# Monte Carlo Estimation

CMPUT 366: Intelligent Systems

P&M §8.6

# Logistics

- Assignment #2 due **Monday, Feb 28 at 11:59pm**
  - Submit via eClass
- Next week is **reading week**
  - No lectures
  - No lab
- *After* reading week (**Mon, Feb 28**), lectures will be **in person**
  - CCIS L1-160

# Recap: Bayesian Learning

- In Bayesian Learning, we learn a **distribution** over models instead of a **single model**
- When the model is **conjugate**, posterior probabilities can be computed **analytically**
  - **Today:** non-conjugate models!
- We can make predictions by **model averaging** to compute the **posterior predictive distribution**

# Lecture Outline

1. Recap & Logistics
2. Prior Distributions as Bias
3. Estimation via Sampling
4. Sampling from Hard-to-Sample Distributions

# Prior Distributions as Bias

- Suppose I'm comparing two models, $\theta_1$ and $\theta_2$ such that

$$\Pr(D \mid \theta_1) = Pr(D \mid \theta_2)$$

- **Question:** Which model has higher **posterior probability** $\Pr(\theta_i \mid D)$?
- Priors are a way of encoding **bias**: they tell use which models to prefer when the data doesn't

# Priors for Pseudocounts

- Recall that when $p(\theta) = \mathrm{Beta}(a, b)$, posterior probability is

$$p(\theta \mid n_1, n_0) = \mathrm{Beta}(a + n_1, b + n_0)$$

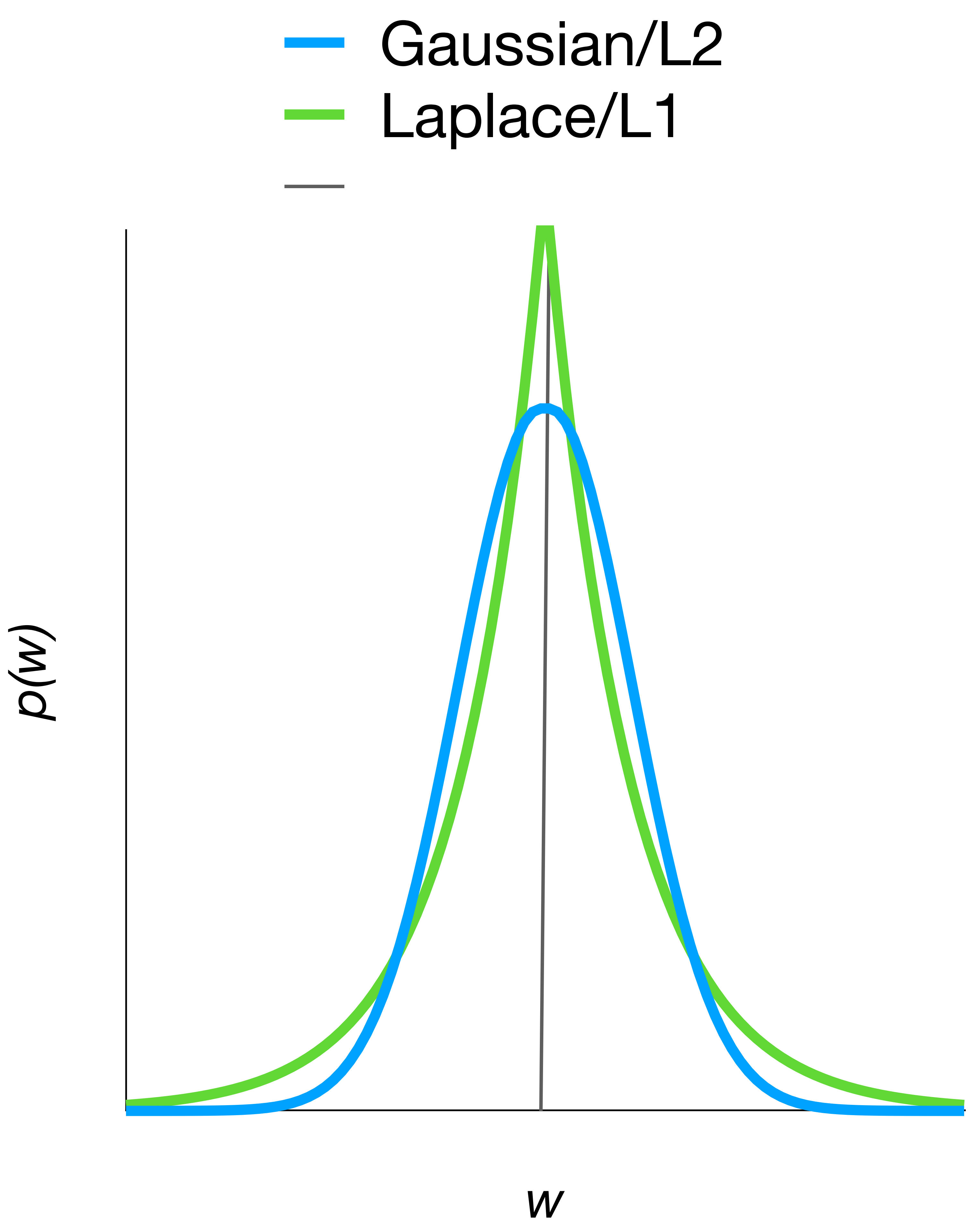
- We can straightforwardly encode **pseudocounts** as prior information in Beta-Binomial and Dirichlet-Multinomial models

