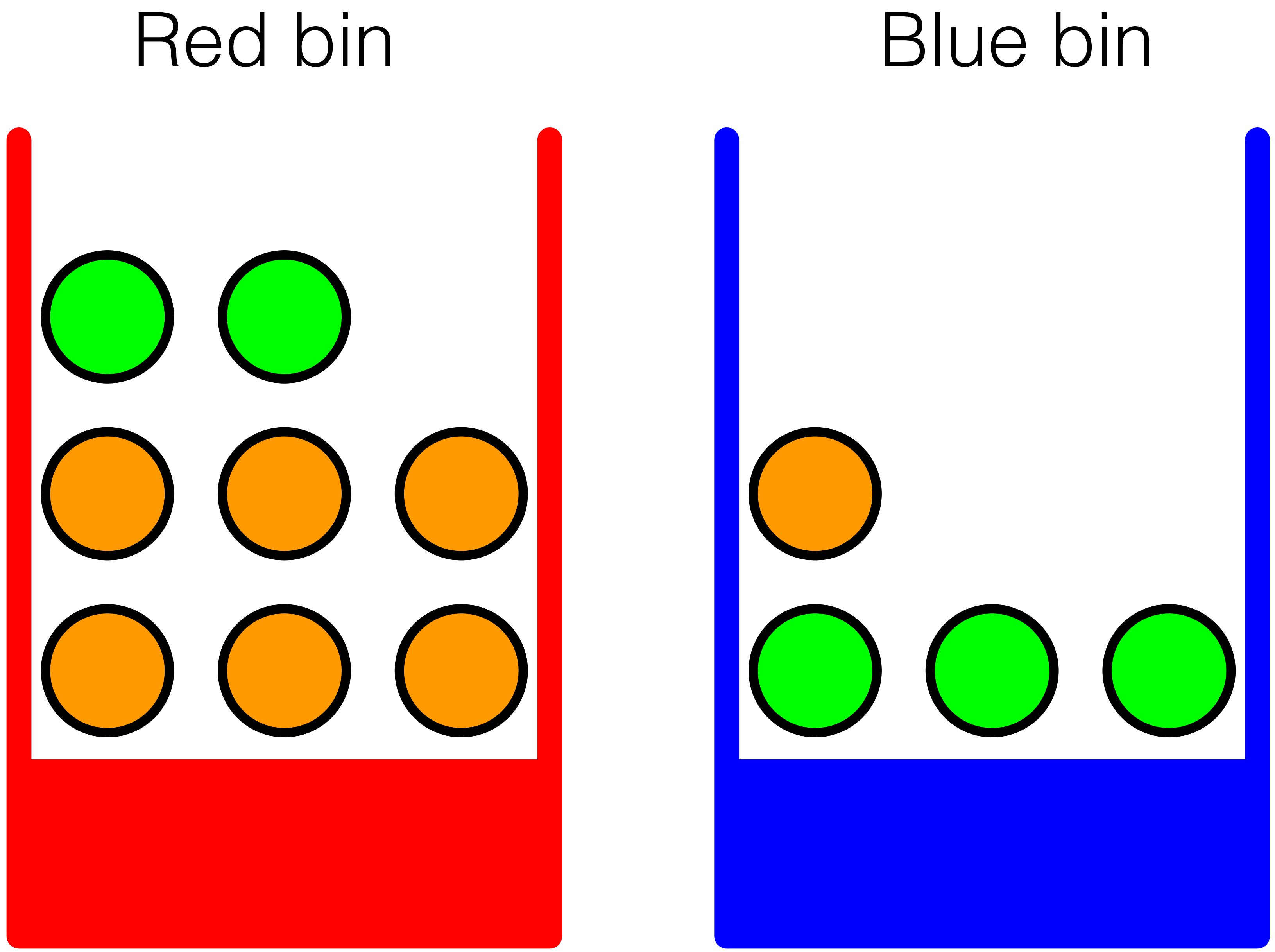
Introduction to Inference

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

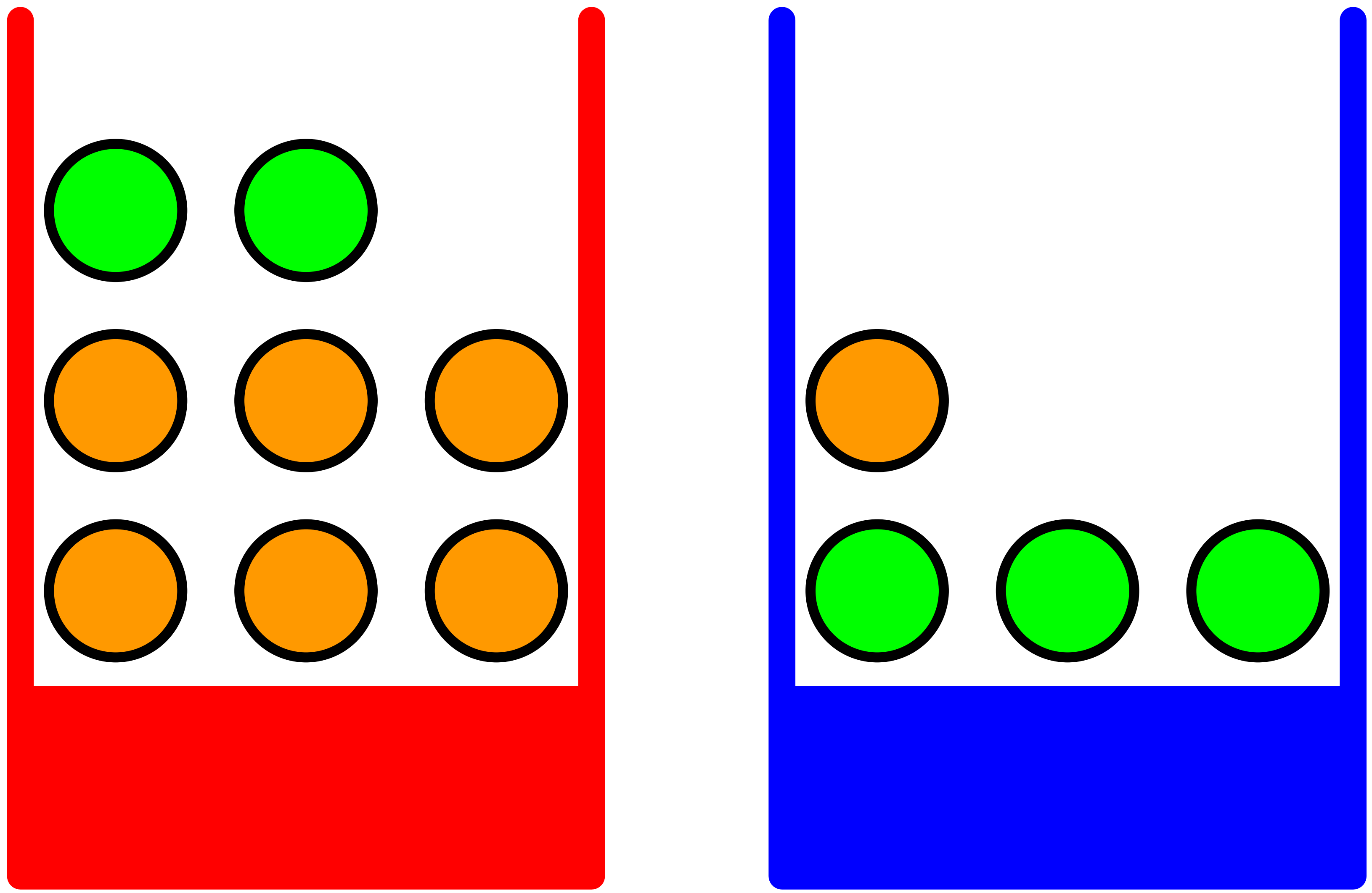


Simple example: discrete probability

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

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

p(apple|red) = 1/4 | p(apple|blue) = 3/4



