

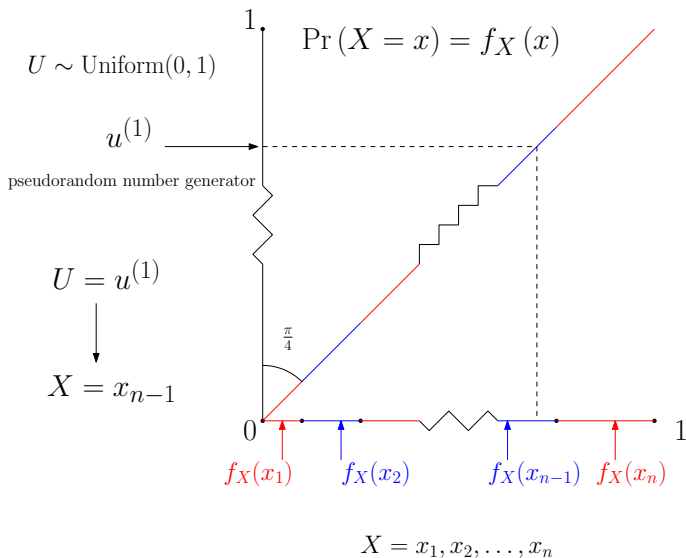
VE414 Lecture 9

Jing Liu

UM-SJTU Joint Institute

October 8, 2019

- Note for discrete random variables with a finite support, sampling is easy:



Q: Suppose we can sample directly from any standard distributions, i.e. those in Appendix 2. How to sample from other distributions with a known cdf?

Theorem

Suppose $U \sim \text{Uniform}(0, 1)$ and F is a one-dimensional cumulative distribution, then $X = F^{-1}(U)$ has the distribution defined by F , where

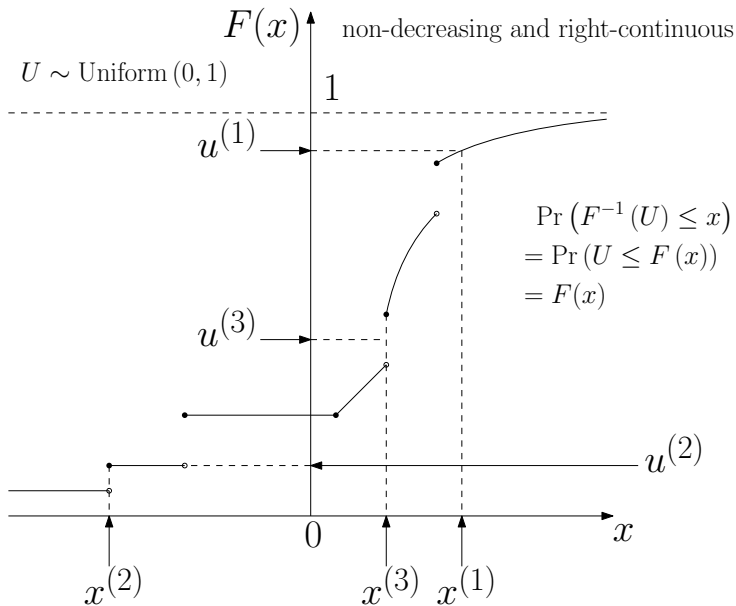
$$F^{-1}(u) = \inf\{x: F(x) \geq u\}$$

- Recall $\inf(\mathcal{S})$ denote the greatest lower bound of \mathcal{S} .
- For continuous and strictly increasing F , it is the value of x such that

