Mathematical Kaleidoscope III:
Auxillary Variable Methods

Anders Malthe Westerkam[1], Malte Bødker[2], and Toke Christian Zinn[3]

[1]amw@es.aau.dk
[2]maltebn@math.aau.dk
[3]tokecz@math.aau.dk

October 31, 2023

Introduction

Suppose that for a given random variable X, we are interested in computing a certain functional or statistic. For example, we may be interested in computing $\mathbb{E}[f(X)]$ for some appropriate function f, for which the expected value is well-defined. Of course, if we can derive the law of $f(X)$ and subsequently compute the expected value, then we have solved the problem.

However, certain problems may be hard to solve analytically. In these cases, one often leverages results that yield an appropriate approximation, which is shown to converge. Continuing the example, we could for instance utilize, say, the strong or weak law of large numbers to approximate $\mathbb{E}[f(X)]$ by $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ for some *sufficiently large* $n \in \mathbb{N}$, where $(X_i)_{i=1}^{n}$ are assumed to be independent and identically distributed samples from the same distribution as X. Of course, the practical problem is obtaining samples from the law of X.

For certain distributions, direct sampling methods may be numerically intractable. However, the problem may become numerically feasible by introducing so-called *auxiliary variables.* Conditioning on the auxiliary variables, sampling the value of interest may be feasible, and one can then recover statistical quantities, such as, e.g., the expected value as above, through marginalization.

In these notes, we summarise the lectures and exercises presented by Jesper Møller in his lectures on "*Auxiliary variable methods for distributions with intractable normalizing constants, with a view to simulation-based Bayesian inference*".

Notation

Throughout the text, we assume that S is a *finite set.* That is, $|S| = n$ for some $n \in \mathbb{N}$, where $|\cdot|$ denoted the cardinality; we may enumerate the elements by $\{1,\ldots,n\}$ in this case.

For each $s \in S$, let $\mathscr{X}_s$ be a countable set. That is, $\mathscr{X}_s$ is either finite or countably infinite; formally, there exists a bijection between $\mathscr{X}_s$ and $\mathbb{N}$ or a finite subset thereof. Moreover, let

$$\mathscr{X} = \prod_{s\in S} \mathscr{X}_s. \tag{1}$$

For each $s \in S$, let X_s be an $\mathscr{X}_s$ valued random variable. Finally, by the collection $(X_s)_{s\in S}$ be the random tuple in $\mathscr{X}$ induced by each X_s for $s \in S$. We denote by

$$p(x_1,\ldots,x_n) = \mathbb{P}(X_1 = x_1,\ldots,X_n = x_n) \tag{2}$$

the probability mass function of the random tuple $(X_s)_{s\in S}$.

Point-processes on Finite Sets

In the lectures, we defined a (simple) point process on a finite set S as a *random set.* Formally, we let $\mathscr{X}_s \equiv \{0,1\}$ for all $s \in S$ and, by extension, the variables X_s are $\{0,1\}$ valued random variables for all $s \in S$. Then, we define the random set

$$X = \{s \mid X_s = 1\}. \tag{3}$$

Within this setting, we can think of each random variable X_s as an indicator for whether s is included in X; if $X_s = 1$, then the proposition $s \in X$, and conversely if $X_s = 0$, then $s \notin X$ for $s \in S$. By extension of (2), we may define the probability mass function for X by

$$p(A) = \mathbb{P}(X = A), \quad A \subseteq S;$$

notice that the above is well-defined X takes image in the power-set of S, which has cardinality 2^n. Hence, there is a finite number of elements, and we can assign a probability mass to each element, without resolving to measure theoretic arguments. We say that X is a Poisson process with rate $\beta > 0$, if X is simple, and

$$p(A) = \frac{1}{Z}\beta^{|A|}, \quad |A| \subseteq S,$$

where Z is a normalizing constant. The normalizing constant, Z, may be tedious to compute; suppose X is a point process in some finite set S with $|S| = n$ for some large $n \in \mathbb{N}$ with probability mass function p. Then, we know that