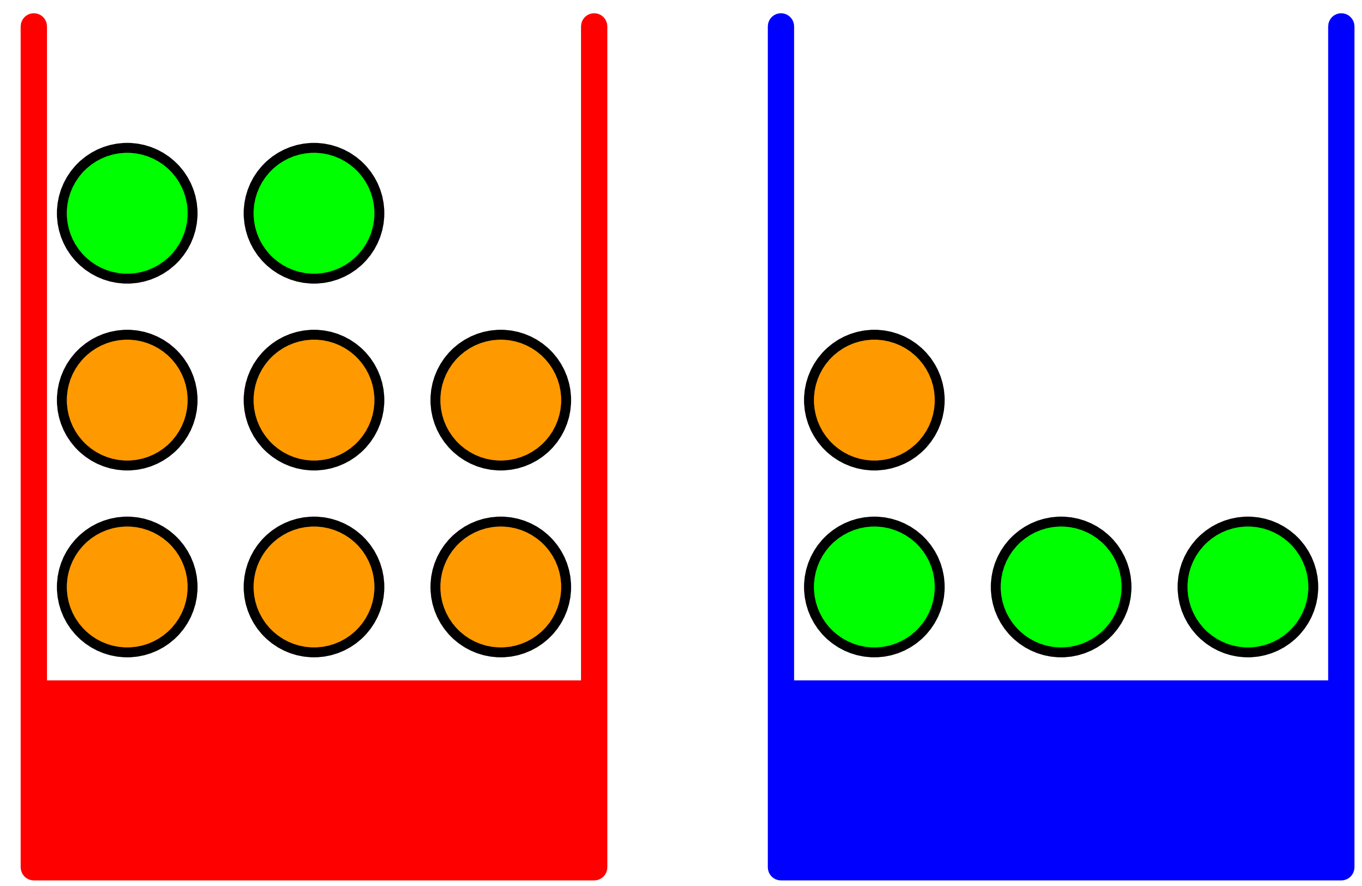
Simple example: discrete probability

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

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

p(red bin) = 2/5
p(apple|red) = 1/4

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



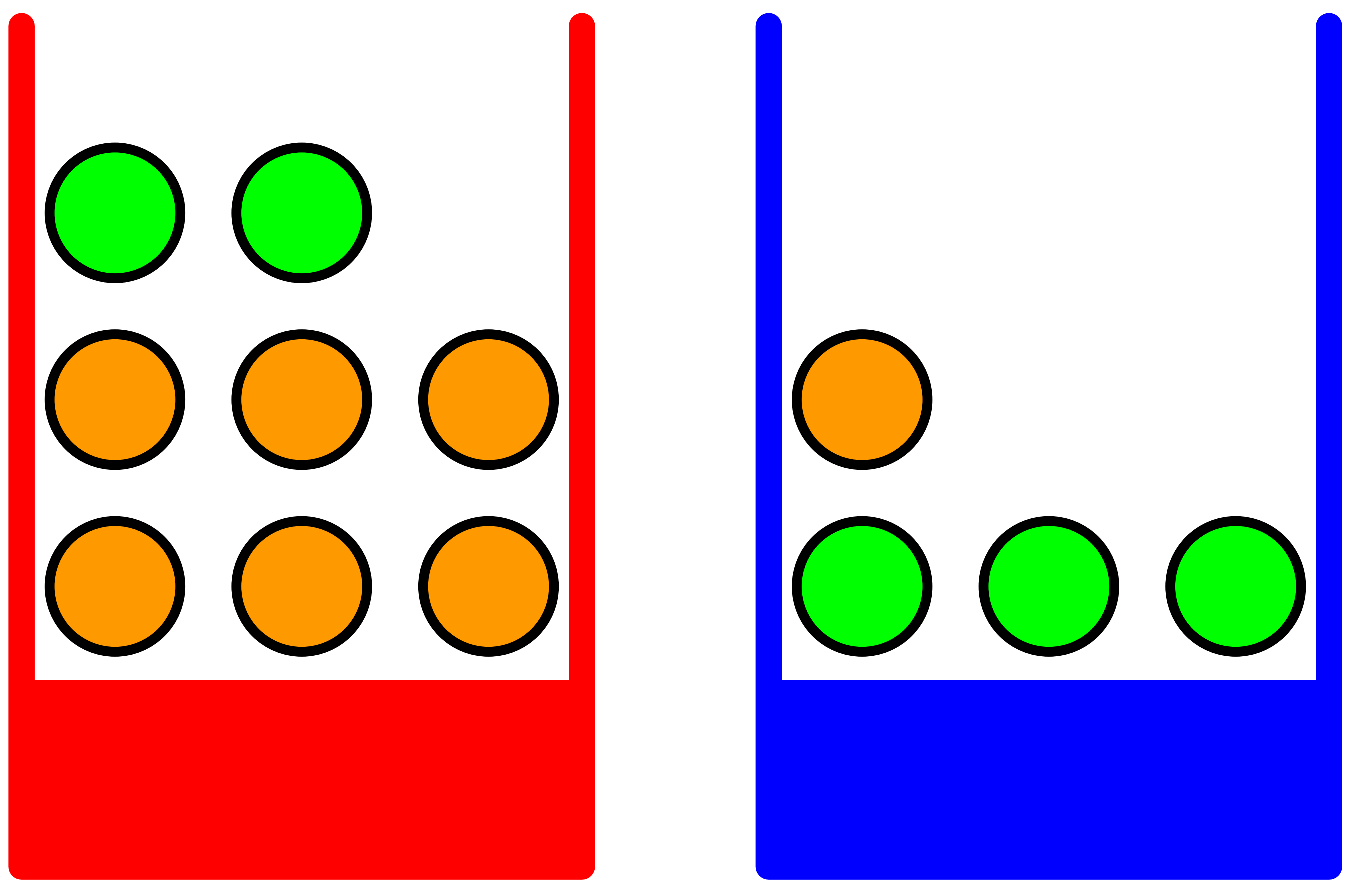
Simple example: discrete probability

