Improved Bayesian model selection through non-equilibrium stochastic processes

Kerol Roussin DONTEU DJOUMESSI (kerol@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by:

Dr. Daniel Nickelsen
African Institute for Mathematical Science, South Africa
Dr. Bubacarr Bah
African Institute for Mathematical Science, South Africa

22 September 2020

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

Bayesian model evidence is probably the best tool to select the best model out of a set of candidate models. However, determining the model evidence becomes virtually impossible for complex models with many parameters, as it involves the marginalization of likelihood and prior in parameter space. Many works have been investigated to estimate the model evidence in such cases, often making use of Monte Carlo methods. Most of these methods, however, require stationarity and suffer from slow equilibration times. One approach to circumvent diverging equilibration times is to employ the Jarzynski equality, which links equilibrium quantities with non-equilibrium observables in stochastic thermodynamics. With the appropriate stochastic process and observable, the equilibrium quantity becomes the model evidence, such that the Jarzynski equality becomes an estimator for the model evidence. While this approach has been used before using Markov Chain Monte Carlo, the actual stochastic differential equations have to our knowledge not yet been analyzed. In this work, we use the connection between Jarzynski's equality, thermodynamic integration and the stochastic differential equations to simulate the high dimensional process which leads to an approximation of model evidence. Since this approximation involves both parameter and protocol influencing the convergence of Jarzynski's equality method, we also find the best parameter and protocol for a small duration of process. After simulations, the confidence interval error analysis shows that with small duration, the best parameter is $\gamma = 1$ while the *linear protocol* is the best.

Key words : stochastic process, model evidence, Jarzynski equality, model evidence.

L'inférence Bayésienne est probablement le meilleur outil pour sélectionner le meilleur modèle parmi un ensemble de modèles candidats. Cependant, déterminer l'évidence d'un modèle devient pratiquement impossible pour les modèles complexes avec de nombreux paramètres, car cela necessite une intégration sur un espace multidimensionnel. Plusieurs méthodes ont été proposé pour estimer l'évidence du modèle, dont celles basées sur l'algorithme de Monte-Carlo. Cependant, la plupart de ces méthodes nécessitent une quasi-stationnairité et souffrent de temps d'équilibrage lents. Une approche pour contourner les temps d'équilibrage divergents consiste à utiliser l'égalité de Jarzynski et la thermodynamique stochastique pour cette estimation. Cependant, les équations différentielles stochastiques n'ont pas encore été analysées à notre connaissance pour cette approximation. Dans ce travail, nous utilisons la connexion entre l'égalité, l'intégration thermodynamique et les équations différentielles stochastiques pour simuler le processus multidimentionel conduisant à l'approximation de l'évidence du modèle. Puisque cette approximation implique deux paramètres influençant l'estimation, nous sélectionnons le meilleur paramètre et protocole nécessaire à cette approximation sur une petite durée. Les resultats de simulations montrent qu'avec une petite durée, le meilleur paramètre est $\gamma = 1$ tandis qu'un *protocole lent* est meilleur.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Kerol Roussin DONTEU DJOUMESSI, 22 September 2020

Contents

1. Introduction

1.1 Background of the study

The main advantage of many statistical models or machine learning methods is being able to extract useful information from observed data in order to obtain predictive power. Whatever the data and fitting procedures used, the main task is to find the model that best matches the data from a set of candidate models. This task of finding the best model is known as model selection and lies at the heart of data analysis and machine learning methods. It is also becoming central in many other scientific studies such as economics, engineering, finance, biology, information theory, neuroscience, computer science, signal processing, and many others (Ding et al., 2018). To address the issue of model selection, a considerable number of methods such as Bayesian methods have been proposed, exhibiting varying performance and following different philosophies.

Although being known for almost three centuries (Bayes, 1763), Bayesian methods have only been used in an impressive array of applications in recent decades due to the need of sufficient computational power. Model comparison with Bayesian model evidence sometimes requires integration over a high dimensional parameter space which cannot be done analytically. In this case, using the power of today's typical computer, efficient algorithms such as Markov Chain Monte Carlo (MCMC) sampling have been used for some time to deal with this problem of high parameter space (Andrieu et al., 2001; Beck and Yuen, 2004; Marshall et al., 2006).

In general, probabilistic models are defined as the likelihood of observed data conditioned on values of unknown parameters (Knuth et al., 2015). These models differ in number and nature of parameters, and the prior knowledge we have on the parameters are encoded in the prior probabilities. Thus, for a given model $\mathcal{M}$, multiplying likelihood and prior distribution, and integrating over all parameters yields the marginal likelihood or the model evidence of observed data under $\mathcal{M}$. Application of the Bayes rule on model selection yields the posterior value which quantifies the probability that the data comes from this model. So, among the potential models, the one with the highest posterior probability will be the best model to describe the observed data.

The biggest challenge of Bayesian model selection is the marginalization due to the high-dimensional parameter space for which analytic integration is impossible, leading to numerical integration methods (Von Der Linden et al., 1999; Knuth et al., 2015; Durstewitz, 2017). A number of approximate methods have been investigated to tackle the marginalization problem such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Due to multi-modal and high dimensional optimization function in parameters space, these methods fail sometimes (MacKay, 1992). However, using stochastic processes to reduce computational cost, sampling techniques have improved the numerical estimation of model evidence in large-dimensional parameter spaces, leading to the study of other methods such as Annealed Importance Sampling (AIS) (Neal et al., 2011; Lee et al., 2016; Celeux et al., 2018).

In order to reduce the high dimensional integration space, some methods based on thermodynamic integration have been used to turn the high dimensional integration into a one dimensional integral. This is done by averaging this high dimensional integration quantity along a thermodynamic equilibrium process where the result is identical to the log-model evidence (Crooks, 2007; Blythe, 2008). However, keeping the process in equilibrium is computationally very expensive and sometimes infeasible because of multi-modality. Using AIS to reduce the computational time instead of keeping the process in equilibrium, the process is driven arbitrary and repeated several times. Each repetition results in different values for model evidence; the final approximation value of the evidence is given by the exponential average. AIS

has been connected to so-called fluctuation theorems (FTs), which is a central theorem in the field of stochastic thermodynamic.

The drawback of using AIS or FTs is that the exponential average is dominated by rare events. Thus, sampling requires more than $10,000$ repetitions of the non-equilibrium process (Jarzynski, 2006; Favaro et al., 2015). Another problem is the choice of the protocol that drives the non-equilibrium process. If this protocol is not chosen carefully, the estimator may get a high variance. Therefore, finding a general method to optimize the non-equilibrium protocol is a still a subject of active research even in stochastic thermodynamic fields (Then and Engel, 2008).

1.2 Motivation

FTs are used extensively to estimate thermodynamic quantities like the *Helmholtz* free energy in stochastic systems (Hummer and Szabo, 2001). However, not all methods used in stochastic thermodynamics have been employed to face the challenges of Bayesian model selection, which can be attributed to the lack of a shared mathematical description. Here, we propose to use stochastic differential equations as a shared mathematical description. Therefore, analyzing AIS as stochastic processes is a first step to benefit from more techniques developed in the field of stochastic thermodynamics. Moreover, one of these techniques is asymptotically exact expressions for the tails of the probability distributions of thermodynamic observable (Nickelsen and Engel, 2012). Application of this technique to Bayesian model selection would eliminate the issues of poor sampling of rare events crucial for AIS. Another such aspect is to find protocols to optimize various aspects of non-equilibrium processes in stochastic thermodynamic.

1.3 Aims and objectives

One approach to circumvent diverging equilibration times is to employ the Jarzynksi equality (JE). The JE links equilibrium quantities with non-equilibrium observables in stochastic thermodynamics. With the appropriate stochastic process and observable, the equilibrium quantity becomes the model evidence, such that the Jarzynski equality becomes an estimator for the model evidence. While this approach has been used before with the MCMC, the actual stochastic differential equations to the best of our knowledge have not yet been analysed. The aim of this essay is to formulate and simulate the high-dimensional stochastic process using the JE to estimate the model evidence for a multivariate unimodal Gaussian example.

1.4 Essay layout

The remain part of this essay is composed of the following.

Chapter 2 gives a literature review on model selection criteria and on the use of FTs and JE with Bayesian model evidence.

Chapter 3 gives the methods applied and the mathematical models. We mainly present the methodology used to estimate the model evidence using JE.

Chapter 4 provides the implementation part and results obtained, as well as discussion.

Chapter 5 gives a brief summary, future work and conclusions of the findings from the study.

2. Literature review

In statistics, to determine the correlation between a series of observations, the first step is to use mathematical tools to build models in order to understand these observations and then, extract knowledge from them. Once the models have been constructed, the next step is to apply some criteria to perform model selection among the candidate models. Model selection is a part of scientific research, which consist of choosing among the candidate models the one that best describes the observed data. From all the candidate models, the crucial question is how to choose the best model. A good model selection technique will balance goodness of fit with simplicity (Aho et al., 2014). However, the complexity of a model can be measured in terms of the number of parameters in the model or in terms of computational time. In this chapter we seek to review the theoretical background on Bayesian model selection with the model evidence and other criteria.

2.1 Criteria for model selection

The purpose of the model selection process is to evaluate the performance of different models in order to choose the best approximation that describes a given set of observed data (Friedman et al., 2001). To choose the set of candidate models, there are many methods such as data transformation, exploration data analysis, model specification, scientific methods, and more others (Cherkassky and Shao, 1998). Once the set of candidate models are known, there are many criteria to perform model selection such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Cross Validation (CV), the Bayesian model evidence and many others (Ding et al., 2018).

AIC is a fine technique based on in-sample fit to estimate the likelihood of a model to predict or estimate future values (Akaike, 1973). Using AIC, the best model is the one that has the lowest AIC among all the candidates. Bayesian or Schwarz information criterion (SIC) is another popular method which performs model selection by measuring the trade-off between model fit and complexity of the model (Schwarz et al., 1978). The idea behind it is to compute the BIC for all the candidate models and select one with the lowest BIC. It is partially based on the likelihood function and closely related to the AIC. The CV, or out-of-sample testing, denotes a class of model selection methods widely used in machine learning practice. It is one of various similar model validation techniques for evaluating how the results of a statistical analysis will generalize to an independent dataset (Stone, 1974; Cawley and Talbot, 2010). The CV is mainly used in real prediction problems when we want to estimate how accurately a predictive model will work in practice. The Bayesian model evidence, also called marginal likelihood or prior-predictive value, is another model selection technique naturally motivated by the Bayes factor which is the likelihood ratio of the model evidence of two competing models (Ly et al., 2016; Ding et al., 2018). Among candidate models, one with the largest prior-predictive value is favoured over other models.

