# Chapter 3

# Background Theory

## 3.1 Review of Probability

### 3.1.1 Expected Value

Suppose that there is one random variable $x$ whose *PDF* (Probability Density Function)/*PMF* (Probability Mass Function) is $\Pr(x)$ , and *probability space* is $S$. The expected value of a function $f$ is defined as:

$$\mathrm{E}_{\Pr(x)}\left[f(x)\right] = \begin{cases} \sum_{x_i \in S} \Pr(x_i) \cdot f(x_i) & \text{for discrete} \\ \int_{x \in S} \Pr(x) \cdot f(x) \cdot dx & \text{for continuous} \end{cases} \tag{3.1}$$

One special case of expected value is $\mu$, the *mean* of a random variable $x$:

$$\mu = \overline{\mathrm{E}_{\Pr(x)}\left[x\right]} = \begin{cases} \sum_{x_i \in S} \Pr(x_i) \cdot x_i & \text{for discrete} \\ \int_{x \in S} \Pr(x) \cdot x \cdot dx & \text{for continuous} \end{cases} \tag{3.2}$$

### 3.1.2 Variance

Given a random variable $x$ with expected value $\mathrm{E}_{\Pr(x)}\left[f(x)\right]$ of function $f$ under probability distribution $\Pr(x)$, the *variance* of function $f$ is defined as:

$$\mathrm{Var}_{\Pr(x)}\left[f(x)\right] = \mathrm{E}_{\Pr(x)}\left[(f(x) - \mathrm{E}_{\Pr(x)}\left[f(x)\right])^2\right] \tag{3.3}$$

The equation above can be expended:

$$\begin{aligned}
\mathrm{Var}_{\Pr(x)}\left[f(x)\right] &= \mathrm{E}_{\Pr(x)}\left[(f(x) - \mathrm{E}_{\Pr(x)}\left[f(x)\right])^2\right] \\
&= \mathrm{E}_{\Pr(x)}\left[f(x)^2 - 2 \cdot f(x) \cdot \mathrm{E}_{\Pr(x)}\left[f(x)\right] + (\mathrm{E}_{\Pr(x)}\left[f(x)\right])^2\right] \\
&= \mathrm{E}_{\Pr(x)}\left[f(x)^2\right] - 2 \cdot \mathrm{E}_{\Pr(x)}\left[f(x)\right] \cdot \mathrm{E}_{\Pr(x)}\left[f(x)\right] + (\mathrm{E}_{\Pr(x)}\left[f(x)\right])^2 \\
&= \mathrm{E}_{\Pr(x)}\left[f(x)^2\right] - (\mathrm{E}_{\Pr(x)}\left[f(x)\right])^2
\end{aligned} \tag{3.4}$$

### 3.1.3 Indicator Function and Dirac Function

#### Indicator Function

Definition: $Ind_A : X \to \{0,1\}$

$$Ind_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{3.5}$$

The indicator function defined on set $X$, is used to indicate whether an element $x$, $x \in X$, is in a subset $A$ of $X$ or not, where $A \subseteq X$. If the domain $X$ is a *probability space* with *PDF* (Probability Density Function)/ PMF (Probability Mass Function) $\Pr(x)$, and $A$ is a *measurable set*, then the expected value of indicator function $Ind_A$ can be derived by applying $Ind_A$ to Eq. 3.1:

$$\mathrm{E}_{\Pr(x)}\left[Ind_A(x)\right] = \begin{cases} \sum_{x \in X} \Pr(x) \cdot Ind_A(x) = \sum_{x \in A} \Pr(x) = \Pr(x \in A) & \text{for discrete} \\ \int_{x \in X} \Pr(x) \cdot Ind_A(x) \cdot dx \int_{x \in A} \Pr(x) \cdot dx = \Pr(x \in A) & \text{for continuous} \end{cases} \tag{3.6}$$

#### Dirac Function

#### Dirac Measure

Dirac measure $\delta_x$ is a probability measure on a set $X$ defined for a given $x$, $x \in X$, and any subset $A$ of $X$, where $A \subseteq X$

$$\delta_x(A) = Ind_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{3.7}$$

, where $Ind_A(x)$ is indicator function Eq. 3.5

Similarly, the expected value of dirac measure $\delta_x$ can be derived like Eq. 3.6.

$$\mathrm{E}_{\Pr(x)}\left[\delta_x(A)\right] = \begin{cases} \sum_{x \in X} \Pr(x) \cdot \delta_x(A) = \sum_{x \in A} \Pr(x) = \Pr(x \in A) & \text{for discrete} \\ \int_{x \in X} \Pr(x) \cdot \delta_x(A) \cdot dx = \int_{x \in A} \Pr(x) \cdot dx = \Pr(x \in A) & \text{for continuous} \end{cases} \tag{3.8}$$

### 3.1.4 Conditional Probability

**Definition 3.1.1.** Probability of $A$ given $B$:

$$\Pr(A \mid B) = \frac{\Pr(A, B)}{\Pr(B)} \tag{3.9}$$

Rearranging the formula of Eq. 3.9 to get the *product rule*:

$$\Pr(A, B) = \Pr(A \mid B) \cdot \Pr(B) \tag{3.10}$$

#### Conditional Probability for Multiple Variables

The following equations would be used frequently:

$$\Pr(A, B \mid C) = \Pr(A \mid B, C) \cdot \Pr(B \mid C) \tag{3.11}$$

$$\Pr(A \mid B, C) = \frac{\Pr(B \mid A, C) \cdot \Pr(A \mid C)}{\Pr(B \mid C)} \tag{3.12}$$

*Proof.* of Eq. 3.11:

From definition of conditional probability, we know that:

$$\Pr(A, B \mid C) = \frac{\Pr(A, B, C)}{\Pr(C)} \tag{3.13}$$

$$\Pr(A \mid B, C) = \frac{\Pr(A, B, C)}{\Pr(B, C)} \tag{3.14}$$

Then, we can get the following equation from Eq. 3.14:

$$\Pr(A, B, C) = \Pr(A \mid B, C) \cdot \Pr(B, C) \tag{3.15}$$

Finally, applying Eq. 3.15 to Eq. 3.13 :

$$\Pr(A, B \mid C) = \frac{\Pr(A \mid B, C) \cdot \Pr(B, C)}{\Pr(C)} = \Pr(A \mid B, C) \cdot \frac{\Pr(B, C)}{\Pr(C)} = \Pr(A \mid B, C) \cdot \Pr(B \mid C)$$

■

*Proof.* of Eq. 3.12:

$$\Pr(B \mid A, C) = \frac{\Pr(A, B, C)}{\Pr(A, C)} \tag{3.16}$$

From above, we can get the following equation:

$$\Pr(A, B, C) = \Pr(B \mid A, C) \cdot \Pr(A, C) \tag{3.17}$$

Finally, applying Eq. 3.17 to Eq. 3.14:

$$\Pr(A \mid B, C) = \frac{\Pr(B \mid A, C) \cdot \Pr(A, C)}{\Pr(B, C)} = \frac{\Pr(B \mid A, C) \cdot \Pr(A \mid C) \cdot \Pr(C)}{\Pr(B \mid C) \cdot \Pr(C)} = \frac{\Pr(B \mid A, C) \cdot \Pr(A \mid C)}{\Pr(B \mid C)}$$

■

**Chain Rule**

The *product rule* of conditional probability Eq. 3.10 can be extended to $n$ variables actually. We can derive the formula from considering some easy cases first.

First, we extend the *product rule* to three variables:

from Eq. 3.15: $\Pr(A,B,C) = \Pr(A \mid B,C) \cdot \Pr(B,C)$

from Eq. 3.10: $\Pr(B,C) = \Pr(B \mid C) \cdot \Pr(C)$

Then, we can get the *product rule* for three variables by applying Eq. 3.10 to Eq. 3.15:

$$\Pr(A,B,C) = \Pr(A \mid B,C) \cdot \Pr(B,C) = \Pr(A \mid B,C) \cdot \Pr(B \mid C) \cdot \Pr(C) \tag{3.18}$$

Next, extending the *product rule* to four variables:

$$\Pr(A,B,C,D) = \Pr(A \mid B,C,D) \cdot \Pr(B,C,D) \tag{3.19}$$

Same as Eq. 3.18, the $\Pr(B,C,D)$ can be written in:

$$\Pr(B,C,D) = \Pr(B \mid C,D) \cdot \Pr(C \mid D) \cdot \Pr(D) \tag{3.20}$$

Therefore, Eq. 3.19 can be rewritten into

$$\Pr(A,B,C,D) = \Pr(A \mid B,C,D) \cdot \Pr(B \mid C,D) \cdot \Pr(C \mid D) \cdot \Pr(D) \tag{3.21}$$

Now, we can generalize the *product rule* to $n$ variables: $\Pr(A_1,A_2,A_3,...,A_n)$

$= \Pr(A_n \mid A_{n-1},...,A_2,A_1) \cdot \Pr(A_{n-1},...,A_2,A_1)$

$= \Pr(A_n \mid A_{n-1},...,A_2,A_1) \cdot \Pr(A_{n-1} \mid A_{n-2},...,A_2,A_1) \cdot \Pr(A_{n-2},...,A_2,A_1)$

$= \Pr(A_n \mid A_{n-1},...,A_2,A_1) \cdot \Pr(A_{n-1} \mid A_{n-2},...,A_2,A_1) \cdot \Pr(A_{n-2} \mid A_{n-3},...,A_2,A_1) \cdot \Pr(A_{n-3},...,A_2,A_1)$

$= ...$

$= \Pr(A_n \mid A_{n-1},...,A_2,A_1) \cdot \Pr(A_{n-1} \mid A_{n-2},...,A_2,A_1) \cdot ... \cdot \Pr(A_k \mid A_{k-1},...,A_2,A_1) \cdot ... \cdot \Pr(A_1)$

To summarize, the equation above can be written as the following formula, called *chain rule*

$$\Pr(A_{1:n}) = \Pr(\bigcap_{k=1}^{n} A_k) = \prod_{k=1}^{n} \Pr(A_k \mid A_{1:k-1}) = \prod_{k=1}^{n} \Pr(A_k \mid \bigcap_{j=1}^{k-1} A_j) \tag{3.22}$$

, where $A_{1:n} = \bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap .... \cap A_n$.

In fact, the *chain rule* can be easily proved by *induction*.

*Proof.* of Eq. 3.22:

1. **Basis**

   when $k = 1$: $\Pr(A_1) = \Pr(A_1)$

   when $k = 2$: $\Pr(A_1, A_2) = \Pr(A_1) \cdot \Pr(A_2 \mid A_1)$

2. **Inductive step**

   Assume the formula Eq. 3.22 holds when $k = m$ (for some unspecified value of $m$):

   $\Pr(A_1, A_2, \ldots, A_m) = \prod_{k=1}^{m} \Pr(A_k \mid \bigcap_{j=1}^{k-1} A_j)$

   when $k = m+1$:

$$
\begin{aligned}
\Pr(A_1, A_2, \ldots, A_m, A_{m+1}) &= \Pr(A_{m+1} \mid A_1, A_2, \ldots, A_m) \cdot \Pr(A_1, A_2, \ldots, A_m) \\
&= \Pr(A_{m+1} \mid A_1, A_2, \ldots, A_m) \cdot \prod_{k=1}^{m} \Pr(A_k \mid \bigcap_{j=1}^{k-1} A_j) \\
&= \Pr(A_{m+1} \mid \bigcap_{j=1}^{m} A_j) \cdot \prod_{k=1}^{m} \Pr(A_k \mid \bigcap_{j=1}^{k-1} A_j) \\
&= \Pr(A_{(m+1)} \mid \bigcap_{j=1}^{(m+1)-1} A_j) \cdot \prod_{k=1}^{m} \Pr(A_k \mid \bigcap_{j=1}^{k-1} A_j) \\
&= \prod_{k=1}^{m+1} \Pr(A_k \mid \bigcap_{j=1}^{k-1} A_j)
\end{aligned}
\tag{3.23}
$$

3. **Conclusion**

   $\Longrightarrow$ $\forall k \in \mathbb{N}$, formula Eq. 3.22 holds. Since both the basis and the inductive step have been performed, by mathematical induction, formula Eq. 3.22 holds for all natural numbers $k$

■

### 3.1.5 Independence

Two events $A$, $B$ are independent if the occurrence of $B$ does not affect the probability of $A$, and vice versa. Formally saying,

**Definition 3.1.2.** $A$, $B$ are independent if

$$\Pr(A \mid B) = \Pr(A) \tag{3.24}$$

Indeed, the definition can be also written in:

$$\Pr(A,B) = \Pr(A) \cdot \Pr(B) \tag{3.25}$$

Actually, Eq. 3.25 holds if and only if Eq. 3.24 holds.

*Proof.*

$$\Pr(A \mid B) = \frac{\Pr(A,B)}{\Pr(B)} = \Pr(A) \iff \Pr(A,B) = \Pr(A) \cdot \Pr(B) \tag{3.26}$$

■

**More than two events**

A finite set of events $\{A_i\}, i = 1,2,\ldots,n$, are *pairwise independent* if and only if every pair of events is independent, then

$$\Pr(A_p, A_q) = \Pr(A_p) \cdot \Pr(A_q) \tag{3.27}$$

, for all distinct pairs of indices $p, q$.

On the other hands, if the finite set of events $\{A_i\}, i = 1,2,\ldots,n$, are *mutually independent*, then

$$\Pr(A_{1:n}) = \prod_{i=1}^{n} \Pr(A_i) \tag{3.28}$$

, where $A_{1:n} = \bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap .... \cap A_n$.

### 3.1.6 Conditional Independence

**Definition 3.1.3.** $A, B$ are conditionally independent given $C$:

$$\Pr(A \mid B, C) = \Pr(A \mid C) \tag{3.29}$$

This means that the occurrence of $B$ does not affect the probability of $A$ given $C$, and vice versa.

The definition may be also written in:

$$\Pr(A, B \mid C) = \Pr(A \mid C) \cdot \Pr(B \mid C) \tag{3.30}$$

In fact, Eq. 3.29 holds if and only if Eq. 3.30 holds. This can be easily proved by applying Eq. 3.29 in Eq. 3.11 $\implies \Pr(A, B \mid C) = \Pr(A \mid C) \cdot \Pr(A \mid C)$

### 3.1.7 Bayesian Rule

From definition of conditional probability, we know that:

$$\Pr(A \mid B) = \frac{\Pr(A, B)}{\Pr(B)} \tag{3.31}$$

$$\Pr(B \mid A) = \frac{\Pr(A, B)}{\Pr(A)} \tag{3.32}$$

$$\Pr(A, B) = \Pr(B \mid A) \cdot \Pr(A) \tag{3.33}$$

Thus,

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)} \tag{3.34}$$

Suppose that $A$ is the event what we concern and $B$ is an observation event. We can use $B$ to predict $A$ by calculating the probability of $A$ given $B$. Actually, Eq. 3.34 can be interpreted as the following:

$$\overbrace{\Pr(A \mid B)}^{\text{posterior}} = \frac{\overbrace{\Pr(B \mid A)}^{\text{likelihood}} \cdot \overbrace{\Pr(A)}^{\text{prior}}}{\underbrace{\Pr(B)}_{\text{evidence}}} \tag{3.35}$$

- **Prior**:Prior belief, the probability of target event $A$ happen.
- **Likelihood**:The probability of observation $B$ happen given that target event $A$ happen.
- **Evidence**:The probability of observation $B$ happen, which is used as a *normalization factor*

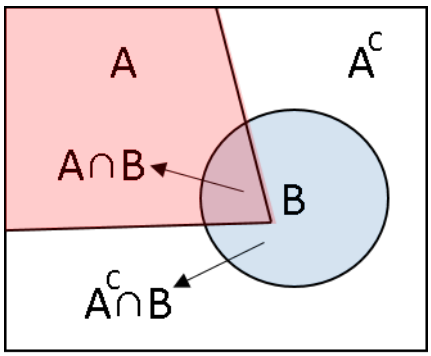



Figure 3.1: The probability space of two events $A$, $B$

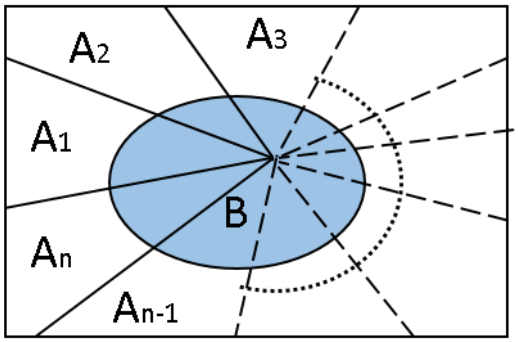



Figure 3.2: The mutually exclusive and exhaustive probability space

- **Posterior**:The probability of target event $A$ happen given the observation $B$

If the relationship of events $A$, $B$ is the shown in Fig. 3.1, then $B = B \cap A + B \cap A^C$, where $A^C$ is the *complement* of $A$. Thus,

$$\begin{aligned} \Pr(B) &= \Pr(A,B) + \Pr(A^C,B) \\ &= \Pr(B \mid A) \cdot \Pr(A) + \Pr(B \mid A^C) \cdot \Pr(A^C) \end{aligned} \tag{3.36}$$

$$\begin{aligned} \Pr(A \mid B) &= \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)} \\ &= \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B \mid A) \cdot \Pr(A) + \Pr(B \mid A^C) \cdot \Pr(A^C)} \end{aligned} \tag{3.37}$$

More general saying, if $\{A_k\}$ are *mutually exclusive and exhaustive* for $k = 1, 2, \ldots, n$, such as Fig. 3.2, then it satisfies that

$$\begin{cases} \Pr(A_i \cap A_j) = 0 & \text{if } i \neq j \\ \Pr\left(\bigcup_{i=1}^{n} A_i\right) = \Pr(A_1 \cup A_2 \cup .... \cup A_n) = 1 \end{cases} \tag{3.38}$$

Therefore, $\Pr(B)$ can be written in following equations:

$$\Pr(B) = \sum_{k=1}^{n} \Pr(A_k, B) = \sum_{k=1}^{n} \Pr(B \mid A_k) \cdot \Pr(A_k) \tag{3.39}$$

, for discrete or

$$\Pr(B) = \int \Pr(A, B) dA = \int \Pr(B \mid A) \cdot \Pr(A) dA \tag{3.40}$$

, for continuous

Appling Eq. 3.39 or Eq. 3.40 in Eq. 3.34:

$$\Pr(A_k \mid B) = \frac{\Pr(B \mid A_k) \cdot \Pr(A_k)}{\Pr(B)} = \frac{\Pr(B \mid A_k) \cdot \Pr(A_k)}{\sum_{k=1}^{n} \Pr(B \mid A_k) \cdot \Pr(A_k)} \tag{3.41}$$

, for discrete or

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)} = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\int \Pr(B \mid A) \cdot \Pr(A) dA} \tag{3.42}$$

, for continuous

We should notice that

$$\sum_{k=1}^{n} \Pr(A_k \mid B) = \sum_{k=1}^{n} \frac{\Pr(A_k, B)}{\Pr(B)} = \frac{1}{\Pr(B)} \cdot \sum_{k=1}^{n} \Pr(A_k, B) = \frac{1}{\Pr(B)} \cdot \Pr(B) = 1 \tag{3.43}$$

, for discrete or

$$\int \Pr(A \mid B) \cdot dA = \int \frac{\Pr(A, B)}{\Pr(B)} \cdot dA = \frac{1}{\Pr(B)} \cdot \int \Pr(A, B) dA = \frac{1}{\Pr(B)} \cdot \Pr(B) = 1 \tag{3.44}$$

for continuous.

This is because $\Pr(A \mid B)$ is a *PDF*/*PMF* , and integral of *PDF*/sum of *PMF* should be 1.

**Normalizing Constant**

Normalizing constant is a constant by which a non-negative function must be multiplied so the area under its graph is 1 for continuous , or the sum of this function is 1 for discrete. This is useful to make some function to be a *PDF*(Probability Density Function) or *PMF*(Probability Mass Function). For example, a function $f$ is defined as:

$$f(x) = e^{-x^2/2}, \text{ where } x \in X = (-\infty, \infty) \tag{3.45}$$

The area under function $f$'s graph is:

$$\int_{-\infty}^{\infty} f(x) \cdot dx = \int_{-\infty}^{\infty} e^{-x^2/2} \cdot dx = \sqrt{2\pi}$$

If constant $C$ is defined as $\frac{1}{\sqrt{2\pi}}$, and function $g$ is defined as

$$g(x) = C \cdot f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}, \text{ where } x \in X = (-\infty, \infty)$$

, then the area under function $g$'s graph is:

$$\int_{-\infty}^{\infty} g(x) \cdot dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \cdot dx = 1$$

Actually, the function $g$ is the *PDF* of the *standard normal distribution* whose expected value(means) is 0 and variance is 1 , and $C$ is normalizing constant of function $f$. More general saying, normalizing constant is defined as:

$$\text{normalizing constant} = \frac{1}{\int_{x \in X} f(x) \cdot dx} \tag{3.46}$$

The *PDF*, function $g$, made from function $f$ is:

$$\text{function } g\text{:}PDF \text{ made from function } f = \frac{f(x)}{\int_{x \in X} f(x) \cdot dx} \tag{3.47}$$

The case above is for continuous. On the other hand, for one discrete example, a discrete function $f$ is defined as :

$$f(x) = x, \text{ where } x \in X = \{1,2,3,4\} \tag{3.48}$$

The sum of function $f$ is :

$$\sum_{x \in X} f(x) = \sum_{x=1}^{4} f(x) = 10 \tag{3.49}$$

If constant $D$ is defined as $\frac{1}{10}$ , and function $g$ is defined as

$$g(x) = D \cdot f(x) = \frac{1}{10} \cdot x, \text{ where } x \in X = \{1,2,3,4\}$$

, then sum of function $g$ is:

$$\sum_{x \in X} g(x) = \sum_{x \in X} \frac{1}{10} \cdot x = \sum_{x=1}^{4} \frac{1}{10} \cdot x = 1$$

The function $g$ is a *PMF* whose probability increase 0.1 from $x$ to $x+1$, and $D$ is normalizing constant of function $f$. More general saying, normalizing constant is defined as:

$$\text{normalizing constant} = \frac{1}{\sum_{x \in X} f(x)} \tag{3.50}$$

The *PMF*, function $g$, made from function $f$ is:

$$\text{function } g\text{:}PMF \text{ made from function } f = \frac{f(x)}{\sum_{x \in X} f(x)} \tag{3.51}$$

Now, we apply normalizing constant to Bayes' rule. Suppose there is a finite set of events $\{A_i\}$, $i = 1,2,\ldots,n$, and one observation event $B$. From Eq. 3.34, the probability of event $A_i$ happen given the observation $B$ is:

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A_i) \cdot \Pr(A_i)}{\Pr(B)} \tag{3.52}$$

The $\Pr(A_i)$ is prior probability of the event $A_i$ happen , $\Pr(B \mid A_i)$ is the probability of the observation assuming the event $A_i$ happen , and $\Pr(B)$ is the probability producing observation data. If $\Pr(B \mid$

$A_i)\cdot\Pr(A_i)$ is viewed as the $f(x)$ in the example of normalizing constant mentioned above , then the $g(x)$ is $\Pr(A_i \mid B)$ and normalizing constant is $\frac{1}{\Pr(B)}$.

$$\overbrace{\Pr(A_i \mid B)}^{\text{function } g} = \frac{\overbrace{\Pr(B \mid A_i)\cdot\Pr(A_i)}^{\text{function } f}}{\underbrace{\Pr(B)}_{\text{normalizing constant}}} \tag{3.53}$$

However, most of the time, $\Pr(B)$ is difficult to calculate, so we only know that:

$$\Pr(A_i \mid B) \propto \Pr(B \mid A_i)\cdot\Pr(A_i)$$

Since $\Pr(A_i \mid B)$ is a probability, the sum over all possible(*mutually exclusive and exhaustive*) events $A_i$ is 1.

$$\sum_{i=1}^{n}\Pr(A_i \mid B) = \sum_{i=1}^{n}\frac{\Pr(B \mid A_i)\cdot\Pr(A_i)}{\Pr(B)} = \frac{1}{\Pr(B)}\cdot\sum_{i=1}^{n}\Pr(B \mid A_i)\cdot\Pr(A_i) = 1 \tag{3.54}$$

Actually, Eq. 3.54 implies:

$$\sum_{i=1}^{n}\Pr(B \mid A_i)\cdot\Pr(A_i) = \Pr(B) \tag{3.55}$$

Thus,

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A_i)\cdot\Pr(A_i)}{\Pr(B)} = \frac{\Pr(B \mid A_i)\cdot\Pr(A_i)}{\sum_{i=1}^{n}\Pr(B \mid A_i)\cdot\Pr(A_i)}$$

The formula above is same as Eq. 3.41, and the *normalizing constant* is:

$$\frac{1}{\Pr(B)} = \frac{1}{\sum_{i=1}^{n}\Pr(B \mid A_i)\cdot\Pr(A_i)} \tag{3.56}$$

In fact, this formula fit the model Eq. 3.51.