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

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

$p(\text{red bin}) = 2/5$
$p(\text{apple}|\text{red}) = 1/4$

$p(\text{blue bin}) = 3/5$
$p(\text{apple}|\text{blue}) = 3/4$



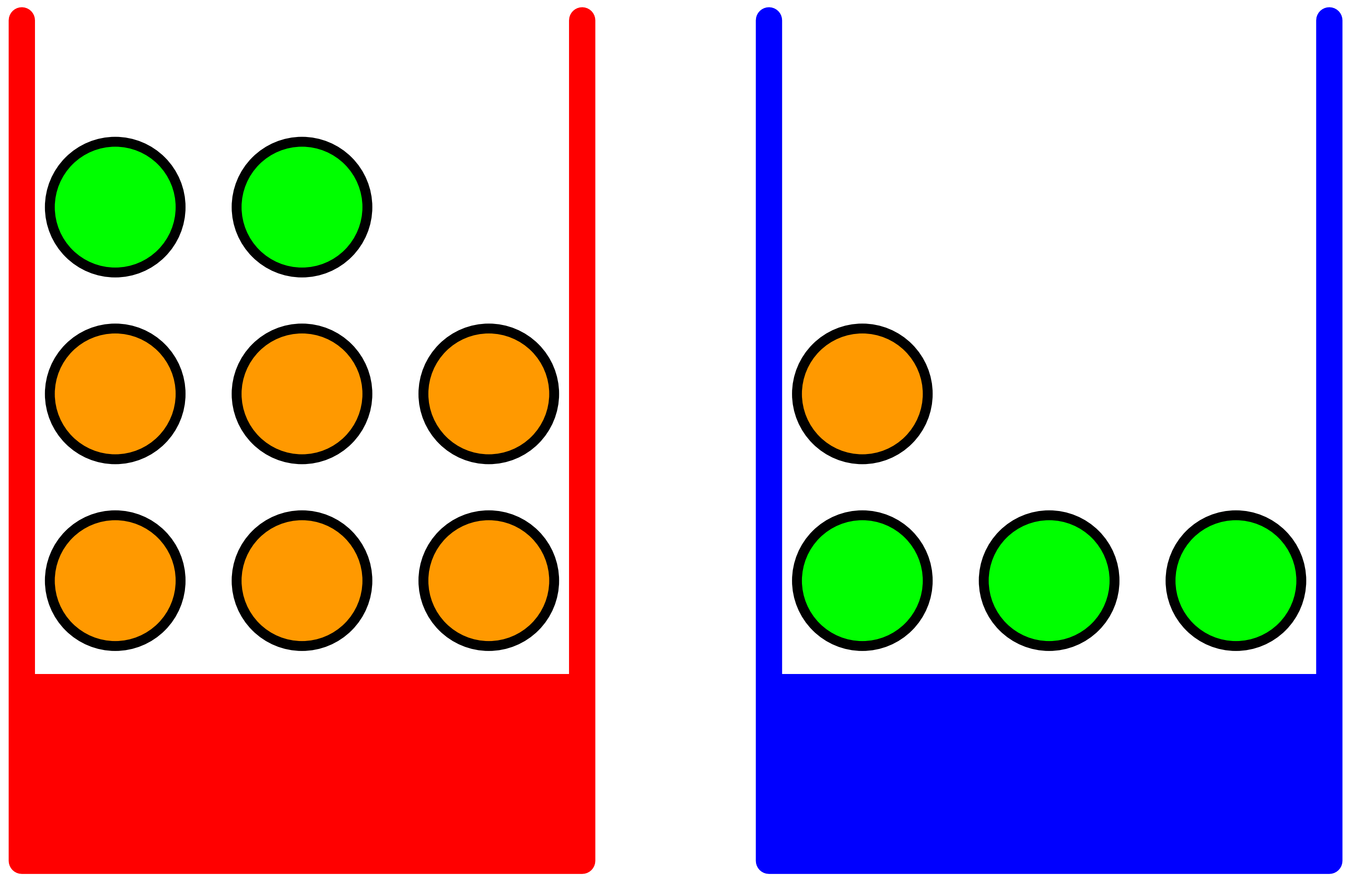
Simple example: discrete probability

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

