Ph.D. in Pure and Applied Mathematics

Ph.D. Thesis

Politecnico di Torino

Dipartimento di Scienze Matematiche "G.L. Lagrange"

Università degli Studi di Torino

Dipartimento di Matematica "G. Peano"

Generative Models as Out-of-equilibrium Particle Systems: the case of Energy-Based Models

Candidate:
Davide Carbone

Supervisors:
Prof. Lamberto Rondoni
Prof. Eric Vanden-Eijnden

XXXVI cycle
Academic years 2020-2021 / 2021-2022 / 2022-2023

A Mamma, Papà e Andrea

Abstract

At first glance, the two topics addressed in each Part of the present PhD thesis may seem orthogonal: indeed, generative models and linear response theory appear to have little overlap. However, many interesting research topic are linked by common themes after a deep analysis. In the present case, the central narrative thread revolves around Nonequilibrium Statistical Physics. The first fundamental tool from that field that has been used in the present work is Jarzynski identity (C. Jarzynski, 1997); it provides a connection between microscale and macroscale, relating microscopic work along trajectories and free energy, respectively. On the other hand, Onsager reciprocal relations (ORR) represent a milestone in that area (L. Onsager, 1931): they serve as a bridge between a microscopic property (time reversal symmetry) and a macroscopic one (response tensors).
In the first Part, we show how *recent* theoretical results in Statistical Physics can be very instrumental in state-of-the-art applications; generative model represent a substantial research challenge since they are already used in everyday life, even if we are far from having a complete theoretical picture about them. In a nutshell, we propose a novel training algorithm for Energy-Based Models (EBMs), which is a class of diffusion generative models strongly inspired by Statistical Physics, namely by Boltzmann-Gibbs ensemble; in light of this relation, a key strength of EBMs compared to other models is their interpretability. Standard procedures, such as those based on Constrastive Divergence, heavily relies on approximations of the real loss objective already in an ideal setup. Because of that, the practical implementation of such methods usually requires a lot of empirical tricks, often not theoretically justified. In contrast, our proposal is exact; furthermore, no extra bias is introduced by discretization in time and the algorithm provides for free additional information on the trained EBM (i.e. the normalization constant of the trained probabilistic model). Our contribution is based on Jarzynski identity in continuous time and Annealed Importance Sampling in discrete time.
To provide insights into the structure, this section is organized into four chapters. The first chapter offers a historical introduction to generative models, focusing on Energy-Based Models (EBMs) in relation to Statistical Physics and Data Science. The second chapter covers essential technical preliminaries necessary for contextualizing our work. This includes defining EBMs and exploring their purpose, as well as their relationship with other state-of-the-art generative models such as Variational Auto-Encoders, Generative Adversarial Networks, Diffusion-Based Models, and Nor-

malizing Flows. We aim for a unifying approach to highlight similarities and differences between these unsupervised models. The third chapter delves into the relationship between EBMs and the sampling problem. Given that EBM training relies on the ability to sample from a Boltzmann-Gibbs ensemble, we discuss key sampling routines such as the Metropolis-Hastings Algorithm, Unadjusted Langevin Algorithms (ULA), and Metropolis Adjusted Langevin Algorithms (MALA). In the final section, we emphasize the connection between EBMs and Statistical Physics. This serves to justify the adoption of the Boltzmann-Gibbs ensemble and provides important context for utilizing the Jarzynski identity in the main result of this thesis.
The third chapter contains the main novel theoretical result we propose about EBM training. The core idea is the use of nonequilibrium sampling, that is sequential Monte Carlo in discrete time, to efficiently compute the gradient of cross-entropy. Such quantity is necessary to perform KL divergence minimization, or equivalently maximization of log-likelihood, which is the standard approach in statistical learning. We present continuous and discrete time versions of our algorithm, as well as algorithmic aspects having particular relevancy in practical applications. In the last chapter we present experimental result to validate our theoretical findings; we investigate our training routine as opposed to standard procedures like Contrastive Divergence (CD) algorithm. We show as already for Gaussian Mixture Model, our proposal evidently outperforms CD. Similar results are obtained for real image datasets as MNIST and CIFAR-10.
In the second Part, we show that *established* theoretical findings in Statistical Physics can be still object of refinements. ORRs basically provide information on the structure response tensors; the main request for such relations to hold is canonical time reversal symmetry, i.e. the invariance of the equations of motion under the inversion of velocities. Our work demonstrates how we can relax this condition by expanding upon the definition of time reversal symmetry. This expansion enables us to prove that the set of symmetries leading to time reversal invariance is broader. The experimental validity of ORRs has been proven in many contexts where canonical time reversal seems to not hold. Thus, our result contributes to explain some of these examples. Regarding the organization of the treatment, we present the two published papers on the topic, being the second a substantial extension of the first preliminary work.

Contents

Part I

Main Content. Generative Models as Out-of-equilibrium Particle Systems: the case of Energy-Based Models

Part of the work described in this section has also been previously published in:

- D. Carbone, M. Hua, S. Coste, and E. Vanden-Eijnden. *Generative models as out-of-equilibrium particle systems: training of Energy-Based Models using Non-Equilibrium Thermodynamics*. To appear in: Proceedings of the 2nd International Conference on Nonlinear Dynamics and Applications (ICNDA), 2024
- D. Carbone, M. Hua, S. Coste, and E. Vanden-Eijnden. *Efficient Training of Energy-Based Models Using Jarzynski Equality*. In: Proceedings of 37th Conference on Neural Information Processing Systems (NeurIPS), 2023

Chapter 1

Introduction

1.1 Generative models

The problem of description of data through a mathematical model is very old, being the basis of scientific method. The set of measurements use to fulfill the role that contemporary data scientists now refer to as a dataset. Presently, the model can take the form of an exceedingly complex neural network, but the underlying extrapolation remains akin to P.S. Laplace's famous deterministic statement[1]: "An intellect which at a certain moment would know all forces [data] that set nature in motion [...] would be uncertain and the future just like the past could be present before its eyes". One can easily extend this reasoning, asserting that the more data one possesses, the more robust and detailed the model that can be constructed atop them. This leads to enhanced predictions and greater stability concerning unforeseen behaviors.
This line of thought was boosted in the previous century with the advent of automatic calculators, and the velocity of development becomes astounding. For instance, consider the remarkable computational power difference between your smartphone and the computer used for the Apollo program by NASA in the 1960s[2]. Hence, the quest for data has become an indispensable aspect of contemporary science.
To delve deeper into this issue, let us construct a historical metaphor. One of the early modern achievements in observational astronomy is Kepler's laws. The genesis of such results is deeply rooted in a vast collection of observational data amassed by T. Brahe[3]. Kepler's formulation was, in fact, motivated by the necessity to explain these astronomical measurements. In a simplified analogy, we observe the dichotomy between the "model," embodied by Kepler, and the "dataset," represented in this narrative by Brahe. Since the 17th century, these two actors have played equally fundamental roles in the advancement of science, taking turns on the stage with the same importance. Consider, for instance, the pivotal role played by Faraday's experiments

[1] Pierre-Simon Laplace. *A philosophical essay on probabilities*. Courier Corporation, 2012.

[2] URL: https://www.linkedin.com/pulse/smartphone-today-has-more-computing-power-than-nasas-1960-offermann

[3] URL: https://www.britannica.com/science/history-of-science/Tycho-Kepler-and-Galileo

in understanding electromagnetism[4], long before Maxwell's laws. Or, conversely, the impact of theory of Relativity[5] way before its experimental confirmation.
In recent years, particularly during the 2000s, we have witnessed a profound paradigm shift represented by the Big Data Era[6]. Thanks to the aforementioned technological advancements in computer science, the volume of generated scientific (and not) data has dramatically increased, resulting from advancements in simulation and storage capabilities. Furthermore, there has been a growing collection of data on human activities, including images, text, sounds, and more.
Returning to the historical analogy, it is akin to Brahe suddenly providing Kepler with a thousand times the amount of data that the latter was accustomed to. This shift posed a methodological problem in what we now refer to as data science, and this is where the machine learning approach came into play[7]. The models required to process Big Data had already been theoretically studied since the invention of the perceptron[8]. Their application was constrained by computational power in the last century, but, as a peculiar example of convergent development, they became the primary tools in the toolbox of data scientists in the 2000s, simultaneously to the appearance of Big Data on the stage.
There is indeed a discontinuity that deserves more attention: the increasing collection of data *generated* by humans. The term *Big Data* is sometimes limited to images, sounds, videos, text, and metadata resulting from human activities, not just on the internet. Unlike scientific measurements, having access to an extensive quantity of information produced by humans opened Pandora's box, prompting the natural question: *can we build artificial intelligence by leveraging Big Data*? In other words, can we construct a machine capable of *generating* data as humans do, by training it in some smart way? Data here is to be understood in a broad sense, encompassing new theorems, art pieces, images, videos, and even novels.
Generative models represent, in this sense, the most recent breakthrough in technological advancement towards intelligent-like machines. It is complicated to provide a general definition, and there are already many available from different sources[9,10]. However, if we informally focus on those already known to the general public, such as Generative Pre-Trained Transformers (GPT)[11], the common traits of most definitions are few. Firstly, generative models require a substantial amount of data for training, in

[4] Jim Al-Khalili. The birth of the electric machines: a commentary on Faraday (1832)'Experimental researches in electricity'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **373**: 20140208, 2015.

[5] URL: `https://www.britannica.com/science/relativity/Intellectual-and-cultural-impact-of-relativity`

[6] URL: `https://medium.com/swlh/big-data-era-84b488491a8d`

[7] Alexander L Fradkov. Early history of machine learning. *IFAC-PapersOnLine*, **53**: 1385–1390, 2020.

[8] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5**: 115–133, 1943.

[9] URL: `https://www.techtarget.com/searchenterpriseai/definition/generative-modeling#:~:text=Generative%20modeling%20is%20the%20use,can%20be%20calculated%20from%20observations.`

[10] URL: `https://www.nvidia.com/en-us/glossary/generative-ai/`

[11] URL: `https://www.nytimes.com/2022/12/10/technology/ai-chat-bot-chatgpt.html`

addition to the selection of a precise architecture, which goes far beyond the original perceptron. Secondly, the training is probably not biologically inspired, i.e., we do not learn through backpropagation[12], which is the most commonly used training technique in machine learning. For completeness, it is worth noting that this thesis is still debated in neuroscience[13]. Thirdly, a generative model is not necessarily informative about the data distribution; for instance, ChatGPT could achieve astounding results in text generation, but the training machine does not provide knowledge about some general features of text generated by humans.
Returning to the historical metaphor: nowadays, we are able to build "BraheGPT," which can generate and gather new plausible measurements about the orbits of planets in unobserved planetary systems after training on observed data from the solar system. However, it is not Kepler; deductive reasoning is not necessary to generate new data instances, although it remains fundamental to understanding the world. Von Neumann would certainly adapt his famous statement[14] about overfitting to modern data science, cautioning against the ability to generate examples without a general picture.
Prominent data scientists, such as Yann LeCun, have recently emphasized that the use of interpretable generative models is crucial for achieving a "unified world model for AI capable of planning"[15]. This thesis becomes imperative in the realm of computational sciences, where qualitative generation alone is insufficient as a benchmark to evaluate model performance. In sectors like Molecular Dynamics, Biochemistry, and similar fields, the model must convey substantial information about the dataset. The generative models that excel in terms of interpretability, which form the main focus of the present work, are precisely the *Energy-Based Models* (EBMs). These models offer a unique advantage in their ability to provide insights into the underlying mechanisms of the data they generate. In areas such as Molecular Dynamics and Biochemistry, where understanding the intricate relationships within the dataset is crucial, the interpretability of EBMs stands out.
In adopting EBMs, researchers and practitioners gain not only the capacity to generate high-quality data but also a clearer understanding of the factors influencing the generated outputs. This interpretability is indispensable in domains where the model's ability to convey meaningful information about the dataset is paramount. As the pursuit of a unified world model for AI continues, the emphasis on interpretable generative models, particularly EBMs, plays a pivotal role in bridging the gap between data generation and comprehensive understanding.

[12] Stephen Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive science*, **11**: 23–63, 1987.
[13] Timothy P Lillicrap et al. Backpropagation and the brain. *Nature Reviews Neuroscience*, **21**: 335–346, 2020.
[14] Freeman Dyson et al. A meeting with Enrico Fermi. *Nature*, **427**: 297–297, 2004.
[15] URL: `https://www.zdnet.com/article/metas-ai-luminary-lecun-explores-deep-learnings-energy-frontier/`

1.2 A long story: from Boltzmann-Gibbs ensemble to the advent of EBMs

After providing a historical overview of generative models, this section is dedicated to exploring the origin and development of Energy-Based Models (EBMs). As we delve into this discussion, it becomes evident that the theoretical foundation of such generative models exists under different names at the intersection of various fields, including statistical physics, probability theory, computer science, and sampling, among others. In this section, we emphasize a historical perspective to shed light on the evolutionary trajectory of EBMs. While we touch upon the overarching theories, more in-depth theoretical discussions are reserved for subsequent chapters. We believe that this review serves as a valuable resource for readers across diverse fields enabling them to construct a comprehensive understanding of what constitutes an Energy-Based Model by tracing the genesis of this topic.

The first ingredient of the story is the Boltzmann-Gibbs measure, a fundamental concept in statistical mechanics, and has its origins in the works of Ludwig Boltzmann and Josiah Willard Gibbs during the late 19th century. These two influential physicists independently contributed to the development of statistical mechanics, providing a bridge between the microscopic behavior of particles and macroscopic thermodynamic properties.

Ludwig Boltzmann made significant strides in understanding the statistical nature of gases, introducing what is now known as the Boltzmann distribution[16]. Boltzmann's statistical approach, which related the statistical weight of different microscopic configurations to their entropy, laid the groundwork for the probabilistic description of thermodynamic systems.

Josiah Willard Gibbs, in parallel with Boltzmann, extended these ideas to develop the canonical ensemble, introducing what is commonly referred to as the Gibbs measure[17]. He provides a mathematical framework for calculating thermodynamic properties based on the statistical distribution of particles in a given system. The Boltzmann-Gibbs measure, which emerged from the synthesis of these ideas, describes the probability distribution of particles in different energy states at *thermal equilibrium* at temperature T. It has become a cornerstone of statistical mechanics, applicable to diverse physical systems, including gases, liquids, and solids. We informally recall its definition: given the state of the system $x \in \Omega$, where Ω is the so-called phase space, and an energy function $U : \Omega \to \mathbb{R}^+$, we can express the associated probability density function

$$\rho(x) \propto e^{-\beta U(x)} \tag{1.2.1}$$

where $\beta = 1/k_B T$, k_B being the Boltzmann constant. A detailed mathematical description will be provided in the next chapters.

[16] Ludwig Boltzmann. Studien über das Gleichgewicht der lebendigen Kraft zwischen bewegten materiellen Punkten [Studies on the balance of living force between moving material points]. *Wiener Berichte*, **58**: 517–560, 1868.

[17] Josiah Willard Gibbs. *Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics.* C. Scribner's sons, 1902.

The analysis of the impact of Boltzmann-Gibbs ensemble on physics would require a full monography per se; for the sake of the present work, we directly advance to 1924, when E. Ising presented his PhD thesis[18]. The so called Ising model is a fundamental mathematical model in statistical mechanics. It serves as a simplified yet powerful representation of magnetic systems, particularly in understanding the behavior of spins in a lattice Λ — for simplicity, we can imagine a graph with N nodes. In the Ising model, each lattice site is associated with a magnetic spin, which can take two possible values, usually denoted as "up" or "down", that is $\Omega = \{-1, 1\}^N$. The interactions between spins are typically modeled using a simple energy function, namely

$$U_{Is}(x) = -\sum_{\langle ij \rangle} J_{ij} x_i x_j - \mu \sum_j h_j x_j \tag{1.2.2}$$

Let us briefly clarify the notation: $i, j \in \Lambda$ are indexes of sites in the lattice; $\langle ij \rangle$ indicates that the sum is restricted to first neighbours and J_{ij} is the strength of the interaction. The field h_i instead individually acts on each site and μ is just a constant that traditionally corresponds to magnetic moment. In laymen terms, each magnetic spin interacts with its first neighbours and with an external field. The alignment of spins is encouraged.
In considering (1.2.1) as associated to U_{Is}, the primary focus is often on the behavior of the system as a function of temperature. In a nutshell, at high temperatures, thermal fluctuations dominate, and the system exhibits no long-range order. As the temperature decreases, there is a critical point at which the system undergoes a phase transition, leading to spontaneous magnetization and the emergence of long-range order.
For some decades the interest for Ising model and its extensions was confined to physics. The motivation for invoking such a model in the present work is the following: in the 80s a fundamental connection between Ising model and data science manifested through Hopfield networks[19] and Boltzmann machines[20]. Both can be viewed as an Ising lattice where interactions are not confined to first neighbors. Apart from the initial summation, which, for the former, extends to $\forall i, j \in \Lambda$ rather than just $\langle ij \rangle$, the energy function bears resemblance to (1.2.2). From a statistical physics standpoint, the distinction between a Hopfield network and Boltzmann machines lies solely in the temperature value.
The purpose of the former is pattern recognition and associative memory tasks. A distinctive feature of Hopfield networks is their proficiency in storing and retrieving patterns through symmetric connections between neurons, that is Ising sites, in the network. In practice, when provided with a set of network configurations $y^\lambda \in \Omega$ representing patterns, denoted by $\lambda = 1, \ldots, n$, one constructs the coupling J as

[18] URL: `https://www.hs-augsburg.de/~harsch/anglica/Chronology/20thC/Ising/isi_fm00.html`
[19] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, **79**: 2554–2558, 1982.
[20] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive science*, **9**: 147–169, 1985.

follows:

$$J_{ij} = \frac{1}{n} \sum_{\lambda=1}^{n} y_i^\lambda y_j^\lambda \tag{1.2.3}$$

This involves employing the Hebbian rule[21] "neurons wire together if they fire together"[22], but further specifications[23] are available. This phase is commonly referred to as the *training* of the network. Subsequently, one can define a retrieval iterative dynamics starting from any configuration $x^{k=0} \in \Omega$, as exemplified by the equation:

$$x^{k+1} = \text{sgn}(Jx^k + h) \qquad k \in \mathbb{N} \tag{1.2.4}$$

Here, J represents the coupling matrix defined element-wise in (1.2.3), and h is a bias vector that influences the preferences for 'up' or 'down'. It is noteworthy that in a Hopfield network, there is no use of the Boltzmann-Gibbs ensemble; the objective is to construct a dynamical system with prescribed attractors, which are the minima of $U(x)$ by design.

Boltzmann Machines share the same structure and energy function but the goal extends beyond the mere retrieval of patterns; it is to model their overall *distribution*. To illustrate this concept, consider a finite set of n natural images of cats and dogs. A meticulously designed Hopfield Network could perfectly retrieve any of these examples. On the contrary, a trained Boltzmann Machine aspires to generate new instances of cats and dogs, capturing, in a sense, the distribution of such images. The objective appears to be on a different level of difficulty: although possibly big, the cardinality of the set of patterns is finite; the number of possible variations of cats and dogs is not. Thus, one can immediately guess why the training and generation phases (n.b. it is no more just a retrieval) are completely different w.r.t. Hopfield Networks. The take home message is the hypothesis that the distribution of the given patterns can be described by a Boltzmann-Gibbs ensemble associated to the energy of the Hopfield Network at temperature T.

It is convenient to consider Boltzmann Machines as a specific instance of Energy-Based Models, a term introduced by Hinton et al.[24], to describe both training and generation phases. EBMs differ from Boltzmann Machines in the use of a generic parametric energy $U_\theta(x)$ instead of the usual choice made for the latter. Here, $\theta \in \Theta$ needs to be selected and trained so that the Boltzmann-Gibbs ensemble ρ_θ associated with $U_\theta(x)$ "fits well" the distribution of the given patterns, which we refer to as ρ_*. After training the EBM, the generative phase involves *sampling* equilibrium configurations from ρ_θ. Specifically, a Boltzmann Machine corresponds to an EBM with the choice of $U(x)$ as the energy of a Hopfield Network and $\theta = J$.

[21] Donald O Hebb. Organization of behavior. new york: Wiley. *J. Clin. Psychol*, **6**: 335–307, 1949.

[22] Siegrid Löwel and Wolf Singer. Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Science*, **255**: 209–212, 1992.

[23] Amos Storkey. "Increasing the capacity of a hopfield network without sacrificing functionality" in: *Artificial Neural Networks—ICANN'97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceeedings 7*. Springer 1997. 451–456

[24] Yee Whye Teh et al. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, **4**: 1235–1260, 2003.

Despite their conceptual simplicity, both training and generation represent fundamental open problems that intersect multiple research fields. In essence, sampling from a Boltzmann-Gibbs ensemble is a challenging task in general, and unfortunately, it is necessary even during the training phase. For this reason, the use of Boltzmann Machines was limited to toy models until the proposal of the Contrastive Divergence algorithm by Hinton[25].
This procedure, along with its generalizations, made it possible to apply EBMs to practical problems. Moreover, thanks to the adoption of a deep neural network[26] as U_θ, the interest towards this class of generative models critically increased and in 2010s the use of EBMs for state-of-the-art tasks became standard. However, all that glitters is not gold. Despite its success in generating high-quality individual samples, the use of Contrastive Divergence is known to be biased. For instance, it could happen that individual images are correctly generated, but ensemble properties as the relative proportion of the two species is incorrect. Although Hinton et al. originally claimed that this bias is generally small[27], numerous counterexamples have been shown in the more than 20 years since their original paper. The absence of novel paradigm shifts, coupled with the rise of alternative generative models (e.g., diffusion-based ones[28]), has reduced attention on EBMs and consequently on Boltzmann Machines.
In order to put a tile in the opposite direction, we will present an alternative proposal for EBM training. Our proposal is based on the interpretation of the training as the evolution of a physical particle system out of equilibrium. The organization of the work will be the following:

- **Preliminaries**. A section in which we present the general problem of EBM training and its relation with other state-of-the-art generative models. Moreover, we highlight the connection of EBM with Boltzmann-Gibbs ensemble, in relation to the problem of sampling as well as to its foundational aspects in Statistical Physics.

- **Efficient Training of EBMs using Jarzynski equality**. A section devoted to the theoretical part of our novel proposal. Firstly, our training proposal in continuous time is presented together with a physical interpretation. Then, the subsequent discrete time version is defined. The chapter is concluded with the notable theoretical example of Gaussian Mixture Model.

- **Practical Implementation and Numerics**. A section devoted to numerical aspects and experimental simulations. Firstly, the needs for resampling is discussed and analyzed. Then, simulations are presented both on synthetic data (Gaussian Mixture Model) and on real data (MNIST and CIFAR-10) dataset.

[25] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, **14**: 1771–1800, 2002.
[26] Jianwen Xie et al. "A theory of generative convnet" in: *International Conference on Machine Learning*. PMLR 2016. 2635–2644
[27] Miguel A Carreira-Perpinan and Geoffrey Hinton. "On contrastive divergence learning" in: *International workshop on artificial intelligence and statistics*. PMLR 2005. 33–40
[28] Yang Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations" in: *International Conference on Learning Representations*. 2020.

Chapter 2

Preliminaries

2.1 Basic definitions and assumptions

In this Section, we provide the basic formal definition of Energy-Based Model. We will adopt the notation and the presented assumption throughout the present work. First of all, the problem we consider can be formulated as follows: we assume that we are given $n \in \mathbb{N}$ data points $\{x_i^*\}_{i=1}^n$ in $\mathbb{R}^d$ drawn from an unknown probability distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, with a positive probability density function (PDF) $\rho_*(x) > 0$ (also unknown). This is a standard problem in statistical learning, where *learning from data* here refers to the ability to fit the data distribution and to generate new examples. More precisely, our aim is to estimate $\rho_*(x)$ via an energy-based model (EBM), i.e. to find a suitable energy function in a parametric class, $U_\theta : \mathbb{R}^d \to [0, \infty)$ with parameters $\theta \in \Theta$, such that the associated Boltzmann-Gibbs PDF

$$\rho_\theta(x) = Z_\theta^{-1} e^{-U_\theta(x)}; \qquad Z_\theta = \int_{\mathbb{R}^d} e^{-U_\theta(x)} dx \tag{2.1.1}$$

is an approximation of the target density $\rho_*(x)$. Actually, any probability density function can be written as a Boltzmann Gibbs ensemble for a particular choice of $U(x)$. The normalization factor Z_θ is known as the partition function in statistical physics[1], see also Section 2.4, and as the evidence in Bayesian statistics[2]. This factor is hard to estimate, especially in high dimension, see the following Box for further details.

[1] Evgenii Mikhailovich Lifshitz and Lev Petrovich Pitaevskii. *Statistical physics: theory of the condensed state*. vol. 9 Elsevier, 2013.

[2] Farhan Feroz and Mike P Hobson. Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, **384**: 449–463, 2008.

Estimation of Partition Function

Even if U_θ is known, an explicit analytical computation of the partition function is generally unfeasable. If the dimension d is big enough, the integral defining Z_θ cannot be computed using standard quadrature methods. The only possibility is Monte-Carlo sampling[3]. To employ such method, one can express the partition function as an expectation $\mathbb{E}_0$ with respect to a chosen probability density function ρ_0, i.e.

$$Z_\theta = \mathbb{E}_0\left[\frac{e^{-U_\theta}}{\rho_0}\right] \tag{2.1.2}$$

The selected density must be known pointwise in $\mathbb{R}^d$, including the normalization constant, and it should be easy to sample from. If these conditions are met, one can compute the partition function by simply replacing the expectation in (2.1.2) with the corresponding empirical average computed using samples drawn from ρ_0. Unfortunately, finding a probability density that satisfies these properties is challenging. For a general choice that is not tailored to e^{-U_θ}, the estimator is likely to be very poor, characterized by a very large, or even infinite, coefficient of variation.

[3] Jun S Liu. *Monte Carlo strategies in scientific computing*. vol. 75 Springer, 2001.

One advantage of EBMs is that they provide generative models that do not require the explicit knowledge of Z_θ. In Section 2.3 we will present some routines that can in principle be used to sample ρ_θ knowing only U_θ – the design of such methods is an integral part of the problem of building an EBM.
To proceed we need some assumptions on the parametric class of energy:

Assumption 2.1.1. *For all $\theta \in \Theta$:*

1. $U_\theta \in C^2(\mathbb{R}^d)$; $\quad \exists L \in \mathbb{R}_+ : \|\nabla\nabla U_\theta(x)\| \leq L \quad \forall x \in \mathbb{R}^d$;
2. $\exists a \in \mathbb{R}_+$ *and a compact set* $\mathcal{C} \in \mathbb{R}^d$: $x \cdot \nabla U_\theta(x) \geq a|x|^2 \quad \forall x \in \mathbb{R}^d \setminus \mathcal{C}$.

The need for the first assumption will be discussed in Section 2.3: it is related to well-posedness and convergence properties of the dynamics used for sampling, i.e. Langevin dynamics and its specifications. The second assumption guarantees that $Z_\theta < \infty$ (i.e. we can associate a PDF ρ_θ to U_θ via (2.1.1) for any $\theta \in \Theta$). We provide now two important definitions:

Definition 2.1.1 (Convexity)**.** *A function $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$ is convex if given $0 < \lambda < 1$ and $x_1, x_2 \in \mathbb{R}^d$ such that $x_1 \neq x_2$, the following is true*

$$\varphi\left(tx_1 + (1-t)x_2\right) \leq t\varphi\left(x_1\right) + (1-t)\varphi\left(x_2\right) \tag{2.1.3}$$

Definition 2.1.2 (Log-Concavity)**.** *A density function ρ with respect to Lebesgue measure on $(\mathbb{R}^d, \mathcal{B}^d)$ is log-concave if $\rho = e^{-\varphi}$ where φ is convex.*

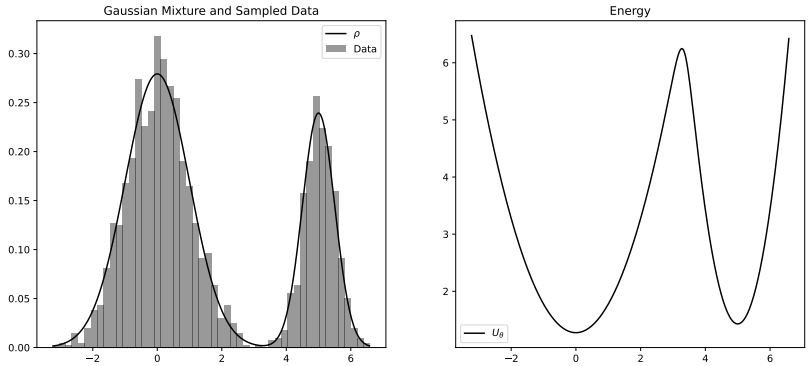


Figure 2.1: *Gaussian Mixture.* Plot of PDF with sampled histogram and associated energy U_θ.

A non-convex function could have more than one local but not global minima; conversely, a non-log-concave probability density could have more than local maxima, which are called *modes*. It is important to stress that Assumption (2.1.1) *does not* imply that U_θ is convex (i.e. that ρ_θ is log-concave): in fact, we will be most interested in situations where U_θ has multiple local minima so that ρ_θ is multimodal. We will elaborate on the topic in Section 2.3. It is well known as for optimization problems, non-convex cases are the most complicated. Similarly, sampling from a non-log-concave probability density function (PDF) can be extremely challenging. Another assumption we will adopt is:

Assumption 2.1.2. *Without loss of generality* $\exists \theta_* \in \Theta \ : \ \rho_{\theta_*} = \rho_*$, *that is* ρ_* *is in the parametric class of* ρ_θ.

Throughout this thesis, our aims are primarily to identify θ_* and to sample ρ_{θ_*}; in the process, we will also show how to estimate Z_{θ_*}.

Example 1. Let us present a simple example to visualize the relation between convexity and log-concavity. In Figure 2.1 we plot side by side the PDF of a Gaussian mixture in 1D and the associated potential

$$U_\theta(x) = \log \left[p \exp\left(-\frac{(x-\mu_1)^2}{\sigma_1^2} \right) + (1-p) \exp\left(-\frac{(x-\mu_2)^2}{\sigma_2^2} \right) \right] \tag{2.1.4}$$

where $\theta = \{p, \mu_{1,2}, \sigma_{1,2}\}$. The specific values are $p = 0.7$, $\mu_1 = 0$, $\mu_2 = 5$, $\sigma_1 = 1$ and $\sigma_2 = 0.5$. It is clear the correspondence between minima of $U_\theta(x)$ and maxima, that is modes, of ρ.

2.1.1 Cross-entropy minimization

Once we defined an EBM, we need to measure its quality with respect to the data distribution. Possibly, this would provide a way to train its parameters. Hence, we

define some important quantities:

Definition 2.1.3. *Consider two probability densities on $\mathbb{R}^d$ and absolutely continuous with respect to Lebesgue measure, namely ρ_1 and ρ_2. We define*

1. **Cross Entropy**
$$H(\rho_1, \rho_2) = -\int_{\mathbb{R}^d} \log \rho_2(x)\rho_1(x)dx \tag{2.1.5}$$

2. **Kullback-Leibler divergence**[4]
$$D_{KL}(\rho_1 \parallel \rho_2) = \int_{\mathbb{R}^d} \rho_1(x) \ \log\left(\frac{\rho_1(x)}{\rho_2(x)}\right) \ dx \tag{2.1.6}$$

3. **Entropy**
$$H(\rho_1) = -\int_{\mathbb{R}^d} \log \rho_1(x)\rho_1(x)dx \tag{2.1.7}$$

The KL divergence is a widely used estimator for the dissimilarity between probability measures. It satisfies the non-negativity condition

$$D_{\mathrm{KL}}(\rho_1 \parallel \rho_2) \geq 0, \qquad D_{\mathrm{KL}}(\rho_1 \parallel \rho_2) = 0 \iff \rho_1 = \rho_2 \text{ a.e.} \tag{2.1.8}$$

However, it is not a proper distance since it is not symmetric and it does not satisfy triangular inequality. The following trivial lemma relates the three quantities we introduced in Definition 2.1.3:

Lemma 2.1.1. *The following equality holds for any choice of PDFs ρ_1 and ρ_2*

$$H(\rho_1, \rho_2) = H(\rho_2) + D_{KL}(\rho_2 \parallel \rho_1) \tag{2.1.9}$$

One can also use the cross-entropy of the model density ρ_θ relative to the target density ρ_* as an estimate of diversity between the two PDFs; in such case, 2.1.5 simplifies becoming

$$H(\rho_*, \rho_\theta) = \log Z_\theta + \int_{\mathbb{R}^d} U_\theta(x)\rho_*(x)dx \tag{2.1.10}$$

Because of 2.1.9, the difference between the cross-entropy and the KL divergence is $H(\rho_*)$, a term that depends just on the data distribution. Hence, the optimal parameters θ^* are solution of an optimization problem on Θ, namely

$$\theta^* = \arg\min_{\theta\in\Theta} D_{\mathrm{KL}}(\rho_*||\rho_\theta) = \arg\min_{\theta\in\Theta} H(\rho_*, \rho_\theta), \tag{2.1.11}$$

meaning that the entropy of ρ_* plays no active role in solving such minimization problem. There is a subtle issue in this reasoning: unlike KL divergence, the cross-entropy is not bounded from below, and in particular $H(\rho, \rho) \coloneqq H(\rho) \neq 0$. That is, we should

[4]Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, **22**: 79–86, 1951.

compute $H(\rho_*)$ to estimate the minimum value of cross-entropy. Unfortunately, most of the empirical estimators to be used when ρ_* is known through samples suffer in high dimension[5]. Solving (2.1.11) is equivalent to maximum likelihood method, a widely used practice in parametric statistics[6].
The use of cross-entropy avoids the very problematic computation of $H(\rho_*)$, but in 2.1.10 the estimation of Z_θ is also needed. However, the most common routines for cross-entropy minimization are gradient-based: they rely on the gradient of $\partial_\theta H(\rho_*, \rho_\theta)$ and not on the cross-entropy itself. The former can be computed using the identity $\partial_\theta \log Z_\theta = -\int_{\mathbb{R}^d} \partial_\theta U_\theta(x)\rho_\theta(x)dx$, obtaining

$$\begin{aligned}\partial_\theta H(\rho_*, \rho_\theta) &= \int_{\mathbb{R}^d} \partial_\theta U_\theta(x)\rho_*(x)dx - \int_{\mathbb{R}^d} \partial_\theta U_\theta(x)\rho_\theta(x)dx \\ &\coloneqq \mathbb{E}_*[\partial_\theta U_\theta] - \mathbb{E}_\theta[\partial_\theta U_\theta].\end{aligned} \tag{2.1.12}$$

This is a crucial expression for the present work. In fact, we can highlight the core theme of the whole thesis:

Remark 2.1.1 (Fundamental problem for EBM training)**.** *Estimating $\partial_\theta H(\rho_*, \rho_\theta)$ requires calculating the expectation $\mathbb{E}_\theta[\partial_\theta U_\theta]$. In contrast $\mathbb{E}_*[\partial_\theta U_\theta]$ can be readily estimated on the data.*

Typical training methods, e.g. based on the so-called Constrastive Divergence[7] and its specifications, resort to various approximations to calculate the expectation $\mathbb{E}_\theta[\partial_\theta U_\theta]$— see Section 3.1 for more discussion about these methods. While these approaches have proven successful in many situations, they are prone to training instabilities that limit their applicability. The cross-entropy is more stringent, and therefore better, than objectives like the Fisher divergence used to train other generative models: for example, unlike the latter, it is sensitive to the relative probability weights of modes on ρ_* separated by low-density regions[8] — we will elaborate in Section 2.2.

2.2 EBMs among generative models

In this section, our objective is to provide a brief overview of the other main generative models available on the market, possibly in relation to Energy-Based Models. The aim is to construct a convenient general framework for the reader, with detailed specifications not being the focus of this section. Let us establish a general classification of the methods we will discuss. As outlined in the introduction, creating a generative model involves developing a computational tool capable of generating new instances representative of a given dataset. Taking the example of image generation, starting

[5] Ziqiao Ao and Jinglai Li. Entropy estimation via uniformization. *Artificial Intelligence*, 103954, 2023.
[6] Stephen M Stigler. The epic story of maximum likelihood. *Statistical Science*, 598–620, 2007.
[7] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, **14**: 1771–1800, 2002.
[8] Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

with a dataset of dogs, a generative model can produce new images of dogs. Even in this simple example, determining whether a generated sample is "good" or not can be far from obvious. A good generative model should possess two key properties: (1) ease of training and (2) ease of generation. Unfortunately, demanding the best of all possible worlds is often impractical, and a trade-off is frequently necessary to balance these two properties.
The concept of a generative model is relatively new and strictly related to the rise of Big Data. Before the advent of modern computer science, generating data (for inference, modeling) was identified with collecting measures. The advent of computer simulations laid the first stone towards generating data from a model. Let us mention Fermi-Pasta-Ulam-Tsingou[9], which is usually referred to as one of the first uses of computers to simulate a physical model. In statistics, this concept of generating data from a given model is called "sampling" (see Section 2.3). The change of paradigm towards generating data *from data* became possible when sufficient computational power and memory were available. Generative AI is following a path similar to the internet: originally limited to academic purposes[10], it now permeates everyday life. Thanks, for instance, to Generative Pre-Trained Transformers (such as ChatGPT[11]), we seem to be closer to creating a machine capable of generating data, text, sounds, and more, as humans do. The debate about artificial general intelligence capable of surpassing humans is already spreading[12–14].
We will now review the technical details of state-of-the-art generative models. At the end, we will also highlight the relation with Energy-Based Models if applicable.

2.2.1 Variational Autoencoders

As we can infer from the name, to present a variational autoencoder (VAE)[15] we firstly need to summarize what an autoencoder (AE) is[16]. Let us focus on Figure 2.2: it is a Deep Neural Network (DNN) designed to replicate an input vector $x \in \mathbb{R}^d$, after the application of two NN in sequence. The left segment of the AE, known as the encoder $e(x)$, generates a low-dimensional latent representation $z \in \mathbb{R}^L$, with $L \leq d$, at the bottleneck layer. The right segment, referred to as the decoder $d(z)$, endeavors to reconstruct x from z. During the training phase, the true output is compared with $d(e(x))$ in order to perform backpropagation and train the nets. During the test phase,

[9] Enrico Fermi et al. *Studies of the nonlinear problems* tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 1955

[10] URL: `https://www.livinginternet.com/i/ii_arpanet.htm`

[11] URL: `https://www.nytimes.com/2022/12/10/technology/ai-chat-bot-chatgpt.html`

[12] Evgeny Morozov. The True Threat of Artificial Intelligence. *International New York Times*, NA–NA, 2023.

[13] Ragnar Fjelland. Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, **7**: 1–9, 2020.

[14] Frederik Federspiel et al. Threats by artificial intelligence to human health and human existence. *BMJ global health*, **8**: e010435, 2023.

[15] Laurent Girin et al. Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*, 2020.

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, **313**: 504–507, 2006.

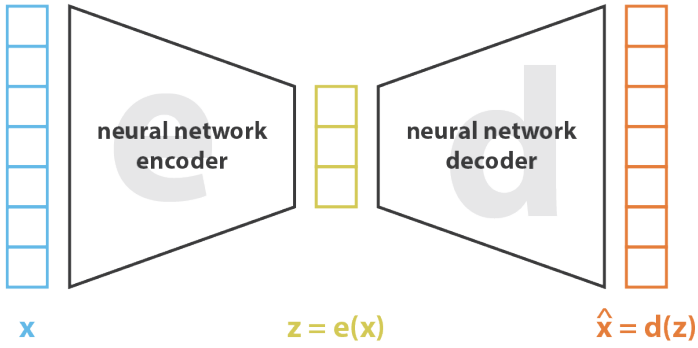


Figure 2.2: Representation of an autoencoder[17]

$\hat{x}$ is used as an estimated value of x, that is $\hat{x} \approx x$. An AE can be seen as a trainable compression protocol: once trained, encoder and decoder are separate parts that can be used separately, for instance before and after a data transmission procedure. In practice, their use is widely diffused in Machine Learning application: it is common to put extra layers acting in the latent space, for instance for a supervised tasks[18]. Up to this point, everything operates deterministically: during testing, when the AE is provided with a specific input vector, it consistently produces the corresponding output.
The subsequent specification of AE are the Variational Autoencoders[19]. While in AE we had two deterministic functions $e(x)$ and $d(z)$, in VAE encoder and decoder are two probabilistic models: an *inference* model and a *generative* model. Despite this classification, VAE are usually referred to as generative models in toto. Let us clarify in formulae the construction.
We consider the joint parametric probability density $\rho_\theta(x, z)$ on $\mathbb{R}^d \times \mathbb{R}^L$, where the parameters $\theta \in \Theta$ are the weights of a neural network (NN). Specifically, using the definition of joint PDF, we write

$$\rho_\theta(x, z) = \rho_\theta(x|z)\rho(z) \tag{2.2.1}$$

The prior distribution $\rho(z)$ is usually assumed to be a multivariate gaussian distribution $\mathcal{N}(z, \mathbf{0}_L, \boldsymbol{I}_L)$, with zero mean vector $\mathbf{0}_L$ and identity $\boldsymbol{I}_L$ as covariance. The

[18] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, 353–374, 2023.
[19] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes" in: *2nd International Conference on Learning Representations, ICLR 2014*. 2014.

parametric conditional PDF $\rho_\theta(x|z)$ is the *decoder network* and can be designed case by case: the simplest and traditional choice is a gaussian

$$\rho_\theta(x|z) = \mathcal{N}(x, \boldsymbol{\mu}_\theta(z), \text{diag}\{\boldsymbol{\sigma}^2_\theta(z)\}) \tag{2.2.2}$$

with parametric mean $\boldsymbol{\mu}_\theta(z)$ and diagonal covariance matrix $\text{diag}\{\boldsymbol{\sigma}^2_\theta(z)\}$ (for instance modelled through appropriate NN). Other possibilities have been studied to tackle different kind of data, for instance audio[20].
Following this formal definition, the marginal distribution of the data x will be

$$\rho_\theta(x) = \int_{\mathbb{R}^L} \rho_\theta(x|z)\rho(z)dz \tag{2.2.3}$$

Similarly to EBM training, we need to select the optimal parameters θ^* that minimize a selected measure of discrepancy between the model and the true data distribution ρ_*, as usual known just through samples. The procedure is analogous to (2.1.11): KL divergence is used to evaluate this diversity,

$$\theta^* = \arg\min_{\theta\in\Theta} D_{\text{KL}}(\rho_*(x) \parallel \rho_\theta(x)) = \arg\max_{\theta\in\Theta} \mathbb{E}_*[\log \rho_\theta(x)] \tag{2.2.4}$$

Differently from EBMs, the right-hand side is traditionally written as an expectation: it is the marginal log-likelihood of the model[6]. It is just a matter of notation — the optimization objectives are the same. When having a dataset $\mathcal{X} = \{x_i \in \mathbb{R}^d\}_{i=1}^N$, one could estimate the expectation via the empirical average $\sum_{i=1}^N \log \rho_\theta(x_i)/N$. However, the log-likelihood is defined via (2.2.3), and such an integral is often analytically intractable. That is, one has no direct access to $\log \rho_\theta(x)$ explicitly. The proposed solution to overcome this issue is based on a variational approach. Let us present a crucial definition and a lemma:

Definition 2.2.1 (ELBO). *Let $\mathcal{F}$ denote a variational family defined as a set of PDFs over the latent variables z. For any $q(z) \in \mathcal{F}$, the* **Evidence Lower Bound (ELBO)** *(also known as* variational free energy*) $\mathcal{L} : \Theta \times \mathcal{F} \times R^d \to \mathbb{R}$ is defined as*

$$\mathcal{L}(\theta, q(z); x) = \mathbb{E}_{q(z)}[\log \rho_\theta(x, z) - \log q(z)] \tag{2.2.5}$$

Lemma 2.2.1. *The following properties hold true:*

1. *Decomposition of marginal log-likelihood[21].*

$$\log \rho_\theta = \mathcal{L}(\theta, q(z); x) + D_{KL}(q(z) \parallel \rho_\theta(z|x)) \tag{2.2.6}$$

2. *Bound on marginal log-likelihood.*

$$\begin{aligned} &\mathcal{L}(\theta, q(z); x) \leq \log \rho_\theta(x) \\ &\mathcal{L}(\theta, q(z); x) = \log \rho_\theta(x) \iff q(z) = \rho_\theta(z|x) \end{aligned} \tag{2.2.7}$$

[20] Laurent Girin et al. "Notes on the use of variational autoencoders for speech and audio spectrogram modeling" in: *DAFx 2019-22nd International Conference on Digital Audio Effects*. 2019. 1–8

[21] Radford M Neal and Geoffrey E Hinton. "A view of the EM algorithm that justifies incremental, sparse, and other variants" in: *Learning in graphical models*. Springer, 1998. 355–368

Proof. The proof of (1) is trivial:

$$\begin{aligned}\mathcal{L}(\theta, q(z); x) + D_{\mathrm{KL}}(q(z) \parallel \rho_\theta(z|x)) = \mathbb{E}_{q(z)}[\log \rho_\theta(x, z) - \log q(z)] \\ + \mathbb{E}_{q(z)}[\log q(z) - \log \rho_\theta(z|x)] = \mathbb{E}_{q(z)}\left[\log\left(\frac{\rho_\theta(x, z)}{\rho_\theta(z|x)}\right)\right] = \log \rho_\theta(x)\end{aligned} \tag{2.2.8}$$

where we used the definition of conditional probability and the fact that the expectation is computed in the latent space. (2) is a direct consequence of (2.2.6) since the KL divergence is non-negative and identically zero just when $q(z) = \rho_\theta(z|x)$. □

Thanks to such results, an estimate of the log-likelihood can be obtained using the Expectation-Maximization (EM) algorithm[22]: (E) step corresponds to solve the unconstrained variational problem at fixed θ

$$q_*(z) = \arg\max_{q \in \mathcal{F}} \mathcal{L}(\theta, q(z); x) \tag{2.2.9}$$

while (M) step to maximization of ELBO w.r.t. θ at fixed $q(z)$. To be precise, the output of the (E) steps is conditioned on x, which is $q(z) = q(z|x)$. It can be theoretically proven that under suitable condition such an algorithm converges to the optimum and satisfies the equality in (2.2.7).
For now there is no evident advantage: solving an explicit variational optimization problem can be unfeasible as the computation of (2.2.3). But further simplifications are possible: in so-called *fixed-form variational* inference[23], the variational family $\mathcal{F}$ is constrained to be any parametric family of PDFs $q_\lambda(z|x)$ dependent on $\lambda \in \Lambda$; e.g. for the gaussian family $q_\lambda(z|x) = \mathcal{N}(z; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ we have $\lambda = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. The advantage is that one can perform the (E) step as optimizing λ and not in a function class, and possibly find

$$\lambda^* = \arg\max_{\lambda} \mathcal{L}(\theta, \lambda; x) \tag{2.2.10}$$

Since we have to deal with a dataset of N data point, we rewrite

$$\mathcal{L}(\theta, \lambda; \mathcal{X}) = \sum_{i=1}^{N} \mathcal{L}(\theta, \lambda_i; x_i) \tag{2.2.11}$$

and ideally perform gradient-based optimization routines both in (E) and (M) step. But we immediately notice that optimizing the "local" λ_i for each sample if N is big is very impractical: for instance, for the gaussian class in dimension d we should update N means and covariance matrices, that is $Nd^2(d+1)/2$ scalars.
Thus, a last assumption is necessary to practically train the generative model, leading to the so-called *amortized variational inference*. It corresponds to assume that there

[22] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, **39**: 1–22, 1977.

[23] Antti Honkela et al. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *The Journal of Machine Learning Research*, **11**: 3235–3268, 2010.

exists a parametric map f_ϕ such that $\lambda_i = f_\phi(x_i)$. In this way, the definitive learning objective for EM algorithm is

$$\mathcal{L}(\theta, \lambda; \mathcal{X}) = \sum_{i=1}^{N} \mathcal{L}(\theta, \phi; x_i) = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(z_i|x_i)}[\log \rho_\theta(x_i, z_i) - \log q_\phi(z_i|x_i)] \tag{2.2.12}$$

Summarizing this first part, we started from the problem of training the decoder network $\rho_\theta(x|z)$ and we had to face the issue of computing the marginal log-likelihood. Thanks to a reformulation of the problem, we could explicit an equivalent objective (2.2.12). Given $q_\phi(z|x)$, that is the approximation of the intractable posterior $\rho_\theta(z|x)$, $\mathcal{L}(\theta, \lambda; x)$ can be attacked via EM algorithm, i.e. alternatively optimizing θ and ϕ.
VAE approach can be seen as a particular case of amortized variational inference in which $q_\phi(z|x)$ is approximated via a neural network, which by analogy with AE is denoted as *encoder network*. Similarly to the decoder network, a widely used choice is a gaussian, i.e.

$$\rho_\phi(z|x) = \mathcal{N}(z, \boldsymbol{\mu}_\phi(x), \mathrm{diag}\{\boldsymbol{\sigma}^2_\phi(x)\}) \tag{2.2.13}$$

where mean and covariance are modelled by a NN. The proposal to train a VAE[19] is to perform gradient-based optimization on the joint set of parameters $\{\theta, \phi\}$ with (2.2.12) as objective. Since the encoder and decoder are in cascade, the joint training can be suboptimal[24] with respect to the alternating routine in EM algorithm.
Despite this drawback, using the definition of KL divergence and conditional probability, we rewrite (2.2.12) as

$$\mathcal{L}(\theta, \lambda; \mathcal{X}) = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(z_i|x_i)}[\log \rho_\theta(x_i|z_i)] - \sum_{i=1}^{N} D_{\mathrm{KL}}[q_\phi(z_i|x_i) \parallel \rho(z)] \tag{2.2.14}$$

The two summations can be easily interpreted: the first one is related to reconstruction accuracy and measures the fidelity of encoding and decoding chain; the second one is a regularization term that forces the posterior (encoder) to be close to the prior, which is a set of independent gaussians — ideally, each z entry should encode an independent characteristic of the data.
Regarding the actual implementation of a gradient routine, the whole point of ELBO reformulation was the intractability of the marginal likelihood. Thus, we have to ensure to not have the same issue for $\mathcal{L}$. The regularization term has an analytical expression for the usual mentioned choices for $q_\phi(z|x)$ and $\rho(z)$ (e.g. if it is the KL divergence between gaussian densities). Thus, the computation of the gradient of that summation w.r.t. to θ or ϕ is not a problem for backpropagation algorithm (n.b. we are dealing with NN). On the other hand, the first summation in analytically intractable: the only possibility is the use of a Monte Carlo estimate using samples $\{z^{(r)_i}\}_{r=1}^{R}$ drawn from $q_\phi(z_i|x_i)$. Sampling from a gaussian encoder is an easy task, but unfortunately it is not a differentiable operation and it poses an obstacle to perform backpropagation w.r.t. ϕ. The solution to this last issue is the following reparametrization trick:

$$z_i^{(r)} = \boldsymbol{\mu}_\phi(x_i) + \mathrm{diag}\{\boldsymbol{\sigma}^2_\phi(x)\}^{\frac{1}{2}} \epsilon^{(r)} \qquad \epsilon^{(r)} \sim \mathcal{N}(\mathbf{0}_L, \boldsymbol{I}_L) \tag{2.2.15}$$

[24] Junxian He et al. "Lagging Inference Networks and Posterior Collapse in Variational Autoencoders" in: *International Conference on Learning Representations*. 2018.

which allows to effectively compute the gradient w.r.t. ϕ. The resulting empirical estimate of $\mathcal{L}(\theta, \lambda; \mathcal{X})$ is

$$\hat{\mathcal{L}}(\theta, \lambda; \mathcal{X}) = \sum_{i=1}^{N} \frac{1}{R} \sum_{i=r}^{R} \log \rho_\theta(x_i | z_i^{(r)}) - \sum_{i=1}^{N} D_{\mathrm{KL}}[q_\phi(z_i|x_i) \parallel \rho(z)], \tag{2.2.16}$$

which is the objective for the joint optimization of θ and ϕ.
After some manipulation, we conclude that VAEs can be trained on log-likelihood objective. The main strength appears to be the ease of generation, since for common choices of encoder and decoder such task reduces to sample from a gaussian distribution. In fact, the main drawbacks[25–27] of VAEs lays in the training phase. First of all, VAEs have several hyperparameters (e.g., the choice of prior, a possible imbalanced weighting of the reconstruction and regularization terms) that can significantly impact their performance. Finding the optimal set of hyperparameters can be a challenging task. The assumed simple structure of the latent space in VAEs might not capture the complex dependencies present in the data, limiting the expressiveness of the learned representations. Plus, achieving perfect disentanglement remains a challenge. The latent variables might still be entangled, making it challenging to control specific factors independently. Empirically, it is observed that VAEs sometimes generate blurry samples or suffer from mode collapse, where the model focuses on capturing only a few modes of the data distribution, neglecting others. In general it seems to be an issue related to their limited capacity: they might struggle with capturing complex and high-dimensional data distributions effectively, especially when compared to other generative models.

2.2.2 Generative Adversarial Networks

Generative adversarial networks[28] (GANs) are a class of generative models which take inspiration from game theory. They consist of two neural networks (see Figure 2.3), namely a *generator* G and a *discriminator* D, trained simultaneously through the so-called adversarial training. Given a dataset $\mathcal{X}$ sampled from the unknown data distribution ρ_*, the generator is devoted to generate synthetic data that ideally resembles the training data. On the other hand, the discriminator has to discern between fake and true samples. In this sense, G and D are adversary: the generator aims to produce realistic data to fool the discriminator, while the discriminator strives to correctly classify real and fake data. Thus, the training ends when the discriminator becomes unable to effectively distinguish between real and generated samples. Let us present the mathematical formulation: firstly we define a prior $\rho_z(z)$, which is a PDF

[25] Ruoqi Wei et al. Variations in variational autoencoders-a comparative evaluation. *Ieee Access*, **8**: 153651–153670, 2020.
[26] Achraf Oussidi and Azeddine Elhassouny. "Deep generative models: Survey" in: *2018 International conference on intelligent systems and computer vision (ISCV)*. IEEE 2018. 1–8
[27] Saptarshi Sengupta et al. A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowledge-Based Systems*, **194**: 105596, 2020.
[28] Ian Goodfellow et al. Generative adversarial nets. *Advances in neural information processing systems*, **27**: 2014.

