UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Approcci di meccanica statistica agli algoritmi di ottimizzazione per l'apprendimento automatico

Statistical-mechanics approach to optimization algorithms for machine learning

Relatore

Prof. Marco Baiesi

Correlatore

Dr. Danilo Forastiere

Laureando

Vescovi Giacomo

Anno Accademico 2023/2024

Contents

Abstract

The project concerns the characterization of the properties of some optimization algorithms (which are called Stochastic Gradient Descent) using out-of-equilibrium statistical mechanics methods.

In particular, under some hypotheses, the dynamics of learning of the weights in some paradigmatic models (e.g. linear models or neural networks) can be modeled using stochastic differential equations that do not satisfy the Fluctuation Dissipation relation, which characterizes the systems studied by equilibrium statistical mechanics. The project aims to investigate the relationship between dynamic theories of statistical mechanics and machine learning models. The methodology with which the project will be conducted is twofold. Firstly, an in-depth bibliographic search will be conducted, exploring the most recent and relevant publications, summarizing the state of the art of the theory currently available and secondly, simple numerical experiments will be implemented aimed at testing the hypotheses underlying the theory itself.

Il progetto riguarda la caratterizzazione delle proprietà di alcuni algoritmi di ottimizzazione (chiamati Stochastic Gradient Descent) utilizzando metodi di meccanica statistica fuori dall'equilibrio.

In particolare, sotto alcune ipotesi, la dinamica di apprendimento dei pesi in alcuni modelli paradigmatici (ad esempio modelli lineari o reti neurali) può essere modellata utilizzando equazioni differenziali stocastiche che non soddisfano la relazione di Fluttuazione Dissipazione, che caratterizza i sistemi studiati dalla meccanica statistica di equilibrio. L'obiettivo del progetto è investigare la relazione tra le teorie dinamiche della meccanica statistica e i modelli di apprendimento automatico. La metodologia con cui il progetto sarà condotto è duplice. In primo luogo, verrà effettuata una ricerca bibliografica approfondita, esplorando le pubblicazioni più recenti e rilevanti, riassumendo lo stato dell'arte della teoria attualmente disponibile, e in secondo luogo, saranno implementati semplici esperimenti numerici volti a testare le ipotesi alla base della teoria stessa.

Chapter 1

Introduction

1.1 Introduction

Over the past few decades, machine learning has gained immense popularity and has become a vast field of study. With increasing resources available for computation and modern parallelization techniques a new paradigm emerged in machine learning, namely deep learning, where artificial neural networks are trained to solve various tasks and have found wide applications. Despite the richness of the field, all problems in supervised machine learning (which will be considered from now on) start with some basic ingredients, which will be treated in detail in chapter 2: a dataset comprised of the features of the system we want to study and the relative labels, which contain information that we want to predict. A model function that predicts a feature of the system from a given input, where the prediction can be adjusted by tuning a set of parameters in the model. Finally, a *cost function* that tells us how well the model makes predictions by comparing it to the labeled data. To find the optimal set of parameters we perform a minimization of the cost function. It is known that the gradient of a function points toward its maximum, so it can be employed to find minima by moving along the opposite direction: this method is called gradient descent (GD). However, the number of necessary parameters for neural networks to work made it clear how the gradient descent approach succumbs to the *curse of dimensionality* [4]: rugged landscapes, saddle points and local minima are the norm, and training models with GD becomes infeasible. Moreover, the size of the datasets needed for training made the computation of the gradient very time-consuming.

To solve these problems, in 1991 Bottou introduced the Stochastic Gradient Descend (SGD) algorithm [7], which consisted of computing the gradient only on a randomly chosen subset of all the data at each iteration. This approach and its variants have been fruitful in the last two decades, giving rise to new approaches to study the training of machine learning models.

One of the most prominent (and most relevant to physics) ways to probe the theoretical properties of SGD is an appropriate limiting procedure that turns the SGD into a Stochastic Differential Equation (SDE) due to the introduction of noise due

to random batch sampling [19],[1]. One of the possible inquiries regards the time it takes for a point in the parameter space to escape a local minimum. The same problem exists in statistical physics and chemistry, where thermal noise allows a particle to move between local minima of the energy by jumping an energy barrier. It turns out that if we assume isotropy and uniformity for the noise, then the escape rate is proportional to $e^{(-\Delta E/T)}$ (*Arrhenius' law*) [18]. In general, systems where detailed balance holds (or, equivalently, the fluctuation-dissipation relation FDR) have the Gibbs distribution as stationary distribution, and exhibit this kind of exiting behavior. It has recently been shown in [21] that by a suitable approximation of the covariance in the SGD, one can derive a novel SDE that does not satisfy detailed balance and by approximating it near a local minimum where detailed balance holds, derive an escape rate law which depends polynomially on $\frac{L(\theta^s)}{L(\theta^*)}$, L being the loss function, θ^s the saddle point through which the escape happens and θ^* the local minimum.

We begin by introducing the problem of training a model through supervised learning, giving the basic notation and definitions. We then move to the analogy between the problem in machine learning and statistical physics: we remark the importance of detailed balance and establish its relationship with the Gibbs distribution and the FDR. Using the dynamical approach to nonequilibrium statistical mechanics we then solve the escape problem in one dimension, informally discussing its generalizations to higher ones.

We then model the dynamics of SGD: we show how one can derive the novel covariance matrix proposed in [21] and we derive the structure of the SDE from the SGD through a technique proposed and rigorously proven in [19]. Once we have established the structure of the SDE, we solve it by a random time change [22] by reducing ourselves to a SDE with additive noise which follows detailed balance near critical points. A proof of the new formula for the escape rate is proposed in one dimension and its implications are discussed. Finally, we numerically investigate the dynamics of the SDE and SGD, the validity of the covariance approximation, the escape time formula and the stationary distribution derived in appendix C of [21].

Appendix A treats the necessary mathematical machinery, while Appendix B shows a different approach to the escape problem in statistical mechanics, presented by [24].

1.2 introduzione

Negli ultimi decenni, il machine learing ha guadagnato un'immensa popolarità ed è diventato un vasto campo di studio. Con la crescente disponibilità di risorse per il calcolo e le moderne tecniche di parallelizzazione, è emerso un nuovo approccio, quello del *deep learning*, dove reti neurali artificiali vengono allenate per svolgere vari compiti, con notevoli risultati. Nonostante la vastità del campo, tutti i problemi nel *supervised machine learning* (il solo che verrà considerato da ora in poi) iniziano con alcuni ingredienti di base, che verranno trattati in dettaglio nel capitolo 2: un dataset composto dalle caratteristiche del sistema che vogliamo studiare e dalle relative etichette, le quali contengono le informazioni che vogliamo predirre. Un

modello (descritto tramite una funzione), che dato un input ritorna una stima della grandezza che vogliamo predirre, dove il risultato è parametrizzato dai parametri che vogliamo ottimizzare. Infine, una *cost function* che ci dice la qualità delle previsioni del modello confrontandolo con i dati etichettati. Per trovare i valori ottimali dei parametri eseguiamo una minimizzazione della funzione di costo. È noto che il gradiente di una funzione punta verso il suo massimo, quindi può essere impiegato per trovare minimi muovendosi nella direzione opposta: questo metodo è chiamato *Gradient Descent* (GD). Tuttavia, l'elevato numero di parametri necessari per far funzionare la rete neurale ha reso evidente come l'approccio GD sia succube della *curse of dimensionality* [4]: paesaggi accidentati, punti di sella e minimi locali sono la norma, e addestrare modelli con il GD diventa molto difficile. Inoltre, le dimensioni dei dataset necessari per l'addestramento hanno reso il calcolo del gradiente molto dispendioso in termini di tempo.

Per risolvere questi problemi, nel 1991 Bottou ha introdotto l'algoritmo di *Stochastic Gradient Descent* (SGD) [7]: questo consiste nel calcolare il gradiente solo su un sottoinsieme scelto casualmente di tutti i dati ad ogni iterazione. Questo metodo e le sue varianti sono stati fruttuosi negli ultimi due decenni, dando origine a nuovi approcci per lo studio dell'addestramento dei modelli di apprendimento automatico.

Uno dei modi più rilevanti (e più legati alla fisica) per sondare le proprietà teoriche della SGD è un'appropriata procedura di limite che trasforma lo SGD in un'Equazione Differenziale Stocastica (Stochastic Differential Equation, SDE), grazie all'introduzione di rumore attraverso il campionamento casuale. [19],[1]. Una delle questioni più importanti riguarda il tempo che impiega un punto nello spazio dei parametri per uscire dai minimi locali. Lo stesso problema esiste in fisica statistica e in chimica, dove il rumore termico consente a una particella di spostarsi tra minimi locali dell'energia saltando una barriera energetica. Se assumiamo isotropia e uniformità per il rumore, allora l'*escape rate* è proporzionale a $e^{(-\Delta E/T)}$ (*legge di Arrhenius*) [18]. In generale, i sistemi in cui vale il bilancio dettagliato (o, equivalentemente, la relazione di fluttuazione-dissipazione FDR) ammettono la distribuzione di Gibbs come distribuzione stazionaria e esibiscono questo tipo di dipendenza. È stato dimostrato di recente in [21] che mediante un'adeguata approssimazione della covarianza nella SGD, si può derivare una nuova SDE che non soddisfa il bilancio dettagliato e, approssimandola vicino a un minimo locale dove quest'ultimo vale, derivare una legge per l'escape rate che dipende polinomialmente da $\frac{L(\theta^s)}{L(\theta^*)}$, dove L è la funzione di costo, θ^s il punto di sella attraverso il quale avviene la fuga e θ^* il minimo locale.

Nel seguente elaborato iniziamo introducendo il problema dell'addestramento di un modello tramite l'apprendimento supervisionato, fornendo la notazione e le definizioni di base. Passiamo poi all'analogia tra il problema nel contensto del machine learning e della fisica statistica: sottolineiamo l'importanza del bilancio dettagliato e stabiliamo il suo rapporto con la distribuzione di Gibbs e la FDR. Utilizzando l'approccio dinamico alla meccanica statistica fuori dall'equilibrio risolviamo quindi il problema di fuga in una dimensione, discutendone qualitativamente la generalizzazione a dimensioni superiori.

Passiamo quindi allo studio della dinamica dello SGD: mostriamo come si possa derivare la nuova matrice di covarianza proposta in [21] e deriviamo la struttura della SDE dalla SGD attraverso una tecnica proposta e dimostrata rigorosamente in [19]. Una volta stabilita la struttura della SDE, la risolviamo mediante un cambiamento di tempo casuale [22] riducendoci a una SDE con rumore additivo che, vicino ai minimi, segue il bilancio dettagliato. Viene proposta una dimostrazione della nuova formula per il tasso di fuga in una dimensione e se ne discutono le implicazioni. Infine, investighiamo numericamente la dinamica della SDE e dello SGD, la validità dell'approssimazione della covarianza, la formula del tempo di fuga e la distribuzione stazionaria derivata nell'appendice C di [21].

L'Appendice A tratta gli strumenti matematici necessari, mentre l'Appendice B mostra un diverso approccio al problema di fuga in meccanica statistica, presentato da [24].

Chapter 2

Machine learning models and their training

This section relies for the most on [20].

Despite the richness of the field, almost every machine learning problem starts with a few basic ingredients and a simple recipe. The main elements are

1. A *dataset* $\mathcal{D} = (X, y)$ where $x \in M_{N\times d}$ is a matrix containing the observations and $y \in \mathbb{R}^N$ contains the observation's labels. The dataset is randomly divided into a *training* dataset and a *testing* dataset (usually 10% of the whole set).
2. The *model* f(x; θ), a function $f : x \to y$ which associates an input $x \in \mathbb{R}^d$ to an output $y \in \mathbb{R}$, which is also a function of the parameters $\theta \in \mathbb{R}^P$
3. the *cost function* $\mathcal{C}(y, f(X; \theta))$ which is used to evaluate how well the model performs on the observation y.

