SAMPLING ALGORITHMS

Sampling Algorithms

In general

- A **sampling algorithm** is an algorithm that outputs samples $x_1, x_2, \dots$ from a given distribution P or density p.
- Sampling algorithms can for example be used to approximate expectations:

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

Inference in Bayesian models

Suppose we work with a Bayesian model whose posterior Π cannot be computed analytically.

- We will see that it can still be possible to *sample* from Π.
- Doing so, we obtain samples $\theta_1, \theta_2, \dots$ distributed according to Π.
- This reduces posterior estimation to a density estimation problem (i.e. estimate Π from $\theta_1, \theta_2, \dots$).

PREDICTIVE DISTRIBUTIONS

Posterior expectations

If we are only interested in some statistic of the posterior of the form $\mathbb{E}_{\Pi}[f(\Theta)]$ (e.g. the posterior mean $\mathbb{E}_{\Pi}[\Theta]$, we can again approximate by

$$\mathbb{E}_{\Pi}[f(\Theta)] \approx \frac{1}{m}\sum_{i=1}^{m} f(\theta_i) \; .$$

Example: Predictive distribution

The **posterior predictive distribution** is our best guess of what the next data point x_{n+1} looks like, given the posterior under previous observations:

$$p(x_{n+1}|x_1,\ldots,x_n) := \int_{\mathcal{T}} p(x_{n+1}|\theta)\Pi(\theta|x_1,\ldots,x_n)d\theta \; .$$

This is one of the key quantities of interest in Bayesian statistics.

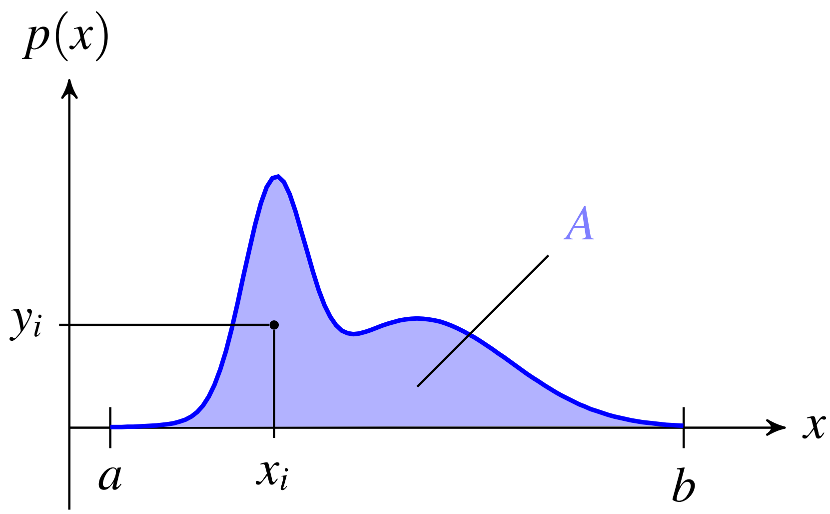
Computation from samples

The predictive is a posterior expectation, and can be approximated as a sample average:

$$p(x_{n+1}|x_{1:n}) = \mathbb{E}_{\Pi}[p(x_{n+1}|\Theta)] \approx \frac{1}{m}\sum_{i=1}^{m} p(x_{n+1}|\theta_i)$$

BASIC SAMPLING: AREA UNDER CURVE

Say we are interested in a probability density p on the interval $[a, b]$.



Key observation

Suppose we can define a uniform distribution U_A on the blue area A under the curve. If we sample

$$(x_1, y_1), (x_2, y_2), \ldots \sim_{\text{iid}} U_A$$

and discard the vertical coordinates y_i, **the x_i are distributed according to p**,

