x=0.872; y=[1 1 1]	x=0.495; y=[1 0 1]
x=0.023; y=[0 0 0]	x=0.461; y=[1 0 1]	x=0.597; y=[1 0 0]
x=0.634; y=[0 1 1]	x=0.687; y=[1 1 1]	x=0.603; y=[0 0 1]
x=0.983; y=[1 1 1]	x=0.519; y=[0 1 1]	x=0.280; y=[0 0 0]
x=0.698; y=[1 1 1]	x=0.800; y=[0 1 1]	x=0.693; y=[0 1 1]
x=0.905; y=[1 0 1]	x=0.363; y=[0 1 0]	x=0.635; y=[0 1 0]
x=0.006; y=[0 0 0]	x=0.042; y=[1 0 0]	x=0.383; y=[0 1 1]
x=0.490; y=[1 0 1]	x=0.535; y=[0 1 1]	x=0.134; y=[0 0 0]
x=0.962; y=[1 1 1]	x=0.935; y=[1 1 1]	x=0.722; y=[0 0 0]
x=0.436; y=[1 1 0]	x=0.449; y=[0 0 0]	x=0.395; y=[1 1 0]

Conditioning by "guess and check"

What if we want to sample from a conditional distribution? The conceptually simplest form is via a type of rejection sampling:

1. Use the ancestral sampling procedure to simulate from the generative process, drawing a sample (x, y) from the joint distribution $p(x, y)$
2. For a given y, e.g. $[0, 1, 1]$, to estimate the posterior $p(x|y = [0, 1, 1])$, we say that x is a sample from the posterior if its corresponding jointly sampled $y = [0, 1, 1]$.

Conditioning by "guess and check"

What if we want to sample from a conditional distribution? The conceptually simplest form is via a type of rejection sampling:

1. Use the ancestral sampling procedure to simulate from the generative process, drawing a sample (x, y) from the joint distribution $p(x, y)$
2. For a given y, e.g. $[0, 1, 1]$, to estimate the posterior $p(x|y = [0, 1, 1])$, we say that x is a sample from the posterior if its corresponding jointly sampled $y = [0, 1, 1]$.

Warning: problematic as the number of observations in y gets large. . . !

Conditioning by "guess and check"

What if we want to sample from a conditional distribution? The conceptually simplest form is via a type of rejection sampling:

1. Use the ancestral sampling procedure to simulate from the generative process, drawing a sample (x, y) from the joint distribution $p(x, y)$
2. For a given y, e.g. $[0, 1, 1]$, to estimate the posterior $p(x|y = [0, 1, 1])$, we say that x is a sample from the posterior if its corresponding jointly sampled $y = [0, 1, 1]$.

Warning: problematic as the number of observations in y gets large. . . !

Exercise: why are the corresponding x values samples from $p(x|y)$?

Continuous-valued 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.

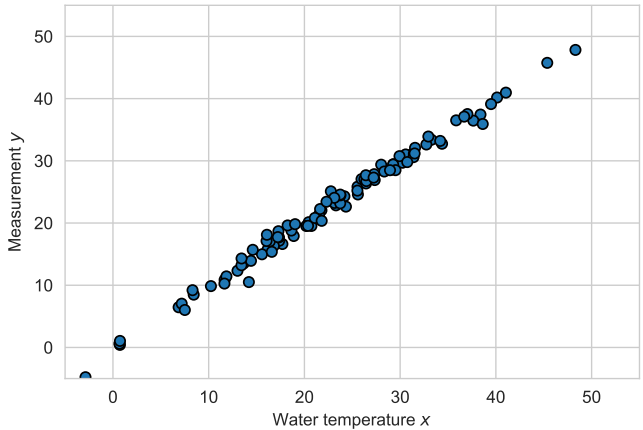
$$x \sim \mathcal{N}(22, 10)$$
$$y|x \sim \mathcal{N}(x, 1)$$

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

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

Exercise: What's wrong with the "rejection" method now?

Ancestral sampling

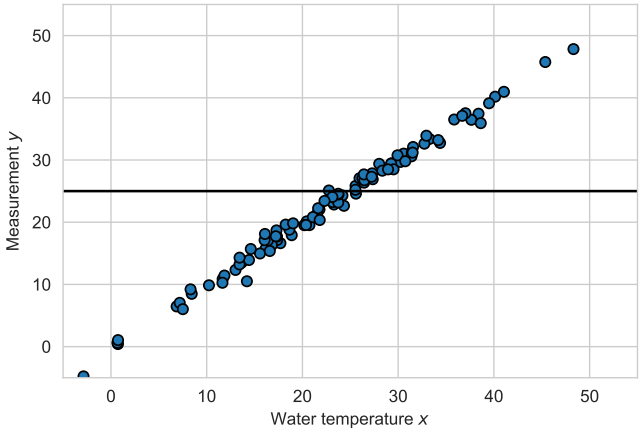


We can draw samples from the joint distribution (x, y) by ancestral sampling.