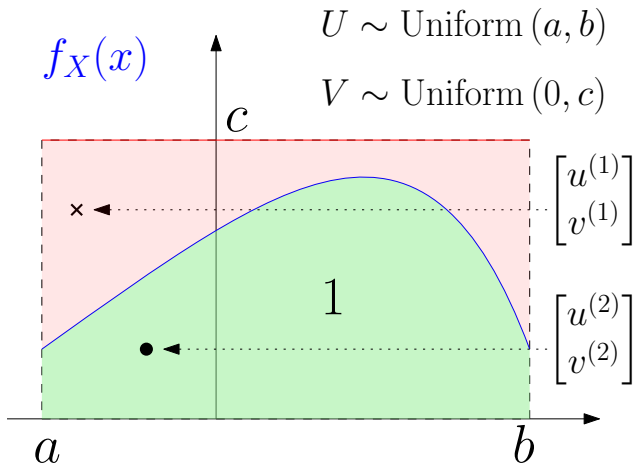
$$F(x) = u$$

- When there is discontinuity, it is the smallest “value” of x such that

$$F(x) \geq u$$

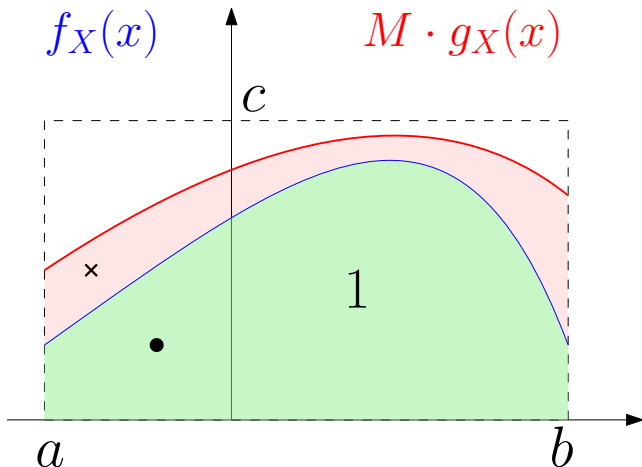


Q: How about a continuous distribution that we have its pdf but not its cdf?



Q: Why is this random sampling scheme more efficient when $bc - ac$ is small?

- The smaller the red region the better since samples from it are wasted.



Q: How to reduce the rejections given a constant $M \in (1, \infty)$ and a pdf g_X ?

Q: How about a distribution over any \mathcal{S} that we have its pdf up to a constant?

- Let $f_Y(y)$ be a pdf up to a multiplicative constant, i.e. A is **unknown**.

$$q_Y(y) = Af_Y(y)$$

where f_Y is the distribution from which we want to sample from.

- Suppose we have a constant $1 < M < \infty$ and a computable distribution

$$g_Y$$

that has the same support \mathcal{S} as f_Y such that

$$q_Y(y) \leq M \cdot g_Y(y) \quad \text{for all } y \in \mathcal{S}$$

then we can sample from the distribution f_Y using **rejection sampling** via the distribution g_Y if we have a way to sample from g_Y . The distribution f_Y is known as the **target distribution**, g_Y is known as a **proposal distribution**.

- Let α denote the event that

$$V \leq \frac{f_Y(U)}{M \cdot g_Y(U)}$$

where $U \sim g_Y$ and $V \sim \text{Uniform}(0, 1)$, thus

$$f_{U,V}(u, v) = \begin{cases} g_Y(u) \cdot 1, & \text{for } (u, v) \in \mathcal{S} \times [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

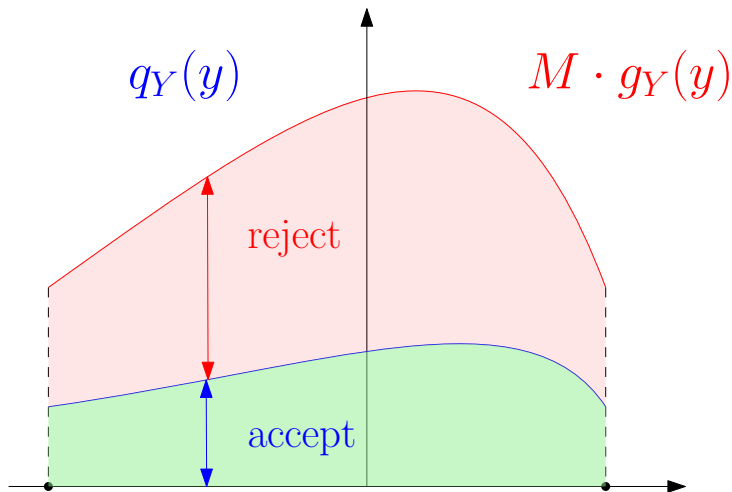
- Let U^* denote the random variable U when α has happened.

$$F_{U^*}(u^*) = \frac{\int_{-\infty}^{u^*} \int_{-\infty}^{q_Y(u)/(M \cdot g_Y(u))} f_{U,V}(u, v) dv du}{\frac{1}{M} \int_{-\infty}^{\infty} q_Y(u) du} = \frac{\int_{-\infty}^{u^*} q_Y(u) du}{A}$$

$$\implies f_{U^*}(u^*) = f_Y(u^*)$$

from which we can conclude the random samples from this scheme follow the distribution defined by the pdf f_Y .

