1 Importance Sampling for Priority Queue

1.1 Event-based simulation and Importance Sampling

Suppose there are K types of customers with priorities $w_1 > w_2 > \cdots > w_K$. And we are interested in the stationary waiting time of flow k' larger than a threshold $\gamma_{k'}$.

Algorithm 1: Change of Measure for Priority Queue

Consider a K-type priority queue, with arrival rates $\lambda_{1,0}, \ldots, \lambda_{K,0}$, and service rates $\mu_{1,0}, \ldots, \mu_{K,0}$.
Input: # of regenerative cycles N, tilted arrival rates $\lambda_{1,1}, \ldots, \lambda_{K,1}$, and tilted service rates $\mu_{1,1}, \ldots, \mu_{K,1}$.
for $j = 1, 2, \ldots, N$ **do**

- Initialize the current time $t = 0$, the packet index $m_1, \ldots, m_K = 0, \ldots, 0$, the number of packets being served $n_1, \ldots, n_K = 0, \ldots, 0$, the measure phase $\phi = 1$ and the current log-likelihood ratio $L = 0$.
- Generate arrival times of the first packet of each type $A_{1,1}, \ldots, A_{K,1}$ following the distributions $\exp(\lambda_{1,\phi}), \ldots, \exp(\lambda_{K,\phi})$, respectively.
- **while** *True* **do**
  - Find the next event: type-k arrives (E_1), type-k starts to be served (E_2), type-k departs (E_3).
  - **if** *type-k arrives (E_1)* **then**
    - A type-k packet comes with $m_k := m_k + 1$.
    - **if** *k is the target: $k = k'$* **then**
      - Compute the lower bound of the waiting time W_{k,m_k} for this packet, and change the phase ϕ to 0 if $W_{k,m_k} > \gamma_k$. (Switching 1). Note that for the highest priority, this is the exact waiting time.
    - **else if** *k has a higher priority: $k < k'$* **then**
      - Update the lower bound of waiting times of packets of flow k'. And change the phase ϕ to 0 if the following packet of flow k' exceeds $\gamma_{k'}$ (Switching 2) or there exists a packet i of flow k' with $W_{k',i} > \gamma_{k'}$ (Switching 3).
      - Need to be checked: Does the switching policy affect the likelihood ratio of each packet?
    - Generate the packet length (service time) S_{k,m_k} following the distribution $\exp(\mu_{k,\phi})$, and record the service phase $\phi^s_{k,m_k} := \phi$.
    - Generate the inter-arrival times A_{k,m_k+1} following the distributions $\exp(\lambda_{k,\phi})$, and record the arrival phase $\phi^a_{k,m_k+1} := \phi$.
  - **if** *type-k packet i starts to be served (E_2)* **then**
    - Add the arrival part for likelihood ratio:
      $$L := L + \text{logpdf}(\lambda_{k,0}, A_{k,i}) - \text{logpdf}(\lambda_{k,\phi^a_{k,i}}, A_{k,i}).$$
    - Assign the log-likelihood ratio for this packet: $L_{k,i} := L + \text{tail}(n_1, \ldots, n_K)$.
    - **if** *Current phase ϕ is 0 but the service phase $\phi^s_{k,i}$ is 1* **then**
      - Re-sample the service time $S_{k,i}$ from the distribution $\exp(\mu_{k,\phi})$, and record phase $\phi^s_{k,i} := \phi$.
    - Add the service part for likelihood ratio:
      $$L := L + \text{logpdf}(\mu_{k,0}, S_{k,i}) - \text{logpdf}(\mu_{k,\phi^s_{k,i}}, S_{k,i}).$$
  - **if** *type-k departs (E_3)* **then**
    - A type-k packet departs with $n_k := n_k + 1$.
    - Check whether the system is idle. If so, break the loop.
- For each type k, compute:
  $$R_k = \sum_{i=1}^{n_k} 1_{\{W_{k,i} > \gamma_k\}} e^{L_{k,i}}$$

Compute the mean and confidence interval.

1.2 Filtration and Likelihood Ratio

Let $A_{k,i}$ be the inter-arrival time between packet $i-1$ and packet i of flow k, and $S_{k,i}$ be the service time for packet i of flow k. Note that there is no packet at time 0, so $A_{k,1}$ is just the arrival time for packet 1 of flow k.

