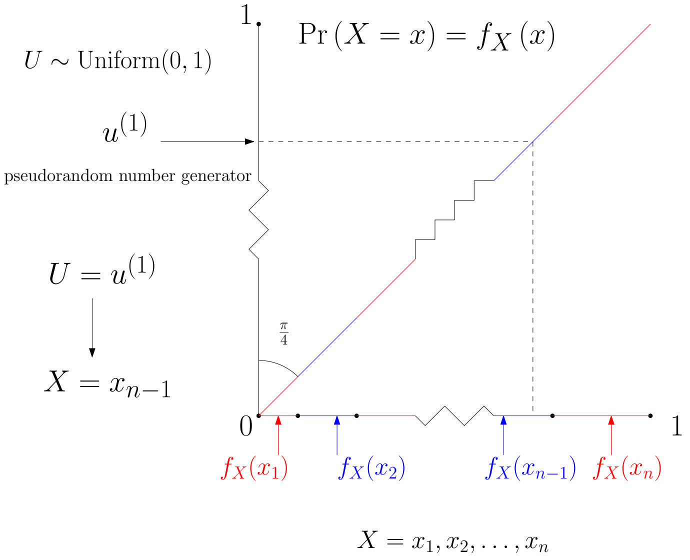
VE414 Lecture 9

Jing Liu

UM-SJTU Joint Institute

June 11, 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 3. How to sample from other distributions with a known cdf/pdf?

Theorem

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

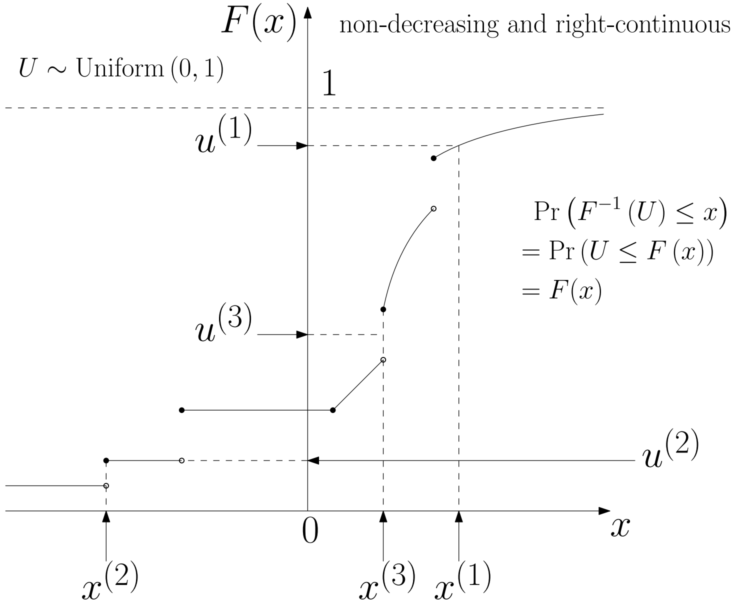
$$F^{-1}(u) = \inf\{x\colon F(x) \geq u\}$$

- 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

$$F(x) = u$$

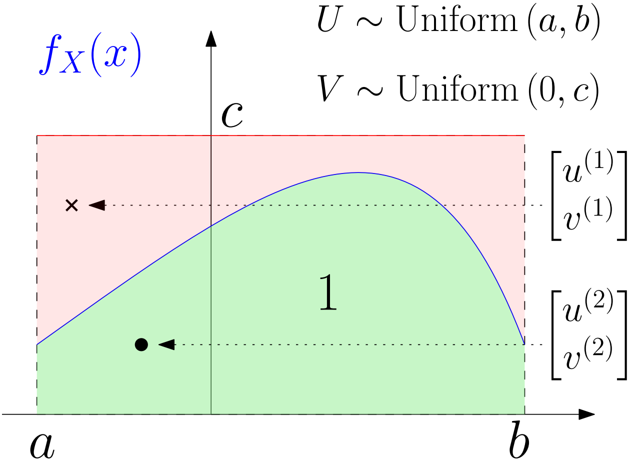
- When there is discontinuity, it is the smallest "value" of x such that