$$\sum_{A\subseteq S} p(A) = \sum_{A\subseteq S} \frac{1}{Z}\beta^{|A|} = 1.$$

Notice then, that

$$Z = \sum_{A\subseteq S} \beta^{|A|} = \sum_{A\in\wp(S)} \beta^{|A|},$$

where $\wp(S)$ denotes the power-set of S. Note, that $|\wp(S)| = 2^{|S|}$, and so if $|S|$ is large, then the sum may easily become numerically intractable as a result of the overwhelming number of terms[1]. Hence, we want to get rid of the normalizing constant. In certain applications, we can utilize auxiliary variables for exactly this purporse.

1 Area-interaction Point Processes

Suppose that the finite set S is endowed with a metric, $d : S \times S \to [0,\infty]$; in some texts, one required d to be finite. However, one can leverage the so-called standard bounded metric, $\bar{d}$ defined by

$$\bar{d}(s,t) = \min\left(d(s,t), 1\right), \qquad s, s' \in S.$$

In fact, one can show that d and $\bar{d}$ are topologically equivalent. For any subset $A \subseteq S$ and $R > 0$ we define

$$A_{\oplus R} = \left\{ s \in S \;\middle|\; \min_{a\in A} d(a,s) \leq R \right\},$$

where we use the convention that $\min_{a\in\emptyset} d(a,s) = \infty$ for all $s \in S$; hence $\emptyset_{\oplus R} = \emptyset$. Notice, that for a non-empty set $A \subseteq S$, the set $A_{\oplus R}$ is an enlargement of A by including all the points that have a distance of at

[1] This relates to the 3rd subquestion of Q5.

most R to *some* element $a \in A$.

For instance, suppose that $S = \{1, \dots, n\}$ and associate for each $s \in S$ the angle

$$\theta_s = \frac{2\pi s}{n}. \tag{4}$$

Then, we can associate each element $s \in S$ with a point on a circle defined by $(\cos(\theta_s), \sin(\theta_s))$; we visualize in Figure 1. Next, we endow S with a metric $d : S \times S \to S$ defined by

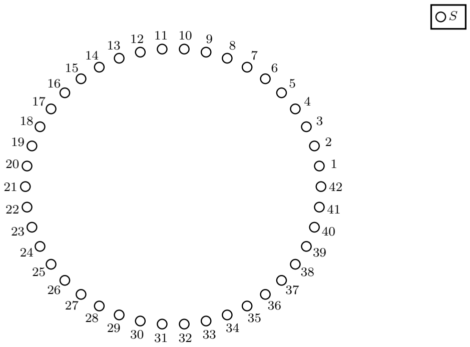

Figure 1: The points $(\cos(\theta_s), \sin(\theta_s))$ for $s \in S = \{1, \dots, 42\}$, where θ_s is defined by (4).

$$d(s,t) = \min(t - s, s + n - t), \qquad s \le t,$$

or equivalently

$$d(s,t) = \min(\max(s,t) - \min(s,t), \min(s,t) + n - \max(s,t)).$$

Within the circle embedding, we may think of the the points as a graph with vertices $(\cos(\theta_s), \sin(\theta_s))$ for $s \in S$ and edges between each $(\cos(\theta_s), \sin(\theta_s))$ and $(\cos(\theta_{s+1}), \sin(\theta_{s+1}))$ for each $s \in \{1, \dots, n-1\}$, aswell as an edge between $(\cos(\theta_n), \sin(\theta_n))$ and $(\cos(\theta_1), \sin(\theta_1))$; see Figure 2. Then, $d(s,t)$ corresponds to the number of edges one must traverse to travel from s to t in the undirected graph. Now, recall that $s \in A_{\oplus 1}$ if

