Project 3 : Student number 491531 and 477789

Problem 1 : Product quality and temperatures

A company wants to find a temperature, t, that maximize the product quality, $x(t)$.

The product quality is initially assumed to be a first order stationary process with mean $E[X_t] = \mu = 50$, $Var(x(t)) = 4^2$ and the matern type correlation function $Corr(x(t), x(s)) = (1 + \phi_m|t-s|)e^{\phi_m|t-s|}$. Where $\phi_m = 0.2$ is the decay parameter for the correlation function.

The product quality has been experimentally tested at five temperature values. The experimental data is given in the table below.

$x(t)$	50.1	39.1	54.7	42.1	40.9
t	19.4	29.7	36.1	50.7	71.9

Table 1: Experimental product quality $x(t)$ at temperature t

a)

Figure 1 below is a visualization of 30 realizations of the Gaussian process described above, conditioned on the experimental data. The product quality was modelled over a regular grid of temperatures in the range $t \in [10, 80]$ degrees with a spacing of 0.5 degrees. Realizations of the process was drawn according to

$$\mathbf{X} = \mu + L\mathbf{z}$$

where μ is the mean vector and z is a vector consisting of randomly drawn variables from the standard normal distribution. L in the equation above is a the lower cholesky matrix of the covariance matrix, Σ, defined by

$$\Sigma = L^{-1}L$$

$$\Sigma_{i,j} = \sigma^2(1 + \phi_m H_{i,j})e^{\phi_m H_{i,j}},$$

where $H_{i,j}$ is a matrix of distances between all pairs of points on the time grid from index i to j.

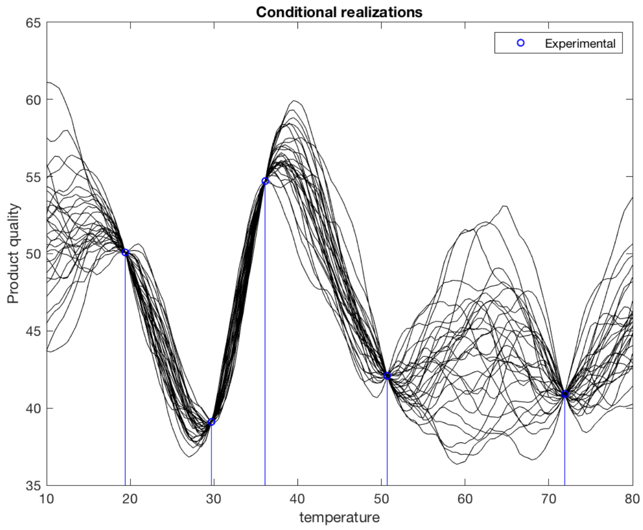

Figure 1: 30 realizations of the Gaussian process conditioned on the experimental data. The experimental data is marked as blue circles.

An interesting feature of Figure 1 is that the variance seems to be minimal around the experimental data points and increase with distance. The maximum variance between two observed data points is thus halfway between the data points. This makes perfectly sense as the uncertainty in the product quality is minimal at the actual observations and increase away from them.

The conditional mean is plotted in Figure 2, along with the 90% prediction interval. As expected, one can observe that the prediction interval roughly follows the uncertainty in the realizations. As the prediction interval encapsulates the realizations, the prediction interval is widest where the variance in product quality is greatest.