$$F(x) \geq u$$



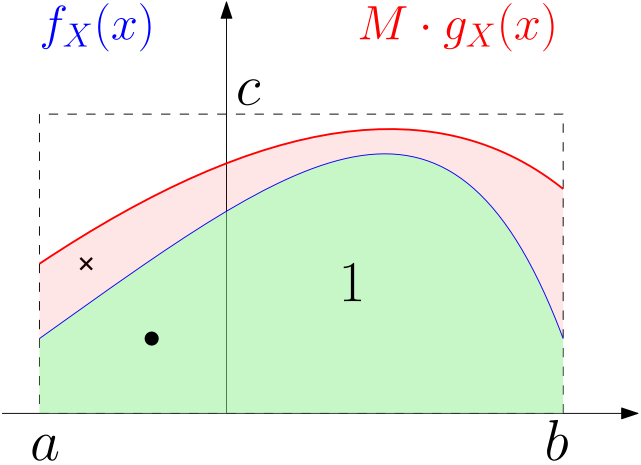
$F(x)$
non-decreasing and right-continuous
$U \sim \text{Uniform}\,(0,1)$
1
$u^{(1)}$
$\Pr\left(F^{-1}(U) \le x\right)$
$= \Pr\left(U \le F(x)\right)$
$= F(x)$
$u^{(3)}$
$u^{(2)}$
0
x
$x^{(2)}$
$x^{(3)}$
$x^{(1)}$

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 more 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 continuous distribution 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

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

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

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

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

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

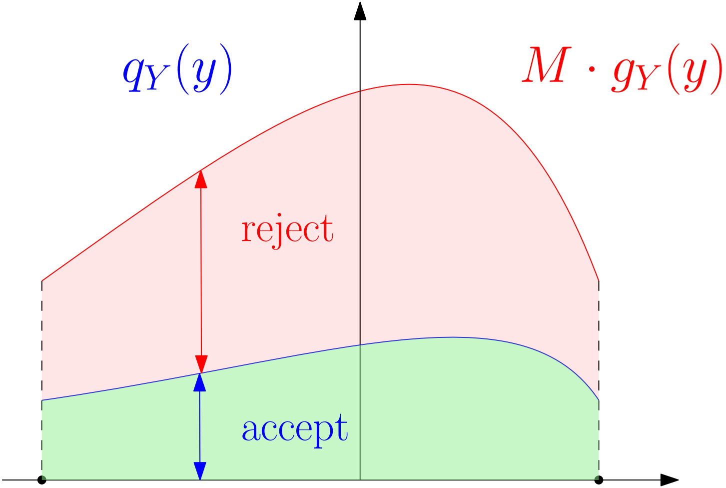
- Let U^* denote the random variable U when α has happened.

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

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

from which we can conclude the random samples generated from the last sampling 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 $[y_i]$

```
Function rejection(g_Y(y), q_Y(y) M, and n):
    i ← 0;
    while i ≠ n do
        v ~ Uniform(0, 1) ;                          /* draw uniform */
        y ~ g_Y ;                                    /* draw from g_Y */
        if v ≤ q_Y(y) / (M g_Y(y)) then
            i ← i + 1;
            y_i ← y;
        end if
    end while
    return [y_i] ;                                   /* samples */
end
```

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

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

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.

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

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

- We want to choose the smallest M such that

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

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

- Suppose we are interested in the following expectation,

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

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

$$\mu = \mathrm{E}\left[h(Y)\right]$$

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 $\mathcal{D}$,

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

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 reply on the following ratios

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

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

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

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.

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

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

- 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 $\mathbf{Y} = \{Y_1, Y_2, \ldots\}$ be a random process of potentially ∞-dimension with

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

where K is a random variable over positive integers.

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

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

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

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

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

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