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

p(red bin) = 2/5
p(apple|red) = 1/4

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



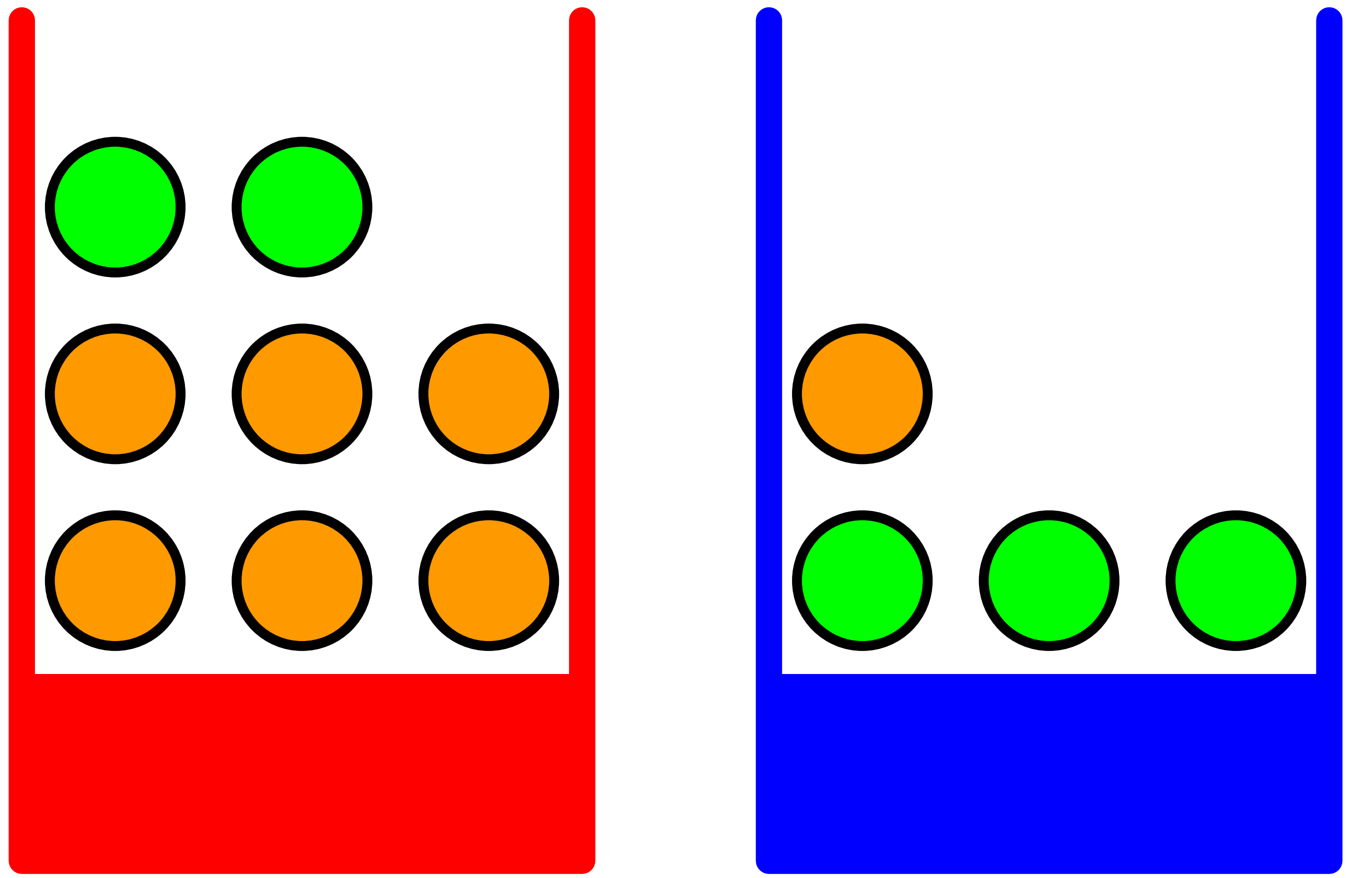
Simple example: discrete probability

"First I pick a bin, then I pick a single ball 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) = 1/4

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 | 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:**

$$p(y) = \sum_{x} p(y, x) \quad p(x) = \sum_{y} p(y, x)$$

- **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:

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

Mini–exercise

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

Use the sum and product rules!

