8 Simulation of random variables and Monte Carlo methods: motivations

The main topic of these notes are sampling (or simulation) and Monte Carlo methods and their broad extensions and applications. What we mean by simulation and Monte Carlo methods are roughly either:

(1) Sampling from a target distribution π on some measurable state space $(\mathsf{X}, \mathcal{X})$ (most often $\mathbb{R}^d$); or equivalently from a random variable distributed according to π.

(2) Regarding Monte Carlo, we aim at estimating either some integrals with respect to π, $\int f \mathrm{d}\pi$, for some function $f : \mathsf{X} \to \mathbb{R}$; or equivalently we aim at estimating expectations of $f(X)$, $\mathbb{E}[f(X)]$.

In addition, we focus here on the case where π admits a density with respect to a reference σ-finite measure λ (most often the Lebesgue measure on $\mathbb{R}^d$ or the counting measure if X is discrete), still denoted by π in general: for any measurable and bounded function $f : \mathsf{X} \to \mathbb{R}$,

$$\int_{\mathsf{X}} f(x)\mathrm{d}\pi(x) = \int_{\mathsf{X}} f(x)\pi(x)\mathrm{d}\lambda(x) \,.$$

We start these notes with some general introduction to Monte Carlo methods and provide first generic motivations. Formal definitions of sampling and Monte Carlo methods are postponed to the next Chapter 9.

8.1 Buffon's needle problem

Buffon's needle problem was raised in the 18th century by Georges-Louis Leclerc, Comte de Buffon:[(1)]

[(1)] Histoire de l'Acad. Roy. des. Sciences (1733), 43–45; Histoire naturelle, générale et particulière Supplément 4 (1777), p. 46.

> Suppose we have a floor made of parallel vertical strips of wood, each the same width, and we drop a needle onto the floor. What is the probability that the needle will lie across a line between two strips?

As we will see, the solution for the sought probability p, in the case $l < t$, where l is the needle length and t is the width of the strips, is $p = 2l/(\pi t)$.

This result can be used to design a *Monte Carlo method* for approximating the number π, although that was not the original motivation for de Buffon's question.[(2)]

We now present the stochastic model associated with this experiment. We suppose the wood is represented by $[0,1]^2$. We consider the couple of independent random variables (X,Θ). The random variables (X,Y) represents the xy-coordinate of the center of the needle and follows the uniform distribution over $[0,1]^2$ while Θ represents the acute angle between the needle and one of the vertical line and follows the uniform distribution over $[0,\pi/2]$.

Then, if we denote by $D = \min(X - \lfloor X/t \rfloor t, \lceil X/t \rceil t - X)$ the distance from the center of the needle to the closest parallel line, it is easy to show that it follows the uniform distribution over $[0,t/2]$.

Now there are two cases.

Case 1: Short needle ($l \leq t$)

The needle crosses a line if $D \leq (l/2)\sin\Theta$[(3)].

Integrating the joint probability density function gives the probability that the needle will cross a line:

$$p = \mathbb{P}(D \leq (l/2)\sin\Theta) = \int_0^{\frac{\pi}{2}} \int_0^{\frac{l}{2}\sin\theta} \frac{4}{t\pi}\, dx\, d\theta = \frac{2l}{t\pi}\,.$$

Case 2: Long needle ($l > t$)

The needle crosses a line if $D \leq (l/2)\sin\Theta$ again but we need to take into account that $D \leq t/2$. Therefore,

$$\begin{aligned} p = \mathbb{P}(D \leq (l/2)\sin\Theta) &= \left(\int_0^{\arcsin(t/l)} \int_0^{\frac{l}{2}\sin\theta} \frac{4}{t\pi} dx\, d\theta\right) + \left(\int_{\arcsin(t/l)}^{\frac{\pi}{2}} \frac{2}{\pi} d\theta\right) \\ &= \frac{2l}{t\pi} - \frac{2}{t\pi}\left(\sqrt{l^2 - t^2} + t\arcsin(t/l)\right) + 1\,. \end{aligned}$$

This suggests that we could conduct this experiment to estimate π. However, we aim to

(2) Behrends, E. (2014). Buffon: Hat er Stöckchen geworfen oder hat er nicht?.