## 3.2 Bayesian Network

A Bayesian network, Bayes network, belief network, Bayes(ian) model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents

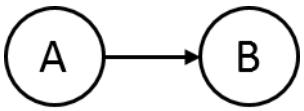



Figure 3.3: The basic Bayesian model

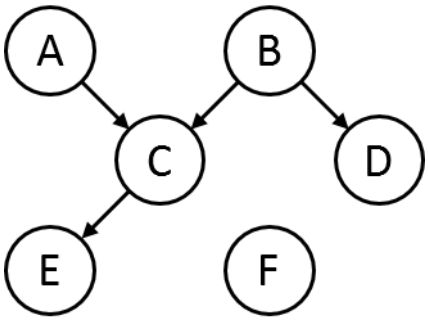



Figure 3.4: The general Bayesian model

a set of random variables and their conditional dependencies via a directed acyclic graph (DAG).

### 3.2.1 Bayes Model

#### Basic Bayes Model

The basic Bayes model is depicted in Fig. 3.3, where the directed arrow represents the *causality*. Fig. 3.3 means that $A$ *causes* $B$. In this Bayes model, $A$ is the *parent* of $B$. In other words, $B$ is the *child* of $A$.

#### General Bayes Model

In general Bayes model, the relations between nodes are more than parents and children. The followings are common relations between them.

- **Descendant**:Nodes that can be reached from other nodes called descendants
- **Ancestor**:Nodes from which the nodes can be reached on a directed path

For example, in Fig. 3.4, the relations between nodes can be organized as following table:

| | |
|---|---|
| parent ; child | $(A;C),(B;C),(B;D),(C;E)$ |
| ancestor ; descendants | $(A;C,E),(B;C,D,E),(C;E)$ |

The interesting part of this example is that **no one** causes $F$. This means that all other nodes can not affect $F$. In other words, $F$ is independent of all other nodes. More formal saying, the Bayes model follows the rules:

**Definition 3.2.1.** There are *no loops* in Bayes networks since no descendant can be its own ancestor

**Definition 3.2.2.** Each variable(node) is *conditionally independent* of all its *non-descendants* in graph given the value of *all its parents*

Thus, the independence among the nodes are easy to be recognized since conditional relationships are clearly defined by directed arrows in the graph. Take Fig. 3.4 as an example,

- Given $A$ and $B,C$ is conditional independent of $D,F$
- Given $C,E$ is conditional independent of $A,B,D,F$
- Given $B,D$ is conditional independent of $A,C,E,F$
- ...

The joint probability is defined as the probability that a series of events will happen concurrently. If we calculate the joint probability of Fig. 3.4:

$$\begin{aligned}
&\Pr(A,B,C,D,E,F) \\
&\quad = \Pr(A)\cdot\Pr(B \mid A)\cdot\Pr(C \mid A,B)\cdot\Pr(D \mid A,B,C)\cdot\Pr(E \mid A,B,C,D)\cdot\Pr(F \mid A,B,C,D,E) \\
&\qquad \text{(By Chain Rule Eq. 3.22)} \\
&\quad = \Pr(A)\cdot\Pr(B)\cdot\Pr(C \mid A,B)\cdot\Pr(D \mid B)\cdot\Pr(E \mid C)\cdot\Pr(F) \\
&\qquad \text{(By conditional independence in fig 3.4)} \qquad (3.57)
\end{aligned}$$

In general, the joint probability in Bayes network is:

$$\Pr(A_{1:n}) = \Pr(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} \Pr(A_i \mid parents(A_i)) \qquad (3.58)$$

, where $A_{1:n} = \bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \ldots \cap A_n$ and $parents(A_i)$ is the parents of $A_i$.

[5] can provide more knowledge about Bayesian networks , which is a student-contributed open-source electronic textbook of Chemical Engineering Process Dynamics and Controls covering the materials used in senior level controls course at University of Michigan.

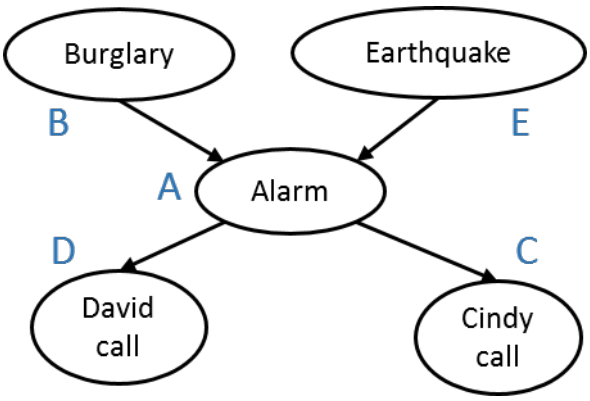



Figure 3.5: The example of Bayesian network

### 3.2.2 Inference from Bayesian Networks

For Bayesian networks, the *Conditional Probability Distribution* (CPD) at each node must be specified. *CPD* can be represented as *Conditional Probability Table* (CPT), which lists the probability that the child node takes on each of its different values for each combination of values of its parents.

Take Fig. 3.5 as an example, the *CPT* is given by following:

1. *CPT* of B

| $B=T$ | $B=F$ |
|---|---|
| 0.01 | 0.99 |

2. *CPT* of E

| $E=T$ | $E=F$ |
|---|---|
| 0.03 | 0.97 |

3. *CPT* of A

| $B$ | $E$ | $A=T$ | $A=F$ |
|---|---|---|---|
| $F$ | $F$ | 0.02 | 0.98 |
| $F$ | $T$ | 0.25 | 0.75 |
| $T$ | $F$ | 0.92 | 0.08 |
| $T$ | $T$ | 0.95 | 0.05 |

4. *CPT* of D

| $A$ | $D=T$ | $D=F$ |
|---|---|---|
| $F$ | 0.15 | 0.85 |
| $T$ | 0.72 | 0.28 |

5. *CPT* of C

| $A$ | $C=T$ | $C=F$ |
|---|---|---|
| $F$ | 0.27 | 0.73 |
| $T$ | 0.96 | 0.04 |

In this example, the joint probability $\Pr(A,B,C,D,E)$ is:

$$\Pr(A,B,C,D,E) = \Pr(B)\cdot\Pr(E)\cdot\Pr(A \mid B,E)\cdot\Pr(C \mid A)\cdot\Pr(D \mid A) \qquad \text{(By Eq. 3.58)} \qquad (3.59)$$

By the equation above and *CPT* of $A,B,C,D,E$, we can get the following joint probability table:

| $A$ | $B$ | $C$ | $D$ | $E$ | $\Pr(A,B,C,D,E)$ |
|---|---|---|---|---|---|
| $F$ | $F$ | $F$ | $F$ | $F$ | $0.99\cdot0.97\cdot0.98\cdot0.73\cdot0.85 = 0.583948827$ |
| $F$ | $F$ | $F$ | $F$ | $T$ | $0.99\cdot0.03\cdot0.75\cdot0.73\cdot0.85 = 0.0138216375$ |
| $F$ | $F$ | $F$ | $T$ | $F$ | $0.99\cdot0.97\cdot0.98\cdot0.73\cdot0.15 = 0.103049793$ |
| $F$ | $F$ | $F$ | $T$ | $T$ | $0.99\cdot0.03\cdot0.75\cdot0.73\cdot0.15 = 0.0024391125$ |
| $F$ | $F$ | $T$ | $F$ | $F$ | $0.99\cdot0.97\cdot0.98\cdot0.27\cdot0.85 = 0.215981073$ |
| $F$ | $F$ | $T$ | $F$ | $T$ | $0.99\cdot0.03\cdot0.75\cdot0.27\cdot0.85 = 0.0051121125$ |
| $F$ | $F$ | $T$ | $T$ | $F$ | $0.99\cdot0.97\cdot0.98\cdot0.27\cdot0.15 = 0.038114307$ |
| $F$ | $F$ | $T$ | $T$ | $T$ | $0.99\cdot0.03\cdot0.75\cdot0.27\cdot0.15 = 0.0009021375$ |
| $F$ | $T$ | $F$ | $F$ | $F$ | $0.01\cdot0.97\cdot0.08\cdot0.73\cdot0.85 = 0.000481508$ |
| $F$ | $T$ | $F$ | $F$ | $T$ | $0.01\cdot0.03\cdot0.05\cdot0.73\cdot0.85 = 0.0000093075$ |
| $F$ | $T$ | $F$ | $T$ | $F$ | $0.01\cdot0.97\cdot0.08\cdot0.73\cdot0.15 = 0.000084972$ |
| $F$ | $T$ | $F$ | $T$ | $T$ | $0.01\cdot0.03\cdot0.05\cdot0.73\cdot0.15 = 0.0000016425$ |
| $F$ | $T$ | $T$ | $F$ | $F$ | $0.01\cdot0.97\cdot0.08\cdot0.27\cdot0.85 = 0.000178092$ |
| $F$ | $T$ | $T$ | $F$ | $T$ | $0.01\cdot0.03\cdot0.05\cdot0.27\cdot0.85 = 0.0000034425$ |
| $F$ | $T$ | $T$ | $T$ | $F$ | $0.01\cdot0.97\cdot0.08\cdot0.27\cdot0.15 = 0.000031428$ |
| $F$ | $T$ | $T$ | $T$ | $T$ | $0.01\cdot0.03\cdot0.05\cdot0.27\cdot0.15 = 0.0000006075$ |

| $A$ | $B$ | $C$ | $D$ | $E$ | $\Pr(A,B,C,D,E)$ |
|---|---|---|---|---|---|
| $T$ | $F$ | $F$ | $F$ | $F$ | $0.99 \cdot 0.97 \cdot 0.02 \cdot 0.04 \cdot 0.28 = 0.0002151072$ |
| $T$ | $F$ | $F$ | $F$ | $T$ | $0.99 \cdot 0.03 \cdot 0.25 \cdot 0.04 \cdot 0.28 = 0.00008316$ |
| $T$ | $F$ | $F$ | $T$ | $F$ | $0.99 \cdot 0.97 \cdot 0.02 \cdot 0.04 \cdot 0.72 = 0.0005531328$ |
| $T$ | $F$ | $F$ | $T$ | $T$ | $0.99 \cdot 0.03 \cdot 0.25 \cdot 0.04 \cdot 0.72 = 0.00021384$ |
| $T$ | $F$ | $T$ | $F$ | $F$ | $0.99 \cdot 0.97 \cdot 0.02 \cdot 0.96 \cdot 0.28 = 0.0051625728$ |
| $T$ | $F$ | $T$ | $F$ | $T$ | $0.99 \cdot 0.03 \cdot 0.25 \cdot 0.96 \cdot 0.28 = 0.00199584$ |
| $T$ | $F$ | $T$ | $T$ | $F$ | $0.99 \cdot 0.97 \cdot 0.02 \cdot 0.96 \cdot 0.72 = 0.0132751872$ |
| $T$ | $F$ | $T$ | $T$ | $T$ | $0.99 \cdot 0.03 \cdot 0.25 \cdot 0.96 \cdot 0.72 = 0.00513216$ |
| $T$ | $T$ | $F$ | $F$ | $F$ | $0.01 \cdot 0.97 \cdot 0.92 \cdot 0.04 \cdot 0.28 = 0.0000999488$ |
| $T$ | $T$ | $F$ | $F$ | $T$ | $0.01 \cdot 0.03 \cdot 0.95 \cdot 0.04 \cdot 0.28 = 0.000003192$ |
| $T$ | $T$ | $F$ | $T$ | $F$ | $0.01 \cdot 0.97 \cdot 0.92 \cdot 0.04 \cdot 0.72 = 0.0002570112$ |
| $T$ | $T$ | $F$ | $T$ | $T$ | $0.01 \cdot 0.03 \cdot 0.95 \cdot 0.04 \cdot 0.72 = 0.000008208$ |
| $T$ | $T$ | $T$ | $F$ | $F$ | $0.01 \cdot 0.97 \cdot 0.92 \cdot 0.96 \cdot 0.28 = 0.0023987712$ |
| $T$ | $T$ | $T$ | $F$ | $T$ | $0.01 \cdot 0.03 \cdot 0.95 \cdot 0.96 \cdot 0.28 = 0.000076608$ |
| $T$ | $T$ | $T$ | $T$ | $F$ | $0.01 \cdot 0.97 \cdot 0.92 \cdot 0.96 \cdot 0.72 = 0.0061682688$ |
| $T$ | $T$ | $T$ | $T$ | $T$ | $0.01 \cdot 0.03 \cdot 0.95 \cdot 0.96 \cdot 0.72 = 0.000196992$ |

This joint probability table can help us to calculate the probability we interest. For instance, if we've got a missed call from Cindy, the probability of burglary can be calculated by:

$$\Pr\big(B = T \;\big|\; C = T\big) = \frac{\Pr\big(B = T, C = T\big)}{\Pr\big(C = T\big)} \tag{3.60}$$

$$\begin{aligned}
\Pr\big(B = T, C = T\big) &= \sum_A \sum_D \sum_E \Pr\big(A, B = T, C = T, D, E\big) \\
&= 0.000178092 + 0.0000034425 + 0.000031428 + 0.0000006075 \\
&\quad + 0.0023987712 + 0.000076608 + 0.0061682688 + 0.000196992 \\
&= 0.00905421
\end{aligned} \tag{3.61}$$

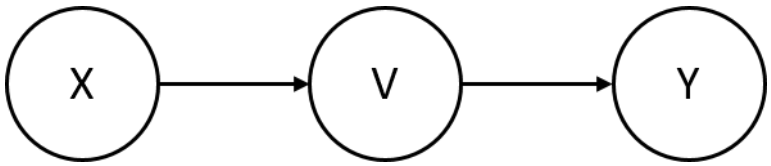



Figure 3.6: Indirect causal effect

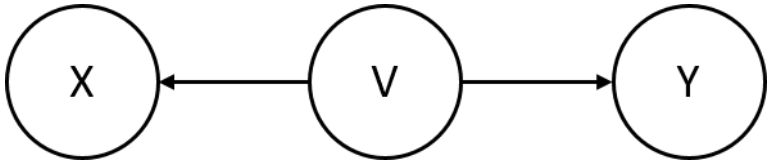



Figure 3.7: Common cause

$$\begin{aligned}\Pr(C=T) &= \sum_A \sum_B \sum_D \sum_E \Pr(A,B,C=T,D,E) \\ &= 0.000178092 + 0.0000034425 + 0.000031428 + 0.0000006075 \\ &+ 0.0023987712 + 0.000076608 + 0.0061682688 + 0.000196992 \\ &+ 0.215981073 + 0.0051121125 + 0.038114307 + 0.0009021375 \\ &+ 0.0051625728 + 0.00199584 + 0.0132751872 + 0.00513216 \\ &= 0.2947296 \end{aligned} \tag{3.62}$$

Therefore, by Eq. 3.60, the probability of burglary given missed call from Cindy, $\Pr(B=T \mid C=T)$, can be calculated by $\frac{0.00905421}{0.2947296} = 0.03072039591$.

### 3.2.3 D-Separation and D-Connection

Bayesian networks encode the dependencies and independencies between variables. There are three basic models of Bayes network[3].

1. Tail-to-head, or serial pattern: *indirect causal effect* or *indirect evidential effect*(Fig. 3.6)
2. Head-to-head, or converging pattern: *common cause*(Fig. 3.7)
3. Tail-to-tail, or diverging pattern(collider on the path): *common effect*(Fig. 3.8)

In *indirect causal effect* and *common cause* model, $X$ and $Y$ are **conditionally independent** given $V$. On the other hand, $X$ and $Y$ are marginally independent in *common effect* model. However, once $V$ is known, they are **conditionally dependent**. This is also called *explaining away*.

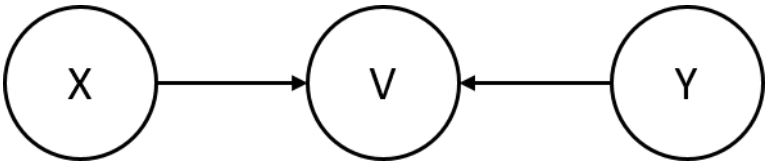



Figure 3.8: Common effect

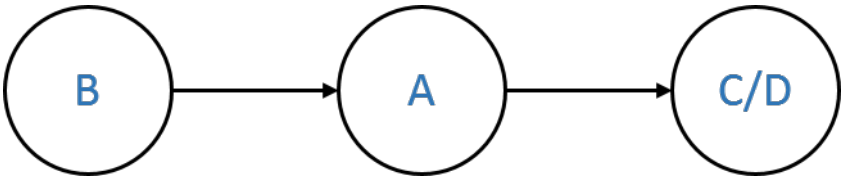



Figure 3.9: Indirect causal effect example


Figure 3.10: Common cause example

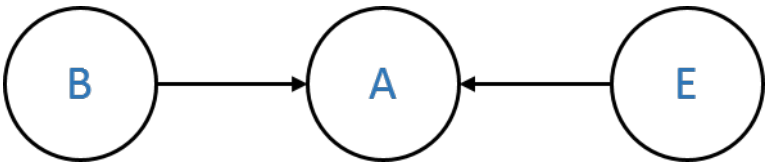



Figure 3.11: Common effect example


Figure 3.12: One famous example of common effect

D-separation is a graphical property of Bayesian networks. The $D$ standards for the dependence. If two sets of nodes $X$, $Y$ are *d-separated* relative to a set of variables $V$ (excluding $X$ and $Y$) in Bayesian networks , the corresponding variable sets $X$ and $Y$ are **independent** conditional on $V$ in all probability distributions. This means that the knowledge about $X$ gives you no extra information about $Y$ once you have know $V$. Likewise, once you have knowledge about $V$, $X$ adds nothing to what you know about $Y$. Take some subsets Fig. 3.9,3.10 in Fig. 3.5 as examples. If we know event $B$(burglary) happen , then we are very likely to get the phone call from Cindy or David because they may hear the alarm. The probability of event $C$(Cindy call) or $D$(David call) would be increased when we know the $B$ is happen. In another example, if we get a phone call from David, it might because he hear the alarm of home security system (but we don't now the alarm is triggered or not). For same reason, Cindy might call you. Thus, the probability getting the phone call from Cindy would be increased if we got the phone call from David. In the first example, event $B$ is dependent of $C$ or $D$. In the second example, event $C$ is dependent of $D$. However, in both examples, once we know the alarm is triggered or not (event $A$), then $B$ is conditionally independent of $C$ or $D$. Also, $C$ is conditionally independent of $D$.

On the other hand, If two sets of nodes $X$, $Y$ are *d-connected* relative to a set of variables $V$ (excluding $X$ and $Y$) in Bayesian networks, the corresponding variable sets $X$ and $Y$ are **dependent** conditional on $V$. This can be explained by one famous example. In Fig. 3.12, there are two **independent** causes of the car refusing start. One is that the car has no gas, and another is the car has a dead battery. The knowledge about the battery is charged add nothing information about whether there is gas. Nevertheless, if we get the information that the battery is charged after knowing the car can not start, then there must be no gas. Therefore, the event of no gas and dead battery are **conditionally dependent** given the information that the car cannot start. In a like manner, the earthquake and the burglary are two **independent** occasions to trigger the alarm of a home security system in the example of the subset Fig. 3.11 of Fig. 3.5. The knowledge about the earthquake cannot help us know burglary happens or not. However, if we know the earthquake does not happen after we know the alarm is trigger, then the probability of burglary would be increased.

