# FAST SIMULATION OF MULTISTAGE CLONAL EXPANSION MODELS

LUKAS KÖSTLER

*Technical University of Munich (TUM)*

ABSTRACT. We derive and demonstrate a method to simulate Multistage Clonal Expansion (MSCE) models. The method is faster than previous ones while retaining similar accuracy.

## 1. DERIVATION

1.1. **Poisson process with random start time.** We will consider a non-homogeneous Poisson process with rate $\lambda(t) > 0$ which has a random start time $\tau \geq 0$ with distribution $p$. For fixed $\tau$ the number of occurrences at time $t \geq \tau$ given start time $\tau$, $N(t)|\tau$ has the characteristic function [3, chapter 4, eqn. (2.1)]

$$\begin{aligned} \varphi_{N(t)|\tau}(u) &= \exp\left[m(t-\tau)\left\{e^{iu}-1\right\}\right] \\ m(t) &= \int_0^t \lambda(s)\,\mathrm{d}s\,. \end{aligned} \tag{1.1}$$

The by the law of total expectation, the characteristic function of $N(t)$ is given by

$$\begin{aligned} \varphi_{N(t)}(u) &= E\left[\varphi_{N(t)|\tau}(u)\right] \\ &= \int_0^t \exp\left[m(t-\tau)\left\{e^{iu}-1\right\}\right] p(\tau)\ \mathrm{d}\tau + \int_t^\infty p(\tau)\ \mathrm{d}\tau\,. \end{aligned}$$

Comparing this expression to a Poisson process with different rate yields the first theorem.

**Theorem 1.1.** *Let $N(t)$ follow a non-homogeneous Poisson process with rate $\lambda(t) \geq 0$ starting at a random time $\tau \geq 0$ with probability density function $p(\tau)$. Let $M(t)$ follow a non-homogeneous Poisson process starting at time $0$ with rate*

$$\nu(t) = \int_0^t \lambda(t-\tau)\,p(\tau)\,\mathrm{d}s = (\lambda * p)(t)\ \ . \tag{1.2}$$

*Then, for the characteristic functions $\varphi_{N(t)}$ and $\varphi_{M(t)}$ there holds*

$$\max_{t\in[0,\sigma]}\left|\varphi_{N(t)} - \varphi_{M(t)}\right| \leq 2\left(\exp\left(2m(\sigma)\right) - 2m(\sigma) - 1\right) = O(m(\sigma)^2) \tag{1.3}$$

*where $m$ is the mean value function of $N(t)$ as defined in eqn.* (1.1).

*E-mail address*: lukas.koestler@tum.de.

*Date*: July 9, 2018.

*Proof.* Let $h$ be the mean value function (eqn. (1.1)) of $M(t)$, then there holds

$$\frac{d}{dt}\int_0^t m(t-\tau)\,p(\tau)\,\mathrm{d}\tau = \int_0^t m'(t-\tau)\,p(\tau)\,\mathrm{d}\tau = \int_0^t \lambda(t-\tau)\,p(\tau)\,\mathrm{d}\tau = \frac{d}{dt}\nu(t)\ .$$

We used the Leibniz integration rule and $m(0)=0$ in the first step. Because $h(0)=0$ has to hold, we get

$$h(t) = \int_0^t m(t-\tau)\,p(\tau)\,\mathrm{d}\tau = (m*p)(t)$$

For $\varphi_{N(t)}(u)$, using $k := \left\{e^{iu}-1\right\}$, we obtain

$$\begin{aligned}\varphi_{N(t)}(u) &= \int_0^t \exp\left[m(t-\tau)\,k\right]p(\tau)\ \mathrm{d}\tau + \int_t^\infty p(\tau)\ \mathrm{d}\tau\\ &= 1 + k\int_0^t m(t-\tau)\,p(\tau)\ \mathrm{d}\tau + \sum_{j=2}^\infty \frac{k^j}{j!}\int_0^t m(t-\tau)^j\,p(\tau)\ \mathrm{d}\tau\,.\end{aligned}$$

For $\varphi_{M(t)}(u)$ we obtain

$$\varphi_{M(t)}(u) = \exp(h(t)\,k) = 1 + kh(t) + \sum_{j=2}^\infty \frac{k^j}{j!}h(t)^j\ .$$