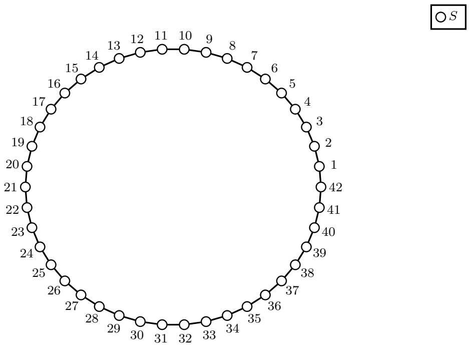

Figure 2: The graph induced by adding edges between $(\cos(\theta_s), \sin(\theta_s))$ and $(\cos(\theta_{s+1}), \sin(\theta_{s+1}))$ for each $s \in \{1, \dots, n-1\}$, aswell as an edge between $(\cos(\theta_n), \sin(\theta_n))$ and $(\cos(\theta_1), \sin(\theta_1))$.

there exists an $a \in A$ such that $d(a,t) \leq R$. Hence, $A_{\oplus R}$ corresponds to the set of vertices A as well as as the vertices which can be reached from a vertex in A by traversing at most R edges; we visualize in Figure 3.

Now, suppose that we are interested in simulating a point process on S with the probability mass function

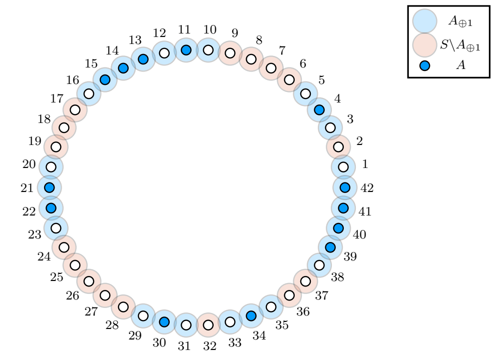

Figure 3: The set $A_{\oplus 1}$ for a given set A visualized in the circle.

$$p(A) = \frac{1}{Z}\beta^{|A|}\gamma^{-|A_{\oplus R}|}, \quad A \subseteq R, \tag{5}$$

where β, γ, and R are positive constants and Z is a normalizing constant. Now, simulating from (5) is hard for two reasons. First, if n is large, it is not feasible to store all outcomes in memory, and in particular computing Z is numerically intractable, c.f. the earlier remarks. Secondly, as the name suggests, there is an *area interaction* enforced by the relationship between β and γ. Of course, the trivial case $\gamma = 1$ is simple, as (5) becomes $\frac{1}{Z}\beta^{|A|}$ which is a Poisson process with parameter β, for which each X_s follows an i.i.d. Bernoulli distribution with parameter $\frac{\beta}{1+\beta}$ for all $s \in S$, which we can easily simulate by sampling each X_s independently[2].

To solve these problems, we introduce an auxiliary variable.

Widom-Rowlinson Model

For X following an area-interaction process, we will now consider an auxiliary variable approach leading to the so-called Widom-Rowlinson model.

Within the Widom-Rowlinsom model, we also consider a finite metric space (S,d). Let Y_1 and Y_2 denote two independent Poisson processes on S and condition on $d(Y_1, Y_2) = \min_{s\in Y_1, t\in Y_2} d(s,t) > R$ for some $R > 0$ called the hard-core; intuitively, the condition separates the two otherwise independent Poisson process, forcing them to keep at least a distance of R to one-another.

[2] This comment answers part 1 of Q5 and part 3 of Q1

Now, let us derive the probability mass function for the conditioned process. Note, that

$$\begin{aligned}\mathbb{P}(Y_1 = A_1, Y_2 = A_2 \mid d(Y_1, Y_2) > R) &= \frac{\mathbb{P}(Y_1 = A_1, Y_2 = A_2, d(Y_1, Y_2) > R)}{\mathbb{P}(d(Y_1, Y_2) > R)} \\ &\propto \mathbb{P}(Y_1 = A_1, Y_2 = A_2, d(Y_1, Y_2) > R) \\ &= \mathbb{P}(Y_1 = A_1, Y_2 = A_2, d(A_1, A_2) > R) \\ &= \mathbb{P}(Y_1 = A_1, Y_2 = A_2)\mathbb{1}[d(A_1, A_2) > R] \\ &= \mathbb{P}(Y_1 = A_1)\mathbb{P}(Y_2 = A_2)\mathbb{1}[d(A_1, A_2) > R],\end{aligned}$$