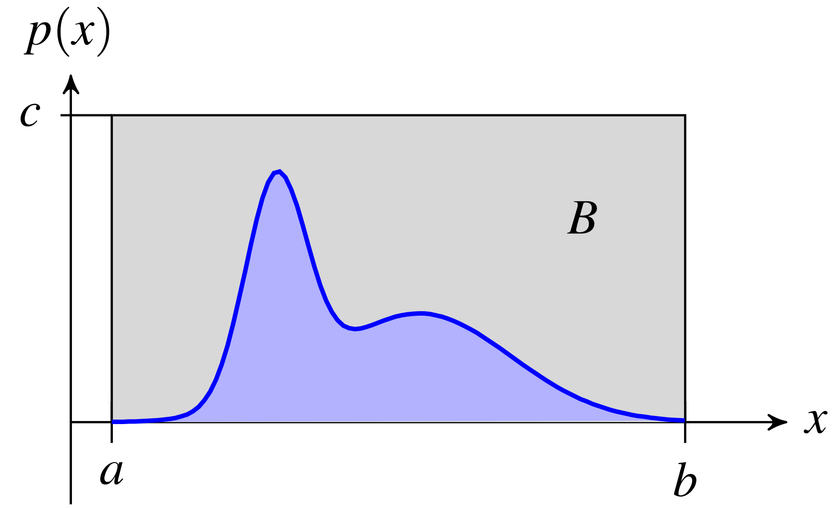
$$x_1, x_2, \ldots \sim_{\text{iid}} p \ .$$

Problem: Defining a uniform distribution is easy on a rectangular area, but difficult on an arbritrarily shaped one.

REJECTION SAMPLING ON THE INTERVAL

Solution: Rejection sampling

We can enclose p in box, and sample uniformly from the box B.



- We can sample (x_i, y_i) uniformly on B by sampling

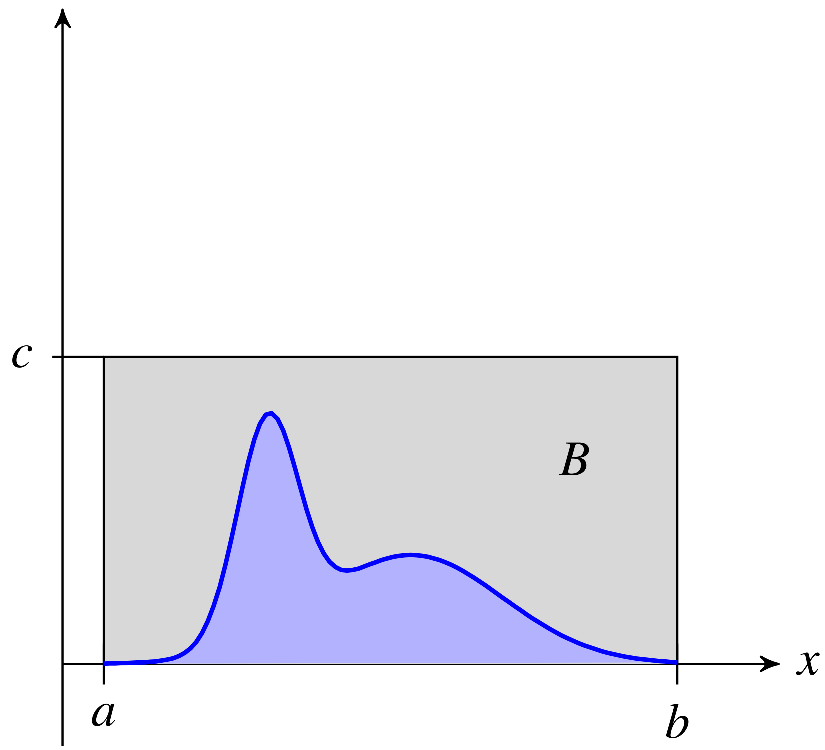
$$x_i \sim \text{Uniform}[a, b] \qquad \text{and} \qquad y_i \sim \text{Uniform}[0, c] \; .$$