The examples above can be modeled into three basic models mentioned on the top of this subsection. Intuitively, a path is active if it carries information, or dependence. Two variables $X$ and $Y$ might be connected by lots of paths in the Baysian network. $X$ and $Y$ are *d-separated* if all the paths

that connect them are *inactive*. In other word, if no path between $X$ and $Y$ is *active*, then they are *d-separated*. A path is active when every nodes on the path is active. Paths, and vertices on these paths, are active or inactive relative to a set of other vertices $V$. Considering that we know nothing about $V$, then we can model the examples above by following:

1. In the first pattern, tail to head, $X$ is an indirect cause of $Y$ (Fig. 3.9 to Fig. 3.6).

2. In the second pattern, head to head, $V$ is the common cause of $X$ and $Y$ (Fig. 3.10 to Fig. 3.7).

3. In the third pattern, $V$ is the common effect of $X$ and $Y$, but no causal connection between them. (Fig. 3.11 to Fig. 3.8)

The causal situations of first two cases give rise to dependence between $X$ and $Y$, and both of these undirected paths are *active* in the theory of d-separation. In the third pattern, there is no information passed from $X$ and $Y$. Thus, the path in the third case is *inactive*.

Now, considering that we have the information about $V$. When $V$ is known, the status of these three cases with respect to being *active* or *inactive* flip-flops. That is, the path in first two patterns are *inactive* and the path in the third one is *active*.

Imagine that the nodes on path are switches and suppose that the conditioning set $V$ is empty (no information about $V$). If the arrows between the node collide on path (tail-to-tail pattern), then the switch is **off**. Otherwise, the switch is **on**. In contrast, if the conditioning set $V$ is known, then the switches **toggled**.

To summarize the rules we learned, the *active path* can be defined as:

1. Tail to head:pattern: active iff $V$ not observed

2. Head to head, or converging pattern: active iff $V$ not observed

3. Tail to tail, or diverging pattern (non-collider on the path): active iff either $V$ or one of its descendants is observed.

[20] is a good tutorial is to learn the *d-separation* and *d-connection*. There are more detail about Bayesian networks can be found there.

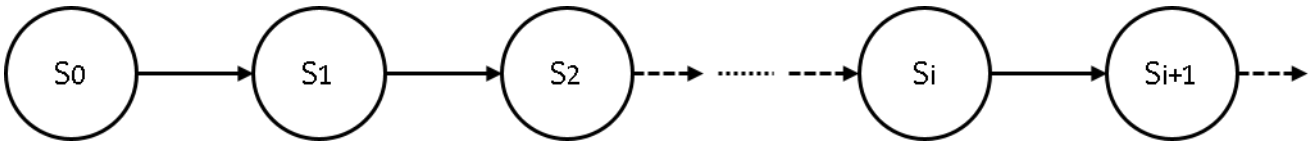



Figure 3.13: The first-order Markov model

**D-Separation**

**Definition 3.2.3.** $X$ and $Y$ are *d-separated* given $V$ if there is no *active path* between any node $x \in X$ and $y \in Y$ given $V$.

**D-Connection**

**Definition 3.2.4.** $X$ and $Y$ are *d-connected* by $V$ in Bayes network if and only if there exists an *inactive path* $U$ between some vertex in $X$ and some vertex in $Y$ such that for every collider $C$ on $U$, either $C$ or a descendent of $C$ is in $V$, and no non-collider on $U$ is in $V$.

By the definition above, another definition of *D-separation* is: $X$ and $Y$ are *d-separated* by $V$ in Bayes network if and only if they are not *d-connected* by $V$.



## 3.3 Hidden Markov Model

### 3.3.1 Markov Model (First Order)

**Definition 3.3.1.** A sequence of random states $\{S_t\} = S_0, S_1, S_2, \ldots, t$ denotes the time, where the future states are independent of all past states given current state. Formally,

$$\Pr\left(S_{t+1} \mid S_0, S_1, S_2, \ldots, S_t\right) = \Pr\left(S_{t+1} \mid S_t\right), \tag{3.63}$$

where $\Pr\left(S_{t+1} \mid S_t\right)$ is the *state transition probability*.

In fact, markov model can be considered as one simple Bayes network in Fig. 3.13. From Fig. 3.13, Eq. 3.63 is clear by the Bayes network's Def. 3.2.2. Also, the joint probability of states

$\Pr(S_{0:n})$ is:

$$\begin{aligned}\Pr(S_{0:n}) &= \Pr(\bigcap_{t=0}^{n} S_t) \\ &= \Pr(S_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1}),\end{aligned} \tag{3.64}$$

where $S_{0:n} = \bigcap_{t=0}^{n} S_t = S_0 \cap S_1 \cap S_2 \cap \ldots \cap S_n$.

*Proof.*

$$\begin{aligned}\Pr(S_{0:n}) &= \Pr(S_0) \cdot \Pr(S_1 \mid S_0) \cdot \Pr(S_2 \mid S_1, S_0) \cdot \ldots \cdot \Pr(S_n \mid S_{n-1}, S_{n-2}, \ldots, S_1, S_0) \text{ (Chain Rule Eq. 3.22)} \\ &= \Pr(S_0) \cdot \Pr(S_1 \mid S_0) \cdot \Pr(S_2 \mid S_1) \cdot \ldots \cdot \Pr(S_n \mid S_{n-1}) \text{ (definition Eq. 3.63)} \\ &= \Pr(S_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1})\end{aligned} \tag{3.65}$$

■

### 3.3.2 Hidden Markov Model

Markov Model is a powerful abstraction for time series data, but it fail to handle some common scenario. However, one challenge of Markov Model is to reason the state if it is unobserved. *Hidden Markov Model* (HMM) can handle this scenario if some outcome generated by each state can be observed. *HMM* is a Markov model in which the system is assumed to be a Markov process with unobserved(hidden) states with visible outputs. It is a useful tool for representing probability distribution over sequences of the observations. Formally, *HMM* follows the definitions:

**Definition 3.3.2.** the output $Z_t$ generated by *hidden* state $S_t$ at time $t$ is independent of the states and outputs at all other time indices given state $S_t$

$$\Pr(Z_t \mid S_0, S_1, \ldots, S_t, Z_0, Z_1, \ldots, Z_{t-1}) = \Pr(Z_t \mid S_{0:t}, Z_{0:t-1}) = \Pr(Z_t \mid S_t) \tag{3.66}$$

**Definition 3.3.3.** the next state $S_{t+1}$ at time $t+1$ is only related to the state $S_t$ at time $t$

$$\Pr(S_{t+1} \mid S_0, S_1, \ldots, S_t, Z_0, Z_1, \ldots, Z_t) = \Pr(S_{t+1} \mid S_{0:t}, Z_{0:t}) = \Pr(S_{t+1} \mid S_t) \tag{3.67}$$

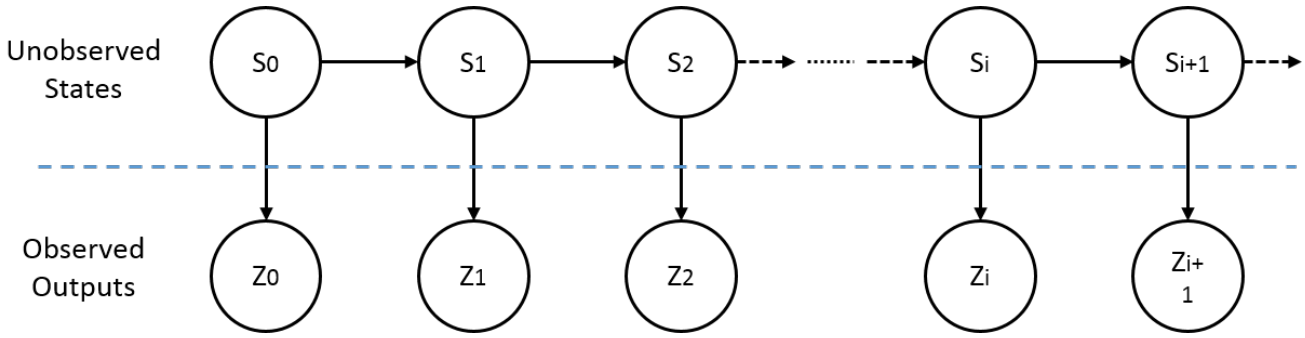



Figure 3.14: The hidden Markov model

The definition of Markov Model Eq. 3.63 holds if Eq. 3.67 holds. That is, $Eq.\ 3.67 \implies Eq.\ 3.63$

In *HMM*, the joint probability of states $\Pr(S_{0:n})$ is same as Eq. 3.64:

$$\Pr(S_{0:n}) = \Pr(S_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1}) \tag{3.68}$$

, where $S_{0:n} = \bigcap_{t=0}^{n} S_t = S_0 \cap S_1 \cap S_2 \cap \ldots \cap S_n$.

Also, *HMM* can be considered as one simple Bayes network in Fig. 3.14. The joint probability of states and outputs $\Pr(S_{0:n}, Z_{0:n})$ is:

$$\begin{aligned} \Pr(S_{0:n}, Z_{0:n}) &= \Pr(\bigcap_{t=0}^{n} S_t \cap Z_t) \\ &= \Pr(S_0) \cdot \Pr(Z_0 \mid S_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1}) \cdot \Pr(Z_t \mid S_t) \end{aligned} \tag{3.69}$$

, where $S_{0:n} = \bigcap_{t=0}^{n} S_t = S_0 \cap S_1 \cap S_2 \cap \ldots \cap S_n$ and $Z_{0:n} = \bigcap_{t=0}^{n} Z_t = Z_0 \cap Z_1 \cap Z_2 \cap \ldots \cap Z_n$.

Actually, this formula can be easily derived by considering some simple cases.

*Proof.* of Eq. 3.69:

Consider the joint probability $\Pr(S_{0:n}, Z_{0:n})$ by Chain-Rule Eq. 3.22:

$$\begin{aligned} \Pr(S_{0:n}, Z_{0:n}) &= \Pr(S_0, Z_0, S_1, Z_1, \ldots, S_n, Z_n) \\ &= \Pr(S_0, Z_0) \cdot \Pr(S_1, Z_1 \mid S_0, Z_0) \cdot \Pr(S_2, Z_2 \mid S_0, Z_0, S_1, Z_1) \\ &\cdot \ldots \cdot \Pr(S_n, Z_n \mid S_0, Z_0, S_1, Z_1, \ldots, S_{n-1}, Z_{n-1}) \\ &= \Pr(S_0, Z_0) \cdot \prod_{t=1}^{n} \Pr(S_t, Z_t \mid S_{0:t-1}, Z_{0:t-1}) \end{aligned} \tag{3.70}$$

Now, we focus on $\Pr(S_t, Z_t \mid S_{0:t-1}, Z_{0:t-1})$ of Eq. 3.70.

when $t = 1$:

$$\begin{aligned}
\Pr(S_1, Z_1 \mid S_0, Z_0) &= \frac{\Pr(S_0, Z_0, S_1, Z_1)}{\Pr(S_0, Z_0)} \text{ (By difinition)} \\
&= \frac{\Pr(S_0, Z_0) \cdot \Pr(S_1 \mid S_0, Z_0) \cdot \Pr(Z_1 \mid S_0, S_1, Z_0)}{\Pr(S_0, Z_0)} \text{ (By Chain-Rule Eq. 3.22)} \\
&= \Pr(S_1 \mid S_0, Z_0) \cdot \Pr(Z_1 \mid S_0, S_1, Z_0) \qquad (3.71)
\end{aligned}$$

Similarly, when $t = 2$:

$$\begin{aligned}
\Pr(S_2, Z_2 \mid S_0, S_1, Z_0, Z_1) &= \frac{\Pr(S_0, S_1, S_2, Z_0, Z_1, Z_2)}{\Pr(S_0, S_1, Z_0, Z_1)} \text{ (By difinition)} \\
&= \frac{\Pr(S_0, S_1, Z_0, Z_1) \cdot \Pr(S_2 \mid S_0, S_1, Z_0, Z_1) \cdot \Pr(Z_2 \mid S_0, S_1, S_2, Z_0, Z_1)}{\Pr(S_0, S_1, Z_0, Z_1)} \text{ (By Chain-Rule)} \\
&= \Pr(S_2 \mid S_0, S_1, Z_0, Z_1) \cdot \Pr(Z_2 \mid S_0, S_1, S_2, Z_0, Z_1) \qquad (3.72)
\end{aligned}$$

In general, when $t = k$:

$$\begin{aligned}
\Pr(S_k, Z_k \mid S_{0:k-1}, Z_{0:k-1}) &= \frac{\Pr(S_{0:k}, Z_{0:k})}{\Pr(S_{0:k-1}, Z_{0:k-1})} \text{ (By difinition)} \\
&= \frac{\Pr(S_{0:k-1}, Z_{0:k-1}) \cdot \Pr(S_k \mid S_{0:k-1}, Z_{0:k-1}) \cdot \Pr(Z_k \mid S_{0:k}, Z_{0:k-1})}{\Pr(S_{0:k-1}, Z_{0:k-1})} \text{ (By Chain-Rule)} \\
&= \Pr(S_k \mid S_{0:k-1}, Z_{0:k-1}) \cdot \Pr(Z_k \mid S_{0:k}, Z_{0:k-1}) \qquad (3.73)
\end{aligned}$$

Applying *HMM*'s rules 3.66 and 3.67 to Eq. 3.73, we can get

$$\begin{aligned}
\Pr(S_k, Z_k \mid S_{0:k-1}, Z_{0:k-1}) &= \Pr(S_k \mid S_{0:k-1}, Z_{0:k-1}) \cdot \Pr(Z_k \mid S_{0:k}, Z_{0:k-1}) \\
&= \Pr(S_k \mid S_{k-1}) \cdot \Pr(Z_k \mid S_k) \qquad (3.74)
\end{aligned}$$

Next, rewitting Eq. 3.70 by Eq. 3.74:

$$\begin{aligned}
\Pr(S_{0:n}, Z_{0:n}) &= \Pr(S_0, Z_0) \cdot \prod_{t=1}^{n} \Pr(S_t, Z_t \mid S_{0:t-1}, Z_{0:t-1}) \\
&= \Pr(S_0, Z_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1}) \cdot \Pr(Z_t \mid S_t) \\
&= \Pr(S_0) \cdot \Pr(Z_0 \mid S_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1}) \cdot \Pr(Z_t \mid S_t) \qquad (3.75)
\end{aligned}$$

Consequently, Eq. 3.69 is proved by above. ■

On the other hand, the probability over sequences of outputs given the states, $\Pr(Z_{0:n} \mid S_{0:n})$ can be calculated by Eq. 3.9:

$$\begin{aligned}
\Pr(Z_{0:n} \mid S_{0:n}) &= \frac{\Pr(S_{0:n}, Z_{0:n})}{\Pr(S_{0:n})} \\
&= \frac{\Pr(S_0, Z_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1}) \cdot \Pr(Z_t \mid S_t) \text{ (By Eq. 3.69)}}{\Pr(S_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1}) \text{ (By Eq. 3.68)}} \\
&= \frac{\Pr(S_0, Z_0)}{\Pr(S_0)} \cdot \prod_{t=1}^{n} \Pr(Z_t \mid S_t) \\
&= \Pr(Z_0 \mid S_0) \cdot \prod_{t=1}^{n} \Pr(Z_t \mid S_t) \\
&= \prod_{t=0}^{n} \Pr(Z_t \mid S_t)
\end{aligned} \tag{3.76}$$

This result is very intuitive because the outputs depended on the states. Meanwhile, the current state is only related with the previous state.

Summarizing the formulas we derive:

From Eq. 3.68: $\Pr(S_{0:n}) = \Pr(S_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1})$

From Eq. 3.69: $\Pr(S_{0:n}, Z_{0:n}) = \Pr(S_0) \cdot \Pr(Z_0 \mid S_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1}) \cdot \Pr(Z_t \mid S_t)$

From Eq. 3.76: $\Pr(Z_{0:n} \mid S_{0:n}) = \prod_{t=0}^{n} \Pr(Z_t \mid S_t)$

### 3.3.3 Recursive Property of HMM

*Hidden Markov Model* (HMM) has recursive properties in its formulas.

$$\begin{aligned}
\Pr(S_{0:n}) &= \Pr(S_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1}) \\
&= \Pr(S_n \mid S_{n-1}) \cdot \Pr(S_0) \cdot \prod_{t=1}^{n-1} \Pr(S_t \mid S_{t-1}) \\
&= \Pr(S_n \mid S_{n-1}) \cdot \Pr(S_{0:n-1})
\end{aligned} \tag{3.77}$$

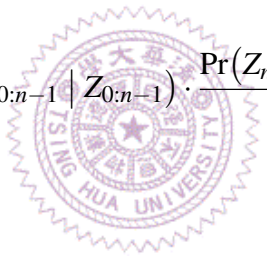
Similarly,

$$\begin{aligned}\Pr(Z_{0:n} \mid S_{0:n}) &= \prod_{t=0}^{n} \Pr(Z_t \mid S_t) \\ &= \Pr(Z_n \mid S_n) \cdot \prod_{t=0}^{n-1} \Pr(Z_t \mid S_t) \\ &= \Pr(Z_n \mid S_n) \cdot \Pr(Z_{0:n-1} \mid S_{0:n-1}) \end{aligned} \tag{3.78}$$

$$\begin{aligned}\Pr(S_{0:n}, Z_{0:n}) &= \Pr(S_0) \cdot \Pr(Z_0 \mid S_0) \cdot \prod_{t=1}^{n} \Pr(S_t \mid S_{t-1}) \cdot \Pr(Z_t \mid S_t) \\ &= \Pr(S_n \mid S_{n-1}) \cdot \Pr(Z_n \mid S_n) \cdot \Pr(S_0) \cdot \Pr(Z_0 \mid S_0) \cdot \prod_{t=1}^{n-1} \Pr(S_t \mid S_{t-1}) \cdot \Pr(Z_t \mid S_t) \\ &= \Pr(S_n \mid S_{n-1}) \cdot \Pr(Z_n \mid S_n) \cdot \Pr(S_{0:n-1}, Z_{0:n-1}) \end{aligned} \tag{3.79}$$

In fact, $\Pr(S_{0:n} \mid Z_{0:n})$ can also be computed recursively by

$$\Pr(S_{0:n} \mid Z_{0:n}) = \Pr(S_{0:n-1} \mid Z_{0:n-1}) \cdot \frac{\Pr(Z_n \mid S_n) \cdot \Pr(S_n \mid S_{n-1})}{\Pr(Z_n \mid Z_{0:n-1})} \tag{3.80}$$



*Proof.* of Eq. 3.80:

$$\begin{aligned}\Pr(S_{0:n} \mid Z_{0:n}) &= \Pr(S_{0:n} \mid Z_n, Z_{0:n-1}) \\ &= \frac{\Pr(Z_n \mid S_{0:n}, Z_{0:n-1}) \cdot \Pr(S_{0:n} \mid Z_{0:n-1})}{\Pr(Z_n \mid Z_{0:n-1})} \text{ (By Eq. 3.12)} \\ &= \frac{\Pr(Z_n \mid S_n) \cdot \Pr(S_{0:n} \mid Z_{0:n-1})}{\Pr(Z_n \mid Z_{0:n-1})} \text{ (By Eq. 3.66)} \end{aligned} \tag{3.81}$$

$$\begin{aligned}\Pr(S_{0:n} \mid Z_{0:n-1}) &= \Pr(S_n, S_{0:n-1} \mid Z_{0:n-1}) \\ &= \Pr(S_n \mid S_{0:n-1}, Z_{0:n-1}) \cdot \Pr(S_{0:n-1} \mid Z_{0:n-1}) \text{ (By Eq. 3.11)} \\ &= \Pr(S_n \mid S_{n-1}) \cdot \Pr(S_{0:n-1} \mid Z_{0:n-1}) \text{ (By Eq. 3.67)} \end{aligned} \tag{3.82}$$

Applying Eq. 3.83 to Eq. 3.81:

$$\begin{aligned}\Pr(S_{0:n} \mid Z_{0:n}) &= \frac{\Pr(Z_n \mid S_n) \cdot \Pr(S_{0:n} \mid Z_{0:n-1})}{\Pr(Z_n \mid Z_{0:n-1})} \\ &= \frac{\Pr(Z_n \mid S_n) \cdot \Pr(S_n \mid S_{n-1}) \cdot \Pr(S_{0:n-1} \mid Z_{0:n-1})}{\Pr(Z_n \mid Z_{0:n-1})} \\ &= \Pr(S_{0:n-1} \mid Z_{0:n-1}) \cdot \frac{\Pr(Z_n \mid S_n) \cdot \Pr(S_n \mid S_{n-1})}{\Pr(Z_n \mid Z_{0:n-1})}\end{aligned} \tag{3.83}$$

The Eq. 3.80 is proved above. ■

## 3.4 Monte-Carlo Method

### 3.4.1 Law of Large Numbers

In probability theorem, *law of large number*[4] describes the phenomena that: If we repeat one same experiment a large numbers of times, the average of the results is close to its expected value. The *law of large number* has two versions. they are called the *strong law of large numbers* , and the *weak law of large numbers*. Suppose that $\{X_i\} = X_1, X_2, \ldots, X_n$ is a sequence of *Independent and identically distributed random variables* (i.i.d). with finite expected value $\mathrm{E}[X_i] = \mu_X$ and finite non-zero variance $\mathrm{Var}[X_i] = \sigma_X^2$ , for all i. There is no correlation between the random variables $\{X_i\}$ because $\{X_i\}$ is a sequence of *i.i.d*. The average of $\{X_i\}$ is:

$$\overline{X_{1:n}} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i \tag{3.84}$$

The *weak law of large number* is defined as:

$$\lim_{n \to \infty} \Pr\left(\left|\overline{X_{1:n}} - \mu_X\right| > \varepsilon\right) = 0 \tag{3.85}$$

, $\varepsilon \in \mathbb{R}$ is any real number. The *strong law of large number* is defined as:

$$\Pr\left(\lim_{n \to \infty} \overline{X_{1:n}} = \mu_X\right) = 1 \tag{3.86}$$

*Proof.* of Eq. 3.85:

$$\begin{aligned}\mathrm{Var}\left[\overline{X_{1:n}}\right] &= \mathrm{Var}\left[\frac{1}{n}\cdot(X_1+X_2+\ldots+X_n)\right] = \frac{1}{n^2}\cdot\mathrm{Var}\left[X_1+X_2+\ldots+X_n\right] \\ &= \frac{1}{n^2}\cdot n\cdot\mathrm{Var}\left[X_i\right] = \frac{\mathrm{Var}\left[X_i\right]}{n} = \frac{\sigma_X^2}{n}\end{aligned} \tag{3.87}$$

$$\begin{aligned}\mathrm{E}\left[\overline{X_{1:n}}\right] &= \mathrm{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n}X_i\right] = \frac{1}{n}\cdot\mathrm{E}\left[\sum_{i=1}^{n}X_i\right] = \frac{1}{n}\cdot\sum_{i=1}^{n}\mathrm{E}\left[X_i\right] \\ &= \frac{1}{n}\cdot n\cdot\mathrm{E}\left[X_i\right] = \mathrm{E}\left[X_i\right] = \mu_X\end{aligned} \tag{3.88}$$

Using *Chebyshev's inequality*

$$\Pr\left(\left|Y-\mathrm{E}\left[Y\right]\right|\geq k\right)\leq\frac{\mathrm{Var}\left[Y\right]}{k^2}\iff\Pr\left(\left|Y-\mu_Y\right|\geq k\right)\leq\frac{\sigma_Y^2}{k^2} \tag{3.89}$$

, where $Y$ is a random variable with finite expected value $\mathrm{E}\left[Y\right]=\mu_Y$ and finite non-zero variance $\mathrm{Var}\left[Y\right]=\sigma_Y^2$ , $k\in\mathbb{R}$ is any real number.