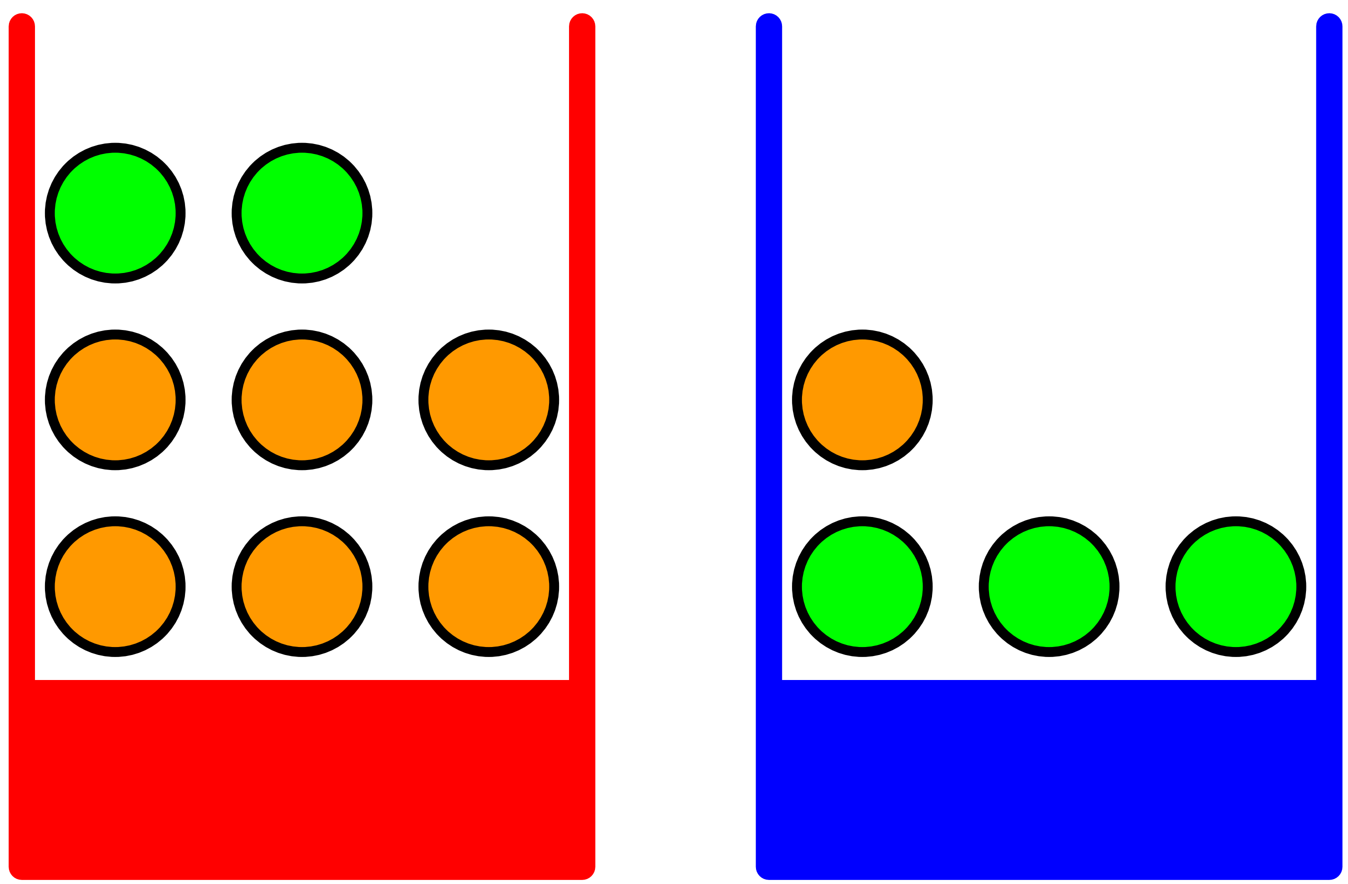
Simple example: discrete probability

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

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

p(red bin) = 2/5
p(apple|red) = 1/4

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



TODO: actually show worked math

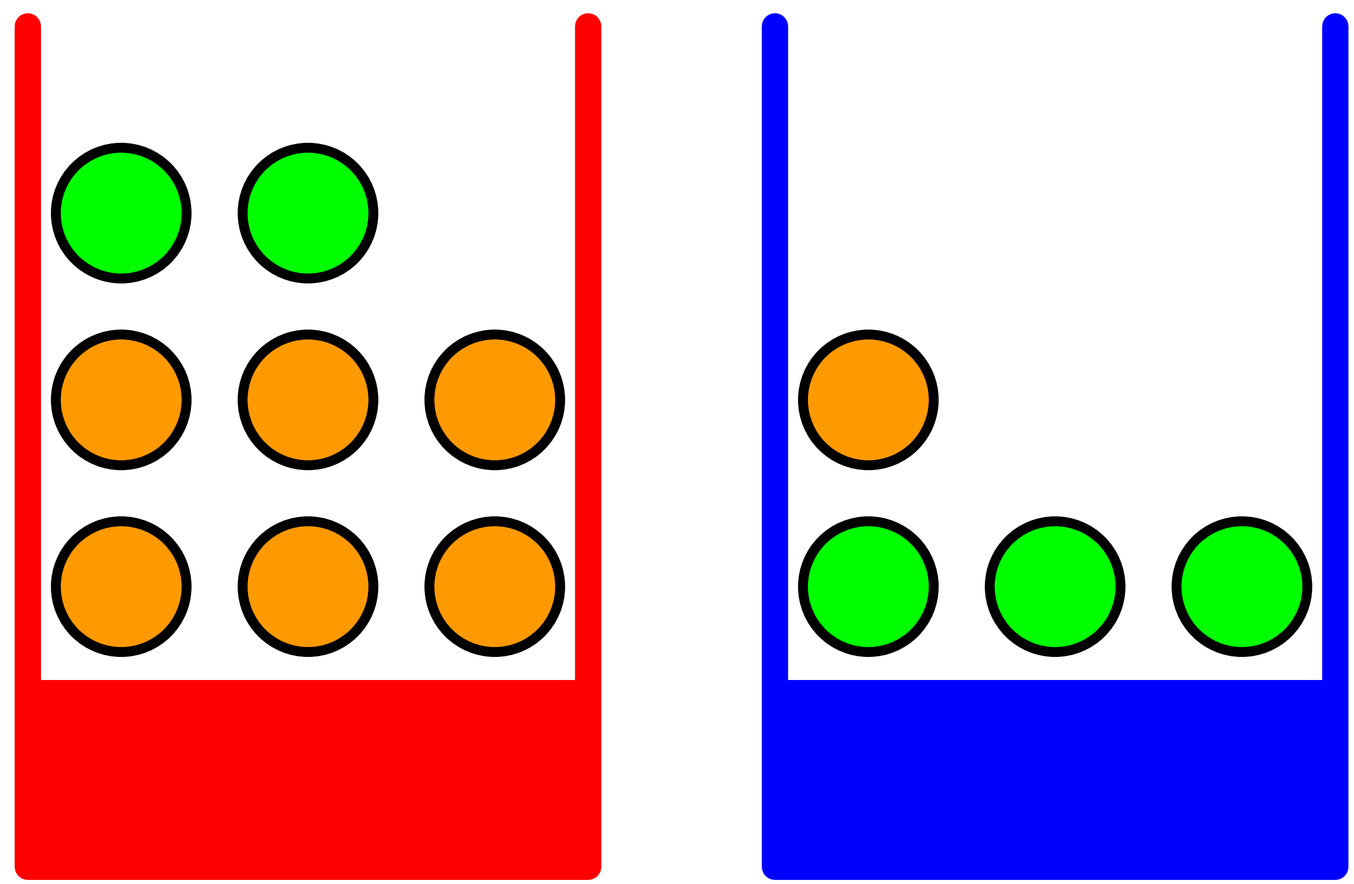
Simple example: discrete probability

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

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