"Guess and check" with continuous distributions

The black bar shows the measurement, $y = 25$.

Question: How many of these samples from the joint have $y = 25$?



Normalizing constant

Bayes' rule:

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

Sum rule, in the denominator:

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

The "sum" rule is now an integral. This is (in general) intractable, and typically we can only evaluate the posterior up to the normalizing constant.

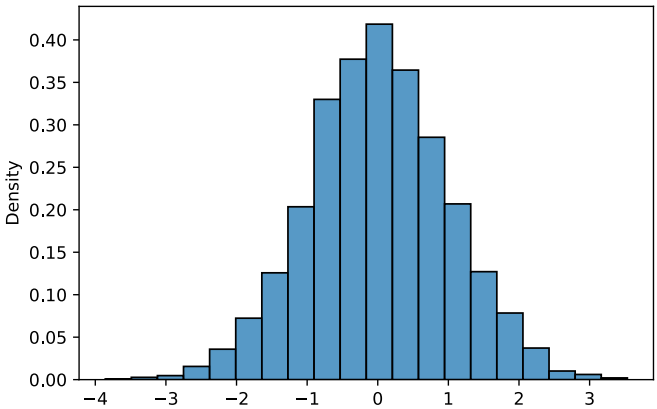
What do we actually want?

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

Posterior Likelihood Prior

- Typically we aren't actually interested in the posterior density (in this case, $p(x|y)$) as our end goal
- Instead, we generally want to know expected values of some function $f(x)$ under the posterior

Histograms are expectations



"Probability that x falls in bin k" $= \mathbb{E}_{p(x)}[\mathbb{I}[x \in [\ell_k, r_k)]]$

Importance sampling

One option is to sidestep sampling from the posterior $p(x|y = 25)$ entirely, and just try to estimate expectations.

In **importance sampling**, we use samples from one distribution to approximate expectations in another. This won't usually scale to high-dimensional problems, but will get us started.

Importance sampling in general (1/2)

Let's say we have some **target distribution** $\pi(\mathbf{x})$ which we do not know how to sample from, but would like to estimate $\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})]$.

However, we do know how to sample from a (hopefully related) **proposal distribution** $q(\mathbf{x})$.

$$\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})] = \int \pi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}$$

Importance sampling in general (1/2)

Let's say we have some **target distribution** $\pi(\mathbf{x})$ which we do not know how to sample from, but would like to estimate $\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})]$.

However, we do know how to sample from a (hopefully related) **proposal distribution** $q(\mathbf{x})$.

$$\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})] = \int \pi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \frac{q(\mathbf{x})}{q(\mathbf{x})} \pi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}$$

Importance sampling in general (1/2)

Let's say we have some **target distribution** $\pi(\mathbf{x})$ which we do not know how to sample from, but would like to estimate $\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})]$.

However, we do know how to sample from a (hopefully related) **proposal distribution** $q(\mathbf{x})$.

$$\begin{aligned}\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})] = \int \pi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} &= \int \frac{q(\mathbf{x})}{q(\mathbf{x})} \pi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \int q(\mathbf{x}) \frac{\pi(\mathbf{x})}{q(\mathbf{x})} f(\mathbf{x}) \mathrm{d}\mathbf{x}\end{aligned}$$

Importance sampling in general (1/2)

Let's say we have some **target distribution** $\pi(\mathbf{x})$ which we do not know how to sample from, but would like to estimate $\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})]$.

However, we do know how to sample from a (hopefully related) **proposal distribution** $q(\mathbf{x})$.

$$\begin{aligned}\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})] = \int \pi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} &= \int \frac{q(\mathbf{x})}{q(\mathbf{x})} \pi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \int q(\mathbf{x}) \frac{\pi(\mathbf{x})}{q(\mathbf{x})} f(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbb{E}_{q(\mathbf{x})}[W(\mathbf{x}) f(\mathbf{x})],\end{aligned}$$

where

$$W(\mathbf{x}) = \frac{\pi(\mathbf{x})}{q(\mathbf{x})}.$$

Importance sampling in general (1/2)

Let's say we have some **target distribution** $\pi(\mathbf{x})$ which we do not know how to sample from, but would like to estimate $\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})]$.