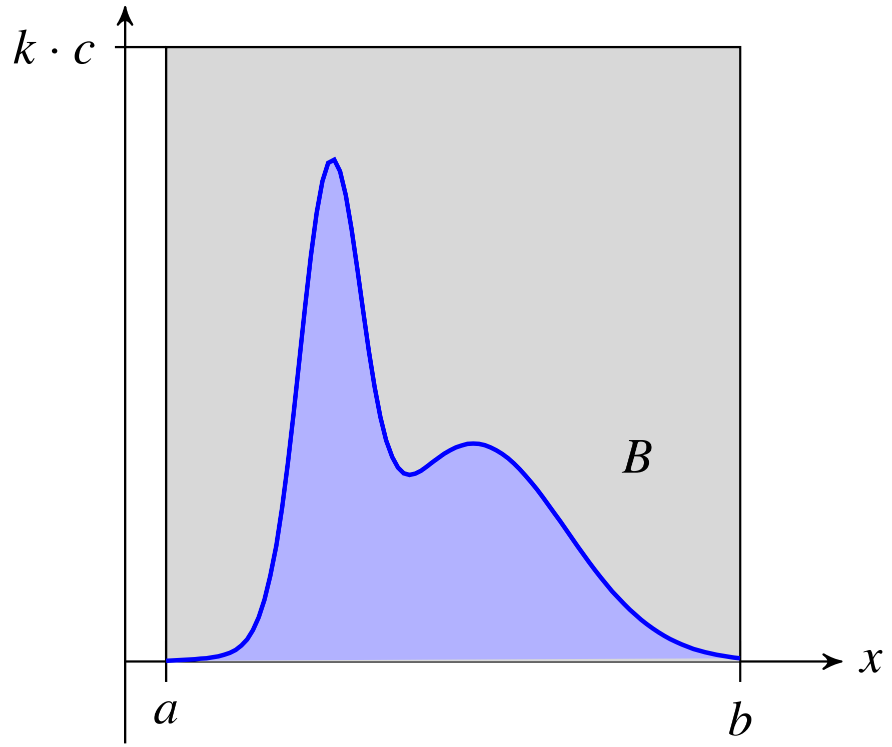
- If $(x_i, y_i) \in A$ (that is: if $y_i \leq p(x_i)$), keep the sample.
  Otherwise: discard it ("reject" it).

Result: The remaining (non-rejected) samples are uniformly distributed on A.

SCALING

This strategy still works if we scale the vertically by some constant $k > 0$:





We simply sample $y_i \sim \text{Uniform}[0, kc]$ instead of $y_i \sim \text{Uniform}[0, c]$.

Consequence

For sampling, it is sufficient if p is known only up to normalization
(i.e. if only the shape of p is known).

Distributions Known up to Scaling

Sampling methods usually assume that we can evaluate the target distribution p up to a constant. That is:

$$p(x) \quad = \quad \frac{1}{\tilde{Z}}\,\tilde{p}(x)\ ,$$

and we can compute $\tilde{p}(x)$ for any given x, but we do not know $\tilde{Z}$.

We have to pause for a moment and convince ourselves that there are useful examples where this assumption holds.

Example 1: Simple posterior

For an arbitrary posterior computed with Bayes' theorem, we could write

$$\Pi(\theta|x_{1:n}) = \frac{\prod_{i=1}^{n} p(x_i|\theta)q(\theta)}{\tilde{Z}} \qquad \text{with} \qquad \tilde{Z} = \int_{\mathcal{T}} \prod_{i=1}^{n} p(x_i|\theta)q(\theta)d\theta\ .$$

Provided that we can compute the numerator, we can sample without computing the normalization integral $\tilde{Z}$.

Example 2: Bayesian Mixture Model

Recall that the posterior of the BMM is (up to normalization):

$$\Pi(c_{1:K}, \theta_{1:K}|x_{1:n}) \propto \prod_{i=1}^{n}\Big(\sum_{k=1}^{K} c_k p(x_i|\theta_k)\Big)\Big(\prod_{k=1}^{K} q(\theta_k|\lambda, y)\Big) q_{\text{Dirichlet}}(c_{1:K})$$

We already know that we can discard the normalization constant, but can we evaluate the non-normalized posterior $\tilde{\Pi}$?

- The problem with computing $\tilde{\Pi}$ (as a function of unknowns) is that the term $\prod_{i=1}^{n}\Big(\sum_{k=1}^{K} \dots\Big)$ blows up into K^n individual terms.
- If we *evaluate* $\tilde{\Pi}$ for specific values of c, x and θ, $\sum_{k=1}^{K} c_k p(x_i|\theta_k)$ collapses to a single number for each x_i, and we just have to multiply those n numbers.

