Markov Chains and Algorithmic Applications: WEEK 8

1 Sampling

1.1 Introduction

In this lecture we are interested in finding good sampling techniques to obtain samples from a probability distribution. In other words, given a probability distribution \( \pi \) on \( S \), how can we pick a random \( i \in S \) such that \( \Pr(i) = \pi_i \)?

But why would we want to do this?

Example 1.1 (Monte Carlo Integration). Suppose we want to compute \( \mathbb{E}(f(X)) \), with \( X \sim \pi \) (i.e. \( \Pr(X = i) = \pi_i, \ i \in S \)). By the definition of expectation we have

\[
\mathbb{E}(f(X)) = \sum_{i \in S} f(i)\pi_i
\]

(1)

Depending on the set \( S \), the above expression can be too expensive to compute exactly (i.e. computing it requires exponential time in \(|S|\)).

Instead of evaluating (1), we can compute the following approximation: take \( M \) i.i.d. samples \( X_1, \ldots, X_M \) from distribution \( \pi \) and compute

\[
\frac{1}{M} \sum_{k=1}^{M} f(X_k)
\]

(2)

Given some conditions on \( f(x) \), the law of large numbers guarantees

\[
\frac{1}{M} \sum_{k=1}^{M} f(X_k) \xrightarrow{M \to \infty} \mathbb{E}(f(X)) \quad \text{almost surely}
\]

But how big should \( M \) be for the approximation to be good? The variance of (2) is given by

\[
\text{Var} \left( \frac{1}{M} \sum_{k=1}^{M} f(X_k) \right) = \frac{1}{M} \text{Var}(f(X_1)) = \mathcal{O}\left( \frac{1}{M} \right)
\]

so \( \frac{1}{M} \sum_{k=1}^{M} f(X_k) \approx \mathbb{E}(f(X)) \pm \frac{C}{\sqrt{M}} \). We see that a good approximation requires taking \( M \) quite large.

A “simple” way to obtain samples is as follows:

Example 1.2 (“Simple” Sampling). Let \( X \) be a \( \pi \)-distributed random variable on \( S = \mathbb{N} \). If we can generate a continuous \( U(0,1) \) random variable \( U \), then we decide

\[
X = \begin{cases} 
0 & 0 \leq U \leq \pi_0, \\
1 & \pi_0 < U \leq \pi_0 + \pi_1, \\
\vdots & \\
i & \sum_{j=0}^{i-1} \pi_j < U \leq \sum_{j=0}^{i} \pi_j \\
\vdots & 
\end{cases}
\]

Hence \( \Pr(X = i) = \pi_i \).

As simple as the above sampling scheme seems, terms of the form \( \sum_{j=0}^{i} \pi_j \) (cdf of \( X \)) can be difficult to compute because we need to know each term \( \pi_j \) exactly: for \( \pi_j \) of the form \( \frac{h(j)}{Z} \), the normalization constant \( Z = \sum_{j \in S} h(j) \) can be non-trivial to compute depending on \( S \), as we will see below.

For the rest of the lecture, we will detail alternative sampling methods to try to side-step the issues above.
1.2 Importance Sampling

Consider again the Monte Carlo integration problem given above: our aim here is to find a better estimate of $E(f(X))$.

For this purpose, take another distribution $\psi = (\psi_i, i \in S)$ from which we know how to sample and let us define the coefficients $w_i = \frac{\pi_i}{\psi_i}$. Then

$$E(f(X)) = \sum_{i \in S} f(i)\pi_i = \sum_{i \in S} f(i)w_i\psi_i = E(f(Y)w(Y))$$

with $Y \sim \psi$. Since we know how to sample from $\psi$, we can approximate $E(f(Y)w(Y))$ by choosing $M$ i.i.d. samples $Y_1, \ldots, Y_M$ from $\psi$ and computing $\frac{1}{M} \sum_{k=1}^{M} f(Y_k)w(Y_k)$. We then have

$$\text{Var} \left( \frac{1}{M} \sum_{k=1}^{M} f(Y_k)w(Y_k) \right) = \frac{1}{M} \text{Var} (f(Y_1)w(Y_1))$$

As we did not assume anything in particular about the distribution $\psi$, we can choose it so as to minimize the variance of $f(Y_1)w(Y_1)$, which improves the approximation of the expectation (but note that the order in $M$ remains the same).

**Remark 1.3.** Why is this method called *importance sampling*? It turns out that the distribution $\psi$ minimizing the above variance puts more weight than $\pi$ itself on the states $i$ with a large probability $\pi_i$, and less weight on those with a small probability $\pi_i$: only the “important” states are therefore sampled with this method.

1.3 Rejection Sampling

Consider yet again the Monte Carlo integration problem (i.e. for $X \sim \pi$, compute $E(f(X))$), but assume now that we are unable to sample directly from $\pi$ (essentially because of the computation cost of this operation).

The idea behind rejection sampling is the following:

1. Take a distribution $\psi$ on $S$ from which samples can be easily produced (e.g. take $\psi$ uniform).
2. Take a sample $X$ from $\psi$.
3. Accept $X$ with some probability, or reject it with the complement probability.

Formally, let $\psi = (\psi_i, i \in S)$ be a distribution from which we can sample and define weights $\bar{w}_i = \frac{1}{c} \frac{\pi_i}{\psi_i}$ with $c = \max_{i \in S} \frac{\pi_i}{\psi_i} (\geq 1)$. The weights $\bar{w}_i$ play the role here of acceptance probabilities. Then

$$P(X = i) = \psi_i \bar{w}_i = \frac{\pi_i}{c}$$

$$P(X \text{ is rejected}) = 1 - \sum_{i \in S} P(X = i) = 1 - \sum_{i \in S} \frac{\pi_i}{c} = 1 - \frac{1}{c}$$

We therefore have

$$E(f(X)) \approx \frac{1}{M'} \sum_{k=1:X_k \text{ accepted}}^{M} f(X_k)$$

where $M'$ is the number of accepted samples among the $X_1, \ldots, X_M$.

The disadvantage of rejection sampling is that it may end up requiring much more samples than needed due to the sample rejection process (especially when the distance between $\pi$ and $\psi$ is large, i.e. when $c$ is large).