Among the model selection methods, enumerated above, the Bayesian model evidence is the most difficult to calculate. In practice, the evidence can be computationally intensive as it often includes integration in high dimensional parameter space. This integration has only become feasible owing to computational resources of modern computers with the development of efficient algorithms like Markov Chain Monte-Carlo (MCMC) sampling which allows us to better explore the parameter space (Andrieu et al., 2001; Didelot et al., 2011). Another way to face this high dimensionality problem is by using thermodynamic integration (TI), where the high dimensional integral is turned into a one dimensional integral of an average quantity along a thermodynamic equilibrium process (Kirkwood, 1935; Neal,

1993). The problem with TI is that it is computationally expensive to keep the process in equilibrium due to multimodality. So, instead of using the MCMC method to sample, it could be important to find the exact stochastic process that corresponds to TI. Using the idea of the Jarzynski Equality (JE) with MCMC is sometimes referred to as Annealed Importance Sampling (AIS) in which the process is not kept in equilibrium but needs to be repeated several times, such that the exponential of these repetitions converges to the true model evidence (Neal, 1993; Favaro et al., 2015). Recently, AIS has been connected to the Fluctuation Theorems (FTs) of stochastic thermodynamics, which has allowed a number of variations and further improves the method (Neal, 2001). In addition, the JE connected to FTs is usually used to estimate equilibrium thermodynamic quantities like free energy differences from non-equilibrium thermodynamic processes. So, JE and TI can be used for model evidence estimation since in TI the free-energy difference becomes the log model evidence (Ahlers and Engel, 2008).

2.2 Bayesian model evidence

The particular virtues of Bayesian methods for statistical inference is the ability to perform model selection (Ahlers and Engel, 2008). Recent years have seen interesting advances in the statistical mechanics of processes arbitrarily far from equilibrium, which has given rise to the emerging field of stochastic thermodynamics (Jarzynski, 1997, 2011; **?**; Seifert, 2012). Central to this field are so-called FTs which, may be used to determine free-energy differences from non-equilibrium trajectories (Pohorille et al., 2010). Due to the close relation between free-energy estimates and the log-evidence calculations, the connection of FTs and JE provides new possibilities for Bayesian inference by allowing both to calculate an estimate value of the marginal likelihood, and a detailed method for error analysis of the resulting approximation in the field of Bayesian data analysis (Ahlers and Engel, 2008; Favaro et al., 2015). In an inference problem, the aspect of non-equilibrium is presented by the use of non-stationary Markov processes, explicitly time dependent, which do not rely on repeated equilibrations (Favaro et al., 2015).

In Bayesian model selection, MCMC methods denote a family of simulation algorithms intended to provide sequences of dependent sample which are marginally distributed according to a distribution of interest (Casella and Robert, 1999; Gilks, 2005). The sequential Monte Carlo sampler is a variant of the Monte Carlo technique which can be employed to sample from complex distributions (Del Moral et al., 2006). It uses sampling and re-sampling techniques in order to efficiently produce samples from a distribution of interest. Without connection to thermodynamic processes, the Monte Carlo methods have been largely used to approximate the model evidence by just trying to sample from a distribution without explicit connection to a formulation in terms of stochastic differential equations. Recent works use the straight Monte-Carlo estimation and TI including some variant of the JE to approximate model evidence.

A stochastic differential equation (SDE) is a differential equation in which one or more of the terms are stochastic processes, resulting in a solution which is also a stochastic process. SDEs are used to model various phenomena such as unstable stock prices or physical systems subject to thermal fluctuations (Arnold, 1974). Therefore, the formulation of the stochastic process also allows other techniques developed in stochastic thermodynamics in addition to the JE, such as an asymptotic analysis of rare events, which is interesting as the convergence of the JE relies on sampling of rare events.

In the next chapter, we establish the link between the thermodynamic process, the JE and the evidence.

3. Mathematical model

In this chapter, we explain the complexity of solving the model selection problem with high-dimensional parameter space, and we present methods and mathematical tools used to solve this issue.

3.1 Problem formulation

3.1.1 Description.

Suppose we are given data d, and our task is to come up with a model $\mathcal{M}$ that involves some parameter x. Here, the model refers to a probability distribution over the data d. This model will be given in terms of the data likelihood $\mathcal{P}_{li}(d|x,\mathcal{M})$, which clearly depends on the model $\mathcal{M}$ with parameter x. In usual maximum-likelihood estimation (MLE), we would decide for some model $\mathcal{M}_0$ and maximize $\mathcal{P}_{li}(d|x,\mathcal{M}_0)$ on x in order to find the parameter that best describes the data. However, to find the best model $\mathcal{M}$, MLE would not work since $\mathcal{M}$ is a non-numeric variable. In this chapter, we present a way to select the best working model $\mathcal{M}$ for given data d using the Bayesian approach.

3.1.2 Bayesian model selection.

Bayesian model selection uses the rules of probability theory and can be applied to situations where we have multiple competing models and need to select the best model. In this work, we consider the probability density (or mass) function (PDF) of a parameter x being the true value, for given data d and model $\mathcal{M}$. According to Bayes's theorem, the model's posterior probability applied on the parameter x is given by

$$\mathcal{P}(x|d,\mathcal{M}) = \frac{\mathcal{P}_{li}(d|x,\mathcal{M}) \cdot \mathcal{P}_{pr}(x|\mathcal{M})}{q(d|\mathcal{M})}, \tag{3.1.1}$$

where $\mathcal{P}(x|d,\mathcal{M})$ is the posterior probability of parameter x given data d and model $\mathcal{M}$, $\mathcal{P}_{li}(d|x,\mathcal{M})$ denotes the likelihood to observe d given x and $\mathcal{M}$. The prior knowledge about the parameters x is encoded by $\mathcal{P}_{pr}(x|\mathcal{M})$, whereas $q(d|\mathcal{M})$ stands for the normalization factor also known as the *model evidence* (or marginal likelihood, or prior predictive value).

To estimate x for a given $\mathcal{M}$ and d, we would maximize $\mathcal{P}(x|d,\mathcal{M})$, which is therefore known as maximum posterior (MAP) estimation and it is equivalent to MLE for constant prior (since with constant prior, the posterior is proportional to the likelihood). Since the parameters depend on the model, different models $\mathcal{M}$ will lead to different estimate values for x, and we want to choose the best model. As a starting point, we again use the Bayes theorem this time at the model level. Applied on the model, the *Bayes theorem* is given by

$$\mathcal{P}(\mathcal{M}|d) = \frac{\mathcal{P}_{li}(d|\mathcal{M}) \cdot \mathcal{P}_{pr}(\mathcal{M})}{q(d)}, \tag{3.1.2}$$

with

$$q(d) = \sum_{\mathcal{M}} \mathcal{P}_{li}(d|\mathcal{M}) \cdot \mathcal{P}_{pr}(\mathcal{M}), \tag{3.1.3}$$

From (3.1.2), when we ignore the normalization factor due to the uncertainty on candidate models, the *Bayes theorem* applied on the model becomes

$$\mathcal{P}(\mathcal{M}|d) \sim \mathcal{P}_{li}(d|\mathcal{M}) \cdot \mathcal{P}_{pr}(\mathcal{M}). \tag{3.1.4}$$

The best model would maximize the *model-posterior* $\mathcal{P}(\mathcal{M}|d)$, which corresponds to maximizing the model evidence $\mathcal{P}_{li}(d|\mathcal{M})$ for the constant model-prior. For complex problems with high-dimensional

parameters space, however, it is often virtually impossible to calculate analytically the true value of model evidence which is given by the following integral

$$\mathcal{P}_{li}(d|\mathcal{M}) = \int \mathcal{P}_{li}(d|x,\mathcal{M}) \cdot \mathcal{P}_{pr}(x|\mathcal{M})dx \tag{3.1.5}$$

$$= q(d|\mathcal{M}). \tag{3.1.6}$$

We develop and implement a method based on *Jarzynski's equality* (JE) to approximate the model evidence in Equation (3.1.5) for complex models. In the rest of this work, we present the *JE method* for model evidence followed by an implementation.

3.2 Basic equations

For a good application of Bayesian inference in real life problems, efficient numerical methods are crucial. In addition, normalization factors of distributions such as the model evidence $q(d|\mathcal{M})$ are more difficult to obtain using Monte Carlo methods than the corresponding average (Newman and Barkema, 2006). It is therefore useful to replace the integration in (3.1.5) by some functions of such averages. One way to do it is to use a variant of *thermodynamic integration* (TI) by defining an auxiliary distribution (Kirkwood, 1935; Favaro et al., 2015).

3.2.1 Thermodynamic integration.

Approximation of the model evidence is well known to be challenging. A standard approach is based on TI, but a key concern is that the thermodynamic integral can suffer from high variability (we can get very different results when repeating the same process) in many applications (Oates et al., 2016).

For this approximation, we define the following auxiliary distribution :

$$\mathcal{P}_\beta(x) = \frac{1}{Z(\beta)}\mathcal{P}_{li}^\beta(x) \cdot \mathcal{P}_{pr}(x), \quad \text{where} \quad \beta \in [0,1], \tag{3.2.1}$$

and the normalization factor $Z(\beta)$ is defined by :

$$Z(\beta) = \int \mathcal{P}_{li}^\beta(x) \cdot \mathcal{P}_{pr}(x)dx \tag{3.2.2}$$

where $\mathcal{P}_\beta(x)$ changes from the prior to the product of prior and likelihood which is the integral in (3.1.5) as β changes from 0 to 1. Due to the normalization of the prior distribution we have $Z(0) = 1$ and $Z(1) = q(d|\mathcal{M})$, which is the desired evidence.