In order to find the best parameters for our model we minimize the cost function over the training set, obtaining the estimator

$$\hat{\theta} = \arg\min_{\theta} \mathcal{C}\left(y_{train}, f(X_{train}; \theta)\right). \tag{2.1}$$

We then compute the *in-sample error* $E_{in} = \mathcal{C}(y_{train}, f(X_{train}; \hat{\theta}))$ and the *out-of sample* error $E_{in} = \mathcal{C}(y_{test}, f(X_{test}; \hat{\theta}))$.

2.1 Minimize the cost function

We now face the task of minimizing the cost function. A first approach could be to employ Newton's method for updating the parameters: after some computation one gets the recurrence relation

$$\theta_{t+1} = \theta_t - H^{-1}(\theta_t)\nabla E(\theta_t). \tag{2.2}$$

This method, however, is not practical because calculating the Hessian is computationally very expensive and because, even if we employ a method for approximating

the Hessian in a more cost-effective way, we still have to store and invert a matrix with n^2 entries.

We then turn to a naive method using only the computation of the gradient on the whole dataset. The parameter update is now $\theta_{t+1} = \theta_t - \eta_t \nabla E(\theta_t)$, where η_t is the *learning rate*, which regulates the size of the steps we take in the gradient's direction. The method employing the Hessian updates the learning rate automatically, whereas in the gradient method, we need to specify how to update it at each iteration. However, gradient descent (GD) methods still have some limitations that need to be addressed:

1. Since the parameters space is high dimensional, very rugged and highly non-convex, if GD converges it converges to local mimima.
2. At each iteration GD computes the gradient over the entire dataset. This is inefficient and computationally expensive.
3. Even with random initialization, if GD ends up in a saddle point it can take an exponential time to exit from it [11].

To overcome this problem we introduce stochasticity in the GD, leading to the *stochastic gradient descent* (SGD): we *randomly* divide the dataset into n/B subsets (called *minibatches*) of cardinality B, and at each iteration we compute the gradient on a different minibatch. A full iteration over the dataset is called an *epoch*.

Consider now the dataset $\mathcal{D} = \{(x^{(\mu)}, y^{(\mu)}) : \mu = 1, 2, ..., N\}$, where $x^{(\mu)} \in \mathbb{R}^d$ is a data vector and $y^{(\mu)} \in \mathbb{R}$ its label. The network output is denoted by $f(\theta, x) \in \mathbb{R}$, where $\theta \in \mathbb{R}^P$ denotes the trainable parameters. For the loss function we use the mean square loss

$$L(\theta) = \frac{1}{2N} \sum_{\mu=1}^{N} [f(\theta, x^{(\mu)}) - y^{(\mu)}]^2 =: \frac{1}{N} \sum_{\mu=1}^{N} l_\mu(\theta) \tag{2.3}$$

We then divide the dataset into minibatches $\{B_k\}_{k \in K}, |B_k| = B$, getting the SGD algorithm

$$\theta_{k+1} = \theta_k - \eta \nabla L_{B_k}(\theta_k) \tag{2.4}$$

$$L_{B_k}(\theta_k) = \frac{1}{B} \sum_{\mu \in B_k} l_\mu(\theta) \tag{2.5}$$

Chapter 3

Nonconvex optimization in physics

One way to study the convergence of the discrete dynamics (2.4) is, by taking a suitable limit, to map the problem to a continuous SDE and analyze the convergence of its solutions. This approach has been first introduced in machine learning and theoretically justified by [19]. In the dynamical approach to statistical physics, the same equations model the dynamics of a particle in contact with a thermal bath; the relaxation to equilibrium is a widely studied problem, with, for example, the theoretical explanation of Arrhenius' law due to Kramer [18].

Since, while training the model, we are interested in finding the global minimum in the parameter space to minimize the cost, we have a natural analogy with the study of the dynamics of a particle moving in a given potential. SGD introduces stochasticity in the dynamics, so the appropriate setting is the dyanimcal approach to nonequilibrium statistical mechanics, which allows one not only to compute statistical properties at equilibrium but to describe aspects of the nonequilibrium dynamics. The necessary tools from stochastic analysis are reviewed in appendix A.

In particular, we are interested in the efficiency of the training algorithm, and one of the metrics is the average time it takes to escape from a local minimum in the presence of noise, the so-called *escape problem.*
Of great insight is the study of a system subject to a double-well potential in one dimension. Consider, for now, the general settings of the overdamped Langevin equation

$$d\theta_t = -\mu \nabla L(\theta_t) dt + \sqrt{2D} dW_t \tag{3.1}$$

Where $L : \mathbb{R}^d \to \mathbb{R}$ is sufficiently smooth and grows sufficiently fast at infinity. The followings then hold:

1. The process is Markov and ergodic, so it admits an invariant (or *stationary*) distribution ρ, the *Gibbs* distribution (for a proof, see [24]).

2. The system is reversible with respect to ρ and thus satisfies the detailed balance condition [6].

   This is not surprising, given that the SDE without the noise term is a time-invariant differential equation, and the diffusion coefficient is constant.

3.1 Detailed balance in physics and in machine learning

A fundamental concept that will be later used in the study of SGD's dynamics is **detailed balance**. We give the following definitions based on [1]:

Definition 3.1.0.1 (**Probability current density**)**.** Consider the multidimensional Ito SDE

$$d\theta(t) = \mu(\theta(t))\nabla L(\theta(t))dt + Z((t))dW(t). \tag{3.2}$$

Then, given $D = \frac{1}{2}Z(\theta)Z^T(\theta)$, it follows that (see appendix A) the corresponding Fokker-Planck equation is

$$\frac{\partial}{\partial t}P(\theta, t) = -\nabla \cdot J(\theta, t), \tag{3.3}$$

where we have introduced the **probability current density**

$$J(\theta, t) := -P(\theta, t)\mu(\theta)\nabla L(\theta) - \nabla \cdot [D(\theta)P(\theta, t)] \tag{3.4}$$

with $(\nabla \cdot D(\theta))_\alpha := \sum_\gamma \partial_\gamma A_{\alpha\gamma}(\theta)$. $\nabla L(\theta)$ is the *force* scaled by the *mobility* $\mu(\theta)$, while $D(\theta)$ is the *diffusion matrix*, an N $\times N$ symmetric matrix.

Definition 3.1.0.2 (**equilibrium of a stochastic system**)**.** We define the *stationary distribution* $P^s(\theta)$ as the solution of (3.3) with $\partial_t P(\theta, t) = 0$. If it holds that $J^s(\theta) = 0$ identically, then we say that stationary state is an *equilibrium state*, and we say that *detailed balance* holds.

As can be seen in [1], we say that a system is in equilibrium if the *Einstein relation* (or *fluctuation-dissipation relation FDR*)

$$\mu(\theta) = \beta D(\theta) \tag{3.5}$$

holds. Consider now the *Gibbs* probability distribution

$$P^s_G(\theta) \propto e^{-\beta L(\theta)} \tag{3.6}$$

It will be shown that in a system that admits the Gibbs distribution as the stationary distribution the FDR holds.

Indeed, consider equation (3.1), where now $D(\theta) = D$ is a constant matrix. Imposing the Gibbs distribution as the stationary distribution in (3.3), we get the equation

$$0 = D\nabla^2 P^s_G(\theta) + \mu(\nabla[P^s_G(\theta)] \cdot \nabla[L(\theta)] + P^s_G(\theta)\nabla^2[L(\theta)]) \tag{3.7}$$

$$= P^s_G(\theta)(D - \frac{\mu}{\beta})(-\|\nabla L(\theta)\|^2 + T\nabla^2 L(\theta)). \tag{3.8}$$

If we now impose the identity for an arbitrary potential, relation (3.5) follows. It's now straightforward to see that (3.4) vanishes identically under the validity of the FDR if the stationary distribution is Gibbs'. For a generic SDE further assumptions are required and we may still have currents even if the distribution is Gibbs'.

3.2 Solution of the escape problem

Consider now two local minima of L, x_0 and y_0, and consider the *communication height*

$$H(x_0, y_0) := \inf_{\gamma: x_0 \to y_0} (\sup_{z \in \gamma} L(z)).$$

It can be shown (see [6]) that the communication height is reached at a unique point z_0 such that $\mathrm{H}(x_0, y_0) = L(z_0)$, which in d dimensions is a critical point where the potential decreases in one direction and increases in the other d-1.

We now give the definitions of *first exit time* and *mean first exit time*:

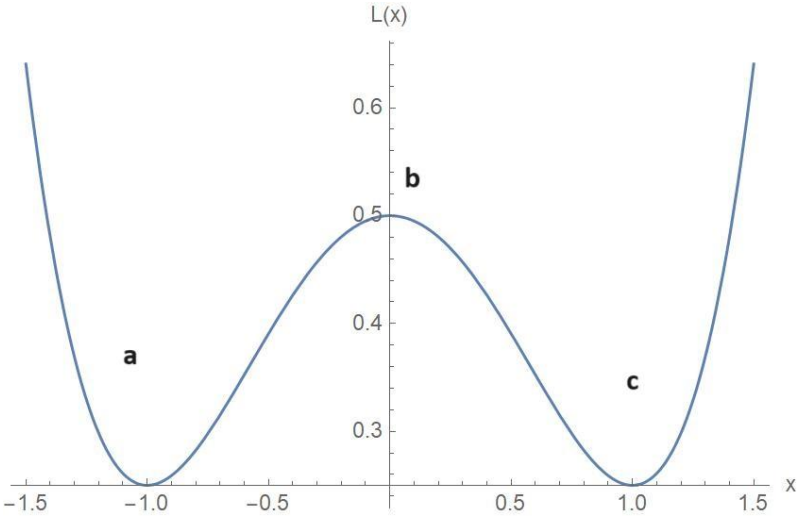
Definition 3.2.0.1 ((Mean) first exit time). Consider a process X_t^x satisfying (A.10) with initial condition $X_0^x = x$ and consider a bounded subset $D \subset \mathbb{R}^d$. We define the mean first exit time (MFET, or mean first passage time MFPT) as

$$\tau(x) := \mathbb{E}[\tau_D^x] = \mathbb{E}[\underbrace{\inf\{t \geqslant 0 : X_t^x \notin D\}}_{\tau_D^x} | X_0^x = x)], \tag{3.9}$$

where τ_D^x is the first exit time.

Consider the following situation for the one-dimensional analogue of (3.1): we have two local minima at a and c, with a local maximum at b, a<b<c, and we solve the equation in the interval (a,b) (see Figure 3.1). Since the potential is confining, we consider a as a reflecting barrier, and since b can be crossed to go into c, b is an absorbing barrier. A solution due to [24], which is described in detail in appendix, gives us the escape rate κ (the so-called ***Kramer escape rate***) from a local minimum through the generator of the stochastic process. An approach that relies more on the physics of the problem and is more suitable for carrying on explicit computations goes as follows, and is a special case of the ones used in [10] and [5].

Figure 3.1: Potential with two local minima and a local maximum



A point particle is placed inside a thermal bath in a position corresponding to one of the minima of the bistable potential L previously described. The damping slows the particle while the thermal fluctuations give the particle the energy to possibly cross the potential barrier.