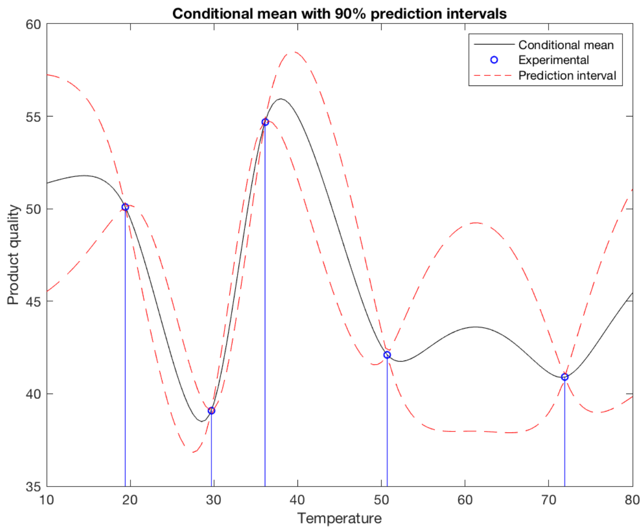

Figure 2: Conditional mean (black solid) along with a 90% prediction interval (red dashed).

b)

The probability of exceeding a certain product quality a, given a certain observations x_b, is calculated as follows

$$P(X > a|x_b) = 1 - \Phi\left(\frac{a - E[X|x_b]}{\sqrt{Var[X|x_b]}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution for a standard gaussian pdf evaluated at $(\cdot)$ and $E[X|x_b]$ and $Var[X|x_b]$ the conditional mean and variance, respectively.

The probabilities of exceeding a product quality of $a = 57$, given the observations, is plotted in Figure 3 below.

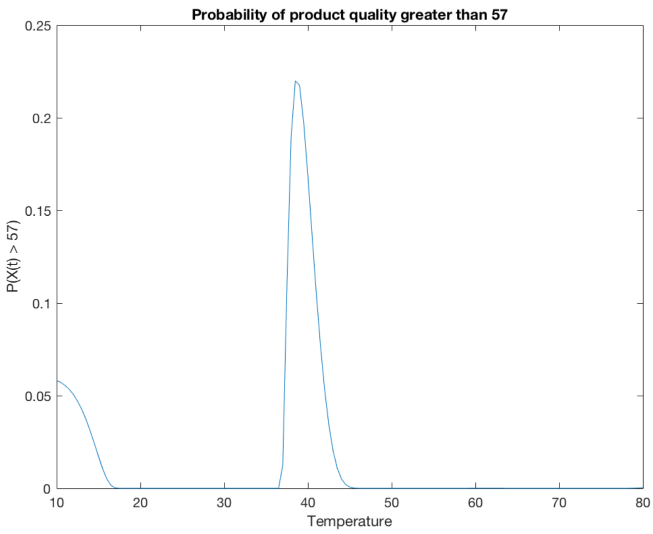

Figure 3: Probability that the product quality exceeds 57 at every temperature in the range $t \in [10, 80]$ degrees.

One can observe that the probability of exceeding a product quality of 57 is greatest in a temperature interval around 40 degrees. One can thus be more certain of getting a product quality greater than 57 if one uses temperatures around 40 degrees, based on this model.

c)

The company decided to perform another experiment, yielding the following experimental data. The additional observation is marked with bold type in the table below.

x(t)	50.1	39.1	54.7	**49.7**	42.1	40.9
t	19.4	29.7	36.1	**40.7**	50.7	71.9

Table 2: Experimental data

Similar visuals were made for these observations, i.e. conditional realizations, conditional mean and a 90% prediction interval and probabilities of exceeding a certain product quality.

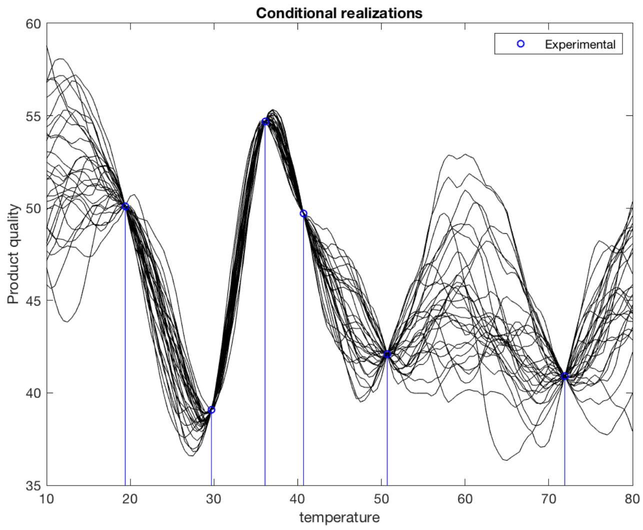

Figure 4: 30 realizations of the Gaussian process conditioned on the additional experimental data.