(3) We can also consider Θ uniform on $[0,\pi]$ and model the angle from the side of the needle the closest to a vertical, with that vertical. In that case we obtain two sets of conditions: $D \leq (l/2)\sin\Theta$ and $\Theta \leq \pi/2$ or $D \leq (l/2)\cos\Theta$ and $\Theta > \pi/2$ but we verify easily that the two probabilities are equal

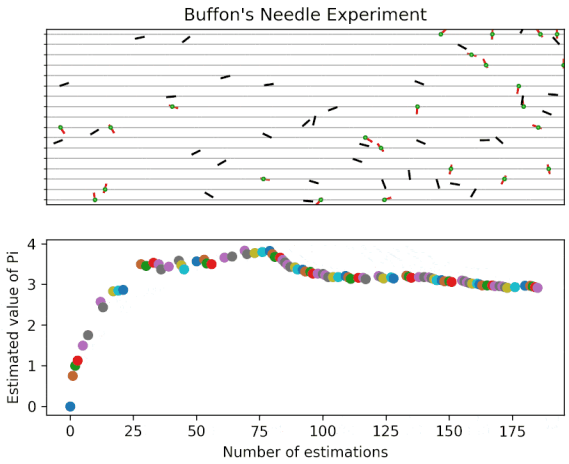


Figure 8.1: Simulation of Bouffon's Needle Experiment

replace this physical experiments by numerical simulations on a computer, i.e., sampling from D and Θ! This illustrated in Figure 8.1

8.2 Traffic modeling

We consider a first application of Monte Carlo methods for traffic modeling. We consider here a very simple model, referred to as the Nagel-Schreckenberg traffic [NS92], of N vehicles. This model belongs to the class of cellular automaton; . These N vehicles are described as N couples $\{(x_i, v_i)\}_{i=1}^N$, where $x_i \in \mathbb{Z}_n$ is the position of the vehicle and $v_i \in \mathbb{N}$ is its velocity. Here $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z} \simeq \{0, \dots, n-1\}$ is the the ring of integers modulo n; we refer to Section 10.A for details on this notation and the related object, in particular congruential relation. However, we give here the *canonical* distance on $\mathbb{Z}_n$: for $x, y \; in \mathbb{Z}_n$,

$$|x-y|_n = \min(|x-y|, |n-(x-y)|)\,. \tag{8.1}$$

Then, we define a stochastic model which describes the time evolution $(\{(x_i(t), v_i(t))\}_{i=1}^N)_{t\in\mathbb{N}}$ of these N vehicles as follows. We first suppose that $\{(x_i(0))\}_{i=1}^N$ are ordered, i.e., $x_1(0) < \dots < x_N(0)$, and $v_i(0) = 0$ for any $i \in \{1, \dots, N\}$. Given a maximal velocity $v_{\mathrm{m}} \in \mathbb{N}$ and a deceleration probability $p \in [0,1]$, at step $t \in \mathbb{N}$, given the current states of the system, $\{(x_i(t), v_i(t))\}_{i=1}^N$, we define $\{(x_i(t+1), v_i(t+1))\}_{i=1}^N$ as follows: Sequentially on $i \in \{1, \dots, N\}$,

1. We define the potential maximum velocity $\tilde{v}_i(t+1) = \min(v_i(t)+1, v_{\mathrm{m}}, D_i(t)-1)$, where $D_i(t) = |x_{i+1}(t) - x_i(t)|_n$ is the distance between the i-th vehicles.

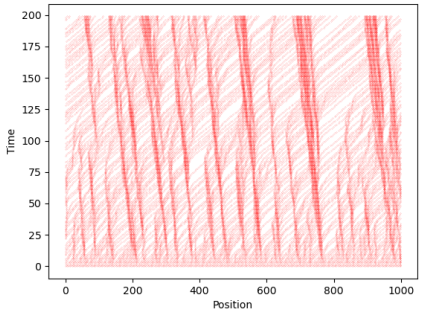
2. Then, with probability p, independently from the past, we set $v_i(t+1) = \min(0, v_i(t)-1)$ and $v_i(t+1) = \tilde{v}_i(t+1)$ otherwise.