- E.g., for pseudocounts $k_1$ and $k_0$,

$$p(\theta) = \mathrm{Beta}(1 + k_1, 1 + k_0)$$

# Priors for Regularization

- Some **regularizers** can be encoded as priors also
- **L2 regularization** is equivalent to a **Gaussian** prior on the weights: $p(w) = \mathcal{N}(w \mid m, s)$
- **L1 regularization** is equivalent to a **Laplacian** prior on the weights: $p(w) = \exp(|w|)/2$



# Estimation via Sampling

- Suppose that we are able to generate independent random **samples** from a random variable $X$
- How can we use those random samples to estimate the **expected value** of $X$?
    - or some function $h$ of $X$; but that in general is just a different random variable $Y = h(X)$
- **Question:** But first, why would we *want* to?

# Estimation from a Sample

**Law of Large Numbers:**

As the number $n$ of independent samples $x_1, x_2, \ldots, x_n$ from a random variable $X$ with distribution $f(x)$ approaches infinity, the **sample average** approaches the **expected value** of $X$.

$$\mathbb{E}[X] = \sum_x f(x)x \approx \frac{1}{n}\sum_{i=1}^{n} x_i$$

Since $Y = h(X)$ is also a random variable, this generalizes to arbitrary **functions** of $X$:

$$\mathbb{E}[h(X)] = \sum_x f(x)h(x) \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

# Probabilities from a Sample

- **Question:** How can we use a sample to estimate the **probability** of a **proposition** $\alpha$?
- Probability of a proposition is just the expectation of its **indicator function**:

$$I_\alpha[x] = \begin{cases} 1 & \text{if } \alpha(x), \\ 0 & \text{otherwise.} \end{cases}$$

- So estimate that expectation as with any other function:

$$\Pr(\alpha) = \mathbb{E}\left(I_\alpha[X]\right) = \sum_x f(x) I_\alpha[x] \approx \frac{1}{n} \sum_x I_\alpha[x].$$

# Probably Approximately Correct

- We never actually have an **infinite** number of sampled values
- How do we know when we have **enough** samples?

**Hoeffding's inequality:**
Suppose $0 \leq X \leq 1$, and $s$ is the sample average from $n$ independent samples from $X$. Then

$$\Pr(|\mathbb{E}[X] - s| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

- For any given **error margin** $\epsilon$ and number of samples $n$, we can plug into this formula and get a **PAC bound**.
  - Can also go the other way: plug in the acceptable error bound to RHS, and derive the **number of samples** $n$ needed
- This generalizes to arbitrary **bounded** random variables $a \leq X \leq b$.