Note in Figure 4 that the variance in product quality decrease drastically in the interval between the additional observation and the former, as compared to Figure 1. This is expected, as the distance between the observations that we condition on is decreased in this situation, thus decreasing the uncertainty between the corresponding observations in our model.

Figure 5 below shows the conditional mean along with a 90% prediction interval with the additional observation. Note that the predication interval between the additional observation and the former extends over a smaller interval than without the additional experiment. Thus, one can be more certain to find the actual value in the interval, with the additional observation.

It is clear that the uncertainty, and the width of the prediction interval, is a function of distance between experimental data that the realizations are conditioned on. A greater (horizontal) distance, i.e temperature interval, results in a greater uncertainty, whereas a smaller temperature interval between observations yields less uncertainty in the product quality.

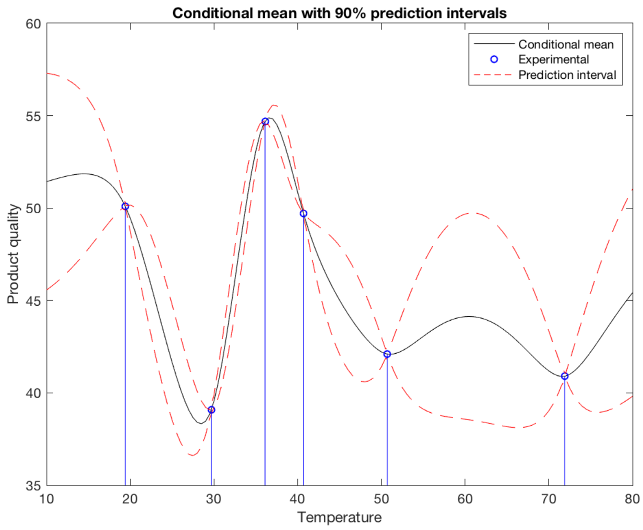

Figure 5: The conditional mean along with a 90% prediction interval for the augmented dataset

Similar to task b), the probability of a exceeding a product quality of 57 were computed for the augmented data set for each temperature in the range we are considering. The probabilities are plotted in Figure 6 below.

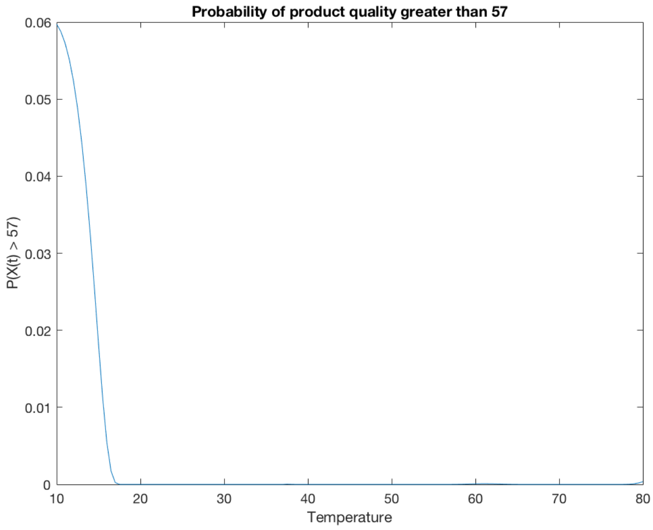

Figure 6: Probability of exceeding a product quality of 57 at each temperature in the range $t \in [10, 80]$ degrees.