Let $t\in[0,\sigma]$ then there holds

$$\begin{aligned}\left|\varphi_{N(t)}-\varphi_{M(t)}\right| &= \Big|\sum_{j=2}^\infty \frac{k^j}{j!}\int_0^t m(t-\tau)^j\,p(\tau)\ \mathrm{d}\tau - \sum_{j=2}^\infty \frac{k^j}{j!}h(t)^j\ \Big|\\ &\le \sum_{j=2}^\infty \frac{|k|^j}{j!}\Big|\int_0^t m(t-\tau)^j\,p(\tau)\ \mathrm{d}\tau - \left(\int_0^t m(t-\tau)\,p(\tau)\ \mathrm{d}\tau\right)^j\ \Big|\\ &\le \sum_{j=2}^\infty \frac{|k|^j}{j!}\left(|m(\sigma)|^j + |m(\sigma)|^j\right)\\ &= 2\sum_{j=2}^\infty \frac{(2m(\sigma))^j}{j!}\\ &= 2\left(\exp(2m(\sigma)) - 2m(\sigma) - 1\right)\ .\end{aligned}$$

We used that $\int_0^t p(\tau)\,\mathrm{d}\tau = 1$ and that $m(\cdot)$ is positive and monotonically increasing. □

*Remark* 1.2. Theorem 1.1 is useful if $m(\sigma)\ll 1$ because it then implies that a Poisson process with random start time can be viewed (and simulated) as a Poisson process with rate $h=(m*p)$. This makes intuitive sense, because if $m(\sigma)$, i.e. the expected number of occurrences for $N(\sigma)$ starting at 0, is much smaller than 1 the correlation that is introduced through the random starting time is negligible.

The formulas for the mean and the variance are

$$\begin{aligned}
E\left[N\left(\sigma\right)\right] &= h\left(\sigma\right) \ , \\
E\left[M\left(\sigma\right)\right] &= h\left(\sigma\right) \ , \\
Var\left(N\left(\sigma\right)\right) &= h\left(\sigma\right) + \int_0^\sigma m^2\left(t-\tau\right)p\left(\tau\right)\mathrm{d}\tau - \left(\int_0^\sigma m\left(t-\tau\right)p\left(\tau\right)\mathrm{d}\tau\right)^2 \ , \\
Var\left(M\left(\sigma\right)\right) &= h\left(\sigma\right) \ .
\end{aligned}$$

While the mean is consistent, the difference in variance is of second order in $m$.

1.2. **Two Stage Poisson Process.** We will consider a two stage Poisson process. The first process has rate $\nu\left(t\right)$ and each occurrence of the first process is the starting point of a second-stage process with rate $\lambda\left(t\right)$. We are interested in the number $N\left(t\right)$ of occurrences from the first process and the arrival times of the second-stage processes. It is vital that we do not need the arrival times of the first process.

**Theorem 1.3.** *Let $N\left(t\right)$ be the number of occurrences of a non-homogeneous Poisson process with rate $\nu\left(t\right)$, mean value function $\eta\left(t\right)$ starting at time $0$. Let $u_1, \ldots, u_{N(t)}$ denote the arrival times of this process.*

*For each $j = 1, \ldots, N\left(t\right)$ let $Y\left(t, u_j\right)$ denote the number of occurrences of a non-homogeneous Poisson process with rate $\lambda\left(t\right)$, mean value function $m\left(t\right)$ starting at time $u_j$.*

*The process*

$$Y\left(t\right) = \sum_{j=1}^{N(t)} Y\left(t, u_j\right)$$

*is called a filtered Poisson process [3, chapter 4, eqn. (5.42)]. For $t \in [0, \sigma]$, if we neglect terms of order $O(m\left(\sigma\right)^2)$, there holds:*

i) *Conditioned on $N\left(\sigma\right)$ the process $Y\left(t\right)$ follows a Poisson process with rate*

$$\mu_N\left(t\right) = \frac{N\left(\sigma\right)}{\eta\left(\sigma\right)}\left(\nu * \lambda\right)\left(t\right) \qquad \forall t \in [0, \sigma] \ .$$

*Proof.* By Proposition 2.206 in [1, p. 147] (*actually this only guarantees the property for a homogeneous process. I am quite certain that this also holds for the non-homogeneous case but I am lacking a source.*) we have that conditioned on $N\left(\sigma\right)$ the distribution for $u_j$ (note that the $u_j$ are not ordered) is given by $p\left(u\right) = \nu\left(u\right)/\eta\left(\sigma\right)$ and all $u_j$ are i.i.d.. Then we know by Theorem 1.1 that $Y\left(t, u_j\right)$ can be approximated up to order 2 by a Poisson process with rate

$$\mu_j\left(t\right) = \int_0^t \lambda\left(t-u\right)\frac{\nu\left(u\right)}{\eta\left(\sigma\right)}\mathrm{d}u = \frac{\left(\lambda * \nu\right)}{\eta\left(\sigma\right)}\left(t\right) \qquad \forall t \in [0, \sigma] \ .$$