The Fokker-Planck equation for (3.1) is

$$\partial_t P = \partial_x[L'(x)P] + D\,\partial_{x^2}^2 P = -\partial_x J \tag{3.10}$$
$$J = -L'(x)P - D\,\partial_x P \tag{3.11}$$

where $P = P(x,t)$ converges to the Gibbs distribution for $t \to +\infty$, $D = kT$ and $\mu = 1$. Now, the escape rate tells us the probability of escaping the potential well given the starting point being the minimum, and multiplying it by the probability p of being in the minimum we get J, i.e.

$$J = \kappa p. \tag{3.12}$$

Due to the structure of J we can cast it in the form

$$J = -\frac{1}{kT} e^{-\frac{L(x)}{kT}}\,\partial_x(e^{\frac{L(x)}{kT}}) \tag{3.13}$$

Indeed,

$$J = -\frac{1}{kT} e^{-\frac{L(x)}{kT}}\,\partial_x(e^{\frac{L(x)}{kT}}) \tag{3.14}$$
$$= -De^{-\frac{L(x)}{kT}}\left[e^{\frac{L(x)}{kT}}\frac{L'^{(x)}}{kT}P + e^{\frac{L(x)}{kT}}P'\right] \tag{3.15}$$

and the result follows from the FD relation $D = kT$.

In the steady state $\partial_t P \approx 0$ and J is independent of the position. Moving the prefactor of the derivative in (3.13) to the LHS and integrating between a and c we get

$$[e^{\frac{L(x)}{KT}}P]_a^c = -\frac{1}{kT}J\int_a^b e^{\frac{L(x')}{kT}}dx'. \tag{3.16}$$

Noticing that $P(x = b) \approx 0$, we can divide both sides by the integral and find J.

Moving to the computation of p, we notice that if $L(b) - L(a) \gg kT$ (i.e. P(x) is sharply peaked), then within the potential well the current is almost zero and, solving the differential equation, the following approximation holds:

$$P(x) = P(a)e^{-\frac{L(x)-L(a)}{kT}}. \tag{3.17}$$

We can then compute the probability by integrating over the potential well and, due to the assumption $T \to 0$, extending it over the real line and considering only the first two terms of the expansion around a:

$$p = \int_{well} P(a)e^{-\frac{L(x)-L(a)}{kT}} \tag{3.18}$$
$$\approx P(a)e^{\frac{L(a)}{kT}}\int_{\mathbb{R}} e^{(-L(a)-L''(a)\frac{(x-a)^2}{2})/kT} \tag{3.19}$$
$$= P(a)\sqrt{\frac{2\pi kT}{L''(a)}} \tag{3.20}$$

by the same argument, we can approximate the integral of the RHS of (3.16) by an expansion around b. Piecing everything together we get

$$k = \frac{1}{2\tau} = \frac{\sqrt{\omega_a \omega_b}}{2\pi} e^{-\beta \Delta L}, \tag{3.21}$$

where $\Delta L = [L(b) - L(a)]$ and ω_a, ω_b are the second derivatives of L evaluated respectively at a and b.

Remarkably in higher dimensions, despite the leading coefficient changing and the techniques getting more advanced, the exponential factor remains unchanged, and time reversibility follows. From a physical point of view, at equilibrium probability currents vanish due to detailed balance and the process converges to the stationary Gibbs distribution, which is independent of the dimension of the problem.

Chapter 4

Dynamics of SGD

4.1 introduction

In chapter 3 we saw that, if the noise in a SDE is uniform and isotropic, then the escape rate from a local minimum follows an exponential law. In machine learning, with the previous assumptions, one gets to the same conclusions, where one has $\Delta L = L(\theta^s) - L(\theta^*)$, L being the loss function, θ^* a local minimum and θ^s a saddle point. If however, as argued in [Mori], we describe the dynamics with an inhomogeneous diffusion coefficient, then the escape rate drastically changes, being determined by the *logarithmic* loss barrier $\Delta \log L$.

4.2 The problem

Consider the dataset $\mathcal{D} = \{(x^{(\mu)}, y^{(\mu)} : \mu = 1, 2, ..., N\}$, where $x^{(\mu)} \in \mathbb{R}^d$ is a data vector and $y^{(\mu)} \in \mathbb{R}$ its label. The network output is denoted by $f(\theta, x) \in \mathbb{R}$, where $\theta \in \mathbb{R}^P$ denotes the trainable parameters. For the loss function we use the mean square loss

$$L(\theta) = \frac{1}{2N} \sum_{\mu=1}^{N} [f(\theta, x^{(\mu)} - y^{(\mu)}]^2 =: \frac{1}{N} \sum_{\mu=1}^{N} l_\mu(\theta) \tag{4.1}$$

We divide the dataset into minibatches $\{B_k\}_{k \in K}, |B_k| = B$ and we get the SGD algorithm

$$\theta_{k+1} = \theta_k - \nabla L_{B_k}(\theta_k) \tag{4.2}$$

$$L_{B_k}(\theta_k) = \frac{1}{B} \sum_{\mu \in B_k} l_\mu(\theta) \tag{4.3}$$

We can model the noise as $\xi_k = -[\nabla L_{B_k}(\theta_k) - \nabla L(\theta_k)]$, where $\xi = W'$ formally denotes a white noise process and, as a SDE in the continuous-time limit, $dW(t) = \xi(t)dt$. If we plug $\nabla L_{B_k}(\theta_k)$ into (2.4) we get the recurrence relation

$$\theta_{k+1} = \theta_k - \eta \nabla L(\theta_k) + \eta \xi_k \tag{4.4}$$

We now wish to construct a corresponding SDE to study the dynamics, and thus need to compute the covariance matrix of ξ_k. The following holds:

Proposition 4.2.0.1. *With the definitions given above, we have*

$$\Sigma(\theta) \approx \frac{2L(\theta)}{B} H(\theta^*) \tag{4.5}$$

where $H(\theta) = \nabla^2 L(\theta)$.

Proof.

$$\Sigma(\theta) := \langle (\xi_k - \langle \xi_k \rangle)(\xi_k - \langle \xi_k \rangle)^T \rangle \tag{4.6}$$

$$= \langle \xi_k \xi_k^T \rangle \tag{4.7}$$

$$= \frac{1}{B} \frac{N-B}{N-1} \left(\frac{1}{N} \sum_{\mu=1}^{N} (\nabla l_\mu - \nabla L)(\nabla l_\mu - \nabla L)^T \right) \tag{4.8}$$

Where equation (4.8) has been derived in [16], [15]. Now we do the following: we expand the product and we notice that $\nabla l_\mu \nabla L^T = \nabla L \nabla l_\mu^T$, that $N \nabla L = \sum_{\mu=1}^{N} \nabla l_\mu$ and that ∇L can be taken out of the summation. In the limit N$\gg$B, we get

$$\Sigma(\theta) \approx \frac{1}{B} \left(\frac{1}{N} \sum_{\mu=1}^{N} \nabla l_\mu \nabla l_\mu^T - \nabla L \nabla L^T \right) \tag{4.9}$$

It has been shown in [25] that around critical points

$$\nabla L \nabla L^T \ll \frac{1}{N} \sum_{\mu=1}^{N} \nabla l_\mu \nabla l_\mu^T \tag{4.10}$$

so that the second term in (4.9) is negligible. Using the chain rule for ∇l_μ we have the following relations

$$\Sigma(\theta) \approx \frac{2}{BN} \sum_{\mu=1}^{N} l_\mu \nabla f(\theta, x^{(\mu)}) \nabla f(\theta, x^{(\mu)})^T \tag{4.11}$$

$$=: \frac{2}{BN} \sum_{\mu=1}^{N} l_\mu C_f^{(\mu)} \tag{4.12}$$

$$\approx \frac{2L(\theta)}{NB} \sum_{\mu=1}^{N} C_f^{(\mu)} \tag{4.13}$$

In (4.13) we used the **decoupling approximation** assumption, so that $\mathbb{E}[l_\mu C_f^{(\mu)}] = \mathbb{E}[l_\mu]\mathbb{E}[C_f^{(\mu)}]$. Being an original introduction due to [21], it's verified numerically in a later section.

To show the relation between (4.13) and the Hessian of L, we directly compute $\nabla^2 L$:

$$\nabla^2 L = \frac{1}{N} \sum_{\mu=1}^{N} \{ C_f^{(\mu)}(\theta) + [f(\theta, x^{(\mu)}) - y^{(\mu)}] \nabla^2 f(\theta, x^{(\mu)}) \}. \tag{4.14}$$

The dynamics of the SGD near a local minimum is governed by large eigenvalues so that (see [23]) the term proportional to $\nabla^2 f(\theta, x^{(\mu)})$ is negligible. By evaluating

$H(\theta^*)$ and assuming $C_f^{(\mu)}(\theta) = C_f^{(\mu)}(\theta^*)$ we get the proportionality between $\Sigma(\theta)$ and $H(\theta^*)$

□

This approximation is the core of the approach used by [21] to solve the escape rate problem for the SGD. There are two important observations:

1. The noise is aligned with the Hessian, which describes the curvature of the loss landscape: this implies, as will be shown when deriving the SDE governing the dynamics, that in the flat directions the SGD noise does not arise. This freezes the dynamics along said directions, so that the dimension of the parameter space is reduced. This result is well known in the literature: it has been first pointed out by [26] that in a parameter-dependent diffusion coefficient, SGD exponentially favors flat minima.

2. The covariance is proportional to $L(\theta)$, a new result of [21]. It will be numerically verified in section 5.2 and will be crucial in the solution of the SDE associated with SGD.

For a generic SDE with drift vector $b(X_t)$ and diffusion matrix $\sigma(X_t)$, the associated Euler-Maruyama discretization is

$$X_{k+1} = X_k + \Delta t b(X_k) + \sqrt{\Delta t}\sigma(X_k)Z_k, Z_k \sim \mathcal{N}(0, \mathbb{I}) \tag{4.15}$$

We recall a standard result in statistics, for which a proof can be found in [14]

Theorem 4.2.0.1 (**Normal random vector**). *Let $X \in \mathbb{R}^k$ be a random vector. Then X is called normal if the following holds:*

$$X \sim \mathcal{N}(\mu, \Sigma) \iff \exists A \in \mathbb{R}^{k\times l} \mu \in \mathbb{R}^k | X = AZ + \mu, Z_n \sim \mathcal{N}(0, \mathbb{I}) \forall l = 1, ..., l, \tag{4.16}$$

Where $\Sigma = AA^T$.