where the first equation follows by definition. The proportionaly then simply ignores the fixed divisor. Subsequently, the joint distribution of $\mathbb{P}(Y_1 = A_1, Y_2 = A_2, d(Y_1, Y_2) > R)$ can be written as $\mathbb{P}(Y_1 = A_1, Y_2 = A_2, d(A_1, A_2) > R)$, as we consider a joint distribution. However, the term $d(A_1, A_2) > R$ is no longer random, in the sense that it is either true or false for a given A_1 and A_2. Hence, we can rewrite $\mathbb{P}(Y_1 = A_1, Y_2 = A_2, d(A_1, A_2) > R) = \mathbb{P}(Y_1 = A_1, Y_2 = A_2)\mathbb{1}[d(A_1, A_2) > R]$. The final equality then follows from the independence of Y_1 and Y_2 prior to the hard-core condition[3]. Notice, that if we condition on $Y_2 = A$, we have

$$\mathbb{P}(Y_1 = A_1 \mid Y_2 = A, d(Y_1, Y_2) > R) \propto \beta_1^{|A_1|}\mathbb{1}[d(A_1, A) > R],$$

for all $A_1 \subseteq S$. However, if $A_1 \subseteq S \backslash A_{\oplus R}$, then clearly $d(A_1, A) > R$; it corresponds to the set-difference between S and $A_{\oplus R}$, where the latter is defined by all the points s for which there exists an $a \in A$ such that $d(s, a) \leq R$. Hence, for s in the complement $S \backslash A_{\oplus R}$ we must have $d(s, a) > R$ for all $a \in A$. Therefore, we have for $A_1 \subseteq S \backslash A_{\oplus R}$ we have

$$\mathbb{P}(Y_1 = A_1 \mid Y_2 = A, d(Y_1, Y_2) > R) \propto \beta_1^{|A_1|},$$

which coincides with the probability mass function of a Poisson process on S with parameter β_1. Hence, conditioned on $Y_2 = A$, we have that Y_1 is a Poisson process with parameter β_1 on $S \backslash A_{\oplus R}$. Analogous arguments show that conditioned on $Y_1 = A$, we have that Y_2 is a Poisson process on $S \backslash A_{\oplus R}$ with parameter β_2.[4].

In particular, since Y_1 and Y_2 were independent, but we later condition on the hard-core condition, we can first simulate Y_1, which was assumed to be a Poisson process on S with parameter β_1, which yields and outcome $Y_1 = A$. Then, we can subsequently simulate Y_2 on $S \backslash A_{\oplus R}$ as a Poisson process with parameter β_2. Then, the pair (Y_1, Y_2) will follow the joint distribution obeying the hard-core condition.[5]

Now, let us derive the conditional probability mass function of Y_1 given the hard-core condition. To this end, notice that

$$\begin{aligned}\mathbb{P}(Y_1 = A \mid d(Y_1, Y_2) > R) &= \sum_{B \in \wp(S)} \mathbb{P}(Y_1 = A, Y_2 = B \mid d(Y_1, Y_2) > R) \\ &\propto \sum_{B \in \wp(S)} \beta_1^{|A|}\beta_2^{|B|}\mathbb{1}[d(A, B) > R] \\ &= \beta_1^{|A|} \sum_{B \in \wp(S)} \beta_2^{|B|} \cdot \mathbb{1}[d(A, B) > R] \\ &= \beta_1^{|A|} \sum_{B \subseteq S \backslash A_{\oplus R}} \beta_2^{|B|} \\ &= \beta_1^{|A|} \cdot \sum_{n=0}^{|S \backslash A_{\oplus R}|} \binom{|S \backslash A_{\oplus R}|}{n} \beta^n \\ &= \beta_1^{|A|}(1 + \beta_2)^{|S \backslash A_{\oplus R}|} \\ &\propto \beta_1^{|A|}(1 + \beta_2)^{-|A_{\oplus R}|},\end{aligned}$$

[3] This answers the 1st question in Q2

[4] This answers the 2nd point of Q2.

