Numéro National de Thèse : 2016SACLX043

THESE DE DOCTORAT
DE L'UNIVERSITE PARIS-SACLAY

préparée à

L'ÉCOLE POLYTECHNIQUE

École Doctorale N°574
Mathématiques Hadamard (EDMH)

Spécialité de doctorat : Mathématiques appliquées

par

Mr. Gang LIU

Rare Event Simulation by Shaking Transformations and NISR Method for Dynamic Programming Problems

Thèse présentée et soutenue à Palaiseau, le 23 novembre 2016

Composition du jury :

Tony LELIEVRE	(Ecole des Ponts ParisTech)	Président du jury
Stefano DE MARCO	(Ecole Polytechnique)	Examinateur
Pierre DEL MORAL	(INRIA)	Examinateur
Gersende FORT	(CNRS, Télécom ParisTech)	Examinatrice
Emmanuel GOBET	(Ecole Polytechnique)	Directeur de thèse
Arnaud GUYADER	(Université Pierre et Marie Curie)	Rapporteur
Mike LUDKOVSKI	(University of California Santa Barbara)	Rapporteur
Eric MOULINES	(Ecole Polytechnique)	Examinateur

UNIVERSITY PARIS-SACLAY

Abstract

Ecole Polytechnique and CNRS
Centre de Mathématiques Appliquées

Doctor of Philosophy

Rare Event Simulation by Shaking Transformation and Non-intrusive Stratified Resampling Method for Dynamic Programming

by Gang LIU

This thesis covers two different subjects: rare event simulation and non-intrusive stratified regression method for dynamic programming problems.

In the first part, we design a Markovian transition called *shaking transformations* on the path space, which enables us to propose IPS and POP methods for rare event simulation, based respectively on interacting particle system and the ergodicity of Markov chain. Efficient designs of shaking transformation in finite and infinite dimensional cases are proposed. We also design an adaptive version of the POP method, which demands less information on the model to be well implemented. Theoretical analysis is given on the convergence of these methods. Besides, we demonstrate how these techniques can be applied to perform sensitivity analysis of rare event statistics on model parameters and to make approximative sampling of rare event. Many numerical examples are discussed to show the performance of our methods and how to appropriately choose method parameters.

In the second part, we aim at numerically solving certain dynamic programming problems. Different from usual settings, we don't have access to full detail of the underlying model and only a relatively small-sized set of root sample is available. To solve the problem with the limited information at hand, we propose a stratified resampling regression method. More precisely, we shall use given the root sample to reconstruct other paths and perform local regression on the stratified spaces. Non-asymptotic error estimations are given and we demonstrate the performance of our methods in several numerical examples.

During my thesis, I have also worked to create a financial software in a startup project inside CMAP Ecole Polytechnique . This work is not reported here due to confidentiality issues.

Acknowledgements

First of all, I would like to thank my PhD supervisor, Emmanuel Gobet, who has always been encouraging me and giving me valuable advice during my PhD project. I have learned a lot of things from him during our collaboration.

It is a great pleasure to have spent three years in Centre de Mathématiques Appliquées in Ecole Polytechnique. I want to express my thankfulness to all my colleagues here, for all the enjoyable collaborations and discussions on numerous topics. I have benefited a lot from our inspiring research environment. I also want to thank Vincent Chapellier, Wilfried Edmond, Nasséra Naar and Alexandra Noiret for their great help in dealing with administrative procedures.

I would like to thank all the members of my defense jury : Stefano De Marco, Pierre Del Moral, Gersende Fort, Emmanuel Gobet, Arnaud Guyader, Tony Lelievre, Mike Ludkovski and Eric Moulines, for their precious time and valuable comments, which have greatly improved the quality of this thesis.

I want to thank Ecole Polytechnique and Chaire Risques Financiers for financially supporting my PhD work and the conference expenditures. I also want to thank IMPA institute for partly supporting my visit during RIO conference.

Last but not least, my great thankfulness gos to my wife Lili Zhu, who has been my constant support during the three years we share together, and to our parents, who give us their understanding and support for pursuing our goals.

Contents

List of Abbreviations

BS	Black Scholes
BSDE	Backward Stochastic Differential Equation
CDF	Cumulative Distribution Function
CPP	Compound Poisson Process
DPE	Dynamic Programming Equation
fBM	Fractional Brownian Motion
FKPP	Fisher-Kolmogorov–Petrovsky–Piscounov
i.i.d.	independent and identically distributed
IPS	Interacting Particle System
IV	Implied Volatility
MCMC	Markov Chain Monte Carlo
MDP	Multi-step Dynamic Programming
MH	Metropolis-Hastings
NISR	Non-Intrusive Stratified Resampler
OLS	Ordinary Least Square
OU	Ornstein-Uhlenbeck
PDE	Partial Differential Equation
POP	Parallel One Path
RESTART	REpetitive Simulation Trials After Reaching Thresholds
RMC	Regression Monte Carlo
SDE	Stochastic Differential Equation

Chapter 1

Thesis Summary

This thesis contains two different subjects: rare event simulation and non-intrusive stratified resampling method for dynamic programming, each of which is covered in a separate part of this thesis. In this beginning chapter, we will briefly introduce these two problems, give a short review of literature and summarize our contributions. Complementary introductions and literature reviews will be given in respective parts for both subjects.

1.1 Rare event simulation

Rare event simulation concerns the study of extreme events, which have very small probabilities of taking place but imply serious consequences once they happen. Examples of rare events are: insurance company default (Subsection 3.1.1), communication network collapse (Subsection 3.1.2), random graph atypical configuration (Subsection 3.1.3) and black swan events in finance (Subsection 3.1.5). Other applications of rare event can also be found in Section 3.1.

1.1.1 Probabilistic formulation

The probabilistic study of rare event usually starts with the following framework:

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we consider a random variable (measurable mapping) $X : \Omega \mapsto \mathbb{S}$, where $\mathbb{S}$ is some general state space, and a measurable set $A \subsetneq \mathbb{S}$. The set A is chosen such that the probability that X lies in A is extremely small, in which case we call $\{X \in A\}$ a rare event. We are mainly interested in achieving the following goals

- to estimate the rare event probability $\mathbb{P}(X \in A)$
- to sample from the conditional distribution $X|X \in A$
- to estimate the conditional expectation on rare event $\mathbb{E}(\varphi(X)|X \in A)$, for bounded measurable functions $\varphi : \mathbb{S} \mapsto \mathbb{R}$
- to evaluate the sensitivity of these rare event statistics with respect to model parameters

In the setting of rare event, $\mathbb{P}(X \in A)$ is usually less than 10^{-4}. We always assume that $\mathbb{P}(X \in A) > 0$. Remark that this formulation is very general, in the sense that the random object X under investigation can be correlated stochastic processes, random graphs and other complicated random systems.

1.1.2 Literature review

Monte Carlo methods are mostly used to estimate probability and expectations. The simplest version is the plain Monte Carlo method, which is based on the law of large numbers and central limit theorem: if we make N independent and identically distributed (i.i.d.) copy $(X_n)_{1\leqslant n\leqslant N}$ of X and compute the empirical proportion of copies which lie in A, then it converges to $\mathbb{P}(X \in A)$ as N goes to infinity and the central limit theorem gives corresponding confidence intervals for our estimators. Unfortunately plain Monte Carlo method fails to work efficiently in the case of rare event. Since the probability for the event $\{X \in A\}$ is very small, a large number of simulations are needed to have one realization of this event in average. Therefore, the computational cost is prohibitively high to obtain a satisfying accuracy for our estimator. Mathematically, it means that the relative variance for our estimator is too high to deliver good estimation.

More precisely, if we make N independent and identically distributed (i.i.d.) copy $(X_n)_{1\leqslant n\leqslant N}$ of X and set $p = \mathbb{P}(X \in A)$ and define the empirical occupation measure by $\hat{p}_N = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_{X_n \in A}$, then by central limit theorem, we have

$$\sqrt{N}(\hat{p}_N - p) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

where $\sigma^2 = p(1-p)$. Thus, when N is large, approximately we can get a 95% confidence interval for p:

$$(\hat{p}_N - 1.96\sqrt{\frac{p(1-p)}{N}}, \hat{p}_N + 1.96\sqrt{\frac{p(1-p)}{N}})$$

.

This may look good at first glance, since the length of this interval is equal to $3.92\sqrt{\frac{p(1-p)}{N}}$, which is small with large N and small p. But if we look at the relative length (i.e. relative error) in percentage of p, it is equal to $3.92\sqrt{\frac{(1-p)}{Np}} \approx 3.92\sqrt{\frac{1}{Np}}$. If for example $p = 10^{-8}$, even if we use 10 million simulations of X, at the end we get a confidence interval of relative length more than 10, so the final conclusion may be that p is between 0 and 10^{-7}. This information is completely useless in our problem, since the possible error goes far beyond our tolerance.

One technique to overcome this problem of having too few realizations of our target event is the importance sampling, see Subsection 3.2.2. Instead of making simulations under the initial probability measure, we propose another probability measure under which the event of

interest $\{X \in A\}$ is more likely to happen, thus fewer simulations are needed to have the same amount of realizations in average. Of course by proposing a new probability measure some bias is induced in the estimator and a correction/weight term needs to be added to provide an unbiased estimator. This method is very efficient when the new probability is easy to simulate, see [120, 19, 67, 72, 51]. But in general, the new probability measure is not easy to find for complex systems and specific study needs to be conducted for different models.

Another idea to deal with rare event simulation is splitting, see Subsection 3.2.3. Instead of making straightforward estimation of $\mathbb{P}(X \in A)$, we define a series of nested subsets

$$\mathbb{S} := A_0 \supset \cdots \supset A_k \supset \cdots \supset A_n := A,$$

and make estimations of each conditional probability $\mathbb{P}(X \in A_{k+1} | X \in A_k)$, then their product gives an estimation of our rare event probability. Studies on splitting methods can be found for example in [84, 93]. The convergence of adaptive splitting method has been shown in [32].

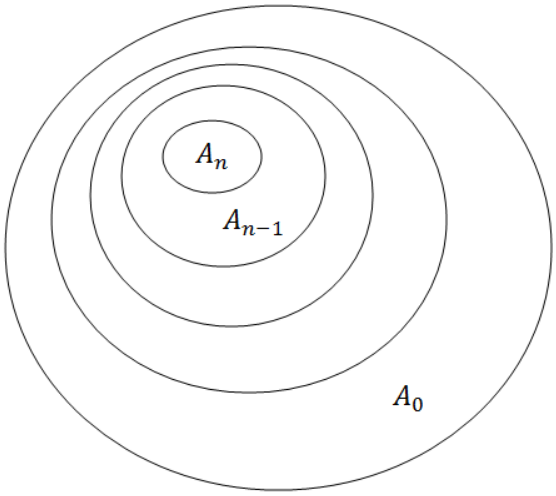


FIGURE 1.1: Nested subsets in splitting

One problem of original splitting method is that a lot of simulation time is spent in areas which are far from the rare event zone. To overcome this problem, RESTART method is proposed, see Subsection 3.2.4. The uniform splitting rule in original splitting method is modified in the RESTART method such that simulation efforts are concentrated in important areas, see [13, 125, 126, 86, 102]. This makes estimation more efficient in many cases, but due to the non-uniform splitting rules, theoretical analysis becomes more difficult.

More recently, another group of methods called IPS is proposed, see [42, 43, 31, 30, 29, 27] . It is based on the theory of interacting particle system and it follows also the spirit of splitting methods, i.e. the final estimator is given as a product of several conditional probability estimators, see Subsection 3.2.5. IPS method imitates the procedure of natural

selection and contains selection and mutation steps. The convergence of adaptive IPS method has been shown in [33].

There are many other tools we can use for rare event simulation problems, such as cross-entropy method [119, 24], large deviation theory [44, 40], etc. Some other interesting works are [23] on generalized splitting method and [92] on switching diffusions. A lot of other works can be found on the websites of the recent two International Workshops on Rare Event Simulation, RESIM 2014 and RESIM 2016.

1.1.3 Our contributions

General applicability According to our numerical experiments, when all the methods can be easily implemented, importance sampling is usually the most efficient one. But specific techniques are needed to address each problem and simulations under the importance sampling measure become time consuming when the model at hand is complicated. We aim at designing a new methodology which needs few adjustments when applied with different models.

Static point view on path space, no Markovian assumption and IPS method The splitting idea applies more generally than importance sampling, combined with different kinds of techniques. But it still relies on several assumptions. When splitting, RESTART or IPS methods are applied with a dynamic model, Markovian assumptions are usually needed. If the system under consideration is not given as a Markovian one, state augmentation techniques can help sometimes. But this makes algorithms a bit more cumbersome and markovianization is not always possible. In this thesis, we overcome this problem by adopting a static pointview to address dynamic model. Thus we don't need any Markovian assumptions. This static point view also brings other benefits. When combined with interacting particle system theory, it gives rise to a new kind of IPS method, see 5.2.2. When one applies this version of IPS method on dynamic models, discretization is no longer an issue to worry about. The error explosion with existing IPS methods when the time step goes to zero is not observed with our new version of IPS method. This static point view is implicitly implied in our presentation of shaking transformation in Section 5.2, which treats random variables and stochastic process in a uniform way.

We will illustrate the above explanations by a simple example. Suppose that we are dealing with the realization of an Ornstein-Uhlenbeck process Z_t between time $t = 0$ and $t = 1$ and we want to compute the probability that the maximal value of Z_t during this period is larger than 10. We take discrete time grid $t_i = \frac{i}{N}$ for some N. If we apply the IPS method with dynamic point view, such as in [27], we are going to make i.i.d simulations for time t_1, and select those paths which goes upwards faster than others, apply the mutation step to simulate for time t_2, then again select those paths which goes upwards faster than others.

We repeat this procedure until time t_N. With the dynamic point view, the selection and mutation step is conducted during the simulation of Brownian paths. However, if we apply the static point view, the entire path of Z_t between time 0 and 1 are treated as an indivisible point and no partial paths will be simulated. We will make simulation of entire paths, select those closer to the rare event zone and apply mutation transformation on the path space. The following graph illustrates both strong and slight shaking transformations: the blue one is the initial path and the green one is obtained by applying a slight shaking transformation on the entire path while the red one is obtained by applying a strong shaking transformation on the entire path.

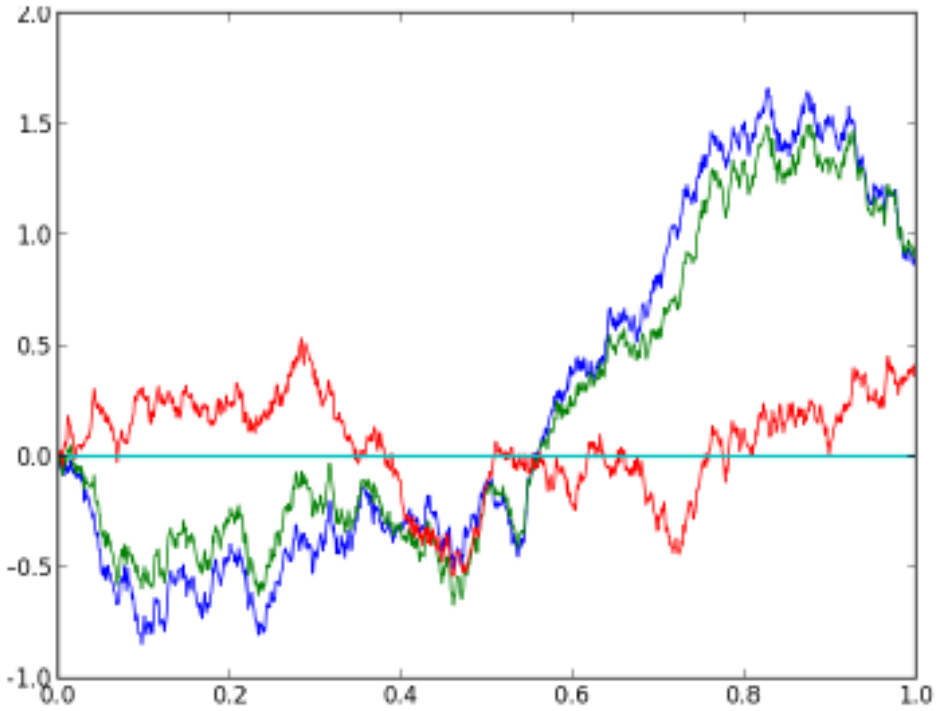


FIGURE 1.2: Gaussian shaking transformation applied on the entire path of an Ornstein-Uhlenlbeck process with different shaking parameters: path before shaking in blue, paths after shaking in red and green

Independent conditional probability estimators by POP method Another issue with splitting idea is the strong interdependence between different conditional probability estimators. Although the final estimator is given as a product, thus avoiding the particular simulation difficulty related to rare event, the elements in the product have strong correlations. It is desirable to have independent estimations of each conditional probability and we naturally expect this to make improvement on numerical performances. We manage to achieve this goal using the theory of ergodicity of Markov chain. More precisely, we will design a Markov chain whose empirical occupation measure approximates the

conditional distribution of the system under investigation and different Markov chains for different conditional distribution run separately. As we shall see, this method does not only give independent estimators for each conditional probability, but also allows parallel implementations, thus further improving the numerical performance. We call this new method POP (Parallel-One-Path) method, which is presented in Subsection 5.2.3, together with implementation remarks.

Briefly speaking, POP methods rely on the Birkhoff's point-wise ergodic theorem. If we can design an ergodic Markov chain $(Z_n)_{n\geqslant 1}$ which has $X|X \in A_k$ as its stationary distribution, then

$$\frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_{Z_n \in A_{k+1}}$$

will converge to $\mathbb{P}(X \in A_{k+1}|X \in A_k)$ as N goes to infinity. We manage to find such a Markov chain and its constructions for different A_k can be made independent. This is made possible by a reversible Markovian transition on path space. Combined with rejection simulation technique, it provides another Markovian transition which preserve the conditional distribution of X.

A slightly more convenient implementation is given in Algorithm 2, where very weak interdependence exists. But this interdependence can be eliminated with negligible computation efforts, as explained in Subsection 5.2.3.

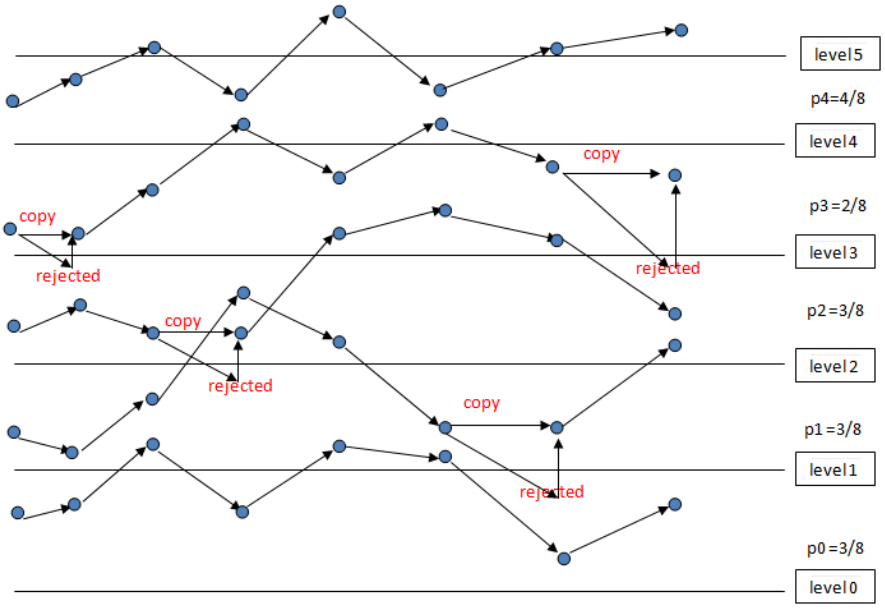


FIGURE 1.3: POP method illustration, computing probability that X as a point lies above level 5

Shaking transformations Under our static point view, the IPS and POP methods are implemented with some reversible Markovian transitions, which we call shaking transformations. Shaking is the key word which summarizes the spirit of our methods, i.e. going through the rare event zone by slight perturbations and keeping the conditional distribution unchanged.

Without further study, Metropolis-Hasting type transitions are natural candidates for shaking transformation in finite dimensional cases. But these transitions are unwieldy for implementation and calibration. They are not numerically efficient either, as explained in Subsection 5.4.4. In this thesis, We propose a particular way of designing our shaking transformation in Section 5.4, which is very easy to implement and to calibrate for achieving best numerical performance. This is done in specific ways for Poisson and geometric distributions. In many continuous random variable cases, this is made possible by an elegant distribution identity called Gamma-Beta algebra:

$$(\Gamma_{a_1,b}, \Gamma_{a_2,b}) \stackrel{d}{=} (B_{a_1,a_2}\Gamma_{a_1+a_2,b}, (1 - B_{a_1,a_2})\Gamma_{a_1+a_2,b})$$

where Γ and B represent independent Gamma and Beta distributions with respective parameters. Using this identity, we can design shaking transformation for Gamma distribution, thus for exponential distribution, which is a particular case of Gamma distribution. Then using the identity

$$U \stackrel{\mathrm{d}}{=} \exp(-\mathtt{Exp}(1))$$

where U is a uniform distribution in [0,1] and $\mathtt{Exp}(1)$ is an exponential distribution, we can design shaking transformations for uniform distribution:

$$\mathcal{K}(U) = U^{\mathtt{Beta}(1-p,p)} \exp(-\mathtt{Gamma}(p,1))$$

The same idea generalizes to the case of many other distributions such as Cauchy distribution and $\chi^2(k)$ distributions. Our density-free way of constructing shaking transformation makes the generalization of our methodology to infinite dimensional cases natural and immediate. For example we can design shaking transformations for compound Poisson process using its classic coloring/superposition decomposition. As will be explained in Section 5.5, our shaking transformation is in some sense a good Metropolis-Hasting type transformation, written with an implicit transition density, as it avoids one rejection step. These identities in law we find are also interesting in themselves.

Sensitivity analysis Another issue we will address in Section 5.7 is the sensitivity of rare event statistics with respect to model parameters. Combined with Malliavin calculus, the idea underlying POP method gives us a surprisingly easy way to evaluate the sensitivity. More precisely, we will show that in many cases rare event relative sensitivity is

easier to evaluate than the rare event probability itself:

$$\frac{\partial_\theta \mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta \in A}\right)}{\mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta \in A}\right)} = \frac{\mathbb{E}\left(\mathcal{I}(Z^\theta, \Phi^\theta) \mid Z^\theta \in A\right)}{\mathbb{E}\left(\Phi^\theta \mid Z^\theta \in A\right)}.$$

We want to compute sensitivity as defined on the left-hand side of the above equation. By Malliavin calculus and Bayesian formula, this sensitivity can also be written as on the right-hand side for some function $\mathcal{I}(Z^\theta, \Phi^\theta)$ to be precised. If we can design a Markov chain to approximate $\mathbb{E}\left(\phi \mid Z^\theta \in A\right)$ for any random variable ϕ, then both $\mathbb{E}\left(\mathcal{I}(Z^\theta, \Phi^\theta) \mid Z^\theta \in A\right)$ and $\mathbb{E}\left(\Phi^\theta \mid Z^\theta \in A\right)$ can be estimated using this Markov chain. Thus only one Markov chain is used to estimated this sensitivity, while we need several Markov chains to estimate rare event probability. See Proposition 5.7.1 and Theorem 5.7.2 and explanations therein.

Adaptive POP method and IPS with more resamplings To make good choice of intermediate subsets, we propose in Section 5.6 an adaptive version of POP method. We also propose in Section 5.9 an IPS method with more resamplings and fewer particles and give numerical experiments to show this improves performance.

Convergence analysis for our methods Convergence analysis on our methods are also given. We will show the L^2 convergence of our POP and IPS methods under several assumptions in Subsection 5.2.4. Then the almost sure convergence of POP method is proved in Section 5.5 for all the finite dimensional cases, without additional assumptions. The almost sure convergence of our adaptive POP method is also demonstrated in Subsection 5.6.2, using results from Markov chain quantile estimations, see Theorem 5.6.1.

Results on Gaussian shaking transformations The shaking transformation in Gaussian framework is particularly interesting. We will show in Subsection 5.3.2 that in one dimensional case, explicit analytical property of Gaussian shaking can be shown via Hermite polynomials, see Lemma 5.3.1. The L^2 convergence of Gaussian shaking transformation in the most general setting, including infinite dimensional cases are also demonstrated in Subsection 5.3.3 using the generalized Gebelein inequality, see Theorem 5.3.3.

Rare event sampling How to apply our techniques to make rare event sampling is discussed in Section 5.8. This is used in financial stress testing and interesting simulation such as Brownian watermelon, see Sections 6.7, 6.8 and 6.13.

At last, many numerical examples are discussed to show the applications of our methods and how to well choose method parameters in Chapter 6. Among others, we have examples on:

- Maximum and oscillation of Ornstein-Uhlenbeck process.
- Insurance ruin probability with compound Poisson model.
- Jackson network in queuing system.
- Atypical configuration of Erdös-Rényi random graph.
- Hawkes process on self-exciting phenomena.
- Model misspecification and robustness in option hedging.
- Measuring default probability in credit portfolios.
- Fractional Brownian motion for modeling volatility.
- Sensitivities for out-of-money options.
- Population survival probability.
- Simulation of Brownian watermelon.

1.2 NISR method for dynamic programming

1.2.1 Problem formulation

Stochastic dynamic programming equations are classic equations arising in the resolution of nonlinear evolution equations, like in stochastic control (see [124, 14]), optimal stopping (see [96, 70]) or non-linear PDEs (see [34, 66]). In a discrete-time setting they take the form:

$$\begin{aligned} Y_N &= g_N(X_N), \\ Y_i &= \mathbb{E}\left(g_i(Y_{i+1}, \ldots, Y_N, X_i, \ldots, X_N) \mid X_i\right), \quad i = N-1, \ldots, 0, \end{aligned}$$

for some functions g_N and g_i which depend on the non-linear problem under consideration. Here $X = (X_0, \ldots, X_N)$ is a Markov chain valued in $\mathbb{R}^d$, entering also in the definition of the problem. The aim is to compute the value function y_i such that $Y_i = y_i(X_i)$.

1.2.2 Literature review

Among the popular methods to solve this kind of problem, we are concerned with Regression Monte Carlo (RMC) methods. One essential part of these methods is the approximation of conditional expectations: given i.i.d. copies $(O_m, R_m)_{1 \leqslant m \leqslant M}$ of two random variables O and R, we want to compute the conditional expectation $f(O) = E(R|O)$. How

to apply regression methods to compute conditional expectations is explained in [75]. Basically a global regression is performed on a dictionary Φ of basis functions and the estimator $\hat{f}$ is taken as

$$\hat{f} := \arg\inf_{\phi\in\Phi} \frac{1}{M}\sum_{m=1}^{M} |R_m - \phi(O_m)|^2$$

And we have the error estimation

$$\mathbb{E}\left(|f-\hat{f}|_{L_2(\mu^M)}\right) \leqslant \sigma^2\frac{K}{M} + \min_{\phi\in\Phi}|f-\hat{f}|_{L_2(\mu)} \tag{1.2.1}$$

where K is the dimension of the vector space Φ, μ is the distribution of O and $\sigma^2 := \sup_o Var(R|O=o)$. For more details, see [67]. We see that if K is large, then we have to take much larger M to achieve good estimations. This issue is particularly serious when dealing with high-dimensional problems, since global approximation in high dimensions needs function dictionaries of large dimension.

To implement RMC methods, we will need M simulated paths of X, say $(X^1,\dots,X^M) =: X^{1:M}$, as input data. These data will help us to build estimations of y_i. Instead of the single period regression above for two variables O and R, we are going to make multi period regressions. Suppose we already have estimations for $y_{i+1}, y_{i+2}, \cdots, y_N$, the simulation-based approximations $y_i^{M,\mathcal{L}}$ for y_i is provided using Ordinary Least Squares (OLS) within a vector space of functions $\mathcal{L}$:

$$y_i^{M,\mathcal{L}} = \arg\inf_{\varphi\in\mathcal{L}} \frac{1}{M}\sum_{m=1}^{M}\left|g_i(y_{i+1}^{M,\mathcal{L}}(X_{i+1}^m),\dots,y_N^{M,\mathcal{L}}(X_N^m),X_i^m,\dots,X_N^m) - \varphi(X_i^m)\right|^2.$$

Basically, we replace $y_{i+1}, y_{i+2}, \cdots, y_N$ in the problem definition by their estimators and we choose among all the functions in $\mathcal{L}$ the one which minimizes the above error. If we use the same $X^{1:M}$ to make all the estimations, some interdependence is introduced among them. Sometimes when extra simulations are easy to make, we can apply re-simulation technique and use independent input data for each regression. This reduces correlations and makes error analysis easier. Remark that in the above regression, we aim at estimating the entire function y_i through one regression. In order to have good performance, we wish to have samples of X_i^m well spread throughout the entire space. If these sample points are too concentrated in one area, then global estimation of y_i may lack accuracy in the area which is not sufficiently represented by X_i^m. Thus a large number of simulation is needed to have good performance with this version of RMC methods.

This Regression Monte Carlo methodology has been investigated in [66] to solve Backward Stochastic Differential Equations associated to semi-linear partial differential equations (PDEs) [106], with some tight error estimates. Generally speaking, it is well known that the number of

simulations M has to be much larger than the dimension of the vector space $\mathcal{L}$ and thus the number of coefficients we are seeking.

1.2.3 Our contribution

In contradistinction, in this thesis, we want to solve the problem in another situation where we are not allowed to make as many simulations as we want. More precisely, we are faced with the case where M is relatively small (a few hundreds) and the paths are not sampled by the user but are directly taken from historical data ($X^{1:M}$ is called **root sample** in this situation), in the spirit of [110]. This is the most realistic situation when we collect data and when the model which fits the data is unknown. In short, we want to solve the problem using limited observed data, without information on a fully specified model and without the permission to make simulations.

Thus, as main differences with the aforementioned references:

- We do not assume that we have full information about the model for X. We need to have information of the model structure but we do not need to have full knowledge on the values of model parameters. Without a fully specified model, naturally we do not assume that we can generate as many simulations as needed to have convergent Regression Monte Carlo methods.

- The size M of the learning samples $X^1, \dots, X^M$ is relatively small, which discards the use of a direct RMC with large dimensional $\mathcal{L}$.

To overcome these major obstacles, we elaborate on two ingredients:

1. First, we partition $\mathbb{R}^d$ in strata $(\mathcal{H}_k)_k$, so that the estimated functions can be computed locally on each stratum $\mathcal{H}_k$. We perform local regression in each *small* stratum. Since the estimation is local, this allows to use only a small dimensional approximation space $\mathcal{L}_k$, and therefore we have only a small number of coefficients to estimate and it puts a lower constraint on M, the number of paths we need, due to our error control similar to Equation (1.2.1). In general, this stratification technique breaks the properties for having well-behaved error propagation due to the resampling technique introduced in the next paragraph, but we manage to provide a precise analysis in order to be able to aggregate the error estimates in different strata. To address the problem of error propagation and complete the convergence analysis, we use a probabilistic distribution ν that has good norm-stability properties with X (see Assumptions 7.3.2 and 7.4.2).

2. Second, by assuming a mild model condition on X, such as arithmetic/geometric Brownian motion, Ornstein-Uhlenbeck process and Lévy process with inhomogeneous coefficients, we are able to resample from the root sample of size M, a *training sample* of

M simulations suitable for the stratum $\mathcal{H}_k$. This resampler is non intrusive in the sense that it only requires to know the form of the model but not its coefficients: for example, we can handle models with independent increments (discrete inhomogeneous Lévy process)

$$U := (X_{i+1} - X_i)_{0 \leqslant i \leqslant N-1}, \qquad \theta_{ij}(x, U) := x + \sum_{i \leqslant k < j} U_k$$

or Ornstein-Uhlenbeck processes

$$\mathbf{X}_t = \mathbf{x}_0 - \int_0^t A(\mathbf{X}_s - \bar{\mathbf{X}}_s)\mathrm{d}s + \int_0^t \Sigma_s \mathrm{d}W_s$$

See Examples 7.2.1-7.2.2-7.2.3-7.2.4 for more details. We call this scheme NISR (Non Intrusive Stratified Resampler), it is described in Definition 7.2.1 and Proposition 7.2.1. To perform local regression mentioned above, we need to have samples located in each stratum. But this is hardly guaranteed by the given historical data due to the small number of paths. Our path resampling technique overcomes this problem by constructing paths starting from each stratum using historical data. This introduces a strong correlations in different regression steps and makes the error analysis more difficult. But we manage to get good error analysis using results based on the concept of covering numbers and uniform concentration inequalities on function dictionary, an introduction of which can be found in [75].

The following picture illustrates how the path construction is done in an additive model, see Section 7.2 for detailed explanations.

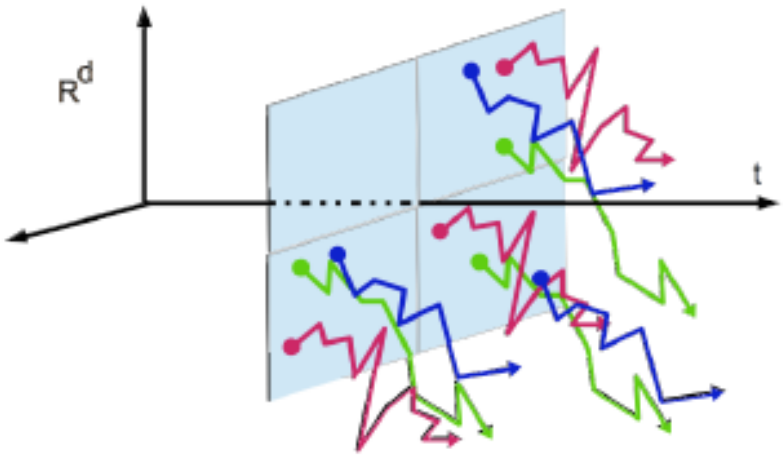


FIGURE 1.4: Description of the use of the root paths to produce new paths

The resulting regression scheme is, to the best of our knowledge, completely new. To sum up, the contributions of this work are the following:

- We design a non-intrusive stratified resample (NISR) scheme that allows to sample from M paths of the root sample restarting from any stratum $\mathcal{H}_k$. See Section 7.2.
- We combine this with regression Monte Carlo schemes, in order to solve one-step ahead dynamic programming equations (Section 7.3), discrete backward stochastic differential equations (BSDEs) and semi-linear PDEs (Section 7.4).
- In Theorems 7.3.4 and 7.4.1, we provide quadratic error estimates, with C_{g_i} as the bound of $|g_i|$, L_{g_i} as the Lipschitz constant of g_i, $\nu(\mathcal{H}_k)$ as the probability mass of ν in $\mathcal{H}_k$, $T_{i,k} := \inf_{\varphi\in\mathcal{L}_k} |y_i - \varphi|^2_{\nu_k}$ and $\nu(T_{i,.}) := \sum_{k=1}^K \nu(\mathcal{H}_k)T_{i,k}$.

  Theorem. *Assume Assumptions 7.2.2-7.2.3-7.3.2-7.3.3 and define $y_i^{(M)}$ as in Algorithm 4. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{E}\left(|y_i^{(M)} - y_i|^2_\nu\right) \leqslant 4(1+\varepsilon)L^2_{g_i}\underline{C}_{(7.3.1)}\mathbb{E}\left(|y_{i+1}^{(M)} - y_{i+1}|^2_\nu\right) + 2\sum_{k=1}^K \nu(\mathcal{H}_k)T_{i,k}$$
$$+ 4c_{(7.3.8)}(M)\frac{|y_i|^2_\infty}{M} + 2(1+\frac{1}{\varepsilon})\frac{\dim(\mathcal{L})}{M}(C_{g_i} + L_{g_i}|y_{i+1}|_\infty)^2 + 8(1+\varepsilon)L^2_{g_i}c_{(7.3.7)}(M)\frac{|y_{i+1}|^2_\infty}{M}$$

  Theorem. *Assume Assumptions 7.2.2-7.2.3-7.3.3-7.4.2 and define $y_i^{(M)}$ as in Algorithm 4. Set*

$$\bar{\mathcal{E}}(Y,M,i) := \mathbb{E}\left(|y_i^{(M)} - y_i|^2_\nu\right) = \sum_{k=1}^K \nu(\mathcal{H}_k)\mathbb{E}\left(|y_i^{(M)} - y_i|^2_{\nu_k}\right).$$

  Define

$$\delta_i = 4c_{(7.3.8)}(M)\frac{|y_i|^2_\infty}{M} + 2\nu(T_{i,.}) + 16\frac{1}{N}\sum_{j=i+1}^{N-1} L^2_{f_j}c_{(7.3.7)}(M)\frac{|y_j|^2_\infty}{M} +$$
$$+ 4\frac{\dim(\mathcal{L})}{M}\left(|y_N|_\infty + \frac{1}{N}\sum_{j=i+1}^{N}(C_{f_j} + L_{f_j}|y_j|_\infty)\right)^2.$$

  Then, letting $L_f := \sup_j L_{f_j}$, we have

$$\bar{\mathcal{E}}(Y,M,i) \leqslant \delta_i + 8\underline{C}_{(7.4.1)}L^2_f \exp\left(8\underline{C}_{(7.4.1)}L^2_f\right)\frac{1}{N}\sum_{j=i+1}^{N-1}\delta_j.$$

which essentially state that

$$\text{quadratic error on } y_i \leqslant \text{ approximation error } + \text{ statistical error} + \text{ interdependency error .}$$

The approximation error is related to the best approximation of y_i on each stratum $\mathcal{H}_k$, and averaged over all the strata. The statistical error is bounded by C/M with a constant C which does not depend on the number of strata: only relatively small M is necessary to get low statistical errors. This is in agreement with the motivation that the root sample has a relatively small size. The interdependency error is an unusual issue, it is related to the strong dependency between regression problems (because they all use the same root sample). The analysis as well as the framework are original. The error estimates take different forms according to the problem at hand (Section 7.3 or Section 7.4).

With this new method, we are able to solve Fisher-Kolmogorov–Petro vsky–Piscounov (FKPP) equations rising in phase transition problems in ecology and optimal stopping problems in dimension 2 with small number of paths and good accuracy, see Chapter 8.

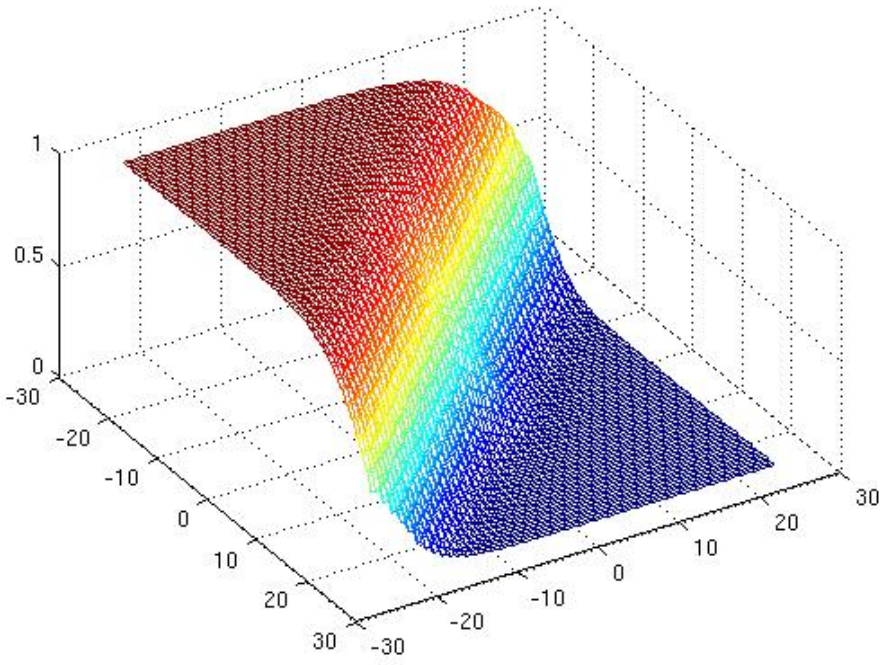


FIGURE 1.5: Estimated solution of FKPP equation with $M = 40$ root samples

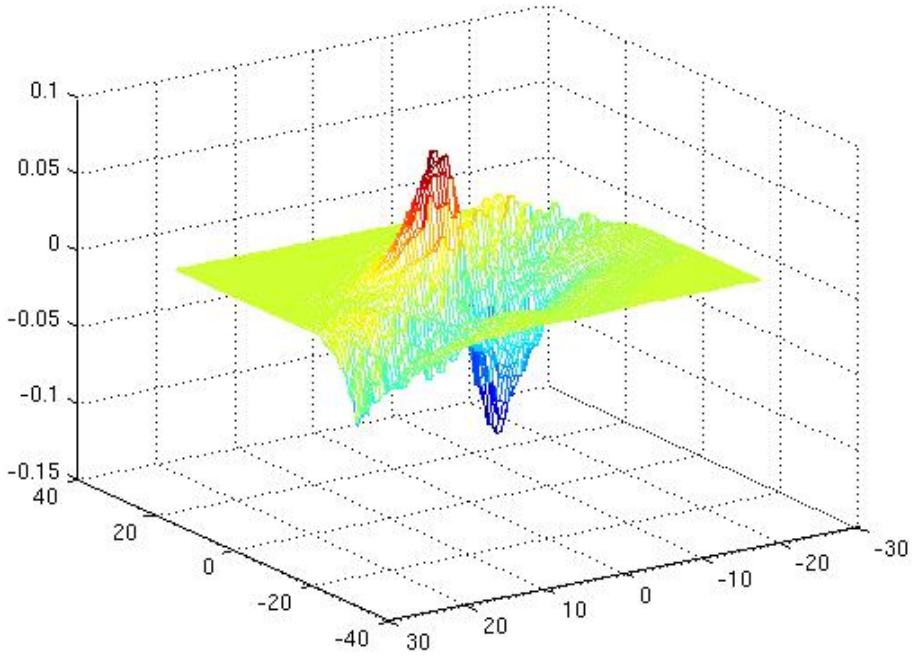


FIGURE 1.6: Estimation error of FKPP equation with $M = 40$ root samples

1.3 Perspectives

Several interesting problems related to this thesis remain to be explored:

- The convergence of Gaussian shaking transformation has been proved in the most general case in Theorem 5.3.3. But the general convergence of Gaussian shaking transformation with rejection remains to be proved in infinite dimensional case. In the numerical example in Section 6.11, we see that on the contrary to intuition, the optimal performance of our estimators do not deteriorate when the event under study becomes rarer and rarer. We suspect this to be related to the geometric property of rare event zone and probability distribution under study.

- The shaking transformation for Poisson process allows us to apply our methodology with jump models. The convergence in this case remains to be analyzed.

- In our numerical examples, we see that the numerical performance of our shaking transformation is closely related to the values of shaking parameter and of rejection rate. How to design a way to optimize these parameters in an automatic and adaptive way remains to be studied.

- We mentioned the application of our methodology in the context of stress test. Further applications on more realistic models and related convergence analysis remain to be conducted.
- Our NISR method may be still valid in the context of Markovian model with only one observation over a long period. Theoretical analysis and numerical tests in this case is not studied in this thesis.

1.4 List of publications

- Emmanuel Gobet and Gang Liu. "Rare event simulation using reversible shaking transformations". In: SIAM Journal on Scientific-Computing 37.5 (2015), A2295–A2316.
- Ankush Agarwal, Stefano De Marco, Emmanuel Gobet, and Gang Liu. "Rare event simulation related to financial risks: efficient estimation and sensitivity analysis". Submitted preprint
- Ankush Agarwal, Stefano De Marco, Emmanuel Gobet, and Gang Liu. "Study of new rare event simulation schemes and their application to extreme scenario generation". Submitted preprint
- Emmanuel Gobet, Gang Liu, and Jorge Zubelli. "A Non-intrusive stratified resampler for regression Monte Carlo: application to solving non-linear equations". Submitted preprint
- Emmanuel Gobet and Gang Liu. Program TEMPO calcul under the protection of Agence pour la Protection des Programmes.

Chapter 2

Résumé en français

Cette thèse contient deux sujets différents: la simulation d'événements rares et la résolution numérique des programmations dynamiques par des méthodes non-intrusives et stratifiées, dont chacun est couvert dans une partie distincte de cette thèse. Dans ce chapitre, en commençant, nous allons présenter brièvement ces deux problèmes, donner une courte revue de la littérature et résumer nos contributions. Des introductions et revues de la littérature complémentaires seront données dans les parties respectives pour les deux sujets.

2.1 Simulation d'événement rare

La simulation d'événements rares concerne l'étude des phénomènes extrêmes, qui ont de très faibles probabilités d'avoir lieu, mais impliquent des conséquences graves une fois qu'ils se produisent. Des exemples d'événements rares sont les suivants: faillite des compagnies d'assurance (Subsection 3.1.1), défaut des réseaux de communications (Subsection 3.1.2), configuration atypique des graphes aléatoires (Subsection 3.1.3) et événement cygne noir dans la finance (Subsection 3.1.5). D'autres applications des événements rares peuvent également être trouvées dans la Section 3.1.

2.1.1 Formulation probabiliste

L'étude probabiliste de l'événement rare commence habituellement avec le cadre suivant:

Étant donné un espace probabiliste $(\Omega, \mathcal{F}, \mathbb{P})$, on considère une variable aléatoire (une application mesurable mesurable) $X : \Omega \mapsto \mathbb{S}$, où $\mathbb{S}$ est un espace d'état général, et un sous-ensemble mesurable $A \subsetneq \mathbb{S}$. Ce sous-ensemble A est pris tel que la probabilité que X se trouve dans A soit extrêmement petite, où on dit que $\{X \in A\}$ est un événement rare. On est principalement intéressé à réaliser les buts suivants:

- Estimer la probabilité de l'événement rare $\mathbb{P}(X \in A)$
- Échantillonner selon la distribution conditionnelle $X|X \in A$
- Estimer l'espérance conditionnellement sur l'événement rare $\mathbb{E}(\varphi(X)|X \in A)$, pour une fonction bornée et mesurable $\varphi : \mathbb{S} \mapsto \mathbb{R}$

- Évaluer les sensibilités des événements rares par rapport aux paramètres des modèles

Dans le cadre des événements rares, $\mathbb{P}(X \in A)$ est typiquement plus petite que 10^{-4}. On suppose toujours que $\mathbb{P}(X \in A) > 0$. Remarquons que cette formulation est très générale, dans le sens où la variable aléatoire X sous considération pourrait être des processus stochastiques corrélés, des graphes aléatoires et d'autres systèmes aléatoires compliqués.

2.1.2 Revue de littérature

Les méthodes de Monte Carlo sont principalement utilisées pour estimer la probabilité et les espérances. La version la plus simple est la méthode de Monte Carlo simple, qui est basée sur la loi des grands nombres et le théorème de la limite centrale: si nous faisons N copies $(X_n)_{1\leqslant n\leqslant N}$ indépendantes et identiquement distribuées (i.i.d.) de X et calculons la proportion empirique des copies qui se trouvent dans A, alors elle converge vers $\mathbb{P}(X \in A)$ lorsque N tend vers l'infini et le théorème de la limite centrale donne des intervalles de confiance correspondants pour nos estimateurs. Malheureusement cette version simple de méthode de Monte Carlo ne parvient pas à donner de bons résultats dans le cas d'événement rare. Étant donné que la probabilité de l'événement $\{X \in A\}$ est très faible, un grand nombre de simulations est nécessaire pour avoir une réalisation de cet événement en moyenne. Par conséquent, le coût de calcul est prohibitif pour obtenir une précision satisfaisante pour notre estimateur. Mathématiquement, cela signifie que la variance relative de notre estimateur est trop élevée pour fournir une bonne estimation.

Plus précisément, on prend N copies i.i.d. $(X_n)_{1\leqslant n\leqslant N}$ de X. On note $p = \mathbb{P}(X \in A)$ et on définit la mesure empirical d'occupation par $\hat{p}_N = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_{X_n\in A}$, alors par le théorème central limite, on a

$$\sqrt{N}(\hat{p}_N - p) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

où $\sigma^2 = p(1-p)$. Donc quand N est grand, approximativement on a un intervalle de confiance de 95% pour la valeur de p:

$$(\hat{p}_N - 1.96\sqrt{\frac{p(1-p)}{N}}, \hat{p}_N + 1.96\sqrt{\frac{p(1-p)}{N}})$$

.

Cela semble très bien, comme la longueur de cet intervalle est égale à $3.92\sqrt{\frac{p(1-p)}{N}}$, qui est petite pour grand N et petit p. Mais si on regarde la longueur relative en pourcentage de p, elle est égale à $3.92\sqrt{\frac{(1-p)}{Np}} \approx 3.92\sqrt{\frac{1}{Np}}$. Si par exemple $p = 10^{-8}$, même si on utilise 10 millions de simulations de X, à la fin on a un intervalle de confiance d'une longueur relative plus grande que 10, donc la conclusion est que p est entre 0 et

10^{-7}. Cet information est complètement inutile pour notre problème, parce que l'erreur est trop grande.

Une technique pour surmonter ce problème d'avoir trop peu de réalisations de notre événement d'intérêt est l'échantillonnage d'importance, voir Subsection 3.2.2. Au lieu de faire des simulations sous la mesure de probabilité initiale, on propose une autre probabilité sous laquelle notre événement d'intérêt $\{X \in A\}$ est plus probable de se produire, donc moins de simulations sont nécessaires pour avoir le même nombre de réalisations en moyenne. Bien sûr en proposant une nouvelle mesure de probabilité un biais est introduit dans l'estimateur et un terme de correction/poids est nécessaire pour donner un estimateur sans biais à la fin. Cette méthode est très efficace lorsque la nouvelle mesure de probabilité est facile à simuler, voir [120, 19, 67, 72, 51]. Mais en général, cette nouvelle mesure de probabilité n'est pas si facile à trouver pour des systèmes compliqués et des études spécifiques sont nécessaires pour des modèles différents.

Une autre idée pour la simulation d'événements rares est celle de splitting, voir Subsection 3.2.3. Au lieu de faire des estimations directes de $\mathbb{P}(X \in A)$, on définit une suite de cascade des sous-ensembles

$$\mathbb{S} := A_0 \supset \cdots \supset A_k \supset \cdots \supset A_n := A,$$

et faire des estimations pour chaque probabilité conditionnelle $\mathbb{P}(X \in A_{k+1} | X \in A_k)$, alors leur produit donne une estimation de la probabilité de notre événement rare. Quelques études sur la méthode de splitting peuvent être trouvées dans [84, 93]. La convergence de la méthode de splitting adaptative a été démontrée dans [32].

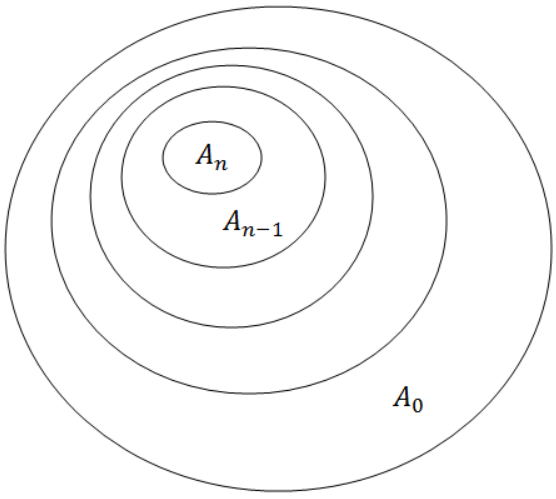


FIGURE 2.1: cascade des sous-ensembles avec splitting

Un problème avec la méthode de splitting initiale est que beaucoup de temps de simulation est passé dans les zones qui sont loin de la zone d'événement rare. Pour surmonter ce problème, la méthode de

RESTART est proposée, voir Subsection 3.2.4. La règle de splitting uniforme dans la méthode initiale est modifiée dans RESTART telle que les efforts de simulation sont concentrés dans des zones importants, voir [13, 125, 126, 86, 102]. Cela rend l'estimation plus efficace dans beaucoup de cas, mais en raison des règles de splitting non uniformes, l'analyse théorique devient plus difficile.

Plus récemment, un autre groupe de méthodes appelées IPS est proposé, voir [42, 43, 31, 30, 29, 27]. Il est basé sur la théorie du système de particules en interaction et il suit aussi l'esprit des méthodes de splitting, à savoir l'estimateur final est donné comme un produit de plusieurs estimateurs de probabilités conditionnelles, voir Subsection 3.2.5. La méthode IPS imite la procédure de sélection naturelle et contient des étapes de sélection et de mutation. La convergence de la méthode IPS adaptative a été démontrée dans [33].

Il y a beaucoup d'autres outils que nous pouvons utiliser pour la simulation d'événements rares, par exemple la méthode d'entropie croisée [119, 24], la théorie des grandes déviations [44, 40], etc. Quelques autres travaux intéressants sont [23] sur la méthode de splitting généralisée et [92] sur des diffusions changeant de régimes. Beaucoup d'autres travaux peuvent être consultés sur les sites Internet des deux derniers ateliers internationaux sur la simulation d'événements rares, RESIM 2014 et RESIM 2016.

2.1.3 Nos contributions

Applicabilité générale Selon nos expériences numériques, lorsque toutes les méthodes peuvent être facilement implémentées, l'échantillonnage d'importance est souvent le plus efficace. Mais des techniques spécifiques sont nécessaires pour gérer chaque problème et la simulation sous la mesure d'échantillonnage d'importance devient très consommatrice en temps de calcul lorsque le modèle est compliqué. Nous visons à la conception d'une nouvelle méthodologie qui a besoin de peu d'ajustements lorsqu'elle est appliquée sur des modèles différents.

Point de vue statique sur l'espace des trajectoires, pas d'hypothèse markovienne et la méthode IPS Combinée avec différents types de techniques, l'idée de splitting applique plus généralement que l'échantillonnage d'importance. Mais il repose toujours sur plusieurs hypothèses. Lorsque le splitting, RESTART ou les méthodes IPS sont appliqués avec un modèle dynamique, les hypothèses markoviennes sont généralement nécessaires. Si le système considéré n'est pas donné comme un modèle markovien, on peut parfois augmenter l'espace d'état. Mais cela rend les algorithmes un peu plus pénibles et en plus la markovianisation n'est pas toujours possible. Dans cette thèse, nous surmontons ce problème en adoptant un point de vue statique sur les modèles dynamiques. Ainsi, nous n'avons plus besoin des hypothèses markoviennes. Ce point de vue statique apporte aussi d'autres avantages. Lorsqu'il est combiné avec la théorie des

systèmes de particules en interaction, elle donne une nouvelle méthode IPS, voir 5.2.2. Lorsqu'on applique cette version de la méthode IPS sur les modèles dynamiques, la discrétisation du temps n'est plus un problème dont il faut se soucier. L'explosion d'erreur avec les méthodes IPS existantes lorsque la discrétisation de temps tend vers zéro n'est plus observée avec notre nouvelle version de la méthode IPS. Ce point de vue statique est implicitement introduit dans notre présentation de transformation de shaking dans la Section 5.2, qui traite des variables aléatoires et processus stochastiques d'une manière unifiée.

Nous allons illustrer les explications ci-dessus par un exemple simple. Supposons que nous avons affaire à la réalisation d'un processus d'Ornstein-Uhlenbeck Z_t entre le temps $t = 0$ et $t = 1$ et nous voulons calculer la probabilité que la valeur maximale de Z_t au cours de cette période soit plus grande que 10. Nous prenons le grille du temps $t_i = \frac{i}{N}$ pour un certain N. Si nous appliquons la méthode IPS avec le point de vue dynamique comme dans [27], nous allons faire des simulations i.i.d. pour le temps t_1, et sélectionner les chemins qui vont vers le haut plus vite que d'autres, appliquer l'étape de mutation pour simuler pour le temps t_2, puis sélectionner à nouveau ces chemins qui vont vers le haut plus vite que d'autres. Nous répétons cette procédure jusqu'au temps t_n. Avec ce point de vue dynamique, les étapes de sélection et mutation sont effectuées au cours de la simulation des chemins browniens. Par contre, si nous appliquons le point de vue statique, la trajectoire entière de Z_t entre le temps 0 et 1 est traitée comme un point indivisible et aucun chemin partiel sera simulé. Nous allons faire la simulation de chemins entiers, sélectionner ceux qui sont plus près de la zone d'événement rare et appliquer la transformation de mutation sur l'espace de chemin. Le graphe suivant montre des exemples de transformations de shaking. Le chemin bleu est celui initial. Le chemin vert est obtenu en appliquant une transformation de shaking légère sur le chemin bleu entier alors que le chemin rouge est obtenu en appliquant une transformation de shaking forte sur le chemin bleu entier.

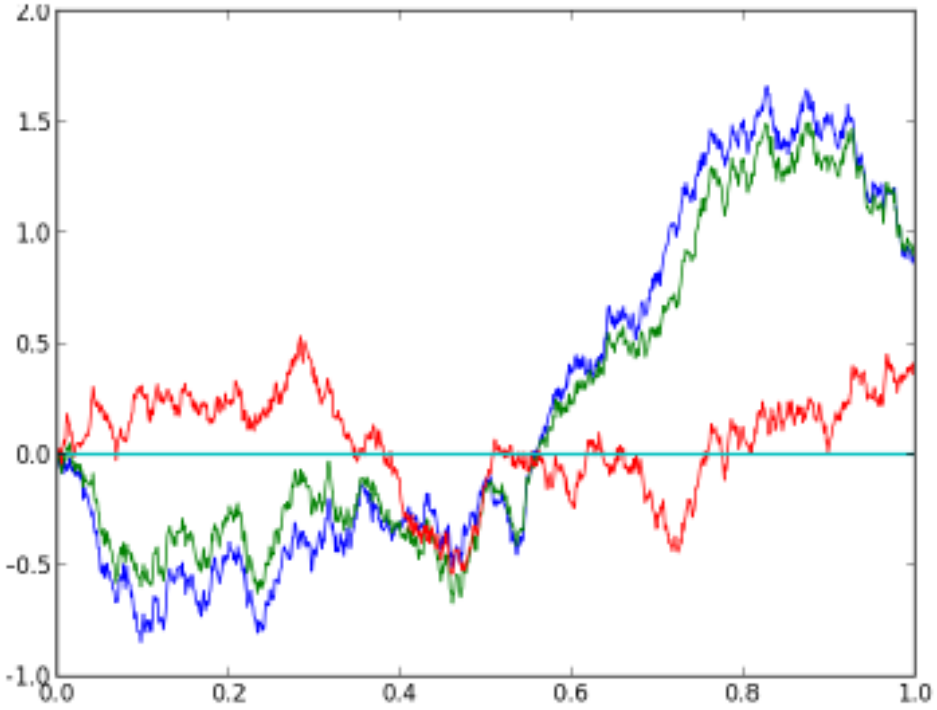


FIGURE 2.2: Appliquer la transformation de shaking gaussienne sur le chemin entier d'un processus d'Ornstein-Uhlenlbeck avec des différents paramètres de shaking. Bleu: chemin avant shaking, vert et rouge: chemins après shaking

Estimateurs indépendants des probabilités conditionnelles par la méthode de POP Un autre problème avec l'idée de splitting est la forte interdépendance entre les différents estimateurs des probabilités conditionnelles. Bien que l'estimateur final soit donné comme un produit, évitant ainsi la difficulté particulière de simulation liée à l'événement rare, les éléments dans le produit ont de fortes corrélations. Il est souhaitable de disposer d'estimations indépendantes pour chaque probabilité conditionnelle et nous nous attendons naturellement à voir des améliorations sur les performances numériques. Nous avons réussi à atteindre cet objectif en utilisant la théorie de l'ergodicité des chaînes de Markov. Plus précisément, nous allons concevoir une chaîne de Markov dont la mesure d'occupation empirique se rapproche de la distribution conditionnelle du système étudiée et des différentes chaînes de Markov pour différentes distributions conditionnelles évoluent séparément. Comme nous le verrons, cette méthode ne donne pas seulement des estimateurs indépendants pour chaque probabilité conditionnelle, mais permet également des implémentations parallèles, ce qui améliore encore la performance numérique. Nous appelons cette nouvelle méthode POP (Parallel-One-Path), qui est présentée dans Subsection 5.2.3, ainsi que des remarques sur l'implémentation.

Brièvement, la méthode de POP repose sur le théorème ergodique trajectoriel de Birkhoff. Si nous pouvons concevoir une chaîne de Markov ergodique $(Z_n)_{n\geqslant 1}$ qui a comme distribution stationnaire $X|X \in A_k$, alors

$$\frac{1}{N}\sum_{n=1}^{N} 1_{Z_n \in A_{k+1}}$$

converge vers $\mathbb{P}(X \in A_{k+1}|X \in A_k)$ lorsque N tend vers l'infini. Nous avons réussi à trouver une telle chaîne de Markov et ses constructions pour les différents A_k peuvent être indépendantes. Cela est rendu possible par une transition markovienne dans l'espace des trajectoires. Combinée avec la technique de rejet, elle donne une autre transition markovienne qui préserve la loi conditionnelle de X.

Une implémentation un peu plus commode est donnée dans l'algorithme 2, où l'interdépendance très faible existe. Mais cette interdépendance peut être éliminée avec des efforts de calcul négligeables, comme expliqué dans Subsection 5.2.3.

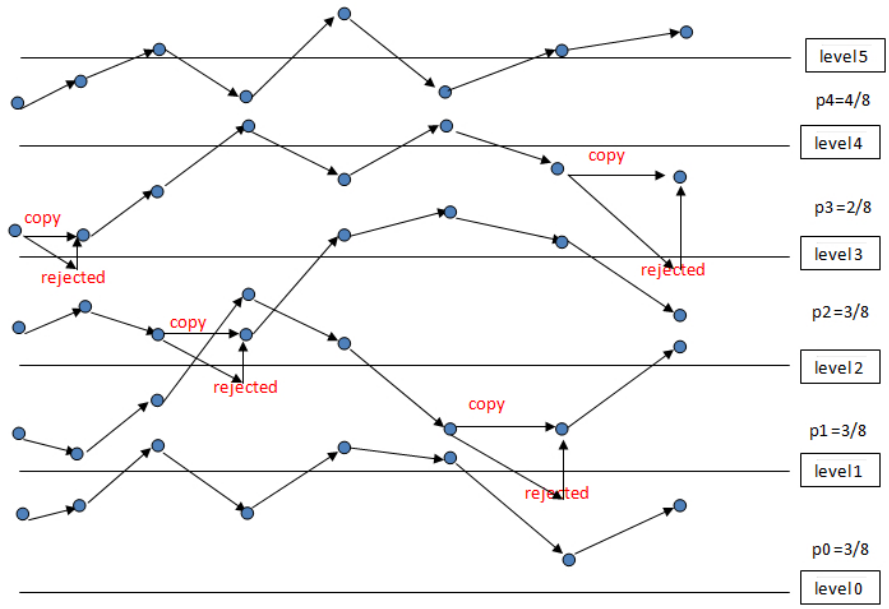


FIGURE 2.3: Illustration de la méthode de POP pour calculer la probabilité qu'un point X se trouve au-dessus du niveau 5

Transformations de shaking Sous notre point de vue statique, les méthodes de IPS et de POP sont implémentées avec des transitions markoviennes réversibles, que nous appelons des transformations de shaking. Shaking est le mot clé qui résume l'esprit de nos méthodes, à savoir passer au travers de la zone d'événement rare par de légères perturbations et préserver la distribution conditionnelle initiale.

Au premier coup d'œil, des transitions de type Metropolis-Hasting sont des candidats naturels pour faire la transformation de shaking dans

le cas fini-dimensionnel. Mais ces transitions ne sont pas directes à implémenter et le calibration de la force de shaking devient difficile avec elles. Elles ne sont pas efficaces d'un point de vue numérique non plus, comme expliqué dans Subsection 5.4.4. Dans cette thèse, nous proposons une façon particulière de concevoir notre transformation de shaking dans la Section 5.4, qui est très facile à implémenter et à calibrer pour atteindre la meilleure performance numérique. Cela se fait des manières spécifiques pour les variables Poisson et géométriques. Pour beaucoup de variables aléatoires continues, ceci est rendu possible par une identité de distribution élégante appelée l'algèbre de Gamma-Beta:

$$(\Gamma_{a_1,b}, \Gamma_{a_2,b}) \stackrel{d}{=} (B_{a_1,a_2}\Gamma_{a_1+a_2,b}, (1 - B_{a_1,a_2})\Gamma_{a_1+a_2,b})$$

où Γ et B représentent des distributions indépendantes de Gamma et de Beta avec des paramètres respectifs. Par cette identité, on peut construire des transformations de shaking pour la loi Gamma, et donc pour la loi exponentielle, qui est un cas particulier des lois Gamma. Ensuite, par l'identité

$$U \stackrel{\mathrm{d}}{=} \exp(-\texttt{Exp}(1))$$

où U est une variable uniforme dans $[0, 1]$ et $\texttt{Exp}(1)$ est une variable exponentielle, on peut construire des transformation de shaking pour la loi uniforme:

$$\mathcal{K}(U) = U^{\texttt{Beta}(1-p,p)} \exp(-\texttt{Gamma}(p, 1))$$

La même idée se généralise sur beaucoup d'autres variables comme les variables Cauchy et les variables $\chi^2(k)$. Notre façon de construire la transformation de shaking sans utiliser la fonction de densité rend la généralisation de notre méthodologie aux cas infini-dimensionnel immédiate. Par exemple on peut construire des transformations de shaking pour le processus de Poisson composé en utilisant sa décomposition classique par coloriage et superposition. Comme cela sera expliqué dans Section 5.5, notre transformation de shaking est en quelque sorte une bonne transformation de type Metropolis-Hasting, donnée avec une densité de transition implicite, car elle évite une étape de rejet. Ces identités probabilistes que nous trouvons sont également intéressants en elles-mêmes.

L'analyse de sensibilité Une autre question que nous aborderons dans Section 5.7 est la sensibilité de statistiques des événements rares par rapport aux paramètres du modèle. Combiné avec le calcul de Malliavin, l'idée sous-jacente dans la méthode POP nous donne un moyen étonnamment facile d'évaluer la sensibilité. Plus précisément, nous allons montrer que dans beaucoup de cas la sensibilité relative de l'événement

rare est plus facile à évaluer que la probabilité de l'événement rare:

$$\frac{\partial_\theta \mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta \in A}\right)}{\mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta \in A}\right)} = \frac{\mathbb{E}\left(\mathcal{I}(Z^\theta, \Phi^\theta) \mid Z^\theta \in A\right)}{\mathbb{E}\left(\Phi^\theta \mid Z^\theta \in A\right)}.$$

On veut calculer la sensibilité comme donnée par le terme ci-dessus à gauche. Par le calcul de Malliavin et la formule de Bayes, cette sensibilité peut aussi s'écrire comme le terme à droite, avec certaine $\mathcal{I}(Z^\theta, \Phi^\theta)$ à préciser. Si on peut construire une chaîne de Markov qui estime $\mathbb{E}\left(\phi \mid Z^\theta \in A\right)$ pour toutes les variables ϕ, $\mathbb{E}\left(\mathcal{I}(Z^\theta, \Phi^\theta) \mid Z^\theta \in A\right)$ et $\mathbb{E}\left(\Phi^\theta \mid Z^\theta \in A\right)$ peuvent être estimées à la fois par cette chaîne de Markov. Donc seulement une chaîne de Markov est utilisée pour estimer la sensibilité, alors que plusieurs chaînes sont utilisées pour estimer la probabilité des événements rares. Voir Proposition 5.7.1 et Théorème 5.7.2 et les explications entre eux.

La méthode de POP adaptative et IPS avec plus de resamplings Pour faire un bon choix de sous-ensembles intermédiaires, nous proposons dans Section 5.6 une version adaptative de la méthode de POP. Nous proposons également dans Section 5.9 une méthode IPS avec plus de resamlings et moins de particules et nous donnons des expériences numériques pour montrer les performances.

Les analyses de convergence de nos méthodes Les analyses de convergence de nos méthodes sont également données. Nous allons montrer la L^2 convergence de nos méthodes de POP et IPS sous plusieurs hypothèses dans Subsection 5.2.4. Ensuite, la convergence presque sûre de la méthode de POP est prouvée dans Section 5.5 pour tous les cas fini-dimensionnels, sans hypothèses supplémentaires. La convergence presque sûre de notre méthode de POP adaptative est également démontrée dans Subsection 5.6.2, en utilisant les résultats sur l'estimation de quantile par la chaîne de Markov, voir Théorème 5.6.1

Résultats sur les transformations de shaking gaussien La transformation de shaking dans le cadre gaussien est particulièrement intéressante. Nous montrerons dans Subsection 5.3.2 que dans le cas uni-dimensionnel, une propriété explicite de shaking gaussienne peut être démontrée via des polynômes Hermite, voir Lemme 5.3.1. Le L^2 convergence des shakings gaussiens dans le cas général, y compris les cas infini-dimensionnels, sont également démontrées dans Subsection 5.3.3 en utilisant l'inégalité Gebelein généralisée, voir Théorème 5.3.3.

L'échantillonnage d'événement rare Comment appliquer nos techniques pour faire l'échantillonnage d'événement rare est discuté dans Section 5.8. Notre méthode de simulation est utilisée dans les stress tests financiers et les applications intéressantes comme la pastèque brownien, voir Sections 6.7, 6.8 and 6.13.

Enfin, de nombreux exemples numériques sont discutés pour montrer les applications de nos méthodes et comment bien choisir les paramètres, dans Chapitre 6. On donne des exemples sur

- Maximum et oscillation d'un processus Ornstein-Uhlenbeck.
- Probabilité de ruine en assurance avec un modèle du processus de Poisson composé.
- Réseau de Jackson dans le système de files d'attente.
- Configuration atypique d'un graphe aléatoire du type Erdös-Rényi.
- Processus de Hawkes sur des phénomènes auto-excitants.
- Mauvaise spécification et robustesse des modèles dans le hedging des options.
- Probabilité de défaut d'un portefeuille de crédit.
- Mouvement brownien fractionnaire pour modéliser la volatilité.
- Sensibilités des options en dehors de la monnaie.
- Probabilité de survie en dynamique de population.
- Simulation de pastèque brownien.