It's now straightforward to see that, if we want to have equation (4.4), we need to set $\Delta t = \eta, b(\theta_k) \sim -\nabla L(\theta_k)$ and $\sigma(\theta_k) = \sqrt{\eta\Sigma(\theta_k)}$: we want the two processes (4.15) and (2.4) to be equal. Substituting $\nabla L(\theta_k)$ using $\xi_k = -[\nabla L_{B_k}(\theta_k) - \nabla L(\theta_k)]$ we get the relations

$$\Delta t = \eta \tag{4.17}$$
$$b(X) = -\nabla L(\theta_k) \tag{4.18}$$
$$\sqrt{\Delta t}\, \sigma(X_k)\, Z_k = \eta\, \xi_K. \tag{4.19}$$

Using (4.17), theorem (4.2.0.1) and diving by $\sqrt{\Delta t}$ we have

$$\sigma(X_k)Z_k = \sqrt{\Delta t}\mathbb{I}\, \xi_k = \sqrt{\Delta t\Sigma}\, Z_k \tag{4.20}$$

It's a result of [19] that this equation is indeed a good continuous approximation of the discrete dynamics in the limit $\eta \to dt$. The Hessian matrix encodes the curvature of the loss landscape, so the dynamics along flat directions is frozen because of the low noise.
We thus have the following SDE

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{\frac{2\eta L(\theta_t)}{B}H(\theta^*)}\, dW_t \tag{4.21}$$

Notice how equation (4.21) induces a dynamics that does not satisfy detailed balance because Einstein's relation fails. However, in solving the escape rate problem for the SDE detailed balance is recovered in the following way: near a local minimum the gradient vanishes and the first nonzero term of the Taylor expansion is the hessian of the loss function. Recalling that the diffusion matrix is proportional to the square of the coefficient governing the influence of the Wiener process, we have

$$-\nabla L(\theta_t) \approx -H(\theta^*)(x - x^*) \propto \frac{\eta L(\theta^*)}{B} H(\theta^*)(x - x^*) \tag{4.22}$$

To solve (4.21), we introduce the random time change

$$\tau = \int_0^{t(\tau)} dt' L(\theta_{t'}) \iff d\tau = L(\theta_t)dt \tag{4.23}$$

Given that $dW_t \sim \mathcal{N}(0, \mathbb{I}dt)$ and by definition $d\tilde{W}_t \sim \mathcal{N}(0, \mathbb{I}d\tau)$, using the substitution (4.23) and by theorem (4.2.0.1) we get $d\tilde{W}_\tau = \sqrt{L(\theta_t)}\, dW_t$ and we deduce the differential equation

$$d\tilde{\theta}_\tau = -\frac{1}{L(\tilde{\theta}_\tau)} \nabla L(\tilde{\theta}_\tau) d\tau + \sqrt{\frac{2\eta}{B} H(\theta^*)}\, d\tilde{W}_t \tag{4.24}$$

$$= -[\nabla \log L(\tilde{\theta}_\tau)] d\tau + \sqrt{\frac{2\eta}{B} H(\theta^*)}\, d\tilde{W}_t \tag{4.25}$$

Which possesses simpler additive noise. This implies that near minima of the potential detailed balance holds. It's now clear the importance of the *logarithmic loss* landscape $U(\theta) = \log L(\theta)$

4.3 Solution of the escape problem from a local minima

Like the statistical mechanics problem, the escape rate is computed in the weak-noise limit $\frac{\eta}{B} \to 0$. Some further assumptions are

1. The parameters θ obey the quasi-stationary distribution of (4.25) i.e. the stationary distribution of the SDE restricted to the basin of attraction of the local minimum.
2. The escape from θ^* is dominated by the most probable escape path (MPEP) [12], which is aligned with an eigenvector of the Hessian. As highlighted in the previous sections, in gradient systems the MPEP crosses a saddle point θ^s [12] (in general it may be a generic unstable minimum).
3. We have n nonzero eigenvalues of $H(\theta^*)$ called **outliers**; the remaining P-n eigenvalues are vanishingly small. This is indeed empirically confirmed in [23]
4. The n outlier eigenvectors are constant within the basin of attraction, and the SGD dynamics is restricted to the n-dimensional subspace. This is consequence of the dynamics being frozen along flat directions.

The main result in [21] is then the solution to the escape rate problem:

Theorem 4.3.0.1 (**Mori, Ziyin, Liu, Ueda**). *Let the model parameter θ_t evolve according to the SDE in equation (4.25). Under the previous assumptions, the escape rate κ asymptotically behaves as*

$$\kappa \sim \frac{\sqrt{h_e^* |h_e^s|}}{2\pi} \left[\frac{L(\theta^s)}{L(\theta^*)}\right]^{-\left(\frac{B}{\eta h_e^*}+1-\frac{n}{2}\right)} \tag{4.26}$$

as $\frac{\eta}{B} \to 0^+$ where θ^s is the saddle on the MPEP, h_e^* is the eigenvalue of the Hessian at θ^* along the MPEP, and h_e^s is the negative eigenvalue of the hessian at θ^s along the MPEP.

Proof. (*Sketch*)
We give a one-dimensional analouge of the proof presented in [21], which, since the MPEP follows the direction of one of the eigenvectors of the Hessian, can easily be extended. Thanks to the change of coordinates equation (4.25) is now of the form (3.1) with the potential $U = \log L(\tilde{\theta}_\tau)$ and with the substitution $\frac{1}{\beta} \to \frac{\eta H(\theta^*)}{B}$. We can use the same technique as in chapter 3 with some adjustments. We proceed in the same way: noticing that $\frac{\partial^2 U}{\partial z^2}(\theta^*) = \frac{h^*}{L(\theta^*)}$ (where $\frac{\partial^2 L}{\partial z^2}(\theta^*) = h^*$) we deduce $p = P(\theta^*)\sqrt{\frac{2\pi\eta L(\theta^*)}{B}}$.

For the probability current, having again $\frac{\partial^2 U}{\partial z^2}(\theta^s) = \frac{h^s}{L(\theta^s)}$ and using (3.16) we get

$$J_\tau = \frac{\frac{\eta h^*}{B} e^{\frac{B}{\eta h^*}U(\theta^*)} P(\theta^*)}{\sqrt{\frac{2\pi\eta h^* L(\theta^s)}{|h^s| B}} e^{\frac{B}{\eta h^*}U(\theta^s)}} = \sqrt{\frac{\eta h^* |h^s|}{2\pi B L(\theta^s)}} e^{-\frac{B}{\eta h^*}\Delta U} P(\theta^*). \tag{4.27}$$

Up to this point, the derivation is valid for the re-scaled time variable $\tau = \int_0^{t(\tau)} dt' L(\theta_{t'})$. For the quasi-stationary distribution in the weak-noise limit $L(\theta_t) \approx L(\theta^*)$ so that $\tau \sim L(\theta^*)t$, the current is re-scaled as well (see [1]) and we have

$$\kappa = L(\theta^*)\frac{J_\tau}{p} = \frac{\sqrt{h^* h^s}}{2\pi}\sqrt{\frac{L^2(\theta^*)}{L(\theta^*)L(\theta^s)}} e^{-\frac{B}{\eta h^*}\Delta U} = \frac{\sqrt{h^* h^s}}{2\pi}\left(\frac{L(\theta^s)}{L(\theta^*)}\right)^{-(\frac{1}{2}+\frac{B}{\eta h^*})} \tag{4.28}$$

That is a special case of the formula derived by [21]. The insights into the dimensionality of the problem are lost in exchange for an easier derivation.

□

We notice some of the most important properties of the formula stated above and compare it to (B.6), keeping in mind that $\kappa \propto \frac{1}{\tau}$:

1. κ increases with h_e^* and n: the formula is consistent with and gives a theoretical explanation to the observation that sharp minima (minima with larger h_e^*) are unstable. Moreover, it gives a new theoretical insight into the fact that minima with low effective dimension n are preferred.

2. By imposing $\kappa > 0$ and choosing the largest possible eigenvalue we get a quantitative higher bound n_c for the effective dimension

$$n < n_c := 2\left(\frac{B}{\eta h_{max}} + 1\right). \tag{4.29}$$

The restriction to the effective dimension subspace implies a reduction of the capacity of the network, giving evidence of the presence of implicit regularization in SGD, which prevents overfitting [20].

3. The new result, in contrast with the one where the escape rate follows Arrhenius' law and following intuition describes how, in the low noise limit $\frac{\eta}{B} \to 0$, it takes an infinite amount of time to escape a local minimum. In particular, for a global minimum $L(\theta^*) = 0$ and the diffusion matrix vanishes, in accordance with the prediction.

Chapter 5

Numerical results

5.1 The objective function

This section is devoted to the numerical verification of the approximations and the results exposed in chapter 4. For the analysis, the following objective function has been selected:

$$f_{mod}(a, b; x) = a^2 + x\left(1 + \frac{b}{1 + e^{-ax}}\right). \tag{5.1}$$

The data is then generated on the condition of

$$f(x) := f_{mod}(1, 1; x) = 1 + x\left(1 + \frac{5}{1 + e^{-x}}\right) \tag{5.2}$$

being the true model we want to obtain through optimization with the true parameters $(a, b) = (1, 5)$, which can be seen in Figure 5.1. The object function is a piece-wise linear function that has been chosen to maintain both the complexity of a nonlinear model (leading to a non-convex optimization problem) and the intuition gained from the direct visualization of the loss landscape(see Figure 5.2). A regularization term has been introduced to smoothen the loss landscape. Furthermore, the loss landscape possesses a deeper, flatter and a sharper, more shallow minimum as can be seen in Figure 5.3.

The data to be used in the analysis has been then generated as follows: the interval $x \in [-15, 15]$ has been equally divided in 30 points and, for each point, 20 instances of f(x) have been evaluated. Each value has then been perturbed by a value drawn form a uniform distribution on the interval $[-3, 3]$.

5.2 Covariance matrix approximation

We begin by verifying the validity of approximation (4.5). A key step in the proof is the assumption of being near a local minimum, so we sample the difference between the actual value and the approximation near the minimum in (a=1,b=5), i.e. we compute

$$Diff(\theta) := N[\Sigma(\theta) - \frac{2L(\theta)H(\theta^*)}{B}] \tag{5.3}$$

Figure 5.1: Model evaluated at the true parameters a=1, b=5

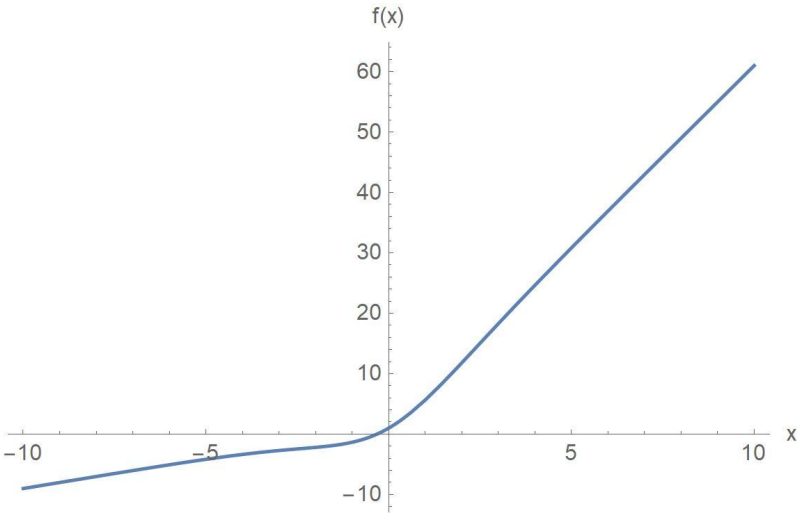
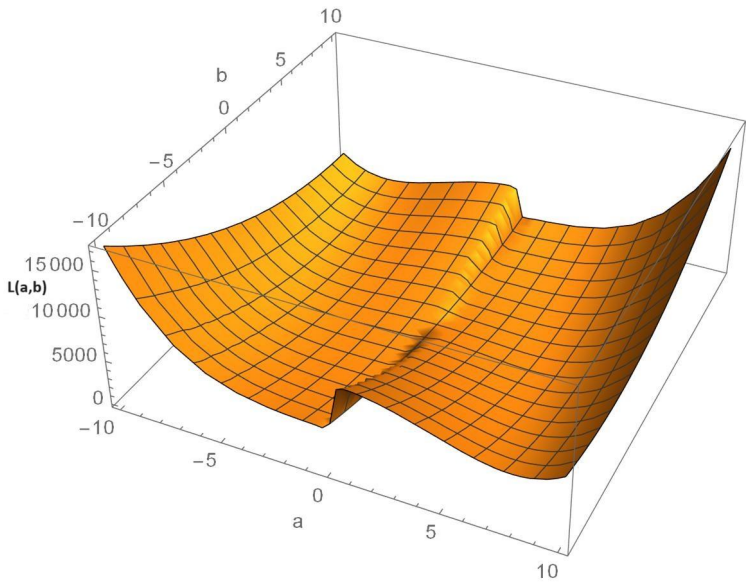