- Again the rejection-acceptance ratio depends on how small the red region is.



Algorithm 1: REJECTION SAMPLING

Input : functions $g_Y(y)$, $q_Y(y)$, constant M , number of samples n

Output : sample array $[y_i]$

```
1 Function rejection( $g_Y(y)$ ,  $q_Y(y)$   $M$ , and  $n$ ):
2    $i \leftarrow 0$ ;
3   while  $i \neq n$  do
4      $v \sim \text{Uniform}(0, 1)$  ;                                /* draw uniform */
5      $y \sim g_Y$  ;                                           /* draw from  $g_Y$  */
6     if  $v \leq \frac{q_Y(y)}{Mg_Y(y)}$  then
7        $i \leftarrow i + 1$ ;
8        $y_i \leftarrow y$ ;
9     end if
10  end while
11  return  $[y_i]$  ;                                           /* samples */
12 end
```

Q: How to sample from a truncated normal distribution, truncated at $c = 0$,

$$f_Y(y) \propto I_{y>c} \phi_Y(y)$$

where $\phi_Y(y)$ is the standard normal pdf and $I_{y>c}$ is the indicator function.

- Assuming we can sample from ϕ_Y , we take a sample from ϕ_Y and simply retain the portion of the sample that is bigger than zero.
- However, it will become inefficient if the truncation c is pushed to the right.
- It is better to use a distribution over $(0, \infty)$ that has long tail if c is big, e.g.

$$g_Y(y) = \lambda \exp(-\lambda y)$$

Q: Which exponential distribution should we use? And how about M ?

- We want to choose the smallest M such that

$$\frac{\phi(y)}{1 - \Phi(c)} \leq M \lambda e^{\lambda y} \quad \text{for all } y \geq c$$

once M is available λ should be chosen to maximum $\Pr(\alpha)$.

- Suppose we are interested in the following expectation,

$$\mu = \mathbb{E}[h(Y)] = \int_{-\infty}^{\infty} h(y) f_Y(y) dy = \int_{\mathcal{D}} h(y) f_Y(y) dy$$

where the target distribution $f_Y(y)$ is cannot be sampled directly.

Q: Given what we have learn so far, what would you do to estimate μ ?

- Suppose we have a distribution

$$g_Y(y)$$

from what we can sample from over \mathcal{D} , then

$$\begin{aligned} \mu &= \int_{\mathcal{D}} h(y) f_Y(y) dy = \int_{\mathcal{D}} h(y) f_Y(y) \frac{g_Y(y)}{g_Y(y)} dy \\ &= \int_{\mathcal{D}} h(y) \frac{f_Y(y)}{g_Y(y)} g_Y(y) dy = \mathbb{E} \left[h(y) \frac{f_Y(y)}{g_Y(y)} \right] \end{aligned}$$

Q: Why is this useful?

- The importance sampling estimate of

$$\mu = \mathbb{E}[h(Y)]$$

is given by the following evaluated at samples $Y_i \sim g_Y$,

$$\hat{\mu}_g = \frac{1}{n} \sum_{i=1}^n h(y_i) \frac{f_Y(y_i)}{g_Y(y_i)}$$

which means $h(y)$ and $f_Y(y)$ as well as $g_Y(y)$ must be computable.

Q: What happens if $g_Y(y) = 0$ for some $y \in \mathcal{D}$?

- The proposal distribution g_Y does not have to be positive everywhere in \mathcal{D} ,

$$g_Y(y) > 0 \quad \text{whenever} \quad h(y)f_Y(y) \neq 0.$$

is sufficient for it to work.