2.2 Méthode de NISR pour la programmation dynamique

2.2.1 Formulation du problème

Les équations des programmations dynamiques stochastiques sont des équations classiques qui apparaissent dans la résolution des équations d'évolution non-linéaires, comme dans le contrôle stochastique (voir [124, 14]), l'arrêt optimal (voir [96, 70]) et les EDPs non-linéaires (voir [34, 66]). Dans un cadre discret, le problème est formulé ainsi:

$$\begin{aligned} Y_N &= g_N(X_N), \\ Y_i &= \mathbb{E}\left(g_i(Y_{i+1}, \dots, Y_N, X_i, \dots, X_N) \mid X_i\right), \quad i = N-1, \dots, 0, \end{aligned}$$

pour certaines fonctions g_N et g_i qui dépendent du problème non-linéaire étudié. Ici $X = (X_0, \dots, X_N)$ est une chaîne de Markov qui prend valeurs dans $\mathbb{R}^d$ et qui est donnée dans la définition du problème. Le but est de calculer les fonctions y_i telles que $Y_i = y_i(X_i)$.

2.2.2 Revue de littérature

Parmi les méthodes populaires pour résoudre ce genre de problème, on s'intéresse aux méthodes de régression Monte Carlo (RMC). Une partie essentielle de ces méthodes est l'approximation des espérances conditionnelles: étant données des copies i.i.d. $(O_m, R_m)_{1\leqslant m\leqslant M}$ de deux variables aléatoires O et R, on veut calculer l'espérance conditionnelle $f(O) = E(R|O)$. Comment appliquer des méthodes de régression pour calculer l'espérance conditionnelle est expliqué dans [75]. Essentiellement, une régression globale est effectuée sur un dictionnaire Φ de fonctions de base et l'estimateur $\hat{f}$ est donné par

$$\hat{f} := \arg\inf_{\phi\in\Phi} \frac{1}{M}\sum_{m=1}^{M} |R_m - \phi(O_m)|^2$$

On a une estimation d'erreur

$$\mathbb{E}\left(|f-\hat{f}|_{L_2(\mu^M)}\right) \leqslant \sigma^2\frac{K}{M} + \min_{\phi\in\Phi}|f-\hat{f}|_{L_2(\mu)} \tag{2.2.1}$$

où K est la dimension de l'espace vectoriel Φ, μ est la distribution de O et $\sigma^2 := \sup_o Var(R|O=o)$. Pour plus de détails, voir [67]. On voit ici que si K est grand, il est nécessaire de prendre M encore beaucoup plus grand pour avoir de bonnes estimations. Ce problème est particulièrement sévère lorsqu'on regarde des applications en grande dimension, parce que l'approximation globale dans les grandes dimensions nécessite un dictionnaire des fonctions de grande dimension.

Pour implémenter les méthodes de RMC, on va avoir besoin de simuler des trajectoires de X, notés $(X^1,\dots,X^M) =: X^{1:M}$, comme les données d'entrée des méthodes. Et ces données vont être utilisées pour construire des estimations de y_i. Au lieu de la régression sur une seul étape, on va faire des régressions sur plusieurs étapes. Supposons qu'on a déjà des estimations pour $y_{i+1}, y_{i+2}, \cdots, y_N$, l'approximation $y_i^{M,\mathcal{L}}$ pour y_i basée sur les simulations est donnée en utilisant la solution aux moindres carrées (Ordinary Least Square, OLS) dans un espace vectoriel des fonctions $\mathcal{L}$:

$$y_i^{M,\mathcal{L}} = \arg\inf_{\varphi\in\mathcal{L}} \frac{1}{M}\sum_{m=1}^{M}\left|g_i(y_{i+1}^{M,\mathcal{L}}(X_{i+1}^m),\dots,y_N^{M,\mathcal{L}}(X_N^m),X_i^m,\dots,X_N^m) - \varphi(X_i^m)\right|^2.$$

Essentiellement, on remplace $y_{i+1}, y_{i+2}, \cdots, y_n$ dans la définition du problème par leurs estimateurs et on choisit parmi toutes les fonctions dans $\mathcal{L}$ celle qui minimise l'erreur ci-dessus . Comme on utilise les mêmes $X^{1:M}$ pour faire toutes les estimations, une certaine interdépendance est introduite parmi eux. Parfois, lorsque des simulations supplémentaires sont faciles à faire, on peut appliquer la technique de re-simulation et alors utiliser des simulations indépendantes pour chaque régression.

Cela réduit les corrélations et rend l'analyse des erreurs plus facile. Remarquons que dans la régression ci-dessus, on vise à estimer l'ensemble de la fonction y_i par une seule régression. Afin d'avoir une bonne performance, on souhaite avoir des échantillons de X_i^m suffisamment répartis dans tout l'espace. Si les points échantillonnés sont trop concentrés dans certaines zones, l'estimation globale de y_i peut manquer de précision dans la zone qui n'est pas assez représentée par X_i^m. Ainsi, un grand nombre de simulation est nécessaire pour avoir une bonne performance avec cette version de méthodes RMC.

Ces méthodes de Régression Monte Carlo ont été étudiées dans [66] pour résoudre les équations différentielles stochastiques rétrogrades associés à des équations aux dérivées partielles semi-linéaires [106], avec des estimations d'erreur fines. Généralement, il est bien connu que le nombre de simulations M doit être beaucoup plus grand que la dimension de l'espace vectoriel $\mathcal{L}$ et donc le nombre de coefficients qu'on cherche.

2.2.3 Nos contributions

A la différence des méthodes présentées précédemment, dans cette thèse, nous voulons résoudre le problème dans une autre situation où nous ne sommes pas autorisés à faire autant de simulations que nous voulons. Plus précisément, nous sommes confrontés au cas où M est relativement faible (quelques centaines) et les chemins ne sont pas échantillonnées par nous-même mais sont directement pris à partir de données historiques ($X^{1:M}$ est appelé **root sample** dans cette situation), dans l'esprit de [110]. C'est la situation la plus réaliste lorsque nous collectons des données et quand le modèle qui correspond aux données est inconnu. En bref, nous voulons résoudre le problème en utilisant les données observées avec une taille limitée, sans information sur un modèle entièrement spécifié et sans autorisation de faire des simulations du modèle.

Ainsi, les différences principales avec les références mentionnées cidessus sont:

- Nous ne supposons pas que nous avons des informations complètes sur le modèle pour X. Nous avons besoin de connaître le type de modèle, mais nous n'avons pas besoin de connaître les valeurs des paramètres du modèle. Sans un modèle entièrement spécifié, naturellement, nous ne supposons pas que nous pouvons générer autant de simulations que nécessaire.

- La taille M des échantillons d'apprentissage $X^1, \dots, X^M$ est relativement petite, ce qui écarte la possibilité d'appliquer RMC directement sur un dictionnaire de fonctions $\mathcal{L}$ avec grande dimension.

Pour surmonter ces obstacles majeurs, nous combinons sur deux ingrédients:

1. Tout d'abord, nous divisons $\mathbb{R}^d$ en des strates $(\mathcal{H}_k)_k$ tel que les fonctions estimées peuvent être calculées localement sur chaque strate $\mathcal{H}_k$. Nous effectuons la régression locale dans chaque *petit* strate. Comme l'estimation est locale, nous avons besoin d'utiliser seulement une espace d'approximation $\mathcal{L}_k$ avec petite dimension, et donc nous avons seulement un petit nombre de coefficients à estimer et cela met une contrainte souple sur M, le nombre de chemins dont nous avons besoin, grâce à des estimations d'erreur similaires à l'Équation (2.2.1). En général, cette technique de stratification ne conserve pas les propriétés garantissant la propagation d'erreur bien réglée à cause de la technique de ré-échantillonnage introduit dans le paragraphe suivant, mais nous réussissons à fournir une analyse précise pour agréger les estimations d'erreur dans les différentes strates. Pour résoudre le problème de la propagation d'erreur, nous utilisons une distribution de probabilité ν qui a de bonnes propriétés de stabilité de normes avec X (voir Hypothèses 7.3.2 et 7.4.2).

2. Deuxièmement, en faisant des hypothèses souples sur le modèle de X, comme des mouvements browniens arithmétiques ou géométriques, processus d'Ornstein-Uhlenbeck et processus de Lévy avec des coefficients non-homogènes, nous sommes en mesure de ré-échantillonner à partir des root samples des *échantillons d'apprentissage* de taille M partant de chaque strate $\mathcal{H}_k$. Ce ré-échantillonnage est non intrusif dans le sens où il ne nécessite que de connaître le type du modèle mais pas ses coefficients: par exemple, nous pouvons prendre des modèles avec des incréments indépendants (processus de Lévy inhomogène et discret)

$$U := (X_{i+1} - X_i)_{0 \leqslant i \leqslant N-1}, \qquad \theta_{ij}(x, U) := x + \sum_{i \leqslant k < j} U_k$$

ou des processus Ornstein-Uhlenbeck

$$\mathbf{X}_t = \mathbf{x}_0 - \int_0^t A(\mathbf{X}_s - \bar{\mathbf{X}}_s)\mathrm{d}s + \int_0^t \Sigma_s \mathrm{d}W_s$$

Voir Exemples 7.2.1-7.2.2-7.2.3-7.2.4 pour plus de détails. Nous appelons ce schéma NISR (Non Intrusif Stratified Resampler), il est décrit dans Définition 7.2.1 et Proposition 7.2.1. Pour effectuer la régression locale mentionnée ci-dessus, nous avons besoin d'échantillonner dans chaque strate. Mais cela n'est pas garanti par les données historiques observées en raison du petit nombre de chemins. Notre

technique de ré-échantillonnage surmonte ce problème en construisant des chemins à partir de chaque strate et en utilisant des données historiques. Cela introduit une forte corrélation dans différentes étapes de régression et rend l'analyse d'erreur plus difficile. Mais nous parvenons à obtenir une bonne analyse des erreurs en utilisant les résultats basés sur le concept de nombre de recouvrement et les inégalités de concentration uniforme sur les dictionnaires des fonctions, dont on peut trouver une introduction dans [75].

L'image suivante montre comment la construction du chemin se fait dans un modèle additif, voir Section 7.2 pour des explications détaillées.

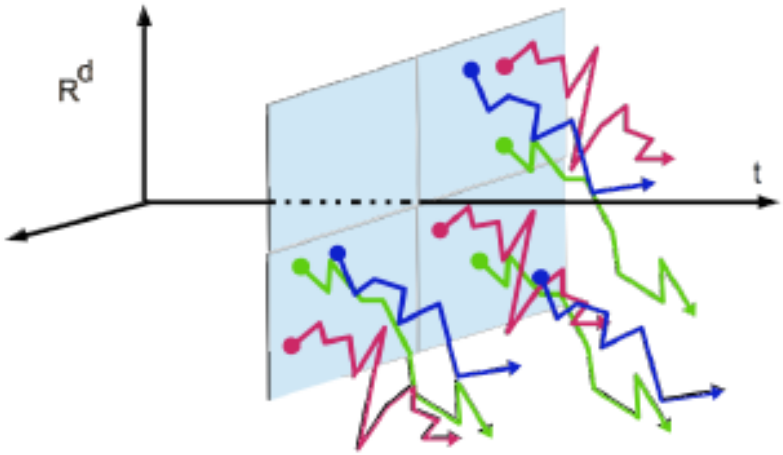


FIGURE 2.4: Description sur comment utiliser les root samples pour construire de nouvelles trajectoires

Le schéma de régression ainsi proposé est, à notre connaissance, complètement nouveau. Pour résumer, les contributions de ce travail sont les suivants:

- Nous concevons un schéma de rééchantillonnage stratifié non intrusif qui permet d'échantillonner à partir de root samples des chemins partant de chaque strate $\mathcal{H}_k$. Voir Section 7.2.
- Nous combinons ceci avec des régressions Monte Carlo, en vue de résoudre les équations des programmations dynamiques (Section 7.3), équations discrètes différentielles stochastiques rétrogrades (EDSR) et EDPs semi-linéaires (Section 7.4).
- Dans les Théorèmes 7.3.4 et 7.4.1, nous donnons l'estimation des erreurs quadratiques de la manière suivante, avec C_{g_i} comme la borne de $|g_i|$, L_{g_i} comme la constante de Lipschitz de g_i, $\nu(\mathcal{H}_k)$ comme le mass de probabilité de ν dans $\mathcal{H}_k$, $T_{i,k} := \inf_{\varphi \in \mathcal{L}_k} |y_i - \varphi|^2_{\nu_k}$ et $\nu(T_{i,\cdot}) := \sum_{k=1}^{K} \nu(\mathcal{H}_k) T_{i,k}$.

Theorem. *Supposons les Hypothèses 7.2.2-7.2.3-7.3.2-7.3.3 et prenons $y_i^{(M)}$ comme dans Algorithme 4. Alors, pour tout $\varepsilon > 0$, on a*

$$\mathbb{E}\left(|y_i^{(M)} - y_i|_\nu^2\right) \leqslant 4(1+\varepsilon)L_{g_i}^2 \underline{C}_{(7.3.1)} \mathbb{E}\left(|y_{i+1}^{(M)} - y_{i+1}|_\nu^2\right) + 2\sum_{k=1}^{K} \nu(H_k) T_{i,k}$$
$$+ 4c_{(7.3.8)}(M)\frac{|y_i|_\infty^2}{M} + 2(1+\frac{1}{\varepsilon})\frac{\dim(\mathcal{L})}{M}(C_{g_i} + L_{g_i}|y_{i+1}|_\infty)^2 + 8(1+\varepsilon)L_{g_i}^2 c_{(7.3.7)}(M)\frac{|y_{i+1}|_\infty^2}{M}$$

Theorem. *Supposons les Hypothèses 7.2.2-7.2.3-7.3.3-7.4.2 et prenons $y_i^{(M)}$ comme dans Algorithme 4. Fixons*

$$\bar{\mathcal{E}}(Y, M, i) := \mathbb{E}\left(|y_i^{(M)} - y_i|_\nu^2\right) = \sum_{k=1}^{K} \nu(\mathcal{H}_k)\mathbb{E}\left(|y_i^{(M)} - y_i|_{\nu_k}^2\right).$$

Définissons

$$\delta_i = 4c_{(7.3.8)}(M)\frac{|y_i|_\infty^2}{M} + 2\nu(T_{i,\cdot}) + 16\frac{1}{N}\sum_{j=i+1}^{N-1} L_{f_j}^2 c_{(7.3.7)}(M)\frac{|y_j|_\infty^2}{M} +$$
$$+ 4\frac{\dim(\mathcal{L})}{M}\left(|y_N|_\infty + \frac{1}{N}\sum_{j=i+1}^{N}(C_{f_j} + L_{f_j}|y_j|_\infty)\right)^2.$$

Alors, avec $L_f := \sup_j L_{f_j}$, on a

$$\bar{\mathcal{E}}(Y, M, i) \leqslant \delta_i + 8\underline{C}_{(7.4.1)}L_f^2 \exp\left(8\underline{C}_{(7.4.1)}L_f^2\right)\frac{1}{N}\sum_{j=i+1}^{N-1}\delta_j.$$

qui essentiellement consiste à dire

erreur quadratique de $y_i \leqslant$ erreur d'approximation + erreur statistique
+ erreur d'interdépendance .

L'erreur d'approximation est liée à la meilleure approximation de y_i sur chaque strate $\mathcal{H}_k$, et la moyenne sur toutes les strates. L'erreur statistique est bornée par C/M avec une constante C qui ne dépend pas du nombre de strates: un relativement petit M est suffisant pour obtenir des erreurs statistiques faibles. Ceci est en accord avec la motivation que le root samples a une taille relativement petite. L'erreur de l'interdépendance est une question inhabituelle, elle est liée à la forte dépendance entre les problèmes de régression (parce qu'ils utilisent tous les même root samples). L'analyse dans le cadre est ainsi originale. Les estimations d'erreur prennent des formes différentes selon les problèmes considérés (Section 7.3 ou Section 7.4).

Avec cette nouvelle méthode, on est capable de résoudre les équations de Fisher-Kolmogorov–Petrovsky–Piscounov (FKPP) liées aux problèmes de transition de phases en écologie et les problèmes d'arrêt optimal en dimension 2 avec un petit nombre de trajectoires et de bonne précisions, voir Chapitre 8.

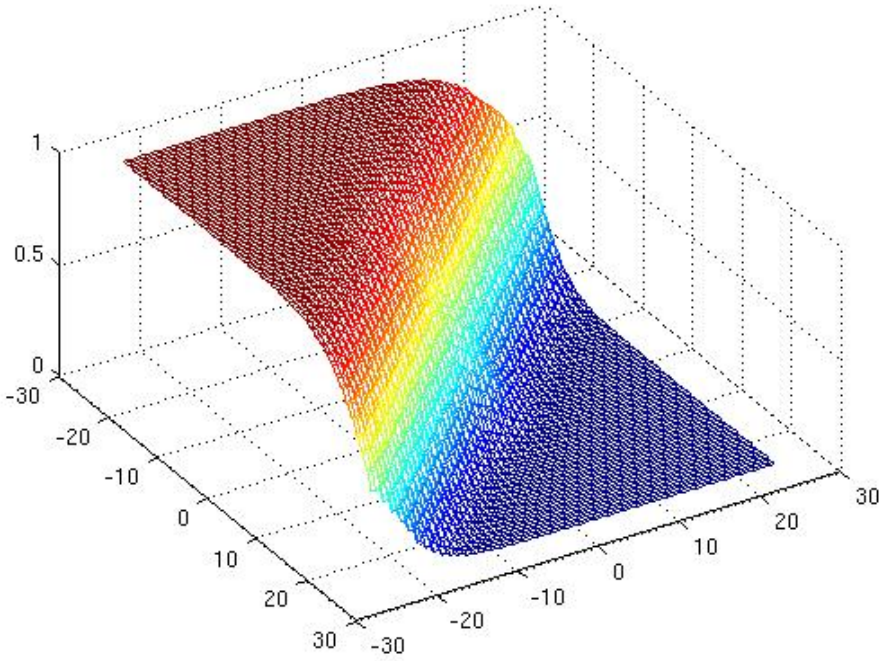


FIGURE 2.5: Solution estimée de l'équation FKPP avec $M = 40$

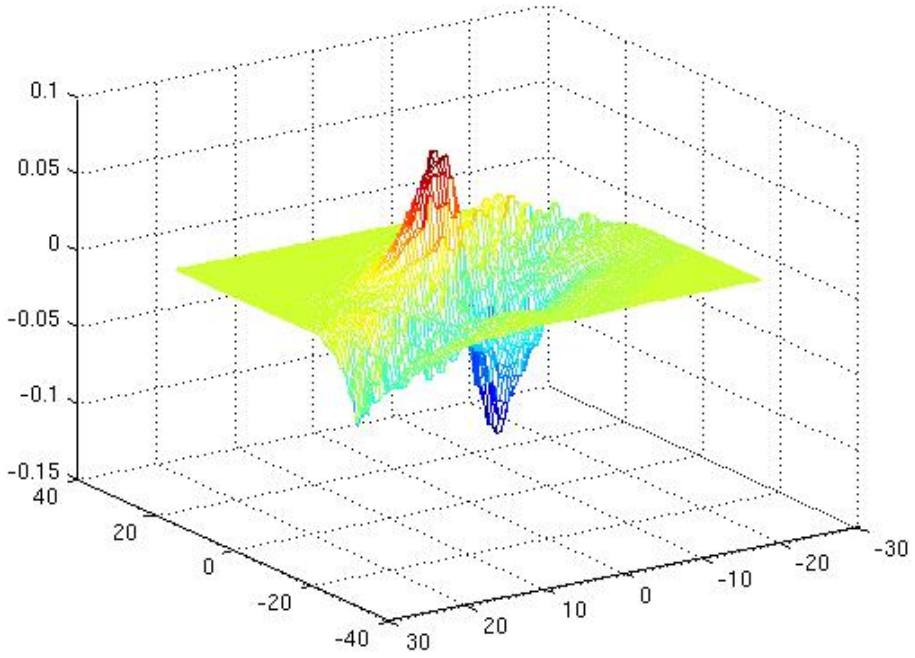


FIGURE 2.6: Erreur d'estimation de l'équation FKPP avec $M = 40$

2.3 Perspectives

Certains sujets intéressants liés à cette thèse restent à être explorés:

- La convergence de transformation de shaking gaussienne a été démontrée dans le cas le plus général dans Théorème 5.3.3. Mais la convergence générale de transformation de shaking gaussienne avec rejet reste à être prouvée dans le cas infini-dimensionnel. Dans l'exemple numérique du Section 6.11, on voit que contrairement à l'intuition, la meilleure performance de nos estimateurs ne se détériore pas lorsque l'événement étudié devient de plus en plus rare. Nous soupçonnons que c'est grâce à des propriétés géométriques de la zone d'événement rare et de la distribution de probabilité considéré.

- La transformation de shaking pour les processus de Poisson nous permet d'appliquer notre méthodologie sur les modèles de saut. La convergence dans ces cas reste à être analysée.

- Dans nos exemples numériques, on voit que la performance numérique de notre transformation de shaking est très liée aux valeur du paramètre de shaking et du taux de rejet. Comment élaborer une manière automatique d'optimiser ces paramètres reste à être étudié.

- On a mentionné l'application de notre méthodologie dans le contexte de stress test. Des applications plus profondes sur des modèles plus réalistes et des analyses de convergence reste à être explorés.
- Notre méthode de NISR pourrait marcher aussi dans le contexte du modèle markovien avec une seule observation sur une longue période. L'analyse théorique et des tests numériques n'ont pas été faits dans cette thèse.

2.4 Liste de publications

- Emmanuel Gobet et Gang Liu. "Rare event simulation usingreversible shaking transformations". In: SIAM Journal on Scientific-Computing 37.5 (2015), A2295–A2316.
- Ankush Agarwal, Stefano De Marco, Emmanuel Gobet, et Gang Liu. "Rare event simulation related to financial risks: efficient estimation and sensitivity analysis". Preprint soumis
- Ankush Agarwal, Stefano De Marco, Emmanuel Gobet, et Gang Liu. "Study of new rare event simulation schemes and their application to extreme scenario generation". Preprint soumis
- Emmanuel Gobet, Gang Liu, et Jorge Zubelli. "A Non-intrusive stratified resampler for regression Monte Carlo: application to solving non-linear equations". Preprint soumis
- Emmanuel Gobet and Gang Liu. Logiciel TEMPO calcul sous la protection de l'Agence pour la Protection des Programmes.

Part I

Rare Event Simulation

Chapter 3

Introduction

The analysis of rare events is an important issue which arises in economy, engineering, life sciences and many other fields. Various applications can be found in actuarial risks [5], communication network reliability [115], aircraft safety [111], social networks and epidemics analysis[21] and other domains, see [25, 117] and references therein. Before talking in more details about the practical implications and different numerical methods for rare event simulation, we will at first give several examples.

3.1 Examples of rare events

3.1.1 Insurance company default

The capital reserve of an insurance company is modeled by

$$R_t = x + ct - \sum_{k=1}^{N_t} Z_k$$

where x is the initial reserve, c is the premium rate, N is a Poisson process with intensity λ and $(Z_k)_k$ are amounts of claims in case of accident or natural disaster [5]. The amounts of claims $(Z_k)_k$ are supposed to follow different probability distributions in different settings. From the perspective of the company manager, we would like to compute

$$\mathbb{P}\left(\min_{0\leqslant t\leqslant T} R_t \leqslant 0\right)$$

i.e. the probability of bankruptcy before T. This probability can also be written in terms of hitting time

$$\mathbb{P}\left(\tau_0 \leqslant T\right) \text{ where } \tau_0 = \inf\{t, R_t \leqslant 0\}$$

This information could help to identify the level of risks that the company is dealing with. There have been many studies on this problem, named ruin probability in insurance science, see for instance [108, 121]. Moreover, the manager may also be wondering: what are the typical scenarios leading to the default of the company?

3.1.2 Communication network reliability

Suppose we have a 2-nodes Jackson network (see [115, Chapter 4] for definition). All the costumers arrive at node 1 and when they are served they go to node 2. The customers' arrival times are jump times of a Poisson process with intensity λ. The serving time at node 1 and at node 2 are respectively exponential variables with parameters μ_1 and μ_2.

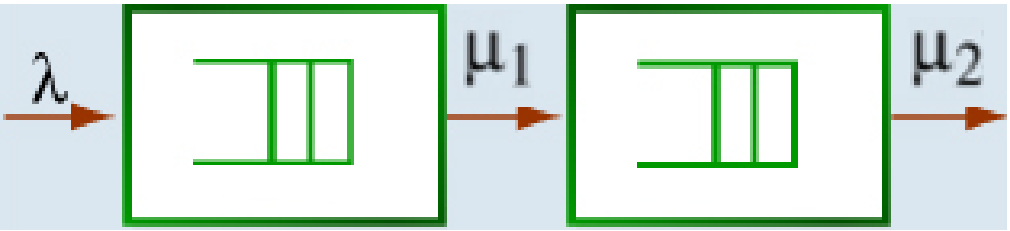


FIGURE 3.1: 2-nodes Jackson network

Our purpose is to compute

$$\mathbb{P}(\max_{0\leqslant t\leqslant T} M_t > K)$$

where M_t denotes the number of customers in the system at time t. In other words, we want to know the probability that at some time before T, the number of customers in the system reaches a fixed level K

In a communication network, those to-be-served customers are information packages transmitted between distant servers. Since each server has a given capacity, when the number of packages waiting to be processed goes out of the limit, the system is saturated and the communication network breaks down. That is what happens during a network collapse, such as in the recent Brussels attacks[1].

3.1.3 Random graph

Random graph is the common tool to model the structure of social network, for instance friend groups on Facebook, and the dynamics of a financial network, such as the effect of risk contagion.

An Erdös-Rényi random graph [21] is a graph with V vertices where every pair of vertices are connected with probability q, independently of each other. It constitutes a toy model for the study of social networks and epidemic. The graph is presented by the upper triangular matrix $X := (X_{ij})_{1\leqslant i<j\leqslant V}$, where

$$X_{ij} = \begin{cases} 1, & \text{if vertices } i \text{ and } j \text{ are connected} \\ 0, & \text{otherwise} \end{cases}$$

[1] http://www.independent.co.uk/life-style/gadgets-and-tech/news/brussels-attacks-phone-networks-zaventem-airport-explosion-maelbeek-metro-live-updates-a6945571.html

If vertices i, j and k are all connected to each other, they form a triangle. Thus the number of triangles in the graph is given by

$$T(X) := \sum_{1 \leqslant i < j < k \leqslant V} X_{ij} X_{jk} X_{ik}.$$

We easily check that

$$\mathbb{E}(T(X)) = \frac{V(V-1)(V-2)}{6} q^3$$

We would like to consider the probability of the deviation event

$$\{T(X) > \frac{V(V-1)(V-2)}{6} t^3\} \text{ for } t > q$$

This problem has attracted recent interest in [35] with theoretical results and in [16] for numerical computation.

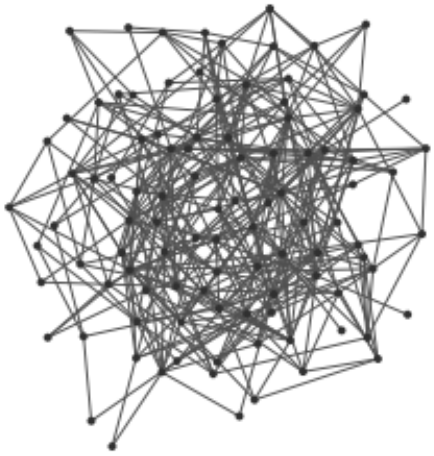


FIGURE 3.2: An example of random graph

3.1.4 Combinatorics, counting problem

The previous random graph problem can also be formulated from a different perspective. Among all the possible configurations, how many configurations contain more than $\frac{V(V-1)(V-2)}{6} t^3$ triangles? This is a combinatorics problem where an explicit answer is not obvious and a direct counting is prohibitively difficult since the total number of configurations is $2^{\frac{V(V-1)}{2}}$. But by taking $q = \frac{1}{2}$, thus attributing to each configuration equal chance to appear, we can get an explicit expression of the number of all such configurations as

$$\mathbb{P}\left(T(X) > \frac{V(V-1)(V-2)}{6} t^3\right) 2^{\frac{V(V-1)}{2}} \tag{3.1.1}$$

Therefore an estimation of $\mathbb{P}\left(T(X) > \frac{V(V-1)(V-2)}{6}t^3\right)$ gives an approximation of (3.1.1). Some other applications of rare event simulation techniques in this direction can be found in [118, 22].

3.1.5 Finance

Rare events find many applications in financial settings, where they are given an elegant name *black swan*(see [123]). A typical example of financial rare event is the financial crisis. During the last thirty years, financial crises have repeatedly occurred, ranging from the Black Monday in 1987 to the recent Chinese stock market crash in 2015, passing through the financial crisis of 2007-2008 triggered by over-valued subprime mortgages. As a consequence, banks, insurance companies and regulators are paying more and more attention to the quantification of risk in all its forms (market risk, credit risk, operational risk) and to its management, in particular in the tails and extremes. In the 90's, the value at risk (VaR) appeared as a common choice of metric to measure the risk in the tails at a given probability (typically 95%) and has been promoted by the Basel committee. Then, convex risk measures, like Expected Shortfall, have emerged to better account for the severity of potential losses and not only for their frequency. More recently, increasing attention has also been paid to stress tests and systemic risk, which are related to extreme scenarios used to evaluate the resilience of individual banks or entire banking system in case of pre-specified unlikely events. All these justify the recent increasing interest in analysis of rare events in finance.

Far from being exhaustive, we give a few examples showing implications of rare events in finance, which are diverse enough to cover a wide range of possible situations.

- The first relevant issue is model risk [39], i.e. the impact of using a misspecified model for hedging financial positions (see Section 6.7 for a detailed example following the analysis of [50, 28]).

- One may also be concerned with credit risk, for which a typical problem is to estimate certain default probabilities required for pricing Credit Default Swaps. An example inspired from [27, 26] will be considered in Section 6.8.

- A third interesting problem is to estimate far-from-the-money implied volatilities (IV). Due to the lack of data for extreme values, standard calibration methods fail to apply here. In the recent work [59, 58], the volatility of S&P 500 index is modeled by fractional Brownian motion (fBM). A study of deep tail of IV in this kind of model will be discussed in Section 6.9

- Another example of market risk comes from the evaluation of deep out-of-the-money options. In Section 6.10, we provide an example

by considering options written on a portfolio of assets and estimate sensitivities with respect to different portfolio and model parameters.

The above examples may give a feeling of how rare event simulation intervenes in different domains of real life. These illustrations are of course far from exhausted and we will have future discussion on them in later chapters. Instead of lengthening the list, we will go to next section and talk about the numerical approaches to address rare event problems and their respective advantages and shortcomings. More examples of rare events will be given as the story goes on and we show the implementations of different methods.

3.2 Existing methods in the literature

Different methods for rare event simulation are grouped under the name *Monte Carlo*, which is originally the name of a casino in Monaco[2]. We will start by specifying the probabilistic setting, which takes a general form to include both finite and infinite dimensional situations. The state space is described by a measurable space $(\mathbb{S}, \mathcal{S})$, where $(\mathbb{S}, d_{\mathbb{S}})$ is a metric space and $\mathcal{S}$ is the Borel sigma-field generated by its open sets. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we consider a random variable (measurable mapping) $X : \Omega \mapsto \mathbb{S}$ and a measurable set $A \subsetneq \mathbb{S}$, then the rare event under investigation is defined by $\{X \in A\}$. We are mainly interested in achieving the following goals

- to estimate the rare event probability $\mathbb{P}(X \in A)$
- to sample from the conditional distribution $X|X \in A$
- to estimate the conditional expectation on rare event $\mathbb{E}(\varphi(X)|X \in A)$, for bounded measurable functions $\varphi : \mathbb{S} \mapsto \mathbb{R}$
- to evaluate the sensitivity of these rare event statistics with respect to model parameters

In the setting of rare event, $\mathbb{P}(X \in A)$ is usually less than 10^{-4}. To avoid uninteresting situations, from now on we always assume that $\mathbb{P}(X \in A) > 0$.

3.2.1 Plain Monte Carlo method and why it fails

The plain Monte Carlo method is based on the central limit theorem: if we make N independent and identically distributed (i.i.d.) copy $(X_n)_{1 \leqslant n \leqslant N}$

[2]For more details about the story behind this puzzling name, see the wikipedia page for *Monte Carlo method*

of X and compute the empirical proportion of copies which lie in A, then it converges to $\mathbb{P}(X \in A)$ as N goes to infinity. More precisely, set $p = \mathbb{P}(X \in A)$ and define the empirical occupation measure by

$$\hat{p}_N = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}_{X_n \in A}$$

then by central limit theorem, we have

$$\sqrt{N}(\hat{p}_N - p) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

where σ^2 is the variance of the random variable $1_{X \in A}$, i.e. $\sigma^2 = p(1 - p)$. Thus, when N is large, approximately we can get a 95% confidence interval

$$(\hat{p}_N - 1.96\sqrt{\frac{p(1-p)}{N}}, \hat{p}_N + 1.96\sqrt{\frac{p(1-p)}{N}})$$

for p.

This may look good at first glance, since the length of this interval is equal to $3.92\sqrt{\frac{p(1-p)}{N}}$, which is small with large N and small p. But if we look at the relative length in percentage of p, it is equal to

$$3.92\sqrt{\frac{(1-p)}{Np}} \approx 3.92\sqrt{\frac{1}{Np}}$$

If for example $p = 10^{-8}$, even if we use 10 million simulations of X, at the end we get a confidence interval of relative length more than 10, so the final conclusion may be that p is between 0 and 10^{-7}. This information is completely useless in our problem, since the possible error goes far beyond our tolerance.

3.2.2 Importance sampling

The problem with plain Monte Carlo method is that the probability of having a realization of X inside A is too small under the original probability distribution. One way to fix the problem is to design another probability distribution under which the event $\{X \in A\}$ is more likely to happen. The idea will be illustrated with the following simple example.

Example 3.2.1. *X is a normal variable with mean m and variance σ^2 under probability $\mathbb{P}$, we define another probability by*

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp(\lambda(X - m) - \frac{\lambda^2}{2}\sigma^2)$$

then under probability $\mathbb{Q}$ X is a normal variable with mean $m + \lambda\sigma^2$ and variance σ^2. This can be easily proven by computing the Laplace transform of X

under $\mathbb{Q}$. Let's take $m = 0$ and $\sigma = 1$, i.e. X is a standard normal variable. We want to estimate the probability that X is greater than 10. To do that, we shall design the probability $\mathbb{Q}$ in the above way with $\lambda = 10$, thus X has the mean 10 under probability $\mathbb{Q}$, then we have

$$\begin{aligned}\mathbb{P}(X > 10) &= \mathbb{E}^{\mathbb{P}}(1_{X>10}) \\ &= \mathbb{E}^{\mathbb{Q}}(\exp(50 - 10X)1_{X>10})\end{aligned}$$

Thus we are going to make N i.i.d. simulations $(X_n)_{1\leqslant n\leqslant N}$ of X under $\mathbb{Q}$ and write our estimator of $\mathbb{P}(X > 10)$ in the following way

$$\begin{aligned}\mathbb{P}(X > 10) &= \mathbb{E}^{\mathbb{Q}}(\exp(50 - 10X)1_{X>10}) \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\exp(50 - 10X_n)1_{X_n>10}\end{aligned}$$

A confidence interval for $\mathbb{P}(X > 10)$ can be found similarly as in section 3.2.1

Importance sampling is a very powerful method to address rare event simulation problem and it has various applications in different setting of contexts. More importance sampling techniques on other probability distributions and their related optimality analysis can be found in [120, 19, 67]

However, importance sampling is not always applicable, its feasibility depends much on the particularity of the model at hand. Various studies have been carried to apply this method on more sophisticated setting, see for instance [72], and it is still a active research field. But we are not going to explore in this direction. Before closing this subsection, we will just recall another importance sampling technique, applied on compound Poisson processes. Later we shall compare performances of new methods against importance sampling in this setting.

Theorem 3.2.1 ([51]). *Let $X_t = \sum_{i=1}^{N_t} Y_i$ be a compound Poisson process with jump intensity λ and jump distribution ν, given a terminal time T, we define*

$$Z_T = \exp\Big(\sum_{i=1}^{N_t} f(Y_i) - \int_{\mathbb{R}}(e^{f(x)} - 1)\lambda T\nu(\mathrm{d}x)\Big)$$

then under the probability $\mathbb{Q}$ define by $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = Z_T$, $(X_t)_{0\leqslant t\leqslant T}$ is a compound Poisson process with jump intensity $\lambda\int_{\mathbb{R}} e^{f(x)}\nu(\mathrm{d}x)$ and jump distribution $\frac{e^{f(x)}\nu(d\times)}{\int_{\mathbb{R}} e^{f(x)}\nu(d\times)}$

3.2.3 Splitting

The earliest reference to our knowledge of splitting[3] method is [84], as said therein mentioned by Dr. von Neumann.

The idea of splitting method is the following: we will use some importance function to divide the space into zones of different importance levels, with the rare event zone being the most important one. So we will get a series of nested subsets

$$\mathbb{S} := A_0 \supset \cdots \supset A_k \supset \cdots \supset A_n := A, \tag{3.2.1}$$

We shall illustrate the procedure of splitting method with an example

Example 3.2.2. *Suppose we have a continuous stochastic process X starting from 0, and the rare event is defined by $\{\tau_{10} < \tau_{-1}\}$, where τ_a is the stopping time that X reaches level a. To implement the splitting method, we define nested subsets $A_k = \{\tau_k < \tau_{-1}\}, k = 1, 2, \cdots, 10$.*

We will start by simulating M trajectories of X, labeled with 0, gradually as time goes on. For each trajectory that enters A_1, it is split into R_1 sub-trajectories, labeled with 1, each of which will evolve independently. Similarly, each time one trajectory label with $k-1$ goes from a less important zone A_{k-1} to a more important zone A_k, it is split into R_k independent sub-trajectories, labeled with k. Then finally we will count the number of trajectories that have reached the rare event zone A, denoted by M_A, and the rare event probability is estimated as

$$\frac{M_A}{M \prod_{k=1}^{n} R_k}$$

The procedure is illustrated in the following figure:

[3] Due to the vast amount of literature, *splitting method* may not have exactly the same meaning in different places, it needs to be understood in the context

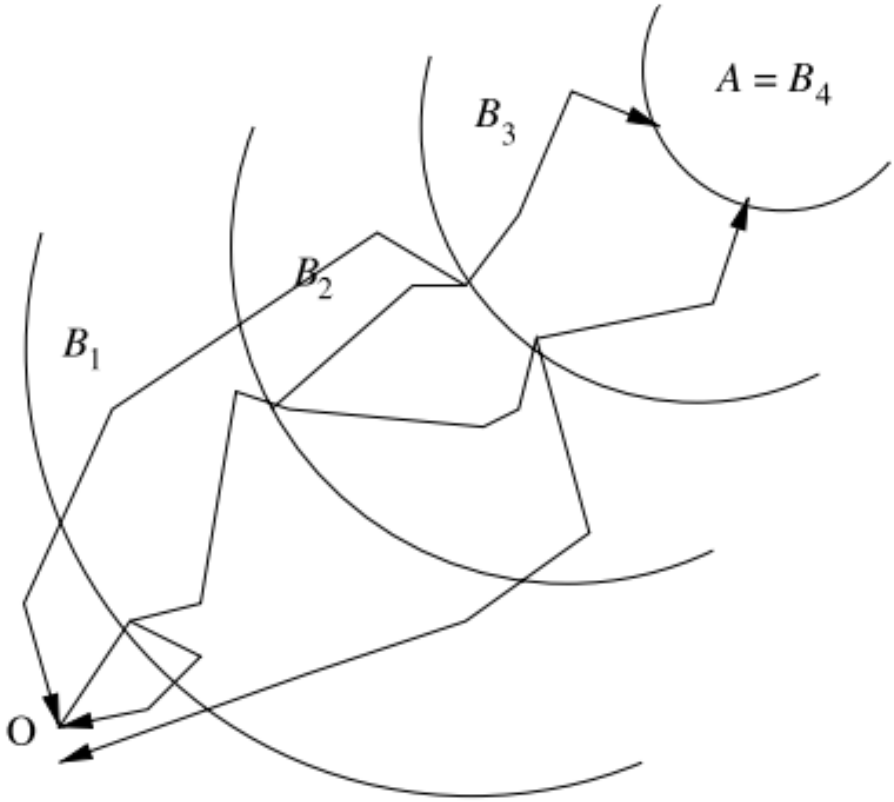


FIGURE 3.3: Splitting illustration from [93]

One drawback in the implementation of splitting method is that the trajectory does not only go to more and more important zone, it can also evolve in the opposite direction, thus much simulation time could be wasted in the unimportant zone and this leads to a loss of efficiency. To moderate this negative affect, the authors of [84] propose a random decision on the simulation continuation at each time one trajectory goes from a more important zone to a less important one, and then add weights to remain unbiased.

To achieve good performance with the splitting method, one needs to be careful about the choice of importance function, the choice of nested subsets and the values of R_k given a fixed computation budget. This problem has been studied in various context, see [93] for example where the author gives explicit formula on these choices.

Without enough information on the problem at hand, it may be difficult to choose appropriate nested subsets. Adaptive splitting method can be used in this kind of situation. Analysis on its convergence has been studied in [32].

3.2.4 RESTART

RESTART (REpetitive Simulation Trials After Reaching Thresholds) method was firstly introduced in [13]. Then the fame and enthusiasm around

this method get a great spur with the paper [125] in 1991, where the authors coined the name for this method and made more theoretical analysis. Sometimes the term RESTART is used as a synonym of splitting. But as is explained in [126], there exist differences between the splitting method in the above section and the RESTART method.

RESTART method shares the spirit of splitting. But different from splitting method where all the trajectory are treated equally, there exist main trajectory and retrials trajectory in RESTART method. A detailed explanation of RESTART method can be found in [126]. We take example 3.2.2 to explain the two main differences: firstly, for a retrial trajectory labeled with k, it has no right to descend(in the sense of importance zone) into A_{k-1} with RESTART method. Once it goes back into A_{k-1}, this retrial trajectory is killed. This feature of RESTART method avoids spending too much simulation time in the unimportant zones. But if we only added this modification to the splitting method, there could be a bias in the final estimator, since those killed trajectories could have had a chance to reach A if they had been kept. To counterbalance this effect, we give more privilege to the main trajectory: they are given more chance to split, i.e. when a main trajectory labeled with k descend into A_{k-1} then ascends to A_k, it could split again whereas in the splitting method it is not allowed to split in this case.

The RESTART method has been applied in a lot of models, see [86, 102] for instance.

3.2.5 Interacting particle system

More recently, methods using interacting particle systems(IPS) have been developed. A systematic account of interacting particle system can be found in the book [42]. To our knowledge, the first application of IPS theory in the simulation of Markov chain related rare events is proposed in [43], then it has been discussed in [31, 30, 29] among many others. The convergence of adaptive IPS method has been shown in [33].

The idea of interacting particle system is to use the empirical measure of a large particle system to approach a target measure. In the case of rare event, a large system is simulated according to a initial distribution, then these particles which are more closer to the rare event zone are given more possibility to be selected and evolved into the next stage. This selection-mutation procedure resembles the evolution process of genes. That is why sometime IPS methods are also called genetic algorithms. IPS method is applied to static distributions in [29],. When IPS is applied to stochastic process, this selection-mutation procedure is usually done during the evolution of the process, such as [27]. This requires some Markovian assumption on the underlying model, and when the process is not Markovian some state augmentation technique is needed.

3.2.6 Others

The above methods do not cover all the study in rare event simulation at all. There are many other tools we can use such as cross-entropy method [119, 24], large deviation theory [44, 40], etc. Some other interesting works are [23] on generalized splitting method and [92] on switching diffusions. A lot of other works can be found on the websites of the recent two International Workshops on Rare Event Simulation, RESIM 2014[4] and RESIM 2016[5].

3.3 Our methods

Methods like importance samplings are very efficient, but lacks generality of application. Specific techniques need to be developed according to each model's particularity. Splitting, RESTART and IPS methods all enjoy wide applications, but most of the time they are developed on Markovian assumptions and take a dynamic point of view when dealing with stochastic process. This may leads to some implementation difficulty and sometime the error may explode as the discretization step length of process tends to zero. We aim at developing methods which can be applied generally and with less assumptions. Of course we are also looking for better numerical performance. Our contributions are summarized in the following.

- We design a Markovian transition called *shaking transformations* on the paths space, which enables us to propose IPS and POP methods for rare event simulation, based respectively on interacting particle system and the ergodicity of Markov chain. These methods work in a Black-Box manner, which ask little information on the system at hand to be implemented.

- We propose a particular way of designing shaking transformation, which is very easy to implement and to calibrate for achieving best numerical performance. Our density-free way of constructing shaking transformation makes the generalization to infinite dimensional cases natural and immediate. As will be explained later, our shaking transformation is in some sense a good Metropolis-Hasting type transformation, written with an implicit transition density, as it avoids one rejection step. These identities in law we find are also interesting in themselves.

- The perspective of transformation on the path spaces allows us to apply the static point view with interacting particle system to deal with dynamic problem. Different from existing particle methods

[4] http://www.tinbergen.nl/conference/10th-international-workshop-rare-event-simulation/

[5] http://www.eurandom.nl/events/workshops/2016/SAM_PROB+ANALYSIS/WS3.html

using dynamics point view, the performance of our IPS method will not deteriorate as time discretization becomes finer and finer.

- The POP method we propose runs in a parallel way, and can give independent estimator for each conditional probability, thus achieves very good numerical performance.
- The almost sure convergence of POP method is proved in all the finite dimensional cases.
- We prove one interesting property of Gaussian shaking combined with Hermite polynomials, as well as the convergence of Gaussian shaking transformation in the most general setting, including infinite dimensional cases.
- We propose an adaptive version of the POP method, which demands less information on the model to be well implemented. We also prove its consistency.
- We demonstrate how these techniques can be applied to make sensitivity analysis of rare event statistics on model parameters and to make approximative sampling of rare event.
- A variant of IPS method is proposed with extra resampling and fewer particles.
- Many numerical examples are discussed to show the applications of our methods and how to well choose method parameters.

Most of the materials in this first part are contained in our papers [68, 2, 3].

Chapter 4

Brief review of ergodicity and IPS theories

This chapter gives a brief review of the theories of Markov chain ergodicity and of interacting particle system, which lay the foundation of the numerical methods for rare event simulation that we are going to develop in the next chapter.

4.1 Ergodic theory for Markov chain

Ergodic theory is by itself a very deep and active research domain, whose development includes the work of one recent Field medalist Artur Avila. This goes of course far beyond the scope of the current PhD thesis. We are just going to look at the application of classic ergodic theory in a probabilistic framework and how this gives rise to a powerful set of simulation methods called MCMC (Markov Chain Monte Carlo) methods.

4.1.1 Ergodic theory

According to [107], the first ergodic theorem is probably a nowadays widely known result in [127].

Theorem 4.1.1 ([127]). *Let α be a irrational number, and define a sequence of numbers between 0 and 1 by*

$$x_n = n\alpha \mod 1$$

then for any interval $I \in [0,1]$, denoting its length by $|I|$, we have

$$\frac{1}{n}\sum_{k=1}^{n} 1_{x_k \in I} \to |I|$$

That is, given the first n numbers in the sequence, if we look at the proportion which is inside the interval I, this proportion tends to the length of this interval. However, if α is rational, the result is no longer true, as x_n only has finite possible values in this case.

To appreciate the powerful theory lurking behind this theorem, we need several definitions.

Definition 4.1.1. *Given a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and a measurable mapping $T : \Omega \to \Omega$, T is said to be measure-preserving if*

$$\forall B \in \mathcal{B}, \mathbb{P}(T^{-1}B) = \mathbb{P}(B)$$

We also say that $\mathbb{P}$ is a T-invariant measure.

Take $f \in L^1(\Omega, \mathcal{B}, \mathbb{P})$ and $\mathcal{F}$ anther filtration smaller than $\mathcal{B}$, then if T is measure-preserving, we have

$$\mathbb{E}\left(f|\mathcal{F}\right) \circ T = \mathbb{E}\left(f \circ T|T^{-1}\mathcal{F}\right)$$

Definition 4.1.2. *Given a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and a measure-preserving mapping $T : \Omega \to \Omega$, we say that T is ergodic if for any $B \in \mathcal{B}$, $T^{-1}B = B$ implies $\mathbb{P}(B) = 0$ or 1. We also say that $\mathbb{P}$ is an ergodic measure for T*

We are then going to state two general ergodic theorem: von Neumann's mean ergodic theorem and Birkhoff's point-wise ergodic theorem.

Theorem 4.1.2. *Given a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and a measure-preserving mapping $T : \Omega \to \Omega$, take $f \in L^2(\Omega, \mathcal{B}, \mathbb{P})$, then we have*

$$\frac{1}{n}\sum_{k=0}^{n-1} f \circ T^n \xrightarrow{L^2} \mathbb{E}\left(f|\mathcal{G}\right)$$

where $\mathcal{G} = \{B \in \mathcal{B} : T^{-1}B = B \text{ a.e.}\}$. When T is ergodic, $\mathbb{E}\left(f|\mathcal{G}\right) = \mathbb{E}\left(f\right)$.

Theorem 4.1.3. *Given a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and a measure-preserving mapping $T : \Omega \to \Omega$, take $f \in L^1(\Omega, \mathcal{B}, \mathbb{P})$, then we have*

$$\frac{1}{n}\sum_{k=0}^{n-1} f \circ T^n \to \mathbb{E}\left(f|\mathcal{G}\right) \text{ a.e.}$$

where $\mathcal{G}$ is defined in the same way as above.

4.1.2 Ergodic theory for Markov chain

To apply the above theorems on Markov chain, we firstly need to precise what probability space we are working with. Suppose we are dealing with a one-dimensional Markov chain $(X_n)_{n\geqslant 0}$ taking values in $\mathbb{R}$, then we will take $\Omega = \mathbb{R}^{\mathbb{N}}$ such that each element ω in Ω is written as $\omega = (x_0, x_1, x_2, \cdots, x_n, \cdots)$. The probability measure on Ω is induced by our Markov chain X, i.e. for any cylinder

$$\mathcal{C} = \mathbb{R} \times \cdots \times \mathbb{R} \times B_{i_1} \times \mathbb{R} \times \cdots \times \mathbb{R} \times B_{i_2} \times \cdots \times B_{i_r}$$

where i_r's are time indices, we define

$$\mathbb{P}(\omega \in \mathcal{C}) = \mathbb{P}(X_{i_1} \in B_{i_1}, X_{i_2} \in B_{i_2}, \cdots, X_{i_r} \in B_{i_r})$$

Define T as the shift operator, i.e. for $\omega = (x_0, x_1, x_2, \cdots)$

$$T(\omega) = (x_1, x_2, x_3, \cdots)$$

and take $f(\omega) = f(x_0)$, then if T is measure preserving and ergodic, we have for example

$$\frac{1}{n}\sum_{k=0}^{n-1} f(x_k) \to \mathbb{E}\left(f(x_0)\right) \text{ a.e.}$$

i.e.

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \to \mathbb{E}\left(f(X_0)\right) \text{ a.e.} \tag{4.1.1}$$

where we replace ω by the realization of our Markov chain $(X_n)_{n\geqslant 0}$.

The above convergence tells us, if our Markov chain satisfies the measure-preserving and ergodicity conditions, its time average will be close to its space average at the starting time.

Naturally we are now wondering what are the necessary and sufficient conditions for T to be measure preserving and ergodic.

Given a Borel set $C \in \mathbb{R}$, take $B = C \times \mathbb{R}^N$, we have $T^{-1}B = \mathbb{R} \times C \times \mathbb{R}^N$. Thus for T to be measure preserving, we need to have $\mathbb{P}(X_0 \in C) = \mathbb{P}(X_1 \in C)$. Or more generally, X_n has the same distribution as X_{n+1} for any positive integer n. That is equivalent to saying the law of X_n is invariant with respect to n. We call such a Markov chain *stationary*. Therefore one of the important conditions for (4.1.1) to hold is that our Markov chain $(X_n)_{n\geqslant 0}$ is stationary.

Next, we need to look into the implication of the ergodic condition. The Markov chain interpretation of this condition is less direct. Instead we are going to look at its consequence: i.e. the time average converges to the space average.

We will suppose the stationary condition always holds and denote by π the stationary distribution, i.e. $X_n \sim \pi, \forall n$, then (4.1.1) can be slightly rewritten as

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \to \mathbb{E}_\pi\big(f\big) \text{ a.e.} \tag{4.1.2}$$

to emphasize the dependence on π

Take a Borel set C such that $\pi(C) > 0$ and $f(\cdot) = \mathbf{1}_{\cdot\in C}$, the above convergence gives $\frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_{X_k\in C} \to \pi(C)$ almost surely. Obviously, there has to be some k such that $\mathbb{P}(X_k \in C) > 0$ for this to hold.

This does look like a over-complication, since $\mathbb{P}(X_k \in C) = \pi(C)$ according to our assumption. The above reasoning is just one way to prepare the introduction of definitions in the following. Actually, the

power in real applications of Markov chain ergodicity is best demonstrated when the initially distribution of X_0 is different from π, which is difficult to simulate under some situations.

Definition 4.1.3. *Given a Markov chain taking values in some general state space $\mathbb{S}$, with the transition kernel $P(x, dy)$, π is said to be a stationary (or invariant) distribution of this Markov chain if $\pi = \pi P$*

Definition 4.1.4. *A Markov chain X_n is said to be η-irreducible, if for any measurable set B such that $\eta(B) > 0$, we have that for any $x \in \mathbb{S}$, there exist some $n \geqslant 1$, possible depending on x such that, $\mathbb{P}_x(X_n \in B) > 0$, where $\mathbb{P}_x$ represents the probability induced by the Markov chain starting at x.*

Convergence of occupation measure

The following is a classic result and one of the most important facts on Markov chain:

Theorem 4.1.4 ([7]). *Given a Markov chain having a stationary distribution π and being η-irreducible for some η, for any positive function f, we have*

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_j) \to \int f(x)\pi(dx), \mathbb{P}_x - a.s.$$

as n goes to infinity, for $\pi - a.a.x$

The proof of this theorem can be found in standard Markov chain textbooks, such as [98]. More recently, a short proof based on Theorem 4.1.3 is given in [7]

Another result for which a proof can be found in [7] is the following:

Theorem 4.1.5 ([7]). *If the transition kernel of the Markov chain is given by*

$$\mathbb{P}_x(X_1 \in dy) = (1 - a(x))\delta_x(dy) + a(x, y)q(x, y)\eta(dy) \tag{4.1.3}$$

where $a(x) > 0, \forall x \in \mathbb{S}$ and X is η-irreducible and has a stationary distribution π, then , we have

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_j) \to \int f(x)\pi(dx), \mathbb{P}_x - a.s.$$

as n goes to infinity, for any $x \in \mathbb{S}$.

Remark that instead of having the convergence for almost every point, this time the convergence is true for *every* starting point in $\mathbb{S}$

To have a convergence rate estimation, we shall need stronger assumptions such as the small set condition, see [94] for example.

Convergence of marginal distribution

Instead of using the entire path of Markov chain, sometimes we may be just interested in its marginal distribution. With the previous results in the mind, it's not surprising to discover that the marginal distribution of the Markov chain will also converge to its stationary distribution, under some conditions. In [98], we can find results in this direction using coupling renewal process. A shorter proof is provided more recently in [76] on the conditions for exponential convergence rate. The marginal convergence of Markov chain will enable us to make approximative simulation of rare event scenarios, which is useful for applications such as financial stress test.

Assumption 4.1.1. *There exists a positive function V and constants $K \geqslant 0$ and $\gamma \in (0,1)$ such that for any $x \in \mathbb{S}$*

$$PV(x) \leqslant \gamma V(x) + K$$

Assumption 4.1.2. *There exists a constant $\alpha \in (0,1)$ and a probability measure ν so that*

$$\inf_{x \in C} P(x, \cdot) \geqslant \alpha \nu(\cdot)$$

with $C = V^{-1}([0,R])$ for some $R > \frac{2K}{1-\gamma}$ where K and γ are the constants from the above assumption

Theorem 4.1.6 ([76]). *If Assumption 4.1.1 and 4.1.2 hold, then P admits a unique invariant measure μ. Furthermore, there exists constant $C > 0$ such that*

$$\|P^n \phi - \mu(\phi)\| \leqslant C\gamma^n \|\phi - \mu(\phi)\|$$

for every bounded measurable function ϕ, where $\|\phi\| = \sup_x \frac{\phi(x)}{1+V(x)}$

Convergence of empirical quantile

What we saw above are the convergence of the occupation measure and the marginal distribution of Markov chain. What will happen if we use the Markov chain realization to estimate the quantile of its stationary distribution? Such a study has been conducted in [46], where one can find convergence rates under different assumptions. We give one example of their results.

Suppose the Markov chain $(X_n)_{n \geqslant 0}$starts with its stationary distribution π and the transition kernel $P(x, dy)$. Given a measurable function g taking value in $\mathbb{R}$, we define a random variable V by $V = g(W)$ where $W \sim \pi$. Let F_V denote the distribution function of V and for a given $q \in (0,1)$, we define the quantile

$$\xi_q := \inf\{v : F_V(v) \geqslant q\}$$

We suppose that F_V is absolutely continuous and has continuous density function which is strictly positive at the point ξ_q.

In addition, we suppose there exists a constant $\alpha > 0$ and a probability measure ν and an integer $n_0 \geqslant 1$ such that for any $x \in \mathbb{S}$ we have

$$P^{n_0}(x, \cdot) \geqslant \alpha\nu(\cdot)$$

Let $\hat{\xi}_{n,q}$ be the empirical quantile of $\{g(X_0), g(X_1), \cdots g(X_{n-1})\}$, then we have the following theorem

Theorem 4.1.7 ([46]). *Under the above assumptions, we have for any $\epsilon > 0$ and $\delta \in (0, 1)$*

$$\mathbb{P}\left(|\hat{\xi}_{n,q} - \xi_q| > \epsilon\right) \leqslant 2\exp\big(-\frac{\alpha^2(n\gamma - 2\frac{n_0}{\alpha})}{2nn_0^2}\big) \tag{4.1.4}$$

for $n \geqslant \frac{2n_0}{\alpha\gamma}$ where $\gamma = \min\{F_V(\xi_q + \epsilon) - q, \delta(q - F_V(\xi_q - \epsilon))\}$

More discussions on this will be given later in Section 5.6 when we introduce our adaptive rare event simulation techniques

The ergodicity of Markov chain gives rise to a powerful set of simulation techniques, called MCMC(Markov Chain Monte Carlo) methods. It finds application in various domains, from statistical mechanics to molecular engineering. We will make use of this theory to address our rare event simulation problem.

4.2 Interacting Particle System

The theory of interacting particle system is more recently developed compared to that of Markov chain ergodicity. A systematic account of this theory can be found in the book [42]. We will just present the non-asymptotic convergence result in [30], which will help us to show the convergence rate of rare event simulation techniques in the next chapters. We follow the presentation framework in [30].

Suppose we have a Markov chain X defined by the transition kernel $M_n(x, dy)$ from the state space E to itself and the initial distribution η_0. G_n and f are bounded function defined on E. We define measures γ_n and η_n in the following way, where $0 \leqslant G_n \leqslant 1$:

$$\gamma_n(f) = \mathbb{E}(f(X_n)\prod_{k=0}^{n-1} G_k(X_k)) \tag{4.2.1}$$

$$\eta_n(f) = \frac{\gamma_n(f)}{\gamma_n(1)} \tag{4.2.2}$$

Then the relation between probability measures η_{n+1} and η_n are given in terms of the Boltzmann-Gibbs transformation Ψ_{G_n}:

$$\Psi_{G_n}(\eta_n)(\mathrm{d}x) := \frac{G_n(x)\eta_n(dx)}{\eta_n(G_n)} \tag{4.2.3}$$

$$\eta_{n+1} = \Psi_{G_n}(\eta_n)M_{n+1} \tag{4.2.4}$$

Given the above notations, we can check that, for $\alpha \in [0,1]$ we define a selection-type transition kernel

$$S_{\alpha G_n,\eta_n}(x,dy) = \alpha G_n(x)\delta_x(dy) + (1-\alpha G_n(x))\Psi_{G_n}(\eta_n)(dy) \tag{4.2.5}$$

and the McKean transition

$$K^{\alpha}_{n+1,\eta_n} = S_{\alpha G_n,\eta_n}M_{n+1}$$

then we have

$$\eta_{n+1} = \eta_n K^{\alpha}_{n+1,\eta_n} \tag{4.2.6}$$

4.2.1 Selection-Mutation simulation

Our aim is to make approximative simulation of the probability measure η_n. To do that, we shall simulate a large particle system evolving in interaction according to Equation 4.2.5 and 4.2.6.

Let's fix the system size as N, i.e. there will be constantly N particles inside the system. At first, we will simulate N i.i.d. particles $X_0^{(N)} = (X_0^{(N,i)})_{1\leqslant i\leqslant N}$ according to the initial distribution η_0. Then suppose we already have a particle system $X_n^{(N)} = (X_n^{(N,i)})_{1\leqslant i\leqslant N}$ approximating η_n, we will apply the following procedure to get the particle system $X_{n+1}^{(N)}$ at generation $n+1$:

- Selection step: for each particle $X_n^{(N,i)}$ in the system $X_n^{(N)}$, apply the selection transition kernel $S_{\alpha G_n,\eta_n^N}$. That is, with probability $\alpha G_n(x)$ the particle remains the same, otherwise, it will be replaced by another particle in the system selected according to the weight function G_n. Thus we have

  $$\hat{X}_n^{(N,i)} \sim S_{\alpha G_n,\eta_n^N}, i = 1,2,\cdots,N \tag{4.2.7}$$

  where η_n^N is the empirical measure represented by the system $X_n^{(N)}$, i.e.

  $$\eta_n^N = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_n^{(N,i)}}$$

- Mutation step: for each selected particle $\hat{X}_n^{(N,i)}$, apply the mutation kernel M_{n+1} to get $X_{n+1}^{(N,i)}$. Thus we have

$$X_{n+1}^{(N,i)} \sim M_{n+1}(\hat{X}_n^{(N,i)}, \mathrm{d}x), i = 1, 2, \cdots, N \tag{4.2.8}$$

At the end we get an approximation of η_{n+1} by

$$\eta_{n+1}^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_{n+1}^{(N,i)}}$$

The above selection-mutation procedure imitates the natural gene evolution process. That's why this method is also called genetic algorithm sometimes.

To implement this procedure, we need to have $\eta_n^N(G_n) > 0$. This is not always guaranteed, so we need to define the stopping time

$$\tau^N = \inf\{n \geqslant 0 : \eta_n^N(G_n) = 0\}$$

For each $n > \tau^N$, we will simply write $\eta_n^N = 0$ by convention.

Using the identity

$$\gamma_n(1) = \prod_{p=0}^{n-1} \eta_p(G_p)$$

we can also get an estimation for $\gamma_n(1)$ by

$$\gamma_n^N(1) = \prod_{p=0}^{n-1} \eta_p^N(G_p)$$

4.2.2 Non-asymptotic estimation

To present the non-asymptotic error estimation in [30], we need to introduce some technical notations.

Let $A_n = G_n^{-1}((0, +\infty))$. For each $x \in A_n$, define

$$\hat{G}_n(x) := M_{n+1}(G_{n+1})(x)$$

and for each $x \in A_{n-1}$, define

$$\hat{M}_n(x, dy) := \frac{M_n(x, dy)G_n(y)}{M_n(G_n)(x)}$$

Suppose for each n, there exist finite constants $\tilde{\delta}_n$ and $\hat{\delta}_n$ such that

$$\sup_{(x,y)\in A_n^2} \frac{G_n(x)}{G_n(y)} \leqslant \tilde{\delta}_n$$

and

$$\sup_{(x,y)\in A_n^2} \frac{\hat{G}_n(x)}{\hat{G}_n(y)} \leqslant \hat{\delta}_n$$

We define another constant $\hat{\delta}_p^{(m)}$ by

$$\hat{\delta}_p^{(m)} := \prod_{q=p}^{p+m-1} \hat{\delta}_q$$

Suppose in addition that there exists some integer m greater than 1 and a sequence of numbers $\hat{\beta}_p^{(m)} \in [1, +\infty)$ such that for any $p \geqslant 0$ and any $(x, x') \in A_p^2$ we have

$$\hat{M}_{p,p+m}(x, dy) \leqslant \hat{\beta}_p^{(m)} \hat{M}_{p,p+m}(x', dy)$$

with

$$\hat{M}_{p,p+m} := \hat{M}_{p+1}\hat{M}_{p+2}\cdots\hat{M}_{p+m}$$

With all the above assumptions and technical notations, we have the following theorem

Theorem 4.2.1 ([30]). *For $N > \sum_{s=0}^{n} \frac{\tilde{\delta}_s \hat{\delta}_s^{(m)} \hat{\beta}_s^{(m)}}{\eta_s(A_s)}$, we have*

$$\mathbb{E}\left(\left(\frac{\gamma_n^N(1)}{\gamma_n(1)} - 1 \right)^2 \right) \leqslant \frac{4}{N} \sum_{s=0}^{n} \frac{\tilde{\delta}_s \hat{\delta}_s^{(m)} \hat{\beta}_s^{(m)}}{\eta_s(A_s)} \tag{4.2.9}$$

More recently, adaptive version of IPS method has been analysis in [33].

Chapter 5

IPS & POP with path shaking transformations

5.1 Problem formulation

Let's recall our problem to solve. Our state space is a measurable space $(\mathbb{S}, \mathcal{S})$, where $(\mathbb{S}, d_{\mathbb{S}})$ is a metric space and $\mathcal{S}$ is the Borel sigma-field generated by its open sets. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we consider a random variable (measurable mapping) $X : \Omega \mapsto \mathbb{S}$ and a measurable set $A \subsetneq \mathbb{S}$, then the rare event under investigation is defined by $\{X \in A\}$. We are mainly interested in achieving the follwoing goals

- to estimate the rare event probability $\mathbb{P}(X \in A)$
- to sample from the conditional distribution $X|X \in A$
- to estimate the conditional expectation on rare event $\mathbb{E}(\varphi(X)|X \in A)$, for bounded measurable functions $\varphi : \mathbb{S} \mapsto \mathbb{R}$
- to evaluate the sensitivity of these rare event statistics with respect to model parameters

In the setting of rare event, $\mathbb{P}(X \in A)$ is usually less than 10^{-4}. We assume that $\mathbb{P}(X \in A) > 0$.

We shall always divide the entire space into nested subsets

$$\mathbb{S} := A_0 \supset \cdots \supset A_k \supset \cdots \supset A_n := A, \tag{5.1.1}$$

Since $\mathbb{P}(X \in A) > 0$, in view of the inclusion (5.1.1) we have $\mathbb{P}(X \in A_k) > 0$ for any k, which justifies the decomposition

$$\mathbb{P}(X \in A) = \prod_{k=1}^{n} \mathbb{P}(X \in A_k | X \in A_{k-1}). \tag{5.1.2}$$

Both of the methods we are going to present share the spirit of splitting, in the sens that instead of directly estimating $\mathbb{P}(X \in A)$ we are going to estimate each conditional probability $\mathbb{P}(X \in A_k|X \in A_{k-1})$ separately.

We emphasis that we take a static point view in our problem formulation. So in case that X is a stochastic process or a random graph evolving with time, we don't look into the time dynamic of X. That is,

we never deal with partial realization of X. Only the entire path of X is concerned in our formulation. We shall see in the following how this point view gives rise to new methods.

5.2 Reversible shaking transformation and algorithms

5.2.1 Shaking transformation and invariance of conditional distribution

In this section, the standing assumption is

(K) There is a measurable mapping $K : \mathbb{S} \times \mathbb{Y} \mapsto \mathbb{S}$, where $(\mathbb{Y}, \mathcal{Y})$ is a measurable space, and a $\mathbb{Y}$-valued random variable Y independent of X such that the following identity in distribution holds:

$$(X, K(X, Y)) \stackrel{\mathrm{d}}{=} (K(X, Y), X). \tag{5.2.1}$$

To simplify notation when there is no ambiguity, we simply write $\mathcal{K}(\cdot) := K(\cdot, Y)$. Identity (5.3.15) reads as a time-reversibility condition and it is equivalent to the balance equation used in [29]. This implies that X and $\mathcal{K}(X)$ have the same distribution. In our algorithms, $\mathcal{K}$ will serve to build Markov chains with invariant distributions given by the distribution of X and that of X restricted to A_k (see Definitions 5.2.2 and 5.2.3). The exact form of K and Y is specific to the model at hand, how to construct K in different settings is explained in Section 5.4.

As we shall understand when we go through all the numerical examples in Chapter 6, we expect the random transformation $X \mapsto \mathcal{K}(X)$ to *slightly modify values of* X while preserving its distribution. Only in this way can X move invariantly throughout the rare event zone and explore all the possible configurations. This motivates the label of *shaking transformation*.

Based on $\mathcal{K}(\cdot)$, for each intermediate subset we define a shaking transformation with rejection as follows.

Definition 5.2.1. *Let* $k \in \{0, 1, \cdots, n-1\}$. *Under* **(K)**, *define*

$$M_k^{\mathcal{K}} : \begin{cases} \mathbb{S} \times \mathbb{Y} \to \mathbb{S}, \\ (x, y) \mapsto K(x, y)\mathbf{1}_{K(x,y) \in A_k} + x\mathbf{1}_{K(x,y) \notin A_k}. \end{cases} \tag{5.2.2}$$

We set $\mathcal{M}_k^{\mathcal{K}}(.) := M_k^K(., Y)$ *where* Y *is the generic random variable defined in* **(K)**.

In [29, last equation of p.796 and first equation of p.797], transformations like (5.3.15) and (5.2.2) are used to design an interacting particle

algorithm for rare events related to random variables in $\mathbb{R}^d$. Here we generalize it to the general state space $\mathbb{S}$. Proposition 5.2.1 and Theorem 5.2.2 when $\mathbb{S} = \mathbb{R}^d$ have similar counterparts in [29] whose proofs make use of explicit Markov transition kernels. Here in order to generalize, we follow a different presentation, focusing on a ω-wise transformation rather than one with explicit transition density, which is more adapted to **(K)** and to our general state space setting, especially for infinite dimensional applications.