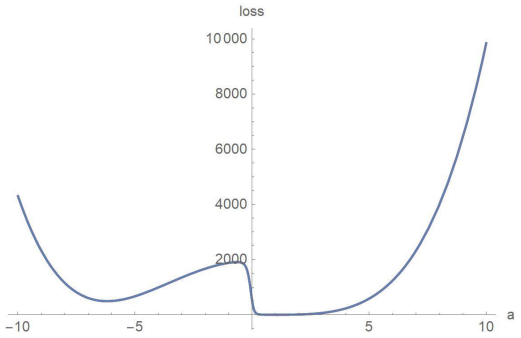
Figure 5.2: Loss landscape near the two minima



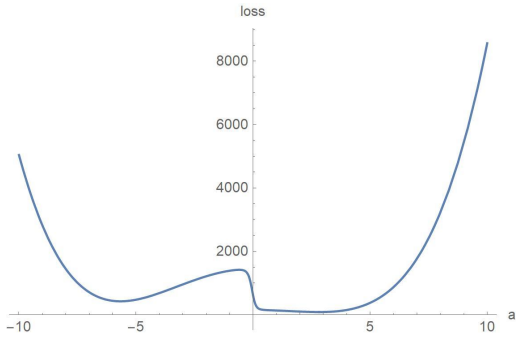
where $N[\cdot]$ is the Frobenius norm $N[A] := \sqrt{\sum_i \sum_j |A_{i,j}|^2}$. As shown in Figure 5.4, the difference is various orders smaller near the minimum, confirming the validity only in this region. Using the insight from Table 5.1, we notice how there seems to be a preference for the flatter minimum by computing the ratio between the difference and the loss.

	real minima	spurious minima	saddle
loss	2.6	424.5	1059.4
difference	$2.7 \cdot 10^5$	$1.44 \cdot 10^8$	$4.8 \cdot 10^8$

Table 5.1: Computation of (5.3) in some relevant points



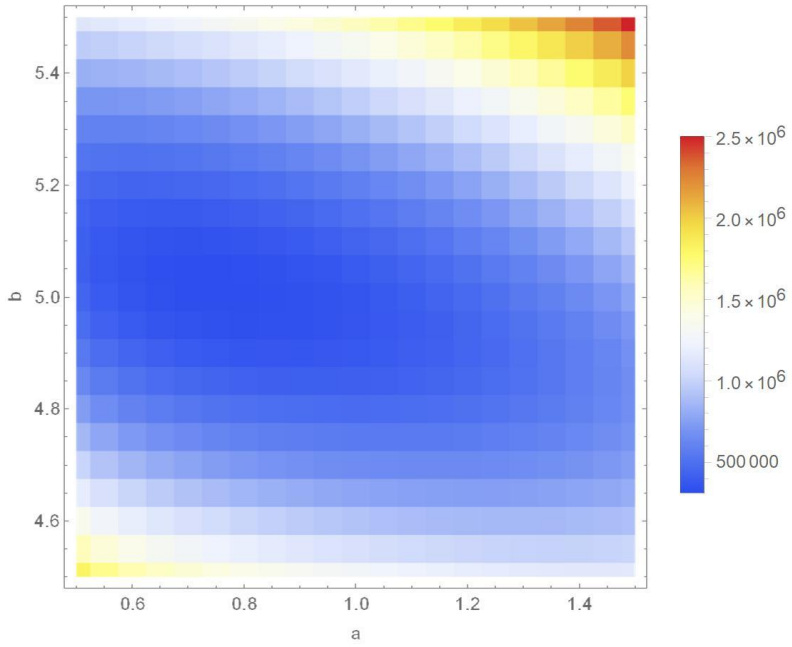
((a)) Section of the loss function with b equal to the real minimum: in blue the section with b=5, in red the section with b found numerically



((b)) Section of the loss function with b equal to the negative (spurious) minimum, found numerically with Mathematica

Figure 5.3: Minima of the loss function

Figure 5.4: Heat map of the norm of $\Sigma(\theta) - \frac{2L(\theta)}{B}$, where on the axes we have the two parameters a and b



5.3 SDE vs SGD

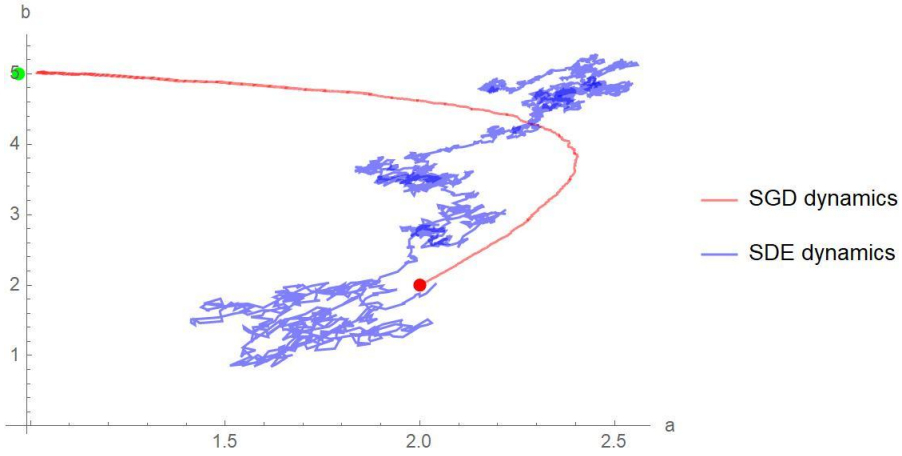
Next, we verify the validity of the continuous-time approximation (4.21) of the discrete dynamics (2.4), and so we indirectly verify the novel decoupling approximation introduced by [21]. Given that the continuous approximation holds in the limit $\frac{\eta}{B} \to 0$, we fixed B=10 and $\eta = 10^{-4}$ in the dynamics of both (4.21) and (2.4). To explore a bigger portion of the parameter space, we sampled six paths with six randomly chosen starting points: we then computed the average and the standard deviation of the points the dynamics converged to. The results are presented in Table 5.2 for the SDE and in Table 5.3 for the SGD, while some realizations of the dynamics can be seen in Figure 5.6. The minimization has been implemented by stopping the training when the conditions $|\tilde{a} - a_{real}| \leq 0.05, |\tilde{b} - b_{real}| \leq 0.05$ were

met. We make the following observations:

1. The ratio $\frac{\eta}{B} = \frac{10^{-4}}{10} = 10^{-5}$ is the only order of magnitude for which we have been able to make a comparison: for higher ratios the SDE diverges, supporting the validity of the approximation only in the small learning rate limit. For lower ratios the SDE converges, but the facts that the time step has to follow $\Delta t = \eta$ (as shown in (4.17)) and that computing the gradient at each iteration is computationally expensive make the simulation harder to complete. The SGD instead performs well in a wide range of learning rates, as can be seen for example in Figure 5.5.

2. The dynamics induced by the SDE is more noisy and tends to follow a less smooth path compared to the SGD dynamics.

Figure 5.5: Comparison of realization of the SDE and the SGD with batch size B=1 and learning rate $\eta = 10^{-5}$. The SDE evolution has been stopped after 4000 iterations. The starting point is (2,2)

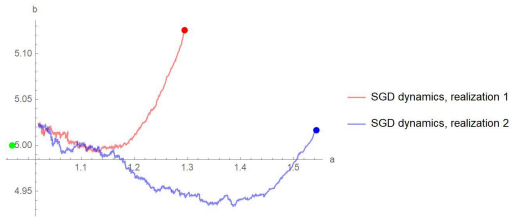


ensamble path	1	2	3	4	5	6
(a_0, b_0)	(1.29,5.12)	(1.54,5.01)	(1.62,5.7)	(1.42,3.05)	(0.83,5.9)	(1.49,6.63)
(a_{fin}, b_{fin})	(1.00,5.03)	(1.01,4.98)	(0.95,5.04)	(1.02,5.00)	(0.93,4.98)	(0.99,5.03)
iterations	217	372	1542	2451	1818	441

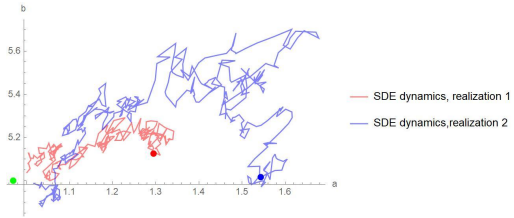
Table 5.2: Performance of SDE dynamics, with B=10 and $\eta = 10^{-4}$. The time steps in the simulations follow the condition $\Delta t = \eta$ as required by the limiting procedure. The true minimum, found with Mathematica, is $(a_{real}, b_{real}) = (0.97, 5.00)$

ensamble path	1	2	3	4	5	6
(a_0, b_0)	(1.29,5.12)	(1.54,5.01)	(1.62,5.7)	(1.42,3.05)	(0.83,5.9)	(1.49,6.63)
(a_{fin}, b_{fin})	(1.02,5.02)	(1.02,5.02)	(1.02,5.02)	(1.02,5.01)	(0.92,5.04)	(0.99,5.05)
iterations	1097	1744	1473	2164	3224	469

Table 5.3: Performance of SGD dynamics, with B=10 and $\eta = 10^{-4}$. The true minimum, found with Mathematica, is $(a_{real}, b_{real}) = (0.97, 5.00)$



((a)) Two realizations of the stochastic process defined by (2.4) (SGD) taken from Table 5.3 . The red and blue points represent the randomly selected starting points, while the green point is the real minimum (0.97,5.00) computed with Mathematica.



((b)) Two realizations of the stochastic process defined by (4.21). The red and blue points represent the randomly selected starting points, while the green point is the real minimum (0.97,5.00) computed with Mathematica.

Figure 5.6: Realization of the SGD and the SDE

5.4 Escape rate

We now examine the behavior of the escape rate from a local minimum in the following way: we start the dynamics of the SDE from the spurious minimum (-5.6,3.2) and we let it evolve until it reaches the basin of attraction of the real minimum (0.96,5.00). We measure the times it takes for the dynamics to cross the saddle point, where the basin of attraction for the spurious minimum has been modeled as the region $a \leq 0.5, b \leq 4$. Computing the escape time as $\tau = 1/\kappa$, we notice how from the escape rate formula (4.26) we can derive, for n=2, the relation

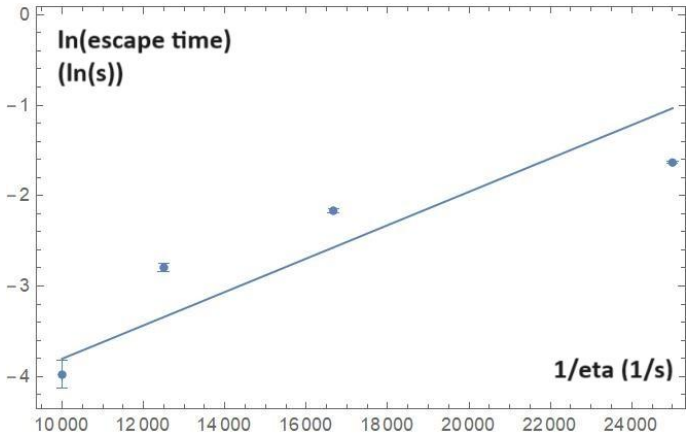
$$\log[\tau] = -\log\left[\frac{\sqrt{h_e^* |h_e^s|}}{2\pi}\right] + \frac{B}{\eta h_e^*} \log\left[\frac{L(\theta^s)}{L(\theta^*)}\right]. \tag{5.4}$$

So, the approach would be to compute the logarithm of the escape time and then employ a linear regression on the pairs $(1/\eta, \log \tau)$. To compute the error we proceed as follow: we assume a uniform distribution for the measurement for τ and we assum to knwo η with certainty, we propagate the error and we deduce the uncertainties in Table 5.7. From the linear regression, we get the results presented in Table 5.4. The theoretical value for the intercept is $m' = \frac{B}{h_e^*} = \frac{1}{h_e^*} = 0.007$. After comparing the theoretical and numerical results, it is clear that the two are not compatible. Through some testing, it was found that the limited region of learning rate exploration and computational time are the most probable reasons for the discrepancy. To obtain a more significant result, a solution could have been to either sample more escape times for the same escape rates or compute the escape rate for lower learning rates. However, both of these solutions required a significant amount of additional time.

	value	standard deviation
m $(s \cdot \log(s))$	-5.6	0.8
q (log(s))	0.00018	0.00006

Table 5.4: Linear regression of the form $\log(\tau) = a + b\frac{1}{\eta}$ for the data presented in Table 5.5.