Then, applying $\overline{X_{1:n}}$ to Eq. 3.89

$$\Pr\left(\left|\overline{X_{1:n}}-\mathrm{E}\left[\overline{X_{1:n}}\right]\right|\geq k\right)\leq\frac{\mathrm{Var}\left[\overline{X_{1:n}}\right]}{k^2}\iff\Pr\left(\left|\overline{X_{1:n}}-\mu_X\right|\geq k\right)\leq\frac{\sigma_X^2}{n\cdot k^2} \tag{3.90}$$

Finally, replacing $k$ by $\varepsilon$, the equation above can be rewritten into:

$$\Pr\left(\left|\overline{X_{1:n}}-\mu_X\right|\geq\varepsilon\right)\leq\frac{\sigma_X^2}{n\cdot\varepsilon^2} \tag{3.91}$$

When $n\to\infty$, $\frac{\sigma_X^2}{n\cdot\varepsilon^2}\to 0$ , so Eq. 3.85 can be proved. ■

### 3.4.2 Monte Carlo Estimate

**Definition 3.4.1.** *Monte Carlo* is the art of approximating an expectation by sample mean of a function of simulated random variables.

Considering a random variable $x$ with *PDF*/*PMF* $\Pr(x)$ and *probability space* is $S$. The expected value of a function $g$, $\mathrm{E}_{\Pr(x)}\left[g(x)\right]$, is shortly denoted by $\mathrm{E}\left[g(x)\right]$ here and it is defined as Eq. 3.1. If the expected value is hard to calculate, then we can approximate this value by the empirical measure

, namely *Monte Carlo estimate*. The following contents explain the method of *Monte Carlo estimate*.

First, we can get the average of function $g$ over $N$ samples from empirical measures:

$$\overline{g(x)} = \frac{1}{N} \cdot \sum_{i=1}^{N} g(x_i) \tag{3.92}$$

The expected value of average of function $g$ is

$$\mathrm{E}\left[\overline{g(x)}\right] = \mathrm{E}\left[\frac{1}{N} \cdot \sum_{i=1}^{N} g(x_i)\right] = \frac{1}{N} \cdot \mathrm{E}\left[\sum_{i=1}^{N} g(x_i)\right] = \frac{1}{N} \cdot \sum_{i=1}^{N} \cdot \mathrm{E}\left[g(x_i)\right] = \mathrm{E}\left[g(x)\right] \tag{3.93}$$

Next, by *weak law of large number* Eq. 3.85, and Eq. 3.93, we can get:

$$\lim_{N\to\infty} \Pr\left(\left|\overline{g(x)} - \mathrm{E}\left[\overline{g(x)}\right]\right| > \varepsilon\right) = \lim_{N\to\infty} \Pr\left(\left|\overline{g(x)} - \mathrm{E}\left[g(x)\right]\right| > \varepsilon\right) = 0 \tag{3.94}$$

, $\varepsilon \in \mathbb{R}$ is any real number.

Finally, from Eq. 3.94, we know that

$$\overline{g(x)} = \frac{1}{N} \cdot \sum_{i=1}^{N} g(x_i) \overset{N\to\infty}{\approx} \mathrm{E}\left[g(x)\right] \tag{3.95}$$

, the $\overline{g(x)}$ is called *Monte Carlo estimator* of $\mathrm{E}\left[g(x)\right]$.

On the other hands, one character of *Monte Carlo estimator* is that its variance shrinks $\propto \frac{1}{N}$ because

$$\mathrm{Var}\left[\overline{X_{1:N}}\right] = \frac{\mathrm{Var}\left[X_i\right]}{N} \propto \frac{1}{N} \text{ (By Eq. 3.87)} \tag{3.96}$$

The most interesting thing is that if we replace $g(x)$ by dirac measure $\delta_x(A)$ Eq. 3.7, then Eq. 3.95 can lead the following formula:

$$\overline{\delta_x(A)} = \frac{1}{N} \cdot \sum_{i=1}^{N} \delta_{x_i}(A) \overset{N\to\infty}{\approx} \mathrm{E}_{\Pr(x)}\left[\delta_x(A)\right] = \Pr(x \in A) \text{ (By Eq. 3.8)} \tag{3.97}$$

### 3.4.3 Monte Carlo Method

*Monte Carlo* method(or Monte-Carlo experiment)[6] is a board class of computational algorithms, which repeatedly sample from simulations randomly to obtain the distribution of an unknown probabilistic entity. The *Monte Carlo* methods vary, but tend to follow this pattern:

1. Define the domain of possible inputs
2. Sample randomly from the probability distribution over the domain defined
3. Perform a deterministic computation on the inputs and repeated sampling in large numbers of times
4. Aggregate the results

## 3.5 Importance Sampling

Give a random variable $x$ with probability distribution $\Pr(x)$ in *probability space* S , the expected value of function $f$ can be "approximated" by the average of $f(x)$ computed from the samples $\{x_1, x_2, \ldots, x_n\}$ generated from $\Pr(x)$ by *Monte Carlo estimate* Eq. 3.95 if the expected value is hard to be calculated by the formula Eq. 3.1.

However, if probability distribution $\Pr(x)$ is "intractable" that we can not easily draw the samples from $\Pr(x)$ , how can the expected value be approximated? An alternative classical method is called *importance sampling*

[2] is a online course note of computational statistics and data analysis offered at the University of Waterloo , which can provide lots of knowledge about *importance sampling*. Also, another useful way to learn more about *importance sampling* is to read [18], which is an online electric book authored by Prof. Art B. Owen at Stanford university. *Sequential importance sampling* is a method to apply *importance sampling* in a sequence of probability distribution. [10], the chapter 1 in book *Sequential Monte Carlo Methods in Practice* authored by Arnaud Doucet, et al , is a good tutorial for beginner to learn *sequential importance sampling*. Furthermore, more mathematical detail and varieties of *sequential Monte Carlo* can be founded in [12, 11, 7].

### 3.5.1 Basic Idea : Unnormalized/Unbiased Importance Sampling

In order to approximate the expected value of function $f$, $f(x)$, under one specific probability distribution $\Pr(x)$ by *Monte Carlo estimate* Eq. 3.95, the samples generated from $\Pr(x)$ are essential. Nevertheless, if it is not easy to draw samples from *target distribution* $\Pr(x)$, we introduce another distribution $\pi(x)$ called *proposal distribution* ( or *importance distribution* ) that we will draw samples

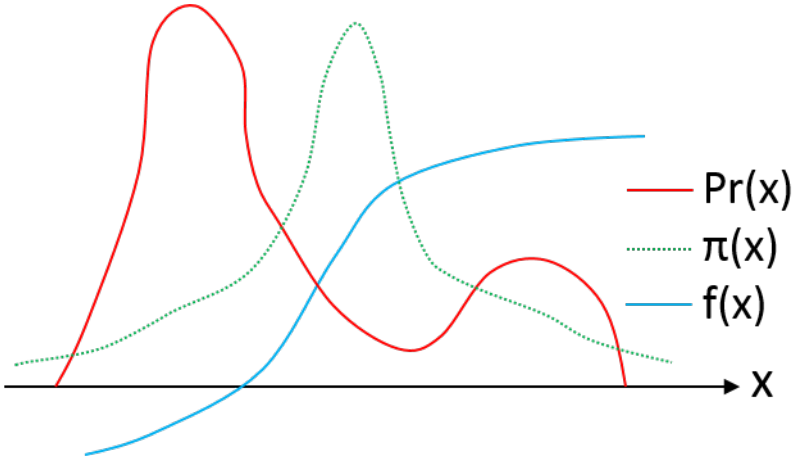



Figure 3.15: Importance sampling

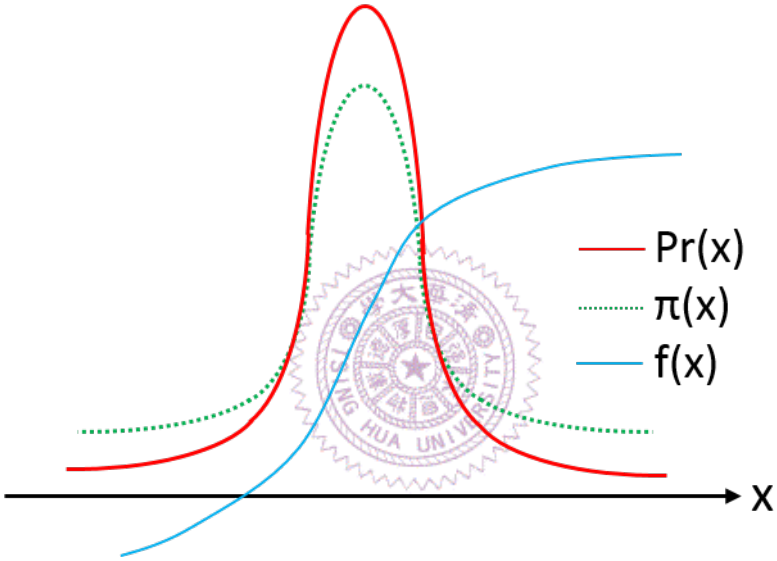



Figure 3.16: The good proposal distribution

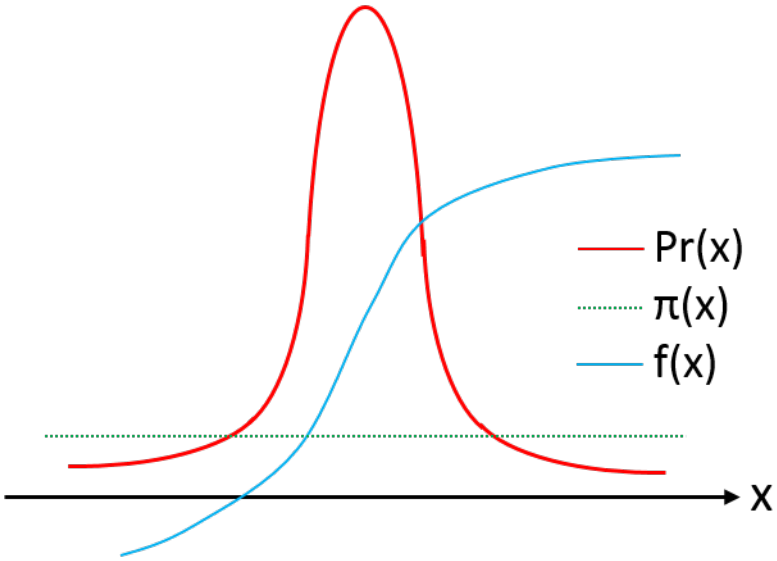



Figure 3.17: The bad proposal distribution

$\{x_1, x_2, \ldots, x_n\}$ from instead without much effort. One required condition of $\pi(x)$ is:

$$\Pr(x) > 0 \implies \pi(x) > 0 \tag{3.98}$$

Thus,

$$\begin{aligned} \mathrm{E}_{\Pr(x)}[f(x)] &= \int_{x \in S} f(x) \cdot \Pr(x) \cdot dx \\ &= \int_{x \in S} [f(x) \cdot \frac{\Pr(x)}{\pi(x)}] \cdot \pi(x) \cdot dx \\ &= \mathrm{E}_{\pi(x)}[f(x) \cdot \frac{\Pr(x)}{\pi(x)}] \end{aligned} \tag{3.99}$$

, for continuous

By *Monte Carlo estimate* Eq. 3.95,

$$\mathrm{E}_{\Pr(x)}[f(x)] = \mathrm{E}_{\pi(x)}[f(x) \cdot \frac{\Pr(x)}{\pi(x)}] \overset{N \to \infty}{\approx} \frac{1}{N} \cdot \sum_{i=1}^{N} f(x_i) \cdot \frac{\Pr(x_i)}{\pi(x_i)} \tag{3.100}$$

Defining *importance weight* $r(x)$ as

$$r(x) = \frac{\Pr(x)}{\pi(x)} \tag{3.101}$$

, then the Eq. 3.100 can be rewritten into:

$$\mathrm{E}_{\Pr(x)}[f(x)] = \mathrm{E}_{\pi(x)}[f(x) \cdot r(x)] \overset{N \to \infty}{\approx} \frac{1}{N} \cdot \sum_{i=1}^{N} f(x_i) \cdot r(x_i) \tag{3.102}$$

From the above, if we know the specific relation between $\Pr(x)$ and $\pi(x)$ , $r(x)$, then the equations can help us to estimate the expected value $\mathrm{E}_{\Pr(x)}[f(x)]$. However, $r(x)$ can not be described elaborately in most of the time. One approach is that we can choose one good *proposal distribution*, $\pi(x)$, **as close as possible** to being proportional to $\Pr(x)$ or $f(x) \cdot \Pr(x)$.

In conclusion, the steps to estimate expected value by *importance sampling* follows:

1. Define the expected value of the function $f$ you want, $\mathrm{E}_{\Pr(x)}[f(x)]$
2. Choose one *proposal distribution* $\pi(x) \propto \Pr(x)$ or $|f(x)| \cdot \Pr(x)$
3. Generate samples $\{x_1, x_2, \ldots, x_i, \ldots, x_N\}$ from $\pi(x)$

4. Calculate the *importance weights* $r(x_i)$ for all $i = 1, 2, \ldots, N$

5. Estimate the expected value by $\frac{1}{N} \cdot \sum_{i=1}^{N} f(x_i) \cdot r(x_i)$

The steps above can be summarized into Alg. 1.

**Algorithm 1:** Importance Sampling

**input**: $f(x)$, *target distribution* $\Pr(x)$
**output**: Expected value $\mathrm{E}_{\Pr(x)}[f(x)]$

Set number of samples $N$;
Choose one *proposal distribution* $\pi(x)$. Desirably, $\pi(x) \propto \Pr(x)$ or $|f(x)| \cdot \Pr(x)$;
Set $sum \leftarrow 0$;
**for** $i \in [1, N]$ **do**
    Generate sample $x_i$ from *proposal distribution* $\pi(x)$;
    Calculate the importance weight $r(x_i)$ of $x_i$ , $r(x_i) = \frac{\Pr(x_i)}{\pi(x_i)}$;
    $sum = sum + r(x_i) \cdot f(x_i)$;
**end**
**return** $\frac{sum}{N}$

### 3.5.2 Deeper Look into Importance Sampling

*Importance sampling* gives more weight for those samples $x_i$ for which $\Pr(x_i) > \pi(x_i)$ and gives less weight for those samples $x_i$ for which $\Pr(x_i) < \pi(x_i)$. By *Importance sampling*, the expected value can be "approximated" by iteratively producing weighted samples $\{[x_1, r(x_1)], [x_2, r(x_2)], \ldots, [x_n, r(x_n)]\}$, where $x_i$ is the sample point and $r(x_i)$ is the weight of sample $x_i$. Ideally, we should pick a good approximation $\pi(x)$ to $\Pr(x)$, which leads *importance weight* $r(x_i)$, $r(x_i) = \frac{\Pr(x_i)}{\pi(x_i)}$ (by Eq. 3.101), are roughly equal for all i. Otherwise, the variance of those samples are likely too large. In other words, the sum in Eq. 3.102, $\sum_{i=1}^{N} f(x_i) \cdot r(x_i)$, may be dominated by a few of samples whose weight $r(x_i)$ are larger than others, and cause the inaccuracy of estimation of this expected value. Therefore, it is a good idea to inspect the variance of those samples.

To choose a good *proposal distribution* $\pi(x)$ to **minimize** variance of samples, we should take a look to the variance of those samples first.

$$\begin{aligned}
\mathrm{Var}_{\pi(x)}[f(x) \cdot r(x)] &= \mathrm{E}_{\pi(x)}[(f(x) \cdot r(x))^2] - (\mathrm{E}_{\pi(x)}[f(x) \cdot r(x)])^2 \text{ (By Eq. 3.4)} \\
&= \mathrm{E}_{\pi(x)}[(f(x) \cdot r(x))^2] - (\mathrm{E}_{\Pr(x)}[f(x)])^2 \text{ (By Eq. 3.99)} \qquad (3.103)
\end{aligned}$$

$\mathrm{E}_{\Pr(x)}\left[f(x)\right]$ does not depend on $\pi(x)$. Therefore, minimizing $\mathrm{E}_{\pi(x)}\left[(f(x)\cdot r(x))^2\right]$ is what we only need.

$$\begin{aligned}\mathrm{E}_{\pi(x)}\left[(f(x)\cdot r(x))^2\right] &= \int_{x\in S}\pi(x)\cdot f(x)^2\cdot r(x)^2\cdot dx \\ &= \int_{x\in S}\pi(x)\cdot f(x)^2\cdot\frac{\Pr(x)^2}{\pi(x)^2}\cdot dx \\ &= \int_{x\in S}f(x)^2\cdot\frac{\Pr(x)^2}{\pi(x)}\cdot dx \end{aligned} \tag{3.104}$$

The following equation can help us to find the minimal value of $\mathrm{E}_{\pi(x)}\left[(f(x)\cdot r(x))^2\right]$

$$\begin{aligned}\mathrm{Var}_{\pi(x)}\left[\left|f(x)\cdot r(x)\right|\right] &= \mathrm{E}_{\pi(x)}\left[\left|f(x)\cdot r(x)\right|^2\right]-(\mathrm{E}_{\pi(x)}\left[\left|f(x)\cdot r(x)\right|\right])^2 \\ &= \mathrm{E}_{\pi(x)}\left[(f(x)\cdot r(x))^2\right]-(\mathrm{E}_{\pi(x)}\left[\left|f(x)\cdot r(x)\right|\right])^2\geq 0\end{aligned} \tag{3.105}$$

$$\implies \mathrm{E}_{\pi(x)}\left[(f(x)\cdot r(x))^2\right]\geq(\mathrm{E}_{\pi(x)}\left[\left|f(x)\cdot r(x)\right|\right])^2$$

Thus, $\mathrm{E}_{\pi(x)}\left[(f(x)\cdot r(x))^2\right]$ has minimal value when $\mathrm{E}_{\pi(x)}\left[(f(x)\cdot r(x))^2\right]=(\mathrm{E}_{\pi(x)}\left[\left|f(x)\cdot r(x)\right|\right])^2$

$$\begin{aligned}(\mathrm{E}_{\pi(x)}\left[\left|f(x)\cdot r(x)\right|\right])^2 &= (\int_{x\in S}\left|f(x)\cdot r(x)\right|\cdot\pi(x)\cdot dx)^2 \\ &= (\int_{x\in S}\left|f(x)\cdot\frac{\Pr(x)}{\pi(x)}\right|\cdot\pi(x)\cdot dx)^2 \\ &= (\int_{x\in S}\left|f(x)\right|\cdot\frac{\Pr(x)}{\pi(x)}\cdot\pi(x)\cdot dx)^2 \\ &= (\int_{x\in S}\left|f(x)\right|\cdot\Pr(x)\cdot dx)^2\end{aligned} \tag{3.106}$$

From the equations above, we can get the following conclusions:

1. *Proposal distribution* should be chosen such that is has a **thicker** tail than *target distribution*. By Eq. 3.104, we know that : $\pi(x)\to 0 \implies \mathrm{E}_{\pi(x)}\left[(f(x)\cdot r(x))^2\right]\to\infty$ ,this occur if $\pi(x)$ has a *thinner tail* than $\Pr(x)$. That is, $f(x)^2\cdot\frac{\Pr(x)^2}{\pi(x)}$ could be infinitely large when $\pi(x)$ has a *thinner tail* than $\Pr(x)$. Therefore, the $\pi(x)$ should be chosen such that is has a thicker tail than $\Pr(x)$.

2. *Proposal distribution* $\pi(x)$ should be similar to shape of $|f(x)| \cdot \Pr(x)$. To put it in another way, $\pi(x) \propto |f(x)| \cdot \Pr(x)$. The choice of *proposal distribution* $\pi(x)$ minimizing variance of samples is $\frac{|f(x)| \cdot \Pr(x)}{\int_{x \in S} |f(x)| \cdot \Pr(x) \cdot dx}$. $\mathrm{E}_{\pi(x)}\left[(f(x) \cdot r(x))^2\right]$ has minimal value when Eq. 3.104 = Eq. 3.106. If you replace $\frac{|f(x)| \cdot \Pr(x)}{\int_{x \in S} |f(x)| \cdot \Pr(x) \cdot dx}$ in Eq. 3.104, you can see what you want. However, this is impossible or very difficult to compute in practice. A alternative way is to choose one $\pi(x) \propto |f(x)| \cdot \Pr(x)$

On the other hand, we should notice that the *importance weight* $r(x)$ can be utilized to form a *control variate*. The following equations can demonstrate this quality more.

$$\mathrm{E}_{\pi(x)}\left[r(x)\right] = \int_{x \in S} r(x) \cdot \pi(x) \cdot dx = \int_{x \in S} \frac{\Pr(x)}{\pi(x)} \cdot \pi(x) \cdot dx = \int_{x \in S} \Pr(x) \cdot dx = 1 \tag{3.107}$$

In addition, by *Monte Carlo estimate* Eq. 3.95, we know:

$$\mathrm{E}_{\pi(x)}\left[r(x)\right] = 1 \overset{N \to \infty}{\approx} \frac{1}{N} \cdot \sum_{i=1}^{N} r(x_i) \tag{3.108}$$

Therefore, the average of importance weight $r(x)$ can be used as a *control variate* whose expectation is known to be one.