One observes that the peak probability of exceeding a product quality of 57 is shifted on the augmented dataset, relative to the original. The original dataset (see Figure 3) experienced a peak probability roughly around 40 degrees, whereas the augmented dataset in Figure 6 above yields a peak probability at 10 degrees. This is a result of the additional measurement, which was conducted in close proximity of the peak probability for the former set, yielding a product quality that was quite lower than 57. The model can then be more certain that temperatures in that area yields product qualities lower than 57.

Say that the company can afford yet another experiment, at what temperature should they conduct it? The company has a goal of of producing products with a quality of at least 57, so with this in mind, the experiment should be conducted around a temperature in which in the probability of exceeding a product quality of 57 is greatest. Based on our modelling, this falls around 10 degrees.

However, if we want to decrease the overall uncertainty in our model, we would recommend to conduct an experiment in which the intermediate temperature intervals of all experiments are minimized. Minimizing these intervals will result in an overall smaller uncertainty, as the prediction intervals are a function of horizontal distance between experimental temperatures. This will clearly improve our model, and might help the company to maximize profits in the long run.

Problem 2 : Brownian motion for stock price

We are considering a randomized Stock prize, which is following a Brownian motion. We assume that the process increments $x(t_i)-x(t_{i-1})$, are Gaussian distributed with 0 mean, and variance 0.75^2. We let t_i, where $i \in [0, 1, ...]$ detone days after Jan 1st.

a)

In order to calculate the probability of the prize being greater than \$50 after 120 days, we can consider the process up until that point a Gaussian distribution, that yields from the sum of 120 independent Gaussian distributions. Denoting the event at May 1st as a stochastic variable B_1, consisting of 120 independent increments, $Z_i, i \in [1, ..., 120]$, which are Gaussian distributed, we can find the mean $E[B_1]$ and variance $Var[B_1]$

$$E[B_1] = E[\sum_{i=1}^{120} Z_i]$$

$$E[B_1] = \sum_{i=1}^{120} E[Z_i]$$

As stated, the increments are all Gaussian distributed with mean 0, thus $E[Z_i] = 0$ for all i. Which leads to:

$$E[B_1] = 0$$

Basic computations for variance is then applied to find the total variance:

$$Var[B_1] \ = Var[\sum_{i=1}^{120} Z_i]$$

$$Var[B_1] = \sum_{i=1}^{120} Var[Z_i]$$

Again, since the increments are all Gaussian distributed with variance 0.75^2, we get a total variance of:

$$Var[B_1] = 120 \cdot 0.75^2$$

Since we have the initial value of 40\$ at Jan 1st, B can we expressed as a Gaussian variable N$\sim$(40,$120 \cdot 0.75^2$), where the standard deviation is $\sigma = \sqrt{120} \cdot 0.75$. By using the cumulative distribution function for normal distributions, we can easily calculate the cumulative probability P($B_1 \leq 50$). Then, we simply take $1 - P(B_1 \leq 50)$ to find the wanted probability P($B_1 >$ 50). These calculations resulted in a probability of 0.11.

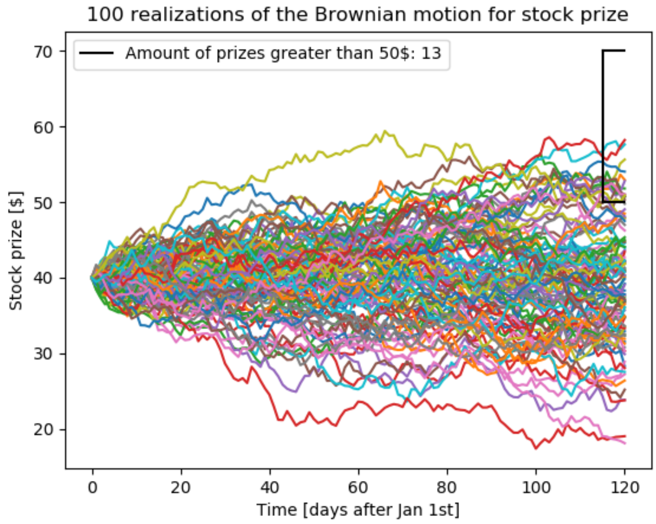

Figure 7: 100 realizations of the Stock prize from Jan 1st to May 1st.