Moreover, the connection to the model evidence is given by

$$\begin{aligned} \ln q(d|\mathcal{M}) &= \ln Z(1), \\ &= \ln Z(1) - \ln Z(0), \\ &= \int_0^1 \frac{\partial}{\partial\beta} \ln Z(\beta)d\beta, \\ &= \int_0^1 \frac{Z'(\beta)}{Z(\beta)}d\beta. \end{aligned} \tag{3.2.3}$$

From (3.2.2), $Z(\beta)$ is given by

$$Z(\beta) = \int e^{\beta \ln(\mathcal{P}_{li}(x))} \cdot \mathcal{P}_{pr}(x)dx, \tag{3.2.4}$$

and the derivative of $Z(\beta)$ with respect to β is given by

$$Z'(\beta) = \int \ln(\mathcal{P}_{li}(x)) \cdot \mathcal{P}_{li}^{\beta}(x) \cdot \mathcal{P}_{pr}(x) dx. \tag{3.2.5}$$

The substitution of (3.2.5) into (3.2.3) yields

$$\ln q(d|\mathcal{M}) = \int_0^1 \frac{1}{Z(\beta)} \left(\int \ln(\mathcal{P}_{li}(x)) \mathcal{P}_{li}^{\beta}(x) \mathcal{P}_{pr}(x) dx \right) d\beta. \tag{3.2.6}$$

Using the standard average sample rule

$$\langle g(x) \rangle_{p(x)} = \int g(x) p(x) dx, \tag{3.2.7}$$

where x denotes a continuous random variable, p is the distribution for x, and g stands for any continuous function.

Equation (3.2.6) becomes, with (3.2.2),

$$\ln q(d|\mathcal{M}) = \int_0^1 \langle \ln(\mathcal{P}_{li}(x)) \rangle_{p_\beta(x)} d\beta. \tag{3.2.8}$$

The result (3.2.8) turns the high-dimensional integral defined in (3.1.5) into a one-dimensional integral, and gives the model evidence in term of the logarithmic function. This result is also preferred for direct use of posterior $\mathcal{P}(x|d, \mathcal{M})$ due to small values typically occurring for likelihood $\mathcal{P}_{li}(d|x, \mathcal{M})$. In practical applications, we choose a sufficient number of β-values from the interval $[0, 1]$. Therefore, enough x-values must be sampled from the auxiliary distribution $\mathcal{P}_\beta(x)$ to ensure a robust ensemble average for each value of β. In the next subsection, we present how we used a stochastic thermodynamic process and the JE to derive the model evidence.

3.2.2 Stochastic thermodynamics.

Stochastic Differential Equations (SDEs) are used to describe fluctuating thermodynamic processes (Arnold, 1974). A SDE can be written as follows

$$\dot{x}(t) = f(x(t), t) + \sigma \xi(t), \tag{3.2.9}$$

where $x(t)$ is a stochastic process, $\dot{x}$ denotes the time derivative $\frac{dx(t)}{dt}$, $f(x, t)$ is a force depending on position x at time t, $\xi(t)$ is the Gaussian noise, and σ defines the magnitude of the noise term.

The type of SDE used above is also known as *the Langevin equation*. Due to its stochastic nature, the stochastic process $x(t)$, solutions of the SDE (3.2.9) are always different. Moreover, the distribution of x at a time t is unique and fully described by a PDF $p(x, t)$ which changes with time ($p(x) = p(x, t)$).

The PDF $p(x, t)$ obeys the PDE known as the *Fokker-Planck equation* defined by

$$\dot{p}(x, t) = -\nabla[f(x, t) \cdot p(x, t)] + \frac{\sigma^2}{2} \nabla p(x, t), \tag{3.2.10}$$

where operator ∇ denotes the derivative, $\dot{p}(x, t)$ describes how this $p(x, t)$ evolves in time. The stationary solution of (3.2.10) is given by

$$p_0(x, t) = \frac{1}{Z(t)} e^{-\frac{2}{\sigma^2} V(x,t)}, \tag{3.2.11}$$

where $V(x, t)$ denotes the potential for an arbitrary point in time such that the force is the gradient and defined by

$$f(x, t) = -\nabla V(x, t). \tag{3.2.12}$$

We note that the process x is said to be in equilibrium, if the statistics of $x(t)$ follows the stationary distribution $p_0(x, t)$ at all times.

3.2.3 Jarzynski's Equality for model evidence.

The JE is an equation in statistical mechanics that relates free energy differences between two states and the irreversible work along an ensemble of trajectories joining the same states (Jarzynski, 1997). For a process to be in equilibrium, the trajectory must be infinitely slow ($T \longrightarrow \infty$, where T is the duration of process). The TI method introduced above is equivalent to an equilibrium stochastic process, where β changes with time from $\beta(0) = 0$ at $t = 0$ to $\beta(t) = 1$ at $t = T$. By finding an appropriate thermodynamic process, the JE can be used to approximate the model evidence.

The auxiliary distribution $\mathcal{P}_\beta(x)$ in (3.2.1) can be rewritten in the form of the stationary solution (3.2.11), of the *Fokker-Planck equation* by using the following step

$$\begin{aligned}\mathcal{P}_\beta(x) &= \frac{1}{Z(\beta)}\mathcal{P}_{li}^{\beta}(x)\cdot\mathcal{P}_{pr}(x),\\ &= \frac{1}{Z(\beta)}e^{\beta\ln\mathcal{P}_{li}(x)+\ln\mathcal{P}_{pr}(x)},\\ &= \frac{1}{Z(\beta)}e^{-\varphi(x,t)},\end{aligned} \tag{3.2.13}$$

with

$$\varphi(x(t),t) = -\beta(t)\ln\mathcal{P}_{li}(x) - \ln\mathcal{P}_{pr}(x), \tag{3.2.14}$$

where φ must define a stationary distribution.

Compared to the stationary solution of the *Fokker-Planck equation*, the stochastic process $x(t)$ defined in (3.2.9) is given such that we have

$$-\frac{\sigma^2}{2}\nabla\varphi(x,t) = f(x,t). \tag{3.2.15}$$

Substitution of $\nabla\varphi(x,t)$ in (3.2.15) yields

$$\frac{2f(x,t)}{\sigma(t)^2} = \beta(t)\frac{p'_{li}(x)}{p_{li}(x)} + \frac{p'_{pr}(x)}{p_{pr}(x)}. \tag{3.2.16}$$

Equation (3.2.16) indicates the possibility of letting the noise magnitude $\sigma(t)^2$ to depend on time to realize the time-dependence through $\beta(t)$.

An analogue equation of the JE for model evidence is given by

$$q(d|\mathcal{M}) = \langle e^{-R[x(\cdot)]}\rangle_{x(\cdot)}, \tag{3.2.17}$$

where $x(\cdot)$ denotes a complete stochastic trajectory sampled from the stochastic process while the average is over many realizations of $x(\cdot)$.

The *functional* $R[x(\cdot)]$ in (3.2.17) is defined by

$$R[x(\cdot)] = \int_0^T \frac{\partial}{\partial t}\varphi(x(t),t)dt. \tag{3.2.18}$$

Since (3.2.16) links both $f(x,t)$ and σ, this gives some freedom in defining the stochastic process. In general, from (3.2.16), the stochastic process is defined by

$$f(x,t) = \beta(t)^\gamma\frac{p'_{li}(x)}{p_{li}(x)} + \beta(t)^{\gamma-1}\frac{p'_{pr}(x)}{p_{pr}(x)}, \tag{3.2.19}$$

with

$$\sigma^2 = 2\beta(t)^{\gamma-1}, \tag{3.2.20}$$

where σ denotes the noise terms, f is the force, and γ is an arbitrary parameter.

To exclude divergence for $t = 0$, we require γ to be greater or equal to one ($\gamma \geq 1$). The JE described in (3.2.17) requires the stochastic process defined in (3.2.9) to start in equilibrium (the initial value x_0 at $t = 0$ must be sampled from the prior distribution $\mathcal{P}_{pr}$).

From (3.2.20) and (3.2.19), one open question is the choice of protocol $\beta(t)$. Therefore, $\beta(t)$ is free to choose as long as it meets requirements below

$$\beta(t) = \begin{cases} 0 & \text{at} \quad t = 0, \\ 1 & \text{at} \quad t = T. \end{cases} \tag{3.2.21}$$

Any parameter $\gamma \geq 1$, and protocol $\beta(t)$ such that $\beta(0) = 0$ and $\beta(T) = 1$ would work, but would probably perform differently. Therefore, finding a well performing value for γ is another interesting open question for the Jarzynski method. In this work, we discuss different protocols $\beta(t)$, and parameters γ.

Note that fixed time step length Δt and a large value of T leads to equilibrium, for which we have

$$\lim_{T \longrightarrow \infty} R[x(\cdot)] = -\ln q(d|\mathcal{M}). \tag{3.2.22}$$

In this limit, all protocols should be equivalent. However, some protocols would give robust results for moderate T. Since JE for model evidence in (3.2.17) involves a stochastic process, we use the Euler scheme for differential equations as described by Mannella (2002) to simulate the corresponding SDE in (3.2.9), where the lowest order scheme is sufficient.

3.2.4 Euler scheme for stochastic differential equation.

Stochastic processes are used to model a variety of different physical situations. The most common situation is that the "solution" of some quantities connected to the stochastic model cannot be found theoretically. In this case, a solution is to simulate it on a computer, also called digital simulation. (Mannella, 2002).

The SDE considered by *Mannella* has the generic form

$$\dot{x}_i = f_i(\vec{x}, t) + \sigma \cdot \xi(t), \quad \langle \xi(t) = 0 \rangle, \quad \langle \xi(t)\xi(s) = \delta(t - s) \rangle, \tag{3.2.23}$$

where we assume that the stochastic process ξ is Gaussian and that only one stochastic force is present. Restricting the SDE to a one dimensional model, the new equation has the form

$$\dot{x} = f(x, t) + \sigma \cdot \xi(t). \tag{3.2.24}$$

After expanding $f(x, t)$ in x, using the Gaussian properties of ξ for the stochastic integral, and assuming that the time step h is sufficiently small, we arrive at the following *Euler scheme*

$$x_{n+1} = x_n + \sigma\sqrt{h}Y + f(x_n, t)h, \tag{3.2.25}$$

where Y is a stochastic Gaussian variable with average 0, and standard deviation 1.

The contributions of up to linear order in h are collected recursively.

In the following section, we assume that the likelihood and the prior follow a *Gaussian or normal distribution*, and we implement *Jarzynski's equality* method to compute a numerical approximation of the model evidence.

