**From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Mathematics

2025

# Long-Time Dynamics of Differential Equations in Physics and AI

Jin, Kexin

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp014j03d307p

*U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf*

# Long-Time Dynamics of Differential Equations in Physics and AI

Kexin Jin

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Mathematics
Adviser: Alexandru Ionescu

March 2025

© Copyright by Kexin Jin, 2025.

All rights reserved.

# Abstract

This thesis consists of two parts: the study of long-time dynamics of differential equations derived from 1) physics, and 2) machine learning.

In the first part, we study the cubic nonlinear Schrödinger (NLS) equation with periodic boundary conditions defined on $[0, L]$. By proving a vanishing property of the normal form transformation and applying it to the quintic resonance interactions, we obtain a description of the dynamics for a time up to $T = L^2/\varepsilon^4$, where $\epsilon$ is the size of the initial data. Since $T$ is the characteristic time of wave turbulence, this result implies the absence of wave turbulence behavior of the 1D cubic NLS. In the proof, we develop a correspondence between Feynman diagrams and terms in normal forms, which allows us to calculate the coefficients inductively. Notably, our approach can be adapted to other integrable systems with minimal difficulty.

In the second part, we consider differential equations in machine learning algorithms. We develop a novel approach to model discrete-time machine learning optimization algorithms as continuous-time dynamics, including stochastic gradient descent (SGD) and its variants. We propose the stochastic gradient process consists in a gradient flow minimizing an indexed target function that is coupled with a continuous-time index process determining the index. Index processes are, e.g., reflected diffusions, pure jump processes, or other Lévy processes on compact spaces. We analyze the approximation properties of the stochastic gradient process and study its long-time behavior and ergodicity under constant and decreasing learning rates. We illustrate the applicability of the stochastic gradient process in a polynomial regression problem with noisy functional data, as well as in a physics-informed neural network.

# Acknowledgements

I would like to express my sincere appreciation to my Ph.D. advisor, Professor Alexandru Ionescu, for his encouragement, patience, and mentorship throughout my Ph.D. This journey would not have been possible without his constant support, especially during the challenging times of the COVID-19 pandemic, when uncertainty and obstacles made progress particularly difficult. Professor Ionescu is truly one of the best mathematicians I have ever had the privilege of working with, and I feel incredibly fortunate to have had him as my advisor. His expertise, guidance, and dedication have been invaluable, and I am grateful for everything he has taught me both academically and personally.

I am also deeply grateful to my co-advisor, Professor Tristan Buckmaster. I sincerely appreciate his unwavering encouragement and insightful guidance during moments of uncertainty and challenge, which provided me with the confidence and direction to persevere. Moreover, Professor Buckmaster has shown me how modern machine learning and mathematics can beautifully intersect, expanding my perspective and shaping my research in profound ways. His ability to connect these fields has been instrumental in guiding my work, and I am truly thankful for his mentorship and support.

I would like to extend my heartfelt thanks to my Master's thesis advisor, Professor Jalal Shatah, whose mentorship has been a cornerstone of my academic journey. He not only introduced me to the fascinating world of mathematics and partial differential equations but also provided me with invaluable guidance and support. Beyond his academic mentorship, Professor Shatah has been a source of emotional support, treating me like family during my time in the US. His encouragement, wisdom, and care have been pivotal in shaping both my research and my experience as an international student.

I am incredibly grateful to Professor Elad Hazan, who introduced me to the world of machine learning optimization. His class was the first I took in machine learning, and it ignited my passion for the field. Professor Hazan's expertise, clear teaching, and thought-provoking insights have deeply influenced my academic path. His mentorship has not only

shaped my understanding of optimization but has also provided me with the foundation to pursue advanced research in this area. I am truly thankful for the opportunity to learn from him.

I would also like to express my gratitude to the faculty at the Courant Institute of Mathematical Sciences for their continuous help and encouragement. Their guidance played a crucial role in inspiring me to pursue and begin this PhD journey. My thanks go to Professors Scott Armstrong, Yuri Bakhtin, Gérard Ben Arous, Paul Bourgade, Jeff Cheeger, Pierre Germain, Sinan Gunturk, Fang-Hua Lin, Eyal Lubetzky, Henry McKean, Sylvia Serfaty, and Vlad Vicol.

I am profoundly grateful to my dear friend and collaborator, Jonas Latz, whose unwavering support, collaboration, and infectious humor have made this journey not only productive but also truly enjoyable. His intellectual contributions have been invaluable, and his willingness to engage in thoughtful discussions has enriched my research in countless ways. Beyond our professional partnership, Jonas's friendship has provided much-needed encouragement during challenging times, making this journey even more memorable.

I am deeply grateful to my dear friend Chenguang Liu, with whom I have shared a friendship since our college days. Chenguang has always been there to offer his help, guidance, and encouragement throughout my academic journey. His insightful perspectives, collaborative spirit, and unwavering support have been invaluable to the progress of my research. I greatly appreciate his willingness to always lend a hand, and our partnership has made this journey both productive and rewarding. I am fortunate to have him as a friend and collaborator.

I am deeply grateful to Xiao Ma, who has been an incredible collaborator and friend since we met at Princeton. We have spent countless late nights working side by side, when the building was quiet and everyone else had gone home. These late-night sessions, filled with problem-solving and idea exchange, were some of the most rewarding moments of my PhD journey. Xiao's dedication, insight, and continuous support have greatly contributed to the

progress of my research. I am truly thankful for his friendship, his valuable collaboration, and the memories we've created together.

I would like to extend my heartfelt thanks to my dear friends Xuecheng Wang, Fan Zheng, Mengran Wang, Ruijie Fang, Can Cui, Wenhao Zhan, Xue Wang, Zihan Ding, and many others I've had the pleasure of encountering, for their friendship and unwavering support throughout this journey. Your presence has been a constant source of joy, comfort, and motivation. Whether we were sharing laughs, discussing ideas, or simply offering each other encouragement during tough times, you all made this experience much richer and more enjoyable. I am truly fortunate to have such supportive and caring friends, and I deeply appreciate the bonds we've built over the years.

Finally, I am deeply indebted to my family for their love, support, and encouragement throughout this entire journey. Their belief in me, even during the most challenging times, has been a constant source of strength. To my parents, your sacrifices and dedication have shaped the person I am today. Your encouragement has always inspired me to push through obstacles, and your endless patience has allowed me to focus on my work. I am truly blessed to have such a loving and supportive family, and I am forever grateful for the sacrifices you've made and the love you've given me every step of the way.

To my family and friends.