However, we do know how to sample from a (hopefully related) **proposal distribution** $q(\mathbf{x})$.

$$\begin{aligned}\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})] = \int \pi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} &= \int \frac{q(\mathbf{x})}{q(\mathbf{x})} \pi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \int q(\mathbf{x}) \frac{\pi(\mathbf{x})}{q(\mathbf{x})} f(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbb{E}_{q(\mathbf{x})}[W(\mathbf{x}) f(\mathbf{x})],\end{aligned}$$

where

$$W(\mathbf{x}) = \frac{\pi(\mathbf{x})}{q(\mathbf{x})}.$$

This allows us to rewrite an expectation w.r.t. $\pi(\mathbf{x})$ as an expectation w.r.t. $q(\mathbf{x})$.

Importance sampling in general (2/2)

In importance sampling, expectations are computed using *weighted* samples from $q(\mathbf{x})$, rather than *unweighted* (or exact) samples from $\pi(\mathbf{x})$.

Importance sampling in general (2/2)

In importance sampling, expectations are computed using *weighted* samples from $q(\mathbf{x})$, rather than *unweighted* (or exact) samples from $\pi(\mathbf{x})$.

- Sample values $\mathbf{x}^{(i)} \sim q(\mathbf{x})$, $i = 1, \ldots, N$

Importance sampling in general (2/2)

In importance sampling, expectations are computed using *weighted* samples from $q(\mathbf{x})$, rather than *unweighted* (or exact) samples from $\pi(\mathbf{x})$.

- Sample values $\mathbf{x}^{(i)} \sim q(\mathbf{x})$, $i = 1, \ldots, N$
- Compute "importance weights"

$$W(\mathbf{x}^{(i)}) = \frac{\pi(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}$$

Importance sampling in general (2/2)

In importance sampling, expectations are computed using *weighted* samples from $q(\mathbf{x})$, rather than *unweighted* (or exact) samples from $\pi(\mathbf{x})$.

- Sample values $\mathbf{x}^{(i)} \sim q(\mathbf{x})$, $i = 1, \ldots, N$
- Compute "importance weights"

$$W(\mathbf{x}^{(i)}) = \frac{\pi(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}$$

- Estimate the expectation as

$$\mathbb{E}_{\pi(\mathbf{x})}[f(\mathbf{x})] = \mathbb{E}_{q(\mathbf{x})}[W(\mathbf{x})f(\mathbf{x})] \approx \frac{1}{N}\sum_{i=1}^{N} W(\mathbf{x}^{(i)})f(\mathbf{x}^{(i)}).$$

Self-normalized importance sampling (1/3)

Small setback: for our posterior $\pi(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})$, we can only evaluate $W(\mathbf{x})$ up to a constant.

Workaround: define unnormalized weights

$$w(\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{x})}.$$

Self-normalized importance sampling (1/3)

Small setback: for our posterior $\pi(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})$, we can only evaluate $W(\mathbf{x})$ up to a constant.

Workaround: define unnormalized weights

$$w(\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{x})}.$$

First, note that these unnormalized weights can be used to approximate the model evidence:

$$p(\mathbf{y}) = \int p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{x} = \int q(\mathbf{x}) \frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{x})} \mathrm{d}\mathbf{x} = \mathbb{E}_{q(\mathbf{x})}[w(\mathbf{x})].$$

Self-normalized importance sampling (2/3)

Rearranging the posterior expectation,

$$\mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[f(\mathbf{x})] = \int p(\mathbf{x}|\mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} f(\mathbf{x}) \mathrm{d}\mathbf{x}$$

Self-normalized importance sampling (2/3)

Rearranging the posterior expectation,

$$\mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[f(\mathbf{x})] = \int p(\mathbf{x}|\mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} f(\mathbf{x}) \mathrm{d}\mathbf{x}$$
$$= \frac{1}{p(\mathbf{y})} \int p(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x}$$

Self-normalized importance sampling (2/3)

Rearranging the posterior expectation,

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[f(\mathbf{x})] &= \int p(\mathbf{x}|\mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})} f(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \frac{1}{p(\mathbf{y})} \int p(\mathbf{x},\mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \frac{1}{p(\mathbf{y})} \int \frac{q(\mathbf{x})}{q(\mathbf{x})} p(\mathbf{x},\mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x}\end{aligned}$$

Self-normalized importance sampling (2/3)