3.3 Model evidence for univariate Gaussian distribution

3.3.1 Generalities.

In probability theory, a normal or *Gaussian* distribution is a type of continuous probability distribution for a real-valued random variable (Altman and Bland, 1995). In statistics, the normal distribution is the most important probability distribution because it fits many natural phenomena. Normal distributions are also used in social and natural sciences to represent real-valued random variables whose distributions are unknown (Ziegel, 2002; Lyon, 2014). Their importance is partly due to the central limit theorem (CLT). According to the CLT, when the number of samples increases, the average of many samples of a random variable with finite mean and variance is equivalent to a random variable whose distribution converges to a normal distribution.

The general formula for the probability density function of the Gaussian distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{3.3.1}$$

where μ is the location parameter, also known as the mean or expectation of the distribution, while σ is the scale parameter also known as standard deviation. The variance of the distribution is σ^2.

3.3.2 Univariate Gaussian distribution model.

In statistics, univariate distribution is a probability distribution of only one random variable (Tong, 2012). Univariate Gaussian distribution is a Gaussian distribution with one variable. When a distribution has a single peak or mode it is called a unimodal distribution (Medgyessy, 1972).

We consider a unimodal univariate Gaussian model on which we have one data point d and one parameter x. The data likelihood distribution is given by

$$\mathcal{P}_{li}(d|x) = \frac{1}{\sqrt{2\pi}\nu_{li}} \exp\left(-\frac{(d-x)^2}{2\nu_{li}^2}\right), \tag{3.3.2}$$

with fixed variance $\sigma^2 = \nu_{li}^2$ and unknown mean x as parameter. The prior distribution is given by

$$\mathcal{P}_{pr}(x) = \frac{1}{\sqrt{2\pi}\nu_{pr}} \exp\left(-\frac{1}{2\nu_{pr}^2}x^2\right), \tag{3.3.3}$$

with mean zero and standard deviation ν_{pr}.

In general, when the prior and likelihood distributions of the mean parameter follow a *Gaussian* distribution, the model evidence $q(d|\mathcal{M})$ can be computed analytically from its definition in (3.1.5), by using the integral of the *Gaussian* function.

For the simulation of the SDE, it is convenient to write the likelihood and the prior distribution as

$$\mathcal{P}_{li}(x) = \frac{1}{Z_{li}} e^{-V(x)}, \tag{3.3.4}$$

$$\mathcal{P}_{pr}(x) = \frac{1}{Z_{pr}} e^{-U(x)}, \tag{3.3.5}$$

where

$$V(x) = \frac{(d-x)^2}{2\nu_{li}^2}, \quad \text{and} \quad Z_{li} = \sqrt{2\pi}\nu_{li}; \tag{3.3.6}$$

$$U(x) = \frac{x^2}{2\nu_{pr}^2}, \qquad \text{and} \quad Z_{pr} = \sqrt{2\pi}\nu_{pr}. \tag{3.3.7}$$

On one hand, by substituting (3.3.4) and (3.3.5) in (3.2.19), we can easily verify that the general Equation (3.2.20) becomes

$$f(x,t) = -\beta(t)^{\gamma} V'(x) - \beta(t)^{\gamma-1} U'(x), \tag{3.3.8}$$

where the noise terms is given by (3.2.20), with

$$V'(x) = -(d-x) \text{ and } U'(x) = x.$$

On the other hand, by substituting (3.2.14) in (3.2.18), we obtain

$$R[x(\cdot)] = \int_0^T \dot{\beta}(t) V\left(x(t)\right) dt + \ln Z_{li}. \tag{3.3.9}$$

Proof. From Equations (3.2.18) and (3.2.14), we have

$$\begin{aligned} R[x(\cdot)] &= \int_0^T \frac{\partial}{\partial t} \varphi(x(t),t) dt, \\ &= -\int_0^T \dot{\beta}(t) \ln \mathcal{P}_{li}(x(t)) dt, \\ &= \int_0^T \dot{\beta}(t) V(x) dt + \int_0^T \dot{\beta}(t) \ln Z_{li} dt, \\ &= \int_0^T \dot{\beta}(t) V(x(t)) dt + \ln Z_{li}. \end{aligned}$$

Note that the integral of R is evaluated along the trajectory $x(t)$. In the general case, whenever the likelihood distribution $P_{li}(x)$ can be rewritten using $V(x)$ and Z_{li} as shown in (3.3.4), Equation (3.2.18) becomes like (3.3.9) and the model evidence is computed by running the stochastic process on (3.2.18), taking the average of several trajectories.

3.4 Model evidence generalization for mutivariate Gaussian model

In statistics and probability theory, multivariate Gaussian distribution, or joint normal distribution is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions (Tong, 2012). The multivariate normal distribution is often used to describe, at least approximately, any set of (possibly) correlated real-valued random variables each of which clusters around a mean value (Tong, 2012).

3.4.1 Prior and likelihood function.

To generalize the model for unimodal multivariate Gaussian distribution, we assume a higher dimension of M parameters, and N data points for each parameter as follows

$$\vec{x} = (x_1, x_2, \ldots, x_M), \quad d = \begin{bmatrix} d_{11} & \ldots & d_{1M} \\ \vdots & \ddots & \vdots \\ d_{N1} & \ldots & d_{NM} \end{bmatrix}.$$

Under this generalization, the likelihood distribution becomes

$$\mathcal{P}_{li}(\vec{x}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi v_{li}^2}} \exp\left(-\frac{1}{2v_{li}^2}(d_{ij} - x_j)^2\right), \tag{3.4.1}$$

$$= (2\pi v_{li}^2)^{-\frac{NM}{2}} \exp\left(-\frac{1}{2v_{li}^2} \sum_{j=1}^{N} \sum_{j=1}^{M} (d_{ij} - x_j)^2\right). \tag{3.4.2}$$

For easy manipulation, we can rewrite this likelihood function as

$$\mathcal{P}_{li}(\vec{x}) = \frac{1}{Z_{li}} e^{-V(\vec{x})}, \tag{3.4.3}$$

where

$$V(\vec{x}) = \frac{1}{2v_{li}^2} \sum_{j=1}^{N} \sum_{j=1}^{M} (d_{ij} - x_j)^2, \text{ and } Z_{li} = (2\pi v_{li}^2)^{-\frac{NM}{2}}. \tag{3.4.4}$$

For a multivariate Gaussian distribution, the prior distribution becomes

$$\mathcal{P}_{pr}(x) = \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi v_{pr}^2}} \exp\left(-\frac{1}{2v_{pr}^2} x_j^2\right), \tag{3.4.5}$$

$$= (2\pi v_{pr}^2)^{-\frac{M}{2}} \exp\left(-\frac{1}{2v_{pr}^2} \sum_{j=1}^{M} x_j^2\right). \tag{3.4.6}$$

We can also rewrite this prior distribution as

$$\mathcal{P}_{pr}(\vec{x}) = \frac{1}{Z_{pr}} e^{-U(\vec{x})}, \tag{3.4.7}$$

where

$$U(\vec{x}) = \frac{1}{2v_{pr}^2} \sum_{j=1}^{M} x_j^2, \text{ and } Z_{pr} = (2\pi v_{pr}^2)^{-\frac{M}{2}}. \tag{3.4.8}$$

3.4.2 Model evidence.

The generalization of the model evidence on a higher dimensional Gaussian model turns the general Equation (3.2.19) into

$$\vec{f}(\vec{x}, t) = -\beta(t)^{\gamma} \nabla V(\vec{x}) - \beta(t)^{\gamma-1} \nabla U(\vec{x}), \tag{3.4.9}$$

where the force $\vec{f}$ becomes a M-dimensional vector, and ∇ operator denotes the gradient. The noise terms σ is given by (3.2.20).

For any $k \in \{1, \ldots, M\}$, the computation of $\nabla V(\vec{x})$ is as follows

$$\begin{aligned}
[\nabla V(\vec{x})]_k &= \frac{1}{2v_{li}^2} \frac{\partial}{\partial x_k} \sum_{i=1}^{N} \sum_{j=1}^{M} (d_{ij} - x_j)^2, \\
&= \frac{1}{2v_{li}^2} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\partial}{\partial x_k} (d_{ij} - x_j)^2, \\
&= -\frac{1}{v_{li}^2} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\partial x_j}{\partial x_k} (d_{ij} - x_j), \\
&= -\frac{1}{v_{li}^2} \sum_{i=1}^{N} \sum_{j=1}^{M} \delta_{jk} (d_{ij} - x_j), \\
&= -\frac{1}{v_{li}^2} \sum_{i=1}^{N} (d_{ik} - x_k),
\end{aligned} \tag{3.4.10}$$

where δ_{jk} denotes the *Kronecker symbol* which is 1 if the variables are equal, and 0 otherwise, that is

$$\delta_{jk} = \begin{cases} 0 & \text{if} \quad j \neq k, \\ 1 & \text{if} \quad j = k. \end{cases} \tag{3.4.11}$$

For any $k \in \{1, \ldots, M\}$, the computation of $\nabla U(\vec{x})$ is similarly

$$\begin{aligned} [\nabla U(\vec{x})]_k &= \frac{1}{2v_{pr}^2} \frac{\partial}{\partial x_k} \sum_{j=1}^{M} x_j^2, \quad \text{with } k \in \{1, \ldots, M\} \\ &= \frac{1}{2v_{pr}^2} \sum_{j=1}^{M} \frac{\partial}{\partial x_k} x_j^2, \\ &= \frac{1}{v_{pr}^2} \sum_{j=1}^{M} \delta_{kj} x_j, \\ &= \frac{1}{v_{pr}^2} x_k. \end{aligned} \tag{3.4.12}$$

By substituting (3.4.10) and (3.4.12) into (3.2.19), the new expression of the force is given by

$$f_k(\vec{x}, t) = \beta(t)^{\gamma} \frac{1}{v_{li}^2} \sum_{i=1}^{N} (d_{ik} - x_k) - \beta(t)^{\gamma - 1} \frac{1}{v_{pr}^2} x_k, \tag{3.4.13}$$

and from (3.2.25), the stochastic process described by f_k becomes

$$\vec{x}_{n+1,k} = \vec{x}_{n,k} + \sigma Y \sqrt{h} + f_k(\vec{x}_{n,k}, t)h. \tag{3.4.14}$$

The R value remains a scalar and becomes

$$R[\vec{x}(\cdot)] = \int_0^T \dot{\beta}(t) V(\vec{x(t)}) dt + \ln Z_{li}. \tag{3.4.15}$$

3.4.3 Analytical value for model evidence.

One advantage of using a Gaussian model as example is that Gaussian intregration provides a way to compute the high dimensional integral in (3.1.5) which stands for the true value of the model evidence. Having this analytical value for the model evidence will allow to compare it with the estimation value given by JE. For a multivariate Gaussian model with M parameters and N data points, the analytic value of the model evidence is given by

$$q(d|\mathcal{M}) = \left(\frac{v_l^2}{(2\pi v_l^2)^N . (N v_p^2 + v_l^2)} \right)^{\frac{M}{2}} \cdot \exp\left(\frac{1}{2v_l^2} \sum_{j=1}^{M} \left[\frac{v_p^2}{N v_p^2 + v_l^2} \left(\sum_{i=1}^{N} d_{ij} \right)^2 - \sum_{i=1}^{N} d_{ij}^2 \right] \right), \tag{3.4.16}$$

where v_l and v_p respectively denote the variance of likelihood and prior distribution.