# Contents


Abstract . . . . . 3

Acknowledgements . . . . . 4

List of Tables . . . . . 12

List of Figures . . . . . 13

**1 Introduction** **15**

1.1 The Schrödinger equation . . . . . 16

1.2 Machine learning optimization . . . . . 17

1.3 Organization . . . . . 18

**2 The cubic Schrödinger equation** **19**

2.1 Introduction . . . . . 19

2.1.1 The times scales of weakly nonlinear large box limit . . . . . 21

2.1.2 Modified scattering . . . . . 23

2.1.3 A vanishing property of normal form transformation . . . . . 24

2.1.4 Other relevant techniques . . . . . 26

2.1.5 Organization and notations . . . . . 26

2.2 Normal form transformation . . . . . 29

2.2.1 Extraction of the modified phase . . . . . 30

2.2.2 Normal form transformation of the cubic term . . . . . 32

2.2.3 Higher order normal form transformations . . . . . 36

2.3 The Feynman diagrams and cancellations of resonances . . . . . 42
2.3.1 The partition trees . . . . . 43
2.3.2 Feynman diagrams and repeated normal form transformations . . . . 45
2.3.3 Correspondence of trees and terms . . . . . 49
2.3.4 A vanishing property of normal form coefficients . . . . . 54
2.4 The long time dynamics of 1D cubic NLS . . . . . 63
2.4.1 Ideas of the proof . . . . . 63
2.4.2 Identification of the main term . . . . . 68
2.4.3 Estimate of errors in normal form transformation . . . . . 70
2.4.4 Dynamics of $d_K$ . . . . . 90
2.4.5 A refined estimate of $|d_K|^2$ . . . . . 94
2.4.6 Wrapping up . . . . . 101

**3 Continuous-time stochastic gradient descent: the stochastic gradient process 104**

3.1 Introduction . . . . . 104
3.1.1 Problem setting and state of the art . . . . . 106
3.1.2 The stochastic gradient process . . . . . 108
3.2 The index process: Feller processes and geometric ergodicity . . . . . 112
3.2.1 Example 1: Lévy processes with two-sided reflection . . . . . 116
3.2.2 Example 2: Continuous-time Markov processes with finite states . . . 117
3.2.3 Example 3: Continuous-time Markov processes with countable states 119
3.2.4 Example 4: Multidimensional processes . . . . . 120
3.3 Stochastic gradient processes with constant learning rate . . . . . 122
3.3.1 Approximation of the full gradient flow . . . . . 125
3.3.2 Longtime behavior and ergodicity . . . . . 132
3.4 Stochastic gradient processes with decreasing learning rate . . . . . 142
3.5 From continuous dynamics to practical optimization. . . . . 149

3.5.1 Discretization of the index process . . . 149
3.5.2 Discretization of the gradient flow . . . 151
3.6 Applications . . . 152
3.6.1 Polynomial regression with functional data . . . 153
3.6.2 Solving partial differential equations using neural networks (NNs) . . . 159
3.7 Conclusions and outlook . . . 166

**4 The stochastic gradient-momentum process 167**
4.1 Introduction . . . 167
4.1.1 Optimization and continuous dynamics . . . 167
4.1.2 Momentum . . . 169
4.1.3 Momentum-based stochastic optimization and the Adam algorithm . . . 172
4.1.4 The stochastic gradient-momentum process . . . 173
4.1.5 Contributions and main results . . . 178
4.2 Homogeneous-in-time . . . 181
4.2.1 Momentum with and without subsampling . . . 182
4.2.2 Subsampling with and without momentum . . . 184
4.3 Heterogeneous-in-time . . . 189
4.3.1 Losing momentum over time . . . 189
4.3.2 Losing, both, momentum and randomness over time . . . 202
4.4 discretization of the continuous-time system . . . 209
4.4.1 Step-size choice . . . 211
4.4.2 randomized version of the method . . . 213
4.5 Numerical Experiments . . . 214
4.5.1 One-dimensional non-smooth stationary point . . . 215
4.5.2 Strongly convex example: quadratic objective . . . 216
4.5.3 Non-convex example: polynomial function . . . 217
4.5.4 Convolutional Neural Network (CNN) . . . 218

4.6 Conclusions . . . . . 220

**A Chapter 4 Appendix** **223**

A.1 Longtime behavior of the underdamped gradient flow . . . . . 223

A.2 Decreasing momentum: the fixed-sample case . . . . . 225

A.3 Other auxiliary results . . . . . 232

A.4 Kushner's perturbed test function theory . . . . . 239

**Bibliography** **241**

# List of Tables

3.1 Accuracy of the estimation in the polynomial regression model . . . . . 158

4.1 Comparison of train and test accuracy (%) with different hyperparameters over the CIFAR-10 dataset. . . . . 221

# List of Figures


2.1 An example of trees defined by a chain of refinements . . . . . 44

2.2 Trees with labels omitted . . . . . 45

2.3 An example of Feynman diagram . . . . . 48

2.4 Another example of Feynman diagram . . . . . 49

3.1 Cartoon of the stochastic gradient process on the discrete index set . . . . . 107

3.2 Cartoon of the stochastic gradient process with continuous index set . . . . . 109

3.3 Polynomial regression example . . . . . 153

3.4 Estimation results of the polynomial regression problem using stochastic gradient descent . . . . . 154

3.5 Estimation results of the polynomial regression problem using the stochastic gradient process with reflected Brownian motion . . . . . 156

3.6 Estimation results of the polynomial regression problem using the stochastic gradient process with pure jump index process . . . . . 157

3.7 Histograms of single simulations of the index process . . . . . 158

3.8 Comparison of the mean rel_err of the stochastic methods for time $t \leq 200$. . 159

3.9 Loss vs iteration and its log scale for SGD, SGPC, and SGPD . . . . . 165

3.10 Solutions at time $t = 0.1, 0.5, 0.9$ . . . . . 165

3.11 Test error at time $t = 0.1, 0.5, 0.9$ . . . . . 165

4.1 Plot of the potential $\Phi$. . . . . 170

4.2 Minimization example . . . . . 175
4.3 Schematic comparing gradient flow, underdamped gradient flow, stochastic gradient process, and stochastic gradient-momentum process in terms of the particle mass $m$ and learning rate $\nu$. . . . . 180
4.4 Convergence analysis of 1D example . . . . . 216
4.5 Convergence rate comparison . . . . . 217
4.6 Decay of objective comparison . . . . . 218
4.7 Train loss comparison for SGD ($\eta = 0.01$), classical momentum ($\eta = 0.01, \rho = 0.9$), and hSGMP for CNN on CIFAR-10 . . . . . 220
4.8 Train loss comparison for SGD ($\eta = 0.01$), classical momentum ($\eta = 0.01, \rho = 0.9$), hSGMP ($\alpha = 10, \varepsilon = 0.1, h = 1$), and dmSGMP ($\alpha = 10, \varepsilon_0 = 0.1, m = m_0 0.995^k, h = 1$) for CNN on CIFAR-10 . . . . . 221

A.1 Plot of function $\bar{\Phi}$ in Example 3. . . . . 232

# Chapter 1

# Introduction

Studying real-world problems in a rigorous manner often requires the use of mathematics. Modeling large and complex systems involves both the adaptation of established mathematical concepts and the development of novel approaches. This process includes describing the fundamental relationships and key features of a system through equations and other mathematical structures. Such models serve as a bridge between abstract mathematical theory and practical applications, enabling us to simulate, analyze, and predict system behavior under varying conditions.