Rearranging the posterior expectation,

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[f(\mathbf{x})] &= \int p(\mathbf{x}|\mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})} f(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \frac{1}{p(\mathbf{y})} \int p(\mathbf{x},\mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \frac{1}{p(\mathbf{y})} \int \frac{q(\mathbf{x})}{q(\mathbf{x})} p(\mathbf{x},\mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \frac{\mathbb{E}_{q(\mathbf{x})}[w(\mathbf{x}) f(\mathbf{x})]}{\mathbb{E}_{q(\mathbf{x})}[w(\mathbf{x})]}\end{aligned}$$

Self-normalized importance sampling (2/3)

Rearranging the posterior expectation,

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[f(\mathbf{x})] &= \int p(\mathbf{x}|\mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} f(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \frac{1}{p(\mathbf{y})} \int p(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \frac{1}{p(\mathbf{y})} \int \frac{q(\mathbf{x})}{q(\mathbf{x})} p(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \frac{\mathbb{E}_{q(\mathbf{x})}[w(\mathbf{x}) f(\mathbf{x})]}{\mathbb{E}_{q(\mathbf{x})}[w(\mathbf{x})]} \approx \frac{\frac{1}{N} \sum_{i=1}^{N} w(\mathbf{x}^{(i)}) f(\mathbf{x}^{(i)})}{\frac{1}{N} \sum_{i=1}^{N} w(\mathbf{x}^{(i)})}.
\end{aligned}$$

This is called **self-normalized importance sampling**: use the unnormalized importance weights $w(\mathbf{x})$, with the same sampled values from $q(\mathbf{x})$, to estimate both the numerator and the denominator.

Self-normalized importance sampling (3/3)

In summary: even if we can't compute importance weights $W(\mathbf{X}^{(i)})$, we can still define

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

Self-normalized importance sampling (3/3)

In summary: even if we can't compute importance weights $W(\mathbf{X}^{(i)})$, we can still define

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

This has the added bonus of providing an estimator of the model evidence $p(\mathbf{y})$ automatically. (This is useful for comparing different models!)

Self-normalized importance sampling (3/3)

In summary: even if we can't compute importance weights $W(\mathbf{X}^{(i)})$, we can still define

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

This has the added bonus of providing an estimator of the model evidence $p(\mathbf{y})$ automatically. (This is useful for comparing different models!)

It can also be understood as defining an approximate posterior distribution

$$p(\mathbf{x}|\mathbf{y}) \approx \hat{p}(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^{N} W_i \delta_{\mathbf{x}^{(i)}}(\mathbf{x}),$$

where $\delta_{\mathbf{x}^{(i)}}(\mathbf{x})$ is a "point mass", or "Dirac delta function", a degenerate probability density that assigns probability only to the particular value $\mathbf{x}^{(i)}$.

Likelihood weighting

We already have a very simple choice of proposal distribution: the prior $p(\mathbf{x})$.

Likelihood weighting

We already have a very simple choice of proposal distribution: the prior $p(\mathbf{x})$.

$$w(\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{x})} = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})} = p(\mathbf{y}|\mathbf{x})$$

Likelihood weighting

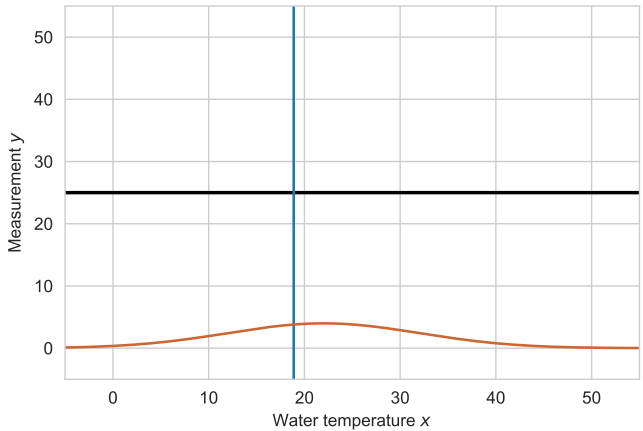
We already have a very simple choice of proposal distribution: the prior $p(\mathbf{x})$.

$$w(\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{x})} = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})} = p(\mathbf{y}|\mathbf{x})$$

The algorithm then resembles rejection sampling, except instead of sampling both the latent variables and the observed variables, we only sample the latent variables.