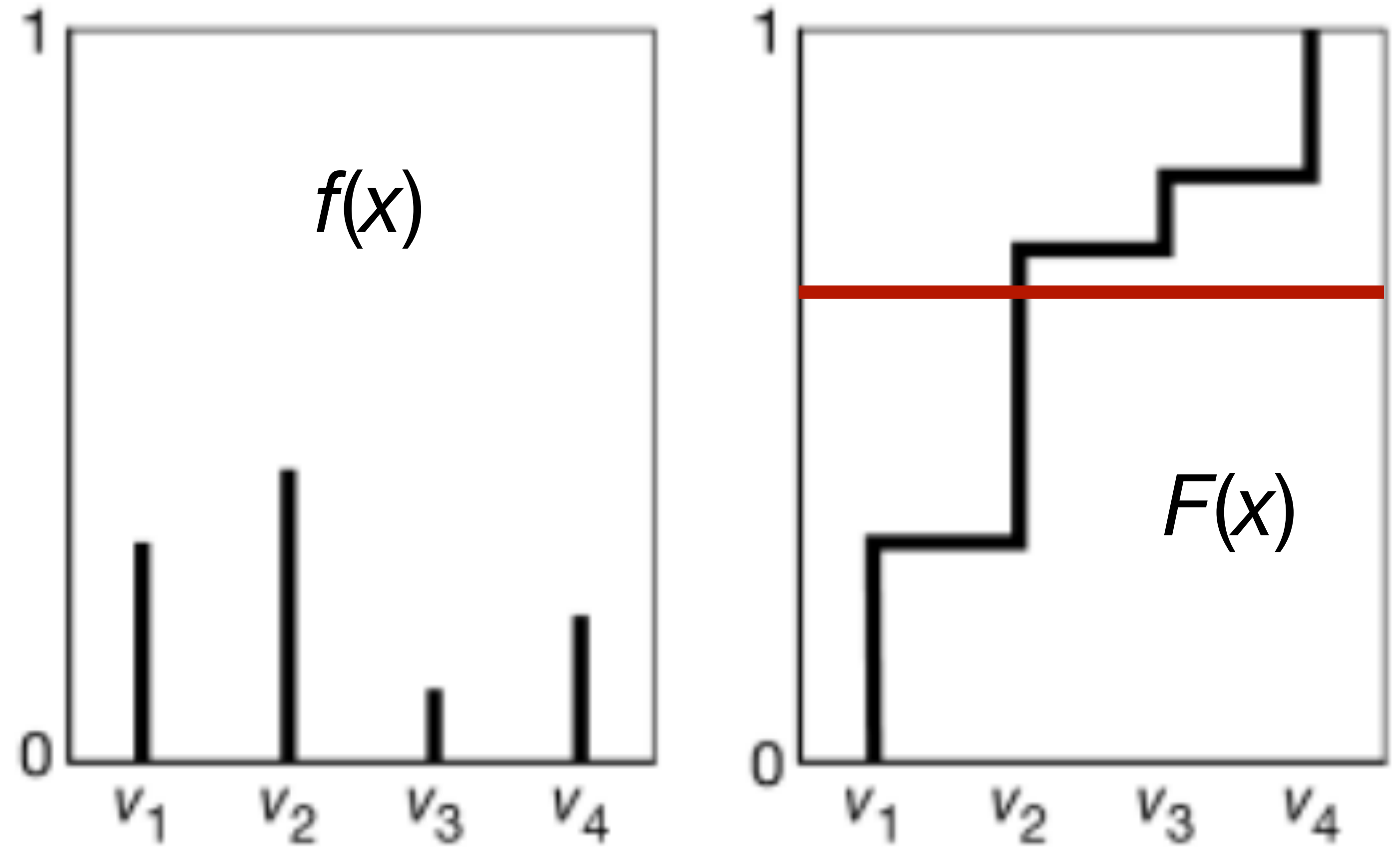
# Generating Samples from a Single Variable

How can we generate samples from a distribution?