Proposition 5.2.1. *Let $k \in \{0, 1, \cdots, n-1\}$. The distribution of X conditionally on $\{X \in A_k\}$ is invariant w.r.t. the random transformation $\mathcal{M}_k^{\mathcal{K}}$: i.e. for any bounded measurable $\varphi : \mathbb{S} \to \mathbb{R}$ we have*

$$\mathbb{E}\big(\varphi(\mathcal{M}_k^{\mathcal{K}}(X))|X \in A_k\big) = \mathbb{E}\big(\varphi(X)|X \in A_k\big). \tag{5.2.3}$$

The above equality still holds if $\varphi(x)$ is replaced by $\varphi(x, U)$ where U is a random variable independent of X and Y (defining $\mathcal{M}_k^{\mathcal{K}}$).

Proof. From Definition 5.2.1 and **(K)**, we write that

$$\begin{aligned}
&\mathbb{E}\left(\varphi(\mathcal{M}_k^{\mathcal{K}}(X))\mathbf{1}_{X\in A_k}\right)\\
=&\mathbb{E}\left(\varphi(\mathcal{K}(X))\mathbf{1}_{X\in A_k}\mathbf{1}_{\mathcal{K}(X)\in A_k}\right) + \mathbb{E}\left(\varphi(X)\mathbf{1}_{X\in A_k}\mathbf{1}_{\mathcal{K}(X)\notin A_k}\right)\\
=&\mathbb{E}\left(\varphi(X)\mathbf{1}_{\mathcal{K}(X)\in A_k}\mathbf{1}_{X\in A_k}\right) + \mathbb{E}\left(\varphi(X)\mathbf{1}_{X\in A_k}\mathbf{1}_{\mathcal{K}(X)\notin A_k}\right)\\
=&\mathbb{E}\left(\varphi(X)\mathbf{1}_{X\in A_k}\right).
\end{aligned}$$

The equality (5.2.3) readily follows. The extension to random $\varphi(\cdot, U)$ is similar. □

5.2.2 Application to IPS algorithm

We are now in a position to put the rare event probability estimation problem in the framework of interacting particles system, which evolves according to the following dynamics.

Definition 5.2.2. *We define a $\mathbb{S}$-valued Markov chain $(X_i)_{0\leqslant i\leqslant n-1}$, as follows:*

$$X_0 \overset{\mathrm{d}}{=} X \tag{5.2.4}$$

$$X_i := \mathcal{M}_i^{\mathcal{K}}(X_{i-1}) = M_i^K(X_{i-1}, Y_{i-1}) \quad \text{for } 1 \leqslant i \leqslant n-1, \tag{5.2.5}$$

*where $(Y_i)_{0\leqslant i\leqslant n-2}$ is a sequence of independent copies of Y (defined in **(K)**) and independent of X_0.*

The IPS interpretation will follow from the next result.

Theorem 5.2.2. *Let $k \in \{1, \cdots, n\}$. We have:*

$$\mathbb{P}\left(X \in A_k\right) = \mathbb{E}\left(\prod_{i=0}^{k-1} \mathbf{1}_{A_{i+1}}(X_i)\right). \tag{5.2.6}$$

For any bounded measurable function $\varphi : \mathbb{S} \to \mathbb{R}$ we have

$$\mathbb{E}\left(\varphi(X)|X \in A_k\right) = \frac{\mathbb{E}\left(\varphi(X_{k-1})\prod_{i=0}^{k-1}\mathbf{1}_{A_{i+1}}(X_i)\right)}{\mathbb{E}\left(\prod_{i=0}^{k-1}\mathbf{1}_{A_{i+1}}(X_i)\right)}. \tag{5.2.7}$$

The above formula is still valid if $\varphi(x)$ is replaced by $\varphi(x, U)$ (as in Proposition 5.2.1) where U is a random variable independent of $(X, X_0, Y_0, \ldots, Y_{k-2})$.

Proof. We first establish (5.2.7) by induction on k. We start with $k = 1$: obviously

$$\begin{aligned}\mathbb{E}\left(\varphi(X,U)|X \in A_1\right) =& \frac{\mathbb{E}\left(\varphi(X,U)\mathbf{1}_{A_1}(X)\right)}{\mathbb{P}\left(X \in A_1\right)}\\ =& \frac{\mathbb{E}\left(\varphi(X_0,U)\mathbf{1}_{A_1}(X_0)\right)}{\mathbb{E}\left(\mathbf{1}_{A_1}(X_0)\right)}.\end{aligned}$$

Assume now that (5.2.7) holds for k, any function φ and any random variable U allowed, and let us prove (5.2.7) for $k+1$. By a slight abuse of notation, we still write $\varphi(x, U) = \varphi(x)$, where U is independent of $(X, X_0, Y_0, \ldots, Y_{k-1})$. We have

$$\begin{aligned}&\mathbb{E}\left(\varphi(X_k)\prod_{i=0}^{k}\mathbf{1}_{A_{i+1}}(X_i)\right)\\ =&\mathbb{E}\left(\varphi(\mathcal{M}_k^{\mathcal{K}}(X_{k-1}))\mathbf{1}_{A_{k+1}}(\mathcal{M}_k^{\mathcal{K}}(X_{k-1}))\prod_{i=0}^{k-1}\mathbf{1}_{A_{i+1}}(X_i)\right)\\ =&\mathbb{E}\left(\varphi(\mathcal{M}_k^{\mathcal{K}}(X))\mathbf{1}_{A_{k+1}}(\mathcal{M}_k^{\mathcal{K}}(X))|X \in A_k\right)\mathbb{E}\left(\prod_{i=0}^{k-1}\mathbf{1}_{A_{i+1}}(X_i)\right)\end{aligned} \tag{5.2.8}$$

where we have applied the induction hypothesis. Then Proposition 5.2.1 yields

$$\mathbb{E}\left(\varphi(\mathcal{M}_k^{\mathcal{K}}(X))\mathbf{1}_{A_{k+1}}(\mathcal{M}_k^{\mathcal{K}}(X))|X \in A_k\right) \tag{5.2.9}$$

$$\begin{aligned}=&\mathbb{E}\left(\varphi(X)\mathbf{1}_{A_{k+1}}(X)|X \in A_k\right)\\ =&\mathbb{E}\left(\varphi(X)|X \in A_{k+1}\right)\mathbb{E}\left(\mathbf{1}_{A_{k+1}}(X)|X \in A_k\right)\end{aligned} \tag{5.2.10}$$

where the nested property (5.1.1) of A_k's is used. Another application of Proposition 5.2.1 and of (5.2.8) with $\varphi \equiv 1$ shows that

$$\begin{aligned}\mathbb{E}\left(\mathbf{1}_{A_{k+1}}(X)|X \in A_k\right) &= \mathbb{E}\left(\mathbf{1}_{A_{k+1}}(\mathcal{M}_k^{\mathcal{K}}(X))|X \in A_k\right)\\ &= \frac{\mathbb{E}\left(\prod_{i=0}^{k}\mathbf{1}_{A_{i+1}}(X_i)\right)}{\mathbb{E}\left(\prod_{i=0}^{k-1}\mathbf{1}_{A_{i+1}}(X_i)\right)}.\end{aligned} \tag{5.2.11}$$

Substituting (5.2.11) into (5.2.10) and (5.2.8) gives the equality (5.2.7) for $k+1$. Lastly, the proof of (5.2.6) now follows easily from (5.2.11) and

(5.1.2). □

By the above theorem, the rare event probability is written in form of an unnormalized Feynman-Kac measure for interacting particle systems, as we reviewed in Section 4.2 (see [42, 43] for detailed discussions). This enables the use of numerical algorithms for estimating it. In general, an interacting particle (a.k.a. genetic genealogical) algorithm provides a way to estimate

$$\mathbb{E}\left(f(X_0,\dots,X_n)\prod_{i=0}^{n-1}G_i(X_i)\right)$$

where f and G_i are bounded and $(X_i)_{0\leqslant i\leqslant n}$ is a Markov chain. In view of (5.2.6) with $k=n$ the rare event probability corresponds to $f\equiv 1$ and $G_i(\cdot)=\mathbf{1}_{A_{i+1}}(\cdot)$ and the corresponding Markov chain is defined in Definition 5.2.2.

The detailed description of interacting particle algorithms can be found in [43, 30] (see also [29] for $\mathbb{S}=\mathbb{R}^d$). The adaptation to our rare event problem in a general state space $\mathbb{S}$ is made without difficulty. As in [30], we introduce an extra rejection parameter $\alpha\in[0,1]$ which increases the independent resampling effect. Note that in [43, 29], $\alpha=1$.

The algorithm below generates at each time $i\in\{0,\dots,n-1\}$ a sample of M elements in $\mathbb{S}$, whose empirical measure approximates the distribution of X conditionally on $\{X\in A_i\}$. We denote by $(Y_i^{(m)}:1\leqslant m\leqslant M, 0\leqslant i\leqslant n-2)$ (resp. $(U_i^{(m)}:1\leqslant m\leqslant M, 0\leqslant i\leqslant n-2)$) a sequence of independent copies of Y from Assumption **(K)** (resp. of a uniformly distributed random variable on $[0,1]$).

Initialization:
Draw $(X_0^{(M,m)}, m = 1, \cdots, M)$ which are M independent copies of X ;
$p_0^{(M)} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}_{A_1}(X_0^{(M,m)})$;
for $i = 0$ until $n - 2$ **do**
  $I_i = \{m \in \{1, \ldots, M\} \text{ s.t. } X_i^{(M,m)} \in A_{i+1}\}$;
  for $m = 1$ until M **do**
    Selection step:
    if $U_i^{(m)} < \alpha$ and $X_i^{(M,m)} \in A_{i+1}$ **then**
      $\hat{X}_i^{(M,m)} = X_i^{(M,m)}$;
    else
      $\hat{X}_i^{(M,m)} = X_i^{(M,\hat{m})}$ where $\hat{m}$ is drawn independently of everything else and uniformly in the set I_i;
    end
    Mutation step:
    $X_{i+1}^{(M,m)} = M_{i+1}^K(\hat{X}_i^{(M,m)}, Y_i^{(m)})$;
  end
  $p_{i+1}^{(M)} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}_{A_{i+2}}(X_{i+1}^{(M,m)})$;
end
Result: $p^{(M)} = \prod_{i=0}^{n-1} p_i^{(M)}$

Algorithm 1: Interacting Particle System algorithm

In the case $\alpha = 1$, the above algorithm takes the same form as that in [29, Section 2] for random variables in $\mathbb{R}^d$. The difference in our work lies in the general state space for which Feynman-Kac formula (Theorem 5.2.2) are nevertheless valid due to the assumption **(K)**: once obtained these formulas, deriving the above IPS algorithm follows a standard routine. How to apply this algorithm to stochastic processes is explained in Sections 5.4 and 6. The convergence properties of this algorithm are postponed to Subsection 5.2.4.

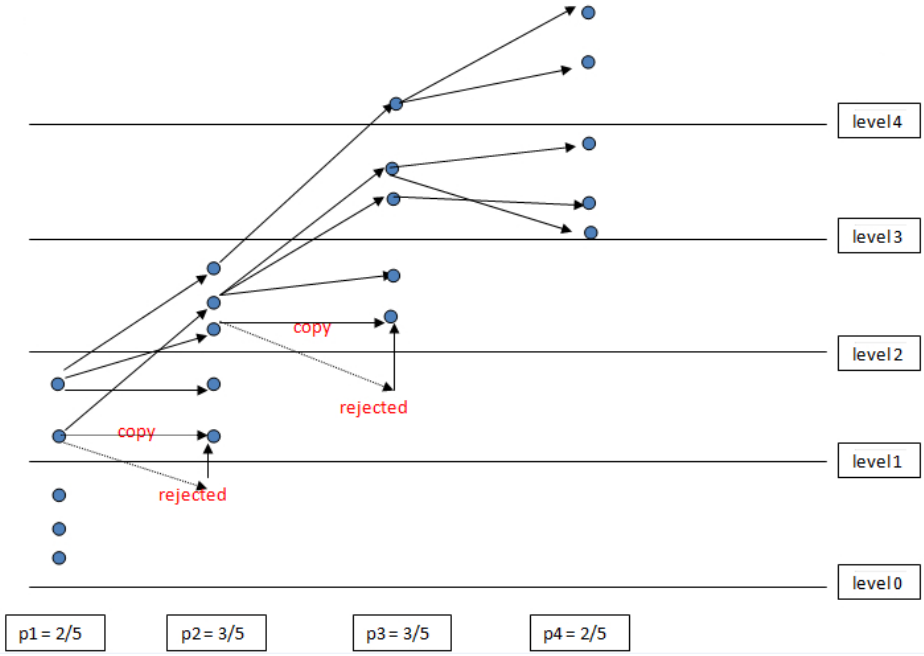


FIGURE 5.1: IPS method illustration, computing probability that X as a point lies above level 4

5.2.3 Application to POP algorithm

In Proposition 5.2.1, we have seen that the distribution of X conditionally on $\{X \in A_k\}$ is invariant with respect to $\mathcal{M}_k^{\mathcal{K}}$. This property allows us to put the problem of computing $\mathbb{P}(X \in A_{k+1} | X \in A_k)$ or $\mathbb{E}(\varphi(X) | X \in A_k)$ in the ergodic Markov chain setting and therefore to compute $\mathbb{P}(X \in A)$ as a consequence of (5.1.2). Before entering into details, we recall that provided $(Z_i)_{i \geqslant 0}$ is a Markov chain on a measurable space with a unique invariant distribution π, the ergodic theorems for Markov chains that we reviewed in Section 4.1 gives

$$\frac{1}{N} \sum_{i=0}^{N-1} f(Z_i) \underset{N \to +\infty}{\longrightarrow} \int f \mathrm{d}\pi \qquad a.s. \tag{5.2.12}$$

for π-a.e. starting point Z_0. Here f is a bounded (or π-integrable) measurable function. See also [98, Chapter 17] or [47, Chapter 7].

For each k we define a Markov chain as follows.

Definition 5.2.3. *For each $k = 0, \ldots, n-1$, given a starting point $X_{k,0}$, define*

$$X_{k,i} := \mathcal{M}_k^{\mathcal{K}}(X_{k,i-1}) = M_k^K(X_{k,i-1}, Y_{k,i-1}) \quad \text{for } i \geqslant 1 \tag{5.2.13}$$

where $(Y_{k,i})_{i \geqslant 0}$ is a sequence of independent copies of Y (defined in **(K)***) and independent of $X_{k,0}$.*

We assume additionally that the sequences $((Y_{k,i})_{i\geqslant 0} : 0 \leqslant k \leqslant n-1)$ are independent. The above process $X_{k,\cdot}$ is a Markov chain in $\mathbb{S}$ and one invariant measure is the distribution of X conditionally on $\{X \in A_k\}$. Then, provided that this is the unique invariant measure, one can use the approximation (as $N \to +\infty$)

$$\mathbb{E}\left(\varphi(X)|X \in A_k\right) \approx \frac{1}{N}\sum_{i=0}^{N-1} \varphi(X_{k,i}), \tag{5.2.14}$$

which for $\varphi \equiv \mathbf{1}_{A_{k+1}}$ yields an approximation of $\mathbb{P}\left(X \in A_{k+1}|X \in A_k\right)$ and therefore of the rare event probability $\mathbb{P}\left(X \in A\right)$[1].

Observe that each conditional probability is computed separately, in parallel for each A_k, on a single path. This gives the reason why we call this method POP for Parallel One-Path. Furthermore, these conditional probabilities can be estimated independently by taking independent initializations (as defined below), i.e., by restarting the initialization from the beginning for each step k with negligible extra time cost since n is usually small. Both the separate and independent evaluations of conditional probabilities are nice properties of POP, and are not shared with other existing algorithms to our knowledge.

The following algorithm evaluating $\mathbb{P}\left(X \in A\right)$ gives a way to automatically initialize each step. Since the initialization is not done with the stationary distribution, in numerical implementation, we could use some burn-in time to reduce its impact.

Initialization: ;
$X_{0,0}$ is a copy of X ;
for $k = 0$ until $n-1$ **do**
  for $i = 1$ until $N-1$ **do**
    $X_{k,i} = M_k^K(X_{k,i-1}, Y_{k,i-1})$;
  end
  $p_k^{(N)} = \frac{1}{N}\sum_{i=0}^{N-1} \mathbf{1}_{A_{k+1}}(X_{k,i})$;
  $i_k = \arg\min\{j : X_{k,j} \in A_{k+1}\}$;
  $X_{k+1,0} = X_{k,i_k}$
end
Result: $p^{(N)} = \prod_{k=0}^{n-1} p_k^{(N)}$

Algorithm 2: Parallel One-Path algorithm

As previously mentioned, the n steps are almost separated, except for initializations, which can also be made independently if we wish. Thus our POP algorithm can be easily parallelized on different processors. For instance, one can use a preliminary run to get all the initial

[1] I realized this application of Markov chain ergodicity on rare event simulation while typing Latex for my supervisor's lecture notes on Markov chain theory, as good surprise often comes at unexpected moments

positions in different subsets/levels. Then all the ergodic time-averages are performed in parallel. We could even use the same copy of Y throughout the different levels to save time used in the generation of random variables Y, even if this will introduce some correlation into conditional probability estimators. This is useful especially when Y is very costly to sample, otherwise it is better to use independent simultions for different levels to achieve better accurary.

Besides, this algorithm can also serve for estimating $\mathbb{E}(\varphi(X)|X \in A)$ using the Markov chain $(X_{n,.})$ and the approximation (5.2.14). This should even be less time-consuming than computing $\mathbb{P}(X \in A)$ since we only need to get a starting point satisfying $X \in A$ and then do POP once at $k = n$ to obtain an empirical distribution of $X|X \in A$.

Lastly, observe that increasing the accuracy of POP algorithm is elementary since it suffices to keep on simulating the n Markov chains $((X_{k,.}) : 0 \leqslant k \leqslant n-1)$ until a larger time horizon N'. This is a significant difference with IPS, for which increasing accuracy implies increasing M and thus re-simulating all the M particles system from the beginning (because of interactions).

The reduction of required memory space with POP method is enormous. With POP only the current particle needs to be stored while with IPS we need to store a large particle system.

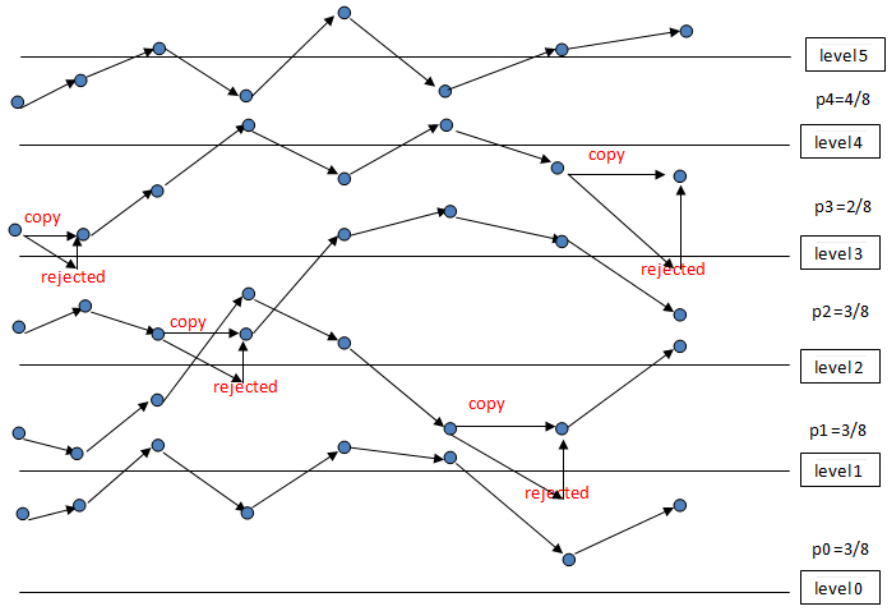


FIGURE 5.2: POP method illustration, computing probability that X as a point lies above level 5

5.2.4 Convergence analysis for both algorithms

Convergence of IPS

Convergence of the IPS algorithm for estimating unnormalized Feynman-Kac measure is well studied in the literature, as the number of particles $M \to +\infty$: under various hypotheses, are proved the law of large number, central limit theorem (at rate $\sqrt{M}$) and non-asymptotic error estimation (fixed M).

Regarding Algorithm 1, it is known that the estimator is unbiased. A non-asymptotic variance control is given in [30], which we reviewed in the last chapter. Since in our algorithm the number of intermediate levels is usually not large, we do not need the assumption $(M)_m$ or $(\hat{M})_m$ in [30] and we can get similar results to Lemma 4.1 and Lemma 4.3 in [30] by only assuming that the following quantity $\hat{\delta}_k$ is finite for each $k = 0, 1, \cdots, n-1$

$$\hat{\delta}_k := \sup_{(x,y)\in A_{k+1}^2} \frac{\mathbb{P}(\mathcal{K}(x) \in A_{k+2}) + \mathbf{1}_{A_{k+2}}(x)\mathbb{P}(\mathcal{K}(x) \notin A_{k+1})}{\mathbb{P}(\mathcal{K}(y) \in A_{k+2}) + \mathbf{1}_{A_{k+2}}(y)\mathbb{P}(\mathcal{K}(y) \notin A_{k+1})} < +\infty \tag{5.2.15}$$

where by convention $A_{n+1} = A_n$. We adapt [30, Corollary 5.2] to our setting.

Theorem 5.2.3. *Under the assumption that all $\hat{\delta}_k$'s are finite, we have the following non-asymptotic control when $M > \sum_{s=0}^{n} \frac{\Delta_s}{\mathbb{P}(X\in A_{s+1}|X\in A_s)}$*

$$\mathbb{E}\left(\left|\frac{p^{(M)}}{p} - 1\right|^2\right) \leqslant \frac{4}{M}\sum_{s=0}^{n} \frac{\Delta_s}{\mathbb{P}\left(X \in A_{s+1} | X \in A_s\right)} \tag{5.2.16}$$

where $\Delta_s = \prod_{k=s}^{n-1} \hat{\delta}_k$ and by convention $\Delta_n := 1$.

Proof. We very closely follow the proof of [30]. The first adjustment comes from a slightly annoying shift of index. Indeed, with notations in the last reference, $G_k(x) = 1_{A_k}(x)$, while with our notations we have $G_k(x) = 1_{A_{k+1}}(x)$. Therefore, to fit as easily as possible with arguments in [30], we set $A_{n+1} = A_n$ so that the last term in the sum becomes $\frac{\Delta_n}{\mathbb{P}(X\in A_{n+1}|X\in A_n)} = 1$. The second adjustment in our setting is that we avoid their assumptions $(M)_m$ or $(\hat{M})_m$ whose role is partly to get better estimates as n is large. We thus just emphasize how to get rid of these assumptions in our work. Firstly, we easily check that $\hat{\delta}_k$ defined in their assumption $(\hat{H})_m$ is the one given in our theorem. Secondly with this estimate at hand, we can prove that (using notation of their Equation (4.3))

$$\sup_{x,y\in A_{k+1}^2} \frac{\hat{Q}_{k,n}(1)(x)}{\hat{Q}_{k,n}(1)(y)} \leqslant \Delta_k.$$

Therefore, the upper bound on the r.h.s. of (4.5) in their Lemma 4.3 becomes Δ_k (by noticing that their $\tilde{\delta}_k = 1$). Lastly, the rest of the proof is similar in that the above estimate propagates to their Corollary 5.2 in the form of our inequality (5.2.16). □

The upper bound of Theorem 5.2.3 is useful to appropriately choose the shaking transformation in order to make the error smaller. Firstly, obviously we have $\hat{\delta}_k \leqslant \sup_{y\in A_{k+1}} \frac{1}{\mathbb{P}(\mathcal{K}(y)\in A_{k+2})}$. In case of slight shaking, $\mathcal{K}(y)$ will differ little from y so the probability of going from A_{k+1} to A_{k+2} is small and $\hat{\delta}_k$ is large. Conversely, in case of strong shaking and since A_{k+2} is expected to be small, $\mathcal{K}(y)$ will be very likely to exit A_{k+2}, resulting in a large $\hat{\delta}_k$. Hence, choosing an intermediate shaking force is presumably the best choice, see later numerical experiments.

Finally, the upper bound (5.2.16) is rather qualitative and seemingly can not be quantitatively computed in general. More general cases are left to further research.

Convergence of POP

Equation (5.2.12) with its assumptions gives

$$\frac{1}{N}\sum_{i=0}^{N-1} \mathbf{1}_{A_{k+1}}(X_{k,i}) \underset{N\to+\infty}{\longrightarrow} \mathbb{P}(X \in A_{k+1}|X \in A_k) \qquad a.s. \tag{5.2.17}$$

for a.e. starting point $X_{k,0}$.

The convergence of ergodic theorem has been much studied in the literature, with results like almost sure convergence, asymptotic and non-asymptotic fluctuations, see for instance [98, 47]. Here we apply the recent work [94, Theorem 3.1] in our rare event setting.

Theorem 5.2.4. *For each k in $\{0, , \cdots, n-1\}$, assume that $(A_k, d_{\mathbb{S}})$ is a Polish space and that the Markov chain $(X_{k,i})_{i\geqslant 0}$ is π_k-irreducible and Harris recurrent, where π_k is the distribution of X conditionally on $\{X \in A_k\}$. If in addition the "small set" condition holds: "there exists a Borel set $F_k \subset A_k$ of positive π_k measure, a positive number $\beta_k > 0$ and a probability measure ν_k such that $P_k(x,\cdot) \geqslant \beta_k\nu_k(\cdot), \forall x \in F_k$" where $P_k(.,.)$ is the transition kernel of $X_{k,.}$, then there exists a constant C_k depending on the model such that*

$$\mathbb{E}\left(\left(p_k^{(N)} - \mathbb{P}(X \in A_{k+1}|X \in A_k)\right)^2\right) \leqslant \frac{C_k}{N}.$$

For application in our rare event examples, the "Polish assumption" is usually satisfied when we consider the space of continuous functions $\mathbb{C}([0,T],\mathbb{R}^d)$ $(T > 0)$ with the uniform convergence topology (example of Subsection 6.1), or the space of càdlàg functions $\mathbb{D}([0,T],\mathbb{R}^d)$ with the Skorohod topology (jump processes in insurance, queuing system and Hawkes process in Subsections 6.2-6.3-6.5) or $\mathbb{R}^{\mathbb{N}}$, see [17] for details. The random graph example in Subsection 6.4 is associated to a finite space and the "Polish assumption" is thus trivial.

But verification of the small set condition is more difficult. In the random graph example, this condition is satisfied obviously since the state space is finite. In general, extra work is still needed to verify the small set condition in each particular example.

Finally recall that our estimated rare event probability $p^{(N)}$ is the product of all $p_k^{(N)}$'s. Since we have already error control for each $p_k^{(N)}$ and since these quantities are bounded, by easy computations we can establish that the convergence rate is also $\sqrt{N}$ for the estimation of rare event probability.

5.3 Gaussian shaking and its properties

5.3.1 Gaussian variable, process and SDE driven by Brownian motion

For a standard Gaussian variable $X := G$ in $\mathbb{R}^d$, a simple shaking transformation is

$$K(G, G') = \rho G + \sqrt{1 - \rho^2} G'$$

with $\rho \in (-1, 1)$ and $Y := G'$ is a independent copy of G. Figure 5.3 gathers two graphs of 100000 independent simulations of $(G, \mathcal{K}(G))$ with their respective marginal histograms (of course close to the Gaussian distribution). The larger the value of ρ, the slighter the shaking, the closer the points to the diagonal.

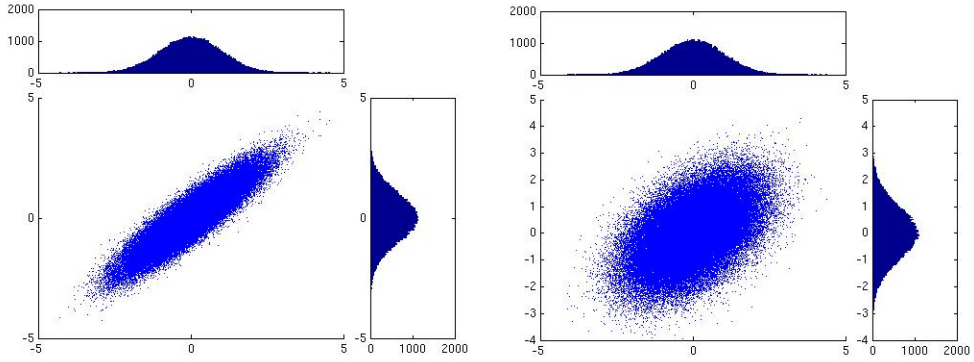


FIGURE 5.3: Shaking Gaussian variables in dimension 1, with $\rho = 0.9$ (left) and $\rho = 0.5$ (right)

The same linear transformation works for Gaussian processes $X := (G_t)_{0 \leqslant t \leqslant T}$ ($T > 0$ fixed), with zero mean and any covariance function. In the case where X is a d-dimensional Brownian motion, one can take slightly more general transformation based on Wiener integrals: namely, for the i-th component of $\mathcal{K}(G)$, take a measurable function $\rho_i : [0, T] \mapsto$

$[-1,1]^d$ with $|\rho_{i,t}| \leqslant 1$ and set

$$\mathcal{K}_i(G) = \left(\int_0^t \rho_{i,s} \cdot \mathrm{d}G_s + \int_0^t \sqrt{1 - |\rho_{i,s}|^2} \mathrm{d}G'_{i,s} \right)_{0 \leqslant t \leqslant T} \tag{5.3.1}$$

where $G' = (G'_1, \dots, G'_d)$ is another independent Brownian motion in $\mathbb{R}^d$. Provided that the matrix $\rho_t = (\rho_{1,t}, \dots, \rho_{d,t})$ is symmetric and $\rho_i \cdot \rho_j \equiv 0$ for all $i \neq j$, this transformation satisfies **(K)**.

With this tool at hand, it is then straightforward to define reversible shaking transformations of solution to a stochastic differential equation of the form

$$\mathrm{d}Z_t = b(t, Z_t)\mathrm{d}t + \sigma(t, Z_t)\mathrm{d}G_t, \quad Z_0 \text{ independent of } G, \tag{5.3.2}$$

where coefficients b and σ fulfill appropriate smoothness and growth conditions in order to have a unique strong solution [113]. Setting $X = (Z_t)_{0 \leqslant t \leqslant T}$ it suffices to define $\mathcal{K}(X)$ as the (strong) solution of

$$\mathrm{d}Z'_t = b(t, Z'_t)\mathrm{d}t + \sigma(t, Z'_t)\mathrm{d}K(G, G')_t, Z'_0 = Z_0$$

where the shaken Brownian motion $K(G, G') = \mathcal{K}(G)$ is defined in (5.3.1): this procedure satisfies **(K)**. This will be applied to the example of Ornstein-Uhlenbeck process in Subsection 6.1. Observe that this method can be directly extended to non-Markovian equations driven by Brownian motion.

5.3.2 Hermite polynomials and one dimensional convergence

▷ **Hermite polynomials** Recall that the j-th Hermite polynomial is defined by

$$\begin{aligned} H_0 &:= 1 \\ H_j(x) &:= \frac{(-1)^j}{j!} e^{\frac{x^2}{2}} \frac{\mathrm{d}^j}{\mathrm{d}x^j}(e^{-\frac{x^2}{2}}), j \geqslant 1 \end{aligned}$$

and it satisfies the following properties:

$$\begin{aligned} H'_j(x) &= H_{j-1}(x) \\ (j+1)H_{j+1}(x) &= xH_j(x) - H_{j-1}(x). \end{aligned}$$

Lemma 5.3.1. *For $\rho \in [-1,1]$, we have*

$$\mathbb{E}\left(H_n(\rho x + \sqrt{1-\rho^2}\mathcal{N}(0,1)) \right) = \rho^n H_n(x)$$

Proof. Obviously, this is true for $n = 0$ and 1. Now suppose the conclusion is true for $1, 2, \cdots, n$, we are going to prove for $n + 1$.

$$\begin{aligned}
&(n+1)\mathbb{E}\left(H_{n+1}(\rho x + \sqrt{1-\rho^2}\mathcal{N}(0,1))\right)\\
=&\mathbb{E}\left((\rho x + \sqrt{1-\rho^2}\mathcal{N}(0,1))H_n(\rho x + \sqrt{1-\rho^2}\mathcal{N}(0,1))\right)\\
&-\mathbb{E}\left(H_{n-1}(\rho x + \sqrt{1-\rho^2}\mathcal{N}(0,1))\right)\\
=&\rho x \rho^n H_n(x) + \mathbb{E}\left((\sqrt{1-\rho^2}\mathcal{N}(0,1))H_n(\rho x + \sqrt{1-\rho^2}\mathcal{N}(0,1))\right)\\
&-\rho^{n-1}H_{n-1}(x)
\end{aligned}$$

It is easy to prove that for h with polynomial growth, we have

$$\mathbb{E}\left(h'(\mathcal{N}(0,1))\right) = \mathbb{E}\left(\mathcal{N}(0,1)h(\mathcal{N}(0,1))\right)$$

Take $h(\cdot) = H_n(\rho x + \sqrt{1-\rho^2}\cdot)$, we have

$$\begin{aligned}
&\mathbb{E}\left((\sqrt{1-\rho^2}\mathcal{N}(0,1))H_n(\rho x + \sqrt{1-\rho^2}\mathcal{N}(0,1))\right)\\
=&(1-\rho^2)\mathbb{E}\left(H_n'((\rho x + \sqrt{1-\rho^2}\mathcal{N}(0,1))\right)\\
=&(1-\rho^2)\rho^{n-1}H_{n-1}(x)
\end{aligned}$$

which enables us to conclude that

$$\begin{aligned}
&(n+1)\mathbb{E}\left(H_{n+1}(\rho x + \sqrt{1-\rho^2}\mathcal{N}(0,1))\right)\\
=&\rho^{n+1}xH_n(x)\rho^{n+1}H_{n-1}(x)\\
=&\rho^{n+1}(n+1)H_{n+1}(x)
\end{aligned}$$

□

▷ Convergence of shaking in dimension one

Theorem 5.3.1. *Assume $Z = f(X)$ is square integrable where $X \sim \mathcal{N}(0,1)$ and f is a given function. If $|\rho| < 1$ and we define*

$$X_1 \sim \mathcal{N}(0,1), X_i = \rho X_{i-1} + \sqrt{1-\rho^2}\mathcal{N}(0,1)$$

and $Z_i = f(X_i)$, then we have

$$\mathbb{E}\left(\left|\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbb{E}\left(Z\right)\right|^2\right) \leqslant \frac{\mathbb{V}\mathrm{ar}\left(Z\right)}{N}\left(\frac{1+|\rho|}{1-|\rho|}\right).$$

Proof. We define $T_\rho f(X) = \mathbb{E}\left(f(\rho X + \sqrt{1-\rho^2}\mathcal{N}(0,1))\right)$ where the expectation is taken only on $\mathcal{N}(0,1)$.

For any square integrable $f : \mathbb{R} \to \mathbb{R}$, since Hermite polynomials form a complete orthonormal system in $L^2(X)$ there exists $a_n \in \mathbb{R}, n \geqslant 1$

such that $f(X) = \mathbb{E}(f(X)) + \sum_{n\geqslant 1} a_n H_n(X)$ in $L^2(X)$. Then the above lemma gives

$$T_\rho(f(X)) = \mathbb{E}(f(X)) + \sum_{n\geqslant 1} \rho^n a_n H_n(X).$$

and we obtain

$$|T_\rho(f(X)) - \mathbb{E}(f(X))|_2^2 \leqslant \sum_{n\geqslant 1} \rho^{2n} \mathbb{E}\left(a_n^2 H_n^2(f(X))\right) \leqslant |\rho|^2 \mathbb{V}\mathrm{ar}(f(X)). \tag{5.3.3}$$

Denote $\mathbb{E}\left(\left|\frac{1}{N}\sum_{i=1}^N Z_i - \mathbb{E}(Z)\right|^2\right)$ by e_N . We have

$$e_N = \frac{1}{N^2}\left[\sum_{1\leqslant k\leqslant N} \mathbb{V}\mathrm{ar}(Z_k) + 2\sum_{1\leqslant k<l\leqslant N} \mathbb{C}\mathrm{ov}(Z_k, Z_l)\right].$$

By the reversibility property, Z_k and Z have the same law. Thus,

$$\sum_{k=1}^N \mathbb{V}\mathrm{ar}(Z_k)) = N\mathbb{V}\mathrm{ar}(Z)$$

On the other hand, for $l > k$, we have by a conditioning argument

$$\begin{aligned}\mathbb{C}\mathrm{ov}(Z_k, Z_l) &:= \mathbb{E}\left(\left(Z_k - \mathbb{E}(Z)\right)\left(Z_l - \mathbb{E}(Z)\right)\right)\\ &= \mathbb{E}\left(\left(Z_k - \mathbb{E}(Z)\right)\left(T_{\rho^{l-k}}(Z_k) - \mathbb{E}(Z)\right)\right)\\ &\leqslant |\rho|^{l-k}\mathbb{V}\mathrm{ar}(Z)\end{aligned}$$

where we have used (5.3.3) in the last line. Consequently, we have

$$\left|\sum_{1\leqslant k<l\leqslant N} \mathbb{C}\mathrm{ov}(f(Z_k), f(Z_l))\right| \leqslant N\frac{|\rho|}{1-|\rho|}\mathbb{V}\mathrm{ar}(f(Z)).$$

Finally, we get

$$e_N \leqslant \frac{\mathbb{V}\mathrm{ar}(f(Z))}{N}\left[1 + 2\frac{|\rho|}{1-|\rho|}\right]$$

which completes the proof. □

5.3.3 Convergence of general Gaussian shaking

For many applications, we will deal with infinite dimensional Gaussian variables, such as stochastic processes or random fields. To apply our methods in these applications, we will define shaking transformation on the underlying Gaussian random source rather than the final random output. The following paragraphs are mainly to give a strict description

on how shaking transformation is applied in these situations. At the end we give a general convergence result.

We adopt the framework in [104] of an isonormal Gaussian process associated with a general Hilbert space $\mathcal{H}$ (a.k.a. the framework of Gaussian Hilbert spaces, see [83]). Namely, we assume that $\mathcal{H}$ is a real separable Hilbert space with scalar product $\langle .,. \rangle_{\mathcal{H}}$ and we consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the stochastic process $X = (X(h) : h \in \mathcal{H})$ is a centered Gaussian family of scalar random variables with

$$\mathbb{E}\left(X(h)X(g)\right) = \langle h, g \rangle_{\mathcal{H}} \quad \text{for any } h, g \in \mathcal{H}.$$

We may refer to X as a path indexed by $h \in \mathcal{H}$. The norm of an element $h \in \mathcal{H}$ is denoted by $\|h\|_{\mathcal{H}}$. The mapping $h \mapsto X(h)$ is linear. Some important examples are as following:

Example 5.3.1 (Finite dimensional Gaussian space)**.** *Let $q \in \mathbb{N}^*$, set $\mathcal{H} := \mathbb{R}^q$ and $\langle h, g \rangle_{\mathcal{H}} := \sum_{j=1}^q h_j g_j$ for any $h, g \in \mathbb{R}^q$, denote by $e^i = (\mathbf{1}_{\{j=i\}} : 1 \leqslant j \leqslant q)$ the i-th element of the canonical basis of $\mathbb{R}^q$. Then $(X(e^1), \ldots, X(e^q))$ is a vector with independent standard Gaussian components.*

Example 5.3.2 (Multidimensional Brownian motion (BM))**.** *Let $q \in \mathbb{N}^*$, denote by $\mathcal{H}$ the $_2$-space $\mathcal{H} := {}_2(\mathbb{R}^+ \times \{1, \ldots, q\}, \mu)$, where the measure μ is the product of the Lebesgue measure times the uniform measure which gives mass one to each point $1, \ldots, q$, and set*

$$\langle h, g \rangle_{\mathcal{H}} := \int_{\mathbb{R}^+ \times \{1,\ldots,q\}} h(x)g(x)\mu(\mathrm{d}x) \quad \textit{for any } h, g \in \mathcal{H}.$$

Define

$$X_t^i := X(\mathbf{1}_{[0,t]\times\{i\}}) \quad \textit{for any } t \geqslant 0, \quad 1 \leqslant i \leqslant q.$$

Then the process $(X_t^1, \ldots, X_t^q : t \geqslant 0)$ is a standard q-dimensional Brownian motion.

Example 5.3.3 (Fractional Brownian motion (fBM))**.** *The fBM with Hurst exponent $H \in (0, 1)$ is a $\mathbb{R}$-valued Gaussian process, centered with covariance function*

$$\mathbb{E}(X_t^{(H)} X_s^{(H)}) = \frac{1}{2}\left(t^{2H} + s^{2H} - |t - s|^{2H}\right) := R_H(t, s), \quad \textit{for any } s, t \geqslant 0.$$

For any fixed $T > 0$, $(X_t^{(H)} : 0 \leqslant t \leqslant T)$ can also be defined within our framework (see [104, Chapter V]). Denote by $\mathcal{H}^0$ the set of step functions on $[0, T]$, and let $\mathcal{H}$ be the Hilbert space defined as the closure of $\mathcal{H}^0$ w.r.t. the scalar product $\langle \mathbf{1}_{[0,t]}, \mathbf{1}_{[0,s]} \rangle_{\mathcal{H}} = R_H(t, s)$. Denote by X the Gaussian process on $\mathcal{H}$ and $(X(\mathbf{1}_{[0,t]}) : 0 \leqslant t \leqslant T)$ defines a fBm $(X_t^{(H)} : 0 \leqslant t \leqslant T)$ with Hurst exponent H.

Of course, we can mix these examples by defining for instance simultaneously standard BM and fBM. On top of this Gaussian model on $\mathcal{H}$, we can define more sophisticated models frequently used in finance for modeling risk. For the sake of convenience of the reader, here we mention two of them.

- Local volatility models [49]:

$$\mathrm{d}S_t = b(t, S_t)\mathrm{d}t + \sigma(t, S_t)\mathrm{d}X_t \tag{5.3.4}$$

  where X is a standard q-dimensional BM and S stands for the price process of d tradable assets.

- Fractional Brownian Motion (fBM) volatility models [38, 59]:

$$\mathrm{d}S_t = \mu_t S_t \mathrm{d}t + \sigma_t S_t \mathrm{d}W_t$$

  where S stands for the price and the random volatility σ_t is defined through a fractional Brownian Motion. To have mean-reverting volatility, we may model σ as a fractional Ornstein-Uhlenbeck process, see [37, 73]. For example, consider the fractional SABR model of [58] where the volatility takes the form

$$\sigma_t = \bar{\sigma} \exp\left(-\frac{1}{2}\alpha^2 t^{2H} + \alpha X_t^{(H)}\right),\ t \in [0, T] \tag{5.3.5}$$

  where $\bar{\sigma}$ and α are positive parameters, and $X^{(H)}$ as in Example 5.3.3.

From now on, we assume that the probability space at hand $(\Omega, \mathcal{F}, \mathbb{P})$ is such that the σ-field $\mathcal{F}$ is generated by $\{X(h) : h \in \mathcal{H}\}$ and for notational simplification, we often identity $\mathcal{H}$ with its orthonormal basis $\mathfrak{b}\mathcal{H} = (\bar{h}_1, \bar{h}_2, \dots)$. To allow rather great generality, we assume that the rare event is defined through two components, some Rare-event Explanatory Variables (REV) and a level-set function, which are parametrized as follows:

a) We consider a random variable taking values in a general metric space $(\mathbf{Z}, \mathcal{Z})$, i.e.

$$Z : \omega \in (\Omega, \mathcal{F}) \mapsto Z(\omega) := \Psi_Z(X(\omega)) \in (\mathbf{Z}, \mathcal{Z}) \tag{5.3.6}$$

   where Ψ_Z is a measurable mapping from $\mathbb{R}^{\mathcal{H}}$ to $\mathbf{Z}$. The random variable Z stands for the REV whose aim is to model the stochasticity of the rare-event.

b) The above REV will be evaluated along a level-set function φ, which completes the definition of the rare event:

$$\varphi : (z, a) \in \mathbf{Z} \times (-\infty, +\infty] \mapsto \varphi(z, a) \in [-\infty, +\infty). \tag{5.3.7}$$

As we will see, non-positive values of $\varphi(z,a)$ correspond to rare-event scenarios whose probabilities we aim to compute. Furthermore, we assume that for any a, $\varphi(.,a)$ is a measurable map in the first component and that $\varphi(.)$ is non-increasing w.r.t. the second variable, i.e.

$$\varphi(z,a) \geqslant \varphi(z,a') \text{ for any } -\infty < a \leqslant a' \leqslant +\infty \text{ and any } z \in \mathbf{Z}. \tag{5.3.8}$$

We take the convention $\varphi(z,+\infty) = -\infty$ for any $z \in \mathbf{Z}$. The property (5.3.8) is crucial for the splitting approach in order to define nested subsets of increasingly rare scenarios (see Equation (5.1.1) later).

c) The rare event under study is described by the critical paths of Z in set A of the form

$$A := \{z \in \mathbf{Z} : \varphi(z,\bar{a}) \leqslant 0\} \tag{5.3.9}$$

for a given level parameter $\bar{a} \in \mathbb{R}$ such that the probability $\mathbb{P}(Z \in A)$ is small.
Attention: This is a slight abuse of notation. Different from Section 5.2, here the rare event zone A is defined in terms of the output variable Z instead of the input variable X.

d) There is an integrable random variable $\Phi : \Omega \mapsto \mathbb{R}$ modeling the output, for which we wish to evaluate the statistics restricted to the event $\{Z \in A\}$, i.e. to compute

$$\mathbb{E}(\Phi \mathbf{1}_{Z \in A}). \tag{5.3.10}$$

Similarly as in Equation (5.1.1), consider level parameters $\bar{a} := a_n < \cdots < a_k < \cdots < a_0 := +\infty$ and set

$$A_k := \{z \in \mathbf{Z} : \varphi(z,a_k) \leqslant 0\} \tag{5.3.11}$$

so that, owing to (5.3.8) we have

$$A := A_n \subset \cdots \subset A_k \subset \ldots A_0 := \mathbf{Z}. \tag{5.3.12}$$

Note that for describing a given rare-event A, there are many possible couples (level set function φ, level set parameter $\bar{a}$). The choice made by the user has an impact on the performance of the methods (see the example on credit-risk in Subsection 6.8). This choice should be made according to the knowledge of the model at hand. Later, we will often refer to $(a_k)_{k=1}^n$ as acceptance level parameters, this terminology is justified by the subsequent Monte-Carlo schemes of Section 5.2. The choice of acceptance levels is discussed later and can be done adaptively (see Section 5.6).

Then above explanation justifies the decompositions

$$\mathbb{E}\left(\Phi \mathbf{1}_{Z\in A}\right) = \mathbb{E}\left(\Phi \mathbf{1}_{Z\in A} \mid Z \in A_{n-1}\right) \prod_{k=1}^{n-1} \mathbb{P}\left(Z \in A_k \mid Z \in A_{k-1}\right) \quad (5.3.13)$$

$$= \mathbb{E}\left(\Phi \mid Z \in A_n\right) \prod_{k=1}^{n} \mathbb{P}\left(Z \in A_k \mid Z \in A_{k-1}\right). \quad (5.3.14)$$

We define the shaking transformation for the infinite dimensional Gaussian variable.

Definition 5.3.1 (Shaker). *Let $\rho := (\rho_h : h \in \mathfrak{b}\mathcal{H}) \in [-1,1]^{\mathfrak{b}\mathcal{H}}$ and define*

$$K : \begin{cases} \mathbb{R}^{\mathcal{H}} \times \mathbb{R}^{\mathcal{H}} & \mapsto \mathbb{R}^{\mathcal{H}} \\ (x, x') := (x_h : h \in \mathfrak{b}\mathcal{H},\ x'_h : h \in \mathfrak{b}\mathcal{H}) & \to (\rho_h x_h + \sqrt{1-\rho_h^2}\, x'_h : h \in \mathfrak{b}\mathcal{H}). \end{cases} \quad (5.3.15)$$

Whenever useful, we will write K_ρ to insist on the dependence on the so-called shaking parameter ρ.

If $X' = (X'(h) : h \in \mathcal{H})$ is an independent copy of X, we simply denote by $\mathcal{K}$ the random transformation from $\mathbb{R}^{\mathcal{H}} \mapsto \mathbb{R}^{\mathcal{H}}$ as

$$\mathcal{K}(x) = K(x, X'). \quad (5.3.16)$$

In the stochastic analysis literature, the above parametrized transformation for a constant parameter $\rho_h = \text{constant} \in (0,1)$ is associated to the Ornstein-Uhlenbeck (or Mehler) semigroup (see [104, Section 1.4]) and simply writes

$$K(x, x') = \rho x + \sqrt{1-\rho^2}\, x', \quad (5.3.17)$$

independently of the choice of the basis $\mathfrak{b}\mathcal{H}$.

We call the transformation (5.3.15) *shaker* and obviously it satisfies the reversibility property in assumption **(K)**

Proposition 5.3.2. *The following identity holds in distribution:*

$$(X, K(X, X')) \overset{\mathrm{d}}{=} (K(X, X'), X).$$

This type of reversibility property is well-known in the Markov Chain Monte-Carlo literature when studying the convergence of Markov chains in large time. Thus, the shaker (5.3.15) preserves the distribution of X (seen now as a stationary measure) and by iterating the transformations and averaging out the outputs in time, we may obtain a numerical evaluation of related expectations (Birkhoff Law of Large Numbers). Actually, we can prove the convergence in $\mathbb{L}_2$ with an explicit error bound. Our proof relies on the generalized Gebelein inequality [60] for the maximal correlation between Gaussian subspaces [83, Chapter 10].

Theorem 5.3.3. *Let $f : \mathbf{Z} \mapsto \mathbb{R}$ be a measurable function and assume that $f(Z) \in {}_2$ where $Z = \Psi_Z(X)$ as in (5.3.6). Define $X_0 = X, X_{k+1} = K_\rho(X_k, X'_k)$*

and $Z_k = \Psi_Z(X_k)$ *where the* X'_k *are independent copies of* X*. Then, for* $|\rho|_\infty := \sup_{h\in \mathfrak{b}\mathcal{H}} |\rho_h| < 1$,

$$\left| \frac{1}{N}\sum_{k=1}^{N} f(Z_k) - \mathbb{E}\left(f(Z)\right) \right|_2^2 \leqslant \frac{\mathbb{V}\mathrm{ar}\left(f(Z)\right)}{N}\left(\frac{1+|\rho|_\infty}{1-|\rho|_\infty}\right), \quad \forall N \geqslant 1. \tag{5.3.18}$$

Proof. Denote by e_N the l.h.s. of the above inequality. We have

$$e_N = \frac{1}{N^2}\left[\sum_{1\leqslant k\leqslant N} \mathbb{V}\mathrm{ar}\left(f(Z_k)\right) + 2\sum_{1\leqslant k<l\leqslant N} \mathbb{C}\mathrm{ov}\left(f(Z_k), f(Z_l)\right)\right].$$

By the reversible shaker property, Z_k and Z have the same law, thus

$$\sum_{k=1}^{N} \mathbb{V}\mathrm{ar}\left(f(Z_k)\right) = N\mathbb{V}\mathrm{ar}\left(f(Z)\right)$$

On the other hand, for $l > k$, we have

$$\begin{aligned} |\mathbb{C}\mathrm{ov}\left(f(Z_k), f(Z_l)\right)| &\leqslant \rho_{X_k,X_l}\sqrt{\mathbb{V}\mathrm{ar}\left(f(Z_k)\right)}\sqrt{\mathbb{V}\mathrm{ar}\left(f(Z_l)\right)} \\ &= \rho_{X_k,X_l}\mathbb{V}\mathrm{ar}\left(f(Z)\right) \end{aligned}$$

where ρ_{X_k,X_l} is the so-called *Renyi maximal correlation coefficient* between X_k and X_l, i.e. the supremum of the correlation between a function g_k of X_k and a function g_l of X_l, the supremum being taken over all functions (g_k, g_l) with squared integrability properties. We claim that

$$\rho_{X_k,X_l} \leqslant |\rho|_\infty^{l-k}. \tag{5.3.19}$$

The proof is provided at the end. With (5.3.19) at hand, we deduce

$$\left| \sum_{1\leqslant k<l\leqslant N} \mathbb{C}\mathrm{ov}\left(f(Z_k), f(Z_l)\right) \right| \leqslant N\frac{|\rho|_\infty}{1-|\rho|_\infty}\mathbb{V}\mathrm{ar}\left(f(Z)\right).$$

Finally, we get

$$e_N \leqslant \frac{\mathbb{V}\mathrm{ar}\left(f(Z)\right)}{N}\left[1 + 2\frac{|\rho|_\infty}{1-|\rho|_\infty}\right],$$

which finishes the proof of (5.3.18).

It remains to justify (5.3.19). This is a consequence of [83, Theorem 10.11]. Indeed, assume without loss of generality that $k = 1$ (for notational convenience). Now define a Gaussian Hilbert space $\mathcal{G}$ for all the variables from shaker iteration $k = 1$ to $l > 1$. For this, set $\mathfrak{H} := \{\mathfrak{h} = (h_1, \ldots, h_l) \in \mathcal{H}^l\}$: endowed with the scalar product $\langle \mathfrak{h}, \mathfrak{g}\rangle_{\mathfrak{H}} = \sum_{i=1}^{l}\langle h_i, g_i\rangle_{\mathcal{H}}$, $\mathfrak{H}$ is a Hilbert space to which we associate the Gaussian process $\mathfrak{X} = \{\mathfrak{X}(\mathfrak{h}) : \mathfrak{h} \in \mathfrak{H}\}$. Let $\mathcal{G}$ denote the Gaussian Hilbert space spanned by $\{\mathfrak{X}(\mathfrak{h}) : \mathfrak{h} \in \mathfrak{H}\}$.

In view of (5.3.15) we observe that (X_1, X_l) can be realized jointly as follows:

$$X_1 = \Big\{\mathfrak{X}(\mathfrak{h}) : \mathfrak{h} = (h, 0, \dots, 0), h \in \mathfrak{b}\mathcal{H}\Big\},$$
$$X_l = \Big\{\mathfrak{X}(\mathfrak{h}) : \mathfrak{h} = (\rho_h^{l-1}h, \rho_h^{l-2}\sqrt{1-\rho_h^2}h, \cdots, \sqrt{1-\rho_h^2}h), h \in \mathfrak{b}\mathcal{H}\Big\}.$$

Let $\mathcal{G}_1$ denote the Gaussian subspace spanned by

$$\{\mathfrak{X}(\mathfrak{h}) : \mathfrak{h} = (h, 0, \dots, 0), h \in \mathcal{H}\}$$

and similarly for $\mathcal{G}_l$. Then, [83, Theorem 10.11] states that ρ_{X_1,X_l} is equal to the norm of the operator $P_{\mathcal{G}_l,\mathcal{G}_1}$ which is defined as the orthogonal projection of $\mathcal{G}$ onto $\mathcal{G}_l$ and then restricted to $\mathcal{G}_1$. This is now an easy exercise to check that $\|P_{\mathcal{G}_l,\mathcal{G}_1}\| \leqslant |\rho|_\infty^{l-1}$. The proof of (5.3.19) is complete. □

5.4 Constructions of shaking transformation

In order to make previous algorithms applicable, we now provide reversible shaking transformations in various situations (for random variables and random processes). Of course, Metropolis-Hastings(MH) and Gibbs type transformations using explicit transition kernels are natural candidates but here, we provide path-wise representations which lead to significant simplifications and which are presumably more suitable for tuning the shaking force (they induce transition kernels which are not explicit, as a difference with the usual MH algorithm). Our path-wise representation also makes the generalization into infinite dimensional cases natural and immediate.

Recall that one has to exhibit a shaking map $K(\cdot,\cdot)$ and a random variable Y such that $(X, K(X,Y)) \stackrel{\mathrm{d}}{=} (K(X,Y), X)$.

5.4.1 Poisson variable and compound Poisson process

For a Poisson variable $X := P \stackrel{\mathrm{d}}{=} \mathtt{Poisson}(\lambda)$ with parameter $\lambda > 0$, a possible transformation is

$$K(P, [\mathtt{Bin}(P, 1-p), \mathtt{Poisson}(p\lambda)]) = \mathtt{Bin}(P, 1-p) + \mathtt{Poisson}(p\lambda)$$

where $p \in (0,1)$, using extra independent Binomial and Poisson random variables, see [87, Chapter 5]. The intuitive interpretation is that we consider the Poisson realization P as a set of points. We remove each point in the set with probability p and add another set of points represented by an independent Poisson variable with parameter $p\lambda$

With the same idea, the above decomposition holds also for compound Poisson process (CPP in short) with parameter (λ, μ), i.e. $X := (P_t)_{0\leqslant t\leqslant T}$ where $P_t = \sum_{k=1}^{N_t} J_k$ where N is a standard Poisson process

with intensity λ and $(J_k)_k$ are i.i.d with distribution μ. Let $p \in (0,1)$: by coloring at random the jumps of N in `red` with probability $1-p$ and in `green` with probability p, we can write $N_t = N_t^{\mathtt{r}} + N_t^{\mathtt{g}}$ and $X_t = X_t^{\mathtt{r}} + X_t^{\mathtt{g}}$, using obvious notations. Then $X^{\mathtt{r}}$ and $X^{\mathtt{g}}$ are two independent CPP with parameters $((1-p)\lambda, \mu)$ and $(p\lambda, \mu)$. Using an extra independent CPP Y distributed as $X^{\mathtt{g}}$, it is easy to check that the following transformation satisfies **(K)**:

$$K(X,Y) = (X_t^{\mathtt{r}} + Y_t)_{0 \leqslant t \leqslant T}. \tag{5.4.1}$$

In Subsection 6.3, we will use this shaking transformation for the example of queuing system with exponential inter-arrival time .

5.4.2 Gamma distribution

For a random variable $X = \Gamma_{a,b}$ with Gamma distribution $\mathtt{Gamma}(a,b)$ ($a > 0, b > 0$) defined by $\mathbb{P}(\Gamma_{a,b} \in \mathrm{d}x) = c_a b^a x^{a-1} e^{-bx} \mathbf{1}_{x>0} \mathrm{d}x$ for a normalizing constant c_a, we can provide a simple transformation based on the so-called *beta-gamma algebra* from [48]. We read about this property from [36][2].

Let $p \in (0,1)$: with the notation of **(K)**, take

$$Y = (\mathtt{Beta}(a(1-p), ap), \mathtt{Gamma}(ap, b))$$

with two extra independent Beta and Gamma distributed random variables, and set

$$\mathcal{K}(\Gamma_{a,b}) = \Gamma_{a,b}\, \mathtt{Beta}(a(1-p), ap) + \mathtt{Gamma}(ap, b). \tag{5.4.2}$$

Then Assumption (**(K)**) is satisfied. This relation helps to construct reversible shaking transformations for other probability distributions. We give the proof in the following.

Proposition 5.4.1 ([48]). *Suppose $a_1, a_2, b > 0$ and $B_{a_1,a_2} \sim \mathtt{Beta}(a_1, a_2)$ we have*

$$(\Gamma_{a_1,b}, \Gamma_{a_2,b}) \overset{d}{=} (B_{a_1,a_2}\Gamma_{a_1+a_2,b}, (1 - B_{a_1,a_2})\Gamma_{a_1+a_2,b}) \tag{5.4.3}$$

where all the random variables are independent of each other.

Corollary 5.4.1. *Equation (5.4.2) satisfies Assumption* **(K)**.

Proof. Take $a_1 = a(1-p)$, $a_2 = ap$ in Equation (5.4.3), we have

$$(\Gamma_{a(1-p),b}, \Gamma_{ap,b}) \overset{d}{=} (B_{a(1-p),ap}\Gamma_{a,b}, (1 - B_{a(1-p),ap})\Gamma_{a,b})$$

[2] I want to thank Société Générale, where I did an internship before starting my PhD and had some idle time to read this book. This interesting exercise seemed very innocent to me at that moment.

Define $\Gamma'_{ap,b}$ as an independent copy of $\Gamma_{ap,b}$. By the trivial identity

$$\Gamma_{a,b} = B_{a(1-p),ap}\Gamma_{a,b} + (1 - B_{a(1-p),ap})\Gamma_{a,b}$$

and Proposition 5.4.1, we get

$$\begin{aligned}
&(\Gamma_{a,b}, \Gamma_{a,b}B_{a(1-p),ap} + \Gamma_{ap,b}) \\
=&(B_{a(1-p),ap}\Gamma_{a,b} + (1 - B_{a(1-p),ap})\Gamma_{a,b}, \Gamma_{a,b}B_{a(1-p),ap} + \Gamma_{ap,b}) \\
\overset{d}{=}&(\Gamma_{a,b}B_{a(1-p),ap} + \Gamma'_{ap,b}, \Gamma_{a,b}B_{a(1-p),ap} + \Gamma_{ap,b})
\end{aligned}$$

Similarly, we get

$$\begin{aligned}
&(\Gamma_{a,b}B_{a(1-p),ap} + \Gamma_{ap,b}, \Gamma_{a,b}) \\
=&(\Gamma_{a,b}B_{a(1-p),ap} + \Gamma_{ap,b}, B_{a(1-p),ap}\Gamma_{a,b} + (1 - B_{a(1-p),ap})\Gamma_{a,b}) \\
\overset{d}{=}&(\Gamma_{a,b}B_{a(1-p),ap} + \Gamma_{ap,b}, \Gamma_{a,b}B_{a(1-p),ap} + \Gamma'_{ap,b})
\end{aligned}$$

Set

$$\begin{aligned}
A &= \Gamma_{a,b}B_{a(1-p),ap} + \Gamma_{ap,b} \\
B &= \Gamma_{a,b}B_{a(1-p),ap} + \Gamma'_{ap,b}
\end{aligned}$$

obviously $(A, B) \overset{d}{=} (B, A)$. This concludes the proof. □

Figure 5.4 represents 100000 independent simulations of $(\Gamma, \mathcal{K}(\Gamma))$ with their respective marginal histograms. The smaller the value of p, the slighter the shaking. On the plots, observe that Γ and $\mathcal{K}(\Gamma)$ have the same distribution (coherently with **(K)**).

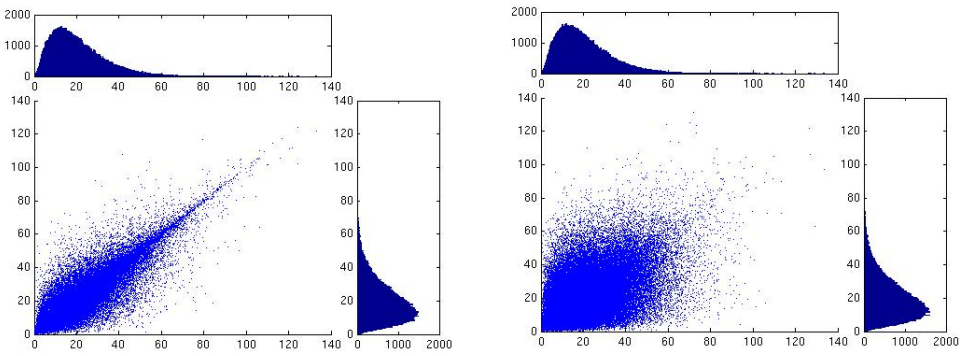


FIGURE 5.4: Shaking `Gamma`$(2.5, 0.12)$ variables with $p = 0.1$ (left) and $p = 0.5$ (right)

5.4.3 Other random variables

General trick for shaking construction Shaking transformations for many other distributions can be found by a change of variable trick: assume that

$$X \overset{\mathrm{d}}{=} f(Z)$$

for some invertible function f and a random variable Z having a shaking transformation $\mathcal{K}_Z$, then

$$\mathcal{K}_X(\cdot) = f(\mathcal{K}_Z(f^{-1}(\cdot)))$$

defines a reversible shaking transformation for X. We list several cases using this trick

▷ **Exponential and $\chi^2(k)$ distributions** In particular, take $a = 1$ in equation (5.4.2) we recover the case of exponential distribution $\texttt{Exp}(b)$. Note also that shaking transformation for $\chi^2(k)$ distribution directly follows from the above since this distribution is the same as that of $2\texttt{Gamma}(\frac{k}{2}, 1)$.

▷ **(s)-distribution** For $X := T$ with the (s)-distribution given by

$$\mathbb{P}(T \in \mathrm{d}t) = \frac{1}{\sqrt{2\pi t^3}} \exp(-\frac{1}{2t}) \mathbf{1}_{t>0} \mathrm{d}t$$

which represents the hitting time of level 1 by a standard Brownian motion, we have (at least) two shaking transformations. Firstly, we can shake Brownian motion as previously explained. Alternatively, we can use the well-known identity

$$T \stackrel{\mathrm{d}}{=} G^{-2}$$

where G is a standard Gaussian random variable and apply the Gaussian shaking transformation.

▷ **Uniform variable on $[0, 1]$** We can rely on the relation with exponential distribution to write

$$X := U \stackrel{\mathrm{d}}{=} \exp(-\texttt{Exp}(1))$$

Let $p \in (0, 1)$: in view of (5.4.2), the following transformation satisfies **(K)**,

$$\mathcal{K}(U) = U^{\texttt{Beta}(1-p,p)} \exp(-\texttt{Gamma}(p, 1)) \tag{5.4.4}$$

with extra independent Beta and Gamma random variables.

An alternative shaker is

$$\mathcal{K}(U) = U + \delta Y \bmod 1$$

where Y is uniformly distributed on $[-1, 1]$ and independent of U with $|\delta|$ small[3]. This choice is presumably efficient only in the case where the values 0 and 1 are identified, since for values of U close to 0 and 1 respectively, the shaker will produce with high probability values close to 1

[3]We thank one referee of our published paper [68] for suggesting this construction

and 0 respectively, thus possibly dramatically changing the system configuration when 0 and 1 are not identified. These are not the properties we may wish for a shaker (changing slightly configurations). Clearly, the choice of shakers is not unique and requires further investigation.

Now that we are in a position to shake uniform distribution, it is easy to shake any distribution on $\mathbb{R}$ having a continuous CDF function F since $F(X) \overset{\mathrm{d}}{=} \texttt{Unif}$. This is useful in the case F and F^{-1} are easily tractable. We do not list all the possibilities.

▷ **Bernoulli and geometric distribution** Suppose X is a Bernoulli variable, i.e. $X = 0$ with probability p and $X = 1$ with probability $1 - p$. The shaking transformation we propose is that if $X = 0$, it will be changed to 1 with probability x and if $X = 1$, it will be changed to 0 with probability y.

In order to satisfy Equation (5.3.15), we need to have

$$px = (1 - p)y$$

If ξ follows a geometric distribution $\mathcal{G}(p)$, i.e.

$$\mathbb{P}(\xi = k) = (1 - p)^k p, \quad k \geqslant 0,$$

The shaking transformation we propose writes in the following way

$$K(\xi) = Y 1_{Y<\xi} + \xi 1_{Y \geqslant \xi, U<1-y} + (\xi + Z + 1) 1_{Y \geqslant \xi, U \geqslant 1-y} \tag{5.4.5}$$

where $Y \sim \mathcal{G}(x)$, $U \sim \mathcal{U}([0, 1])$ and $Z \sim \mathcal{G}(p)$ with shaking parameters x, y satisfying$(1 - p)x = py$.

Proposition 5.4.2. *Equation* (5.4.5) *satisfies Assumption* **(K)**.

Proof. Equation (5.4.5) does not come out of nothing. It is related to the interpretation of geometric distribution as the sum of independent Bernoulli variables.

We will just give a computational proof by verifying

$$\mathbb{P}(\xi = n)\,\mathbb{P}(\xi = n, K(\xi) = m) = \mathbb{P}(\xi = m)\,\mathbb{P}(\xi = m, K(\xi) = n)$$

We can easily check that

$$\mathbb{P}(\xi = n, K(\xi) = m) = \begin{cases} (1 - x)^m x, & m < n \\ y(1 - x)^n (1 - p)^{m-n-1} p & m > n \end{cases}$$

Thus, supposing $m < n$, we have

$$\begin{aligned} \mathbb{P}(\xi = n)\,\mathbb{P}(\xi = n, K(\xi) = m) &= (1 - p)^n p (1 - x)^m x \\ \mathbb{P}(\xi = m)\,\mathbb{P}(\xi = m, K(\xi) = n) &= (1 - p)^m p y (1 - x)^m (1 - p)^{n-m-1} p \end{aligned}$$

$$(1 - p)^n p (1 - x)^m x = (1 - p)^m p y (1 - x)^m (1 - p)^{n-m-1} p$$

implies

$$(1-p)x = py$$

which is exactly what we impose with Equation (5.4.5). □

Alternatively to Equation (5.4.5), we can use the fact that the integer part of a random variable following exponential distribution has the geometric distribution. It writes as follows: we define $\lambda = -\ln(1-p)$ and

$$K(X) = \lfloor (X + \{E\})B + G \rfloor \tag{5.4.6}$$

where $E \sim \texttt{Exp}(\lambda), B \sim \texttt{Beta}(1-x, x)$ and $G \sim \texttt{Gamma}(x, \lambda)$ with shaking parameter $x \in [0,1]$. In words, we add the fractional part to turn the geometric variable into an exponential one (indeed, the integer and decimal parts of a $\texttt{Exp}(\lambda)$ are two independent random variables), then shake the exponential variable using existing formula, then take the integer part of the shaken exponential variable as the shaken geometric variable.

▷ **Other shakings for random variables** In cases where explicit transformation is not available, we can use implicit transformation. Namely, assume for instance that $X := f(Z_1, \cdots, Z_n)$ with independent $(Z_i)_i$, which serves to simulate X through the simulation of $(Z_i)_{1 \leqslant i \leqslant n}$, and suppose that each Z_i has an explicit shaking transformation. Then the implicit shaking transformation for X is

$$\mathcal{K}(X) = f(\mathcal{K}_1(Z_1), \cdots, \mathcal{K}_n(Z_n))$$

where the exact expression of $\mathcal{K}_i$ may be different according to the type of random variables Z_i and each shaking is made independently of the others. For example, this can be applied to X having Beta distribution because of the identity $\texttt{Beta}(a,b) \overset{\mathrm{d}}{=} \frac{\texttt{Gamma}(a,1)}{\texttt{Gamma}(a,1)+\texttt{Gamma}(b,1)}$ with independent Gamma distributions.

5.4.4 Other variations on the shaking

▷ **Randomized shaking** Actually, in the previous examples $K(\cdot,\cdot)$ is often written as $K_\theta(\cdot,\cdot)$ for a parameter θ serving to tune the shaking force. A first remark is that instead of fixing the parameter value of θ, one can also randomize it, which gives rise to another reversible shaking transformation.