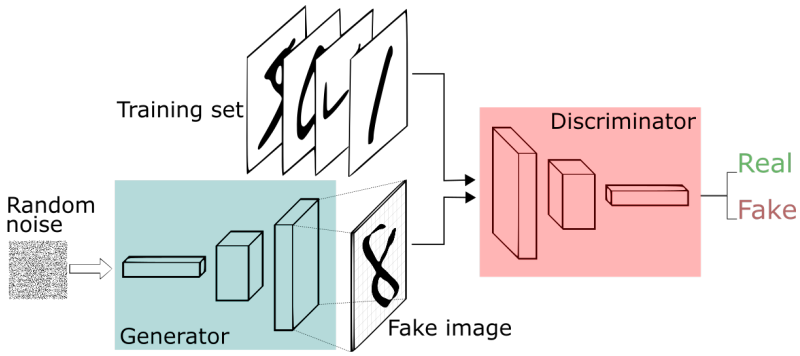


Figure 2.3: Scheme of the structure of GANs[29].

easy to sample from that serve to inject noise into the generator. The latter is a function $G_{\theta_g}(z)$ that is fed with noise and generate "fake" samples that should be similar to samples from ρ_*. The discriminator is a parametric function $D_{\theta_d}(x)$ that gives the probability that a sample x comes from the training set rather than have been generated by G. Both θ_g and θ_d are parameters of a NN. The optimal weights are solution of the following two-player minimax problem:

$$\arg\min_{\theta_g}\arg\max_{\theta_d}\mathbb{E}_*[\log D_{\theta_d}(x)]+\mathbb{E}_{\rho_z}[\log(1-D_{\theta_d}(G_{\theta_g}(z)))] \coloneqq \arg\min_{\theta_g}\arg\max_{\theta_d} V(G,D) \tag{2.2.17}$$

We refer in the following to $\rho_g(x)$ as the distribution of "fake" samples induced by the generator, that is such that

$$\mathbb{E}_{\rho_z}[\log(1-D_{\theta_d}(G_{\theta_g}(z)))] = \mathbb{E}_{\rho_g}[\log(1-D_{\theta_d}(x))] \tag{2.2.18}$$

The empirical idea to solve the minimax game is via an alternating algorithm:

Proposition 2.2.1. *The optimization algorithm for a GAN is made by two alternating steps:*

- **Update of the discriminator**
  1. *Sample* $\{z^{(i)}\}_{i=1}^N$ *(noise) from* ρ_z *and* $\{x^{(i)}\}_{i=1}^N$ *(data) from* ρ_*.
  2. *Compute* $\nabla_{\theta_d} V(G,D)$ *and perform* **gradient ascent** *to update* θ_d.
- **Update of the generator**
  1. *Sample* $\{z^{(i)}\}_{i=1}^N$ *(noise) from* ρ_z.
  2. *Compute* $\nabla_{\theta_g} V(G,D)$ *and perform* **gradient descent** *to update* θ_g.

This proposal is driven by common sense, but a more careful analysis of the minimax game is necessary to ensure convergence of such algorithm. In order to characterize the solutions of this adversarial game, it is necessary to search for the optima. The

method of proof is: (1) classify solutions of optimization of D at fixed G and viceversa and then (2) present a convergence result of the alternating game. Let us start from the update of the discriminator:

Theorem 2.2.1 (Existence of optimal discriminator[28]). *For G fixed, the optimal discriminator D is*

$$D_G^* = \frac{\rho_*(x)}{\rho_*(x) + \rho_g(x)} \tag{2.2.19}$$

Proof. Using (2.2.17) and (2.2.18), we have

$$V(G, D) = \int_\Omega (\rho_*(x) \log D_{\theta_d}(x) + \rho_g(x)(1 - D_{\theta_d}(x)))dx \tag{2.2.20}$$

The function $y \to a\log(y) + b\log(1-y)$ achieve its maximum in $(0,1)$ at $a/(a+b)$ for $(a,b) \neq (0,0)$. Applied to the case in study, the discriminator can be defined just in $Supp(\rho_*(x)) \cup Supp(\rho_g(x))$, hence concluding the proof. □

This lemma ensure that the gradient ascending will eventually reach a maximum, that is

$$C(G) = \arg\max_D V(G, D) = \mathbb{E}_* \left[\frac{\rho_*(x)}{\rho_*(x) + \rho_g(x)} \right] + \mathbb{E}_{\rho_g} \left[\frac{\rho_g(x)}{\rho_*(x) + \rho_g(x)} \right] \tag{2.2.21}$$

Now we need to characterize the solutions of the minimization problem $\arg\min_G C(G)$

Theorem 2.2.2 (Existence of optimal generator[28]). *At fixed $D = D_G^*$, the optimal generator G^* induce a ρ_g such that $\rho_g = \rho_*$. At that point, $C(G^*) = -\log 4$.*

Proof. Regarding the last point, for $\rho_g = \rho_*$ we obtain $D_G^* = 1/2$, that inserted in $C(G)$ gives exactly $-\log 4$. We need to test whether this is a global optimum: we can sum and subtract $-\log 4$ to $C(G)$ obtaining

$$\begin{aligned} C(G) &= -\log(4) + D_{KL}\left(\rho_* \,\middle\|\, \frac{\rho_* + \rho_g}{2}\right) + D_{KL}\left(\rho_g \,\middle\|\, \frac{\rho_* + \rho_g}{2}\right) \\ &= -\log(4) + 2 \cdot JSD(\rho_* \parallel \rho_g) \end{aligned} \tag{2.2.22}$$

where JSD is the Jensen-Shannon divergence[30]. Such quantity has the same non-negativity property of KL divergence, i.e. $JSD(\rho_* \parallel \rho_g) \geq 0$ and $JSD(\rho_* \parallel \rho_g) = 0$ iff $\rho_g = \rho_*$. This proves that $\rho_g = \rho_*$, or more precisely the corresponding generator G^*, is the global minimum for $C(G)$. □

To summarize, we have showed separate theoretical guarantees about convergence of gradient ascent and descent. However, we need to show that alternating those two steps would eventually converge to the global Nash equilibrium of the minimax game, i.e. $\rho_* = \rho_g$. The result is summarized in the following Theorem[28,31] of which omit the proof for the sake of brevity.

[30] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[31] URL: `https://granadata.art/gan-convergence-proof/#/`

Theorem 2.2.3. *If G and D have enough capacity, and at each step of the alternating algorithm, the discriminator is allowed to reach its optimum given G, and ρ_g is updated so as to improve the criterion*

$$\mathbb{E}_*[\log D_G^*(x)] + \mathbb{E}_{\rho_g}[\log(1 - D_G^*(x)] \tag{2.2.23}$$

then ρ_g converges to ρ_*.

Ideally, the theoretical treatment of Generative Adversarial Networks (GANs) concludes with the proof that the proposed minimax game has a unique Nash equilibrium. This equilibrium corresponds to a generator capable of sampling from ρ_*, making it indistinguishable from true samples by the discriminator, performing no better than a random classifier with a probability of 1/2.
We now discuss the main drawbacks[31–34] of GANs. Firstly, practical application of Theorem 2.2.3 reveals immediate limitations. In practice, optimization involving gradients is executed in parameter space on θ_g rather than in functional space on ρ_g. This deviation introduces challenges, as a convex problem in probability space may become non-convex, especially when using deep neural networks to model G: in fact, the induced loss function becomes inherently non-convex. Additionally, a numerical issue arises when attempting to find the perfect discriminator D_G^* at a fixed G; backpropagation to train the generator (specifically because the term $D(G_{\theta_g}(z))$) may yield gradients close to zero by definition at the beginning of training when the generator is very poor.
Regarding practical aspects, GAN training is notorious for its instability. Achieving the right balance between the generator and discriminator can be delicate, leading to oscillations during the training process and making it difficult to converge to a stable solution. This instability often requires careful tuning of hyperparameters, adding an extra layer of complexity to the training process. Additionally, GANs often require large and diverse datasets for training to generalize well.
Also generating samples from a trained GAN poses significant challenges. A critical one is mode collapse, where the generator tends to produce a limited set of outputs, neglecting the diversity present in the training data. This results in generated samples lacking variety and richness. Furthermore, GANs can be computationally intensive, especially when dealing with high-resolution images or complex datasets. This computational demand can be a hindrance for researchers and practitioners with limited resources, both in terms of time and hardware. Ultimately, evaluating the performance of a GAN can be problematic. Common metrics like Inception Score and Frechet Inception Distance have limitations, and there is no universally accepted metric for assessing the quality of generated samples. This lack of clear evaluation criteria makes

[32] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks" in: *Proceedings of the 5th International Conference on Learning Representations Workshop Track*. 2016.

[33] Tim Salimans et al. "Improved techniques for training GANs" in: *Advances in Neural Information Processing Systems*. 2016. 2226–2234

[34] Martin Arjovsky and Léon Bottou. "Towards principled methods for training generative adversarial networks" in: *Neural Information Processing Systems Conference Workshop: Adversarial Training*. 2016.

it challenging to compare different GAN models effectively.
Despite the mentioned issues, the adversarial paradigm represents an important concept in unsupervised learning, in particular in relation with robustness of pre-trained generative models[35], and generally machine learning models.

2.2.3 Diffusion Models

Diffusion generative models[36–38] typically refer to a class of generative models that leverage the concept of diffusion processes. In the context of generative models, diffusion processes involve the transformation of a simple distribution into a more complex one over time. This transformation occurs through a series of steps, each representing a diffusion process. The overarching idea is to initiate the process with a basic distribution, such as Gaussian noise, and iteratively transform it to approximate the target distribution, often representing real data like images. In recent years they have become state of the art in many domains of application, partially substituting GANs[39]. In this section we provide a summary of the main common features of diffusion models, without entering too much in details about every single specification currently available.
As for other generative models, the main ingredient is a dataset $\mathcal{X} = \{x_i\}_{i=1}^N$ where x_i are sampled from an unknown target density $\rho_*(x)$. We will assume $\mathcal{X} \subset \mathbb{R}^d$ for simplicity. Both for VAEs and GANs, the idea is to generate new samples from noise, that is respectively decoding from a gaussian in latent space, or generate from noise via G in GANs. In diffusion models, the objective is again to push samples extracted from a simple distribution, like a gaussian, towards the data distribution.
Since the main content of the following will be strictly related to stochastic calculus[40], let us fix the notation. We will refer to $X_t \in \mathbb{R}^d$ as a stochastic process, that is a sequence of random variables, where $t \in \mathbb{R}$ is the continuous time variable. Differently from deterministic processes, the focus is on the distribution in law of X_t, namely $\rho(x,t)$, and not on the single trajectory. As deterministic trajectories can be solutions of ordinary differential equations (ODEs), a stochastic process can be solution of a stochastic differential equation (SDE).

Proposition 2.2.2 (SDE and Fokker-Plank PDE)**.** *Given the drift $\mu : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ and the diffusion matrix $\sigma : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d,d}$, let us consider the stochastic process X_t*

[35] Aleksander Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks" in: *International Conference on Learning Representations*. 2018.
[36] Jascha Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics" in: *International conference on machine learning*. PMLR 2015. 2256–2265
[37] Ling Yang et al. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, **56**: 1–39, 2023.
[38] Florinel-Alin Croitoru et al. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[39] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, **34**: 8780–8794, 2021.
[40] L Chris G Rogers and David Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*. vol. 2 Cambridge university press, 2000.

solution for $t \in [0,T] \subset [0,+\infty]$ *of the SDE*

$$dX_t = \mu(X_t,t)dt + \sigma(X_t,t)dW_t, \qquad X_0 \sim \rho_0 \tag{2.2.24}$$

where W_t *is a Wiener process. Using Ito convention, the law of* X_t*, namely* $\rho(x,t)$*, satisfies the Fokker-Planck partial differential equation (PDE)*

$$\frac{\partial}{\partial t}\rho(x,t) = -\nabla \cdot [\mu(x,t)\rho(x,t)] + \Delta\left[\frac{\sigma(x,t)^2}{2}\rho(x,t)\right], \qquad \rho(x,0) = \rho_0(x) \tag{2.2.25}$$

This proposition is important to understand the relation between the single random process X_t and its distribution in law. Let us present a simple example to clarify such connection.

Example 2 (Wiener process)**.** Let us consider the case $\mu(x,t) = 0$ and $\sigma(x,t) = 1$ in $d = 1$, that corresponds to the SDE

$$dX_t = dW_t \tag{2.2.26}$$

The solution of the associated Fokker-Planck equation

$$\frac{\partial \rho(x,t)}{\partial t} = \frac{1}{2}\frac{\partial^2 \rho(x,t)}{\partial x^2}, \tag{2.2.27}$$

for a delta initial datum $\rho(x,0) = \delta(x)$ is precisely

$$\rho(x,t) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} \tag{2.2.28}$$

This is a gaussian density with variance proportional to t. That is, the initial concentrated density spreads on the real line.

This brief summary about SDEs is sufficient to provide a consistent definition of generative diffusion model:

Definition 2.2.2 (Generative diffusion model)**.** *Let us consider the data distribution* $\rho_* : \mathbb{R}^d \to \mathbb{R}_+$ *and a base distribution* $\bar{\rho}(x) : \mathbb{R}^d \to \mathbb{R}_+$*. Given a time interval* $[0,T] \in [0,\infty]$*, a generative diffusion model is an SDE with fixed terminal condition*

$$dX_t = \mu(X_t,t)dt + \sigma(X_t,t)dW_t, \qquad X_0 \sim \bar{\rho}, \quad X_T \sim \rho_* \tag{2.2.29}$$

where W_t *is a Wiener process.*

This definition resembles concepts from stochastic optimal control[41]: in fact, the terminal condition is not sufficient to uniquely fix $\mu(X_t,t)$ and $\sigma(X_t,t)$. Under this point of view, the specification of a particular class of diffusion models reduces to a *prescription on how to determine the drift and the diffusion matrix.* In the following we will summarize two highlighted methods present in literature.

[41] Wendell H Fleming and Raymond W Rishel. *Deterministic and stochastic optimal control.* vol. 1 Springer Science & Business Media, 2012.

Score-based diffusion[42]. To explain what is score-based diffusion we need the following preliminaries:

Definition 2.2.3. *Given a PDF $\rho(x)$, the* **score** *is the vector field*

$$s(x) = \nabla \log \rho(x) \tag{2.2.30}$$

Proposition 2.2.3 (Naive score-based diffusion)**.** *For any $\varepsilon > 0$ and $\rho_0(x)$, the choice $\mu(x,t) = \varepsilon s_*(x) = \varepsilon \nabla \log \rho_*(x)$ and $\sigma = \sqrt{2\varepsilon}$ in (2.2.29) satisfies the endpoint condition for $T = \infty$.*

Proof. If we consider the Fokker-Planck PDE associated to (2.2.29) with the selected drift and variance, we have

$$\partial_t \rho(x,t) = \nabla \cdot [-s_*(x)\rho(x,t) + \nabla \rho(x,t)] = \nabla \cdot \left[\rho(x,t) \nabla \log \left(\frac{\rho(x,t)}{\rho_*(x)} \right) \right] \tag{2.2.31}$$

By direct substitution, the stationary probability density $\rho_*(x)$ is a solution. For uniqueness, we need to prove that any solution of the PDE would converge to this solution. A formal argument is based on Jordan-Kinderlehrer-Otto (JKO) variational formulation of Fokker-Planck equation[43], interpreted as a gradient flow in probability space with respect to Wasserstein-2 distance. An alternative way is the following: for any solution $\rho(x,t)$, we can compute the time derivative of the KL divergence between such solution and $\rho_*(x)$. If we define $R = \rho/\rho_*$:

$$\frac{d}{dt} D_{\mathrm{KL}}(\rho \parallel \rho_*) = \frac{d}{dt} \int_{\mathbb{R}^d} \rho \log R \, dx = \int_{\mathbb{R}^d} \partial_t \rho \log R \, dx + \int_{\mathbb{R}^d} \frac{\rho}{R} \partial_t R \, dx \tag{2.2.32}$$

We can use Fokker-Planck equation to substitute $\partial_t \rho$ and integrate by parts:

$$\frac{d}{dt} D_{\mathrm{KL}}(\rho \parallel \rho_*) = \frac{d}{dt} \int_{\mathbb{R}^d} \rho \log R \, dx = -\int_{\mathbb{R}^d} \rho |\nabla \log R|^2 \, dx + \int_{\mathbb{R}^d} \rho_* \partial_t R \, dx \tag{2.2.33}$$

We notice that $\rho_* \partial_t R = \nabla \cdot (\rho \log R)$, hence that the second addend is zero by integration by parts. The conclusion is that

$$\frac{d}{dt} D_{\mathrm{KL}}(\rho \parallel \rho_*) = -\int_{\mathbb{R}^d} \rho |\nabla \log R|^2 \, dx \leq 0, \tag{2.2.34}$$

concluding the proof. □

The result seems to say that we are able to build a diffusion generative models estimating the score of the target. In a data driven context, ρ_* is known just through data points and one has to face the problem of estimating s_*. A possible approach[44] is score matching.

[42] Yang Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations" in: *International Conference on Learning Representations*. 2020.

[43] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, **29**: 1–17, 1998.

[44] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**: 2005.

Definition 2.2.4 (Fisher divergence)**.** *Given two PDFs* $\rho(x)$ *and* $\pi(x)$*, the* **Fisher divergence** *is defined as*

$$D_F(\rho \parallel \pi) = \int_{\mathbb{R}^d} \rho(x) \|\nabla \log \rho(x) - \nabla \log \pi(x)\|^2 \, dx \tag{2.2.35}$$

Even if in some sense D_F seems to measure some distance between two PDFs, it is very different from the KL divergence, see following Box.

Fisher Divergence versus KL divergence

By definition, both KL and Fisher divergence between two PDFs satisfy the non-negativity property, i.e. they are strictly positive, and zero only when the densities are the same. D_F does not depend on normalization constants of the PDFs because of the gradients. This is a double-edged weapon: it is apparently useful in high dimension, where the computation of normalization of a density is impractical (as for instance the partition function for EBMs). But if the distribution is multimodal, the local nature of D_F is very insensible to global characteristics of the densities, as for instance the relative mass in each mode. Let us consider a key example: the distributions we would like to compare are:

$$\begin{aligned} \rho_1(x) &= 0.5\mathcal{N}(x, -5, 1)(x) + 0.5\mathcal{N}(x, 5, 1), \\ \rho_2(x) &= \sigma(z)\mathcal{N}(x, -5, 1) + (1 - \sigma(z))\mathcal{N}(x, 5, 1) \end{aligned} \tag{2.2.36}$$

where $\sigma(z) = 1/(1 + e^{-z})$ is a sigmoid function. The two densities are bimodal gaussian mixture in 1D with same means and variances; the second mixture is balanced with relative mass equal to $1/2$. We would like to compare $D_F(\rho_1 \parallel \rho_2)$ and $D_{\mathrm{KL}}(\rho_1 \parallel \rho_2)$ as functions of z. In Figure 2.4 we plot the two divergences in function of z. We estimate the expectations that define the two divergences using a Monte Carlo estimate, namely

$$\begin{aligned} D_F(\rho_1 \parallel \rho_2) &\approx \sum_{i=1}^{N} \|\nabla \log \rho_1(x_i) - \nabla \log \rho_2(x_i)\|^2 \\ D_{\mathrm{KL}}(\rho_1 \parallel \rho_2) &\approx \sum_{i=1}^{N} \log \left[\frac{\rho_1(x_i)}{\rho_2(x_i)} \right] \end{aligned} \tag{2.2.37}$$

where $x_i \sim \rho_1(x)$. The minimum value is 0 and corresponds to $z = 0$, that is $\rho_1 = \rho_2$. The first difference is that the values of D_F are smaller of several order of magnitude — in general, this could be a problem in practical implementations. Most importantly, the shape of the curve is very different. In this one dimensional example we need $N = 10000$ to appreciate a similar growth, even if D_F curve is more steep. For smaller N, D_F is basically flat for $z \neq 0$. This is related to the absence of points in low density regions, that is where the integrand in D_F gives a non-zero contribution.

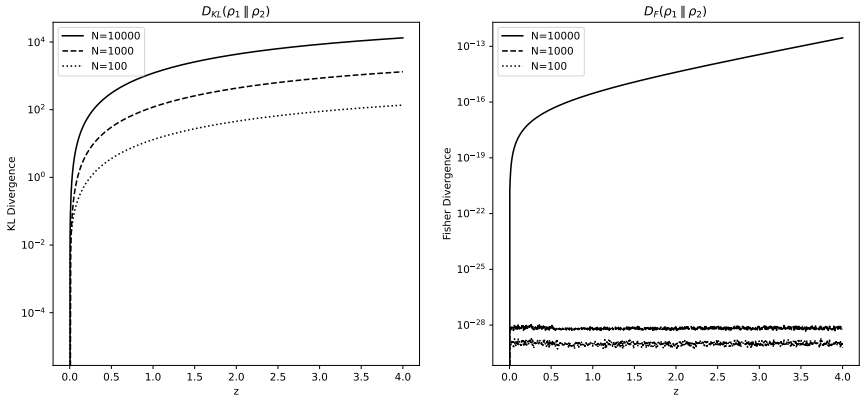


Figure 2.4: Comparison between KL divergence and Fisher divergence for the two bimodal gaussian mixtures in (2.2.36). The variable $z \in (0, \infty)$ is related to the relative mass of the two modes via a sigmoid function $\sigma(z) \in (0, 1)$; the plots for $z < 0$ are analogous by symmetry. Notice the different scales of the y axes. The Monte Carlo estimation is performed using $N = 100, 1000, 10000$ samples.

In score matching, one propose a parametric score s_θ, for instance a neural network, and train such model to match the true score s_*. The loss on which the model is trained, using for instance gradient routines, is

$$\mathcal{L}(s_\theta, \rho_*) = \frac{1}{2} \int_{\mathbb{R}^d} \rho_* \|s_\theta - \nabla \log \rho_*\|^2 \, dx = \mathbb{E}_*[\|s_\theta\|^2 + \mathrm{tr}(J_x s_\theta)] + C_p \tag{2.2.38}$$

where we integrated by parts, $C_p = const$ does not depend on θ and J_x is the Jacobian with respect to x. Denoting with ρ_θ one (n.b. not unique) PDF associated to s_θ, the loss is evidently $D_F(\rho_* \parallel \rho_\theta)$. The right hand side reformulation in (2.2.38) is crucial: the expectation $\mathbb{E}_*$ can be estimated using data points at our disposal, bypassing the problematic term $\nabla \log \rho_*$.
Unfortunately, the naive score-based approach is plagued by two fundamental issues that make it impractical. The first regards the score estimation itself: usually, data at our disposal comes from high density region of ρ_*, that is the estimation of $\mathbb{E}_*$, hence of the score, will be inaccurate outside such areas. The problem is that the initial condition (e.g. noise) of the SDE is usually located far from data. An imprecise drift will critically affect the generation process, leading to unpredictable outcomes. The second regards the difference between the PDE and the practical implementation through (2.2.29). The generation problem is convex in probability space, i.e. ρ_* is the unique asymptotic stationary solution, but the rate of convergence of the law of X_t is critically related to the particular ρ_* in study, in particular in relation with multimodality and slow mixing. We will discuss in details about this issue in Section 2.3.

The next step towards state-of-the-art score-based diffusion is the following lemma[45]:

Lemma 2.2.2. *Any SDE in the form*

$$dX_t = f(X_t, t)dt + g(t)dW_t, \qquad X_0 \sim \rho_1, \quad X_T \sim \rho_2 \tag{2.2.39}$$

with solution $X_t \sim \rho(x,t)$ *admits an associated reversed SDE*

$$dX_s = [f(X_s, s) - g(s)^2 \nabla \log \rho(x,s)]ds + g(s)dW_s, \qquad X_T \sim \rho_2, \quad X_0 \sim \rho_1 \tag{2.2.40}$$

where ds *is a negative infinitesimal time step and* s *flows backward from* T *to* 0. *By convention,* (2.2.39) *is also called forward SDE and* (2.2.40) *backward one.*

Exploiting this result, we can define a score-based diffusion model:

Definition 2.2.5. *A* **score-based generative model** *is the backward SDE* (2.2.40)*, where* $\rho_1 = \rho_*$ *and* $\rho_2 = \bar{\rho}$.

Apparently, the situation is even worse with respect to score matching: the score in (2.2.40) is related to the law of X_t, i.e. it is time dependent and generally not analytically known — score estimation was already an issue for $s_*(x)$. The core idea in score-based diffusion is to extract information about $\nabla \log \rho(x,s)$ from the forward process since the solutions of (2.2.39) and (2.2.40) have the same law, see Figure 2.5. By Definition 2.2.5, the forward process brings *data to noise* and the model to be

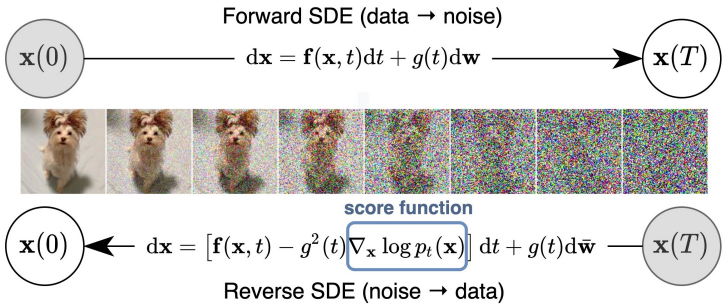


Figure 2.5: Schematic representation of forward and backward process in score-based diffusion. Image taken from Song et al., 2021[42].

learned is a time dependent parametric vector field $s_\theta(x,t)$. The loss used during the forward process to learn the score is:

$$\mathcal{L}_{SM}(s_\theta(x,t)) = \mathbb{E}_{t\sim\mathcal{U}(0,T)}\mathbb{E}_{\rho(x,t)}[\lambda(t)\|s_\theta(x,t) - \nabla \log \rho(x,t)\|^2] \tag{2.2.41}$$

where $\lambda : [0,T] \to \mathbb{R}_+$ is a positive scalar weight function and $U(0,T)$ is the uniform distribution in $(0,T)$. After the same integration by parts used in (2.2.41), there is

[45] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, **12**: 313–326, 1982.

still the problem that computing the hessian is expensive in high dimension, especially if s_θ is a neural network. Several proposal to solve this issue have been proposed and successfully exploited, such as denoising score matching[46] or sliced score matching[47]. Another subtle issue is that generally the forward process would generate pure noise for $T = \infty$ — one could be worried that the truncation at finite time would provide an imprecise estimate of the score at that time scale, that is close to noise, and induce errors during the generative phase. This problem is attacked by practitioners via several tricks but the theoretical results in this sense are not complete.
Let us provide a brief interpretation of why score-based diffusion works better than naive score matching (Proposition 2.2.3). Let us consider the simple case $f(x,t) = 0$ and $g(t) = e^t$; the resulting forward process is perturbing data with gaussian noise at increasing variance scale[8]. That is, time scale corresponds to amount of noise in this setup. We recall that the problem of naive score matching was the lack of data in low density region for the target density. In score-based diffusion one use perturbed data to populate those region and compute the score at each time scale that serves as bridge from $\bar{\rho}$ and the target ρ^*.

Stochastic Interpolants. Another more recent class of diffusion-based generative models are the stochastic interpolants[48]. Let us immediately provide a definition of such objects:

Definition 2.2.6. *Given two probability densities $\rho_1, \rho_2 : \mathbb{R}^d \to \mathbb{R}_+$, a* **stochastic interpolant** *between them is a stochastic process $X_t \in \mathbb{R}^d$ such that*

$$X_t = I(t, X_0, X_1) + \gamma(t) z \qquad t \in [0,1] \tag{2.2.42}$$

where:

- *The function I is of class C^2 on its domain and satisfy the following endpoint conditions*

$$I(i, X_0, X_1) = X_i \qquad i = 0, 1 \tag{2.2.43}$$

as well as

$$\exists C_1 < \infty \,:\, |\partial_t I\,(t, X_0, X_1)| \leq C_1\,|X_0 - X_1| \quad \forall\,(t, X_0, X_1) \in [0,1] \times \mathbb{R}^d \times \mathbb{R}^d \tag{2.2.44}$$

- *$\gamma : [0,1] \to \mathbb{R}$ is such that $\gamma(0) = \gamma(1) = 0$ and $\gamma(t) > 0$ for $t \in (0,1)$.*
- *The pair (X_0, X_1) is sampled from a measure ν that marginalizes on ρ_0 and ρ_1, that is $\nu(dX_0, \mathbb{R}^d) = \rho_0 dX_0$ and $\nu(\mathbb{R}^d, dX_1) = \rho_1 dX_1$.*

[46] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, **23**: 1661–1674, 2011.
[47] Yang Song et al. "Sliced score matching: A scalable approach to density and score estimation" in: *Uncertainty in Artificial Intelligence*. PMLR 2020. 574–584
[48] Michael Samuel Albergo and Eric Vanden-Eijnden. "Building Normalizing Flows with Stochastic Interpolants" in: *The Eleventh International Conference on Learning Representations*. 2022.

- *The variable z is a Gaussian random variable independent from (X_0, X_1), i.e. $z \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$ and $z \perp (X_0, X_1)$*

Let us focus on the case in which $\rho_0 = \bar{\rho}$ to be a simple base distribution (e.g. a Gaussian) and $\rho_1 = \rho_*$, that is the data distribution. Equation (2.2.42) means that if we sample a couple $X_0 \sim \rho_0$ and X_1 from the dataset, the interpolant is a stochastic process that connects the two points. The objective is to build a generative model that, in some sense, learns from the interpolants the way to map samples from $\bar{\rho}$ to ρ_*. The first important result in this sense is the following[48]:

Proposition 2.2.4. *The interpolant X_t is distributed at any time $t \in [0,1]$ following a time dependent density $\rho(x,t)$ such that $\rho(x,0) = \rho_0$ and $\rho(x,1) = \rho_1$, and also satisfies the following transport equation:*

$$\partial_t \rho + \nabla \cdot (b\rho) = 0 \tag{2.2.45}$$

where the vector field $b(x,t)$ is defined by a conditional expectation:

$$b(x,t) = \mathbb{E}\left[\dot{X}_t \mid X_t = x\right] = \mathbb{E}\left[\partial_t I\left(t, X_0, X_1\right) + \dot{\gamma}(t) z \mid X_t = X\right] \tag{2.2.46}$$

Proof. Let $g(k,t) = \mathbb{E}e^{ik\cdot X_t}$ the characteristic function of $\rho(x,t)$, that is

$$g(k,t) = \mathbb{E}e^{ik\cdot(I(t,X_0,X_1)+\gamma(t)z)} \tag{2.2.47}$$

If we compute the time derivative of g, we obtain

$$\partial_t g(k,t) = ik \cdot m(k,t) \tag{2.2.48}$$

where $m(k,t) = \mathbb{E}[(\partial_t I\left(t, X_0, X_1\right) + \dot{\gamma}(t) z) e^{ik\cdot X_t}]$. By definition of conditional expectation,

$$\begin{aligned} m(k,t) &= \int_{\mathbb{R}^d} \mathbb{E}[(\partial_t I\left(t, X_0, X_1\right) + \dot{\gamma}(t) z) e^{ik\cdot X_t} \mid X_t = x] \rho(x,t) dx \\ &= \int_{\mathbb{R}^d} e^{ik\cdot x} b(x,t) \rho(x,t) dx \end{aligned} \tag{2.2.49}$$

where we used the definition of b. If we insert $m(k,t)$ in (2.2.48) and we compute the Fourier anti-transform, we immediately obtain (2.2.45) in real space. □

Other properties of b can be proven, but for the sake of the present summary we will not delve into them. As usual we can identify $\bar{\rho}$ and ρ_* as base and data distributions. Thanks to the previous Proposition we can already define a diffusion-based generative model:

Lemma 2.2.3 (ODE Generative Model)**.** *Given Proposition 2.2.4 and $\rho(x,0) = \bar{\rho}$, the choice $\mu(X_t,t) = b(X_t,t)$ and $\sigma(X_t,t) = 0$ in (2.2.29) satisfies the endpoint condition for $T = 1$.*

Differently from score-based diffusion models, such ODE-based formulation does not involve stochasticity during generation. In fact, the ODE $\dot{X}_t = b(X_t, t)$ can be interpreted as a Normalizing Flow (see Section 2.6) where the pushforward is defined via a transport PDE. Interestingly, the ODE formulation is formally equivalent to an SDE formulation:

Lemma 2.2.4 (SDE Generative Model)**.** *For $\varepsilon > 0$, given Proposition 2.2.4, $\rho(x, 0) = \bar{\rho}$ and the score $s(x, t) = \nabla \log \rho(x, t)$, the choice $\mu(X_t, t) = b(X_t, t) + \varepsilon s(X_t, t)$ and $\sigma(X_t, t) = \sqrt{2\varepsilon}$ in (2.2.29) satisfies the endpoint condition for $T = 1$.*

Proof. Adding and subtracting the score to (2.2.45), we obtain for any $\varepsilon > 0$

$$\partial_t \rho + \nabla \cdot ((b + \varepsilon s - \varepsilon s)\rho) = 0 \tag{2.2.50}$$

But $s\rho = \nabla \rho$, that is

$$\partial_t \rho + \nabla \cdot ((b + \varepsilon s)\rho - \varepsilon \nabla \rho) = 0 \tag{2.2.51}$$

Trivially, the solution of the PDE (2.2.51) is the law of a stochastic process solution of an SDE as in (2.2.29). □

We presented the proof as an example of the standard trick used to convert the diffusion term into a transport term exploiting the score.
We defined the generative model, but similarly to score-based diffusion, we need to clarify how b and s are learned in practice from data. For such purpose, we present the following result:

Proposition 2.2.5. *The vector field $b(x, t)$ is the unique minimizer of the following objective loss*

$$\mathcal{L}_b[\hat{b}] = \int_0^1 \mathbb{E}\left(\frac{1}{2}\left|\hat{b}\left(t, X_t\right)\right|^2 - \left(\partial_t I\left(t, X_0, X_1\right) + \dot{\gamma}(t) z\right) \cdot \hat{b}\left(t, X_t\right)\right) dt \tag{2.2.52}$$

Similarly, the the score $s(x, t)$ the unique minimizer of the following objective loss

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left(\frac{1}{2}\left|\hat{s}\left(t, X_t\right)\right|^2 + \gamma^{-1}(t) z \cdot \hat{s}\left(t, X_t\right)\right) dt \tag{2.2.53}$$

For the sake of the present summary, we will not present the proof[48]. The take home message is that one can now propose two neural networks, namely $b_\theta(x, t)$ and $s_{\theta'}(x, t)$, and train them through backpropagation using (2.2.52) and (2.2.53). The integrals are estimated using random pairs $(X_0, X_1) \sim \nu$ and times $t \sim \mathcal{U}[0, 1]$. As for score-based diffusion, we avoid delving into practical details regarding the implementation of the neural networks. We emphasize the main message: it is feasible to construct a diffusion model defined in a finite time interval that does not solely rely on the score function. In fact, score-based diffusion can be viewed as a specific instance of stochastic interpolation or similar methods (refer to Section 2.2.5 for more details).
Concerning practical aspects, the freedom in choosing the function $I(t, X_0, X_1)$ as well as $\gamma(t)$ can be challenging due to the absence of a general guiding principle.

Unfortunately, the structure of the interpolant and the implementation of b_θ and $s_{\theta'}$ can significantly impact the efficient training of the model. Regarding the generative phase, the SDE and ODE formulations are formally equivalent, but the practical choice is not straightforward. From a numerical perspective, the primary issue lies in the time discretization and integration of the differential equations. The ODE is preferred since integration methods are more stable and precise compared to those for SDEs; this allows for larger time steps and accelerated generation. This is also a significant advantage of stochastic interpolants over score-based diffusion, which is SDE-based. However, the presence of noise appears to be necessary as regularization: in layman's terms, since b is learned and possibly imperfect, any mismatch is "smoothed" in the SDE setting by the presence of noise. The value of ε functions as a hyperparameter in this context.
In conclusion, stochastic interpolants provide a general framework closely related to other diffusion models, such as score-based diffusion, flow matching[49,50], or Schrödinger bridge[51]. However, some common issues of diffusion-based generative models persist: slow generation, dependence on hyperparameters and neural architectures, and data dependence are the primary drawbacks.

2.2.4 Normalizing Flows

The fundamental idea underlying Normalizing Flows[52,53] (NF) is very close to the usual in generative modelling: to transform samples from a straightforward base distribution, often a Gaussian, to data distribution. The main feature of NF is that the transformation is performed through a series of invertible and differentiable transformations.
The core concept revolves around constructing a model capable of learning a sequence of *invertible* operations that can map samples from a simple distribution to the target distribution. In particular, we recall the well-known lemma:

Lemma 2.2.5. *Let us consider a random variable $Z \in \mathbb{R}^d$ and its associated probability density function $\rho_Z(z)$. Given an invertible function $Y = \phi(Z)$ on $\mathbb{R}^d$, the probability density function in the variable Y is defined through*

$$\rho_Y(y) = \rho_Z(g^{-1}(y))|\det J_y\varphi^{-1}(y)| = \rho_Z(\phi^{-1}(y))|\det J_y\phi(\phi^{-1}(y))|^{-1} \tag{2.2.54}$$

where ϕ^{-1} is the inverse of ϕ and J_y is the Jacobian w.r.t. y. The density ρ_Y is also called **pushforward** *of ρ_Z by the function ϕ and denoted by $\phi_\#\rho_Z$.*

[49] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
[50] Yaron Lipman et al. "Flow Matching for Generative Modeling" in: *The Eleventh International Conference on Learning Representations*. 2022.
[51] Valentin De Bortoli et al. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, **34**: 17695–17709, 2021.
[52] Esteban G. Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, **8**: 217–233, 2010.
[53] E. G. Tabak and Cristina V. Turner. A Family of Nonparametric Density Estimation Algorithms. *Communications on Pure and Applied Mathematics*, **66**: 145–164, 2013.

In generative modelling, ρ_Z is identified with the base distribution and its pushforward as the target, i.e. data, distribution. The direction from noise to data is called generative direction, while the other way is called normalizing direction — data are normalized, gaussianized, by the inverse of ϕ. The name Normalizing Flow originates from the latter. In fact, the mathematical foundation of NF is reduced to Lemma 2.2.5.

The whole problem reduces to design the pushforward in a data driven setup, that is where we just have a dataset $\mathcal{X}$ of samples from the target and no access to the analytic form of ρ_*. In order to link NF with other generative models, let us denote with ϕ_θ with $\theta \in \Theta$ the parametric map that characterizes the pushforward $\rho_\theta = (\phi_\theta)_\# \rho_Z$. In practice, this map is usually a neural network and ρ_θ will implicitly depend on it. The optimal parameters θ^* are chosen to be solution of the following optimization problem:

$$\theta^* = \arg\min_{\theta\in\Theta} D_{\mathrm{KL}}(\rho_*(x) \parallel \rho_\theta(x)) = \arg\max_{\theta\in\Theta} \mathbb{E}_*[\log \rho_\theta(x)] \tag{2.2.55}$$

As already stressed, this formulation in term of maximum log-likelihood is equivalent to cross-entropy minimization for EBMs. As for VAEs, the analytical form of ρ_θ is not known: in NF it is implicitly defined through the pushforward. This issue is attacked using Lemma 2.2.5 to rewrite the right hand side in (2.2.55) as

$$\arg\max_{\theta\in\Theta} \mathbb{E}_*[\log \rho_\theta(x)] = \arg\max_{\theta\in\Theta} \mathbb{E}_*[\log \rho_Z(\phi_\theta^{-1}(y)) + \log|\det J_y \phi^{-1}(y)|] \tag{2.2.56}$$

The likelihood of a sample under the base measure is represented as the first term, and the second term, often referred to as the log-determinant or volume correction, accommodates the alteration in volume resulting from the transformation introduced by the normalizing flows. After this manipulation every addend inside the expectation is calculable — the map ϕ and the noise distribution ρ_Z are given (e.g. a gaussian). As usual, the expectation can estimated via Monte Carlo using the finite dataset $\mathcal{X}$ at our disposal. Any gradient based optimization routine can be then exploited to optimize θ. During training, the model adjusts the parameters θ to bring the transformed distribution in close alignment with the true data distribution.

The main limitation in NF is that the pushforward map must be bijective for any θ.

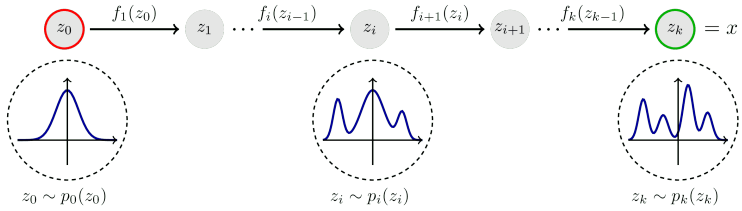


Figure 2.6: Schematic representation of Normalizing Flows, image taken from `https://flowtorch.ai/users/`[54].

Not only that: both forward and inverse operations are required to be computationally feasible to perform generation and normalization. Furthermore, the Jacobian determinant must be tractable to facilitate efficient computation. These requests constrain

the possible neural architectures that one can use to model ϕ_θ. The following lemma provides a decisive tool in this sense.

Lemma 2.2.6. *Let us consider a set of M bijective functions $\{f_i\}_{i=1}^M$. If we denote with $f = f_M \circ f_{M-1} \circ \cdots \circ f_1$ their composition, one can prove that f is bijective and its inverse is*

$$f^{-1} = f_1^{-1} \circ \cdots \circ f_{M-1}^{-1} \circ f_M^{-1} \tag{2.2.57}$$

Moreover, if we denote with $x_i = f_i \circ \cdots \circ f_1(z) = f_{i+1}^{-1} \circ \cdots \circ f_M^{-1}$ and $y = x_M$, we have

$$\det J_y f^{-1}(y) = \prod_{i=1}^{M} \det J_y f_i^{-1}(x_i) \tag{2.2.58}$$

Exploiting this factorization result, the strategy is to compose invertible building blocks $(\phi_\theta)_i$ to construct a function ϕ_θ that is sufficiently expressive. In general, the architecture of Normalizing Flows encompasses various transformations (see Figure 2.6), including simple operations like affine transformations and permutations, as well as more complex functions such as coupling layers. Common flow architectures include RealNVP, Glow, and Planar Flows, each introducing unique ways to parameterize and structure transformations[55].

Regarding drawbacks of NF, one significant limitation lies in the computational cost associated with training NF, particularly as model complexity increases. The requirement for invertibility and the computation of determinants of Jacobian matrices contributes to the time-consuming nature of training, especially in deep architectures. The architectural complexity of NF poses another challenge. Designing an optimal structure and tuning parameters may prove challenging, necessitating experimentation and careful consideration. Moreover, they may face challenges in scaling to extremely high-dimensional spaces, limiting their applicability in certain scenarios. Despite their expressiveness, NF may struggle to capture extremely complex distributions, requiring an impractical number of transformations to model certain intricate data distributions effectively.

Another degree of freedom is the choice of the base distribution ρ_Z, which can impact NF performance. Using a base distribution that does not align well with the true data distribution may hinder the model's ability to accurately capture underlying patterns. Training NF is observed to be less stable compared to other generative models, requiring careful tuning of hyperparameters and training strategies to achieve convergence and avoid issues like mode collapse. Lastly, interpreting the learned representations and transformations within NF can be challenging, which is an obstacle for a straightforward comprehension of how the model captures and represents information.

2.2.5 Comparison and EBMs

In this section, we present a summarized comparison of EBMs and the other generative models. First of all, the main similarity is the objective: maximize the log-likelihood is

[55] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, **43**: 3964–3979, 2020.

	Implementation of Maximum Log-Likelihood
EBM	Cross-Entropy Minimization $\arg\min_{\theta\in\Theta} \mathbb{E}_*[U_\theta] + \log Z_\theta$
VAE	Latent Space $\arg\max_{\theta,\theta'} \sum_{i=1}^N \frac{1}{R}\sum_{i=r}^R \log \rho_\theta(x_i\|z_i^{(r)}) - \sum_{i=1}^N D_{\mathrm{KL}}[\log q_{\theta'}(z_i\|x_i) \parallel \rho(z_i)]$
GAN	Minimax Game $\arg\min_\theta \arg\max_{\theta'} \mathbb{E}_*[\log D_{\theta'}(x)] + \mathbb{E}_{\rho_z}[\log(1 - D_{\theta'}(G_\theta(z)))]$
SBD or SI	Implicit via Transport-Diffusion Equation $\partial_t\rho(x,t) + \nabla\cdot((b_\theta(x,t) + \varepsilon s_{\theta'}(x,t))\rho(x,t) - \varepsilon\nabla\rho(x,t)) = 0$
NF	Volume Correction Factor $\arg\max_{\theta\in\Theta} \mathbb{E}_*[\log \rho_Z(\phi_\theta^{-1}(y)) + \log\|\det J_y\phi_\theta^{-1}(y)\|]$

Table 2.1: Comparison of implementation of maximum log-likelihood for different generative models.

	Generation	Evaluation
EBM	MCMC	Energy function
VAE	Sampling from gaussian	Fidelity of encoding-decoding
GAN	Generator	Discriminator
SBD (or SI)	SDE (or ODE)	Score (or vector field)
NF	pushforward map	Fidelity of Normalizing Flow

Table 2.2: Generation and evaluation for Generative Models.

the general aim. In Table 2.1 we present how this task is instantiated case by case. In generative models, there exists an inherent trade-off between the model's ability to generate data and its alignment with real-world data. Essentially, the paradigm followed in each optimization step involves two key stages: (1) the generation of data using a fixed model, and (2) the evaluation of the model's performance by comparing the generated ("fake") data with the actual dataset. This dual-step process is universally applicable, albeit with variations in implementation. It represents an interpretation of generative models as a balance between their discriminative and sampling capabilities. For conciseness, Table 2.2 provides a summary of specific details for each generative model. The primary distinction among the Generative Models under examination lies in the manner in which they learn. In Normalizing Flows (NFs) and Generative Adversarial Networks (GANs), the focus is on directly learning a deterministic mapping from data to noise. Stochasticity enters the picture primarily through the selection of an initial datum for generation. Conversely, in Variational Autoencoders (VAEs), diffusion models, and Energy-Based Models (EBMs), generation is intrinsically linked to a sampling routine (such as Stochastic Differential Equations for Score-Based Diffusion). This disparity has its advantages and disadvantages: while deterministic generation can be more efficient, any inaccuracies in the learned generator, stemming from finite dataset sizes, tend to be amplified. Empirically, this mirrors the rationale behind employing SDEs rather than ODEs in stochastic interpolants: a noisy evolution serves as a regularizer. However, the magnitude of noise becomes a critical hyperparameter

in diffusion models, as does the structure within latent space for VAEs. Currently, there is no universally applicable recipe for determining the best generative model for a specific problem.
The bidirectional nature of generative models (from noise or latent space to data, and vice versa) is a noteworthy common characteristic, except in the case of GANs, where the generation model lacks invertibility. Interestingly, it appears that more recent generative models, such as Score-Based Diffusion, enhance fidelity by leveraging information acquired during the "noising" process—transforming data into noise. To explore this perspective, the utilization of tools native to Mathematical Physics, particularly those related to stochastic processes, has proven necessary, suggesting that a meticulous examination of Generative Models through the lens of physical processes could be crucial for future advancements.
Now, let's delve into a more detailed mathematical comparison, with a focus on Energy-Based Models (EBMs). Specifically, we demonstrate how, in certain cases, other generative models can be interpreted as EBMs:

- For GANs, if the discriminator is $D_{\theta'}(x) \propto e^{-U_{\theta'}(x)}$, we immediately recover the term $\mathbb{E}_*[U_\theta]$. The training of the generator correspond to learn a perfect sampler, and resemble the use of machine learning to improve MCMC in computational science[56].

- For SBD and SI, if the score is modelled by $s_{\theta'}(x,t) \propto -\nabla U_{\theta'}(x)$ the law of the process solution of the SDE is a Boltzmann-Gibbs ensemble by construction. Thus, the strong analogy is related to the constrained structure imposed to the law of the bridging process between the noise and the data, forced to be a BG ensemble. Regarding the loss, since the model is trained on Fisher divergence or on the interpolants, there is no direct analogy between the losses.

- For NFs, if ϕ_θ is the map associated to a flow that brings $X_0 \sim \rho_Z(\phi_\theta^{-1}(x)) \propto e^{-U_\theta(x)}$ to $X_1 \sim \rho_*$, then the term $\mathbb{E}_*[U_\theta]$ present in EBMs is analogous to $\mathbb{E}_*[\log \rho_Z(\phi_\theta^{-1}(y))$ for NFs. In practice, if the composition of ρ_Z with the normalizing flow can be written as an EBM, there is no difference between the models. This is of course not true in general — it is not given that for any θ, a composition of the inverse map ϕ_θ^{-1} and ρ_Z can be always associated to an EBM parameterized via U_θ.

- For VAEs the situation is a bit more intricate because of the ELBO reformulation. A possible interpretation towards EBMs is to think about encoder and decoder as a forward and backward processes from data to a latent space (possibly independent features, similarly to gaussian noise). One could imagine $\rho_\theta(x_i|z_i^{(r)})$ and $q_{\theta'}(z_i|x_i)$ as EBMs that have to match with some constraint on the z space. In fact, the original EM steps represent an alternating optimization, where θ is not related to θ'. In this sense, VAEs tries to match the forward and backward processes, similarly to SBD where they are the same by construction of the score.

[56] Jiaming Song, Shengjia Zhao, and Stefano Ermon. A-nice-mc: Adversarial training for mcmc. *Advances in neural information processing systems*, **30**: 2017.

A fitting metaphor for generative models is to liken them to bridges connecting a "simple" source, such as noise, to real data. Just like constructing a physical bridge, building a generative model requires understanding the abutments. In the realm of data science, this translates to conducting statistical analysis of the dataset on one side, and selecting the appropriate noise source on the other. Additionally, modifying the docking configuration where the bridge is anchored—equivalent to data preprocessing—is often necessary. This step is crucial, akin to using the correct coordinates to describe a physical system. For instance, molecular configurations may not be easily trainable in standard three-dimensional space due to numerous implicit structural constraints.
Once the groundwork is laid, constructing the bridge begins. Just like real roads, different paths are tailored for different canyons, and similarly, for different data structures. Whether the bridge is bidirectional or not depends on the specific requirements. The key takeaway is always to maximize the log-likelihood between the model and the data distribution, ensuring that the bridge effectively connects the source to the desired destination.

2.3 EBMs and sampling

The challenge of sampling from the Boltzmann-Gibbs (BG) ensemble arises in statistical mechanics, particularly when dealing with complex systems at equilibrium. This ensemble encapsulates the probability distribution of states for a system with numerous interacting particles at a given temperature. The primary obstacle in that context lies in the exponential number of possible states and intricate dependencies among particles, rendering brute-force methods impractical for large systems. A similar difficulty is encountered during EBM training, since the computation of the expectation $\mathbb{E}_\theta$ requires the ability to sample from a Boltzmann-Gibbs density.
Let us restrict to the case in which the energy $U(x)$ is defined on $\mathbb{R}^d$, that corresponds to continuous states in Statistical Physics. Any proposed techniques to efficiently sample from $\rho_{BG}(x) = \exp(-U(x))/Z$ can rely just on $U(x)$ or on its derivatives, even if the computation of many iterated derivatives can be expensive in high dimension. The estimation of the partition function or the shape of the energy landscape are in general unknown — on the contrary, they are the unknowns. Methods as rejecting sampling[57] cannot be used since one has usually access to $U(x)$ and not to the normalized density $\rho_{BG}(x)$. Since the advent of computational science, sampling has been attacked with many methods — a complete and exhaustive review of the existent methods would lead us off-topic. In this Section, we will highlight three common routines for sampling from a BG enseble: Metropolis-Hastings and Unadjusted Langevin Algorithm, and lastly Metropolis Adjusted Langevin Algorithm, a sort of fusion of the first two.
Let us better define the mathematical setting. We consider a space $\Omega \subseteq \mathbb{R}^d$ and a discrete sequence $(t_k)_{k\geq 0} \subset \mathbb{N}$. Then, we consider $X_{t_k} := X_k$ to be a stochastic process in Ω and discrete time. For the sake of simplicity we will always consider absolutely

[57] George Casella, Christian P Robert, and Martin T Wells. Generalized accept-reject sampling schemes. *Lecture Notes-Monograph Series*, 342–347, 2004.

continuous densities with respect to Lebesgue measure.

Definition 2.3.1 (Informal)**.** *Sampling from a BG ensemble consists in defining the process X_k such that $\exists T > 0$, not necessarily unique, for which $X_T \sim \exp(-U(x))/Z$.*

Once we manage to define such a process, and implement it in practice, we have solved the problem of sampling from a BG ensemble. A possible implicit way to define such stochastic process is via a transition kernel. Suppose we are interested in the law of the process X at time $k+1$, that we denote $\rho(X_{k+1})$ with an abuse of notation (n.b. analogous of $\rho(x,t)$ in the context of SDEs and Fokker-Planck equation). By definition of conditional probability, there exists a function $T : \Omega^{n+1} \to \mathbb{R}_+$ such that

$$\rho(X_{k+1}) = \int_{\Omega^d} T(X_{k+1}|X_k, \dots, X_0)\rho(X_k, \dots, X_0) \prod_{i=0}^{n} dX_i \tag{2.3.1}$$

This equation asserts that any property, uniquely defined by the law $\rho(X_{k+1})$ of the system at time $k+1$, depends on the system's state at any $k \geq 0$. Generally, this strict constraint is relaxed by imposing Markovianity[58], which is the property of the transition kernel to depend solely on the present state X_k and not on previous states, i.e.

$$T(X_{k+1}|X_k, \dots, X_0) = T(X_{k+1}|X_k) \coloneqq T(X_k, X_{k+1}) \tag{2.3.2}$$

The sequence $(X_k)n \geq 0$ is called a Markov chain if the associated transition kernel is Markovian. The question now is how to design such a chain to solve the sampling problem. Traditionally, it is simpler to identify a transition kernel for which $\rho_{BG}(x)$ is the unique stationary distribution, i.e., $\rho(X_k) = \rho_{BG}(x)$ for any $n > T$ in Definition 2.3.1. Moreover, the integral definition (2.3.1) is not suitable for applications since one usually evolves X_k and not its law. Typically, it is required that T is associated with an explicit time evolution for the process, namely an explicit mapping $X_{k+1} = F(X_k)$. For historical reasons, let us present the most famous procedure to build the required sampling stochastic process, namely the Metropolis-Hastings algorithm[59,60]. Such techniques stand out as a foundational Markov chain Monte Carlo (MCMC)[61] method. Here, we provide its definition and a sketch of the proof of its properties.

Definition 2.3.2 (Metropolis-Hastings (MH) algorithm)**.** *Let us consider an initial condition $X_0 \sim \rho_0(x)$, where $\rho_0(x)$ simple to sample from (e.g. Gaussian or uniform). Let us consider a conditional probability distribution $g(X_{k+1}|X_k)$, also called proposal distribution, defined on the state space Ω and let $\rho_{BG}(x) = \exp(-U(x))/Z$ the BG ensemble we would like to sample from. Starting at $n = 0$, we define a Markov chain X_k via the following repeated steps:*

[58] Daniel W Stroock. *An introduction to Markov processes*. vol. 230 Springer Science & Business Media, 2013.

[59] Nicholas Metropolis et al. Equation of state calculations by fast computing machines. *The journal of chemical physics*, **21**: 1087–1092, 1953.

[60] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. 1970.

[61] Christophe Andrieu et al. An introduction to MCMC for machine learning. *Machine learning*, **50**: 5–43, 2003.

