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 π on S, how can we pick a random $i \in S$ such that $\mathbb{P}(i) = \pi_i$?

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. $\mathbb{P}(X = i) = \pi_i$, $i \in S$). By the definition of expectation we have

$$\mathbb{E}(f(X)) = \sum_{i \in S} f(i)\pi_i \tag{1}$$

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 (**??**), we can compute the following approximation: take M i.i.d. samples $X_1, \ldots, X_M$ from distribution π and compute

$$\frac{1}{M} \sum_{k=1}^{M} f(X_k) \tag{2}$$

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

$$\frac{1}{M} \sum_{k=1}^{M} f(X_k) \underset{M \to \infty}{\longrightarrow} \mathbb{E}(f(X)) \text{ almost surely}$$

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

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

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 π-distributed random variable on $S = \mathbb{N}$. If we can generate a continuous $\mathcal{U}(0,1)$ random variable U, then we decide

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

Hence $\mathbb{P}(X = i) = \pi_i$.

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 π_j exactly: for π_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 $\mathbb{E}(f(X))$.

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

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

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

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

As we did not assume anything in particular about the distribution ψ, 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 ψ minimizing the above variance puts more weight than π itself on the states i with a large probability π_i, and less weight on those with a small probability π_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 $\mathbb{E}(f(X))$), but assume now that we are unable to sample directly from π (essentially because of the computation cost of this operation).

The idea behind rejection sampling is the following:

1. Take a distribution ψ on S from which samples can be easily produced (e.g. take ψ uniform).
2. Take a sample X from ψ.
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 $\widetilde{w}_i = \dfrac{1}{c}\,\dfrac{\pi_i}{\psi_i}$ with $c = \max_{i \in S} \dfrac{\pi_i}{\psi_i}$ (≥ 1). The weights $\widetilde{w}_i$ play the role here of acceptance probabilities. Then

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

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

We therefore have

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

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 π and ψ is large, i.e. when c is large).