3. Finally, set $x_i(t+1) = x_i(t) + v_i(t)$.

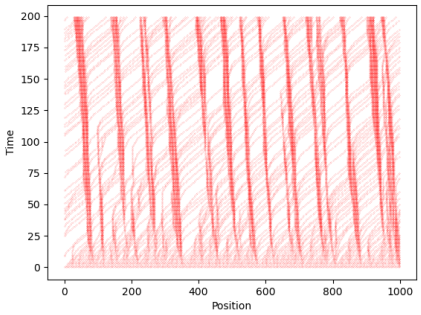
The last step is to update $(x_N(t+1), v_N(t+1))$, which is done similarly but the distance $D_N(t) = |x_1(t+1) - x_N(t)|_n$ is computed using the updated position of the first vehicle.

We can comment each step of this simulation. First, in 1, the driver set increases its velocity to 1 if it is below v_{m} and the distance (minus one) to the next driver to avoid collision, otherwise it sets its velocity to the minimum of these two quantities. In a second step 2, randomness appears which is supposed to model any events which may cause traffic disruption. Here, with probability p, the velocity is reduced by 1 if non-zero. Finally, in 3, based on the velocity $v_i(t+1)$, the i-vehicle moves to $x_i(t) + v_i(t+1)$.

In Figure 8.2 on the left hand side, one simulation are displayed starting with 100 cars represented as black dots on the top row. Traffic jams can be noticed, illustrated by the presence of diagonal bands. The traffic jams move backward as the traffic moves forward.



(a) Simulation of the simple Nagel–Schreckenberg traffic model. The vehicles move from left to right, as time increases from top to bottom.



(b) Simulation of the VDR traffic model

Figure 8.2: Simulation of the two Nagel–Schreckenberg models

There exist several extensions of these two models: including multiple line roads, car crashes events, traffic lights etc. See [HK12] and the references therein. Based on this probabilistic modeling, different questions can be addressed: can we find an "optimal" speed limitation? Does it equilibrium/stationary distribution of the system? Does it exist a limiting distribution of the average system as $N \to +\infty$? All these questions can be solved empirically by simulating models.

Remark 8.1 (Phase transition)**.** In particular, one important phenomenon can highlighted and studied using Monte Carlo methods. This phenomenon is called phase transition and it appears in many fields. Roughly, this phenomenon most often concerns system subject to external conditions or constraints; for example, for physical systems, theses conditions may be the initial state of the system, temperature, magnetic fields, force fields... Phase

transition reflects any abrupt changes in characteristics, also called global observables; in physics again, these characteristics may be energy, magnetism or density... From the author' knowledge, there is no real formal definition of this phenomenon which is not problem specific and it requires technicalities that we do not want to bother the reader for the moment. Therefore, we only provide more of a intuitive definition and illustration of this phenomenon.

We start with an illustration of the phase transition phenomenon on a slightly more complex traffic model than the Nagel–Schreckenberg model. For the Nagel–Schreckenberg traffic model, there is no phase transition. Somehow, the model is too simple. such a phenomenon does not appear. We present here a simple generalization which incorporates a slow-to-start behaviour. by using a velocity-dependent randomization (VDR). We call this model the VDR model [Bar+98]. This model consists in changing the deceleration probability depending on the velocity of the agent. More precisely, we define the time evolution $(\{(\bar{x}_i(t),\bar{v}_i(t))\}_{i=1}^N)_{t\in\mathbb{N}}$ of these N vehicles as follows. We first suppose that $\{(\bar{x}_i(0))\}_{i=1}^N$ are ordered, i.e., $\bar{x}_1(0) < \ldots < \bar{x}_N(0)$, and $\bar{v}_i(0) = 0$ for any $i \in \{1,\ldots,N\}$. Given a maximal velocity $v_m \in \mathbb{N}$ and deceleration probabilities $p, p_0 \in [0,1]$, at step $t \in \mathbb{N}$, given the current states of the system, $\{(\bar{x}_i(t),\bar{v}_i(t))\}_{i=1}^N$, we define $\{(\bar{x}_i(t+1),\bar{v}_i(t+1))\}_{i=1}^N$ as follows: Sequentially on $i \in \{1,\ldots,N\}$,

1. We define the potential maximum velocity $\tilde{v}_i'(t+1) = \min(\bar{v}_i(t)+1, v_m, D_i(t)-1)$, where $D_i(t) = |\bar{x}_{i+1}(t) - \bar{x}_i(t)|_n$ is the distance between the i-th vehicles.