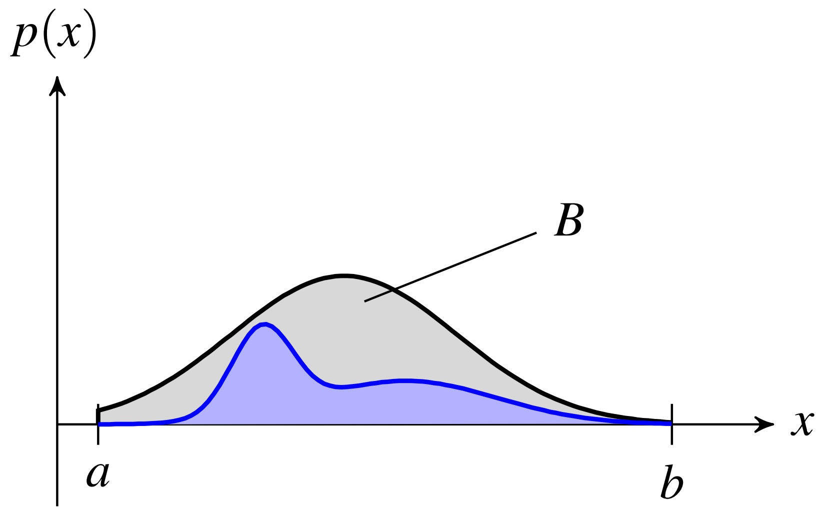
So: Computing $\tilde{\Pi}$ as a formula in terms of unknowns is difficult; evaluating it for specific values of the arguments is easy.

Rejection Sampling on $\mathbb{R}^d$

If we are not on the interval, sampling uniformly from an enclosing box is not possible (since there is no uniform distribution on all of $\mathbb{R}$ or $\mathbb{R}^d$).

Solution: Proposal density

Instead of a box, we use *another distribution* q to enclose p:

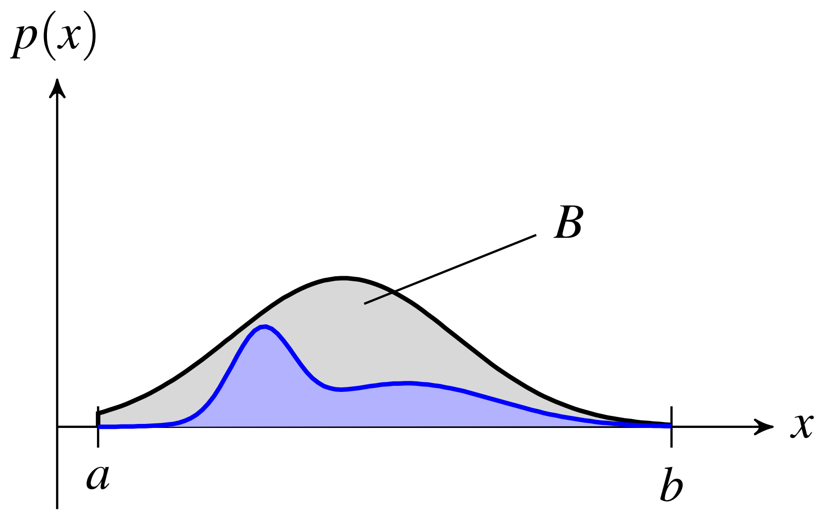


To generate B under q, we apply similar logic as before backwards:

- Sample $x_i \sim q$.
- Sample $y_i \sim \text{Uniform}[0, q(x_i)]$.

q is always a simple distribution which we can sample and evaluate.

REJECTION SAMPLING ON $\mathbb{R}^d$



- Choose a simple distribution q from which we know how to sample.
- Scale $\tilde{p}$ such that $\tilde{p}(x) < q(x)$ everywhere.
- Sampling: For $i = 1, 2, \ldots,$:
  1. Sample $x_i \sim q$.
  2. Sample $y_i \sim \text{Uniform}[0, q(x_i)]$.
  3. If $y_i < \tilde{p}(x_i)$, keep x_i.
  4. Else, discard x_i and start again at (1).
- The surviving samples $x_1, x_2, \ldots$ are distributed according to p.