Lemma 5.4.1. *Assume that $K_\theta(\cdot, Y)$ satisfies* **(K)** *for any θ in a measurable space Θ and that $K_\cdot(\cdot,\cdot)$ defines a measurable function from $\Theta \times \mathbb{S} \times \mathbb{Y}$ into $\mathbb{S}$. Let T be any Θ-valued random variable independent of Y, then $K_T(\cdot, Y)$ satisfies* **(K)**.

The proof is easy and left to the reader. As a consequence, all the shaking transformations presented before can be generalized with a random parameter. This randomization technique will be seen useful in the example of Section 6.6.

▷ **Partial shaking** When the random variable X is built on several independent random variables, it may be relevant to shake only some of them. For instance, consider a general pure jump process (including CPP or renewal process), where $A = (A_n)_{n\geqslant 1}$ represents the inter-arrival times and $B = (B_n)_{n\geqslant 1}$ represents the jump sizes, A and B being independent: the shaking transformation may concern both A and B, or only A (the jump times), or only B (the jump sizes). These alternatives are tested in the subsequent examples on insurance and queuing system. Similarly, for a SDE driven by both Brownian motion and another independent Levy process, we can shake the first driving process or the second, or both.

Another strategy is to apply *randomized partial shaking*. For a model of the form $X := f(Z_i, 1 \leqslant i \leqslant n)$ with independent $(Z_i)_i$, when n large or $n = +\infty$ we can reduce the computational cost by picking at random a subset of coordinates and only shake independently the corresponding random variables. The property of reversible shaking transformation is preserved owing to Lemma 5.4.1. This method will be used in the random graph example of Subsection 6.4.

Metropolis-Hastings type transformations In finite dimensional applications, we can also use Metropolis-Hastings(MH) type transformations, which consists in proposing a potential transition and then use a rejection function to decide with which probability this transition will be accepted. But this transition is usually less efficient than our shaking transformation.

Firstly, suppose we work with a continuous distribution. With MH type transformation, the simulation has a strictly positive probability to be rejected and we stick to the previous position in this case. But with our shaking transformation, we always get a different point so no simulation effort is wasted.

Interestingly, our shaking with rejection can be interpreted as one MH transformation with implicit transition kernel. This again shows the interest of our shaking transformation. If we use standard MH type transition in the shaking with rejection step, then we need to decide twice if the transition proposition will be rejected (once inside MH transition, once for shaking with rejection) while using shaking transformation we only need to decide once.

Secondly, our shaking transformation is easier than MH transformation to be implemented. And since we only have one parameter inside our shaking transformation, it is easy to be calibrated to achieve good numerical performance.

Lastly, the natural generalization to infinite dimensional case provided by our shaking transformation is out of scope for MH transformation since in infinite dimensional case we do not have a density function.

But there are also applications where only MH transformation is available, such as the case where the system is simulated according to a given density function with an unknown normalization constant.

5.5 Almost sure convergence of POP method in finite dimensions

In Subsection 5.2.4 the L^2 convergence of POP method is proven under additional assumptions which are not obvious to verify. In this section, we are going to prove the almost sure convergence of POP method in all the finite dimensional cases without any assumption. Here by finite dimensional cases we mean that X under consideration is a finite dimensional random variable. The proof is based on Theorem 4.1.5 that we reviewed in Chapter 4.

Theorem 5.5.1. *POP method converges almost surely in all the finite dimensional cases where the shaking transformation admits a strictly positive transition density for continuous distributions or a strictly positive transition probability for discrete distributions. In particular, it converges for all the cases presented in Section 5.4.*

Proof. As is shown in Algorithm 2, the final estimator in POP method is given as the product of a fixed number of conditional probability estimators, so it suffices to prove that each conditional probability estimator converges almost surely.

Firstly, we remark that the shaking and rejection transformation at level k, defined in Equation (5.2.2), can be interpreted in the form of Equation (4.1.3), which is well-known as *Metropolis-Hastings sampler*. Unlike usual Metropolis-Hastings sampler where explicit transition densities and acceptance functions are used, we use implicit transition densities and acceptance functions via all the shaking transformations proposed in Section 5.4. The implicit transition density for Gaussian shaking transformations is given as an example after the proof. All the implicit transition densities in Section 5.4 can be written out in a similar way.

Secondly, the assumption $a(x) > 0$ in Theorem 4.1.5 writes in our rare event setting as $\mathbb{P}\left(\Psi_Z(\mathcal{K}(x)) \in A_k\right) > 0$ for any x s.t. $\Psi_Z(x) \in A_k$. This inequality holds true since we assume $0 < \mathbb{P}\left(Z \in A\right) \leqslant \mathbb{P}\left(\Psi_Z(X) \in A_k\right)$, i.e. $\Psi_Z^{-1}(A_k)$ has a strictly positive Lebesgue measure.

Thirdly, notice that the existence of a stationary distribution has been shown previously in Proposition 5.2.1 and we can easily see that the Markov chain in POP method is η-irreducible, due to the strictly positive transition density p. Therefore, from Theorem 4.1.5 we deduce that

in Algorithm 2, $p_k^{(N)}$ converges to $\mathbb{P}(Z \in A_k \mid Z \in A_{k-1})$ almost surely, which concludes the proof. □

In the case of shaking transformation for standard q-dimensional normal variable, we will take

$$\mathcal{K}(x) = K(x, X') = (\rho_i x_i + \sqrt{1-\rho_i^2} X_i')_{1 \leqslant i \leqslant q}$$

with i.i.d. standard Gaussian variables $(X_i')_{1 \leqslant i \leqslant q}$ and $\sup_{1 \leqslant i \leqslant q} |\rho_i| < 1$, the measure η in Theorem 4.1.5 can be taken as the Lebesgue measure on $\mathbb{R}^q$ and the transition density is given by

$$p(x, y) = \exp\left(-\sum_{i=1}^{q} \frac{|y_i - \rho_i x_i|^2}{2(1-\rho_i^2)} \right) (2\pi)^{-q/2} \prod_{i=1}^{q} (1-\rho_i^2)^{-1/2} \tag{5.5.1}$$

Then the acceptance function corresponds to $a(x, y) = 1_{\Psi_Z(y) \in A_k}$ and the local mean acceptance rate to $a(x) = \int_{\mathbb{R}^q} a(x, y) p(x, y) \mathrm{d}y$.

In a similar way, we can easily find the implicit transition density for all the finite dimensional shaking transformations given in Section 5.4.

The above proof can not be extended directly to infinite dimensional cases because of the loss of density function. An attempt towards the convergence of infinite dimensional case has been made in Subsection 5.3.3, with further work remaining to be done.

5.6 Adaptive POP method

5.6.1 Algorithm

For good numerical performance, one may wish that the conditional probabilities at intermediate levels of POP method are of the same order (for example, in [93] it is argued that the equiprobability choice minimizes the variance of splitting algorithms). However, the appropriate choice of intermediate levels to ensure this condition requires a priori knowledge about the nature of the rare event under consideration. In the absence of such knowledge, choosing appropriate intermediate levels is challenging. Here, we propose an adaptive POP method where at each level, except for the last, the conditional probability is fixed to a pre-decided value $p \in (0, 1)$ (typically 10%).

For the ease of exposition, let us suppose that $Z = \Psi_Z(X)$ takes values in $\mathbb{R}^d$ and that the rare event set is of the form

$$A = \{z \in \mathbb{R}^d : \varphi(z) \leqslant \bar{a}\}$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a measurable function and $\bar{a}$ is a given finite threshold. The principle of the adapted version of POP is to set

$$A_k := \{z \in \mathbb{R}^d : \varphi(z) \leqslant a_k\} \tag{5.6.1}$$

with online computations of the acceptance level a_k. Notice that this choice of A_k corresponds to the notations explained in Subsection 5.3.3. For having constant conditional probabilities, we should take a_k as the quantile of $V := \varphi(\Psi_Z(X)) = \varphi(Z)$ at level p^k. This heuristics guides the following notation and definition.

We denote the p-quantile of the distribution of V as

$$Q_p^1 = F_V^{-1}(p) := \inf\{v \in \mathbb{R} : F_V(v) \geqslant p\} \tag{5.6.2}$$

where $F_V(\cdot)$ is the cumulative distribution function of V. We define the conditional quantile function $g_p(\cdot)$ of V in the following way:

$$g_p(q) := \inf\{v \in \mathbb{R} : \mathbb{P}(V \leqslant v \mid V \leqslant q) \geqslant p\} \tag{5.6.3}$$

and also recursively define

$$Q_p^{l+1} := g_p(Q_p^l), l \geqslant 1. \tag{5.6.4}$$

The above formula remains valid for $l = 0$ by setting $Q_p^0 := +\infty$. This is our convention from now on. Moreover, we define

$$r(q) := \mathbb{P}(V \leqslant \bar{a} \mid V \leqslant q), \tag{5.6.5}$$

then the true rare event probability $\alpha = \mathbb{P}(V \in A)$ can be written in a unique way as

$$\alpha = r(Q_p^{L^*})p^{L^*} \tag{5.6.6}$$

where $L^* \in \mathbb{N}$ and $r(Q_p^{L^*}) \in (p, 1]$. We are now in a position to define the POP algorithm with adaptive number of levels (approximation of L^*).

Initialization. We are given a common initialization point x_0 such that $\varphi(\Psi_Z(x_0)) \leqslant \bar{a}$ (we can follow the method given in Algorithm 2.

1st Markov chain. Simulate the first N iterations of the Markov chain based on Equation (5.2.13) with starting state x_0. Then, sort the sample $(V_{N,1}^1, \ldots, V_{N,N}^1)$ in ascending order as

$$V_{N,(1)}^1 \leqslant \ldots \leqslant V_{N,(k)}^1 \leqslant \ldots \leqslant V_{N,(N)}^1$$

and take $k_p^1 \in \{1, \ldots, N\}$ such that

$$k_p^1 - 1 < Np \leqslant k_p^1.$$

Denote by $\hat{Q}_{N,p}^1 = V_{N,(k_p^1)}^1$, the estimate for Q_p^1 based on N samples.

2nd Markov chain. Start the Markov chain in Equation (5.2.13) with initial state x_0 and with cascade event set A_1 corresponding to level $\hat{a}_1 := \hat{Q}_{N,p}^1$ (see (5.6.1)). Again, simulate the first N steps and sort

the sample $(V^2_{N,1}, \ldots, V^2_{N,N})$ in ascending order as

$$V^2_{N,(1)} \leqslant \ldots \leqslant V^2_{N,(k)} \leqslant \ldots \leqslant V^2_{N,(N)}.$$

Take k^2_p such that

$$k^2_p - 1 < Np \leqslant k^2_p$$

and denote by $\hat{a}_2 := \hat{Q}^2_{N,p} = V^2_{N,(k^2_p)}$, the estimate for Q^2_p based on N samples.

Iteration and stopping. Next, repeat the procedure until the $(L_N + 1)$th step where we have $\hat{Q}^{L_N+1}_{N,p} \leqslant \bar{a}$ for the first time. The intermediate sets A_k in (5.6.1) are defined by the acceptance levels $\hat{a}_k = \hat{Q}^k_{N,p}$. Calculate $\hat{r}_N(\hat{Q}^{L_N}_{N,p})$, defined as the proportion of values $(V^{L_N+1}_{N,1}, \ldots, V^{L_N+1}_{N,N})$ which are smaller than $\bar{a}$ with the cascade event set corresponding to acceptance level $\hat{Q}^{L_N}_{N,p}$.

In the case $L_N = 0$, we set by convention $\hat{Q}^0_{N,p} = +\infty$ (similarly to Q^0_p).

Outputs. Compute the probability estimate as

$$\hat{\alpha}_N := \hat{r}_N(\hat{Q}^{L_N}_{N,p}) p^{L_N} \tag{5.6.7}$$

as an approximation of the probability α written in (5.6.6).

Remark 5.6.1. *In the following Theorem 5.6.1, we assume that the initial points of the above Markov chains are fixed (actually all equal to x_0). The deterministic initialization of Markov chain at each level, indeed, partly simplifies the convergence analysis. However in practice, we could advantageously start the l-th level Markov chain from a point close to the acceptance level, i.e. $X_{l,0}$ equal to the x-configuration of one of the $V^l_{N,(1)}, \ldots, V^l_{N,(k^l_p)}$. The choice of $V^l_{N,(1)}$ is the simplest from algorithmic viewpoint, since we only need to update the smallest $V^l_{N,i}$ (with the corresponding X) during the algorithm run. Besides, we observe only a very small impact of initialization on the numerical results.*

The numerical performance of adaptive POP method and its comparison with POP method with prefixed intermediate levels can be found in Section 6.7 and Section 6.8.

5.6.2 Convergence result

In order to prove the consistency of estimator $\hat{\alpha}_N$, we make the following assumptions. We discuss the applicability of these assumptions after the proof of estimator convergence.

Assumption 5.6.1. *The distribution of V admits a density $q \mapsto f(q)$, which is continuous and strictly positive at $q = Q^l_p$ for all $l \in \{1, \ldots, L^*\}$.*

Assumption 5.6.2. *For any $q \in (-\infty, +\infty]$, let $\hat{g}_{N,p}(q)$ denote the quantile estimator for $g_p(q)$ based on N iterations of the Markov chain based on the rejection set $\{x : \varphi(\Psi_Z(x)) > q\}$. For any $l \in \{0, \dots, L^* + 1\}$, there exist an open interval I_l containing Q_p^l (with the convention $Q_p^0 = +\infty$ and $I_0 = \{+\infty\}$) and a function $b : I_l \times \mathbb{N}^* \times (0, +\infty) \to [0, +\infty)$ such that for all $\varepsilon > 0$ and $q \in I_l$*

$$\mathbb{P}\left(|\hat{g}_{N,p}(q) - g_p(q)| > \varepsilon\right) \leqslant b(q, N, \varepsilon).$$

We further assume

$$\sum_{N \geqslant 1} \sup_{q \in I_l} b(q, N, \varepsilon) < +\infty.$$

Assumption 5.6.3. *For any $q \in (-\infty, +\infty]$, let $\hat{r}_N(q)$ denote the mean estimator for $r(q)$ based on N iterations of the Markov chain based on the rejection set $\{x : \varphi(\Psi_Z(x)) > q\}$. For any $l \in \{L^* - 1, L^*\} \cap \mathbb{N}$, there exist an open interval J_l containing Q_p^l (with the convention $J_0 = \{+\infty\}$) and a function $c : J_l \times \mathbb{N}^* \times (0, +\infty) \to [0, +\infty)$ such that for all $\varepsilon > 0$ and $q \in J_l$*

$$\mathbb{P}\left(|\hat{r}_N(q) - r(q)| > \varepsilon\right) \leqslant c(q, N, \varepsilon).$$

We further assume

$$\sum_{N \geqslant 1} \sup_{q \in J_l} c(q, N, \varepsilon) < +\infty.$$

Theorem 5.6.1. *Suppose that Assumptions 5.6.1, 5.6.2 and 5.6.3 hold. Then, $\hat{\alpha}_N$ converges almost surely to $\alpha = \mathbb{P}(Z \in A)$ as $N \to +\infty$.*

Theorem 5.6.2. *With*

$$\sum_{N \geqslant 1} \sup_{q \in I_l} b(q, N, \varepsilon) < +\infty$$

and

$$\sum_{N \geqslant 1} \sup_{q \in J_l} c(q, N, \varepsilon) < +\infty$$

in the above assumptions replaced by less restrictive conditions that

$$\sup_{q \in I_l} b(q, N, \varepsilon) \to 0, \sup_{q \in J_l} c(q, N, \varepsilon) \to 0$$

as $N \to +\infty$, we have that $\hat{\alpha}_N$ converges in probability to $\alpha = \mathbb{P}(Z \in A)$ as $N \to +\infty$.

Proof of Theorem 5.6.1

We split the proof in several steps.

Lemma 5.6.1. *For any $l \in \{1, \dots, L^* + 1\}$ and any $\epsilon > 0$, we have*

$$\sum_{N \geqslant 1} \mathbb{P}\left(|\hat{Q}_{N,p}^l - Q_p^l| > \varepsilon\right) < +\infty. \tag{5.6.8}$$

Thus, $\hat{Q}^l_{N,p}$ converges to Q^l_p almost surely as $N \to +\infty$.

Proof. We proceed by induction on l. Assumption 5.6.2 with empty rejection ($Q^0_p = +\infty$) ensures that (5.6.8) is true for $l = 1$. Now suppose that (5.6.8) is true for some $l \geqslant 1$ and let us prove it for $l+1$. We have

$$\begin{aligned}
&\mathbb{P}\left(|\hat{Q}^{l+1}_{N,p} - Q^{l+1}_p| > \varepsilon\right) \\
\leqslant &\mathbb{P}\left(|\hat{Q}^{l+1}_{N,p} - g_p(\hat{Q}^l_{N,p})| > \varepsilon/2\right) + \mathbb{P}\left(|g_p(\hat{Q}^l_{N,p}) - g_p(Q^l_p)| > \varepsilon/2\right) \\
\leqslant &\mathbb{P}\left(|\hat{Q}^{l+1}_{N,p} - g_p(\hat{Q}^l_{N,p})| > \varepsilon/2, \hat{Q}^l_{N,p} \in I_l\right) \\
+&\mathbb{P}\left(\hat{Q}^l_{N,p} \notin I_l\right) + \mathbb{P}\left(|g_p(\hat{Q}^l_{N,p}) - g_p(Q^l_p)| > \varepsilon/2\right) \\
:= &\mathtt{I} \quad + \mathtt{II} \quad + \mathtt{III}.
\end{aligned} \tag{5.6.9}$$

From Assumption 5.6.2, on $\{\hat{Q}^l_{N,p} \in I_l\}$ we have

$$\mathbb{P}\left(|\hat{Q}^{l+1}_{N,p} - g_p(\hat{Q}^l_{N,p})| > \varepsilon/2 \mid \hat{Q}^l_{N,p}\right) \leqslant b(\hat{Q}^l_{N,p}, N, \varepsilon/2).$$

Thus, the term $\mathtt{I}$ in the right hand side of (5.6.9) is bounded the supremum $\sup_{q \in I_l} b(q, N, \varepsilon/2)$, and still by Assumption 5.6.2, we get

$$\sum_{N \geqslant 1} \mathbb{P}\left(|\hat{Q}^{l+1}_{N,p} - g_p(\hat{Q}^l_{N,p})| > \varepsilon/2, \hat{Q}^l_{N,p} \in I_l\right) < +\infty. \tag{5.6.10}$$

Furthermore, Assumption 5.6.1 implies that the function $g_p(q)$ is continuous at $q = Q^l_p$. This combined with the induction hypothesis at level l implies that the series with general terms given by $\mathtt{II}$ and $\mathtt{III}$ converge similarly to (5.6.10). Therefore, (5.6.8) is proved for $l+1$ and the result follows. □

Corollary 5.6.1. *When $\frac{\log \alpha}{\log p}$ is not an integer, i.e. $Q^{L^*+1}_p < \bar{a} < Q^{L^*}_p$, we have*

$$\mathbb{P}\left(L_N = L^* \text{ for } N \text{ large enough }\right) = 1.$$

Proof. This is a direct consequence from Lemma 5.6.1. □

Lemma 5.6.2. *Assume $L^* \neq 0$. When $\frac{\log \alpha}{\log p}$ is not an integer, we have for any $\epsilon > 0$,*

$$\sum_{N \geqslant 1} \mathbb{P}\left(|\hat{Q}^{L_N}_{N,p} - Q^{L^*}_p| > \epsilon\right) < +\infty.$$

Proof. Firstly, we make a trivial decomposition:

$$\mathbb{P}\left(|\hat{Q}^{L_N}_{N,p} - Q^{L^*}_p| > \epsilon\right) = \mathbb{P}\left(\hat{Q}^{L_N}_{N,p} - Q^{L^*}_p > \epsilon\right) + \mathbb{P}\left(Q^{L^*}_p - \hat{Q}^{L_N}_{N,p} > \epsilon\right). \tag{5.6.11}$$

Recall, that $\hat{Q}^{L_N+1}_{N,p}$ is the first quantile estimation which lies below $\bar{a}$. In the first term in r.h.s of Equation (5.6.11), if $\hat{Q}^{L_N}_{N,p} > Q^{L^*}_p + \epsilon$ and $\hat{Q}^{L_N+1}_{N,p} \leqslant$

$\bar{a}$, then there is no $\hat{Q}^l_{N,p}$ which lies in the interval $]Q_p^{L^*} - \delta, Q_p^{L^*} + \delta[$ with $\delta = \min\{\epsilon, Q_p^{L^*} - \bar{a}\} > 0$. So $\{\hat{Q}^{L_N}_{N,p} - Q_p^{L^*} > \epsilon\}$ implies $\{|\hat{Q}^{L^*}_{N,p} - Q_p^{L^*}| > \delta\}$ and we have

$$\mathbb{P}\left(\hat{Q}^{L_N}_{N,p} - Q_p^{L^*} > \epsilon\right) \leqslant \mathbb{P}\left(|\hat{Q}^{L^*}_{N,p} - Q_p^{L^*}| > \delta\right).$$

Next, we make another decomposition:

$$\begin{aligned}\mathbb{P}\left(Q_p^{L^*} - \hat{Q}^{L_N}_{N,p} > \epsilon\right) \leqslant & \mathbb{P}\left(Q_p^{L^*} - \hat{Q}^{L_N}_{N,p} > \epsilon, |\hat{Q}^{L^*}_{N,p} - Q_p^{L^*}| \leqslant \epsilon\right) \\ & + \mathbb{P}\left(|\hat{Q}^{L^*}_{N,p} - Q_p^{L^*}| > \epsilon\right).\end{aligned}$$

On the joint event in the first probability in the above r.h.s. inequality, we must have $\hat{Q}^{L_N}_{N,p} < \hat{Q}^{L^*}_{N,p}$, and consequently $\hat{Q}^{L^*+1}_{N,p} > \bar{a}$ (by definition of $\hat{Q}^{L_N}_{N,p}$ as the last quantile estimation above $\bar{a}$). Thus, it follows that

$$\begin{aligned}&\mathbb{P}\left(Q_p^{L^*} - \hat{Q}^{L_N}_{N,p} > \epsilon, |\hat{Q}^{L^*}_{N,p} - Q_p^{L^*}| \leqslant \epsilon\right) \\ \leqslant &\mathbb{P}\left(|\hat{Q}^{L^*+1}_{N,p} - Q_p^{L^*+1}| > \bar{a} - Q_p^{L^*+1}\right).\end{aligned}$$

We are able to conclude the proof by collecting the above results and using Lemma 5.6.1 with $l = L^*$, $l = L^* + 1$ and various $\epsilon > 0$. □

Lemma 5.6.3. *When $\frac{\log \alpha}{\log p}$ is not an integer, we have for any $\epsilon > 0$*

$$\sum_{N \geqslant 1} \mathbb{P}\left(|\hat{r}_N(\hat{Q}^{L_N}_{N,p}) - r(Q_p^{L^*})| > \epsilon\right) < +\infty.$$

Consequently, $\hat{r}_N(\hat{Q}^{L_N}_{N,p})$ converges to $r(Q_p^{L^*})$ almost surely as $N \to +\infty$.

Proof. Assume first that $L^* \geqslant 1$. We decompose each probability using the notation of Assumption 5.6.3:

$$\begin{aligned}&\mathbb{P}\left(\left|\hat{r}_N(\hat{Q}^{L_N}_{N,p}) - r(Q_p^{L^*})\right| > \varepsilon\right) \\ &\leqslant \mathbb{P}\left(\left|\hat{r}_N(\hat{Q}^{L_N}_{N,p}) - r(\hat{Q}^{L_N}_{N,p})\right| > \varepsilon/2\right) + \mathbb{P}\left(\left|r(\hat{Q}^{L_N}_{N,p}) - r(Q_p^{L^*})\right| > \varepsilon/2\right) \\ &\leqslant \mathbb{P}\left(\left|\hat{r}_N(\hat{Q}^{L_N}_{N,p}) - r(\hat{Q}^{L_N}_{N,p})\right| > \varepsilon/2, \hat{Q}^{L_N}_{N,p} \in J_{L^*}\right) \\ &+ \mathbb{P}\left(\hat{Q}^{L_N}_{N,p} \notin J_{L^*}\right) + \mathbb{P}\left(\left|r(\hat{Q}^{L_N}_{N,p}) - r(Q_p^{L^*})\right| > \varepsilon/2\right). \qquad (5.6.12)\end{aligned}$$

The first term in the above r.h.s. is bounded by $\sup_{q \in J_{L^*}} c(q, N, \varepsilon/2)$ (arguing as in the proof of Lemma 5.6.1), thus it forms a convergent series owing to Assumption 5.6.3 ; the second term gives also a convergent series in view of Lemma 5.6.2; the last term is handled as the second, by noting that $r(q)$ is continuous at $q = Q_p^{L^*}$ (Assumption 5.6.1). Now

consider the case $L^* = 0$ and write

$$\mathbb{P}\left(\left|\hat{r}_N(\hat{Q}_{N,p}^{L_N}) - r(Q_p^{L^*})\right| > \varepsilon\right) \leqslant \mathbb{P}\left(\left|\hat{r}_N(\hat{Q}_{N,p}^{0}) - r(Q_p^{0})\right| > \varepsilon, L_N = 0\right) + \mathbb{P}\left(L_N \neq 0\right). \tag{5.6.13}$$

The convergence of the series formed by the first probability term in the above r.h.s. directly follows from Assumption 5.6.3. Moreover, by definition of L_N and since $L^* = 0$,

$$\{L_N \neq 0\} \subset \{\hat{Q}_{N,p}^1 > \bar{a}\} \subset \{|\hat{Q}_{N,p}^1 - Q_p^1| > \bar{a} - Q_p^1 > 0\}$$

then we conclude by Lemma 5.6.1 with $l = 1$.

□

Proof of Theorem 5.6.1, when $\log\alpha/\log p$ is not an integer. This is a direct result from Corollary 5.6.1 and Lemma 5.6.3. □

Next, we prove the convergence when $\log\alpha/\log p$ is an integer. This case needs to be dealt with separately as we no longer have almost sure convergence of L_N to L^*. When $\alpha = p^{L^*}$, the estimator can be expressed as

$$\begin{aligned}\hat{\alpha}_N = {} & 1_{\{L_N = L^*-1\}}\hat{r}_N(\hat{Q}_{N,p}^{L^*-1})p^{L^*-1} + 1_{\{L_N = L^*\}}\hat{r}_N(\hat{Q}_{N,p}^{L^*})p^{L^*} \\ & + 1_{\{L_N \notin \{L^*-1, L^*\}\}}\hat{r}_N(\hat{Q}_{N,p}^{L_N})p^{L_N}.\end{aligned}$$

Then, the error of our estimator is given as:

$$\begin{aligned}\hat{\alpha}_N - p^{L^*} = {} & 1_{\{L_N = L^*-1\}}\left(\hat{r}_N(\hat{Q}_{N,p}^{L^*-1}) - p\right)p^{L^*-1} \\ & + 1_{\{L_N = L^*\}}\left(\hat{r}_N(\hat{Q}_{N,p}^{L^*}) - 1\right)p^{L^*} \\ & + 1_{\{L_N \notin \{L^*-1, L^*\}\}}\left(\hat{r}_N(\hat{Q}_{N,p}^{L_N})p^{L_N} - p^{L^*}\right).\end{aligned} \tag{5.6.14}$$

Lemma 5.6.4. *If $\alpha = p^{L^*}$, we have $\mathbb{P}\left(L_N \in \{L^*-1, L^*\}, \text{as } N \text{ is large enough}\right) = 1$.*

Proof. In any case $Q_p^{L^*+1} < \bar{a}$: since $\hat{Q}_{N,p}^{L^*+1}$ converges a.s. to $Q_p^{L^*+1}$ (Lemma 5.6.1), by the definition of L_N we have $L_N + 1 \leqslant L^* + 1$ as $N \to +\infty$. Similarly, provided that $L^* > 1$, $\hat{Q}_{N,p}^{L^*-1}$ converges a.s. to $Q_p^{L^*-1} > \bar{a}$, thus $L_N \geqslant L^* - 1$ as $N \to +\infty$. □

Lemma 5.6.5. *For $l \in \{L^*-1, L^*\} \cap \mathbb{N}$ and any $\epsilon > 0$, we have*

$$\sum_{N \geqslant 1} \mathbb{P}\left(|\hat{r}_N(\hat{Q}_{N,p}^{l}) - r(Q_p^{l})| > \epsilon\right) < +\infty.$$

Thus for such l, $\hat{r}_N(\hat{Q}_{N,p}^{l})$ converges to $r(Q_p^{l})$ almost surely as $N \to +\infty$.

Proof. Similarly to (5.6.12), write

$$\mathbb{P}\left(\left|\hat{r}_N(\hat{Q}^l_{N,p}) - r(Q^l_p)\right| > \varepsilon\right) \leqslant \mathbb{P}\left(\left|\hat{r}_N(\hat{Q}^l_{N,p}) - r(\hat{Q}^l_{N,p})\right| > \varepsilon/2, \hat{Q}^l_{N,p} \in J_l\right) + \mathbb{P}\left(\hat{Q}^l_{N,p} \notin J_l\right) + \mathbb{P}\left(\left|r(\hat{Q}^l_{N,p}) - r(Q^l_p)\right| > \varepsilon/2\right).$$

If $l = 0$, the two last probabilities on the above r.h.s. are 0, since $\hat{Q}^0_{N,p} = Q^0_p = +\infty$, while the first probability forms a convergent series in view of Assumption 5.6.3.
If $l > 0$, we argue as in the proof of Lemma 5.6.1, applying Assumption 5.6.3, Lemma 5.6.1 and the local continuity of $r(\cdot)$. □

Proof of Theorem 5.6.1, when $\log\alpha/\log p$ is an integer. In the r.h.s. of Equation (5.6.14), applying Lemma 5.6.5 for $l = L^* - 1$ and $l = L^*$, we get that $\hat{r}_N(\hat{Q}^{L^*-1}_{N,p}) - p = \hat{r}_N(\hat{Q}^{L^*-1}_{N,p}) - r(Q^{L^*-1}_p)$ converges to zero almost surely and that $\hat{r}_N(\hat{Q}^{L^*}_{N,p}) - 1 = \hat{r}_N(\hat{Q}^{L^*}_{N,p}) - r(Q^{L^*}_p)$ converges to zero almost surely, respectively. Thus, applying Lemma 5.6.4 completes the proof. □

We provide some discussion on the assumptions made to prove Theorem 5.6.1.

- Assumption 5.6.1 is required for the continuity of $g_p(\cdot)$ and $r(\cdot)$ at quantile levels Q^l_p. In [46], this type of condition is also required for Assumption 5.6.2 to hold under some conditions.

- The first parts of Assumptions 5.6.2 and 5.6.3 are related to deviation inequalities of various statistics of ergodic Markov chains. Such inequalities have been shown for instance in [62, 90, 46] for uniformly geometrically, or high order polynomially ergodic Markov chains, when the starting point of the underlying Markov chain is either fixed or distributed with the stationary distribution. Whereas we always initialize the Markov chain at hand at a fixed point x_0, we believe that these assumptions are still reasonable because the marginal distribution of Markov chain converges to the stationary distribution (Proposition 5.8.1).

- The second halves of Assumption 5.6.2 and 5.6.3 are satisfied as soon as some exponential-type inequalities hold locally uniformly. Such exponential-type inequalities hold true under some assumptions, see for instance [62, Theorem 2], [90, Theorem 1] or [46, Theorems 2 and 3]. We require some local uniformity w.r.t. the parameters defining the Markov chain which is valid in the aforementioned references. Thus, we argue that our assumptions appear to be reasonable but it still requires some extra work to check these conditions for the general (possibly infinite-dimensional) Gaussian shaker.

Proof of Theorem 5.6.2

The proof of convergence in probability of our estimator is based on a well-known equivalence relationship: for a sequence of random variable C_n, C_n converges to 0 in probability is equivalent to the fact that for any given subsequence of C_n, we can extract a sub-subsequence such that the sub-subsequence converges to zero almost surely, see [18, Theorem 20.5].

Proof. For any given subsequence $\hat{\alpha}_{s_N}$ of $\hat{\alpha}_N$, we can extract a sub-subsequence $\hat{\alpha}_{s_{t_N}}$ such that

$$\sum_{N\geqslant 1} \sup_{q\in I_l} b(q, s_{t_N}, \varepsilon) < +\infty$$

$$\sum_{N\geqslant 1} \sup_{q\in J_l} c(q, s_{t_N}, \varepsilon) < +\infty$$

Remark the above summations are applied along the series s_{t_N}. Following exactly the same lines in the previous proof, with the sequence $1, 2, \cdots, N, \cdots$ replaced by its sub-subsequence s_{t_N}, we can prove that $\hat{\alpha}_{s_{t_N}}$ converges almost surely when N goes to infinity, thus concluding the proof using the above-mentioned equivalence relationship. □

Remark that in this adaptive version of POP method, the parallel feature of original POP method is lost. To recover this nice feature, the best way to implement POP method given a pratical problem is, in our opinion, the one to be discussed in Section 5.10, where we keep the parallel feature and use the adaptive version to overcome the problem of unusually small conditional probability.

We only provided the consistency study of this adaptive POP method. It is well known that the variance of ergodic Markov chain is difficult to make explicit in general, thus the theoretical study on the optimal variance of adaptive POP method remains to be conducted. However, in Section 6.7 and Section 6.8, we will provide numerical comparisons of different versions of IPS and POP methods with the same computation cost.

5.7 Sensitivity analysis in the Gaussian space

Another issue which is not often addressed in the rare event literature is the analysis of sensitivities of rare event statistics with respect to (w.r.t.) model parameters. This is an important issue especially because the rare events statistics are known to be strongly dependent on the model parameters (see the limit (5.7.1)). Moreover, if the parameters are estimated from the observed data, they typically constitute some error, thus, relating the sensitivity analysis to the concept of model risk. To the best of our knowledge, there are very few contributions on this subject in the rare event setting. We refer to [6] where such study is handled in the case of compound Poisson process using the score function method coupled

with the IS method. Our aim here is to extend the IPS and POP methods to encompass sensitivity analysis. As we consider Gaussian based models, for the sensitivity analysis we rely on the machinery of Malliavin calculus to derive elegant representations of derivatives of expectations of general Gaussian functionals (see e.g. [55, 63, 65, 88]). We will show that this approach suits the POP algorithm based on path configuration since there is no need to Markovianize the sensitivity weights. We note that in order to derive these results, we do not need any semimartingale models and Itô calculus framework.

Assume that the model at hand depends on a real-valued parameter θ, through the definition of Z and Φ so that $\mathbb{E}\left(\Phi \mathbf{1}_{Z\in A}\right)$ now should be written as $\mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta\in A}\right)$. The sensitivity of the above quantity w.r.t. θ is an important issue to account for because the errors in model calibration and estimation procedures could have a significant impact. This concerns the evaluation of model risk (see e.g. [39]). This question is even more delicate when combined with rare-event analysis since it is known that tails are very sensitive to parameter shocks [4]. For instance, if $G_\sigma \overset{\mathrm{d}}{=} \mathcal{N}(0,\sigma^2)$ then

$$\lim_{x\to+\infty} \frac{\mathbb{P}\left(G_\sigma \geqslant x\right)}{\mathbb{P}\left(G_{\sigma'} \geqslant x\right)} = \begin{cases} 0 & \text{if } 0<\sigma<\sigma' \\ +\infty & \text{if } \sigma>\sigma'>0 \end{cases}, \tag{5.7.1}$$

i.e. a small change of parameters may cause a large change of tail-probabilities.

To quantify the impact of θ on $\mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta\in A}\right)$, we may evaluate the derivative w.r.t. θ whenever it exists. However, this quantity may be uninformative in practice since in our rare-event setting, the above expectation is small and likely its derivative too. Alternatively, we suggest to evaluate the relative sensitivity defined by

$$\frac{\partial_\theta \mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta\in A}\right)}{\mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta\in A}\right)} \tag{5.7.2}$$

provided that $\mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta\in A}\right)$ is differentiable in θ and non zero.

Regarding the computational aspects, the derivative $\partial_\theta \mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta\in A}\right)$ can be estimated by the re-simulation method as follows: Take two values of θ which are close to each other, approximate expectation for each value of θ by Monte-Carlo simulations and form the finite difference as an estimator of the derivative. This is known to be not well suited to the case where the functional inside the expectation is irregular in θ which is typically our case because of the indicator function. A better strategy is to represent the derivative as an expectation (known as the likelihood method in the case of explicit distributions, or based on Integration-By-Parts formula in the Malliavin calculus setting [55]) and then evaluate it by simulations. This is our approach which we formulate as an assumption.

(IBP) There exists an open set $\Theta \subset \mathbb{R}$ such that $\theta \mapsto \mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta \in A}\right)$ is differentiable on Θ and for any $\theta \in \Theta$, there is an integrable random variable $\mathcal{I}(Z^\theta, \Phi^\theta)$ such that

$$\partial_\theta \mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta \in A}\right) = \mathbb{E}\left(\mathcal{I}(Z^\theta, \Phi^\theta)\mathbf{1}_{Z^\theta \in A}\right). \tag{5.7.3}$$

Combining this with Equation (5.3.14) gives a simple representation of the relative sensitivity.

Proposition 5.7.1. *Assume* **(IBP)**. *For any $\theta \in \Theta$ such that $\mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta \in A}\right) \neq 0$, we have*

$$\frac{\partial_\theta \mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta \in A}\right)}{\mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta \in A}\right)} = \frac{\mathbb{E}\left(\mathcal{I}(Z^\theta, \Phi^\theta) \mid Z^\theta \in A\right)}{\mathbb{E}\left(\Phi^\theta \mid Z^\theta \in A\right)}. \tag{5.7.4}$$

It is important to observe that this can be directly evaluated by the POP method using the ratio of two time-average approximations of $\mathcal{I}(Z^\theta, \Phi^\theta)$ and Φ^θ respectively, along only one Markov chain defined by applying shaking with rejection with respect to $Z^\theta \in A$. The computations at intermediate levels are unnecessary which very much simplifies the numerical evaluation. When we are concerned by the sensitivity of the rare-event probability, it takes the simple form

$$\begin{aligned}\partial_\theta \left[\log\left(\mathbb{P}\left(Z^\theta \in A\right)\right)\right] :=& \frac{\partial_\theta \mathbb{P}\left(Z^\theta \in A\right)}{\mathbb{P}\left(Z^\theta \in A\right)} \\ =& \mathbb{E}\left(\mathcal{I}(Z^\theta, 1) \mid Z^\theta \in A\right).\end{aligned} \tag{5.7.5}$$

In full generality on the probabilistic setting, the determination of $\mathcal{I}(Z^\theta, \Phi^\theta)$ is difficult but in our Gaussian noise setting, it can be achieved using the Integration by Parts formula of Malliavin calculus. There are numerous situations where one can obtain such a representation for sensitivities (see [55, 63, 65, 88] among others, and [104, Section 6.2] for more references). We establish such a result in the case Z^θ takes values in $\mathbb{R}^d$, and Z^θ, Φ^θ are smooth in θ. Hereafter, we adopt and follow the notation of [104] for the derivative operator D, for the space $\mathbf{D}^{1,2}$ of random variables that are one time Malliavin differentiable with $_2$-integrability, and for the divergence operator δ. We say that a family of random variables $(U^\theta : \theta \in \Theta)$ is in $_p^{loc}$ $(p \geqslant 1)$ if for any $\theta \in \Theta$, there is a open set $V_\theta \subset \Theta$ containing θ such that $\sup_{\theta' \in V_\theta} |U^{\theta'}|$ is bounded by a random variable in $_p$.

Theorem 5.7.2. *Consider $\mathbf{Z} = \mathbb{R}^d$ and let $q > d$. Assume the following conditions:*

(a) *$(\Phi^\theta, \theta \in \Theta)$ is in $_2^{loc}$ and Z^θ has a $_q$-norm bounded locally uniformly in θ;*

(b) *Φ^θ and Z^θ are continuous and differentiable on Θ and their derivatives $(\dot{\Phi}^\theta, \dot{Z}^\theta : \theta \in \Theta)$ are respectively in $_1^{loc}$ and $_2^{loc}$;*

(c) *for any $\theta \in \Theta$, $Z^\theta \in \mathbf{D}^{1,2}$ and the Malliavin covariance matrix $\gamma_{Z^\theta} := (\langle D.Z_i^\theta, D.Z_j^\theta \rangle_{\mathcal{H}})_{1 \leqslant i,j \leqslant d}$ is invertible a.s.;*

(d) *for any* $\theta \in \Theta$, $\Phi^\theta \sum_{j=1}^d (\gamma_{Z^\theta}^{-1} \dot{Z}^\theta)_j D.Z_j^\theta$ *is in the domain of* δ *and* $\dot{\Phi}^\theta + \delta(\Phi^\theta \sum_{j=1}^d (\gamma_{Z^\theta}^{-1} \dot{Z}^\theta)_j D.Z_j^\theta)$ *has a* $_2$*-norm bounded locally uniformly in* θ*;*

(e) *for any* $\theta \in \Theta$ *and any* $i \in \{1, \dots, d\}$, $\sum_{j=1}^d (\gamma_{Z^\theta}^{-1})_{j,i} D.Z_j^\theta$ *is in the domain of* δ *and* $\delta(\sum_{j=1}^d (\gamma_{Z^\theta}^{-1})_{j,i} D.Z_j^\theta)$ *has a* $_q$*-norm bounded locally uniformly in* θ*.*

Then **(IBP)** *is satisfied on* Θ *and*

$$\mathcal{I}(Z^\theta, \Phi^\theta) := \dot{\Phi}^\theta + \delta\left(\Phi^\theta \sum_{j=1}^d (\gamma_{Z^\theta}^{-1} \dot{Z}^\theta)_j D.Z_j^\theta\right).$$

Proof. The proof follows a standard routine inspired by [55, 63, 65, 88] but it requires a careful analysis because of the indicator function. Firstly, properly mollify the indicator function $z \to \mathbf{1}_{\varphi(z,\bar{a}) \leqslant 0}$. Secondly, compute the derivative of the expectation for the mollified function, then, integrate by parts and take the limit w.r.t. the mollified parameter. Mollifying and passing to the limit is the critical part. In [88, Section 6], it has been done for functions which are almost everywhere continuous. Here we do not impose such restrictions.

Step 1. Let us define the measure $\bar{\mu}(\mathrm{d}z) = (1 + |z|)^{-q} \mathrm{d}z$ on $\mathbb{R}^d$ with q as in the statement and as $q > d$, this is a finite measure. Since $\mathbf{1}_A$ is in $_4(\mu)$, there is a sequence $(\xi_k)_{k \in \mathbb{N}}$ of smooth functions with compact support, such that

$$\int_{\mathbb{R}^d} |\mathbf{1}_{z \in A} - \xi_k(z)|^4 (1 + |z|)^{-q} \mathrm{d}z \underset{k \to +\infty}{\longrightarrow} 0. \tag{5.7.6}$$

W.l.o.g. we assume that $0 \leqslant \xi_k \leqslant 1$. Now, define

$$\begin{aligned} u_k(\theta) &:= \mathbb{E}\left(\Phi^\theta \xi_k(Z^\theta)\right), \\ u(\theta) &:= \mathbb{E}\left(\Phi^\theta \mathbf{1}_{Z^\theta \in A}\right), \\ v_k(\theta) &:= \mathbb{E}\left(\mathcal{I}(Z^\theta, \Phi^\theta) \xi_k(Z^\theta)\right), \\ v(\theta) &:= \mathbb{E}\left(\mathcal{I}(Z^\theta, \Phi^\theta) \mathbf{1}_{Z^\theta \in A}\right). \end{aligned}$$

Going forward, we shall establish three results. Firstly,

$$u_k(\theta) \underset{k \to +\infty}{\longrightarrow} u(\theta)$$

for any $\theta \in \Theta$, then, $u_k'(\theta) = v_k(\theta)$ for any $\theta \in \Theta$, and finally, v_k converges to v locally uniformly on Θ. By [45, Statement (8.6.4) Chap. VIII], this proves that u is differentiable on Θ and its derivative is v.

Step 2: Proof of $u_k'(\theta) = v_k(\theta)$**.** We can show

$$u_k'(\theta) = \partial_\theta \mathbb{E}\left(\Phi^\theta \xi_k(Z^\theta)\right) = \mathbb{E}\left(\dot{\Phi}^\theta \xi_k(Z^\theta)\right) + \mathbb{E}\left(\Phi^\theta \sum_{i=1}^d \partial_{z_i} \xi_k(Z^\theta) \dot{Z}_i^\theta\right),$$

from the dominated convergence theorem using the boundedness of $\xi_k, \nabla\xi_k$ and the uniform controls in the assumptions (a)-(b). Further, by the chain rule property, $\xi_k(Z^\theta) \in \mathbf{D}^{1,2}$ with $D.[\xi_k(Z^\theta)] = \sum_{i=1}^d \partial_{z_i}\xi_k(Z^\theta)D.Z_i^\theta$. Moreover by definition of δ as the adjoint operator of D, we have

$$\begin{aligned} v_k(\theta) &= \mathbb{E}\left(\dot{\Phi}^\theta \xi_k(Z^\theta) + \langle \sum_{i=1}^d \partial_{z_i}\xi_k(Z^\theta)D.Z_i^\theta, \Phi^\theta \sum_{j=1}^d (\gamma_{Z^\theta}^{-1}\dot{Z}^\theta)_j D.Z_j^\theta\rangle_{\mathcal{H}}\right) \\ &= \mathbb{E}\left(\dot{\Phi}^\theta \xi_k(Z^\theta) + \Phi^\theta \sum_{i=1}^d \partial_{z_i}\xi_k(Z^\theta) \sum_{j=1}^d (\gamma_{Z^\theta})_{i,j}(\gamma_{Z^\theta}^{-1}\dot{Z}^\theta)_j\right) \\ &= \mathbb{E}\left(\dot{\Phi}^\theta \xi_k(Z^\theta) + \Phi^\theta \sum_{i=1}^d \partial_{z_i}\xi_k(Z^\theta)\dot{Z}_i^\theta\right) \\ &= u_k'(\theta). \end{aligned}$$

Step 3: Proof of $(u_k, v_k) \underset{k\to+\infty}{\longrightarrow} (u, v)$ locally uniformly on Θ. Assume for a while the $_2$-convergence

$$\mathbb{E}\left(|\xi_k(Z^\theta) - \mathbf{1}_{Z^\theta\in A}|^2\right) \underset{k\to+\infty}{\longrightarrow} 0 \quad \text{locally uniformly in } \theta \in \Theta. \tag{5.7.7}$$

Then from above and (d), we deduce that for any $\theta \in \Theta$, there is an open set $V \subset \theta$ such that

$$|v_k(\theta) - v(\theta)| \leqslant \sup_{\theta\in V}\left|\dot{\Phi}^\theta + \delta(\Phi^\theta \sum_{j=1}^d (\gamma_{Z^\theta}^{-1}\dot{Z}^\theta)_j D.Z_j^\theta)\right|_2 \sup_{\theta\in V}\left|\xi_k(Z^\theta) - \mathbf{1}_{Z^\theta\in A}\right|_2 \underset{k\to+\infty}{\longrightarrow} 0$$

The same arguments apply for $u_k - u$. Consequently, it remains to justify (5.7.7).

Under the assumption (e), we have the integration by parts formula at order 1 (derived as in the proof of Step 2), i.e. for any smooth function ζ with compact support and any $i \in \{1, \dots, d\}$,

$$\mathbb{E}\left(\partial_{z_i}\zeta(Z^\theta)\right) = \mathbb{E}\left(\zeta(Z^\theta)\delta\left(\sum_{j=1}^d (\gamma_{Z^\theta}^{-1})_{j,i} D.Z_j^\theta\right)\right).$$

Therefore, from [122, Theorem 5.4] the distribution of Z^θ has a continuous density $p_{Z^\theta}(.)$ w.r.t. the Lebesgue measure, which is uniformly bounded by a function depending only on the $_q$-norms of $\delta\left(\sum_{j=1}^d (\gamma_{Z^\theta}^{-1})_{j,i} D.Z_j^\theta\right), 1 \leqslant$

$i \leqslant d$. In view of (a)-(e), we deduce that for any $\theta \in \Theta$, there is a neighborhood $V \subset \Theta$ of θ such that $\sup_{\theta' \in V} |p_{Z^{\theta'}}|_\infty := C_V < +\infty$, and

$$\begin{aligned}
&\mathbb{E}\left(|\xi_k(Z^{\theta'}) - \mathbf{1}_{Z^{\theta'} \in A}|^2\right) \\
&\leqslant \left(\mathbb{E}\left(|\xi_k(Z^{\theta'}) - \mathbf{1}_{Z^{\theta'} \in A}|^4 (1 + |Z^{\theta'}|)^{-q}\right)\right)^{1/2} \left(\mathbb{E}\left((1 + |Z^{\theta'}|)^q\right)\right)^{1/2} \\
&\leqslant \left(\int_{\mathbb{R}^d} |\xi_k(z) - \mathbf{1}_{z \in A}|^4 (1 + |z|)^{-q} C_V \mathrm{d}z\right)^{1/2} \sup_{\theta' \in V} \left(\mathbb{E}\left((1 + |Z^{\theta'}|)^q\right)\right)^{1/2}.
\end{aligned}$$

Owing to (5.7.6), the above converges to 0 as $k \to +\infty$, uniformly w.r.t. $\theta' \in V$, and (5.7.7) is proved. □

5.8 Rare event sampling and stress test

Besides computing all kinds of rare event statistics, sometimes we may also want to pick some rare event samples. We can not just apply plain Monte Carlo method with rejection because a huge number of simulation will be needed to just have one rare event realization. Inpired by our POP method, in order to make rare event sampling, we can use a well-known result for positive Harris recurrent Markov chain, which is that if the chain is in addition aperiodic, then its marginal distribution converges to its stationary distribution (see for example [99, Theorem 13.0.1]).

Remark that in all the finite dimensional cases, the existence of an implicit positive transition density p ensures that our Markov chain is aperiodic.

We have the following result.

Proposition 5.8.1. *For any fixed $k \in \{0, \dots, n-1\}$, denote by $\mathcal{L}(X_{k,N}^{x_{k,0}})$ the law of $X_{k,N}^{x_{k,0}}$ with initialization at a given point $X_{k,0} = x_{k,0} \in \Psi_Z^{-1}(A_k)$, and denote the distribution of X conditionally on $\{\Psi_Z(X) \in A_k\}$ by π_k. Then, for any $x_{k,0} \in \Psi_Z^{-1}(A_k)$ we have*

$$\|\mathcal{L}(X_{k,N}^{x_{k,0}}) - \pi_k\|_{\mathrm{TV}} \to 0$$

as $N \to +\infty$, where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm.

If some additional conditions are satisfied, we can have explicit convergence rate, such as in Theorem 4.1.6.

The convergence of marginal distributions may have interesting practical use. For example, let X denote the financial random environment that a banking system faces, and Z denote the related risk exposure. In order to test the system resilience, regulators usually design some stress test scenario which means imposing a presumably rare event in A on the banking system and then see how the system reacts to this event. Some references on the design of stress test can be found in [52]. In most stress testing designs, regulators artificially construct one or a few elements

in A. Using the POP method and Proposition 5.8.1, one can sample approximately according to the conditional distributions $X|Z \in A$ and/or $Z|Z \in A$, which gives a more relevant choice of stress test scenarios.

5.9 A variant of IPS method

We recall all the notations in the Subsection 5.2.2.

An intrinsic feature of IPS method is that different generations of particles are correlated with each other. Thus, if the empirical measure is inaccurate at the first generation, it is likely to be inaccurate at the following generation, which amplifies the variance of the final estimator. Here, we propose a way to reduce this dependency between generations. As previously seen, the shaking with rejection transformation is invariant with respect to the conditional distribution of X. Hence, if we apply several iterations of the transformation to obtain the next particles generation, the distribution of the system will be less influenced by the previous state. In order to keep the same computational cost, we reduce the size of the particle system. Thus, we run the proposed version of IPS algorithm with particle size $\lfloor \frac{M}{J} \rfloor$ and the transformation $\mathcal{M}_k^{\mathcal{K}}(X)$ applied J times at each time step. Regarding the convergence analysis, we can show that for a given J, the convergence still holds as $\frac{M}{J}$ goes to infinity. As we see in the numerical experiments, this variant of IPS method has a better performance, compared to the standard algorithm in Subsection 5.2.2 without extra resampling ($J = 1$)[4]. This idea of more iterations to gain more independence between particles have beem mentioned before, for example in [9] and [29]. We provide one numerical test for this variant in Subsection 6.8.2. As it is not the main focus of this thesis, we leave more numerical study on this interesting variant of IPS method for further research.

The modified IPS method follows the following procedure:

[4]We would like to thank Professor Tony LELIEVRE for bringing this variant of IPS method to our attention

Initialization:
For given integers J and M, set $M' = \lfloor \frac{M}{J} \rfloor$;
Draw $(X_0^{(M',m)}, m = 1, \cdots, M')$ which are i.i.d. copies of X ;
$p_0^{(M')} = \frac{1}{M'} \sum_{m=1}^{M'} \mathbf{1}_{A_1}(X_0^{(M',m)})$;
for $i = 0$ until $n-2$ **do**
  for $m = 1$ until M' **do**
    Selection step:
    if $X_i^{(M',m)} \in A_{i+1}$ **then**
      $\hat{X}_{i,0}^{(M',m)} = X_i^{(M',m)}$;
    else
      Pick $\hat{X}_{i,0}^{(M',m)}$ uniformly in $\{X_i^{(M',m)} \in A_{i+1}\}$ and independently of everything else;
    end
    Mutation step:
    for $j = 1$ until J **do**
      $\hat{X}_{i,j}^{(M',m)} = M_{i+1}^K(\hat{X}_{i,j-1}^{(M',m)}, Y_{i,j-1}^{(m)})$ where $Y_{i,j}^{(m)}$ are i.i.d copies of Y;
    end
    $X_{i+1}^{(M',m)} = \hat{X}_{i,J}^{(M',m)}$;
  end
  $p_{i+1}^{(M')} = \frac{1}{M'} \sum_{m=1}^{M'} \mathbf{1}_{A_{i+2}}(X_{i+1}^{(M',m)})$;
end
Result: $p^{(M')} = \prod_{i=0}^{n-1} p_i^{(M')}$

Algorithm 3: IPS method with extra resampling and reduced size for computing $\mathbb{P}(X \in A)$

5.10 Combine parallel and adaptive features in POP method

As we shall see in the numerical examples, POP method is efficient when provided with good choices of nested subsets, which is possible thanks to the aforementioned level adaptive version. However, the advantage of parallelization is lost. Here, we propose a variant with level refinement which allows to recover the possibility of parallelization.

To begin, we fix a threshold value q (for example $q = 0.05$) for each conditional probability under which the estimation using one Markov chain is considered insufficient, and we arbitrarily choose some nested subset A_k (such as an equi-distant partition of the entire space according to some criteria function). Then, we run one Markov chain as defined in Definition 5.2.3 at each level in parallel to estimate all the conditional probabilities $\mathbb{P}(X \in A_k \mid X \in A_{k-1})$. Next, we check if the estimator $p_k^{(N)}$ is larger than q and accept the estimate for each level for which it is true. Otherwise, if $p_{k_0}^{(N)} < q$ for some k_0, we change the algorithm

to the adaptive scheme as given in Section 5.6 to estimate $\mathbb{P}(X \in A_{k_0} \mid X \in A_{k_0-1})$. Notice that the Markov chain already used for level k_0 need not be simulated again. Another scheme is to put new nested subsets between A_{k_0} and A_{k_0-1} and run POP method again in parallel. Finally, the product of all the estimators provides an estimate of the rare event probability.

5.11 Black-Box feature of our methods

It is worth pointing out that, different from importance sampling techniques which depend much on model's particularity, all our methods presented above work with a Black-Box feature.

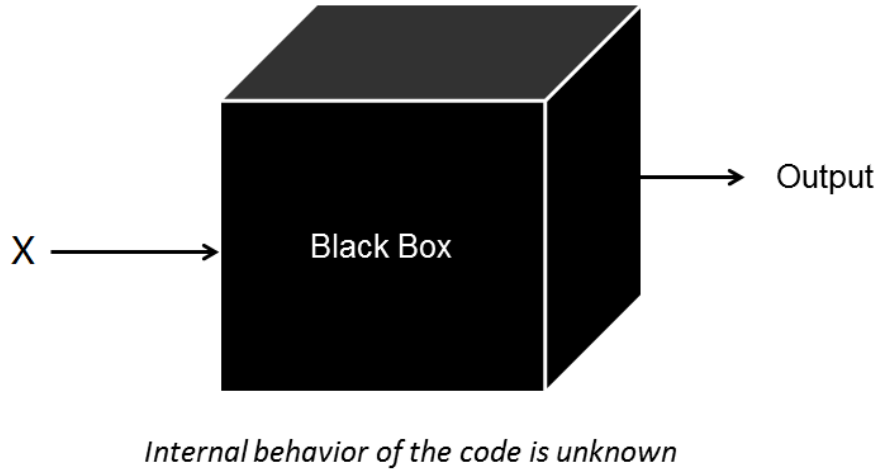


FIGURE 5.5: Black-Box system

As illustrated by Figure 5.5, our X can be the input of a Black-Box system, some commercial computation code for example, and the rare event is defined based on the output of this system. We apply the shaking transformation on the input X on the left side and thus make corresponding change on the output on the right hand side. In this way we completely ignore what happens inside the Black-Box. This makes our method applicable in very general situations.

As is demonstrated in the applications, even if the system is not given as a Black-Box, this feature can be very useful. Sometimes the explicit system involves very complicated dynamics, for example many correlations, such that it is not even easy to artificially design a rare event realization. Applying our method can help to overcome this problem.

Chapter 6

Applications

In this chapter, we will demonstrate the numerical performances of our methods using shaking transformations presented in Section 5.4.

The examples below are chosen according to their importance in applications and also because they are numerically challenging: we choose parameters so that rare-event probabilities are very small, from 10^{-5} to 10^{-10} or even less, moreover for some of them we can compute benchmark values using importance sampling techniques.

We do not provide theoretical results on the optimal choice of shaking transformation, but numerical results with different shaking parameters are given in order to discuss about the robustness of algorithms and to provide intuitive insights about these choices.

Besides, we have not optimized the choice of intermediate levels in these examples. When our methods are run with prefixed intermediate levels, some rough preliminary runs are done to make all conditional probabilities more or less of the same magnitude. However, observe that adding extra intermediate levels in POP can be done directly, without changing the estimation for other levels, thus preliminary runs are not necessary for POP; this is an advantage compared to IPS where one would to resimulate the whole particle system. When our methods are run with the adaptive versions, we pick a prefixed value for the quantile estimation, this value is not optimized either.

Another important remark concerns the memory. While for IPS one has to store all the particles (due to interactions), for POP only one particle per level needs to be stored which constitutes a large memory save.

Lastly we report the means and standard deviations of the algorithms outputs which are evaluated empirically by several runs (50 or 100 runs in most cases). We report the ratio `std/mean` which measures the relative error. Indeed, 50-100 macro-runs may be not sufficient to accurately estimate the standard deviation. However, provided that as M (and N) goes to $+\infty$, the renormalized errors of IPS and POP methods converge towards a Gaussian distribution. The algorithm output (with finite but large M and N) is approximately Gaussian, which supports an empirical estimation of the standard deviation with only 50-100 macro-runs. In addition, our aim is to highlight the impact of different shaking parameters and as we will see, with 50 or 100 runs these differences are already significantly clear.

Usually in POP method, one could use some burn-in time to reduce influence of the initial position of Markov chain. In all the applications using POP method, if there is no extra explanation, by convention we use the first 1 percent transitions as the burn-in time, except for the first level where no burn-in time is needed.

Discussions and comments in more details are given in each example.

6.1 One dimensional Ornstein-Uhlenbeck process

The OU process we consider is given by

$$\mathrm{d}Z_t = -Z_t \mathrm{d}t + \mathrm{d}G_t, \quad Z_0 = 0, \tag{6.1.1}$$

where G is a standard Brownian motion. It is in the form (5.3.2) and in the sequel, we apply the Brownian motion shaking (5.3.1) with constant ρ.

Actually, the following rare events are described in terms of the path of $(Z_t)_{0\leqslant t\leqslant T}$ with $T = 1$. Instead of an exact simulation, we simply use an Euler scheme $\tilde{Z}$ with time step $h := T/m$ for $m = 100$ and piecewise constant path approximation between the times $t_l := lh$. This discretization scheme does not alter significantly the performance of IPS and POP algorithms.

6.1.1 Maximum of OU process

Here the rare event is given by $\{\max_{0\leqslant l\leqslant m} \tilde{Z}_{t_l} > L\}$ with $L = 3.6$. Because of the mean reverting effect, the related probability is rather small. By 10^7 direct Monte Carlo simulations with importance sampling technique under the new probability $\mathrm{d}\mathbb{Q} = \exp(aG_T - \frac{1}{2}a^2T)\mathrm{d}\mathbb{P}$ where $a = 5$, we derive a 99% confidence interval of the requested probability $[0.977, 1.004]\times 10^{-7}$.

In (5.1.1) we take $n = 5$ intermediate sets associated to the levels $L_k = L\sqrt{\frac{k}{n}}, k = 1, \cdots, n$. In the experiments we report, we change the values of ρ, α N and M.

Results For the IPS and POP algorithms, we take respectively $M = 100000$ and $N = 100000$ so that the computational effort is similar. The following tables show results for different values of (α, ρ) for IPS and of ρ for POP. Output statistics (mean, standard deviation) are computed with 50 algorithm runs.

TABLE 6.1: mean, standard deviation and relative error of IPS method with $\alpha = 1$

IPS, $\alpha = 1$	mean	std	std/mean
$\rho = 0.9$	1.06×10^{-7}	5.12×10^{-8}	0.48
$\rho = 0.75$	9.51×10^{-8}	2.15×10^{-8}	0.22
$\rho = 0.5$	9.32×10^{-8}	9.42×10^{-8}	1.01

TABLE 6.2: mean, standard deviation and relative error of IPS method with $\alpha = 0.5$

IPS, $\alpha = 0.5$	mean	std	std/mean
$\rho = 0.9$	1.01×10^{-7}	3.67×10^{-8}	0.36
$\rho = 0.75$	9.81×10^{-8}	1.76×10^{-8}	0.18
$\rho = 0.5$	7.32×10^{-8}	9.18×10^{-8}	1.25

TABLE 6.3: mean, standard deviation and relative error of IPS method with $\alpha = 0$

IPS, $\alpha = 0$	mean	std	std/mean
$\rho = 0.9$	1.01×10^{-7}	3.94×10^{-8}	0.39
$\rho = 0.75$	9.98×10^{-8}	2.46×10^{-8}	0.25
$\rho = 0.5$	8.27×10^{-8}	1.18×10^{-7}	1.42

TABLE 6.4: mean, standard deviation and relative error of POP method

POP	mean	std	std/mean
$\rho = 0.9$	9.80×10^{-8}	6.74×10^{-9}	0.07
$\rho = 0.75$	1.00×10^{-7}	9.52×10^{-9}	0.10
$\rho = 0.5$	1.05×10^{-7}	2.78×10^{-8}	0.27

We first notice, by considering usual confidence intervals, that the probability is estimated coherently regarding the benchmark value (obtained by importance sampling). We note that POP has a better performance compared to IPS (see the column std/mean), whatever the value of ρ is. Regarding the variance, we observe that the impact of α (used for extra resampling) on IPS algorithm is not as significant as that of ρ, which is important for both IPS and POP. The above standard deviations are comparable to the one using importance sampling but our approaches have the advantage to work in a rather general setting. As for computational time, in MATLAB R2013a with Intel i7-4770 CPU 3.40GHz, one run of IPS with $M = 100000$ takes about 26 seconds while one run of POP with $N = 100000$ takes about 27 seconds. We also recall that POP is much more economic in memory since using POP requires to store only the current state of Markov chain while with IPS one needs to store the entire particle system.

In the following table, we report the average and standard deviation of each conditional probability estimator by POP method, based on 50 runs, as well as the averaged rejection rate for each level. Intuitively,

we expect methods in splitting spirit to have low variance when they are run with roughly constant conditional probabilities and that the optimality of a Metropolis-Hastings algorithm, whose spirit is shared by our shaking transformation, is related to the acceptance rate (for highly dimensional Gaussian distribution case, it is proved in [116] that the optimal acceptance rate is precisely 0.234). Thus, reporting these conditional probabilities and rejection rates is interesting to see how far the current implementation is from the optimal conditions, regarding the intermediate sets A_k and the shaking force parameter ρ.

TABLE 6.5: mean, standard deviation and rejection rate of POP method at level 1

POP, level1, $(\times 10^{-2})$	mean	std	rejection rate
$\rho = 0.9$	2.74	0.11	0
$\rho = 0.75$	2.74	0.08	0
$\rho = 0.5$	2.73	0.06	0

TABLE 6.6: mean, standard deviation and rejection rate of POP method at level 2

POP, level2, $(\times 10^{-2})$	mean	std	rejection rate
$\rho = 0.9$	4.07	0.10	42.40
$\rho = 0.75$	4.12	0.10	63.52
$\rho = 0.5$	4.16	0.14	82.40

TABLE 6.7: mean, standard deviation and rejection rate of POP method at level 3

POP, level3, $(\times 10^{-2})$	mean	std	rejection rate
$\rho = 0.9$	4.33	0.12	57.39
$\rho = 0.75$	4.37	0.15	80.60
$\rho = 0.5$	4.34	0.33	94.98

TABLE 6.8: mean, standard deviation and rejection rate of POP method at level 4

POP, level4, $(\times 10^{-2})$	mean	std	rejection rate
$\rho = 0.9$	4.47	0.11	67.15
$\rho = 0.75$	4.44	0.19	89.03
$\rho = 0.5$	4.50	0.51	98.44

TABLE 6.9: mean, standard deviation and rejection rate of POP method at level 5

POP, level5, $(\times 10^{-2})$	mean	std	rejection rate
$\rho = 0.9$	4.53	0.10	74.19
$\rho = 0.75$	4.56	0.19	93.59
$\rho = 0.5$	4.66	1.08	99.50

6.1.2 Oscillation of OU process

Now the rare event is associated to a large oscillation of the OU process, i.e., we compute

$$\mathbb{P}\left(\max_{0\leqslant l\leqslant m}\tilde{Z}_{t_l} > L \text{ and } \min_{0\leqslant l\leqslant m}\tilde{Z}_{t_l} < -L\right)$$

with $L = 1.6$. Note that in this situation *standard* importance sampling techniques with shifted Brownian motion do not work any more. By a crude Monte Carlo algorithm with 7×10^9 simulations, we obtain a 99% confidence interval equal to $[3.97, 4.37] \times 10^{-7}$.

In our IPS and POP approaches, we simply take $L_k = L\sqrt{\frac{k}{5}}$ for $k = 1, \ldots, 5$ and define intermediate events as

$$\{\max_{0\leqslant l\leqslant m}\tilde{Z}_{t_l} > L_k \text{ and } \min_{0\leqslant l\leqslant m}\tilde{Z}_{t_l} < -L_k\}.$$

Results In the following tables the empirical results of IPS and POP algorithms are computed over 100 experiments, respectively with $M = 100000$ and $N = 100000$.

TABLE 6.10: mean, standard deviation and relative error of IPS method with $\alpha = 1$

IPS, $\alpha = 1$	mean	std	std/mean
$\rho = 0.9$	4.01×10^{-7}	1.23×10^{-7}	0.31
$\rho = 0.75$	4.10×10^{-7}	1.67×10^{-7}	0.41
$\rho = 0.5$	2.44×10^{-7}	4.76×10^{-7}	1.95

TABLE 6.11: mean, standard deviation and relative error of IPS method with $\alpha = 0.5$

IPS, $\alpha = 0.5$	mean	std	std/mean
$\rho = 0.9$	3.94×10^{-7}	1.08×10^{-7}	0.27
$\rho = 0.75$	4.12×10^{-7}	1.89×10^{-7}	0.46
$\rho = 0.5$	3.41×10^{-7}	9.89×10^{-7}	2.90

TABLE 6.12: mean, standard deviation and relative error of IPS method with $\alpha = 0$

IPS, $\alpha = 0$	mean	std	std/mean
$\rho = 0.9$	4.18×10^{-7}	1.08×10^{-7}	0.26
$\rho = 0.75$	4.20×10^{-7}	2.02×10^{-7}	0.48
$\rho = 0.5$	2.66×10^{-7}	4.61×10^{-7}	1.73

TABLE 6.13: mean, standard deviation and relative error of POP method

POP	mean	std	std/mean
$\rho = 0.9$	4.14×10^{-7}	2.68×10^{-8}	0.06
$\rho = 0.75$	4.18×10^{-7}	4.60×10^{-8}	0.11
$\rho = 0.5$	4.29×10^{-7}	1.26×10^{-7}	0.29

In the following table, we report the average and standard deviation of each conditional probability estimated by POP, based on 100 runs, as well as the averaged rejection rate for each level.

TABLE 6.14: mean, standard deviation and rejection rate of POP method at level 1

POP, level1, $(\times10^{-2})$	mean	std	rejection rate
$\rho = 0.9$	4.27	0.12	0
$\rho = 0.75$	4.25	0.08	0
$\rho = 0.5$	4.25	0.07	0

TABLE 6.15: mean, standard deviation and rejection rate of POP method at level 2

POP, level2, $(\times10^{-2})$	mean	std	rejection rate
$\rho = 0.9$	4.95	0.13	57.56
$\rho = 0.75$	4.96	0.15	76.15
$\rho = 0.5$	4.91	0.22	88.97

TABLE 6.16: mean, standard deviation and rejection rate of POP method at level 3

POP, level3, $(\times10^{-2})$	mean	std	rejection rate
$\rho = 0.9$	5.55	0.17	71.10
$\rho = 0.75$	5.53	0.23	88.55
$\rho = 0.5$	5.52	0.45	97.15

TABLE 6.17: mean, standard deviation and rejection rate of POP method at level 4

POP, level4, $(\times10^{-2})$	mean	std	rejection rate
$\rho = 0.9$	5.85	0.17	78.86
$\rho = 0.75$	5.85	0.33	93.83
$\rho = 0.5$	5.78	0.86	99.12

TABLE 6.18: mean, standard deviation and rejection rate of POP method at level 5

POP, level5, $(\times10^{-2})$	mean	std	rejection rate
$\rho = 0.9$	6.05	0.20	83.95
$\rho = 0.75$	5.95	0.43	96.50
$\rho = 0.5$	5.68	1.56	99.71

As before, both algorithms seemingly converge, with better results for POP (the std for POP is about 4-5 times smaller than for IPS). Here again, the value of α has less impact than the value of ρ on the variance. We do not investigate further on the optimality of α. In all the following IPS algorithms, we fix α equal to 1 (i.e. we skip the resampling step). For computational time, with the same configuration as in previous example, one run of both IPS and POP with $M, N = 100000$ takes about 28 seconds. In the next examples, we do not compare anymore the computational times since they are quite the same for IPS and POP.

