CS 240: Reasoning Under Uncertainty (Fall 21) Dec 1

Lecture 23

Instructor: *Profs Peter J. Hass and Jie Xiong* SI Worksheet: *Juelin Liu*

1 Introduction

We discussed simulation further in this lecture.

2 Simulation

Simulation is a tool for making decisions under uncertainty when analytical methods do not suffice. We have discussed the case where the analytical result for a Markov Chain is too complex to compute. Instead, we can simulate the result by computing multiple point estimates. Those point estimates can be represented as a normal distribution by the Central Limited Theorem. Then we can use the normal distribution to generate the confidence interval and associated confidence level of the simulation.

2.1 Central Limited Theorem Recap

Recall that we first learn CLT as the following:

Let $X_1, X_2, \ldots$ be s sequence of independent and identically distributed (i.i.d.) random variables (discrete or continuous). All these random variables have the same mean μ and variance δ^2. Its sample mean $\overline{X}_n$, which is also a random variable can be defined as

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i; \quad E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

$$var(\overline{X}_n) = var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) = \frac{\sigma^2}{n}$$

$$std(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}$$

$$Let\ Z_n = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}};\ then\ E(Z_n) = 0;\ Var(Z_n) = 1$$

The Central Limit Theorem states that the CDF of Z_n converges to the CDF of a standard normal random variable denoted as Φ. That is:

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x)$$

Z_n is approximately distributed as $N(0,1)$ for large n. $\overline{X}_n$ is approximately distributed as $N(\mu, \frac{\sigma^2}{n})$.

2.2 CLT in Simulation

In the case of simulation, we are gathering the values of the random variables whose variance is unknown. However, we can estimate the variance of the sample mean using the values collected. Let $Y_1, Y_2, \ldots, Y_n$ be the n samples we gathered.

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i;\ var(\overline{Y}_n) = s_k^2 = \frac{1}{n-1}\sum_{1}^{n}(Y_i - \overline{Y}_n)^2$$

The average of the samples is still a random variable with a distribution similar to a normal random variable. Thus, we can use it to generate an interval for estimating the analytical solution.

The normal distribution is continuous, and the probability that the analytical solution is equal to the sample average is 0 if the sample average is normally distributed. However, we can argue with some probability that the analytical solution should fall in a certain interval. This interval is the confidence interval and the likelihood that the analytical solution falls in this interval is the confidence level. We can reduce the variance of the sample mean to improve the estimation by increasing the number of simulation points we gathered.

2.3 Computing Condifence Interval

The method for computing the $100(1-\delta)\%$ **confidence interval** is the following:

- Choose z_δ such that $P(-z_\delta \leq Z \leq z_\delta) = 1 - \delta$
- By the central limit theorem, $\frac{\overline{Y}_k - \mu}{\sigma/\sqrt{k}}$ approximately a standard normal distribution so that

$$P(-z_\delta \leq \frac{\overline{Y}_k - \mu}{\sigma/\sqrt{k}} \leq z_\delta) \approx 1 - \delta$$

- in this case we estimate the standard deviation as $\sigma = s_k$
- rearrange the formula we get:

$$P(\overline{Y}_k - \frac{z_\delta \sigma}{\sqrt{k}} \leq \mu \leq \overline{Y}_k + \frac{z_\delta \sigma}{\sqrt{k}}) \approx 1 - \delta$$

- so the confidence interval is

$$[\overline{Y}_k - \frac{z_\delta \sigma}{\sqrt{k}}, \overline{Y}_k + \frac{z_\delta \sigma}{\sqrt{k}}]$$

  covers the unknown mean μ with probability $\approx 1 - \delta$

Note that the above method first chooses the confidence level $1 - \delta$ and then computes the associated confidence interval.

2.4 Multiple Systems

In this case, the system we want to simulate includes multiple sub-systems. We need to simulate each sub-system and the question is how to set the confidence level for each sub-system's simulation.

For example, suppose the entire system consists of two systems A and B. We want to achieve a confidence level of $100(1-\delta)\%$ for the entire system. Then the confidence level for A's and B's simulation should be $100(1-\frac{\delta}{2})\%$.

We want the confidence intervals for both systems to be valid simultaneously. However, we are not sure if the simulations for A and B are independent. Even systems A and B are independent, the random variables we used to simulate the two systems can be dependent. (In fact, you can this fact to get a narrower confidence interval when you estimate the difference between two systems with the same amount of computation).

In this case, we can still use Bonferroni's Inequality to compute the confidence level required for each sub-system's simulation. This is because the **Bonferroni's Inequality** works regardless the events are independent or not:

$$P(A \cap B) \geq 1 - P(A^c) - P(B^c) \text{ Bonferroni's Inequality}$$

If we set the confidence level for $L^A \leq \mu_A \leq U^A$ and $L^B \leq \mu_B \leq U^B$ to be $100(1-\frac{\delta}{2})\%$, then

$$P(L^A \leq \mu_A \leq U^A \cap L^B \leq \mu_B \leq U^B) \geq 1 - \frac{\delta}{2} - \frac{\delta}{2} = 1 - \delta$$

Similarly, if there are n sub-systems, then the confidence level for each sub-system's simulation would be $100(1-\frac{\delta}{n})\%$.