Factorization Perspective

The rejection step can be interpreted in terms of probabilities and densities.

Factorization

We factorize the target distribution or density p as

distribution from which we know how to sample

$$p(x) \quad = \quad q(x) \cdot A(x)$$

probability function we can evaluate *once a specific value of x is given*

Sampling from the factorization

$$\text{sampling } x \text{ from } p \quad = \quad \text{sampling } x \text{ from } q + \text{coin flip with probability } A(x)$$

By "coin flip", we mean a binary variable with $\Pr(1) = A(x)$ (ie a Bernoulli variable).

Sampling Bernoulli variables with uniform variables

$$Z \sim \text{Bernoulli}(A(x)) \qquad \Leftrightarrow \qquad Z = \mathbb{I}\{U < A(x)\} \quad \text{where} \quad U \sim \text{Uniform}[0,1] \; .$$

INDEPENDENCE

If we draw proposal samples x_i i.i.d. from q, the resulting sequence of accepted samples produced by rejection sampling is again i.i.d. with distribution p. Hence:

Rejection samplers produce i.i.d. sequences of samples.

Important consequence

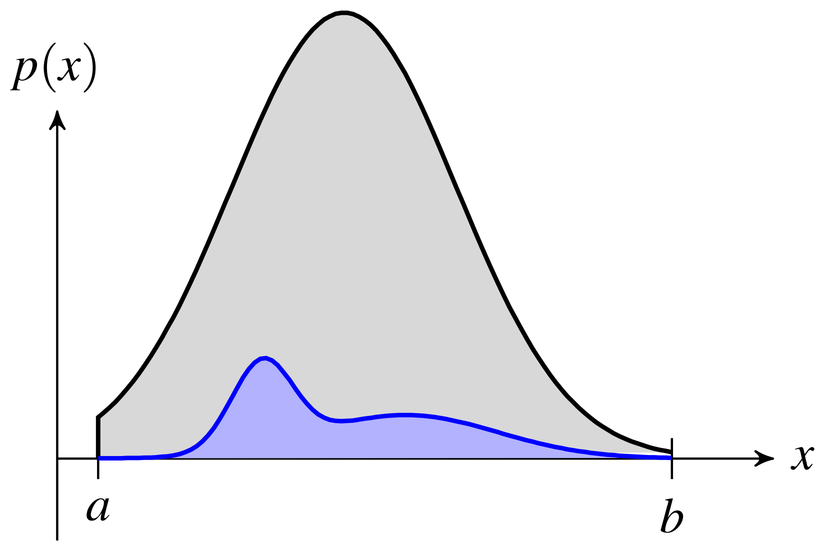
If samples $x_1, x_2, \dots$ are drawn by a rejection sampler, the sample average

$$\frac{1}{m}\sum_{i=1}^{m} f(x_i)$$

(for some function f) is an unbiased estimate of the expectation $\mathbb{E}_p[f(X)]$.

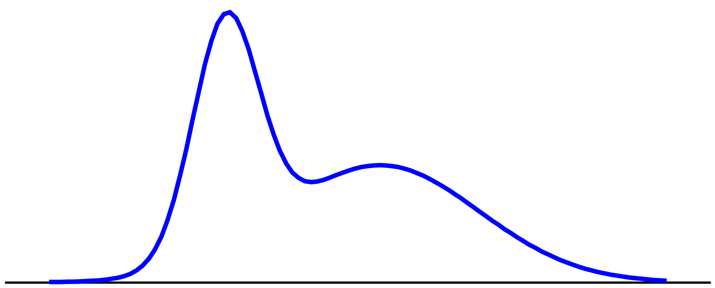
EFFICIENCY

The fraction of accepted samples is the ratio $\frac{|A|}{|B|}$ of the areas under the curves $\tilde{p}$ and q.