Now consider the filtration and likelihood ratio of the newly served packet of flow j, which starts to be served at time t. Let n_k be the number of packets for flow k, which have been served before the current packet. It can be seen that the filtration relies on the 3 parts:

1. Higher priority. For flow k, this includes any packet $1, 2, \dots, n_k$ and the inter-arrival of the following packet $n_k + 1$:

   $$\sigma(A_{k,1}, \dots, A_{k,n_k}, A_{k,n_k+1}, S_{k,1}, \dots, S_{k,n_k}) \tag{1}$$

   The likelihood ratio corresponds to

   $$\frac{d\mathbb{P}(A_{k,1} = a_{k,1}, \dots, A_{k,n_k} = a_{k,n_k}, S_{k,1} = s_{k,1}, \dots, S_{k,n_k} = s_{k,n_k})}{d\mathbb{Q}(A_{k,1} = a_{k,1}, \dots, A_{k,n_k} = a_{k,n_k}, S_{k,1} = s_{k,1}, \dots, S_{k,n_k} = s_{k,n_k})} \cdot \frac{\mathbb{P}(\sum_{i=1}^{n_k} a_{k,i} + A_{k,n_k+1} \geq s)}{\mathbb{Q}(\sum_{i=1}^{n_k} a_{k,i} + A_{k,n_k+1} \geq s)}$$

   where $\mathbb{P}$ is the original probability measure and $\mathbb{Q}$ is the importance probability measure which is used for sampling.

2. Same priority. For flow j, this includes any packet $1, 2, \dots, n_j$ and the inter-arrival of the following packet $n_j + 1$:

   $$\sigma(A_{j,1}, \dots, A_{j,n_j}, A_{j,n_j+1}, S_{j,1}, \dots, S_{j,n_j}) \tag{2}$$

   The likelihood ratio corresponds to

   $$\frac{d\mathbb{P}(A_{k,1} = a_{k,1}, \dots, A_{k,n_k} = a_{k,n_k}, A_{k,n_k+1} = a_{k,n_k+1}, S_{k,1} = s_{k,1}, \dots, S_{k,n_k} = s_{k,n_k})}{d\mathbb{Q}(A_{k,1} = a_{k,1}, \dots, A_{k,n_k} = a_{k,n_k}, A_{k,n_k+1} = a_{k,n_k+1}, S_{k,1} = s_{k,1}, \dots, S_{k,n_k} = s_{k,n_k})} \tag{3}$$

3. Lower priority. For flow k, this includes any packet $1, 2, \dots, n_k$, but if there is no packet served $n_k = 0$, the first one $A_{k,1}$ must need to be considered during the idle system:

   $$\sigma(A_{k,1}, \dots, A_{k,n_k}, S_{k,1}, \dots, S_{k,n_k}, A_{k,1}) \tag{4}$$

   The likelihood ratio corresponds to

   $$\frac{d\mathbb{P}(A_{k,1} = a_{k,1}, \dots, A_{k,n_k} = a_{k,n_k}, S_{k,1} = s_{k,1}, \dots, S_{k,n_k} = s_{k,n_k})}{d\mathbb{Q}(A_{k,1} = a_{k,1}, \dots, A_{k,n_k} = a_{k,n_k}, S_{k,1} = s_{k,1}, \dots, S_{k,n_k} = s_{k,n_k})} 1_{n_k \neq 0} + \frac{\mathbb{P}(A_{k,1} > \min_{i=1,\dots,K} a_{i,1})}{\mathbb{Q}(A_{k,1} > \min_{i=1,\dots,K} a_{i,1})} 1_{n_k = 0}$$

1.3 Cross-Entropy Method

The importance sampling estimator for the numerator is

$$E_{\mathbb{P}}\left[\sum_{i=1}^{\sigma_{k'}} 1_{W_{k',i}>\gamma_{k'}}\right] = E_{\mathbb{Q}}\left[\sum_{i=1}^{\sigma_{k'}} 1_{W_{k',i}>\gamma_{k'}} L_i\right], \tag{5}$$

where $\sigma_{k'}$ is the number of packets of flow k' in a regenerative cycle, and denote $E_{\mathbb{Q}}[H] = E_{\mathbb{Q}}\left[\sum_{i=1}^{\sigma_{k'}} 1_{W_{k',i}>\gamma_{k'}} L_i\right]$. The corresponding zero-variance IS distribution is