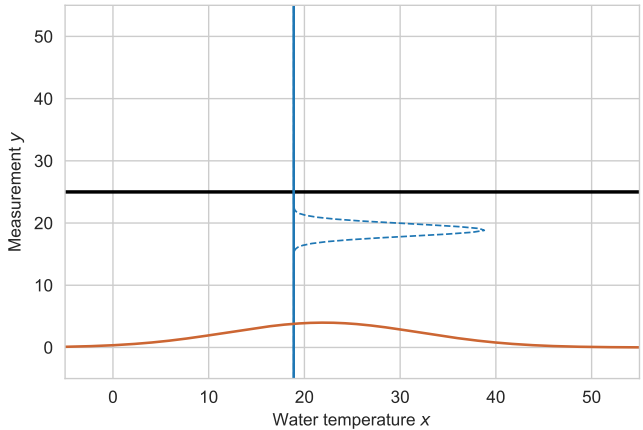
Then instead of a "hard" rejection step, we use the likelihood to assign a "soft" importance weight to the sampled values.

Likelihood weighting schematic



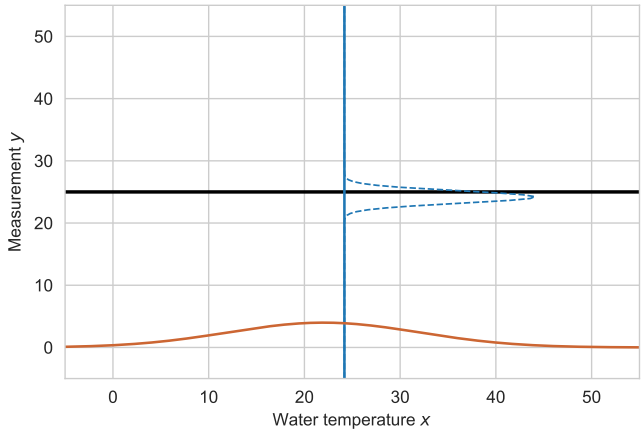
Draw a sample of x from the prior

Likelihood weighting schematic



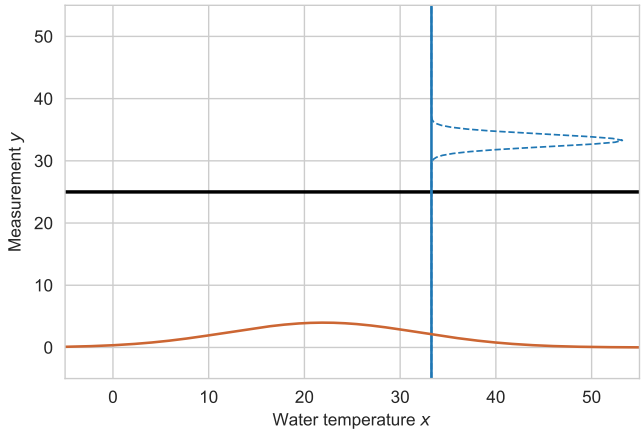
What does $p(y|x)$ look like for this sampled x?

Likelihood weighting schematic



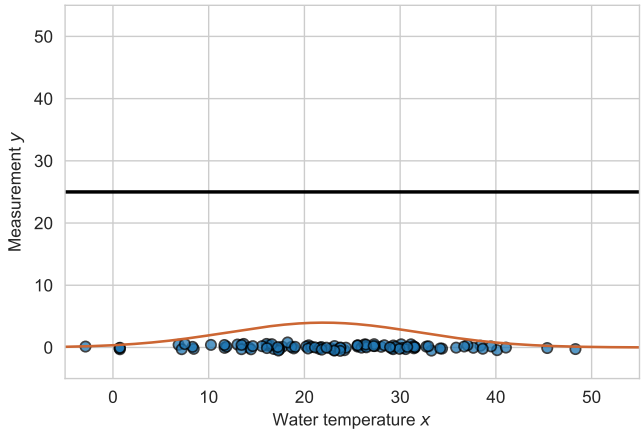
What does $p(y|x)$ look like for this sampled x?

Likelihood weighting schematic



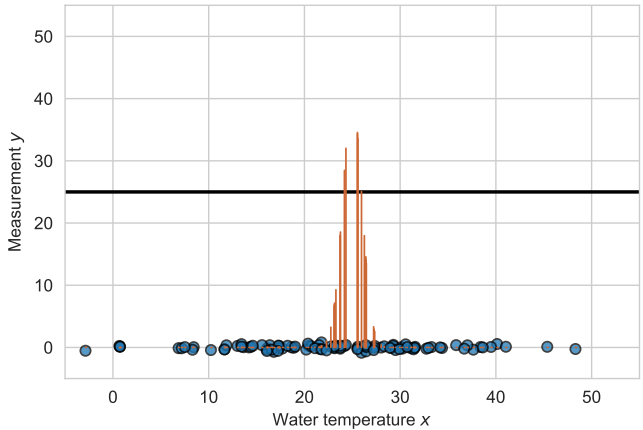
What does $p(y|x)$ look like for this sampled x?

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