Such systems arise not only in traditional disciplines like physics, biology, and chemistry but also in rapidly advancing fields such as finance, medicine, and artificial intelligence. In each of these domains, mathematical models play a critical role. For instance, in finance, models are used to manage risk and optimize investment strategies. In medicine, they assist in understanding disease dynamics and creating personalized treatments. In artificial intelligence, mathematical tools form the foundation of machine learning algorithms, driving progress in areas like natural language processing, computer vision, and robotics.

In this thesis, we focus on problems derived from physics and artificial intelligence. Physics has a rich history of using mathematical models to deepen our understanding of the natural world, with equations such as Newton's laws, Maxwell's equations, and the

Schrödinger equation. In artificial intelligence, the emphasis lies on creating functional models capable of solving specific tasks, often prioritizing performance, scalability, and adaptability over interpretability.

## 1.1 The Schrödinger equation

In physics, equations are often derived using physical laws and symmetry principles, which typically lead to systems of partial differential equations. For example, the Schrödinger equation,

$$\begin{cases} -i\partial_t v + \Delta v = \pm|v|^2 v, \\ v(t=0) = v_0, \end{cases}$$

which predicts the behavior of quantum systems, is constructed using principles such as energy conservation, wave-particle duality, and Planck's relation. This equation has various applications: it can explain the energy levels of electrons, calculate the electronic properties of molecules, and describe molecular structures. In the following chapter, we will focus on the wave turbulence phenomenon and rigorously justify what physicists have conjectured. This will showcase the role of mathematics in advancing our understanding of complex physical phenomena.

In particular, we shall study the nonlinear Schrödinger equation defined on a torus of length $L$ with small initial data. The solution to this equation is analyzed in different time scales, with each time interval allowing for an specific approximation. We study it up to the characteristic time scale of the wave turbulence. By applying the normal form transformation and discuss the corresponding Feynman diagrams, we obtain a precise description of the dynamic.

## 1.2 Machine learning optimization

While mathematical modeling is a classical subject in physics, the fast-developing field of artificial intelligence faces the challenge of a lack of comprehensive mathematical theory. In this context, we focus on optimization algorithms in machine learning. Unlike classical optimization problems, where the objective is to find the minimizer or maximizer of a given function over a specified domain, machine learning optimization introduces an additional layer of complexity. Here, the goal is not only to optimize the performance on the given set (the training set) but also to ensure generalization on an unknown set (the test set). To address this, stochastic noise is often introduced to mitigate overfitting on the training set.

A widely used method in this context is the stochastic gradient descent (SGD) algorithm. In this thesis, we introduce the stochastic gradient process as our model and analyze its long-term behavior. By leveraging the weak form of Harris' theorem, we establish convergence to the invariant measure, providing rigorous convergence guarantees for the associated optimization algorithm. Furthermore, by discretizing the continuous-time dynamic, we derive an subsampling algorithm, offering improved efficiency in practical applications, such as physics-informed neural networks.

In addition, we extend our model to incorporate a variant of SGD—SGD with momentum—using our stochastic gradient process. Momentum, a technique designed to accelerate convergence and smooth optimization trajectories, introduces an additional velocity term that influences the dynamics of the algorithm. By analyzing the relationship between the continuous-time dynamics of standard SGD and SGD with momentum, we gain deeper insights into how these two methods interact. By studying the long-time behavior of the stochastic gradient process with momentum, we rigorously establish the convergence of the dynamics to the minimizer. These findings provide a theoretical foundation for understanding the advantages and trade-offs of incorporating momentum in optimization algorithms.

This serves as our second example of how mathematical methods can effectively address real-world challenges.

## 1.3 Organization

This thesis is organized as follows. Chapter 2 focuses on the analysis of the nonlinear Schrödinger equation and its vanishing property. In Chapter 3, we explore the stochastic gradient process, analyzing its long-term behavior and establishing its ergodic properties. Chapter 4 extends this discussion to the stochastic gradient process with momentum, examining its dynamics and demonstrating its convergence to the standard stochastic gradient process.

# Chapter 2

# The cubic Schrödinger equation

## 2.1 Introduction

We consider the nonlinear Schrödinger (NLS) equation with cubic nonlinearity on $[0, L]$ with periodic boundary conditions:

$$\begin{cases} -i\partial_t v + \frac{1}{2\pi}\partial_{xx} v = \sigma |v|^2 v, & x \in \mathbb{T}_L = [0, L], \\ v(t=0) = \epsilon v_0, \end{cases} \tag{2.1}$$

where $\sigma = \pm 1$ denotes the focusing case or the defocusing case. This equation is a fundamental equation in nonlinear optics and Bose-Einstein condensation. For the case when the equation is posted on $\mathbb{R}$, for example, Zakharov-Shabat [133] using the inverse scattering method discovered many mathematical results. On $\mathbb{T}$, for example, in [15], Bourgain used number theory and restricted norm methods to derive the Strichartz estimates and local well-posedness in $L^2(\mathbb{T})$ (local well-posedness in $H^s(\mathbb{T})$ with $s > \frac{1}{2}$ is obvious and global well-posedness follows from energy conservation laws in the defocusing case). For more refined local well-posedness results see for example [30], [44] and [61].

Once well-posedness is established, the next step is to describe the long time dynamics of these solutions. For example, in [17], Bourgain constructed an invariant measure for 1D

NLS, which is the preliminary of any ergodic theory for this equation; In [16], he applied symplectic nonsqueezing to exclude asymptotic stability of NLS on $\mathbb{T}$. He also initiated the construction of quasi-periodic in time solutions. See, for example, [103] for a recent result of time-quasi-periodic solutions.