### 3.5.3 (Self-)Normalized/Weight Importance Sampling

For many applications, it is possible to apply *importance sampling* to estimate a expected value of function $f$ under $\Pr(x)$ with probability space $S$, $\mathrm{E}_{\Pr(x)}\left[f(x)\right]$, but we only know an unnormalized version $\widehat{\Pr(x)}$ of $\Pr(x)$. For example, we want to estimate the expected value under $\Pr(A \mid B)$ but we only know $\Pr(A,B)$, where $\Pr(A,B)$ is unnormalized version of $\Pr(A \mid B)$ by $\Pr(A \mid B) = \frac{\Pr(A,B)}{\Pr(B)}$. In practice, our actual *target distribution* $\Pr(x)$ is often unnormalized. Here,

$$\Pr(x) = \frac{1}{Z_p} \cdot \widehat{\Pr(x)} \tag{3.109}$$

, where $\frac{1}{Z_p}$ is normalizing constant of $\Pr(x)$. Of course, $Z_p$ can be expressed as $\sum_{x \in S} \Pr(x)$ or $\int_{x \in S} \Pr(x) \cdot dx$ by Eq. 3.50 or Eq. 3.46 (Normalizing constant is already mentioned in Bayesian Rule 3.1.7). However, $\Pr(x)$ is "intractable", so $\int_{x \in S} \Pr(x) \cdot dx$ can not be computed directly, and that is

the reason why we use *importance sampling* to "approximate" the expected value. For handling this problem, another approach called *(self-)normalized importance sampling* (or *weight importance sampling*) was proposed. Starting from defining an *associated unnormalized weight* $\widehat{r(x)}$ as following:

$$\widehat{r(x)} = \frac{\widehat{\Pr(x)}}{\pi(x)} = \frac{Z_p \cdot \Pr(x)}{\pi(x)} = Z_p \cdot r(x) \tag{3.110}$$

After defining the $\widehat{r(x)}$, the expected value of $\widehat{r(x)}$ under $\pi(x)$ can be calculated.

$$\begin{aligned} \mathrm{E}_{\pi(x)}\left[\widehat{r(x)}\right] &= \int_{x\in S} \widehat{r(x)} \cdot \pi(x) \cdot dx = \int_{x\in S} \frac{Z_p \cdot \Pr(x)}{\pi(x)} \cdot \pi(x) \cdot dx \\ &= Z_p \cdot \int_{x\in S} \Pr(x) \cdot dx = Z_p \cdot 1 = Z_p \end{aligned} \tag{3.111}$$

Then, the expected value can be written into:

$$\begin{aligned} \mathrm{E}_{\Pr(x)}\left[f(x)\right] &= \int_{x\in S} f(x) \cdot \Pr(x) \cdot dx = \int_{x\in S} f(x) \cdot \frac{\Pr(x)}{\pi(x)} \cdot \pi(x) \cdot dx \\ &= \int_{x\in S} f(x) \cdot \frac{1}{Z_p} \cdot \frac{\widehat{\Pr(x)}}{\pi(x)} \cdot \pi(x) \cdot dx \text{ (By Eq. 3.109)} \\ &= \frac{1}{Z_p} \cdot \int_{x\in S} f(x) \cdot \frac{\widehat{\Pr(x)}}{\pi(x)} \cdot \pi(x) \cdot dx = \frac{1}{Z_p} \cdot \int_{x\in S} f(x) \cdot \widehat{r(x)} \cdot \pi(x) \cdot dx \text{ (By Eq. 3.110)} \\ &= \frac{1}{Z_p} \cdot \mathrm{E}_{\pi(x)}\left[f(x) \cdot \widehat{r(x)}\right] = \frac{\mathrm{E}_{\pi(x)}\left[f(x) \cdot \widehat{r(x)}\right]}{\mathrm{E}_{\pi(x)}\left[\widehat{r(x)}\right]} \text{ (By Eq. 3.111)} \end{aligned} \tag{3.112}$$

Next, appling *Monte Carlo estimate* Eq. 3.95 in numerator and denominator of Eq. 3.112 separately.

$$\mathrm{E}_{\Pr(x)}\left[f(x)\right] = \frac{\mathrm{E}_{\pi(x)}\left[f(x) \cdot \widehat{r(x)}\right]}{\mathrm{E}_{\pi(x)}\left[\widehat{r(x)}\right]} \overset{N\to\infty}{\approx} \frac{\frac{1}{N} \cdot \sum_{i=1}^{N} f(x_i) \cdot \widehat{r(x_i)}}{\frac{1}{N} \cdot \sum_{j=1}^{N} \widehat{r(x_j)}} = \frac{\sum_{i=1}^{N} f(x_i) \cdot \widehat{r(x_i)}}{\sum_{j=1}^{N} \widehat{r(x_j)}} \tag{3.113}$$

The right side of equation above can be simplified as:

$$\begin{aligned} \frac{\sum_{i=1}^{N} f(x_i) \cdot \widehat{r(x_i)}}{\sum_{j=1}^{N} \widehat{r(x_j)}} &= \frac{\sum_{i=1}^{N} f(x_i) \cdot Z_p \cdot r(x_i)}{\sum_{j=1}^{N} Z_p \cdot r(x_j)} \text{ (By Eq. 3.110)} \\ &= \frac{Z_p \cdot \sum_{i=1}^{N} f(x_i) \cdot r(x_i)}{Z_p \cdot \sum_{j=1}^{N} r(x_j)} = \frac{\sum_{i=1}^{N} f(x_i) \cdot r(x_i)}{\sum_{j=1}^{N} r(x_j)} \end{aligned} \tag{3.114}$$

That is,

$$\mathrm{E}_{\Pr(x)}\left[f(x)\right] \overset{N\to\infty}{\approx} \frac{\sum_{i=1}^{N} f(x_i)\cdot r(x_i)}{\sum_{j=1}^{N} r(x_j)} \tag{3.115}$$

Let us define a *normalized importance weights* $w(x_i)$

$$w(x_i) = \frac{\widehat{r(x_i)}}{\sum_{j=1}^{N}\widehat{r(x_j)}} = \frac{r(x_i)}{\sum_{j=1}^{N} r(x_j)} \tag{3.116}$$

, then Eq. 3.152 can be rewritten into:

$$\mathrm{E}_{\Pr(x)}\left[f(x)\right] \overset{N\to\infty}{\approx} \sum_{i=1}^{N} w(x_i)\cdot f(x_i) \tag{3.117}$$

Actually, *(self-)normalized importance sampling* is a general version of *importance sampling*. If we know *target distribution* $\Pr(x)$ directly , then $Z_p = 1$ and $\widehat{\Pr(x)} = \Pr(x), \widehat{r(x)} = r(x)$ in all the equations above. Therefore, $\mathrm{E}_{\pi(x)}\left[\widehat{r(x)}\right] = \mathrm{E}_{\pi(x)}\left[r(x)\right] = 1 \overset{N\to\infty}{\approx} \frac{1}{N}\cdot\sum_{i=1}^{N} r(x_i)$ by Eq. 3.108. This implies $\sum_{i=1}^{N} r(x_i) \overset{N\to\infty}{\approx} N$. Replace $\sum_{i=1}^{N} r(x_i)$ by $N$ , the Eq. 3.115 : $\mathrm{E}_{\Pr(x)}\left[f(x)\right] = \frac{\sum_{i=1}^{N} f(x_i)\cdot r(x_i)}{\sum_{j=1}^{N} r(x_j)} = \mathrm{E}_{\Pr(x)}\left[f(x)\right] = \frac{1}{N}\cdot\sum_{i=1}^{N} f(x_i)\cdot r(x_i)$ is same as *importance sampling* Eq. 3.102.

In summary, using *Normalized importance sampling* to estimate the expected value follows:

1. Define the expected value of the function $f$ under the *target distribution* $\Pr(x)$ , $\mathrm{E}_{\Pr(x)}\left[f(x)\right]$ , and we may only know an unnomalized version $\widehat{\Pr(x)}$ of $\Pr(x)$.
2. Choose one *proposal distribution* $\pi(x) \propto \widehat{\Pr(x)}$ or $\left|f(x)\right|\cdot\widehat{\Pr(x)}$
   ($\implies \pi(x) \propto \Pr(x)$ or $\left|f(x)\right|\cdot\Pr(x)$)
3. Generate samples $\{x_1, x_2, \ldots, x_i, \ldots, x_N\}$ from $\pi(x)$
4. Calculate the *normalized importance weights* $w(x_i)$ for all $i = 1, 2, \ldots, N$
5. Estimate the expected value by $\sum_{i=1}^{N} w(x_i)\cdot f(x_i)$

The steps above can be summarized into Alg. 2. In practice, one interesting application of *(self-)normalized importance sampling* or *importance sampling* is to estimate an unknown probability

distribution by *dirac* measure:

$$\mathrm{E}_{\Pr(x)}\left[\delta_x(A)\right] = \Pr(x \in A) \overset{N\to\infty}{\approx} \frac{\sum_{i=1}^{N} \delta_{x_i}(A) \cdot \widehat{r(x_i)}}{\sum_{j=1}^{N} \widehat{r(x_j)}} \tag{3.118}$$

**Algorithm 2:** Normalized Importance Sampling

**input**: $f(x)$, unnormalized version $\widehat{\Pr(x)}$ of *target distribution* $\Pr(x)$
**output**: Expected value $\mathrm{E}_{\Pr(x)}\left[f(x)\right]$

//: Initialization
Set number of samples $N$;
Choose one *proposal distribution* $\pi(x) \propto \widehat{\Pr(x)}$ or $\left|f(x)\right| \cdot \widehat{\Pr(x)}$ , which implies $\pi(x) \propto \Pr(x)$ or $\left|f(x)\right| \cdot \Pr(x)$;
Set $sum \leftarrow 0$;
//: Importance Sampling Step
**for** $i \in [1,N]$ **do**
    Generate sample $x_i$ from *proposal distribution* $\pi(x)$;
    Calculate $\widehat{r(x_i)}$ of $x_i$ , $\widehat{r(x_i)} = \dfrac{\widehat{\Pr(x_i)}}{\pi(x_i)}$;
**end**
//: Normalize
**for** $i \in [1,N]$ **do**
    Get *normalized importance weights* of $x_i$ : $w(x_i) = \dfrac{\widehat{r(x_i)}}{\sum_{j=1}^{N} \widehat{r(x_j)}}$;
**end**
**return** $\sum_{i=1}^{N} w(x_i) \cdot f(x_i)$

### 3.5.4 Importance Sampling Diagnostics : Effective Sample Size

*(self-)normalized importance sampling* method estimate the expected value by unequally weighted samples. In a extreme situation, a few of the samples whose weights are possibly larger than others. Thus, the estimation would be dominated by these samples , and it may cause the inaccuracy of estimation. To avoid this situation, we would like to make a diagnosis when the samples may be problematic. *Effective Sample Size* $N_e$ can indicate whether the samples are representable enough or not, where $N_e$ is given by:

$$N_e = \frac{\overline{\widehat{r(x_{1:N})}}^{2}}{\overline{\widehat{r(x_{1:N})}^{2}}} \cdot N \tag{3.119}$$

The numerator $\overline{\widehat{r(x_{1:N})}}^2$ denotes the "square of the average" of $\widehat{r(x_i)}$ from $i=1$ to $N$ and denominator $\overline{\widehat{r(x_{1:N})}^2}$ is the "average of square" of $\widehat{r(x_i)}$ from $i=1$ to $N$. $\frac{\overline{\widehat{r(x_{1:N})}}^2}{\overline{\widehat{r(x_{1:N})}^2}}$ is the ratio between the "square of average of weights" and the "average of the square of weights" This ratio represents the differences between samples' weights.

If the weights of samples $\widehat{r(x_i)}$ is roughly equal to each other , then $\frac{\overline{\widehat{r(x_{1:N})}}^2}{\overline{\widehat{r(x_{1:N})}^2}}$ would be roughly equal to 1. Otherwise, a few of samples' weight $\widehat{r(x_i)}$ are larger than others, would make this ratio much smaller than 1. For example, $\langle 3,4,5,4,3\rangle$ are the weights of five samples. The square of average of weights is $(19/5)^2 = 14.44$ and the average of the square of these samples' weights is $(3^2+4^2+5^2+4^2+3^2)/5 = 75/5 = 15$. The ratio between "the square of average of weights" and "the average of the square of weights" is $14.44/15 \approx 0.963$ , the *Effective Sample Size* $N_e = 0.963 \cdot 5 = 4.815$. On the other hand, if weights of five samples is $\langle 1,1,15,1,1\rangle$, the square of average of weights is $(19/5)^2 = 14.44$ and the average of the square of these samples' weights is $(1^2+1^2+15^2+1^2+1^2)/5 = 229/5 = 45.8$. The ratio between "the square of average of weights" and "the average of the square of weights" is $14.44/45.8 \approx 0.315$ , the *Effective Sample Size* $N_e = 0.315 \cdot 5 = 1.575$.

Briefly, *Effective Sample Size* , described by the original sample size $N$ multiply the ratio that represent the differences between samples' weights , indicates how many samples are effective now. In fact, the formula of *Effective Sample Size* $N_e$ appears in many formats. The following formulas can be derived by the definition above easily. Except for , $N_e$ also can be defined as:

$$\begin{aligned}
N_e &= \frac{\overline{\widehat{r(x_{1:N})}}^2}{\overline{\widehat{r(x_{1:N})}^2}} \cdot N = \frac{(\frac{1}{N} \cdot \sum_{i=1}^N \widehat{r(x_i)})^2}{\frac{1}{N} \cdot \sum_{j=1}^N \widehat{r(x_j)}^2} \cdot N \\
&= \frac{(\frac{1}{N} \cdot \sum_{i=1}^N Z_p \cdot r(x_i))^2}{\frac{1}{N} \cdot \sum_{j=1}^N (Z_p \cdot r(x_j))^2} \cdot N = \frac{Z_p^2 \cdot (\frac{1}{N} \cdot \sum_{i=1}^N r(x_i))^2}{Z_p^2 \cdot \frac{1}{N} \cdot \sum_{j=1}^N r(x_j)^2} \cdot N \\
&= \frac{(\frac{1}{N} \cdot \sum_{i=1}^N r(x_i))^2}{\frac{1}{N} \cdot \sum_{j=1}^N r(x_j)^2} \cdot N = \frac{\overline{r(x_{1:N})}^2}{\overline{r(x_{1:N})^2}} \cdot N
\end{aligned} \tag{3.120}$$

Another formula of $N_e$ is $\frac{(\sum_{i=1}^N r(x_i))^2}{\sum_{j=1}^N r(x_j)^2}$. It can be easily proved by following:

*Proof.*

$$
\begin{aligned}
N_e &= \frac{\overline{\widehat{r(x_{1:N})}}^2}{\widehat{r(x_{1:N})}^2} \cdot N = \frac{(\frac{1}{N} \cdot \sum_{i=1}^N \widehat{r(x_i)})^2}{\frac{1}{N} \cdot \sum_{j=1}^N \widehat{r(x_j)}^2} \cdot N \\
&= \frac{\frac{1}{N^2} \cdot (\sum_{i=1}^N \widehat{r(x_i)})^2}{\frac{1}{N} \cdot \sum_{j=1}^N \widehat{r(x_j)}^2} \cdot N = \frac{(\sum_{i=1}^N \widehat{r(x_i)})^2}{\sum_{j=1}^N \widehat{r(x_j)}^2}
\end{aligned} \tag{3.121}
$$

Next, removing the normalizing constant of $\widehat{r(x_i)}$:

$$
N_e = \frac{(\sum_{i=1}^N \widehat{r(x_i)})^2}{\sum_{j=1}^N \widehat{r(x_j)}^2} = \frac{(\sum_{i=1}^N Z_p \cdot r(x_i))^2}{\sum_{j=1}^N (Z_p \cdot r(x_j))^2} = \frac{Z_p^2 \cdot (\sum_{i=1}^N r(x_i))^2}{Z_p^2 \cdot \sum_{j=1}^N r(x_j)^2} = \frac{(\sum_{i=1}^N r(x_i))^2}{\sum_{j=1}^N r(x_j)^2} \tag{3.122}
$$

■

In fact, the formula above can lead another definition of $N_e$ used frequently. In many textbooks, the $N_e$ is defined by $\frac{N}{\mathrm{Var}_{\pi(x)}[r(x)]+1}$. We now prove it by following:

*Proof.*

$$
\begin{aligned}
N_e = \frac{N \cdot (\frac{1}{N} \cdot \sum_{i=1}^N r(x_i))^2}{\frac{1}{N} \cdot \sum_{j=1}^N r(x_j)^2} &\overset{N\to\infty}{\approx} \frac{N \cdot (\frac{1}{N} \cdot \sum_{i=1}^N r(x_i))^2}{\frac{1}{N} \cdot \sum_{j=1}^N r(x_j)^2} = \frac{N \cdot (\mathrm{E}_{\pi(x)}[r(x)])^2}{\mathrm{E}_{\pi(x)}[r(x)^2]} \text{ (By Eq. 3.95)} \\
&= \frac{N \cdot 1^2}{\mathrm{E}_{\pi(x)}[r(x)^2]} \text{ (By Eq. 3.108)} \\
&= \frac{N}{\mathrm{E}_{\pi(x)}[r(x)^2]}
\end{aligned} \tag{3.123}
$$

By Eq. 3.4

$$
\begin{aligned}
\mathrm{E}_{\pi(x)}[r(x)^2] &= \mathrm{Var}_{\pi(x)}[r(x)] + (\mathrm{E}_{\pi(x)}[r(x)])^2 \\
&= \mathrm{Var}_{\pi(x)}[r(x)] + 1 \text{ (By Eq. 3.108)}
\end{aligned} \tag{3.124}
$$

Thus, the popular version of $N_e$ is:

$$N_e = \frac{N}{\mathrm{E}_{\pi(x)}\left[r(x)^2\right]} = \frac{N}{\mathrm{Var}_{\pi(x)}\left[r(x)\right]+1} \leq N \tag{3.125}$$

■

From above, the ratio of $Ne$, $\frac{\overline{\widehat{r(x_{1:N})}}^2}{\overline{\widehat{r(x_{1:N})}^2}}$, is equal to $\frac{1}{\mathrm{Var}_{\pi(x)}\left[r(x)\right]+1}$. It is clear to see that the ratio represent the difference between samples' weight from $\frac{1}{\mathrm{Var}_{\pi(x)}\left[r(x)\right]+1}$. The more the variance of weight, the smaller the ratio is. Although The formula of $N_e$ varies, the following formula may be the **most suitable for calculating** because it reuses the *normalized importance weight*, which was used before when *normalized importance sampling* step. Therefore, the loading to diagnose the samples are representable or not would be less.

$$\begin{aligned} N_e &= \frac{(\sum_{i=1}^N \widehat{r(x_i)})^2}{\sum_{j=1}^N \widehat{r(x_j)}^2} = \frac{1}{\frac{\sum_{j=1}^N \widehat{r(x_j)}^2}{(\sum_{i=1}^N \widehat{r(x_i)})^2}} = \frac{1}{\sum_{j=1}^N \frac{\widehat{r(x_j)}^2}{(\sum_{i=1}^N \widehat{r(x_i)})^2}} \\ &= \frac{1}{\sum_{j=1}^N (\frac{\widehat{r(x_j)}}{\sum_{i=1}^N \widehat{r(x_i)}})^2} = \frac{1}{\sum_{j=1}^N w(x_j)^2} \text{ (By Eq. 3.116)} \end{aligned} \tag{3.126}$$

## 3.6 Sequential Importance Sampling

So far, we have already learned how to approximate one *target distribution* by *importance sampling*. Now, applying what we learn in a sequence of probability distribution $\{\Pr(x_{1:t})\}$ , where $t \geq 1$, defined on a sequence of measurable spaces. (Notice that the meaning of subscript of $x_{1:t}$ denoted here is different from previous. The $x_{1:t}$ here denotes a sequence of **random variables** $x_1, x_2, \ldots, x_t$. The subscript is the index of random variables. On the contrary, the $x_i$ in previous pages denotes **samples** generated from one random variable $x$). We want to estimate the expected value of function $f$ under $\{\Pr(x_{1:t})\}$,

$$\mathrm{E}_{\Pr(x_{1:t})}\left[f(x_{1:t})\right] = \int \Pr(x_{1:t}) \cdot f(x_{1:t}) \cdot dx_{1:t} \tag{3.127}$$

,but we only know a unnormalized version $\widehat{\Pr(x_{1:t})}$ of $\Pr(x_{1:t})$.

$$\Pr(x_{1:t}) = \frac{1}{Z_{1:t}} \cdot \widehat{\Pr(x_{1:t})} \tag{3.128}$$

, where $\frac{1}{Z_{1:t}}$ is normalizing constant of $\Pr(x_{1:t})$.

To estimate what we want, we can start from defining an *importance weight*

$$r(x_{1:t}) = \frac{\Pr(x_{1:t})}{\pi(x_{1:t})} \tag{3.129}$$

, and the *associated unnormalized weight* is

$$\widehat{r(x_{1:t})} = \frac{\widehat{\Pr(x_{1:t})}}{\pi(x_{1:t})} = Z_{1:t} \cdot r(x_{1:t}) \tag{3.130}$$

By Eq. 3.152 and Eq. 3.115 , $\mathrm{E}_{\Pr(x_{1:t})}\left[f(x_{1:t})\right]$ can be approximated:

$$\mathrm{E}_{\Pr(x_{1:t})}\left[f(x_{1:t})\right] \overset{N\to\infty}{\approx} \frac{\sum_{i=1}^{N} f(x_{1:t}^i) \cdot \widehat{r(x_{1:t}^i)}}{\sum_{j=1}^{N} \widehat{r(x_{1:t}^j)}} = \frac{\sum_{i=1}^{N} f(x_{1:t}^i) \cdot r(x_{1:t}^i)}{\sum_{j=1}^{N} r(x_{1:t}^j)} \tag{3.131}$$

### 3.6.1 Computational Complexity of Importance Sampling

The computational complexity of sampling scheme increases at least with the $t$. Suppose that we want to keep tracking the expected value of function $f$ under $\{\Pr(x_{1:t})\}$, in other words, we want to get $\mathrm{E}_{\Pr(x_1)}\left[f(x_1)\right]$ , $\mathrm{E}_{\Pr(x_{1:2})}\left[f(x_{1:2})\right]$ , $\mathrm{E}_{\Pr(x_{1:3})}\left[f(x_{1:3})\right]$ , ... and $\mathrm{E}_{\Pr(x_{1:t})}\left[f(x_{1:t})\right]$ sequentially upon data at time $t$ is available.

Every time $t$ we want to approximate $\mathrm{E}_{\Pr(x_{1:t})}\left[f(x_{1:t})\right]$ , we must sample $x_{1:t}^i$ from $\pi(x_{1:t})$ , where $i = 1, 2, \ldots, N$, where $N$ is the number of samples. Next, we need to compute the associated unnormalized weight $\widehat{r(x_{1:t})}$ to estimate what we want. The "sampling" here increases with $t$ actually , if $t$ grows large, then we would take a lot of time waiting for generating samples. And this problem may cause the incapability of tracking $\mathrm{E}_{\Pr(x_{1:t})}\left[f(x_{1:t})\right]$ in real time. To address this problem, one approach is that we can choose a *proposal distribution* with recursive structure.