[5] This answers the 3rd question in Q2.

which gives us a marginal distribution for Y_1, albeit still conditional on the hard-core condition. One shows that

$$\mathbb{P}(Y_2 = A \mid d(Y_1, Y_2) > R) \propto \beta_2^{|A|}(1+\beta_1)^{-|A_{\oplus R}|}$$

analogously.[6] Now, notice that the last proportionality coincides exactly with the area-interaction model when $\gamma > 1$; set $\gamma = (1+\beta_2)$ and notice that the proportionality accounts for the normalizing constant.[7]

In particular, we can, conditioned on the hard-core condition, consider Y_2 (or Y_1) as an auxiliary variable for the area-interaction process with $\gamma > 1$; set $X = Y_1$ and let Y_2 be the auxiliary variable, where we implicitly assume that Y_1 and Y_2 are conditioned on the hard-core condition. [8]

1.1 Simulation from the Widom-Rowlinson

Now, we consider a Gibbs sampler for the Widom-Rowlinsom model. We exemplify using the space $S = \{1, \dots, n\}$ which we embed on the circle using the transformation

$$s \mapsto (\cos(\theta_s), \sin(\theta_s)), \quad \theta_s = \frac{2\pi s}{n}.$$

The simulation is then carried out as follows:

(i) Initialize Y_1; we choose to sample Y_1 as a Poisson process on S with parameter β_1.

(ii) Sample Y_2 as a Poisson process with parameter β_2 on $S \backslash (Y_1)_{\oplus R}$.

(iii) Sample Y_1 as a Poisson process with parameter β_1 on $S \backslash (Y_2)_{\oplus R}$.

(iv) Alternate points (ii) and (iii).

2 Markov Random Fields

We now alleviate the need for $\mathscr{X}_s$ to be $\{0, 1\}$ in favor of an arbitrary countable space for each $s \in S$.

In particular, we refer to the vector $(X_s)_{s \in S}$ as a random field, and for simplicity we assume that $\mathbb{P}(X = \{x\}) > 0$ for all $x \in \mathscr{X}$. We are now interested in extending the *Markov* property. In the classical theory for Markov chains, there is an implicit ordering and the Markov condition can then be stated as the probability of the next event given all the previous events reduces to the conditioning on the most recent event. However, if S is, say, a lattice, then there is no natural ordering; one could choose the lexicographic ordering, but it is not nessecarily the natural ordering.

To extend the idea, note that any ordering is also a relation. Hence, we can relax the ordering and replace it with a relation, $\sim$, which satisfies

(i) $s \not\sim s$,

(ii) $s \sim t$ implies $t \sim s$,

for all $s, t \in S$. Notice, that (ii) is violated in the case of an ordering; if $s < t$ then clearly we cannot have $s > t$.

Nonetheless, for each $s \in S$, we can define the set

$$\partial(s) = \{t \in S \mid s \sim t\},$$

[6] This answers all the questions in Q4.

[7] This answers the 2$^{\text{nd}}$ question in Q5.

[8] This answers Q16.

and subsequently, define the set

$$\partial = \{\partial(s) \mid s \in S\}.$$

Then, we say that a random field $(X_s)_{s \in S}$ is a Markov random field, if

$$\mathbb{P}(X_s = x_s \mid X_t = x_t, t \in S \backslash \{s\}) = \mathbb{P}(X_s = x_s \mid X_t = x_t, t \in \partial(s));$$

intuitively, this means that $(X_s)_{s \in S}$ is a Markov random field if each X_s depends only on the *related* points X_t for $t \in \partial(s)$, which is advantageous if, say, the cardinality of S is very large, but the cardinality of $\partial(s)$ is relatively small for each $s \in S$.

For example, in the case of an area interaction