Figure 5.7: Linear regression for the escape time on the data in Table 5.5. The parameters of the interpolation can be found in Table 5.4



eta	iterations	escape time τ (s)	ln(escape time) [ln(s)]	uncertainty
10^{-4}	187	0.018	-4.0	0.1
$8 \cdot 10^{-5}$	765	0.061	-2.79	0.05
$4 \cdot 10^{-5}$	1911	0.114	-2.17	0.03
$4 \cdot 10^{-5}$	4885	0.195	-1.634	0.001

Table 5.5: Data for the escape time from the spurious minimum, where the escape time has been computed as the number of iteration times the learning rate. Assuming a resolution R=0.01 s and a uniform distribution for the values, we have an uncertainty $\sigma_\tau = \frac{0.01}{\sqrt{12\tau}}$.

5.5 Stationary distribution

As shown in Appendix C of [21], given that equation (4.25) has the Gibbs' distribution as the stationary distribution $\tilde{P}_s(\theta) \propto e^{-U(\theta)/T}$, equation (4.21) with the multiplicative noise has the stationary distribution

$$P_s(\theta) \propto L(\theta)^{-1} P_s(\theta). \tag{5.5}$$

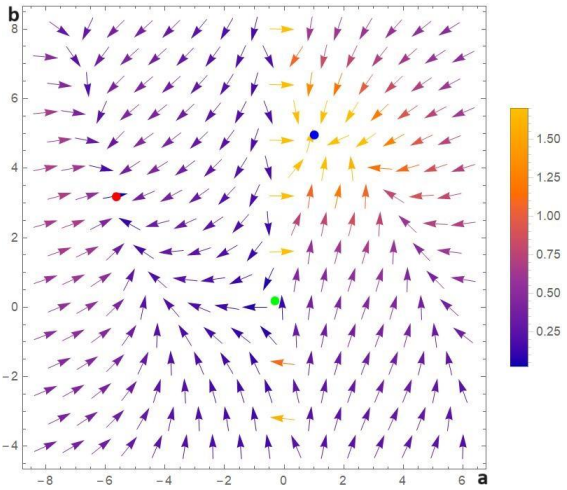
In deriving (5.5) [21] starts from an SDE of the form

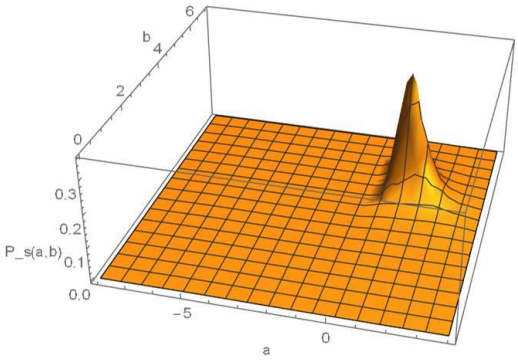
$$d\tilde{\theta}_\tau = -U'(\tilde{\theta}_\tau)d\tau + \sqrt{2T}d\tilde{W}_t \tag{5.6}$$

and derives the form of the old distribution before the time change (4.23). Fixing now the potential $U(\theta) = \log L(\theta)$, the temperature T= 100 K, the batch size b=1 and the learning rate $\eta = 10^{-5}$ we can compute and analyze the probability current densities vector field: we compute it through equation (3.4) and verify the consistency of its qualitative properties. Figure 5.8 shows the results.

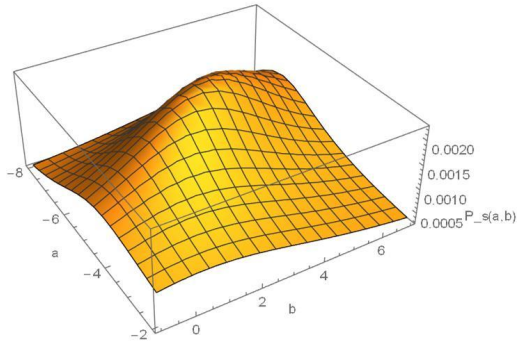
Next, we perform the following test: we let the dynamics evolve from different starting points according to equation (4.21). The hyperparameters are the batchsize b=1 and the learning rate $\eta = 10^{-5}$. We simulate 20 paths with starting point $(a_0, b_0) \in (-6, 6) \times (-1, 6)$ and we compute 1000 time steps, saving the point every tenth iteration. we then produced the normalized histogram in Figure 5.10, which is then compared with the analytic expression $P_s(\theta) = L(\theta)^{(-1/T-1)}$, whose plot can

Figure 5.8: Probability current densities $J[P_s(\theta)](a, b)$, where the coloring represents the magnitude. The red point represents the spurious minimum, the blue point the actual minimum and the green point the saddle point, independently computed in previous sections





((a)) Stationary distribution for the stochastic process (4.21)



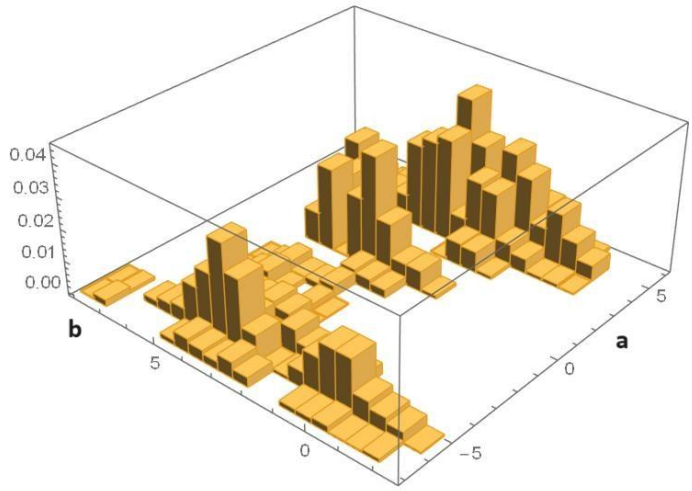
((b)) Zoom of the stationary distribution of (4.21) near the spurious minimum

Figure 5.9: Stationary distribution of (4.21)

be seen in Figure 5.9.

First, we can notice how the equilibrium points computed in the previous experiments with independent methods coincide with the ones of the vector field in Figure 3.4, confirming the validity of the derived stationary distribution.

Figure 5.10: Normalized distribution of the points reached by the dynamics induced by (4.21)



Analyzing now the graph of the stationary distribution presented in Figure 5.9, it's clear how the distribution is more peaked near the real, flatter minimum while it's more smothered near the spurious, sharper minimum. This result agrees with the insight for which SGD is frozen along flat directions and prefers flat minima. Moreover, the result has been validated by the relative frequencies in Figure 5.10: the distribution is more peaked near the flatter minimum and more smeared near the sharper minimum. It must be noticed however that the relative frequencies of the two minima do not coincide between the histogram and the plot of the analytical function: this is probably due to the finite size of the path ensemble.

Chapter 6

Conclusions

Stochastic gradient descent induces a highly complex and very rich dynamics. The reduction of its behaviour to the dynamics induced by a SDE, while very fruitful for a statistical physics approach, severely limits the range of parameters accessible to the analysis in exchange for some very profound insight into certain aspects of the dynamics.

In the previous treatment, the main obstacle to the numerical analysis of the various features of the model has been the computational resources necessary to make simulations. To solve this problem, some refinements can be made: first, the software Mathematica, while great for symbolic computation, does not behave well when treating large amounts of data, so finding a new programming platform could speed up computations. Second, the simulation algorithms implemented can be improved: for example, a further study can be made into rare events sampling methods in stochastic processes (see, for example, [8]). Another way to speed up the algorithms is to precalculate computational-intensive calculations and approximate them on a grid of values during the iterations instead of computing them exactly.

Nevertheless, most of the claims regarding the behaviour of SGD have been verified, at least qualitatively.

Appendix A

Stochastic analysis

A.1 Why do we need a new mathematical framework?

There are various ways through which one can find inconsistencies in the naive approach to the treatment of noise in physical systems used by Langevin, see for example ([13],[17], [2]). One simple argument is that the Gaussian white noise has infinite variance, so that every realization is nowhere differentiable. Another argument can be exposed as follows: consider the Langevin equation

$$\frac{du}{dt} = f(u,t) + g(u,t)\eta(t) \tag{A.1}$$

for Gaussian white noise, and in particular the equation

$$\dot{u}(t) = \eta(t)u(t) \tag{A.2}$$

Choosing a forward Euler method to discretize the derivative, we get the recursive functional equation $u(t_{j+1}) = (1 + \eta(t_j)\Delta t)u(t_j)$, thus we find

$$u(t_N) = (1 + \eta(t_{N-1})\Delta t)...(1 + \eta(t_0)\Delta t)u(t_0) \tag{A.3}$$

Taking the average of equation (A.3), we see that given that $\langle \eta(t_j)\eta(t_k)\rangle = \phi(s_1 - s_2) \longrightarrow \Gamma\delta(s_1 - s_2)$, we can factorize the expectation of the product, and given that $\eta(t)$ has zero mean, we conlude that $\langle u(t_N)\rangle = u(t_0)$. If we then take the limit $\Delta t \longrightarrow 0$ fixig $N\Delta t$, **after** having computed the expected value, we get that $\langle u(t)\rangle = u(t_0)$ i.e. we have no drift.

If one instead takes the continuum limit of A.3 in a Riemann fashion, we get, as for an ordinary differential equation

$$u(t) = u(0)\exp\left\{\int_0^t \eta(s)\,ds\right\} \tag{A.4}$$

and it can be shown (see [2]) that in this case $\langle u(t)\rangle = 0$.

This follows from the fact that the noise-induced drift appears only if a certain prescription for treating the discretization of the formal differential equation A.2 is used. Chapter A.2 will lead to the introduction of *Ito* and *Stratonovich* stochastic differential equations.

A.2 Markov processes and differential Chapman-Kolmogorov equations

There are several ways in which one can approach the subject: following a physics-friendly one, the main reference will be [13], but the following shall be used as well. For a slightly more formal introduction see [22] or [9].

Definition A.2.0.1 (**Stochastic process**)**.** A $(E, \mathcal{E})$-valued *stochastic process* is a family of random variables $\{X_t\}_{t\in I}$ where I is a totally ordered set and $\forall t \in I, X_t : (\Omega, \mathcal{F}, P) \longrightarrow (E, \mathcal{E})$, between a probability space and a measurable space. Moreover, for all $\omega \in \Omega$, the mapping $X(\cdot, \omega) : t \in I \longrightarrow X_t(\omega)$ is called the *path of the process* or *sample path* corresponding to ω.