2. Then, with probability $p_i(t+1) = p_0 \mathbb{1}\{\bar{v}_i(t) = 0\} + p\,\mathbb{1}\{\bar{v}_i(t) > 0\}$, independently from the past, we set $\bar{v}_i(t+1) = \min(0, \bar{v}_i(t) - 1)$ and $\bar{v}_i(t+1) = \tilde{v}_i(t+1)$ otherwise.

3. Finally, set $\bar{x}_i(t+1) = \bar{x}_i(t) + \bar{v}_i(t)$.

The last step is to update $(\bar{x}_N(t+1), \bar{v}_N(t+1))$, which is done similarly but the distance $D_N(t) = |\bar{x}_1(t+1) - \bar{x}_N(t)|_n$ is computed using the updated position of the first vehicle.

Note that the only step which differs from the Nagel–Schreckenberg model is the step 2, where as already mentioned, the deceleration probability depends now if the vehicle is stopped or not. Note that to have a phase transition, it has to be held $p < p_0$.

Simulation of the two introduced Nagel–Schreckenberg models are displayed in Figure 8.2.

We define the averaged flow of the VDR and Nagel–Schreckenberg traffic models as $J = (nN(t_J - t_0))^{-1} \sum_{i=1}^N \sum_{\ell=t_0}^{t_J} v_i(t_J)$, where t_0, t_J are prescribed iterations large enough, and $\rho = N/n$ is the density. In Figure 8.3, the flow of the VDR with respect to ρ are displayed for two initial states: (1) the homogeneous (hom) case where vehicles are set to be equidistant to each other; (2) the jammed (jam) case where the initial positions of the vehicles are set to $x_i(0) = i$ for any $i \in \{1,\ldots,N\}$. We can observe that the averaged flow is not continuous with respect to the density at ρ_2 for the homogeneous initialization. In

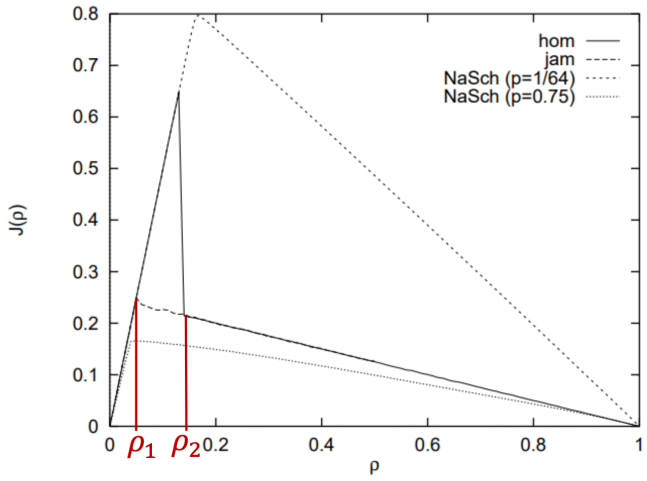


Figure 8.3: Flow of the VDR and Nagel–Schreckenberg traffic models with respect to density ρ for two initial conditions, the jammed condition and the space homogeneous condition. $v_{\mathrm{m}} = 5$, $n = 1000$, $p = 1/64$ and $p_0 = 0.75$, $t_0 = 10000$ and t_J corresponds to 100000 sweeps through the lattice. Credit [Bar+98]

addition, for $\rho \in [\rho_1, \rho_2]$, the average flow is different depending on the initial states, while outside $[\rho_1, \rho_2]$, the average flow is the same for both initial states.