Because the sum of $N\left(\sigma\right)$ independent Poisson process is again a Poisson process, we know that $Y\left(t\right)$ can be approximated up to order 2 by a Poisson process with rate

$$\mu_N\left(t\right) = \frac{N\left(\sigma\right)}{\eta\left(\sigma\right)}\left(\nu * \lambda\right)\left(t\right) \ .$$

□

*Remark* 1.4. Theorem 1.3 is useful if $m(\sigma) \ll 1$ because it then implies that a two stage Poisson process can be simulated as follows. a) Draw $N(\sigma)$ at random from a Poisson distribution with mean $\eta(\sigma)$. b) Simulate a Poisson process with rate

$$\frac{N(\sigma)}{\eta(\sigma)}(\nu * \lambda)(t) \ .$$

The direct solution would be to simulate the first Poisson process fully and obtain arrival times $u_j$. For each arrival time one would simulate a Poisson process with rate $\lambda(t - u_j)$. This means on average $\eta(\sigma)$ many Poisson process simulations. The method proposed here can, under the circumstances described, generate a very good approximation with only two Poisson process simulations. This advantage comes from marginalizing out (approximately) the arrival times $u_j$.

For the Colorectal cancer model from [2] the first rate is of order $10^2$, $\sigma = 50$ and the second rate is of order $10^{-6}$. The direct approach results in approximately 5000 Poisson process simulations. The approximate method yields a theoretic speedup factor of 1000. Also $m(\sigma) \approx 10^{-4}$ and thus the approximation is extremely accurate.

## 2. Numerical Experiments

In this section we present numerical experiments for all theorems presented in this paper.

### 2.1. Poisson process with random start time.

In Figure 1 the model as described in Theorem 1.1 is simulated with a sample size of $10^7$. For $\lambda = 10^{-2}$ the approximation is already very accurate. The relative error in variance is $\lambda/6 \approx 1.7 \times 10^{-3}$ for $\lambda = 10^{-2}$.

### 2.2. Two Stage Poisson Process.

We consider a two-stage Poisson process with homogeneous rates, i.e. $\nu(t) \equiv 10^3$, $\lambda(t) \equiv 10^{-3}$. We choose $\sigma = 2$.

Because the two-stage Poisson process is a filtered Poisson process, we can use the formulas from [3, chapter 4, eqn. (5.43)–(5.45)] to analytically calculate the mean and variance of the true process $N$. For the approximation $M$ we use the law of total expectation and obtain

$$\begin{aligned} E\left[N(\sigma)\right] &= \frac{\nu\lambda\sigma^2}{2}\,, \\ Var\left(N(\sigma)\right) &= \frac{\nu\lambda\sigma^2}{2}\left(1 + \frac{2\sigma\lambda}{3}\right)\,, \\ E\left[M(\sigma)\right] &= \frac{\nu\lambda\sigma^2}{2}\,, \\ Var\left(M(\sigma)\right) &= \frac{\nu\lambda\sigma^2}{2}\left(1 + \frac{\sigma\lambda}{2}\right)\,. \end{aligned} \tag{2.1}$$