$$\mathbb{Q}^*(dx) = \frac{H(x)}{E_{\mathbb{Q}}[H]}\mathbb{Q}(dx), \tag{6}$$

where x is a sample path. The cross-entropy method is to find a measure $\mathbb{Q}_1$ with density h such that

$$\text{argmin}_{\mathbb{Q}_1} \text{KL}(\mathbb{Q}^* \| \mathbb{Q}_1) = \text{argmax}_h \int \log h(x)\mathbb{Q}^*(dx) = \text{argmax}_h \int H(x)\log h(x)\mathbb{Q}(dx). \tag{7}$$

Suppose we only change the measure of phase 1 to a new phase, say $u(1)$ with $u(0) = 0$, and then the density h of one path x sampled from $\mathbb{Q}$ is

$$h(x) = \prod_{k=1}^{K}\left(\prod_{i=1}^{\sigma_k+1} \lambda_{k,u(\phi^a_{k,i})} \exp(-\lambda_{k,u(\phi^a_{k,i})} a_{k,u(\phi^a_{k,i})}) \prod_{i=1}^{\sigma_k} \mu_{k,u(\phi^s_{k,i})} \exp(-\mu_{k,u(\phi^s_{k,i})} s_{k,u(\phi^s_{k,i})})\right), \tag{8}$$

and

$$\log h(x) = \sum_{k=1}^{K}\left(\sum_{i=1}^{\sigma_k+1}\left(\log \lambda_{k,u(\phi^a_{k,i})} - \lambda_{k,u(\phi^a_{k,i})} a_{k,u(\phi^a_{k,i})}\right) + \sum_{i=1}^{\sigma_k}\left(\log \mu_{k,u(\phi^s_{k,i})} - \mu_{k,u(\phi^s_{k,i})} s_{k,u(\phi^s_{k,i})}\right)\right),$$

and hence, the objective is to maximize

$$E_{\mathbb{Q}}\left[H(X)\sum_{k=1}^{K}\left(\sum_{i=1}^{\sigma_k+1}\left(\log \lambda_{k,u(\phi^a_{k,i})} - \lambda_{k,u(\phi^a_{k,i})} A_{k,u(\phi^a_{k,i})}\right) + \sum_{i=1}^{\sigma_k}\left(\log \mu_{k,u(\phi^s_{k,i})} - \mu_{k,u(\phi^s_{k,i})} S_{k,u(\phi^s_{k,i})}\right)\right)\right]. \tag{9}$$

Consider the empirical objective, and then the derivative and expectation can be exchanged:

$$\nabla_{\lambda_{k,u(1)}} \text{obj} = \frac{1}{N}\sum_{j=1}^{N}\left[H(x_j)\sum_{i=1}^{\sigma_k+1}\left(\frac{1}{\lambda_{k,u(1)}} - a^j_{k,u(1)}\right) 1_{\phi^a_{k,i}=1}\right] = 0$$

$$\nabla_{\mu_{k,u(1)}} \text{obj} = \frac{1}{N}\sum_{j=1}^{N}\left[H(x_j)\sum_{i=1}^{\sigma_k}\left(\frac{1}{\mu_{k,u(1)}} - s^j_{k,u(1)}\right) 1_{\phi^s_{k,i}=1}\right] = 0,$$

and hence, the cross-entropy parameters are

$$\lambda_{k,u(1)} = \frac{\sum_{j=1}^{N} H_k(x_j)\sum_{i=1}^{\sigma_k+1} 1_{\phi^a_{k,i}=1}}{\sum_{j=1}^{N} H_k(x_j)\sum_{i=1}^{\sigma_k+1} a^j_{k,u(1)} 1_{\phi^a_{k,i}=1}}, \quad k = 1, \dots, K \tag{10}$$

and

$$\mu_{k,u(1)} = \frac{\sum_{j=1}^{N} H(x_j)\sum_{i=1}^{\sigma_k} 1_{\phi^s_{k,i}=1}}{\sum_{j=1}^{N} H(x_j)\sum_{i=1}^{\sigma_k} s^j_{k,u(1)} 1_{\phi^s_{k,i}=1}}, \quad k = 1, \dots, K. \tag{11}$$