8.3 Statistical mechanics

We consider now the problem of sampling from a statistical mechanics perspective. One of the simplest example is the Ising model which is used to model magnets and lattice gases. We give here an informal presentation of this model, the associated phase transition problem, and do not aim at being exhaustive.

Magnets are described using spin systems on a graph $\mathsf{G} = (\mathsf{V}_n^d, \mathsf{E}_n^d)$. We simply choose here for the set of vertices, the lattice with periodic boundary conditions $\mathsf{V}_n^d = \mathbb{Z}_n^d \simeq (\mathbb{Z}/n\mathbb{Z})^d$, for a dimension $d \in \mathbb{N}^*$ and a number of spins $n \in \mathbb{N}^*$. An element $\mathbf{i} \in \mathsf{V}$ model the position of a particle on the lattice V_n^d. In addition, we choose that $(\mathbf{i},\mathbf{j}) \in \mathsf{E}$ is and only if $|\mathbf{i}-\mathbf{j}|_{n,d} = 1$, where

$$|\mathbf{i}-\mathbf{j}|_{n,d} = \sum_{\ell=1}^{d} |i_\ell - j_\ell|_n \ ,$$

where $|\cdot|_n$ is defined in (8.1), and i_ℓ, j_ℓ denoted the ℓ-th component of $\mathbf{i},\mathbf{j}$. In this context, a choice of a label function $\omega : \mathsf{V} \to \{\pm 1\}$, $(\omega_i)_{i\in\mathsf{V}} \in \Omega_n^d = \{\pm 1\}^{\mathsf{V}_n^d}$ is said to be the configuration or the set of directions of the spins, if $\omega_i = 1$, the i spin is said to be up otherwise it is said down.

To model gases or magnets, statistical mechanics supposes that configurations of the

system are random and associated with a distribution with density with respect to the counting measure on Ω_n^d: for $\omega \in \Omega_n^d$,

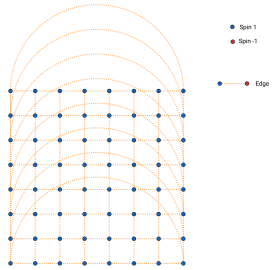
$$\pi_{n,\beta}^d(\omega) = \exp(-\mathscr{H}_{n,\beta,h}(\omega))/\mathcal{Z}_N \,,$$

where

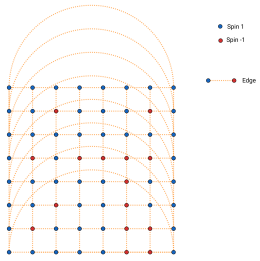
(1) $\mathscr{H}_{n,\beta,h}(\omega) = -\beta \sum_{(\mathbf{i},\mathbf{j})\in \mathsf{E}} \omega_{\mathbf{i}}\omega_{\mathbf{j}} - h\sum_{\mathbf{i}\in \mathsf{V}_n^d} \omega_{\mathbf{i}}$ is called the energy of ω associated to the inverse temperature β and magnetic field h;

(2) the quantity $\mathcal{Z}_n^d = \sum_{\omega'\in\Omega_n^d} \exp(-\mathscr{H}_{n,\beta,h}(\omega'))$ is called the partition function of the system.

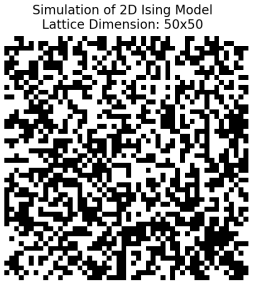
Any random variable σ with distribution $\pi_{n,\beta,h}^d$ is called a spin, i.e., $\mathbb{P}(\sigma = \omega) = \pi_{n,\beta,h}^d(\omega)$, for any $\omega \in \Omega_n^d$. At high temperature, physics modeling assumes that up and down spins have equal probability, whereas as at low temperature, spins are likely either all down or all up. In the latter case, the material is said to be magnetized. Similar considerations can also be made for lattice gases, but not formally introduced here; see [Geo11; Kra06]. In this setting, we can also highlight a phase transition phenomenon. Illustrations of the spin modelling are displayed in Figure 8.4.