In this chapter, we shall describe the long dynamics on a torus of length $L$, $\mathbb{T}_L$. The large box long time dynamics of dispersive PDEs has been studied by many authors, for example, [23], [22], [45], [137], and [138]. In what follows, it is more convenient to consider the transformation $v = \epsilon u$. Without loss of generality, we take $\sigma = 1$. Thus, $u$ satisfies

$$\begin{cases} -i\partial_t u + \frac{1}{2\pi}\partial_{xx} u = \epsilon^2 |u|^2 u, & x \in \mathbb{T}_L, \\ u(t=0) = u_0. \end{cases} \tag{2.2}$$

For $u$, we have the following result.

**Theorem 1.** *Fix $\ell > 1$, $0 < \gamma \ll 1$, and $M > 0$. Let $e(z) := e^{2\pi i z}$. Let $f_0 \in X^{\ell,2}$ and $B = \|f_0\|_{X^{\ell,2}(\mathbb{R})}$. Let $u$ be a solution of NLS (2.2) with initial data $u_0(x) = \frac{1}{L}\sum_{\mathbb{Z}_L} f_0(K)e(Kx)$ and Fourier coefficients $u_K(t)$. Then, for $L$ sufficiently large and $\epsilon^2 L^\gamma$ sufficiently small, depending on $M$, $B$, there exists a constant $C_\gamma$, such that for any $t \in [0, MT_R]$, $T_R = \frac{L^2}{\epsilon^4}$,*

$$\begin{aligned} &\left\| u_K(t) - f_0(K) e\left( \left( K^2 + \frac{\epsilon^2}{\pi L}\|u\|^2_{L^2(\mathbb{T})} + \frac{\epsilon^2}{\pi L^2}|f_0(K)|^4 \right) t - \frac{\epsilon^2}{2\pi L^2}\int_0^t Q_K(f)(s)ds \right) \right\|_{X^\ell(\mathbb{Z}_L)} \\ &\lesssim C_\gamma(\epsilon^2 L^\gamma + \frac{1}{L^{1-\gamma}}). \end{aligned} \tag{2.3}$$

*where $Q_K(f)$ is a complicated but explicit function of $f(K)$ defined in Proposition 2, $f(K)$ is defined by*

$$f(K) = e(\frac{1}{2\pi}|f_0(K)|^4 t) f_0(K),$$

*$\|f\|_{X^{\ell,N}(\mathbb{R})} = \sum_{0\le|\alpha|\le N} \|\langle x\rangle^\ell \nabla^\alpha f(x)\|_{L^\infty(\mathbb{R})}$ and $\|u\|_{X^\ell(\mathbb{T}_L)} = \sup_{K\in\mathbb{Z}_L} \|\langle K\rangle^\ell u_K\|$ where $u_K$ is the Fourier coefficients of $u$.*

**Remark 1.** *Ideally, we would like to assume $\epsilon$ is a small constant independent of $L$ which can be nonzero even if $L \to \infty$, because then this theorem covers the Euclidean scatter result. However, there are some $L^\gamma$ losses with $\gamma$ being arbitrarily small in our proof, so we have to assume $\epsilon^2 < L^{-\gamma}$ and the theorem implies the Euclidean scatter result up to $\gamma$-loss. The solution $u_K(t)$ is well-approximated only if the right hand side of (2.3) is small, which requires $C_{\gamma,B,M}$ is not very large. While $M > 0$ may be chosen arbitrarily, we have a meaningful result only when it does not depend on $L$, $\epsilon$. Thus, the interval $[0, MT_R]$ is essentially of the same magnitude as $[0, T_R]$.*

**Remark 2.** *If we fix a function $u_0$ with $supp(u_0) \subseteq [-1, 1]$, we may think of it as a periodic function on $\mathbb{T}_L = [-\frac{L}{2}, \frac{L}{2}]$. If our initial data is of this form, $f_0(K)$ may be taken as the Fourier transform of $u_0$ on $\mathbb{R}$. More generally, for all spatially localized initial data, we can find $f_0(K)$ as in the assumption of Theorem 1. The existence of $f_0(K)$ is a type of spatial localization assumption.*

**Remark 3.** *Recall that for a periodic function on $[0, L]$, its Fourier coefficients constitute a discrete function defined on $\mathbb{Z}_L$. As $L$ goes to infinity, these Fourier coefficients are defined on a denser and denser set. So one may expect that there is asymptotically a continuous profile of this discrete function. Our result formulates this continuous profile.*

The following parts of the introduction are devoted to explaining three important aspects: the time in which the phenomenon described by our result occurs, what is the phenomenon in particular, and the tools to prove this phenomenon.

### 2.1.1 The times scales of weakly nonlinear large box limit

The first important aspect is the time scale $T_R$. We will explain the scales of $L$ and $\epsilon$ in $T_R = \frac{L^2}{\epsilon^4}$.

We have two trivial time scales:

• The weak nonlinearity time scale: By Duhamel's principle, we have

$$u(x,t) = e^{\frac{it}{2\pi}\partial_x^2}u_0 + \varepsilon^2 \int_0^t e^{\frac{i(t-s)}{2\pi}\partial_x^2}|u|^2 u ds.$$

If $t \in [0, \frac{1}{\varepsilon^2}]$, the local well-posedness can be established by contraction mapping arguments which also gives us $u \sim e^{\frac{it}{2\pi}\partial^2}u_0$. So when $t \in [0, \frac{1}{\varepsilon^2}]$, the dynamics of the solution on torus is similar to that of a linear equation.

• The large box time scale: We may think that in physical space the solution is supported in a ball of radius 1. Thus the dispersion relation $|\xi|^2$ tells us the speed of the wave is less than 1. Hence it takes time $O(L)$ for the solution to see the boundary effect. Therefore, when $t \leq L$, the dynamics of $u$ is similar to that of the NLS equation on $\mathbb{R}$.

In summary, when $t \leq \frac{1}{\epsilon^2}$ or $t \leq L$, the solution can be approximated either by the linear Schrödinger equation on $\mathbb{T}_L$ or the nonlinear Schrodinger equation on $\mathbb{R}$; so the dynamics are trivial. To go beyond these trivial time scales, for instance, Faou-Germain-Hani in [45] and Buckmaster-Germain-Hani-Shatah in [23] derived the *continuous resonance* (CR) equation, which describes the dynamics of NLS on rational torus in dimensions higher than 1 up to the resonant time-scale $L^2/\epsilon^2$. One could interpret the $L^2$ here as the period of the linear solution. In [23], they adapted the Hardy-Littlewood circle method to PDE setting and obtained a equidistribution result for the algebraic variety of resonance interactions. See, for instance, [22] for more analysis of the CR equation. In [24], the authors upgrade the time scale to $\frac{L^{2.65}}{\epsilon^2}$ in the higher dimensional case, given initial data with random phase on generic irrational torus. They derived the wave kinetic equation to describe the evolution of the moduli of the Fourier coefficients of the solution up to this time scale.