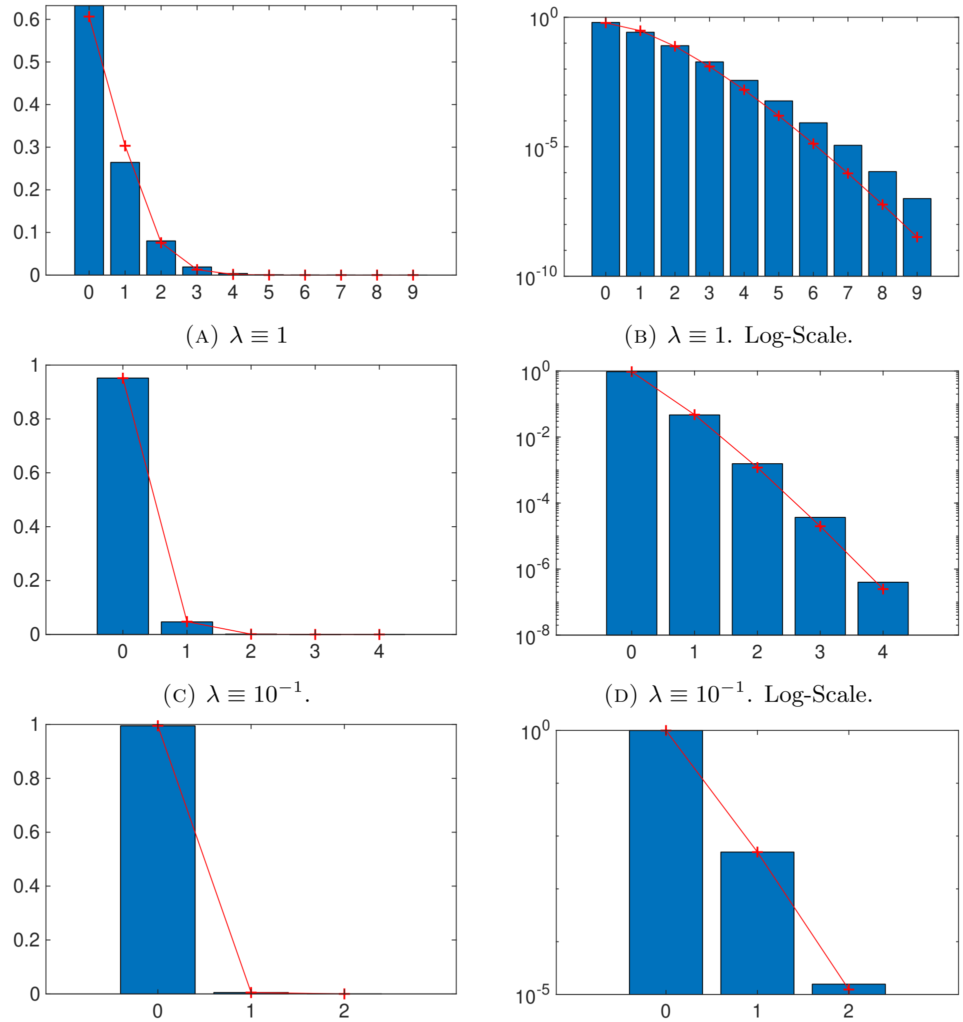
The relative error in variance is thus $\frac{\sigma\lambda}{6} \approx 0.8\%$. The results for the sampling the real and approximate process with $10^6$ samples are shown in Table 1 and Figure 2.

If we would have used a normal Poisson process as approximation, the variance would be exactly the mean and the relative error in Variance would be $\sigma\lambda$, i.e. six times higher. More significant would be that by direct simulation we would not obtain the number of occurrences in the intermediate stage, which is needed for further computation.

The runtime[1] is approximately $4.8 \times 10^{-4}$ seconds per sample for the direct method and $1.4 \times 10^{-6}$ seconds per sample for the approximate method. This is a speed up by a factor of ca. 100. It should be noted that for this experiment only the number of occurrences $N$ was computed and not their arrival times, therefore the speedup is probably even more substantial for the real simulation.

From Equation 2.1 it can be seen that by simulating the approximation with $\hat{\nu} = \frac{3}{4}\nu$ and $\hat{\lambda} = \frac{4}{3}\lambda$ mean and variance of the approximation will be correct. The numerical experiments (Table 2 and Figure 2) indicate that the effect of this change is small.

[1] This experiment was carried out on one core of a Intel Xeon X5680 at 3.33GHz that was launched in 2010.



(A) $\lambda \equiv 1$

(B) $\lambda \equiv 1$. Log-Scale.

(C) $\lambda \equiv 10^{-1}$.

(D) $\lambda \equiv 10^{-1}$. Log-Scale.

(E) $\lambda \equiv 10^{-2}$.

(F) $\lambda \equiv 10^{-2}$. Log-Scale.