p(red bin) = 2/5
p(apple|red) = 1/4

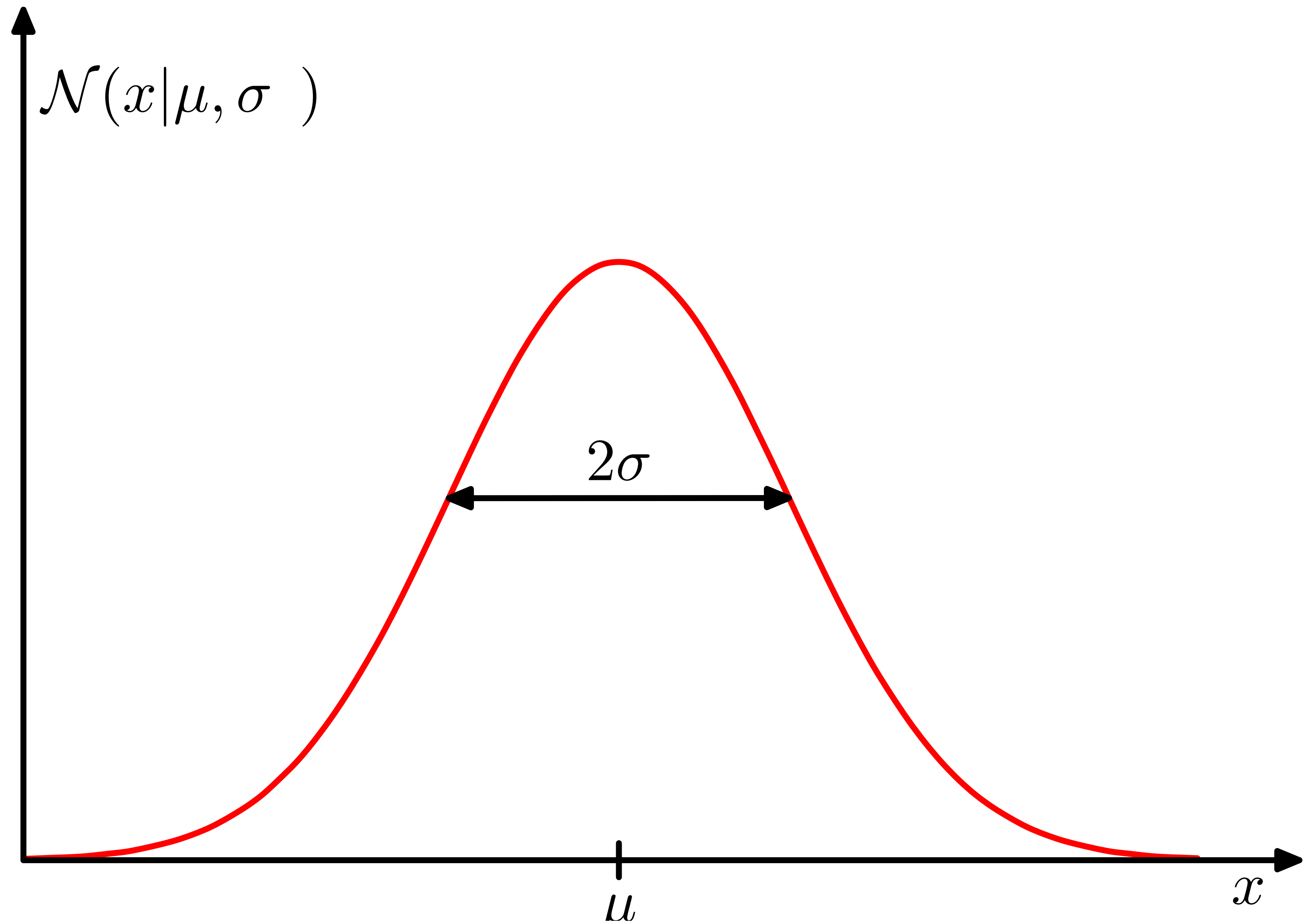
p(blue bin) = 3/5
p(apple|blue) = 3/4



TODO: actually show worked math

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

- Suppose some number *y* is drawn from a normal distribution with s.d. 1, but with an unknown mean
- We suppose the mean is somewhere near zero:

$$\mu \sim \mathcal{N}(\mu|0, 10)$$
$$y|\mu \sim \mathcal{N}(y|\mu, 1)$$

Easy question: what is p(y | μ = 3)?