In this chapter, we examine the one dimensional cubic equation up to the time-scale $L^2/\epsilon^4$. Although typically the long time behavior problem is more difficult in lower dimensional space, we improve the time scale in $\epsilon$ compared with previous results. To get this improvement, we take advantage of the integrability structure of the 1D cubic NLS. Re-

markably, our time scale $T_R = L^2/\epsilon^4$ is the characteristic time scale of the wave turbulence. In this time scale, we find a quasi-periodic behavior of the Fourier coefficients of the solution. We call this behavior modified scattering because of its similarity with the results of modified scattering on Euclidean space. In following two subsections, we discuss this behavior as well as a structure of the equation leading to this behavior, which is related to integrability.

### 2.1.2 Modified scattering

Many researchers have discovered that the solutions of low dimensional dispersive PDEs with small initial data exhibit the modified scattering behavior, i.e. Fourier transform of the solution converges to the linear solution with a phase correction, contrary to the usual scattering results in which the limit is the linear solution without correction, see for example Kato-Pausader [62], Ionescu-Pausader [55], [56]. The asymptotic behavior in Theorem 1 can also be interpreted as modified scattering in the sense that the Fourier coefficients $u_K(t)$ of the solution "converge" to the linear solution $f_0(K)e(K^2t)$ with phase corrections, $\frac{\epsilon^2}{\pi L}\|u\|^2_{L^2(\mathbb{T})}t$, $\frac{\epsilon^2}{\pi L^2}|f_0(K)|^4t$ and $\frac{\epsilon^2}{2\pi L^2}\int_0^t Q_K(f)(s)ds$. This convergence is valid up to time $T_R = \frac{L^2}{\epsilon^4}$, the characteristic time of wave turbulence. Comparing to the wave turbulence behavior of the higher dimensional NLS in [24], the solution of 1D NLS roughly exhibits modified scattering behavior, which is much regular. This coincides with the intuition of physicists that integrable systems are more predictable and have no turbulence behavior.

Recall that in Kato and Pusateri [62], where they derived the modified scattering result of 1D cubic NLS on $\mathbb{R}$. More precisely, they proved a scattering result in the $L^\infty$ norm after doing a change of phase. The phase is of magnitude $log(t)$, which is essentially due to the $t^{-1}$ decay in the time integration of the Duhamel formula. Here, we also use the change of phase argument, which gives the phase correction terms. The last phase correction term $\frac{\epsilon^2}{2\pi L^2}\int_0^t Q_K(f)(s)ds$ is of magnitude $L^+$ after integration. The similarity of $log(t) = t^+$ and $L^+$ indicates that these two results are similar, although the proofs are quite different. This is another reason why we call the behavior of the solution here the modified scattering.

### 2.1.3 A vanishing property of normal form transformation

In this chapter, the basic tool we use to prove modified scattering is the normal form transformation. This technique was introduced by Shatah [113] in its original form of modified Poincaré normal form. A crucial improvement was made by Germain-Shatah-Masmoudi [46] with a new set of ideas called the space-time resonance, which uses integration by parts instead of the series expansion argument. This method unifies the vector field method and the Poincaré normal form transformation. These methods are replaced by integration by parts in space variables and the time variable, respectively. There are also other forms of normal form transformation, for example, the differentiation by parts approach in [12].

The normal form transformations may be applied to exploit the hidden smoothing effect in the nonlinearity to prove low regularity well-posedness results. This approach was initiated by Bourgain [15], then reinterpreted as standard normal form transformation in [18], which also pointed out the connection with $I-$method introduced in [33]. For later development, see for example, [12], [50], [95] and section 4.1 of [126].

Our main idea is to approximate our system by a simpler ODE system (2.140). The right hand side of (2.140) is extracted from the main term given by the normal form transformation. We use the differentiation by parts approach in [12] to perform this transformation. When working with this transformation to extract the main term, the fact that it is an integrable system plays a significant role because it implies the vanishing of most of the normal form coefficients. To prove Theorem 1, we only apply this vanishing property for quintic forms. But more generally, we show this vanishing property holds for arbitrary order normal form transformation. Because of the vanishing properties of higher order terms, it might be possible to control the dynamics up to a much longer time $\frac{L^3}{\epsilon^4}$, but further work is needed. For more explanation, see section 2.4.1. Roughly speaking, we prove the following theorem. For a complete version of this theorem, see Theorem 2.

**Theorem 2'.** *For nonlinear Schrodinger equation (2.2), the coefficients $H^{d+1}_{K_1...K_{2d+3}}$ of its normal form vanish for $(K_1, ..., K_{2d+3})$ outside some lower dimensional submanifolds of the*

*resonance surface $\mathscr{R}$ for any $d \leq P$. Here the resonance surface is defined by*

$$\mathscr{R} = \{(K_1, ..., K_{2d+3}) : K_1 - K_2 + \cdots + K_{2d+3} = K,\ K_1^2 - K_2^2 + \cdots + K_{2d+3}^2 = K^2\} \quad (2.4)$$

The Poincaré normal form has been known by physicists as the Dyson series, where coefficients are called collision coefficients. Collision coefficients are not rigorously defined mathematically, but it is related to normal form coefficients. For example, Zakharov and Schulman [133], [134] (Theorem 2.2.1) used perturbation expansion and S-matrix (scattering matrix) to prove that the collision coefficients vanish for any systems on $\mathbb{R}$, if it possesses an additional conservation law other than energy and momentum. Their method does not apply to bounded domains, like $\mathbb{T}$, because physically there is no scattering on these domains, so it makes no sense to talk about the S-matrix. From mathematical point of view, this failure is because one heuristic identity

$$\lim_{\epsilon \to 0} \frac{e^{iEt}}{E - i\epsilon} = \pi \delta(E),$$

is not true on $\mathbb{T}$. For the vanishing property of the quintic term, [132] provides a calculation for all common integrable system; however, there is no direct generalization of the proof to higher order terms. We prove that what Zakharov claimed (see [134]) is valid up to arbitrary order of normal form transformations, which to our knowledge is given for the first time.

The Feynman diagram is a common tool in the investigation of normal form transformation, see, for instance, [24], [85], [50] and [95]. But for our purpose, we shall develop a new formula (2.54) of normal form coefficients in terms of the Feynman diagrams. Moreover, to prove the vanishing property, we shall carefully exploit the graph structure of the Feynman diagram.