In Figure 6.1, we show empirical variances of 100 experiments results for M and N respectively equal to 100000, 10000, 5000, 3000 and 2000. These variances are not perfectly estimated since we use only 100 runs. Nevertheless, we approximately obtain a linear convergence with respect to $1/M$ and $1/N$, as expected from theoretical results (see Theorems 5.2.3 and 5.2.4).

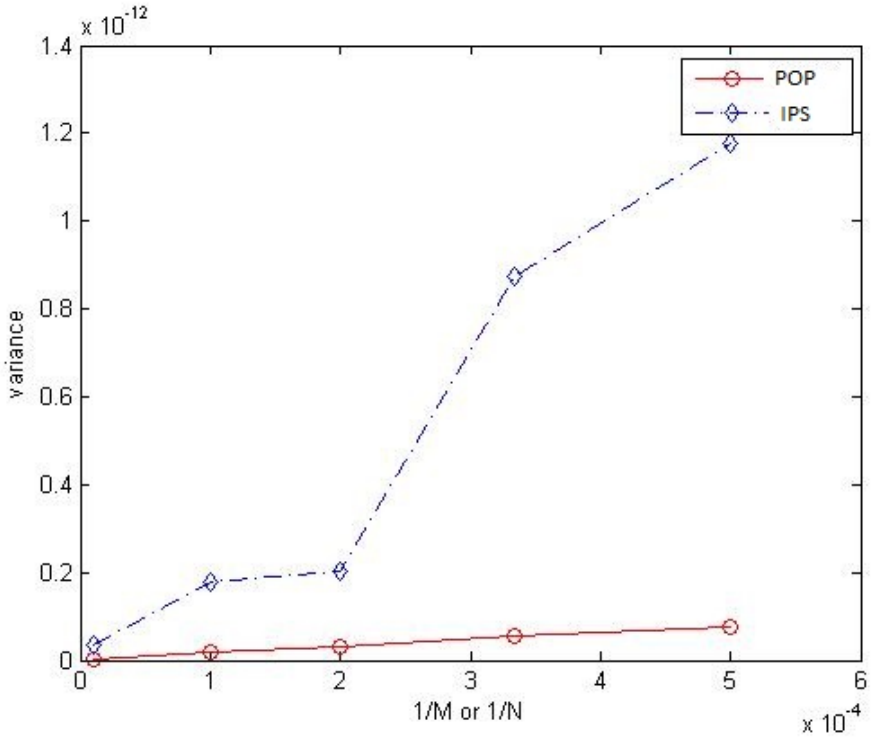


FIGURE 6.1: Variance for IPS and POP methods as a function of $1/M$ and $1/N$ respectively

6.2 Insurance

The capital reserve of an insurance company is modeled by

$$R_t = x + ct - \sum_{k=1}^{N_t} Z_k$$

where x is the initial reserve, c is the premium rate, N is a Poisson process with intensity λ and $(Z_k)_k$ are amounts of claims in case of accident or natural disaster [5]. In the following example, we take $c = 1, \lambda = 0.005, x = 100, T = 1$ and suppose $(Z_k)_k$ are Gamma variables with parameters $(a, b) = (2.5, 0.12)$. We aim at computing the probability of bankruptcy before T, i.e. $\mathbb{P}\left(\min_{0\leqslant t\leqslant T} R_t < 0\right)$.

Using Esscher transformation, we get the 99% confidence interval for this probability: $[1.042, 1.188] \times 10^{-6}$ through 10^5 Monte Carlo simulations under the new probability

$$\mathrm{d}\mathbb{Q} = \exp(\sum_{k=1}^{N_T} f(Z_k) - \int_{\mathbb{R}} (e^{f(y)} - 1)\lambda T \nu(\mathrm{d}y))\mathrm{d}\mathbb{P}$$

where $f(y) = 0.09y$ and $\nu(\mathrm{d}y)$ is the probability measure of $\texttt{Gamma}(a, b)$. We can easily check that the distribution of Z_k is still of Gamma type under this new probability.

We take $n = 5$ intermediate levels, defined by $L_k = x(1 - (\frac{k}{5})^2)$ for $k = 1, \ldots, 5$.

IPS algorithm We apply the partial shaking to the jump sizes (and not to the jump times), i.e. we shake all $(Z_k)_k$ with the shaking transformations for Gamma variables (with parameter p), then we get the following results ($M = 10000$, over 100 times experiments).

	$p = 0.1$	$p = 0.2$	$p = 0.3$
mean	1.25×10^{-6}	1.11×10^{-6}	1.01×10^{-6}
std	2.82×10^{-6}	1.30×10^{-6}	6.46×10^{-7}
std/mean	2.26	1.17	0.64

	$p = 0.4$	$p = 0.5$	$p = 0.6$
mean	1.02×10^{-6}	1.15×10^{-6}	1.09×10^{-6}
std	8.39×10^{-7}	5.15×10^{-7}	4.11×10^{-7}
std/mean	0.82	0.45	0.38

With the Poisson process decomposition shaking with parameter p (M is equal to 10000, over 100 times experiments), results become as follows.

	$p = 0.1$	$p = 0.2$	$p = 0.3$
mean	3.10×10^{-6}	2.02×10^{-6}	2.93×10^{-6}
std	1.76×10^{-5}	1.39×10^{-5}	1.65×10^{-5}
std/mean	5.68	6.85	5.62

	$p = 0.4$	$p = 0.5$	$p = 0.6$
mean	8.63×10^{-8}	9.32×10^{-7}	2.08×10^{-6}
std	1.32×10^{-7}	8.68×10^{-6}	1.42×10^{-5}
std/mean	1.53	9.32	6.81

We observe that Poisson shaking can not even produce a good mean value and that partial shaking on Gamma variables is much better than Poisson decomposition shaking. This can be explained as follows: in this particular insurance reserve example where there are very few jumps with important jump sizes, the Poisson shaking gives large perturbation of the system (opposite to the spirit of slight shaking), since by removing a jump time (and therefore the claim amount at this instant), this may completely change the situation of the company, from being close to bankruptcy to running with good profit.

Obviously, partial shaking involving Gamma variables doesn't cause this kind of problem since we keep every jump time and only modify claim amount. In this sense, the Gamma shaking is more continuous and better suits this example.

Shaking all the inter-arrival and jump variables yields the following results (over 100 experiments with $M = 10000$), which gives larger variance than for Gamma shaking only, as expected.

	$p = 0.1$	$p = 0.2$	$p = 0.3$
mean	9.75×10^{-7}	9.35×10^{-7}	1.22×10^{-6}
std	4.73×10^{-6}	3.63×10^{-6}	7.14×10^{-6}
std/mean	4.85	3.89	5.87

	$p = 0.4$	$p = 0.5$	$p = 0.6$
mean	2.97×10^{-7}	9.75×10^{-7}	1.10×10^{-6}
std	5.90×10^{-7}	8.15×10^{-6}	9.80×10^{-6}
std/mean	1.99	8.36	8.94

POP algorithm When using Poisson shaking or Gamma shaking for our POP algorithm, we have observed that both of them fail. The reason for Poisson shaking is similar to IPS case. As for Gamma shaking, the difference between IPS and POP is that, in IPS we sample M trajectories and we pick those with jumps, while in POP algorithm we have only one trajectory and (in this insurance example) the initial configuration may have no jump with a large probability, yielding that the output of POP algorithm is destined to be 0.

To retrieve good convergence properties, we simply apply the shaking for inter-arrival and jump variables and we get the following results

(over 100 experiments with $M = 10000$), which are slightly more accurate than IPS.

	$p = 0.1$	$p = 0.2$	$p = 0.3$
mean	1.14×10^{-6}	1.11×10^{-6}	1.12×10^{-6}
std	5.08×10^{-7}	4.44×10^{-7}	4.80×10^{-7}
std/mean	0.45	0.40	0.43

	$p = 0.4$	$p = 0.5$	$p = 0.6$
mean	1.05×10^{-6}	1.12×10^{-6}	9.29×10^{-7}
std	6.74×10^{-7}	8.24×10^{-7}	9.52×10^{-7}
std/mean	0.64	0.74	1.02

6.3 Queuing system

Suppose we have a 2-nodes Jackson network (see [115, Chapter 4] for definition and [20] for related numerical algorithms). All the costumers arrive at node 1 and when they are served they go to node 2. The customers' arrival times are jump times of a Poisson process with intensity λ. The serving time at node 1 and at node 2 are respectively exponential variables with parameters μ_1 and μ_2. Our purpose is to compute the probability that at some time before T, the number of customers in the system reaches a fixed level K, i.e. $\mathbb{P}(\max_{0\leqslant t\leqslant T} M_t > K)$ where M_t denotes the number of customers in the system at time t.

Given the Poisson process N representing customers' arrival time, we define two compound Poisson processes

$$Z_t^A = \sum_{k=1}^{N_t} A_k, \qquad Z_t^B = \sum_{k=1}^{N_t} B_k,$$

to which we apply shaking transformations. Here A_k and B_k are respectively the serving times of k-th customer at node 1 and at node 2. We now claim that $\max_{0\leqslant t\leqslant T} M_t = \Phi((Z_t^A)_{0\leqslant t\leqslant T}, (Z_t^B)_{0\leqslant t\leqslant T})$ for a functional Φ, this representation will be the basis for our algorithms. To justify this, denote by a_k the arrival time of the k-th customer (i.e. the k-th jump of N). Then if we note by e_k^1 the instant when the service for k-th customer at node 1 is finished, we can find the following recursive relation:

$$e_{k+1}^1 = \max(a_{k+1}, e_k^1) + A_{k+1}$$

with the initial condition $e_1^1 = a_1 + A_1$. Remark that the service finishing time at node 1 is the customer arrival time at node 2, we have the same recursive relation for e_k^2, the instants when service for k-th customer at

node 2 is finished:

$$e^2_{k+1} = \max(e^1_{k+1}, e^2_k) + B_{k+1}$$

with the initial condition $e^2_1 = e^1_1 + B_1$. Then when the k-th customer enters the system, the number of customers in the system is

$$k - \#\{e^2_j : e^2_j < a_k\}$$

Since the maximal number of customers in the system is possibly reached only when a new customer enters the system we have

$$\max_{0\leqslant t\leqslant T} M_t = \max_{0\leqslant k\leqslant N_T} (k - \#\{e^2_j : e^2_j < a_k\})$$

which leads to our claim that the maximal number of customers in the system before T is determined by the two CPP's $(Z^A_t)_{0\leqslant t\leqslant T}$ and $(Z^B_t)_{0\leqslant t\leqslant T}$.

We take $\lambda = 0.5, \mu_1 = 1, \mu_2 = 1, T = 10$ and $n = 10$ intermediate levels defined as $L_k = K\sqrt{\frac{k}{n}}, k = 1, \cdots, n$. We set $K = 20$. For the benchmark value, we use an importance sampling method (with 10^7 simulations) based on Esscher transformation using the new probability

$$\mathrm{d}\mathbb{Q} = \exp(cN_T - (e^c - 1)\lambda T)\mathrm{d}\mathbb{P}$$

where $c = 1.5$: the resulting 99% confidence interval for $\mathbb{P}(\max_{0\leqslant t\leqslant T} M_t > K)$ is $[4.6380, 5.1210] \times 10^{-10}$. The shaking transformation we use here is defined in (5.4.1), with different values of p.

Results The following IPS and POP results are computed with $M = 10000$ and $N = 10000$ respectively, over 100 times experiments with each parameter.

IPS	$p = 0.1$	$p = 0.3$	$p = 0.5$
mean	4.35×10^{-10}	5.04×10^{-10}	5.58×10^{-10}
std	5.33×10^{-10}	4.26×10^{-10}	2.00×10^{-9}
std/mean	1.23	0.84	3.58

POP	$p = 0.1$	$p = 0.3$	$p = 0.5$
mean	4.93×10^{-10}	5.24×10^{-10}	5.27×10^{-10}
std	1.33×10^{-10}	2.09×10^{-10}	4.62×10^{-10}
std/mean	0.27	0.40	0.88

The POP algorithm provides more accurate results than IPS, and seems more stable as p is modified. If we only shake service times A and B instead of the Poisson process Z^A and Z^B (as in (5.4.1)), both algorithms fail to work, almost systematically the output of algorithm is

0. This is not surprising since by shaking the service time, we will never increase the number of clients in the system.

The POP method has been tested in the case of renewal process where inter-arrival and service times are uniformly distributed. The performance in that case is also good.

6.4 Random graph

A Erdös-Rényi random graph [21] is a graph with V vertices where every pair of vertices are connected with probability q, independently of the others. It constitutes a toy model for the study of social networks, epidemics etc. The graph is presented by the upper triangular matrix $X := (X_{ij})_{1 \leqslant i < j \leqslant V}$ where

$$X_{ij} = \begin{cases} 1, & \text{if vertices } i \text{ and } j \text{ are connected} \\ 0, & \text{otherwise} \end{cases}$$

If vertices i, j and k are all connected to each other, they form a triangle. Thus the number of triangles in the graph is given by

$$T(X) := \sum_{1 \leqslant i < j < k \leqslant V} X_{ij} X_{jk} X_{ik}.$$

We easily check that $\mathbb{E}(T(X)) = \frac{V(V-1)(V-2)}{6} q^3$ and as a rare event, we consider the deviation event

$$\{T(X) > \frac{V(V-1)(V-2)}{6} t^3\}$$

for $t > q$. This problem has deserved recent interest in [35] with theoretical results and in [16] with numerical computations based on importance sampling techniques.

The total number of possible connections is $\frac{V(V-1)}{2}$ and may be rather large even for small graphs. In our case we take $V = 64$, $q = 0.35$ and $t = 0.4$: the corresponding estimation given in [16] is about 2.19×10^{-6}. To reduce the complexity of IPS and POP algorithms, we use the technique of partial shaking, by picking randomly a proportion c of X_{ij} and shake them independently. Regarding the reversible shaking transformation of each Bernoulli random variable X_{ij}, the only possibility is described by a transition matrix $P(x, y)$ ($x, y \in \{0, 1\}^2$) which satisfies the following condition

$$qP(1, 0) = (1 - q)P(0, 1),$$

i.e. $P(0, 1) = \frac{q}{1-q} P(1, 0)$. Since in this example $\frac{q}{1-q} < 1$, $P(1, 0)$ can be any value in $[0, 1]$ and it parametrizes the force of shaking. The larger the value of $P(1, 0)$, the more important the change in the graph configuration. Numerical results are performed with $n = 5$ intermediate levels

given by

$$L_k = \frac{V(V-1)(V-2)}{6} t^3 (\frac{k}{5})^{\frac{1}{5}}$$

with $k = 1, \dots, n$.

Results First, we take $c = 10\%$ and statistics are computed over 50 algorithm experiments. For IPS and POP algorithms, we take respectively $M = 10000$ and $N = 10000$ and we obtain the following results.

IPS - $P(1,0)$	0.25	0.5	0.75	1
mean	1.79×10^{-6}	1.83×10^{-6}	1.92×10^{-6}	2.10×10^{-6}
std	2.29×10^{-6}	1.30×10^{-6}	1.04×10^{-6}	8.79×10^{-7}
std/mean	1.28	0.71	0.54	0.42

POP - $P(1,0)$	0.25	0.5	0.75	1
mean	2.15×10^{-6}	2.05×10^{-6}	2.06×10^{-6}	2.13×10^{-6}
std	5.76×10^{-7}	4.52×10^{-7}	3.23×10^{-7}	3.35×10^{-7}
std/mean	0.27	0.22	0.16	0.16

The performance of POP appears rather stable w.r.t. $P(1,0)$ and systematically better than the IPS method.

Secondly we can modify the value of c by keeping the product $Mc = Nc$ constant (the computational effort remains the same). Taking $c = 1\%$ yields less accurate results we do not report. In the opposite direction, taking $c = 100\%$ fails to work. The question of the best choice of c and $P(0,1)$ according to t, q, V is open.

6.5 Hawkes process

The Hawkes process [77] is a self-exciting counting process $(N_t)_{t\geqslant 0}$ whose intensity evolves as

$$\mathrm{d}\lambda_t = \theta(\mu - \lambda_t)\mathrm{d}t + \mathrm{d}N_t.$$

In the last years, it has become rather popular to model earthquakes activity, high-frequency financial data, information flow on internet (Twitter etc) etc. We guess that this is a challenging model for rare event simulation because of its self-exciting property. Here we set $\theta = 2$, $\mu = 1$, the terminal time $T = 24$ and $\lambda_0 = 1$. We denote all the jump instants before T by $(\tau_j)_{j\geqslant 1}$ and define

$$H = \max\{\tau_j - \tau_i : \tau_{k+1} - \tau_k < 0.5, i \leqslant k < j - 1\},$$

which is the longest period between jump instants during which all jump inter-arrivals are less than 0.5. Our aim is to estimate $\mathbb{P}(H > 11)$,

using 3×10^8 crude Monte Carlo simulations gives a 99% confidence interval $[3.2469, 3.8064] \times 10^{-6}$.

According to [105, Algorithm 2], Hawkes process (and thus H) can be seen as a functional of countable number of uniform variables in $[0, 1]$ which fits our general setting.[1]. Thus we can use the shaking transformation for uniform variables in our algorithms. We define $n = 5$ intermediate sets as $\{H > L_k\}$ where $(L_k)_{k=1,\cdots,5} = [3.5, 5.5, 7.5, 9.5, 11]$. Results over 50 experiments for different shaking coefficients are listed in the following (with $M = N = 10^4$).

IPS	$p = 0.1$	$p = 0.3$	$p = 0.5$
mean	3.30×10^{-6}	5.19×10^{-6}	3.88×10^{-6}
std	2.84×10^{-6}	1.37×10^{-5}	1.60×10^{-5}
std/mean	0.86	2.64	4.12

POP	$p = 0.1$	$p = 0.3$	$p = 0.5$
mean	3.33×10^{-6}	3.51×10^{-6}	2.69×10^{-6}
std	1.25×10^{-6}	2.92×10^{-6}	3.71×10^{-6}
std/mean	0.37	0.83	1.38

We observe good performance of POP (about three times more accurate than IPS). Both algorithms are much more accurate than the crude Monte Carlo method, as expected.

6.6 An example of randomized shaking transformation

We conclude this presentation of numerical experiments by illustrating the benefit of randomization of shaking parameter as explained in Lemma 5.4.1 of Subsection 5.4.4.

We consider the simple problem of estimating $\mathbb{P}(G > 6 \text{ or } G < -5)$, where $X := G$ is a standard Gaussian variable. Of course, one could compute respectively $\mathbb{P}(G > 6)$ and $\mathbb{P}(G < -5)$ then add them up. But this solution requires extra knowledge about the problem that we could not afford in general. Hence for the sake of exposition, we do not use this decomposition.

If we use the POP method on the initial problem with intermediate levels defined by

$$\{G > \sqrt{\frac{k}{5}} \times 6 \text{ or } G < -\sqrt{\frac{k}{5}} \times 5\}, k = 1, \cdots, 5,$$

[1] During implementation, we only need to keep record of uniform variables that have been used.

the results are rather unstable. Over 100 experiments with the shaking $G = \rho G + \sqrt{1-\rho^2}G'$ where $\rho = 0.75$, 23 outputs are of order 10^{-9} and the others are of order 10^{-7}. This is due to the fact that in POP method we average only one path. When shaking level after level, this path tends gradually either towards $\{G > 6\}$ or towards $\{G < -5\}$ and it becomes practically impossible to realize the jump from $\{G > \sqrt{\frac{k}{5}} \times 6\}$ to $\{G < -\sqrt{\frac{k}{5}} \times 5\}$. As a consequence, only one part of the distribution is selected and estimated[2]. The IPS approach is less sensitive to this problem since it is based on a large sample of paths.

To circumvent this problem for POP, we can take a random ρ such that $\rho = 0.75$ with probability 0.8 and $\rho = -0.75$ with probability 0.2: this enables the path to sometimes jump from $\{G > \sqrt{\frac{k}{5}} \times 6\}$ to $\{G < -\sqrt{\frac{k}{5}} \times 5\}$, thus to yield a better performance. Indeed over 100 experiments, with fixed ρ we get mean 2.84×10^{-7} and standard deviation 1.70×10^{-7}, while with the random ρ we get mean 2.81×10^{-7} and standard deviation 6.73×10^{-8}. We recall that

$$\mathbb{P}\left(G > 6 \text{ or } G < -5\right) = 2.8764\times10^{-7}.$$

In more general situations, randomization is certainly beneficial to explore disjoint configurations. The right tuning is a delicate question since too much randomization may alter the benefit of POP method. This issue is left to future investigation.

6.7 Model misspecification and robustness

To address the issue of model risk, we consider the Profit&Loss (PL) when the trader uses a Black-Scholes (BS) model to hedge a European call option while the true dynamics of the underlying S is given by a path-dependent volatility model. Let us suppose that there are two volatility levels $\sigma_-, \sigma_+ \in \mathbb{R}_+\backslash\{0\}$ such that $\sigma_- < \sigma_+$. We propose a discrete-time path-dependent volatility model based on a monitoring period Δ_t (say 1 week) and monitoring dates $t_i = i\Delta_t$, wherein, if the underlying spot price drops below the average of previous four monitored prices, the level of volatility becomes σ_+, otherwise it remains

[2]The same phenomenon occurs using importance sampling techniques and other splitting methods.

constant at σ_-. The asset price is given as

$$S_t = S_0 \exp\big(-\frac{1}{2}\sigma_-^2 t + \sigma_- W_t\big), \quad t < t_4, \tag{6.7.1}$$

$$S_t = \begin{cases} S_{t_i}\exp\big(-\frac{1}{2}\sigma_-^2(t-t_i) + \sigma_-(W_t - W_{t_i})\big) \\ \qquad \text{if } S_{t_i} \geq \frac{1}{4}\sum_{k=1}^4 S_{t_{i-k}} \text{ and } t_i \leqslant t < t_{i+1}, \\ S_{t_i}\exp\big(-\frac{1}{2}\sigma_+^2(t-t_i) + \sigma_+(W_t - W_{t_i})\big) \\ \qquad \text{if } S_{t_i} < \frac{1}{4}\sum_{k=1}^4 S_{t_{i-k}} \text{ and } t_i \leqslant t < t_{i+1}, \end{cases} \quad \text{when } t \geq t_4. \tag{6.7.2}$$

This model corresponds to the usual empirical observation that the underlying volatility is higher when price falls. This is a discrete version of the continuous time model proposed by [74]. Furthermore, we assume the risk-free interest rate to be zero. The resulting model is complete in the sense that any square integrable payoff written on S can be replicated by a self-financing strategy (see [79] for complete models with stochastic volatility). The above model is directly written under the risk-neutral measure $\mathbb{P}$.

Meanwhile, we assume that the trader uses a BS model in which the volatility is constant and equal to σ_-. The call option maturity is $T > 0$, and $[0, T]$ is the trading period under consideration. As the trader assumes a BS model, he/she uses the BS formula to perform delta hedging. For our numerical study, we take $T = 1, n = 50$ Δ_t is s.t. $n\Delta_t = T$) and assume that the trader makes a rebalancing after every period of $5\Delta_t$. At times $t_{5\Delta_t j}, 0 \leqslant j < 10$, the trader holds δ_j assets, so at the maturity his/her PL is given by

$$PL_{\text{trader}} := \mathbb{E}_{\text{trader}}\big[(S_T - K_{\text{strk}})_+\big] + \sum_{j=0}^{9} \delta_j (S_{5\Delta_t(j+1)} - S_{5\Delta_t j}) - (S_T - K_{\text{strk}})_+$$

where δ_j is given from the BS-Delta formula with volatility σ_- and spot $S_{5\Delta_t j}$.

Since the realized volatility is higher than the one used for hedging, the trader may underhedge the option (in continuous time hedging, see [50] for precise results) and may incur large losses due to the model risk. Thus, we wish to estimate the probability $\mathbb{P}\left(PL_{\text{trader}} \leqslant L\right)$.

6.7.1 Large loss probability

In the model of (6.7.1), we set $S_0 = K_{\text{strk}} = 10$, $\sigma_- = 0.2$, $\sigma_+ = 0.27$ and take $L = -2.4$. In our IPS and POP methods, we create the intermediate levels as $L_k := kL/5, k = 1, 2, 3, 4, 5$. The crude Monte Carlo method with 5×10^8 simulations provides a 99% confidence interval for this probability as $[2.93, 3.34] \times 10^{-6}$. The mean estimates and empirical standard deviations of non-adaptive and adaptive estimators based on IPS and POP methods using 100 macro-runs are given in Tables 6.19-6.20. The adaptive algorithms are performed with parameter $p = 10\%$

for the intermediate conditional probability. From Table 6.19, it is clear that POP based estimators provide accurate estimates with a lower standard deviation than IPS based estimators, both schemes being in their non-adaptive versions. In Table 6.20, results with adaptive algorithms are compared, here again POP method yields smaller variances in the estimation.

When comparing standard deviations of Table 6.19 and Table 6.20, we observe that they are similar (for a given shaker parameter ρ). The reader may think that seemingly adaptive versions do not provide any benefit. One should recall that, the non-adaptive versions require a priori choices of levels (here we choose them by preliminary experiments to have roughly equal conditional probabilities) while with the adaptive version we do not need this kind of a priori knowledge and still we obtain efficient estimators. Actually the advantage really stems from the fully adaptive tuning of levels, which is made possible without deteriorating the variances.

In Figure 6.2 (top), we investigate the dependence of the standard deviation (of each conditional probability computed with POP method) w.r.t. the shaking parameter ρ and the level l. We do not report results for $l = 1$ (no rejection) since independent sampling ($\rho = 0$) is obviously the best. We observe that the impact of ρ on the variance is significant: the optimal parameter ρ_l^* minimizing the variance changes from one level to another and ρ_l^* increases with l (the shaking has to become slighter with increasing rarity of the event). These features are easily explained heuristically. Complementary to this, we plot in Figure 6.2 (bottom) the rejection rate, which also depends on ρ and l. It appears that ρ_l^* depends very much on l whereas the associated rejection rate remains rather stable and ranges from 60% to 80%. Since we observe a quick explosion of standard deviation when ρ is too close to 1, we recommend to take ρ such that the rejection rate is above 60% rather than below 60%, to be on the safe side when a finer optimization of ρ is not possible. We shall take it as a rule of thumb for further experiments.

Lastly, in Figure 6.3 we report statistics on standard deviation and rejection rate for the adaptive POP method. We observe similar features as in Figure 6.2. Since different intermediate levels in the adaptive POP method are correlated, we first run the beginning level to get corresponding value of ρ minimizing the standard deviation. Then, we use this fixed value for the corresponding level in the search of optimal ρ of the next level and so on. We can see in Figure 6.3 that with all the values of ρ chosen in this way, the best standard deviation among the final estimators is around 1.5×10^{-7}, which is about 62.5% of standard deviation of the estimator with a constant $\rho = 0.9$ for all the levels.

If we compare Table 6.19 and Table 6.20, we will find the relative errors of non-adaptive and adptive methods are roughly the same. Then why do we need to design adaptive versions of our methods? In fact, these two tables show roughly the same errors because when we run the non-adptive versions of our methods, some preliminary runs are used

to gain knowledge about where to put the intermediate levels. These preliminary runs help to fix intermediary levels near the places found automatically by the adaptive methods. Therefore, the advantage of adaptive methods is that they can find automatically appropriate intermediate levels without preliminary runs, and without introducing extra variance on numerical results according to our tests.

	IPS			POP		
	mean $(\times 10^{-6})$	std. $(\times 10^{-7})$	std./mean	mean $(\times 10^{-6})$	std. $(\times 10^{-7})$	std./mean
$\rho = 0.9$	3.10	5.29	0.17	3.13	2.07	0.07
$\rho = 0.7$	3.23	13.3	0.41	3.11	3.98	0.13
$\rho = 0.5$	2.79	25.9	0.93	3.18	8.44	0.27

TABLE 6.19: *Estimators of $\mathbb{P}(PL_{trader} \leqslant L)$ (mean) for $L = -2.4$ with empirical standard deviation (std.) for non-adaptive IPS and POP methods based on 100 algorithm macro-runs. Each intermediate level estimator in both methods is based on $M = N = 10^5$ simulations.*

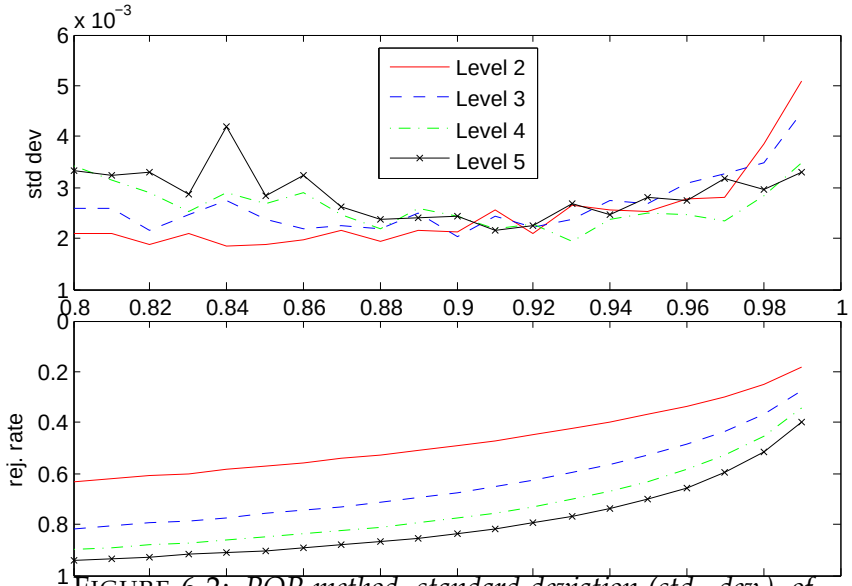


FIGURE 6.2: *POP method, standard deviation (std. dev.) of each conditional probability estimator and corresponding rejection rate (rej. rate), based on 100 macro-runs, for different values of ρ.*

	Adaptive IPS			Adaptive POP		
	mean ($\times 10^{-6}$)	std. ($\times 10^{-7}$)	std./mean	mean ($\times 10^{-6}$)	std. ($\times 10^{-7}$)	std./mean
$\rho = 0.9$	3.06	4.95	0.16	3.18	2.42	0.08
$\rho = 0.7$	2.98	11.1	0.37	3.10	3.71	0.12
$\rho = 0.5$	2.45	23.6	0.96	3.06	7.27	0.24

TABLE 6.20: *Estimators of $\mathbb{P}(PL_{trader} \leqslant L)$ (mean) for $L = -2.4$ with empirical standard deviation (std.) for adaptive IPS and POP methods ($p = 10\%$) based on 100 algorithm macro-runs. Each intermediate level estimator in both methods is based on $M = N = 10^5$ simulations.*

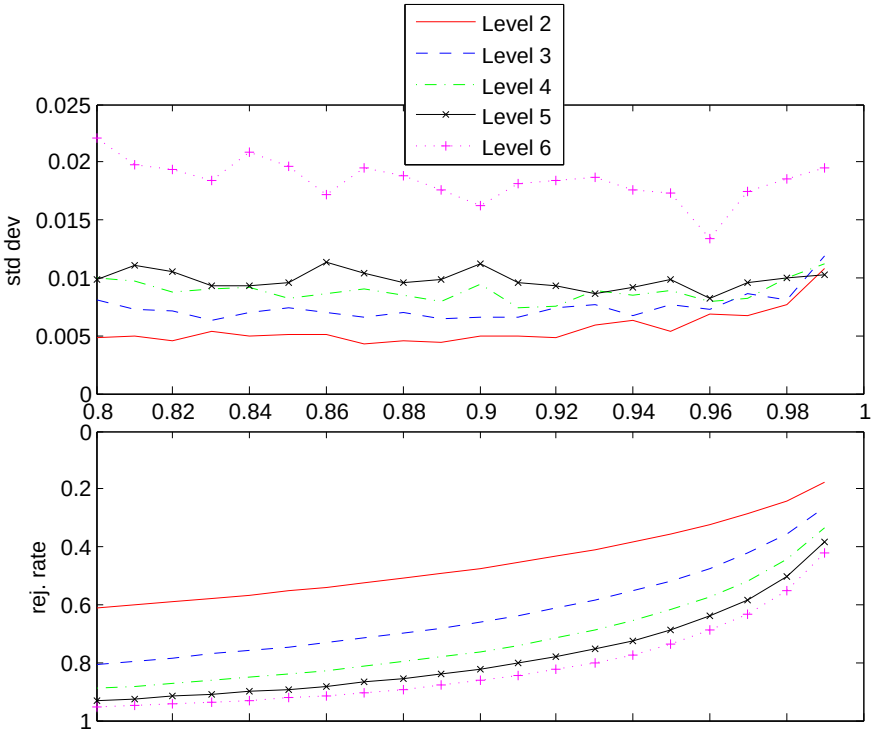


FIGURE 6.3: *Adaptive POP method ($p = 10\%$). Standard deviation (std dev.) and corresponding rejection rate (rej. rate), based on 100 macro-runs, of each quantile estimator $(\hat{Q}^l_{N,p})_{1 \leqslant l \leqslant L^*-1}$ and last level occupation measure estimator $\hat{r}_N(\hat{Q}^{L_N}_{N,p})$, for different values of ρ.*

6.7.2 Stress Testing

Large loss probability $\mathbb{P}(A)$ with $A = \{PL_{\text{trader}} \leqslant L\}$ has been estimated above using POP and IPS methods. Here we are more interested to know what are the typical scenarios which generate large losses. We set $S_0 = K_{\text{strk}} = 10$, $\sigma_- = 0.2$, $\sigma_+ = 0.27$ and take $L = -2.4$. Applying the principle of Section 5.8 on stress-testing, we can get samples from the distribution $X \mid X \in A$; 5 typical scenarios are reported in Figure

6.4. Each scenario is obtained using 10^4 iterations of shaking with rejection (with $\rho = 0.9$). As intuitively expected, typical extreme scenarios exhibit large fluctuations of S.

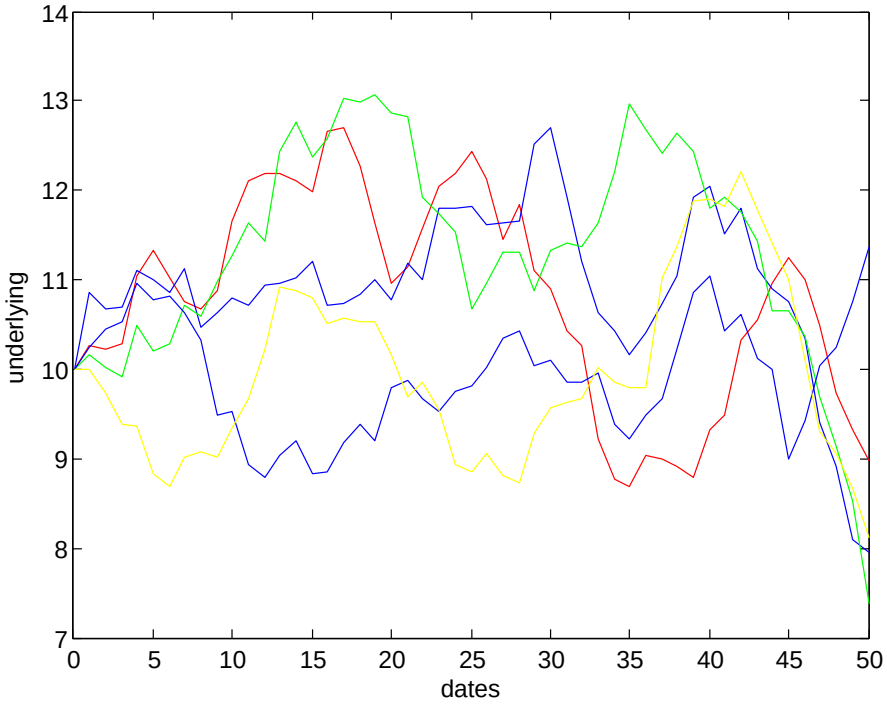


FIGURE 6.4: Typical paths of the underlying stock price which lead to large hedging loss

6.8 Measuring default probabilities in credit portfolios

In this subsection, we consider a credit portfolio based on asset values of N_0 different firms. Let us suppose $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$ is a probability space where $\{W_1, W_2, \ldots, W_{N_0}, W\}$ are $\mathbb{P}$-standard Brownian motions with constant correlations. We denote by $\{\mathcal{F}_t, t \geqslant 0\}$ the $\mathbb{P}$-augmentation of the filtration generated by $\{W_1, W_2, \ldots, W_{N_0}, W\}$. As in [27], we assume that the dynamics of asset values is given by the following system of stochastic differential equations

$$\mathrm{d}S_i(t) = rS_i(t)\mathrm{d}t + \sigma(t)S_i(t)\mathrm{d}W_i(t), \quad i = 1, \ldots, N_0, \tag{6.8.1}$$

where r is the risk-free interest rate, the common stochastic volatility factor $\sigma(t)$ is modeled by a Cox-Ingersoll-Ross model satisfying

$$\mathrm{d}\sigma(t) = \kappa\big(\bar{\sigma} - \sigma(t)\big)\mathrm{d}t + \gamma\sqrt{\sigma(t)}\mathrm{d}W_t, \tag{6.8.2}$$

where $\kappa, \bar{\sigma}$ and γ are positive constants. Brownian motions are correlated as follows:

$$\mathrm{d}\langle W_i, W_j\rangle_t = \rho^W \mathrm{d}t, i \neq j, \quad \mathrm{d}\langle W_i, W\rangle_t = \rho^\sigma \mathrm{d}t, \quad i = 1, \dots, N_0. \tag{6.8.3}$$

Next, we consider the default boundary for each firm i to be a fixed value $B_i \in \mathbb{R}_+$. The time of default for firm i in the portfolio is defined as

$$\tau_i(B_i) := \inf\left\{t \geqslant 0 : S_i(t) \leqslant B_i\right\}.$$

The current methods would directly adapt to the case where the default level B_i is replaced by a time-dependent deterministic function.

In order to evaluate different tranches in a credit portfolio, we are interested to calculate the probability that at least L defaults occur before T, i.e. for $0 < L < N_0$

$$\mathrm{P}(L) = \mathbb{P}\left(\sum_{i=1}^{N_0} \mathbf{1}_{\{\tau_i(B_i)\leqslant T\}} > L\right) \tag{6.8.4}$$

$$= \mathbb{P}\left(\sum_{i=1}^{N_0} \mathbf{1}_{\{\min_t S_i(t)\leqslant B_i\}} > L\right), \tag{6.8.5}$$

Due to the path dependency of the default scheme and of the stochastic volatility model, it is not clear how to find the optimal change of measure to perform importance sampling to estimate $\mathrm{P}(L)$, which motivates the use of alternative simulation techniques.

6.8.1 Default probability

In order to express $\mathrm{P}(L)$ in the form of (5.3.13), we need to create a cascade of decreasing sets $\{A_k\}_{1\leqslant k\leqslant n}$. We define $Z \in \mathbb{R}^{N_0}$ whose i-th component is the minimum of $(S_i(t)/S_i(0))_t$ and we set

$$A_k := \left\{z \in \mathbb{R}^{N_0} : \sum_{i=1}^{N_0} \mathbf{1}_{\{z_i\times(B_i+\frac{k}{n}(S_i(0)-B_i)\leqslant B_i\}} > L\right\}, \quad 1 \leqslant k \leqslant n,$$

which consists in progressively decreasing the default trigger levels. The nested set condition (5.1.1) is then fulfilled. Then, we apply POP and IPS methods to compute all the conditional probabilities $\mathbb{P}\left(Z \in A_{k+1} \mid Z \in A_k\right)$.

Remark 6.8.1. *Another natural way to create the nested sequence of sets is to progressively increase the number of defaults:*

$$\tilde{A}_k := \left\{z \in \mathbb{R}^{N_0} : \sum_{i=1}^{N_0} \mathbf{1}_{\{z_i S_i(0)\leqslant B_i\}} > \frac{k}{n}L\right\}, \quad 1 \leqslant k \leqslant n. \tag{6.8.6}$$

We empirically observe that the choice is in general less accurate. Although we have proven that POP method will eventually converge in all the finite-dimensional cases, how to construct intermediate sets to achieve the best convergence rate remains to be explored.

To perform numerical experiments in the considered model of (6.8.1)-(6.8.2), we fix the parameter values as in Table 6.21.

$S_i(0)$	r	ρ^W	$\sigma(0)$	κ	$\bar{\sigma}$	γ	ρ^σ
90	0.06	0.10	0.4	3.5	0.4	0.7	-0.06

TABLE 6.21: *Parameters for credit portfolio model*

Further, we fix the total number of firms $N_0 = 125$ and threshold level $B_i = B$ for some $B > 0$. Next, we estimate the default probability $\mathrm{P}(L)$ for different values of L over $T = 1$ with 50 time steps per year in the Euler discretization scheme of Deelstra and Delbaen [41]. For $L = 100$ and $B = 36$, the crude Monte Carlo estimator of default probability with 3×10^9 sample paths has a 99% confidence interval as $[4.92, 5.13] \times 10^{-6}$. In Table 6.22, we report the results for IPS and POP based estimators for fixed $n = 5$ levels. For different values of the shaking parameter ρ, it is clear that POP based estimator provides more accurate results than IPS method. In Figure 6.7, using POP based estimator with fixed number of levels $n = 20$ and 10^4 simulations at each level, we also report $\mathrm{P}(L)$ for different levels of default threshold B based on different values of L. Remarkably, it allows to compute very low probabilities (up to 10^{-24}).

Next, we implement IPS and POP based estimators with an adaptive number of levels as discussed in Section 5.6. To estimate $\mathrm{P}(L)$, we fix the conditional probability $\mathbb{P}\left(Z \in A_{k+1} \mid Z \in A_k\right)$ of each, except the last, intermediate level (to be estimated) to $p = 10^{-1}$. In Table 6.23, we can see that both IPS and POP based estimates are within the reported confidence interval of the true value for $\rho = 0.9$. However, the corresponding POP based estimator has a lower standard deviation. When comparing with Table 6.22, variances are roughly unchanged by using the adaptive scheme, but the advantage of this version is to have a fully simulation-based scheme where we do not need to pre-specify the acceptance threshold levels.

In Figures 6.5 and 6.6, we report different detailed statistics w.r.t. the level and the shaking parameter (non-adaptive POP method: standard deviation and rejection rate; adaptive POP method: standard deviation of quantile and occupation measure along with rejection rate). We observe similar behaviors as in the first example of Subsection 6.7. For rare regions (levels $l = 3, 4, 5$), the parameter ρ_l^* minimizing the standard deviation of the l-th conditional probability seems to be associated to rejection rate of 70%. We believe that this (so far empirical) invariance relation between best shaking parameters (for minimal variances) and rejection rate of about $70\% - 80\%$ should give a way to adaptively

choose ρ. This will be further investigated in the future. Again we see that for the adaptive POP method, with different values of ρ minimizing standard deviation in each intermediate level, the standard deviation of the final adaptive estimator of the rare event probability is about 60% of that with a constant $\rho = 0.9$.

The above methodology can also be applied directly to better account for the systemic risk and the illiquidity issues, for example, in the settings of [54] where inter-bank lending is modeled by a system of coupled diffusion processes in a mean-field regime.

	IPS			POP		
	mean ($\times 10^{-6}$)	std. ($\times 10^{-6}$)	std./mean	mean ($\times 10^{-6}$)	std. ($\times 10^{-6}$)	std./mean
$\rho = 0.9$	5.82	4.37	0.75	5.01	0.80	0.16
$\rho = 0.7$	4.92	1.56	0.32	4.99	1.02	0.20
$\rho = 0.5$	4.79	3.80	0.79	5.02	1.94	0.39

TABLE 6.22: *Estimators of default probability (mean) for $L = 100$ and $B = 36$ with empirical standard deviation (std.) for non-adaptive IPS and POP methods based on 100 algorithm macro-runs. Each intermediate level estimator in both methods is based on $M = N = 10^4$ simulations.*

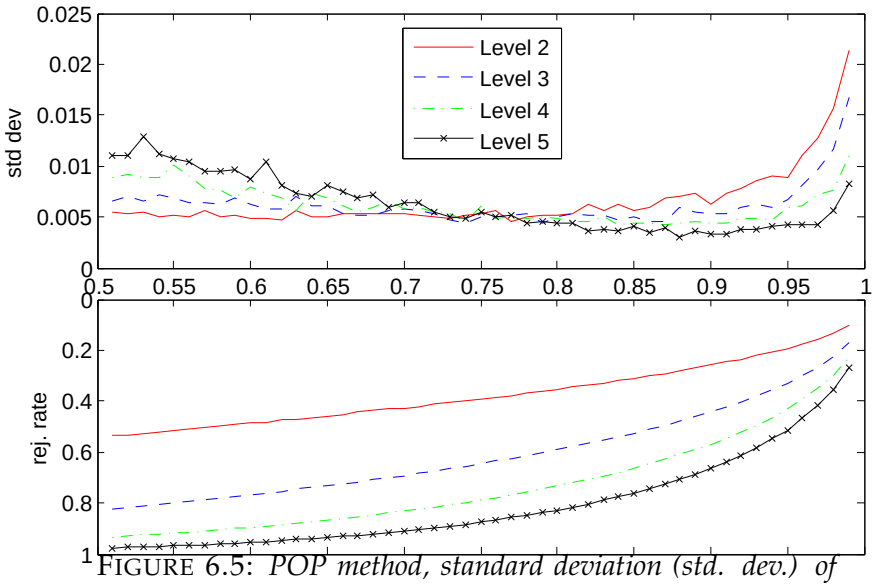


FIGURE 6.5: *POP method, standard deviation (std. dev.) of each conditional probability estimator and corresponding rejection rate (rej. rate), based on 100 macro-runs, for different values of ρ.*

	Adaptive IPS			Adaptive POP		
	mean ($\times 10^{-6}$)	std. ($\times 10^{-6}$)	std./mean	mean ($\times 10^{-6}$)	std. ($\times 10^{-6}$)	std./mean
$\rho = 0.9$	4.93	1.91	0.39	5.16	0.85	0.16
$\rho = 0.7$	5.42	1.58	0.29	4.98	1.02	0.20
$\rho = 0.5$	6.40	5.00	0.78	5.35	2.05	0.38

TABLE 6.23: *Estimators of default probability (mean) for $L = 100$ and $B = 36$ with empirical standard deviation (std.) for adaptive IPS and POP methods ($p = 10\%$) based on 100 algorithm macro-runs. Each intermediate level estimator in both methods is based on $M = N = 10^4$ simulations.*

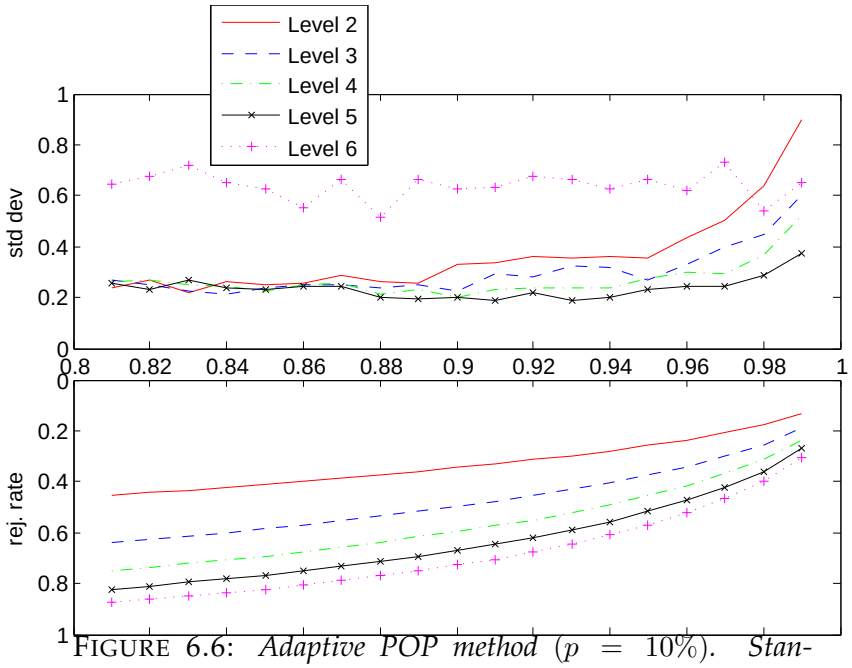
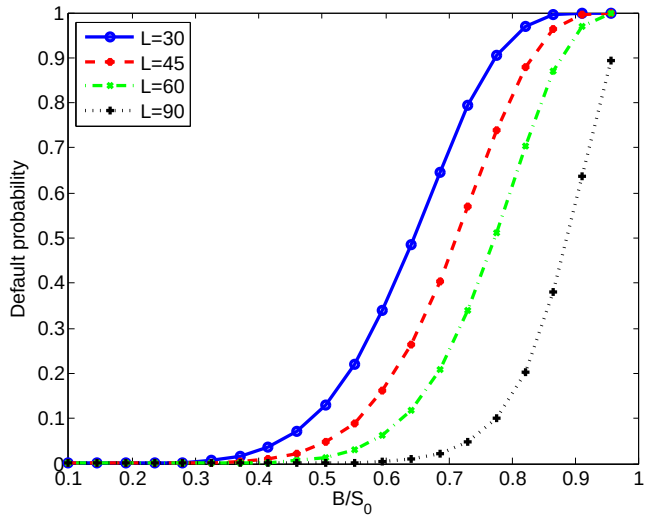
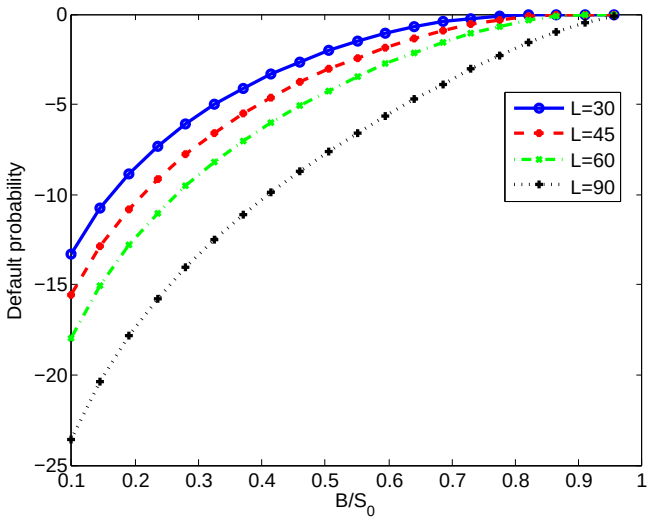


FIGURE 6.6: *Adaptive POP method ($p = 10\%$). Standard deviation (std. dev.) and corresponding rejection rate (rej. rate), based on 100 macro-runs, of each quantile estimator $(\hat{Q}^l_{N,p})_{1 \leqslant l \leqslant L^*-1}$ and last level occupation measure estimator $\hat{r}_N(\hat{Q}^{L_N}_{N,p})$, for different values of ρ. The std. dev. of occupation measure estimator has been scaled by 10 for easier comparison.*



(a)



(b)

FIGURE 6.7: *Plot (a) and log-plot (b) of default probabilities for varying B/S_0.*

6.8.2 Variant of IPS method

In the following Tables 6.24 and 6.25, we present the numerical results obtained by our modified IPS method, i.e. the variant of IPS method with extra sampling and smaller system size, which is presented in Section 5.9.

$\times 10^{-6}$	$J=1$	$J=2$	$J=3$	$J=4$	$J=5$	POP
$\rho=0.9$	2.50	1.21	1.04	1.01	1.03	0.80
$\rho=0.7$	1.74	1.40	1.29	1.25	1.20	1.02
$\rho=0.5$	4.46	3.69	3.56	3.11	3.18	1.94

TABLE 6.24: Empirical standard deviation of IPS estimators of default probability for $L=100$ and $B=36$ based on 1000 algorithm macro-runs, with $M=10^4$ and particle system size equal to $M'=\lfloor \frac{M}{J} \rfloor$. The last column is the empirical standard deviation of POP method using $n=M=10^4$ iterations at each level.

$\times 10^{-6}$	$J=6$	$J=7$	$J=8$	$J=9$	$J=10$	POP
$\rho=0.9$	1.06	1.11	1.19	1.31	1.37	0.80
$\rho=0.7$	1.29	1.28	1.33	1.37	1.43	1.02
$\rho=0.5$	2.90	2.67	2.73	2.63	2.61	1.94

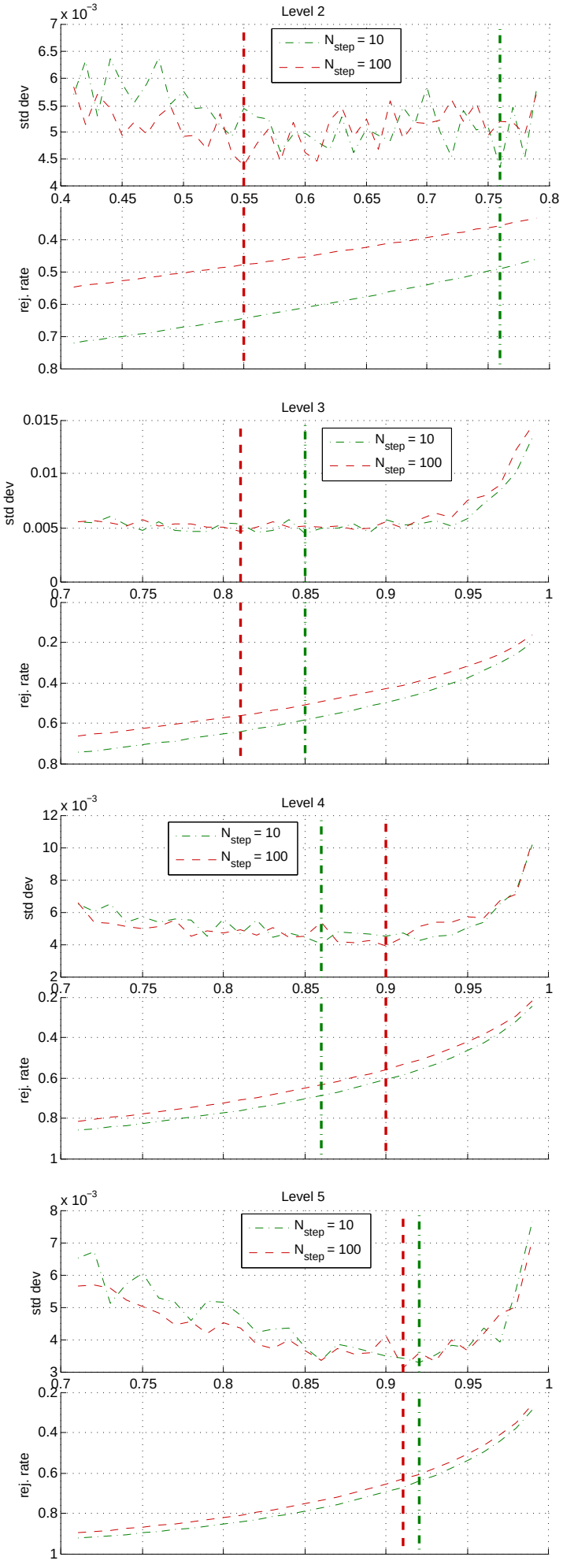
TABLE 6.25: Empirical standard deviation of IPS estimators of default probability for $L=100$ and $B=36$ based on 1000 algorithm macro-runs, with $M=10^4$ and particle system size equal to $M'=\lfloor \frac{M}{J} \rfloor$. The last column is the empirical standard deviation of POP method using $n=M=10^4$ iterations at each level.

As we can see, the IPS method with fewer particles but extra resampling at each step have better performance compared to the case without resampling. Heuristically, a good choice is $J=4$. However, even after extra resampling, POP method yields smaller standard deviation.

6.8.3 Impact of discretization

A different IPS-based method has been proposed by Carmona et al. [27] in order to compute $\mathrm{P}(L)$ (see also [26] for application of this method in other models). We would like to emphasize the main difference between the former IPS approach and our work. The underlying Markov chain for their IPS method is simply the time-discretization of the $(2N_0+1)$-dimensional process $(S_i, \min S_i, \sigma, 1 \leqslant i \leqslant N_0)$. This poses several difficulties for the authors. Firstly, one needs to exhibit a good potential function for the selection of particles which is very delicate because of the high-dimensionality of the problem. Secondly, one needs to choose an appropriate discretization time step Δ_t. This is also intricate, since on the one hand a large number of time steps may help in better selection of the particles in rarer and rarer regions, but on the other hand it slows down the statistical convergence of IPS (the resampling adds noise in the estimation). In our case, we directly consider Markov chains valued on path space, thus avoiding the delicate problem of choosing the

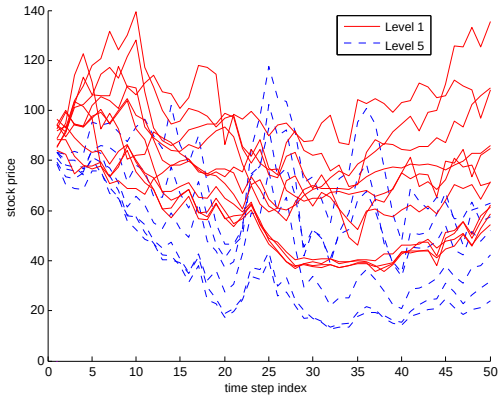
time step Δ_t and the high-dimensional potential function (in our numerical experiments, we have observed that Δ_t has no significant impact on the convergence of our versions of IPS-POP methods when it is small enough). Thus, our approach and results are rather different from those of Carmona et al. [27]. These differences are mainly due to the fact that our method does not require any Markovian assumption on $(S_i, \min S_i, \sigma, 1 \leqslant i \leqslant N_0)$ and could be directly applied to path-dependent models (whenever useful).


Level 2
x 10^-3
N_step = 10
N_step = 100
std dev
rej. rate
Level 3
N_step = 10
N_step = 100
std dev
rej. rate
Level 4
x 10^-3
N_step = 10
N_step = 100
std dev
rej. rate
Level 5
x 10^-3
N_step = 10
N_step = 100
std dev
rej. rate

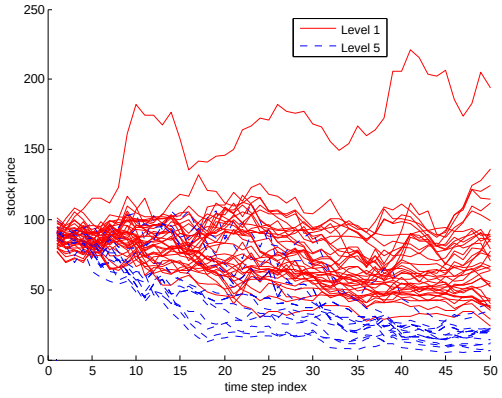
In our version of IPS and POP methods, the underlying random variable X lies in a large-dimensional space $\mathcal{X} = \mathbb{R}^{(N_0+1)N_{\text{step}}}$. Therefore it is important to assess how the large dimension affects the statistical errors of the methods (like in MCMC sampler, for example, see [109]). The important point is that we use reversible transformation directly in the path space. This suggests that our methods are less sensitive to time-discretization. We investigate this problem in Figure 6.8, where we study the impact of N_{step} in POP method with $n = 5$ levels. We report the numerical results only for the POP method, as the qualitative phenomenon for IPS method is the same. The graphs show that the choice of N_{step}, when large, has no significant impact on the optimal value of the shaking parameter ρ, which corresponds to the minimum standard deviation of the conditional probability estimator at a given level. Additionally, both the standard deviation and the rejection rate are quite insensitive to N_{step}. These are important advantages of the methods studied in this work. This allows to decouple the problem of bias reduction and the control of statistical convergence.

6.8.4 Stress Testing

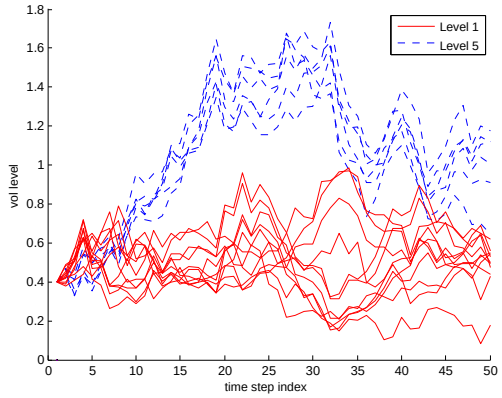
In Figure 6.9, we exhibit extreme scenarios for the asset price of firm 1 and its volatility, in the situation of various defaults (level 1 and last level, $n = 5$, $L = 100$) with two different values of B. Each scenario is obtained by using 10^4 iterations of shaking with rejection (with $\rho = 0.9$). For this, we have used the first and last-level Markov chain $X_{1,\cdot}$ and $X_{5,\cdot}$, respectively, to get samples from the distributions $X \mid X \in A_1$ and $X \mid X \in A$ respectively, as explained in Section 5.8. Such tools may be efficiently exploited by regulators and risk managers for a better risk assessment.



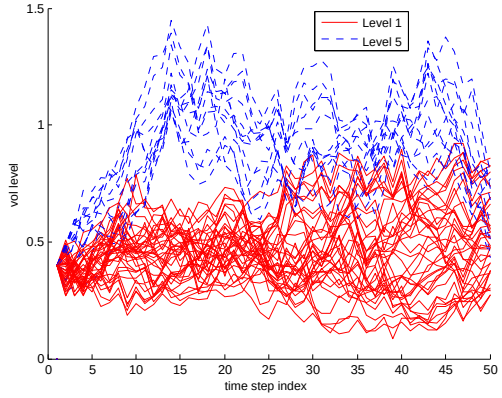
(a) $B = 18$



(b) $B = 36$



(c) $B = 18$



(d) $B = 36$

FIGURE 6.9: Sample paths for the asset price of firm 1 at Level 1 and Level 5 in the POP method and the respective volatility sample paths.

6.9 Fractional Brownian motion for modeling volatility

The fractional Brownian motion (fBM) $(B_t^{(H)})_{t\in\mathbb{R}}$with Hurst exponent $H \in (0,1)$ was defined in Example 5.3.3. For $H \neq 1/2$, it is well known that $B^{(H)}$ is not a semimartingale. In order to represent fBM, we make use of the Mandelbrot and van Ness representation of $B^{(H)}$ as an integral w.r.t. a standard Brownian motion B:

$$B_t^{(H)} = C_H\Big[\int_{-\infty}^{t}\big[(t-s)^{H-\frac{1}{2}} - (-s)_+^{H-\frac{1}{2}}\big]\mathrm{d}B_s$$

with

$$C_H = \sqrt{\frac{2H\Gamma(3/2-H)}{\Gamma(H+1/2)\Gamma(2-2H)}}.$$

Recently, Gatheral and co-authors [59] have successfully employed fBM to model the market observed volatility of stock indexes. In order to demonstrate the application of POP method for models which are not necessarily based on semimartingales, we consider the fractional SABR (fSABR) model proposed by Gatheral et al. [58]. In fSABR, the underlying asset dynamics are given by

$$\frac{\mathrm{d}S_t}{S_t} = \sigma_t \mathrm{d}Z_t, \tag{6.9.1}$$

$$\sigma_t = \bar{\sigma}\exp\left(-\frac{1}{2}\alpha^2 t^{2H} + \alpha B_t^{(H)}\right), \tag{6.9.2}$$

where Z_t is a standard Brownian motions with instantaneous correlation ρ^{BZ} with B_t (i.e. $\mathrm{d}\langle B, Z\rangle_t = \rho^{BZ}\mathrm{d}t$). Under the model (6.9.1), we use POP method to estimate the small-strike tail asymptotic slope of implied variance

$$\beta_L := \limsup_{x\to-\infty}\frac{I^2(x)T}{|x|} \tag{6.9.3}$$

where $I(x)$ is the BS implied volatility of a Vanilla option on S with log-moneyness $x = \log K/S_0$ and maturity T. The estimate of the slope can be, in turn, used to obtain estimate of the critical negative moment $\tilde{q} := \sup\{q : \mathbb{E}[S_T^{-q}] < \infty\}$ from the well-known moment formula [95, Theorem 3.4]

$$\tilde{q} = 1/2\beta_L + \beta_L/8 - 1/2. \tag{6.9.4}$$

We work with the following parameter values: $S_0 = 40, \bar{\sigma} = 0.235, r = 0, T = 1.0$ and $\alpha = 0.5, 1.0$. We use intermediate levels at

$$[32.5, 25, 19.5, 14, 10.5, 7, 5, 3, 2, 1]$$

in the POP method (shaking parameter value = 0.9) with 10^5 simulations[3] at each level in order to estimate the implied volatility at different values of the log-moneyness. The output values are based on 100 independent algorithm macro-runs. We observe on Figure 6.10 that the squared implied volatilities seemingly behave linearly for large negative values of the log-moneyness, which suggests that the $\limsup$ in (6.9.3) is presumably a limit (see Remark 6.9.1 below for a related discussion).

In light of (6.9.3), we could use the most extreme value of the implied variance $I^2(x_{\min})$ (corresponding to $x_{\min} = -3.75$ in Figure 6.10) in order to evaluate β_L. Instead of doing so, we compute the slope β_L by linear interpolation of the two most extreme implied variances $I^2(x_{\min})$ and $I^2(x_{\min} + \Delta x)$. We observe that following one or the other procedure has no significant impact on the results. This yields the estimates of $\tilde{q}$ in Tables 6.26 and 6.27.

ρ^{BZ}	$H = 0.15$	$H = 0.25$	$H = 0.75$	$H = 0.9$
-0.3	2.6133	2.5515	2.8058	2.9753
-0.5	2.4222	2.3823	2.6733	2.8715
-0.7	2.2593	2.2042	2.5465	2.7918
-0.9	2.1235	2.0653	2.4339	2.6919

TABLE 6.26: Estimates of critical negative moment $\tilde{q}$ in fSABR model (6.9.1) using POP method, $\alpha = 0.5$

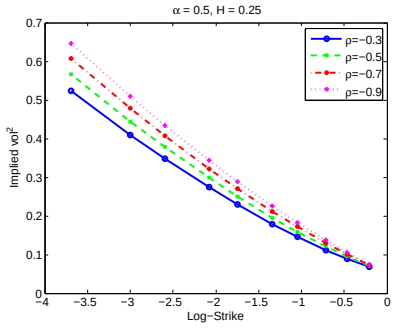
ρ^{BZ}	$H = 0.15$	$H = 0.25$	$H = 0.75$	$H = 0.9$
-0.3	0.8251	0.8267	0.9211	0.9632
-0.5	0.7905	0.7913	0.8950	0.9449
-0.7	0.7597	0.7591	0.8686	0.9277
-0.9	0.7325	0.7297	0.8449	0.9113

TABLE 6.27: Estimates of critical negative moment $\tilde{q}$ in fSABR model (6.9.1) using POP method, $\alpha = 1.0$

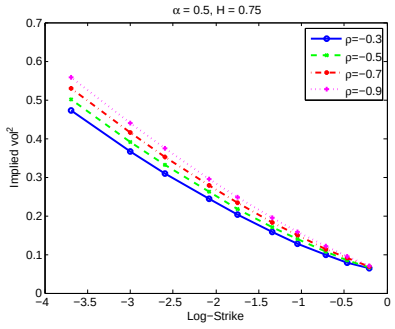
From our numerical results, we can observe that $\tilde{q}$ increases with the value of the correlation ρ^{BZ} in the model. Conversely, $\tilde{q}$ decreases with the value of the parameter α. There is no global monotonicity appearing from the relationship between $\tilde{q}$ and value of $H \in (0, 1)$. On the other hand, one does see (as expected) the emergence of two different regimes for $H < 1/2$ and $H > 1/2$. These observations suggest that it is possible - at least in theory - to calibrate the value of one of these model parameters from extreme implied volatility estimates, for example by

[3] We exactly simulate the skeleton of Z, B and B^H (with a step length of $T/100$) as a correlated Gaussian vector since the covariance matrix of this vector can be computed explicitly.

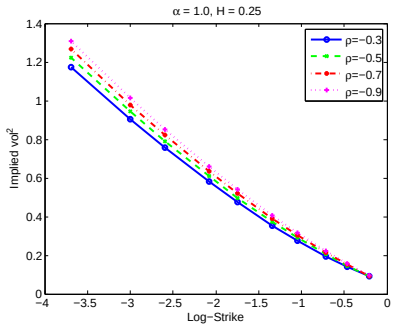
using POP method. Moreover, the plots in 6.10 indicate a 'tilting' effect of the correlation parameter ρ^{BZ} on the whole smile curve, analogous to that in standard stochastic volatility models based on Brownian motion. This indicates that under the fractional model (6.9.1), too, the appropriate value of the correlation parameter can be reasonably inferred from market implied volatilities by observing the slopes of the left- and right hand sides of the smile.



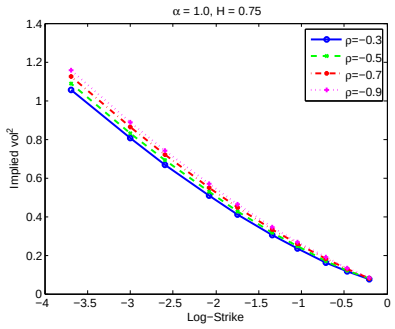
(a)



(b)



(c)



(d)

FIGURE 6.10: *Squared implied volatility as a function of log-strike in the fSABR model* (6.9.1).