1. **Totally order** the domain of the variable
   (can be arbitrary for categorical variables)

2. **Cumulative distribution**: $F(x) = \Pr(X \leq x)$

$$F(x) = \int_{-\infty}^{x} f(z)dz \qquad F(x) = \sum_{x' \leq x} f(x')$$




3. Select a **uniform** random number $y \in [0,1]$

4. Return $x_i = F^{-1}(y)$

# Hard-To-Sample Distributions

Often, we want to sample from distributions that are **hard** to sample from, especially large **joint distributions**

**Question:** Why might a distribution be hard to sample from?

1. Use samples from easier distributions:
    - Rejection Sampling
    - Importance Sampling
2. Go piece by piece through the joint distribution
    - Forward Sampling in a Belief Network
    - Particle Filtering

# Proposal Distributions

- Can we use an **easy-to-sample** distribution $g(x)$ to help us sample from $f(x)$?
  - Very common: We know an **unnormalized** $f^*(x)$, but not the properly normalized distribution $f(x)$:

$$f(x) = \frac{f^*(x)}{\int_{-\infty}^{\infty} f^*(z)dz}$$

- $f(x)$ is the **target distribution**
  - $f^*(x)$ is the **unnormalized target distribution**
- $g(x)$ is the **proposal distribution**

# Rejection Sampling

- Rejection sampling is one way to use a proposal distribution to sample from a target distribution
- *Assumption:* We know a constant $M$ such that