### 3.6.2 Choose One Recursive Proposal Distribution

Selecting a *proposal distribution* which has a recursive structure such as:

$$\begin{aligned}\pi(x_{1:t}) &= \pi(x_{1:t-1}) \cdot \pi(x_t \mid x_{1:t-1}) \\ &= \pi(x_1) \cdot \prod_{k=2}^{t} \pi(x_k \mid x_{1:k-1}) \text{ (By Chain-Rule Eq. 3.22)}\end{aligned} \tag{3.132}$$

admits a **fixed** computational complexity for generating samples at each time $t$. Then the problem of the computational complexity of sampling scheme above will be solved.

Due to the recursive character, we can reuse the samples $\{x^i_{1:k}\}$ generated from $\pi(x_{1:k})$, where $i = 1,2,\ldots,N$ and $N$ is the number of samples, to easily build the samples $\{x^i_{1:k+1}\}$, which should be generated from $\pi(x_{1:k+1})$ directly. Furthermore, the *proposal distribution* with recursive structure leads the *associated unnormalized weight* $\widehat{r(x_{1:t})}$ has recursive structure also.

$$\begin{aligned}\widehat{r(x_{1:t})} &= \frac{\widehat{\Pr(x_{1:t})}}{\pi(x_{1:t})} \\ &= \frac{\widehat{\Pr(x_{1:t})}}{\pi(x_{1:t-1}) \cdot \pi(x_t \mid x_{1:t-1})} \\ &= \frac{\widehat{\Pr(x_{1:t-1})}}{\pi(x_{1:t-1})} \cdot \frac{\widehat{\Pr(x_{1:t})}}{\widehat{\Pr(x_{1:t-1})} \cdot \pi(x_t \mid x_{1:t-1})} \\ &= \widehat{r(x_{1:t-1})} \cdot \frac{\widehat{\Pr(x_{1:t})}}{\widehat{\Pr(x_{1:t-1})} \cdot \pi(x_t \mid x_{1:t-1})} \\ &= \widehat{r(x_{1:t-1})} \cdot \alpha(x_{1:t})\end{aligned} \tag{3.133}$$

, where $\alpha(x_{1:t}) = \dfrac{\widehat{\Pr(x_{1:t})}}{\widehat{\Pr(x_{1:t-1})} \cdot \pi(x_t \mid x_{1:t-1})}$ is called *incremental importance weight*.

Organizing the equation above, then we can get:

$$\begin{aligned}\widehat{r(x_{1:t})} &= \widehat{r(x_{1:t-1})} \cdot \alpha(x_{1:t}) \\ &= \widehat{r(x_{1:t-2})} \cdot \alpha(x_{1:t-1}) \cdot \alpha(x_{1:t}) \\ &= \dots \\ &= \widehat{r(x_1)} \cdot \prod_{k=2}^{t} \alpha(x_{1:k})\end{aligned} \tag{3.134}$$

After computing the *associated unnormalized weight* $\widehat{r(x_{1:t})}$ , we should **normalize** $\widehat{r(x_{1:t})}$ to get *normalized importance weights*

$$w(x^i_{1:t}) = \frac{\widehat{r(x^i_{1:t})}}{\sum_{j=1}^{N} \widehat{r(x^j_{1:t})}} \tag{3.135}$$

for all $i = 1, 2, \dots, N$.

As a result, the expected value $\mathrm{E}_{\Pr(x_{1:t})}\left[f(x_{1:t})\right]$ can be approximated by Eq. 3.117

$$\mathrm{E}_{\Pr(x_{1:t})}\left[f(x_{1:t})\right] \overset{N\to\infty}{\approx} \sum_{i=1}^{N} w(x^i_{1:t}) \cdot f(x^i_{1:t}) \tag{3.136}$$

### 3.6.3 Put It All Together

Suppose that we choose one *proposal distribution* with recursive structure , then the procedure to approximate a sequence of probability distribution $\{\Pr(x_{1:t})\}$ is demonstrated as following :

Assume number of samples is $N$,

1. At time 1, generating samples $x^1_1, x^2_1, \dots, x^N_1$ from $\pi(x_1)$ and computing the *associated unnormalized weight* $\widehat{r(x^i_1)} = \frac{\widehat{\Pr(x^i_1)}}{\pi(x^i_1)}$ for all $i = 1, 2, \dots, N$. After computing $\widehat{r(x^i_1)}$ , normalizing it to get *normalized importance weight* $w(x^i_1) = \frac{\widehat{r(x^i_1)}}{\sum_{j=1}^{N} \widehat{r(x^j_1)}}$ for all i. As a result, the expected value can be approximated by $\mathrm{E}_{\Pr(x_1)}\left[f(x_1)\right] \overset{N\to\infty}{\approx} \sum_{i=1}^{N} w(x^i_1) \cdot f(x^i_1)$.

2. At time 2, instead of generating samples $x^1_{1:2}, x^2_{1:2}, \dots, x^N_{1:2}$ from $\pi(x_{1:2})$ directly , we obtain $x^1_{1:2}, x^2_{1:2}, \dots, x^N_{1:2}$ by sampling $(x^1_2|x^1_1), (x^2_2|x^2_1), \dots, (x^N_2|x^N_1)$ from $\pi(x_2 \mid x^i_1)$ for $i = 1, 2, \dots, N$, because $\pi(x_{1:2}) = \pi(x_1) \cdot \pi(x_2 \mid x_1)$. After sampling, computing the *associated unnormalized*

*weight* $\widehat{r(x^i_{1:2})} = \widehat{r(x^i_1)} \cdot \alpha(x^i_{1:2})$ for all $i = 1,2,\ldots,N$. Next, normalizing $\widehat{r(x^i_{1:2})}$ to get *normalized importance weight* $w(x^i_{1:2}) = \dfrac{\widehat{r(x^i_{1:2})}}{\sum_{j=1}^{N} \widehat{r(x^j_{1:2})}}$ for all i. As a result, the expected value can be approximated by $\mathrm{E}_{\Pr(x_{1:2})}\left[f(x_{1:2})\right] \overset{N\to\infty}{\approx} \sum_{i=1}^{N} w(x^i_{1:2}) \cdot f(x^i_{1:2})$

3. At time 3, we can obtain $x^1_{1:3}, x^2_{1:3}, \ldots, x^N_{1:3}$ by sampling $(x^1_3|x^1_{1:2}), (x^2_3|x^2_{1:2}), \ldots, (x^N_3|x^N_{1:2})$ from $\pi(x_3 \mid x^i_{1:2})$ for $i = 1,2,\ldots,N$, because $\pi(x_{1:3}) = \pi(x_{1:2}) \cdot \pi(x_3 \mid x_{1:2})$. After sampling, computing the *associated unnormalized weight* $\widehat{r(x^i_{1:3})} = \widehat{r(x^i_{1:2})} \cdot \alpha(x^i_{1:3})$ for all $i = 1,2,\ldots,N$. Next, normalizing $\widehat{r(x^i_{1:3})}$ to get *normalized importance weight* $w(x^i_{1:3}) = \dfrac{\widehat{r(x^i_{1:3})}}{\sum_{j=1}^{N} \widehat{r(x^j_{1:3})}}$ for all i. As a result, the expected value can be approximated by $\mathrm{E}_{\Pr(x_{1:3})}\left[f(x_{1:3})\right] \overset{N\to\infty}{\approx} \sum_{i=1}^{N} w(x^i_{1:3}) \cdot f(x^i_{1:3})$

4. In general, at time $k$,

$$\begin{aligned} \pi(x_{1:k}) &= \pi(x_{1:k-1}) \cdot \pi(x_k \mid x_{1:k-1}) \\ &= \pi(x_1) \cdot \pi(x_2 \mid x_1) \cdot \pi(x_3 \mid x_{1:2}) \cdot \ldots \cdot \pi(x_k \mid x_{1:k-1}) \end{aligned} \tag{3.137}$$

So, if we already get samples $x^1_{1:k-1}, x^2_{1:k-1}, \ldots, x^N_{1:k-1}$, then we only need to draw $(x^1_k|x^1_{1:k-1}), (x^2_k|x^2_{1:k-1}), \ldots, (x^N_k|x^N_{1:k-1})$ from $\pi(x_k \mid x^i_{1:k-1})$ for $i = 1,2,\ldots,N$. After sampling, computing the *associated unnormalized weight* $\widehat{r(x^i_{1:k})} = \widehat{r(x^i_{1:k-1})} \cdot \alpha(x^i_{1:k})$ for all $i = 1,2,\ldots,N$. Next, normalizing $\widehat{r(x^i_{1:k})}$ to get *normalized importance weight* $w(x^i_{1:k}) = \dfrac{\widehat{r(x^i_{1:k})}}{\sum_{j=1}^{N} \widehat{r(x^j_{1:k})}}$ for all i. As a result, the expected value can be approximated by $\mathrm{E}_{\Pr(x_{1:k})}\left[f(x_{1:k})\right] \overset{N\to\infty}{\approx} \sum_{i=1}^{N} w(x^i_{1:k}) \cdot f(x^i_{1:k})$

5. Repeat to time $t$

### 3.6.4 Algorithm Description

In conclusion, we can now summarize the steps of *sequential importance sampling* into Alg. 3.

**Algorithm 3:** Sequential Importance Sampling

//: Setting

Set number of samples $N$;

Choose one *proposal distribution* $\pi$ ;

//: time $t = 1$ : Initialization

**for** $i \leftarrow 1$ *to* $N$ **do**

  Generate sample $x_1^i$ from *proposal distribution* $\pi(x_1)$;

  Compute the associated unnormalized weight $\widehat{r(x_1^i)} = \dfrac{\widehat{\Pr(x_1^i)}}{\pi(x_1^i)}$;

**end**

//: Normalize

**for** $i \leftarrow 1$ *to* $N$ **do**

  Get *normalized importance weight* of $x_1^i$ : $w(x_1^i) = \dfrac{\widehat{r(x_1^i)}}{\sum_{j=1}^{N} \widehat{r(x_1^j)}}$;

**end**

//: time $t \geq 2$ : Importance Sampling Step

**for** $i \leftarrow 1$ *to* $N$ **do**

  Sample $(x_t^i | x_{1:t-1}^i)$ from $\pi(x_t \mid x_{1:t-1}^i)$;

  Set $x_{1:t}^i = (x_t^i | x_{1:t-1}^i) \cdot x_{1:t-1}^i$;

  Compute *incremental importance weight* : $\alpha(x_{1:t}^i) = \dfrac{\widehat{\Pr(x_{1:t}^i)}}{\widehat{\Pr(x_{1:t-1}^i)} \cdot \pi(x_t^i \mid x_{1:t-1}^i)}$;

  Update the associated unnormalized weight by $\widehat{r(x_{1:t}^i)} = \widehat{r(x_{1:t-1}^i)} \cdot \alpha(x_{1:t}^i)$;

**end**

//: Normalize

**for** $i \leftarrow 1$ *to* $N$ **do**

  Get *normalized importance weight* of $x_{1:t}^i$ : $w(x_{1:t}^i) = \dfrac{\widehat{r(x_{1:t}^i)}}{\sum_{j=1}^{N} \widehat{r(x_{1:t}^j)}}$;

**end**

## 3.7 Sequential Importance Resampling

One crucial drawback of standard *importance sampling* applied in a sequence of variables $\{x^i_{1:t}\}$ for $i \in [1, N]$ is that the variance of samples' weight increases exponentially with $t$(see [14]).

It appears that we have provided an approach, *sequential importance sampling*, to reduce the computational complexity when applying *importance sampling* in a sequence of variables. Nevertheless, it is important to be aware that the *sequential importance sampling* is nothing but one special case of *importance sampling* in which we restrict ourselves to an *proposal distribution* of form Eq. 3.132, so it still suffers from same drawbacks.

### 3.7.1 Degeneracy Problem

Degeneracy phenomenon is a common problem with *sequential importance sampling* , where a few iterations, most of samples will have negligible weights. That is, the approximation is dominated by a few of samples whose weight are larger than others. Degeneracy is an unavoidable phenomenon because the variance of the *importance weight* can only increase over time, and this was shown in [11]. The degeneracy implies that an awful lot of computation efforts are wasted on the samples with negligible contribution to the final approximation.

### 3.7.2 Selection of Proposal Distribution

The natural strategy to limit *degeneracy phenomenon* consists of selecting a *proposal distribution* which minimize the variance of *importance weight*. However, the optimal *proposal distribution* is difficult to find and it is not always possible to generate samples from this optimal proposal distribution. Selecting optimal proposal distribution is usually infeasible in practice.

### 3.7.3 Resampling

As has previously illustrated, degeneracy of *sequential importance sampling* is unavoidable. A suitable measure of degeneracy phenomenon is the *effective sample size* $N_e$ describe in 3.5.4. $N_e$ represents the number of samples which can impact the approximation. A small $N_e$ indicates that the approximation is dominated by a few of samples, thus it suffers from severe degeneracy. From Eq. 3.126,

$N_e$ can be calculated by:

$$N_e = \frac{1}{\sum_{j=1}^{N} w(x_j)^2} \tag{3.138}$$

To avoid wasting large computation effort on the samples with negligible contribution to the final approximation , one basic idea of approach, namely *resampling*, is to eliminate those samples which have small *normalized importance weight* and to concentrate on other samples with large weight. When $N_e$ is lower than a fixed threshold $N_{thres}$, the *resampling* procedure is executed.

In summary, the three most popular algorithms found in literature are, in descending order of popularity:

- **Systematic Resampling**
- **Residual Resampling**
- **Multinomial Resampling**

Although the *resampling* step reduces the effects of degeneracy problem, it introduce other practical problems.

1. It limits the opportunity of parallel processing because all samples must be combined.
2. The high weighted samples are statically selected many times and leads to a loss of **diversity** among the samples. This problem known as *sample impoverishment* is serve in the case of small processing noise. In fact, in the case of small processing noise, the samples will collapse to a single point within a few iterations.
3. Since the diversity of the paths of the samples is reduced , any smoothed estimates based on the samples' path degenerate.

**Threshold of Effective Sample Size**

In practice, the threshold of $N_e$, $N_{thres}$, is hard to define. Besides, $N_{thres}$ is case-specific. On one hand, if $N_{thres}$ is too small, it can not suppress the *degeneracy problem* efficiently. On the other hand, if $N_{thres}$ is too large, then it will lead a serve *sample impoverishment problem*. Both two situations cause the inaccuracy of approximation.

**Systematic Resampling**

The *systematic resampling* is the most widely-used algorithm in literature as it is extremely easy to implement and outperforms other schemes. The procedure of *systematic resampling* can be illustrated by Fig. 3.18.

The steps of *systematic resampling*:

1. Construct the CDF(*cumulative distribution function*) of weights of particles
2. Sample $u_1$ from *uniform distribution* in range 0 to $\frac{1}{N}$
3. Define $u_j = u_1 + \frac{j-1}{N}$
4. Find $i$ satisfy that $\sum_{k=1}^{i-1} w^k \leq u_j \leq \sum_{k=1}^{i} w^k$ , where $w^k$ is the weight of the particle $k$.
5. Assign new particle set $\{x_j, w_j\} = \{x_i, w_i\}$
6. Repeat previous three steps for $j = 1, 2, 3, \ldots, N$.

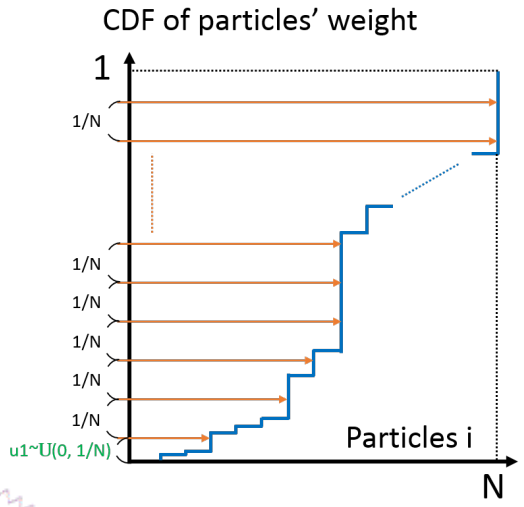



Figure 3.18: Systematic resampling

In Fig. 3.18, it is clear to see that the particles with low weights have lower chances to be duplicated as new particles. After all, the steps of *systematic resampling* can be organized into Alg. 4.

### 3.7.4 Generic Sequential Importance Resampling

Generic version of *sequential importance resampling* is described by Alg. 5.

## 3.8 State Estimation Problem

State estimation is nowadays a widely used class in innumerable practical applications. *state estimation problems*, also designated as *nonstationary inverse problems*, aim to sequentially estimate the unobserved time-varying states of a dynamic system based on prior knowledge about its physical

**Algorithm 4:** Systematic Resampling

**input**: samples $\{x^1, w^1\}, \{x^2, w^2\}, \ldots, \{x^N, w^N\}$
**output**: new samples $\{x^1, w^1\}, \{x^2, w^2\}, \ldots, \{x^N, w^N\}$

Initialize the CDF(cumulative distribution function) : $c_1 = w^1$;
**for** $i \leftarrow 2$ *to* $N$ **do**
    Construct CDF : $c_i = c_{i-1} + w^i$;
**end**
Start at the bottom of the CDF : $i = 1$;
Draw a starting point from uniform distribution : $u_1 \sim \mathbb{U}[0, \frac{1}{N}]$;
**for** $j \leftarrow 1$ *to* $N$ **do**
    Move along the CDF : $u_j = u_1 + \frac{j-1}{N}$;
    **while** $u_j > c_i$ **do**
        $i = i + 1$;
    **end**
    Assign sample : $x^j = x^i$;
    Assign weight : $w^j = w^i$;
**end**
**return** $\{x^1, w^1\}, \{x^2, w^2\}, \ldots, \{x^N, w^N\}$

phenomena and observed data from measuring instrument during evolution of the system. The key issue is modelling the state variables together with observation variables, known as *state model* and *observation model*, and they both are time-dependence. The classical applications include navigation, orbit determination, location tracking, and decision making.

The dynamic system with unobserved states and visible observations can be describe as a sequential *HMM* such as Fig. 3.14. There are two models used for estimating the states of a dynamic system from some observations. One is *state model* $f$ ,and another is *observation model*(or *measurement model*) $h$. The *state model* depicts the transition between states. On the other hand, the *observation model* describes the relation between observations and states. Let $S_t$, $v_t$, $Z_t$, $n_t$ denote the state, state noise , observation and observation noise separately , all in time $t$.

- State Model:

$$S_t = f_t\left(S_{t-1}, v_{t-1}\right) \tag{3.139}$$

- Observation Model:

$$Z_t = h_t\left(S_t, n_t\right) \tag{3.140}$$

**Algorithm 5:** Generic Sequential Importance Resampling

//: Setting

Set number of samples $N$;

Set fixed threshold of *effective sample size* : $N_{thres}$;

Choose one *proposal distribution* $\pi$ ;

//: time $t = 1$ : Initialization

**for** $i \leftarrow 1$ *to* $N$ **do**

  Generate sample $x_1^i$ from *proposal distribution* $\pi(x_1)$;

  Compute the associated unnormalized weight $\widehat{r(x_1^i)} = \dfrac{\widehat{\Pr(x_1^i)}}{\pi(x_1^i)}$;

**end**

//: Normalize

**for** $i \leftarrow 1$ *to* $N$ **do**

  Get *normalized importance weight* of $x_1^i$ : $w(x_1^i) = \dfrac{\widehat{r(x_1^i)}}{\sum_{j=1}^{N} \widehat{r(x_1^j)}}$;

**end**

//: time $t \geq 2$ : Importance Sampling Step

**for** $i \leftarrow 1$ *to* $N$ **do**

  Sample $(x_t^i | x_{1:t-1}^i)$ from $\pi(x_t \mid x_{1:t-1}^i)$;

  Set $x_{1:t}^i = (x_t^i | x_{1:t-1}^i) \cdot x_{1:t-1}^i$;

  Compute *incremental importance weight* : $\alpha(x_{1:t}^i) = \dfrac{\widehat{\Pr(x_{1:t}^i)}}{\widehat{\Pr(x_{1:t-1}^i)} \cdot \pi(x_t^i \mid x_{1:t-1}^i)}$;

  Update the associated unnormalized weight by $\widehat{r(x_{1:t}^i)} = \widehat{r(x_{1:t-1}^i)} \cdot \alpha(x_{1:t}^i)$;

**end**

//: Normalize

**for** $i \leftarrow 1$ *to* $N$ **do**

  Get *normalized importance weight* of $x_{1:t}^i$ : $w(x_{1:t}^i) = \dfrac{\widehat{r(x_{1:t}^i)}}{\sum_{j=1}^{N} \widehat{r(x_{1:t}^j)}}$;

**end**

//: Resampling Step

Calculate the *effective sample size* : $N_e = \dfrac{1}{\sum_{j=1}^{N} w(x_j)^2}$

**if** $N_e < N_{thres}$ **then**

  Resampling using algorithm 4 : $\{x_{1:t}^1\}, \{x_{1:t}^2\}, \ldots, \{x_{1:t}^N\}$ = Resampling($\{x_{1:t}^1\}, \{x_{1:t}^2\}, \ldots, \{x_{1:t}^N\}$) ;

**end**

The following are assumptions of *state estimation problems*:

1. State sequence $S_k$ for $k = 1, 2, \ldots$ is a *Markov* process Eq. 3.63, that is

$$\Pr\left(S_{t+1} \mid S_0, S_1, S_2, \ldots, S_t\right) = \Pr\left(S_{t+1} \mid S_t\right) \tag{3.141}$$

2. State $S_k$ is independent of all the past observations given previous state $S_{k-1}$, that is

$$\Pr\left(S_k \mid S_{k-1}, Z_0, Z_1, Z_2, \ldots, Z_{k-1}\right) = \Pr\left(S_k \mid S_{k-1}\right) \tag{3.142}$$

3. Observation $Z_k$ is independent to all others given the state $S_k$, that is

$$\Pr\left(Z_k \mid S_0, S_1, S_2, \ldots, S_k\right) = \Pr\left(Z_k \mid S_k\right) \tag{3.143}$$

4. state noise $v_i$ and observation noise $n_j$ are mutually independent for all $i$ and $j$.