Remark 6.9.1. *While the formulas (6.9.3)-(6.9.4) always hold when β_L is defined via a* $\limsup$*, it is interesting to notice that there is a (large) class of models for which the limsup can actually be updated to a true limit, thus providing the full asymptotic equivalence $I^2(x)T \sim |x|$ as $x \to -\infty$. This class is fully characterized in Gulisashvili [71, Theorem 3.5]. Recall that a positive measurable function f defined on some neighborhood of infinity is said to be regularly varying with index $\alpha \in \mathbb{R}$ if for every $\lambda > 0$, $\frac{f(\lambda x)}{f(x)} \to \lambda^\alpha$ as $x \to +\infty$. Furthermore, the class of Pareto-type functions is introduced in [71]. Let g be positive measurable functions defined on $(0, c)$ for some $c > 0$: if there exist two functions g_1 and g_2 that are regularly varying with index α and such that $g_1(x^{-1}) \leq g(x) \leq g_2(x^{-1})$ for all $0 < x < c$, then we say that the function g is of weak Pareto-type near zero with index α.*
Gulisashvili [71] proves the following: under the assumption $0 < \tilde{q} < \infty$, the asymptotic formula

$$\lim_{x \to -\infty} \frac{I^2(x)T}{|x|} = \beta_L \tag{6.9.5}$$

holds if and only if the following condition is satisfied:

i) The put price function $P(K) = \mathbb{E}[(K - S_T)^+]$, $K > 0$, is of weak Pareto type near zero with index $\alpha_1 = -\tilde{q} - 1$.

It is possible to relate the property i) in a more direct way to the distribution of the stock price: condition i) holds for the put price if one of the following two conditions is satisfied:

ii) The cdf of the stock price $F(K) = \mathbb{P}(S_T \leq K)$ is of weak Pareto type near zero with index $\alpha_2 = -\tilde{q}$.

iii) The density $p_T(\cdot)$ of the stock price S_T (if it exists) is of weak Pareto type near zero with index $\alpha_3 = -\tilde{q} + 1$.

The implication $iii) \Rightarrow i)$ is proven in [71], Theorem 3.11. The implication $ii) \Rightarrow i)$ can be proven following the lines of the proofs of Theorems 3.11 and 3.7 in [71].

Figure 6.10 suggests that squared implied volatilities behave asymptotically linearly with log-moneyness, and we can therefore conjecture that equation (6.9.5) holds for the fSABR model (6.9.1). An analysis of the cdf or the density function of the stock price, as performed in [73] for a class of models with Gaussian self-similar stochastic volatility, would allow to show that properties $ii)$ and $iii)$ hold true in the fSABR model. We leave such kind of investigation for future research.

6.10 Sensitivities for out-of-the money options

In this example, we consider a d-dimensional Black-Scholes model in which the asset price vector $S = (S^1, S^2, \ldots, S^d)$ is given as

$$\frac{\mathrm{d}S_t^i}{S_t^i} = \mu^i \mathrm{d}t + \sigma^i \mathrm{d}(LW_t)^i \tag{6.10.1}$$

where $\sigma^i > 0$ for all $i = 1, \dots, d$, W is an d-dimensional standard Brownian motion, and L is the symmetric square root of a d-dimensional correlation matrix C, so that $LL^* = C$ (here $L = L^*$). Hereafter we assume that the matrix C (therefore L) is invertible. Denoting by Z^i the log of S^i, one has

$$Z_T^i = Z_0^i + \left(\mu^i - \frac{1}{2}(\sigma^i)^2\right) T + \sigma^i (LW_T)^i \tag{6.10.2}$$

with $Z_0^i = \log(S_0^i)$. Equation (6.10.1) allows to model separately the individual volatility σ^i of each asset and the correlation between the driving Brownian factors. The introduction of a volatility smile on each asset can be achieved simply by switching from constant to local volatility functions $\sigma^i(t, \cdot)$ (which can be separately calibrated to option data on each asset).

We consider a digital-style payoff written on a generalized basket, whose financial evaluation is defined by

$$\mathcal{P} := \mathbb{P}\left(\varphi(Z_T, \bar{a}) \geqslant 0\right)$$

where

$$\varphi(z, \bar{a}) := \sum_{i=1}^{d} \varepsilon_i p_i e^{z^i} - \bar{a} \tag{6.10.3}$$

with $p_i > 0, \varepsilon_i \in \{-1, 1\}$ and $\bar{a} \in \mathbb{R}$. This setting can cover the situation of risk management of an insurance contract (when each asset evolves with its own drift coefficient μ^i), and of course the pricing of a digital option on the basket, which corresponds to set $\mu^i = r$, where r is a risk-free interest rate. We are interested in computing the sensitivities of $\mathcal{P}$ with respect to different model parameters, such as

- p_i in order to assess the influence of the individual weights, possibly in order to reweight the portfolio and lower the risk,
- σ^i in order to quantify the impact of individual volatilities on the tails of the basket,
- $C_{i,j} = (LL^*)_{i,j}$ for $i < j$, in order to study the effect of pair-wise correlations on the product.

In order to obtain explicit sensitivity formulas, we apply Theorem 5.7.2 with $\Phi^\theta = 1$, $Z^\theta = Z_T$ in the context of multidimensional Brownian motion (Example 5.3.2), where θ plays the role of one the model parameters or payoff parameters above. A direct computation shows

$$(D_t Z_T)_{i,j} = \sigma^i L_{i,j} \mathbf{1}_{t \leqslant T} = \Sigma_{i,j} \mathbf{1}_{t \leqslant T} \qquad \gamma_{Z_T} = T\, \Sigma\Sigma^*, \tag{6.10.4}$$

where we denote Σ the matrix $\Sigma = \mathrm{diag}(\sigma)L$, where $\mathrm{diag}(\sigma)_{i,j} = \sigma^i \delta_{i,j}$. Under our assumption, the matrices Σ and γ_{Z_T} are invertible.

In what follows, we denote $A_{i,\cdot}$ (respectively $A_{\cdot,i}$) the i-th row (respectively i-th column) of the matrix A.

6.10.1 Sensitivity by Malliavin calculus

Sensitivity w.r.t. p_i. In view of (6.10.3) we have $\partial_{p_i}\mathcal{P} = \partial_{Z_0^i}\mathcal{P}\frac{1}{p_i}$ and it suffices to compute sensitivity w.r.t. $\theta = Z_0^i$. Clearly $\partial_{Z_0^i} Z_T = e^i$ where e^i is the i-th element of the canonical basis of $\mathbb{R}^d$, therefore the weight $\mathcal{I}(Z^\theta, 1)$ in Theorem 5.7.2 becomes

$$\begin{aligned}
\mathcal{I}(Z^\theta,1) &= \delta\left(\sum_{j=1}^d \left(\gamma_{Z_T}^{-1}\partial_{Z_0^i} Z_T\right)_j D_\cdot Z_T^j\right)\\
&= \delta\left(\sum_{j=1}^d (\gamma_{Z_T}^{-1})_{j,i}(\Sigma \mathbf{1}_{[0,T]})_{j,\cdot}\right) = \delta\left(\sum_{j=1}^d (\gamma_{Z_T}^{-1})_{i,j}(\Sigma \mathbf{1}_{[0,T]})_{j,\cdot}\right)\\
&= \frac{1}{T}\delta\left((\Sigma\Sigma^*)^{-1}\Sigma \mathbf{1}_{[0,T]})_{i,\cdot}\right)\\
&= \frac{1}{T}\delta\left(((\Sigma^*)^{-1})_{i,\cdot}\mathbf{1}_{[0,T]}\right) = \frac{1}{T}\Sigma^{-1}e^i \cdot W_T.
\end{aligned}$$

The computation of the sensitivities with respect to σ^i and $C_{i,j}$ involves quantities of the form $\delta((AW_T)^i \times u^* \mathbf{1}_{[0,T]}(\cdot))$, where A is a $d \times d$ matrix and u a (constant) vector in $\mathbb{R}^d$. We will therefore make use of the following formula

$$\delta((AW_T)^i u^* \mathbf{1}_{[0,T]}(\cdot)) = (AW_T)^i u \cdot W_T - T\,(Au)^i \quad . \tag{6.10.5}$$

Equation (6.10.5) can be proven using the identity $\delta(F\,U_\cdot) = F\delta(U_\cdot) - \langle DF, U\rangle$ which holds for $U \in dom(\delta)$ and $F \in \mathbf{D}^{1,2}$, where we denote $\langle V, U\rangle = \sum_{j=1}^d \int_0^T V_t^j U_t^j dt$.

Sensitivity w.r.t. $\theta = \sigma^i$. We have $\partial_{\sigma^i} Z_T = (-\sigma^i T + (LW_T)^i)e^i$. Since $\partial_{\sigma^i} Z_T$ and e^i are collinear, the computations are very similar to the previous ones, and we obtain

$$\begin{aligned}
\mathcal{I}(Z^\theta,1) &= \delta\left(\sum_{j=1}^d \left(\gamma_{Z_T}^{-1}\partial_{\sigma^i} Z_T\right)_j D_\cdot Z_T^j\right)\\
&= \frac{1}{T}\delta\left((-\sigma^i T + (LW_T)^i)((\Sigma^*)^{-1})_{i,\cdot}\mathbf{1}_{[0,T]}\right)\\
&= -\sigma^i\delta\left(((\Sigma^*)^{-1})_{i,\cdot}\mathbf{1}_{[0,T]}\right) + \frac{1}{T}\delta\left((LW_T)^i((\Sigma^*)^{-1})_{i,\cdot}\mathbf{1}_{[0,T]}\right)\\
&= \Sigma^{-1}e^i \cdot W_T\left(-\sigma^i + \frac{1}{T}(LW_T)^i\right) - (L\Sigma^{-1})_{i,i}
\end{aligned}$$

where we have applied the identity (6.10.5) with $A = L$ and $u^* = ((\Sigma^*)^{-1})_{i,\cdot}$ in the last step.

Sensitivity w.r.t. $\theta = C_{i,j}$, $i < j$. We wish to take partial derivatives of functions defined on the set of correlation matrices

$$\mathbb{C} = \left\{(C_{i,j})_{i,j} : C \in \mathcal{S}^d_{\geq 0}, C_{i,i} = 1, C \text{ invertible}\right\}$$

with respect to each of the entries $C_{i,j}$, $i < j$, where $\mathcal{S}^d_{\geq 0}$ denotes the set of symmetric and positive matrices. This is possible under the invertibility assumption because, given a matrix $C \in \mathbb{C}$ and fixed $i < j$, the whole set $\{C_\varepsilon := C + \varepsilon e^{i,j} + \varepsilon e^{j,i}, \varepsilon \in \mathbb{R}\}$ is contained in $\mathbb{C}$ for ε small enough, where $e^{i,j}$ denotes the matrix such that $(e^{i,j})_{i,j} = 1$ and with zero entries elsewhere.[4]
In particular, for the symmetric square root $L = \sqrt{C}$, its partial derivative $\dot{L} := \partial_{C_{i,j}} L$ solves the Sylvester equation [78, p.58]

$$\dot{L}\, L + L\, \dot{L} = e^{i,j} + e^{j,i} := \dot{C}. \tag{6.10.6}$$

From (6.10.2), we derive $\partial_{C_{i,j}} Z_T = \partial_{C_{i,j}}[\mathrm{diag}(\sigma) L W_T] = \mathrm{diag}(\sigma)\dot{L} W_T$. This yields

$$\begin{aligned}
\mathcal{I}(Z^\theta, 1) &= \delta\left(\sum_{l=1}^d \left(\gamma_{Z_T}^{-1} \partial_{C_{i,j}} Z_T\right)_l D.Z_T^l\right) \\
&= \frac{1}{T}\delta\left(\sum_{l=1}^d \left((\Sigma\Sigma^*)^{-1}\mathrm{diag}(\sigma)\dot{L}W_T\right)_l \Sigma_{l,.}\mathbf{1}_{[0,T]}\right) \\
&= \frac{1}{T}\sum_{l=1}^d ((\Sigma\Sigma^*)^{-1}\mathrm{diag}(\sigma)\dot{L}W_T)_l\, \Sigma^*_{.,l} \cdot W_T - \sum_{l=1}^d ((\Sigma\Sigma^*)^{-1}\mathrm{diag}(\sigma)\dot{L}\Sigma^*_{.,l})_l
\end{aligned}$$

(using (6.10.5) with $A = (\Sigma\Sigma^*)^{-1}\mathrm{diag}(\sigma)\dot{L}$ and $u^* = \Sigma_{l,.}$)

$$\begin{aligned}
&= \frac{1}{T} W_T \cdot (\Sigma^*(\Sigma\Sigma^*)^{-1}\mathrm{diag}(\sigma)\dot{L}W_T) - \mathrm{Tr}((\Sigma\Sigma^*)^{-1}\mathrm{diag}(\sigma)\dot{L}\Sigma^*) \\
&= \frac{1}{T} W_T \cdot L^{-1}\dot{L}W_T - \mathrm{Tr}(L^{-1}\dot{L}).
\end{aligned}$$

Since $W_T \cdot L^{-1}\dot{L}W_T$ is a scalar, it is equal to its transpose $W_T \cdot \dot{L}L^{-1}W_T$, and thus to its average $\frac{1}{2}W_T \cdot (L^{-1}\dot{L} + \dot{L}L^{-1})W_T$. Similarly, $\mathrm{Tr}(L^{-1}\dot{L}) = \frac{1}{2}\mathrm{Tr}(L^{-1}\dot{L} + \dot{L}L^{-1})$. We claim that

$$L^{-1}\dot{L} + \dot{L}L^{-1} = L^{-1}\dot{C}L^{-1}, \tag{6.10.7}$$

[4] The matrices C_ε are clearly symmetric and satisfy $(C_\varepsilon)_{i,i} = 1$. The invertibility of C_ε for ε small enough follows from the continuity of the smallest eigenvalue $\lambda_{\min}$ from $\mathcal{S}^d_{\geq 0}$ into $\mathbb{R}$, $A \mapsto \lambda_{\min}(A)$ (with respect to, say, the topology induced by the Hilbert-Schmidt norm), see [81, Hoffman and Wielandt's theorem, p.368].

which gives the final representation

$$\begin{aligned}\mathcal{I}(Z^\theta,1) &= \frac{1}{2T}W_T\cdot(L^{-1}(e^{i,j}+e^{j,i})L^{-1})W_T - \frac{1}{2}\mathrm{Tr}(L^{-1}(e^{i,j}+e^{j,i})L^{-1}) \\ &= \frac{1}{T}(L^{-1}W_T)^i(L^{-1}W_T)^j - (C^{-1})_{i,j}\end{aligned}$$

where the final formula follows from standard manipulations.

To prove (6.10.7), write the derivative of C_ε and $L_\varepsilon = \sqrt{C_\varepsilon}$ w.r.t. ε: it gives (using the notation $\dot{A} = \partial_\varepsilon A_\varepsilon|_{\varepsilon=0}$)

$$\begin{aligned}-C^{-1}\dot{C}C^{-1} &= \partial_\varepsilon C_\varepsilon^{-1}|_{\varepsilon=0} = \partial_\varepsilon L_\varepsilon^{-2}|_{\varepsilon=0} \\ &= \partial_\varepsilon L_\varepsilon^{-1}|_{\varepsilon=0}L^{-1} + L^{-1}\partial_\varepsilon L_\varepsilon^{-1}|_{\varepsilon=0} \\ &= -L^{-1}\dot{L}L^{-2} - L^{-2}\dot{L}L^{-1}.\end{aligned}$$

Now multiplying by L on the left and right we obtain (6.10.7).

6.10.2 Sensitivity by likelihood method

We can perform a change of variables and directly differentiate the density function to obtain sensitivity with respect to θ. Let us also define $(\mathcal{Y})^i := (\mathcal{Z})^i + \log(p^iS_0^i) + (\mu^i - \frac{1}{2}|\sigma^i|^2)T$, $1 \leqslant i \leqslant d$. We have the following

$$\begin{aligned}\frac{\partial\mathcal{P}}{\partial\theta} =& \frac{\partial}{\partial\theta}\int\phi\left(\sum_{i=1}^d \varepsilon_i p_i S_0^i \exp\big((\mu^i - \frac{1}{2}|\sigma^i|^2)T + z^i\big)\right) \\ &\times \frac{1}{\sqrt{(2\pi T)^d\det C}} \times \exp\Big(-\frac{1}{2T}z\cdot C^{-1}z\Big)\mathrm{d}z \\ =& \frac{\partial}{\partial\theta}\int\phi(\sum_{i=1}^d \varepsilon_i\exp(y^i))\frac{1}{\sqrt{(2\pi T)^d\det C}} \\ &\times\exp\Big(-\frac{1}{2T}\big(y^i - (\mu^i - \frac{1}{2}|\sigma^i|^2)T - \log(p_iS_0^i)\big)_{1\leqslant i\leqslant d}\cdot \\ &\quad C^{-1}\big(y^i - (\mu^i - \frac{1}{2}|\sigma^i|^2)T - \log(p_iS_0^i)\big)_{1\leqslant i\leqslant d}\Big)\mathrm{d}y \\ :=& \frac{\partial}{\partial\theta}\int\phi(\sum_{i=1}^d \varepsilon_i\exp(y^i))p_\theta(y)\mathrm{d}y \\ =& \int\phi(\sum_{i=1}^d \varepsilon_i\exp(y^i))\partial_\theta\log(p_\theta(y))p_\theta(y)\mathrm{d}y \\ =& \mathbb{E}\left(g(S_T)h\Big(\big((\sigma W_T)^i + (\mu^i - \frac{1}{2}|\sigma^i|^2)T + \log(p^iS_0^i)\big)_{1\leqslant i\leqslant d}\Big)\right) \\ :=& \mathbb{E}\left(g(S_T)\Xi\right)\end{aligned} \tag{6.10.8}$$

where

$$
\begin{aligned}
h(y) &:= \partial_\theta \log(p_\theta(y)),\\
\Xi &:= h\Big(\big((\sigma W_T)^i + (\mu^i - \frac{1}{2}|\sigma^i|^2)T + \log(p^i S_0^i)\big)_{1\leqslant i\leqslant d}\Big),\\
\log p_\theta(y) &= -\frac{1}{2}\log\Big((2\pi T)^d\Big) - \frac{1}{2}\log\det(C)\\
&\quad - \frac{1}{2T}\Big(y^i - (\mu^i - \frac{1}{2}|\sigma^i|^2)T - \log(p^i S_0^i)\Big)_{1\leqslant i\leqslant d}\cdot C^{-1}\\
&\quad \times \Big(y^i - (\mu^i - \frac{1}{2}|\sigma^i|^2)T - \log(p^i S_0^i)\Big)_{1\leqslant i\leqslant d}.
\end{aligned}
$$

We can write an explicit formula of the model parameter sensitivity in (6.10.8) for few specific cases as follows, with the notation

$$
D := C^{-1}\Big(y^i - (\mu^i - \frac{1}{2}|\sigma^i|^2)T - \log(p^i S_0^i)\Big)_{i=d}^{i=1}
$$

1. Case $\theta = p^i S_0^i$.

$$
\begin{aligned}
\partial_\theta\big(\log p_\theta(y)\big) &= -\frac{1}{T}\,(D)_i \times \Big(\frac{-1}{p^i S_0^i}\Big),\\
\Xi &= \frac{1}{T}\frac{1}{p^i S_0^i}\Big(C^{-1}\sigma W_T\Big)_i = \frac{1}{T}\frac{1}{p^i S_0^i}\Big((\sigma^*)^{-1}W_T\Big)_i.
\end{aligned}
$$

2. Case $\theta = C_{ii} = |\sigma_i|^2$.

$$
\begin{aligned}
\partial_\theta\big(\log p_\theta(y)\big) &= -\frac{1}{2}\frac{\det(C_{-i,-i})}{\det(C)} - \frac{1}{2}\,(D)_i + \frac{1}{2T}D\cdot\Big(\mathbf{1}_{ii}\Big)D,\\
\Xi &= -\frac{1}{2}\frac{\det(C_{-i,-i})}{\det(C)} - \frac{1}{2}\Big((\sigma^*)^{-1}W_T\Big)_i + \frac{1}{2T}\Big|\Big((\sigma^*)^{-1}W_T\Big)_i\Big|^2
\end{aligned}
$$

   where $C_{-i,-i}$ denotes the matrix C with ith row and column removed and $\mathbf{1}_{ii}$ denotes the $d\times d$ matrix with all zero elements except for at (i,i) position.

3. Case $\theta = C_{ij}, j\neq i$.

$$
\begin{aligned}
\partial_\theta\big(\log p_\theta(y)\big) &= -(-1)^{(i+j)}\frac{1}{2}\frac{\det(C_{-i,-j})}{\det(C)} + \frac{1}{2T}D\Big(\mathbf{1}_{ij}\Big)D,\\
\Xi &= -(-1)^{(i+j)}\frac{1}{2}\frac{\det(C_{-i,-j})}{\det(C)} + \frac{1}{2T}\Big((\sigma^*)^{-1}W_T\Big)_i \times \Big((\sigma^*)^{-1}W_T\Big)_j.
\end{aligned}
$$

6.10.3 Numerical results

In order to illustrate the application of POP method to estimate model sensitivity in the setting of (6.10.1), we consider a two-dimensional example which is similar to the example discussed in [72, Pg. 10]. We take interest rate $r = \mu^i = 0.01$ and for the other parameters $K = 100$, $T = 1$,

$\sigma^1 = 0.25$, $\sigma^2 = 0.225$, correlation parameter $C_{1,2} = 0.9$, $p_1 = 10$, $S_0^1 = 10$, $p_2 = 5$, $S_0^2 = 20$ and estimate the sensitivities of the rare event statistics $\mathbb{E}(K - p_1 S_T^1 - p_2 S_T^2)_+$ with respect to p_1, σ^1 and $C_{1,2}$. For other results and approximations related to deep out-of-the-money options, see for instance [56]. Observe that we choose $\sigma^2 = C_{1,2}\sigma^1$, which corresponds to the critical case described in [72, Theorem 1] where the asymptotics of the density of the basket undergoes a change of regime. It is thus arguably delicate to obtain a tractable analytical approximation via the derivation of the density.

In Tables 6.28, 6.29, 6.30, we compare the results of finite difference method using simple Monte Carlo with common random numbers [57] to those of the POP method with the number of simulations as indicated.

	Sensitivity w.r.t. p_1
POP method (10^6) (mean/std)	-0.7155 (0.0046)
Finite difference (10^6) (mean/std)	-0.7120 (0.0157)
Finite difference (10^9) (99% conf. interval)	$(-0.7155, -0.7129)$

TABLE 6.28: *Estimates of relative sensitivity w.r.t.* p_1.

	Sensitivity w.r.t. σ^1
POP method (10^6) (mean/std)	24.0078 (0.1760)
Finite difference (10^6) (mean/std)	23.9252 (0.4838)
Finite difference (10^9) (99% conf. interval)	$(23.9285, 24.0108)$

TABLE 6.29: *Estimates of relative sensitivity w.r.t.* σ^1

	Sensitivity w.r.t. $C_{1,2}$
POP method (10^6) (mean/std)	3.1058 (0.0253)
Finite difference (10^6) (mean/std)	3.0866 (0.1128)
Finite difference (10^9) (99% conf. interval)	$(3.0801, 3.0990)$

TABLE 6.30: *Estimates of relative sensitivity w.r.t. correlation.*

Here the rare event probability $\mathbb{P}(p_1 S_T^1 + p_2 S_T^2 \leq K)$ is around 1.7×10^{-3}. We deliberately choose such an example in order to show the application of our method in "not-so-rare" situations. Actually, when the rare event probability becomes smaller, the performance of POP method is considerably improved with respect to the simple Monte Carlo.

6.11 Optimal shaking parameter for standard normal distribution

Given all the above tables and figures, it is natural to wonder which shaking parameter will give the minimal standard deviation. Unfortunately, due to the implicit property of many theoretical results on Markov chain ergodicity, we are not able to do much theoretical work on it. Thus, in this section, we are going to conduct some numerical experiments on the simplest case, i.e. one dimensional normal distribution, to gain some intuitive insights.

We denote the 10^{-n} quantile of $\mathcal{N}(0,1)$ by q_n, i.e.

$$\mathbb{P}\left(G < q_n\right) = 10^{-n}$$

where $G \sim \mathcal{N}(0,1)$

6.11.1 Occupation measure estimation

We apply our POP algorithm to estimate

$$\mathbb{P}\left(G < q_{n+1} | G < q_n\right)$$

whose exact value is 0.1 as given by definition.

The initialization of Markov chain is done by applying the shaking transformation and only keeping the result when the value becomes smaller. The computational time of this step is negligible. Once the initial point is available, we will use M iterations to estimate our conditional probability and we also record the rejection rate. Among the M iterations, 1 percent is used as burn-in time to make the initial distribution close to the stationary one and we denote $M_{burn-in} = 0.01M$. The conditional probability estimator and the rejection rate are obtained by observing only the Markov chain after burn-in time. We say that we are at level k when n is equal to k.

The shaking transformation we use is

$$\mathcal{K}(X) = \rho X + \sqrt{1-\rho^2}\mathcal{N}(0,1)$$

we will take ρ going from 0.01 to 0.99 with step length 0.01. In Figures 6.11 - 6.20, we plot for each level the mean and standard deviation of our esitmators for different shaking parameters.

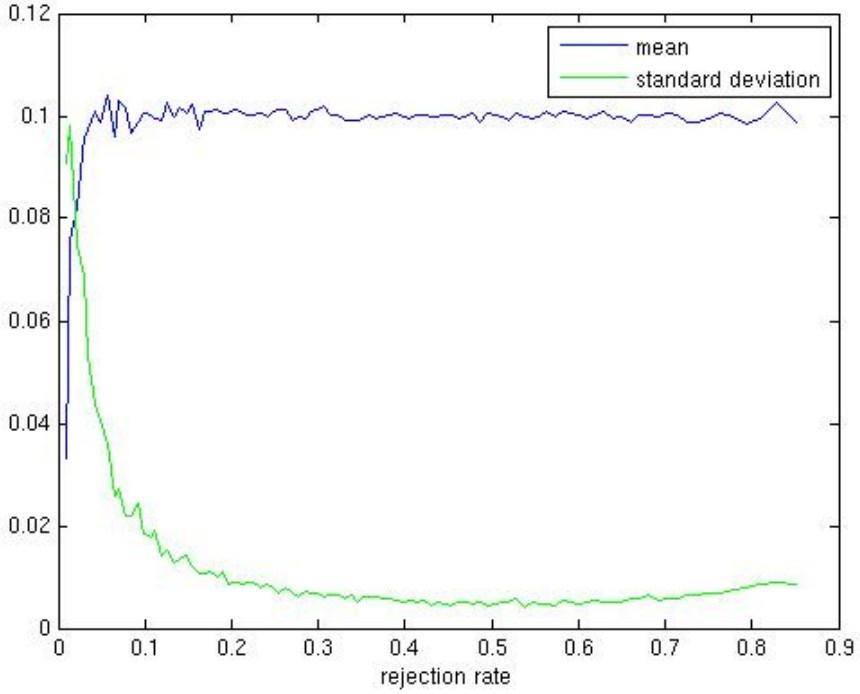


FIGURE 6.11: Level 1, $M = 10^4$, results over 100 macro-runs

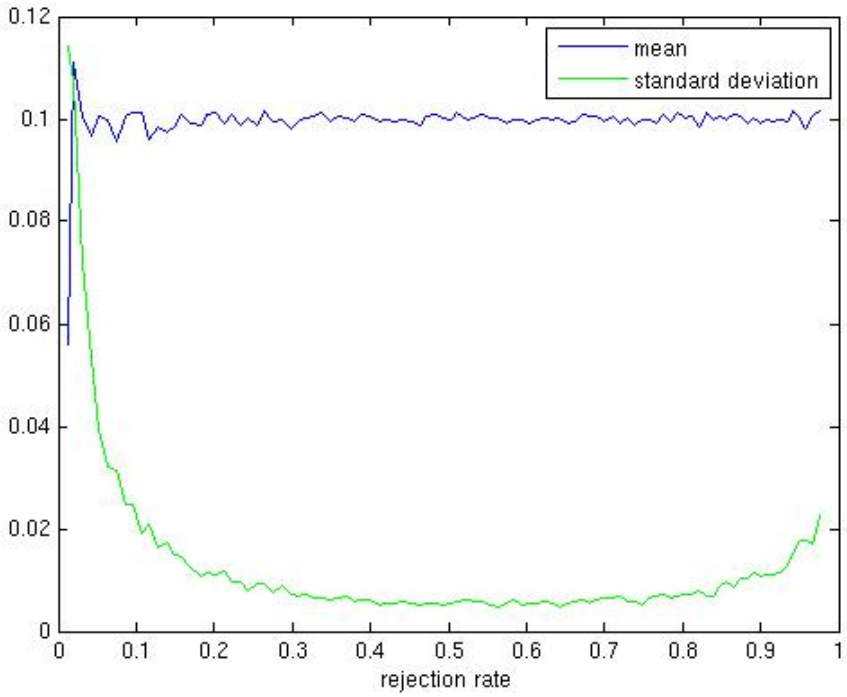


FIGURE 6.12: Level 2, $M = 10^4$, results over 100 macro-runs

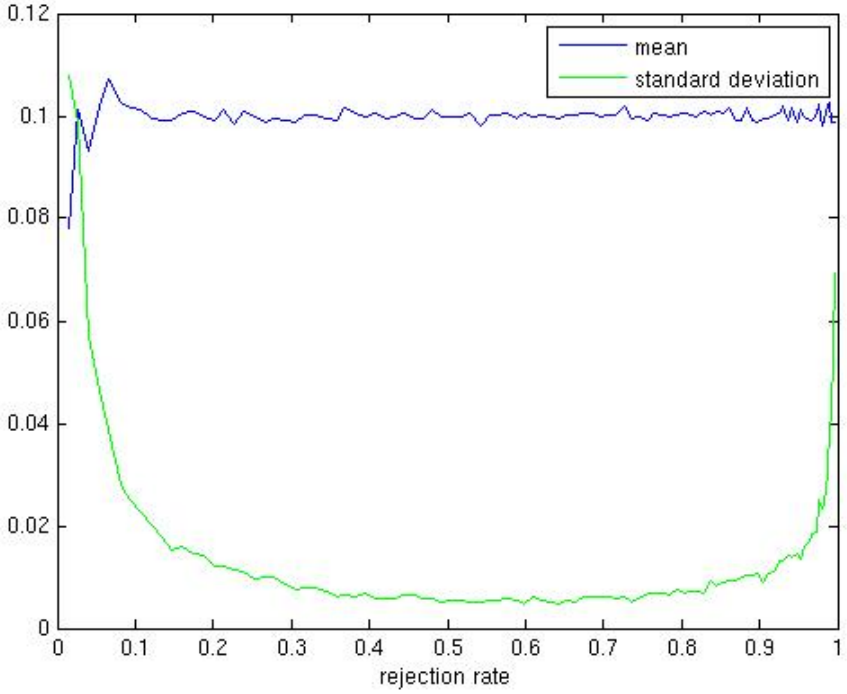


FIGURE 6.13: Level 3, $M = 10^4$, results over 100 macro-runs

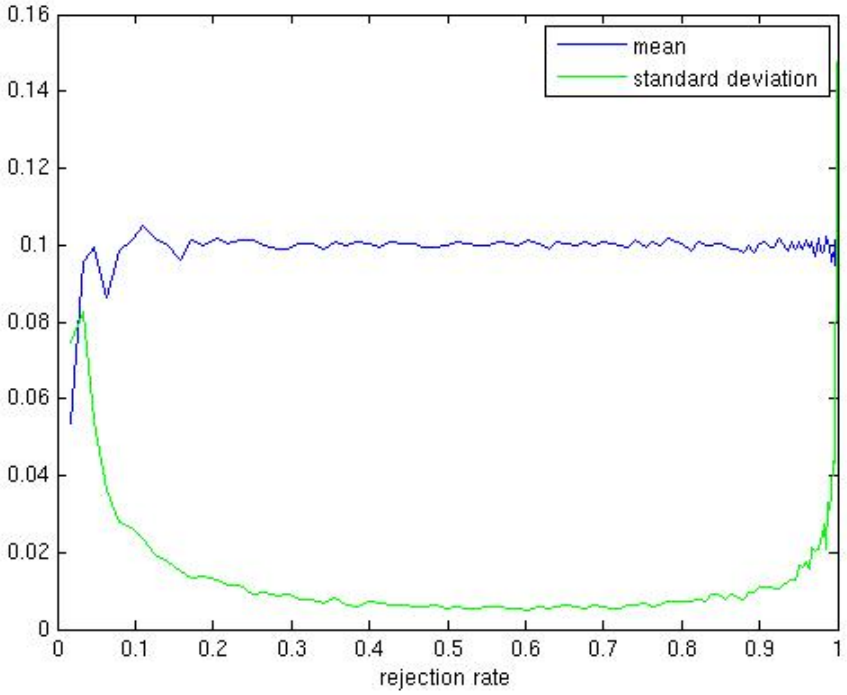


FIGURE 6.14: Level 4, $M = 10^4$, results over 100 macro-runs

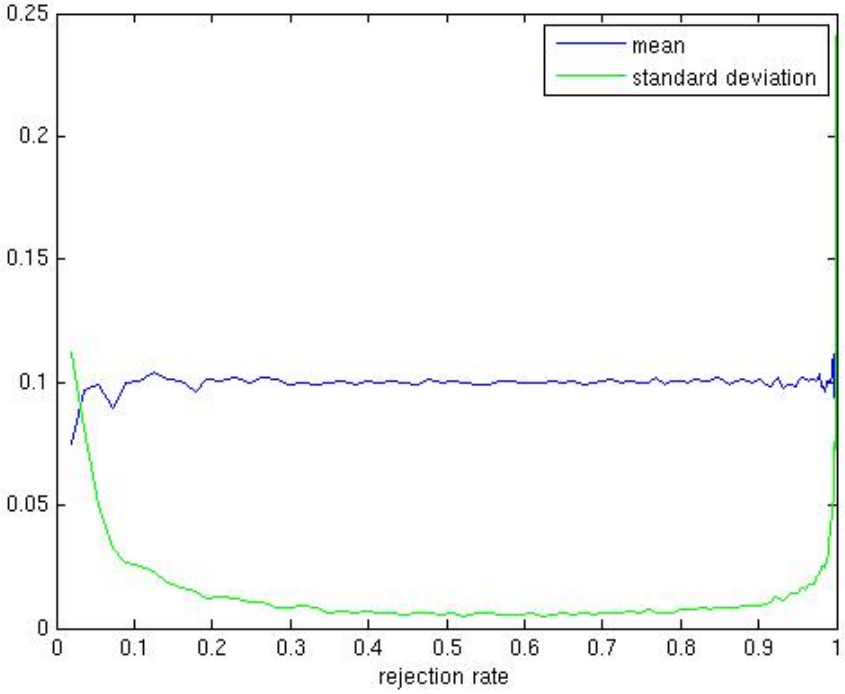


FIGURE 6.15: Level 5, $M = 10^4$, results over 100 macro-runs

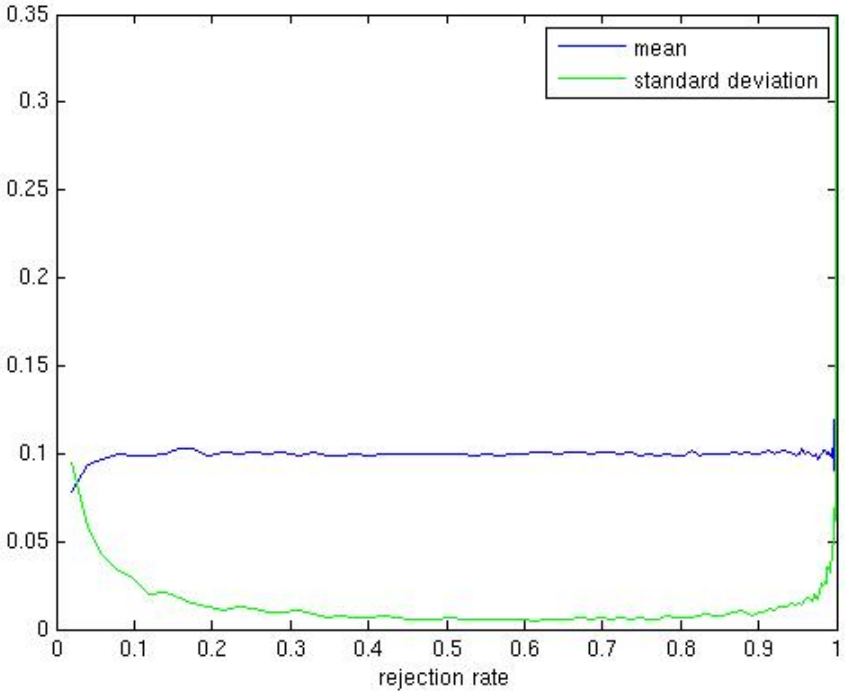


FIGURE 6.16: Level 6, $M = 10^4$, results over 100 macro-runs

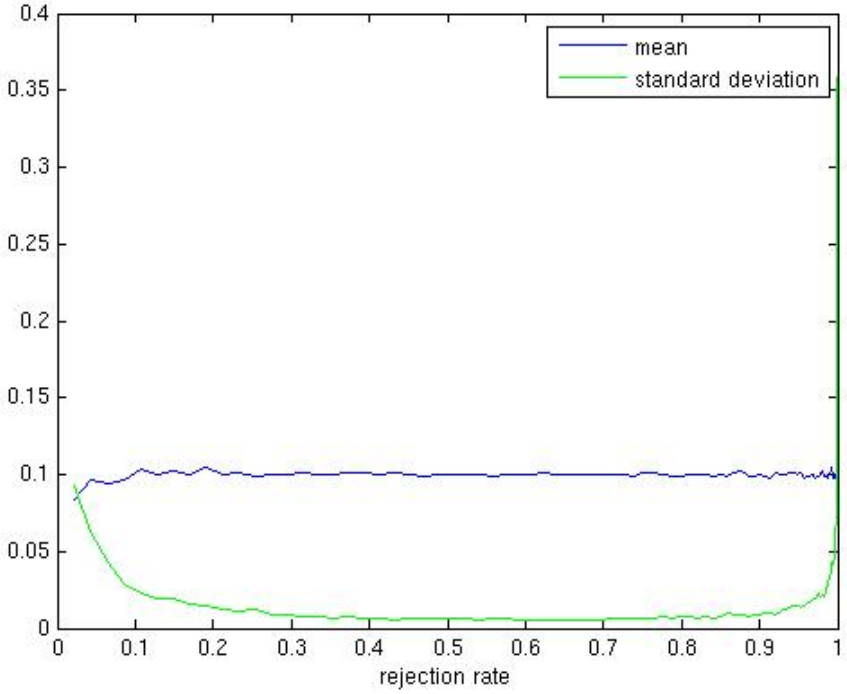


FIGURE 6.17: Level 7, $M = 10^4$, results over 100 macro-runs

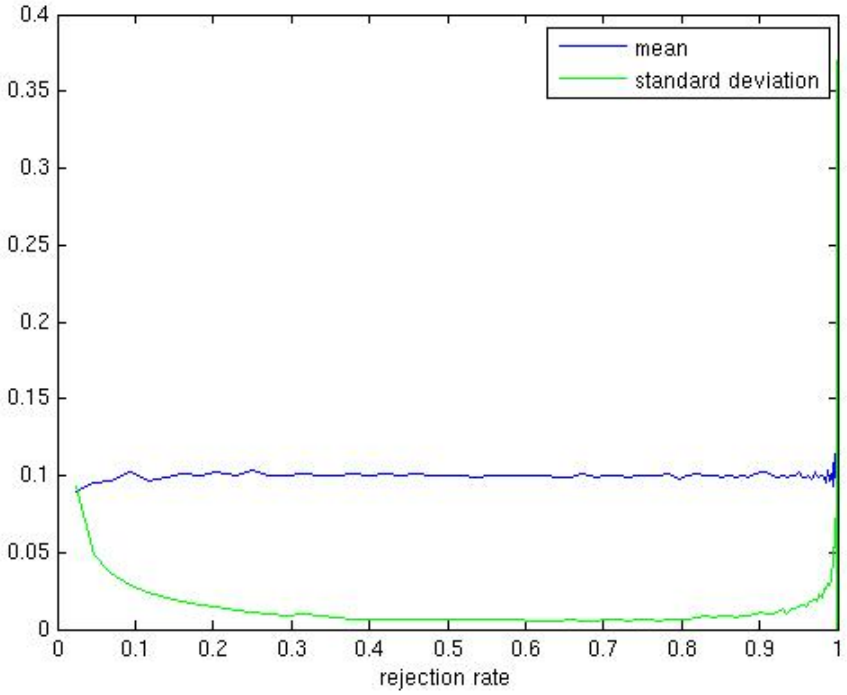


FIGURE 6.18: Level 8, $M = 10^4$, results over 100 macro-runs

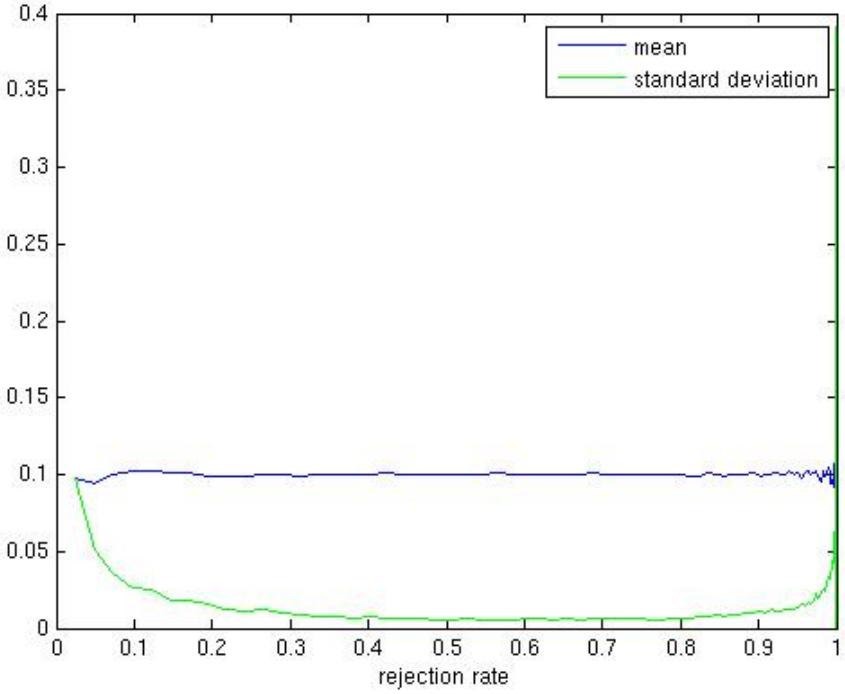


FIGURE 6.19: Level 9, $M = 10^4$, results over 100 macro-runs

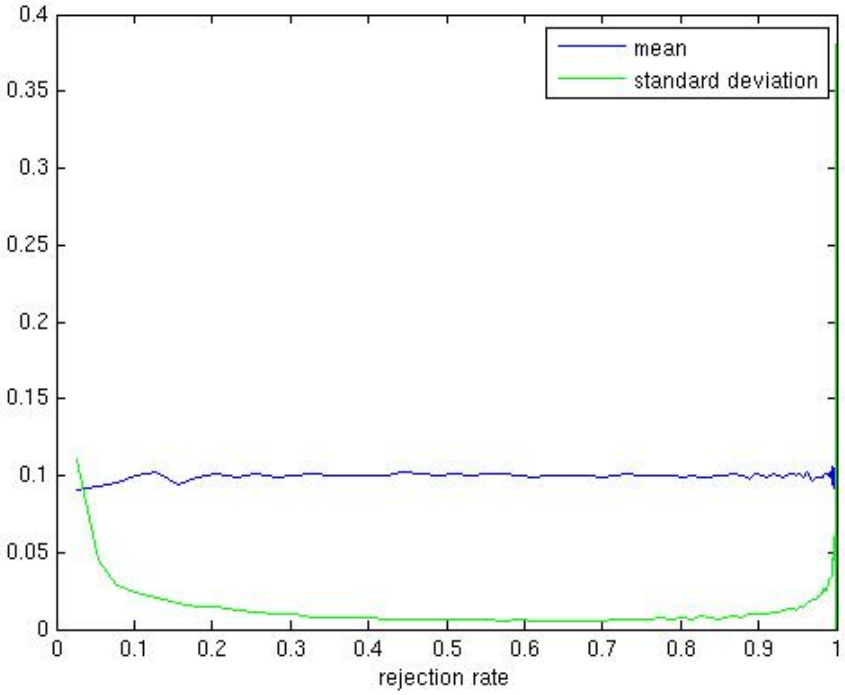


FIGURE 6.20: Level 10, $M = 10^4$, results over 100 macro-runs

The value of ρ which minimizes the standard deviation in each level and the corresponding standard deviation and rejection rate are given in the following table. Remark that since we consider only the empirical standard deviation of 100 marco runs, these values are not exact optimal ones.

	Level 1	Level 2	Level 3	Level 4	Level 5
ρ	0.69	0.78	0.88	0.91	0.92
std	0.0041	0.0048	0.0049	0.0050	0.0046
rejection rate	0.5373	0.6417	0.5980	0.6027	0.6234

	Level 6	Level 7	Level 8	Level 9	Level 10
ρ	0.94	0.93	0.96	0.97	0.98
std	0.0049	0.0052	0.0053	0.0051	0.0048
rejection rate	0.6130	0.6831	0.6083	0.5653	0.5716

In Figures 6.18, 6.19, 6.20, when the shaking is too strong (ρ close to 0), we get a numerical standard deviation 0, which is due to the fact that all the shaking transformations are rejected, so the outputs are constantly zero. Of course this is not what we want, so these 0 outputs are not considered when we give the above tables.

We can observe from these figures and tables that, the shaking forces should be calibrated according to the rejection rate, and the best rejection rates in this example are between 50% and 70%. To achieve the best rejection rates, the shaking force needs to be smaller and smaller when we go deep into the rare event zone.

Surprisingly, in this example, the best standard deviation does not deteriorate when we approach the rare event zone. This is a very good feature which may be related to the geometric property of our rare event zone.

6.11.2 Quantile estimation

Next, we give tables showing the best shaking parameters and rejection rates when we use Markov chain to estimate quantiles, i.e. we apply our POP algorithm to estimate the value of q_{n+1} given the rejection level q_n.

	Level 1	Level 2	Level 3	Level 4	Level 5
ρ	0.66	0.82	0.89	0.93	0.94
std	0.0159	0.0136	0.0122	0.0110	0.0102
rejection rate	0.5645	0.5941	0.5747	0.5517	0.5526

	Level 6	Level 7	Level 8	Level 9	Level 10
ρ	0.92	0.93	0.94	0.95	0.97
std	0.0088	0.0086	0.0086	0.0072	0.0079
rejection rate	0.6700	0.6982	0.6884	0.6887	0.5912

Similarly to the previous subsection, we observe an empirically optimal rejection rate between 50% and 70%

6.12 Population dynamics

The classic probabilistic model for describing the evolution of a population of individuals is the Galton-Watson model, which can also be regarded as a branching process with non random environment, see [8]. Theoretical analysis for these process with random enviroment has also been conducted, such as in [61].

Under this model, the size of the population at date $n+1$ is denoted by

$$Z_{n+1} = \sum_{j=1}^{Z_n} \xi_{n,j},$$

where the number of children $(\xi_{n,j})$ are i.i.d. variables. That is, each individual of the generation n will give birth to a certain number of children, following the same probability distribution independently. And the sum of all their children is the number of individuals at the generation $n+1$

We set $Z_0 = 1$ and $m = \mathbb{E}(Z_1)$ and we want to compute the survival probability

$$\mathbb{P}(Z_n > 0)$$

, which is an important issue for studying and preventing species extinctions, see [85]. Other works on this can be found in [11, 103, 82].

We suppose the reproduction law ξ is geometric $\mathcal{G}(p)$ on $\mathbb{N}$, i.e.

$$\mathbb{P}(\xi = k) = (1-p)^k p, \quad k \geqslant 0,$$

where $p = \frac{1}{m+1}$.

This is a case where we know the exact theoretical value, thus we can make accurate comparisons. Indeed, we know that the moment generating function $\Phi_{Z_n}(z) = \mathbb{E}(z^{Z_n})$ is given by

$$\frac{1}{1-\Phi_{Z_n}(z)} = \begin{cases} \frac{1-m^{-n}}{1-m^{-1}} + \frac{m^{-n}}{1-z} & \text{if } m \neq 1, \\ n + \frac{1}{1-z} & \text{if } m = 1, \end{cases}$$

and that

$$\mathbb{P}(Z_n > 0) = 1 - \Phi_{Z_n}(0).$$

Thus, using these closed-form probabilities, we can compare values obtained by our methods.

Shaking directly via geometric distribution We recall the shaking transformation for the random variable $X \sim \mathcal{G}(p)$ proposed in Subsection 5.4.3 :

$$K(X) = Y 1_{Y<X} + X 1_{Y \geqslant X, U<1-y} + (X + Z + 1) 1_{Y \geqslant X, U \geqslant 1-y}$$

where $Y \sim \mathcal{G}(x)$, $U \sim \mathcal{U}([0,1])$ and $Z \sim \mathcal{G}(p)$ with shaking parameters x, y satisfying$(1-p)x = py$.

In Tables 6.31 and 6.32, we report the results of POP methods over 100 times macro-runs for given parameter values. The intermediate levels are takes as $\{Z_k > 0\}, k = 1, 2, 3, 4, 5$

	$n = 5$	theoretical value
$m = 0.1, x = 0.05$	$9.59\times10^{-6}(4.62\times10^{-6})$	9×10^{-6}
$m = 0.1, x = 0.1$	$9.48\times10^{-6}(3.10\times10^{-6})$	
$m = 0.1, x = 0.2$	$8.96\times10^{-6}(2.24\times10^{-6})$	
$m = 0.1, x = 0.3$	$9.10\times10^{-6}(2.01\times10^{-6})$	
$m = 0.1, x = 0.4$	$9.31\times10^{-6}(2.26\times10^{-6})$	
$m = 0.1, x = 0.5$	$8.44\times10^{-6}(2.61\times10^{-6})$	
$m = 0.1, x = 0.6$	$9.15\times10^{-6}(3.73\times10^{-6})$	
$m = 0.1, x = 0.7$	$9.31\times10^{-6}(6.31\times10^{-6})$	
$m = 0.1, x = 0.8$	$8.67\times10^{-6}(1.08\times10^{-5})$	
$m = 0.1, x = 0.9$	$1.45\times10^{-5}(5.20\times10^{-5})$	
$m = 0.2, x = 0.05$	$2.70 \times10^{-4}(7.61 \times10^{-5})$	2.56×10^{-4}
$m = 0.2, x = 0.1$	$2.57 \times10^{-4}(5.24 \times10^{-5})$	
$m = 0.2, x = 0.2$	$2.51 \times10^{-4}(3.47 \times10^{-5})$	
$m = 0.2, x = 0.3$	$2.58 \times10^{-4}(4.15 \times10^{-5})$	
$m = 0.2, x = 0.4$	$2.58 \times10^{-4}(4.15 \times10^{-5})$	
$m = 0.2, x = 0.5$	$2.65 \times10^{-4}(5.63 \times10^{-5})$	
$m = 0.2, x = 0.6$	$2.64 \times10^{-4}(6.76 \times10^{-5})$	
$m = 0.2, x = 0.7$	$2.36 \times10^{-4}(1.02 \times10^{-4})$	
$m = 0.2, x = 0.8$	$2.46 \times10^{-4}(1.86 \times10^{-4})$	
$m = 0.2, x = 0.9$	$2.05 \times10^{-4}(4.08 \times10^{-4})$	

TABLE 6.31: Survival probability in the subcritical and critical cases ($m \leqslant 1$): format = POP outputs averaged over 100 runs (standard deviation)

Shaking via exponential distribution We then recall the shaking transformation for geometric distribution via exponential variable, as proposed in Subsection 5.4.3: we define $\lambda = -\ln(1-p)$ and

$$K(X) = \lfloor (X + \{E\})B + G \rfloor$$

where $E \sim \mathcal{E}(\lambda), B \sim Beta(1-x, x)$ and $G \sim \Gamma(x, \lambda)$ with shaking parameter $x \in [0, 1]$.

In Table 6.32, we report the results of POP methods over 100 times macro-runs for given parameter values, using this alternative shaking transformation

	$n = 5$	theoretical value
$m = 0.1, x = 0.05$	$8.40\times10^{-6}(2.69\times10^{-6})$	9×10^{-6}
$m = 0.1, x = 0.1$	$8.97\times10^{-6}(2.17\times10^{-6})$	
$m = 0.1, x = 0.2$	$8.95\times10^{-6}(2.55\times10^{-6})$	
$m = 0.1, x = 0.3$	$8.61\times10^{-6}(2.04\times10^{-6})$	
$m = 0.1, x = 0.4$	$8.87\times10^{-6}(2.72\times10^{-6})$	
$m = 0.1, x = 0.5$	$8.48\times10^{-6}(2.99\times10^{-6})$	
$m = 0.1, x = 0.6$	$9.75\times10^{-6}(4.56\times10^{-6})$	
$m = 0.1, x = 0.7$	$8.60\times10^{-6}(5.38\times10^{-6})$	
$m = 0.1, x = 0.8$	$1.04\times10^{-5}(1.04\times10^{-5})$	
$m = 0.1, x = 0.9$	$8.84\times10^{-6}(2.38\times10^{-5})$	
$m = 0.2, x = 0.05$	$2.56\times10^{-4}(5.44\times10^{-5})$	2.56×10^{-4}
$m = 0.2, x = 0.1$	$2.56\times10^{-4}(4.48\times10^{-5})$	
$m = 0.2, x = 0.2$	$2.65\times10^{-4}(3.98\times10^{-5})$	
$m = 0.2, x = 0.3$	$2.57\times10^{-4}(5.04\times10^{-5})$	
$m = 0.2, x = 0.4$	$2.55\times10^{-4}(4.16\times10^{-5})$	
$m = 0.2, x = 0.5$	$2.50\times10^{-4}(4.41\times10^{-5})$	
$m = 0.2, x = 0.6$	$2.63\times10^{-4}(6.46\times10^{-5})$	
$m = 0.2, x = 0.7$	$2.54\times10^{-4}(7.91\times10^{-5})$	
$m = 0.2, x = 0.8$	$2.45\times10^{-4}(1.06\times10^{-4})$	
$m = 0.2, x = 0.9$	$2.48\times10^{-4}(1.60\times10^{-4})$	

TABLE 6.32: Survival probability in the subcritical and critical cases ($m \leqslant 1$): format = POP outputs averaged over 100 runs (standard deviation)

Comparing Tables 6.31 and 6.32, we see that the two different ways of shaking geometric variables do not have significantly difference effects. The second way seems to deteriorate a bit more slowly when the shaking force becomes too strong.

6.13 Brownian watermelon

A Brownian bridge is the trajectory of a standard Brownian motion W_t between time $t = 0$ and $t = 1$, conditionally on $W_0 = W_1 = t$. It is well know that a Brownian bridge has the same distribution as $W_t - tW_1, t \in [0, 1]$, see [114], and thus can be simulated in this way.

If we have several trajectories of independent Brownian bridges and they do not cross each other, we get something which looks like a watermelon. We call it a Brownian watermelon. Thus a Brownian watermelon

is the configuration of several Brownian bridges conditionally on they don't intersect with each other. It has attracted attention in the field of theoretical physics, see [112, 91] for example.

It is not easy to make exact simulation of Brownian watermelon, since the probability that several Brownian bridges don't intersect with each other is very small. Using our shaking transformation, we can make approximative simulation of Brownian watermelon.

We will consider the system of Brownian bridges as a functional of several independent Brownian motion. Thus we can apply the shaking transformation given by Equation (5.3.1) with a constant parameter to shake the entire system of Brownian bridge.

We shall start with a given set of bridges which don't intersect with each other. Then we apply our shaking transformation on it. If some of these bridges cross each other after the shaking transformation, we will reject the result and just make a copy of the initial configuration. Then Proposition 5.8.1 in Section 5.8 ensures that after a large number of iterations, the current bridges configuration will be close to the exact distribution of a Brownian bridge. This procedure is illustration in the following figures.

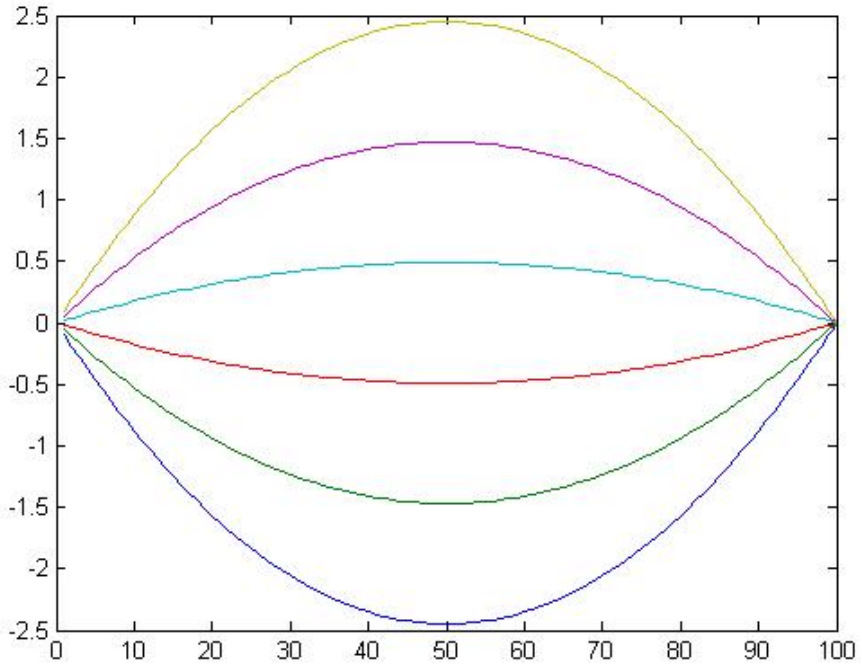


FIGURE 6.21: Initial configuration before we start

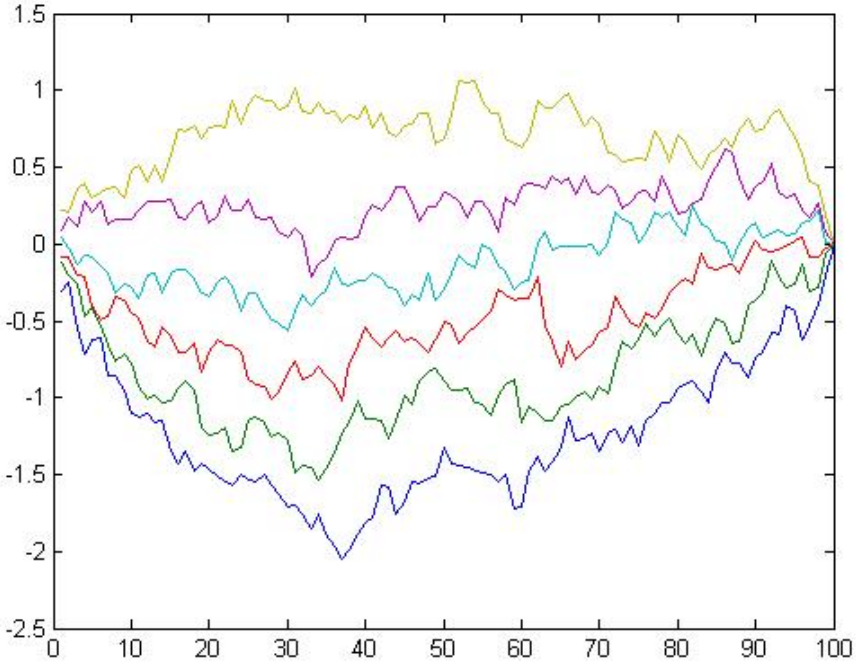


FIGURE 6.22: Final configuration after 10^5 iteration with $\rho = 0.999$

Note that we have taken a ρ which is very close to 1. According to our numerical experiments, more trajectories we have, more slightly we need to shake, otherwise most of the transformations will be rejected. Obviously, we can also compute average values along the way to estimate quantities related to a Brownian watermelon.

Part II

Non-intrusive Stratified Resampling Method for Dynamic Programming

Chapter 7

Non-intrusive stratified resampling

7.1 Introduction

Stochastic dynamic programming equations are classic equations arising in the resolution of nonlinear evolution equations, like in stochastic control (see [124, 14]), optimal stopping (see [96, 70]) or non-linear PDEs (see [34, 66]). In a discrete-time setting they take the form:

$$\begin{aligned} Y_N &= g_N(X_N), \\ Y_i &= \mathbb{E}\left(g_i(Y_{i+1}, \dots, Y_N, X_i, \dots, X_N) \mid X_i\right), \quad i = N-1, \dots, 0, \end{aligned}$$

for some functions g_N and g_i which depend on the non-linear problem under consideration. Here $X = (X_0, \dots, X_N)$ is a Markov chain valued in $\mathbb{R}^d$, entering also in the definition of the problem. The aim is to compute the value function y_i such that $Y_i = y_i(X_i)$.

Among the popular methods to solve this kind of problem, we are concerned with Regression Monte Carlo (RMC) methods that take as input M simulated paths of X, say $(X^1, \dots, X^M) =: X^{1:M}$, and provide as output simulation-based approximations $y_i^{M,\mathcal{L}}$ using Ordinary Least Squares (OLS) within a vector space of functions $\mathcal{L}$:

$$y_i^{M,\mathcal{L}} = \arg\inf_{\varphi \in \mathcal{L}} \frac{1}{M} \sum_{m=1}^{M} \left| g_i(y_{i+1}^{M,\mathcal{L}}(X_{i+1}^m), \dots, y_N^{M,\mathcal{L}}(X_N^m), X_i^m, \dots, X_N^m) - \varphi(X_i^m) \right|^2.$$

This Regression Monte Carlo methodology has been investigated in [66] to solve Backward Stochastic Differential Equations associated to semi-linear partial differential equations (PDEs) [106], with some tight error estimates. Generally speaking, it is well known that the number of simulations M has to be much larger than the dimension of the vector space $\mathcal{L}$ and thus the number of coefficients we are seeking.

In contradistinction, throughout this chapter, we focus on the case where M is relatively small (a few hundreds) and the simulations are not sampled by the user but are directly taken from historical data ($X^{1:M}$ is called **root sample**), in the spirit of [110]. This is the most realistic situation when we collect data and when the model which fits the data is unknown.

Thus, as a main difference with the aforementioned references:

- We do not assume that we have full information about the model for X and we do not assume that we can generate as many simulations as needed to have convergent Regression Monte Carlo methods.
- The size M of the learning samples $X^1, \dots, X^M$ is relatively small, which discards the use of a direct RMC with large dimensional $\mathcal{L}$.

To overcome these major obstacles, we elaborate on two ingredients:

1. First, we partition $\mathbb{R}^d$ in strata $(\mathcal{H}_k)_k$, so that the regression functions can be computed locally on each stratum $\mathcal{H}_k$; for *small* stratum this allows to use only a small dimensional approximation space $\mathcal{L}_k$, and therefore it puts a lower constraint on M. In general, this stratification breaks the properties for having well-behaved error propagation and we provide a precise way to sample in order to be able to aggregate the error estimates in different strata. We use a probabilistic distribution ν that has good norm-stability properties with X (see Assumptions 7.3.2 and 7.4.2).
2. Second, by assuming a mild model condition on X, we are able to resample from the root sample of size M, a *training sample* of M simulations suitable for the stratum $\mathcal{H}_k$. This resampler is non intrusive in the sense that it only requires to know the form of the model but not its coefficients: for example, we can handle models with independent increments (discrete inhomogeneous Levy process) or Ornstein-Uhlenbeck processes. See Examples 7.2.1-7.2.2-7.2.3-7.2.4. We call this scheme NISR (Non Intrusive Stratified Resampler), it is described in Definition 7.2.1 and Proposition 7.2.1.

The resulting regression scheme is, to the best of our knowledge, completely new. To sum up, the contributions of this work are the following:

- We design a non-intrusive stratified resample (NISR) scheme that allows to sample from M paths of the root sample restarting from any stratum $\mathcal{H}_k$. See Section 7.2.
- We combine this with regression Monte Carlo schemes, in order to solve one-step ahead dynamic programming equations (Section 7.3), discrete backward stochastic differential equations (BSDEs) and semi-linear PDEs (Section 7.4).
- In Theorems 7.3.4 and 7.4.1, we provide quadratic error estimates of the form

$$\begin{aligned}\text{quadratic error on } y_i \leqslant\ & \text{approximation error } + \text{ statistical error} \\ & + \text{ interdependency error }.\end{aligned}$$

The approximation error is related to the best approximation of y_i on each stratum $\mathcal{H}_k$, and averaged over all the strata. The statistical error is bounded by C/M with a constant C which does not depend on the number of strata: only relatively small M is necessary to get low statistical errors. This is in agreement with the motivation that the root sample has a relatively small size. The interdependency error is an unusual issue, it is related to the strong dependency between regression problems (because they all use the same root sample). The analysis as well as the framework are original. The error estimates take different forms according to the problem at hand (Section 7.3 or Section 7.4).

- Finally we illustrate the performance of the methods on two types of examples: first, approximation of non-linear PDEs arising in reaction-diffusion biological models (Subsection 8.1) and optimal sequential decision (Subsection 8.2), where we illustrate that root samples of size $M = 20 - 40$ only can lead to remarkably accurate numerical solutions.

This chapter is organized as follows. In Section 7.2 we present the model structure that leads to the non-intrusive stratified resampler for regression Monte Carlo (NISR), together with the stratification. Main notations will be also introduced. The algorithm is presented in a generic form of dynamic programming equations in Algorithm 4. In Section 7.3 we analyze the convergence of the algorithm in the case of one-step ahead dynamic programming equations (for instance optimal stopping problems). Section 7.4 is devoted to the convergence analysis for discrete BSDEs (probabilistic representation of semi-linear PDEs arising in stochastic control problems). Numerical examples are provided in the next chapter. Technical results are postponed to the Appendix. Most of the materials in this part are contained in our paper [69].

7.2 Setting and the general algorithm

7.2.1 General dynamic programming equation

Suppose we have N discrete dates, and we aim at solving numerically the following dynamic programming equation (DPE for short), written in general form:

$$\begin{aligned} Y_N &= g_N(X_N), \\ Y_i &= \mathbb{E}\left(g_i(Y_{i+1:N}, X_{i:N}) \mid X_i\right), \quad 0 \leqslant i < N. \end{aligned}$$

Here, $(X_i)_{0\leqslant i\leqslant N}$ is a Markov chain with state space $\mathbb{R}^d$, $(Y_i)_{0\leqslant i\leqslant N}$ is a random process taking values in $\mathbb{R}$ and we use for convenience the generic short notation $z_{i:N} := (z_i, \ldots, z_N)$. Note that the argument of the conditional expectation is path-dependent, thus allowing greater generality.

Had we considered Y to be multidimensional, the subsequent algorithm and analysis would remain essentially the same.

Later (Sections 7.3 and 7.4), specific forms for g_i will be considered, depending on the model of DPE to solve at hand: it will have an impact on the error estimates that we can derive. However, the description of the algorithm can be the same for all the DPEs, as seen below, and this justifies our choice of unifying the presentation.

Thanks to the Markovian property of X, under mild assumptions we can easily prove by induction that there exists a measurable function y_i such that $Y_i = y_i(X_i)$, our aim is to compute an approximation of the value functions $y_i(.)$ for all i. We assume that a bound on y_i is available.

Assumption 7.2.1 (A priori bound)**.** *The solution y_i is bounded by a constant $|y_i|_\infty$.*

7.2.2 Model structure and root sample

We will represent $y_i(.)$ through its coefficients on a vector space, and the coefficients will be computed thanks to learning samples of X.

Assumption 7.2.2 (Data)**.** *We have the observation of M independent paths of X, which are denoted by $((X_i^m : 0 \leqslant i \leqslant N), 1 \leqslant m \leqslant M)$. We refer to this data as the* root sample*.*

For our needs, we adopt a representation of the *flow of the Markov chain for different initial conditions*, i.e., the Markov chain $X^{i,x}$ starting at different times $i \in \{0, \dots, N\}$ and points $x \in \mathbb{R}^d$. Namely, we write

$$X_j^{i,x} = \theta_{i,j}(x, U), \quad i \leqslant j \leqslant N, \tag{7.2.1}$$

where

- U is some random vector, called random source,
- $\theta_{i,j}$ are (deterministic) measurable functions.

We emphasize that, for the sake of convenience, U is the same for representing all $X_j^{i,x}, 0 \leqslant i \leqslant j \leqslant N, x \in \mathbb{R}^d$.

Assumption 7.2.3 (Noise extraction)**.** *We assume that $\theta_{i,j}$ are known and we can retrieve the random sources $(U^1, \dots, U^M)$ associated to the root sample $X^{1:M} = (X^m : 1 \leqslant m \leqslant M)$, i.e.,*

$$X_j^m = X_j^{0,x_0^m,m} = \theta_{0,j}(x_0^m, U^m).$$

Observe that this assumption is much less stringent than identifying the distribution of the model. We exemplify this now.

Example 7.2.1 (Arithmetic Brownian motion with time dependent parameters). *Let $(t_i : 0 \leqslant i \leqslant N)$ be N times and define the arithmetic Brownian motion by*

$$X_i = x_0 + \int_0^{t_i} \mu_s \mathrm{d}s + \int_0^{t_i} \sigma_s \mathrm{d}W_s$$

where $\mu_t \in \mathbb{R}^d, \sigma_t \in \mathbb{R}^{d\times q}, W_t \in \mathbb{R}^q$ and μ, σ are deterministic functions of time. In this case, the random source is given by

$$U := (X_{i+1} - X_i)_{0\leqslant i\leqslant N-1}$$

and the functions by

$$\theta_{ij}(x, U) := x + \sum_{i\leqslant k<j} U_k.$$

This works since $U_i = \int_{t_i}^{t_{i+1}} \mu_s \mathrm{d}s + \int_{t_i}^{t_{i+1}} \sigma_s \mathrm{d}W_s$. The crucial point is that, in order to extract U from X, we do not assume that μ and σ are known.