Proof. From Equation (3.1.5), we have

$$\begin{aligned}
q(d|\mathcal{M}) &= \int \mathcal{P}_{li}(d|x,\mathcal{M}) \cdot \mathcal{P}_{pr}(x|\mathcal{M}) dx \\
&= \int \prod_{j=1}^{M} dx_j \left[(2\pi v_l^2)^{-\frac{NM}{2}} \exp\left(-\frac{1}{2v_l^2} \sum_{i=1}^{N}\sum_{j=1}^{M} (d_{ij} - x_j)^2 \right) \cdot (2\pi v_p^2)^{\frac{-M}{2}} \exp\left(-\frac{1}{2v_p^2} \sum_{j=1}^{M} x_j^2 \right) \right] \\
&= \prod_{j=1}^{M} (2\pi v_l^2)^{-\frac{N}{2}} \cdot (2\pi v_p^2)^{-\frac{1}{2}} \int dx_j \left[\exp\left(-\frac{1}{2v_l^2} \sum_{i=1}^{N} (d_{ij} - x_j)^2 - \frac{x_j^2}{2v_p^2} \right) \right] \\
&= \prod_{j=1}^{M} (2\pi v_l^2)^{-\frac{N}{2}} \cdot (2\pi v_p^2)^{-\frac{1}{2}} \int dx_j \left[\exp\left(-\frac{1}{2v_l^2} \sum_{i=1}^{N} (d_{ij}^2 - 2d_{ij}x_j + x_j^2) - \frac{x_j^2}{2v_p^2} \right) \right] \\
&= \prod_{j=1}^{M} (2\pi v_l^2)^{-\frac{N}{2}} \cdot (2\pi v_p^2)^{-\frac{1}{2}} \int dx_j \left[\exp\left(-\left(\frac{N}{2v_l^2} + \frac{1}{2v_p^2} \right) x_j^2 + \frac{\sum_{i=1}^{N} d_{ij}}{v_l^2} x_j - \frac{\sum_{i=1}^{N} d_{ij}^2}{2v_l^2} \right) \right] \\
&= \prod_{j=1}^{M} A \int dx_j \exp\left(-ax_j^2 + bx_j + c \right) \\
&= \prod_{j=1}^{M} A \cdot \sqrt{\frac{\pi}{a}} \exp\left(\frac{b^2}{4a} + c \right).
\end{aligned} \tag{3.4.17}$$

By identification, A, a, b and c are given by

$$A = \frac{1}{(2\pi v_l^2)^{\frac{N}{2}} \cdot (2\pi v_p^2)^{\frac{1}{2}}}, \qquad b = \frac{\sum_{i=1}^{N} d_{ij}}{v_l^2}, \tag{3.4.18}$$

$$a = \frac{N}{2v_l^2} + \frac{1}{2v_p^2}, \qquad c = -\frac{\sum_{i=1}^{N} d_{ij}^2}{2v_l^2}. \tag{3.4.19}$$

By substituting term of Equations (3.4.18) and (3.4.19) into (3.4.17), we obtain

$$\begin{aligned}
q(d|\mathcal{M}) &= \prod_{j=1}^{M} \frac{1}{(2\pi v_l^2)^{\frac{N}{2}} . (2\pi v_p^2)^{\frac{1}{2}}} \cdot \sqrt{\frac{2\pi v_l^2 v_p^2}{N v_p^2 + v_l^2}} \exp\left[\frac{1}{2v_l^2} \frac{v_p^2}{N v_p^2 + v_l^2} \left(\sum_{i=1}^{N} d_{ij} \right)^2 - \frac{1}{2v_l^2} \sum_{i=1}^{N} d_{ij}^2 \right] \\
&= \left(\frac{v_l^2}{(2\pi v_l^2)^N . (N v_p^2 + v_l^2)} \right)^{\frac{M}{2}} \exp\left(\frac{1}{2v_l^2} \sum_{j=1}^{M} \left[\frac{v_p^2}{N v_p^2 + v_l^2} \left(\sum_{i=1}^{N} d_{ij} \right)^2 - \sum_{i=1}^{N} d_{ij}^2 \right] \right).
\end{aligned}$$

Finally, we obtain the same result as in (3.4.16).

In this chapter, we have presented the JE for model evidence. In particular, we explained the difficulty of obtaining analytically the model evidence for high dimensional complex models, and we presented the JE method allowing us to estimate this value. We have also introduced the key concepts necessary to understand how JE method work for the model evidence. We then showed how this JE method can be used with a Gaussian model. So, in the next chapter, we will give an implementation of this method followed by a discussion about some parameters and protocols (mainly γ and β value of (3.2.20) and (3.2.19)).

4. Numerical simulations

In the previous chapter, we presented the *Jarzynski equality* (JE) method to approximate the model evidence. We noticed that Equations (3.2.20) and (3.2.19) involved one key parameter γ and protocol β influencing the convergence of the JE method. In this chapter, we implement *Jarzynski's equality* method for the model evidence by taking different protocols for β, and parameters for γ. In this implementation, we assume that the data come from an unimodal multivariate Gaussian distribution.

4.1 Protocols description

With fixed time step length, when the duration T of the stochastic process defined in (3.2.9) becomes large ($T \longrightarrow \infty$), $R[x(\cdot)]$ described in (3.2.18) averages to the log-model evidence. In this case, all protocols and parameters are equivalent. However, with the restriction of computing resources, it becomes more difficult to simulate the model evidence with large value of T. In this case, to face this issue, one solution is to use the protocol $\beta(t)$ and parameter γ which gives the best approximation for moderate T.

For this implementation, we recall, that the stochastic processes used for model evidence are described in (3.2.9), where the general form of the force f is given in (3.2.20).

4.1.1 Gamma parameter.

We investigated three different γ parameter values, $\gamma = 1, 2, 3$. For all these values, the general equation described in (3.2.20) becomes

$$f(x,t) = -\beta(t)^{\gamma}\nabla V(\vec{X}) - \beta(t)^{\gamma-1}U(\vec{X}), \tag{4.1.1}$$

where the noise term is given by

$$\sigma^2 = \sqrt{2\beta(t)^{\gamma-1}}. \tag{4.1.2}$$

For $\gamma = 1$, the noise term is constant whereas for $\gamma > 1$, it becomes time-dependent. Each modification occurring on β or γ yields to a new stochastic process.

4.1.2 Beta protocol.

We recall that β is free to choose from $\mathbb{R}_+$ to $\mathbb{R}_+$, such that $\beta(0) = 0$ and $\beta(T) = 1$. In this essay, we implement four classes of β functions as shown in Figure 4.1. The *"linear class"* denotes all the β-functions in the general form

$$\beta(t) = at, \quad \text{where} \quad a \in \mathbb{R}_+^* \quad \text{and} \quad t \in [0, T]. \tag{4.1.3}$$

For small $\beta(t)$ we have the *"slow class"* which denotes all β-functions which increase slowly initially to 1. When $\beta(t)$ increases quickly but hardly overshoots 1, we have the *"just class"*, whereas the *"fast class"* stands for all β-functions which increase fast by clearly overshooting 1 for some values of t.

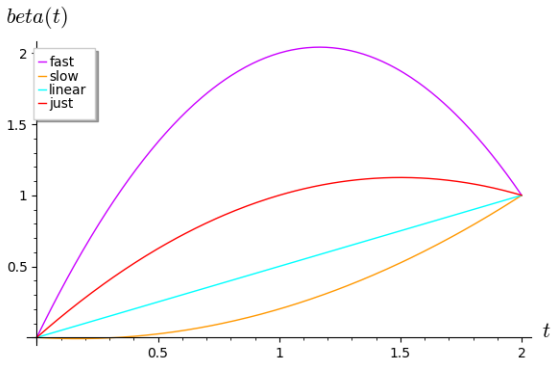


Figure 4.1: β functions.

4.2 Error analysis of the Jarzynski estimator for model evidence

The aim of trying different protocols and parameters is to compare them in order to make a good statement about the best performing choices of the γ parameter, and β protocol. A way to achieve this is by using the convergence plot comparison. Indeed, the idea behind this method is to simulate the stochastic process for many trajectories several times for a given protocol and parameter. Then, for each repetition, derive the model evidence, and plot the solution to get the performance for the given parameter and protocol. However, due to the stochastic nature of processes, we will get different convergence curves every time. So, it might just be by chance, that some protocols and parameter appear to perform better than the others.

Since the model evidence using the JE is approximated by

$$q(d|\mathcal{M}) \approx \langle e^{-R[x(\cdot)]} \rangle, \tag{4.2.1}$$

another way to compare the protocols and parameters is by using the confidence interval. Indeed, Favaro et al. (2015) provide a statistical error analysis method based on confidence intervals for the determination of the model evidence using *Jarzynski's estimation* in a *Bayes* problem. The authors derive a confidence interval for $q(d|\mathcal{M})$ from the central limit theorem. To derive this confidence interval, two main assumptions are made :

1. the sequence $(e^{-R_1}, \ldots, e^{-R_N})$ of the N independent random variables results from the same distribution,
2. the variance ς^2 of that distribution is finite while the expectation is $q(d|\mathcal{M})$.

Under these assumptions, a total number of N values of R are grouped in one block. For this block, the central limit theorem holds and allows to use the normal distribution to determine the confidence interval.