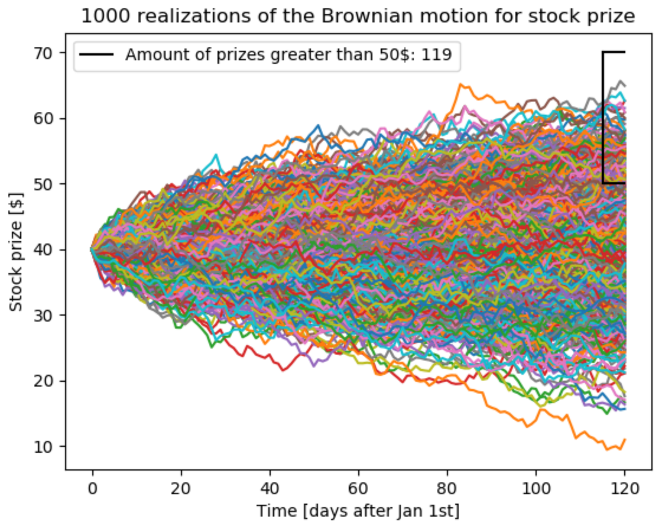

Figure 8: 1000 realizations of the Stock prize from Jan 1st to May 1st.

We then simulated the process 100 times, as well as 1000 times for even more accuracy. These processes were plotted as a function of time, as shown in figure 7 and figure 8, respectively. We observe that the approximated probabilities by simulation corresponds really well to the analytically calculated probabilities.

b)

Since a Brownian motion is a Markovian process, meaning that $P(X_i \mid X_{i-1}, ..., X_0) = P(X_i \mid X_{i-1})$, the increments only depend on the current state. The path of the process up until the current state is irrelevant. Thus, in this problem, we only need to consider the process from March 2nd to May 1st. This yields a similar problem as in the previous task, however now we have an event at May 1st, which can be considered Gaussian distributed, consisting of 60 independent Gaussian distributions. Lets now denote the event at May 1st as B_2, and we can use a similar argument and calculations as previously in order to determine $E[B_2]$ and $\text{Var}[B_2]$.

$$E[B_2] = E[\sum_{i=1}^{60} Z_i]$$

$$E[B_2] = \sum_{i=1}^{60} E[Z_i]$$

Just as previously, the increments are all Gaussian distributed with mean 0, thus $E[Z_i] = 0$ for all i. Which leads to:

$$E[B_2] = 0$$

And using similar calculations for the variance, we obtain:

$$Var[B_2] \ = Var[\sum_{i=1}^{60} Z_i]$$

$$Var[B_2] = \sum_{i=1}^{60} Var[Z_i]$$

Using the same argument that $Var[Z_i] = 0.75^2$, leads to:

$$Var[B_2] = 60 \cdot 0.75^2$$

In this case, the initial condition is \$45 at March 2nd, which leads to B_2: N$\sim$ $(45, 60 \cdot 0.75^2)$, where the standard deviation is $\sigma = \sqrt{60} \cdot 0.75$. $P(B_2 > 50)$ can be calculated identically as in the previous task. $P(B_2 > 50) = 1 - P(B_2 \leq 50)$, where $P(B_2 \leq 50)$ is calculated from the cumulative distribution function. This results in an analytical probability, $P(B_2 > 50)$, of 0.19.