1. *Given X_k, generate a proposal $X_{k+1}^{(p)}$ using the time evolution prescribed by T.*

2. *Compute the acceptance ratio*

$$A(X_{k+1}^{(p)}, X_k) = \arg\min\left\{1, \frac{\rho_{BG}(X_{k+1}^{(p)})g(X_k|X_{k+1}^{(p)})}{\rho_{BG}(X_k)g(X_{k+1}^{(p)}|X_k)}\right\} \tag{2.3.3}$$

3. *Sample a real number $u \sim \mathcal{U}[0,1]$. If $u < A(X_k, X_{k+1}^{(p)})$, accept the proposal and set $X_{k+1} = X_{k+1}^{(p)}$; otherwise, refuse the move and set $X_{k+1} = X_k$. Then, increment n to $n+1$.*

Proposition 2.3.1. *The Markov chain X_k defined via MH algorithm has $\rho_{BG}(x)$ as unique stationary distribution, i.e.*

$$\rho_{BG}(x) = \int_{\Omega^d} T_{MH}(x|x')\rho_{BG}(x')dx', \qquad \forall x, x' \in \Omega \tag{2.3.4}$$

where $T_{MH}(x|x')$ is the transition kernel of MH algorithm.

Proof. We have show that (1) $\rho_{BG}(x)$ is a stationary distribution and (2) it is unique. Regarding (2) we advocate to geometric ergodicity[62]. We present the proof of property (1): firstly, it is equivalent to detailed balance condition[63]

$$\rho_{BG}(x)T_{MH}(x,x') = \rho_{BG}(x')T_{MH}(x',x) \qquad \forall x, x' \in \Omega \tag{2.3.5}$$

The transition kernel is by definition

$$T_{MH}(x,x') = g(x'|x)A(x',x) + \delta(x-x')\left(1 - \int_\Omega A(x,s)g(s|x)ds\right) \tag{2.3.6}$$

where the first addend takes into account the case of accepted move, while the second of the rejected one. Actually, for $x = x'$ the detailed balance condition is trivially true. Then, for $x \neq x'$ we compute the left hand side of (2.3.5)

$$\begin{aligned}\rho_{BG}(x)T_{MH}(x,x') &= \rho_{BG}(x)g(x'|x)A(x',x) \\ &= \rho_{BG}(x)g(x'|x)\arg\min\left\{1, \frac{\rho_{BG}(x')g(x|x')}{\rho_{BG}(x)g(x'|x)}\right\} \\ &= \arg\min\left\{\rho_{BG}(x)g(x'|x), \rho_{BG}(x')g(x|x')\right\}\end{aligned} \tag{2.3.7}$$

The right hand side is symmetric with respect to swap of x with x', hence concluding the proof. □

[62] Kerrie L Mengersen and Richard L Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The annals of Statistics*, **24**: 101–121, 1996.

[63] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*. vol. 2 Springer, 1999.

In practice, convergence is considered achieved when the acceptance ratio is consistently close to 1. In such cases, every newly generated proposal can be regarded as an independent sample obtained from ρ_{BG}.
Despite its popularity, the Metropolis-Hastings algorithm has some limitations. It is sensitive to the choice of the proposal distribution g and its parameters, and improper tuning may result in inefficient exploration. For instance, in the so-called *random walk setting*, g is chosen to be a Gaussian transition kernel, and its variance is a critical hyperparameter in this case. Moreover, the algorithm generates correlated samples, impacting the independence of successive samples and hindering accurate estimation even after convergence. Convergence may be slow in high-dimensional spaces, requiring numerous iterations. In such setups, the algorithm's performance is influenced by the initial state, and initial points far from the basin of the target may impede efficient exploration, leading to an acceptance rate close to zero. Another issue pertains to multimodal distributions, especially those with widely separated modes. They pose a significant challenge for the Metropolis-Hastings algorithm because, depending on the choice of g, jumps between modes can be very rare and may necessitate a very long chain to practically observe convergence.
The second class of Markov chain we would like to review are the Langevin-based algorithms. The basic idea is very close to the definition of naive score-based diffusion in Proposition 2.2.3. For the sake of simplicity let us fix the state space $\Omega = \mathbb{R}^d$.

Proposition 2.3.2. *Let us denote with dW_t a Wiener process. Under Assumption 2.1.1, namely*

$$\exists a \in \mathbb{R}_+ \text{and a compact set } \mathcal{C} \in \mathbb{R}^d : \; x \cdot \nabla U(x) \geq a|x|^2 \quad \forall x \in \mathbb{R}^d \setminus \mathcal{C}, \tag{2.3.8}$$

the Langevin SDE

$$dX_t = -\nabla U(x)dt + \sqrt{2}dW_t \qquad X_0 \sim \rho_0 \tag{2.3.9}$$

have a global solution in law and is ergodic[64–66]*. For any initial condition $\rho_0(x)$ such solution is $\rho_{BG}(x)$.*

Given this result, one can define a Markov chain based on the time discretization of such SDE and use it for sampling[67]. Such procedure is commonly named Unadjusted Langevin Algorithm (ULA)[68].

[64] Bernt Oksendal. *Stochastic Differential Equations*. 6th ed. Springer-Verlag Berlin Heidelberg, 2003.

[65] J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, **101**: 185–232, 2002.

[66] Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, **8**: 483–509, 1990.

[67] Giorgio Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, **180**: 378–384, 1981.

[68] Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 341–363, 1996.

Definition 2.3.3 (ULA). *Given a time step $h > 0$ and a set of i.i.d. gaussian variables $\{\xi_k\} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$, the Unadjusted Langevin Algorithm (ULA) is the Markov chain defined as*

$$X_{k+1} = X_k - h\nabla U_{\theta_k}(X_k) + \sqrt{2h}\,\xi_k, \qquad X_0 \sim \rho_{\theta_0}, \tag{2.3.10}$$

for $k \in \mathbb{N}$.

Under Assumption 2.1.1, the Unadjusted Langevin Algorithm (ULA) is ergodic and possesses a unique global solution. An advantage over the Metropolis-Hastings (MH) algorithm is that the chain is uniquely defined via $U(x)$, and no proposal distribution is necessary. However, it is well-known that ULA represents a biased implementation of Langevin dynamics[69]. For a nonzero time step, the global solution is $\rho_{bias} \neq \rho_{BG}$. Let us illustrate this point with a simple example.

Example 3. Let $U(x) = (x-\mu)^T\Sigma^{-1}(x-\mu)/2 + \log[\det(2\pi\Sigma)]/2$, that is BG ensemble is a gaussian with mean μ and covariance matrix Σ. The associated Langevin SDE is also known as Ornstein-Uhlenbeck (OU) process[70], having a linear drift as peculiarity:

$$dX_t = -\Sigma^{-1}(X_t - \mu)dt + \sqrt{2}dW_t. \tag{2.3.11}$$

It is possible to write an explicit solution using Ito integral, namely

$$X_t - \mu \sim e^{-t\Sigma^{-1}}(X_0 - \mu) + \Sigma^{\frac{1}{2}}\left(\boldsymbol{I}_d - e^{-2t\Sigma^{-1}}\right)^{\frac{1}{2}} Z \tag{2.3.12}$$

for any $t \geq 0$ and where $Z \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$ indipendently from X_0. It means that the law of the process converges exponentially fast to $\mathcal{N}(\mu, \Sigma)$. The associated ULA is

$$X_{k+1} - \mu = \left(\boldsymbol{I}_d - h\Sigma^{-1}\right)(X_k - \mu) + \sqrt{2h}\xi_k. \tag{2.3.13}$$

and the corresponding solution in law is

$$X_k - \mu \sim A_h^k(X_0 - \mu) + \sqrt{2h}\left(\boldsymbol{I}_d - A_h^2\right)^{-\frac{1}{2}}\left(\boldsymbol{I}_d - A_h^{2k}\right)^{\frac{1}{2}} Z \tag{2.3.14}$$

where $A_h = \boldsymbol{I}_d - h\Sigma^{-1}$. Naming $\lambda_{min}(\Sigma) > 0$ the minimum eigenvalue of the covariance matrix, for $0 < h < \lambda_{min}(\Sigma)$ we have $\lim_{k\to\infty} A_h^k = 0$. Thus, for $k \to \infty$

$$X_k \xrightarrow{d} \mu + \sqrt{2h}\left(\boldsymbol{I}_d - A_h^2\right)^{-\frac{1}{2}} Z \tag{2.3.15}$$

This means that the limiting measure for ULA is not ρ_{BG}, but

$$\rho_{bias}(x) = \mathcal{N}\left(\mu, \Sigma\left(\boldsymbol{I}_d - \frac{h}{2}\Sigma^{-1}\right)^{-1}\right)(x) \tag{2.3.16}$$

[69] Andre Wibisono. "Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem" in: *Conference on Learning Theory*. PMLR 2018. 2093–3027

[70] George E Uhlenbeck and Leonard S Ornstein. On the theory of the Brownian motion. *Physical review*, **36**: 823, 1930.

This example illustrates that the Unadjusted Langevin Algorithm (ULA) exhibits bias even for a very simple target density. This phenomenon has been recently analyzed mathematically[69]. The physical interpretation is that detailed balance is broken by construction. Let us elaborate on this point: in Proposition 2.2.2, we demonstrated how a Stochastic Differential Equation (SDE) can be associated with a Partial Differential Equation (PDE). The specific case studied in this section was previously analyzed in Proposition 2.2.3. Specifically, the Boltzmann-Gibbs (BG) density is the unique minimizer of the Kullback-Leibler (KL) divergence functional $D_{\mathrm{KL}}(\rho \parallel \rho_{GB})$. Moreover, the Fokker-Plank PDE corresponds to the gradient flow in $\mathcal{P}(\mathbb{R}^d)$ with respect to the 2-Wasserstein distance $\mathcal{W}_2$[43]. If we split (2.3.10) in two substeps

$$\begin{aligned} X_{k+\frac{1}{2}} &= X_k - h\nabla U(X_k) \\ X_{k+1} &= X_{k+\frac{1}{2}} + \sqrt{2\varepsilon}\xi_k \end{aligned} \tag{2.3.17}$$

we can associate each of them to a precise operation in probability space. In particular, denoting with ρ_i the law of X_i, we obtain

$$\begin{aligned} \rho_{k+\frac{1}{2}} &= (\boldsymbol{I}_d - h\nabla U)_{\#}\rho_k \\ \rho_{k+1} &= \mathcal{N}(\mathbf{0}_d, 2h\boldsymbol{I}_d) \star \rho_{k+\frac{1}{2}} \end{aligned} \tag{2.3.18}$$

We recall the decomposition of the Kullback-Leibler (KL) divergence as $D_{\mathrm{KL}}(\rho \parallel \rho_{GB}) = -H(\rho, \rho_{GB}) - H(\rho)$. In (2.3.18), the first step involves the forward discretization of gradient descent on $-H(\rho, \rho_{GB}) = \mathbb{E}_\rho[U]$, while the second step represents the exact gradient flow for negative entropy in probability space. Therefore, ULA is also referred to as the Forward-Flow method in probability space. The bias arises because the forward gradient descent does not correspond, in probability space, to the adjoint of the flow at iteration $k+1/2$. One possible solution is to use forward-backward combinations, referring to proximal algorithms[71]. In particular, the Forward-Backward (FB) implementation for Langevin dynamics would be

$$\begin{aligned} \rho_{k+\frac{1}{2}} &= (\boldsymbol{I}_d - h\nabla U)_{\#}\rho_k \\ \rho_{k+1} &= \underset{\rho\in\mathcal{P}}{\arg\min} \left\{ -H(\rho) + \frac{1}{2\epsilon}\mathcal{W}_2\left(\rho, \rho_{k+\frac{1}{2}}\right)^2 \right\} \end{aligned} \tag{2.3.19}$$

Similarly, the Backward-Forward (BF) version

$$\begin{aligned} \rho_{k+\frac{1}{2}} &= \left((\boldsymbol{I}_d + h\nabla U)^{-1}\right)_{\#}\rho_k \\ \rho_{k+1} &= \exp_{\rho_{k+\frac{1}{2}}}\left(-h\nabla \log \rho_{k+\frac{1}{2}}\right) \end{aligned} \tag{2.3.20}$$

where exp is the exponential map. Unfortunately, both FB and BF are not implementable in practice, except for the trivial case of gaussian initial data and target ρ_{BG}. The heat flow (the step $k+1/2$) is the most problematic since it concerns steps

[71] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, **1**: 127–239, 2014.

in probability space. Neither forward (n.b. beyond one iteration) nor backward are usable. As a side note, one could imagine to directly perform a single forward or backward step on the KL divergence. Unfortunately, the encountered issues are the same one has for the heat flow, i.e. the hard task appears to be the actual implementation of forward or backward routines in probability space. In conclusion, ULA appears to be the simplest time discretization of Langevin dynamics, since it is practically implementable in general, hence very used for sampling from a BG ensemble. However, it is known to be biased and other methods are studied to eliminate, or at least reduce, such bias.
One possibility we would like to review is Metropolis Adjusted Langevin Algorithm[72] (MALA), which represents a sort of hybrid between MH and ULA.

Definition 2.3.4 (MALA)**.** *Metropolis Adjusted Langevin Algorithm is a particular case of MH algorithm 2.3.2 where the proposal distribution is the transition kernel associated to ULA* (2.3.10)*, namely (for $x \in \mathbb{R}^d$)*

$$g(x' \mid x) = \frac{1}{(2\pi h)^{\frac{d}{2}}} \exp\left(-\frac{1}{4h}\|x' - x + hU(x)\|_2^2\right) \tag{2.3.21}$$

On the other hand, one can interpret MALA as a corrective measure for the breakdown of detailed balance in ULA. While the Metropolis-Hastings algorithm inherently respects detailed balance, implying that MALA becomes asymptotically unbiased for a large number of iterations as $k \to \infty$, certain challenges persist. A primary concern is the sensitivity to the choice of the step size h during the discretization of Langevin dynamics, significantly influencing the efficiency of sampling. When h is too small, it can lead to poor exploration and potentially a very low acceptance rate, while an excessively large h can lead to instability of the chain. Determining an optimal h lacks a general rule, contributing to MALA introducing bias in samples, particularly evident when the target distribution features sharp peaks or multimodal structures. This bias introduces potential inaccuracies in statistical estimates.
In practical applications, MALA may exhibit random walk behavior, especially when step sizes are inadequately tuned, resulting in inefficient exploration and sluggish convergence. The algorithm's performance is further contingent on the choice of initial conditions, and beginning far from high-probability regions may necessitate a considerable number of iterations for meaningful exploration. Additionally, MALA may struggle to adapt to changes in the geometry of the target distribution, particularly when facing varying curvatures or strong anisotropy.
While various more advanced algorithms exist, they often build upon the foundational concepts discussed in this section. Notable among them is Hamiltonian (or Hybrid) Monte Carlo[73] (HMC), an advanced MCMC method inspired by Hamiltonian mechanics. HMC utilizes fictitious Hamiltonian dynamics to propose new states, enhancing exploration, especially in high-dimensional systems. Gibbs sampling[74], another MCMC

[72] Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, **56**: 549–581, 1994.
[73] Simon Duane et al. Hybrid monte carlo. *Physics letters B*, **195**: 216–222, 1987.
[74] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian

approach, iteratively samples from conditional distributions given current variable values on a single dimension, proving effective, particularly in high-dimensional spaces. Parallel Tempering, or Replica Exchange[75], involves running multiple chains at different temperatures concurrently, with periodic swaps between neighboring chains to facilitate improved exploration.
In general, most methods aim to find a chain that produces independent samples from a Boltzmann-Gibbs ensemble, particularly when run for extended periods. A critical issue is measuring the effective bias due to the truncation at finite time of the chain, posing challenges for convergence towards the asymptotic ρ_{BG}. Unfortunately, few general results are available, and they are often limited to specific BG ensembles, such as Gaussian or log-concave densities. This becomes particularly problematic in the context of Energy-Based Models (EBM), as outlined in Remark 2.1.1, where sampling from a BG distribution is required at each step of parameter optimization. Further discussion on this criticality in the context of state-of-the-art EBM training will be provided in Section 3.1.

2.4 EBMs and physics

In this section, we provide a review of the Boltzmann-Gibbs ensemble in relation to Statistical Physics, covering its derivation in equilibrium. The purpose of this treatment is not only to motivate the specific structure of EBMs but also to provide a context for the main topic of the present thesis, which explores the use of nonequilibrium results to train such generative models.
The first step involves the derivation of the Boltzmann-Gibbs ensemble. Here, we present a derivation based on information theory[76,77], offering a posteriori physical interpretation of the quantities we will manipulate. Alternative methods of proof are also available[78]. Consider a physical system whose state is uniquely determined by a variable $x \in \Omega \subseteq \mathbb{R}^d$, where Ω is often referred to as phase space. The connection with information theory is linked to the fundamental problem of Statistical Physics: describing a system as a statistical ensemble, i.e., identifying an observation of x as a sample from an underlying PDF ρ. Like classical statistics, ρ contains a wealth of information about the system, particularly its global properties.
In Physics, this dichotomy translates into the microscopic versus macroscopic realms. Let's envision a simple thought experiment: picture a large city where each of the N inhabitants is given a fair coin, with the coin's state represented by our variable $x \in \{-1, 1\}^N$. Twice a day, everyone has to flip their coin. If we were omniscient,

restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 721–741, 1984.

[75] Robert H Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical review letters*, **57**: 2607, 1986.

[76] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, **106**: 620, 1957.

[77] Edwin T Jaynes. Information theory and statistical mechanics. II. *Physical review*, **108**: 171, 1957.

[78] Giovanni Gallavotti. *Statistical mechanics: A short treatise*. Springer Science & Business Media, 1999.

there would be a way to predict the state x with no error (i.e., the microscopic state) and derive any global (macroscopic) property, such as the sum or product of the state values at each flipping event, with no error. However, in reality, nobody could achieve this; we rely on statistics, the central limit theorem, and so forth. In other words, we know the probability density $\rho(x)$ from which the process is a sampled event. For instance, if N is large enough, we expect the average of the state vector to be 0 for any flipping event, and we can deduce so directly from ρ.
In Statistical Physics, each coin represents a component of a system, such as a particle in a gas, for which a direct measurement of x is unattainable. The goal is to determine ρ so that standard statistical tools can be used to analyze global properties. The challenge that makes Statistical Physics more complex than the simple example above is that the dynamics of individual components can be unknown and inaccessible. Additionally, interactions between components can make the identification of ρ challenging, even if the underlying microscopic dynamics are known.
To address this issue, we recognize that, before formulating any physical model, we need some motivated assumptions—constraints or information—regarding how the system should behave, at least on a macroscopic level. This is the bare minimum; without any information about a system, it is impossible to provide any meaningful analysis. Thus, adopting a claim of epistemic modesty, one can state that we aim to select the model compatible with such constraints that maximizes our ignorance about the system. The mathematical translation of such idea is the *Principle of Maximum Entropy.*

Assumption 2.4.1 (Principle of Maximum Entropy)**.** *Let us consider the unknown $\rho : \Omega \to \mathbb{R}_+$ that describes the probability distribution of the states. We assume ρ to be absolutely continuous w.r.t. Lebesgue measure without loss of generality. Given a vector field $F : \Omega \to \mathbb{R}^d$, $\Lambda \in \mathbb{R}^d$ and a PDF π, a set of constraints is any componentwise (in)equality*

$$I_k[\pi] \coloneqq \int_\Omega F_k(x)\pi(x)dx \leq_k \Lambda_k \qquad k = 1, \dots, d \tag{2.4.1}$$

where the symbol $\leq_k$ can be an equality or inequality. The **Principle of Maximum Entropy** *is*

$$\rho = \underset{\substack{\pi \in \mathcal{P}(\Omega) \\ I_k[\pi] \leq_k \Lambda_k}}{\arg\max}\, H(\pi) \tag{2.4.2}$$

where $H(\rho)$ is the usual differential entropy, cfr. (2.1.7).

This variational formulation identify the "ignorance" about the system with the entropy associated to ρ. It has been proven that the entropy can be characterized in an axiomatic way[79], so that the definition of differential entropy is unique with respect to certain properties. For the sake of the present treatment, let us motivate the Maximum Principle with a simple example.

[79] János Aczél, Bruno Forte, and Che Tat Ng. Why the Shannon and Hartley entropies are 'natural'. *Advances in applied probability*, **6**: 131–146, 1974.

Example 4 (Maximum principle on an interval)**.** Let us consider an interval $\Omega = [a, b] \subset \mathbb{R}$, with $\mathrm{Vol}(\Omega) = \int_a^b dx$. Moreover, the only constraint is that ρ can be normalized and is positive. Thus, we have $I_0[\pi] = \int_a^b \rho(x)dx = 1$ and

$$\rho = \underset{\substack{\pi \in \mathcal{P}(\Omega) \\ I_0[\pi]=1,\ \pi>0}}{\arg\max} \; H(\pi) \tag{2.4.3}$$

We can use Lagrange multipliers to solve a constrained optimization problem, solving the unconstrained optimization of the Lagrangian

$$J(\pi) \coloneqq H(\pi) + \lambda_0 \left(\int_a^b \pi(x)dx - 1 \right) \tag{2.4.4}$$

To find stationary points we can compute the first variational derivative with respect to π and finding its roots, namely solutions of

$$\frac{\delta J(\pi)}{\delta \pi} = -\log \pi - 1 + \lambda_0 = 0 \tag{2.4.5}$$

that is $\rho(x, \lambda_0) = \exp(\lambda_0 - 1)$. To find λ_0 we can substitute $\rho(x, \lambda_0)$ into the constraint, yielding $\lambda_0 = 1 - \log(b - a)$. In conclusion, $\rho(x) = 1/(b - a)$, which also satisfies the positivity request. We have just to check that such stationary point is a maximum. The second variation of $J(\pi)$ evaluated in the stationary point is

$$\frac{\delta^2 J(\pi)}{\delta \pi^2}\Big|_{\pi=\rho} = -\frac{1}{\rho(x)} < 0 \tag{2.4.6}$$

Hence, we conclude that $\rho(x)$ is a maximum. If Ω is discrete such derivation can be easily generalized. The interpretation is straightforward: imagine that Ω is the event space for some random process. Without any knowledge, the simplest possible model is the one that associates the same probability to all the possible outcomes.

At this point we have all the ingredients to present the derivation of Boltzmann-Gibbs ensemble.

Proposition 2.4.1 (Boltzmann-Gibbs ensemble from Maximum Entropy principle)**.** *Given $U(x)$ that satisfies Assumption 2.1.1 and a constant $\bar{U}$, the Boltzmann-Gibbs $\rho_{BG} = e^{\lambda_1 U(x)}/Z$, where $\lambda_1 > 0$, is the unique solution of the variational maximization problem* (2.4.2) *where the set of constraints is*

$$\begin{aligned} I_0[\pi] &\coloneqq \int_\Omega \pi(x)dx = 1 \\ I_1[\pi] &\coloneqq \int_\Omega U(x)\pi(x)dx = \bar{U} \end{aligned} \tag{2.4.7}$$

Proof. The proof proceeds similarly to Example 4. The constrained optimization problem (2.4.2) is associated to the unconstrained one

$$\rho = \underset{\pi \in \mathcal{P}(\Omega)}{\arg\max}\, J(\pi) \coloneqq \underset{\pi \in \mathcal{P}(\Omega)}{\arg\max}\, H(\pi) + \lambda_0 \left(\int_\Omega \pi(x)dx - 1 \right) + \lambda_1 \left(\int_\Omega U(x)\pi(x)dx - \bar{U} \right). \tag{2.4.8}$$

where λ_0, λ_1 are Lagrange multiplier. We compute the first variational derivative and find its roots

$$\frac{\delta J(\pi)}{\delta \pi} = -\log \pi(x) - 1 + \lambda_0 + \lambda_1 U(x) = 0 \tag{2.4.9}$$

That is, $\rho(x) = \exp(\lambda_0 + \lambda_1 U(x) - 1)$. This means that λ_0 can be incorporated in the normalization factor, namely the partition function $Z^{-1} = \exp(\lambda_0 - 1)$, and determined via the constraint $I_0[\rho] = 1$. While λ_1 is implicitly determined by I_1. To check that such solution is a maximum, we compute the second variation obtaining a result analogous to (2.4.6). □

The remaining issue is the identification of λ_1 with $\beta = 1/k_B T$, related to temperature T and Boltzmann dimensional constant $k_B = 1.23 \times 10^{-28}\,\mathrm{J} \cdot \mathrm{K}^{-1}$. The reason is that if we interpret the Boltzmann-Gibbs ensemble with an equilibrium ensemble, the derivation via Maximum Entropy principle must be coherent with Thermodynamics[80]. A complete survey on such field would lead the present treatment off-topic. The take home message is related to a different interpretation of the unconstrained optimization problem (2.4.8), namely

$$\frac{\delta}{\delta \pi}\left(H(\pi) + \lambda_0 \int_\Omega \pi(x)dx + \lambda_1 \int_\Omega U(x)\pi(x)dx \right) = 0 \tag{2.4.10}$$

In particular, the following lemma holds true:

Lemma 2.4.1. *Maximum Entropy principle and its variational formulation are equivalent to*

- *Constrained minimization of energy functional* $\bar{U} = \int_\Omega U(x)\pi(x)dx$.
- *Constrained minimization of Helmholtz Free Energy functional* $F = \bar{U} - TH(\pi)$, *where* $T > 0$ *is the usual thermodynamic temperature.*

Proof. The proof is just related to a redefinition of the Lagrange multipliers. For the minimization of energy, one define $\lambda_1' = -1/\lambda_1$ and $\lambda_0' = -\lambda_0/\lambda_1$, where the sign is just a convention, obtaining

$$\frac{\delta}{\delta \pi}\left(\lambda_1' H(\pi) + \lambda_0' \int_\Omega \pi(x)dx + \int_\Omega U(x)\pi(x)dx \right) = 0 \tag{2.4.11}$$

While for the Helmholtz Free Energy, we have just to impose the thermodynamics constraint[81] $\partial F/\partial S = -T$, that is $\lambda_1 = -1/T$. □

Remark 2.4.1 (Free Energy in EBM training)**.** *In Section 2.1.1 we presented the training procedure for an EBM as the KL divergence minimization. If* ρ_θ *is in the same class of* ρ_*, *the global minimum in probability space corresponds to* $\rho_\theta = \rho_*$, *i.e.*

[80] Clement John Adkins. *Equilibrium thermodynamics*. Cambridge University Press, 1983.
[81] Enrico Fermi. *Thermodynamics*. Courier Corporation, 2012.

KL divergence equal to zero since by definition $D_{KL}(\rho_\theta \parallel \rho_\theta) = 0$. *However, if we expand the such identity, we have*

$$\log Z_\theta + \beta \int_{\mathbb{R}^d} U_\theta(x)\rho_\theta(x)dx - H(\rho_\theta) = 0 \tag{2.4.12}$$

where we used (2.1.10)*, and restored* β *in front of* U_θ *(n.b. we put* $k_B = 1$ *and* $T = 1$ *in Section 2.1.1). If we identify* $F_\theta = -\log Z_\theta$*, we immediately notice that* (2.4.12) *is the definition of Helmholtz Free Energy. In fact, the convergence of the training corresponds to have reached the equilibrium. The equality* $F = \bar{U} - TH(\pi)$ *is not true out of equilibrium — the KL divergence* $D_{KL}(\rho_\theta \parallel \rho_*)$ *is zero iff* $\rho_\theta = \rho_*$*. Moreover, it is even more clear the statement of Lemma 2.4.1: since*

$$\log Z_\theta + \min_{\substack{\pi \in \mathcal{P}(\Omega) \\ I_0[\pi]=1,\ \rho_\theta > 0}} [\beta \int_{\mathbb{R}^d} U_\theta(x)pi(x)dx - H(\pi)] = 0 \tag{2.4.13}$$

and $F = \bar{U} - TH(\pi)$*, at equilibrium* F *is necessarily minimized in correspondence of* $F[\rho_\theta] = -\log Z_\theta$*.*

The requested compatibility between Thermodynamics and the Maximum Entropy principle in Lemma 2.4.1 represents the final ingredient needed to define the Boltzmann-Gibbs probability density associated with a system at equilibrium with a thermal reservoir at temperature T. For the purpose of this thesis, it would be beneficial to elaborate on the physical significance of EBM. Assuming we are dealing with an equilibrium ensemble, we presume that the parameters θ in the energy U_θ have already been determined. Similar to a physical gas where particles move within an energy landscape, different datasets or even individual data points can be envisioned as snapshots of an evolving physical system. The crucial aspect is that from a statistical perspective, the average energy $\bar{U}$ associated with the EBM must remain constant, with fluctuations suppressed as the number of components increases. An example of dynamics consistent with such a constraint is Langevin dynamics. The connection with sampling and Physics becomes evident: sampling is the process of relaxation[82] towards equilibrium. Utilizing our understanding of nature entails designing sampling routines capable of facilitating such relaxation.
We introduced the concept of free energy as a thermodynamic quantity minimized at equilibrium by the Boltzmann-Gibbs ensemble. It serves as the natural link to the Jarzynski equality, a fundamental result in nonequilibrium Statistical Physics central to this thesis. Generally, computing free energy is a extensively studied problem in Chemistry[83], spanning from organic Chemistry to protein folding[84]. However, the concept of free energy appears ubiquitous, extending into seemingly disparate contexts

[82] Denis J Evans, Debra J Searles, and Stephen R Williams. Dissipation and the relaxation to equilibrium. *Journal of Statistical Mechanics: Theory and Experiment*, **2009**: P07029, 2009.

[83] William L Jorgensen. Free energy calculations: a breakthrough for modeling organic chemistry in solution. *Accounts of Chemical Research*, **22**: 184–189, 1989.

[84] Aaron R Dinner et al. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in biochemical sciences*, **25**: 331–339, 2000.

far from computational chemistry, such as autoencoders[85], lattice field theory[86], and neuroscience[87]. Invariably, it is associated with some equilibrium principle, often directly linked to the use of a generalization of the Boltzmann-Gibbs ensemble.
The importance of free energy can be readily understood: the expected value of any observable at equilibrium can be computed if we have access to the normalization constant of the Boltzmann-Gibbs ensemble, which is the partition function $Z = e^{-F}$. However, as demonstrated in Section 2.1.1, computing the partition function, and consequently the free energy, is exceedingly complex using standard Monte Carlo methods for systems with many degrees of freedom, roughly corresponding to dimension d for EBM training. Among the various proposed advanced methods[88], the utilization of the Jarzynski identity[89] stands out as a highlighted result that will be fundamental for this thesis. In Section 3.2, we will present a proof within the specific context of EBM training. The approach of Jarzynski is related to nonequilibrium Statistical Physics and an extended interpretation will be provided in such Section.

[85] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and Helmholtz free energy. *Advances in neural information processing systems*, **6**: 1993.

[86] Kim A Nicoli et al. Estimation of thermodynamic observables in lattice field theories with deep generative models. *Physical review letters*, **126**: 032001, 2021.

[87] Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, **13**: 293–301, 2009.

[88] Gabriel Stoltz, Mathias Rousset, et al. *Free energy computations: A mathematical perspective*. World Scientific, 2010.

[89] C Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, **78**: 2690, 1997.

Chapter 3

Efficient Training of EBMs using Jarzynski equality

3.1 State of the art: Contrastive Divergence

In this section we summarize the most common algorithm used for EBM training, namely Contrastive Divergence. For readers convenience, we fix the notation: in the following, $\rho_\theta(x) = \rho_{\theta(t)}(x) = \exp(-U_{\theta(t)}(x))/Z_\theta$ is the EBM we aim to train. As we showed in Section 2.1, training an EBM reduces to perform gradient-based optimization on cross-entropy. After some manipulation, the gradient of $H(\rho_*, \rho_\theta)$ reduces to

$$\partial_\theta H(\rho_*, \rho_\theta) = \mathbb{E}_*[\partial_\theta U_\theta] - \mathbb{E}_\theta[\partial_\theta U_\theta] \coloneqq -\mathcal{D}. \tag{3.1.1}$$

As we stressed in Remark 2.1.1, the main issue is the estimation of $\mathbb{E}_\theta[\partial_\theta U_\theta]$. An analytical computation is outreach for a generic U_θ, as well as the use of numerical spline methods which are impractical in high dimension. The only possibility is to generate a set of N samples $\{X^i\}_{i=1}^N$ distributed as $\rho_{\theta(t)}$ and exploit a Monte Carlo integration, namely

$$\mathbb{E}_\theta[\partial_\theta U_\theta] \approx \frac{1}{N}\sum_{i=1}^{N} \partial_\theta U_\theta(X^i) \qquad X^i \sim \rho_\theta \tag{3.1.2}$$

We stress that such generation is required at *each optimization step* of θ. Following the treatment in Section 2.3, the basic idea is to couple a gradient-based routine with a Markov Chain devoted to the generation of the needed samples. Without loss of generality, we present the state-of-the-art algorithm using ULA as the sampler.
As mentioned, a problem encountered by standard sampling routines (such as ULA) is related to multimodality; that is, for fixed θ and a general initial condition $\bar{X} \sim \bar{\pi}$ for the Markov Chain, there are no general results on the convergence rate towards the desired equilibrium $X \sim \rho_\theta$. However, if one were to choose a smart initial condition, such an issue is alleviated. For instance, in the ideal case where we could sample from

an initial distribution $\bar{\rho}$ very close to ρ_θ. In this sense, the naive approach in which the sampling routine restarts from the same "simple" distribution, like a Gaussian, for every optimization step, is not well adapted to EBM training. The question then arises: how to select an appropriate initial condition?

The idea of Contrastive Divergence[1] (CD) and Persistent Contrastive Divergence[2] (PCD) in their original formulation is to use the unknown data distribution ρ_* as the initial condition for Markov Chain sampler. This is feasible since we have the dataset; that is, we could simply extract some data points from it and use them as the initial condition of the sampler at every optimization step. To better analyze the two routines, we present CD and PCD in Algorithms 1 and 2, where ULA is chosen as the sampling routine.

Algorithm 1 Contrastive divergence (CD) algorithm

Inputs: data points $\mathcal{X} = \{x_*^i\}_{i=1}^n$ in $\mathbb{R}^d$; energy model U_θ; optimizer step $\mathrm{opt}(\theta, \mathcal{D})$ using θ and the empirical gradient $\mathcal{D}$; initial parameters θ_0; number of walkers $N \in \mathbb{N}_0$ with $N < n$; total duration $K \in \mathbb{N}$; ULA time step h; $P \in \mathbb{N}$.
for $k = 1, \dots, K-1$ **do**
  for $i = 1, ..., N$ **do**
    $X_0^i = RandomSample(\mathcal{X})$
    for $p = 0, ..., P-1$ **do**
      $X_{p+1}^i = X_p^i - h\nabla U_{\theta_k}(X_p^i) + \sqrt{2h}\,\xi_p^i, \qquad \xi_p^i \sim \mathcal{N}(0_d, I_d)$ ▷ ULA
    end for
  end for
  $\tilde{\mathcal{D}}_k = N^{-1}\sum_{i=1}^N \partial_\theta U_{\theta_k}(X_P^i) - n^{-1}\sum_{i=1}^n \partial_\theta U_{\theta_k}(x_*^i)$ ▷ empirical gradient
  $\theta_{k+1} = \mathrm{opt}(\theta_k, \tilde{\mathcal{D}}_k)$ ▷ optimization step
end for
Outputs: Optimized energy U_{θ_K}; set of walkers $\{X_P^i\}_{i=1}^N$

Let us clarify the notation. Each X used for the estimation of the gradient of cross-entropy is named a "walker". Each walker is indexed by a superscript, and the function $RandomSample(\mathcal{X})$ performs a random extraction of N points from $\mathcal{X}$. In CD, the chain for sampling is reinitialized at data at every cycle; in PCD, as for the name, the chain is "persistent", meaning it starts from the data just at the first iteration — after each optimization step, the sampling routine restarts from the samples found at the previous iteration. Traditionally, x^* is referred to as "positive" samples, while the samples from ρ_θ are termed "negative", especially in the community of Boltzmann Machines. The adjective "Contrastive" originates from the minus sign between expectations in (3.1.1): the contribution of negative and positive samples to the variation of cross-entropy is indeed opposite. In fact, the ODE associated to gradient descent

[1] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, **14**: 1771–1800, 2002.

[2] Tijmen Tieleman. "Training restricted Boltzmann machines using approximations to the likelihood gradient" in: *International conference on Machine learning*. 2008. 1064–1071

Algorithm 2 Persistent contrastive divergence (PCD) algorithm

Inputs: data points $\mathcal{X} = \{x_i^*\}_{i=1}^n$ in $\mathbb{R}^d$; energy model U_θ; optimizer step $\mathrm{opt}(\theta, \mathcal{D})$ using θ and the empirical CE gradient $\mathcal{D}$; initial parameters θ_0; number of walkers $N \in \mathbb{N}_0$ with $N < n$; total duration $K \in \mathbb{N}$; ULA time step h.
$X_0^i = \mathit{RandomSample}(\mathcal{X})$ for $i = 1, \dots, N$.
for $k = 1, \dots, K-1$ **do**
  $\tilde{\mathcal{D}}_k = N^{-1} \sum_{i=1}^N \partial_\theta U_{\theta_k}(X_k^i) - n^{-1} \sum_{i=1}^n \partial_\theta U_{\theta_k}(x_*^i)$ ▷ empirical gradient
  $\theta_{k+1} = \mathrm{opt}(\theta_k, \tilde{\mathcal{D}}_k)$ ▷ optimization step
  for $i = 1, ..., N$ **do**
    $X_{k+1}^i = X_k^i - h \nabla U_{\theta_k}(X_k^i) + \sqrt{2h}\, \xi_k^i, \qquad \xi_k^i \sim \mathcal{N}(0_d, I_d)$ ▷ ULA
  end for
end for
Outputs: Optimized energy U_{θ_K}; set of walkers $\{X_K^i\}_{i=1}^N$.

on cross-entropy minimization is

$$\dot{\theta} = \mathbb{E}_\theta[\partial_\theta U_\theta] - \mathbb{E}_*[\partial_\theta U_\theta] \tag{3.1.3}$$

This equation can be interpreted as gradient descent on the energy per positive sample and gradient ascent for the energy per negative sample. It corresponds to increasing the probability of data points in the dataset and decreasing it for the samples obtained from the chain. Stationarity is reached when $\rho_* = \rho_\theta$, so that generated points belong to the same distribution as true data points.

The natural question that arises concerns the convergence of the algorithms. To simplify the treatment, we do not analyze the algorithms for a finite set of walkers, but we study the time evolution of the probability distribution of the walkers $\check{\rho}(t, x)$. Ideally, this should remove any possible spurious bias from the analysis and permit an easier analytical study. We can write down an equation that mimics the evolution of the PDF of the walkers in the CD algorithm in the continuous-time limit. This equation reads:

$$\partial_t \check{\rho} = \alpha \nabla \cdot \big(\nabla U_{\theta(t)}(x) \check{\rho} + \nabla \check{\rho} \big) - \nu(\check{\rho} - \rho_*), \qquad \check{\rho}(t = 0) = \rho_* \tag{3.1.4}$$

with fixed $\alpha > 0$ and where the parameter $\nu > 0$ controls the rate at which the walkers are reinitialized at the data points: the last term in (3.1.4) is a birth-death term that captures the effect of these reinitializations. The solution to this equation is not available in closed from (and $\check{\rho}(t, x) \neq \rho_{\theta(t)}(x)$ in general), but in the limit of large ν (i.e. with very frequent reinitializations), we can show[3] that

$$\check{\rho}(t, x) = \rho_*(x) + \nu^{-1} \alpha \nabla \cdot \big(\nabla U_{\theta(t)}(x) \rho_*(x) + \nabla \rho_*(x) \big) + O(\nu^{-2}). \tag{3.1.5}$$

[3] Carles Domingo-Enrich et al. Dual Training of Energy-Based Models with Overparametrized Shallow Neural Networks. *arXiv preprint arXiv:2107.05134*, 2021.

As a result, the gradient of cross-entropy (3.1.1) is

$$\begin{aligned}
&\int_{\mathbb{R}^d} \partial_\theta U_{\theta(t)}(x)(\rho_*(x) - \check{\rho}(t,x))dx \\
&= -\nu^{-1} \int_{\mathbb{R}^d} \partial_\theta U_{\theta(t)}(x) \nabla \cdot \left(U_{\theta(t)}(x)\rho_*(x) + \nabla \rho_*(x) \right) dx + O(\nu^{-2}) \\
&= \nu^{-1} \int_{\mathbb{R}^d} \left(\partial_\theta \nabla U_{\theta(t)}(x) \cdot \nabla U_{\theta(t)}(x) - \partial_\theta \Delta U_{\theta(t)}(x) \right) \rho_*(x) dx + O(\nu^{-2})
\end{aligned} \tag{3.1.6}$$

The leading order term at the right hand side is precisely ν^{-1} times the gradient with respect to θ of the Fisher divergence

$$\begin{aligned}
&\frac{1}{2} \int_{\mathbb{R}^d} |\nabla U_\theta(x) + \nabla \log \rho_*(x)|^2 \rho_*(x) dx \\
&= \frac{1}{2} \int_{\mathbb{R}^d} \left[|\nabla U_\theta(x)|^2 - 2\Delta U_\theta(x) + |\nabla \log \rho_*(x)|^2 \right] \rho_*(x) dx
\end{aligned} \tag{3.1.7}$$

where Δ denotes the Laplacian and we used

$$\int_{\mathbb{R}^d} \nabla U_\theta(x) \cdot \nabla \log \rho_*(x) \rho_*(x) dx = \int_{\mathbb{R}^d} \nabla U_\theta(x) \cdot \nabla \rho_*(x) dx = -\int_{\mathbb{R}^d} \Delta U_\theta(x) \rho_*(x) dx \tag{3.1.8}$$

This confirms the known fact that the CD algorithm effectively performs GD on the Fisher divergence rather than the cross-entropy[4], similarly to score matching.
Regarding PCD, the associated PDE is (3.1.4) with $\nu = 0$. Again, the solution $\check{\rho}(t,x) \neq \rho_{\theta(t)}(x)$ in general, thus for any finite α, we have $\mathbb{E}_{\check{\rho}}[\partial_\theta U_\theta] \neq \mathbb{E}_\theta[\partial_\theta U_\theta]$. In other words, one cannot be sure to perform true gradient descent on cross-entropy — if we were able to estimate the loss, we could observe non-monotonic behavior. Extensions of standard PCD exploit an initial condition different from ρ_* for the persistent chain, but such approach is plagued by the same issue regarding the convergence rate towards equilibrium.
The takeaway message is that from an analytical standpoint, neither CD nor PCD actually perform a gradient-based optimization of cross-entropy. One important issue is that the presence of bias is related to time scales, in PCD regarding the length of the Markov Chain for sampling, and in CD also for the reinitialization frequency. Even if they are widely adopted in practice, the presence of such criticality even in an ideal setup is far from optimal and critically links the applicability of EBMs to the particular situation under study. We will show in the following some analytical and numerical examples, even very simple ones, in which PCD and CD break down. The overall objective of the next chapters is to present an alternative method that is not biased in the continuous-time limit and for infinitely many particles. Hopefully, a practical implementation of such a method would be more efficient in situations where PCD and CD encounter problems.

[4] Aapo Hyvarinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on neural networks*, **18**: 1529–1531, 2007.

3.2 Continuous time: use of Jarzynski equality

In this section we start the presentation of the novel result object of the present thesis. We use tools from nonequilibrium statistical mechanics[5,6] to write exact expressions for both $\mathbb{E}_\theta[\partial_\theta U_\theta]$ and Z_θ that are amenable to empirical estimation via sequential Monte-Carlo methods[7], thereby enabling gradient descent-type algorithms for the optimization of EBMs (Section 3.4). The first theoretical result is in continuous time:

Proposition 3.2.1. *Assume that the parameters θ are evolved according to some time-differentiable protocol $\theta(t)$ such that $\theta(0) = \theta_0$. Given any $\alpha > 0$, let $X_t \in \mathbb{R}^d$ and $A_t \in \mathbb{R}$ be the solutions of*

$$\begin{cases} dX_t = -\alpha \nabla U_{\theta(t)}(X_t)dt + \sqrt{2\alpha}\, dW_t, & X_0 \sim \rho_{\theta_0}, \\ \dot{A}_t = -\partial_\theta U_{\theta(t)}(X_t) \cdot \dot{\theta}(t), & A_0 = 0. \end{cases} \tag{3.2.1}$$

where $U_\theta(x)$ is the model energy and $W_t \in \mathbb{R}^d$ is a standard Wiener process. Then, for any $t \geq 0$,

$$\mathbb{E}_{\theta(t)}[\partial_\theta U_{\theta(t)}] = \frac{\mathbb{E}[\partial_\theta U_{\theta(t)}(X_t)e^{A_t}]}{\mathbb{E}[e^{A_t}]}, \qquad Z_{\theta(t)} = Z_{\theta_0}\mathbb{E}[e^{A_t}], \tag{3.2.2}$$

where the expectations on the right-hand side are over the law of the joint process (X_t, A_t).

The second equation in (3.2.2) can also be written in term of the free energy $F_\theta = -\log Z_\theta$ as $F_{\theta(t)} = F_{\theta_0} - \log \mathbb{E}[e^{A_t}]$: this is Jarzynski's equality[5]. We stress that it is key to include the weights in (3.2.2) and, in particular, $\mathbb{E}[\partial_\theta U_{\theta(t)}(X_t)] \neq \mathbb{E}_{\theta(t)}[\partial_\theta U_{\theta(t)}]$. This is because the PDF of X_t alone lags behind the model PDF $\rho_{\theta(t)}$: the larger α, the smaller this lag, but it is always there if $\alpha < \infty$, see next Subsection 3.2.1. The inclusion of the weights in (3.2.2) corrects exactly for the bias induced by this lag.
An immediate consequence of Proposition 3.2.1 is that we can evolve $\theta(t)$ by the gradient descent flow over the cross-entropy by solving (3.2.1) concurrently with

$$\dot{\theta}(t) = \frac{\mathbb{E}[\partial_\theta U_{\theta(t)}(X_t)e^{A_t}]}{\mathbb{E}[e^{A_t}]} - \mathbb{E}_*[\partial_\theta U_{\theta(t)}], \qquad \theta(0) = \theta_0 \tag{3.2.3}$$

since by (3.2.2) the right hand side of (3.2.3) is precisely what is needed, which is the identity $-\partial_\theta H(\rho_{\theta(t)}, \rho_*) = \mathbb{E}_{\theta(t)}[\partial_\theta U_{\theta(t)}] - \mathbb{E}_*[\partial_\theta U_{\theta(t)}]$. This is notable difference with respect to Contrastive Divergence. Another important consequence is that if we assume to know Z_{θ_0} we can also track the evolution of partition function, hence cross-entropy (2.1.5) via

$$H(\rho_{\theta(t)}, \rho_*) = \log \mathbb{E}[e^{A_t}] + \log Z_{\theta_0} + \mathbb{E}_*[U_{\theta(t)}]. \tag{3.2.4}$$

Let us present the proof of Proposition 3.2.1.

[5] C Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, **78**: 2690, 1997.

[6] Radford M Neal. Annealed importance sampling. *Statistics and computing*, **11**: 125–139, 2001.

[7] Arnaud Doucet, Nando De Freitas, Neil James Gordon, et al. *Sequential Monte Carlo methods in practice*. vol. 1 Springer, 2001.

Proof. The joint PDF $f(t,x,a)$ of the process (X_t, A_t) satisfying (3.2.1) solves the Fokker-Planck equation (FPE)

$$\partial_t f = \alpha \nabla_x \cdot (\nabla_x U_{\theta(t)} f + \nabla_x f) + \partial_\theta U_{\theta(t)} \cdot \dot{\theta}(t) \partial_a f, \quad f(0,x,a) = Z_{\theta_0}^{-1} e^{-U_{\theta_0}(x)} \delta(a). \tag{3.2.5}$$

Let us derive an equation for

$$\hat{\rho}(t,x) = \int_{-\infty}^{\infty} e^a f(t,x,a) da \tag{3.2.6}$$

To this end, multiply (3.2.5) by e^a, integrate the result over $a \in (-\infty, \infty)$, and use integration by parts for the last term at the right-hand side to obtain:

$$\partial_t \hat{\rho} = \alpha \nabla_x \cdot (\nabla_x U_{\theta(t)} \hat{\rho} + \nabla_x \hat{\rho}) - \partial_\theta U_{\theta(t)} \cdot \dot{\theta}(t) \hat{\rho}, \qquad \hat{\rho}(0,x) = Z_{\theta_0}^{-1} e^{-U_{\theta_0}(x)} \tag{3.2.7}$$

By general results for the solutions of parabolic PDEs[8] such as (3.2.7), we know that the solution to this equation is unique, and we can check by direct substitution that it is given by

$$\hat{\rho}(t,x) = Z_{\theta_0}^{-1} e^{-U_{\theta(t)}(x)}. \tag{3.2.8}$$

This implies that

$$\int_{\mathbb{R}^d} \hat{\rho}(t,x) dx = Z_{\theta_0}^{-1} Z_{\theta(t)}. \tag{3.2.9}$$

Since by definition $\mathbb{E}[e^{A_t}] = \int_{\mathbb{R}^d} \int_{-\infty}^{\infty} e^a f(t,x,a) da dx = \int_{\mathbb{R}^d} \hat{\rho}(t,x) dx$ this establishes the second equation in (3.2.2) . To establish the first notice that

$$\begin{aligned} \frac{\mathbb{E}[\partial_\theta U_{\theta(t)}(X_t) e^{A_t}]}{\mathbb{E}[e^{A_t}]} &= \frac{\int_{\mathbb{R}^d} \int_{-\infty}^{\infty} \partial_\theta U_{\theta(t)}(x) e^a f(t,x,a) da dx}{\int_{\mathbb{R}^d} \int_{-\infty}^{\infty} e^a f(t,x,a) da dx} = \frac{\int_{\mathbb{R}^d} \partial_\theta U_{\theta(t)}(x) \hat{\rho}(t,x) dx}{\int_{\mathbb{R}^d} \hat{\rho}(t,x) dx} \\ &= \frac{Z_{\theta_0}^{-1} \int_{\mathbb{R}^d} \partial_\theta U_{\theta(t)}(x) e^{-U_{\theta(t)}(x)} dx}{Z_{\theta_0}^{-1} Z_{\theta(t)}} = \mathbb{E}_{\theta(t)}[\partial_\theta U_{\theta(t)}] \end{aligned} \tag{3.2.10}$$

□

3.2.1 Details of Jarzynski correction and physical interpretation

Suppose that the walkers satisfy the Langevin equation (first equation in (3.2.1)):

$$dX_t = -\alpha \nabla U_{\theta(t)}(X_t) dt + \sqrt{2\alpha}\, dW_t, \qquad X_0 \sim \rho_{\theta_0}, \tag{3.2.11}$$

where $\theta(t)$ is evolving according to some protocol. The probability density function $\rho(t,x)$ of X_t then satisfies the Fokker-Planck equation

$$\partial_t \rho = \alpha \nabla \cdot \left(\nabla U_{\theta(t)}(x) \rho + \nabla \rho \right), \qquad \rho(t=0) = \rho_{\theta_0} \tag{3.2.12}$$

[8] Lawrence C Evans. *Partial differential equations.* vol. 19 American Mathematical Society, 2022.