5. state noise $v_i$ and $v_j$ are mutually independent , and also mutually independent of the initial state $S_0$ for $i \neq j$.

6. likewise, observation noise $n_i$ and $n_j$ are mutually independent , and also mutually independent of the initial state $S_0$ for $i \neq j$.

The first two assumptions follow the rule of *HMM* Eq. 3.67 and Eq. 3.66 actually, so this dynamic system can be viewed as a *HMM* whose state transitions and outputs rely on the *state model* and *observation model* separately.

On the other hands, *state estimation problems* can be organized into four types[17]:

1. Prediction problem, concerns with the determination of the probability of the next state given all the observations till now , $\Pr\left(S_{t+1} \mid Z_{1:t}\right)$.

2. Filtering problem, concerns with the determination of the probability of the state now given all the observations till now , $\Pr\left(S_t \mid Z_{1:t}\right)$.

3. Fixed-lag smoothing problem, concerns with the determination of the probability of state at time $t$ given all observations beyond time $t$ , $\Pr\left(S_t \mid Z_{1:t+l}\right)$, where $l \geq 1$ is fixed-lag.

4. Whole-domain smoothing problem, concerns with the determination of the probability of state at time $t$ given all observations from initials to finish, $\Pr(S_t \mid Z_{1:n})$ , where $Z_{1:n} = \{Z_1, Z_2, \ldots, Z_n\}$ is the complete sequence of observations from initials to finish.

Generally speaking, filtering problem can be thought of as a **tracking** problem, and smoothing problem can be thought of as a **trajectory estimation** problem. In fact, the prediction and filter problems are what most need to solve in many applications.

### 3.8.1 Bayesian Filter

*State estimation problems* are solved with so-called *Bayesian filter*( or *Bayes filter*), an attempt to utilize all available information. As new information is obtained, it is combined with the previous information to form the basis for statistical procedures by mechanism of *Bayesian theorem*. The goal of *Bayesian filter* is to estimate the posterior probability of state based on all available observations. In formal saying, we want to estimate $\Pr(S_t \mid Z_{1:t})$ given *system model*:

$$\text{System model} \begin{cases} \text{State model} : \Pr(S_t \mid S_{t-1}) \\ \text{Observation model} : \Pr(Z_t \mid S_t) \end{cases} \tag{3.144}$$

*Bayes filter* is one of recursive filter which formed into some related equations with recursive property. The key concept of recursive filter relies on sequential processing. Sequential processing is fired when new data upon arrival. It need not store the complete dataset and need not reprocess all data for each new observations. Conversely, batch processing computes with **all** available data in one step.

Given system model Eq. 3.144, initial state $\Pr(S_0 \mid Z_0) = \Pr(S_0)$ and observations $\{Z_1, Z_2, \ldots, Z_t\}$ with probability space $S$, the computation of *Bayes filter* consist of following two steps:

1. Prediction step: To get *current prior*, $\Pr(S_{t-1} \mid Z_{1:t-1}) \rightarrow \Pr(S_t \mid Z_{1:t-1})$

   - Uses the state model to predict forward
   - Translates/Spreads PDF(*probability density function*) due to noise.

   Assume that $\Pr(S_{t-1} \mid Z_{1:t-1})$ is available at time $t-1$, by *Chapman-Kolmogorov equation*,

$\Pr(S_t \mid Z_{1:t-1})$ can be computed as:

$$\overbrace{\Pr(S_t \mid Z_{1:t-1})}^{\text{Current prior}} = \int_{S_{t-1} \in S} \overbrace{\Pr(S_t \mid S_{t-1})}^{\text{State model}} \cdot \overbrace{\Pr(S_{t-1} \mid Z_{1:t-1})}^{\text{Previous posterior}} \cdot dS_{t-1} \tag{3.145}$$

2. Update step: To get *current posterior*, $\{\Pr(S_t \mid Z_{1:t-1}), Z_t\} \rightarrow \Pr(S_t \mid Z_{1:t})$

   - Update the prediction in light of new data by observation model
   - Tightens the state *PDF*

$$\begin{aligned}
\overbrace{\Pr(S_t \mid Z_{1:t})}^{\text{Current posterior}} &= \frac{\overbrace{\Pr(Z_t \mid S_t)}^{\text{Observation model}} \cdot \overbrace{\Pr(S_t \mid Z_{1:t-1})}^{\text{Current prior}}}{\underbrace{\Pr(Z_t \mid Z_{1:t-1})}_{\text{normalizing constant}}} \\
&= \frac{\overbrace{\Pr(Z_t \mid S_t)}^{\text{Observation model}} \cdot \overbrace{\Pr(S_t \mid Z_{1:t-1})}^{\text{Current prior}}}{\int_{S_t \in S} \underbrace{\Pr(Z_t \mid S_t)}_{\text{Observation model}} \cdot \underbrace{\Pr(S_t \mid Z_{1:t-1})}_{\text{Current prior}} \cdot dS_t} \text{ (By } \textit{Chapman-Kolmogorov}\text{)}
\end{aligned} \tag{3.146}$$

The equation used in update step can be derived by following:

$$\begin{aligned}
\Pr(S_t \mid Z_{1:t}) &= \Pr(S_t \mid Z_t, Z_{1:t-1}) \\
&= \frac{\Pr(Z_t \mid S_t, Z_{1:t-1}) \cdot \Pr(S_t \mid Z_{1:t-1})}{\Pr(Z_t \mid Z_{1:t-1})} \text{ (By Eq. 3.12} \\
&= \frac{\Pr(Z_t \mid S_t) \cdot \Pr(S_t \mid Z_{1:t-1})}{\Pr(Z_t \mid Z_{1:t-1})} \text{ (By definition)}
\end{aligned} \tag{3.147}$$

For interpretation, the prediction and update steps can be described as:

1. Prediction step:

$$\text{Prior} = \int \text{state model} \cdot \text{posterior} \tag{3.148}$$

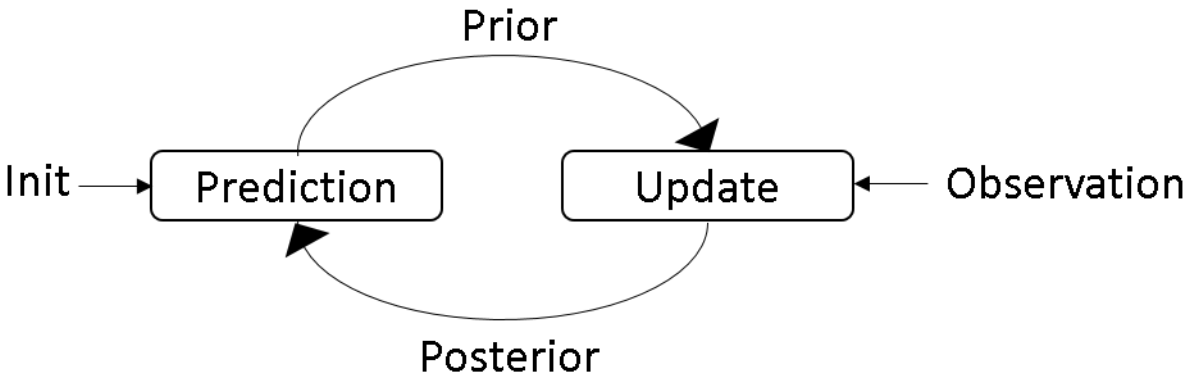



Figure 3.19: The recursive Bayesian filter

2. Update step:

$$\begin{aligned} \text{Posterior} &= \frac{\text{observation model} \cdot \text{prior}}{\text{normalizing constant}} \\ &= \frac{\text{observation model} \cdot \text{prior}}{\int \text{observation model} \cdot \text{prior}} \end{aligned} \tag{3.149}$$

The recursive relation between prediction and update step is like ping-pong, which can be described as Fig. 3.19. More deeply, to see the detail of sequential processing of *Bayes filter*, the flow-process is demonstrated in Fig. 3.20. From the viewpoint of engineering, the sequential processing of *Bayes filter* can be portrayed by two block diagrams in Fig. 3.21.

In fact, the equations used in two steps can be expressed into a close-form equation:

$$\begin{aligned} \Pr(S_t \mid Z_{1:t}) &= \frac{\Pr(Z_t \mid S_t) \cdot \Pr(S_t \mid Z_{1:t-1})}{\Pr(Z_t \mid Z_{1:t-1})} \\ &= \frac{\Pr(Z_t \mid S_t) \cdot \int \Pr(S_t \mid S_{t-1}) \Pr(S_{t-1} \mid Z_{1:t-1}) \cdot dS_{t-1}}{\Pr(Z_t \mid Z_{1:t-1})} \\ &= \frac{\Pr(Z_t \mid S_t) \cdot \int \Pr(S_t \mid S_{t-1}) \Pr(S_{t-1} \mid Z_{1:t-1}) \cdot dS_{t-1}}{\int \Pr(Z_t \mid S_t) \cdot \Pr(S_t \mid Z_{1:t-1}) \cdot dS_t} \end{aligned} \tag{3.150}$$

It's pretty clear to see the recursive property from the equation above: you can calculate the $\Pr(S_t \mid Z_{1:t})$ from $\Pr(S_{t-1} \mid Z_{1:t-1})$. Consequently, $\Pr(S_{t+1} \mid Z_{1:t+1})$ can be calculated by $\Pr(S_t \mid Z_{1:t})$ and so on.

Prediction and update steps give optimal solution for recursive estimation problem. Unfortunately, it's only conceptual solution, because integrals are difficult to be calculated in most of the time. The *Bayes filter* is infeasible in practice.

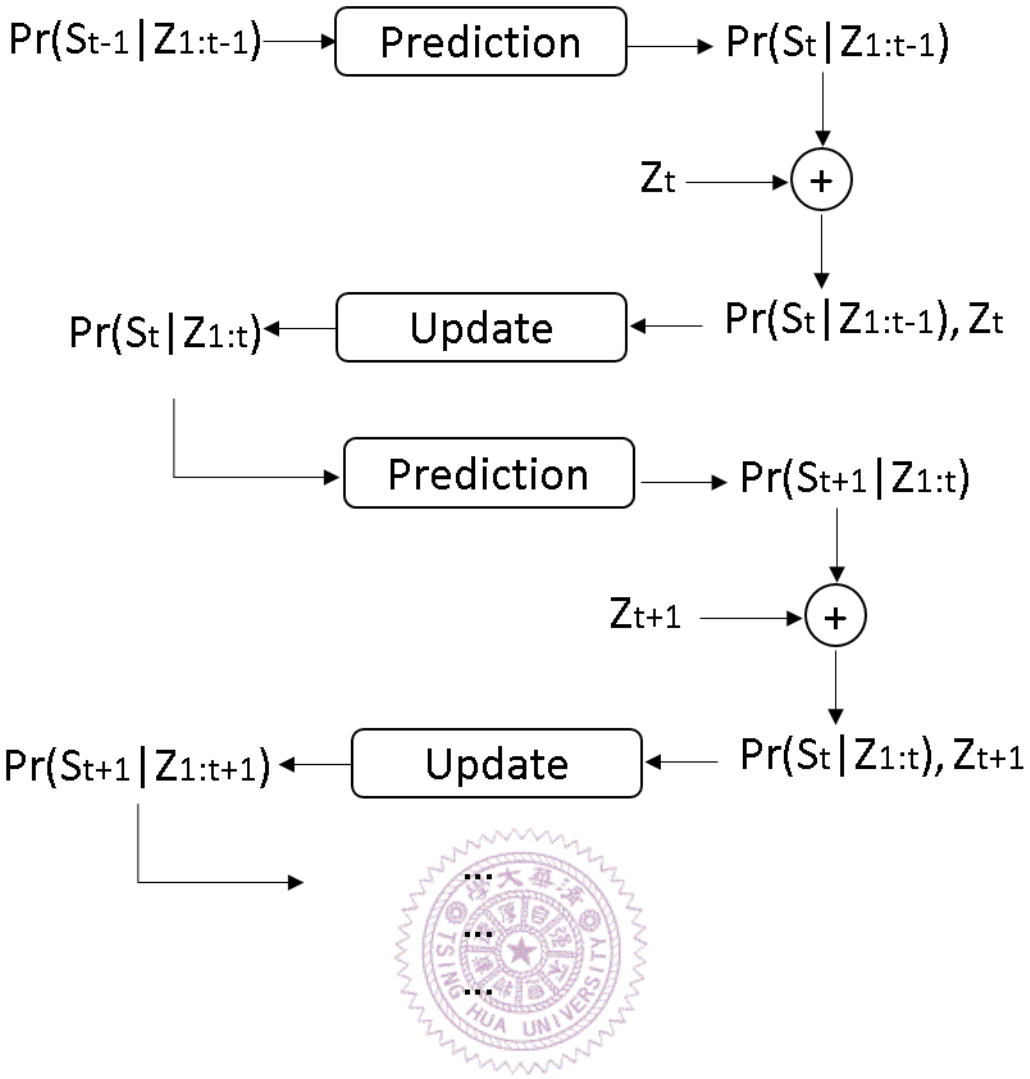



Figure 3.20: Sequential processing in Bayesian filter

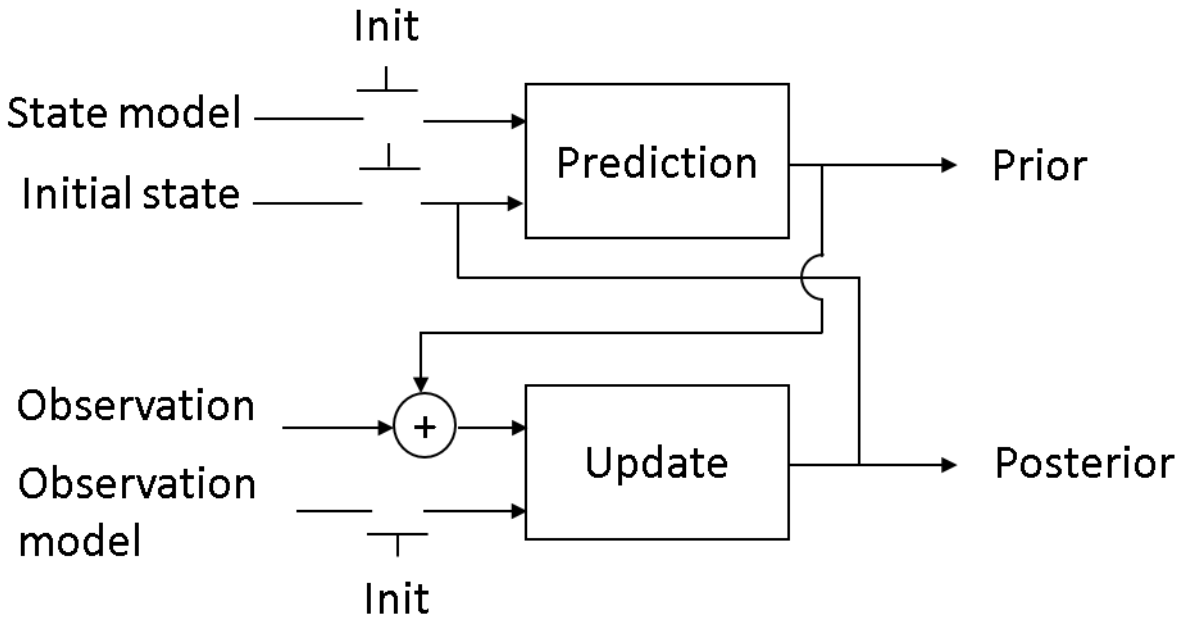



Figure 3.21: The Bayesian filter module

### 3.8.2 Particle Filter

To address the problem of *Bayes filter*, the integrals are difficult to be calculated, one possible approach is *importance sampling*. The *Bayes filter* relies on two recursive steps , prediction and update, to process data sequentially for tracking the newest state. *Importance sampling* is also able to track one sequential data, namely *sequential importance sampling*. However, the variance of samples used in *sequential importance sampling* increase exponentially with time $t$. In other word, after a long run, the *effective sample size* will be very low and cause the inaccuracy of estimation. *Resampling* scheme is one good choice to restrain the variance growing. Combining *resampling* with *sequential importance sampling* form a method called *sequential importance resampling*, which is effective approach for tracking one sequential data. Applying *sequential importance resampling* to estimate the states of dynamic system becomes popular in past twenty years, and this specific application has a proper noun, *particle filter*.

In the following contents, we introduced *particle filter* in brief. The detail of *particle filter* can be founded in [10, 12, 7].

#### Quick Review of Sequential Importance Resampling

*Monte Carlo estimate* is frequently used when the integrals of one *target distribution* $\Pr(x)$ is hard to be calculated. It proposes a method to approximate an expected value under $\Pr(x)$ by taking average of a large number of samples drawn from $\Pr(x)$. However, if it is not easy to generate samples from $\Pr(x)$, then we can choose one *proposal distribution* that we will draw samples from instead without much effort. In practice, we may only know an unnormalized version $\widehat{\Pr(x)}$ of $\Pr(x)$, $\Pr(x) = \dfrac{\widehat{\Pr(x)}}{Z_p}$, where $Z_p = \int \widehat{\Pr(x)} dx$, and $\dfrac{1}{Z_p}$ is namely *normalizing constant*.

Suppose that we want to calculate the expected value of function $f$ under $\Pr(x)$ by sampling from a *proposal distribution* $\pi(x)$

$$\begin{aligned}
\mathrm{E}_{\Pr(x)}[f(x)] &= \int f(x)\cdot \Pr(x)\cdot dx = \int f(x)\cdot \frac{\widehat{\Pr(x)}}{Z_p}\cdot dx = \frac{\int f(x)\cdot \widehat{\Pr(x)}\cdot dx}{\int \widehat{\Pr(x)}\cdot dx} \\
&= \frac{\int f(x)\cdot \dfrac{\widehat{\Pr(x)}}{\pi(x)}\cdot \pi(x)\cdot dx}{\int \dfrac{\widehat{\Pr(x)}}{\pi(x)}\cdot \pi(x)\cdot dx} = \frac{\int f(x)\cdot \widehat{r(x)}\cdot \pi(x)\cdot dx}{\int \widehat{r(x)}\cdot \pi(x)\cdot dx} = \frac{\mathrm{E}_{\pi(x)}[f(x)\cdot \widehat{r(x)}]}{\mathrm{E}_{\pi(x)}[\widehat{r(x)}]}
\end{aligned} \tag{3.151}$$

, where $\widehat{r(x)} = \dfrac{\widehat{\Pr(x)}}{\pi(x)}$.

Appling *Monte Carlo Estimate* Eq. 3.95 in numerator and denominator of Eq. 3.112 separately , then

$$\mathrm{E}_{\Pr(x)}\left[f(x)\right] = \frac{\mathrm{E}_{\pi(x)}\left[f(x)\cdot\widehat{r(x)}\right]}{\mathrm{E}_{\pi(x)}\left[\widehat{r(x)}\right]} \overset{N\to\infty}{\approx} \frac{\frac{1}{N}\cdot\sum_{i=1}^{N} f(x_i)\cdot\widehat{r(x_i)}}{\frac{1}{N}\cdot\sum_{j=1}^{N}\widehat{r(x_j)}} = \frac{\sum_{i=1}^{N} f(x_i)\cdot\widehat{r(x_i)}}{\sum_{j=1}^{N}\widehat{r(x_j)}} = \sum_{i=1}^{N} f(x_i)\cdot w(x_i) \tag{3.152}$$

, where $w(x_i) = \dfrac{\widehat{r(x_i)}}{\sum_{j=1}^{N}\widehat{r(x_j)}}$ called *normalized importance weight*

In many application, what we need to concern is a sequence of random variables $\{x_1, x_2, \ldots, x_t\}$ which are generated at each time $t$ rather than just one random variable $x$. Keeping tracking $\mathrm{E}_{\Pr(x_{1:t})}\left[f(x_{1:t})\right]$ at each time step $t$ become an importance task. Nevertheless, the sampling scheme increases at least with $t$. That is, if $t$ is large, we need to wait a long time to generate the samples $\{x^i_{1:t}\}$ for $i \in [1, N]$, where $N$ is the number of samples, to calculate the $\mathrm{E}_{\Pr(x_{1:t})}\left[f(x_{1:t})\right]$. The approach to overcome this problem namely *sequential importance sampling*. The key idea of *sequential importance sampling* is selecting one *proposal distribution* with **recursive** structure such as

$$\pi(x_{1:t}) = \pi(x_{1:t-1})\cdot\pi(x_t \mid x_{1:t-1}) \tag{3.153}$$

so if we already get samples $x^i_{1:t-1}$ , then we only need to draw $(x^i_t | x^i_{1:k-1})$ from $\pi(x_t \mid x^i_{1:t-1})$, where $i = 1, 2, \ldots, N$. Thus, it admits a **fixed** computational complexity for generating samples at each time $t$. What's more, the recursive *proposal distribution* leads weight $\widehat{r(x_{1:t})}$ has recursive structure also. Other detail is mentioned section 3.6 already.

One drawback of *sequential importance sampling* is that the variance of samples used in *sequential importance sampling* increase exponentially with time $t$, but it can be overcome by combining *resampling* scheme with *sequential importance sampling*, namely *sequential importance resampling*. The detail of *sequential importance resampling* is mentioned section 3.7 already.

### Sequential Importance Resampling in Dynamic System

To address the problem described in the *Bayes filter* section(the integrals is "intractable"), *Monte-Carlo estimate* may be a good approach to approximate the intractable integrals appearing in prediction and update steps of *Bayes filter*.