### 2.1.4 Other relevant techniques

We discuss some relevant techniques that may be of interest. For integrable systems, one may use the Birkhoff normal form and the action-angle variables to make use of the complete integrability structure, see, for example, [49], [59], and [60]. In particular, in [49], Grebert-Kappeler globally constructed the action-angle variables for the 1D defocusing NLS on $[0, 2\pi]$. Long time behavior results may be obtained from this, see, for example, [61]. For more details of the action-angle variables method on $\mathbb{T}$ under the assumption of finite gap property, see [94]. A vanishing property may be easily obtained in the case of the Birkhoff normal form. The vanishing property in our context is not obvious and we prove it by a direct calculation using the Feynman diagram.

For both of the vanishing property and the long time behavior result, it is desirable to have a method that does not rely on complete integrability and conservative quantities explicitly. The hope is to study long time behavior of equations that are not integrable. Using our approach, we derive an explicit expression for the asymptotic state in terms of the initial data, which is particularly favorable in the theory of wave turbulence to exploit the random initial data assumption.

### 2.1.5 Organization and notations

This chapter is organized in the following way: in section 2.2, we perform a change of phase argument, normal form transformation, and obtain recurrence formulas of the coefficients; in section 2.3, we introduce Feynman diagrams, use them to formulate explicit formulas of norm form coefficients, and prove the vanishing property; in section 2.4, we establish our main long time dynamics result.

We will use following notations frequently:

- $\mathbb{T}_L := [0, L]$ is the torus of length $L$.
- $\mathbb{Z}_L = \frac{\mathbb{Z}}{L} = \{\frac{k}{L} : k \in \mathbb{Z}\}$.

- $e(z) = e^{2\pi iz}$ for $z \in \mathbb{C}$.
- Given two quantities $A$ and $B$, we denote $\langle A \rangle = \sqrt{1 + A^2}$ and $A \lesssim B$ if $\exists\, C$, a universal constant, such that $A \leq CB$.
- $L^+$ is the short hand of the quantity $A$ that satisfies $\forall\ \gamma \ll 1$, $\exists\ C_\gamma$ independent of $L$, s.t. $|A| \leq C_\gamma L^\gamma$.
- The Fourier transform of a function $f$ on $\mathbb{R}$ is

$$\text{if } \xi \in \mathbb{R}, \quad \mathcal{F}f(\xi) = \widehat{f}(\xi) = \int_{\mathbb{R}^d} e(-x \cdot \xi) f(x)\, dx,$$

  where $e(z) = e^{2\pi iz}$ for $z \in \mathbb{C}$.
- The inverse Fourier transform of of a function $f$ on $\mathbb{R}$

$$\mathcal{F}^{-1}g(x) = \check{g}(x) = \int_{\mathbb{R}^d} e(x \cdot \xi) g(\xi)\, d\xi.$$

- The Fourier coefficients of a function $f$ defined on torus $\mathbb{T}_L$ is

$$f_K = \int_{\mathbb{T}_L} f(x) e(-K \cdot x)\, dx, \quad K \in \mathbb{Z}_L = \frac{\mathbb{Z}}{L}.$$

  The inversion formula is

$$f(x) = \frac{1}{L} \sum_{K \in \mathbb{Z}_L} f_K e(K \cdot x).$$

  The Plancherel identity is

$$||f||^2_{L^2(\mathbb{T}_L)} = \frac{1}{L} \sum_{K \in \mathbb{Z}_L} |f_K|^2.$$

  We note that there are more than one definition of the Fourier coefficients. For example, $f_K = \frac{1}{L} \int_{\mathbb{T}_L} f(x) e(-K \cdot x)\, dx, \quad K \in \mathbb{Z}_L = \frac{\mathbb{Z}}{L}$. Our choice ensures the

presence of $L^{-1}$ in front of the inversion formula. This is necessary for the sum on the right hand side to converge to an integral.

The functional spaces:

- For function $f$ defined on $\mathbb{R}$,

$$\|f\|_{X^\ell(\mathbb{R})} = \|\langle x\rangle^\ell f(x)\|_{L^\infty(\mathbb{R})}, \quad \|f\|_{X^{\ell,N}(\mathbb{R})} = \sum_{0\le|\alpha|\le N} \|\nabla^\alpha f\|_{X^\ell(\mathbb{R})}.$$

- For function $a_K$ defined on $\mathbb{Z}_L$, $\|a_K\|_{X^\ell} = \sup_{K\in\mathbb{Z}_L} \|\langle K\rangle^\ell a_K\|$. And for function $u$ defined on $\mathbb{T}_L$ with Fourier coefficients $u_K$, $\|u\|_{X^\ell(\mathbb{T}_L)} = \|u_K\|_{X^\ell}$.

Normal form transformation:

- We will encounter following linear form and quadratic form,

$$\mathcal{S}_{2d+1,K}(K_1,\dots,K_{2d+1}) = K_1 - K_2 + \cdots - K_{2d} + K_{2d+1} - K,$$

$$\Omega_{2d+1,K}(K_1,\dots,K_{2d+1}) = K_1^2 - K_2^2 + \cdots - K_{2d}^2 + K_{2d+1}^2 - K^2.$$

  We write $\mathcal{S}_{2d+1,K}$ and $\Omega_{2d+1,K}$ if there is no confusion about the indices.

- We shall modify the phase of $u_K$ for several times, and each time we have a corresponding equation, for $a_K$ see (2.6), for $b_K$ see (2.9), for $d_K$ see (2.10). $c_K$ the variable derived from $d_K$ after applying $P$ times normal form, see (2.43).

- For the cubic quadratic form with time dependent correction,

$$\widetilde{\Omega}_{2d+1,K} = \Omega_{2d+1,K} - \frac{\epsilon^2}{2\pi L^2} D_{2d+1,K}(t),$$

where

$$D_{2d+1,K}(t) = |d_{K_1}|^2 - |d_{K_2}|^2 + \cdots + |d_{K_{2d+1}}|^2 - |d_K|^2.$$

- We use $H^d_{K_1K_2\ldots K_{2d+1}}$, see (2.54), and $G^d_{K_1K_2\ldots K_{2d+1}}$, see (2.55), to denote the coefficients of terms after $d$ times normal form transformation.

- We use $S^d_K(u)$, $H^d_K(u)$, $G^d_K(u)$, $\widetilde{H}^{P+1}_K(u)$, $\widetilde{G}^{P+1}_K(u)$, $F^d_K(u)$, and $E^d_K(u)$ to denote polynomials in $u$ after normal form transformation, see (2.29), (2.36), (2.37), (2.38), (2.39), (2.40), and (2.41).

Feynman diagrams:

- $T$ is a tree in a forest of Feynman diagrams.

- $\mathcal{G}_d$ is the forest of Feynman diagrams after d times normal form transformation.