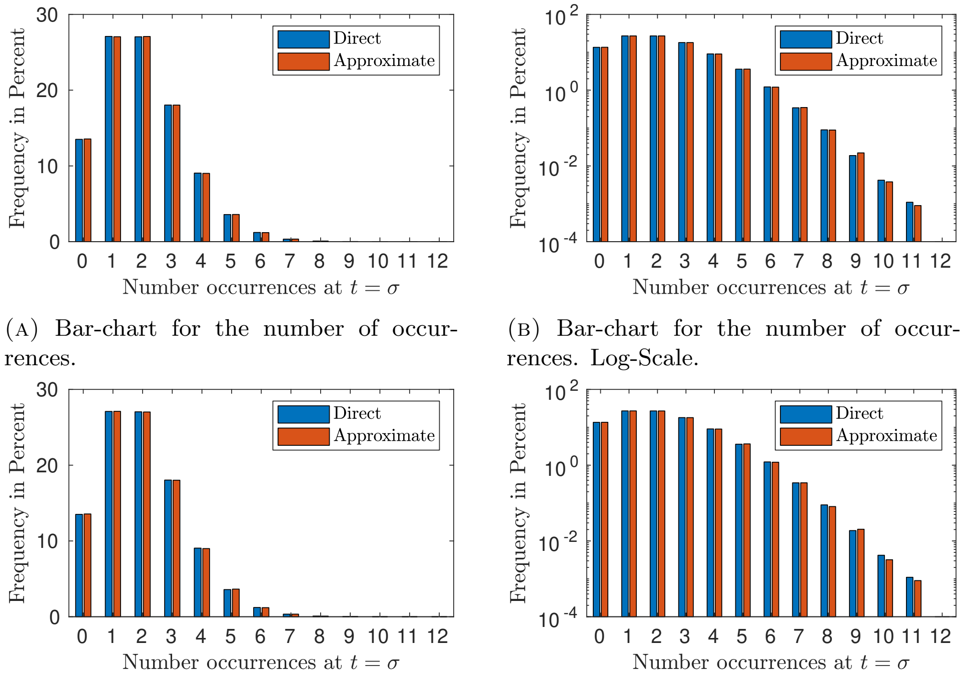
FIGURE 1. $10^7$ samples. $\sigma = 1$. $\tau \sim \text{unif}(0, 1)$. For this example there holds $E[N] = \lambda/2$ and $Var(N) = \lambda/2 + \lambda^2/12$. The blue histogram represents the values from the direct simulation with $10^7$ samples, i.e. the "ground truth". The red line is a Poisson distribution with parameter $E[M] = \lambda/2$, i.e. the approximation. Because the approximation is just a Poisson process, the distribution of $M$ is Poisson and can be obtained without sampling.

| **Statistic** | **Approximation** | **Type** | **Value** |
|---|---|---|---|
| Mean | No | Analytic | 2 |
| Mean | No | Samples | 2.000921e+00 |
| Mean | Yes | Analytic | 2 |
| Mean | Yes | Samples | 1.999512e+00 |
| Variance | No | Analytic | 2.002667e+00 |
| Variance | No | Samples | 2.002378e+00 |
| Variance | Yes | Analytic | 2.002000e+00 |
| Variance | Yes | Samples | 2.002872e+00 |

Table 1. Two stage Poisson process with $\nu \equiv 10^3$, $\lambda \equiv 10^{-3}$, $\sigma = 2$ and a total of $10^6$ samples.

| **Statistic** | **Approximation** | **Type** | **Value** |
|---|---|---|---|
| Mean | No | Analytic | 2 |
| Mean | No | Samples | 2.000921e+00 |
| Mean | Yes | Analytic | 2 |
| Mean | Yes | Samples | 1.998874e+00 |
| Variance | No | Analytic | 2.002667e+00 |
| Variance | No | Samples | 2.002378e+00 |
| Variance | Yes | Analytic | 2.002667e+00 |
| Variance | Yes | Samples | 2.001399e+00 |

Table 2. Two stage Poisson process with $\nu \equiv 10^3$, $\lambda \equiv 10^{-3}$, $\sigma = 2$ and a total of $10^6$ samples. For the approximate simulation $\hat{\nu} = \frac{3}{4}\nu$ and $\hat{\lambda} = \frac{4}{3}\lambda$ were used. Therefore, the analytic mean and variance are identical for direct and approximate simulation.



(A) Bar-chart for the number of occurrences.

(B) Bar-chart for the number of occurrences. Log-Scale.

(C) Bar-chart for the number of occurrences. $\hat{\nu} = \frac{3}{4}\nu$ and $\hat{\lambda} = \frac{4}{3}\lambda$.

(D) Bar-chart for the number of occurrences. $\hat{\nu} = \frac{3}{4}\nu$ and $\hat{\lambda} = \frac{4}{3}\lambda$. Log-Scale.

FIGURE 2. Two stage Poisson process with $\nu \equiv 10^3$, $\lambda \equiv 10^{-3}$. For the direct simulation $\nu$, $\lambda$ were used. For the approximate simulation $\nu$, $\lambda$ were used in (A) and (B) and $\hat{\nu} = \frac{3}{4}\nu$ and $\hat{\lambda} = \frac{4}{3}\lambda$ were used for (C) and (D). $\sigma = 2$ and a total of $10^6$ samples was used.

## References

[1] Vincenzo Capasso and David Bakstein. *An Introduction to Continuous- Time Stochastic Processes*. 2015.

[2] Jihyoun Jeon et al. "Evaluation of screening strategies for pre-malignant lesions using a biomathematical approach". In: *Mathematical biosciences* 213.1 (2008), pp. 56–70.

[3] Emanuel Parzen. *Stochastic processes*. Holden-Day, 1962.