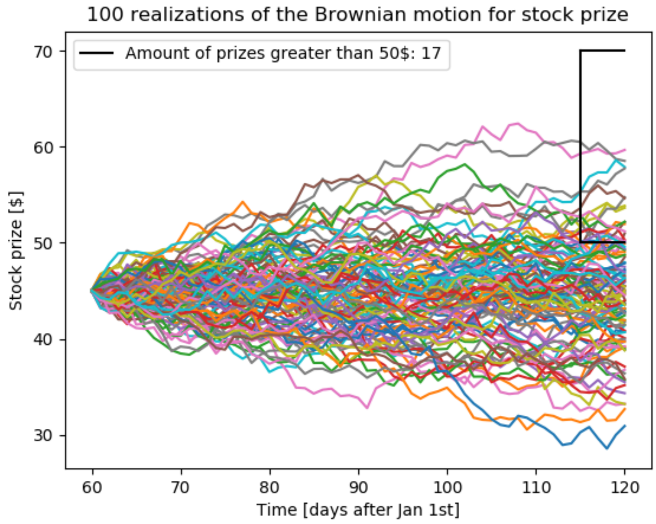

Figure 9: 100 realizations of the Stock prize from March 2nd to May 1st.

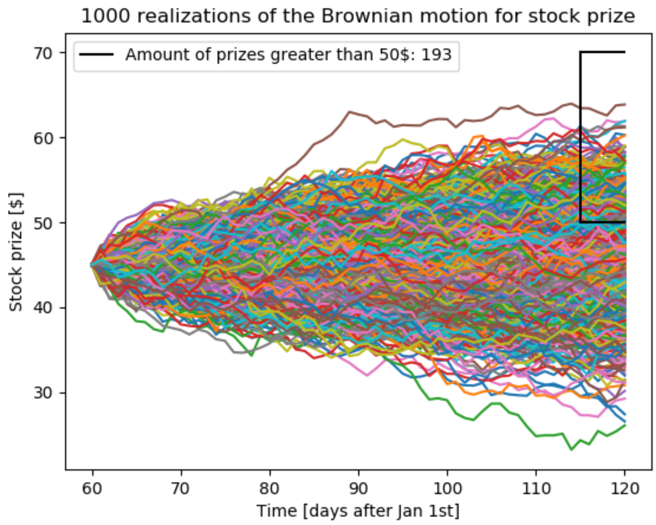

Figure 10: 1000 realizations of the Stock prize from March 2nd to May 1st.

To verify these calculations, we simulated the process from March 2nd to May 1st, both with 100 realizations and 1000 realizations. We observe from figure 9 and 10, that the simulations gives similar a similar probability as we calculated analytically. As expected, the more realizations we did, the closer the results were to the theoretical results.

c)

Now we are interested in studying when the process hits a certain prize for the first time, which we call hitting times. We will define T_a as the first time that the process hits a fixed value a. It is given that the stock prize is $40 at Jan 1st, and we are curious about when the prize has increased with 10%, meaning when it hits $44. For hitting times, we have the given cumulative probability

$$P(T_a \leq t) = 2\left(1 - \Phi(\frac{a}{\sqrt{t\sigma^2}})\right)$$

Which was used to find a cumulative probability distribution as a function of t. The cumulative distribution was also approximated, by simulating the process and observing when the process hit $44. A decision of simulating 1000 realizations was made, because of the increased accuracy compared to 100 realizations. These results were plotted and compared in figure 11. We observe that the results are corresponding satisfactorily.