Example 7.2.2 (Levy process). *More generally, we can set $X_i = \mathbf{X}_{t_i}$ with a time-inhomogeneous Levy process $\mathbf{X}$. Then take*

$$U := (X_{i+1} - X_i)_{0\leqslant i\leqslant N-1}, \qquad \theta_{ij}(x, U) := x + \sum_{i\leqslant k<j} U_k.$$

Example 7.2.3 (Geometric Brownian motion with time dependent parameters). *With the same kind of parameters as for Example 7.2.1, define the geometric Brownian motion (component by component)*

$$X_i = X_0 \exp\left(\int_0^{t_i} \mu_s \mathrm{d}s + \int_0^{t_i} \sigma_s \mathrm{d}W_s\right).$$

Then, we have that

$$U := \left(\log(\frac{X_{i+1}}{X_i})\right)_{0\leqslant i\leqslant N-1}, \qquad \theta_{ij}(x, U) := x \prod_{i\leqslant k<j} \exp(U_k).$$

Example 7.2.4 (Ornstein-Uhlenbeck process with time dependent parameters). *Given N times $(t_i : 0 \leqslant i \leqslant N)$, set $X_i = \mathbf{X}_{t_i}$ where $\mathbf{X}$ has the following dynamics:*

$$\mathbf{X}_t = \mathbf{x}_0 - \int_0^t A(\mathbf{X}_s - \bar{\mathbf{X}}_s)\mathrm{d}s + \int_0^t \Sigma_s \mathrm{d}W_s$$

where A is $d \times d$-matrix, $\mathbf{X}_t$ and $\bar{\mathbf{X}}_t$ are in $\mathbb{R}^d$, Σ_t is a $d \times q$-matrix, $W_t \in \mathbb{R}^q$. $\bar{\mathbf{X}}_t$ and Σ_t are both deterministic functions of time. The explicit solution is

$$\mathbf{X}_t = e^{-A(t-s)}\mathbf{X}_s + e^{-At} \int_s^t e^{Ar}(A\bar{\mathbf{X}}_r \mathrm{d}r + \Sigma_r \mathrm{d}W_r).$$

Assume that we know A: in this case, an observation of $X_{0:N}$ enables to retrieve the random source

$$U := \left(X_j - e^{-A(t_j - t_i)} X_i\right)_{0 \leqslant i \leqslant j \leqslant N}$$

and then

$$\theta_{ij}(x, U) := e^{-A(t_j - t_i)} x + U_{i,j}.$$

The noise extraction works since $U_{i,j} = e^{-At_j} \int_{t_i}^{t_j} e^{Ar}(A\bar{\mathbf{X}}_r \mathrm{d}r + \Sigma_r \mathrm{d}W_r)$.

As illustrated above, through Assumption 7.2.2, all we need to know is the general structure of the Markov chain model but we do not need to estimate all the model parameters, and sometimes none of them (Examples 7.2.1, 7.2.2, 7.2.3). Our approach is non intrusive in this sense.

7.2.3 Stratification and resampling algorithm

On the one hand, we can rely on a root sample of size M only (possibly with a relatively small M, constrained by the available data), which is very little to perform accurate Regression Monte-Carlo methods (usually M has to be much larger than the dimension of approximation spaces, as reminded in introduction).

On the other hand, we are able to access the random sources so that resampling the M paths is possible. The degree of freedom comes from the flexibility of initial conditions (i, x), thanks to the flow representation (7.2.1). We now explain how we take advantage of this property.

The idea is to resample the model paths for different starting points in different parts of the space $\mathbb{R}^d$ and on each part, we will perform a regression Monte Carlo using M paths and a low-dimensional approximation space. These ingredients give the ground reasons for getting accurate results.

Let us proceed to the details of the algorithm. We design a stratification approach: suppose there exist K strata $(\mathcal{H}_k)_{1 \leqslant k \leqslant K}$ such that

$$\mathcal{H}_k \cap \mathcal{H}_l = \emptyset \quad \text{for } k \neq l, \qquad \bigcup_{k=1}^{K} \mathcal{H}_k = \mathbb{R}^d.$$

An example for $\mathcal{H}_k$ is a hypercube of the form $\mathcal{H}_k = \prod_{l=1}^{d} [x_{k,l}^-, x_{k,l}^+)$. Then, we are given a probability measure ν on $\mathbb{R}^d$ and denote its restriction on $\mathcal{H}_k$ by

$$\nu_k(\mathrm{d}x) := \frac{1}{\nu(\mathcal{H}_k)} 1_{\mathcal{H}_k}(x) \nu(\mathrm{d}x).$$

The measure ν will serve as a reference to control the errors. See Paragraph 7.3.1 and Chapter 8 for choices of ν.

Definition 7.2.1 (Non-intrusive stratified resampler, NISR for short)**.** *We define the M-sample used for regression at time i and in the k-th stratum $\mathcal{H}_k$:*

- *let* $(X_i^{i,k,m})_{1\leqslant m\leqslant M}$ *be an i.i.d. sample according to the law* ν_k*;*
- *for* $j = i+1, \dots, N$*, set*

$$X_j^{i,k,m} = \theta_{i,j}(X_i^{i,k,m}, U^m)\,,$$

where $U^{1:M}$ *are the random sources from Assumption 7.2.3.*

In view of Assumptions 7.2.2 and 7.2.3, the random sources $U^1, \dots, U^M$ are independent, therefore we easily prove the following.

Proposition 7.2.1. *The* M *paths* $(X_{i:N}^{i,k,m}, 1 \leqslant m \leqslant M)$ *are independent and identically distributed as* $X_{i:N}$ *with* $X_i \overset{d}{\sim} \nu_k$*.*

7.2.4 Approximation spaces and regression Monte Carlo schemes

On each stratum, we approximate the value functions y_i using basis functions. We can take different kinds of basis functions:

- **LP**$_0$ (partitioning estimate): $\mathcal{L}_k = \mathrm{span}(1_{\mathcal{H}_k})$,
- **LP**$_1$ (piecewise linear): $\mathcal{L}_k = \mathrm{span}(1_{\mathcal{H}_k}, x_1 1_{\mathcal{H}_k}, \cdots, x_d 1_{\mathcal{H}_k})$,
- **LP**$_n$ (piecewise polynomial): $\mathcal{L}_k = \mathrm{span}($ all the polynomials of degree less than or equal to n on $\mathcal{H}_k)$.

To simplify the presentation, we assume hereafter that the dimension of $\mathcal{L}_k$ does not depend on k, we write

$$\dim(\mathcal{L}_k) =: \dim(\mathcal{L}).$$

To compute the approximation of y_i on each stratum $\mathcal{H}_k$, we will use the M samples of Definition 7.2.1. Our NISR-regression Monte Carlo algorithm takes the form:

set $y_N^{(M)}(\cdot) = g_N(\cdot)$
for $i = N-1$ until 0 **do**
  for $k = 1$ until K **do**
    sample $(X_{i:N}^{i,k,m})_{1\leqslant m\leqslant M}$ using the NISR (Definition 7.2.1)
    set $S^{(M)}(x_{i:N}) = g_i(y_{i+1}^{(M)}(x_{i+1}), \dots, y_N^{(M)}(x_N), x_{i:N})$
    compute $\psi_i^{(M),k} = \mathrm{OLS}(S^{(M)}, \mathcal{L}_k, X_{i:N}^{i,k,1:M})$
    set $y_i^{(M),k} = \mathcal{T}_{|y_i|_\infty}(\psi_i^{(M),k})$ where $\mathcal{T}_L$ is the truncation operator, defined by $\mathcal{T}_L(x) = -L \vee x \wedge L$
  end
  set $y_i^{(M)} = \sum_{k=1}^K y_i^{(M),k} 1_{\mathcal{H}_k}$
end

Algorithm 4: General NISR-regression Monte Carlo algorithm

In the above, the Ordinary Least Squares approximation of the response function $\tilde{S} : (\mathbb{R}^d)^{N-i+1} \mapsto \mathbb{R}$ in the function space $\mathcal{L}_k$ using the M sample $X_{i:N}^{i,k,1:M}$ is defined and denoted by

$$\mathrm{OLS}(\tilde{S}, \mathcal{L}_k, X_{i:N}^{i,k,1:M}) = \arg \inf_{\varphi \in \mathcal{L}_k} \frac{1}{M} \sum_{m=1}^{M} |\tilde{S}(X_{i:N}^{i,k,m}) - \varphi(X_i^{i,k,m})|^2.$$

The main difference with the usual regression Monte-Carlo schemes (see [64] for instance) is that here we use the common random numbers $U^{1:M}$ for all the regression problems. This is the effect of resampling. The convergence analysis becomes more delicate because we lose nice independence properties. Figure 7.1 describes a key part in the algorithm, namely the process of using the root paths to generate new paths.

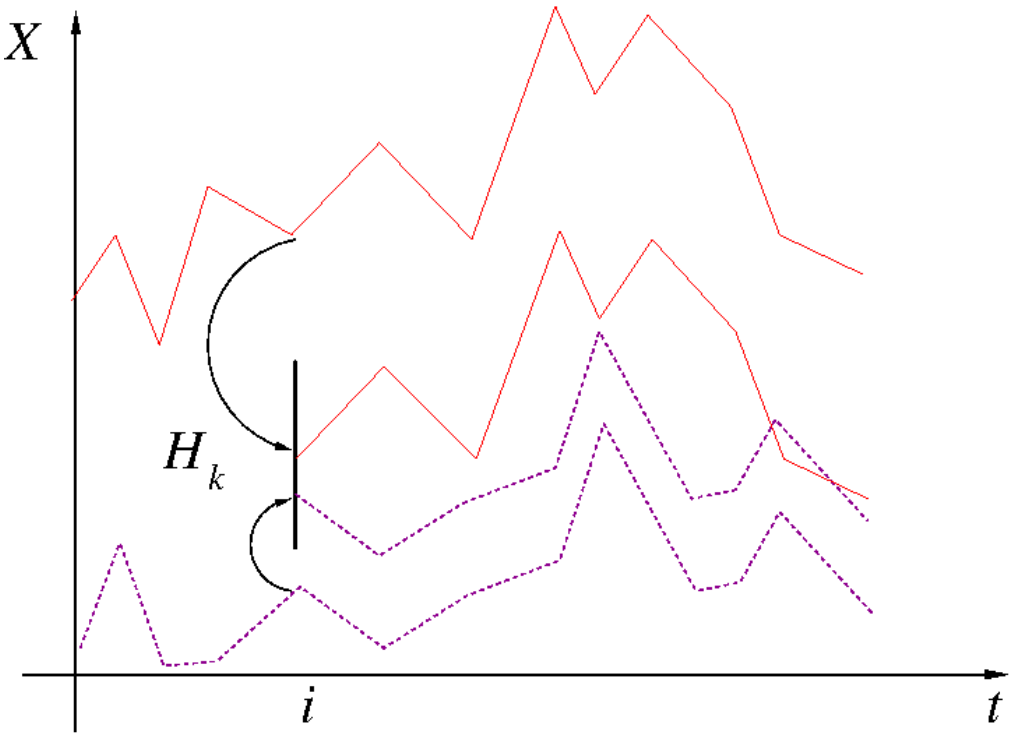


FIGURE 7.1: Description of the use of the root paths to produce new paths in an arbitrary hypercube.

7.3 Convergence analysis in the case of the one-step ahead dynamic programming equation

We consider here the case

$$\begin{aligned} Y_N &= g_N(X_N), \\ Y_i &= \mathbb{E}\left(g_i(Y_{i+1}, X_i, \ldots, X_N) \mid X_i\right), \quad 0 \leqslant i < N, \end{aligned}$$

where we need the value of Y_{i+1} (one step ahead) to compute the value Y_i (at the current date) through a conditional expectation. To compare with Algorithm 4, we take $g_i(Y_{i+1:N}, X_{i:N}) = g_i(Y_{i+1}, X_{i:N})$.

Equations of this form are quite natural when solving optimal stopping problems (see [96, 70]) in the Markovian case. Indeed, if V_i is the related value function at time i, i.e., the essential supremum over stopping times $\tau \in \{i, \dots, N\}$ of a reward process $f_\tau(X_\tau)$, then $V_i = \max(Y_i, f_i(X_i))$ where Y_i is the continuation value defined by

$$Y_i = \mathbb{E}\left(\max(Y_{i+1}, f_{i+1}(X_{i+1})) \mid X_i\right) ,$$

see [124] for instance. This corresponds to our setting with

$$g_i(y_{i+1}, x_{i:N}) = \max(y_{i+1}, f_{i+1}(x_{i+1})) .$$

Similar dynamic programming equations appear in stochastic control problems. See [14].

7.3.1 Standing assumptions

The following assumptions enable us to provide error estimates (Theorem 7.3.4 and Corollary 7.3.1) for the convergence of Algorithm 4.

Assumptions on g_i

Assumption 7.3.1 (Functions g_i)**.** *Each function g_i is Lipschitz w.r.t. the variable y_{i+1}, with Lipschitz constant L_{g_i} and $C_{g_i} := \sup_{x_{i:N}} |g_i(0, x_{i:N})| < +\infty$.*

It is then easy to justify that y_i (such that $y_i(X_i) = Y_i$) is bounded (Assumption 7.2.1).

Assumptions on the distribution ν

We assume a condition on the probability measure ν and the Markov chain X, which ensures a suitable stability in the propagation of errors.

Assumption 7.3.2 (norm-stability)**.** *There exists a constant $\underline{C}_{(7.3.1)} \geqslant 1$ such that for any $\varphi : \mathbb{R}^d \mapsto \mathbb{R} \in L^2(\nu)$ and any $0 \leqslant i \leqslant N-1$, we have*

$$\int_{\mathbb{R}^d} \mathbb{E}\left(\varphi^2(X_{i+1}^{i,x})\right) \nu(\mathrm{d}x) \leqslant \underline{C}_{(7.3.1)} \int_{\mathbb{R}^d} \varphi^2(x)\nu(\mathrm{d}x). \tag{7.3.1}$$

We now provide some examples of distribution ν where the above assumption holds, in connection with Examples 7.2.1, 7.2.2 and 7.2.4.

Proposition 7.3.1. *Let $\alpha = (\alpha^1, \dots, \alpha^d) \in]0, +\infty[^d$ and assume that $X_{i+1}^{i,x} = x + U_i$ (as in Examples 7.2.1 and 7.2.2) with $\mathbb{E}\left(\prod_{j=1}^d e^{\alpha^j |U_i^j|}\right) < +\infty$. Then,*

the tensor-product Laplace distribution

$$\nu(\mathrm{d}x) := \prod_{j=1}^{d} \frac{\alpha^j}{2} e^{-\alpha^j |x^j|} \mathrm{d}x$$

satisfies Assumption 7.3.2.

Proof. The L.H.S. of (7.3.1) writes

$$\begin{aligned}\mathbb{E}\left(\int_{\mathbb{R}^d} \varphi^2(x+U_i)\nu(\mathrm{d}x)\right) &= \mathbb{E}\left(\int_{\mathbb{R}^d} \varphi^2(x) \prod_{j=1}^{d} \frac{\alpha^j}{2} e^{-\alpha^j |x^j - U_i^j|} \mathrm{d}x\right) \\ &\leqslant \mathbb{E}\left(\int_{\mathbb{R}^d} \varphi^2(x) \prod_{j=1}^{d} \frac{\alpha^j}{2} e^{-\alpha^j |x^j| + \alpha^j |U_i^j|} \mathrm{d}x\right)\end{aligned}$$

which leads to the announced inequality (7.3.1) with

$$\underline{C}_{(7.3.1)} := \mathbb{E}\left(\prod_{j=1}^{d} e^{\alpha^j |U_i^j|}\right)$$

. □

Proposition 7.3.2. *Let $k > 0$ and assume that $X_{i+1}^{i,x} = Dx + U_i$ for a diagonal invertible matrix $D := \mathrm{diag}(D^1, \ldots, D^d)$ (a form similar to Example 7.2.4) with $\mathbb{E}\left((1+|U_i|)^{d(k+1)}\right) < +\infty$. Then, the tensor-product Pareto-type distribution*

$$\nu(\mathrm{d}x) := \prod_{j=1}^{d} \frac{k}{2} (1+|x^j|)^{-k-1} \mathrm{d}x$$

satisfies Assumption 7.3.2.

Proof. The L.H.S. of (7.3.1) equals

$$\begin{aligned}&\mathbb{E}\left(\int_{\mathbb{R}^d} \varphi^2(Dx+U_i)\nu(\mathrm{d}x)\right) \\ &= \mathbb{E}\left(\int_{\mathbb{R}^d} \varphi^2(x) \det(D^{-1}) \prod_{j=1}^{d} \frac{k}{2} (1+|(x^j - U_i^j)/D^j|)^{-k-1} \mathrm{d}x\right) \\ &\leqslant \int_{\mathbb{R}^d} \varphi^2(x) \det(D^{-1}) \prod_{j=1}^{d} \frac{k}{2} \left(\mathbb{E}\left((1+|(x^j - U_i^j)/D^j|)^{-d(k+1)}\right)\right)^{1/d} \mathrm{d}x.\end{aligned} \tag{7.3.2}$$

On the set $\{|U_i^j| \leqslant |x^j|/2\}$ we have $(1+|(x^j - U_i^j)/D^j|) \geqslant (1+(|x^j| - |U_i^j|)/D^j|) \geqslant (1+|x^j|/(2D^j))$. On the complementary set $\{|U_i^j| > |x^j|/2\}$, the random variable inside the j-th expectation in (7.3.2) is bounded by 1 and furthermore

$$\mathbb{P}\left(|U_i^j| > |x^j|/2\right) \leqslant \frac{\mathbb{E}\left((1+2|U_i^j|)^{d(k+1)}\right)}{(1+|x^j|)^{d(k+1)}}.$$

By gathering the two cases, we observe that we have shown that the j-th expectation in (7.3.2) is bounded by $\mathrm{Cst}(1+|x^j|)^{-d(k+1)}$, for any x^j, whence the advertised result. □

Remarks.

- Since we will apply the inequality (7.3.1) only to functions in a finite dimensional space, the norm equivalence property of finite dimensional space may also give the existence of a constant $\underline{C}_{(7.3.1)}$. But the constant built in this way could depend on the finite dimensional space (and may blow up when its dimension increases) while here the constant is valid for any φ.
- The previous examples on ν are related to distributions with independent components: this is especially convenient when one has to sample ν restricted to hypercubes $\mathcal{H}_k$, since we are reduced to independent one-dimensional simulations.
- In Proposition 7.3.2, had the matrix D been symmetric instead of diagonal, we would have applied an appropriate rotation to the density ν.

Covering number of an approximation space

To analyze how the M-samples $(X_{i:N}^{i,k,m}, 1 \leqslant m \leqslant M)$ from NISR approximates the exact distribution of $X_{i:N}$ with $X_i \overset{d}{\sim} \nu_k$ over test functions in the space $\mathcal{L}_k$, we will invoke concentration of measure inequalities (uniform in $\mathcal{L}_k$). This is possible thanks to complexity estimates related to $\mathcal{L}_k$, expressed in terms of covering numbers. Note that the concept of covering numbers is mainly used to introduce Assumption 7.3.3 and it intervenes in the main theorems only through the proof of Proposition 7.3.5.

We briefly recall the definition of a covering number of a dictionary of functions $\mathcal{G}$, see [75, Chapter 9] for more details. For a dictionary $\mathcal{G}$ of functions from $\mathbb{R}^d$ to $\mathbb{R}$ and for M points $x^{1:M} := \{x^{(1)}, \dots, x^{(M)}\}$ in $\mathbb{R}^d$, an *ε-cover* ($\varepsilon > 0$) of $\mathcal{G}$ w.r.t. the L^1-empirical norm $\|g\|_1 := \frac{1}{M}\sum_{m=1}^M |g(x^{(m)})|$ is a finite collection of functions $g_1, \dots, g_n$ such that for any $g \in \mathcal{G}$, we can find a $j \in \{1, \cdots, n\}$ such that $\|g - g_j\|_1 \leqslant \varepsilon$. The smallest possible integer n is called the *ε-covering number* and is denoted by $\mathcal{N}_1(\varepsilon, \mathcal{G}, x^{1:M})$.

Assumption 7.3.3 (Covering the approximation space)**.** *There exist three constants*

$$\alpha_{(7.3.3)} \geqslant \frac{1}{4}, \qquad \beta_{(7.3.3)} > 0, \qquad \gamma_{(7.3.3)} \geqslant 1$$

such that for any $B > 0, \varepsilon \in (0, \frac{4}{15}B]$ and stratum index $1 \leqslant k \leqslant K$, the minimal size of an ε-covering number of $\mathcal{T}_B\mathcal{L}_k := \{\mathcal{T}_B\varphi : \varphi \in \mathcal{L}_k\}$ is bounded as follows:

$$\mathcal{N}_1(\varepsilon, \mathcal{T}_B\mathcal{L}_k, x^{1:M}) \leqslant \alpha_{(7.3.3)}\Big(\frac{\beta_{(7.3.3)}B}{\varepsilon}\Big)^{\gamma_{(7.3.3)}} \tag{7.3.3}$$

independently of the points sample $x^{1:M}$.

We assume that the above constants do not depend on k, mainly for the sake of simplicity. In the error analysis (see also Proposition 7.5.1), the constants $\alpha_{(7.3.3)}$ and $\beta_{(7.3.3)}$ appear in $\log$ and thus, they have a small impact on error bounds. On the contrary, $\gamma_{(7.3.3)}$ appears as a multiplicative factor and we seek to have the smallest estimate.

Proposition 7.3.3. *In the case of approximation spaces $\mathcal{L}_k$ like $\mathbf{LP}_0$, $\mathbf{LP}_1$ or $\mathbf{LP}_n$, Assumption 7.3.3 is satisfied with the following parameters: for any given $\eta > 0$, we have*

	$\alpha_{(7.3.3)}$	$\beta_{(7.3.3)}$	$\gamma_{(7.3.3)}$
$\mathbf{LP}_0$	1	$7/5$	1
$\mathbf{LP}_1$	3	$[4c_\eta 6^\eta]^{1/(1+\eta)}e$	$(d+2)(1+\eta)$
$\mathbf{LP}_n$	3	$[4c_\eta 6^\eta]^{1/(1+\eta)}e$	$((d+1)^n+1)(1+\eta)$

where $c_\eta = \sup_{x \geqslant \frac{45e}{2}} x^{-\eta}\log(x)$.

The proof is postponed to the Appendix.

7.3.2 Main result: Error estimate

We are now in the position to state a convergence result, expressed in terms of the quadratic error of the best approximation of y_i on the stratum $\mathcal{H}_k$:

$$T_{i,k} := \inf_{\varphi\in\mathcal{L}_k} |y_i - \varphi|^2_{\nu_k} \quad \text{where} \quad |\varphi|^2_{\nu_k} := \int_{\mathbb{R}^d} |\varphi|^2(x)\nu_k(\mathrm{d}x).$$

Our goal is to find an upper bound for the error $\mathbb{E}\left(|y_i^{(M)} - y_i|^2_\nu\right)$ where

$$|\varphi|^2_\nu := \int_{\mathbb{R}^d} |\varphi|^2(x)\nu(\mathrm{d}x).$$

Note that the above expectation is taken over all the random variables, including the random sources $U^{1:M}$, i.e., we estimate the quadratic error averaged on the root sample.

Theorem 7.3.4. *Assume Assumptions 7.2.2-7.2.3-7.3.2-7.3.3 and define $y_i^{(M)}$ as in Algorithm 4. Then, for any $\varepsilon > 0$, we have*

$$\begin{aligned}\mathbb{E}\left(|y_i^{(M)} - y_i|^2_\nu\right) \leqslant\ & 4(1+\varepsilon)L^2_{g_i}\underline{C}_{(7.3.1)}\mathbb{E}\left(|y_{i+1}^{(M)} - y_{i+1}|^2_\nu\right) + 2\sum_{k=1}^{K}\nu(\mathcal{H}_k)T_{i,k} \\ & + 4c_{(7.3.8)}(M)\frac{|y_i|^2_\infty}{M} + 2(1+\frac{1}{\varepsilon})\frac{\dim(\mathcal{L})}{M}(C_{g_i} + L_{g_i}|y_{i+1}|_\infty)^2 \\ & + 8(1+\varepsilon)L^2_{g_i}c_{(7.3.7)}(M)\frac{|y_{i+1}|^2_\infty}{M}.\end{aligned}$$

We emphasize that whenever useful, the constant $4(1+\varepsilon)L^2_{g_i}\underline{C}_{(7.3.1)}$ could be reduced to $(1+\delta)(1+\varepsilon)L^2_{g_i}\underline{C}_{(7.3.1)}$ (for any given $\delta > 0$) by

slightly adapting the proof: namely, the term $4 = 2^2$ comes from two applications of deviation inequalities stated in Proposition 7.5.1. These inequalities are valid with $(1+\delta)^{\frac{1}{2}}$ instead of 2, up to modifying the constants $c_{(7.5.2)}(M)$ and $c_{(7.5.3)}(M)$.

As a very significant difference with usual Regression Monte-Carlo methods (see [66, Theorem 4.11]), in our algorithm there is no competition between the bias term (approximation error) and the variance term (statistical error), while in usual algorithms as the dimension of the approximation space $K\dim(\mathcal{L})$ goes to infinity, the statistical term (of size $\frac{K\dim(\mathcal{L})}{M}$) blows up. This significant improvement comes from the stratification which gives rise to decoupled and low-dimensional regression problems.

Since $y_N^{(M)} = y_N$, we easily derive global error bounds.

Corollary 7.3.1. *Under the assumptions and notations of Theorem 7.3.4, there exists a constant $C_{(7.3.4)}(N)$ (depending only on N, $\sup_{0\leqslant i<N} L_{g_i}$, $\underline{C}_{(7.3.1)}$), such that for any $j \in \{0,\dots,N-1\}$,*

$$\mathbb{E}\left(|y_j^{(M)} - y_j|_\nu^2\right) \leqslant C_{(7.3.4)}(N)\sum_{i=j}^{N-1}\left[\sum_{k=1}^{K}\nu(H_k)T_{i,k}\right. \tag{7.3.4}$$
$$\left.\frac{1}{M}\left(c_{(7.3.8)}(M)|y_i|_\infty^2 + \dim(\mathcal{L})(C_{g_i} + L_{g_i}|y_{i+1}|_\infty)^2 + L_{g_i}^2 c_{(7.3.7)}(M)|y_{i+1}|_\infty^2\right)\right].$$

It is easy to see that if $4(1+\varepsilon)L_{g_i}^2\underline{C}_{(7.3.1)} \leqslant 1$, then interestingly $C_{(7.3.4)}(N)$ can be taken uniformly in N. This case corresponds to a small Lipschitz constant of g_i. In the case $4(1+\varepsilon)L_{g_i}^2\underline{C}_{(7.3.1)} \gg 1$, the above error estimates deteriorate quickly as N increases. We shall discuss that in Section 7.4 which deals with BSDEs and where we propose a different scheme that allows both large Lipschitz constant and large N.

7.3.3 Proof of Theorem 7.3.4

Let us start by setting up some useful notations:

$$S(x_{i:N}) := g_i(y_{i+1}(x_{i+1}), x_{i:N}), \qquad \psi_i^k := \mathrm{OLS}(S, \mathcal{L}_k, X_{i:N}^{i,k,1:M}),$$
$$|f|_{i,k,M}^2 := \frac{1}{M}\sum_{m=1}^{M} f^2(X_{i:N}^{i,k,m})$$

(or $|f|_{i,k,M}^2 := \frac{1}{M}\sum_{m=1}^{M} f^2(X_i^{i,k,m})$ when f depends only on one argument).

We first aim at deriving a bound on $\mathbb{E}\left(|y_i^{(M)} - y_i|_{i,k,M}^2\right)$. First of all, note that

$$|y_i^{(M)} - y_i|_{i,k,M}^2 = \left|\mathcal{T}_{|y_i|_\infty}(\psi_i^{(M),k}) - \mathcal{T}_{|y_i|_\infty}(y_i)\right|_{i,k,M}^2 \leqslant |\psi_i^{(M),k} - y_i|_{i,k,M}^2$$

since the truncation operator is 1-Lipschitz. Now we define

$$\mathbb{E}\left(S(X_{i:N}^{i,k,m})|X_i^{i,k,1:M}\right) = \mathbb{E}\left(S(X_{i:N}^{i,k,m})|X_i^{i,k,m}\right) = y_i(X_i^{i,k,m}) \tag{7.3.5}$$

where the first equality is due to the independence of the paths $(X_{i:N}^{i,k,m}, 1 \leqslant m \leqslant M)$ (Proposition 7.2.1) and where the last equality stems from the definition of y_i.

According to [66, Proposition 4.12] which allows to interchange conditional expectation and OLS, we have

$$\mathbb{E}\left(\psi_i^k(\cdot)|X_i^{i,k,1:M}\right) = \mathrm{OLS}(y_i, \mathcal{L}_k, X_{i:N}^{i,k,1:M}).$$

Since the expected values $\left(\mathbb{E}\left(\psi_i^k(X_i^{i,k,m})|X_i^{i,k,1:M}\right)\right)_{1\leqslant m\leqslant M}$ can be seen as the projections of $(y_i(X_i^{i,k,m}))_{1\leqslant m\leqslant M}$ on the subspace of $\mathbb{R}^M$ spanned by $\{(\varphi(X_i^{i,k,m}))_{1\leqslant m\leqslant M}, \varphi \in \mathcal{L}_k\}$ and $(\psi_i^{(M),k}(X_i^{i,k,m}))_{1\leqslant m\leqslant M}$ is an element in this subspace, Pythagoras theorem yields

$$\begin{aligned}
|\psi_i^{(M),k} - y_i|_{i,k,M}^2 &= \left|\psi_i^{(M),k} - \mathbb{E}\left(\psi_i^k(\cdot)|X_i^{i,k,1:M}\right)\right|_{i,k,M}^2 \\
&\quad + \left|\mathbb{E}\left(\psi_i^k(\cdot)|X_i^{i,k,1:M}\right) - y_i\right|_{i,k,M}^2 \\
&= \left|\psi_i^{(M),k} - \mathbb{E}\left(\psi_i^k(\cdot)|X_i^{i,k,1:M}\right)\right|_{i,k,M}^2 + \inf_{\varphi\in\mathcal{L}_k} |\varphi - y_i|_{i,k,M}^2.
\end{aligned}$$

For any given $\phi \in \mathcal{L}_k$, we have

$$\begin{aligned}
\mathbb{E}\left(\inf_{\varphi\in\mathcal{L}_k} |\varphi - y_i|_{i,k,M}^2\right) &\leqslant \mathbb{E}\left(|\phi - y_i|_{i,k,M}^2\right) \\
&= \mathbb{E}\left(\frac{1}{M}\sum_{m=1}^{M} |\phi(X_i^{i,k,m}) - y_i(X_i^{i,k,m})|^2\right) \\
&= \int_{\mathbb{R}^d} |\phi(x) - y_i(x)|^2 \nu_k(\mathrm{d}x).
\end{aligned}$$

Taking the infimum over all functions ϕ on the R.H.S. gives

$$\mathbb{E}\left(\inf_{\varphi\in\mathcal{L}_k} |\varphi - y_i|_{i,k,M}^2\right) \leqslant T_{i,k}.$$

So, for any $\varepsilon > 0$, we have

$$\begin{aligned}
\mathbb{E}\left(|\psi_i^{(M),k} - y_i|_{i,k,M}^2\right) &\leqslant T_{i,k} + (1+\varepsilon)\mathbb{E}\left(|\psi_i^{(M),k} - \psi_i^k|_{i,k,M}^2\right) \\
&\quad + (1+\frac{1}{\varepsilon})\mathbb{E}\left(\left|\psi_i^k - \mathbb{E}\left(\psi_i^k(\cdot)|X_i^{i,k,1:M}\right)\right|_{i,k,M}^2\right).
\end{aligned}$$

By [66, Proposition 4.12], the last term is bounded by $\frac{\dim(\mathcal{L})}{M}(C_{g_i} + L_{g_i}|y_{i+1}|_\infty)^2$

where $(C_{g_i}+L_{g_i}|y_{i+1}|_\infty)^2$ clearly bounds the conditional variance of $S(X_{i:N}^{i,k})$. This is the statistical error contribution. Here, we have used the independence of $(X_{i:N}^{i,k,m}, 1 \leqslant m \leqslant M)$ (Proposition 7.2.1).

The control of the term $\mathbb{E}\left(|\psi_i^{(M),k} - \psi_i^k|_{i,k,M}^2\right)$ is possible due to the linearity and stability of OLS [66, Proposition 4.12]:

$$|\psi_i^{(M),k} - \psi_i^k|_{i,k,M}^2 \leqslant |S^{(M)} - S|_{i,k,M}^2 \leqslant L_{g_i}^2 \frac{1}{M}\sum_{m=1}^{M}(y_{i+1}^{(M)} - y_{i+1})^2(X_{i+1}^{i,k,m}),$$

where we have taken advantage of the Lipschitz property of g_i w.r.t. the component y_{i+1}. So far we have shown

$$\mathbb{E}\left(|y_i^{(M)} - y_i|_{i,k,M}^2\right) \leqslant T_{i,k} + (1+\varepsilon)L_{g_i}^2\mathbb{E}\left(\frac{1}{M}\sum_{m=1}^{M}(y_{i+1}^{(M)} - y_{i+1})^2(X_{i+1}^{i,k,m})\right) + (1+\frac{1}{\varepsilon})\frac{\dim(\mathcal{L})}{M}(C_{g_i} + L_{g_i}|y_{i+1}|_\infty)^2. \tag{7.3.6}$$

This shows a relation between the errors at time i and time $i+1$, but measured in different norms. In order to retrieve the same $L^2(\nu)$-norm and continue the analysis, we will use the norm-stability property (Assumption 7.3.2) and the following result about concentration of measures. The proof is a particular case of Proposition 7.5.1 in the Appendix, with $\psi(x) = (-2|y_{i+1}|_\infty \vee x \wedge 2|y_{i+1}|_\infty)^2$, $B = |y_{i+1}|_\infty, \mathcal{K} = \mathcal{L}_k, \eta = y_{i+1}$.

Proposition 7.3.5. *Define* $(c_{(7.3.7)}(M), c_{(7.3.8)}(M))$ *by considering* $c_{(7.5.2)}(M)$ *and* $c_{(7.5.3)}(M)$ *from Proposition 7.5.1 with the values* $(\alpha_{(7.3.3)}, \beta_{(7.3.3)}, \gamma_{(7.3.3)})$ *instead of* (α, β, γ). *Then we have*

$$\mathbb{E}\left(\frac{1}{M}\sum_{m=1}^{M}(y_{i+1}^{(M)} - y_{i+1})^2(X_{i+1}^{i,k,m})\right) \leqslant 2\mathbb{E}\left(|y_{i+1}^{(M)}(X_{i+1}^{i,\nu_k}) - y_{i+1}(X_{i+1}^{i,\nu_k})|^2\right) + 4c_{(7.3.7)}(M)\frac{|y_{i+1}|_\infty^2}{M}, \tag{7.3.7}$$

$$\mathbb{E}\left(|y_i^{(M)} - y_i|_{\nu_k}^2\right) \leqslant 2\mathbb{E}\left(|y_i^{(M)} - y_i|_{i,k,M}^2\right) + 4c_{(7.3.8)}(M)\frac{|y_i|_\infty^2}{M}. \tag{7.3.8}$$

Multiply both sides of Equation (7.3.7) by $\nu(H_k)$, sum over k, and use the norm-stability property (Assumption 7.3.2): it readily follows that

$$\begin{aligned}\sum_{k=1}^{K}\nu(H_k)\mathbb{E}\left(\frac{1}{M}\sum_{m=1}^{M}(y_{i+1}^{(M)} - y_{i+1})^2(X_{i+1}^{i,k,m})\right) &\leqslant 2\mathbb{E}\left(|y_{i+1}^{(M)}(X_{i+1}^{i,\nu}) - y_{i+1}(X_{i+1}^{i,\nu})|^2\right) + 4c_{(7.3.7)}(M)\frac{|y_{i+1}|_\infty^2}{M} \\ &\leqslant 2\underline{C}_{(7.3.1)}\mathbb{E}\left(|y_{i+1}^{(M)} - y_{i+1}|_\nu^2\right) + 4c_{(7.3.7)}(M)\frac{|y_{i+1}|_\infty^2}{M}.\end{aligned}$$

Similarly, we can get from Equation (7.3.8) that

$$\mathbb{E}\left(|y_i^{(M)} - y_i|_\nu^2\right) \leqslant 2\sum_{k=1}^{K} \nu(H_k)\mathbb{E}\left(|y_i^{(M)} - y_i|_{i,k,M}^2\right) + 4c_{(7.3.8)}(M)\frac{|y_i|_\infty^2}{M}.$$

Finally, by combining the above estimates with (7.3.6), we get

$$\mathbb{E}\left(|y_i^{(M)} - y_i|_\nu^2\right) \leqslant 4c_{(7.3.8)}(M)\frac{|y_i|_\infty^2}{M} + 2\sum_{k=1}^{K} \nu(H_k)T_{i,k} + 2(1+\frac{1}{\varepsilon})\frac{\dim(\mathcal{L})}{M}(C_{g_i}$$
$$+ L_{g_i}|y_{i+1}|_\infty)^2 + 2(1+\varepsilon)L_{g_i}^2\left(2\underline{C}_{(7.3.1)}\mathbb{E}\left(|y_{i+1}^{(M)} - y_{i+1}|_\nu^2\right) + 4c_{(7.3.7)}(M)\frac{|y_{i+1}|_\infty^2}{M}\right).$$

This links $\mathbb{E}\left(|y_i^{(M)} - y_i|_\nu^2\right)$ with $\mathbb{E}\left(|y_{i+1}^{(M)} - y_{i+1}|_\nu^2\right)$ as announced. □

7.4 Convergence analysis for the solution of BSDEs with the MDP representation

Let us consider the semi-linear final value problem for a parabolic PDE of the form

$$\begin{cases} \partial_t u(t,x) + \frac{1}{2}\Delta u(t,x) + f(t,u(t,x),x) = 0, & t < 1, x \in \mathbb{R}^d, \\ u(1,x) = g(x). \end{cases}$$

This is a simple form of the Hamilton-Jacobi-Bellman equation of stochastic control problems [97]. Under fairly mild assumptions (see [106] for instance), the solution to the above PDE is related to a Backward Stochastic Differential Equation $(\mathcal{Y}, \mathcal{Z})$ driven by a d-dimensional Brownian motion W. Namely,

$$\mathcal{Y}_t = g(W_1) + \int_t^1 f(s, \mathcal{Y}_s, W_s)\mathrm{d}s - \int_t^1 \mathcal{Z}_s \mathrm{d}W_s$$

and $\mathcal{Y}_t = u(t, W_t)$, $\mathcal{Z}_t = \nabla u(t, W_t)$. Needless to say, the Laplacian Δ and the process W could be replaced by a more general second order operator and its related diffusion process, and that f could depend on the gradient Z as well. We stick to the above setting which is consistent with this work. Taking conditional expectation reduces to

$$\mathcal{Y}_t = \mathbb{E}\left(g(W_1) + \int_t^1 f(s, \mathcal{Y}_s, W_s)\mathrm{d}s \mid W_t\right).$$

There are several time discretization schemes of $\mathcal{Y}$ (explicit or implicit Euler schemes, high order schemes [34]) but here we follow the Multi-Step Forward Dynamic Programming (MDP for short) Equation of [66],

which allows a better error propagation compared to the One-Step Dynamic Programming Equation:

$$Y_i = \mathbb{E}\left(g_N(X_N) + \frac{1}{N}\sum_{j=i+1}^{N} f_j(Y_j, X_j, \dots, X_N)|X_i \right) = y_i(X_i), \quad 0 \leqslant i < N.$$

Here, we consider a more general path-dependency on f_j, actually this does not affect the error analysis. In comparison with Algorithm 4, we take

$$g_i(y_{i+1:N}, x_{i:N}) = g_N(x_N) + \frac{1}{N}\sum_{j=i+1}^{N} f_j(y_j, x_{j:N}).$$

In [10] similar discrete BSDEs appear but with an external noise. That corresponds to time-discretization of Backward Doubly SDEs, which in turn are related to stochastic semi-linear PDEs.

7.4.1 Standing assumptions

We shall now describe the main assumptions that are needed in the methodology proposed in this paper.

Assumptions on f_i and g_N

Assumption 7.4.1 (Functions f_i and g_N)**.** *Each f_i is Lipschitz w.r.t. y_i, with Lipschitz constant L_{f_i} and $C_{f_i} = \sup_{x_{i:N}} |f_i(0, x_{i:N})| < +\infty$. Moreover g_N is bounded.*

The reader can easily check that y_i is bounded.

Assumptions on the distribution ν

Assumption 7.4.2 (norm-stability)**.** *There exists a constant $\underline{C}_{(7.4.1)} \geqslant 1$ such that for any $\varphi \in L^2(\nu)$ and any $0 \leqslant i < j \leqslant N$, we have*

$$\int_{\mathbb{R}^d} \mathbb{E}\left(\varphi^2(X_j^{i,x})\right) \nu(\mathrm{d}x) \leqslant \underline{C}_{(7.4.1)} \int_{\mathbb{R}^d} |\varphi(x)|^2 \nu(\mathrm{d}x). \tag{7.4.1}$$

It is straightforward to extend Propositions 7.3.1 and 7.3.2 to fulfill the above assumption.

7.4.2 Main result: error estimate

We express the error in terms of the best local approximation error and the averaged one:

$$T_{i,k} := \inf_{\varphi \in \mathcal{L}_k} |y_i - \varphi|^2_{\nu_k}, \qquad \nu(T_{i,\cdot}) := \sum_{k=1}^{K} \nu(\mathcal{H}_k) T_{i,k}.$$

In this discrete time BSDE context, Theorem 7.3.4 becomes the following.

Theorem 7.4.1. *Assume Assumptions 7.2.2-7.2.3-7.3.3-7.4.2 and define $y_i^{(M)}$ as in Algorithm 4. Set*

$$\bar{\mathcal{E}}(Y,M,i) := \mathbb{E}\left(|y_i^{(M)} - y_i|_\nu^2\right) = \sum_{k=1}^{K} \nu(\mathcal{H}_k)\mathbb{E}\left(|y_i^{(M)} - y_i|_{\nu_k}^2\right).$$

Define

$$\delta_i = 4c_{(7.3.8)}(M)\frac{|y_i|_\infty^2}{M} + 2\nu(T_{i,.}) + 16\frac{1}{N}\sum_{j=i+1}^{N-1} L_{f_j}^2 c_{(7.3.7)}(M)\frac{|y_j|_\infty^2}{M} + \\ + 4\frac{\dim(\mathcal{L})}{M}\left(|y_N|_\infty + \frac{1}{N}\sum_{j=i+1}^{N}(C_{f_j} + L_{f_j}|y_j|_\infty)\right)^2.$$

Then, letting $L_f := \sup_j L_{f_j}$, we have

$$\bar{\mathcal{E}}(Y,M,i) \leqslant \delta_i + 8\underline{C}_{(7.4.1)}L_f^2 \exp\left(8\underline{C}_{(7.4.1)}L_f^2\right)\frac{1}{N}\sum_{j=i+1}^{N-1}\delta_j.$$

The above general error estimates become simpler when the parameters are uniform in i.

Corollary 7.4.1. *Under the assumptions of Theorem 7.4.1 and assuming that C_{f_i}, L_{f_i} and $|y_i|_\infty$ are bounded uniformly in i and N, there exists a constant $C_{(7.4.2)}$ (independent of N and of approximation spaces $\mathcal{L}_k$) such that*

$$\bar{\mathcal{E}}(Y,M,i) \leqslant C_{(7.4.2)}\left(\frac{c_{(7.3.8)}(M) + c_{(7.3.7)}(M) + \dim(\mathcal{L})}{M} + \nu(T_{i,.}) + \frac{1}{N}\sum_{j=i+1}^{N-1}\nu(T_{j,.})\right). \tag{7.4.2}$$

We observe that this upper bound is expressed in a quite convenient form to let $N \to +\infty$ and $K \to +\infty$. As a major difference with the usual Regression Monte Carlo schemes, the impact of the statistical error (through the parameter M) is not affected by the number K of strata.

7.4.3 Proof of Theorem 7.4.1

We follow the arguments of the proof of Theorem 7.3.4 with the following notation:

$$
\begin{aligned}
S(x_{i:N}) &:= g_N(x_N) + \frac{1}{N}\sum_{j=i+1}^{N} f_j(y_j(x_j), x_{j:N}),\\
S^{(M)}(x_{i:N}) &:= g_N(x_N) + \frac{1}{N}\sum_{j=i+1}^{N} f_j(y_j^{(M)}(x_j), x_{j:N}),\\
\psi_i^k &:= \mathrm{OLS}(S, \mathcal{L}_k, X_{i:N}^{i,k,1:M}), \qquad \psi_i^{(M),k} := \mathrm{OLS}(S^{(M)}, \mathcal{L}_k, X_{i:N}^{i,k,1:M}).
\end{aligned}
$$

The beginning of the proof is similar and we obtain (here, there is no need to optimize ε and we take $\varepsilon = 1$)

$$
\begin{aligned}
\mathbb{E}\left(|y_i^{(M)} - y_i|_{i,k,M}^2\right) &\leqslant \mathbb{E}\left(|\psi_i^{(M),k} - y_i|_{i,k,M}^2\right)\\
&\leqslant T_{i,k} + 2\mathbb{E}\left(|\psi_i^{(M),k} - \psi_i^k|_{i,k,M}^2\right) + 2\mathbb{E}\left(\left|\psi_i^k - \mathbb{E}\left(\psi_i^k(\cdot)|X_i^{i,k,1:M}\right)\right|_{i,k,M}^2\right).
\end{aligned}
$$

The last term is a statistical error term, which can be controlled as follows:

$$
\mathbb{E}\left(\left|\psi_i^k - \mathbb{E}\left(\psi_i^k(\cdot)|X_i^{i,k,1:M}\right)\right|_{i,k,M}^2\right) \leqslant \frac{\dim(\mathcal{L})}{M}\left(|y_N|_\infty + \frac{1}{N}\sum_{j=i+1}^{N}(C_{f_j} + L_{f_j}|y_j|_\infty)\right)^2
$$

where $(\dots)^2$ is a rough bound of the conditional variance of $S(X_{i:N}^{i,k})$.

We handle the control of the term $\mathbb{E}\left(|\psi_i^{(M),k} - \psi_i^k|_{i,k,M}^2\right)$ as in Theorem 7.3.4 but the results are different because the dynamic programming equation differs:

$$
\begin{aligned}
\mathbb{E}\left(|\psi_i^{(M),k} - \psi_i^k|_{i,k,M}^2\right) &\leqslant \mathbb{E}\left(|S^{(M)} - S|_{i,k,M}^2\right)\\
&\leqslant \mathbb{E}\left(\frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{N}\sum_{j=i+1}^{N-1} L_{f_j}|y_j^{(M)} - y_j|(X_j^{i,k,m})\right)^2\right)\\
&\leqslant \mathbb{E}\left(\frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{j=i+1}^{N-1} L_{f_j}^2|y_j^{(M)} - y_j|^2(X_j^{i,k,m})\right).
\end{aligned}
$$

We multiply the above by $\nu(H_k)$, sum over k, apply the *extended* Proposition 7.3.5 valid also for the problem at hand, and the Assumption 7.4.2. Then, it follows that

$$
\sum_{k=1}^{K}\nu(H_k)\mathbb{E}\left(|\psi_i^{(M),k} - \psi_i^k|_{i,k,M}^2\right) \leqslant 2\frac{1}{N}\sum_{j=i+1}^{N-1} L_{f_j}^2\left(\underline{C}_{(7.4.1)}\mathbb{E}\left(|y_j^{(M)} - y_j|_\nu^2\right) + 2c_{(7.3.7)}(M)\frac{|y_j|_\infty^2}{M}\right).
$$

On the other hand, from Equation (7.3.8) we have

$$\mathbb{E}\left(|y_i^{(M)} - y_i|_\nu^2\right) \leqslant 2\sum_{k=1}^{K} \nu(H_k)\mathbb{E}\left(|y_i^{(M)} - y_i|_{i,k,M}^2\right) + 4c_{(7.3.8)}(M)\frac{|y_i|_\infty^2}{M}.$$

Now collect the different estimates: it writes

$$\begin{aligned}\bar{\mathcal{E}}(Y,M,i) := \mathbb{E}\left(|y_i^{(M)} - y_i|_\nu^2\right) &\leqslant 4c_{(7.3.8)}(M)\frac{|y_i|_\infty^2}{M} + 2\sum_{k=1}^{K}\nu(H_k)T_{i,k} \\ &+ 8\frac{1}{N}\sum_{j=i+1}^{N-1} L_{f_j}^2\left(\underline{C}_{(7.4.1)}\mathbb{E}\left(|y_j^{(M)} - y_j|_\nu^2\right) + 2c_{(7.3.7)}(M)\frac{|y_j|_\infty^2}{M}\right) + \\ &+ 4\frac{\dim(\mathcal{L})}{M}\left(|y_N|_\infty + \frac{1}{N}\sum_{j=i+1}^{N}(C_{f_j} + L_{f_j}|y_j|_\infty)\right)^2 \\ &:= \delta_i + 8\underline{C}_{(7.4.1)}\frac{1}{N}\sum_{j=i+1}^{N-1} L_{f_j}^2\bar{\mathcal{E}}(Y,M,j).\end{aligned}$$

It takes the form of a discrete Gronwall lemma, which easily allows to derive the following upper bound (see [10, Appendix A.3]):

$$\bar{\mathcal{E}}(Y,M,i) \leqslant \delta_i + 8\underline{C}_{(7.4.1)}\frac{1}{N}\sum_{j=i+1}^{N-1}\Gamma_{i,j}L_{f_j}^2\delta_j,$$

$$\text{where } \Gamma_{i,j} := \begin{cases}\prod_{i<k<j}(1 + 8\underline{C}_{(7.4.1)}\frac{1}{N}L_{f_k}^2), & \text{for } i+1<j,\\ 1, & \text{otherwise.}\end{cases}$$

Using now $L_f = \sup_j L_{f_j}$, we get $\Gamma_{i,j} \leqslant \exp\left(\sum_{i<k<j} 8\underline{C}_{(7.4.1)}\frac{1}{N}L_{f_k}^2\right) \leqslant \exp(8\underline{C}_{(7.4.1)}L_f^2)$. This completes the proof. □

7.5 Appendix

7.5.1 Proof of Proposition 7.3.3

Consider first the case of the partitioning estimate ($\mathbf{LP}_0$) and let $\varepsilon \in (0, \frac{4}{15}B]$. We use an ε-cover in the L^∞-norm, which simply reduces to cover $[-B, B]$ with intervals of size 2ϵ. A solution is to take the interval center defined by $h_j = -B + \varepsilon + 2\varepsilon j$, $0 \leqslant j \leqslant n$, where n is the smallest integer such that $h_n \geqslant B$ (i.e., $n = \lceil \frac{B}{\varepsilon} - \frac{1}{2}\rceil$). Thus, we obtain

$$\mathcal{N}_1(\varepsilon, \mathcal{T}_B\mathcal{L}_k, x^{1:M}) \leqslant n + 1 \leqslant \frac{B}{\varepsilon} + \frac{3}{2} \leqslant \frac{7}{5}\frac{B}{\varepsilon}$$

where we use the constraint on ε.

In the case of general vector space of dimension K, from [75, Lemma 9.2, Theorem 9.4 and Theorem 9.5], we obtain

$$\mathcal{N}_1(\varepsilon, \mathcal{T}_B\mathcal{K}, x^{1:M}) \leq 3\Big(\frac{4eB}{\varepsilon}\log\big(\frac{6eB}{\varepsilon}\big)\Big)^{K+1}$$

whenever $\varepsilon < B/2$. For ε as in the statement of Assumption 7.3.3, we have $\frac{6eB}{\varepsilon} \geqslant \frac{45e}{2}$. Let $\eta > 0$, since $\log(x) \leqslant c_\eta x^\eta$ for any $x \geqslant \frac{45e}{2}$ with $c_\eta = \sup_{x \geqslant \frac{45e}{2}} \frac{\log(x)}{x^\eta}$, we get

$$\mathcal{N}_1(\varepsilon, \mathcal{T}_B\mathcal{K}, x^{1:M}) \leqslant 3\left([4c_\eta 6^\eta]^{1/(1+\eta)}\frac{eB}{\varepsilon}\right)^{(K+1)(1+\eta)}.$$

For $\mathbf{LP}_1$ and $\mathbf{LP}_n$, we have respectively $K = d+1$ and $K = (d+1)^n$, therefore the announced result. Whenever useful, the choice $\eta = 1$ gives $\beta_{(7.3.3)} \leqslant 3.5$.

For the partitioning estimate (case $\mathbf{LP}_0$), we could also use this estimate with $K = 1$ but with the first arguments, we get better parameters (especially for γ). □

7.5.2 Probability of uniform deviation

Lemma 7.5.1 ([66, Lemma B.2]). *Let $\mathcal{G}$ be a countable set of functions $g : \mathbb{R}^d \mapsto [0, B]$ with $B > 0$. Let $\mathcal{X}, \mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(M)}$ ($M \geqslant 1$) be i.i.d. $\mathbb{R}^d$ valued random variables. For any $\alpha > 0$ and $\varepsilon \in (0, 1)$ one has*

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{\frac{1}{M}\sum_{m=1}^M g(\mathcal{X}^{(m)}) - \mathbb{E}\left(g(\mathcal{X})\right)}{\alpha + \frac{1}{M}\sum_{m=1}^M g(\mathcal{X}^{(m)}) + \mathbb{E}\left(g(\mathcal{X})\right)} > \varepsilon\right)$$
$$\leqslant 4\mathbb{E}\left(\mathcal{N}_1\big(\frac{\alpha\varepsilon}{5}, \mathcal{G}, \mathcal{X}^{1:M}\big)\right)\exp\big(-\frac{3\varepsilon^2\alpha M}{40B}\big),$$
$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{\mathbb{E}\left(g(\mathcal{X})\right) - \frac{1}{M}\sum_{m=1}^M g(\mathcal{X}^{(m)})}{\alpha + \frac{1}{M}\sum_{m=1}^M g(\mathcal{X}^{(m)}) + \mathbb{E}\left(g(\mathcal{X})\right)} > \varepsilon\right)$$
$$\leqslant 4\mathbb{E}\left(\mathcal{N}_1\big(\frac{\alpha\varepsilon}{8}, \mathcal{G}, \mathcal{X}^{1:M}\big)\right)\exp\big(-\frac{6\varepsilon^2\alpha M}{169B}\big).$$

7.5.3 Expected uniform deviation

Proposition 7.5.1. *For finite $B > 0$, let $\mathcal{G} := \{\psi\big(\mathcal{T}_B\phi(\cdot) - \eta(\cdot)\big) : \phi \in \mathcal{K}\}$, where $\psi : \mathbb{R} \to [0, \infty)$ is Lipschitz continuous with $\psi(0) = 0$ and Lipschitz constant L_ψ, $\eta : \mathbb{R}^d \to [-B, B]$, and $\mathcal{K}$ is a finite K-dimensional vector space of functions with*

$$\mathcal{N}_1(\varepsilon, \mathcal{T}_B\mathcal{K}, \mathcal{X}^{1:M}) \leqslant \alpha\Big(\frac{\beta B}{\varepsilon}\Big)^\gamma \quad \textit{for } \varepsilon \in (0, \frac{4}{15}B] \tag{7.5.1}$$

for some positive constants α, β, γ with $\alpha \geqslant 1/4$ and $\gamma \geqslant 1$. Then, for $\mathcal{X}^{(1)}, \dots, \mathcal{X}^{(M)}$ i.i.d. copies of $\mathcal{X}$, we have

$$\mathbb{E}\left(\sup_{g\in\mathcal{G}}\Big(\frac{1}{M}\sum_{m=1}^{M} g(\mathcal{X}^{(m)}) - 2\int_{\mathbb{R}^d} g(x)\mathbb{P}\circ\mathcal{X}^{-1}(\mathrm{d}x)\Big)_+\right) \leqslant c_{(7.5.2)}(M)\frac{BL_\Psi}{M}$$
$$\text{with } c_{(7.5.2)}(M) := 120\left(1+\log(4\alpha)+\gamma\log\left((1+\frac{\beta}{16})M\right)\right), \tag{7.5.2}$$

$$\mathbb{E}\left(\sup_{g\in\mathcal{G}}\Big(\int_{\mathbb{R}^d} g(x)\mathbb{P}\circ\mathcal{X}^{-1}(\mathrm{d}x) - \frac{2}{M}\sum_{m=1}^{M} g(\mathcal{X}^{(m)})\Big)_+\right) \leqslant c_{(7.5.3)}(M)\frac{BL_\Psi}{M}$$
$$\text{with } c_{(7.5.3)}(M) := \frac{507}{2}\left(1+\log(4\alpha)+\gamma\log\left((1+\frac{8\beta}{169})M\right)\right). \tag{7.5.3}$$

Proof. The idea is to adapt the arguments of [66, Proposition 4.9].

▷ We first show (7.5.2). Set $\mathcal{Z} := \sup_{g\in\mathcal{G}}\Big(\frac{1}{M}\sum_{m=1}^{M} g(\mathcal{X}^{(m)}) - 2\int_{\mathbb{R}^d} g(x)\mathbb{P}\circ\mathcal{X}^{-1}(\mathrm{d}x)\Big)_+$. Let us find an upper bound for $\mathbb{P}(\mathcal{Z}>\varepsilon)$ in order to bound $\mathbb{E}(\mathcal{Z}) = \int_0^\infty \mathbb{P}(\mathcal{Z}>\varepsilon)\,\mathrm{d}\varepsilon$. Using the equality

$$\mathbb{P}(\mathcal{Z}>\varepsilon) = \mathbb{P}\left(\exists g\in\mathcal{G} : \frac{\frac{1}{M}\sum_{m=1}^{M} g(\mathcal{X}^{(m)}) - \int_{\mathbb{R}^d} g(x)\mathbb{P}\circ\mathcal{X}^{-1}(\mathrm{d}x)}{2\varepsilon + \int_{\mathbb{R}^d} g(x)\mathbb{P}\circ\mathcal{X}^{-1}(\mathrm{d}x) + \frac{1}{M}\sum_{m=1}^{M} g(\mathcal{X}^{(m)})} > \frac{1}{3}\right),$$

and that the elements of $\mathcal{G}$ take values in $[0, 2BL_\psi]$, it follows from Lemma 7.5.1 that

$$\mathbb{P}(\mathcal{Z}>\varepsilon) \leq 4\mathbb{E}\left(\mathcal{N}_1(\frac{2\varepsilon}{15}, \mathcal{G}, \mathcal{X}^{1:M})\right)\exp\Big(-\frac{\varepsilon M}{120BL_\psi}\Big).$$

Define $\mathcal{T}_B\mathcal{K}$ as in Proposition 7.5.1. Since $|\psi\big(\phi_1(x)-\eta(x)\big) - \psi\big(\phi_2(x)-\eta(x)\big)| \leq L_\psi|\phi_1(x)-\phi_2(x)|$ for all $x\in\mathbb{R}^d$ and all (ϕ_1,ϕ_2), it follows that

$$\mathcal{N}_1(\frac{2\varepsilon}{15}, \mathcal{G}, \mathcal{X}^{1:M}) \leq \mathcal{N}_1(\frac{2\varepsilon}{15L_\psi}, \mathcal{T}_B\mathcal{K}, \mathcal{X}^{1:M}).$$

Due to Equation (7.5.1), we deduce that

$$\mathbb{P}(\mathcal{Z}>\varepsilon) \leq 4\alpha\Big(\frac{15\beta BL_\psi}{2\varepsilon}\Big)^\gamma \exp\Big(-\frac{\varepsilon M}{120BL_\psi}\Big) \tag{7.5.4}$$

whenever $\frac{2\varepsilon}{15L_\psi} \leqslant \frac{4}{15}B$, i.e., $\varepsilon \leq 2BL_\Psi$. On the other hand, $\mathbb{P}(\mathcal{Z}>\varepsilon) = 0$ for all $\varepsilon > 2BL_\Psi$. Setting $a = \frac{15\beta BL_\psi}{2}$, $b = \frac{1}{120BL_\psi}$, it follows from (7.5.4) that

$$\mathbb{P}(\mathcal{Z}>\varepsilon) \leq 4\alpha\Big(\frac{a}{\varepsilon}\Big)^\gamma \exp(-bM\varepsilon), \qquad \forall\,\varepsilon>0.$$

Fix ε_0 to be some finite value (to be determined later) such that

$$\varepsilon_0 \geq \frac{a}{M(1+ab)}. \tag{7.5.5}$$

It readily follows that

$$\begin{aligned}\mathbb{E}\left(\mathcal{Z}\right) = \int_0^\infty \mathbb{P}\left(\mathcal{Z} > \varepsilon\right) \mathrm{d}\varepsilon &\leq \varepsilon_0 + \int_{\varepsilon_0}^\infty 4\alpha\Big(\frac{a}{\varepsilon}\Big)^\gamma \exp(-bM\varepsilon)\mathrm{d}\varepsilon \\ &\leq \varepsilon_0 + \frac{4\alpha}{bM}\big(M(1+ab)\big)^\gamma \exp(-bM\varepsilon_0).\end{aligned}$$

We choose $\varepsilon_0 = \frac{1}{bM}\log\Big(4\alpha\big((1+ab)M\big)^\gamma\Big)$: It satisfies (7.5.5) since

$$\frac{1}{bM}\log\Big(4\alpha\big((1+ab)M\big)^\gamma\Big) \geqslant \frac{a}{M}\frac{\log(1+ab)}{ab} \geqslant \frac{a}{M}\frac{1}{1+ab}$$

(use $\alpha \geqslant 1/4$, $\gamma \geqslant 1$, $M \geqslant 1$ and $\log(1+x) \geq x/(1+x)$ for all $x \geqslant 0$). Moreover, this choice of ε_0 implies that

$$\begin{aligned}\mathbb{E}[\mathcal{Z}] &\leq \frac{1}{bM}\Big(1 + \log(4\alpha) + \gamma\log\big((1+ab)M\big)\Big) \\ &= \frac{120BL_\psi}{M}\left(1 + \log(4\alpha) + \gamma\log\left((1+\frac{\beta}{16})M\right)\right).\end{aligned} \tag{7.5.6}$$

The inequality (7.5.2) is proved.

▷ We now justify (7.5.3) by similar arguments. Set $\mathcal{Z} := \sup_{g\in\mathcal{G}}\Big(\int_{\mathbb{R}^d} g(x)\mathbb{P}\circ\mathcal{X}^{-1}(\mathrm{d}x) - \frac{2}{M}\sum_{m=1}^M g(\mathcal{X}^{(m)})\Big)_+$. From Lemma 7.5.1, we get

$$\mathbb{P}\left(\mathcal{Z} > \varepsilon\right) \leq 4\mathbb{E}\left(\mathcal{N}_1(\frac{\varepsilon}{12}, \mathcal{G}, \mathcal{X}^{1:M})\right)\exp\Big(-\frac{2\varepsilon M}{507BL_\psi}\Big).$$

Since $\mathcal{N}_1(\frac{\varepsilon}{12}, \mathcal{G}, \mathcal{X}^{1:M}) \leq \mathcal{N}_1(\frac{\varepsilon}{12L_\psi}, \mathcal{T}_B\mathcal{K}, \mathcal{X}^{1:M})$ and thanks to (7.5.1), we derive

$$\mathbb{P}\left(\mathcal{Z} > \varepsilon\right) \leq 4\alpha\Big(\frac{12\beta BL_\psi}{\varepsilon}\Big)^\gamma \exp\Big(-\frac{2\varepsilon M}{507BL_\psi}\Big) \tag{7.5.7}$$

whenever $\frac{\varepsilon}{12L_\psi} \leqslant \frac{4}{15}B$. For other values of ε the above probability is zero, therefore (7.5.7) holds for any $\varepsilon > 0$. The end of the computations is now very similar to the previous case: we finally get the inequality (7.5.6) for the new $\mathcal{Z}$ with adjusted values $a = 12\beta BL_\psi$, $b = \frac{2}{507BL_\psi}$. Thus inequality (7.5.3) is thus proved. □

Chapter 8

Numerical Tests

We shall now illustrate the methodology presented in the previous chapter in two numerical examples coming from practical problems. The first one concerns a reaction-diffusion PDE connected to spatially distributed populations, whereas the second one deals with a stochastic control problem.

8.1 An Application to Reaction-Diffusion Models in Spatially Distributed Populations

In this section we consider a biologically motivated example to illustrate the strength of the stratified resampling regression methodology presented in the previous sections. We selected an application to spatially distributed populations that evolve under reaction diffusion equations. Besides the theoretical challenges behind the models, it has recently attracted attention due to its impact in the spread of infectious diseases [100, 101] and even to the modeling of Wolbachia infected mosquitoes in the fight of disease spreading *Aedes aegypti* [12, 80].

The use of reaction diffusion models to describe the population dynamics of a single species or genetic trait expanding into new territory dominated by another one goes back to the work of R. A. Fisher [53] and A. Kolmogorov et al. [89]. The mathematical model behind it is known as the (celebrated) Fisher-Kolmogorov-Petrovski-Piscounov (FKPP) equation.

In a conveniently chosen scale it takes the form, in dimension 1,

$$\partial_t u + \partial_x^2 u + au(1-u) = 0\,, \quad u(T,x) = h(x), x \in \mathbb{R}, t \leq T\,, \tag{8.1.1}$$

where $u = u(t,x)$ refers to the proportion of members of an invading species in a spatially distributed population on a straight line. The equation is chosen with time running backwards and as a final value problem to allow direct connection with the standard probabilistic interpretation.

It is well known [1] that for any arbitrary positive C, if we define

$$h(x) := \left(1 + C\exp\left(\pm\frac{\sqrt{6a}}{6}x\right)\right)^{-2} \tag{8.1.2}$$

then

$$u(t,x) = \left(1 + C\exp\left(\frac{5a}{6}(t-T) \pm \frac{\sqrt{6a}}{6}x\right)\right)^{-2}$$

is a traveling wave solution to Equation (8.1.1). We fix $a = C = 1$ in this example. The behavior of $h(x)$ as $x \to \pm\infty$ is either one or zero according to the sign chosen inside the exponential. Thus describing full dominance of the invading species or its absence.

The probabilistic formulation goes as follows: Introduce the system, as in Section 7.4,

$$\begin{aligned} dP_s &= \sqrt{2}dW_s \\ dY_s &= -f(Y_s)ds + Z_s dW_s \\ Y_T &= u(T,P_T) = h(P_T)\,. \end{aligned} \tag{8.1.3}$$

where $f(x) = ax(1-x)$ and $Z_s = \sqrt{2}\partial_x u(s,P_s)$.

Then, the process $Y_t = \mathbb{E}\left[Y_T + \int_t^T f(Y_s)ds|P_t\right]$ satisfies $Y_t = u(t,P_t)$.

To test the algorithms presented herein, we shall start with the following more general parabolic PDE

$$\partial_t W + \sum_{1\le i,j\le d} A_{ij}\partial_{y_i}\partial_{y_j}W + aW(1-W) = 0\,,\ t \le T, \text{ and } y \in \mathbb{R}^d\,. \tag{8.1.4}$$

Here, the matrix A is chosen as an arbitrary *positive-definite* constant $d\times d$ matrix. Furthermore, we choose, for convenience, the final condition

$$W(T,y) = h(y'\boldsymbol{\Sigma}^{-1}\theta)\,, \tag{8.1.5}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}' = \sqrt{A}$ and θ is arbitrary unit vector. We stress that this special choice of the final condition has the sole purpose of bypassing the need of solving Equation (8.1.5) by other numerical methods for comparison with the present methodology. Indeed, the fact that we are able to exhibit an explicit solution to Equation (8.1.4) with final condition (8.1.5) allows an easy checking of the accuracy of the method. We also stress that the method developed in this work does not require an explicit knowledge of the diffusion coefficient matrix A of Equation (8.1.4) since we shall make use of the observed paths. Yet the knowledge of the function $W \mapsto aW(1-W)$ is crucial.

It is easy to see that if $u = u(t,x)$ satisfies Equation (8.1.1) with final condition given by Equation (8.1.2) then

$$W(t,y) := u(t,y'\boldsymbol{\Sigma}^{-1}\theta) \tag{8.1.6}$$

satisfies Equation (8.1.4) with final condition (8.1.5).

An interpretation of the methodology proposed here is the following: If we were able to observe the trajectories performed by a small number of free individuals according to the diffusion process associated to Equation (8.1.4), even if we did not know the explicit form of the diffusion (i.e., we did not have a good calibration of the covariance matrix)

we could use such trajectories to produce a reliable solution to the final value problem (8.1.4) and (8.1.5).

We firstly present some numerical results in dimension 1 (Tables 8.1-8.2-8.3). We have tested both the one-step (Section 7.3) and multi-step schemes (Section 7.4). The final time T is fixed to 1 and we use time discretization $t_i = \frac{i}{N}T, 0 \leqslant i \leqslant N$ with $N = 10$ or 20. We divide the real line $\mathbb{R}$ into K subintervals $(I_i)_{1\leqslant i\leqslant K}$ by fixing $A = 25$ and dividing $[-A, A]$ into $K - 2$ equal length intervals and then adding $(-\infty, -A)$ and $(A, +\infty)$. We implement our method by using piecewise constant estimation on each interval. Then finally we get a piecewise constant estimation of $u(0, y)$, noted as $\hat{u}(0, y)$. Then we approximate the squared $L^2(\nu)$ error of our estimation by

$$\sum_{1\leqslant k\leqslant K} |u(0, y_k) - \hat{u}(0, y_k)|^2 \nu(I_k)$$

where y_k is chosen as the middle point of the rectangle if I_k is finite and the boundary point if I_k is infinite. We take $\nu(\mathrm{d}x) = \frac{1}{2}e^{-|x|}\mathrm{d}x$ and we use the restriction of ν on I_k to sample initial points. The squared $L^2(\nu)$ norm of $u(0, \cdot)$ is around 0.25. Finally remark that the error of our method includes three parts: time discretization error, approximation error due to the use of piecewise constant estimation on hypercubes and statistical error due to the randomness of trajectories. In the following tables, M is the number of trajectories that we use (i.e., the root sample).

We observe in Tables 8.1 and 8.3 that the approximation error (visible for small K) contributes much more to the global error for the one-step scheme, compared to the multi-step one. This observation is in agreement with those of [15, 66].

When N gets larger with fixed K and M (Tables 8.1 and 8.2), we may observe an increase of the global error for the one-step scheme, this is coherent with the estimates of Corollary 7.3.1.