Focusing on "tracking" problem, we now concern on the expected value of function $f$ under the probability of unobserved states given all observations till now, $\Pr(S_{1:t} \mid Z_{1:t})$. Assume that we are able to generate $N$ *independent and identically(i.i.d)* random variables, namely *particles*, $\{S^i_{1:t}\}$, where $i = 1, 2, \ldots, N$ according to $\Pr(S_{1:t} \mid Z_{1:t})$. Then, by *Monte Carlo estimate* Eq. 3.95

$$\mathrm{E}_{\Pr(S_{1:t}|Z_{1:t})}\left[f(S_{1:t})\right] = \int f(S_{1:t}) \cdot \Pr(S_{1:t} \mid Z_{1:t}) \cdot dS_{1:t} \overset{N\to\infty}{\approx} \frac{1}{N} \cdot \sum_{i=1}^{N} f(S^i_{1:t}) \tag{3.154}$$

Unfortunately, it is seldom possible to obtain samples from $\Pr(S_{1:t} \mid Z_{1:t})$ directly, so we can use *importance sampling* to achieve that. Let us introduce another distribution $\pi(S_{1:t} \mid Z_{1:t})$ called *proposal distribution* that we can draw samples $\{S^1_{1:t}, S^2_{1:t}, \ldots, S^N_{1:t}\}$ from instead without much effort. In practice, we only know an unnormalized version $\widehat{\Pr(S_{1:t} \mid Z_{1:t})}$ of $\Pr(S_{1:t} \mid Z_{1:t})$ most of the time.

$$\Pr(S_{1:t} \mid Z_{1:t}) = \frac{\widehat{\Pr(S_{1:t} \mid Z_{1:t})}}{C_{1:t}} \tag{3.155}$$

, where *normalizing constant* $C_{1:t} = \int \widehat{\Pr(S_{1:t} \mid Z_{1:t})} \cdot dS_{1:t}$ (the integral of $\Pr(S_{1:t} \mid Z_{1:t})$ should be 1. This is easy to see if you replace $S_{1:t}$,$Z_{1:t}$ into $A$, $B$ of Eq. 3.44).

$$\begin{aligned}
\mathrm{E}_{\Pr(S_{1:t}|Z_{1:t})}\left[f(S_{1:t})\right] &= \int f(S_{1:t}) \cdot \Pr(S_{1:t} \mid Z_{1:t}) \cdot dS_{1:t} = \int f(S_{1:t}) \cdot \frac{\widehat{\Pr(S_{1:t} \mid Z_{1:t})}}{C_{1:t}} \cdot dS_{1:t} \\
&= \frac{\int f(S_{1:t}) \cdot \widehat{\Pr(S_{1:t} \mid Z_{1:t})} \cdot dS_{1:t}}{\int \widehat{\Pr(S_{1:t} \mid Z_{1:t})} \cdot dS_{1:t}} = \frac{\int f(S_{1:t}) \cdot \frac{\widehat{\Pr(S_{1:t} \mid Z_{1:t})}}{\pi(S_{1:t} \mid Z_{1:t})} \cdot \pi(S_{1:t} \mid Z_{1:t}) \cdot dS_{1:t}}{\int \frac{\widehat{\Pr(S_{1:t} \mid Z_{1:t})}}{\pi(S_{1:t} \mid Z_{1:t})} \cdot \pi(S_{1:t} \mid Z_{1:t}) \cdot dS_{1:t}} \\
&= \frac{\int f(S_{1:t}) \cdot \widehat{r(S_{1:t})} \cdot \pi(S_{1:t} \mid Z_{1:t}) \cdot dS_{1:t}}{\int \widehat{r(S_{1:t})} \cdot \pi(S_{1:t} \mid Z_{1:t}) \cdot dS_{1:t}} = \frac{\mathrm{E}_{\pi(S_{1:t}|Z_{1:t})}\left[f(S_{1:t}) \cdot \widehat{r(S_{1:t})}\right]}{\mathrm{E}_{\pi(S_{1:t}|Z_{1:t})}\left[\widehat{r(S_{1:t})}\right]}
\end{aligned} \tag{3.156}$$

, where $\widehat{r(S_{1:t})} = \frac{\widehat{\Pr(S_{1:t} \mid Z_{1:t})}}{\pi(S_{1:t} \mid Z_{1:t})}$

That is to say, if we can generate particles $\{S^1_{1:t}, S^2_{1:t}, \ldots, S^N_{1:t}\}$ according to $\pi(S_{1:t} \mid Z_{1:t})$, then we

can apply *Monte Carlo estimate*Eq. 3.95 in numerator and denominator separately.

$$\begin{aligned}\mathrm{E}_{\Pr\left(S_{1:t}\middle|Z_{1:t}\right)}\left[f\left(S_{1:t}\right)\right] &= \frac{\mathrm{E}_{\pi\left(S_{1:t}\middle|Z_{1:t}\right)}\left[f\left(S_{1:t}\right)\cdot\widehat{r\left(S_{1:t}\right)}\right]}{\mathrm{E}_{\pi\left(S_{1:t}\middle|Z_{1:t}\right)}\left[\widehat{r\left(S_{1:t}\right)}\right]} \\ &\overset{N\to\infty}{\approx} \frac{\frac{1}{N}\cdot\sum_{i=1}^{N} f\left(S_{1:t}^{i}\right)\cdot\widehat{r\left(S_{1:t}^{i}\right)}}{\frac{1}{N}\cdot\sum_{j=1}^{N}\widehat{r\left(S_{1:t}^{j}\right)}} = \frac{\sum_{i=1}^{N} f\left(S_{1:t}^{i}\right)\cdot\widehat{r\left(S_{1:t}^{i}\right)}}{\sum_{j=1}^{N}\widehat{r\left(S_{1:t}^{j}\right)}} = \sum_{i=1}^{N} f\left(S_{1:t}^{i}\right)\cdot w\left(S_{1:t}^{i}\right)\end{aligned} \tag{3.157}$$

, where *normalized importance weight* $w\left(S_{1:t}^{i}\right) = \dfrac{\widehat{r\left(S_{1:t}^{i}\right)}}{\sum_{j=1}^{N}\widehat{r\left(S_{1:t}^{j}\right)}}$

The computation above needs to get all observations $\{Z_1, Z_2, \ldots, Z_t\}$ before approximating $\mathrm{E}_{\Pr\left(S_{1:t}\middle|Z_{1:t}\right)}\left[f\left(S_{1:t}\right)\right]$ at time $t$. Thus, once observation $Z_{t+1}$ becomes available at time $t+1$, one need to **recompute** the observation data $\{Z_1, Z_2, \ldots, Z_t\}$. This computation would occupy a lot of memory usage because we must need to save all observation data. In addition, the computational complexity of *importance sampling* for generating particles $\{S_{1:t}^{1}, S_{1:t}^{2}, \ldots, S_{1:t}^{N}\}$ increase at least with time $t$.

One good approach is to choose a tricky *proposal distribution* that can be recursively computed by itself because we already know the target distribution $\Pr\left(S_{1:t} \mid Z_{1:t}\right)$ is recursive by Eq. 3.80.

$$\Pr\left(S_{1:t} \mid Z_{1:t}\right) = \Pr\left(S_{1:t-1} \mid Z_{1:t-1}\right)\cdot\frac{\Pr\left(Z_t \mid S_t\right)\cdot\Pr\left(S_t \mid S_{t-1}\right)}{\Pr\left(Z_t \mid Z_{1:t-1}\right)} \tag{3.158}$$

Suppose we choose one *proposal distribution* $\pi\left(S_{1:t} \mid Z_{1:t}\right)$ which satisfies that

$$\begin{aligned}\pi\left(S_{1:t} \mid Z_{1:t}\right) &= \pi\left(S_{1:t-1} \mid Z_{1:t-1}\right)\cdot\pi\left(S_t \mid S_{1:t-1}, Z_{1:t}\right) \\ &= \pi\left(S_1 \mid Z_1\right)\cdot\prod_{k=2}^{t}\pi\left(S_k \mid S_{1:k-1}, Z_{1:k}\right)\end{aligned} \tag{3.159}$$

(Notice that this just an example, you are not necessary to choose proposal distribution equal to this) Then, we can reuse the particles $\{S_{1:t-1}^{1}, S_{1:t-1}^{2}, \ldots, S_{1:t-1}^{N}\}$ to generate particles $\{S_{1:t}^{1}, S_{1:t}^{2}, \ldots, S_{1:t}^{N}\}$ by sampling particles $\{(S_t^i \mid S_{1:t-1}^{i}, Z_{1:t})\}$ from $\pi\left(S_t \mid S_{1:t-1}^{i}, Z_{1:t}\right)$ and set $\{S_{1:t}^{i}\} = \{S_{1:t-1}^{i}\}\cdot\{(S_t^i | S_{1:t-1}^{i}, Z_{1:t})\}$, where $i \in [1, N]$.

The associated unnormalized weigh $\widehat{r(S^i_{1:t})}$ of this *proposal distribution* is:

$$\begin{aligned}
\widehat{r(S_{1:t})} &= \frac{\widehat{\Pr(S_{1:t} \mid Z_{1:t})}}{\pi(S_{1:t} \mid Z_{1:t})} = C_{1:t} \cdot \frac{\Pr(S_{1:t} \mid Z_{1:t})}{\pi(S_{1:t} \mid Z_{1:t})} \\
&= C_{1:t} \cdot \frac{1}{\Pr(Z_t \mid Z_{1:t-1})} \cdot \frac{\Pr(S_{1:t-1} \mid Z_{1:t-1}) \cdot \Pr(Z_t \mid S_t) \cdot \Pr(S_t \mid S_{t-1})}{\pi(S_{1:t-1} \mid Z_{1:t-1}) \cdot \pi(S_t \mid S_{1:t-1}, Z_{1:t})} \\
&= C_{1:t} \cdot \frac{1}{\Pr(Z_t \mid Z_{1:t-1})} \cdot \frac{\widehat{\Pr(S^i_{1:t-1} \mid Z_{1:t-1})}}{C_{1:t-1}} \cdot \frac{\Pr(Z_t \mid S_t) \cdot \Pr(S_t \mid S_{t-1})}{\pi(S_{1:t-1} \mid Z_{1:t-1}) \cdot \pi(S_t \mid S_{1:t-1}, Z_{1:t})} \\
&= K_t \cdot \frac{\widehat{\Pr(S_{1:t-1} \mid Z_{1:t-1})}}{\pi(S_{1:t-1} \mid Z_{1:t-1})} \cdot \frac{\Pr(Z_t \mid S_t) \cdot \Pr(S_t \mid S_{t-1})}{\pi(S_t \mid S_{1:t-1}, Z_{1:t})} \\
&= K_t \cdot \widehat{r(S_{1:t-1})} \cdot \frac{\Pr(Z_t \mid S_t) \cdot \Pr(S_t \mid S_{t-1})}{\pi(S_t \mid S_{1:t-1}, Z_{1:t})}
\end{aligned} \tag{3.160}$$

, where $K_t = \dfrac{C_{1:t}}{C_{1:t-1} \cdot \Pr(Z_t \mid Z_{1:t-1})}$.

Therefore, you can say that $\widehat{r(S^i_{1:t})} \propto \widehat{r(S^i_{1:t-1})} \cdot \dfrac{\Pr(Z_t \mid S^i_t) \cdot \Pr(S^i_t \mid S^i_{t-1})}{\pi(S^i_t \mid S^i_{1:t-1}, Z_{1:t})}$ has recursive structure also. Next, the *normalized importance weight* can be computed by:

$$w(S^i_{1:t}) = \frac{\widehat{r(S^i_{1:t})}}{\sum_{j=1}^N \widehat{r(S^j_{1:t})}} = \frac{K_t \cdot \widehat{r(S^i_{1:t-1})} \cdot L^i_t}{\sum_{j=1}^N K_t \cdot \widehat{r(S^j_{1:t-1})} \cdot L^j_t} = \frac{\widehat{r(S^i_{1:t-1})} \cdot L^i_t}{\sum_{j=1}^N \widehat{r(S^j_{1:t-1})} \cdot L^j_t} \tag{3.161}$$

, where $L^i_t = \dfrac{\Pr(Z_t \mid S^i_t) \cdot \Pr(S^i_t \mid S^i_{t-1})}{\pi(S^i_t \mid S^i_{1:t-1}, Z_{1:t})}$.

The equation above shows that $K_t$ would be eliminated in computation, so no matter what the value $K_t$ is, it does not influence the estimation. In practice, the value of $K_t$ is usually assigned by 1, and it allows $\widehat{r(S^i_{1:t})}$ can be updated by $\widehat{r(S^i_{1:t})} = \widehat{r(S^i_{1:t-1})} \cdot \dfrac{\Pr(Z_t \mid S^i_t) \cdot \Pr(S^i_t \mid S^i_{t-1})}{\pi(S^i_t \mid S^i_{1:t-1}, Z_{1:t})}$. Therefore, the *incremental importance weight* here is equal to $\dfrac{\Pr(Z_t \mid S^i_t) \cdot \Pr(S^i_t \mid S^i_{t-1})}{\pi(S^i_t \mid S^i_{1:t-1}, Z_{1:t})}$.

It appears that we have provided an approach, *sequential importance sampling*, to reduce the computational complexity when applying *importance sampling* in a sequence of variables. However, one crucial drawback of *importance sampling* is *degeneracy problem*: after a few iterations, most of particles will have negligible weight. This implies the approximation is dominated by a few of samples whose weight are larger than others and an awful lot of computation effort are wasted on the samples with negligible contribution to the final approximation. Degeneracy is an unavoidable phenomenon because the variance of the *importance weight*. The natural strategy to limit *degeneracy phenomenon* consists of selecting a *proposal distribution* which minimize the variance of *importance*

*weight*.

**Theorem 3.8.1.** $\pi\left(S_t \mid S_{1:t-1}^i, Z_{1:t}\right) = \Pr\left(S_t \mid S_{t-1}^i, Z_t\right)$ *in the proposal distribution Eq. 3.159 minimize the variance of weight* $\widehat{r\left(S_{1:t}^i\right)}$ *conditional upon* $S_{1:t-1}^i$ *and* $Z_{1:t}$

This is already proved by [11] actually. However, it is not always possible to generate particles from such optimal proposal distribution, so another approach called *resampling* is proposed. Combining *resampling* scheme with *sequential importance sampling*, one useful method to approximate the expected value or probability is presented to the public, known as *sequential importance resampling* (or *Sequential Monte Carlo*), which is popular over the past twenty years.

### Generic Particle Filter

On the whole, the algorithm description of *SIR* in dynamic system estimation given $\widehat{\Pr\left(S_1 \mid Z_1\right)} = \widehat{\Pr\left(S_1\right)}$ and $\pi\left(S_1 \mid Z_1\right) = \pi\left(S_1\right)$ is described by Alg. 6.

### Bootstrap Filter

*Bootstrap Filter* is an important particular case of particle filter, which adopt the joint probability of states $\Pr\left(S_{1:t}\right)$ as *proposal distribution*

$$\begin{aligned}\pi\left(S_{1:t} \mid Z_{1:t}\right) = \Pr\left(S_{1:t}\right) &= \Pr\left(S_1\right) \cdot \prod_{k=2}^{t} \Pr\left(S_k \mid S_{k-1}\right) \text{ (By Eq. 3.68)} \\ &= \Pr\left(S_t \mid S_{t-1}\right) \cdot \Pr\left(S_{1:t-1}\right) \text{ (By Eq. 3.77)} \end{aligned} \tag{3.162}$$

Thus, if we already get particles $\{S_{1:t-1}^1, S_{1:t-1}^2, \ldots, S_{1:t-1}^N\}$, then we only need to draw $\{(S_t^i \mid S_{t-1}^i)\}$ from $\Pr\left(S_t \mid S_{t-1}\right)$ and set $\{S_{1:t}^i\} = \{S_{1:t-1}^i\} \cdot \{(S_t^i | S_{t-1}^i)\}$, where $i \in [1, N]$. Notice that $\Pr\left(S_t \mid S_{t-1}\right)$ here is actually our *state model* of dynamic system. This means that we just need to update the states to generate particles.

**Algorithm 6:** Generic Particle Filter

//: Setting

Set number of samples $N$;

Set fixed threshold of *effective sample size* : $N_{thres}$;

Choose one *proposal distribution* $\pi$ ;

//: time $t = 1$ : Initialization

**for** $i \leftarrow 1$ *to* $N$ **do**

  Generate particle $S_1^i$ from *proposal distribution* $\pi(S_1)$;

  Compute the associated unnormalized weight $\widehat{r(S_1^i)} = \dfrac{\widehat{\Pr(S_1^i)}}{\pi(S_1^i)}$;

**end**

//: Normalize

**for** $i \leftarrow 1$ *to* $N$ **do**

  Get *normalized importance weight* of $S_1^i$ : $w(S_1^i) = \dfrac{\widehat{r(S_1^i)}}{\sum_{j=1}^{N} \widehat{r(S_1^j)}}$;

**end**

//: time $t \geq 2$ : Importance Sampling Step

**for** $i \leftarrow 1$ *to* $N$ **do**

  Sample $(S_t^i | S_{1:t-1}^i, Z_{1:t})$ from $\pi(S_t \mid S_{1:t-1}^i, Z_{1:t})$;

  Set $S_{1:t}^i = S_{1:t-1}^i \cdot (S_t^i | S_{1:t-1}^i, Z_{1:t})$

  Compute *incremental importance weight* : $\alpha(S_{1:t}^i) = \dfrac{\Pr(Z_t \mid S_t^i) \cdot \Pr(S_t^i \mid S_{t-1}^i)}{\pi(S_t^i \mid S_{1:t-1}^i, Z_{1:t})}$;

  Update the associated unnormalized weight by $\widehat{r(S_{1:t}^i)} = \widehat{r(S_{1:t-1}^i)} \cdot \alpha(S_{1:t}^i)$;

**end**

//: Normalize

**for** $i \leftarrow 1$ *to* $N$ **do**

  Get *normalized importance weight* of $S_{1:t}^i$ : $w(S_{1:t}^i) = \dfrac{\widehat{r(S_{1:t}^i)}}{\sum_{j=1}^{N} \widehat{r(S_{1:t}^j)}}$;

**end**

//: Resampling Step

Calculate the *effective sample size* : $N_e = \dfrac{1}{\sum_{j=1}^{N} w(S_{1:t}^j)^2}$

**if** $N_e < N_{thres}$ **then**

  Resampling using algorithm 4 : $\{S_{1:t}^1\}, \{S_{1:t}^2\}, \ldots, \{S_{1:t}^N\}$ = Resampling($\{S_{1:t}^1\}, \{S_{1:t}^2\}, \ldots, \{S_{1:t}^N\}$) ;

**end**

The associated unnormalized weight $\widehat{r(S^i_{1:t})}$ of this *proposal distribution* is:

$$\begin{aligned}
\widehat{r(S_{1:t})} &= \frac{\Pr(\widehat{S_{1:t} \mid Z_{1:t}})}{\pi(S_{1:t} \mid Z_{1:t})} = \frac{\Pr(\widehat{S_{1:t} \mid Z_{1:t}})}{\Pr(S_{1:t})} = C_{1:t} \cdot \frac{\Pr(S_{1:t} \mid Z_{1:t})}{\Pr(S_{1:t})} \\
&= C_{1:t} \cdot \frac{1}{\Pr(Z_t \mid Z_{1:t-1})} \cdot \frac{\Pr(S_{1:t-1} \mid Z_{1:t-1}) \cdot \Pr(Z_t \mid S_t) \cdot \Pr(S_t \mid S_{t-1})}{\Pr(S_t \mid S_{t-1}) \cdot \Pr(S_{1:t-1})} \\
&= C_{1:t} \cdot \frac{1}{\Pr(Z_t \mid Z_{1:t-1})} \cdot \frac{\Pr(\widehat{S^i_{1:t-1} \mid Z_{1:t-1}})}{C_{1:t-1}} \cdot \frac{\Pr(Z_t \mid S_t)}{\Pr(S_{1:t-1})} \\
&= K_t \cdot \frac{\Pr(\widehat{S_{1:t-1} \mid Z_{1:t-1}})}{\Pr(S_{1:t-1})} \cdot \Pr(Z_t \mid S_t) \\
&= K_t \cdot \widehat{r(S_{1:t-1})} \cdot \Pr(Z_t \mid S_t) \propto \widehat{r(S_{1:t-1})} \cdot \Pr(Z_t \mid S_t)
\end{aligned} \tag{3.163}$$

, where $K_t = \dfrac{C_{1:t}}{C_{1:t-1} \cdot \Pr(Z_t \mid Z_{1:t-1})}$.

The value of $K_t$ dose not impact the estimation because it would be eliminated in computation. Thus, $\widehat{r(S^i_{1:t})}$ can be updated by $\widehat{r(S^i_{1:t})} = \widehat{r(S^i_{1:t-1})} \cdot \Pr(Z_t \mid S^i_t)$ owing to $K_t$ is assigned by 1. The $\Pr(Z_t \mid S_t)$ here is the *observation model*. It means that the weights of the particles rely on the observations only. Therefore, the *incremental importance weight* here is equal to $\Pr(Z_t \mid S^i_t)$. All things considered, the algorithm of *bootstrap filter* is Alg. 7.

**Algorithm 7:** Bootstrap Filter

//: Setting

Set number of samples $N$;

Set fixed threshold of *effective sample size* : $N_{thres}$;

//: time $t = 1$ : Initialization

**for** $i \leftarrow 1$ *to* $N$ **do**

  Generate particle $S_1^i$ from $\Pr(S_1)$;

  Set the associated unnormalized weight $\widehat{r(S_1^i)} = \dfrac{\widehat{\Pr(S_1^i)}}{\Pr(S_1^i)} = C_1$;

**end**

//: Normalize

**for** $i \leftarrow 1$ *to* $N$ **do**

  Get *normalized importance weight* of $S_1^i$ : $w(S_{1:t}^i) = \dfrac{\widehat{r(S_1^i)}}{\sum_{j=1}^N \widehat{r(S_1^j)}} = \dfrac{C_i}{\sum_{j=1}^N C_j}$;

**end**

//: time $t \geq 2$ : Importance Sampling Step

**for** $i \leftarrow 1$ *to* $N$ **do**

  Sample $(S_t^i | S_{t-1}^i)$ from $\Pr(S_t \mid S_{t-1}^i)$;

  Set $S_{1:t}^i = S_{1:t-1}^i \cdot (S_t^i | S_{1:t-1}^i)$

  Update the associated unnormalized weight by $\widehat{r(S_{1:t}^i)} = \widehat{r(S_{1:t-1}^i)} \cdot \Pr(Z_t \mid S_t^i)$;

**end**

//: Normalize

**for** $i \leftarrow 1$ *to* $N$ **do**

  Get *normalized importance weight* of $S_{1:t}^i$ : $w(S_{1:t}^i) = \dfrac{\widehat{r(S_{1:t}^i)}}{\sum_{j=1}^N \widehat{r(S_{1:t}^j)}}$;

**end**

//: Resampling Step

Calculate the *effective sample size* : $N_e = \dfrac{1}{\sum_{j=1}^N w(S_{1:t}^j)^2}$

**if** $N_e < N_{thres}$ **then**

  Resampling using algorithm 4 : $\{S_{1:t}^1\}, \{S_{1:t}^2\}, \ldots, \{S_{1:t}^N\}$ = Resampling($\{S_{1:t}^1\}, \{S_{1:t}^2\}, \ldots, \{S_{1:t}^N\}$) ;

**end**