For the model evidence error analysis, we compute the confidence limits $\hat{D}_{\pm}(N)$ defined by

$$\hat{D}_{\pm}(N) = -\ln\left[1 \pm \sqrt{\frac{2}{N}} \frac{\hat{\zeta}(N) \cdot \mathrm{erf}^{-1}(\alpha)}{\langle e^{-R[x(\cdot)]} \rangle}\right], \tag{4.2.2}$$

where $\hat{\zeta}$ denotes the estimated variance of the block, α is the confidence interval, and erf^{-1} stands to the inverse error function. The mean square error is given by

$$\xi^2(N) = \langle [\ln\langle e^{-R} \rangle_N - \ln q(d|\mathcal{M})]^2 \rangle, \tag{4.2.3}$$

where ξ^2 is the expected squared difference between the *Jarzynski estimator* and the true value of the model evidence. The variance ζ^2 of e^{-R} values is approximated by the sample variance as follows

$$\hat{\zeta}^2(N) = \frac{1}{N-1}\sum_{i=1}^{N}(e^{-R_i} - \langle e^{-R}\rangle_N)^2. \tag{4.2.4}$$

In (4.2.2), the confidence level α can be selected as one deems fit, but the ordinary choices are 95%, 99%, 99.5% and 99.9%.

By using the confident limit defined in (4.2.2), with *Jarzynski's estimator* provided in (4.2.1), the final confidence interval is given by

$$CI_{\pm} = \hat{D}_{\pm}(N) + \langle e^{-R[x(\cdot)]}\rangle. \tag{4.2.5}$$

To choose the best parameter and protocol, we compare the confidence intervals for the different β and γ, the smaller will be the better. To evaluate the estimation, we just use the confidence interval resulting from the error analysis of *Jarzynski's estimator*. More information about this error analysis method of *Jarzynski's estimator* can be found on Favaro et al. (2015).

4.3 General description of the algorithm

For numerical approximation of the model evidence using *Jarzynski's estimator* method, the proposed general algorithm has two main steps. On the first step (Algorithm 1), we compute the R-values (e^{-R}) for many trajectories and on the second step (Algorithm 2), we use these values to derive both the model evidence and the confidence interval for different number of R-values.

The first step is described as follows

Algorithm 1: JARZYNSKI'S ESTIMATOR FOR MODEL EVIDENCE

Data: $time$, $tstep$, $Xval$, std_{li}, std_{pr}, $\beta(t)$, γ, obs_data
Result: $R\text{-}values$

Function SDEs$(Data)$:
- initialization() ;
- **for** $k \leftarrow 0$ **to** $Nstep$ **do**
  - $t \longleftarrow timeVal$;
  - $\nabla V(\vec{X}) = -\frac{1}{v_{li}^2}\sum_{i=1}^{N}(d_{ij} - X_j)$;
  - $\nabla U(\vec{X}) = \frac{1}{v_{pr}^2}X_j$;
  - $f(\vec{X}, t) = -\beta(t)^{\gamma}\cdot\nabla V(\vec{X}) - \beta(t)^{\gamma-1}\cdot\nabla U(\vec{X})$;
  - $\sigma = \sqrt{2\beta(t)^{\gamma-1}}$;
  - $\vec{X} = \vec{X} + \sigma\cdot Y_1\sqrt{h} + f(\vec{X}, t)\cdot h$;
  - save the current value of $\vec{X}$ in $Xsol$ for current t ;
- $V(\vec{X}) = \frac{1}{2v_{li}^2}\sum_{j=1}^{N}\sum_{j=1}^{M}(d_{ij} - \vec{X}sol_j)^2$;
- $\ln Z_{li} = -\frac{NM}{2}\ln(2\pi v_{li}^2)$;
- $R[x(\cdot)] = \int_0^T \dot{\beta}(t)V(\vec{X})dt + \ln Z_{li}$;
- **return** e^{-R}

We recall that Algorithm 1 is used for the multivariate unimodal Gaussian model. In this case, the likelihood and the prior distribution function can be respectively written in terms of $V(\vec{X}), Z_{li}$ and

$U(\vec{X}), Z_{pr}$ as shown in (3.4.3) and (3.4.7). We can also use this algorithm for any other likelihood and prior distribution function which can be rewritten as shown in those equations. However, for other distributions, we just have to adapt the algorithm to the new distribution by replacing the likelihood P_{li}, and prior distribution P_{pr} by the corresponding function, and then compute their respective derivatives P'_{li} and P'_{pr} and plug the new results in the general Equation (3.2.19). Therefore, Algorithm 2 for confidence interval remains the same whatever the likelihood and the prior distributions.

The second step is described as follows

Algorithm 2: CONFIDENCE INTERVAL FOR JARZYNSKI'S ESTIMATOR

Data: $time$, $tstep$, $Xval$, std_{li}, std_{pr}, $\beta(t)$, γ, obs_data
Result: $R\text{-}values$

begin
- $R = SDEs(time, tstep, Xval, std_{li}, std_{pr})$;
- split $lenght(R)$ in N values ;
- **for** $k \in N$ **do**
  - $R_k \longleftarrow$ the first K values of R ;
  - $\zeta^2(K) = \frac{1}{K-1}\sum_{i=1}^{K}(e^{-R_i} - \langle e^{-R}\rangle_K)^2$;
  - $\hat{D}(K)_+ = -\ln\left[1 - \sqrt{\frac{2}{K}}\frac{\hat{\zeta}(K)\cdot erf^{-1}(\alpha)}{\langle e^{-R[x(\cdot)]}}\right]$;
  - $\hat{D}(K)_- = -\ln\left[1 + \sqrt{\frac{2}{K}}\frac{\hat{\zeta}(K)\cdot erf^{-1}(\alpha)}{\langle e^{-R[x(\cdot)]}}\right]$;
  - $CI_+ = \langle e^{-R[x(\cdot)]}\rangle_K + \hat{D}(K)_+$;
  - $CI_- = \langle e^{-R[x(\cdot)]}\rangle_K + \hat{D}(K)_-$;
  - save CI_+, CI_- and $\langle e^{-R[x(\cdot)]}\rangle_K$
- plot the true value of model evidence ;
- plot the model evidence approximation for each number or R-values ;
- plot the confidence interval for each approximate value of model evidence.

4.4 Simulations

4.4.1 Tools and simulation environment.

The *Jarzynski estimator* for model evidence presented in this document has been simulated using the free open source mathematics software $SageMath$. For array manipulation, we used the $Numpy$ library which provides mathematical functions to operate on arrays, and $Matplotlib$ library for plots. All these simulations have been done on a computer with the following configuration: Intel Core i5 7th Gen CPU 2.50GHz, 6 GB RAM on a linux operating system.

4.4.2 Simulation process.

To select the best performing parameter and protocol, we estimated the model evidence by generating randomly 10×5 artificial data from a multivariate unimodal normal distribution with 5 parameters, 0 mean and 0.5 as variance. Since the mean parameters are unknown, we set the variance of the mean in the prior distribution to 2.5 in order to cover a lot of plausible values.

We simulated the process with several trajectories simultaneously and plugged them into the JE which involves an exponential average that leads to a good approximation of the model evidence. We also fixed the duration of the process at $T = 2$ with time step $h = 0.001$. To reduce the uncertainty due to

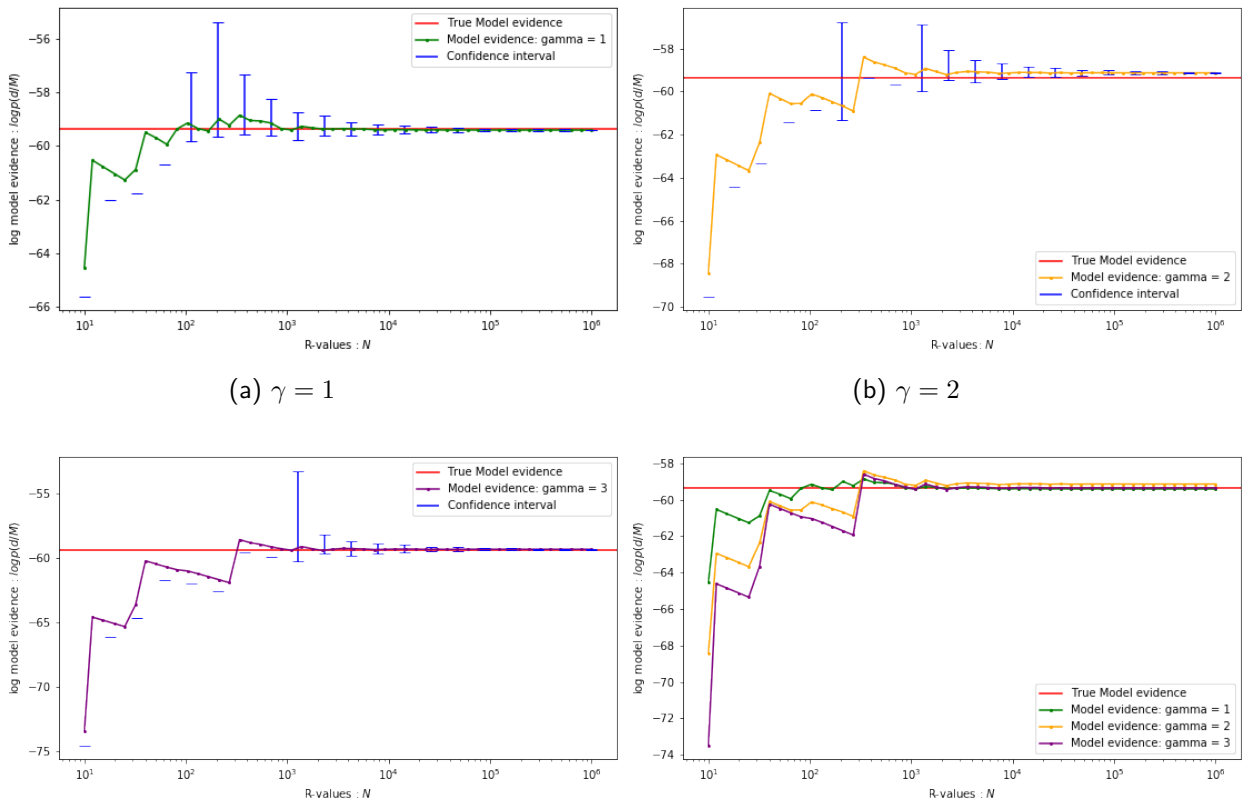
the stochastic process, we used the same random data points and initial parameters like initial random values on the stochastic process for every β protocol and γ parameter. Since each continuous function can be approximated with a polynomial function, we implemented the *Lagrange* interpolation algorithm to get the analytical function corresponding to each protocol given the initial values. Finally, we used the $Numpy$ trapezoidal function to compute the numerical integration on (3.3.9) which leads to the model evidence.

