Conditional Bernoulli

Riyad Kettaf
Clément de Lardemelle

April 2025

Rejection sampling

Justification of Rejection Sampling for Conditional Bernoulli Distribution

Let $X = (X_1, \ldots, X_N)$ be a vector of independent Bernoulli random variables, with $X_i \sim \text{Bernoulli}(p_i)$, and define the joint distribution:

$$g(x) = \prod_{i=1}^{N} p_i^{x_i}(1-p_i)^{1-x_i}, \quad x \in \{0,1\}^N.$$

Let the target distribution be the law of X conditioned on the event $\sum_{i=1}^N X_i = k$. We define the conditional distribution f by:

$$f(x) = \mathbb{P}(X = x \mid \sum_{i=1}^{N} X_i = k) = \frac{g(x) \cdot \mathbf{1}\left\{\sum_{i=1}^N x_i = k\right\}}{\mathbb{P}\left(\sum_{i=1}^N X_i = k\right)}.$$

We aim to sample from f using rejection sampling with proposal g.

Rejection Sampling Scheme

The rejection sampler proceeds as follows:

1. Sample $x \sim g$, i.e., generate $X_i \sim \text{Bernoulli}(p_i)$ independently.
2. Accept x if $\sum_{i=1}^N x_i = k$; otherwise, reject and repeat.

Validity of the Rejection Sampler

This algorithm is valid because:

- The support of the target distribution f is contained within the support of the proposal g.
- For any $x \in \{0,1\}^N$ such that $\sum x_i = k$, we have

$$f(x) = \frac{g(x)}{\mathbb{P}\left(\sum_{i=1}^N X_i = k\right)}.$$

- This implies that for all such x,

$$f(x) \leq M \cdot g(x), \quad \text{where } M = \frac{1}{\mathbb{P}\left(\sum_{i=1}^{N} X_i = k\right)} < \infty.$$

This satisfies the standard requirement of rejection sampling: there exists a constant M such that

$$f(x) \leq M \cdot g(x) \quad \text{for all } x.$$

Main drawback of this approach

Exact Expression of the Poisson-Binomial PMF

Let $X_1, \ldots, X_N$ be independent Bernoulli random variables with success probabilities $p_1, \ldots, p_N$. Define the sum:

$$S = \sum_{i=1}^{N} X_i.$$

Then the probability mass function of S, for an integer $k \in \{0, 1, \ldots, N\}$, is given by:

$$\mathbb{P}(S = k) = \sum_{\substack{A \subseteq \{1,\ldots,N\} \\ |A|=k}} \prod_{i \in A} p_i \prod_{j \notin A} (1 - p_j).$$

While this formula is exact, it involves a sum over $\binom{N}{k}$ subsets and does not provide a clear analytical understanding of how $\mathbb{P}(S = k)$ behaves as a function of the size N or the distribution of the probabilities $p = (p_1, \ldots, p_N)$.

Central Limit Theorem Approximation

Let:

$$\mu_N = \sum_{i=1}^{N} p_i, \qquad \sigma_N^2 = \sum_{i=1}^{N} p_i(1 - p_i)$$

be the mean and variance of S.

We consider the standardized sum:

$$Z_N = \frac{S - \mu_N}{\sigma_N}.$$

To apply the Central Limit Theorem for independent but not identically distributed random variables, we invoke the **Lindeberg–Feller Central Limit Theorem**:

Theorem (Lindeberg–Feller CLT). *Let $\{X_i\}$ be a sequence of independent random variables with finite means μ_i and variances σ_i^2. Define $S_N = \sum X_i$, $\mu_N = \sum \mu_i$, and $\sigma_N^2 = \sum \sigma_i^2$. If for every $\varepsilon > 0$,*

$$\frac{1}{\sigma_N^2} \sum_{i=1}^{N} \mathbb{E}\left[(X_i - \mu_i)^2 \cdot \mathbf{1}_{\{|X_i - \mu_i| > \varepsilon \sigma_N\}}\right] \xrightarrow{N \to \infty} 0,$$

then

$$\frac{S_N - \mu_N}{\sigma_N} \xrightarrow{d} \mathcal{N}(0,1).$$

In our case, $X_i \sim \text{Bernoulli}(p_i)$, hence $|X_i - \mu_i| \leq 1$, and

$$\mathbb{E}\left[(X_i - p_i)^2 \cdot \mathbf{1}_{\{|X_i - p_i| > \varepsilon \sigma_N\}}\right] = 0 \quad \text{for large enough } N,$$

since the indicator becomes zero when $\sigma_N > \varepsilon^{-1}$. Therefore, the Lindeberg condition is satisfied automatically for Bernoulli variables.