Similarly to what we discussed about PCD, the solution to this equation is not available in closed form, and in particular $\rho(t,x) \neq \rho_{\theta(t)}(x)$; $\rho(t,x)$ is only close to $\rho_{\theta(t)}(x)$ if we let $\alpha \to \infty$, so that the walkers X_t move much faster than the parameters $\theta(t)$, but this limit is not easily achievable in practice (as convergence of the FPE solution to its equilibrium is very slow in general if the potential $U_{\theta(t)}$ is complicated). As a result $\mathbb{E}[\partial_\theta U_{\theta(t)}(X_t)] \neq \mathbb{E}_{\theta(t)}[\partial_\theta U_{\theta(t)}]$, implying the necessity to include the weights in the expectation (3.2.2). Specifically, we showed in the proof of Proposition 3.2.1 that the PDF of the walker X_t in (3.2.1), satisfies an equation like (3.2.12) with an additional birth-death contribution added to it:

$$\partial_t \rho = \alpha \nabla \cdot \big(\nabla U_{\theta(t)}(x)\rho + \nabla \rho\big) - \partial_\theta U_{\theta(t)}(x) \cdot \dot{\theta}(t)\rho + \lambda(t)\rho, \qquad \rho(t=0) = \rho_{\theta_0} \tag{3.2.13}$$

where $\lambda(t) = \int_{\mathbb{R}^d} \partial_\theta U_{\theta(t)}(x) \cdot \dot{\theta}(t)\rho dx$ is a Lagrange multiplier enforcing normalization of ρ.
Let us provide a physical intuition of the Jarzynski correction. Seeting $\alpha = 1$ in the SDE (3.2.11) for simplicity, we can compute the total differential of the energy $U_{\theta(t)}(X_t)$ using Ito's lemma:

$$\begin{aligned} dU_{\theta(t)}(X_t) = &\underbrace{\partial_\theta U_{\theta(t)}(X_t) \cdot \dot{\theta}(t) dt}_{d\mathcal{W}_{prot}(t)} \underbrace{-|\nabla U_{\theta(t)}(X_t)|^2 dt}_{d\mathcal{Q}_{diss}(t)} + \\ &+ \underbrace{\Delta U_{\theta(t)}(X_t) dt + \sqrt{2}\nabla U_{\theta(t)}(X_t) \cdot dW_t}_{d\mathcal{Q}_{therm}(t)} \end{aligned} \tag{3.2.14}$$

In the context of stochastic thermodynamics[9] this differential can be interpreted as the instantaneous variation of the total energy along a single realization of the SDE; thus, we expect the validity of a first-law-like relation, and we split the different contribution in work $\mathcal{W}$ and heat $\mathcal{Q}$ in the second line of (3.2.14). Let us analyze each term to motivate the notation: the first

$$d\mathcal{W}_{prot}(t) = \partial_\theta U_{\theta(t)}(X_t) \cdot \dot{\theta}(t) dt \tag{3.2.15}$$

is due to the change of the potential caused by a deliberate modification of the control parameter $\theta(t)$; it is then natural to interpret this contribution as the external work performed on the system, in analogy to experimental setups where $\theta(t)$ is a physical actor that modifies the potential, for instance the intensity of the laser in an optical trap[10]. The second term

$$d\mathcal{Q}_{diss}(t) = -|\nabla U_{\theta(t)}(X_t)|^2 dt \tag{3.2.16}$$

is the energy dissipated due to the deterministic part of the dynamic; mathematically, it is a consequence of the fact that $\dot{X}_t = -\nabla U_{\theta(t)}(X_t)$, for fixed $\theta(t)$, is a gradient

[9] Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on progress in physics*, **75**: 126001, 2012.
[10] Johannes Berner et al. Oscillating modes of driven colloids in overdamped systems. *Nature communications*, **9**: 999, 2018.

dynamical system[11]; for such ODE the energy U_θ is not a conserved quantity. Physically, (3.2.16) represents the instantaneously dissipated heat along the trajectory due to the interaction with the external potential; in real experiments, this potential can be for instance due to some external field applied to the system. Let us stress that this is an example in which a conservative force, i.e. the gradient of a scalar potential, leads to energy dissipation. Usually, one thinks about friction as prototype of energy dissipation phenomenon, but in the overdamped setup one can have also a "conservative source of dissipation."
Without the stochastic contribution to the dynamics, we can visualize the system as a particle moving downhill in a landscape: it follows the steepest descent, but the landscape itself is changing. Let us stress that $\dot{X}_t = -\nabla U_{\theta(t)}(X_t)$ is a first order ODE: in this sense we should imagine a particle subject to a peculiar time dependent and conservative friction. Without the external work, the particle would eventually stop ($\dot{X}_t = 0$); with the change of the potential, there exists a time dependent asymptotic solution:

Lemma 3.2.1. *Let us consider $U_{\theta(t)}(x)$ to be convex for any t. Given any two distinct solutions $X_1(t)$ and $X_2(t)$ of the ODE $\dot{X}_t = -\nabla U_{\theta(t)}(X_t)$ (i.e. solutions for different initial data), we have*

$$\frac{1}{2}\frac{d}{dt}(X_1 - X_2)^2 \leq 0 \tag{3.2.17}$$

Hence, by contraction arguments there exists a unique asymptotic solution ($\bar{X}(t) = \arg\min_x U_{\theta(t)}(x)$ is **not** *a solution).*

Proof. If we expand (3.2.17), we obtain

$$\frac{1}{2}\frac{d}{dt}(X_1 - X_2)^2 = (X_1 - X_2)\cdot(\dot{X}_1 - \dot{X}_2) = -(X_1 - X_2)\cdot(\nabla U_{\theta(t)}(X_1) - \nabla U_{\theta(t)}(X_2)), \tag{3.2.18}$$

where we used the ODE. Now, we can exploit convexity to conclude the proof

$$-(X_1 - X_2)(\nabla U_{\theta(t)}(X_1) - \nabla U_{\theta(t)}(X_2)) \leq -(X_1 - X_2)^T \nabla\nabla U_{\theta(t)}(X_1)(X_1 - X_2) \leq 0 \tag{3.2.19}$$

□

Finally, we can consider the stochastic contribution, that is the term

$$d\mathcal{Q}_{therm}(t) = \Delta U_{\theta(t)}(X_t)dt + \sqrt{2}\nabla U_{\theta(t)}(X_t)\cdot dW_t \tag{3.2.20}$$

Such part would be zero in absence of the stochastic term in the SDE and it represents the energy pumped into the system by a thermostat at temperature T. In computational experiments, this temperature has no other meaning than an hyperparameter that we set to $T = 1$ in the present context. On the other hand, in real experiments the temperature is associated to some physical property of the thermostat, e.g. the kinetic energy of particles in it. Whatever is the source of this energy pumping, its physical

[11] Stephen Smale. On gradient dynamical systems. *Annals of Mathematics*, 199–206, 1961.

effect is to contrast the dissipation due to the gradient nature of the deterministic part of the dynamics, so that the system remains at equilibrium.
To better understand this balance, let us consider the case when there is no external drive, i.e. $\theta(t) = \theta = cst$, and let us look at the average energy

$$\mathcal{U}(t) \coloneqq \mathbb{E}[U_\theta(X_t)]. \tag{3.2.21}$$

In absence of drive, we can establish convergence towards equililibrium, meaning that the probability density function $\rho(t,x)$ of X_t converges towards the stationary solution $\rho_\theta(x) = \exp(-U_\theta(x))/Z_\theta$. This also means that $\lim_{t\to\infty} d\mathcal{U}/dt = 0$, which also implies, after taking the time derivative of (3.2.21) and using (3.2.14) with $d\mathcal{W}_{prot} = 0$ (since $\theta(t) = \theta = cst$)

$$0 = \lim_{t\to\infty} \mathbb{E}\left[d\mathcal{Q}_{diss}(t) + d\mathcal{Q}_{therm}(t)\right] \tag{3.2.22}$$

To close the circle, the Jarzynski equality $F_{\theta(t)} = F_{\theta_0} - \log \mathbb{E}[e^{A_t}]$ provides an alternative way to measure the free energy, that is the partition function, for BG ensemble. It exploits a nonequilibrium process, in the sense that the potential in the Langevin SDE is time dependent. The ensemble is driven out of equilibrium since the law of the process X_t solution of the SDE in (3.2.1) is not $\rho_{\theta(t)}$ expect that for $t = 0$. However, if we reweight the particle distribution using the exponential of minus the cumulative work obtained from (3.2.15), then the "reweighted ensemble" is always in equilibrium with $\rho_{\theta(t)}$. Moreover, the weights can be used to calculate the free energy difference, and so the target free energy if we know F_{θ_0}. In Figure 3.1 we present a schematic simplified comparison: we consider the translation of a single well potential U_θ, and we plot how reweighting is supposed to correct the mismatch between the evolved ensemble and the desired ρ_θ.

3.3 Discrete time

A discrete time version of Proposition 3.2.1 is necessary in order to practically implement an algorithm. The derivation is closely related to methods of Annealing Importance Sampling[6]. The main result is the following:

Proposition 3.3.1. *Assume that the parameters θ are evolved by some time-discrete protocol $\{\theta_k\}_{k\in\mathbb{N}_0}$ and that (2.1.1) hold. Given any $h \in (0, L)$, let $X_k \in \mathbb{R}^d$ and $A_k \in \mathbb{R}$ be given by the iteration rule*

$$\begin{cases} X_{k+1} = X_k - h\nabla U_{\theta_k}(X_k) + \sqrt{2h}\,\xi_k, & X_0 \sim \rho_{\theta_0}, \\ A_{k+1} = A_k - \alpha_{k+1}(X_{k+1}, X_k) + \alpha_k(X_k, X_{k+1}), & A_0 = 0, \end{cases} \tag{3.3.1}$$

where $U_\theta(x)$ is the model energy, $\{\xi_k\}_{k\in\mathbb{N}_0}$ are independent $N(0_d, I_d)$, and we defined

$$\alpha_k(x,y) = U_{\theta_k}(x) + \tfrac{1}{2}(y-x)\cdot\nabla U_{\theta_k}(x) + \tfrac{1}{4}h|\nabla U_{\theta_k}(x)|^2 \tag{3.3.2}$$

Then, for all $k \in \mathbb{N}_0$,

$$\mathbb{E}_{\theta_k}[\partial_\theta U_{\theta_k}] = \frac{\mathbb{E}[\partial_\theta U_{\theta_k}(X_k)e^{A_k}]}{\mathbb{E}[e^{A_k}]}, \qquad Z_{\theta_k} = Z_{\theta_0}\mathbb{E}\left[e^{A_k}\right] \tag{3.3.3}$$

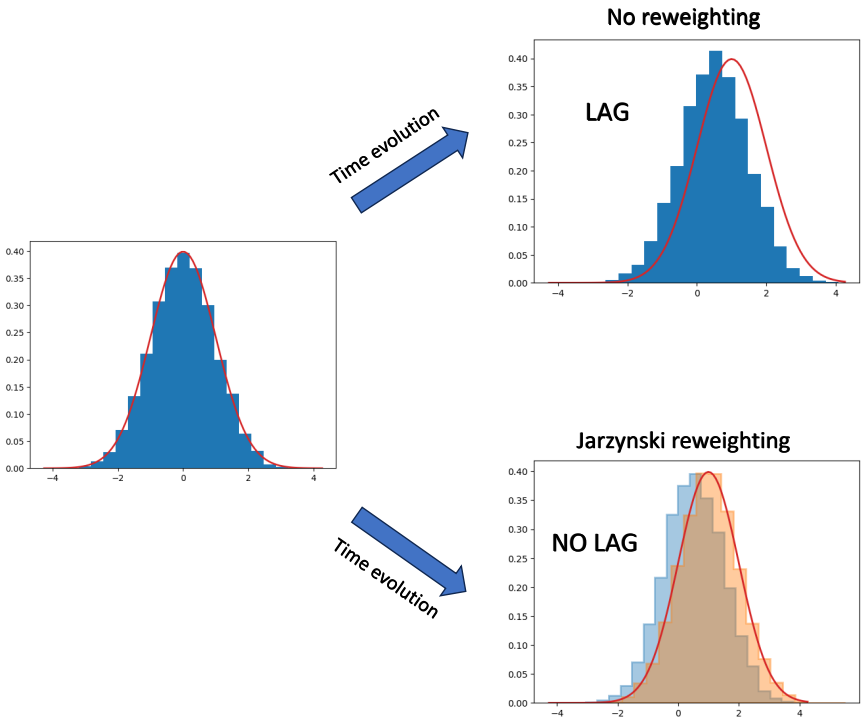


Figure 3.1: Schematic comparison of standard ULA and Jarzynski corrected algorithm for a translation of a single well potential. The blue one are the unweighted samples evolved via ULA, while the orange one are the reweighted ones.

where the expectations on the right-hand side are over the law of the joint process (X_k, A_k).

As in continuous time, we stress that the inclusion of the weights in (3.3.3) is the key, as $\mathbb{E}[\partial_\theta U_{\theta_k}(X_k)] \neq \mathbb{E}_{\theta_k}[\partial_\theta U_{\theta_k}]$ in general. We also stress that (3.3.3) holds *exactly* despite the fact that for $h > 0$ the iteration step for X_k in (3.3.1) is that of the unadjusted Langevin algorithm (ULA) with no Metropolis correction. That is, the inclusion of the weights A_k exactly corrects for the biases induced by both the slow mixing and the time-discretization errors in ULA.
Proposition 3.3.1 shows that we can evolve the parameters by gradient descent over the cross-entropy by solving (3.3.1) concurrently with

$$\theta_{k+1} = \theta_k + \gamma_k \mathcal{D}_k, \qquad \mathcal{D}_k = -\partial_\theta H(\rho_{\theta_k}, \rho_*) = \frac{\mathbb{E}[\partial_\theta U_{\theta_k}(X_k) e^{A_k}]}{\mathbb{E}[e^{A_k}]} - \mathbb{E}_*[\partial_\theta U_{\theta_k}], \tag{3.3.4}$$

where $\gamma_k > 0$ is the learning rate and $k \in \mathbb{N}_0$ with θ_0 given. We can also replace the gradient step for θ_k in (3.3.4) by any update optimization step (via AdaGrad, ADAM, etc.) that uses as input the gradient $\mathcal{D}_k$ of the cross-entropy evaluated at θ_k to get θ_{k+1}. Again, assuming that we know Z_{θ_0} we can track the evolution of the cross-entropy via

$$H(\rho_{\theta_k}, \rho_*) = \log \mathbb{E}[e^{A_k}] + \log Z_{\theta_0} + \mathbb{E}_*[U_{\theta_k}]. \tag{3.3.5}$$

Let us present the proof of Proposition 3.3.1.

Proof. The iteration rule for A_k in (3.3.1) implies that

$$A_k = \sum_{q=1}^{k} \left(\alpha_{q-1}(X_{q-1}, X_q) - \alpha_q(X_q, X_{q-1}) \right), \qquad k \in \mathbb{N}. \tag{3.3.6}$$

For $k \in \mathbb{N}_0$, let

$$\beta_k(x, y) = (2\pi h)^{-d/2} \exp\Big(-\frac{1}{4h} \left| y - x + h \nabla U_{\theta_k}(x) \right|^2 \Big) \tag{3.3.7}$$

be the transition probability density of the ULA update in (3.3.1), i.e. $\beta_k(X_k, X_{k+1})$ is the probability density of X_{k+1} conditionally on X_k. By the definition of A_k, we have

$$\begin{aligned} \exp(A_k) &= \prod_{q=1}^{k} \exp\left(\alpha_{q-1}(X_{q-1}, X_q) - \alpha_q(X_q, X_{q-1}) \right) \\ &= e^{-U_{\theta_k}(X_k) + U_{\theta_0}(X_0)} \prod_{q=1}^{k} \frac{\beta_q(X_q, X_{q-1})}{\beta_{q-1}(X_{q-1}, X_q)} \end{aligned} \tag{3.3.8}$$

where in the second line we added and subtracted $|X_q - X_{q-1}|^2/4h$ and used the definition of $\alpha_k(x, y)$ given in (3.3.2). Since the joint probability density function of

the path $(X_0, X_1, \ldots, X_k)$ at any $k \in \mathbb{N}$ is

$$\varrho(x_0, x_1, \ldots, x_k) = Z_{\theta_0}^{-1} e^{-U_{\theta_0}(x_0)} \prod_{q=1}^{k} \beta_{q-1}(x_{q-1}, x_q) \tag{3.3.9}$$

we deduce from (3.3.8) and (3.3.9) that, given an $f : \mathbb{R}^d \to \mathbb{R}$, we can express the expectation $\mathbb{E}[f(X_k)e^{A_k}]$ as the integral

$$\begin{aligned}
&\mathbb{E}[f(X_k)e^{A_k}] \\
&= \int_{\mathbb{R}^{dk}} f(x_k) e^{-U_{\theta_k}(x_k) + U_{\theta_0}(x_0)} \prod_{q=1}^{k} \frac{\beta_q(x_q, x_{q-1})}{\beta_{q-1}(x_{q-1}, x_q)} \varrho(x_0, x_1, \ldots, x_k) dx_0 \cdots dx_k \\
&= Z_{\theta_0}^{-1} \int_{\mathbb{R}^{dk}} f(x_k) e^{-U_{\theta_k}(x_k)} \prod_{q=1}^{k} \beta_q(x_q, x_{q-1}) dx_0 \cdots dx_k
\end{aligned} \tag{3.3.10}$$

Since $\int_{\mathbb{R}^d} \beta_k(x, y) dy = 1$ for all $k \in \mathbb{N}_0$ and all $x \in \mathbb{R}^d$, we can perform the integrals over x_0, then x_1, etc. in this expression to be left with

$$\mathbb{E}[f(X_k)e^{A_k}] = Z_{\theta_0}^{-1} \int_{\mathbb{R}^d} f(x_k) e^{-U_{\theta_k}(x_k)} dx_k \tag{3.3.11}$$

Setting $f(x) = 1$ in this expression gives

$$\mathbb{E}[e^{A_k}] = Z_{\theta_0}^{-1} \int_{\mathbb{R}^d} e^{-U_{\theta_k}(x_k)} dx_k = Z_{\theta_0}^{-1} Z_{\theta_k} \tag{3.3.12}$$

which implies the second equation in (3.3.3); setting $f(x_k) = \partial_\theta U_{\theta_k}(x_k)$ in (3.3.11) gives

$$\mathbb{E}[\partial_\theta U_{\theta_k}(X_k) e^{A_k}] = Z_{\theta_0}^{-1} \int_{\mathbb{R}^d} \partial_\theta U_{\theta_k}(x_k) e^{-U_{\theta_k}(x_k)} dx_k = Z_{\theta_0}^{-1} Z_{\theta_k} \mathbb{E}_{\theta_k}[\partial_\theta U_{\theta_k}] \tag{3.3.13}$$

which can be combined with (3.3.12) to arrive at the first equation in (3.3.3). □

Since the proofs of the main propositions 3.2.1 and 3.3.1 exploit different mathematical tools, a natural question is whether the continuous limit $h \to 0$ of (3.3.1) is (3.2.1). For the SDE, this sanity check is trivial by definition of Langevin equation. On the other hand, the limit of the second equation in (3.3.1) is less obvious. Here we present the derivation of the consistency result. For the sake of simplicity, we use the compact notation $U_{\theta_k} \coloneqq U_k$ and $\nabla\nabla U_k$ to denote the jacobian of the gradient. Then, we expand $A_{k+1} - A_k$ in power series of h, that is

$$\begin{aligned}
A_{k+1} - A_k = U_k(X_k) - U_{k+1}(X_{k+1}) + (X_{k+1} - X_k) \cdot \nabla U_k(X_k) + \\
+ \frac{1}{2}(X_{k+1} - X_k)^T \nabla\nabla U_k(X_k)(X_{k+1} - X_k) + O(h^{3/2})
\end{aligned} \tag{3.3.14}$$

where

$$U_k(X_k) - U_{k+1}(X_{k+1}) = -\partial_t U_t(X_t)|_{t=k} h - (X_{k+1} - X_k) \cdot \nabla U_k(X_k) + \\ - \frac{1}{2}(X_{k+1} - X_k)^T \nabla\nabla U_k(X_k)(X_{k+1} - X_k) + O(h^{3/2}) \tag{3.3.15}$$

and for which we used the definition A_{k+1} given in (3.3.1). After simplifications, we obtain

$$A_{k+1} - A_k = -\partial_t U_t(X_t)|_{t=hk} h + O(h^{3/2}) \tag{3.3.16}$$

As we expected, the only term of $O(h)$ is $-\partial_t U_t(X_t)|_{t=hk} h$, that is for $h \to 0$ we recover the ODE in (3.2.1). Moreover, we can interpret the definition of A_{k+1} from a different angle as a first order approximation of the integral solution of the ODE in (3.2.1), namely

$$A(t+h) = A(t) - \int_t^{t+h} \partial_\theta U_{\theta_s}(X_s) \cdot \dot{\theta}(s) ds \tag{3.3.17}$$

3.4 Algorithmic aspects

In spite of a practical implementation of the proposed routine, we introduce N independent pairs of walkers and weights, $\{X_k^i, A_k^i\}_{i=1}^N$, which we evolve independently using (3.3.1) for each pair. To evolve θ_k from some prescribed θ_0 we can then use the empirical version of (3.3.4):

$$\theta_{k+1} = \theta_k + \gamma_k \tilde{\mathcal{D}}_k, \tag{3.4.1}$$

where $\tilde{\mathcal{D}}_k$ is the estimator for the gradient in θ of the cross-entropy:

$$\tilde{\mathcal{D}}_k = \frac{\sum_{i=1}^N \partial_\theta U_{\theta_k}(X_k^i) \exp(A_k^i)}{\sum_{i=1}^N \exp(A_k^i)} - \frac{1}{n}\sum_{j=1}^n \partial_\theta U_{\theta_k}(x_*^j), \tag{3.4.2}$$

These steps are summarized in Algorithm 3, which is a specific instance of a sequential Monte-Carlo algorithm[12]. During the calculation, we can monitor the evolution of the partition function and the cross-entropy using as estimators

$$\tilde{Z}_{\theta_k} = Z_{\theta_0} \frac{1}{N}\sum_{i=1}^N \exp(A_k^i), \qquad \tilde{H}_k = \log \tilde{Z}_{\theta_k} + \frac{1}{n}\sum_{j=1}^n U_{\theta_K}(x_*^j) \tag{3.4.3}$$

We can also use mini-batches of $\{X_k^i, A_k^i\}_{i=1}^N$ and the data set $\{x_*^j\}_{j=1}^n$ at every iteration, and switch to any optimizer step that uses θ_k and $\mathcal{D}_k$ as input to get the updated θ_{k+1}. In Algorithm 4 we present such mini-batched version of Algorithm 3, where we do not update the positions of every walker at every iteration. Instead, we evolve only a small portion of the walkers, while keeping the other walkers frozen and only updating their weights using the information from the model update (which uses only the model energy and does not require back-propagation). This mini-batched version

[12] Jun S Liu. *Monte Carlo strategies in scientific computing*. vol. 75 Springer, 2001.

Algorithm 3 Sequential Monte-Carlo training with Jarzynski correction

Inputs: data points $\{x_*^i\}_{i=1}^n$; energy model U_θ; optimizer step $\mathrm{opt}(\theta, \mathcal{D})$ using θ and the empirical CE gradient $\mathcal{D}$; initial parameters θ_0; number of walkers $N \in \mathbb{N}_0$; set of walkers $\{X_0^i\}_{i=1}^N$ sampled from ρ_{θ_0}; total duration $K \in \mathbb{N}$; ULA time step h; set of positive constants $\{c_k\}_{k\in\mathbb{N}}$.
$A_0^i = 0$ for $i = 1, \dots, N$.
for $k = 0, \dots, K-1$ **do**
  $p_k^i = \exp(A_k^i) / \sum_{j=1}^N \exp(A_k^j)$ ▷ normalized weights
  $\tilde{\mathcal{D}}_k = \sum_{i=1}^N p_k^i \partial_\theta U_{\theta_k}(X_k^i) - n^{-1} \sum_{j=1}^n \partial_\theta U_{\theta_k}(x_*^j)$ ▷ empirical CE gradient
  $\theta_{k+1} = \mathrm{opt}(\theta_k, \tilde{\mathcal{D}}_k)$ ▷ optimization step
  for $i = 1, ..., N$ **do**
    $X_{k+1}^i = X_k^i - h\nabla U_{\theta_k}(X_k^i) + \sqrt{2h}\,\xi_k^i, \qquad \xi_k^i \sim \mathcal{N}(0_d, I_d)$ ▷ ULA
    $A_{k+1}^i = A_k^i - \alpha_{k+1}(X_{k+1}^i, X_k^i) + \alpha_k(X_k^i, X_{k+1}^i)$ ▷ weight update
  end for
  Resample the walkers and reset the weights if $\mathrm{ESS}_{k+1} < c_{k+1}$, see (3.4.4). ▷ resampling step
end for
Outputs: Optimized energy U_{θ_K}; set of weighted walkers $\{X_K^i, A_K^i\}_{i=1}^N$ sampling ρ_{θ_K}; partition function estimate $\tilde{Z}_{\theta_K} = Z_{\theta_0} N^{-1} \sum_{i=1}^N \exp(A_K^i)$; CE estimate $\log \tilde{Z}_{\theta_K} + n^{-1} \sum_{j=1}^n U_{\theta_K}(x_*^j)$

is more computationally efficient and leads to convergence in training with much fewer steps of ULA (see Chapter 4 for experiments). More importantly, the mini-batched version of the algorithm, as compared to the full-batched version, enlarges the total sample size and therefore improves the sample variety with even less computational cost. The only sources of error in these algorithms come from the finite sample sizes, $N < \infty$ and $n < \infty$. Regarding n, we may need to add a regularization term in the loss to avoid overfitting: this is standard in machine learning applications. Regarding N and the presence of a weighted average, we need to make sure that the effective sample size[13] of the walkers remains sufficient during the evolution. This is nontrivial since the A_k^i's will spread away from zero during the optimization, implying that the weights $\exp(A_k^i)$ will become non-uniform, thereby reducing the effective sample size. This is a known issue with sequential Monte-Carlo algorithms that can be alleviated by resampling as discussed next.
A standard quantity to monitor the effective sample size is the ratio between the square of the empirical mean of the weights and their empirical variance, i.e.

$$\mathrm{ESS}_k = \frac{\left(N^{-1} \sum_{i=1}^N \exp(A_k^i)\right)^2}{N^{-1} \sum_{i=1}^N \exp(2A_k^i)} \in (0, 1] \tag{3.4.4}$$

[13] James Berger, MJ Bayarri, and LR Pericchi. The effective sample size. *Econometric Reviews*, **33**: 197–217, 2014.

Algorithm 4 Mini-batched Sequential Monte-Carlo training with Jarzynski correction

Inputs: data points $\{x_i^*\}_{i=1}^n$; energy model U_θ; optimizer step $\mathrm{opt}(\theta, \mathcal{D})$ using θ and the empirical CE gradient $\mathcal{D}$; initial parameters θ_0; number of walkers $N \in \mathbb{N}_0$; batch size $N' \in \mathbb{N}_0$ with $N' < N$, set of walkers $\{X_0^i\}_{i=1}^N$ sampled from ρ_{θ_0}; total duration $K \in \mathbb{N}$; ULA time step h; set of positive constants $\{c_k\}_{k\in\mathbb{N}}$.
$A_0^i = 0$ for $i = 1, \dots, N$.
for $k = 1 : K - 1$ **do**
  $p_k^i = \exp(A_k^i) / \sum_{j=1}^N \exp(A_k^j)$ ▷ normalized weights
  $\tilde{\mathcal{D}}_k = \sum_{i=1}^N p_k^i \partial_\theta U_{\theta_k}(X_k^i) - n^{-1} \sum_{j=1}^n \partial_\theta U_{\theta_k}(x_*^j)$ ▷ empirical CE gradient
  $\theta_{k+1} = \mathrm{opt}(\theta_k, \tilde{\mathcal{D}}_k)$ ▷ optimization step
  Randomly select a mini-batch $\{X_k^j\}_{j\in B}$ with $\#B = N'$ from the set of walkers $\{X_k^i\}_{i=1}^N$
  for $j \in B$ **do**
    $X_{k+1}^j = X_k^j - h\nabla U_{\theta_k}(X_k^j) + \sqrt{2h}\,\xi_k^j, \qquad \xi_k^j \sim \mathcal{N}(0_d, I_d)$ ▷ ULA
    $A_{k+1}^j = A_k^j - \alpha_{k+1}(X_{k+1}^j, X_k^j) + \alpha_k(X_k^j, X_{k+1}^j)$ ▷ weight update
  end for
  for $j \notin B$ **do**
    $X_{k+1}^j = X_k^j$ ▷ no update of the walkers
    $A_{k+1}^j = A_k^j - U_{\theta_{k+1}}(X_k^j) + U_{\theta_k}(X_k^j)$ ▷ weight update
  end for
  Resample the walkers and reset the weights if $\mathrm{ESS}_{k+1} < c_{k+1}$, see (3.4.4). ▷ resampling step
end for
Outputs: Optimized energy U_{θ_K}; set of weighted walkers $\{X_K^i, A_K^i\}_{i=1}^N$ sampling ρ_{θ_K}; partition function estimate $\tilde{Z}_{\theta_K} = Z_{\theta_0} N^{-1} \sum_{i=1}^N \exp(A_K^i)$; CE estimate $\log \tilde{Z}_{\theta_K} + n^{-1} \sum_{j=1}^n U_{\theta_K}(x_*^j)$

The effective sample size of the N walkers is $\mathrm{ESS}_k N$. Initially, since $A_0^i = 0$, $\mathrm{ESS}_0 = 1$, but it decreases with k. At each iteration k_r such that $\mathrm{ESS}_{k_r} < c_{k_r}$, where $\{c_k\}_{k\in\mathbb{N}}$ is a set of predefined positive constants in $(0, 1)$, we then:

1. Resample the walkers $X_{k_r}^i$ using $p_{k_r}^i = e^{A_{k_r}^i} / \sum_{j=1}^N e^{A_{k_r}^j}$ as probability to pick walker i;

2. Reset $A_{k_r}^i = 0$;

3. Use the update $Z_{\theta_k} = Z_{\theta_{k_r}} N^{-1} \sum_{i=1}^N \exp(A_k^i)$ for $k \geq k_r$ until the next resampling step.

Resampling schemes are necessary to tackle the decay of effective sample size (3.4.4). For the sake of completeness, here we recall three of the most widely used routines:

multinomial[14], stratified[15], and systematic resampling[16], and refer the reader to the review[17] for more details.
Given a set of normalized scalar weights $\{p_i\}_{i=1}^N \in [0,1]$ with $\sum_{i=1}^N p_i = 1$ associated to the N walkers, we define the cumulative sum

$$P_n = \sum_{i=1}^{n} p_i, \qquad n = 1, \dots, N \tag{3.4.5}$$

All three methods prescribe a way to choose a set $\{u_n\}_{n=1}^N \in (0,1]$ used to perform the resampling in the following way: for every $n, m \in \{1, \dots, N\}$, the m-th particle is chosen during the n-th extraction if

$$P_{m-1} < u_n < P_m \tag{3.4.6}$$

Let us now specify how the set of u_n is selected in the cases in study. We denote by $\mathcal{U}(a,b]$ as usual the uniform probability distribution on the interval $(a,b]$.

- **Multinomial resampling.** Sample $u_n^{\text{mult}} \sim \mathcal{U}(0,1]$, independently for every $n \in \{1, \dots, N\}$. This approach leads to a large number of possible resampled configurations, which is not desirable in practice as it increases the variance of the estimator.
- **Stratified resampling.** Partition the interval $(0,1]$ into N sub-intervals, or strata, of size $1/N$; then, sample $u_n^{\text{str}} \sim \mathcal{U}((n-1)/N, n/N]$ independently for each $n = \{1, \dots, N\}$. This approach picks a single u_n in each stratus, thereby reducing the number of possible resampled configurations.
- **Systematic resampling.** Partition the interval $(0,1]$ into N sub-intervals, or strata, of size $1/N$; then, sample $u_1^{\text{sys}} \sim \mathcal{U}(0, 1/N]$, and $u_n^{\text{sys}} = u_1^{\text{sys}} + (n-1)/N$ for $n > 1$. This method also reduces the number of possible resampled configurations.

Note that there are various modifications of these three methods[17], all of them meeting the so-called unbiasedness condition, that is, the i-particle is expected to be sampled in average Np_i times. However, these extensions do not lead to a critical lowering of the number of possible resampled configurations compared to systematic resampling. For this reason in our numerical experiments we only tested the three methods above. Note also that, regardless of the resampling routine one uses, a fundamental role is played by the variance of the weights: if the cumulative sum is dominated by one or

[14] Neil J Gordon, David J Salmond, and Adrian FM Smith. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation" in: *IEE proceedings F (radar and signal processing)*. vol. 140 IET 1993. 107–113

[15] Genshiro Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of computational and graphical statistics*, **5**: 1–25, 1996.

[16] James Carpenter, Peter Clifford, and Paul Fearnhead. Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, **146**: 2–7, 1999.

[17] Tiancheng Li, Miodrag Bolic, and Petar M Djuric. Resampling methods for particle filtering: classification, implementation, and strategies. *IEEE Signal processing magazine*, **32**: 70–86, 2015.

very few weights, i.e. resampling is triggered too late, it does not remedy the suppression of population variability. Other criteria, based e.g. on the entropy of the weights, are also possible[18].
As a side remark, the application of the proposed algorithm could be in principle extended in the context of Restricted Boltzmann Machines[19]. As we briefly mentioned in the Introduction, Boltzmann Machines are physics inspired generative neural network used in unsupervised learning which can be interpreted as EBMs. In fact, their energy function in the notation of the present work is

$$U_\theta(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^{D}\sum_{j=1}^{J} v_i W_{ij} h_j - \sum_{j=1}^{J} b_j h_j - \sum_{i=1}^{D} c_i v_i, \tag{3.4.7}$$

where $\boldsymbol{v} = \{v_i\}_{i=1}^{D}$ are the so-called visible units, $\boldsymbol{h} = \{h_i\}_{i=1}^{J}$ are the hidden units and $\theta = \{\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}\}$ are the learnable parameters. The layers of visible and hidden units have with connections between them but not within the same layer (see Figure 3.2). RBMs

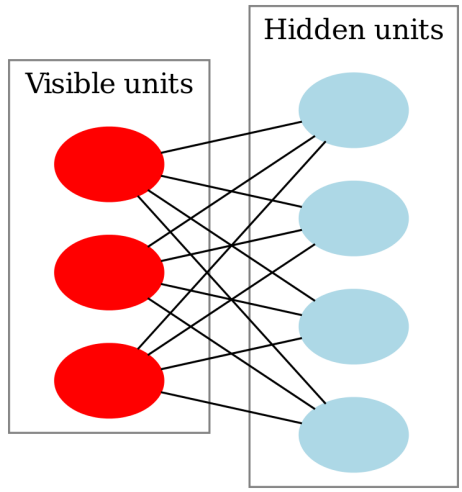


Figure 3.2: Schematic representation of Restricted Boltzmann Machines.

are trained to learn the underlying patterns in the data by adjusting the weights of connections to reconstruct input data. They are proficient in modeling complex probability distributions and are used in various applications such as collaborative filtering, feature learning, and dimensionality reduction. The model one is interested

[18] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli*, **18**: 252–278, 2012.
[19] Nan Zhang et al. An overview on restricted Boltzmann machines. *Neurocomputing*, **275**: 1186–1199, 2018.

to is actually the density marginalized on the hidden layers

$$\rho_\theta(\boldsymbol{v}) = \frac{1}{Z_\theta} \sum_{\boldsymbol{h}} e^{-U_\theta(\mathbf{v},\mathbf{h})} \tag{3.4.8}$$

Due to the distinct roles of hidden and visible layers, and the challenge of marginalization, RBMs utilize a variational approach similar to Variational Autoencoders for training, as the marginal density is analytically intractable. Subsequently, the contrastive divergence algorithm is employed for efficient gradient-based optimization. We will not delve into further technical details here; however, it is noteworthy that, for the convenience of the RBMs community, Contrastive Divergence could potentially be replaced by our proposed method.

Chapter 4

Applications and Numerics

4.1 Theoretical Application: generative phase of diffusion models

In this section we present a theoretical application of Jarzynski reweighting in a more general context. Let us consider the following SDE for $\varepsilon > 0$

$$dX_t = [b(t, X_t) + \varepsilon s(t, X_t)]dt + \sqrt{2\varepsilon}\, dW_t \tag{4.1.1}$$

where $b(t,x)$ and $s(t,x)$ are respectively the drift and the score as defined in stochastic interpolation. That is

$$s(t,x) = -\nabla U_{\theta(t)}(x) \coloneqq -\nabla\phi(t,x) \tag{4.1.2}$$

and $X_t \sim \rho(x,t) = \exp(-\phi(t,x))/Z_t$ where $\nabla \cdot (b\rho) = \partial_t \rho$. In the light of the definition of Jarzynski correction, notice how the role of $b(t,x)$ in stochastic interpolant framework[1] is, in some sense, to substitute the Jarzynski correction — as Jarzynski himself suggested[2] before the advent of generative models, the estimation of free energy differences would be perfect if one knew the exact b. The meaning of such field is to keep the ensemble in equilibrium with Boltzmann-Gibbs density associated to the time dependent potential. In fact, we recall as the Fokker-Planck PDE associated to (4.1.1) is

$$\partial_t \rho + \nabla \cdot ((b + \varepsilon s)\rho - \varepsilon \nabla \rho) = 0 \tag{4.1.3}$$

meaning that $\rho(x,t) = \exp(-\phi(t,x))/Z_t$ is a solution for the initial datum $\rho(x,0) = \exp(-\phi(0,x))/Z_t$. There is no need for Jarzynski correction in presence of b.
The Euler-Maruyama time discretization of (4.1.2) yields

$$X_{k+1} = X_k + [b(t_k, X_t) + \varepsilon s(t_k, X_t)]h + \sqrt{2h\varepsilon}\, \xi_k, \qquad X_0 \sim \rho_0, \tag{4.1.4}$$

[1] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[2] Suriyanarayanan Vaikuntanathan and Christopher Jarzynski. Escorted free energy simulations: Improving convergence by reducing dissipation. *Physical Review Letters*, **100**: 190601, 2008.

For simplicity, given any function $f(t_k, x)$ we will use the shorthand $f_k(x) \coloneqq f(t_k, x)$ w.r.t. dependence in time. As we mentioned in Section 2.2.3, one critical issue in diffusion-based generative models is that generation corresponds to integration of an SDE (or ODE). The trade-off between a smaller discretization error and an higher computational cost is critical in the choice of the integrator. In spite of Chapter 3, we now show how is possible to correct the error induced by Euler-Maruyama of the exact SDE (4.1.4) using Jarzynski weights.

Proposition 4.1.1. *Assume that the parameters θ are evolved by some time-discrete protocol $\{\theta_k\}_{k\in\mathbb{N}_0}$ and that (2.1.1) hold. Given any $h \in (0, L)$, let $X_k \in \mathbb{R}^d$ and $A_k \in \mathbb{R}$ be given by the iteration rule*

$$\begin{cases} X_{k+1} = X_k + [b(t_k, X_t) + \varepsilon s(t_k, X_t)]h + \sqrt{2h\varepsilon}\,\xi_k, & X_0 \sim \rho_0, \\ A_{k+1} = A_k + \zeta_k(X_k, X_{k+1}) & A_0 = 0, \end{cases} \tag{4.1.5}$$

where $\{\xi_k\}_{k\in\mathbb{N}_0}$ are independent $N(0_d, I_d)$ and $\nabla \cdot (b\rho) = \partial_t \rho$, and we defined

$$\begin{aligned} \zeta_k(x, y) = \phi_k(x) - \phi_{k+1}(y) + \frac{1}{2}(x - y) \cdot \left(\frac{[b_k(x) - b_{k+1}(y)]}{\varepsilon} + [s_k(x) + s_{k+1}(y)] \right) \\ + \frac{h}{4} \left(\frac{|b_k(x)|^2 - |b_{k+1}(y)|^2}{\varepsilon} + \varepsilon[|s_k(x)|^2 - |s_{k+1}(y)|^2] \right) + \\ + \frac{h}{2}(b_k(x) \cdot s_k(x) + b_{k+1}(y) \cdot s_{k+1}(y)) \end{aligned} \tag{4.1.6}$$

Then, for all $k \in \mathbb{N}_0$,

$$\mathbb{E}_{\theta_k}[\partial_\theta U_{\theta_k}] = \frac{\mathbb{E}[\partial_\theta U_{\theta_k}(X_k) e^{A_k}]}{\mathbb{E}[e^{A_k}]}, \qquad Z_{\theta_k} = Z_{\theta_0} \mathbb{E}\left[e^{A_k}\right] \tag{4.1.7}$$

where the expectations on the right-hand side are over the law of the joint process (X_k, A_k).

Proof. The proof is a generalization of the one of Proposition 3.3.1. The iteration rule for A_k in (4.1.5) implies that

$$A_k \coloneqq \sum_{q=1}^{k} \left(\alpha_{q-1}(X_{q-1}, X_q) - \lambda_q(X_q, X_{q-1}) \right), \qquad k \in \mathbb{N}. \tag{4.1.8}$$

where

$$\alpha_k(x, y) = \phi(t_k, x) - \frac{1}{2\varepsilon}(y - x) \cdot [b_k(x) + \varepsilon s_k(x)] + \frac{h}{4\varepsilon}|b_k(x) + \varepsilon s_k(x)|^2 \tag{4.1.9}$$

and

$$\lambda_k(x, y) = \phi(t_k, x) - \frac{1}{2\varepsilon}(y - x) \cdot [-b_k(x) + \varepsilon s_k(x)] + \frac{h}{4\varepsilon}| - b_k(x) + \varepsilon s_k(x)|^2 \tag{4.1.10}$$

For $k \in \mathbb{N}_0$, let

$$\beta_k^{forw}(x,y) = (2\pi h\varepsilon)^{-d/2} \exp\Big(-\frac{1}{4h\varepsilon} |y - x - h[b_k(x) + \varepsilon s_k(x)]|^2 \Big) \tag{4.1.11}$$

be the transition probability density of the ULA update in (4.1.5), i.e. $\beta_k^{forw}(X_k, X_{k+1})$ is the probability density of X_{k+1} conditionally on X_k. Similarly,

$$\beta_k^{back}(x,y) = (2\pi h\varepsilon)^{-d/2} \exp\Big(-\frac{1}{4h\varepsilon} |y - x - h[-b_k(x) + \varepsilon s_k(x)]|^2 \Big) \tag{4.1.12}$$

Notice the minus sign before b in the backward transition kernel. By the definition of A_k, we have

$$\begin{aligned} \exp(A_k) &= \prod_{q=1}^{k} \exp\left(\alpha_{q-1}(X_{q-1}, X_q) - \lambda_q(X_q, X_{q-1})\right) \\ &= e^{-U_{\theta_k}(X_k) + U_{\theta_0}(X_0)} \prod_{q=1}^{k} \frac{\beta_q^{back}(X_q, X_{q-1})}{\beta_{q-1}^{forw}(X_{q-1}, X_q)} \end{aligned} \tag{4.1.13}$$

where in the second line we added and subtracted $|X_q - X_{q-1}|^2/4h\varepsilon$ and used the definition of $\alpha_k(x,y)$ and $\lambda_k(x,y)$ given in (4.1.9) and (4.1.10). Since the joint probability density function of the path $(X_0, X_1, \ldots, X_k)$ at any $k \in \mathbb{N}$ is

$$\varrho(x_0, x_1, \ldots, x_k) = Z_{\theta_0}^{-1} e^{-U_{\theta_0}(x_0)} \prod_{q=1}^{k} \beta_{q-1}^{forw}(x_{q-1}, x_q) \tag{4.1.14}$$

we deduce from (4.1.13) and (4.1.14) that, given an $f : \mathbb{R}^d \to \mathbb{R}$, we can express the expectation $\mathbb{E}[f(X_k)e^{A_k}]$ as the integral

$$\begin{aligned} &\mathbb{E}[f(X_k)e^{A_k}] \\ &= \int_{\mathbb{R}^{dk}} f(x_k) e^{-U_{\theta_k}(x_k) + U_{\theta_0}(x_0)} \prod_{q=1}^{k} \frac{\beta_q^{back}(x_q, x_{q-1})}{\beta_{q-1}^{forw}(x_{q-1}, x_q)} \varrho(x_0, x_1, \ldots, x_k) dx_0 \cdots dx_k \\ &= Z_{\theta_0}^{-1} \int_{\mathbb{R}^{dk}} f(x_k) e^{-U_{\theta_k}(x_k)} \prod_{q=1}^{k} \beta_q^{back}(x_q, x_{q-1}) dx_0 \cdots dx_k \end{aligned} \tag{4.1.15}$$

Since $\int_{\mathbb{R}^d} \beta_k(x,y)dy = 1$ for all $k \in \mathbb{N}_0$ and all $x \in \mathbb{R}^d$, we can perform the integrals over x_0, then x_1, etc. in this expression to be left with

$$\mathbb{E}[f(X_k)e^{A_k}] = Z_{\theta_0}^{-1} \int_{\mathbb{R}^d} f(x_k) e^{-U_{\theta_k}(x_k)} dx_k \tag{4.1.16}$$

Setting $f(x) = 1$ in this expression gives

$$\mathbb{E}[e^{A_k}] = Z_{\theta_0}^{-1} \int_{\mathbb{R}^d} e^{-U_{\theta_k}(x_k)} dx_k = Z_{\theta_0}^{-1} Z_{\theta_k} \tag{4.1.17}$$

which implies Jarzynski identity; combining (4.1.16) and (4.1.17) we obtain the fundamental relation

$$\mathbb{E}_{t_k}[f] = \frac{\mathbb{E}[f(X_k)e^{A_k}]}{\mathbb{E}[e^{A_k}]} \tag{4.1.18}$$

□

To get an insight of the role of b we can expand $A_{k+1} - A_k$ in series of h. We notice as $O(h^{1/2})$ vanishes since

$$\begin{aligned} A_{k+1} - A_k = &\underbrace{-\partial_t\phi_k(X_k)h + (X_{k+1} - X_k)\cdot s_k(X_k) + \frac{1}{2}(X_{k+1} - X_k)^T J_k^s(X_k)(X_{k+1} - X_k)}_{\phi_k(X_k) - \phi_{k+1}(X_{k+1})} \\ &-(X_{k+1} - X_k)\cdot s_k(X_k) + \frac{1}{2\varepsilon}(X_{k+1} - X_k)^T J_k^b(X_k)(X_{k+1} - X_k) \\ &-\frac{1}{2}(X_{k+1} - X_k)^T J_k^s(X_k)(X_{k+1} - X_k) + hb_k(X_k)\cdot s_k(X_k) + O(h^{3/2}) \end{aligned} \tag{4.1.19}$$

where J_b and J_s are the Jacobian of b and s. The blue and the red terms simplify and we get

$$\begin{aligned} A_{k+1} - A_k = -\partial_t\phi_k(X_k)h + \frac{1}{2\varepsilon}(X_{k+1} - X_k)^T J_k^b(X_k)(X_{k+1} - X_k) + \\ + hb_k(X_k)\cdot s_k(X_k) + O(h^{3/2}) \end{aligned} \tag{4.1.20}$$

We can use (4.1.4) and $\xi_k^T J_k^b(X_k)\xi_k \overset{d}{=} \nabla\cdot b_k(X_k)$ (in law) to further manipulate the expression

$$A_{k+1} - A_k = -\partial_t\phi_k(X_k)h + hb_k(X_k)\cdot s_k(X_k) + h\nabla\cdot b_k(X_k) + O(h^{3/2}) \tag{4.1.21}$$

In conclusion, we notice how $\partial_t\rho_t(x) + \nabla\cdot(b(t,x)\rho_t(x)) = 0$ implies

$$-\partial_t\phi_k(X_k) + b_k(X_k)\cdot s_k(X_k) + \nabla\cdot b_k(X_k) = 0 \tag{4.1.22}$$

that yields to the expected $A_{k+1} - A_k = O(h^{3/2})$. This computation shows that we have no Jarzynski correction in the continuous limit. On the contrary, for any finite time step h, the weights have a non trivial evolution, meaning that their presence is necessary to correct time discretization error.

Remark 4.1.1. *Given an SDE, any sufficiently smooth drift can be always be decomposed in a gradient field and a divergence-free field thanks to Helmholtz decomposition*[3]*. Hence, the treatment of the present section can be also interpreted per se as a corrector for Euler-Maruyama integrator.*

[3] Erhard Glötzl and Oliver Richters. Helmholtz decomposition and potential functions for n-dimensional analytic vector fields. *Journal of Mathematical Analysis and Applications*, **525**: 127138, 2023.

Remark 4.1.2. *The derivation we presented can be generalized to any transition kernel, even containing dissipative contributions, as far as it transforms the BG ensembles associated to ϕ_k at time k with the one at time $k+1$. In this sense a further investigation of different dynamics in place of basic Langevin Algorithm (e.g. Kinetic Langevin Diffusion, etc.) could represent a notable follow-up for the present results.*

4.2 Highlighted example: Gaussian Mixture

In this section we propose the analytical study of an exemplary toy model. Despite its apparent simplicity, notable differences between CD, PCD and our Jarzynski corrected algorithm will already arise. As target density we take the so-called Gaussian Mixture Model (GMM) in one dimension

$$\rho_*(x) = \frac{e^{-\frac{1}{2}|x-a|^2} + e^{-z_*-\frac{1}{2}|x-b|^2}}{\sqrt{2\pi}(1+e^{-z_*})}, \tag{4.2.1}$$

which is a superposition of two gaussians. Here $z_* \in \mathbb{R}$ parameterizes the mass of the second mode relative to the first and is the sole parameter of interest. The proportion of samples in both modes is indeed

$$p_* := \frac{1}{1+e^{-z_*}}, \qquad q_* := 1 - p_* = \frac{e^{-z_*}}{1+e^{-z_*}}. \tag{4.2.2}$$

Our point is that when both modes are separated by very low-density regions, learning z_* without weight correction leads to an incorrect estimation of the mode probabilities ("no-learning") or to mode collapse depending on the initialization of the learning procedure, whereas using the Jarzynski correction does not. The possible interest of such situation is evident: in real applications multimodality is in fact very common. From now on, we will suppose that the modes are separated in the following sense:

$$|a-b| = 10, \tag{4.2.3}$$

which will be enough for our needs. The more separated they are, the stronger our quantitative bounds will be. This request mimics a common difficult situation for sampling, hence for EBM training.
The parametrization for our model potential U_z is consistent with (4.2.1):

$$U_z(x) = -\log\left(e^{-\frac{1}{2}|x-a|^2} + e^{-z-\frac{1}{2}|x-b|^2}\right) \tag{4.2.4}$$

and the associated partition function and free energy are

$$Z_z = \sqrt{2\pi}\left(1+e^{-z}\right), \qquad F_z = -\log Z_z = -\log(1+e^{-z}) - \frac{1}{2}\log(2\pi). \tag{4.2.5}$$

The normalized probability density associated with $U_z(x)$ is thus $\rho_z(x) = e^{-U_z(x)+F_z}$. Gradient descent on the cross entropy leads to the following continuous-time dynamics:

$$\dot{z}(t) = \mathbb{E}_{z(t)}[\partial_z U_{z(t)}] - \mathbb{E}_*[\partial_z U_{z(t)}]. \tag{4.2.6}$$

For simplicity, we will start this ODE at $z(0) = 0$. It corresponds to a proportion of $\frac{1}{2}$ for both modes.
The gradient descent (4.2.6) is an ideal situation where the expectations $\mathbb{E}_{z(t)}, \mathbb{E}_*$ can be exactly analyzed. In practice however, the two terms of (4.2.6) are estimated; the second term using a finite number of training data $\{x_*^i\}_{i=1}^n$, and the first one using a finite number of walkers $\{X_t^i\}_{i=1}^N$, with associated weights $\{e^{A_t^i}\}_{i=1}^N$. For simplicity we set $N = n$ and the empirical GD dynamics is thus

$$\dot{z}(t) = \frac{\sum_{i=1}^n e^{A_t^i} \partial_z U_{z(t)}(X_t^i)}{\sum_{i=1}^n e^{A_t^i}} - \frac{\sum_{i=1}^n \partial_z U_{z(t)}(x_*^i)}{n} \tag{4.2.7}$$

and the walkers evolve under the Langevin dynamic

$$dX_t^i = -\alpha \nabla U_{z(t)}(X_t^i)dt + \sqrt{2\alpha} dW_t^i \tag{4.2.8}$$

for some fixed $\alpha > 0$. Now the nature of the algorithm varies depending on how the walkers are initialized and the Jarzynski weights are evolved.

1. The standard PCD algorithm sets $X_0^i = x_*^i$, that is, the walkers are initialized at the data points, and the weights are not evolved, that is $A_t^i = 0$ at any time.

2. Alternatively, the walkers could be initialized at samples of the initial model: $X_0^i \sim \rho_{z(0)}$, with the weights not evolved. We refer to this algorithm as the *umweighted* procedure. The Markov chain is still persistent.

3. Algorithm 4 corresponds to initializing the walkers at samples of the initial model, $X_0^i \sim \rho_{z(0)}$, and uses the Jarzynski rule (3.2.1) for the weights updates.

For simplicity we analyze the outcome of these algorithms in the continuous-time set-up, i.e. using (4.2.7). This is an idealization of the actual algorithms, but it makes the analysis more transparent.

Highlights of Perturbative analysis

We now show that in the three cases above, the dynamics (4.2.7) can be seen as a perturbation of a simpler differential system whose qualitative behaviour fits with our numerical simulations. These systems depend on the initialization of the walkers and are thus prone to small stochastic fluctuations. We introduce:

- $\hat{q}_*$ the proportion of training data $\{x_*^i\}$ that are close to mode b (a more precise definition is given in Subsection 4.2.1), and we define $\hat{z}_*$ as satisfying $\hat{q}_* = e^{-\hat{z}_*}/(1 + e^{-\hat{z}_*})$;

- $\hat{q}(0)$ the proportion of walkers at initialization that are close to b, and $\hat{p}(0) = 1 - \hat{q}(0)$.

Practically, $\hat{q}_*$ is a random variable centered at q_* and with fluctuations of order $n^{-1/2}$, and $\hat{q}(0)$ is centered at $e^{-z(0)}/(1 + e^{-z(0)}) = \frac{1}{2}$ with fluctuations of the same order.

In the limit where n is large, they can be neglected, but in more realistic training settings, the use of mini-batches leads to small but non-negligible fluctuations, as will be clear in Equation (4.2.12).
The arguments in Subsection 4.2.1 lead to the following approximations:

- In the model-initialized algorithm without Jarzynski correction ('unweighted'), (4.2.7) is a perturbation of

$$\dot{z}_{\mathrm{unw}}(t) = \hat{q}(0) - \hat{q}_*. \tag{4.2.9}$$

  This system has no stable fixed point since the RHS no longer depends on $z_{\mathrm{unw}}(t)$, leading to a linear drift $z_{\mathrm{unw}}(t) = (\hat{q}(0)-\hat{q}_*)t$ and thus to a divergence of $z_{\mathrm{unw}}(t)$. Consequently, the mass of the second mode, $q(t) = e^{-z_{\mathrm{unw}}(t)}/(1 + e^{-z_{\mathrm{unw}}(t)})$, converges to 0 or 1. However, on longer time scales, this drift leads to a sudden transfer of all walkers in one the modes, then to a complete reversal of the drift of z_{unw} which then diverges in the opposite direction, leading to a succession of alternating mode-collapses; see also Figure 4.4 and Remark 4.2.3 below.

- In the continuous-time version of the standard PCD algorithm, (4.2.7) is a perturbation of the same ODE as (4.2.9). However, in this context, since the initial data $\{X_0^i\}$ and the training data $\{x_*^i\}$ are identical, we have $\hat{q}(0) = \hat{q}_*$, and

$$\dot{z}_{\mathrm{pcd}}(t) = 0. \tag{4.2.10}$$

  The parameters do not evolve and the system is stuck at $z_{\mathrm{pcd}}(0)$ ('no-learning'). Note however that in this version of PCD, the walkers are initialized at the *full* training data. In practice, the number of walkers is smaller than the number of training data so that one often uses a small batch of training data to initialize them; in this case we can still have $\hat{q}(0) \neq \hat{q}_*$, falling back to the first case above.

- The continuous-version of Contrastive-Divergence is equivalent to the well-known Score-Matching technique. In this context, (4.2.7) is a perturbation of

$$\dot{z}_{\mathrm{cd}}(t) = 0, \tag{4.2.11}$$

  leading to the same "no-learning" phenomenon.

- In the model-initialized algorithm using the Jarzynski correction, (4.2.7) is a perturbation of

$$\dot{z}_{\mathrm{jar}}(t) = \frac{\hat{q}(0)e^{-z_{\mathrm{jar}}(t)}}{\hat{p}(0) + \hat{q}(0)e^{-z_{\mathrm{jar}}(t)}} - \frac{e^{-\hat{z}_*}}{1 + e^{-\hat{z}_*}} \tag{4.2.12}$$

  This system has a unique stable point $\tilde{z}_*$ satisfying $e^{-\tilde{z}_*} = \hat{p}(0)e^{-\hat{z}_*}/\hat{q}(0)$, hence

$$\tilde{z}_* = \hat{z}_* + \log\left(\frac{\hat{q}(0)}{\hat{p}(0)}\right). \tag{4.2.13}$$

  Note that the second term at the right hand side is a small correction of order $O(n^{-1/2})$ since $\hat{q}(0)/\hat{p}(0) = 1 + O(n^{-1/2})$.