(a) Example of a spin configuration for a size box $n = 8$ at low temperature and graph structure



(b) Example of a spin configuration for a size box $n = 8$ at high temperature and graph structure



(c) Example of a spin configuration for a size box $n = 50$

Figure 8.4: Examples of spin configurations

Define the magnetization of the system at temperature β as

$$m_n^d(\beta,h) = \int_{\Omega_n^d} \left\{ \sum_{\mathbf{i}\in \mathsf{V}_n^d} \omega_{\mathbf{i}} \right\} \mathrm{d}\pi_{n,\beta,h}^d(\omega)\,. \tag{8.2}$$

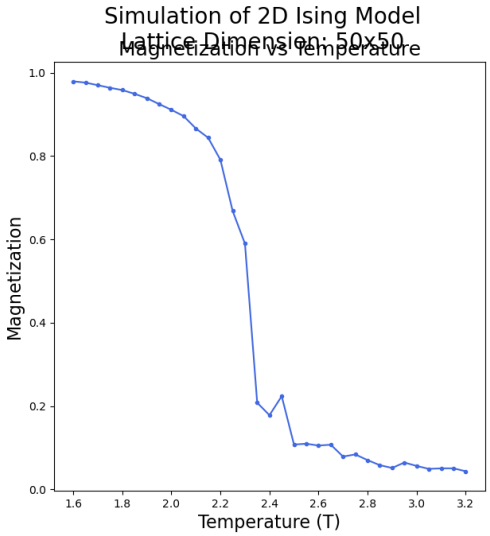


Figure 8.5: Average magnetization

It can be shown that $\lim_{h\downarrow 0}\lim_{n\to+\infty} m_n^d(\beta,h) = m_\star^d(\beta)$ exists and is called the average magnetization density. If there exists $\beta_c^d \in \mathbb{R}_+^*$ such that for any $0 \le \beta < \beta_c^d$, $m_\star^d(\beta) = 0$ and for any $\beta_c^d < \beta$, $m_\star^d(\beta) > 0$, we say that a phase transition of first order occurs. In this context, the quantity β_c^d is the called the critical inverse temperature and estimating it is of prime interest in statistical mechanics.

There is no phase transition for $d = 1$ but there is for $d \ge 2$. In particular, we can show that $\beta_c^2 = 2^{-1}\operatorname{arcsinh}(1) \approx 0.441$. For $\beta \le \beta_c^d$, we say that the system is paramagnetic and ferromagnetic otherwise.

As emphasized previously, estimating β_c^d is important in physics. To this end, Monte Carlo methods can be used. A naive but illustrative way is to numerically estimate $m_n^d(\beta_i)$ by Monte Carlo for n large and a thin grid, i.e., for an increasing sequence of inverse temperatures $\{\beta_i\}_{i=1}^{\ell}$ such that $\max_i \left|\beta_{i+1} - \beta_i\right|$ is small, since $m_n^d(\beta_i)$ is an integral by (8.2), and observe around which region a discontinuity appears. Result of such approach is displayed in Figure 8.5. Note that we use here a Metropolis-Hasting algorithm that we will see in this course.

A careful reader may wonder why using Monte Carlo methods since $m_n^d(\beta_i)$ has an explicit expression. The main reason is that for n and d large explicitly computing the partition function $\mathcal{Z}_n^d$ is computationally impossible since it implies to sum over 2^{dn} terms.

8.4 Mathematical finance

We consider here the discrete formulation of an American option. The price of an asset is modelled here as a random process $(V_n)_{n\in\mathbb{N}}$ defined by induction: given V_0,

$$V_{n+1} = V_n \exp(X_{n+1}),$$

where $(X_n)_{n\in\mathbb{N}^*}$ are supposed to be i.i.d. $\mathrm{N}(m,\sigma^2)$.