- $A_1$, $A_2$, $A_3$ are defined in (2.23) and $B_1$, $B_2$, $B_3$, $B_4$ are defined in (2.21).

- $*$, $\mathcal{A}_k(T)$, $\mathcal{B}_k(T)$, $\Omega_{2k+3,K}(T)$, $l^T_k$ and $S_{T_*}$ are defined in Lemma 1.

## 2.2 Normal form transformation

In this section, we shall obtain recurrence formulas of coefficients of the normal form transformation. Meanwhile, we extract the main term and the error terms from the nonlinearity. We remove the cubic resonance interaction by change of phase, but there will be a term left which forces us to further change phase with dependence on $K$. Later, we do normal form transformation in the same way as in [23]. The main difficulty here is to deal with the extra term coming from the $K$ dependent phase correction. To deal with this, we will define coefficients $G^d_{K_1\cdots K_{2d+1}}$ in addition to coefficients $H^d_{K_1\cdots K_{2d+1}}$ introduced in [23].

### 2.2.1 Extraction of the modified phase

In this subsection, we first introduce the profile by removing the linear part. Then we separate the cubic nonlinearity into resonance and non-resonance interaction. While we keep the non-resonance interaction which will be removed later by normal form, the resonance interaction will be removed by changing phase. In Fourier space, we have $u(t,x) = \frac{1}{L}\sum_{K\in\mathbb{Z}_L} u_K(t)e(Kx)$, where $K \in \mathbb{Z}_L = \frac{\mathbb{Z}}{L}$, $u_K$ satisfies

$$-i\partial_t u_K + 2\pi K^2 u_K = \frac{\epsilon^2}{L^2}\sum_{\mathcal{S}_{3,K}=0} u_{K_1}\overline{u_{K_2}}u_{K_3}, \tag{2.5}$$

where $K_i \in \mathbb{Z}_L$, for $i = 1, 2, 3$.

Consider the profile $a_K = e(-K^2 t)u_K$, $a_K$ satisfies

$$-i\partial_t a_K = \frac{\epsilon^2}{L^2}\sum_{\mathcal{S}_{3,K}=0} a_{K_1}\overline{a_{K_2}}a_{K_3}e(\Omega_{3,K}t). \tag{2.6}$$

If we take $\epsilon = 0$, the profile is a constant function in time. The profile is simply obtained by removing the effect of linear evolution, so it should grow much slower than $u_K$. But due to the presence of nonlinearity, variation of the profile still happens. The resonance interaction is the most important part of the nonlinearity which derives this variation. In the cubic term, we separate the dynamics of $a_K$ into resonance and non-resonance parts,

$$-i\partial_t a_K = \frac{\epsilon^2}{L^2}\sum_{\substack{\mathcal{S}_{3,K}=0\\ \Omega_{3,K}=0}} a_{K_1}\overline{a_{K_2}}a_{K_3} + \frac{\epsilon^2}{L^2}\sum_{\substack{\mathcal{S}_{3,K}=0\\ \Omega_{3,K}\neq 0}} a_{K_1}\overline{a_{K_2}}a_{K_3}e(\Omega_{3,K}t). \tag{2.7}$$

Since for any fixed $K$,

$$\begin{cases} \mathcal{S}_{3,K}(K_1,K_2,K_3) = 0 \\ \Omega_{3,K}(K_1,K_2,K_3) = 0 \end{cases} \tag{2.8}$$

only has degenerated solutions $\{K_1, K_3\} = \{K, K_2\}$. The equation of $a_K$ can be written as:

$$
\begin{aligned}
-i\partial_t a_K &= \frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}=0}} a_{K_1}\overline{a_{K_2}}a_{K_3} + \frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} a_{K_1}\overline{a_{K_2}}a_{K_3} e(\Omega_{3,K}t) \\
&= \frac{\epsilon^2}{L^2}\left(2\sum_{K_1\in\mathbb{Z}_L}|a_{K_1}|^2 - |a_K|^2\right)a_K + \frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} a_{K_1}\overline{a_{K_2}}a_{K_3} e(\Omega_{3,K}t),
\end{aligned}
$$

Then multiplying the profile by a phase factor, one can remove the cubic resonance interaction. Notice that $\sum_{K_1\in\mathbb{Z}_L}|a_{K_1}|^2 = L\|u\|_{L^2}^2$ is conserved and let

$$
b_K = e(-\frac{\epsilon^2}{\pi L^2}\sum_{K_1\in\mathbb{Z}_L}|a_{K_1}|^2 t)a_K,
$$

we have

$$
\begin{aligned}
-i\partial_t b_K =& e\left(-\frac{\epsilon^2}{\pi L^2}\sum_{K_1\in\mathbb{Z}_L}|a_{K_1}|^2 t\right)\left(-i\partial_t a_K - \frac{2\epsilon^2}{L^2}\sum_{K_1\in\mathbb{Z}_L}|a_{K_1}|^2 a_K\right) \\
=& -\frac{\epsilon^2}{L^2}|b_K|^2 b_K + \frac{\epsilon^2}{L^2}\sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} b_{K_1}\overline{b_{K_2}}b_{K_3}e(\Omega_{3,K}t).
\end{aligned} \tag{2.9}
$$

Let

$$
d_K := b_K e(\frac{\epsilon^2}{2\pi L^2}\int_0^t |b_K(s)|^2 ds) = a_K e(-\frac{\epsilon^2}{\pi L}\|u\|_{L^2}^2 t + \frac{\epsilon^2}{2\pi L^2}\int_0^t |a_K(s)|^2 ds),
$$

note that $|d_K| = |b_K| = |a_K| = |u_K|$.

Define

$$
\widetilde{\Omega}_{3,K} = \Omega_{3,K} - \frac{\epsilon^2}{2\pi L^2}(|d_{K_1}|^2 - |d_{K_2}|^2 + |d_{K_3}|^2 - |d_K|^2)
$$

and

$$
D_{3,K}(t) = |d_{K_1}|^2 - |d_{K_2}|^2 + |d_{K_3}|^2 - |d_K|^2.
$$

Then $d_K$ satisfies

$$-i\partial_t d_K = \frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} d_{K_1}\overline{d_{K_2}}d_{K_3} e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds). \tag{2.10}$$

So far, we have removed the cubic resonance interaction.

### 2.2.2 Normal form transformation of the cubic term

The normal form transformation means the process of separating the resonance and non-resonance interaction of the nonlinearity and using differentiation by parts argument on the nonresonance interaction. The nonresonance interaction are small due to cancellation of the phase factor. In the same spirit as oscillatory integral, an appropriate way of exploiting this cancellation is by integration by parts or differentiation by parts.