In continuous time, the Jarzynski correction described in Proposition 3.2.1 exactly realizes (4.2.6). However, even for simple mixtures of Gaussian densities, the expectations in (4.2.1) do not have a simple closed-form that would allow for an exact solution. That being said, when a and b are sufficiently well-separated, the system can be seen as a perturbation of a simpler system whose solution can be analyzed.
First, we note that

$$\partial_z U_z(x) = \frac{e^{-z} e^{-\frac{|x-b|^2}{2}}}{e^{-U_z(x)}}. \tag{4.2.14}$$

The main idea of the approximations to come is that $\partial_z U_z(x)$ is almost zero when x is far from b (and in particular, close to a), and is almost 1 if x is close to b. The next lemma quantifies this; from now on we will adopt the notation

$$I_a = [a-4, a+4] \qquad \text{and} \qquad I_b = [b-4, b+4]. \tag{4.2.15}$$

Lemma 4.2.1. *Under Assumption* (4.2.3)*, for any* $v \in \mathbb{R}$*,*

- *if* $x \in I_a$*, then* $\partial_z U_v(x) \leq e^{-v-10}$*;*
- *if* $x \in I_b$*, then* $|\partial_z U_v(x) - 1| \leq e^{v-10}$*.*

Proof. From (4.2.14), we see that if $x \in I_a$ then $|x-b| > 6$ and consequently $e^{-(x-b)^2/2} \leq e^{-18}$. But the denominator of (4.2.14) is itself greater than $e^{-(x-a)^2/2}$ which is itself greater than $e^{-4^2/2} = e^{-8}$ since $|x-a| < 4$. This gives the first bound and the second is proved similarly. □

We recall that if $\xi \sim \mathcal{N}(0,1)$, then $\mathbb{P}(|\xi| > t) \leq e^{-t^2/2}/t$, hence

$$\mathbb{P}(|\xi| > 4) \leq e^{-4^2/2}/4 \leq 0.0001. \tag{4.2.16}$$

Lemma 4.2.2. *Let* $u, v \in \mathbb{R}$*. Under Assumption* (4.2.3)*, we have*

$$\left| \mathbb{E}_u[\partial_z U_z|_{z=v}] - \frac{e^{-u}}{1+e^{-u}} \right| \leq \varepsilon \tag{4.2.17}$$

where $|\varepsilon| \leq 0.0002 + 2e^{-10} e^{|v|}$*.*

Proof. The integral is exactly given by

$$\frac{1}{1+e^{-u}} \mathbb{E}\left[\frac{e^{-v-(\xi_a-b)^2/2}}{e^{-U_v(\xi_a)}} \right] + \frac{e^{-u}}{1+e^{-u}} \mathbb{E}\left[\frac{e^{-v-(\xi_b-b)^2/2}}{e^{-U_v(\xi_b)}} \right] \tag{4.2.18}$$

where ξ_a, ξ_b denote two Gaussian random variables with respective means a, b. By (4.2.16), ξ_a and ξ_b are respectively contained in $I_a = [a-4, a+4]$ and $I_b = [b-4, b+4]$ with probability greater than 0.999. Let us examine the first term of (4.2.18). The fraction inside the expectation is always smaller than 1, hence by Lemma 4.2.1,

$$\partial_z U_z(\xi_a)|_{z=v} \leq \mathbf{1}_{\xi_a \notin I_a} + \mathbf{1}_{\xi_a \in I_a}\, \partial_z U_z(\xi_a)|_{z=v} \leq \mathbf{1}_{\xi_a \notin I_a} + e^{-v-10}.$$

Consequently, by (4.2.16), the expectation is smaller than $0.0001+e^{-v-10}$. The second expectation in (4.2.18) is equal to

$$1 - \mathbb{E}\left[\frac{e^{-(X_b-a)^2/2}}{e^{-U_v(\xi_b)}}\right]$$

and by the same kind of analysis, the expectation here is smaller than $0.0001 + e^{v-10}$. Gathering the bounds yields the result. □

In particular, the right hand side of (4.2.6) can be approximated by $\frac{e^{-z(t)}}{1+e^{-z(t)}} - \frac{e^{-z_*}}{1+e^{-z_*}}$ up to an error term smaller than $0.0004+2e^{-10}(e^{|z(t)|}+e^{|z_*|})$. If $z(t), z_*$ are contained in a small interval $[-C, C]$ with, say, $C < 5$, this error term is uniformly small in time, and one might see (4.2.6) as a perturbation of the following system:

$$\dot{z}(t) = \frac{e^{-z(t)}}{1+e^{-z(t)}} - \frac{e^{-z_*}}{1+e^{-z_*}}, \tag{4.2.19}$$

a system with only one fixed point at $z(t) = z_*$, the ground-truth solution.

4.2.1 Mode collapse in absence of Jarzynski correction

Now let us analyze in a similar fashion the dynamics without reweighting. Here, (4.2.6) is replaced by

$$\dot{z}(t) = \mathbb{E}_{z(t)}[\partial_z U_{z(t)}(X_t)] - \mathbb{E}_*[\partial_z U_{z(t)}], \tag{4.2.20}$$

where the process X_t solves

$$dX_t = -\alpha \nabla U_{z(t)}(X_t)dt + \sqrt{2\alpha}dW_t, \qquad X_0 \sim \rho_{z(0)}$$

The probability density function $\rho(t, x)$ of X_t satisfies a Fokker-Planck equation

$$\partial_t \rho = \alpha \nabla \cdot \left(\nabla U_{z(t)}(x)\rho + \nabla \rho\right), \quad \rho(t=0) = \rho_{z(0)}$$

which, in full generality, is hard to solve exactly, and thus exact expressions for the first term of (4.2.20) are intractable. However, depending on whether X_0 is close to a or b, the process X_t can be well approximated by an Ornstein-Uhlenbeck process, hence $\rho(t, x)$ can itself be approximated by a Gaussian mixture.

Proposition 4.2.1. *Suppose that* (4.2.3) *holds and that*

$$\exists T, C \in \mathbb{R}_+ \text{ such that for all } t \in [0, \alpha^{-1}T], \qquad z(t) \in [-C, C]. \tag{4.2.21}$$

Then one has $D_{\mathrm{KL}}(\rho(0)|\rho(t)) \leq \delta t$ *where* $\delta = 0.000025 + 100e^{-20}e^{2C}$.

In other words, $\rho(t)$ is approximately constant, up to reasonnable time scales $t = O(1/\delta)$.

Proof. X_0 is drawn from $\rho_{z(0)}$, a Gaussian mixture; the probability of it being sampled from a Gaussian with mean a is $e^{-z(0)}/(1+e^{-z(0)}) = 1/2$. We will work conditionally on this event $\mathcal{E}_a$ and we will note $\rho^a(t)$ the density of X_0 conditional on $\mathcal{E}_a$; thus, $\rho^a(0) = \mathcal{N}(a,1)$. We set $V(x) = \frac{1}{2}|x-a|^2$ so that $\nabla V(x) = (x-a)$ and we consider the following Ornstein-Uhlenbeck process:

$$dY_t = -\alpha \nabla V(Y_t)dt + \sqrt{2\alpha}dW_t, \qquad Y_0 = X_0$$

whose density will be denoted $\tilde{\rho}^a(t)$. We use classical bounds on the divergence between $\rho^a(t)$ and $\tilde{\rho}^a(t)$. For example, the bounds in [Lemma 2.20][1] directly apply and yield

$$D_{\mathrm{KL}}(\tilde{\rho}^a(t)|\rho^a(t)) \leq \frac{1}{4}\int_0^t \mathbb{E}\left[|\nabla U_{z(s)}(Y_s) - \nabla V(Y_s)|^2\right] ds.$$

Since Y_t is nothing but an Ornstein-Uhlenbeck at equilibrium, $Y_t \sim \mathcal{N}(a,1)$ for all $t \geq 0$. The term inside the integral is a Gaussian expectation and will be shown to be small:

$$\mathbb{E}\left[|\nabla U_{z(s)}(Y_s) - \nabla V(Y_s)|^2\right] \leq 0.0001 + 400e^{-2z(t)-20}. \tag{4.2.22}$$

Consequently,

$$D_{\mathrm{KL}}(\tilde{\rho}^a(t)|\rho^a(t)) \leq t\frac{0.0001}{4} + 100e^{-20}\int_0^t e^{-2z(s)}ds.$$

Under (4.2.21), the overall bound remains smaller than t times $0.000025 + 100e^{-20}e^{2C}$ as requested, thus proving that $\tilde{\rho}^a(t) = \mathcal{N}(a,1)$ and $\rho^a(t)$ are close with the same quantitative bound

Similarly, $\rho^b(t)$, the density of X_t conditional on X_0 being sampled from a Gaussian with mean b, is close to $\mathcal{N}(b,1)$ with the same quantitative bounds.

Overall, using the chain rule for KL divergences,

$$D_{\mathrm{KL}}(\tilde{\rho}(t)|\rho(t)) \leq \mathbb{P}(\mathcal{E}_a)D_{\mathrm{KL}}(\tilde{\rho}^a(t)|\rho^a(t)) + \mathbb{P}(\overline{\mathcal{E}_a})D_{\mathrm{KL}}(\tilde{\rho}^b(t)|\rho^b(t)) \leq \delta t.$$

In other words, $\rho(t)$ is close to a mixture of two Gaussians with modes centered at a, b, and the probability of belonging to the first mode is the probability of X_0 belonging to the first mode, that is, $e^{-z(0)}/(1+e^{-z(0)}) = 1/2$. □

Proof of (4.2.22). We have

$$\nabla U_z(x) = \frac{(x-a)e^{-|x-a|^2/2} + (x-b)e^{-|x-b|^2/2-z}}{U_z(x)}. \tag{4.2.23}$$

Using Lemma 4.2.1 and the fact that if $x \in I_a$ then $|x-b| < 16$ and $|x-a| < 4$, we get $|\nabla U_v(x) - (x-a)| \leq 20\varepsilon$ with $\varepsilon \leq e^{-v-10}$. Consequently,

$$\begin{aligned}\mathbb{E}\left[|\nabla U_{z(s)}(Y_s) - \nabla V(Y_s)|^2\right] &\leq \mathbb{P}(Y_t \notin I_a) + (20e^{-v-10})^2 \\ &\leq 0.0001 + 400e^{-2v-20}.\end{aligned}$$

□

As a consequence of the Proposition 4.2.1, the first term of (4.2.20) can be approximated by $\mathbb{E}_{z(0)}[\partial_z U_{z(t)}(X_t)] = \mathbb{E}_{z(0)}[\partial_z U_{z(t)}]$, which in turn can be approximated by $e^{-z(0)}/(1+e^{-z(0)})$ thanks to (4.2.17). Overall, (4.2.20) is therefore a perturbation of the system

$$\dot{z}(t) = \frac{e^{-z(0)}}{1+e^{-z(0)}} - \frac{e^{-z_*}}{1+e^{-z_*}} = \frac{1}{2} - q_* =: \gamma.$$

Since the right hand side no longer depends on $z(t)$, this system leads to a constant drift of $z(t)$, that is $z(t) = \gamma t$, leading to mode collapse since $(1+e^{-z(t)})^{-1} \approx (1+e^{\gamma t})^{-1}$ goes to either 0 or 1.

Contrastive Divergence and Score-Matching. The continuous-time limit of Contrastive Divergence (Algorithm 1) is equivalent to Score-Matching minimization. The objective function becomes the Fisher divergence,

$$\begin{aligned}\mathrm{SM}(z) &= \mathbb{E}_*[|\nabla \log \rho_z(X) - \nabla \log \rho_{z_*}(X)|^2] \\ &= \mathbb{E}_*[|\nabla U_z(X) - \nabla U_{z_*}(X)|^2],\end{aligned}$$

which is in theory intractable due to the presence of the unknown parameter z_*; a well-known computation[4] shows that the gradient $\partial_z \mathrm{SM}(z)$ can be estimated using the training samples without resorting to z_*.
Now, the dynamics (4.2.6) is replaced by

$$\dot{z}(t) = \partial_z \mathbb{E}_*[|\nabla U_{z(t)}(X) - \nabla U_{z_*}(X)|^2] \tag{4.2.24}$$

$$= p_* \partial_z \mathbb{E}[|\nabla U_{z(t)}(\xi_a) - \nabla U_{z_*}(\xi_a)|^2] + q_* \partial_z \mathbb{E}[|\nabla U_{z(t)}(\xi_b) - \nabla U_{z_*}(\xi_b)|^2] \tag{4.2.25}$$

where here again $\xi_x \sim \mathcal{N}(x,1)$. From (4.2.22) and the triangle inequality, we have

$$\begin{aligned}&\left(\mathbb{E}[|\nabla U_{z(t)}(\xi_a) - \nabla U_{z_*}(\xi_a)|^2]\right)^{1/2} \leq \\ &\left(\mathbb{E}[|\nabla U_{z(t)}(\xi_a) - (\xi_a - a)|^2]\right)^{1/2} + \left(\mathbb{E}[|\nabla U_{z_*}(\xi_a) - (\xi_a - a)|^2]\right)^{1/2} \\ &\leq 0.0002 + 800 e^{-2z(t)-20}.\end{aligned}$$

and the same approximation holds for the second part in (4.2.25). Overall, we get that for any reasonable z, $\mathrm{SM}(z) \approx 0$: that is, every z minimizes the score. A similar analysis leads to $\partial_z \mathrm{SM}(z) \approx 0$. Consequently, $\dot{z}(t) \approx 0$: Score Matching and Contrastive Divergence leads to "no-learning".

4.2.2 Empirical gradient descent analysis

The gradient descent (4.2.6) represented an ideal situation where the expectations $\mathbb{E}_{z(t)}, \mathbb{E}_*$ can be exactly analyzed. In practice, the two terms of (4.2.6) are estimated; the second term using a finite number of training data $\{x_*^i\}$, and the first one using

[4] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**: 2005.

a finite number of walkers $\{X_t^i\}$, with associated weights $e^{A_t^i}$ which are either evolved using the Jarzynski rule, or simply set to 1 in the PCD algorithm. Our goal in this section is to explain how these finite-size approximations do not substantially modify the previous analysis and lead to the behaviour presented in 4.2. For simplicity we keep the time continuous.
We recall (4.2.7):

$$\dot{z}(t) = \frac{\sum_{i=1}^N e^{A_t^i} \partial_z U_{z(t)}(X_t^i)}{\sum_{i=1}^N e^{A_t^i}} - \frac{\sum_{i=1}^n \partial_z U_{z(t)}(x_*^i)}{n}. \tag{4.2.26}$$

Jarzynski correction and correct estimation of the empirical weigths. The continuous-time dynamics of the walkers and weigths in our method is given by

$$dX_t^i = -\alpha \nabla U_{z(t)}(X_t^i)dt + \sqrt{2\alpha} dW_t \tag{4.2.27}$$

$$\dot{A}_t^i = -\partial_z U_{z(t)}(X_t^i)\dot{z}(t). \tag{4.2.28}$$

Let $\hat{n}_*^a$ be the number of training data in I_a and $\hat{p}_* = \hat{n}_*^a/n$ their proportion, and similarly $\hat{q}_*$ the proportion in I_b, and $r = 1 - \hat{p}_* - \hat{q}_*$. By elementary concentration results, the remainder $1 - \hat{p}_* - \hat{q}_*$ can be neglected: with high probability, it is smaller than, eg, 0.0001. We will note $\hat{z}_*$ the parameter satisfying $\hat{q}_* = \frac{e^{-\hat{z}_*}}{1+e^{-\hat{z}_*}}$. Using Lemma 4.2.1, the second term in (4.2.26) is approximated by $\hat{q}_*$. Now let us turn to the first term in (4.2.26). Still using Lemma 4.2.1, we see that the first term in (4.2.26) is well approximated by

$$\frac{\sum_{i:\ x_t^i \in I_b} e^{A_t^i}}{\sum_{i=1}^n e^{A_t^i}}. \tag{4.2.29}$$

The second equation in (4.2.27) entails $e^{A_t^i} = \exp\left(-\int_0^t \partial_z U_{z(s)}(X_s^i)\dot{z}(s)ds\right)$. Now let us use Lemma 4.2.1: if $X_s^i \in I_a$, then $\partial_z U_{z(s)}(X_s^i) \approx 0$. Conversely, if $X_s^i \in I_b$, then $\partial_z U_{z(s)}(X_s^i) \approx 1$. Moreover, Proposition 4.2.1 and its proof essentially show that if X_0^i belongs to the first well (close to a), then with high probability so does X_s^a for every s, and in particular $X_s^a \in I_a$ with high probability for every s. Consequently,

$$e^{A_t^i} \approx \exp\left(-\int_0^t 0ds\right) = 1 \qquad \text{if } X_0^i \in I_a, \tag{4.2.30}$$

$$e^{A_t^i} \approx \exp\left(-\int_0^t \dot{z}(s)ds\right) = \exp\left(-z(t) + z(0)\right) = \exp\left(-z(t)\right) \qquad \text{if } X_0^i \in I_b. \tag{4.2.31}$$

As already explained, the dynamics (4.2.27) leaves approximately constant the number of walkers in both modes; consequently, the proportion $\hat{q}(t)$ of walkers X_t^i in I_b remains well approximated by the initial proportion, which is $\hat{q}(0)$, and we obtain that (4.2.29) is well approximated by

$$\frac{\hat{q}(0)e^{-z(t)}}{\hat{p}(0) + e^{-z(t)}\hat{q}(0)} \tag{4.2.32}$$

where we noted $\hat{p}(0) = 1 - \hat{q}(0)$. Note that since $z(0) = 0$, with high probability $\hat{p}(0)$ and $\hat{q}(0)$ are close to $1/2$. The random variable $\hat{p}(0)/\hat{q}(0)$ is thus close to 1.
Overall, we obtain that the system (4.2.26) is a perturbation of the following system:

$$\dot{z}(t) = \frac{\hat{q}(0)e^{-\hat{z}(t)}}{\hat{p}(0) + \hat{q}(0)e^{-\hat{z}(t)}} - \frac{e^{-\hat{z}_*}}{1 + e^{-\hat{z}_*}}. \tag{4.2.33}$$

This system has a unique stable point equal to

$$\tilde{z}_* := \hat{z}_* + \log(\hat{q}(0)/\hat{p}(0)). \tag{4.2.34}$$

Remark 4.2.1. *If the algorithm had been started at $z(0) \neq 0$, one could check that the stable point would become $\hat{z}_* + \log(\hat{q}(0)/\hat{p}(0)) + z(0)$.*

Freezing the weights leads to mode collapse. If the walkers still evolve under the Langevin dynamics in (4.2.27) but the weigths are frozen at $e^{A_t^i} = e^0 = 1$ ('unweighted' algorithm), then (4.2.26) becomes

$$\dot{z}(t) = \frac{\sum_{i=1}^N \partial_z U_{z(t)}(X_t^i)}{N} - \frac{\sum_{i=1}^n \partial_z U_{z(t)}(x_*^i)}{n}. \tag{4.2.35}$$

Keeping the same notation as in the last subsection, the second term is still approximated by $\hat{q}_*$, but this time the first term is instead approximated by

$$\frac{\sum_{i:X_t^i \in I_b} 1}{n} = \hat{q}(t) \approx \hat{q}(0).$$

Consequently, (4.2.35) is a perturbation of the system

$$\dot{z}(t) = \hat{q}(0) - \frac{e^{-\hat{z}_*}}{1 + e^{-\hat{z}_*}} =: \hat{\gamma}, \tag{4.2.36}$$

which no longer depends on t and thus leads to $z(t) = \hat{\gamma}t$ and to mode collapse.

The PCD algorithm leads to no-learning. In the preceding paragraph, at initialization, the walkers X_0^i are distributed according to the initial model $\rho_{z(0)}$. In the PCD algorithm 2, the walkers are instead initialized directly at the training data $\{x_*^i\}_{i=1}^n$. However, in this case, the analysis of the preceding paragraph remains essentially the same, with a single difference: the initial proportion of walkers that are close to a, noted $\hat{q}(0)$, is now exactly $\hat{q}_*$. Thus, (4.2.26) becomes a perturbation of

$$\dot{z}(t) = \hat{q}_* - \hat{q}_* = 0. \tag{4.2.37}$$

The parameters remain constantly equal to its initial value $z(0)$, i.e. there is no learning.

The CD algorithm leads to no-learning. The Continuous-time Contrastive-Divergence algorithm is minimizing the Stein score and is equivalent to Score-Matching as mentioned above: the direction of the gradient of the log-likelihood is that of the gradient of the Stein Score, leading to no-learing. With the estimation given by the training samples, the analysis is exactly the same as above:

$$\begin{aligned}\dot{z}(t) &= \frac{1}{n}\sum_{i=1}^{n} \partial_z |\nabla U_z(x_*^i) - \nabla U_{z_*}(x_*^i)|^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} 2\partial_z \nabla U_z(x_*^i) \times (\nabla U_z(x_*^i) - \nabla U_{z_*}(x_*^i)).\end{aligned}$$

If $x_*^i \in I_a$, then as explained in the proof of Lemma 4.2.22, $\nabla U_s(x_*^i) \approx (x_*^i - a)$ for every s, hence $\partial_z \nabla U_z(x_*^i) \times (\nabla U_z(x_*^i) - \nabla U_{z_*}(x_*^i)) \approx 0$. The same holds for $x_*^i \in I_b$, leading to (4.2.26) being a perturbation of the system

$$\dot{z}(t) = 0. \tag{4.2.38}$$

4.2.3 On mode-collapse oscillations

Most of the approximations performed earlier rely on (4.2.21), that is, the learned parameter $z(t)$ remains in a compact set.

However, in the unweighted algorithm, this is no longer the case at large time scales, since $z(t)$ diverges from (4.2.36); in particular, the approximations from Lemma 4.2.1 become meaningless. In fact, Proposition 4.2.1 is no longer relevant. The core of Proposition 4.2.1 rests upon the fact that if a walker X_t^i is close to a, then its dynamics (4.2.27) is close to an Ornstein-Uhlenbeck process since $\nabla U_{z(t)}(X_t^i) \approx (X_t^i - a)$. This fails when $z(t)$ has a large absolute value. Let us suppose for instance that $z(t)$ is very small (negative), so that $e^{-z(t)}$ is very large, $|z(t)| \gg |x-b|^2$. In (4.2.23), the first term of the numerator is dominated by the second term. Overall, we get

$$\nabla U_{z(t)}(X_t^i) \approx (X_t^i - b),$$

and this is valid for all X_t^i. Consequently, *all* the walkers now undergo an Ornstein-Uhlenbeck process *centered at* b and in particular, the walkers that are close to a are exponentially fast transferred to the region close to b. At this point, the first term in (4.2.35) becomes close to 1, leading to the approximated system $\dot{z}(t) = 1 - \hat{q}_*$: $z(t)$ oscillates back to $+\infty$, until the same phenomenon happens again and all the walkers transfer to the region close to a.

This leads to an oscillating behavior that can be observed on longer time scales (see Figure 4.4 for example). We do not think that this phenomenon is relevant to real-world situations since most learning algorithms are trained for a limited time period.

4.3 Numerics

The code used to perform the numerical experiments is available at a public GitHub repository[5].

4.3.1 Synthetic Data: Gaussian Mixture

In this section, we use a GMM synthetic model to numerically illustrate the advantages of our approach. Specifically, we assume that the data is drawn from the Gaussian mixture density with two modes given by

$$\rho_*(x) = Z_*^{-1}\left(e^{-\frac{1}{2}|x-a_*|^2} + e^{-\frac{1}{2}|x-b_*|^2-z_*}\right), \qquad Z_* = (2\pi)^{d/2}\left(1+e^{-z_*}\right) \tag{4.3.1}$$

where $a_*, b_* \in \mathbb{R}^d$ specify the means of the two modes and $z_* \in \mathbb{R}$ controls their relative weights $p_* = 1/(1+e^{-z^*})$ and $q_* = 1 - p_* = e^{-z_*}/(1+e^{-z^*})$. The values of a_*, b_*, z_* are carefully chosen such that the modes are well separated and the energy barrier between the modes is high enough such that jumps of the walkers between the modes are not observed during the simulation with ULA. Consistent with (4.3.1) we use an EBM with

$$U_\theta(x) = -\log\left(e^{-\frac{1}{2}|x-a|^2} + e^{-\frac{1}{2}|x-b|^2-z}\right), \tag{4.3.2}$$

where $\theta = (a, b, z)$ are the parameters to be optimized. We choose this model as it allows us to calculate the partition function of the model at any value of the parameters, $Z_\theta = (2\pi)^{d/2}(1+e^{-z})$. We use this information as a benchmark to compare the prediction with those produced by our method.
In our numerical experiments, we set $d = 50$, use $N = 10^5$ walkers with a mini-batch of $N' = 10^4$ and $n = 10^5$ data points. We initialize the model at $\theta_0 = (a_0, b_0, z_0)$ with a_0 and b_0 drawn from an $N(0, \epsilon^2 I_d)$ with $\epsilon = 0.1$ and $z_0 = 0$, meaning that the initial ρ_{θ_0} is close to the PDF of an $N(0, I_d)$. The training is performed using Algorithm 4 with $h = 0.1$ and fixed learning rates $\gamma_k = 0.2$ for a_k and b_k and $\gamma_k = 1$ for z_k. We perform the resampling step by monitoring ESS_k defined in (3.4.4) with constant $1/c_k = 1.05$ and using the systematic method. We also compare our results to those obtained using ULA with these same parameters (which is akin to training with the PCD algorithm) and with those obtained with the CD algorithm: in the latter case, we evolve the walkers by ULA with $h = 0.1$ for 4 steps between resets at the data points, and we adjust the learning rates by multiplying them by a factor 10. In all cases, we use the full batches of walkers, weights, and data points to estimate the empirical averages. We also use (3.4.3) to estimate the cross-entropy $H(\rho_{\theta_k}, \rho_*)$ during training by our method (CD and PCD do not provide estimates for these quantities), and in all cases compare the result with the estimate

$$\tilde{H}_k = \log\left((2\pi)^{d/2}\left(1+e^{-z_*}\right)\right) - \frac{1}{n}\sum_{j=1}^{n}\log\left(e^{-\frac{1}{2}|x_*^j-a_k|^2} + e^{-\frac{1}{2}|x_*^j-b_k|^2-z_k}\right) \tag{4.3.3}$$

[5] URL: `https://github.com/Davidedaca/EBMs_Jarzynski`

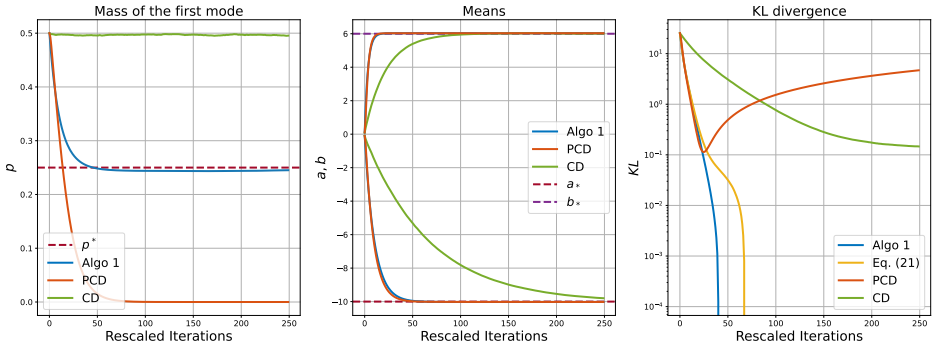


Figure 4.1: *GMM experiments:* Evolution of the parameters and the cross entropy during training by Algorithm 4, PCD, and CD. Average of 20 runs. *Left panels:* evolution of $p_k = 1/(1+e^{-z_k})$; *middle panel:* evolution of a_k and b_k; *right panel:* evolution of the Kullback-Leibler divergence. All three methods capture the location of the modes accurately, but only ours get the relative weights of these modes accurately (whereas PCD leads to mode collapse, and CD to an inaccurate estimate). Our method is also the only one that allows for direct estimation of the cross-entropy during training, and the only one performing GD on this cross-entropy–for better visualization we subtract the entropy of the target $H(\rho_*)$ and plot the Kullback-Leibler divergence instead of the cross-entropy.

The results are shown in Figure 4.1. As can be seen, all three methods learn well the values of a_* and b_* specifying the positions of the modes. However, only our approach learns the value of z specifying their relative weights. In contrast, the PCD algorithm leads to mode collapse, consistent with the theoretical explanation given in Section 4.2, and the CD algorithm returns a biased value of z, consistent with the fact that it effectively uses the Fisher divergence as the objective. The results also show that the cross-entropy decreases with our approach, but bounces back up with the PCD algorithm and stalls with the CD algorithms: this is consistent with the fact that only our approach actually performs the GD on the cross-entropy, which, unlike the other algorithms, our approach estimates accurately during the training.

Additional Details about GMM experiments. Here we give some additional details and numerical results about GMM. The two wells of the teacher are aligned along the first dimension, fixing the first component of the means to be $a_*^1 = -10$ and $b_*^1 = 6$, and $a_*^\alpha = b_*^\alpha = 0$ for any $\alpha = 2, \dots, d$; we also set $z_* = -\log(3)$, corresponding to a mass $p_* = 1/(1+e^{-z_*}) = 0.25$ of the mode centered at a.

All the simulations are performed in $d = 50$, with a time step of $h = 0.1$ for the ULA update. The number of data points is $n = 10^4$. The setup of the teacher is the same for every simulation we display here and the optimization step is performed with full batch gradient descent with learning rate constant in time.

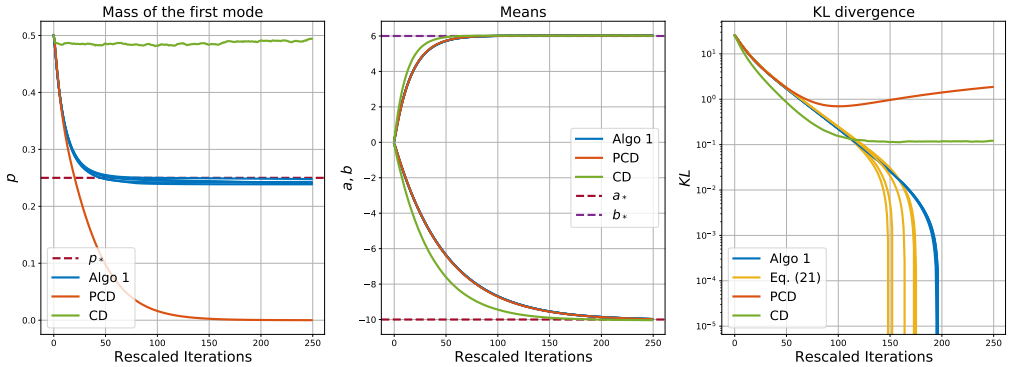


Figure 4.2: *GMM experiments:* Evolution of the parameters and the cross entropy during training by Algorithm 3, PCD, and CD. W.r.t. Algorithm 3, we display the results for five different thresholds in the resampling step. *Left panels:* evolution of $p_k = 1/(1+e^{-z_k})$; *middle panel:* evolution of a_k and b_k; *right panel:* evolution of the Kullback-Leibler divergence.

GMM for different choices of c_k. Results are shown in Figures 4.2 and 4.3. As initial conditions we select $a_0^1 = -10^{-1}$, $b_0^1 = 10^{-1}$, $a_0^\alpha \sim 10^{-2}\mathcal{N}(0,1)$ and $b_0^\alpha \sim 10^{-2}\mathcal{N}(0,1)$ for any $\alpha = 2, \dots, d$; this perturbation around $a = b = 0$ is prescribed to avoid numerical degeneracy. For z, we fix $z_0 = 0$. We run Algorithm 3, as well as the CD and PCD Algorithms 1 and 2 using $N = 10^4$ walkers for $K = 8 \cdot 10^3$ iterations. We use a different learning rate for z and a, b, namely $\gamma_z = 0.125$ and $\gamma_{a,b} = 0.2\gamma_z$. Moreover, these values are multiplied by a factor 10 in CD. With regard to the resampling step in Algorithm 3, we choose a threshold $c_k = c$ which is fixed in time. We display the result for five possible values of this hyperparameter, namely $c = [0.1, 0.2, 0.4, 0.8, 1.0]$; these are related to thresholds in the effective sampling size (ESS) via $\mathrm{ESS}_{\mathrm{thresh}} = N/(c+1)$. With regard to Algorithm 3, we choose $M = 1$ and $N' = N$, that is the full batch version; in the CD Algorithm 2 we choose $P = 4$ for the number of ULA steps between restarts. Every run is performed five times and the average between them is shown in figures.

Mode collapse in GMM for PCD. Results are shown in Figure 4.4. In the same setup as above we select as initial conditions $a_0^1 = -10^{-1}$, $b_0^1 = 10^{-1}$, $a_0^\alpha \sim 10^{-2}\mathcal{N}(0,1)$ and $b_0^\alpha \sim 10^{-2}\mathcal{N}(0,1)$ for any $\alpha \in [2, d]$; for z, we fix $z_0 = 0$. The learning rate is chosen to be $\gamma_z = 5$ for z and $\gamma_{a,b} = 0.2\gamma_z$ for the means. The time step of ULA is $h = 0.2$. Since the objective is to show mode collapse in the PCD algorithm, we run just Algorithm 2 for $K = 10^4$ iterations and $N = 10^4$ walkers.

The need for resampling. In absence of resampling, we observe a dramatic deterioration of ESS (Figure 4.5).

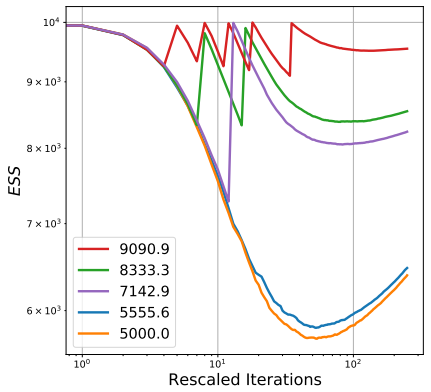


Figure 4.3: *GMM experiments:* Evolution of the effective sample size (ESS) for five different choices of threshold associated to c, constant in time. Bimodal student.

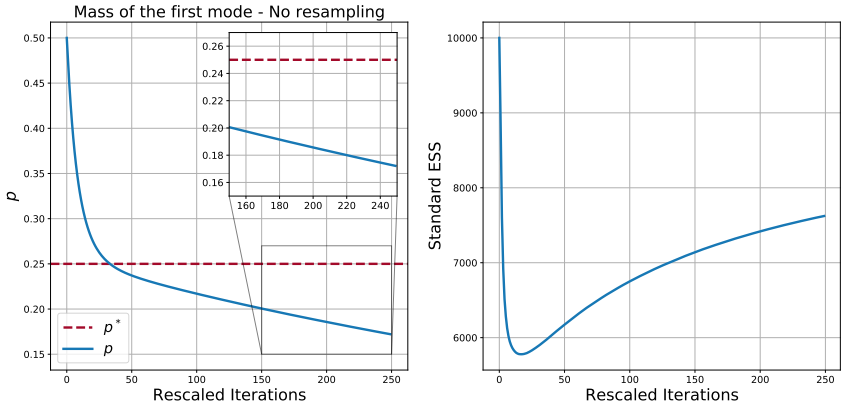


Figure 4.5: GMM in 50d without resampling step, full batch experiment with $N = 10^5$ walkers, average of 30 runs, $4.8 \cdot 10^5$ iterations. *Left Panel:* relative mass of the first mode. *Right panel:* evolution of ESS

To investigate how this issue is solved by resampling, we compare the three routines presented in Section 3.4, using three pre-specified lags between the resampling steps. We use the same hyperparameters (learning rate, target distribution, etc.) as in Figure 4.5. The results are shown in Figure 4.6: looking at the upper panels, we see that stratified and systematic resampling are more stable than multinomial; moreover, if the resampling step is performed too infrequently (green lines), the method converges to a result where the target relative mass is off. Looking at the lower panels, we see that the minimum statistical error is obtained with systematic resampling; moreover, the behaviour of uncertainty appears to be more stable than stratified.
As a side note, systematic resampling requires just a single random number, contrarily

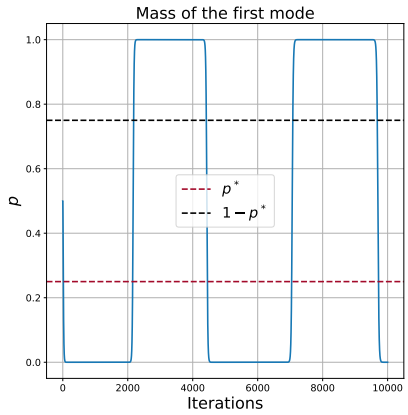


Figure 4.4: Mode collapse and oscillations in PCD. Evolution of the probability $p_k = 1/(1+e^{-z_k})$.

to multinomial and stratified which need N, making the former also more efficient from a computational point of view. This consideration, plus the experimental results we just discussed, strongly motivate the adoption of such method for the resampling step in our proposed algorithm.

Discussion. GMM in high dimension are challenging for the standard CD and PCD algorithms given in 1 and 2. The experimental results of this section also confirm the theoretical analysis prsented in Appendix 4.2.1 below: CD is performing GD but on Fisher divergence rather than cross entropy, and as a result it incorrectly estimates the mass of the modes since Fisher divergence is insensitive to this quantity. On the other hand, PCD causes cycles of mode collapse (see Figure 4.4 and the left panel of Figure 4.2), and the KL divergence does not decreases monotonically, since the protocol is not ensured to be gradient descent (right panel in Figure 4.2).
In this example, our Algorithm 3 outperforms these standard methods as it is an implementation GD on cross-entropy. In particular, the estimation of the relative mass with Algorithm 3 is more accurate than with the CD and PCD algorithms (see Figure 4.2). Moreover, the computation of KL divergence via Jarzynski weights is fairly precise; in Figure 4.2 the estimated KL and the exact one overlaps beyond the minimum values reached in with the CD and PCD algorithms. With regard to Figure 4.3, the choice of the threshold for resampling does not appear to be decisive in this regime of hyperparameters; in fact, looking at the evolution of θ and of KL divergence, the overall behavior of Algorithm 3 is not dramatically influenced by the choice of c.

4.3.2 Real data: MNIST

Next, we perform empirical experiments on the MNIST dataset to answer the following question: when it comes to high-dimensional datasets with multiple modes, can

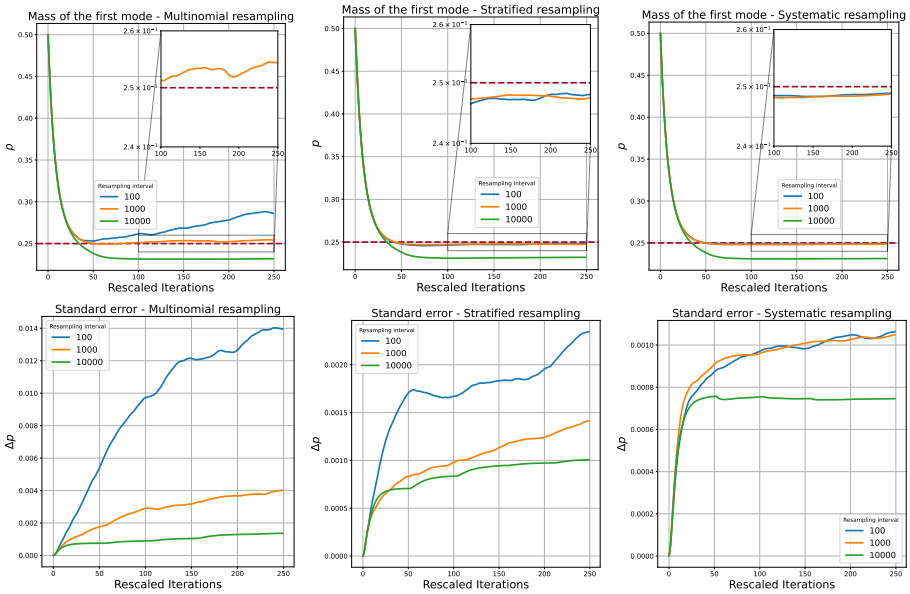


Figure 4.6: GMM in 50d with three resampling routines, full batch experiment with $N = 10^5$ walkers, 50 runs per each resampling interval, $4.8 \cdot 10^5$ iterations. *Upper panels:* relative mass of the first mode for each resampling method. *Lower panels:* standard error computed using the empirical standard deviation via $\sigma_{emp}/\sqrt{50}$.

our method produces an EBM that generates high-quality samples and captures the relative weights of the modes accurately? To this end, we select a subset of MNIST consisting of only three digits: 2, 3, and 6. Then, we choose 5600 images of label 2, 2800 images of label 3, and 1400 images of label 6 from the training set (for a total of $n = 9800$ data points), so that in this manufactured dataset the digits are in have respective weights 4/7, 2/7, and 1/7. We train two EBMs to represent this data set, the first using our Algorithm 4, and the second using the PCD Algorithm 2. We represent the energy using a simple six-layer convolutional neural network with the swish activation and about $77K$ parameters. We use the ADAM optimizer for the training with a learning rate starting from 10^{-4} and linearly decaying to 10^{-10} until the final training step. The sample size of the walkers is set to $N = 1024$ and it is fixed throughout the training.

The results are shown in Figure 4.7. Both our Algorithm 3 and the PCD Algorithm 2 generate images of reasonable quality; Jarzynski weights of the generated samples are directly related to the image quality: this is shown in the left panel of Figure 4.10, where we display images along with their Jarzynski weights. Moreover, the right panel of Figure 4.10 indicates that resampling at the end of training using these weights helps improve sample quality. Examples of generated images are shown in Figure 4.11.

However, the real difference between the methods comes when we look at the relative proportion of the digits generated in comparison to those in the data set. In Figure 4.8,

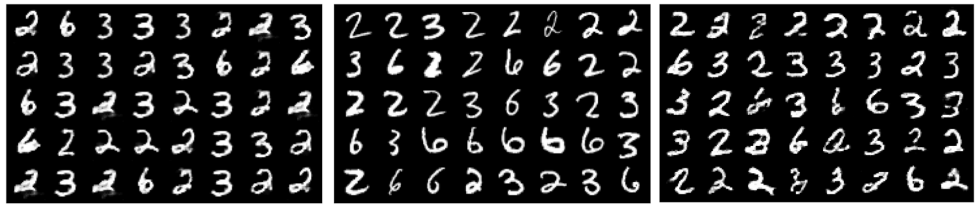


Figure 4.7: *MNIST: Left panel:* Examples of images generated by our method right after resampling in the last epoch. *Middle panel:* Images randomly selected from the test dataset of MNIST. *Right panel:* Examples of images generated by training using the persistent contrastive divergence (PCD) algorithm.

we show the relative error on the weight estimation of each mode obtained by using a classifier pre-trained on the data set. Although the EBM trained with the PCD algorithm can discover all the modes and generate images of all three digits present in the training set, it cannot accurately recover the relative proportions of each digit. In contrast, our method is successful in both mode exploration and weight recovery.

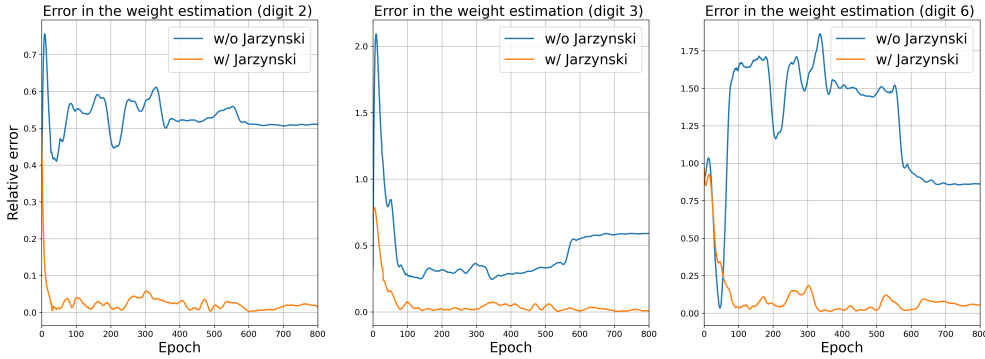


Figure 4.8: *MNIST:* Relative error of the weight estimation of the three modes (i.e. three digits). Our method outperforms the PCD algorithm in terms of recovery of the weight of each mode.

Unlike in the numerical experiments done on Gaussian mixtures with teacher-student models, for the MNIST dataset, we cannot estimate the KL divergence throughout the training as we do not know the normalization constant of the true data distribution. Nevertheless, we can still use the formula (4.3.3) to track the cross-entropy between the data distribution and the walker distribution a plot of which is given in the right panel Figure 4.9 for the mini-batched training algorithm (Algorithm 4).
In these experiments with MNIST, we used the adaptive resampling scheme described after equation 3.4.4. In practice, we observed that few resamplings are needed during training, and they can often by avoided altogether if the learning rate is sufficiently small. For example, in the experiment reported in the left panel of Figure 4.9) a single

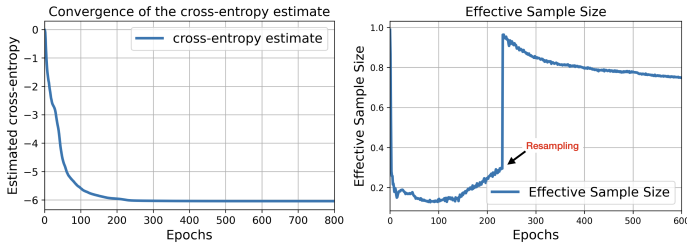


Figure 4.9: *MNIST dataset: Left Panel:* Convergence of the cross-entropy estimated by using formulae (3.4.3) with the mini-batched algorithm (Algorithm 4). *Right Panel:* Evolution of the effective sample size (ESS) defined in (3.4.4) – here resampling was started after 240 epochs (with $c_k = 0$ before and $c_k = 0.5$ afterwards), and occurred only once immediately after being switched on.

resampling step was made. We also found empirically that the results are insensitive to the choice of of the parameters c_k used for resampling: the rule of thumb we found is to not resample at the beginning of training, to avoid possible mode collapse, and use resampling towards the end, to improve the image quality.

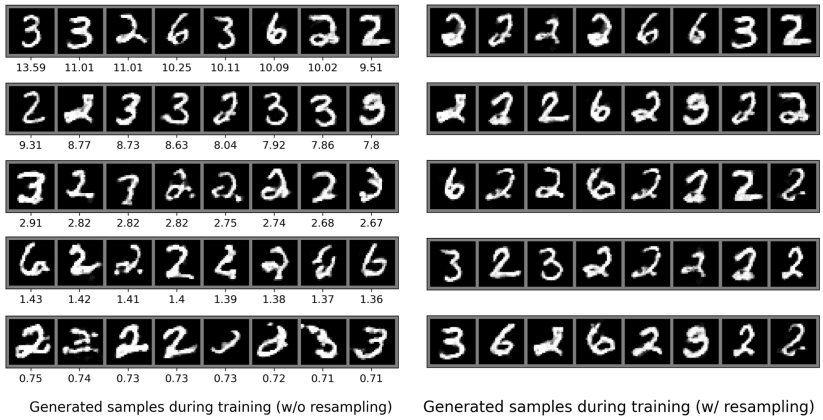


Figure 4.10: *MNIST dataset: Left panel:* Images randomly chosen during the training from the entire set of generated samples with their associate Jarzynski weights. From left to right, top to bottom, the higher the Jarzynski weight is, the better the image quality is. *Right panel:* Images obtained under the same training conditions, after resampling and continued training for 120 epochs.

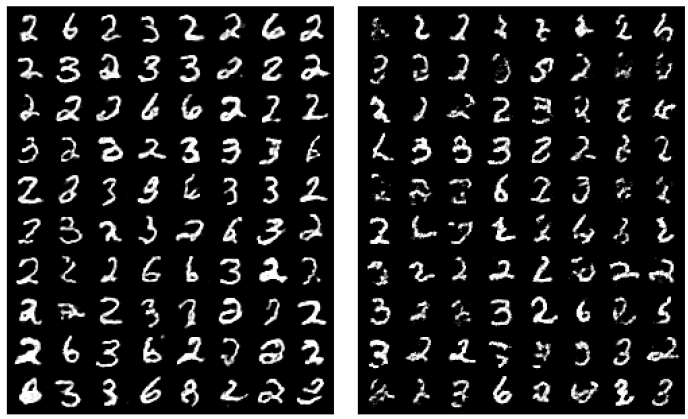


Figure 4.11: *MNIST dataset: Left panel*: images generated by the mini-batched version of the Algorithm 4. *Right panel*: Images generated from the PCD with mini-batches.

4.3.3 Real data: CIFAR-10

We perform an empirical evaluation of our method on the full CIFAR-10 (32×32) image dataset. We use the a setup present in literature[6], with the same neural architecture with $n_f = 128$ features, and compare the results obtained with our approach with mini-batching (Algorithm 4) to those obtained in a recent work[7] with PCD and PCD with data augmentation (which consists of a combination of color distortion, horizontal flip, rescaling, and Gaussian blur augmentations, to help the mixing of the MCMC sampling and stabilize training). The hyperparameters are the same in all

Method	*FID*	*Inception Score (IS)*
PCD with mini-batches	38.25	5.96
PCD with mini-batches and data augmentation	36.43	6.54
Algorithm 4 with multinomial resampling	**32.18**	**6.88**
Algorithm 4 with systematic resampling	**30.24**	**6.97**

Table 4.1: *CIFAR-10 dataset:* Comparison of FID and Inception Score (IS) for PCD and Algorithm 4. Experiments performed using the neural architecture in[6] to model the energy.

cases: we take $N = 4096$ Langevin walkers with a mini-batch size $N' = 256$. We use the Adam optimizer with learning rate $\eta = 10^{-4}$ and inject a Gaussian noise of standard deviation $\sigma = 3 \times 10^{-2}$ to the dataset while performing gradient clipping

[6] Erik Nijkamp et al. "Learning non-convergent non-persistent short-run MCMC toward energy-based model" in: *Advances in Neural Information Processing Systems*. vol. 32 2019.

[7] Yilun Du et al. "Improved Contrastive Divergence Training of Energy-Based Models" in: *International Conference on Machine Learning*. PMLR 2021. 2837–2848

Figure 4.12: *CIFAR-10 dataset: Left panel*: images generated by training with Algorithm 4. *Right panel*: Images generated from the PCD with mini-batches.

in Langevin sampling for better performance. All the experiments were performed on a single A100 GPU. Training for 600 epochs took about 34 hours with the PCD algorithm (w/ and w/o data augmentation) and about 36 hours with our method. Some of the images generated by our method are shown in Figure 4.12. We also quantitatively evaluate the performance of our models with the commonly used metrics (e.g. FID and Inception Score): the results are given in Table 4.1. They indicate that our method can achieve slightly better performance than the PCD algorithms w/ and w/o data augmentation at a similar computational cost. Furthermore, these results on CIFAR-10 suggest that our method scales well to complicated training tasks on more realistic data sets. A key component of the success of our method is the resampling step. In Figure 4.13 we plot the Effective Sample Size (ESS): each peak in Figure 4.13 indicates a step of resampling of all the samples. We note that training EBMs with mini-batches has the side effect of a rapid loss of the ESS, which measures the sample quality. So, doing resampling with a proper criterion and a reliable resampler is necessary. In conclusion, we would like to emphasize that the objective of the experiments conducted was to assess the potential scalability of our method, rather than

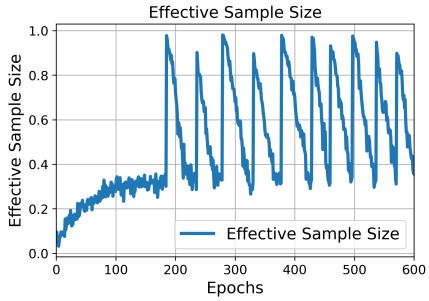


Figure 4.13: *CIFAR-10 dataset:* The Effective Sample Size (ESS) during the training with Algorithm 4. Each peak in the plot implies one step of resampling. Notice that the number of resampling in CIFAR-10 experiments is significantly larger than the one in MNIST experiments (4.9), which suggests the necessity of resampling in the scalability of our method.

striving to achieve state-of-the-art benchmarks. It's worth noting that when dealing with complex data, particularly images, various techniques are commonly employed, including sophisticated sampling and post-processing methods. Therefore, the results of the CIFAR-10 experiments should be viewed conceptually: they demonstrate the feasibility of integrating our algorithm into the realm of deep learning.

Part II

Additional work: Time Reversal Symmetry for Classical, Nonrelativistic Quantum and Spin Systems in Presence of Magnetic Fields

The work described in this Part has also been previously published in:

- D. Carbone and L. Rondoni. *Necessary and Sufficient Conditions for Time Reversal Symmetry in Presence of Magnetic Fields*. In: Symmetry 12.8 (2020), p. 1336.
- D. Carbone, P. De Gregorio and L. Rondoni. *Time reversal symmetry for classical, non-relativistic quantum and spin systems in presence of magnetic fields*. Annals of Physics 2022, 441, 168853.

The bibliography style of the present Part is kept as in the original papers, i.e. not in footnote format.

Chapter 5

Time reversal symmetry for classical, non-relativistic quantum and spin systems in presence of magnetic fields

5.1 Introduction

The relation between time reversal invariance (TRI) and Onsager reciprocal relations [1, 2], for systems coupled with a magnetic field is a topic well investigated since Casimir's article [3]. A cardinal contribution was given by Kubo in Refs. [4–6] who used the usual time reversal operation

$$\mathcal{T}_B(\boldsymbol{r}, \boldsymbol{p}, t; \boldsymbol{B}) = (\boldsymbol{r}, -\boldsymbol{p}, -t; -\boldsymbol{B}) \tag{5.1.1}$$

for the correlator of two classical observables ϕ and ψ in the stationary state, where $\boldsymbol{r}, \boldsymbol{p}$ collectively represent coordinates and momenta of the particles of the system of interest. He obtained the following chain of equalities:

$$\langle \phi(0)\psi(t)\rangle_{\boldsymbol{B}} = \eta_\phi \eta_\psi \langle \phi(0)\psi(-t)\rangle_{-\boldsymbol{B}} = \eta_\phi \eta_\psi \langle \phi(t)\psi(0)\rangle_{-\boldsymbol{B}} \tag{5.1.2}$$

Here the factors η_ψ and η_ϕ are, respectively, the signatures of the *observables* ψ and ϕ, i.e., of two generic functions defined on the phase space, with regard to the transformation $\mathcal{T}_B$. Moreover, the angular brackets represent the average with respect to the equilibrium probability distribution in phase space.
Generalized time reversal transformations different from $\mathcal{T}_B$ are already given by Lax in Ref. [7], but in the previous century the statement that crystallized in the literature was that only $\mathcal{T}_B$ allows the reciprocal relations to hold. Unfortunately, this only leads to a relation between two different systems as stressed by the subscripts in (6.1.2), one

with magnetic field $\boldsymbol{B}$ and the other with opposite field, which leads to Casimir's modification of Onsager reciprocal relations. As a consequence, the predictive power of these relations is quite limited, compared to that of the original relations.

Recently, however, a different perspective has been adopted in Refs. [8–10] for classical systems coupled with a constant magnetic field along an axis and in Ref. [11] for a magnetic field dependent on one space coordinate. In particular, it was shown that suitable time reversal operations exist that yield (6.1.2) without the inversion of the field. Furthermore, the quantum case, in the presence of a constant magnetic field has been similarly treated in Ref. [12].

As we will show in detail, the generalized time reversal transformations that were investigated do not exhaust the set of all possible operations leading to TRI. The first objective of this paper is to identify the most general time reversal operation compatible with a classical Hamiltonian system. After this, we analyze the minimal coupling with a generic magnetic field, formulating sufficient conditions for the magnetic field and for the force potential that make the Onsager reciprocal relations hold.