In finance, there exists a product called an American option which allows the buyer of this option to sell the given asset at price K for any stage $n \in \{0,\dots,N\}$, for a given terminal stage $N \in \mathbb{N}^*$. More precisely, for any $n \in \{0,\dots,N\}$, the buyer can sell its asset at price $\max(K,V_n)$ for some $K>0$. We set the price of the option also to K for simplicity here and therefore the potential benefit for the buyer is $\max(V_n-K,0)$. Then, to maximize its average profit, the buyer aims to estimate $\mathrm{argmax}_{n\in\{0,\dots,N\}}\mathbb{E}[\max(V_n-K,0)]$. Set $u_n = \mathbb{E}[\max(V_n-K,0)]$. We distinguish two cases

- If $m+\sigma^2/2 \geq 0$. Then using the tower property and the Jensen inequality, we have

$$u_{n+1} \geq \mathbb{E}[\max(V_n\mathbb{E}[\mathrm{e}^{X_{n+1}}|\mathcal{F}_n]-K,0)],$$

where $(\mathcal{F}_n)_{n\in\mathbb{N}}$ is the filtration generated by $(X_n)_{n\in\mathbb{N}^*}$ with $\mathcal{F}_0=\sigma(V_0)$. By independence, $\mathbb{E}[\mathrm{e}^{X_{n+1}}|\mathcal{F}_n] = \mathbb{E}[\mathrm{e}^{X_{n+1}}] = \mathrm{e}^{m+\sigma^2/2} \geq 1$ and we deduce that $(u_n)_{n\in\{0,\dots,N\}}$ is non-decreasing and $\mathrm{argmax}_{n\in\{0,\dots,N\}}\mathbb{E}[\max(V_n-K,0)] = N$.

- If $m+\sigma^2/2 < 0$, the problem at hand is much more complex. A naive but illustrative solution is to estimate numerically the sequence $(u_n)_{n\in\{0,\dots,N\}}$ using Monte Carlo methods since the u_n are expectations.

Remark 8.2. Of course, this is a simplified presentation of the problem associated to an American option. We only mention the more complex problem at hand. Denoting by $\mathcal{T}_{\mathcal{F}}$ the set of stopping times associated with $(\mathcal{F}_n)_{n\in\mathbb{N}}$. The problem which the buyer aim to solve is to estimate $\sup_{\tau\in\mathcal{T}_{\mathcal{F}}}\mathbb{E}[\max(V_\tau-K,0)]$. In this setting, we model that the buyer can take its decision to sell its asset at stage n given all information before.

8.5 Bayesian statistics

In Bayesian statistics that we will present in Chapter 7, the problem of sampling a distribution π or estimating some integrals with respect to it, is at the core of the inference process. Here we provide an informal discussion on this subject and refer to the corresponding chapter of these notes for a more formal introduction.

Recall that in *frequentist* statistics, based on i.i.d. observations $X_1,\dots,X_n$ with density f_θ with respect to some reference measure λ, say the Lebesgue measure, the parameter $\theta \in \Theta \subset \mathbb{R}^d$ is supposed to be deterministic and is inferred using a one point estimator, for

example the maximum likelihood estimator:

$$\hat{\theta}_n \in \operatorname*{argmax}_{\vartheta \in \Theta} \prod_{i=1}^{n} f_\vartheta(X_i)\,.$$

In Bayesian statistics, θ is not supposed to be deterministic but random. Roughly, it is suppose that the parameter θ is a realization of a fixed and chosen prior distribution with density p with respect to the Lebesgue measure. In that situation, we can define the posterior distribution as the conditional distribution of θ given the observations $X_1, \dots, X_n$. This conditional distribution has for density (with respect to the Lebesgue measure):

$$\pi_n(\vartheta) = p(\vartheta) \prod_{i=1}^{n} f_\vartheta(X_i)/Z_n\,, \quad Z_n = \int_{\mathbb{R}^d} p(\vartheta) \prod_{i=1}^{n} f_\vartheta(X_i) \mathrm{d}\theta\,.$$

Note that in most cases of interest, the normalizing constant Z_n is intractable which brings further challenges in the problem of sampling and estimating integrals with respect to π_n. Yet, Bayesian inference is primarily based on integrals with respect to π_n. For example, one estimator for θ is the Bayes estimator defined as the mean of π_n:

$$\hat{\theta}_n^{\mathrm{B}} = \int \vartheta \pi_n(\vartheta) \mathrm{d}\vartheta\,.$$