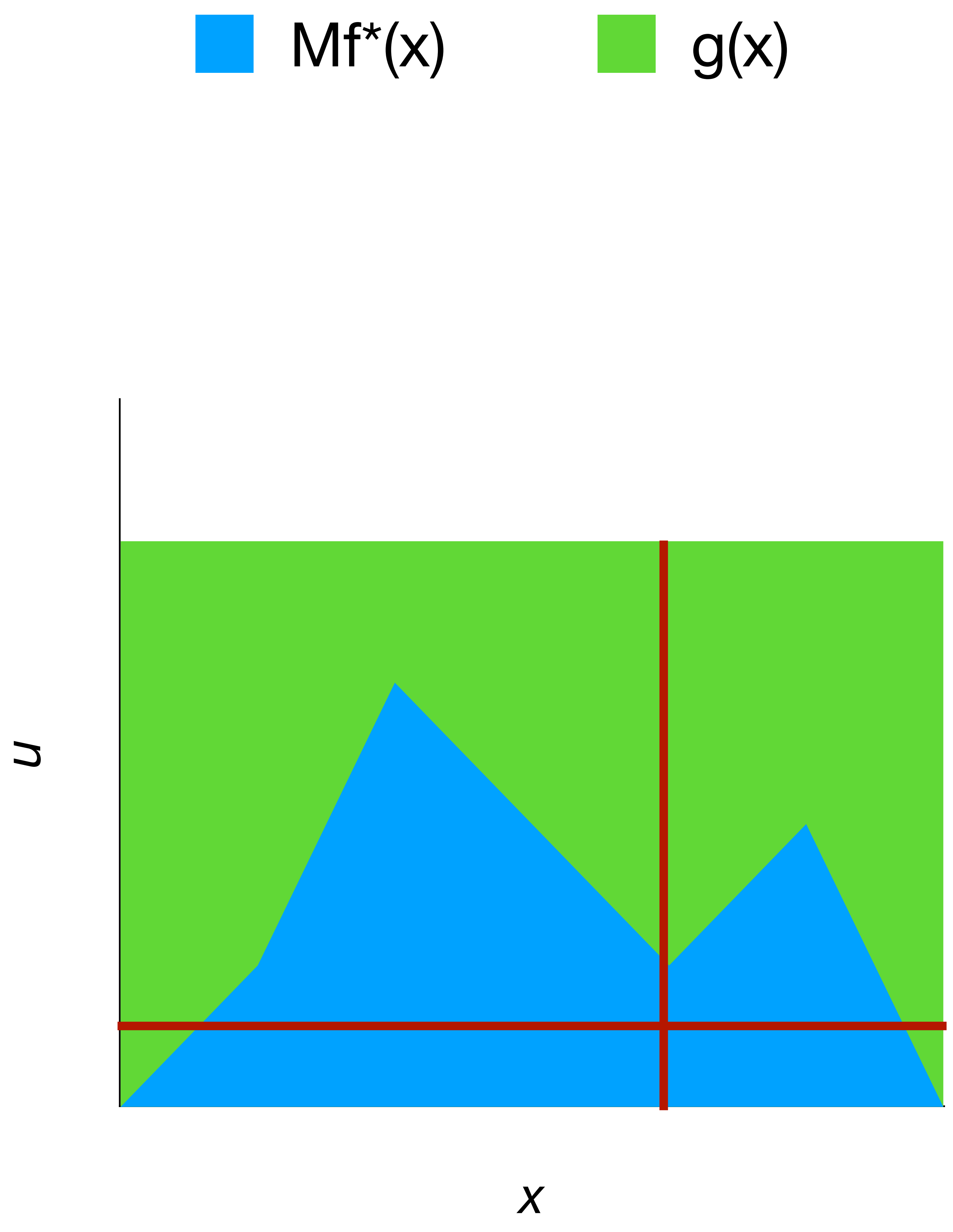
$$\forall x : Mf^*(x) \leq g(x)$$

  - Much **easier** to find $M$ than to find the constant that makes the integral come out to **exactly** 1

- **Repeat** until "enough" samples accepted:

1. **Sample** $x \sim g(x)$ from the **proposal distribution**
2. **Sample** $u \sim \text{Uniform}[0,1]$
3. **If** $u \leq \left[ Mf^*(x) \,/\, g(x) \right]$, **accept** $x$ (add it to samples)

   **Else reject**



# Importance Sampling

- Rejection sampling works, but it can be **wasteful**
  - Lots of samples get rejected when proposal and target distributions are very **different**
- What if we took a **weighted average** instead?
  1. Sample $x_1, x_2, \ldots, x_n$ from $g(x)$
  2. **Weight** each sample $x_i$ by $w_i = \dfrac{Mf^*(x_i)}{g(x_i)}$
  3. **Estimate** is $\dfrac{1}{\sum_j w_j} \sum_{x_i \sim g} w_i x_i$

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x f(x)x \\
&= \sum_x \frac{g(x)}{g(x)} f(x)x \\
&= \sum_x g(x)\frac{f(x)}{g(x)} x \\
&\approx \frac{1}{n}\sum_{x_i \sim g} \frac{f(x_i)}{g(x_i)} x_i
\end{aligned}$$

# Forward Sampling in a Belief Network

- Sometimes we know how to sample **parts** of a large joint distribution in terms of other parts
  - E.g., belief networks: $P(X, Y, Z) = P(X)P(Y)P(Z \mid X, Y)$
  - We might be able to **directly** sample from each **conditional distribution** but not from the **joint distribution**
- Forward sampling:
  1. **Select** an ordering of variables consistent with the factoring
  2. **Repeat** until enough samples generated:

     **For** each variable $X$ in the ordering:

     **Sample** $x_i \sim P(X \mid pa(X))$

# Particle Filtering

- **Forward sampling** generates a value for each variable, then moves on to the next sample
- **Particle filtering** swaps the order:
  - Generate $n$ values for variable $X$, then $n$ values for variable $Y$, etc.
  - Especially useful when there is no fixed number of variables (e.g., in sequential models)
- Each sample is called a **particle**. Update its **weight** each time a value is sampled.
- Periodically **resample** from the particles with replacement, resetting weights to 1
  - High-probability particles likely to be **duplicated**
  - Low-probability particles likely to be **discarded**
- Resampling means the particles cover the distribution better

# Rejection Sampling with Propositions

- How do we condition on some **propositional evidence** $\alpha$?

e.g., $\alpha(x) = (x_1 > 0 \wedge x_4 \leq 12)$

- Repeat until enough samples accepted

  1. **Sample** $x$ from the **full joint distribution**
     (e.g., using **forward sampling** or **particle sampling**)

  2. **If** $\alpha(x)$, then **accept** $x$
     **Else** **reject**

- Another view of this procedure:

  1. **Approximate** the full joint distribution

  2. **Condition** on evidence $\alpha$

# Summary

- Often we **cannot directly estimate** probabilities or expectations from our model
- **Monte Carlo estimates**: Use a **random sample** from the distribution to estimate expectations by sample averages
- Two families of techniques for hard to sample distributions:
  1. Use an easier-to-sample **proposal** distribution instead
  2. Sample parts of the model **sequentially**