This theoretical result is relevant also in the context of quantum mechanics, that will be dealt with in a future paper. For exemplary instance, in Ref. [13] Büttiker and collaborators analyzed quantum systems using the "tenfold way" developed by Zirnbauer in Ref. [14], which is founded on the idea that the validity of the Onsager reciprocal relations necessarily requires microreversibility, i.e., Onsager's notion that: "if the velocities of all the particles present are reversed simultaneously the particles will retrace their former paths, reversing the entire succession of configurations", which is to say that $\mathcal{T}(\boldsymbol{r}, \boldsymbol{p}, t) = (\boldsymbol{r}, -\boldsymbol{p}, -t)$ holds. As demonstrated in Refs. [8–10], this is not always required for statistical properties, because other symmetries may as well do. In this paper we show that further generalized time reversal operations exist that can be used in Linear Response Theory and beyond.

In Section 5.2, we derive and discuss our results about time reversal invariant (TRI) systems, in presence of magnetic fields, and we introduce our methods of investigation. In particular, we provide sufficient conditions for the magnetic fields that allow TRI. In Section 5.3, we summarize such results and outline future developments.

5.2 Theory and Results

This section is organized as follows: Section 5.2.1 summarizes previous results on TRI in presence of a magnetic field and its relevance for the Onsager reciprocal relations and other statistical equalities. Section 5.2.2 identifies the general form of a TRI operation for a system coupled with a magnetic field $\mathbf{B}$, and gives sufficient conditions on $\mathbf{B}$ for such operations to exist. This is connected with the question of gauge freedom, which is analyzed in Section 5.2.3. Section 5.2.4 closes the loop concerning sufficient conditions, expressing them directly from the point of view of the magnetic field. Finally, various examples of potentials are used to illustrate our theoretical results.

5.2.1 Onsager Reciprocal Relations and T-Symmetry

A dynamical system $S^t : \Omega \to \Omega$, on a phase space Ω with $t \in \mathbb{R}$, is called TRI if there exists a map $\mathcal{M} : \Omega \to \Omega$, such that:

$$\mathcal{M}S^t = S^{-t}\mathcal{M}\,, \quad \text{and} \quad \mathcal{M}^2 = I \tag{5.2.1}$$

The operator S^t is the time evolution operator on the phase space, which moves every initial condition $\Gamma \in \Omega$ to the corresponding evolved phase point $S^t\Gamma \in \Omega$. As S^t and S^{-t} are operators related to the same dynamics, forward in one case and backward in the other, $\mathcal{M}$ in (5.2.1) has to preserve the equations of motion and so the Hamiltonian, cf. Section 5.2.2.
As shown for instance in Ref. [8], the canonical time reversal operation, i.e., $\mathcal{M}(\boldsymbol{r}, \boldsymbol{p}) = (\boldsymbol{r}, -\boldsymbol{p})$, does not verify Equation (5.2.1) when S^t describes the evolution of a system in a magnetic field. While the equations of motion are preserved by $\mathcal{T}_B$, i.e., by inverting momenta and magnetic field together with time, that operation means dealing with different systems, subject to different magnetic fields, rather than with a single system in given magnetic field. Thus, one only obtains relations such as the Onsager–Casimir ones, (6.1.2), that do not quantify the properties of a system of interest: they merely link non-quantified properties of two different systems in two different magnetic fields. Given the observables $\phi, \psi : \Omega \to \mathbb{R}$, their correlator with respect to a probability distribution in phase space, ρ, is defined by:

$$\langle \phi(0)\psi(t)\rangle_{\boldsymbol{B}} = \int_\Omega dX \rho(X)\phi(X)\psi(S^t X) \tag{5.2.2}$$

In case an operation $\mathcal{M}$ verifying Equation (5.2.1) exists, Onsager reciprocal relations hold, as can be demonstrated analyzing the correlator (5.2.2). This can be seen through the following steps: first, $\mathcal{M}$ is used to change variable within the integral, setting $X = \mathcal{M}Y$, whose Jacobian determinant is 1, because $\mathcal{M}$ is an isometry. It follows that:

$$\langle \phi(0)\psi(t)\rangle_{\boldsymbol{B}} = \int_\Omega dY \rho(\mathcal{M}Y)\phi(\mathcal{M}Y)\psi(S^t\mathcal{M}Y) \tag{5.2.3}$$

Suppose that ϕ and ψ respectively possess signatures η_ϕ and η_ψ under the action of $\mathcal{M}$, and that the probability density ρ is even under $\mathcal{M}$, as appropriate for an equilibrium distribution of a Hamiltonian particles system, such as the canonical ensemble. This leads to the result showed in Ref. [8]:

$$\langle \phi(0)\psi(t)\rangle_{\boldsymbol{B}} = \eta_\phi\eta_\psi \int_\Omega dY \rho(Y)\phi(Y)\psi(S^{-t}Y) = \eta_\phi\eta_\psi\langle \phi(0)\psi(-t)\rangle_{\boldsymbol{B}} \tag{5.2.4}$$

Using the invariance for time translation of the equilibrium state, i.e., translating forward by a time t the last term of (5.2.4), we come to the final result:

$$\langle \phi(0)\psi(t)\rangle_{\boldsymbol{B}} = \eta_\phi\eta_\psi\langle \phi(t)\psi(0)\rangle_{\boldsymbol{B}} \tag{5.2.5}$$

This is related to the Onsager theory of linear response as follows: given the macroscopic observables α_i, $i = 1, ..., n$, and entropy $\mathcal{S}$ of a system subjected to (relatively)

small thermodynamic forces X_j, $j = 1, ..., n$, one may write:

$$\dot{\alpha_i} = \sum_j L_{ij} X_j \qquad X_j = \frac{\partial \mathcal{S}}{\partial \alpha_j}\,; \quad i, j = 1, ..., n \tag{5.2.6}$$

where the linear transport coefficients are obtained via the Green–Kubo integrals of the corresponding correlators (see Ref. [15]). Therefore, the symmetry properties of L_{ij} descend from those of $\langle \alpha_i(0)\alpha_j(t)\rangle$. If η_i and η_j are the signatures of the macroscopic observables, we have:

$$\langle \alpha_i(0)\alpha_j(t)\rangle_{\boldsymbol{B}} = \eta_i\eta_j \langle \alpha_i(t)\alpha_j(0)\rangle_{\boldsymbol{B}}\,; \quad i, j = 1, ..., n \tag{5.2.7}$$

that, after integration in time, yield the Onsager reciprocal relations:

$$L_{ij} = \eta_i\eta_j L_{ji}\,; \quad i, j = 1, ..., n \tag{5.2.8}$$

Our goal is to identify the general form of a time reversal transformation, as well as the conditions under which Onsager symmetry may be obtained in presence of a magnetic field.

5.2.2 Dynamics and Transformations

Consider a system of particles coupled with an external static magnetic field and subject to forces expressed by a potential. The corresponding Hamiltonian writes:

$$H = \sum_{i=1}^{N} \left[\frac{(\boldsymbol{p}_i - q_i \boldsymbol{A}(x_i, y_i, z_i))^2}{2m_i} \right] + U(\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{C}) \tag{5.2.9}$$

where N is the number of particles, q_i and m_i are the charge and the mass of the i-th particle, the first addend is the coupling to the magnetic field and $U(\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{C})$ is the force potential. In general, U depends on $2dN$ coordinates $(\boldsymbol{X}, \boldsymbol{P})$, if each particle has got d degrees of freedom, but it may also depend on a set of parameters $\boldsymbol{C}$. Without loss of generality, let us assume that the particles move in 3-dimensional space and that $d = 3$. In the following we are going to use $A_k(x_i, y_i, z_i)$, with $k = 1, 2, 3$, to denote the components of the vector potential $\boldsymbol{A}(x_i, y_i, z_i)$.
Let us begin identifying the possible time reversal operations for a Hamiltonian system, in general. Later, we will focus on those that are not broken by magnetic field.

Proposition 5.2.1. *Take the 6-dimensional space of a single particle, with coordinates and momenta (x, y, z, p_x, p_y, p_z), and let I be the identity operator on this space. The general form of a time reversal operator $\mathcal{T}$, for classical Hamiltonian dynamics, writes:*

$$\mathcal{T}(x, y, z, p_x, p_y, p_z) = P(s_1 x, s_2 y, s_3 z, -s_1 p_x, -s_2 p_y, -s_3 p_z) \tag{5.2.10}$$

where P is a permutation of coordinates and of their conjugate momenta, such that $P^2 = I$, and s_i, which equals 1 or -1, takes opposite values in front of coordinates and momenta.

Proof. That P^2 be the identity and that s_i be ± 1 is imposed by the fact that $\mathcal{T}^2 = I$, i.e., that a time reversal transformation must be involutional. That a coordinate and its respective momentum have opposite sign is imposed by the form of the Hamilton equations:

$$\begin{cases} \dfrac{\partial H}{\partial p_i} = \dot{x}^i \\ \dfrac{\partial H}{\partial x^i} = -\dot{p}_i \end{cases} \tag{5.2.11}$$

In fact, assuming that the Hamiltonian itself verifies TRI, an overall minus sign arises when time is reversed. Then, in order to preserve the form of the equations of motion, a minus sign has to distinguish x^i from its conjugate momentum p_i. □

Note that P in Equation (5.2.10) is not a permutation of six elements but it acts in a block diagonal way on the coordinates and in the same way on the momenta. For instance, assuming P swaps x and y, it does the same with the corresponding momenta:

$$(x, y, z, p_x, p_y, p_z) \xrightarrow{P} (y, x, z, p_y, p_x, p_z) \tag{5.2.12}$$

This action comes in addition to the compulsory alternation of signs between coordinates and conjugated momenta produced by the s_i factors.
In order to enumerate how many different time reversal transformations exist, let us represent them in matrix form. As positions and momenta are bound to be distinguished by a minus sign, it suffices to consider the 3-dimensional space of positions, hence to consider a 3×3 matrix, $\mathcal{M}_d$. The action of $\mathcal{T}$ on the corresponding momenta will be given by $-\mathcal{M}_d$.
First, suppose P is the identity, so that $\mathcal{M}_d$ takes the diagonal form:

$$\mathcal{M}_d = \begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{pmatrix} \tag{5.2.13}$$

In this case, there are eight possible choices for $\mathcal{T}$, as shown in Ref. [9]. For example, the usual time reversal operation that preserves the coordinates and reverses the momenta corresponds to $s_1 = s_2 = s_3 = 1$.
If, on the other hand, $P \neq I$, the total number of permutations of three elements is the order of the discrete group S_3, i.e., $3! = 6$. But the cyclical and the counter-cyclical permutations are not involutions, and only the swap permutations remain:

$$\mathcal{M}_{xy} = \begin{pmatrix} 0 & s_P & 0 \\ s_P & 0 & 0 \\ 0 & 0 & s_3 \end{pmatrix} \tag{5.2.14}$$

$$\mathcal{M}_{yz} = \begin{pmatrix} s_1 & 0 & 0 \\ 0 & 0 & s_P \\ 0 & s_P & 0 \end{pmatrix} \tag{5.2.15}$$

$$\mathcal{M}_{xz} = \begin{pmatrix} 0 & 0 & s_P \\ 0 & s_2 & 0 \\ s_P & 0 & 0 \end{pmatrix} \tag{5.2.16}$$

where $s_P = \pm 1$ and the subscript on $\mathcal{M}$ identifies the swap. The non-zero elements in the 2×2 permutation blocks must own the same sign to ensure that the transformation squared is the identity. This amounts to 12 transformations: four for each of the matrices (5.2.14), (5.2.15) and (5.2.16). Adding these to the previous 8 transformations, we obtain a total of 20 generalized time reversal transformations, that can be used to derive the Onsager reciprocal relations, following e.g., the approach of Ref. [9].

For the invariance of the Hamiltonian, let us directly consider the magnetic field, $\mathbf{B} \neq 0$. First, let the particles of the system be coupled to $\mathbf{B}$ only, so that $U(\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{C}) = 0$. As there are 20 possible transformations for each particle subspace, one can choose a time reversal operation among 20^N. For instance, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two matrices that represent two suitable transformations on 6-dimensional subspaces; one may combine them in a single transformation O acting on the entire phase space as:

$$O(\boldsymbol{X}, \boldsymbol{P}) = (\mathcal{M}_1 \boldsymbol{x}_1, -\mathcal{M}_1 \boldsymbol{p}_1, ..., \mathcal{M}_2 \boldsymbol{x}_k, -\mathcal{M}_2 \boldsymbol{p}_k, ..., \mathcal{M}_2 \boldsymbol{x}_N, -\mathcal{M}_2 \boldsymbol{p}_N) \tag{5.2.17}$$

where a special combination of the two operations has been chosen. By definition, O automatically satisfies the conditions (5.2.1), and can be used under the Kubo correlation integral.

To find involutions that act on the entire phase space, not as block diagonal single particle matrices, one may consider non-diagonal time reversal operations, that act on the Hamiltonian (5.2.9) exchanging coordinates and momenta of different particles. However, because in general particles have different masses, $m_i \neq m_j$ for $i \neq j$, such operations do not qualify as time reversal involutions. For example, consider the following transformation:

$$(\boldsymbol{x}_1, ..., \boldsymbol{x}_j, \boldsymbol{x}_{j+1}, ..., \boldsymbol{x}_N, \boldsymbol{p}_1, ..., \boldsymbol{p}_j, \boldsymbol{p}_{j+1}, ..., \boldsymbol{p}_N) \xrightarrow{\mathcal{M}_{nd}} (\boldsymbol{x}_1, ..., \boldsymbol{x}_{j+1}, \boldsymbol{x}_j, ..., \boldsymbol{x}_N, -\boldsymbol{p}_1, ..., -\boldsymbol{p}_{j+1}, -\boldsymbol{p}_j, ..., -\boldsymbol{p}_N) \tag{5.2.18}$$

where $\boldsymbol{x}_1 = (x_1, y_1, z_1)$. Writing the summation in Equation (5.2.9) as:

$$\sum_{i=1}^{N} \left[\frac{(\boldsymbol{p}_i - q_i \boldsymbol{A}(\boldsymbol{x}_i))^2}{2m_i} \right] = ... + \frac{(\boldsymbol{p}_j - q_j \boldsymbol{A}(x_j, y_j, z_j))^2}{2m_j} + \frac{(\boldsymbol{p}_{j+1} - q_{j+1} \boldsymbol{A}(x_{j+1}, y_{j+1}, z_{j+1}))^2}{2m_{j+1}} + ... \tag{5.2.19}$$

the transformation (5.2.18) yields:

$$... + \frac{(\boldsymbol{p}_j + q_j \boldsymbol{A}(x_j, y_j, z_j))^2}{2m_{j+1}} + \frac{(\boldsymbol{p}_{j+1} + q_{j+1} \boldsymbol{A}(x_{j+1}, y_{j+1}, z_{j+1}))^2}{2m_j} + ... \tag{5.2.20}$$

As the transformation (5.2.18) does not act on the masses, Equation (5.2.20) may differ from the corresponding term in Equation (5.2.19), even in cases in which $\boldsymbol{A}(x_j, y_j, z_j) = \boldsymbol{A}(x_{j+1}, y_{j+1}, z_{j+1})$: the Hamiltonian is not invariant under the action of $\mathcal{M}_{nd}$. Depending on the values of the particles masses, certain swaps may be allowed or not. In the following, we limit our investigation to the case that excludes particles swaps.

Considering the 20 operations listed above, (5.2.13), (5.2.14), (5.2.15) and (5.2.16), let us now relate them to the functional form of the vector potential of Equation (5.2.9). Neglecting for sake of simplicity the particle index i, we have:

$$(\boldsymbol{p} - q\boldsymbol{A})^2 = (p_x - qA_1)^2 + (p_y - qA_2)^2 + (p_z - qA_3)^2 \tag{5.2.21}$$

Under the action of the map (5.2.10) with $P = I$, this yields:

$$(-s_1 p_x - qA_1(s_1 x, s_2 y, s_3 z))^2 + (-s_2 p_y - qA_2(s_1 x, s_2 y, s_3 z))^2 + (-s_3 p_z - qA_3(s_1 x, s_2 y, s_3 z))^2 \tag{5.2.22}$$

and imposing that the result equals the expression (5.2.21),

$$(p_x - qA_1)^2 + (p_y - qA_2)^2 + (p_z - qA_3)^2 = (p_x + qs_1 A_1^T)^2 + (p_y + qs_2 A_2^T)^2 + (p_z + qs_3 A_3^T)^2 \tag{5.2.23}$$

where the A_k^T is the transformed component $A_k(s_1 x, s_2 y, s_3 z)$, the Hamiltonian verifies TRI. We can thus write:

Proposition 5.2.2. *The necessary and sufficient algebraic conditions for the validity of Equation* (5.2.23) *are given by:*

$$A_k^T = -s_k A_k \qquad k = 1, 2, 3 \tag{5.2.24}$$

Proof. On the one hand, if (5.2.24) holds, substitution immediately yields (5.2.23). Vice versa, starting from the validity of (5.2.23), one notes that the squares of $\boldsymbol{p}$ and $\boldsymbol{A}$ are squared norms of vectors in $\mathbb{R}^3$, hence are invariant under rotations, as the generalized time reversal operations are. Consequently, the following equality holds:

$$-p_x A_1 - p_y A_2 - p_z A_3 = p_x s_1 A_1^T + p_y s_2 A_2^T + p_z s_3 A_3^T \tag{5.2.25}$$

As each A_k only depends on (x, y, z), and the conjugate momenta are independent, one may vary at will the values of (p_x, p_y, p_z) in (5.2.25). Setting to zero two of them, one gets (5.2.24) for the third. Repeating, for the other pairs, (5.2.24) is obtained. □

Actually, TRI in presence of a magnetic field is less demanding than that, because it suffices that (5.2.23) holds up to a gauge transformation. In other words, (5.2.21) can be generally replaced by:

$$[\boldsymbol{p} - q(\boldsymbol{A} + \nabla G)]^2 = [p_x - q(A_1 + \partial_x G)]^2 + [p_y - q(A_2 + \partial_y G)]^2 + [p_z - q(A_3 + \partial_z G)]^2 \tag{5.2.26}$$

where G is a suitable scalar function that can be introduced without affecting the dynamics.

Proposition 5.2.3. *Admitting possible gauge transformations, the necessary and sufficient algebraic conditions for the time reversal invariance of Equation* (5.2.23) *are expressed by:*

$$A_k^T = -s_k(A_k + \partial_i G) \qquad k = 1, 2, 3 \text{ and } i = x, y, z \tag{5.2.27}$$

Proof. The reasoning used in the proof of Proposition 5.2.2 can be repeated. Introducing $A_i + \partial_i G$ in place of A_i, in the left hand side of Equation (5.2.23), we get:

$$\begin{aligned}&(p_x - q(A_1 + \partial_x G))^2 + (p_y - q(A_2 + \partial_y G))^2 + (p_z - q(A_3 + \partial_z G))^2 = \\ &(p_x + qs_1 A_1^T)^2 + (p_y + qs_2 A_2^T)^2 + (p_z + qs_3 A_3^T)^2\end{aligned} \tag{5.2.28}$$

Then, direct substitution shows that (5.2.27) implies (5.2.28). The inverse implication follows from the fact that Equation (5.2.28) has to hold for any value of the coordinates and the momenta. In particular, considering the case $p_x = p_y = p_z = 0$, we have:

$$(\boldsymbol{A} + \nabla G)^2 = [\boldsymbol{A}^T]^2 \tag{5.2.29}$$

and trivially the following:

$$-p_x(A_1 + \partial_x G) - p_y(A_2 + \partial_y G) - p_z(A_3 + \partial_z G) = p_x s_1 A_1^T + p_y s_2 A_2^T + p_z s_3 A_3^T \tag{5.2.30}$$

The thesis follows separately considering pairs in which two among p_x, p_y and p_z vanish. □

As an example, take a constant magnetic field along the z axis, which corresponds to a vector potential $\boldsymbol{A}(x, y, z) = A_0(0, x, 0) = (0, A_0 x, 0)$, and choose the Coulomb gauge. Then (5.2.24) reduces to $s_1 x = -s_2 x$ for any value of x, that is:

$$s_1 = -s_2 \tag{5.2.31}$$

In this case, the number of diagonal time reversal operations that preserve TRI is four, Ref. [9]. Indeed, every constraint on the values of s_i halves the number of available reversal operations. Then, applying the transformation (5.2.14) to (5.2.21) yields (the same can be repeated for (5.2.15) and (5.2.16)):

$$(-s_P p_x - qA_2(s_P y, s_P x, s_3 z))^2 + (-s_P p_y - qA_1(s_P y, s_P x, s_3 z))^2 + (-s_3 p_z - qA_3(s_P y, s_P x, s_3 z))^2 \tag{5.2.32}$$

and in the same way as Proposition 5.2.2 we derive three necessary and sufficient conditions

$$\begin{cases} A_1(s_P y, s_P x, s_3 z) = -s_P A_2(x, y, z) \\ A_2(s_P y, s_P x, s_3 z) = -s_P A_1(x, y, z) \\ A_3(s_P y, s_P x, s_3 z) = -s_z A_3(x, y, z) \end{cases} \tag{5.2.33}$$

In the singular case $\boldsymbol{A}(x, y, z) = (0, A_0 x, 0)$, (5.2.33) reduces to $0 = \pm s_P x$, which clearly has no solution for $s_P = \pm 1$; on the other hand, one observes that the same magnetic field corresponds to the vector potential $\boldsymbol{A}(x, y, z) = A_0/2(-y, x, 0)$, that instead leads to

$$\begin{cases} -s_P x = -s_P x \\ s_P y = s_P y \end{cases} \tag{5.2.34}$$

which has solution. In other words, the four transformations in the form (5.2.14) continue to hold. The point is that one can use the gauge freedom to replace (5.2.33),

and write:

$$\begin{cases} A_1(s_P y, s_P x, s_3 z) = -s_P(A_2(x,y,z) + \partial_y G) \\ A_2(s_P y, s_P x, s_3 z) = -s_P(A_1(x,y,z) + \partial_x G) \\ A_3(s_P y, s_P x, s_3 z) = -s_z(A_3(x,y,z) + \partial_z G) \end{cases} \tag{5.2.35}$$

In the next section, we discuss in detail the role of the gauge.

5.2.3 Gauge

By definition, the gauge choice has no physical consequences. In our case, the dynamics does not change if the vector potential $\boldsymbol{A}$ is replaced by $\boldsymbol{A} + \nabla G$, with $G : \mathbb{R}^3 \to \mathbb{R}$ a scalar function. As commonly done in this kind of magnetostatic problems, we choose the Coulomb gauge:

$$\nabla \cdot \boldsymbol{A} = 0 \tag{5.2.36}$$

The consequence of this on the physical field $\boldsymbol{B}$, hence on the conditions for TRI, can be illustrated starting from the diagonal transformations and recasting (5.2.27) in the following fashion:

$$(s_1 A_1^T, s_2 A_2^T, s_3 A_3^T) = -(A_1 + \partial_x G, A_2 + \partial_y G, A_3 + \partial_z G) = -(\boldsymbol{A} + \nabla G) \tag{5.2.37}$$

where we used the fact that (5.2.10) has to be an involution.
One can view Equation (5.2.10) (with $P = I$ in the diagonal case) as a transformation on the vector field $V(\mathbb{R}^3)$ of which $\boldsymbol{A}$ is an element, that transforms as a vector and not as a pseudo-vector. Hence, the necessary conditions (5.2.24) imply that $\boldsymbol{A}$ transformed as a vector field in $\mathbb{R}^3$ under a diagonal operation $\mathcal{M}' : V(\mathbb{R}^3) \to V(\mathbb{R}^3)$ has to equal $-\boldsymbol{A}$ up to a gauge transformation, and $\boldsymbol{B}$ is then mapped to $-\boldsymbol{B}$.
The same applies to the non diagonal transformations: we rewrite (5.2.35) as

$$\begin{pmatrix} A_1' \\ A_2' \\ A_3' \end{pmatrix} = -\mathcal{M}_{xy} \begin{pmatrix} A_1 + \partial_x G \\ A_2 + \partial_y G \\ A_3 + \partial_z G \end{pmatrix} \tag{5.2.38}$$

where $A_k' = A_k(s_P y, s_P x, s_3 z)$. As the inverse of the matrix $\mathcal{M}_{xy}$ equals the matrix itself, multiplying Equation (5.2.38) side by side by $\mathcal{M}_{xy}$ the consequence is again to transform $\boldsymbol{A}$ into $-\boldsymbol{A}$ up to a gauge transformation. The same obviously holds for $\mathcal{M}_{xz}$ and $\mathcal{M}_{yz}$.
The gauge freedom can be accounted for by introducing the equivalence classes $[\boldsymbol{A}]$ of the vector potentials that lead to the same magnetic fields, i.e., whose elements differ by the gradient of an at least twice differentiable scalar function $G(x, y, z)$. We denote by $[\boldsymbol{A}]_R$ an element of the class $[\boldsymbol{A}]$, that corresponds to a particular choice of G. We can now state the following:

Proposition 5.2.4. *A generalized time reversal operation $\mathcal{M}$ of form* (5.2.10)*, that acts on all particles 6-dimensional subspaces, preserves TRI in the presence of a magnetic vector potential $\boldsymbol{A}$ if and only if the associated transformation defined on the 3-dimensional vector field space, $\mathcal{M}' : V(\mathbb{R}^3) \to V(\mathbb{R}^3)$, obeys:*

$$\mathcal{M}'\boldsymbol{A} = \mathcal{M}_M(A_1(\mathcal{M}_M \boldsymbol{x}), A_2(\mathcal{M}_M \boldsymbol{x}), A_3(\mathcal{M}_M \boldsymbol{x})) = [-\boldsymbol{A}]_R \tag{5.2.39}$$

with $\mathcal{M}_M$ one 3-dimensional specific matrix representation verifying $\mathcal{M}_M^2 = I$.

When this is verified, the Hamiltonian is preserved up to a gauge transformation and the corresponding equations of motion are in turn verified.

Proof. The direct implication directly comes from Equations (5.2.37) and (5.2.38), where the invariance of the equations of motion leads to the condition (5.2.39). Vice versa, assuming there is an involution $\mathcal{M}'$ verifying Equation (5.2.39), with 3-dimensional matrix representation $\mathcal{M}_M$, one can introduce the transformation $\mathcal{M} \equiv (\mathcal{M}_M \boldsymbol{x}, -\mathcal{M}_M \boldsymbol{p})$, which preserves the structure of the Hamilton equations under time reversal, because it alternates signs. Furthermore, the Hamiltonian is unchanged under the application of $\mathcal{M}$ to every particle space, since $\mathcal{M}(\boldsymbol{p} - q\boldsymbol{A}(\boldsymbol{x}))^2 = (-\mathcal{M}_M \boldsymbol{p} - q\boldsymbol{A}(\mathcal{M}_M \boldsymbol{x}))^2$ by definition. Using (5.2.39) and $\mathcal{M}_M^2 = I$ we obtain $\mathcal{M}(\boldsymbol{p} - q\boldsymbol{A}(\boldsymbol{x}))^2 = (\boldsymbol{p} - q[\boldsymbol{A}(\boldsymbol{x})]_R)^2$. □

Remark 5.2.1. *Applying $\mathcal{M}$ as a variable change in the integral* (5.2.3) *deeply differs from inverting $\boldsymbol{B}$. The coordinates swap operated by $\mathcal{M}$ may amount to a mere rearrangement of the order in which the contributions to the integral coming from the different regions of the phase space are summed up, that does not affect the total. That depends on the functions that are integrated. For instance, given an average electric current from left to right, corresponding to a forward trajectory of particles, its time reverse may exist even if the particles do not trace backward the configurations of the forward trajectory; a reversed average of momenta suffices.*

Remark 5.2.2. *Remark 5.2.1 rests on the hypothesis that all coordinate transformations of interest map the domain of integration on itself. Depending on the geometry of interest, a coordinate change may kick some particle out of the volume occupied by the system under investigation. As long as one remains within the realm of infinite homogeneous systems, or far from possible boundaries, as common in response theory, this is not an issue. In general, one has to consider case by case whether the phase space is invariant under the chosen time reversal mapping. If the dynamics is not translation invariant, making all time reversal symmetries fail, in principle one obtains a method to experimentally find a violation of Onsager reciprocal relations.*

To test the condition of Proposition 5.2.4, it suffices to check that the curl of $\boldsymbol{A}$ and of $\mathcal{M}'\boldsymbol{A}$ corresponds to $\boldsymbol{B}$ and $-\boldsymbol{B}$, respectively. For example, take a constant magnetic field with gauge choices $\boldsymbol{A}_1(x, y, z) = (0, A_0 x, 0)$ and $\boldsymbol{A}_2(x, y, z) = (-A_0 y, 0, 0)$, which are elements of the same class $[\boldsymbol{A}]$. Applying the transformation of Equation (5.2.14) with $s_P = 1$ and $s_3 = 1$, one obtains $\boldsymbol{A}_2'(x, y, z) = (0, -A_0 x, 0)$ that does not equal $-\boldsymbol{A}_2(x, y, z)$, but equals $-\boldsymbol{A}_1(x, y, z)$, showing that it nevertheless belongs to the class $[-\boldsymbol{A}]$. Thus, the transformation of Equation (5.2.14) satisfies the necessary condition (5.2.39) for TRI.

5.2.4 Magnetic field

Proposition 5.2.4 can be formulated in an equivalent form that does not involve gauge freedom:

Proposition 5.2.5. *A generalized time reversal operation $\mathcal{M}$ of form* (5.2.10)*, that acts on all particles 6-dimensional subspaces, preserves TRI in the presence of a magnetic field $\boldsymbol{B}$ if and only if the associated transformation defined on the 3-dimensional vector field space, $\mathcal{M}' : V(\mathbb{R}^3) \to V(\mathbb{R}^3)$, obeys:*

$$\mathcal{M}'\boldsymbol{B} = det(\mathcal{M}_M)\mathcal{M}_M(B_1(\mathcal{M}_M\boldsymbol{x}), B_2(\mathcal{M}_M\boldsymbol{x}), B_3(\mathcal{M}_M\boldsymbol{x})) = -\boldsymbol{B} \tag{5.2.40}$$

with $\mathcal{M}_M$ the 3-dimensional specific matrix representation verifying $\mathcal{M}_M^2 = I$.

Proof. The derivation is trivial because (5.2.39) and (5.2.40) are equivalent statements by definition of a magnetic field as curl of vector potential, which transforms as a pseudo-vector in 3D space. □

Again, TRI preserves the Hamiltonian, up to a gauge choice, as well as the corresponding equations of motion. This perspective is particularly useful in classical mechanics, in which only the magnetic field matters, because the equations of motion are the fundamental element of the theory.
Now, given a magnetic field $\boldsymbol{B}(\boldsymbol{x})$, the necessary conditions for a transformation to preserve TRI are obtained from Equations (5.2.13), (5.2.14), (5.2.15) or (5.2.16). To do that for the 20 transformations we have got, let us express $\boldsymbol{B}$ in the basis $\hat{\boldsymbol{i}}, \hat{\boldsymbol{j}}, \hat{\boldsymbol{k}}$ of the 3-dimensional space as:

$$\boldsymbol{B} = B_1(\boldsymbol{x})\hat{\boldsymbol{i}} + B_2(\boldsymbol{x})\hat{\boldsymbol{j}} + B_3(\boldsymbol{x})\hat{\boldsymbol{k}} \tag{5.2.41}$$

and take the diagonal transformations with matrix representation (5.2.13). Following the rule (5.2.40), $\boldsymbol{B}$ transforms as:

$$\boldsymbol{B}' = s_1s_2s_3[s_1B_1(s_1x, s_2y, s_3z)\hat{\boldsymbol{i}} + s_2B_2(s_1x, s_2y, s_3z)\hat{\boldsymbol{j}} + s_3B_3(s_1x, s_2y, s_3z)\hat{\boldsymbol{k}}] \tag{5.2.42}$$

Then, the necessary matching conditions between the magnetic field components and the transformation follow from the second equality of (5.2.40), and write:

$$\begin{cases} B_1(x,y,z) = -s_2s_3B_1(s_1x, s_2y, s_3z) \\ B_2(x,y,z) = -s_1s_3B_2(s_1x, s_2y, s_3z) \\ B_3(x,y,z) = -s_1s_2B_3(s_1x, s_2y, s_3z) \end{cases} \tag{5.2.43}$$

Therefore, given the magnetic field, one can verify by inspection which of the eight diagonal transformations yield TRI. The same reasoning can be repeated for the non diagonal transformations, with representations (5.2.14), (5.2.15) or (5.2.16), whose application to (5.2.41) implies:

$$\boldsymbol{B}'_{xy} = -s_3[s_PB_2(s_Py, s_Px, s_3z)\hat{\boldsymbol{i}} + s_PB_1(s_Py, s_Px, s_3z)\hat{\boldsymbol{j}} + s_3B_3(s_Py, s_Px, s_3z)\hat{\boldsymbol{k}}] \tag{5.2.44}$$

$$\boldsymbol{B}'_{yz} = -s_1[s_1B_1(s_1x, s_Pz, s_Py)\hat{\boldsymbol{i}} + s_PB_3(s_1x, s_Pz, s_Py)\hat{\boldsymbol{j}} + s_PB_2(s_1x, s_Pz, s_Py)\hat{\boldsymbol{k}}] \tag{5.2.45}$$

$$\boldsymbol{B}'_{xz} = -s_2[s_PB_3(s_Pz, s_2y, s_Px)\hat{\boldsymbol{i}} + s_2B_2(s_Pz, s_2y, s_Px)\hat{\boldsymbol{j}} + s_PB_1(s_Pz, s_2y, s_Px)\hat{\boldsymbol{k}}] \tag{5.2.46}$$

where the subscripts identify the transformation. This derives from the fact that the determinant of the matrices (5.2.14), (5.2.15) and (5.2.16) equals the opposite of the diagonal element: $-s_P^2 s_i = -s_i$. Then, the necessary matching conditions for the 12 non-diagonal reversal operators write:

$$\begin{cases} B_1(x,y,z) = s_3 s_P B_2(s_P y, s_P x, s_3 z) \\ B_2(x,y,z) = s_3 s_P B_1(s_P y, s_P x, s_3 z) \\ B_3(x,y,z) = B_3(s_P y, s_P x, s_3 z) \end{cases} \tag{5.2.47}$$

$$\begin{cases} B_1(x,y,z) = B_1(s_1 x, s_P z, s_P y) \\ B_2(x,y,z) = s_1 s_P B_3(s_1 x, s_P z, s_P y) \\ B_3(x,y,z) = s_1 s_P B_2(s_1 x, s_P z, s_P y) \end{cases} \tag{5.2.48}$$

$$\begin{cases} B_1(x,y,z) = s_2 s_P B_3(s_P z, s_2 y, s_P x) \\ B_2(x,y,z) = B_2(s_P z, s_2 y, s_P x) \\ B_3(x,y,z) = s_2 s_P B_1(s_P z, s_2 y, s_P x) \end{cases} \tag{5.2.49}$$

This concludes the case of systems with $U(\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{C}) = 0$ in the Hamiltonian. For $U(\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{C}) \neq 0$, TRI requires also the following:

$$\mathcal{M}U(\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{C}) = U(\mathcal{M}_C \boldsymbol{X}, -\mathcal{M}_C \boldsymbol{P}, \boldsymbol{C}) = U(\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{C}) \tag{5.2.50}$$

where $\mathcal{M}$ is a time reversal transformation on the phase space, obtained by applying a given $\mathcal{M}_C$ to the coordinates, and alternating signs with the momenta. Let us begin introducing a force $\boldsymbol{E}$ deriving from a scalar potential Φ that depends only on coordinates, so that $-\nabla\Phi = \boldsymbol{F}$, and the Hamiltonian reads:

$$H = \sum_i^N \left[\frac{[\boldsymbol{p}_i - q_i \boldsymbol{A}(x_i, y_i, z_i))]^2}{2m_i} + \Phi(x_i, y_i, z_i) \right] \tag{5.2.51}$$

Given a transformation $\mathcal{M}$ that satisfies the conditions of Proposition 5.2.5, the Hamiltonian (5.2.51) results invariant under the application of $\mathcal{M}$ if:

$$\mathcal{M}\Phi(\boldsymbol{X}) = \Phi(\mathcal{M}_C \boldsymbol{X}) = \Phi(\boldsymbol{X}) \tag{5.2.52}$$

and $\mathcal{M}_C$ is used as in Equation (5.2.50) (*n.b.* this includes the notable case of the coupling with an electric field). In the following Section, we investigate notable examples of force potentials.

5.2.5 Force Potentials

In this Section we consider physically relevant inter-particle potentials. Without loss of generality, we take a constant magnetic fields along the z axis, i.e., $\boldsymbol{B} = (0, 0, 1)$, which breaks four of the eight diagonal time reversal symmetries. In turn, the conditions (5.2.47), (5.2.48) and (5.2.49) imply that only the four non diagonal operations (5.2.14) yield TRI, producing a total of eight time reversal symmetries.

example Take a central potential, e.g., the Coulomb potential between charged particles:

$$U(\boldsymbol{X},\boldsymbol{P},\boldsymbol{C}) = \sum_{i<j}^{N} f_{ij}(\boldsymbol{C}) u(r_{ij})\,; \quad r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}\,, \tag{5.2.53}$$

r_{ij} being the distance between particle i and particle j, $\boldsymbol{C}$ a vector of parameters, and f_{ij} a function of such parameters. This potential satisfies the condition (5.2.52) because each of the 20 available transformations $\mathcal{M}_C$ is an element of the orthogonal group $O(3)$. In particular, one may take block diagonal operators with 3×3 blocks given by (5.2.13), (5.2.14), (5.2.15) or (5.2.16). As a consequence, r_{ij} is left unchanged by the action of $\mathcal{M}$ on the phase space. Moreover, $\mathcal{M}_C$ does not act on the space of the parameters $\boldsymbol{C}$, leaving each f_{ij} invariant.
While very simple, the potentials of this form are most common and useful; in particular, interactions between structureless objects are commonly modelled by central forces, such as those derived from Lennard–Jones, Morse, Coulomb, gravitational and Yukawa potentials.

example The Coulomb ring-shaped (or Hartmann) potential treated in Ref. [16]

$$U(x_i, y_i, z_i) = -\frac{Z}{\sqrt{x_i^2 + y_i^2 + z_i^2}} + \frac{1}{2} Q \frac{1}{x_i^2 + y_i^2} \qquad Q > 0\,,\ Z > 0 \tag{5.2.54}$$

is used in quantum mechanics, and can be used to model a force field that is not purely central, thanks to its second addend, that depends on the square distance from z axis. Here, the term $x_i^2 + y_i^2$ is invariant under the action of the 8 possible diagonal transformations; in particular, we have:

$$(s_1 x_i)^2 + (s_2 y_i)^2 = x_i^2 + y_i^2 \tag{5.2.55}$$

In addition, for the non-diagonal transformations of the form (5.2.14), we have:

$$(s_P y_i)^2 + (s_P x_i)^2 = x_i^2 + y_i^2 \tag{5.2.56}$$

In conclusion, this kind of potential does not add restrictions to TRI, other than those imposed by the magnetic field.

example A different kind of potentials, used, e.g., in molecular dynamics, depends on momenta. For instance, in Ref. [17], classical Fermion-like particles are simulated with the following potential:

$$U(\boldsymbol{p}_i) = \frac{E_p}{1 + e^{b_p(|\boldsymbol{p}_i|^2 - 1)}} \tag{5.2.57}$$

where E_p and b_p are dimensional constants, while $\boldsymbol{p}_i = (p_i^x, p_i^y, p_i^z)$. In this case, the particles are decoupled, but they are subject to an external momentum dependent force. TRI, hence its consequences such as Onsager reciprocal relations, may hold even in a system like this, if the functional form of the magnetic field allows, because $|\boldsymbol{p}_i|$ is invariant under rotations.

example The Polarisable Ion Model (PIM) potential, is particularly interesting in molecular dynamics studies, to take into account certain intermolecular interactions cf. Refs. [18, 19]. In the case of an N particles system, it is expressed by:

$$U = U_{charge} + U_{dispersion} + U_{repulsion} + U_{polarization} \tag{5.2.58}$$

where

$$U_{charge} = \sum_i \sum_{j>i} \frac{q_i q_j}{r_{ij}} \tag{5.2.59}$$

is the Coulomb electric potential,

$$U_{dispersion} = -\sum_i \sum_{j>i} \left(\frac{C_6^{ij}}{(r_{ij})^6} f_6^{ij}(r_{ij}) + \frac{C_8^{ij}}{(r_{ij})^8} f_8^{ij}(r_{ij}) \right) \tag{5.2.60}$$

is due to dipole-dipole and dipole-quadrupole dispersion,

$$U_{repulsion} = \sum_i \sum_{j>i} B_{ij} e^{-\alpha_{ij} r_{ij}} \tag{5.2.61}$$

is a short-range repulsion term, and

$$\begin{aligned} U_{polarization} = & \sum_i \sum_{j>i} \left(\frac{q_i \boldsymbol{r_{ij}} \cdot \boldsymbol{\mu_j}}{(r_{ij})^3} f_4^{ij}(r_{ij}) - \frac{q_j \boldsymbol{r_{ij}} \cdot \boldsymbol{\mu_i}}{(r_{ij})^3} f_4^{ji}(r_{ij}) \right) + \\ & \sum_i \sum_{j>i} \left(\frac{\boldsymbol{\mu_i} \cdot \boldsymbol{\mu_j}}{(r_{ij})^3} - \frac{3(\boldsymbol{r_{ij}} \cdot \boldsymbol{\mu_i})(\boldsymbol{r_{ij}} \cdot \boldsymbol{\mu_j})}{(r_{ij})^5} \right) + \sum_i \frac{|\boldsymbol{\mu_i}|^2}{2\alpha_i} \end{aligned} \tag{5.2.62}$$

is the polarization interaction term, with $\boldsymbol{\mu_i}$ the induced dipole moment of the molecule i. While the parts in Equations (5.2.59), (5.2.60) and (5.2.61) are like the potential (5.2.53), and are invariant under any time reversal operation, the term in Equation (5.2.62) is hard to control, since it is defined recursively: for any particle i, $\boldsymbol{\mu_i}$ in principle depends on the coordinates and on the dipole momenta of all the other particles. Explicitly expressing this dependence is problematic, and the verification of Equation (5.2.50) so far remains out of reach. In fact, this potential is only analyzed through approximations and numerically.

5.3 Conclusions

In this article, we have generalized the results of Refs. [8–11], increasing the number of time reversal symmetries that concern mechanical systems in general, and systems in magnetic field, in particular. We focused on block diagonal transformations, composed by operations acting on the 6-dimensional subspace of each particle, and we have introduced suitable equivalence classes to account for the corresponding gauge invariance. We then obtained sufficient conditions for TRI to hold in presence of a magnetic field, which imply, for instance, Onsager reciprocal relations. Substantially

enlarging the range of applicability of TRI, we contribute to understand why violations of such relations to date are not reported, despite the presence of magnetic fields.
The next step will be to investigate the necessary conditions for the validity of Onsager reciprocal relations. Indeed, as Ref. [11] states, the discovery of a violation of Onsager reciprocal relations may lead to the never observed situation of non-dissipative currents. This may be a dynamically indirect reason why Onsager reciprocal relations cannot be broken, at least in classical systems where the evidence of superconductivity was never found.
In the final part of this paper, we have illustrated the application of our results to notable potentials. Such a few examples do not exhaust the set of possible situations in which TRI holds or is violated, both theoretically and experimentally. However, it covers typical situations and constitutes a guide for further investigations of the Onsager reciprocal relations.
As pointed out by one of the anonymous referees, electromagnetism is inherently relativistic, hence in future works we may investigate the extension of our present results to the relativistic case. As a matter of fact, regarding the time reversal operations on the single particle subspace, thus of any set of non-interacting particles, a formal extension of our involutions is immediate, although not necessarily conceptually satisfactory, given the role of time in Minkowski space. Moreover, Statistical Mechanics relations, such as those considered in this paper, require interacting particles. This makes the subject most intriguing and challenging [20–22].

Chapter 6

Time reversal symmetry for classical, non-relativistic quantum and spin systems in presence of magnetic fields

6.1 Introduction

For particles systems constituting a macroscopic object, the standard derivation of the validity of the Onsager reciprocal relations is based on the time reversal invariance (TRI) of the dynamics [1, 2]. Magnetic fields have long been thought to break TRI, and to require a modification of the Onsager relations, that fundamentally weakens their predictive power and practical relevance [3]. Unlike the original relations, that represent a property of a single system in a given magnetic field, the modified relations refer, indeed, to a relation between two independent systems, separately coupled to opposite magnetic fields.

Kubo continued this investigation in his fundamental papers on linear response, illustrating the effect of the time reversal operator on the correlator of two classical, or quantum, observables ϕ and ψ. Beginning with a classical system in a magnetic field $\mathbf{B}$, with configuration coordinates $\mathbf{r}$, and momenta $\mathbf{p}$, Kubo introduced the following operator:

$$\mathcal{T}_B(\mathbf{r}, \mathbf{p}, t; \mathbf{B}) = (\mathbf{r}, -\mathbf{p}, -t; -\mathbf{B}) \tag{6.1.1}$$

where t is the time variable. This operator implements a kind of time reversal, because it associates every trajectory of the system, with one that traces back itself, in configuration space. He then proved that

$$\langle \phi(0)\psi(t)\rangle_{\mathbf{B}} = \eta_\phi \eta_\psi \langle \phi(0)\psi(-t)\rangle_{-\mathbf{B}} = \eta_\phi \eta_\psi \langle \phi(t)\psi(0)\rangle_{-\mathbf{B}} \tag{6.1.2}$$

where the angular brackets represent averaging with respect to the equilibrium phase space probability distribution in presence of the magnetic field $\mathbf{B}$, or $-\mathbf{B}$, depending on the index; and η_ϕ and η_ψ are respectively the signatures of ϕ and ψ under the action of $\mathcal{T}_B$, cf. [4–6].

Recently, it has however been shown that generalized time reversal symmetries hold, and can be used to prove the validity of the Onsager relations, as well as other relations that require TRI, such as the fluctuation relations, also in presence of magnetic fields, [7–13].

Generalized time reversal transformations different from $\mathcal{T}_B$ were known much earlier, cf. Lax [14] and other authors, but the view that crystallized in the XX century and early XXI century literature was that Onsager relations and fluctuation relations require invariance of the dynamics under the Kubo transformation. Unfortunately, that is not enough to quantify a given property of a given system, but it merely links different properties of two different systems, one with magnetic field $\mathbf{B}$ and the other with opposite, $-\mathbf{B}$, field. In this respect, the Onsager-Casimir relations, in particular, fundamentally differ from the Onsager relations.

This point of view has been superseded in works concerning classical systems coupled to a constant magnetic field along a fixed direction, first in Ref.[7], and in Ref.[11] for space dependent fields. It was shown that there exist time reversal symmetries which are preserved by a magnetic field. The list of such symmetries has been substantially extended in Ref.[15].

The quantum case with a constant magnetic field has been first treated in Ref.[10], for spinless particles. In reality, the concept of generalized TRI was well understood much earlier in the quantum mechanical literature [14]. For instance, in the so-called "ten fold formalism" [16], it is used as a classification tool for the symmetries of quantum systems. Nevertheless, when discussing Onsager relations, Lax reiterates the Casimir reasoning, based on the inversion of the magnetic field [14].

The basis of this approach, using Onsager's words, is the following: *"if the velocities of all the particles present are reversed simultaneously the particles will retrace their former paths, reversing the entire succession of configurations"*, which is known as the microreversibility condition [1]. Continuing, Onsager also states: *"We like to think that the dynamical laws which govern the world of atoms are also reversible. The information that we have about the atoms affords considerable support for this belief of ours, and we have no serious counter-indications, if any. If the dynamical laws of an isolated molecular system are reversible the kinetic theory requires that in the long run every type of motion must occur just as often as its reverse, because the congruence of the two types of motion makes them apriori equivalent"*.

Onsager then develops a reasoning, in which "opposite" phenomena occur with same frequency, opposite meaning the reversed succession of configurations. In absence of magnetic fields or rotating frames, this comes for free in Hamiltonian dynamics, hence it is a perfectly fine assumption.

Therefore, Ref.[16], which investigates the Onsager relations and the symmetries of quantum mechanical systems, takes microreversibility as its starting point: *"In this classification, the possible symmetries that the Hamiltonian H satisfies are (i) TRS : $H = THT^{-1}$, with $T = -iK$, with the complex conjugation operator K in the spinless*

case, and $T = -i\sigma_y K$ for spin-1/2 fermions, with the Pauli matrix σ_y acting in spin space".
Nevertheless, while this approach is sufficient, it is way stronger than required for statistical relations, such as Onsager's. What suffices is that the correlation functions, hence averages over trajectories, enjoy a given symmetry that holds even if trajectories do not come in pairs that trace the same set of configurations in opposite chronological order. One may argue that conditions weaker than microreversibility do not correspond to the *true* time reversibility, whatever that means, but the effect on statistical properties corresponding to macroscopic observables may well be the same. Something less than the traditional reversibility still suffices for the purpose of statistical mechanics or of thermodynamics.
In this paper, we show formally and via examples that:

- there exists a more general class of time reversal transformation in Classical Mechanics. Their compatibility with the magnetic field can be studied through a particular compatibility condition;
- this new generalized time reversal operations, including the ones found in Ref.[15], can be applied to quantum systems;
- the spin is compatible with TRI in presence of an external magnetic field;
- one application of the above results implies an interesting symmetry of the diffusion tensor.

6.2 Theory

As recalled in Sect.6.1, sufficient conditions for the validity of Onsager reciprocal relations are known, and they include a set of generalized time reversal symmetries. On the contrary, very little is known about *necessary* conditions. First, in Sec.6.2.1, we recall the main facts about time reversal transformation in classical mechanics, introducing a whole new class of operations. Then, we illustrate quantum mechanical time reversal operators, in the case of spinless particles, extending the treatment of Sec.6.2.1 to the quantum realm, cf. Sec.6.2.2. In Sec.6.2.3 we investigate the issue of time reversal in the context of quantum systems composed by particle of spin 1/2. In Sec.6.2.4 we summarize the concept of Kubo canonical quantum correlators, as well as the derivation of Onsager relations from a generalized time reversal symmetry obtained in Ref.[10]. Subsequently, we illustrate the compatibility conditions between the different time reversal operations and a generic magnetic field. We recall the sufficient conditions of Ref.[15] that lead to Onsager relations in classical mechanics, and we extend them to the operations found in Sec.6.2.1 and to the case of particles with spin. The striking result of this Section is that a time reversal operator that works in the spinless case, works also for particles of spin 1/2. Finally, in Sec.6.2.5 we study the constraints imposed by TRI on the diffusion tensor in a particular physical setup, in the wake of the reasoning of Ref.[8].

6.2.1 Classical mechanics

The microscopic state of a system made of N classical particles is represented by the collection of coordinates and momenta of all its particles, that constitutes a point Γ in the phase space Ω. In case of dynamics determined by a Hamiltonian H, the time evolution of the coordinates and momenta is prescribed by the following equations of motion:

$$\begin{cases} \dfrac{\partial H}{\partial p_i} = \dot{x}^i \\ \dfrac{\partial H}{\partial x^i} = -\dot{p}_i \end{cases} \qquad i = 1, ..., 3N \tag{6.2.1}$$

Denote by $S^t : \Omega \to \Omega$ the operator that expresses the solutions of the equations of motion up to time t, *i.e.* $S^t\Gamma$ is the state at time t, if it was Γ a time 0. Following Abraham Ref.[17], the dynamics are called time reversal invariant (TRI) if there exists a linear antisymplectic involution $\mathcal{M} : \Omega \to \Omega$ (an operator such that $\mathcal{M}^2 = I$, the identity) such that:

$$\mathcal{M}S^t\Gamma = S^{-t}\mathcal{M}\Gamma\,, \quad \forall\Gamma \in \Omega\,,\ \forall t \in \mathbb{R} \tag{6.2.2}$$

The reason for this terminology is that right multiplication by $\mathcal{M}$ before application to Γ yields:

$$S^{-t} = \mathcal{M}S^t\mathcal{M} \tag{6.2.3}$$

which means that the backward time evolution is conjugated to the forward time evolution *via* the involution $\mathcal{M}$, in such a way that one backward trajectory can be obtained by properly applying $\mathcal{M}$ to the forward trajectory.
Microreversibility corresponds to TRI under the involution $\mathcal{M}$ defined by $\mathcal{M}(\mathbf{x}, \mathbf{p}) = (\mathbf{x}, -\mathbf{p})$, but we call time reversal invariant all the dynamics for which one involution obeying Eq.(6.2.2) exists.
From the point of view of the equations of motion, we can define an extended operation $\mathcal{T} \equiv \mathcal{M} \circ \mathbb{T}$, where $\mathbb{T}$ maps the time parameter t to $-t$ in Eqs.(6.2.1). The system verifies TRI under $\mathcal{M}$ if the application of $\mathcal{T}$ to the equations of motion leaves them unchanged. Note that the concept of time reversal defined by Eq.(6.2.2) concerns trajectories in phase space.

Definition 6.2.1. *A matrix operator P is symplectic or, respectively, antisymplectic on a $2n$-dimensional phase space Ω if*

$$P^T\omega P = \omega \quad \text{or} \quad P^T\omega P = -\omega \tag{6.2.4}$$

where ω is the antisymmetric matrix

$$\omega = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \tag{6.2.5}$$

and I_n is the $n \times n$ identity matrix.

This can now be used to define the time reversal transformations of Hamiltonian systems.

Definition 6.2.2. *A time reversal transformation $\mathcal{M}$, for a Hamiltonian system, is a linear operator acting on the space Ω in such a way that:*

- *It is an involution, that is $\mathcal{M}^2 = I$.*
- *It is an antisymplectic linear operator.*

As in Ref.[15], we begin considering transformations that separately act on the 6-dimensional subspaces of Ω that concern single particles. In this case, a time reversal transformation on Ω can be written as:

$$\mathcal{M}(x, y, z, p_x, p_y, p_z) = P(s_1 x, s_2 y, s_3 z, -s_1 p_x, -s_2 p_y, -s_3 p_z) \tag{6.2.6}$$

where $s_i = \pm 1$, and P is a permutation obeying $P^2 = I$, which acts in the same way on coordinates and on the corresponding momenta. Our first result is the following theorem.

Theorem 6.2.1. *Consider a system of N particles subject to an external magnetic field, described by the Hamiltonian*

$$H = \sum_{i=1}^{N} \left[\frac{(\boldsymbol{p}_i - q_i \boldsymbol{A}(\boldsymbol{x}_i))^2}{2m_i} \right] \tag{6.2.7}$$

The matrix representation of the antisymplectic operator $\mathcal{M}$ representing a time reversal operation for such a system is block diagonal on the phase space Ω and takes the form:

$$\mathcal{M} = \begin{pmatrix} A & 0 \\ 0 & -A \end{pmatrix} \tag{6.2.8}$$

with $M = 3N$, $A \in O(M)$ and $A^2 = I$.