The code for our simulations is free access and available on $GitHub$ through the link (Simulations, 2020)

4.4.3 Results.

As results, we compare the model evidence given by using each protocol and parameter with the true value calculated using analytic expression provided in (3.4.16). The plots in Figures 4.2 and 4.3 show the performance of *Jarzynski's estimator* with various configurations, followed by the error analysis for an increasing number N of considered R values. We therefore use the confidence limits $\hat{D}_{\pm}(N)$ as error bars, which are found to always cover or be close to the analytic result. For a number of R-values, we can notice that there is no upper limit for the confidence interval as shown in the Table 4.2. This reflects the fact that in these cases, the upper limits are infinite due to small sample sizes.

In Figures 4.2d and 4.4, we summarize the performance of each γ parameter, and β protocol. The model evidence and confidence interval values for some number of R-values are summarized for different protocols and parameters in the Tables 4.1 and 4.2.



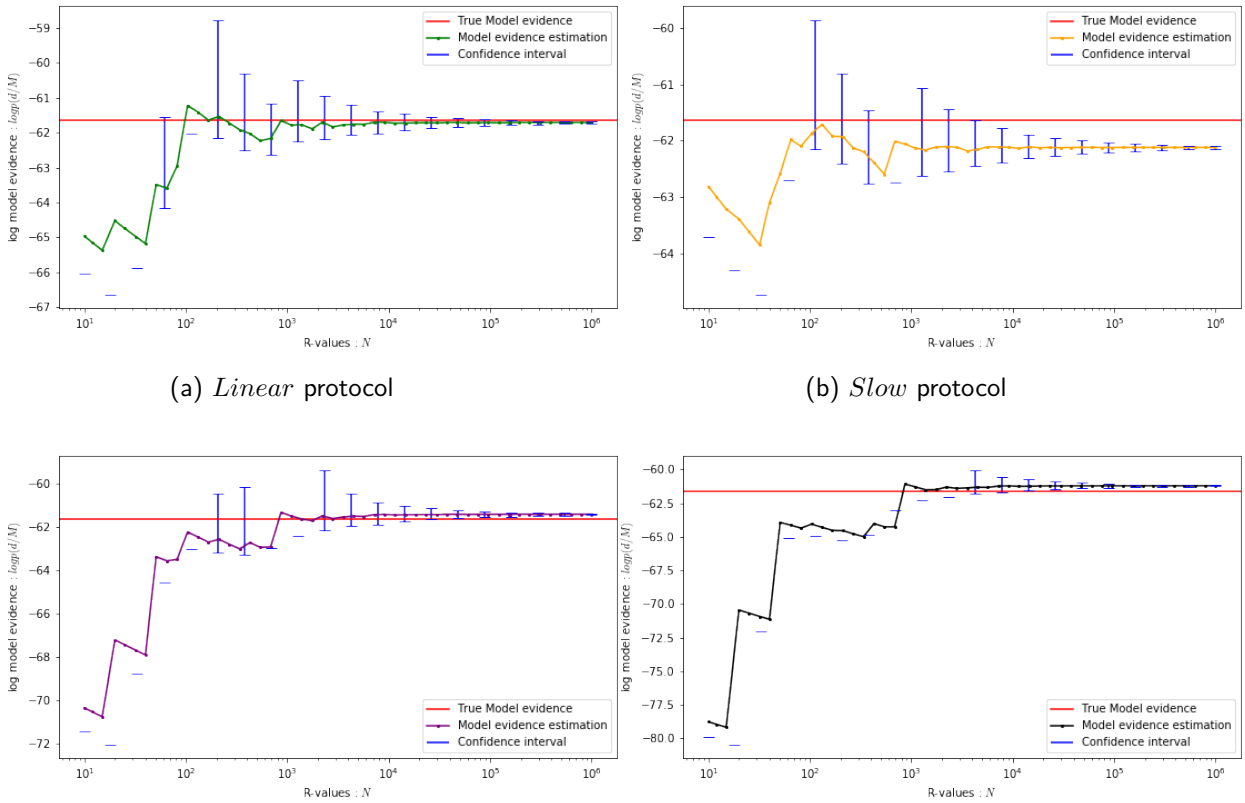
(a) $\gamma = 1$ (b) $\gamma = 2$

(c) $\gamma = 3$ (d) $\gamma = 1, 2, 3$

Figure 4.2: Convergence plot for γ parameter

	$\gamma = \mathbf{1}$	$\gamma = \mathbf{2}$	$\gamma = \mathbf{3}$	N
CI_-	-59.564439	-59.400170	-59.621843	10^3
CI_+	-59.217683	-58.778565	-58.954993	
$\lvert CI_- \rvert - \lvert CI_+ \rvert$	**0.346755**	**0.621604**	**0.666850**	
Estimate value	$\mathbf{-59.406016}$	$\mathbf{-59.136908}$	$\mathbf{-59.343004}$	
True value	$\mathbf{-59.372934}$			
Error	**0.033082**	**0.2360250**	**0.049929**	

Table 4.1: Confidence interval of the log-model evidence ($\log q(d|\mathcal{M})$) for $gamma$ parameter.



(a) $Linear$ protocol

(b) $Slow$ protocol

(c) $Just$ protocol

(d) $Overshoot$ protocol

Figure 4.3: Convergence plot for β protocol

	Linear	**Slow**	**Just**	**Overshoot**	N
CI_-	-62.336735	-62.711384	-62.256695	-62.082912	10^3
CI_+	-59.702955	-60.468171			
$\|CI_-\| - \|CI_+\|$	**2.633780**	**2.243213**			
Estimate value	$\mathbf{-61.712934}$	$\mathbf{-62.119092}$	$\mathbf{-61.425976}$	$\mathbf{-61.193843}$	
CI_-	-61.739938	-62.144314	-61.466090	-61.238136	10^6
CI_+	-61.685180	-62.09321	-61.384184	-61.147497	
$\|CI_-\| - \|CI_+\|$	**0.054757**	**0.051096**	**0.08190590**	**0.090639**	
Estimate value	$\mathbf{-61.712934}$	$\mathbf{-62.119092}$	$\mathbf{-61.425976}$	$\mathbf{-61.193843}$	
True value	$\mathbf{-61.636547}$				
Error	**0.0763865**	**0.482545**	**0.210571**	**0.442704**	

Table 4.2: Confidence interval for the *log-model evidence* with β protocol

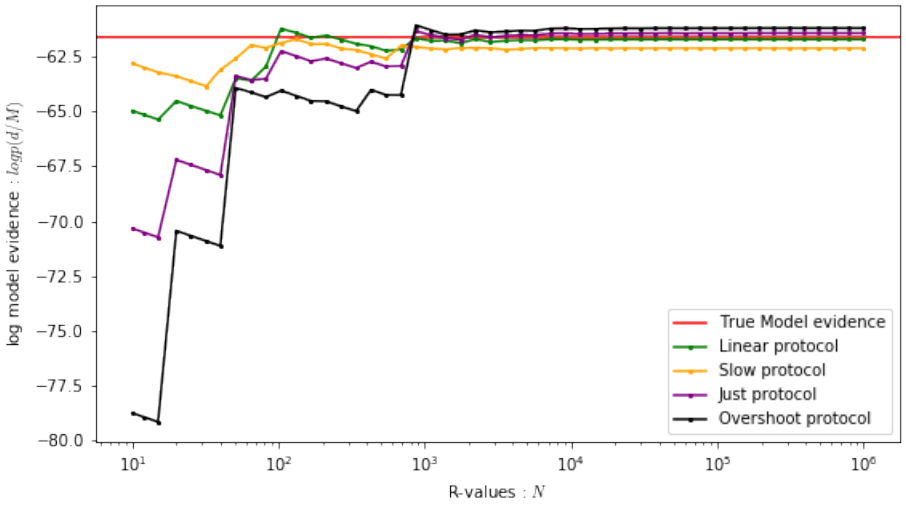


Figure 4.4: β protocol : summary.

4.5 Discussion

From the results of the simulations, we notice that for all parameters and protocols, the estimate values of model evidence converge close to the true value. Up to $N = 10^3$ trajectories, the method converges to a constant evidence value for each γ parameter. The convergence plot in Figure 4.2d clearly shows that for small number of trajectories or R values (before the convergence), 1 is the best value for the parameter γ at almost each point. For these small values of trajectories, we can also notice that the increase of γ leads to a less accurate approximation from the true value. For a large number of trajectories $N > 10^3$, the estimate value is almost constant and for each parameter we observe that the estimate becomes linear and more close to the true value. However, from confidence interval on Table 4.1, we see that $\gamma = 1$ gives the best approximation both in terms of confident interval and estimate error define as the absolute difference between the analytic and the estimate value. After many

simulations, we made sure that this result is not random due to the stochastic process, and that $\gamma = 1$ is effectively the best value for γ parameter in terms of the smallest confidence interval and smallest error compared to the analytical value. Moreover, we notice that with $\gamma = 1$ the confidence interval always contains the analytical value which is not always the case with other values for γ.

The plots in Figure 4.3 show that each protocol leads to a valid approximation of the model evidence. However, we realize that for a small number of trajectories, the confidence interval provided by each protocol contains the true value. We can also notice that the increase in the number of trajectories leads to convergence and reduces the confidence interval. When the convergence occurs (with large trajectory numbers), the estimated value is closer to the analytic value but outside the confidence interval. From the Table 4.2, we observe that the slow protocol gives the best approximation with respect to confidence interval, while the *linear protocol* gives the best approximation in terms of the estimate error. We can also see that this statement is the same with both small and large numbers of trajectories. After many simulations, we ensured again that these results are not by chance due to the stochastic nature of the process. This brings us to draw the conclusion that while the *slow protocol* is the one which leads to the smallest confidence interval but not the closest to the true value, the *linear protocol* is the best choice as it gives an estimate value closest to the true value with still a reasonably narrow confidence interval.

From Figures 4.2 and 4.3, and Tables 4.1 and 4.2 we noticed that for a large number of trajectories, the confidence interval given by some β protocols and γ parameters does not contain the true value of the model evidence. This result can be explained by the fact that the confidence intervals are estimates themselves, and as such may suffer from the same bias as the model evidence estimate. However, as the JE holds exactly, the bias should vanish by eliminating possible sources of error.