	$K = 10$	$K = 20$	$K = 50$	$K = 100$	$K = 200$	$K = 400$
$M = 20$	0.0993	0.0253	0.0038	0.0014	0.0014	0.0019
$M = 40$	0.0997	0.0252	0.0034	9.01e-04	5.16e-04	6.17e-04
$M = 80$	0.0993	0.0249	0.0029	6.15e-04	3.92e-04	3.91e-04
$M = 160$	0.0990	0.0248	0.0029	3.15e-04	1.57e-04	1.71e-04
$M = 320$	0.0990	0.0248	0.0028	2.47e-04	1.02e-04	1.19e-04
$M = 640$	0.0990	0.0246	0.0028	2.26e-04	5.46e-05	4.94e-05

TABLE 8.1: Average squared L^2 errors with 50 macro runs, $N = 10$, one-step scheme.

	$K = 10$	$K = 20$	$K = 50$	$K = 100$	$K = 200$	$K = 400$
$M = 20$	0.1031	0.0299	0.0073	0.0018	0.0011	0.0012
$M = 40$	0.1031	0.0294	0.0066	0.0014	7.86e-04	7.28e-04
$M = 80$	0.1027	0.0293	0.0065	0.0010	3.18e-04	3.86e-04
$M = 160$	0.1027	0.0294	0.0064	8.91e-04	2.46e-04	1.04e-04
$M = 320$	0.1026	0.0293	0.0064	8.39e-04	1.42e-04	7.03e-05
$M = 640$	0.1027	0.0292	0.0063	8.04e-04	8.16e-05	5.60e-05

TABLE 8.2: Average squared L^2 errors with 50 macro runs, $N = 20$, one-step scheme.

	$K = 10$	$K = 20$	$K = 50$	$K = 100$	$K = 200$	$K = 400$
$M = 20$	0.0484	0.0066	0.0017	0.0015	0.0011	0.0013
$M = 40$	0.0488	0.0058	8.45e-04	5.81e-04	6.35e-04	5.68e-04
$M = 80$	0.0478	0.0053	4.33e-04	2.96e-04	3.45e-04	4.06e-04
$M = 160$	0.0481	0.0051	2.98e-04	2.23e-04	1.71e-04	1.08e-04
$M = 320$	0.0479	0.0051	1.79e-04	6.48e-05	8.38e-05	1.04e-04
$M = 640$	0.0478	0.0050	1.50e-04	6.49e-05	6.66e-05	5.70e-05

TABLE 8.3: Average squared L^2 errors with 50 macro runs, $N = 10$, multi-step scheme.

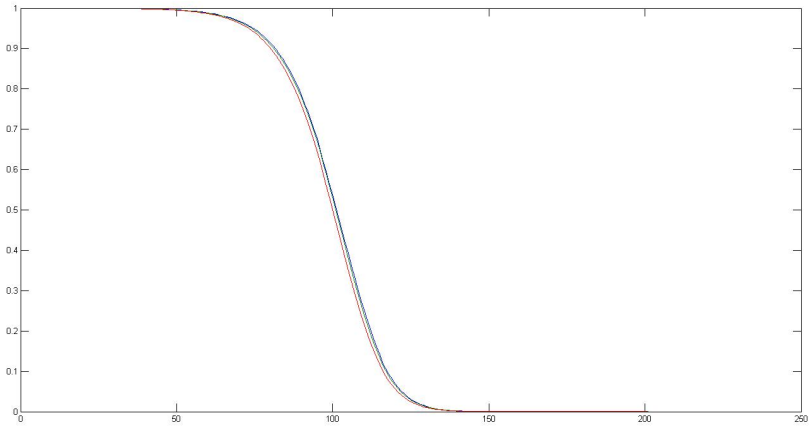


FIGURE 8.1: Upper and lower bounds of exaxt solution and piecewise constant estimation with $M = 50$ and $K = 200$

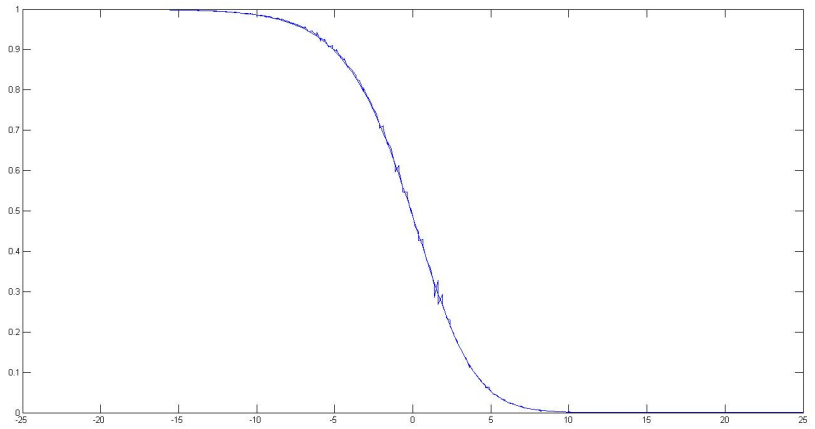


FIGURE 8.2: exact solution and linear estimation with $M = 50$ and $K = 200$

Table 8.4 below describes numerical results in dimension 2. The final time T is fixed to 1 and we use the time discretization $t_i = \frac{i}{N}T, 0 \leqslant i \leqslant N$ with $N = 10$. We divide the real line $\mathbb{R}$ into K subintervals $(I_i)_{1 \leqslant i \leqslant K}$ by fixing $A = 25$ and dividing $[-A, A]$ into $K - 2$ equal length intervals and then adding $(-\infty, -A)$ and $(A, +\infty)$. We take $\Sigma = [1, \beta; \beta, 1]$ with $\beta = 0.25$ and $\theta = \frac{[1;1]}{\sqrt{2}}$. We implement our method by using piecewise constant estimation on each finite (or infinite) rectangle $I_i \times I_j$. Then finally we get a piecewise constant estimation of $W(0, y)$, noted as $\hat{W}(0, y)$. Then we approximate the squared $L^2(\nu \otimes \nu)$ error of our estimation by

$$\sum_{1 \leqslant k_1 \leqslant K, 1 \leqslant k_2 \leqslant K} |W(0, y_{k_1}, y_{k_2}) - \hat{W}(0, y_{k_1}, y_{k_2})|^2 \nu \otimes \nu(I_{k_1} \times I_{k_2})$$

where (y_{k_1}, y_{k_2}) is chosen as the middle point of the rectangle if $I_{k_1} \times I_{k_2}$ is finite and the boundary point if one or both of I_{k_1} and I_{k_2} are infinite. We take $\nu(\mathrm{d}x) = \frac{1}{2}e^{-|x|}\mathrm{d}x$ and we use the restriction of $\nu \otimes \nu$ on $I_i \times I_j$ to sample initial points. The squared $L^2(\nu \otimes \nu)$ norm of $W(0, \cdot, \cdot)$ is around 0.25.

	$K = 10$	$K = 20$	$K = 50$	$K = 100$	$K = 200$
$M = 20$	0.0592	0.0167	0.0027	0.0018	0.0010
$M = 40$	0.0588	0.0163	0.0022	5.34e-04	5.00e-04
$M = 80$	0.0588	0.0160	0.0019	3.74e-04	2.98e-04
$M = 160$	0.0586	0.0160	0.0018	3.08e-04	9.16e-05
$M = 320$	0.0586	0.0159	0.0017	1.1e-04	9.24e-05

TABLE 8.4: Average squared L^2 errors with 50 macro runs, $N = 10$, one-step scheme.

As for the previous case in dimension 1, we observe that when K is small, it is useless to increase M. This is because in such case the approximation error is dominant. But when K is large enough, the performance of our method improves when M becomes larger, since this time it is the statistical error which becomes dominant and larger M means smaller statistical error.

In the perspective of a given root sample (M fixed), it is recommended to take K large: indeed, in agreement with Theorems 7.3.4 and 7.4.1, we observe from the numerical results that the global error decreases up to the statistical error term (depending on M but not K). In this way, for $M = 20$ (resp. $M = 40$) the relative squared L^2 error is about 0.4% (resp. 0.22%).

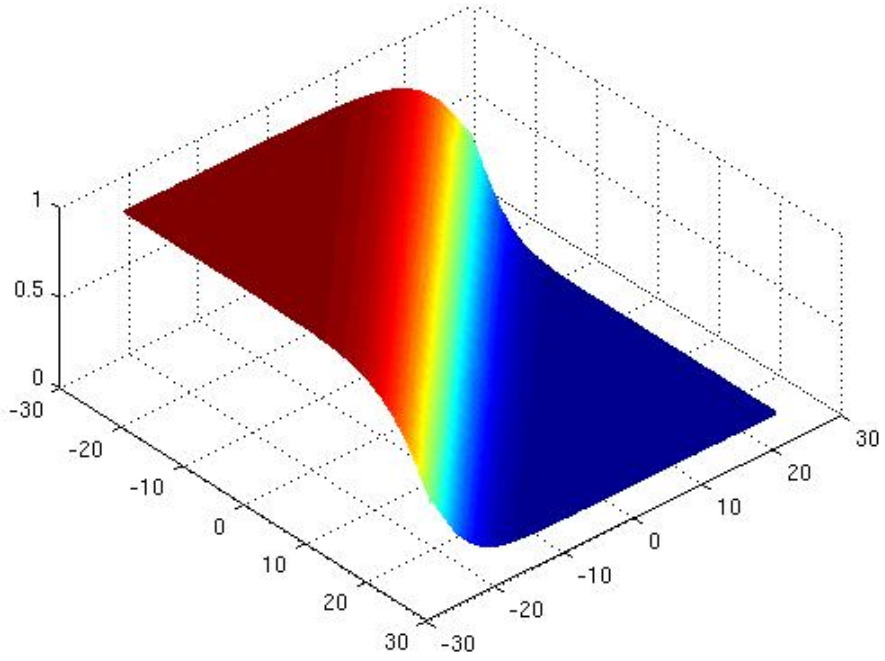


FIGURE 8.3: Estimated solution with $M = 320$ and $K = 200$

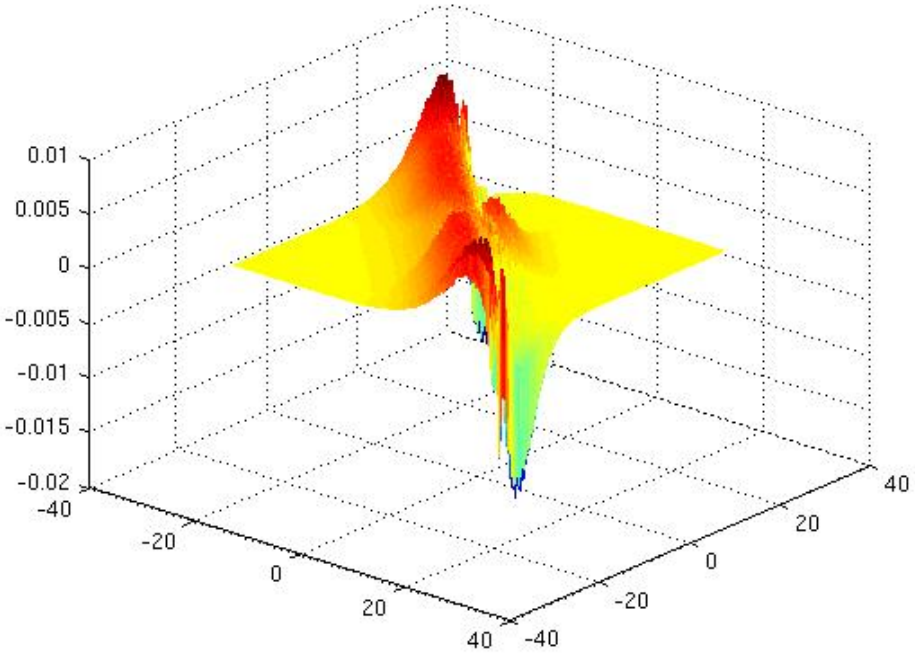


FIGURE 8.4: Estimation error with $M = 320$ and $K = 200$

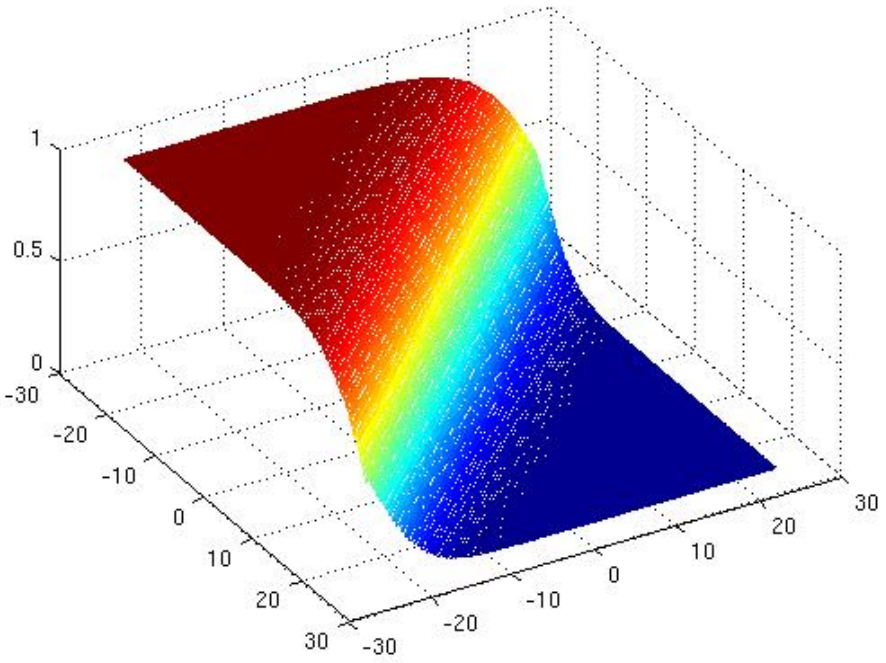


FIGURE 8.5: Estimated solution with $M = 40$ and $K = 100$

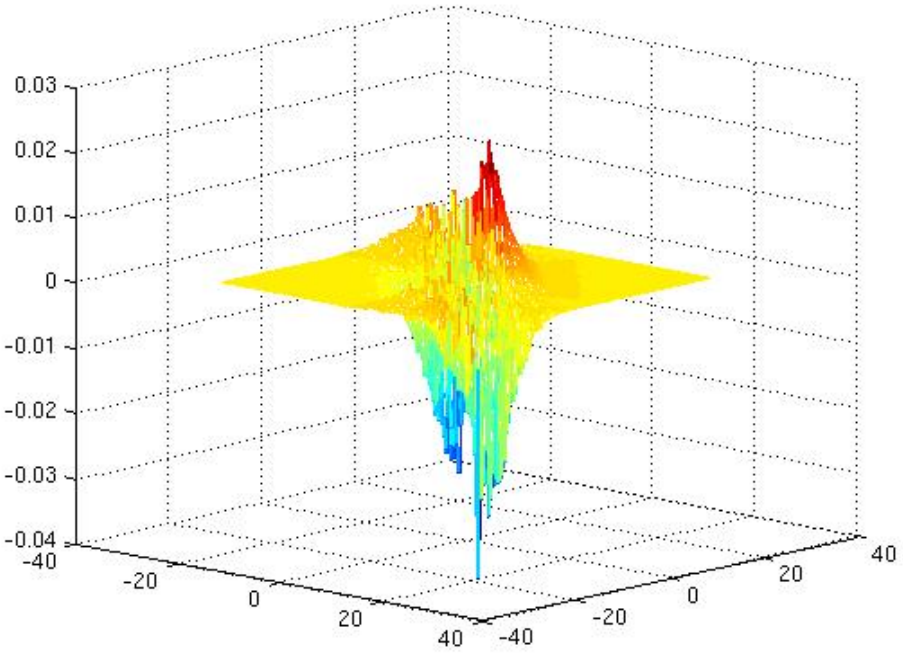


FIGURE 8.6: Estimation error with $M = 40$ and $K = 100$

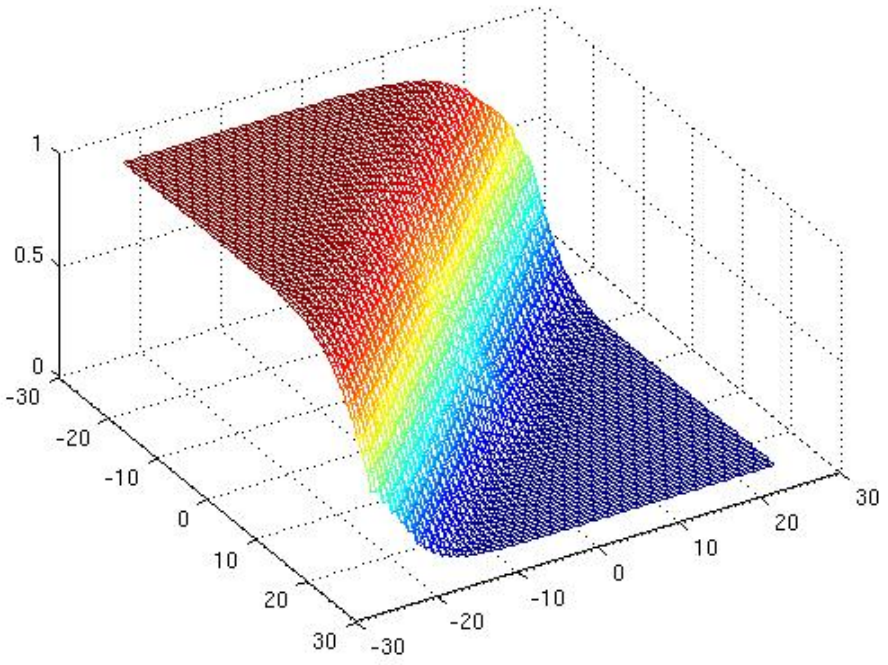


FIGURE 8.7: Estimated solution with $M = 40$ and $K = 50$

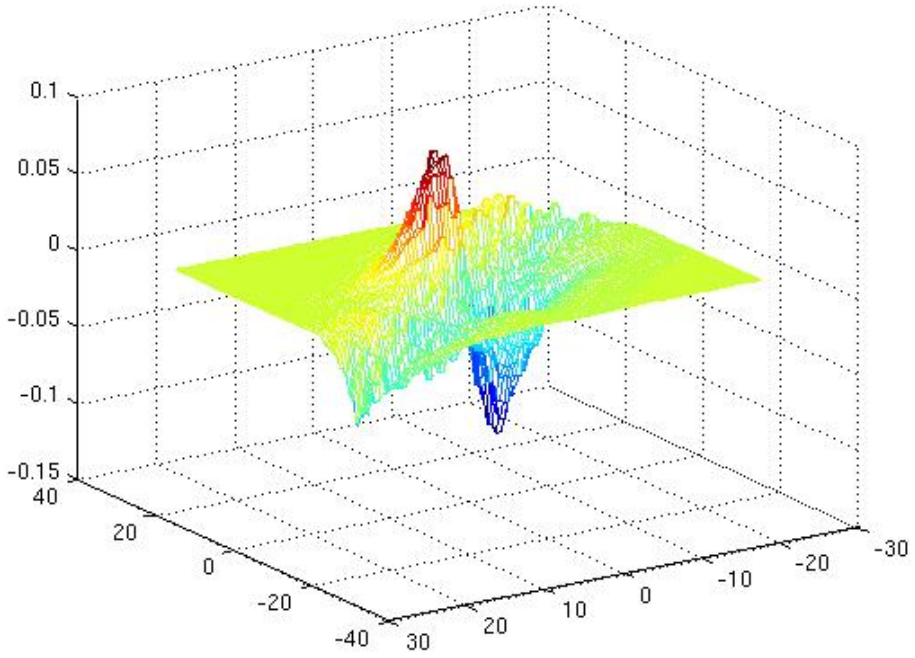


FIGURE 8.8: Estimation error with $M = 40$ and $K = 50$

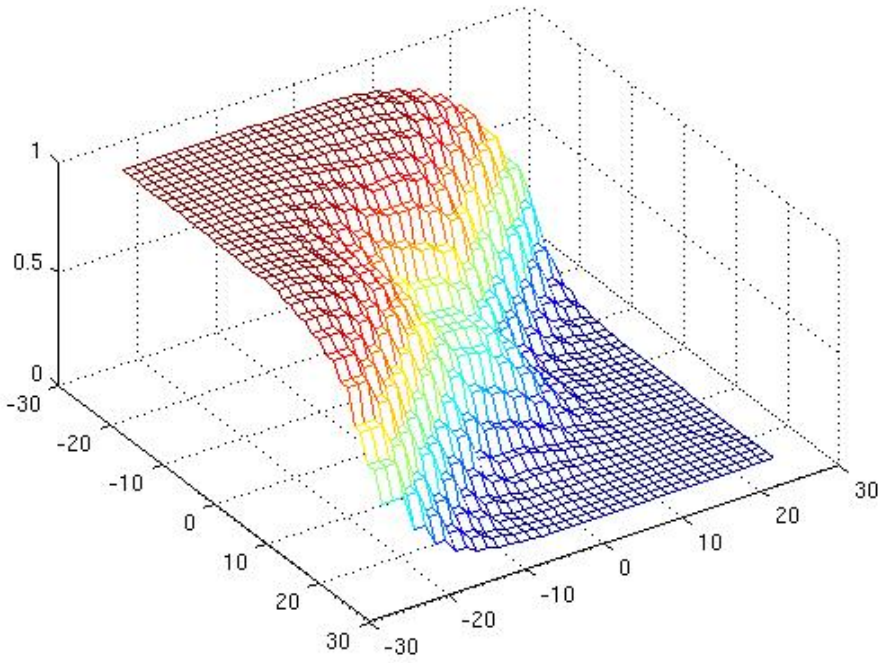


FIGURE 8.9: Estimated solution with $M = 40$ and $K = 20$

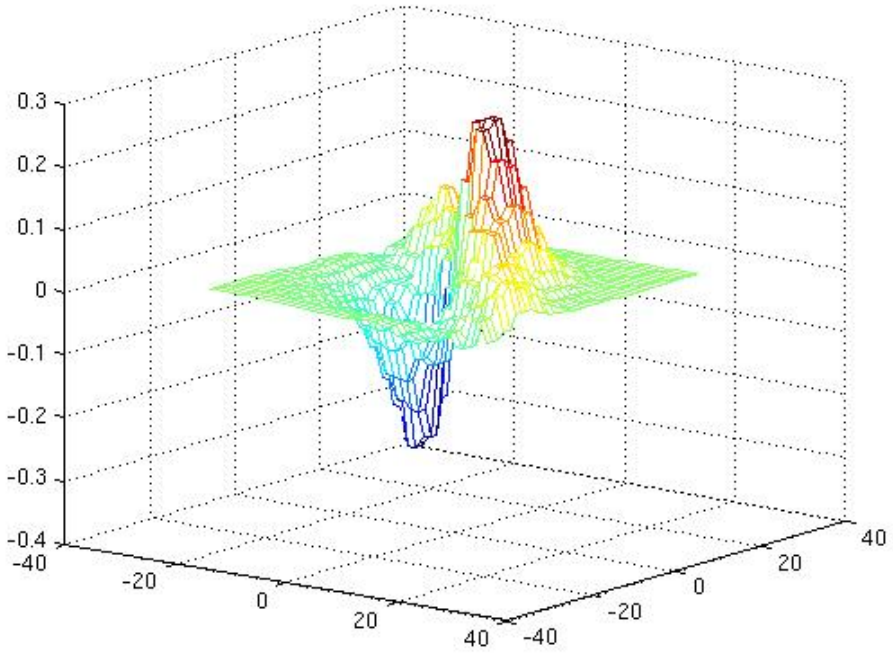


FIGURE 8.10: Estimation error with $M = 40$ and $K = 20$

How to derive FKPP equation

Suppose the population density at point x and time t is $u(t, x)$. Then given a arbitray volume Ω, the number of individuals inside Ω is given by

$$N(t) = \int_{\Omega} u(t, x)dx$$

As time goes on, the number of individuals inside Ω will be affected by two different sources. The first change is internal and it comes from the new births inside Ω, which depends on the base population and also on the limited natural resource supporting the population inside a fixed domain. When the base population is small, there are not many new births and then the birth number increases with the population number until this tendency is inversed by the limited resource. We model this internal change by

$$\int_{\Omega} au(t, x)(1 - u(t, x))dx$$

Remark that a is the only value that we need to know to apply our NISR method. Besides, there is another change which is external, i.e. individuals coming and leaving across the boundary of Ω. We suppose that this movement is proportional to the gradient of $u(t, x)$ with the coefficient D. Thus is external change is given by

$$\int_{\partial\Omega} D\nabla u(t, x) \cdot n(t, x)dS$$

where $n(t,x)$ is the unit outward orthogonal vector. By divergence theorem it is equal to

$$\int_\Omega D\Delta u(t,x)dx$$

So finally we have

$$\begin{aligned}\frac{\mathrm{d}N(t)}{\mathrm{d}t} &= \int_\Omega \frac{\mathrm{d}u(t,x)}{\mathrm{d}t}dx \\ &= \int_\Omega au(t,x)(1-u(t,x))dx + \int_\Omega D\Delta u(t,x)dx\end{aligned}$$

Since the domain Ω is arbitrary, we get

$$\frac{\partial u(t,x)}{\partial t} = au(t,x)(1-u(t,x)) + D\Delta u(t,x)$$

By a change of variable $s = T-t$, we get the FKPP equation as stated in this section.

Remark that if we keep the constant D, then Equation (8.1.3) will become

$$dP_s = \sqrt{2D}dW_s$$

which means, when we apply our algorithm based on observed trajectories of population individuals, we are making implicitly an assumption that the diffusion coefficient is proportional to the variance of individual movement, which may be simplistic but still is a reasonable assumption.

8.2 Travel agency problem: when to offer travels, according to currency and weather forecast...

In this section we illustrate the stratified resampler methodology in the solution of an optimal investment problem. The underlying model will have two sources of stochasticity, one related to the weather and the other one to the exchange rate. The corresponding stochastic processes shall be denoted by X_t^1 and X_t^2.

We envision the following situation: A travel agency wants to launch a campaign for the promotion of vacations in a warm region abroad during the Fall-Winter season. Such travel agency would receive a fixed value $\underline{c}$ in local currency from the customers and on the other hand would have to pay the costs $c = c(\exp(X^2_{\tau+1/12}))$ in a future time $\tau+1/12$, where τ is the launching time of the campaign and $X^2_{\tau+1/12}$ is the prevailing logarithm of exchange rate one month after the launching, with the time unit set to be one year. The initial time $t = 0$ is by convention October 1st. In other words, the costs are fixed to the traveler and variable

for the agency. A pictorial description of the cost function is presented in Figure 8.11.

The effectiveness of the campaign will depend on the local temperature $(t-0.25)^2 \times 240 + X_t^1$ (in Celsius) and will be denoted by $q((t-0.25)^2 \times 240 + X_t^1)\exp(-|t-1/6|)$, where $(t-0.25)^2 \times 240$ represents the seasonal component and X_t^1 represents the random part. Its purpose is to capture the idea that if the local temperature is very low, then people would be more interested in spending some days in a warm region, whereas if the weather is mild then people would just stay at home. A pictorial description of the function q is presented in Figure 8.11. The second part of this function $\exp(-|t-1/6|)$ is created to represent the fact that there are likely more registrations at beginning of December for the period of new year holidays.

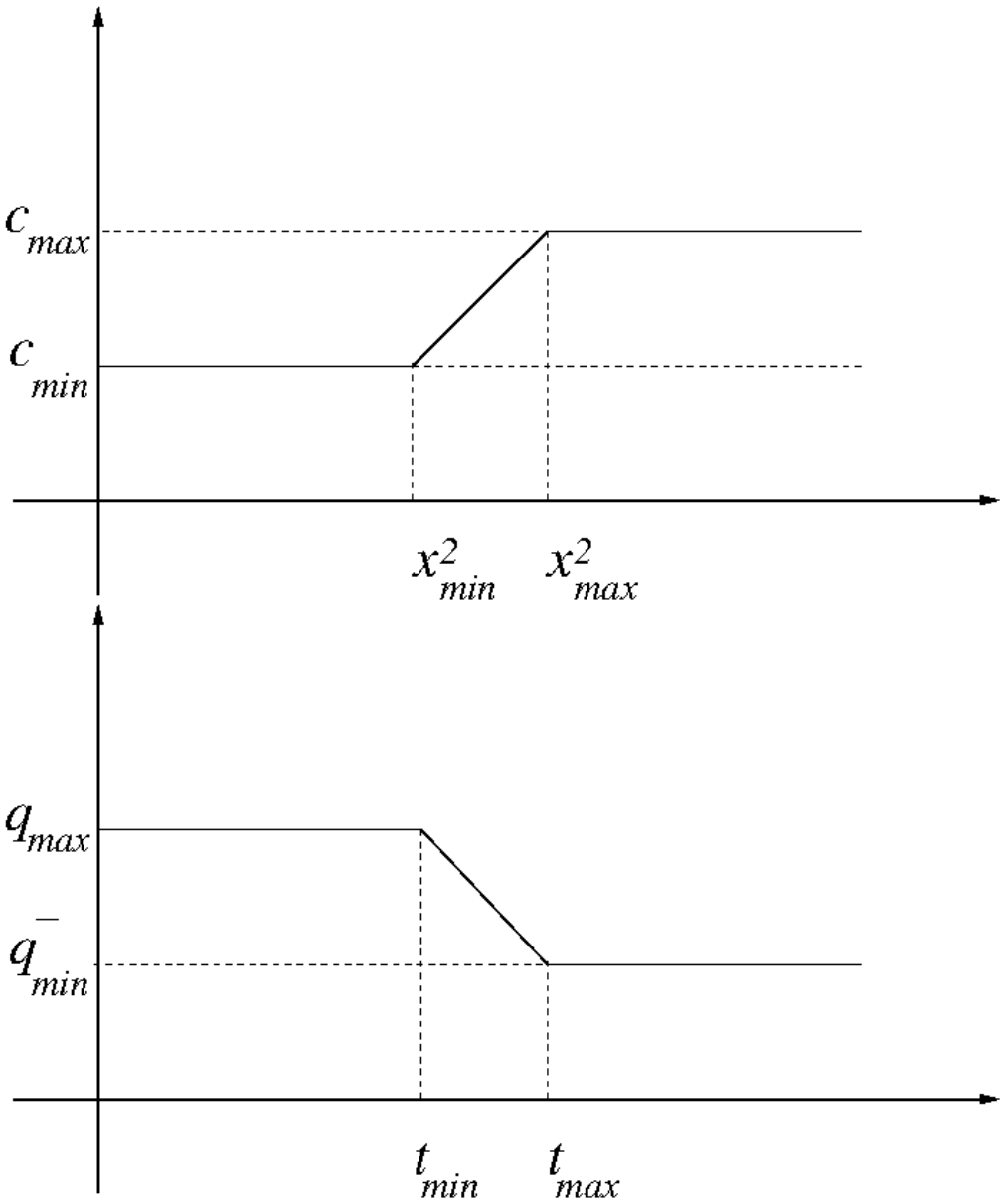


FIGURE 8.11: Pictorial description of the cost function c (left) and of the campaign effectiveness q (right).

Thus, our problem consists of finding the function v defined by

$$v(X_0^1, X_0^2) = \operatorname*{ess\,sup}_{\tau \in \mathcal{T}} \mathbb{E}\left(q((\tau - 0.25)^2 \times 240 + X_\tau^1) e^{-|\tau - 1/6|} \left(\underline{c} - c(e^{X^2_{\tau + 1/12}}) \right) \mid X_0^1, X_0^2 \right)$$
$$= \operatorname*{ess\,sup}_{\tau \in \mathcal{T}} \mathbb{E}\left(q((\tau - 0.25)^2 \times 240 + X_\tau^1) e^{-|\tau - 1/6|} \left(\underline{c} - \mathbb{E}\left(c(e^{X^2_{\tau + 1/12}}) \mid X_\tau^2 \right) \right) \mid X_0^1, X_0^2 \right),$$

where $\mathcal{T}$ denotes the set of stopping times valued in the weeks of the Fall-Winter seasons $\{\frac{k}{48}, k = 0, 1, \cdots, 24\}$, which corresponds to possible weekly choices for the travel agency to launch the campaign. The above function v models the optimal expected benefit for the travel agency and the optimal τ gives the best launching time. We shall assume, for

	$K = 10$	$K = 20$	$K = 50$	$K = 100$
$M = 20$	0.1827	0.0512	0.0349	0.0269
$M = 40$	0.1982	0.0361	0.0249	0.0114
$M = 80$	0.2063	0.0325	0.0051	0.0047
$M = 160$	0.1928	0.0264	0.0058	0.0067

TABLE 8.5: Average squared L^2 errors with 20 macro runs. Simple regression.

simplicity, that the processes X^1 and X^2 are uncorrelated since we do not expect much influence of the weather on the exchange rate or vice-versa.

The problem is tackled by formulating it as a dynamic programming one related to optimal stopping problems (as exposed in Section 7.3) using a mean-reversion process for the underlying process X^1 and a drifted Brownian motion for X^2. Their dynamics are given as follows:

$$\mathrm{d}X_t^1 = -aX_t^1\mathrm{d}t + \sigma_1\mathrm{d}W_t, X_0^1 = 0, \qquad X_t^2 = -\frac{\sigma_2^2}{2}t + \sigma_2 B_t.$$

The cost function c is chosen piecewise linear so that we can get $\mathbb{E}(c(e^{X^2_{\tau+1/12}})|X^2_\tau)$ explicitly as a function of X^2_τ using the Black-Scholes formula in mathematical finance. Thus we can run our method in two different ways: either using this explicit expression and apply directly the regression scheme of Section 7.3; or first estimating

$$\mathbb{E}(c(e^{X^2_{\tau+1/12}})|X^2_\tau)$$

by stratified regression then plugging the estimate in our method again to get a final estimation. We refer to these two different ways as simple regression and nested regression. The latter case corresponds to a coupled two-component regression problem (that could be mathematically analyzed very similarly to Section 7.3).

The parameter's values are given as: $a = 2, \sigma_1 = 10, \sigma_2 = 0.2, \underline{c} = 3, x^2_{min} = e^{-0.5}, x^2_{max} = e^{0.5}, c_{min} = 1, c_{max} = c_{min} + x^2_{max} - x^2_{min}, t_{min} = 0, t_{max} = 15, q_{min} = 1, q_{max} = 4$. We use the restriction of $\mu(\mathrm{d}x) = \frac{k}{2}(1+|x|)^{-k-1}\mathrm{d}x$ with $k = 6$ to sample point for X^1 and the restriction of $\nu(\mathrm{d}x) = \frac{1}{2}e^{-|x|}\mathrm{d}x$ to sample points for X^2. Note that $k = 6$ means that, in the error estimation, more weight is distributed to the region around $X_0^1 = 0$, which is the real interesting information for the travel agency.

We will firstly run our method with $M = 320$ and $K = 300$ to get a reference value for v then our estimators will be compared to this reference value in a similar way as in the previous example. The squared $L^2(\mu \otimes \nu)$ norm of our reference estimation is 32.0844. The results are displayed in the Tables 8.5 and 8.6.

As in Subsection 8.1 and in agreement with Theorem 7.3.4, we observe an improved accuracy as K and M increases, independently of each other. The relative error is rather small even for small M.

	$K = 10$	$K = 20$	$K = 50$	$K = 100$
$M = 20$	0.1711	0.0458	0.0436	0.0252
$M = 40$	0.1648	0.0361	0.0130	0.0169
$M = 80$	0.1534	0.0273	0.0109	0.0085
$M = 160$	0.1510	0.0296	0.0048	0.0058

TABLE 8.6: Average squared L^2 errors with 20 macro runs. Nested regression.

Interestingly, the nested regression algorithm (which is the most realistic scheme to use in practice when the model is unknown) is as accurate as the scheme using the explicit form of the internal conditional expectation $\mathbb{E}\left(c(e^{X^2_{\tau+1/12}}) \mid X^2_\tau\right)$. Surprisingly, the simple regression scheme takes much more time than the nested regression one because of the numerous evaluations of the Gaussian CDF in the Black-Scholes formula.

More numerical results are given in the following graphs.

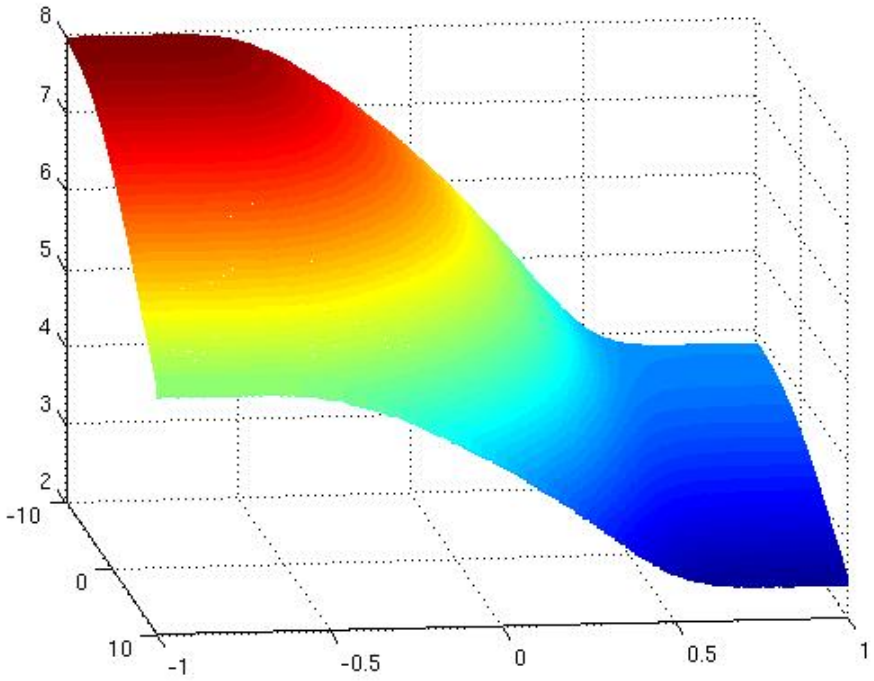


FIGURE 8.12: Reference value for v at time $t = 0$ obtained with $K = 200$ and $M = 320$

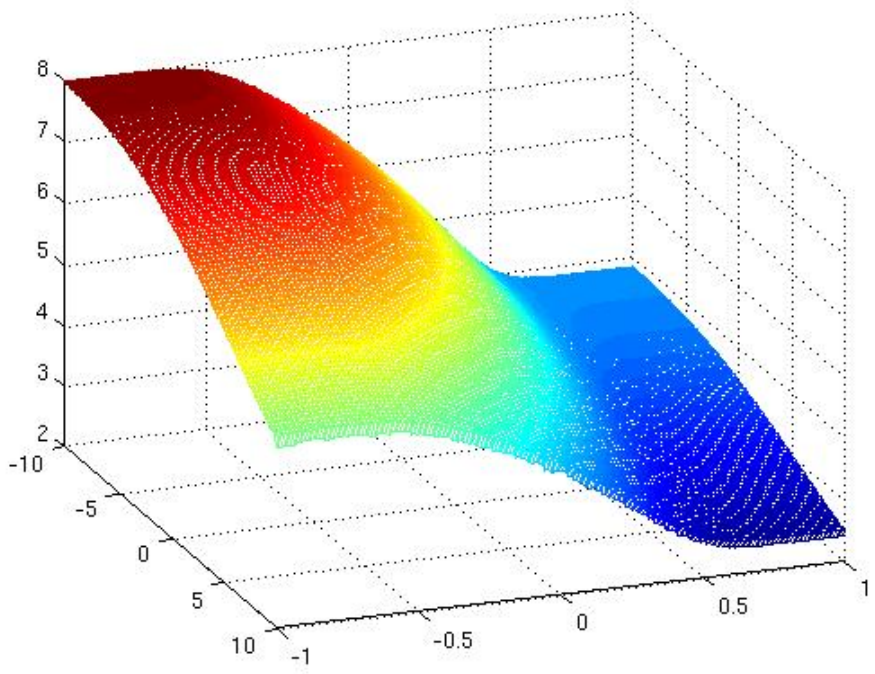


FIGURE 8.13: Estimated value for v at time $t = 0$ obtained with $K = 100$ and $M = 40$

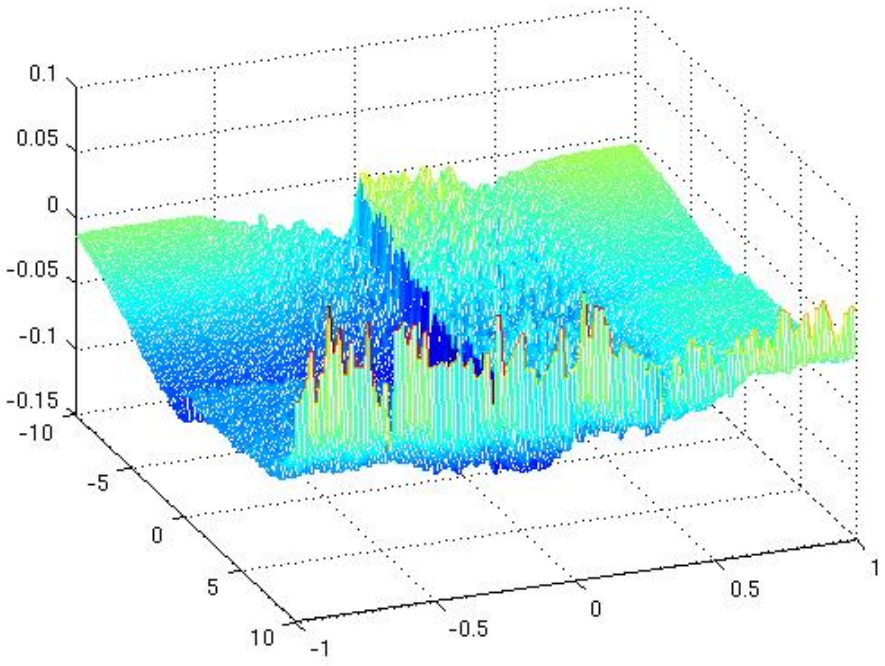


FIGURE 8.14: Estimation error for v at time $t = 0$ obtained with $K = 100$ and $M = 40$

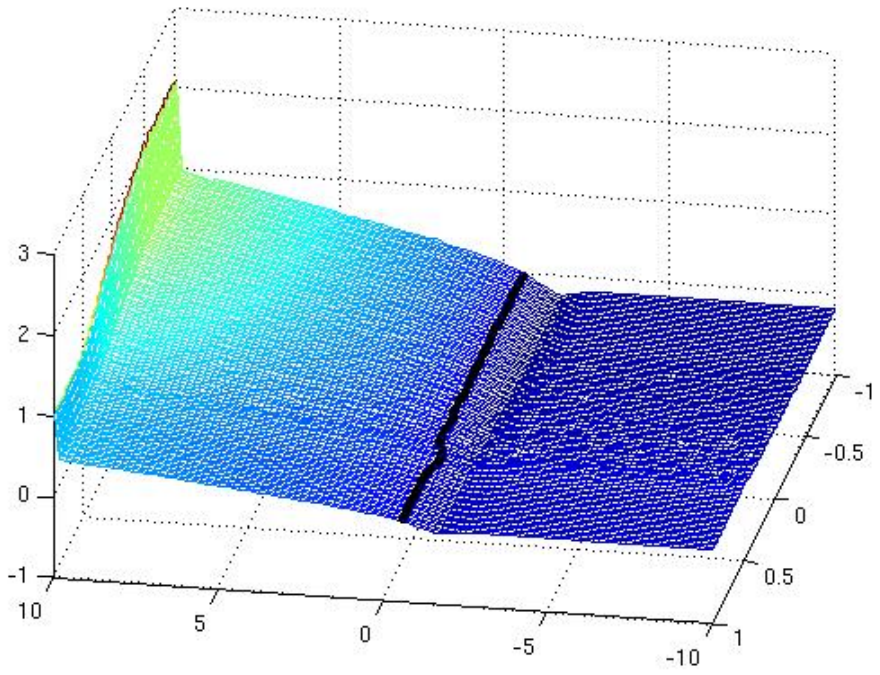


FIGURE 8.15: Difference between continuation value and current payoff at end November

8.3 Conclusion

As we can see from the above examples, our method works well with relatively small root samples, especially in the case where the number of hypercubes is large. We have thus proposed an efficient method to solve dynamic programming problem where only historical data is available and no additional simulation is possible. It still remains to be explored how this methodology can be further generalized to more complicated models, such as local volatility model.

Bibliography

[1] M.J. Ablowitz and A. Zeppetella. "Explicit solutions of Fisher's equation for a special wave speed". In: *Bull. Math. Biol.* 41.6 (1979), pp. 835–840. ISSN: 0092-8240. DOI: 10.1016/S0092-8240(79)80020-8. URL: http://dx.doi.org/10.1016/S0092-8240(79)80020-8.

[2] A. Agarwal, S. De Marco, E. Gobet, and G. Liu. "Rare event simulation related to financial risks: efficient estimation and sensitivity analysis". In: *Preprint available at https://hal-polytechnique.archives-ouvertes.fr/hal-01219616* (2015).

[3] Ankush Agarwal, Stefano De Marco, Emmanuel Gobet, and Gang Liu. "Study of new rare event simulation schemes and their application to extreme scenario generation". In: *Preprint available at https://hal-polytechnique.archives-ouvertes.fr/hal-01249625* (2016).

[4] Carol Alexander and José María Sarabia. "Quantile Uncertainty and Value-at-Risk Model Risk". In: *Risk Analysis* 32.8 (2012), pp. 1293–1308.

[5] S. Asmussen and H. Albrecher. *Ruin probabilities*. second. Advanced Series on Statistical Science & Applied Probability, 14. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2010.

[6] S. Asmussen and R.Y. Rubinstein. "Sensitivity analysis of insurance risk models via simulation". In: *Management Science* 45.8 (1999), pp. 1125–1141.

[7] Søren Asmussen and Peter W. Glynn. "A new proof of convergence of MCMC via the ergodic theorem". In: *Statist. Probab. Lett.* 81.10 (2011), pp. 1482–1485. ISSN: 0167-7152. DOI: 10.1016/j.spl.2011.05.004. URL: http://dx.doi.org/10.1016/j.spl.2011.05.004.

[8] Krishna B. Athreya and Peter E. Ney. *Branching processes*. Die Grundlehren der mathematischen Wissenschaften, Band 196. Springer-Verlag, New York-Heidelberg, 1972, pp. xi+287.

[9] S.K. Au and J.L. Beck. "Estimation of small failure probabilities in high dimensions by Subset Simulation". In: *Probabilistic Engineering Mechanics* 16.4 (2001), pp. 263–277.

[10] A. Bachouch, E. Gobet, and A. Matoussi. "Empirical Regression Method for Backward Doubly Stochastic Differential Equations". In: *To appear in SIAM ASA Journal on Uncertainty Quantification* (2015).

[11] Vincent Bansaye and Julien Berestycki. "Large deviations for branching processes in random environment". In: *arXiv preprint arXiv:0810.4991* (2008).

[12] N. H. Barton and M. Turelli. "Spatial Waves of Advance with Bistable Dynamics: Cytoplasmic and Genetic Analogues of Allee Effects". In: *The American Naturalist* 178.3 (2011), E48–E75. ISSN: 00030147, 15375323. URL: http://www.jstor.org/stable/10.1086/661246.

[13] A.J. Bayes. "Statistical techniques for simulation models". In: *Australian computer journal* 2.4 (1970), pp. 180–184.

[14] D. Belomestny, A. Kolodko, and J. Schoenmakers. "Regression methods for stochastic control problems and their convergence analysis". In: *SIAM Journal on Control and Optimization* 48.5 (2010), pp. 3562–3588.

[15] C. Bender and R. Denk. "A forward scheme for backward SDEs". In: *Stochastic Processes and their Applications* 117.12 (2007), pp. 1793–1823.

[16] S. Bhamidi, J. Hannig, C.Y. Lee, and J. Nolen. "The importance sampling technique for understanding rare event in Erdos-Renyi random graphs". In: *arXiv preprint arXiv:1302.6551* (2013).

[17] P. Billingsley. *Convergence of probability measures*. second. A Wiley-Interscience Publication. New York: John Wiley & Sons Inc., 1999.

[18] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[19] J. Blanchet, P. Glynn, and K. Leder. "On Lyapunov inequalities and subsolutions for efficient importance sampling". In: *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 22.3 (2012), p. 13.

[20] Jose Blanchet, Kevin Leder, and Yixi Shi. "Analysis of a splitting estimator for rare event probabilities in Jackson networks". In: *Stoch. Syst.* 1.2 (2011), pp. 306–339. ISSN: 1946-5238. DOI: 10.1214/11-SSY026. URL: http://dx.doi.org/10.1214/11-SSY026.

[21] B. Bollobás. *Random graphs*. second. Vol. 73. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2001.

[22] Z. I. Botev and D. P. Kroese. "An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting". In: *Methodol. Comput. Appl. Probab.* 10.4 (2008), pp. 471–505. ISSN: 1387-5841. DOI: 10.1007/s11009-008-9073-7. URL: http://dx.doi.org/10.1007/s11009-008-9073-7.

[23] Z. I Botev and D. P Kroese. "Efficient Monte Carlo simulation via the generalized splitting method". In: *Statistics and Computing* 22.1 (2012), pp. 1–16.

[24] Zdravko I Botev, Dirk P Kroese, and Thomas Taimre. "Generalized cross-entropy methods with applications to rare-event simulation and optimization". In: *Simulation* 83.11 (2007), pp. 785–806.

[25] J.A. Bucklew. *Introduction to Rare Event Simulation*. Springer, 2004.

[26] R. Carmona and S. Crépey. "Particle methods for the estimation of credit portfolio loss distributions". In: *Int. J. Theor. Appl. Finance* 13.4 (2010), pp. 577–602.

[27] R. Carmona, J.P. Fouque, and D. Vestal. "Interacting particle systems for the computation of rare credit portfolio losses". In: *Finance Stoch.* 13.4 (2009), pp. 613–633.

[28] P. Carr and D. Madan. "Towards a theory of volatility trading". In: *Option Pricing, Interest Rates and Risk Management, Handbook in Mathematical Finance* (2001), pp. 458–476.

[29] F. Cérou, P. Del Moral, T. Furon, and A. Guyader. "Sequential Monte Carlo for rare event estimation". In: *Stat. Comput.* 22.3 (2012), pp. 795–808.

[30] F. Cérou, P. Del Moral, and A. Guyader. "A nonasymptotic theorem for unnormalized Feynman-Kac particle models". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 47.3 (2011), pp. 629–649.

[31] F. Cérou, P. Del Moral, F. Le Gland, and P. Lezaud. "Genetic genealogical models in rare event analysis". In: *ALEA Lat. Am. J. Probab. Math. Stat.* 1 (2006), pp. 181–203.

[32] F. Cérou and A. Guyader. "Adaptive multilevel splitting for rare event analysis". In: *Stoch. Anal. Appl.* 25.2 (2007), pp. 417–443.

[33] Frederic Cerou and Arnaud Guyader. "Fluctuation analysis of adaptive multilevel splitting". In: *arXiv preprint arXiv:1408.6366* (2014).

[34] J.F. Chassagneux and D. Crisan. "Runge-Kutta schemes for backward stochastic differential equations". In: *Ann. Appl. Probab.* 24.2 (2014), pp. 679–720.

[35] S. Chatterjee and S.R.S. Varadhan. "The large deviation principle for the Erdős-Rényi random graph." In: *Eur. J. Comb.* 32.7 (2011), pp. 1000–1017.

[36] L. Chaumont and M. Yor. *Exercises in probability*. second. Cambridge Series in Statistical and Probabilistic Mathematics. A guided tour from measure theory to random processes, via conditioning. Cambridge University Press, Cambridge, 2012.

[37] P. Cheridito, H. Kawaguchi, and M. Maejima. "Fractional Ornstein-Uhlenbeck processes". In: *Electron. J. Probab* 8.3 (2003), p. 14.

[38] F. Comte and E. Renault. "Long memory in continuous-time stochastic volatility models". In: *Mathematical Finance* 8.4 (1998), pp. 291–323.

[39] R. Cont. "Model uncertainty and its impact on the pricing of derivative instruments". In: *Mathematical finance* 16.3 (2006), pp. 519–547.

[40] Thomas Dean and Paul Dupuis. "Splitting for rare event simulation: A large deviation approach to design and analysis". In: *Stochastic processes and their applications* 119.2 (2009), pp. 562–587.

[41] G. Deelstra and F. Delbaen. "Convergence of discretized stochastic (interest rate) processes with stochastic drift term". In: *Applied stochastic models and data analysis* 14.1 (1998), pp. 77–84.

[42] P. Del Moral. *Feynman-Kac formulae: Genealogical and Interacting Particle Systems with applications*. New-York: Springer, 2004.

[43] P. Del Moral and J. Garnier. "Genealogical Particle analysis of rare events". In: *Annals of Applied Probability* 15 (2005), pp. 2496–2534.

[44] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*. Vol. 38. Springer Science & Business Media, 2009.

[45] J. Dieudonné. *Eléments d'analyse*. Tome 1, (chapitres I à XI), Fondements de l'Analyse moderne. Paris: Jacques Gabay, 1990.

[46] C.R. Doss, J.M. Flegal, G.L. Jones, and R.C. Neath. "Markov chain Monte Carlo estimation of quantiles". In: *Electronic Journal of Statistics* 8.2 (2014), pp. 2448–2478.

[47] R. Douc, E. Moulines, and D. Stoffer. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. CRC Press, 2014.

[48] Daniel Dufresne. "Algebraic properties of beta and gamma distributions, and applications". In: *Advances in Applied Mathematics* 20.3 (1998), pp. 285–299.

[49] B. Dupire. "Pricing with a smile". In: *Risk* 7.1 (1994), pp. 18–20.

[50] N. El Karoui, M. Jeanblanc-Picqué, and S.E. Shreve. "Robustness of the Black and Scholes formula". In: *Math. Finance* 8.2 (1998), pp. 93–126.

[51] F Escher. "On the probability function in the collective theory of risk". In: *Skand. Aktuarie Tidskr.* 15 (1932), pp. 175–195.

[52] European Central Bank. "EU banks' liquidity stress testing and contingency funding plans". In: *ECB Publications* http://www.ecb.europa.eu/pub/pdf/other/eubanksliquiditystresstesting200811e.pdf (2008).

[53] R. A. Fisher. "The wave of advance of advantageous genes". In: *Annals of Eugenics* 7 (1937), pp. 353–369.

[54] J-P. Fouque and T. Ichiba. "Stability in a model of inter-bank lending". In: *SIAM Journal on Financial Mathematics* 4.1 (2014), pp. 784–803.

[55] E. Fournié, J.M. Lasry, J. Lebuchoux, P.L. Lions, and N. Touzi. "Applications of Malliavin calculus to Monte Carlo methods in finance". In: *Finance and Stochastics* 3.4 (1999), pp. 391–412.

[56] P.K. Friz, J. Gatheral, A. Gulisashvili, A. Jacquier, and J. Teichmann, eds. *Large Deviations and Asymptotic Methods in Finance*. Springer Proceedings in Mathematics and Statistics. Berlin: Springer-Verlag, 2015.

[57] S. Gal, R. Y Rubinstein, and A. Ziv. "On the optimality and efficiency of common random numbers". In: *Mathematics and computers in simulation* 26.6 (1984), pp. 502–512.

[58] J. Gatheral and T. Jaisson. "Fractional volatility models". In: *Presentation at Bloomberg seminar* http://mfe.baruch.cuny.edu/wp-content/uploads/2012/09/FractionalVolatility2014BBQ.pdf (2014).

[59] J. Gatheral, T. Jaisson, and M. Rosenbaum. "Volatility is rough". In: *Preprint available at SSRN 2509457* (2014).

[60] H. Gebelein. "Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung". In: *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 21.6 (1941), pp. 364–379.

[61] Jochen Geiger, Götz Kersting, and Vladimir A Vatutin. "Limit theorems for subcritical branching processes in random environment". In: *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*. Vol. 39. 4. 2003, pp. 593–620.

[62] P.W. Glynn and D. Ormoneit. "Hoeffding's inequality for uniformly ergodic Markov chains". In: *Statistics & probability letters* 56.2 (2002), pp. 143–146.

[63] E. Gobet. "Revisiting the Greeks for European and American options". In: *Stochastic processes and applications to mathematical finance*. Ed. by J. Akahori, S. Ogawa, and S. Watanabe. World Scientific, 2004, pp. 53–71.

[64] E. Gobet, J. Lopez-Salas, P. Turkedjiev, and C. Vasquez. "Stratified regression Monte-Carlo scheme for semilinear PDEs and BSDEs with large scale parallelization on GPUs". In: *In revision for SIAM Journal of Scientific Computing, Hal preprint hal-01186000* (2015).

[65] E. Gobet and R. Munos. "Sensitivity analysis using Itô-Malliavin calculus and martingales. Application to stochastic control problem." In: *SIAM Journal of Control and Optimization* 43:5 (2005), pp. 1676–1713.

[66] E. Gobet and P. Turkedjiev. "Linear regression MDP scheme for discrete backward stochastic differential equations under general conditions". In: *Math. Comp.* 85.299 (2016), pp. 1359–1391.

[67] Emmanuel Gobet. *Méthodes de Monte-Carlo et processus stochastiques: du linéaire au non-linéaire*. Editions de l'Ecole Polytechnique, 2013.

[68] Emmanuel Gobet and Gang Liu. "Rare event simulation using reversible shaking transformations". In: *SIAM Journal on Scientific Computing* 37.5 (2015), A2295–A2316.

[69] Emmanuel Gobet, Gang Liu, and Jorge Zubelli. "A Non-intrusive stratified resampler for regression Monte Carlo: application to solving non-linear equations". In: *Preprint available at https://hal-polytechnique.archives-ouvertes.fr/hal-01291056* (2016).

[70] Robert B Gramacy and Michael Ludkovski. "Sequential design for optimal stopping problems". In: *SIAM Journal on Financial Mathematics* 6.1 (2015), pp. 748–775.

[71] A. Gulisashvili. "Asymptotic Equivalence in Lee's moment Formulas for the Implied Volatility, asset price models without moment explosions, and Piterbarg's conjecture". In: *International Journal of Theorical and Applied Finance* 15 (2012), pp. 1–34.

[72] A. Gulisashvili and P. Tankov. "Tail behavior of sums and differences of log-normal random variables". In: *ArXiv preprint arXiv:1309.3057* (2013).

[73] A. Gulisashvili, F. Viens, and X. Zhang. "Extreme-Strike Asymptotics for General Gaussian Stochastic Volatility Models". In: *arXiv preprint arXiv:1502.05442* (2015).

[74] J. Guyon. "Path-Dependent Volatility". In: *Preprint available at http://ssrn.com/abstract=2425048* (2014).

[75] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics, 2002.

[76] M. Hairer and J.C. Mattingly. "Yet another look at Harris? ergodic theorem for Markov chains". In: *Seminar on Stochastic Analysis, Random Fields and Applications VI*. Springer. 2011, pp. 109–117.

[77] A. G. Hawkes. "Spectra of some self-exciting and mutually exciting point processes". In: *Biometrika* 58 (1971), pp. 83–95.

[78] Nicholas J Higham. *Functions of matrices: theory and computation*. Siam, 2008.

[79] D.G. Hobson and L.C.G. Rogers. "Complete models with stochastic volatility". In: *Math. Finance* 8.1 (1998), pp. 27–48.

[80] A.A. Hoffmann, B.L. Montgomery, J. Popovici, I. Iturbe-Ormaetxe, P.H. Johnson, F. Muzzi, M. Greenfield, M. Durkan, Y.S. Leong, Y. Dong, et al. "Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission". In: *Nature* 476.7361 (2011), pp. 454–457.

[81] R. A Horn and C. R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[82] Vincent AA Jansen and Jin Yoshimura. "Populations can persist in an environment consisting of sink habitats only". In: *Proceedings of the National Academy of Sciences* 95.7 (1998), pp. 3696–3698.

[83] S. Janson. *Gaussian Hilbert spaces*. Vol. 129. Cambridge university press, 1997.

[84] H. Kahn and T.E Harris. "Estimation of particle transmission by random sampling". In: *National Bureau of Standards applied mathematics series* 12 (1951), pp. 27–30.

[85] David A Keith, H Resit Akçakaya, Wilfried Thuiller, Guy F Midgley, Richard G Pearson, Steven J Phillips, Helen M Regan, Miguel B Araújo, and Tony G Rebelo. "Predicting extinction risks under climate change: coupling stochastic population models with dynamic bioclimatic habitat models". In: *Biology Letters* 4.5 (2008), pp. 560–563.

[86] Christian Kelling. "A framework for rare event simulation of stochastic Petri nets using "RESTART"". In: *Proceedings of the 28th conference on Winter simulation*. IEEE Computer Society. 1996, pp. 317–324.

[87] J.F.C. Kingman. *Poisson processes*. Vol. 3. Oxford Studies in Probability. Oxford Science Publications. New York: The Clarendon Press Oxford University Press, 1993.

[88] A. Kohatsu-Higa and K. Yasuda. "Estimating multidimensional density functions using the Malliavin-Thalmaier formula". In: *SIAM J. Numer. Anal.* 47 (2009), pp. 1546–1575.

[89] A. Kolmogoroff, I. Pretrovsky, and N. Piscounoff. *Étude de l'équation de la diffusion avec croissance de la quantite de matière et son application à un problème biologique.* French. Bull. Univ. État Moscou, Sér. Int., Sect. A: Math. et Mécan. 1, Fasc. 6, 1-25. 1937.

[90] I. Kontoyiannis, L.A. Lastras-Montano, and S.P. Meyn. "Relative entropy and exponential deviation bounds for general Markov chains". In: *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*. IEEE. 2005, pp. 1563–1567.

[91] Christian Krattenthaler, Anthony J Guttmann, and Xavier G Viennot. "Vicious walkers, friendly walkers and Young tableaux: II. With a wall". In: *Journal of Physics A: Mathematical and General* 33.48 (2000), p. 8835.

[92] J. Krystul, F. Le Gland, and P. Lezaud. "Sampling per mode for rare event simulation in switching diffusions". In: *Stochastic Process. Appl.* 122.7 (2012), pp. 2639–2667.

[93] A. Lagnoux. "Rare event simulation". In: *Probab. Engrg. Inform. Sci.* 20.1 (2006), pp. 45–66.

[94] K. Łatuszyński, B. Miasojedow, and W. Niemiro. "Nonasymptotic bounds on the estimation error of MCMC algorithms". In: *Bernoulli* 19.5A (2013), pp. 2033–2066.

[95] R.W. Lee. "The moment formula for implied volatility at extreme strikes". In: *Mathematical Finance* 14.3 (2004), pp. 469–480.

[96] Michael Ludkovski. "A simulation approach to optimal stopping under partial information". In: *Stochastic Processes and their Applications* 119.12 (2009), pp. 4061–4087.

[97] J. Ma and J. Yong. *Forward-Backward Stochastic Differential Equations*. A course on stochastic processes. Lecture Notes in Mathematics, 1702, Springer-Verlag, 1999.

[98] S. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. second. Cambridge University Press, Cambridge, 2009.

[99] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Cambridge university press, 2009.

[100] J. D. Murray. *Mathematical biology. I*. Third. Vol. 17. Interdisciplinary Applied Mathematics. An introduction. Springer-Verlag, New York, 2002, pp. xxiv+551. ISBN: 0-387-95223-3.

[101] J. D. Murray. *Mathematical biology. II*. Third. Vol. 18. Interdisciplinary Applied Mathematics. Spatial models and biomedical applications. Springer-Verlag, New York, 2003, pp. xxvi+811. ISBN: 0-387-95228-4.

[102] Maurizio Naldi and F Calonico. "A comparison of the GEVT and RESTART techniques for the simulation of rare events in ATM networks". In: *Simulation Practice and Theory* 6.2 (1998), pp. 181–196.

[103] Olle Nerman. "On the maximal generation size of a non-critical galton-watson process". In: *Scandinavian Journal of Statistics* (1977), pp. 131–135.

[104] D. Nualart. *Malliavin calculus and related topics*. second. (with corrections on the webpage of the author). Springer Verlag, 2006.

[105] Y. Ogata. "On Lewis' simulation method for point processes". In: *Information Theory, IEEE Transactions on* 27.1 (1981), pp. 23–31.

[106] E. Pardoux and A. Rascanu. *Stochastic Differential Equations, Backward SDEs, Partial Differential Equations*. Vol. 69. Stochastic Modelling and Applied Probability. Springer-Verlag, 2014.

[107] William Parry. *Topics in ergodic theory*. Vol. 75. Cambridge Tracts in Mathematics. Reprint of the 1981 original. Cambridge University Press, Cambridge, 2004, pp. x+110. ISBN: 0-521-60490-7.

[108] Philippe Picard and Claude Lefèvre. "The probability of ruin in finite time with discrete claim size distribution". In: *Scandinavian Actuarial Journal* 1997.1 (1997), pp. 58–69.

[109] N.S. Pillai, A.M. Stuart, and A.H. Thiery. "Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions". In: *Annals of Applied Probability* 22.6 (2012), pp. 2320–2356.

[110] M. Potters, J.-P. Bouchaud, and D. Sestovic. "Hedged Monte Carlo: low variance derivative pricing with objective probabilities". In: *Physica A. Statistical Mechanics and its Applications* 289.3-4 (2001), pp. 517–525.

[111] M. Prandini and O.J. Watkins. "Probabilistic Aircraft Conflict Detection". In: *HYBRIDGE WP3: Reachability analysis for probabilistic hybrid systems* (2005).

[112] Joachim Rambeau and Grégory Schehr. "Distribution of the time at which N vicious walkers reach their maximal height". In: *Physical Review E* 83.6 (2011), p. 061146.

[113] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*. third. Comprehensive Studies in Mathematics. Berlin: Springer, 1999.

[114] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*. Third. Vol. 293. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1999, pp. xiv+602. ISBN: 3-540-64325-7. DOI: 10.1007/978-3-662-06400-9. URL: http://dx.doi.org/10.1007/978-3-662-06400-9.

[115] P. Robert. *Stochastic networks and queues*. French. Vol. 52. Applications of Mathematics (New York). Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2003.

[116] G. O Roberts, A. Gelman, and W. R Gilks. "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *The annals of applied probability* 7.1 (1997), pp. 110–120.

[117] G. Rubino and B. Gerardo, eds. *Rare event simulation using Monte Carlo methods*. Wiley, Chichester, 2009.

[118] R. Y. Rubinstein and D. P. Kroese. *The cross-entropy method*. Information Science and Statistics. A unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning. Springer-Verlag, New York, 2004, pp. xx+300. ISBN: 0-387-21240-X. DOI: 10.1007/978-1-4757-4321-0. URL: http://dx.doi.org/10.1007/978-1-4757-4321-0.

[119] Reuven Y Rubinstein. "Combinatorial optimization, cross-entropy, ants and rare events". In: *Stochastic optimization: algorithms and applications* 54 (2001), pp. 303–363.

[120] R.Y. Rubinstein and D.P. Kroese. *Simulation and the Monte Carlo method*. Second. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2008.

[121] Didier Rullière and Stéphane Loisel. "Another look at the Picard-Lefèvre formula for finite-time ruin probabilities". In: *Insurance: Mathematics and Economics* 35.2 (2004), pp. 187–203.

[122] I. Shigekawa. *Stochastic analysis*. Vol. 224. American Mathematical Soc., 2004.

[123] N. N. Taleb. *The black swan: The impact of the highly improbable*. Random House, 2007.

[124] J.N. Tsitsiklis and B. Van Roy. "Regression Methods for Pricing Complex American-Style Options". In: *IEEE Transactions on Neural Networks* 12.4 (2001), pp. 694–703.

[125] M. Villén-Altamirano and J. Villén-Altamirano. "RESTART: a method for accelerating rare event simulations". In: *Proceedings of the 13th International Teletraffic Congress, Copenhagen*. 1991, pp. 71–76.

[126] Manuel Villén-Altamirano and José Villén-Altamirano. "The rare event simulation method RESTART: efficiency analysis and guidelines for its application". In: *Network performance engineering*. Springer, 2011, pp. 509–547.

[127] Hermann Weyl. "Über die gleichverteilung von zahlen mod. eins". In: *Mathematische Annalen* 77.3 (1916), pp. 313–352.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Simulation des événements rares par transformation de shaking et NISR méthode pour la programmation dynamique

Mots clés : événement rare, splitting, ergodicité, programmation dynamique, stratification

Résumé :
Cette thèse contient deux sujets différents: la simulation d'événements rares et la résolution des programmations dynamiques par méthode de régression empirique stratifiée. Dans la première partie, on construit une transition markovienne appelée *transformation de shaking* sur l'espace des trajectoires, qui nous permet de proposer les méthodes IPS et POP, basées respectivement sur le système des particules en interaction et sur l'ergodicité de chaîne de Markov. Des constructions efficaces de transformations de shaking ont été proposées. On a aussi élaboré une version adaptative de la méthode POP. Les analyses de convergence pour ces méthodes sont également données. En plus, on montre comment ces techniques peuvent être utilisées pour calculer la sensibilité des événements rares. De nombreux exemples numériques sont donnés pour montrer la performance de nos méthodes. Dans la deuxième partie, notre but est de résoudre certains problèmes de programmation dynamique. À la différence du contexte usuel, on n'a pas accès à toutes les informations du modèle et seulement un petit ensemble d'observations historiques sont disponibles. On propose une méthode de régression par stratification et ré-échantillonnage. Plus précisément, on utilise les données historiques pour reconstruire d'autres trajectoires et faire des régressions locales sur l'espace stratifié. L'estimation des erreurs non-asymptotiques est établie avec certains exemples numériques.

Title : Rare event simulation by shaking transformations and NISR method for dynamic programming problems

Keywords : rare event, splitting, ergodicity, dynamic programming, stratification

Abstract :
This thesis covers two different subjects: rare event simulation and stratified regression method for dynamic programming problems. In the first part, we design a Markovian transition called *shaking transformations* on the path space, which enables us to propose IPS and POP methods, based respectively on interacting particle system and the ergodicity of Markov chain. Efficient designs of shaking transformation are proposed. We also design an adaptive version of the POP method. Theoretical analysis is given on the convergence of these methods. Besides, we demonstrate how these techniques can be applied to perform sensitivity analysis of rare event statistics and to make approximative sampling of rare event. Many numerical examples are discussed to show the performance of our methods. In the second part, we aim at numerically solving certain dynamic programming problems. Different from usual settings, we don't have access to full detail of the underlying model and only a relatively small-sized set of root sample is available. We propose a stratified resampling regression method. More precisely, we shall use given the root sample to reconstruct other paths and perform local regression on the stratified spaces. Non-asymptotic error estimations are given with several numerical examples.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France