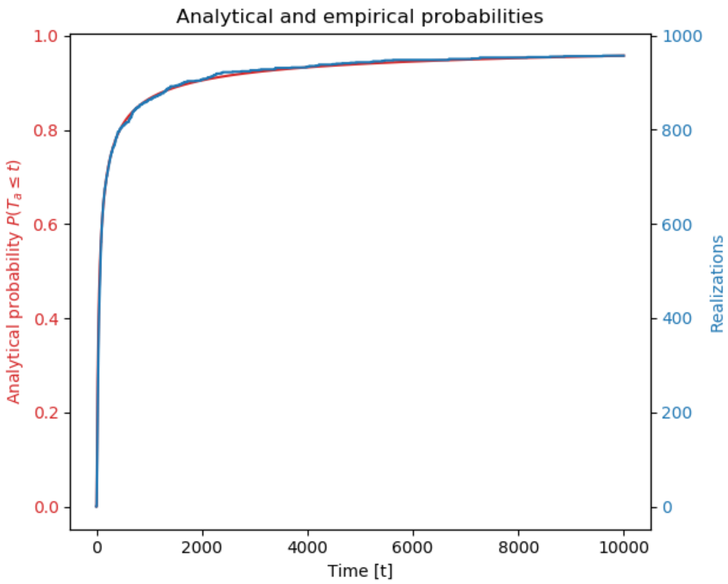

Figure 11: 1000 realizations of the cumulative hitting times.

As well as the cumulative distribution, we also wanted to study the probability density function. This can be derived by differentiating the cumulative distribution function with regards to t.

$$P(T_a = t) = \frac{d}{dt}\left(P(T_a \leq t)\right) = \frac{d}{dt}\left(2\left(1 - \Phi(\frac{a}{\sqrt{t\sigma^2}})\right)\right)$$

where $\Phi(\frac{a}{\sqrt{t\sigma^2}}) = \frac{1}{2}\left(1 + erf(\frac{a}{\sqrt{t\sigma^2}})\right)$

By using that

$$\frac{d}{dt}\left(erf(t)\right) = \frac{2}{\sqrt{\pi}} \cdot e^{-t^2}$$

and applying the chain rule, we obtain

$$P(T_a = t) = \frac{a}{\sqrt{2t^3\pi\sigma^2}} \cdot e^{-\frac{a^2}{2t\sigma^2}}$$

where σ is the standard deviation.

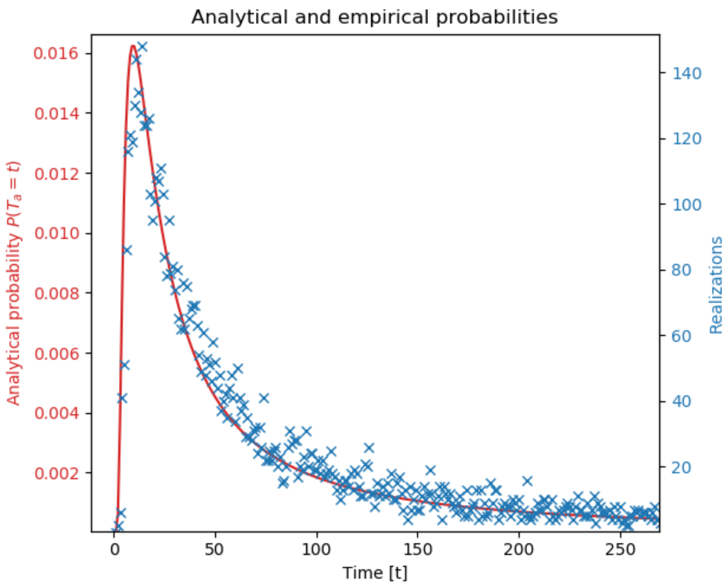

Figure 12: 10000 realizations of the density.

This probability was also approximated by simulating the process, and observing the day when the prize hit \$44. The analytical probability $P(T_a = t)$ was then plotted and compared to the simulated hitting times, and presented in figure 12. A choice of drawing 10000 realizations was made, in order to easily see the correlation in analytical and empirical results. By summing up the probability density over all t $\in [1, ..., 10000]$, we obtain the same probability as $P(T_a \leq 10000)$, which also verifies the probability density function, as it is the derivative of the cumulative distribution function. The density in figure 12 is just illustrated up to $t \approx 300$, for better representation. This should still be pretty representative, as $\approx 75\%$ of the hitting times should be within this time interval.