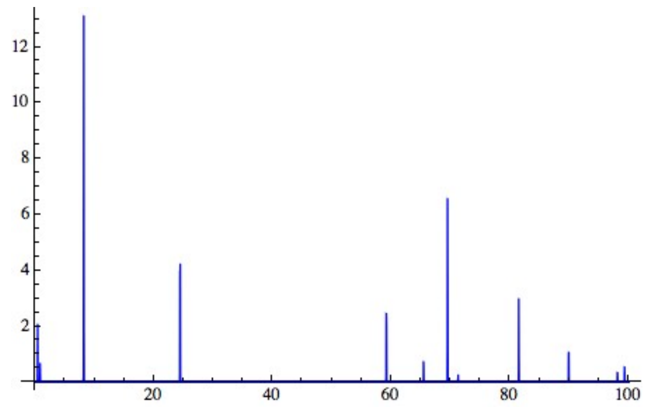


If q is not a reasonably close approximation of p, we will end up rejecting a lot of proposal samples.

AN IMPORTANT BIT OF IMPRECISE INTUITION



Example figures for sampling methods tend to look like this.



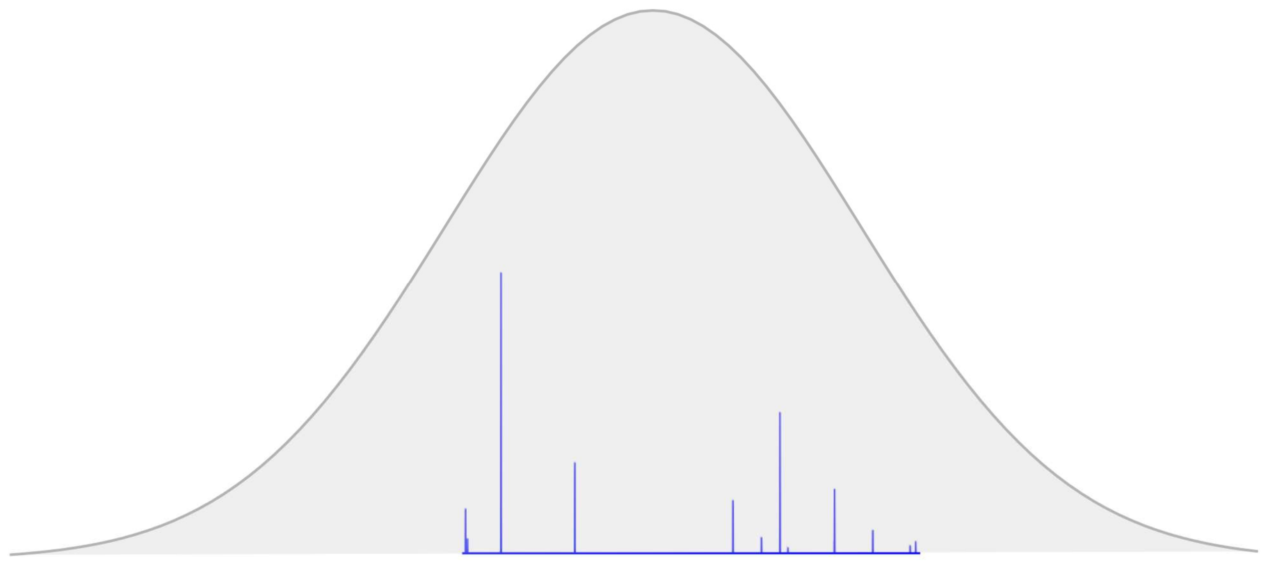
A high-dimensional distribution of correlated RVs will look rather more like this.

Sampling is usually used in multiple dimensions. Reason, roughly speaking:

- Intractable posterior distributions arise when there are several *interacting* random variables. The interactions make the joint distribution complicated.
- In one-dimensional problems (1 RV), we can usually compute the posterior analytically.
- Independent multi-dimensional distributions factorize and reduce to one-dimensional case.

Warning: Never (!!!) use sampling if you can solve analytically.

WHY IS NOT EVERY SAMPLER A REJECTION SAMPLER?



We can easily end up in situations where we accept only one in 10^6 (or 10^{10}, or 10^{20},...) proposal samples. Especially in higher dimensions, we have to expect this to be not the exception but the rule.