Observation A.2.0.1. All paths of the process belong to the space E^I of functions defined on I and with values in E. It is possible to equip E^T with a suitable σ-algebra $\mathcal{B}^T$ to turn our family of trajectories into a random function $X : (\Omega, \mathcal{F}) \longrightarrow (E^T, \mathcal{B}^T)$

If a process is *separable* [9], then to completely determine it one can simply specify the set $\{p(x_1, t_1; x_2, t_2; ...x_n, t_n) \mid n \in N\}$ from which one get all the conditional probability distributions.

Even if in practice we will not deal with this kind of mathematical technicality, it's still useful for clarity to introduce the concept of

Definition A.2.0.2 (**Filtration**)**.** In the context of measure theory, a *filtration* is an indexed family $(\mathcal{F}_t)_{t\in R_+}$ of increasing subalgebras of a given algebra $\mathcal{F}$. A process is said to be *adapted to the filtration* if, for every t, X_t is measurable with respect to $\mathcal{F}_t$.

The family $\mathcal{F}_s^X$ with $\mathcal{F}_s^X := \sigma(\{X_u\}_{0\leq u\leq s})$, $\sigma(X_i) := \{X_i^{-1}(B) : B \in \mathcal{F}\}$ is said to be the *natural filtration with respect to the process* $(\mathcal{F}_t)_{t\in R_+}$.

We are now ready to state the *Markov property* for a general random process:

Definition A.2.0.3 (**Markov property**)**.** With the notation above, a $(E, \mathcal{E})$-valued stochastic process $\{X_t\}_{t\in I}$ adapted to the filtration $(\mathcal{F}_t)_{t\in R_+}$ is said to be *Markov* if, $\forall\, A \in \mathcal{E}$, $s, t \in I$ with $s < t$, $P(X_t \in A|\mathcal{F}_s) = P(X_t \in A|X_s)$. If the index set I is discrete, then the condition reduces to $P(X_n = x_n|X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_n = x_n|X_{n-1} = x_{n-1})$

Observation A.2.0.2. We notice how the Markov property simply describes the *memory-less* nature of the process and implies that every joint probability can be factored as product of conditional probabilities, i.e.

$$p(x_1, t_1; ...; x_n, t_n) = p(x_n, t_n) \prod_0^n p(x_{i-1}, t_{i-1}|x_i, t_i) \tag{A.5}$$

with $t_1 \geqslant t_2 \geqslant .... \geqslant t_n$

With the aid of Markovianity, we can easily prove that:

Proposition A.2.0.1 (**Chapman-Kolmogorov equation**). *With the notation above, the conditional probabilities of a Markov process obey the relation*

$$p(x_1, t_1|x_3, t_3) = \int p(x_1, t_1|x_2, t_2)p(x_2, t_2|x_3, t_3)\, dx_2$$

Observation A.2.0.3. Notice how, while the random variable X(t) for fixed t may take values in the continuum, this in general doesn't imply that the sample paths are continuous. This holds if the probability density function of the process satisfies the *Lindeberg condition*: $\forall \epsilon > 0$,

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|x-z|>\epsilon} p(x, t + \Delta t|z, t)\, dx = 0$$

uniformly in z,t and Δt.

We would now like to find processes that satisfy the Chapman-Kolmogorov equation: this is a highly complex functional equation for the conditional probabilities. As it's customary, one approach is to derive a differential relation and then find (some of) the solutions: the ***forward (or backward) Chapman-Kolmogorov equation***.

In general the relation can be derived for the conditional expectation $u(x, s) = \mathbb{E}[f(X_t)|X_s = x]$, where, wishing to derive a relation independent of the particular function, f ranges over all twice differentiable functions. If the probability measure admits a density (as will be the case), then one can deduce the relation for the conditional probability $p(x, t|y, t')$.

Theorem A.2.0.1 (**Forward differential Chapman-Kolmogorov equation**). *Let $x, y, z \in \mathbb{R}^n$, and consider the following definitions/conditions, $\forall \epsilon > 0$:*

1. $$\lim_{\Delta t \to 0} p(x, t + \Delta t|z, t)/\Delta t = W(x|z, t)$$

   uniformly in x,z,t and for $|x - z| \geqslant \epsilon$.

2. $$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|x-z|<\epsilon} (x_i - z_i)p(x, t + \Delta t|z, t)\, dx = A_i(z, t) + O(\epsilon)$$

3. $$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|x-z|<\epsilon} (x_i - z_i)(x_j - z_j)p(x, t + \Delta t|z, t)\, dx = B_{ij}(z, t) + O(\epsilon)$$

If the previous conditions hold, then for a process obeying the Chapman-Kolmogorov

equation one has:

$$\partial_t p(z,t|y,t') = -\sum_i \frac{\partial}{\partial z_i}[A_i(z,t)p(z,t|y,t')] + \sum_{i,j} \frac{1}{2}\frac{\partial^2}{\partial z_i \partial z_j}[B_{ij}(z,t)p(z,t|y,t')] + \int dx[W(z|x,t)p(x,t|y,t') - W(x|z,t)p(z,t|y,t')] \quad \text{(A.6)}$$

The key points of the proof are the absolute continuity, the differentiability and the fact that the process is not singular for x=z. Equation (A.7) is one of the most important in the study of stochastic processes, so we'll point out some of the key features.

Observation A.2.0.4.

1. One can ask whether the solutions of the *differential* Chapman-Kolmogorov equation are solutions of the Chapman-Kolmogorov equation: it can be shown that under the initial condition p(z,t|y,t)=$\delta(y-z)$ and under some additional boundary conditions it is the case.

2. The definitions preceding the differential equation reflect different aspects of the stochastic process: 1 constitutes the non-continuous part of the process which is responsible for the jumps in the path: we can deduce it by the fact that if condition in observation A.2.0.3 is satisfied, then we have W(x|y,t)=0 identically. 2 and 3 represent respectively the *drift* and *diffusion coefficients*: the first one is responsible for the deterministic behaviour of the system, so that when the last two terms are identically zero we get the *Liouville equation.* The second term is responsible for the stochastic continuous behaviour: to see this heuristically, we can take the initial condition p(z,t|y.t)=$\delta(z-y)$ and approximate the differential equation for p(z,t+Δt|y,t) at zeroth-order. The solution to said equation is a gaussian distribution with mean y+A(y,t)Δt and with covariance matrix B(y,t)Δt. The full example can be found in [13], and the formalization will lead to the concept of *Stochastic differential equation.*

3. We have the *backwards differential Chapman-Kolmogorov equation*

$$\partial_{t'} p(x,t|y,t') = -A_i(y,t')\sum_i \frac{\partial}{\partial y_i}[p(x,t|y,t')] - \sum_{i,j}\frac{1}{2}B_{ij}(y,t')\frac{\partial^2}{\partial y_i \partial y_j}[p(x,t|y,t')] + \int dz[W(z|y,t')p(x,t|y,t') - W(z|y,t')p(x,t|z,t')] \quad \text{(A.7)}$$

one can obtain the equation from various techniques given the forward Chapman-Kolmogorov equation.

Definition A.2.0.4 (**Fokker-Planck equation**)**.** Consider a time homogeneous random process $X(t) \in \mathbb{R}^d$ with drift vector b(t) and diffusion matrix Σ(t). Consider the ($\mathbb{R}-$ *valued*) probability density p(x,t)$\in$ $C^{2,1}(R^d \times R^+)$. We can then get

the *Fokker-Planck* equation by imposing continuity in the backwards differential Chapman-Kolmogorov equation, getting, in compact form:

$$\frac{\partial p}{\partial t} = \nabla \cdot (-A(x)p - \frac{1}{2}\nabla \cdot (B(x)p)) \tag{A.8}$$

If we specify the initial value p(x,0)=$\rho_0(x)$ then we can solve the initial value problem and get the transition probability density.

The Fokker-Planck equation enables us to calculate transition probability densities, which in turn we can use to compute expectation values of observables of the process

A valuable definition for the behaviour of a stochastic system at equilibrium, which will frequently use, is the following:

Definition A.2.0.5 (**Stationary distribution and Stationary process**). A stochastic process is said to be *stationary* if

$$\forall k \in \mathbb{N}, t \in R^+$$
$$P(X_{t1} \in A_1, ..., X_{tk} \in A_k) = P(X_{t_1+s} \in A_1, ..., X_{t_k+s} \in A_k)\, \forall s \tag{A.9}$$

Observation A.2.0.5. As a consequence of (A.2.0.5) we have that the one time probability distribution p_s is independent of t, the two-time joint probability distribution is $p_s(x_1, t_1 - t_2; x_2, 0)$ and the conditional probability is $p_s(x_1, t_1 - t_2|x_2, 0)$, dependent only on the time differences.

A.3 Stochastic differential equations

We now want to explicitly address stochastic differential equations of the form

$$\frac{dX(t)}{dt} = A(t, X(t)) + B(t, X(t))\xi(t) \tag{A.10}$$

where $X(t) \in \mathbb{R}^d$, $b : [0, T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$, $\sigma : [0, T] \times \mathbb{R}^d \mapsto \mathbb{R}^{d\times m}$. $\xi(t) = \frac{dW}{dt}$ is the white noise process, the formal derivative of the brownian motion. We are now faced with the fact that the brownian motion in nowhere differentiable (as it can easily be seen given that brownian increments are characterized by $(W(s)-W(t)) \sim \mathcal{N}(0, s-t) \forall s > t, s, t \in \mathbb{R}^+$). One of the possible approaches is to cast the differential equation into a suitable form and find weaker solutions to the new problem; namely, we define equation (A.10) as the stochastic integral equation

$$X(t) = x + \int_0^t A(s, X(s))ds + \int_0^t B(s, X(s))dW(s) \tag{A.11}$$

with the equivalent *formal* expression as

$$dX(t) = A(t, X(t))dt + B(t, X(t))dW(t) \tag{A.12}$$

The next step consist in giving a method to construct integrals of the form

$$I(t) := \int_0^t h(t')dW(t') \tag{A.13}$$

Due to the nature of Brownian increments (specifically tied to Hölder's continuity of the integrand and the integrator), we can't define the integral (A.13) in a Riemann–Stieltjes way uniquely.

Definition A.3.0.1 (**Riemann sum approximation; Ito and Stratonovich prescriptions**)**.** Given the interval [0,T] $\subset \mathbb{R}$, we consider the partition $t_k = k\Delta t$, k = 0,..., K-1, $K\Delta t = T$; defining the parameter $\lambda \in [0, 1]$ and setting $\tau_k = (1-\lambda)t_k + \lambda t_{k+1}$, k=0,...,K-1, we define the integral (A.13) as the Riemann sum approximation

$$I(t) := \lim_{k \longrightarrow \infty} \sum_{k=0}^{K-1} f(\tau_k)(W(t_{k+1}) - W(t_k)) \tag{A.14}$$

where $f(\cdot)$ is square integrable and adapted to the filtration generated by the brownian motion. $\lambda = 0$ corresponds to the *Ito prescription* I_I, while $\lambda = 1/2$ to the *Stratonovich* one I_S.

One can easily see ([24]) that if there exist a $\delta > 0$ such that $\mathbb{E}(f(s) - f(t))^2 \leq C|t - s|^{1+\delta}$, then the convergence is independent of λ's value.

It turns out that the Ito and Stratonovich prescriptions are equivalent, i.e. we can convert an Ito SDE into a Stratonovich SDE and vice versa. This, as many other results in stochastic calculus like *Ito's lemma*, stems from the following property:

Proposition A.3.0.1 (**Behaviour of** dW^{N+2})**.** *Given a nonanticipating function G(t), i.e. a function which is statistically independent from increments of the Brownian motion, we have the following equalities:*

$$\begin{aligned}\int_{t_0}^{t} G(t')dW(t')^{N+2} = \text{ms}- \lim_{n \longrightarrow \infty} \sum_{i=0}^{n} G_{i-1}\Delta W_i^{2+N} = 0 \ \ if \ N \gneqq 0 \\ = \int_{t_0}^{t} G(t')dt' \ \ if \ N = 0 \\ - \\ \int_{t_0}^{t} G(t')dt'dW(t') = 0\end{aligned} \tag{A.15}$$

For a proof, see [13].