Proof. We start from a generic matrix $\mathcal{M} \in GL(2M, \mathbb{R})$, *i.e.*

$$\mathcal{M} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{6.2.9}$$

where A, B, C and D are $M \times M$ blocks. As stressed in Refs.[7, 15], a time reversal operation defined as in Definition 6.2.2 yields TRI if and only if

$$H(\mathcal{M}\Gamma) = H(\Gamma), \quad \forall\, \Gamma \in \Omega \tag{6.2.10}$$

Thus, the Hamiltonian (6.2.7) requires $\mathcal{M}$ to not mix coordinates and momenta. In the case it does, the minimal coupling term $(\boldsymbol{p}_i - q_i \boldsymbol{A}(\boldsymbol{x}_i))^2$ is not preserved by the operation. In formulae, the matrix must be block diagonal:

$$\mathcal{M} = \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} \tag{6.2.11}$$

As $\mathcal{M}$ must preserve the norm in the $2N$-dimensional phase space, because of the presence of the square in the minimal coupling term, $\mathcal{M}$ must be orthogonal and,

consequently one has $A \in O(M)$ and $D \in O(M)$, with $M = 3N$. Moreover, the transformation must be antisymplectic by Definition 6.2.2. Then, we impose the constraint:

$$\mathcal{M}^T \omega \mathcal{M} = -\omega\,; \quad \text{i.e.} \quad \begin{pmatrix} A^T & 0 \\ 0 & D^T \end{pmatrix} \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix} \tag{6.2.12}$$

which leads to the constraint $A^T D = -I$. Then,

$$\mathcal{M} = \begin{pmatrix} A & 0 \\ 0 & -(A^T)^{-1} \end{pmatrix} \tag{6.2.13}$$

Finally, because $\mathcal{M}$ is block diagonal, and $\mathcal{M}^2 = I$, Eq.(6.2.11) implies $A^2 = I$ and $D^2 = I$. Ultimately, since A is orthogonal, that is $AA^T = I$, and $A^2 = I$, we obtain that A is a symmetric matrix, *i.e.* $A = A^T$. The proof is then complete, observing that

$$(A^T)^{-1} = A^{-1} = A \tag{6.2.14}$$

□

This means that the number of time reversal operations is the number of matrices A that are orthogonal and involutory. In particular, let us consider $3N \times 3N$ matrices whose entries are 0 or 1, called binary matrices (more general matrices will be considered later, in Section 6.2.4). Among binary matrices we have the following:

Definition 6.2.3. *A permutation matrix is a matrix obtained permuting the row of an identity matrix of the same dimension.*

We are now going to find the time reversal matrices using some results of Group Theory [18]. First, note the following proposition.

Proposition 6.2.1. *The involutory binary matrices A are in one-to-one correspondence with the linear representations of the cyclic group $\mathbb{Z}_2$ over a $3N$-dimensional vector space.*

Proof. By definition, the group $\mathbb{Z}_2$ is the set $\{e, a\}$ with e the group identity and $a^2 = e$. A linear representation $\mathcal{R}(a)$ on a $3N$-dimensional vector space is a $3N \times 3N$ matrix such that $\mathcal{R}(a)^2 = \mathcal{R}(e) = I_{3N}$. Moreover, S_2, the symmetric group on 2 elements, coincides with $\mathbb{Z}_2$ and, in general, a representation of an element of S_k must be a permutation, hence an orthogonal, matrix. Then, a matrix $\mathcal{R}(a)$ has the required property, and the possible A and the $\mathcal{R}(a)$ are in one-to-one correspondence. □

Considering that one such matrix A is a permutation, it is a representation of an element of S_{3N}, like $\mathcal{R}(a)$. Now, let us introduce the cycle decomposition of a permutation of S_{3N}, denoted by:

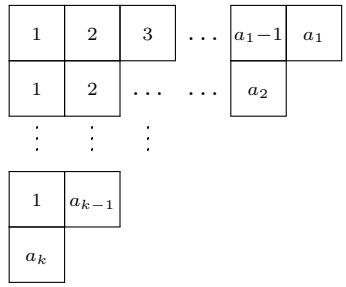
$$\underbrace{(\cdot)\ldots(\cdot)}_{r_1}\underbrace{(\cdot\cdot)\ldots(\cdot\cdot)}_{r_2}\ldots\underbrace{(\cdot\ldots\cdot)}_{r_{3N}} \tag{6.2.15}$$

which is invariant under group conjugation, and allows us to uniquely label each conjugation class with a set of integers $\{r_1, ..., r_{3N}\}$ satisfying the following constraint:

$$\sum_{l=1}^{3N} l r_l = 3N \tag{6.2.16}$$

There is a useful way to represent the conjugation classes:

Definition 6.2.4 (Young tableaux). *There is a one-to-one correspondence between a conjugation class $\{r_l\}$ of the group S_k and the following graphical representations known as Young tableau,*



(6.2.17)

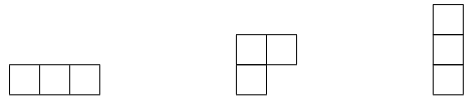
where the number of squares a_l in each of the k rows obeys the following rules:

1. $a_{l+1} \leq a_l \quad \forall l = 1, ..., k.$
2. *The total number of squares in the tableau equals k:* $\sum_{l=1}^{k} a_l = k$

Then, the one-to-one correspondence between the class $\{r_l\}$ and the corresponding Young tableau is given by:

$$\begin{cases} r_l = a_l - a_{l+1} \ , & \forall\, l = 1, ..., k-1 \\ r_k = a_k \end{cases} \tag{6.2.18}$$

For instance, S_3 can be associated with any of the three following Young tableaux:



(6.2.19)

that respectively correspond to three sets $\{r_l\}$: $\{3,0,0\}$, $\{1,1,0\}$ and $\{0,0,1\}$. Incidentally, this also constitutes one method for identifying the partitions of a fixed integer n (see Cap. 5.2 in Ref.[18]). Now, the constraint (6.2.16) takes an interesting form, thanks to the following Corollary.

Corollary 6.2.1. *The conjugation classes $\{r_i\}$ of S_{3N} containing elements represented by involutory binary matrices obey*

$$r_l = 0 \quad \forall l \geq 3 \tag{6.2.20}$$

Proof. This is a consequence of Proposition 6.2.1: the conjugation class represented by A contains elements a such that $a^2 = e$, that is its cycle decomposition can be made only of cycles of order 1 or 2. □

To compute the number $|\{r_i\}|$ of elements in a conjugation class $\{r_i\}$, we use the following known fact [18]:

Theorem 6.2.2. *The number of elements in the conjugation class* $\{r_l\}$ *is*

$$|\{r_l\}| = \frac{k!}{\prod_{i=1}^{k} i^{r_i} r_i!} \tag{6.2.21}$$

which leads us the fundamental result of this Section:

Corollary 6.2.2 (Number of generalized time reversal operations)**.** *Setting* $M = 3N$*, the maximum number of generalized time reversal operations involving binary matrices, for a system of* N *particles described by the Hamiltonian* (6.2.7)*, is given by:*

$$\tilde{\Delta}_{even}(M) = \sum_{r_2=0}^{M/2} \frac{M!\, 2^{M-2r_2}}{(M-2r_2)!\, r_2!} \tag{6.2.22}$$

if M *is even and by*

$$\tilde{\Delta}_{odd}(M) = \sum_{r_2=0}^{(M-1)/2} \frac{M!\, 2^{M-2r_2}}{(M-2r_2)!\, r_2!} \tag{6.2.23}$$

if M *is odd.*

Proof. The total number of time reversal operations that satisfy Corollary 6.2.1 is the sum of the number of elements in the conjugation classes, *i.e.*

$$\tilde{\Delta} \equiv \sum_{\{r_l\}} |\{r_l\}| \tag{6.2.24}$$

under the condition (6.2.20) which, together with (6.2.16), yields:

$$r_1 + 2r_2 = M \tag{6.2.25}$$

Then, using (6.2.21) and (6.2.25) in (6.2.24) we obtain

$$\Delta_{even}(M) = \sum_{r_2=0}^{M/2} \frac{M!}{(M-2r_2)! r_2!\, 2^{r_2}} \tag{6.2.26}$$

for M even, while

$$\Delta_{odd}(M) = \sum_{r_2=0}^{(M-1)/2} \frac{M!}{(M-2r_2)! r_2!\, 2^{r_2}} \tag{6.2.27}$$

for M odd, where the sum stops at $(M-1)/2$, because of the constraint in Eq.(6.2.25).

Now, consider for instance one conjugation class $\{r_l\}$, hence a single addend of the summation in these formulae. Each cycle in a decomposition like (6.2.15) is defined up to a minus sign. For example, take the matrix representation

$$\begin{pmatrix} 0 & s_p & 0 \\ s_P & 0 & 0 \\ 0 & 0 & s_3 \end{pmatrix} \tag{6.2.28}$$

associated to the time reversal operation that acts as a block diagonal on each particle subspace as

$$\mathcal{M}(x, y, z, p_x, p_y, p_z) = (s_P y, s_P x, s_3 z, -s_P p_y, -s_P p_x, -s_3 p_z) \tag{6.2.29}$$

That corresponds to a permutation belonging to the conjugation class $\{r_l\} = \{1, 1, 0\}$ of S_3: it is evident as the 2×2 block that corresponds to the cycle of order 2 carries a sign s_P, and similarly the cycle of order 1 for s_3. In general, we can multiply by ± 1 each cycle block in the representative matrix. This proves that we have to multiply each addend of the summations in (6.2.26) and (6.2.27) by a factor $2^{r_1+r_2}$, obtaining:

$$\tilde{\Delta}_{even}(M) = \sum_{r_2=0}^{M/2} \frac{M!\, 2^{r_1+r_2}}{(M-2r_2)! r_2!\, 2^{r_2}} \tag{6.2.30}$$

$$\tilde{\Delta}_{odd}(M) = \sum_{r_2=0}^{(M-1)/2} \frac{M!\, 2^{r_1+r_2}}{(M-2r_2)! r_2!\, 2^{r_2}} \tag{6.2.31}$$

Finally, recalling relation (6.2.25) we get the two formulae (6.2.22) and (6.2.23). □

To illustrate this result, let us apply it to the 6-dimensional subspace of a system with $N = 1$. Here, $M = 3$ so we use (6.2.23) and we obtain:

$$\tilde{\Delta}_{odd}(3) = \sum_{r_2=0}^{1} \frac{M!\, 2^{M-2r_2}}{(M-2r_2)!\, r_2!} = \frac{3!\, 2^3}{3!\, 0!} + \frac{3!\, 2^1}{1!\, 1!} = 8 + 12 = 20 \tag{6.2.32}$$

Correctly, this result coincides with the one obtained in Ref.[15] for single particle subspaces of N particle systems. This kind of operations swap different coordinates and momenta of a given particle. For example, they include the time reversal transformation

$$(x_1, ..., x_j, x_{j+1}, ..., x_M, p_1, ..., p_j, p_{j+1}, ..., p_M) \xrightarrow{\mathcal{M}_{nd}} (x_1, ..., x_{j+1}, x_j, ..., x_M, -p_1, ..., -p_{j+1}, -p_j, ... \tag{6.2.33}$$

that permutes coordinates x_j and x_{j+1} of a given particle. In fact, this operation corresponds to a matrix A of form

$$\begin{pmatrix} I_{j-1} & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & I_{M-j-1} \end{pmatrix} \tag{6.2.34}$$

where I_k is the $k \times k$ identity matrix and B is a 2×2 block such as

$$B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{6.2.35}$$

This matrix is the representation of a permutation that belongs to the conjugation class $\{M-2, 1\}$, that is the permutations composed by $M-2$ cycles of order 1 and one cycle of order 2.
As last remark, the generalization of the concept of time reversal leads to a notable outcome on the rule of transformation of other classical observables, as for example angular momentum. Let us recall its definition:

Definition 6.2.5. *For a classical Hamiltonian system, the total angular momentum is defined by*

$$\boldsymbol{L} = \sum_{i=1}^{N} \boldsymbol{x}^i \times \boldsymbol{p}_i \tag{6.2.36}$$

The canonical time reversal operation $(\boldsymbol{X}, \boldsymbol{P}) \mapsto (\boldsymbol{X}, -\boldsymbol{P})$ trivially transforms $\boldsymbol{L}$ into $-\boldsymbol{L}$. The same holds for the 20 time reversal operations that do not permute coordinates of different particles. Restricting to the single particle subspaces $(\boldsymbol{x}^i, \boldsymbol{p}_i)$, and denoting the transformation rule by:

$$(\boldsymbol{x}^i, \boldsymbol{p}_i) \to (R\boldsymbol{x}^i, -R\boldsymbol{p}_i) \tag{6.2.37}$$

where Theorem 6.2.1 implies $R \in O(3)$, we obtain:

$$\boldsymbol{L} \mapsto \boldsymbol{L}' = -\sum_{i=1}^{N} R\boldsymbol{x}^i \times R\boldsymbol{p}_i = -\boldsymbol{L} \tag{6.2.38}$$

where we used the fact that the cross product is unchanged by coherent rotation of both factors. On the other hand, transformations with matrix representation like (6.2.8), where A does not separately act on each particle subspace, do not reverse $\boldsymbol{L}$. Consider, for example, the following generalized time reversal operation:

$$(x_1, ..., x_j, x_{j+1}, ..., x_{3N}, p_1, ..., p_j, p_{j+1}, ..., p_{3N}) \xrightarrow{\mathcal{M}_{nd}} (x_{3N}, ..., x_j, x_{j+1}, ..., x_1, -p_{3N}, ..., -p_j, -p_{j+1}, ..., \tag{6.2.39}$$

which swaps x_1 and x_{3N} and coherently acts on momenta as prescribed by Theorem 6.2.1. Then, consider the z component of $\boldsymbol{L}$. By definition of cross product we have:

$$L_z = x^1 p_2 - x^2 p_1 + ... + x^{3N-2} p_{3N-1} - x^{3N-1} p_{3N-2} \tag{6.2.40}$$

which becomes

$$L_z = -x^{3N} p_2 + x^2 p_{3N} + ... - x^{3N-2} p_{3N-1} + x^{3N-1} p_{3N-2} \neq -L_z \tag{6.2.41}$$

under the application of $\mathcal{M}_{nd}$. The same holds for L_x and L_y. The conclusion is the following:

Proposition 6.2.2. *There exist generalized time reversal operations that do not reverse the sign of the classical mechanics angular momentum.*

Note that the time reversal operation $\mathcal{M}_{nd}$ works, in particular for a system of free equal mass particles, whose Hamiltonian is expressed by $H = \sum_{i=1}^{3N} p_i^2/2m$.
We now extend to Quantum Mechanics, the present reasoning. Indeed, in Ref.[10], only 8 time reversal operations had been identified, that classically take the form

$$\mathcal{T}(x, y, z, p_x, p_y, p_z, t) = (s_1 x, s_2 y, s_3 z, -s_1 p_x, -s_2 p_y, -s_3 p_z, -t) \tag{6.2.42}$$

6.2.2 The time reversal operator

Let us begin with spinless nonrelativistic particles. The fundamental axiom is that the state of the system is represented by a wave function, that obeys the Schrödinger equation:

$$i\frac{\partial \psi(x, y, z, t)}{\partial t} = H\psi(x, y, z, t) \tag{6.2.43}$$

where our units imply $\hbar = 1$. As in Classical Mechanics, we have an evolution equation that is invariant under time reversal. The corresponding Hilbert space is the quantum counterpart of the classical phase space. Then, we follow Wigner's approach to define time reversal transformations, cf. Refs.[19, 20]. In the following we will consider operators and their matrix representations; for the sake of simplicity, the same symbol will be used.

Definition 6.2.6 (Spinless particles)**.** *An operator on the Hilbert space of the wave functions of an N particle system is called a time reversal operator $\mathcal{T}$ if it obeys:*

- *$\mathcal{T} = UK$, with U a unitary operator and K the complex conjugation operator, i.e. it is antilinear*
- *It is an involution in the quantum sense, that is $\mathcal{T}^2 = \pm I$, where I is the identity operator*
- *It is kinematically admissible, that is it preserve the canonical commutation relations $[x_n, p_m] = i\delta_{nm}$*

To proceed analogously to the classical case, we exploit the following result [21]:

Proposition 6.2.3. *The matrix representation of a time reversal operator $\mathcal{T}$ for a system of N particles belongs to $Sp(6N, \mathbb{C})$.*

For instance, in the case of a system with only two coordinates, the operators x_1 and x_2, and of a time reversal operator such that:

$$\begin{aligned} \mathcal{T} x_1 \mathcal{T}^{-1} &= x_2 \\ \mathcal{T} x_2 \mathcal{T}^{-1} &= x_1 \end{aligned} \tag{6.2.44}$$

the matrix representation of $\mathcal{T}$ in terms of coordinates is given by:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{6.2.45}$$

The Stone-Von Neumann Theorem [22] now states that in coordinates representation, *i.e.* for a wave function defined as coordinate dependent, the operators x^i and p_i uniquely act as:

$$\begin{cases} x^i\psi(x_0) = x_0^i\psi(x_0) \\ p_i\psi(x_0) = -i\dfrac{\partial\psi}{\partial x^i}(x_0) \end{cases} \tag{6.2.46}$$

One can equivalently define the action of coordinate and momentum operators in momentum space, via Fourier transform.

Proposition 6.2.4. *The complex conjugation preserves the canonical commutation relation.*

Proof. Using Eq. (6.2.46), we can obtain the rule of transformation of the coordinate and momentum operators:

$$Kp_iK^{-1} = -KiK^{-1}\frac{\partial}{\partial x_i} = i\frac{\partial}{\partial x^i} = -p_i \tag{6.2.47}$$

and

$$Kx^iK = x^i \tag{6.2.48}$$

The proof of the canonical nature of K is then trivial, since applying K to both side of the canonical commutation relations

$$\left[x^i, p_j\right] = i\delta_j^i \tag{6.2.49}$$

one obtains

$$\left[Kx^iK, Kp_jK\right] = KiK\delta_j^i \tag{6.2.50}$$

that is equivalent to (6.2.49) by definition of K. Then, the $6N\times6N$ matrix representing K on the $6N$-dimensional space of symbolic vectors $(\boldsymbol{X}, \boldsymbol{P})$, is given by:

$$K = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \tag{6.2.51}$$

which is trivially antisymplectic, and acts separately on coordinate and momentum operators □

Counting the generalized time reversal operations is now reduced to finding alternatives to the matrix representation of K. Table 6.1 illustrates the parallel with Classical Mechanics.

Recall that a fundamental aspect of Classical mechanics is that $\mathbb{T}$ acts on the Hamiltonian structure as an antisymplectic operator, and then $\mathcal{M}$ is also antisymplectic in order to obtain a symplectic transformation. Now, Proposition 6.2.3 says that in Quantum Mechanics $\mathcal{T}$ is represented by a symplectic matrix. Furthermore, K is represented by an antisymplectic matrix, that is it plays the role of $\mathbb{T}$. So, the situation is very similar to the classical one, and U is antisymplectic.

	Classical Mechanics	Quantum Mechanics
Time Reversal	$\mathcal{M}$	$\mathcal{T} = UK$
Commutation Relations	$\{x^i, p_j\} = \delta^i_j$	$\left[x^i, p_j\right] = i\delta^i_j$

Table 6.1: Comparative scheme

Theorem 6.2.3. *Take the $3N$-dimensional vectors $\boldsymbol{X}$ and $\boldsymbol{P}$ of respectively the coordinate and momentum operators of a quantum system made of N particles. Given a time reversal operator $\mathcal{T} = UK$, the matrix representation of U on the $6N$-vectors $(\boldsymbol{X}, \boldsymbol{P})$ takes the form*

$$U = \begin{pmatrix} A & 0 \\ 0 & -A \end{pmatrix} \tag{6.2.52}$$

where A is a symmetric or an antisymmetric matrix.

Proof. As in Theorem 6.2.1, $\mathcal{T}$ cannot swap coordinates and momenta of different particles, because of the minimal coupling term in the Hamiltonian. Then, the matrix representation of U takes the form

$$U = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \tag{6.2.53}$$

As in the proof of Theorem 6.2.1, we obtain $A^\dagger B = -I$. Then, we can write:

$$U = \begin{pmatrix} A & 0 \\ 0 & -(A^\dagger)^{-1} \end{pmatrix} \tag{6.2.54}$$

where we used the fact that the antisymplectic condition for $Sp(6N, \mathbb{C})$ is expressed by:

$$U^\dagger \omega U = -\omega \tag{6.2.55}$$

and involves the dagger operation instead of the transpose. Now, U is unitary by Definition 6.2.6, and so we have $A = (A^\dagger)^{-1}$, which leads to Eq. (6.2.52).
To prove that A is symmetric or antisymmetric, use the involutory property $\mathcal{T}^2 = UKUK = UU^* = \pm I$, which implies $AA^* = \pm I$. Moreover, U is unitary, hence $AA^\dagger = I$. Then, for $AA^* = I$ we obtain $A^T = A$, *i.e.* A symmetric, whilst for $AA^* = -I$ we have $A^T = -A$, and is A antisymmetric. □

Note that K is of the form (6.2.52), and corresponds to the the canonical time reversal operation, that preserves coordinates and reverses momenta. If we restrict to the part of U acting on the coordinate (and then separately on momentum) operators, we find that a complex transformation of $\boldsymbol{X}$ and $\boldsymbol{P}$ cannot be used. Suppose that

$$\mathcal{T}\boldsymbol{P}\mathcal{T}^{-1} = P' = \mathfrak{Re}(\boldsymbol{P'}) + \mathfrak{Im}(\boldsymbol{P'})i \tag{6.2.56}$$

and consider a system of free particles with Hamiltonian $H = \boldsymbol{P}^2/2m$. Applying the time reversal operator, we have:

$$H' = \mathcal{T}H\mathcal{T}^{-1} = \frac{(\boldsymbol{P}')^2}{2m} = \frac{\mathfrak{Re}(\boldsymbol{P}')^2 + 2\mathfrak{Im}(\boldsymbol{P}')\mathfrak{Re}(\boldsymbol{P}')i - \mathfrak{Im}(\boldsymbol{P}')^2}{2m} \tag{6.2.57}$$

where the Hamiltonian is preserved only if $2\mathfrak{Im}(\boldsymbol{P}')\mathfrak{Re}(\boldsymbol{P}') = 0$, since the H is real in $\boldsymbol{X}$ and $\boldsymbol{P}$. But if $\mathfrak{Im}(\boldsymbol{P}') \neq 0$, as assumed, that requires $\mathfrak{Re}(\boldsymbol{P}') = 0$. But then $H' < 0$ should equal $H > 0$, which is absurd. This reasoning can be extended to the case of an external magnetic field. Because $H \propto \sum_{i=0}^{N}(\boldsymbol{p}_i - q\boldsymbol{A}(\boldsymbol{x}_i))^2$, there is again an addend proportional to $\boldsymbol{P}^2$ and the proof applies. Therefore, we may adopt the following:

Assumption 6.2.1. *The unitary operation U associated to a time reversal operator $\mathcal{T} = UK$ acting on the coordinate (and momentum) operators is real.*

This assumption immediately leads to:

Corollary 6.2.3. *If A is real and symmetric, it belongs to $O(3N)$, as in Classical Mechanics.*

Proof. In the real case, the constraints $AA^\dagger = I$ and $A = A^T$ imply $AA^T = I$, which means that $A \in O(3N)$. □

We can then straightforwardly repeat the arguments developed for the classical case, starting from Proposition 6.2.1. In particular, we consider permutation matrices also in this quantum framework. Therefore, the generalized time reversal operators can be counted using Corollary 6.2.2. Moreover, let A be real but antisymmetric. In this case, Theorem 6.2.3 requires $A^2 = -I$. Then, consider the following definition.

Definition 6.2.7. *Let $\mathbb{Z}_4$ be a finite cyclic group with generator a (a^4 is the group identity e), and with the real 2×2 matrix representation given by:*

$$\mathcal{R}_4(e) = I \qquad \mathcal{R}_4(a) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \qquad \mathcal{R}_4(a^2) = -I \qquad \mathcal{R}_4(a^3) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{6.2.58}$$

Then, the following holds.

Proposition 6.2.5. *If the number of particles N is odd, A is not real and antisymmetric. If N is even, the number of time reversal operators $\tilde{\Delta}$ (that is the number of possible A's) equals*

$$\tilde{\Delta} = \frac{M!}{(M/2)!} \quad \textit{with} \quad M = 3N \tag{6.2.59}$$

Proof. The condition $A^2 = -I$ means that $A \in GL(3N, (R))$ is a linear representation of the element a or a^3 of the group $\mathbb{Z}_4$ over a $3N$-dimensional vector space. Moreover, $A^T = -A$ implies that it is antisymmetric. Then, A must contain only 2×2 blocks, like $\mathcal{R}_4(a)$ and $\mathcal{R}_4(a^3)$, with vanishing diagonal elements. So, for odd N, the $3N \times 3N$ matrix cannot be built using 2×2 blocks.

In the case of N even, it suffices to use Eq.(6.2.21) with $k = M$ and $r_i = 0$, for $i \neq 2$. Indeed, Corollary 6.2.1 holds, and antisymmetry implies that the diagonal elements vanish, excluding cycles of order 1. Moreover, it is trivial to make a one-to-one correspondence between $\mathcal{R}_4(a)$ and the representative of the non trivial element of $\mathbb{Z}_2$:

$$\mathcal{R}(a) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{6.2.60}$$

Then, one must count the elements of a conjugation class built only with cycles of order 2, which can be done using Eq.(6.2.21) and considering that the number of cycles of order 2 over M elements is $M/2$. The result is:

$$\Delta = \frac{M!}{2^{M/2}(M/2)!} \tag{6.2.61}$$

Finally, unlike the case of Classical mechanics, in which $\mathcal{M}^2 = I$, here we also admit the opposite sign for the 2×2 block, that corresponds to $\mathcal{R}_4(a^3)$. Multiplying Δ by $2^{M/2}$, we finally get Eq.(6.2.59). □

The difference between Classical and Quantum cases is easily revealed by the following example. The Hamiltonian of a system made of 2 free quantum particles moving in 1-dimension is given by $H = p_1^2/2m + p_2^2/2m$. Consider a time reversal operator $\mathcal{T}$ whose unitary part is represented on the vectors of momentum operators $\boldsymbol{P} = (p_1, p_2)$ by the matrix

$$Q = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \tag{6.2.62}$$

While admitted in quantum mechanics, it is not acceptable in classical mechanics, because $Q^2 = -I$.
Having exhausted our treatment of spinless particles coupled with an external magnetic field, we now turn to the case of particles with spin 1/2.

6.2.3 Time reversal with spin

For nonrelativistic spin systems, the equation of motion is the Pauli equation:

$$\left\{ \sum_i^N \frac{1}{2m_i} [\boldsymbol{\sigma}_i \cdot (\boldsymbol{p}_i - q_i \boldsymbol{A}(x_i, y_i, z_i))^2 \right\} |\psi\rangle = i \frac{\partial}{\partial t} |\psi\rangle \tag{6.2.63}$$

where $|\psi\rangle$ represents the total state of the system of N particles and the spin operator $\boldsymbol{\sigma}_i$ of the i-th particle is expressed by

$$\boldsymbol{\sigma}_i = I \otimes ... \otimes \underbrace{\boldsymbol{\sigma}}_{i} \otimes ... \otimes I \tag{6.2.64}$$

where $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ contains the Pauli matrices as Cartesian components and I is the identity operator. The Pauli vector identity:

$$(\boldsymbol{\sigma} \cdot \boldsymbol{a})(\boldsymbol{\sigma} \cdot \boldsymbol{b}) = \boldsymbol{a} \cdot \boldsymbol{b} + i\boldsymbol{\sigma} \cdot (\boldsymbol{a} \times \boldsymbol{b}) \tag{6.2.65}$$

yields

$$\sum_i^N \frac{1}{2m_i}\left[(\boldsymbol{p}_i - q_i\boldsymbol{A}(x_i,y_i,z_i))^2 - q_i\boldsymbol{\sigma}_i\cdot\boldsymbol{B}\right]|\psi\rangle = i\frac{\partial}{\partial t}|\psi\rangle \tag{6.2.66}$$

where $\boldsymbol{B} = \nabla\times\boldsymbol{A}$. For the time reversal operator, we may adopt the following form:

$$\mathcal{T} = (U_c\otimes U_s)K \tag{6.2.67}$$

where U_c denotes the part acting on the coordinate and momentum operators, as in Sec.6.2.2; while U_s is the part acting on the spin operators. Both U_c and U_s must be unitary, in order to represent time reversal operators. Since $\boldsymbol{\sigma}_i$ separately acts on single particles spin operators, in the following we omit the subscript i, considering, without loss of generality, the effect of $\mathcal{T}$ on $\boldsymbol{\sigma}$.
In this framework, time reversal has been traditionally defined as an operator $\mathcal{T}$ that preserves the commutation relations and, analogously to classical mechanics, yields [20]:

$$\mathcal{T}\boldsymbol{\sigma}\mathcal{T}^{-1} = -\boldsymbol{\sigma} \tag{6.2.68}$$

which directly implies:

$$U_s = \sigma_y \tag{6.2.69}$$

Unfortunately, this condition does not allow us to treat particles with spin [10]. On the other hand, Proposition 6.2.2 suggests we may relax some restriction. Therefore, we adopt the following, minimal, definition that does not include condition (6.2.68) and that actualizes Definition 6.2.6 to particles with spin 1/2.

Definition 6.2.8 (Particles with spin 1/2)**.** *An antilinear operator $\mathcal{T}$ verifying Definition 6.2.6 is called a time reversal operator for particles with spin, if it preserves the commutation relations of Pauli matrices:*

$$[\sigma_i,\sigma_j] = i\varepsilon_{ij}{}^k\sigma_k \tag{6.2.70}$$

where $i,j,k = x,y,z$ and $\varepsilon_{ij}{}^k$ is the Levi-Civita symbol.

Note that only σ_y is complex and so the complex conjugation operator K preserves (6.2.70). To proceed, let us recall two results of Lie Group Theory [23], taking $\mathbb{R}^3$ as a Lie algebra generated by three vectors $(\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3)$:

Theorem 6.2.4. *There exists a linear invertible map $\Phi:\mathfrak{su}(2)\to\mathbb{R}^3$ such that*

$$\Phi\left(\left[\frac{i}{2}\sigma_i,\frac{i}{2}\sigma_j\right]\right) = \Phi\left(\frac{i}{2}\sigma_i\right)\times\Phi\left(\frac{i}{2}\sigma_j\right);\quad \textit{with}\quad \Phi\left(\frac{i}{2}\sigma_i\right) = \boldsymbol{e}_i \tag{6.2.71}$$

where $\times$ (the usual cross product) is the Lie product.

Theorem 6.2.5. *An element of the unitary group $U\in U(2)$ can be expressed as*

$$U = e^{\frac{i}{2}\lambda^0}\left(I\cos\frac{\lambda}{2} + i\frac{\lambda^j\sigma_j}{\lambda}\sin\frac{\lambda}{2}\right) = e^{\frac{i}{2}\lambda^0}V \tag{6.2.72}$$

where $\lambda^j\in\mathbb{R}$ for $j=0,1,2,3$, $\lambda = |\boldsymbol{\lambda}|$, for $\boldsymbol{\lambda} = (\lambda^1,\lambda^2,\lambda^3)$, and $V\in SU(2)$.

Theorem 6.2.4 allows us to find a U_s that preserves the commutation relations (6.2.70), by finding a transformation that preserves a right handed triad in $\mathbb{R}^3$.

Lemma 6.2.1. *Linear transformations mapping a right handed basis in $\mathbb{R}^3$ into another belong to $SO(3)$.*

Another useful fact is that given $V \in SU(2)$ as in (6.2.72), one has:

$$\sigma_y U \sigma_y = U^* \tag{6.2.73}$$

where U^* is the complex conjugate of U. Then, thanks to the properties of the Pauli matrices, we obtain:

$$\sigma_y \left(I \cos\frac{\lambda}{2} + i\frac{\lambda^j \sigma_j}{\lambda} \sin\frac{\lambda}{2} \right) \sigma_y = I\cos\frac{\lambda}{2} + i\frac{\lambda^1\sigma_y\sigma_x\sigma_y + \lambda^2\sigma_y\sigma_y\sigma_y + \lambda^3\sigma_y\sigma_z\sigma_y}{\lambda} \sin\frac{\lambda}{2} = \left(I\cos\frac{\lambda}{2} - i \right. \tag{6.2.74}$$

Another useful classical result of group theory is the following:

Theorem 6.2.6. *The quotient group $SU(2)/\mathbb{Z}^2$ is isomorphic to the group $SO(3)$. The isomorphism is defined by:*

$$U^\dagger \sigma_i U = [\Lambda(U)]_i^j \sigma_j \tag{6.2.75}$$

where $\Lambda(U)$ is a 3×3 special orthogonal matrix associated to $U \in SU(2)/\mathbb{Z}^2$ and we used Einstein summation rule.

For example, the operator U_d acting on the spin operators of a single particle as

$$\begin{aligned} U_d \sigma_x U_d^{-1} &= s_x \sigma_x \\ U_d \sigma_y U_d^{-1} &= s_y \sigma_y \\ U_d \sigma_z U_d^{-1} &= s_z \sigma_z \end{aligned} \tag{6.2.76}$$

can be associated with the matrix

$$\Lambda(U_d) = \begin{pmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{pmatrix} \tag{6.2.77}$$

and the whole equivalence class $\{U_d, -U_d\}$ is is represented by $\Lambda(U_d)$.
Let us now analyze the case in which $\Lambda(U)$ is a permutation and an involution, hence it is not a cyclic or counter-cyclic permutation of the basis vectors. Proceeding analogously to the case of time reversal operations acting on coordinates, we obtain the following theorem concerning the spin space.

Theorem 6.2.7. *Given an operator $\mathcal{T} = UK$, where K is complex conjugation, a time reversal operator acting on the spin space is obtained if U is unitary and obeys:*

$$\begin{aligned} U_{xy}^{(1)} &= \theta(\sigma_z \mp iI) \\ U_{yz}^{(1)} &= \theta(\sigma_x \mp iI) \\ U_{xz}^{(2)} &= \theta(\sigma_x \pm \sigma_z) \\ U_d &= \sigma_i \end{aligned} \tag{6.2.78}$$

where $\theta \in \mathbb{C}$, $|\theta| = 1/\sqrt{2}$, *and* $i = x, y, z$.

Proof. First, we consider the diagonal transformations U_d that do not permute the Pauli matrices:

$$\begin{aligned} U_d \sigma_x U_d^{-1} =& s_x \sigma_x \\ U_d \sigma_y U_d^{-1} =& s_y \sigma_y \\ U_d \sigma_z U_d^{-1} =& s_z \sigma_z \end{aligned} \tag{6.2.79}$$

where $s_i = \pm 1$, $i = x, y, z$, to ensure the involution nature of the operation. Moreover, since the triad must remain right handed we need that $s_x s_y s_z = 1$. Then, apart from the trivial case $s_x = s_y = s_z = 1$, there are three possibilities. First, the case $s_y = 1$ and $s_x = s_z = -1$ that corresponds to the canonical situation, since Eqs.(6.2.79) become

$$\begin{aligned} U_d \sigma_x U_d^{-1} =& -\sigma_x \\ U_d \sigma_y U_d^{-1} =& \sigma_y \\ U_d \sigma_z U_d^{-1} =& -\sigma_z \end{aligned} \tag{6.2.80}$$

whose solution is $U_d = \sigma_y$. The second case is $s_x = 1$ and $s_y = s_z = -1$, which leads to $U_d = \sigma_x$. The third case, $U_d = \sigma_z$, holds for the $s_z = 1$ and $s_x = s_y = -1$. This leads to $\mathcal{T}^2 = U_d K U_d K = \pm I$ as requested. Regarding $U_d = \sigma_x$ and $U_d = \sigma_z$ the action of K is irrelevant since these matrices are real, hence:

$$\mathcal{T}^2 = \sigma_x^2 = \sigma_z^2 = I \tag{6.2.81}$$

Lastly, the case $U_d = \sigma_y$, gives $\mathcal{T}^2 = -\sigma_y^2 = -I$.
Concerning transformations that permute the order of the basis elements, and so of the Pauli matrices, we have the following possibilities: if we want to permute σ_x and σ_y we take either:

$$\begin{aligned} U_{xy}^{(1)} \sigma_x (U_{xy}^{(1)})^{-1} =& \mp \sigma_y \\ U_{xy}^{(1)} \sigma_y (U_{xy}^{(1)})^{-1} =& \pm \sigma_x \\ U_{xy}^{(1)} \sigma_z (U_{xy}^{(1)})^{-1} =& \sigma_z \end{aligned} \tag{6.2.82}$$

or

$$\begin{aligned} U_{xy}^{(2)} \sigma_x (U_{xy}^{(2)})^{-1} =& \pm \sigma_y \\ U_{xy}^{(2)} \sigma_y (U_{xy}^{(2)})^{-1} =& \pm \sigma_x \\ U_{xy}^{(2)} \sigma_z (U_{xy}^{(2)})^{-1} =& -\sigma_z \end{aligned} \tag{6.2.83}$$

To permute σ_y and σ_z we take either:

$$\begin{aligned} U_{yz}^{(1)} \sigma_x (U_{yz}^{(1)})^{-1} =& \sigma_x \\ U_{yz}^{(1)} \sigma_y (U_{yz}^{(1)})^{-1} =& \mp \sigma_z \\ U_{yz}^{(1)} \sigma_z (U_{yz}^{(1)})^{-1} =& \pm \sigma_y \end{aligned} \tag{6.2.84}$$

or

$$\begin{aligned} U^{(2)}_{yz}\sigma_x(U^{(2)}_{yz})^{-1} &= -\,\sigma_x \\ U^{(2)}_{yz}\sigma_y(U^{(2)}_{yz})^{-1} &= \pm\,\sigma_z \\ U^{(2)}_{yz}\sigma_z(U^{(2)}_{yz})^{-1} &= \pm\,\sigma_y \end{aligned} \tag{6.2.85}$$

To permute σ_x and σ_z we take either:

$$\begin{aligned} U^{(1)}_{xz}\sigma_x(U^{(1)}_{xz})^{-1} &= \mp\,\sigma_z \\ U^{(1)}_{xz}\sigma_y(U^{(1)}_{xz})^{-1} &= \sigma_y \\ U^{(1)}_{xz}\sigma_z(U^{(1)}_{xz})^{-1} &= \pm\,\sigma_x \end{aligned} \tag{6.2.86}$$

or

$$\begin{aligned} U^{(2)}_{xz}\sigma_x(U^{(1)}_{xz})^{-1} &= \pm\,\sigma_z \\ U^{(2)}_{xz}\sigma_y(U^{(1)}_{xz})^{-1} &= -\,\sigma_y \\ U^{(2)}_{xz}\sigma_z(U^{(1)}_{xz})^{-1} &= \pm\,\sigma_x \end{aligned} \tag{6.2.87}$$

Here, we required that the triad remains right handed, so the representation on the associated vectors $\boldsymbol{e}_i$ (see Theorem 6.2.4) is a matrix of $SO(3)$. For instance, Eqs.(6.2.82) are associated with a matrix

$$O = \begin{pmatrix} 0 & \mp 1 & 0 \\ \pm 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{6.2.88}$$

that is special orthogonal. In fact, it is the linear application of $SO(3)$ that maps $(\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z)$ to $(\mp\boldsymbol{e}_y, \pm\boldsymbol{e}_x, \boldsymbol{e}_z)$. Now, use Theorem 6.2.5. Because U_{xy} must be unitary, we express it as the following linear combination:

$$U_{xy} = \alpha\sigma_x + \beta\sigma_y + \gamma\sigma_z + \delta I \qquad \alpha, \beta, \gamma, \delta \in \mathbb{C} \tag{6.2.89}$$

whose coefficients obey Eq. (6.2.72). Substituting (6.2.89) in the third equation of Eqs.(6.2.82), we obtain the condition:

$$\alpha\sigma_x\sigma_z + \beta\sigma_y\sigma_z + \gamma I + \delta\sigma_z = \alpha\sigma_z\sigma_x + \beta\sigma_z\sigma_y + \gamma I + \delta\sigma_z \tag{6.2.90}$$

and, recalling that the Pauli matrices satisfy the fundamental relation

$$\sigma_i\sigma_j = \delta_{ij}I + i\epsilon_{ijk}\sigma_k \tag{6.2.91}$$

Eq(6.2.90) writes

$$-i\alpha\sigma_y + i\beta\sigma_x + \gamma I + \delta\sigma_z = i\alpha\sigma_y - i\beta\sigma_x + \gamma I + \delta\sigma_z \tag{6.2.92}$$

Since the set $\{\sigma_i, I\}$ is a basis for the algebra $\mathfrak{u}(2)$, the coefficient of the linear combinations on both sides of Eq.(6.2.92) must coincide, which means $\alpha = \beta = 0$, and $U_{xy} = \gamma\sigma_z + \delta I$. Substituting in the first of Eqs.(6.2.82) one gets:

$$\gamma\sigma_z\sigma_x + \delta\sigma_x = \mp\gamma\sigma_y\sigma_z \mp \delta\sigma_y \tag{6.2.93}$$

that is

$$i\gamma\sigma_y + \delta\sigma_x = \mp i\gamma\sigma_x \mp \delta\sigma_y \tag{6.2.94}$$

which implies $\delta = \mp i\gamma$. To ensure that the corresponding matrix

$$U_{xy}^{(1)} = \gamma\sigma_z \mp i\gamma I = \gamma(\sigma_z \mp iI) \tag{6.2.95}$$

is unitary, we need $UU^\dagger = I$, that is:

$$\gamma^2(\sigma_z \mp iI)(\sigma_z \pm iI) = 2\gamma^2 I = I \tag{6.2.96}$$

which means: $|\gamma| = 1/\sqrt{2}$. The last thing to check, for a time reversal operator, is that it is an involution, *i.e.* that $\mathcal{T}^2 = UKUK = \pm I$ holds. This is indeed the case:

$$U_{xy}^{(1)} K U_{xy}^{(1)} K = \gamma^2(\sigma_z \mp iI)(\sigma_z \pm iI) = I \tag{6.2.97}$$

The same reasoning can be repeated for the remaining permutations, which leads to:

$$\begin{aligned} U_{xy}^{(2)} &= \theta(\sigma_x \pm \sigma_y) \\ U_{yz}^{(1)} &= \theta(\sigma_x \mp iI) \\ U_{yz}^{(2)} &= \theta(\sigma_y \pm \sigma_z) \\ U_{xz}^{(1)} &= \theta(\sigma_y \pm iI) \\ U_{xz}^{(2)} &= \theta(\sigma_x \pm \sigma_z) \end{aligned} \tag{6.2.98}$$

where θ is a complex number with modulus $|\theta| = 1/\sqrt{2}$. One may easily check that $\mathcal{T}^2 = \pm I$ in all cases:

$$\begin{aligned} U_{xy}^{(2)} K U_{xy}^{(2)} K &= \theta^2(\sigma_x \pm \sigma_y)(\sigma_x \mp \sigma_y) = \mp i\sigma_z \neq \pm I \\ U_{yz}^{(1)} K U_{yz}^{(1)} K &= \theta^2(\sigma_x \mp iI)(\sigma_x \pm iI) = I \\ U_{yz}^{(2)} K U_{yz}^{(2)} K &= \theta^2(\sigma_y \pm \sigma_z)(-\sigma_y \pm \sigma_z) = \pm i\sigma_x \neq \pm I \\ U_{xz}^{(1)} K U_{xz}^{(1)} K &= \theta^2(\sigma_y \pm iI)(-\sigma_y \mp iI) = \mp 2i\sigma_y \neq \pm I \\ U_{xz}^{(2)} K U_{xz}^{(2)} K &= \theta^2(\sigma_x \pm \sigma_z)(\sigma_x \pm \sigma_z) = I \end{aligned} \tag{6.2.99}$$

The Theorem is proven, taking $U_{xy}^{(1)}$, $U_{yz}^{(1)}$ and $U_{xz}^{(2)}$ as the non diagonal time reversal operators on spin space. □

Now, swapping a pair of coordinates requires at most 2 matrices. The total, considering also U_d, makes 9 possible matrices U_s. Because, each of them is defined up to a complex number of modulus 1, it seems that infinitely many are allowed. But whenever U_s is applied to an operator on the spin space, giving $U_s f(\sigma_i) U_s^{-1}$, that number is eliminated. Therefore, there are 9 operators acting differently on Pauli matrices.
This exhausts our present investigation of generalized time reversal invariance, for quantum nonrelativistic systems, with and without half integer spin. In the following, we apply our results to the theory of the Onsager relations, looking for generalizations of the compatibility conditions found in Ref.[15], between time reversal transformations and a generic magnetic field.

6.2.4 Compatibility condition between time reversal operations and magnetic field

The reasoning developed by Kubo about correlation functions and the entailing transport coefficients [4], requires the existence of a time reversal operation, that commutes with the Hamiltonian of the system of interest. Because the quantum observables are defined as hermitian operators, one cannot follow the classical procedure to derive correlator relations. Given two observables ϕ and ψ, the mean value $\langle\phi(0)\psi(t)\rangle$ is ill-defined, as the combined operator $\phi(0)\psi(t)$ is not guaranteed to be hermitian. Therefore, Kubo introduced the *canonical* correlator in Response Theory as:

$$\langle\phi(0);\psi(t)\rangle = \frac{1}{\beta}\int_0^\beta d\lambda \;\; \mathbf{Tr}[\rho e^{\lambda H}\phi(0)e^{-\lambda H}\psi(t)] \tag{6.2.100}$$

where ρ is the density operator, H the hamiltonian, and **Tr** the trace over the Hilbert. Kubo then proved that this is indeed a real quantity. Note that nonrelativistic systems with spin are allowed: it suffices to include the spin degrees of freedom within the Hilbert space. Another possible definition is to symmetrize the correlator as $\langle\phi(0),\psi(t)\rangle = \langle\phi(0)\psi(t)\rangle + \langle\psi(t)\phi(0)\rangle$, but Kubo showed that 6.2.100 is sufficiently general. The sufficient condition for the validity of the Onsager relations can then be formulated as *e.g.* Ref.[10]:

Proposition 6.2.6. *Consider a quantum mechanical particle system in a magnetic field* $\mathbf{B}$*, with Hamiltonian* H *and density matrix* ρ*. Let* $\mathcal{T} = UK$ *be a time reversal operator that commutes with* H*, and let* ϕ *and* ψ *be two observables with signatures* η_ϕ *and* η_ψ *with respect to* $\mathcal{T}$*,* i.e

$$\mathcal{T}\phi\mathcal{T}^{-1} = \eta_\phi\phi \qquad \mathcal{T}\psi\mathcal{T}^{-1} = \eta_\psi\psi \tag{6.2.101}$$

Then, the equality:

$$\langle\phi(0);\psi(t)\rangle_{\mathbf{B}} = \eta_\phi\eta_\psi\langle\phi(t);\psi(0)\rangle_{\mathbf{B}} \tag{6.2.102}$$

holds.

This is the core of the Onsager reciprocal relations in Quantum mechanics in the presence of a magnetic field. Therefore, we now investigate how particular magnetic fields affect the number of time reversal operations consistent with Eq.(6.2.102). Consider the following Hamiltonian for a system of spinless particles coupled with a potential vector $\mathbf{A}$:

$$H = \sum_{i=1}^{N}\left[\frac{(\mathbf{p}_i - q_i\mathbf{A}(x_i,y_i,z_i))^2}{2m_i}\right] \tag{6.2.103}$$

where N is the number of particles, q_i and m_i are the charge and the mass of the i-th particle, and

$$(\mathbf{p} - q\mathbf{A})^2 = -\nabla^2 + iq\nabla\cdot\mathbf{A} + iq\mathbf{A}\cdot\nabla + q^2\mathbf{A}^2 \tag{6.2.104}$$

The commonly used Coulomb gauge $\nabla\cdot\boldsymbol{A} = 0$ that importantly effects on this expression eliminating the second addend. As well-known, the vector potential associated to

a magnetic field is defined up to the gradient of a scalar function since $\boldsymbol{B} = \nabla \times \boldsymbol{A}$. Thus, we can define an equivalence class containing the vector potentials that originate the same magnetic field. In the following, we refer to $[\boldsymbol{A}(\boldsymbol{x})]_R$ to denote a representative of the class containing $\boldsymbol{A}(\boldsymbol{x})$.
After this clarifications, let us step back to the classical case: we now show the core theorem that regards the compatibility between a particular time reversal operation and a generic magnetic field (and so vector potential). We point out as in the following we are going to refer to $V(\mathbb{R}^n)$ as the space of vector fields on $\mathbb{R}^n$.

Theorem 6.2.8 (Compatibility conditions for $\boldsymbol{A}$ in Classical Mechanics)**.** *Consider a system of N particles of equal mass m and charge q. Let $\mathcal{M}$ be a generalized time reversal operator acting on the coordinates as $(\mathcal{M}A)(\boldsymbol{X}) = A(\mathcal{M}_M \boldsymbol{X})$, where $\mathcal{M}_M \in O(3N)$. Denote by $\boldsymbol{\mathcal{A}}(\boldsymbol{X}) = (\boldsymbol{A}(\boldsymbol{x}_1), ..., \boldsymbol{A}(\boldsymbol{x}_N))$ the $3N$ dimensional vectors of coordinates, and by $[\boldsymbol{\mathcal{A}}(\boldsymbol{X})]_R = ([\boldsymbol{A}(\boldsymbol{x}_1)]_R, ..., [\boldsymbol{A}(\boldsymbol{x}_N)]_R)$ the corresponding equivalence class. Introduce the reversal operator $\mathcal{M}' : V(\mathbb{R}^{3N}) \to V(\mathbb{R}^{3N})$ defined by:*

$$(\mathcal{M}'\boldsymbol{\mathcal{A}})(\boldsymbol{X}) \equiv \mathcal{M}_M \boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X}) \tag{6.2.105}$$

The operator $\mathcal{M}$ yields TRI in the presence of a magnetic vector potential $\boldsymbol{A}$, if and only if

$$(\mathcal{M}'\boldsymbol{\mathcal{A}})(\boldsymbol{X}) = [-\boldsymbol{\mathcal{A}}(\boldsymbol{X})]_R \tag{6.2.106}$$

Proof. Recall that TRI holds if there is a time reversal transformation that preserves the equations of motion, as well as the Hamiltonian up to a gauge transformation. In our case, the equations of motion are invariant under a gauge transformation, as can be easily seen. Then, in the presence of the minimal coupling with a magnetic field, the condition $H(\mathcal{M}\Gamma) = H(\Gamma)$ is verified for a gauge transformation, for any Γ. Therefore, the Hamiltonian can be written as

$$H(\boldsymbol{X}, \boldsymbol{P}) = \frac{1}{2m}\sum_i [\boldsymbol{p}_i - q\boldsymbol{A}(\boldsymbol{x}_i)]^2 = \frac{1}{2m}[\boldsymbol{P} - q\boldsymbol{\mathcal{A}}(\boldsymbol{X})]^2 = \frac{1}{2m}[\boldsymbol{P} + q\mathcal{M}_M \boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X})] \cdot [\boldsymbol{P} + q\mathcal{M}_M \boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X}) \tag{6.2.107}$$

where $\boldsymbol{P}$ and $\mathcal{A}$ are $3N$-dimensional vectors. By Theorem 6.2.1, a time reversal operation for systems subject to a magnetic field must act separately on coordinates and momenta, namely the transformation must be expressed by:

$$(\boldsymbol{X}, \boldsymbol{P}) \xrightarrow{\mathcal{M}} (\mathcal{M}_M \boldsymbol{X}, -\mathcal{M}_M \boldsymbol{P}) \tag{6.2.108}$$

with $\mathcal{M}_M \in O(3N)$. This yields

$$\begin{aligned} H(\mathcal{M}_M \boldsymbol{X}, -\mathcal{M}_M \boldsymbol{P}) &= \frac{1}{2m}[-\mathcal{M}_M \boldsymbol{P} - q\boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X})]^2 \\ &= \frac{1}{2m}[\boldsymbol{P} + q\mathcal{M}_M \boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X})]^2 \end{aligned} \tag{6.2.109}$$

where we used the fact that the scalar product is invariant under rotations such as $\mathcal{M}_M$. Now, let $\mathcal{G}$ be the scalar function involved in the gauge choice, define $\nabla_{3N}\mathcal{G}(\boldsymbol{X}) =$

$(\nabla G(\boldsymbol{x}_1), ..., \nabla G(\boldsymbol{x}_N))$, and let us introduce the equivalence class of vector potentials defined by:

$$[\boldsymbol{A}] = \{\boldsymbol{A} \ | \ \nabla \times \boldsymbol{A} = \boldsymbol{B}\} \tag{6.2.110}$$

This is the set of vector potentials corresponding to a given magnetic field $\boldsymbol{B}$. One representative element of this class, denoted by a subscript R, is expressed by:

$$[\boldsymbol{\mathcal{A}}(\boldsymbol{X})]_R = ([\boldsymbol{A}(\boldsymbol{x}_1)]_R, ..., [\boldsymbol{A}(\boldsymbol{x}_N)]_R) \tag{6.2.111}$$

Now, taking $\mathcal{M}_M \boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X}) = -(\boldsymbol{\mathcal{A}}(\boldsymbol{X}) + \nabla_{3N} \mathcal{G}(\boldsymbol{X}))$, substitution shows that (6.2.107) and (6.2.109) are equal, up to a gauge transformation. The transformation leaves the equations of motion unchanged. If, on the other hand, (6.2.107) and (6.2.109) are equivalent for any value of $\boldsymbol{P}$ and $\boldsymbol{X}$, that is:

$$\begin{aligned} \boldsymbol{P}^2 + 2q\boldsymbol{P} \cdot (\boldsymbol{\mathcal{A}}(\boldsymbol{X}) + \nabla_{3N} \mathcal{G}(\boldsymbol{X})) + q^2 (\boldsymbol{\mathcal{A}}(\boldsymbol{X}) + \nabla_{3N} \mathcal{G}(\boldsymbol{X}))^2 \\ = \boldsymbol{P}^2 - 2q\boldsymbol{P} \cdot \mathcal{M}_M \boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X}) + q^2 (\boldsymbol{\mathcal{M}_M \mathcal{A}}(\mathcal{M}_M \boldsymbol{X}))^2 , \quad \forall \boldsymbol{P} \end{aligned} \tag{6.2.112}$$

we can choose $\boldsymbol{P} = 0$ obtaining

$$q^2 (\boldsymbol{\mathcal{A}}(\boldsymbol{X}) + \nabla_{3N} \mathcal{G}(\boldsymbol{X}))^2 = q^2 (\boldsymbol{\mathcal{M}_M \mathcal{A}}(\mathcal{M}_M \boldsymbol{X}))^2 \tag{6.2.113}$$

Then, Eq.(6.2.112) reduces to

$$2q\boldsymbol{P} \cdot (\boldsymbol{\mathcal{A}}(\boldsymbol{X}) + \nabla_{3N} \mathcal{G}(\boldsymbol{X})) = -2q\boldsymbol{P} \cdot \mathcal{M}_M \boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X}) \tag{6.2.114}$$

Therefore, $\mathcal{M}_M \boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X}) = -(\boldsymbol{\mathcal{A}}(\boldsymbol{X}) + \nabla_{3N} \mathcal{G}(\boldsymbol{X}))$ holds. □

This result generalizes the compatibility conditions of Ref.[15], which focused on the 20 transformations that separately act on each particle subspace. The result of Ref.[15] can be interpreted as a corollary of Theorem 6.2.8.