Thus, by the CLT,

$$\frac{S - \mu_N}{\sigma_N} \xrightarrow{d} \mathcal{N}(0,1),$$

which implies that for large N,

$$\mathbb{P}(S = k) \approx \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left(-\frac{(k - \mu_N)^2}{2\sigma_N^2}\right),$$

at least when k is in a neighborhood of μ_N.

Behavior as a Function of N and p

As $N \to \infty$, the total variance

$$\sigma_N^2 = \sum_{i=1}^{N} p_i(1 - p_i)$$

also grows to infinity, provided that the probabilities p_i are bounded away from 0 and 1.

As a consequence, since the normal approximation is:

$$\mathbb{P}(S = k) \approx \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left(-\frac{(k - \mu_N)^2}{2\sigma_N^2}\right),$$

the denominator increases without bound and the exponential term remains bounded. Therefore,

$$\mathbb{P}(S = k) \to 0 \quad \text{as } N \to \infty.$$

This reflects the fact that the mass of the distribution becomes increasingly spread out over a wider range of values as N increases.

Conclusion

For the reasons discussed above, the rejection sampling approach becomes increasingly inefficient as N grows large or when the success probabilities p_i are extreme (i.e., close to 0 or 1). In such cases, the variance σ_N^2 may grow slowly or the target event $\{S = k\}$ may become very unlikely. Since the acceptance rate of rejection sampling is exactly $\mathbb{P}(S = k)$, which tends to zero as $N \to \infty$, the expected number of trials before acceptance increases drastically.

Exact Sampling Algorithm from Appendix A (Pseudocode and Complexity)

We wish to sample a binary vector $x = (x_1, \ldots, x_N) \in \{0,1\}^N$ with independent components $X_i \sim \text{Bernoulli}(p_i)$, conditioned on the constraint

$$\sum_{i=1}^{N} X_i = I.$$

The following algorithm computes such a sample exactly using a dynamic programming table $q(i,n)$, which represents:

$$q(i,n) = \mathbb{P}\left(\sum_{m=n}^{N} X_m = i\right).$$

Pseudocode

1. **Initialization of $q(0,n)$:**
   Set $q(0, N+1) = 1$. Then for $n = N, \ldots, 1$,

   $$q(0,n) = (1-p_n) \cdot q(0, n+1).$$

2. **Recursive computation for $i = 1, \ldots, I$:**
   Set $q(i, N+1) = 0$ for all $i > 0$. Then for each $i = 1, \ldots, I$ and for $n = N, \ldots, 1$,

   $$q(i,n) = (1-p_n) \cdot q(i, n+1) + p_n \cdot q(i-1, n+1),$$

   with the convention that $q(-1, n+1) = 0$.

3. **Sampling step:**
   Initialize $x = []$, and let $i = I$ be the number of remaining ones to place. For $n = 1, \ldots, N$,

   - Compute the conditional probability:

     $$r = \frac{p_n \cdot q(i-1, n+1)}{q(i,n)}.$$

   - Sample $x_n \sim \text{Bernoulli}(r)$.
   - If $x_n = 1$, decrement $i \leftarrow i - 1$.

   At the end, return the sample $x = (x_1, \ldots, x_N)$.

Complexity Justification

- The initialization step for $q(0,n)$ uses a recursion from N down to 1, with a single multiplication per step. This takes $O(N)$ time.
- The recursive step fills in the table $q(i,n)$ for $i = 1, \ldots, I$ and $n = N, \ldots, 1$. This is a table of size $(I+1) \times (N+1)$, and each entry requires constant time. Assuming I can be as big as N, this step has worst case time complexity of $O(N^2)$.

- The sampling step runs in $O(N)$ time, since it samples each component once.
- **Overall complexity:** The dominant cost comes from the recursive table construction, leading to a total time complexity of:
$$O(N^2).$$

MCMC Sampler

Proof of Invariance of the Target Distribution for the MCMC Sampler

Let $x \in \{0,1\}^N$ be a binary vector with exactly k ones, and let $\pi(x)$ denote the target distribution, namely the conditional Bernoulli distribution given $\sum_{i=1}^N x_i = k$. We define

$$\pi(x) \propto \prod_{i=1}^{N} \left(\frac{p_i}{1-p_i} \right)^{x_i},$$

subject to $\sum_{i=1}^N x_i = k$. For convenience, define

$$w_i = \frac{p_i}{1-p_i}.$$

Then, for two states x and x' differing only by a swap between two coordinates i_0 and i_1 (with $x_{i_0} = 0$, $x_{i_1} = 1$ and $x'_{i_0} = 1$, $x'_{i_1} = 0$), the ratio of their probabilities is

$$\frac{\pi(x')}{\pi(x)} = \frac{w_{i_0}}{w_{i_1}}.$$

The Proposal and Acceptance Mechanism

At each MCMC step, the algorithm:

1. Randomly selects an index i_0 from the set $\{i : x_i = 0\}$ and an index i_1 from the set $\{i : x_i = 1\}$.
2. Proposes the state x' obtained by swapping the entries at i_0 and i_1.
3. Accepts x' with probability
$$\alpha(x, x') = \min\left(1, \frac{w_{i_0}}{w_{i_1}}\right).$$

Since the proposal mechanism is *symmetric* (i.e., the probability of proposing x' from x is the same as proposing x from x' because both involve uniformly selecting a pair of indices), we have:

$$q(x \to x') = q(x' \to x).$$

Detailed Balance Condition

The detailed balance condition requires that for every pair of states x and x',

$$\pi(x)P(x \to x') = \pi(x')P(x' \to x),$$

where $P(x \to x')$ is the overall transition probability from x to x'.

Since the proposal probabilities are symmetric, we have:

$$P(x \to x') = q(x \to x')\alpha(x, x') \quad \text{and} \quad P(x' \to x) = q(x' \to x)\alpha(x', x).$$

Thus, it suffices to show that

$$\pi(x)\alpha(x, x') = \pi(x')\alpha(x', x).$$

There are two cases to consider:

Case 1: If $\frac{w_{i_0}}{w_{i_1}} \leq 1$, then

$$\alpha(x, x') = \frac{w_{i_0}}{w_{i_1}} \quad \text{and} \quad \alpha(x', x) = 1.$$

Therefore,

$$\pi(x)\alpha(x, x') = \pi(x)\frac{w_{i_0}}{w_{i_1}},$$

and since

$$\frac{\pi(x')}{\pi(x)} = \frac{w_{i_0}}{w_{i_1}},$$

we obtain

$$\pi(x)\alpha(x, x') = \pi(x') = \pi(x') \cdot 1 = \pi(x')\alpha(x', x).$$

Case 2: If $\frac{w_{i_0}}{w_{i_1}} > 1$, then

$$\alpha(x, x') = 1 \quad \text{and} \quad \alpha(x', x) = \frac{w_{i_1}}{w_{i_0}}.$$

Thus,

$$\pi(x')\alpha(x', x) = \pi(x')\frac{w_{i_1}}{w_{i_0}},$$

and using the same probability ratio, we have

$$\pi(x')\frac{w_{i_1}}{w_{i_0}} = \pi(x).$$

Hence,

$$\pi(x')\alpha(x', x) = \pi(x) = \pi(x) \cdot 1 = \pi(x)\alpha(x, x').$$

In both cases, the detailed balance condition is satisfied. Therefore, the MCMC sampler leaves the target distribution $\pi(x)$ invariant.

Conclusion

Since the Markov chain is constructed so that:

1. The proposal distribution is symmetric,
2. The acceptance probability is chosen to satisfy detailed balance,
3. The state space is irreducible (every state can be reached from any other through a series of swaps) and aperiodic (due to the possibility of rejecting a move),

we conclude that the Markov chain converges to the target distribution $\pi(x)$.

RQMC

Randomized QMC Acceleration

Instead of using independent uniform variables, we use a *randomized quasi-Monte Carlo* (RQMC) sequence — typically a scrambled Sobol sequence — to generate the vector $(U_1, \ldots, U_N)$. This preserves the structure of the dynamic algorithm while introducing a low-discrepancy structure into the sampling process.

Why RQMC Leads to Faster Convergence

- In the original (Appendix A) algorithm, the convergence of empirical estimates (e.g., frequencies over $\mathcal{S}_{N,k}$) is limited by the variance inherent to i.i.d. uniform sampling:

$$\text{MSE}_{\text{MC}} = \mathcal{O}\left(\frac{1}{n}\right).$$

- RQMC replaces the i.i.d. uniform draws with a **structured**, low-discrepancy input sequence, leading to improved uniformity of the sample space coverage. For smooth-enough conditional transitions, this translates into:

$$\text{MSE}_{\text{RQMC}} = \mathcal{O}\left(\frac{1}{n^{2-\epsilon}}\right), \quad \text{for any } \epsilon > 0.$$