The previous result only holds in the the Ito prescription, given that the proof requires the independence of G_{i-1} and ΔW which is not guaranteed by the other hypothesis. This prompts a discussion about which choice is more suitable ([3]):

- Intuitively, Ito's formulation reflects the construction of the continuous stochastic process as the limit of a discrete one. The previous result allows us to easily compute expected values. On the other end, as we'll shortly see, the chain rule and the integral calculus of polynomials have to be modified.
- Stratonovich's formulation does not have the the proof-wise advantages of Ito's, but it obeys the rules of ordinary calculus. Moreover if instead of Brownian motion we choose a noise source with finite time correlation scale, then

we get a sequence of random ODE's whose limit is a Stratonovich SDE (for a rigorous proof see [24]).

Now, given a d-dimensional Ito SDE $dX_t = A(X_t)_I dt + B(X_t)_I dW_t$, using the definition and the above formula is straightforward to prove that

$$(A_t)_S = A(X_t)_I + h(X_t) \tag{A.16}$$

(from which we can deduce the inverse relation linking a Stratonovich SDE to an Ito SDE), where

$$h_i(t) = 1/2 \sum_{j=1}^{d} \sum_{k=1}^{m} B_{jk}(x) \frac{\partial B_{ik}}{\partial x_j}(x) \tag{A.17}$$

We can now state one of the most important results in stochastic calculus

Proposition A.3.0.2 (**Ito's lemma**). *Let X_t be the solution of the SDE with drift vector $A(X_t)$ and diffusion matrix $B(X_t)$. Then, under some technical conditions (see [24]), the process $V(X_t)$ satisfies*

$$dV(t, X_t) = \frac{\partial X_t}{\partial t} dt + \sum_{i=1}^{d} \frac{\partial V}{\partial x_i} dX_i + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2 V}{\partial x_i \partial x_j} dX_i dX_j \tag{A.18}$$

Where the properties in (A.3.0.1) hold.

On a side note, it's straightforward to show (see [13]) the complete equivalence between a Fokker-Planck equation with drift vector $A(t, X_t)$ and diffusion matrix $B(t, X_t)$ and the SDE $A(t, X_t)dt + \sqrt{2B(x, X_t)}dX_t$ in the Ito framework.

Appendix B

Derivation of Kramer's law in one dimension

Proposition B.0.0.1 (**Boundary value problem for the MFET**). *Consider a bounded region $D \subset \mathbb{R}^d$ with ∂D smooth and absorbing. Then the MFET τ satisfies the boundary value problem*

$$-\mathscr{L}\tau = 1 \text{ if } x \in \overset{\circ}{D} \tag{B.1}$$

$$\tau = 0 \text{ if } x \in \partial D \tag{B.2}$$

for a proof, see [Pavliotis].

Now, for a general diffusion process in one dimension we have $\mathscr{L} = b(x)\frac{d}{dx} + \frac{1}{2}\sigma(x)\frac{d^2}{dx^2}$ and, for the specific process (3.1) we can write $\mathscr{L} = -\mathrm{L}'(x)\frac{d}{dx} + \frac{1}{\beta}\frac{d^2}{dx^2}$ with $D = \frac{1}{\beta}$. The potential L(x) is a double well, which can analytically be described by, for example, $L(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2 + \frac{1}{2}$. Thanks to the special structure of the drift term we can cast the boundary value problem in the form

$$-\frac{1}{\beta}e^{\beta L(x)}\frac{d}{dx}\left(e^{-\beta L(x)}\frac{d}{dx}\tau\right) = 1 \tag{B.3}$$

This equation can be solved by a double integration. We have the following situation for the escape problem, which we repeat for the sake of clarity (see Figure 3.1): we have two local minima at a and b, with a local maximum at b, a<b<c, and we solve the equation in the interval (a,b). Since the potential is confining, we consider a as a reflecting barrier, and since b can be crossed to go into c, b is an absorbing barrier. We get

$$\tau(x) = \beta\int_x^b dy e^{-\beta L(y)}\int_a^y dz e^{-\beta L(z)} \approx \beta\int_x^b dy e^{-\beta L(y)}\int_{-\infty}^y dz e^{-\beta L(z)} \tag{B.4}$$

where the second approximation holds since the potential grows sufficiently fast at infinity. If $\Delta L\beta = [L(b) - L(a)]\beta \gg 1$ i.e. in the low noise limit (from a physical point of view where $\beta = \frac{1}{k_b T}$, a low temperature limit), the second integral is dominated by the value of the potential around a: one can analytically study

the behaviour of $\beta L(x)$ and quantify ΔL or simply notice that β "stretches" L(x), so that, when considering the integral of $e^{-\beta L(z)}$ over (a,b), the mass (over $\mathbb{R}$) is concentrated around the minimum of L The same argument can be made for $e^{\beta L(z)}$ and the maximum, with the exception that, not being interested in the half of the function after b, using symmetry we integrate over $\mathbb{R}$ and divide by two the result. Expanding the potential around the two points up to order two, we get

$$\tau(x) \approx \frac{1}{2}\int_{-\infty}^{+\infty} e^{\beta L(b)} e^{-\frac{\beta\omega_b^2}{2}(y-b)^2} dy \int_{-\infty}^{+\infty} e^{-\beta L(a)} e^{-\frac{\beta\omega_a^2}{2}(y-b)^2} dy = \frac{\pi}{\omega_a\omega_b} e^{\beta\Delta L}. \quad \text{(B.5)}$$

Solving the gaussian integrals we get an estimate for the MFET independent of the starting position x, provided that $x \in (a,b)$. If a particle reaches b, it has only a 50% chance of crossing into c, so the ***Kramer escape rate*** is

$$k = \frac{1}{2\tau} = \frac{\omega_a\omega_b}{2\pi} e^{-\beta\Delta L}. \quad \text{(B.6)}$$

Bibliography

[1] Shishir Adhikari et al. *Machine learning in and out of equilibrium*. 2023. arXiv: 2306.03521 [cs.LG].

[2] Daniel Arovas. *Lecture Notes on Nonequilibrium Statistical Physics (A Work in Progress)*. URL: https://courses.physics.ucsd.edu/2013/Fall/physics210b/LECTURES/STOCHASTIC.pdf.

[3] Daniel Arovas. *The Itô Stratonovich integrals*. URL: https://www.robots.ox.ac.uk/~lsgs/resources/ito-strat.pdf.

[4] Richard Bellman. *Dynamic Programming*. Dover Publications, 1957. ISBN: 9780486428093.

[5] Arjun Berera et al. "Formulating the Kramers problem in field theory". In: *Physical Review D* 100.7 (Oct. 2019). ISSN: 2470-0029. DOI: 10.1103/physrevd.100.076005. URL: http://dx.doi.org/10.1103/PhysRevD.100.076005.

[6] Nils Berglund. *Kramers' law: Validity, derivations and generalizations*. 2013. arXiv: 1106.5799 [math.PR].

[7] Léon Bottou. "Stochastic Gradient Learning in Neural Networks". In: 1991. URL: https://api.semanticscholar.org/CorpusID:12410481.

[8] Freddy Bouchet, Joran Rolland, and Jeroen Wouters. "Rare Event Sampling Methods". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29.8 (Aug. 2019), p. 080402. ISSN: 1054-1500. DOI: 10.1063/1.5120509. eprint: https://pubs.aip.org/aip/cha/article-pdf/doi/10.1063/1.5120509/14620316/080402_1_online.pdf. URL: https://doi.org/10.1063/1.5120509.

[9] Vincenzo Capasso and David Bakstein. "Fundamentals of Probability". In: *An Introduction to Continuous-Time Stochastic Processes: Theory, Models, and Applications to Finance, Biology, and Medicine*. New York, NY: Springer New York, 2015. ISBN: 978-1-4939-2757-9. DOI: 10.1007/978-1-4939-2757-9_1. URL: https://doi.org/10.1007/978-1-4939-2757-9_1.

[10] Abhishek Dhar. *On Kramer's escape rate problem*. URL: https://home.icts.res.in/~abhi/notes/kram.pdf.

[11] Simon S. Du et al. *Gradient Descent Can Take Exponential Time to Escape Saddle Points*. 2017. arXiv: 1705.10412 [math.OC].

[12] M. I. Freidlin and A. D. Wentzell. "Stability Under Random Perturbations". In: *Random Perturbations of Dynamical Systems*. New York, NY: Springer New York, 1998. ISBN: 978-1-4612-0611-8. DOI: 10.1007/978-1-4612-0611-8_10. URL: https://doi.org/10.1007/978-1-4612-0611-8_10.

[13] C. W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*. Third. Vol. 13. Springer Series in Synergetics. Berlin: Springer-Verlag, 2004, pp. xviii+415. ISBN: 3-540-20882-8.

[14] Allan Gut. "Multivariate Random Variables". In: *An Intermediate Course in Probability*. New York, NY: Springer New York, 2009. ISBN: 978-1-4419-0162-0. DOI: 10.1007/978-1-4419-0162-0_1. URL: https://doi.org/10.1007/978-1-4419-0162-0_1.

[15] Wenqing Hu et al. *On the diffusion approximation of nonconvex stochastic gradient descent*. 2018. arXiv: 1705.07562 [stat.ML].

[16] Stanisław Jastrzębski et al. *Three Factors Influencing Minima in SGD*. 2018. arXiv: 1711.04623 [cs.LG].

[17] N.G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier Science Publishers, Amsterdam, 1992.

[18] H.A. Kramers. "Brownian motion in a field of force and the diffusion model of chemical reactions". In: *Physica* 7.4 (1940), pp. 284–304. ISSN: 0031-8914. DOI: https://doi.org/10.1016/S0031-8914(40)90098-2. URL: https://www.sciencedirect.com/science/article/pii/S0031891440900982.

[19] Qianxiao Li, Cheng Tai, and Weinan E. *Stochastic modified equations and adaptive stochastic gradient algorithms*. 2017. arXiv: 1511.06251 [cs.LG].

[20] Pankaj Mehta et al. "A high-bias, low-variance introduction to Machine Learning for physicists". In: *Physics Reports* 810 (2019). A high-bias, low-variance introduction to Machine Learning for physicists, pp. 1–124. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2019.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0370157319300766.

[21] Takashi Mori et al. *Power-law escape rate of SGD*. 2022. arXiv: 2105.09557 [cs.LG].

[22] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications (Universitext)*. 6th. Springer, Jan. 2014. ISBN: 3540047581. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20%5C&path=ASIN/3540047581.

[23] Vardan Papyan. *The Full Spectrum of Deepnet Hessians at Scale: Dynamics with SGD Training and Sample Size*. 2019. arXiv: 1811.07062 [cs.LG].

[24] Grigorios A. Pavliotis. "Introduction to Stochastic Processes". In: *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. New York, NY: Springer New York, 2014. ISBN: 978-1-4939-1323-7. DOI: 10.1007/978-1-4939-1323-7_1. URL: https://doi.org/10.1007/978-1-4939-1323-7_1.

[25] Ravid Shwartz-Ziv and Naftali Tishby. *Opening the Black Box of Deep Neural Networks via Information*. 2017. arXiv: 1703.00810 [cs.LG].

[26] Zeke Xie, Issei Sato, and Masashi Sugiyama. *A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima*. 2021. arXiv: 2002.03495 [cs.LG].