Corollary 6.2.4. *Take a block diagonal transformation $\mathcal{M}_M$ separately acting on each particle subspace, denoting by $\mathcal{M}_m$ one such 3×3 block. Then*

$$\mathcal{M}' \boldsymbol{A} = \mathcal{M}_m \boldsymbol{A}(\mathcal{M}_m \boldsymbol{X}) = [-\boldsymbol{A}]_R \tag{6.2.115}$$

Proof. It trivially follows form the definition of $\boldsymbol{\mathcal{A}}$ given in Theorem 6.2.8, and from Eq.(6.2.106), where $\mathcal{M}_M$ is a block diagonal acting on a single 3-dimensional particle subspace. □

Proposition 6.2.7. *Take the Hamiltonian* (6.2.103) *in which $(m_i, q_i) \neq (m_j, q_j)$ if $i \neq j$ are particle indices. Then, a time reversal operator yielding TRI is given by a block diagonal matrix A, whose blocks separately act on single particle subspaces.*

Proof. By *reductio ad absurdum*, suppose a valid non diagonal time reversal operation $\mathcal{M}_{nd}$ exists for the system described by the Hamiltonian (6.2.103). Without loss of generality, assume that the operation acts on the $M \times M$ coordinates (with $M = 3N$) swapping those of particle j with those of particle $j+1$:

$$(\boldsymbol{x}_1, ..., \boldsymbol{x}_j, \boldsymbol{x}_{j+1}, ..., \boldsymbol{x}_N, \boldsymbol{p}_1, ..., \boldsymbol{p}_j, \boldsymbol{p}_{j+1}, ..., \boldsymbol{p}_N) \xrightarrow{\mathcal{M}_{nd}} (\boldsymbol{x}_1, ..., \boldsymbol{x}_{j+1}, \boldsymbol{x}_j, ..., \boldsymbol{x}_N, -\boldsymbol{p}_1, ..., -\boldsymbol{p}_{j+1}, -\boldsymbol{p}_j, . \tag{6.2.116}$$

By hypothesis, $H(\mathcal{M}_{nd}\Gamma) = H(\Gamma)$, because $\mathcal{M}_{nd}$ yields TRI. It suffices to check the contributions due to particles j and $j+1$, because all other contributions are unchanged. Thus we can write:

$$\sum_{i=1}^{N}\left[\frac{(\boldsymbol{p}_i - q_i\boldsymbol{A}(\boldsymbol{x}_i))^2}{2m_i}\right] = ... + \frac{(\boldsymbol{p}_j - q_j\boldsymbol{A}(\boldsymbol{x}_j))^2}{2m_j} + \frac{(\boldsymbol{p}_{j+1} - q_{j+1}\boldsymbol{A}(\boldsymbol{x}_{j+1}))^2}{2m_{j+1}} \tag{6.2.117}$$

$$... + \frac{(\boldsymbol{p}_{j+1} + q_j\boldsymbol{A}(\boldsymbol{x}_{j+1}))^2}{2m_j} + \frac{(\boldsymbol{p}_j + q_{j+1}\boldsymbol{A}(\boldsymbol{x}_j))^2}{2m_{j+1}} \tag{6.2.118}$$

But $m_i \neq m_j$ or $q_i \neq q_j$ for $\forall i \neq j$ by hypothesis. Then, in general, one has $H(\mathcal{M}_{nd}\Gamma) \neq H(\Gamma)$, unless special cases are considered, because the coordinates and the momenta may take any value in $\mathbb{R}^3$. This is absurd. As the reasoning can be repeated varying i and j in (6.2.116), $\mathcal{M}_{nd}$ cannot mix the different single particles coordinates and momenta; it must be block diagonal, with each block acting on a single particle subspace. □

This result may seem to be a drastic limitation on the range of TRI, but in reality it is not. Systems with particles of same charge and mass are the most widely studied, both theoretically and experimentally.

In Ref.[15], the compatibility conditions is expressed also in terms of the magnetic field, instead of the vector potential. Here, we present an alternative derivation of that result result. We point out as in the following we are going to refer to $V_p(\mathbb{R}^n)$ as the space of pseudovector fields on $\mathbb{R}^n$.

Theorem 6.2.9 (Restricted compatibility conditions for $\boldsymbol{B}$ in Classical Mechanics)**.** *Consider a system of N particles of equal mass m and charge q. Let $\mathcal{M}$ be a generalized time reversal operator acting on the coordinates as $\mathcal{M}\boldsymbol{X} = (\mathcal{M}_m\boldsymbol{x}_1, ..., \mathcal{M}_m\boldsymbol{x}_N)$, where $\mathcal{M}_m \in O(3)$. Introduce the pseudovector field rotation operator $\mathcal{M}' : V_p(\mathbb{R}^3) \to V_p(\mathbb{R}^3)$ defined by:*

$$(\mathcal{M}'\boldsymbol{B})(\boldsymbol{x}) = \det(\mathcal{M}_m)\mathcal{M}_m\boldsymbol{B}(\mathcal{M}_m\boldsymbol{x}) \tag{6.2.119}$$

The operator $\mathcal{M}$ yields TRI in the presence of a magnetic field $\boldsymbol{B}$, if and only if

$$(\mathcal{M}'\boldsymbol{B})(\boldsymbol{x}) = -\boldsymbol{B}(\boldsymbol{x}) \tag{6.2.120}$$

Proof. We have to prove that (6.2.115) and (6.2.120) are equivalent. Let us introduce the notation:

$$\nabla_{\boldsymbol{x}} \equiv (\partial_x, \partial_y, \partial_z) \tag{6.2.121}$$

hence $\nabla_{\mathcal{M}_m\boldsymbol{x}}$ is the gradient with respect to the coordinates rotated by $\mathcal{M}_m$. Then, keeping in mind that $\boldsymbol{B} = \nabla \times \boldsymbol{A}$, the application of the transformation $\mathcal{M}'$ defined by the orthogonal matrix $\mathcal{M}_m$ yields [24]:

$$\begin{aligned}\mathcal{M}'(\nabla_{\boldsymbol{x}} \times \boldsymbol{A}) &= (\mathcal{M}_m\nabla_{\mathcal{M}_m\boldsymbol{x}}) \times (\mathcal{M}_m\boldsymbol{A} \circ \mathcal{M}_m) \\ &= \det(\mathcal{M}_m)\mathcal{M}_m(\nabla_{\mathcal{M}_m\boldsymbol{x}} \times \boldsymbol{A} \circ \mathcal{M}_m) \\ &= \det(\mathcal{M}_m)\mathcal{M}_m\boldsymbol{B} \circ \mathcal{M}_m\end{aligned} \tag{6.2.122}$$

because an orthogonal matrix such as $\mathcal{M}_m$ yields:

$$\nabla_{\mathcal{M}_m \boldsymbol{x}} \times \boldsymbol{A} \circ \mathcal{M}_m = \nabla \times \boldsymbol{A} = \boldsymbol{B}\,, \quad \mathcal{M}_m \nabla_{\mathcal{M}_m \boldsymbol{x}} = \nabla_{\boldsymbol{x}} \tag{6.2.123}$$

Now, assuming Eq.(6.2.115) holds, we have:

$$\det(\mathcal{M}_m)\mathcal{M}_m \boldsymbol{B} \circ \mathcal{M}_m = \nabla_{\boldsymbol{x}} \times [-\boldsymbol{A}]_R = -\boldsymbol{B} \tag{6.2.124}$$

Vice versa, assuming Eq.(6.2.120) holds, we start from

$$\det(\mathcal{M}_m)\mathcal{M}_m \nabla_{\mathcal{M}_m \boldsymbol{x}} \times \boldsymbol{A} \circ \mathcal{M}_m = -\nabla_{\boldsymbol{x}} \times \boldsymbol{A} \tag{6.2.125}$$

Moreover, using the first and second lines of (6.2.122), we get

$$\nabla_{\boldsymbol{x}} \times (\mathcal{M}_m \boldsymbol{A} \circ \mathcal{M}_m) = -\nabla_{\boldsymbol{x}} \times \boldsymbol{A} \tag{6.2.126}$$

and so, considering the fact that the rotor of a gradient is null, we obtain the thesis (6.2.115), where gauge freedom is included. □

In the presence of magnetic fields, Theorem 6.2.8 seems to directly apply only in the case of the block diagonal time reversal operations. Unlike Theorem 6.2.8, Theorem 6.2.9 cannot be immediately generalised to the cases with a magnetic field, because in $\mathbb{R}^{3N}$ we miss the analogue of the relation $\boldsymbol{A} = \nabla \times \boldsymbol{B}$, which holds in $\mathbb{R}^3$. Operators similar to the curl can be defined in spaces other than $\mathbb{R}^3$, see for example Ref.[25], but we cannot claim that $\boldsymbol{\mathcal{A}} = \nabla \times \mathcal{B}$, for $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{3N}$. Finally coming to the quantum context, the following holds.

Proposition 6.2.8. *The statements of Theorems 6.2.8 and 6.2.9 extend to Quantum Mechanics on compact spaces, when $\boldsymbol{X}$ is identified with the position operator, cf. Theorem 6.2.3.*

Proof. Let us start from Theorem 6.2.8: regarding (6.2.107) and (6.2.108) the only differences are that $\boldsymbol{P} = -i\nabla$ and that the case $\mathcal{M}_M^2 = -I$ is admitted. The proof can be repeated until (6.2.112), since using Coulomb gauge and (6.2.104) we do not have the 2 in front of the mixed product. As a side note, the Coulomb gauge condition involves a scalar product too and so it invariant under rotation.
At this point, since we are using operators, we cannot impose $\boldsymbol{P} = 0$ as done in Classical Mechanics. Nevertheless, (6.2.112) is an equivalence between operators and so it must holds for any wave function defined on a certain domain. If the domain is compact, we can consider the case of the constant without issues of normalization; the application to it of the differential operator $\boldsymbol{P}$ yields the constraint

$$q^2(\boldsymbol{\mathcal{A}}(\boldsymbol{X}) + \nabla_{3N}\mathcal{G}(\boldsymbol{X}))^2 = q^2(\boldsymbol{\mathcal{M}_M}\boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X}))^2 \tag{6.2.127}$$

Finally, we obtain an expression similar to (6.2.114):

$$(\boldsymbol{\mathcal{A}}(\boldsymbol{X}) + \nabla_{3N}\mathcal{G}(\boldsymbol{X})) \cdot \boldsymbol{P} = -\mathcal{M}_M \boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X}) \cdot \boldsymbol{P} \tag{6.2.128}$$

where we used the correct ordering of the momentum operators. Also in this case the only way to verify this operator equality is to have $\mathcal{M}_M \boldsymbol{\mathcal{A}}(\mathcal{M}_M \boldsymbol{X}) = [-\boldsymbol{\mathcal{A}}]_R$.

Coming to Theorem 6.2.9, its extension is trivial since the proof involves only functions of the coordinate operator which acts in a multiplicative way in coordinate representation; in this way, we can proceed as if we were manipulating numbers and not operators, easily reproducing the demonstration. □

This is our main result for the spinless case. Two observations and a fundamental example are in order.

Remark 6.2.1. The compatibility condition on magnetic fields of Theorem 6.2.9 seems to be of limited applicability, compared to the one about vector potentials, since it is limited to transformations separately acting on single particle subspaces. Nevertheless, that is the most interesting case, since particles usually interact with each other. For instance, in the case of the central interaction potential:

$$V = \sum_{i<j} v(\boldsymbol{x}_i - \boldsymbol{x}_j) \tag{6.2.129}$$

with given pair potential $v(\boldsymbol{x}) = v(|\boldsymbol{x}|)$, treated in Ref.[15], the time reversal operations must be block diagonal. It must separately and identically act on each particle subspace, otherwise $v(\boldsymbol{x}_i - \boldsymbol{x}_j)$ would not be preserved, in general. If the time reversal operation is such a block diagonal operator Q, it can be combined with the operator P_{ij} that swaps particle i with particle j, obtaining a new involution QP_{ij}: $QP_{ij}QP_{ij} = I$. The same applies to the case of half integer spin particles, and it is not necessary to repeat the reasoning. Obviously, an interaction term, such as (6.2.129), prevents the use of the time reversal operations of Proposition 6.2.5, that cannot be expressed by 3×3 block diagonal matrices. Thus, in the following we neglect the case with $Q^2 = -I$.

Remark 6.2.2. The action of the operator $\mathcal{T}$ of Proposition 6.2.6 deeply differs from that of the traditional one, which includes the inversion of the magnetic field, $\mathcal{T}_B$, cf. Eq.(6.1.1). Nevertheless, $\mathcal{T}$ suffices for the validity of the Onsager relations, because they arise from a phase space integration and, pairing differently the contributions to the correlator, the integral may still yield the same result. Statistical mechanical relations are typically of this kind. Therefore, they do not require microreversibility. Weaker conditions suffice, such as those discussed here, that we may call *statistical time reversibility*, cf. also Refs.[26–28]. This equally applies to the quantum case, if trajectories are replaced by the time evolution of states in the Hilbert space. In fact, the invariance of the Hamiltonian and of the equation of motion ensures that

$$|\psi'(t')\rangle = \mathcal{T}\,|\psi(t')\rangle\ , \quad \text{with} \quad t' = -t$$

is a physical state if $|\psi(t)\rangle$ is. Roughly speaking, in the quantum case it is just a matter of rearranging the contributions to the trace in (6.2.100).

Let us reconsider, in the light of the above result, the case of a constant magnetic field.

Theorem 6.2.10. *Take a system of particles interacting via the central potential* (6.2.129). *Let this system be subjected to a constant external magnetic field,* $\boldsymbol{B} =$

(B_1, B_2, B_3). *The dynamics is invariant under infinitely many time reversal operations: those whose action on each single particle subspace is represented by:*

$$M = \begin{pmatrix} a & b & 0 \\ b & -a & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{6.2.130}$$

where $a, b \in \mathbb{R}$, $a^2 + b^2 = 1$ *and* $\det(M) = -1$.

Proof. The time evolution of the system does not depend on the coordinates frame, so we can choose the axis z along the direction of $\boldsymbol{B}$, and write $B = (0, 0, 1)$, up to a dimensional constant. Consider the time reversal transformations whose action on each single particle subspace is represented by Eq.(6.2.130). By definition, $M \in O(3)$ and $M^2 = I$, therefore M represents the action on coordinates of a time reversal operation, as follows from Theorem 6.2.1. Furthermore, the compatibility condition of Theorem 6.2.9 is trivially verified:

$$\det(M) M\boldsymbol{B} = (0, 0, -1) = -\boldsymbol{B} \tag{6.2.131}$$

Taking $a = \cos\theta$ and $b = \sin\theta$, with $\theta \in [0, 2\pi)$, infinitely many possible choices are allowed for M. □

This result seems to boldly contradict the traditional opinion that any magnetic field breaks TRI. With hindsight, it is not so surprising, when the existence of generalized time reversal symmetries has been ascertained. In particular, a constant magnetic field directed along along the z axis preserves all the symmetries that differ by rotations of the xy plane, if it preserves one of them.
In the case of particles with spin, the Pauli Hamiltonian (6.2.63) contains the minimal coupling term, for which Proposition 6.2.8 applies. However, magnetic field and spins are also coupled, with the spins belonging to a different space. Then, a time reversal operation ought to take the form:

$$\mathcal{T} = (U_c \otimes U_s^1 \otimes ... \otimes U_s^N)K \tag{6.2.132}$$

where U_c acts on coordinate and momentum operators, while U_s^i acts on the spin operator of particle i. The situation differs from that considered in Theorem 6.2.8, because we now have:

$$\mathcal{T} H \mathcal{T}^{-1} \propto \boldsymbol{B}(U_c \boldsymbol{x}_i U_c^{-1}) \cdot U_s^i K \boldsymbol{\sigma}_i K U_s^i \tag{6.2.133}$$

where H is the Hamiltonian and $\boldsymbol{B}$ the magnetic field. This way, the minimal coupling and the spin couplings with $\boldsymbol{B}$ are considered independently of each other: a time reversal operation compatible with the minimal coupling leaves the spin field coupling unchanged.

Remark 6.2.3. *The transformation U_c acts on coordinates and not on $\boldsymbol{B}$ as a vector field, that is the components of $\boldsymbol{B}$ transform as*

$$U_c B_i U_c^{-1} \equiv B_i(U_c \boldsymbol{x}_i U_c^{-1}) \tag{6.2.134}$$

Note that statistical relations could be used to distinguish systems whose particles possess spin or from those which do not, if the extra conditions implied by the presence of spin reduced the number of suitable time reversal operators. In fact, for transformations like (6.2.132), and with block diagonal U_c separately acting on single particles subspaces, this is not possible.

Theorem 6.2.11. *The compatibility condition of Theorem 6.2.9 suffices for TRI to hold also in case of particles with spin that obey the Pauli equation* (6.2.63).

Proof. Write the Pauli hamiltonian as:

$$H = H_{mc} + H_{sc} \tag{6.2.135}$$

where H_{mc} refers to the minimal coupling, of the spinless case, and $H_{sc} = \sum_{j=1}^{N} \boldsymbol{\sigma}_j \cdot \boldsymbol{B}(\boldsymbol{x}_j)$ is the spin-field coupling, with $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$. Given that we are analyzing transformations separately acting on single particle spaces, we study a single addend of the summation.
By hypothesis, there exists a time reversal operator $\mathcal{T} = (U_c \otimes U_s)K$ whose action on coordinates commutes with H_{mc}, *i.e.* Eq.(6.2.120) holds or, equivalently,

$$\boldsymbol{B}(\mathcal{M}_m \boldsymbol{x}) = -\frac{1}{\det(\mathcal{M}_m)} \mathcal{M}_m^{-1} \boldsymbol{B}(\boldsymbol{x}) \tag{6.2.136}$$

Because of Theorem 6.2.9, $\mathcal{M}_m$ is the 3-dimensional orthogonal and involutory matrix representation of the action of the time reversal operator on a single particle position operator. Under the action of U_c, one has:

$$U_c(\boldsymbol{\sigma} \cdot B(\boldsymbol{x}))U_c^{-1} = \boldsymbol{\sigma} \cdot (-P\boldsymbol{B}(\boldsymbol{x})) = -P\boldsymbol{\sigma} \cdot \boldsymbol{B}(\boldsymbol{x}) \tag{6.2.137}$$

where $P\boldsymbol{\sigma} = P_i^j \sigma_j$, and

$$P = \frac{1}{\det(\mathcal{M}_m)} \mathcal{M}_m^{-1} \tag{6.2.138}$$

which is a special orthogonal matrix. By Theorem 6.2.6, there exists a special unitary matrix U such that

$$U\sigma_i U^{-1} = P_i^j \sigma_j \tag{6.2.139}$$

Then, because $P^2 = I$ by definition of time reversal operation, we have $U^2 = \pm I$, since the isomorphism maps $\pm I \in SU(2)$ to $I \in SO(3)$. Moreover, letting K be the complex conjugation operator, we have

$$\sigma_y K \sigma_i K \sigma_y = -\sigma_i \tag{6.2.140}$$

where $i, j = 1, 2, 3 = x, y, z$. Now, take $U_s = U\sigma_y$. By definition, the time reversal operator acts on spin operators as

$$\mathcal{T}\sigma_i\mathcal{T}^{-1} = U\sigma_y K \sigma_i K \sigma_y U^{-1} = -P_i^j \sigma_j \tag{6.2.141}$$

where we used the decomposition of the Hilbert space of the system as the direct product of the spin space and of the coordinate space. This allows us to separately change the coordinate system in each space, and to write:

$$\sum_{k=1}^{N} \boldsymbol{\sigma}_k \cdot \boldsymbol{B}(\boldsymbol{x}_k) = \sum_{k=1}^{N} \Big[\boldsymbol{\sigma}_x \otimes \boldsymbol{B}_1(\boldsymbol{x}_k) + \boldsymbol{\sigma}_y \otimes \boldsymbol{B}_2(\boldsymbol{x}_k) + \boldsymbol{\sigma}_z \otimes \boldsymbol{B}_3(\boldsymbol{x}_k) \Big] \tag{6.2.142}$$

As the choice of the z axis in the spin space is not related to that of the z axis in coordinate space, U_s does not depend on the choice of U_c. This closes the circle: for any valid U_c, we can always find U, and so U_s, in order to keep H_{sc} invariant. In particular, we obtain:

$$\mathcal{T}\boldsymbol{\sigma} \cdot B(\boldsymbol{x})\mathcal{T}^{-1} = -P\boldsymbol{\sigma} \cdot -P\boldsymbol{B}(\boldsymbol{x}) = \boldsymbol{\sigma} \cdot \boldsymbol{B}(\boldsymbol{x}) \tag{6.2.143}$$

□

Lemma 6.2.2. *The Hamiltonian is not invariant under the action of $U_s = \sigma_y$ if the magnetic field is constant. To be recover the original Hamiltonian, the magnetic field must be manually inverted.*

Proof. The choice $U_s = \sigma_y$ corresponds to $U = I$, that is $P = I$ and $\mathcal{M}_m = \pm I$. This yields to two constraints on the magnetic field: $\boldsymbol{B}(\boldsymbol{x}) = -\boldsymbol{B}(\boldsymbol{x})$ choosing the plus and $\boldsymbol{B}(-\boldsymbol{x}) = -\boldsymbol{B}(\boldsymbol{x})$ with the minus, both absurd if $\boldsymbol{B}(\boldsymbol{x}) = \boldsymbol{B}$ is constant. But inverting $\boldsymbol{B}$ obviously restores the original Hamiltonian. □

Lemma 6.2.3. *The operator $U_s = U\sigma_y$ is such that*

$$U_s K U_s K = \pm I \tag{6.2.144}$$

Proof. By direct computation

$$U_s K U_s K = U\sigma_y K U \sigma_y K = U\sigma_y U^* \sigma_y^* = -U\sigma_y U^* \sigma_y = -U^2 \tag{6.2.145}$$

where we used (6.2.73). This complete the proof. □

In particular, taking $U = I$ in order to lie in the case of Lemma 6.2.2, we find the usual relation $\sigma_y K \sigma_y K = -I$.
This means that a class of time reversal operations, the ones separately acting on the single particle subspaces, can *always* be absorbed by an internal change of coordinates in spin space. Consequently, one finds that, *contrary* to what is commonly believed, the presence of spin does not prevent time reversibility.

6.2.5 Application of generalized TRI

Let us consider the calculation of Ref.[9], concerning the diffusion tensor for a particle system in the presence of a constant magnetic field along the z axis:

$$D_{\alpha\beta} = \langle v_i^\alpha(0) v_j^\beta(t) \rangle \quad \forall i, j\,; \quad \alpha, \beta = x, y, z \tag{6.2.146}$$

where $\alpha, \beta = x, y, z$ and i, j are particle labels. In particular, Ref.[8] proved that the correlator of velocities (6.2.146) vanishes if there are two time reversal operations that map $p_j^\beta(t)$ respectively in $p_k^\gamma(-t)$ and $-p_k^\gamma(-t)$ (the case with $i = j$ and $\alpha = \beta$ simultaneously being excluded), while they act in the same way on $p_i^\alpha(t)$. The proof simply observes that two expressions must be satisfied at once:

$$\begin{cases} \langle v_i^\alpha(0) v_j^\beta(t)\rangle = (\pm 1)(+1)\langle v_i^\alpha(0) v_k^\gamma(-t)\rangle \\ \langle v_i^\alpha(0) v_j^\beta(t)\rangle = (\pm 1)(-1)\langle v_i^\alpha(0) v_k^\gamma(-t)\rangle \end{cases} \tag{6.2.147}$$

which only happens if the correlator vanishes. The case $i = j$ and $\alpha = \beta$ is not included, since the autocorrelation $\langle v_i^\alpha(0) v_i^\alpha(t)\rangle$ is always mapped in another autocorrelation with plus sign. This example shows how practical the apparently abstract notion of TRI can be. In the present case, it allows a direct evaluation of transport coefficients. For the classical case, we may now extend the result of [9] considering a more general form of magnetic field.

Proposition 6.2.9. *Let a system of coupled particles be subjected to an external magnetic field directed along the z-axis,* $\boldsymbol{B} = B(x, y)\hat{\boldsymbol{k}}$. *Assume* $B(x, y) = B(x, -y)$ *and* $B(x, y) = B(y, x)$. *Then, the diffusion tensor obeys:*

$$D_{xy} = -D_{yx} \tag{6.2.148}$$

Proof. Notice first that Eq.(6.2.120) is verified by a magnetic field like the one in this Proposition. For example, consider the time reversal operation on the single particle subspace defined as

$$\mathcal{M}(x, y, z, p_x, p_y, p_z) = (y, x, z, p_y, p_x, p_z) \tag{6.2.149}$$

which implies

$$\mathcal{M}_m = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{6.2.150}$$

with the notation of Theorem 6.2.9. Equation (6.2.120) then holds for such a transformation and the magnetic field. Applying the same to all the particle subspaces and computing the correlators we obtain

$$\langle v_i^x(0) v_j^y(t)\rangle = \langle v_i^y(0) v_j^x(-t)\rangle = -\langle v_i^y(0) v_j^x(t)\rangle \tag{6.2.151}$$

that means $D_{xy} = -D_{yx}$. □

Examples of magnetic fields that satisfy Proposition 6.2.9 are $B(x, y) = const$ and $B(x, y) = B(x^2 + y^2)$, respectively the case of a constant magnetic field and of a magnetic field depending on the distance from the z axis. The result cannot be obtained in term of the diagonal operations only, because the diagonal operations cannot disentangle the pairs of subscripts and superscripts (i, x) and (j, y).

6.3 Results and Discussion

In this paper we extended the list of generalized time reversal transformations, to include operators that swap the coordinates of different particles. This is allowed by the fact that a point in a $6N$-dimensional space does not distinguish the nature of its single components. The formal definition of TRI does not prevent the use of this kind of operations. Their importance arises in particular when one studies systems of particles with the same mass and charge. These are important in the present context, since investigations of the Onsager relations rarely deal with more than two species of particles.
Moreover, we extended this treatment to the quantum mechanical framework, in the wake of the work of Ref.[10], about the 8 diagonal time reversal operations on the single particle subspace. In particular, we investigated the swap operation in a context of nonrelativistic Quantum Mechanics.
We then defined generalized TRI for systems of particles with spin 1/2, described by the Pauli equation. In previous works, this was considered out of reach [10], because spin was believed to irremediably break TRI. On the contrary, we found combined sufficient conditions, concerning the generalized time reversal transformations and the form of the magnetic field, for the validity of TRI. That allow us to derive Onsager relations, as well as other relations requiring TRI, to hold in quantum systems coupled with an external magnetic field.
This is interesting not only because of the experimental and theoretical relevance of such statistical mechanical relations, but also for the method used: we took full advantage of the fact that relations such as Onsager's are statistical relations. This allows many different paths to the same result and, in particular, microreversibility results unnecessary for Onsager and Fluctuation Relations.
We then developed an application to the calculation of the diffusion tensor, which uses different time reversal symmetries to conclude that certain correlators identically vanish.
The investigation, however, is not over. As recalled in Ref.[11], violations of Onsager relations, which would imply the existence non-dissipative currents, have never been observed. And indeed Ref.[11] proved that those relations hold in cases in which one would have expected they are violated. At the same time, or theory does not exhaust all possible cases.
Our results may also also pave the way to a new understanding of symmetries in quantum systems. The consequences of the generalized TRI have indeed rarely been fully investigated.

Bibliography

References for Chapter 1

[1] Pierre-Simon Laplace. *A philosophical essay on probabilities*. Courier Corporation, 2012.

[2] URL: https://www.linkedin.com/pulse/smartphone-today-has-more-computing-power-than-nasas-1960-offermann

[3] URL: https://www.britannica.com/science/history-of-science/Tycho-Kepler-and-Galileo

[4] Jim Al-Khalili. The birth of the electric machines: a commentary on Faraday (1832)'Experimental researches in electricity'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **373**: 20140208, 2015.

[5] URL: https://www.britannica.com/science/relativity/Intellectual-and-cultural-impact-of-relativity

[6] URL: https://medium.com/swlh/big-data-era-84b488491a8d

[7] Alexander L Fradkov. Early history of machine learning. *IFAC-PapersOnLine*, **53**: 1385–1390, 2020.

[8] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5**: 115–133, 1943.

[9] URL: https://www.techtarget.com/searchenterpriseai/definition/generative-modeling#:~:text=Generative%20modeling%20is%20the%20use,can%20be%20calculated%20from%20observations.

[10] URL: https://www.nvidia.com/en-us/glossary/generative-ai/

[11] URL: https://www.nytimes.com/2022/12/10/technology/ai-chat-bot-chatgpt.html

[12] Stephen Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive science*, **11**: 23–63, 1987.

[13] Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, **21**: 335–346, 2020.

[14] Freeman Dyson et al. A meeting with Enrico Fermi. *Nature*, **427**: 297–297, 2004.

[15] URL: `https://www.zdnet.com/article/metas-ai-luminary-lecun-explores-deep-learnings-energy-frontier/`

[16] Ludwig Boltzmann. Studien über das Gleichgewicht der lebendigen Kraft zwischen bewegten materiellen Punkten [Studies on the balance of living force between moving material points]. *Wiener Berichte*, **58**: 517–560, 1868.

[17] Josiah Willard Gibbs. *Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics*. C. Scribner's sons, 1902.

[18] URL: `https://www.hs-augsburg.de/~harsch/anglica/Chronology/20thC/Ising/isi_fm00.html`

[19] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, **79**: 2554–2558, 1982.

[20] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive science*, **9**: 147–169, 1985.

[21] Donald O Hebb. Organization of behavior. new york: Wiley. *J. Clin. Psychol*, **6**: 335–307, 1949.

[22] Siegrid Löwel and Wolf Singer. Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Science*, **255**: 209–212, 1992.

[23] Amos Storkey. "Increasing the capacity of a hopfield network without sacrificing functionality" in: *Artificial Neural Networks—ICANN'97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceeedings 7*. Springer 1997. 451–456

[24] Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, **4**: 1235–1260, 2003.

[25] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, **14**: 1771–1800, 2002.

[26] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. "A theory of generative convnet" in: *International Conference on Machine Learning*. PMLR 2016. 2635–2644

[27] Miguel A Carreira-Perpinan and Geoffrey Hinton. "On contrastive divergence learning" in: *International workshop on artificial intelligence and statistics*. PMLR 2005. 33–40

[28] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-Based Generative Modeling through Stochastic Differential Equations" in: *International Conference on Learning Representations*. 2020.

References for Chapter 2

[1] Evgenii Mikhailovich Lifshitz and Lev Petrovich Pitaevskii. *Statistical physics: theory of the condensed state*. vol. 9 Elsevier, 2013.

[2] Farhan Feroz and Mike P Hobson. Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, **384**: 449–463, 2008.

[3] Jun S Liu. *Monte Carlo strategies in scientific computing*. vol. 75 Springer, 2001.

[4] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, **22**: 79–86, 1951.

[5] Ziqiao Ao and Jinglai Li. Entropy estimation via uniformization. *Artificial Intelligence*, 103954, 2023.

[6] Stephen M Stigler. The epic story of maximum likelihood. *Statistical Science*, 598–620, 2007.

[7] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, **14**: 1771–1800, 2002.

[8] Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

[9] Enrico Fermi, P Pasta, Stanislaw Ulam, and Mary Tsingou *Studies of the nonlinear problems* tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 1955

[10] URL: `https://www.livinginternet.com/i/ii_arpanet.htm`

[11] URL: `https://www.nytimes.com/2022/12/10/technology/ai-chat-bot-chatgpt.html`

[12] Evgeny Morozov. The True Threat of Artificial Intelligence. *International New York Times*, NA–NA, 2023.

[13] Ragnar Fjelland. Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, **7**: 1–9, 2020.

[14] Frederik Federspiel, Ruth Mitchell, Asha Asokan, Carlos Umana, and David McCoy. Threats by artificial intelligence to human health and human existence. *BMJ global health*, **8**: e010435, 2023.

[15] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*, 2020.

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, **313**: 504–507, 2006.

[17] URL: `https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73`

[18] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, 353–374, 2023.

[19] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes" in: *2nd International Conference on Learning Representations, ICLR 2014*. 2014.

[20] Laurent Girin, Fanny Roche, Thomas Hueber, and Simon Leglaive. "Notes on the use of variational autoencoders for speech and audio spectrogram modeling" in: *DAFx 2019-22nd International Conference on Digital Audio Effects*. 2019. 1–8

[21] Radford M Neal and Geoffrey E Hinton. "A view of the EM algorithm that justifies incremental, sparse, and other variants" in: *Learning in graphical models*. Springer, 1998. 355–368

[22] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, **39**: 1–22, 1977.

[23] Antti Honkela, Tapani Raiko, Mikael Kuusela, Matti Tornio, and Juha Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *The Journal of Machine Learning Research*, **11**: 3235–3268, 2010.

[24] Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. "Lagging Inference Networks and Posterior Collapse in Variational Autoencoders" in: *International Conference on Learning Representations*. 2018.

[25] Ruoqi Wei, Cesar Garcia, Ahmed El-Sayed, Viyaleta Peterson, and Ausif Mahmood. Variations in variational autoencoders-a comparative evaluation. *Ieee Access*, **8**: 153651–153670, 2020.

[26] Achraf Oussidi and Azeddine Elhassouny. "Deep generative models: Survey" in: *2018 International conference on intelligent systems and computer vision (ISCV)*. IEEE 2018. 1–8

[27] Saptarshi Sengupta, Sanchita Basak, Pallabi Saikia, Sayak Paul, Vasilios Tsalavoutis, Frederick Atiah, Vadlamani Ravi, and Alan Peters. A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowledge-Based Systems*, **194**: 105596, 2020.

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, **27**: 2014.

[29] URL: `https://sthalles.github.io/intro-to-gans/`

[30] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[31] URL: `https://granadata.art/gan-convergence-proof/#/`

[32] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks" in: *Proceedings of the 5th International Conference on Learning Representations Workshop Track*. 2016.

[33] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. "Improved techniques for training GANs" in: *Advances in Neural Information Processing Systems*. 2016. 2226–2234

[34] Martin Arjovsky and Léon Bottou. "Towards principled methods for training generative adversarial networks" in: *Neural Information Processing Systems Conference Workshop: Adversarial Training*. 2016.

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks" in: *International Conference on Learning Representations*. 2018.

[36] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. "Deep unsupervised learning using nonequilibrium thermodynamics" in: *International conference on machine learning*. PMLR 2015. 2256–2265

[37] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, **56**: 1–39, 2023.

[38] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[39] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, **34**: 8780–8794, 2021.

[40] L Chris G Rogers and David Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*. vol. 2 Cambridge university press, 2000.

[41] Wendell H Fleming and Raymond W Rishel. *Deterministic and stochastic optimal control*. vol. 1 Springer Science & Business Media, 2012.

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-Based Generative Modeling through Stochastic Differential Equations" in: *International Conference on Learning Representations*. 2020.

[43] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, **29**: 1–17, 1998.

[44] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**: 2005.

[45] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, **12**: 313–326, 1982.

[46] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, **23**: 1661–1674, 2011.

[47] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. "Sliced score matching: A scalable approach to density and score estimation" in: *Uncertainty in Artificial Intelligence*. PMLR 2020. 574–584

[48] Michael Samuel Albergo and Eric Vanden-Eijnden. "Building Normalizing Flows with Stochastic Interpolants" in: *The Eleventh International Conference on Learning Representations*. 2022.

[49] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[50] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. "Flow Matching for Generative Modeling" in: *The Eleventh International Conference on Learning Representations*. 2022.

[51] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, **34**: 17695–17709, 2021.

[52] Esteban G. Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, **8**: 217–233, 2010.

[53] E. G. Tabak and Cristina V. Turner. A Family of Nonparametric Density Estimation Algorithms. *Communications on Pure and Applied Mathematics*, **66**: 145–164, 2013.

[54] URL: `https://flowtorch.ai/users/`

[55] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, **43**: 3964–3979, 2020.

[56] Jiaming Song, Shengjia Zhao, and Stefano Ermon. A-nice-mc: Adversarial training for mcmc. *Advances in neural information processing systems*, **30**: 2017.

[57] George Casella, Christian P Robert, and Martin T Wells. Generalized accept-reject sampling schemes. *Lecture Notes-Monograph Series*, 342–347, 2004.

[58] Daniel W Stroock. *An introduction to Markov processes*. vol. 230 Springer Science & Business Media, 2013.

[59] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, **21**: 1087–1092, 1953.

[60] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. 1970.

[61] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, **50**: 5–43, 2003.

[62] Kerrie L Mengersen and Richard L Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The annals of Statistics*, **24**: 101–121, 1996.

[63] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*. vol. 2 Springer, 1999.

[64] Bernt Oksendal. *Stochastic Differential Equations*. 6th ed. Springer-Verlag Berlin Heidelberg, 2003.

[65] J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, **101**: 185–232, 2002.

[66] Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, **8**: 483–509, 1990.

[67] Giorgio Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, **180**: 378–384, 1981.

[68] Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 341–363, 1996.

[69] Andre Wibisono. "Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem" in: *Conference on Learning Theory*. PMLR 2018. 2093–3027

[70] George E Uhlenbeck and Leonard S Ornstein. On the theory of the Brownian motion. *Physical review*, **36**: 823, 1930.

[71] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, **1**: 127–239, 2014.

[72] Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, **56**: 549–581, 1994.

[73] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, **195**: 216–222, 1987.

[74] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 721–741, 1984.

[75] Robert H Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical review letters*, **57**: 2607, 1986.

[76] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, **106**: 620, 1957.

[77] Edwin T Jaynes. Information theory and statistical mechanics. II. *Physical review*, **108**: 171, 1957.

[78] Giovanni Gallavotti. *Statistical mechanics: A short treatise*. Springer Science & Business Media, 1999.

[79] János Aczél, Bruno Forte, and Che Tat Ng. Why the Shannon and Hartley entropies are 'natural'. *Advances in applied probability*, **6**: 131–146, 1974.

[80] Clement John Adkins. *Equilibrium thermodynamics*. Cambridge University Press, 1983.

[81] Enrico Fermi. *Thermodynamics*. Courier Corporation, 2012.

[82] Denis J Evans, Debra J Searles, and Stephen R Williams. Dissipation and the relaxation to equilibrium. *Journal of Statistical Mechanics: Theory and Experiment*, **2009**: P07029, 2009.

[83] William L Jorgensen. Free energy calculations: a breakthrough for modeling organic chemistry in solution. *Accounts of Chemical Research*, **22**: 184–189, 1989.

[84] Aaron R Dinner, Andrej Šali, Lorna J Smith, Christopher M Dobson, and Martin Karplus. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in biochemical sciences*, **25**: 331–339, 2000.

[85] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and Helmholtz free energy. *Advances in neural information processing systems*, **6**: 1993.

[86] Kim A Nicoli, Christopher J Anders, Lena Funcke, Tobias Hartung, Karl Jansen, Pan Kessel, Shinichi Nakajima, and Paolo Stornati. Estimation of thermodynamic observables in lattice field theories with deep generative models. *Physical review letters*, **126**: 032001, 2021.

[87] Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, **13**: 293–301, 2009.

[88] Gabriel Stoltz, Mathias Rousset, et al. *Free energy computations: A mathematical perspective*. World Scientific, 2010.

[89] C Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, **78**: 2690, 1997.

References for Chapter 3

[1] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, **14**: 1771–1800, 2002.

[2] Tijmen Tieleman. "Training restricted Boltzmann machines using approximations to the likelihood gradient" in: *International conference on Machine learning*. 2008. 1064–1071

[3] Carles Domingo-Enrich, Alberto Bietti, Marylou Gabrié, Joan Bruna, and Eric Vanden-Eijnden. Dual Training of Energy-Based Models with Overparametrized Shallow Neural Networks. *arXiv preprint arXiv:2107.05134*, 2021.

[4] Aapo Hyvarinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on neural networks*, **18**: 1529–1531, 2007.

[5] C Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, **78**: 2690, 1997.

[6] Radford M Neal. Annealed importance sampling. *Statistics and computing*, **11**: 125–139, 2001.

[7] Arnaud Doucet, Nando De Freitas, Neil James Gordon, et al. *Sequential Monte Carlo methods in practice*. vol. 1 Springer, 2001.

[8] Lawrence C Evans. *Partial differential equations*. vol. 19 American Mathematical Society, 2022.

[9] Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on progress in physics*, **75**: 126001, 2012.

[10] Johannes Berner, Boris Müller, Juan Ruben Gomez-Solano, Matthias Krüger, and Clemens Bechinger. Oscillating modes of driven colloids in overdamped systems. *Nature communications*, **9**: 999, 2018.

[11] Stephen Smale. On gradient dynamical systems. *Annals of Mathematics*, 199–206, 1961.

[12] Jun S Liu. *Monte Carlo strategies in scientific computing*. vol. 75 Springer, 2001.

[13] James Berger, MJ Bayarri, and LR Pericchi. The effective sample size. *Econometric Reviews*, **33**: 197–217, 2014.

[14] Neil J Gordon, David J Salmond, and Adrian FM Smith. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation" in: *IEE proceedings F (radar and signal processing)*. vol. 140 IET 1993. 107–113

[15] Genshiro Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of computational and graphical statistics*, **5**: 1–25, 1996.

[16] James Carpenter, Peter Clifford, and Paul Fearnhead. Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, **146**: 2–7, 1999.

[17] Tiancheng Li, Miodrag Bolic, and Petar M Djuric. Resampling methods for particle filtering: classification, implementation, and strategies. *IEEE Signal processing magazine*, **32**: 70–86, 2015.

[18] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli*, **18**: 252–278, 2012.

[19] Nan Zhang, Shifei Ding, Jian Zhang, and Yu Xue. An overview on restricted Boltzmann machines. *Neurocomputing*, **275**: 1186–1199, 2018.

References for Chapter 4

[1] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[2] Suriyanarayanan Vaikuntanathan and Christopher Jarzynski. Escorted free energy simulations: Improving convergence by reducing dissipation. *Physical Review Letters*, **100**: 190601, 2008.

[3] Erhard Glötzl and Oliver Richters. Helmholtz decomposition and potential functions for n-dimensional analytic vector fields. *Journal of Mathematical Analysis and Applications*, **525**: 127138, 2023.

[4] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**: 2005.

[5] URL: `https://github.com/Davidedaca/EBMs_Jarzynski`

[6] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. "Learning non-convergent non-persistent short-run MCMC toward energy-based model" in: *Advances in Neural Information Processing Systems*. vol. 32 2019.

[7] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. "Improved Contrastive Divergence Training of Energy-Based Models" in: *International Conference on Machine Learning*. PMLR 2021. 2837–2848

References for Chapter 5

[1] L. Onsager. Reciprocal relations in irreversible processes. I. *Physical review*, **37**: 405, 1931. DOI: `https://doi.org/10.1103/PhysRev.37.405`

[2] L. Onsager. Reciprocal relations in irreversible processes. II. *Physical review*, **38**: 2265, 1931. DOI: `https://doi.org/10.1103/PhysRev.38.2265`

[3] H. B. G. Casimir. On Onsager's Principle of Microscopic Reversibility. *Reviews of Modern Physics*, **17**: 343, 1945. DOI: `https://doi.org/10.1103/RevModPhys.17.343`

[4] R. Kubo. Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems. *Journal of the Physical Society of Japan*, **12**: 570–586, 1957. DOI: `https://doi.org/10.1143/JPSJ.12.570`

[5] R. Kubo *Some Aspects of the Statistical-Mechanical Theory of Irreversible Processes (Lecture Notes in Theoretical Physics) ed WE Brittin and LG Dunham* 1959 DOI: `https://doi.org/10.1143/JPSJ.12.570`

[6] R. Kubo. The fluctuation-dissipation theorem. *Reports on progress in physics*, **29**: 255, 1966. DOI: `https://doi.org/10.1088/0034-4885/29/1/306`

[7] M. J. Lax. *Symmetry Principles in Solid State and Molecular Physics*. vol. 29 1 John Wiley and Sons, 1972. 255 DOI: https://doi.org/10.1088/0034-4885/29/1/306

[8] S. Bonella, G. Ciccotti, and L. Rondoni. Time reversal symmetry in time-dependent correlation functions for systems in a constant magnetic field. *EPL (Europhysics Letters)*, **108**: 60004, 2015. DOI: https://doi.org/10.1209/0295-5075/108/60004

[9] A. Coretti, S. Bonella, L. Rondoni, and G. Ciccotti. Time reversal and symmetries of time correlation functions. *Molecular Physics*, **116**: 3097–3103, 2018. DOI: https://doi.org/10.1080/00268976.2018.1464674

[10] S. Bonella, A. Coretti, L. Rondoni, and G. Ciccotti. Time-reversal symmetry for systems in a constant external magnetic field. *Physical Review E*, **96**: 012160, 2017. DOI: https://doi.org/10.1103/PhysRevE.96.012160

[11] R. Luo, G. Benenti, G. Casati, and J. Wang. Onsager reciprocal relations with broken time-reversal symmetry. *Phys. Rev. Research*, **2**: 022009, 2020. DOI: https://doi.org/10.1103/PhysRevResearch.2.022009

[12] P. De Gregorio, S. Bonella, and L. Rondoni. Quantum Correlations under Time Reversal and Incomplete Parity Transformations in the Presence of a Constant Magnetic Field. *Symmetry*, **9**: 120, 2017. DOI: https://doi.org/10.3390/sym9070120

[13] P. Jacquod, R. S. Whitney, J. Meair, and M. Büttiker. Onsager relations in coupled electric, thermoelectric, and spin transport: The tenfold way. *Physical Review B*, **86**: 155118, 2012. DOI: https://doi.org/10.1103/PhysRevB.86.155118

[14] M. R. Zirnbauer. Symmetry classes. *arXiv preprint arXiv:1001.0722*, **86**: 155118, 2010. DOI: https://arxiv.org/abs/1001.0722v1

[15] U. Marini Bettolo Marconi, A. Puglisi, L. Rondoni, and A. Vulpiani. Fluctuation–dissipation: Response theory in statistical physics. *Physics Reports*, **461**: 111–195, 2008. DOI: https://arxiv.org/abs/1001.0722v1

[16] F. Yaşuk, C. Berkdemir, and A. Berkdemir. Exact solutions of the Schrödinger equation with non-central potential by the Nikiforov–Uvarov method. *Journal of Physics A: Mathematical and General*, **38**: 6579, 2005. DOI: https://arxiv.org/abs/1001.0722v1

[17] P. Cordero and E.S. Hernández. Momentum-dependent potentials: Towards the molecular dynamics of fermionlike classical particles. *Physical Review E*, **51**: 2573, 1995. DOI: https://doi.org/10.1103/PhysRevE.96.012160

[18] Y. Ishii, S. Kasai, M. Salanne, and N. Ohtori. Transport coefficients and the Stokes–Einstein relation in molten alkali halides with polarisable ion model. *Molecular Physics*, **113**: 2442–2450, 2015. DOI: https://doi.org/10.1103/PhysRevE.96.012160

[19] S. Tesson, M. Salanne, B. Rotenberg, S. Tazi, and V. Marry. Classical polarizable force field for clays: Pyrophyllite and talc. *The Journal of Physical Chemistry C*, **120**: 3749–3758, 2016. DOI: https://doi.org/10.1103/PhysRevE.96.012160

[20] Jörn Dunkel and Peter Hänggi. Relativistic brownian motion. *Physics Reports*, **471**: 1–73, 2009. DOI: https://doi.org/10.1103/PhysRevE.96.012160

[21] R mi Hakim. *Introduction to relativistic statistical mechanics: classical and quantum*. vol. 471 1 World scientific, Apr. 2011. 1–73 DOI: https://doi.org/10.1103/PhysRevE.96.012160

[22] A. Aliano, L. Rondoni, and G.P. Morriss. Maxwell-Jüttner distributions in relativistic molecular dynamics. *The European Physical Journal B-Condensed Matter and Complex Systems*, **50**: 361–365, 2006. DOI: https://doi.org/10.1103/PhysRevE.96.012160

References for Chapter 6

[1] L. Onsager. Reciprocal relations in irreversible processes. I. *Physical review*, **37**: 405, 1931. DOI: https://doi.org/10.1103/PhysRev.37.405

[2] L. Onsager. Reciprocal relations in irreversible processes. II. *Physical review*, **38**: 2265, 1931. DOI: https://doi.org/10.1103/PhysRev.38.2265

[3] H. B. G. Casimir. On Onsager's Principle of Microscopic Reversibility. *Reviews of Modern Physics*, **17**: 343, 1945. DOI: https://doi.org/10.1103/RevModPhys.17.343

[4] R. Kubo. Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems. *Journal of the Physical Society of Japan*, **12**: 570–586, 1957. DOI: https://doi.org/10.1143/JPSJ.12.570

[5] R. Kubo *Some Aspects of the Statistical-Mechanical Theory of Irreversible Processes (Lecture Notes in Theoretical Physics) ed WE Brittin and LG Dunham* 1959 DOI: https://doi.org/10.1143/JPSJ.12.570

[6] R. Kubo. The fluctuation-dissipation theorem. *Reports on progress in physics*, **29**: 255, 1966. DOI: https://doi.org/10.1088/0034-4885/29/1/306

[7] S. Bonella, G. Ciccotti, and L. Rondoni. Time reversal symmetry in time-dependent correlation functions for systems in a constant magnetic field. *EPL (Europhysics Letters)*, **108**: 60004, 2015. DOI: https://doi.org/10.1209/0295-5075/108/60004

[8] A. Coretti, S. Bonella, L. Rondoni, and G. Ciccotti. Time reversal and symmetries of time correlation functions. *Molecular Physics*, **116**: 3097–3103, 2018. DOI: https://doi.org/10.1080/00268976.2018.1464674

[9] S. Bonella, A. Coretti, L. Rondoni, and G. Ciccotti. Time-reversal symmetry for systems in a constant external magnetic field. *Physical Review E*, **96**: 012160, 2017. DOI: https://doi.org/10.1103/PhysRevE.96.012160

[10] P. De Gregorio, S. Bonella, and L. Rondoni. Quantum Correlations under Time Reversal and Incomplete Parity Transformations in the Presence of a Constant Magnetic Field. *Symmetry*, **9**: 120, 2017. DOI: https://doi.org/10.3390/sym9070120

[11] R. Luo, G. Benenti, G. Casati, and J. Wang. Onsager reciprocal relations with broken time-reversal symmetry. *Phys. Rev. Research*, **2**: 022009, 2020. DOI: https://doi.org/10.1103/PhysRevResearch.2.022009

[12] A. Coretti, L. Rondoni, and S. Bonella. Fluctuation relations for systems in a constant magnetic field. *Physical Review E*, **102**: 030101, 2020. DOI: https://doi.org/10.1103/PhysRevResearch.2.022009

[13] A. Coretti, L. Rondoni, and S. Bonella. Fluctuation relations for dissipative systems in constant external magnetic field: theory and molecular dynamics simulations. *Entropy*, **23**: 146, 2021. DOI: https://doi.org/10.1103/PhysRevResearch.2.022009

[14] M. J. Lax. *Symmetry Principles in Solid State and Molecular Physics*. vol. 23 2 John Wiley and Sons, Apr. 1972. 146 DOI: https://doi.org/10.1103/PhysRevResearch.2.022009

[15] D. Carbone and L. Rondoni. Necessary and Sufficient Conditions for Time Reversal Symmetry in Presence of Magnetic Fields. *Symmetry*, **12**: 1336, 2020. DOI: https://doi.org/10.1103/PhysRevResearch.2.022009

[16] P. Jacquod, R. S. Whitney, J. Meair, and M. Büttiker. Onsager relations in coupled electric, thermoelectric, and spin transport: The tenfold way. *Physical Review B*, **86**: 155118, 2012. DOI: https://doi.org/10.1103/PhysRevB.86.155118

[17] R. Abraham, A. Ralph, J. E. Marsden, M.E. Jerrold, U.J.E. Marsden, T.S. Ratiu, T.S. Ratiu, and R. Cushman. *Foundations Of Mechanics*. vol. 86 15 Basic Books, Apr. 1978. 155118 DOI: https://doi.org/10.1103/PhysRevB.86.155118

[18] W.K. Tung. *Group theory in physics: an introduction to symmetry principles, group representations, and special functions in classical and quantum physics*. vol. 86 15 World Scientific Publishing Company, Apr. 1985. 155118 DOI: https://doi.org/10.1103/PhysRevB.86.155118

[19] E. Wigner. Über die Operation der Zeitumkehr in der Quantenmechanik. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, **1932**: 546–559, 1932. DOI: https://doi.org/10.1007/978-3-662-02781-3_15

[20] R. G. Sachs. *The physics of time reversal*. vol. 1932 15 University of Chicago Press, Apr. 1987. 546–559 DOI: https://doi.org/10.1007/978-3-662-02781-3_15

[21] A. Anderson. Canonical transformations in quantum mechanics. *Annals of Physics*, **232**: 292–331, 1994. DOI: https://doi.org/10.1007/978-3-662-02781-3_15

[22] J. Von Neumann. On the uniqueness of the Schrödinger operators. *Math. Ann*, **104**: 570–578, 1931. DOI: https://doi.org/10.1007/978-3-662-02781-3_15

[23] R. Gilmore. *Lie groups, Lie algebras, and some of their applications.* vol. 104 2 Courier Corporation, Apr. 2012. 570–578 DOI: https://doi.org/10.1007/978-3-662-02781-3_15

[24] E. A. Milne. *Vectorial mechanics.* vol. 104 2 Interscience, Apr. 1948. 570–578 DOI: https://doi.org/10.1007/978-3-662-02781-3_15

[25] L. Conlon. *Differentiable manifolds.* vol. 104 2 Springer Science & Business Media, Apr. 2008. 570–578 DOI: https://doi.org/10.1007/978-3-662-02781-3_15

[26] M. Colangeli, R. Klages, P. De Gregorio, and L. Rondoni. Steady state fluctuation relations and time reversibility for non-smooth chaotic maps. *Journal of Statistical Mechanics: Theory and Experiment*, **2011**: P04021, 2011. DOI: https://doi.org/10.1007/978-3-662-02781-3_15

[27] M. Colangeli and L. Rondoni. Equilibrium, fluctuation relations and transport for irreversible deterministic dynamics. *Physica D: Nonlinear Phenomena*, **241**: 681–691, 2012. DOI: https://doi.org/10.1007/978-3-662-02781-3_15

[28] P. A. Adamo, M. Colangeli, and L. Rondoni. Role of ergodicity in the transient Fluctuation Relation and a new relation for a dissipative non-chaotic map. *Chaos, Solitons & Fractals*, **83**: 54–66, 2016. DOI: https://doi.org/10.1007/978-3-662-02781-3_15