Finally, we noticed that the model evidence provided by the JE method differs slightly from the analytical value. This difference could be explained on the one hand by the error due to the *Euler* scheme used to approach the solution of SDE defined on 3.2.23 and on the other hand by the error resulting from the trapezoidal scheme used to estimate the integral defined on 3.2.18.

5. Conclusion and future work

5.1 Conclusion

The model evidence is considered by Bayesian statisticians as a good standard for model selection. However, in many applications such as in deep neural networks, the computation of the model evidence often takes place in a very high dimensional parameter space for which analytical integration is impossible. In these cases, efficient Markov Chain Monte Carlo algorithms connected to thermodynamic integration schemes enable efficient computation of the Bayesian model evidence. However, these computations become very expensive, or even infeasible in case of multimodality.

The aim of this essay was to estimate the model evidence by using the high dimensional stochastic process, and the *Jarzynski Equality*. To do this, we used that in thermodynamic integration, the free-energy difference becomes the log model evidence. This allowed us to link *Jarzynski's Equality* and stochastic processes to estimate the model evidence from non-equilibrium thermodynamic processes. We realized that the stochastic differential equation (*Langevin equation*) used to compute the stochastic process for this approximation involved both a parameter γ and a protocol β influencing the convergence of the *Jarzynski Equality* method for model evidence estimation. To validate the method, we considered a multivariate unimodal Gaussian distribution for both likelihood and prior distribution for which we could compute the analytical value of the model evidence, and we implemented the Jarzynski estimation method to approximate the model evidence by taking different protocols and parameters. We also used the confidence interval method provided by *Favaro et al.* for error analysis to select the best parameter and protocol.

After many simulations with random values, we realized that $\gamma = 1$ is the best value for the γ parameter both in terms of the smallest confidence interval and the smallest error compared to the true value. For the β protocol, we realized that the *slow protocol* leads to the smallest confidence interval while the *linear protocol* is the best which gives an estimate value closest to the true value and a small confidence interval.

5.2 Future work

Due to limited time, we were not able to make a concrete application of the proposed method. For this, as future work, we would like to apply this method for the case of the Gaussian mixture models in order to find the number K of components in the mixture distribution. Indeed, Gaussian mixture models are probabilistic models for representing normally distributed sub-populations within an overall population. For data d generated from a mixture model, where the number K of components is unknown, we claim that *Jarzynski's* method for model evidence can be tried to find the presumably unknown number K of clusters by calculating the model evidence $q(d|\mathcal{M})$ for different K. The number K that gives the greatest evidence value will probably be the true number of clusters.

The size of the confidence interval is a measure for the convergence of the exponential average involved in *Jarzynski's equality*, where the estimate may still suffer from a significant bias (a relevant difference to the true value). Since we have seen that have the smallest confidence interval for the estimation of the model evidence does not necessarily mean that the estimate is close to the true value, another future work would be to find out why some protocols work well in terms of 'confidence intervals, but poor in accuracy, such that these high-bias protocols could be avoided in a real application.

Acknowledgements

I would like to sincerely express my appreciation to the following :

- Almighty God, who gave me the opportunity to be at AIMS and to successfully complete my studies;
- AIMS and its funders especially Prof. Neil Turok, who gave me the opportunity of doing a structured master in applied mathematics at Stellenbosch University. It has helped me to acquire more skills and knowledge;
- The AIMS management team especially the Director Prof. Barry Green, the Academic Director Dr. Simukai Utete, and the Facicilities & Logistic Manager Igsaan who are doing a great job at AIMS regarding all the students despite the COVID-19 situation ;
- My supervisors, Dr. Daniel Nickelsen who was like a friend, and Dr. Bubacarr Bah. Thank you for the input contributed and the continuous support during my study. Your patience, guidance, helpful comments, and encouragement you gave through my thesis period showed me the right way. It made me realize and unlock my potential. Thank you very much;
- My father, my mother, my brothers and sisters. Thank you very much for the advice, support and motivation throughout this time;
- My English teacher Noluvuyo for her help in my English all the time. She was like a friend. Thank you so much;
- My Tutors, especially Fara, Rahinnatou, Rock and Alice who gave me a listening ear. Thank you really for all your help, guidance and advice. You were always available for me;
- My AIMS Christian fellowship, and my AIMS friends especially Arinze, Mwale and Thabang, I am deeply indebted to you for making our time on campus worthwhile;
- My Cameroonian friends and family for their endless support, it really helped me on this long road;
- All the other people who contributed from far or near to the success of this warm adventure at AIMS.

References

Ahlers, H. and Engel, A. Prior-predictive value from fast growth simulations. *The European Physical Journal B*, 62(3):357–364, 2008.

Aho, K., Derryberry, D., and Peterson, T. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, 95(3):631–636, 2014.

Akaike, H. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, 1973.

Altman, D. G. and Bland, J. M. Statistics notes: the normal distribution. *BMJ*, 310(6975):298, 1995.

Andrieu, C., Djuric, P. M., and Doucet, A. Model selection by MCMC computation. *Signal Process.*, 81(1):19–37, 2001. doi: 10.1016/S0165-1684(00)00188-2. URL https://doi.org/10.1016/S0165-1684(00)00188-2.

Arnold, L. Stochastic differential equations. *New York*, 1974.

Bayes, T. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, (53):370–418, 1763.

Beck, J. L. and Yuen, K.-V. Model selection using response measurements: Bayesian probabilistic approach. *Journal of Engineering Mechanics*, 130(2):192–203, 2004.

Blythe, R. Reversibility, heat dissipation, and the importance of the thermal environment in stochastic models of nonequilibrium steady states. *Physical Review Letters*, 100(1):010601, 2008.

Casella, G. and Robert, C. P. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 1999.

Cawley, G. C. and Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

Celeux, G., Frühwirth-Schnatter, S., and Robert, C. P. Model selection for mixture models—perspectives and strategies. *Handbook of Mixture Analysis*, pages 121–160, 2018.

Cherkassky, V. and Shao, X. Model selection for wavelet-based signal estimation. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)*, volume 2, pages 843–848. IEEE, 1998.

Crooks, G. E. Measuring thermodynamic length. *Physical Review Letters*, 99(10):100602, 2007.

Del Moral, P., Doucet, A., and Jasra, A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

Didelot, X., Everitt, R. G., Johansen, A. M., Lawson, D. J., et al. Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6(1):49–76, 2011.

Ding, J., Tarokh, V., and Yang, Y. Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34, 2018.

Durstewitz, D. Model complexity and selection. In *Advanced Data Analysis in Neuroscience*, pages 73–83. Springer, 2017.

Favaro, A., Nickelsen, D., Barykina, E., and Engel, A. Prior-predictive value from fast-growth simulations: Error analysis and bias estimation. *Physical Review E*, 91(1):012127, 2015.

Friedman, J., Hastie, T., and Tibshirani, R. *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York, 2001.

Gilks, W. R. Markov Chain Monte Carlo. *Encyclopedia of Biostatistics*, 4, 2005.

Hummer, G. and Szabo, A. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Sciences*, 98(7):3658–3661, 2001.

Jarzynski, C. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690, 1997.

Jarzynski, C. Rare events and the convergence of exponentially averaged work values. *Physical Review E*, 73(4):046105, 2006.

Jarzynski, C. Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale. *Annu. Rev. Condens. Matter Phys.*, 2(1):329–351, 2011.

Kirkwood, J. G. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935.

Knuth, K. H., Habeck, M., Malakar, N. K., Mubeen, A. M., and Placek, B. Bayesian evidence and model selection. *Digital Signal Processing*, 47:50–67, 2015.

Lee, J. E., Robert, C. P., et al. Importance sampling schemes for evidence approximation in mixture models. *Bayesian Analysis*, 11(2):573–597, 2016.

Ly, A., Verhagen, J., and Wagenmakers, E.-J. Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72:19–32, 2016.

Lyon, A. Why are normal distributions normal? *The British Journal for the Philosophy of Science*, 65 (3):621–649, 2014.

MacKay, D. J. The evidence framework applied to classification networks. *Neural Computation*, 4(5): 720–736, 1992.

Mannella, R. Integration of stochastic differential equations on a computer. *International Journal of Modern Physics C*, 13(09):1177–1194, 2002.

Marshall, P., Rajguru, N., and Slosar, A. Bayesian evidence as a tool for comparing datasets. *Physical Review D*, 73(6):067302, 2006.

Medgyessy, P. On the unimodality of discrete distributions. *Periodica Mathematica Hungarica*, 2(1-4): 245–257, 1972.

Murphy, K. P. *Machine Learning: a Probabilistic Perspective*. MIT press, 2012.

Neal, R. M. *Probabilistic inference using Markov Chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, Ontario, Canada, 1993.

Neal, R. M. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Neal, R. M. et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2 (11):2, 2011.

Newman, M. and Barkema, G. *Monte Carlo Methods in Statistical Physics, 4 a edición, 68-70*. Oxford University Press, New York, United States, 2006.

Nickelsen, D. and Engel, A. Asymptotics of work distributions: the pre-exponential factor. *The European Physical Journal B*, 82(3-4):207–218, 2011.

Nickelsen, D. and Engel, A. Asymptotic work distributions in driven bistable systems. *Physica Scripta*, 86(5):058503, 2012.

Oates, C. J., Papamarkou, T., and Girolami, M. The controlled thermodynamic integral for Bayesian model evidence evaluation. *Journal of the American Statistical Association*, 111(514):634–645, 2016.

Pohorille, A., Jarzynski, C., and Chipot, C. Good practices in free-energy calculations. *The Journal of Physical Chemistry B*, 114(32):10235–10253, 2010.

Schwarz, G. et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Seifert, U. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, 2012.

Simulations. Simulation code for Bayesian model selection with model evidence. https://github.com/kerol-djoums/AIMS_Project_Code/blob/master/SDEs_for_model_evidence_simulation.ipynb, 2020.

Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

Then, H. and Engel, A. Computing the optimal protocol for finite-time processes in stochastic thermodynamics. *Physical Review E*, 77(4):041105, 2008.

Tong, Y. L. *The multivariate Normal Distribution*. Springer Science & Business Media, 2012.

Von Der Linden, W., Preuss, R., and Dose, V. The prior-predictive value: A paradigm of nasty multi-dimensional integrals. In *Maximum Entropy and Bayesian Methods Garching, Germany 1998*, pages 319–326. Springer, 1999.

Ziegel, E. R. *Statistical Inference*. Taylor & Francis, 2002.