Hard question: what is p(μ | y = 3)?

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:

$$p(\mu|y=3) = \frac{p(\mu)p(y=3|\mu)}{p(y=3)}$$

Sum rule, in the denominator:

$$p(y=3) = \int p(\mu)p(y=3|\mu)\mathrm{d}\mu$$

How do we really do this?

Bayes' rule — up to an unknown *normalizing constant*

$$p(\mu|y=3) \propto p(\mu)p(y=3|\mu)$$

Write out the joint distribution as a function of μ:

$$p(\mu)p(y=3|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{1}{2}(3-\mu)^2\right\} \times \frac{1}{\sqrt{200\pi}} \exp\left\{\frac{1}{200}(\mu)^2\right\}$$

Now: if you squint at this for a while (combine terms in the exponent, and complete the square…), you will see that this is also a normal distribution — though not normalized

How do we really do this?

$$p(\mu)p(y=3|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{1}{2}(3-\mu)^2\right\} \times \frac{1}{\sqrt{200\pi}} \exp\left\{\frac{1}{200}(\mu)^2\right\}$$

TODO: actually show worked math for completing the square

Show posterior concentration as we add more data

TODO: there needs to be a slide somewhere showing what happens to posterior distributions as you add more data (concentration relative to the prior)

No, but how do we *really* do this?

- In the previous example, we got lucky: because the conditional distribution $p(\mu \mid y = 3)$ turned out to *also be a normal distribution*, we were able to find it analytically.
- **This will rarely be the case.**
- For most interesting models, we turn to Monte-Carlo methods.

Monte Carlo inference

General problem:

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

- 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 | 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:

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

Mini-exercise

TODO: insert some sort of mini-exercise here to make sure people understand expectation intuitively

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)

- A few options include:

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

- From our example with Gaussians, suppose we know how to sample from a normal distribution already.

$$\mu \sim \mathcal{N}(\mu|0, 10)$$
$$y|\mu \sim \mathcal{N}(y|\mu, 1)$$

  We can sample *y* by literally simulating from the generative process: we first draw a sample μ, and then we sample a value of *y*.

- We can use this approach to estimate expectations with respect to *p(y)*.

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 μ and a sample of y. These are drawn together from the joint distribution $p(y, \mu)$.

- To estimate the posterior $p(\mu \mid y = 3)$, we say that μ is a sample from the posterior if its corresponding value $y = 3$.

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

Conditioning via importance sampling

- One option is to sidestep sampling from the posterior *p(μ | y = 3)* entirely, and draw from some proposal distribution q(*μ)* instead.

$$\mathbb{E}[f(x)] = \int f(x)p(x|y)\mathrm{d}x = \int f(x)p(x|y)\frac{q(x)}{q(x)}\mathrm{d}x = \int f(x)W(x)q(x)\mathrm{dx} = \mathbb{E}_q[f(x)W(x)]$$

The idea here is we can now approximate this expectation with *weighted samples* from $q(x)$.

Algorithmically, this works as follows: we define an unnormalized *importance weight* function

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

We then draw samples $x_i \sim q(x)$ for $i = 1, \ldots, N$ and approximate expectations with

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

TODO: make less hideous slide

$$\mathbb{E}[f(x)] \approx \sum_{i=1}^{N} W_i f(x_i)$$

Conditioning via importance sampling

- A we already have very simple proposal distribution we know how to sample from: the prior *p(μ)*.

- 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.

Conditioning via MCMC

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

- 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. Technically, this works by constructing a Markov chain whose stationary distribution is the target distribution we are trying to sample from. For the moment do not worry about *why* MCMC works; first, here is how to implement it algorithmically.

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

- There is a lot of freedom in choosing different sorts of creative proposal distributions, but a simple and typical class of proposal distributions for real-valued latent variables takes a value x and adds a small amount of Gaussian noise along one or more of its dimensions.

TODO: make less hideous slide

Conditioning via MCMC

Assuming we are trying to sample from a posterior distribution $p(x|y) \propto p(x, y)$, we define an *acceptance ratio*

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

After we propose some new value x', we then *accept* it with probability $A(x \to x')$ and "move" to the new position x', otherwise we *reject* it and stay at x.

This entire sequence of values at every iteration (including the duplicated values after a reject step) are jointly a sample from the posterior distribution $p(x|y)$.

If we choose a proposal distribution $q(x'|x)$ that is symmetric, such that $q(x'|x) = q(x|x')$, then the acceptance ratio simplifies to a ratio of the joint distributions using the new and old values of x. A simple intuitive interpretation of the algorithm in this case is as a sort of noisy hill-climbing on $p(x, y)$; "better" values of x' are accepted always, and "worse" values of x' are accepted "sometimes".

TODO: make less hideous slide