- In practice, the value y^* will not occur in our sample if $g_Y(y^*) = 0$.

Q: How to use importance sampling if $f_Y(y)$ is only known up to a constant?

- If $f_Y(y)$ is not computable, i.e. A is not available

$$q_Y(y) = A f_Y(y)$$

then we have to rely on the following ratios

$$\mu = \frac{\int_{\mathcal{D}} h(y) q_Y(y) dy}{\int_{\mathcal{D}} q_Y(y) dy} = \frac{\int_{\mathcal{D}} h(y) \frac{q_Y(y)}{g_Y(y)} g_Y(y) dy}{\int_{\mathcal{D}} \frac{q_Y(y)}{g_Y(y)} g_Y(y) dy}$$

which can be estimated according to the basic concept of Monte Carlo using

$$\frac{\frac{1}{n} \sum_{i=1}^n h(y_i) \frac{q_Y(y_i)}{g_Y(y_i)}}{\frac{1}{n} \sum_{i=1}^n \frac{q_Y(y_i)}{g_Y(y_i)}} = \frac{\sum_{i=1}^n h(y_i) w_i}{\sum_{i=1}^n w_i} \quad \text{where} \quad w_i = \frac{q_Y(y_i)}{g_Y(y_i)}$$

are called **importance ratios**.

- If $g_Y(y)$ is chosen so that the following function is roughly constant,

$$h(y) \frac{q_Y(y)}{g_Y(y)}$$

then obtaining fairly precise estimates requires fewer samples than otherwise.

- Importance sampling is not reliable if the importance ratios vary a lot

$$w_i = \frac{q_Y(y_i)}{g_Y(y_i)}$$

- The worst case is when w_i and $g_Y(y_i)$ go in opposite direction, if so, we should really try some other proposal distribution.
- Plotting $\ln(w_i)$ with $g_Y(y_i)$, and comparing individual w_i with its average are traditional methods of assessing whether the estimates are poor.
- Avoid using the importance sampling estimates with a large number of small w_i with a few really big w_i .

- Importance sampling and rejection sampling are quite similar ideas. Both of them distort a sample from one distribution in order to sample from another.

$$\text{Rejection sampling:} \quad V \leq \frac{q_Y(Y)}{M g_Y(Y)} \quad \text{Explicitly}$$

$$\text{Importance sampling:} \quad w(Y) = \frac{q_Y(Y)}{g_Y(Y)} \quad \text{Implicitly}$$

- The difference can be best understood in terms of the trade-off and the type of problems that the two methods are usually used and are good at.
- Although rejection sampling provides more, it becomes inefficient when the target distribution becomes complicated, especially as the dimension grows.
- Let $\mathbf{Y} = \{Y_1, Y_2, \dots\}$ be a **random process** of potentially ∞ -dimension with

$$f_{\mathbf{Y}}(\mathbf{Y}) = \prod_{j \geq 1}^k f_j(y_j \mid y_1, \dots, y_{j-1}) \quad \text{where} \quad f_1(y_1)$$

where K is a random variable over positive integers.

- Even given some appropriate proposal distribution that we can sample from

$$g_{\mathbf{Y}}(\mathbf{Y}) = \prod_{j=1}^k g_j(y_j \mid y_1, \dots, y_{j-1}) \quad \text{where } g_1(y_1)$$

it is not clearly what M should we use in the rejection sampling

$$V \leq \frac{f_{\mathbf{Y}}(\mathbf{Y})}{M g_{\mathbf{Y}}(\mathbf{Y})}$$

since the exact pdfs are not simple, despite being simple enough to compute.

- Using importance sampling, we simply estimate

$$\mu = \mathbb{E}[h(\mathbf{Y})] \quad \text{by} \quad \hat{\mu}_g = \frac{1}{n} \sum_{i=1}^n h(\mathbf{y}_i) w_{k_i}$$

where the set of vectors $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ is a sample from $g_{\mathbf{Y}}$ and

$$w_{k_i} = \prod_{j=1}^{k_i} \frac{f_j(y_j \mid y_1, \dots, y_{j-1})}{g_j(y_j \mid y_1, \dots, y_{j-1})}$$