As in the oscillatory integral setting, if we have $\int_{\mathbb{R}} e(\lambda\phi(x))f(x)dx$, as $\lambda$ goes to infinity, the oscillation in $x$ of the exponential factor $e(\lambda\phi(x))$ is getting stronger and stronger. So the value of this integral should be smaller and smaller. A common idea of exploiting the oscillation is to do the integration by parts argument in $x$, i.e. $e(\lambda\phi(x)) = \frac{\partial_t e(\lambda\phi(x))}{2\pi i\lambda\phi'(x)}$. Each time of applying this argument we gain one $\frac{1}{\lambda}$ , provided no stationary points $x_0$ s.t. $\phi'(x_0) = 0$.

Now our situation is very similar.We have a term like $d_{K_1}\overline{d_{K_2}}d_{K_3}e(\Omega_{3,K}t)$ which is oscillating rapidly when $t$ is large. We apply the identity $\frac{\partial_t e(\Omega_{3,K}t)}{\Omega_{3,K}}$ and get

$$d_{K_1}\overline{d_{K_2}}d_{K_3}e(\Omega_{3,K}t) = \partial_t\left(\frac{d_{K_1}\overline{d_{K_2}}d_{K_3}e(\Omega_{3,K}t)}{\Omega_{3,K}}\right) - \frac{e(\Omega_{3,K}t)}{\Omega_{3,K}}\partial_t(d_{K_1}\overline{d_{K_2}}d_{K_3}).$$

The first term can be absorbed by $-i\partial_t d_K$ on the left hand side and the second term will give us $\epsilon^2$ if we substitute the equation of $\partial_t d_K$. But notice that because we have changed the phase, we have $e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds)$ instead of $e(\Omega_{3,K}t)$. We do differentiation by parts on $e(\Omega_{3,K}t)$

since we have the restriction that $e(\Omega_{3,K}t) \neq 0$ and leave a factor $e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{3,K}(s)ds)$. When the time derivative hits this factor, we gain $\epsilon^2/L^2$.

$$
\begin{aligned}
-i\partial_t d_K =& \frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} d_{K_1}\overline{d_{K_2}}d_{K_3} e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds) && (2.11)\\
=& \frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} d_{K_1}\overline{d_{K_2}}d_{K_3} \frac{\partial_t e(\Omega_{3,K}t)}{2\pi i\Omega_{3,K}} e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{3,K}(s)ds) && (2.12)\\
=& \frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} \partial_t\big(d_{K_1}\overline{d_{K_2}}d_{K_3} \frac{1}{2\pi i\Omega_{3,K}} e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds)\big) && (2.13)\\
&-\frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} \frac{1}{2\pi i\Omega_{3,K}}\partial_t \left(d_{K_1}\overline{d_{K_2}}d_{K_3} e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{3,K}(s)ds)\right) e(\Omega_{3,K}t) && (2.14)
\end{aligned}
$$

Absorb the second term on the right hand side which is of the form $\partial_t(...)$ by the left hand side,

$$
\begin{aligned}
&-i\partial_t\Big(d_K - \frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} d_{K_1}\overline{d_{K_2}}d_{K_3} \frac{1}{\Omega_{3,K}} e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds)\Big)\\
&= -\frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} \frac{1}{2\pi i\Omega_{3,K}}\partial_t \left(d_{K_1}\overline{d_{K_2}}d_{K_3} e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{3,K}(s)ds)\right) e(\Omega_{3,K}t)\\
&= -\frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} \frac{1}{2\pi i\Omega_{3,K}}\partial_t \left(d_{K_1}\overline{d_{K_2}}d_{K_3}\right) e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds)\\
&+ \frac{\epsilon^4}{L^4} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} \frac{1}{2\pi\Omega_{3,K}} d_{K_1}\overline{d_{K_2}}d_{K_3} D_{3,K}(t) e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds)
\end{aligned}
\tag{2.15}
$$

Let

$$
c_K^1 = d_K - \frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} d_{K_1}\overline{d_{K_2}}d_{K_3} \frac{1}{2\pi\Omega_{3,K}} e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds), \tag{2.16}
$$

then

$$
\begin{aligned}
-i\partial_t c_K^1 = &-\frac{\epsilon^2}{L^2}\sum_{\substack{\mathcal{S}_{3,K}=0\\ \Omega_{3,K}\neq 0}} \frac{1}{2\pi i\Omega_{3,K}}\partial_t\left(d_{K_1}\overline{d_{K_2}}d_{K_3}\right)e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds)\\
&+\frac{\epsilon^4}{L^4}\sum_{\substack{\mathcal{S}_{3,K}=0\\ \Omega_{3,K}\neq 0}} \frac{1}{2\pi\Omega_{3,K}}d_{K_1}\overline{d_{K_2}}d_{K_3}D_{3,K}(t)e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds)
\end{aligned}
\tag{2.17}
$$

When the time derivative hits one of the $d_{K_j}$, for example when it hits $d_{K_3}$, we have

$$
\begin{aligned}
&-\frac{\epsilon^2}{L^2}\sum_{\substack{\mathcal{S}_{3,K}=0\\ \Omega_{3,K}\neq 0}} \frac{1}{2\pi i\Omega_{3,K}}d_{K_1}\overline{d_{K_2}}\partial_t d_{K_3}e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds)\\
=&-\frac{\epsilon^2}{L^2}\sum_{\substack{K_1-K_2+K_3=K\\ K_1^2-K_2^2+K_3^2\neq K^2}} \frac{e(\int_0^t \widetilde{\Omega}_{3,K_3}(s)ds)}{2\pi i\Omega_{3,K}(K_1,K_2,K_3)}\frac{\epsilon^2}{L^2}d_{K_1}\overline{d_{K_2}}\Big(\sum_{K_3=K_3'-K_4+K_5} d_{K_3'}\overline{d_{K_4}}d_{K_5}\Big)e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds)
\end{aligned}
\tag{2.18}
$$

Relabel the indices, we have

$$
\begin{aligned}
&-\frac{\epsilon^2}{L^2}\sum_{\substack{K_1-K_2+K_3=K\\ K_1\neq K_2,K_1\neq K}} \frac{1}{2\pi i\Omega_{3,K}(K_1,K_2,K_3)}\frac{\epsilon^2}{L^2}d_{K_1}\overline{d_{K_2}}\times\\
&\qquad\qquad\times\Big(\sum_{K_3=K_3'-K_4+K_5} d_{K_3'}\overline{d_{K_4}}d_{K_5}e(\int_0^t(\widetilde{\Omega}_{3,K_3}(s)+\widetilde{\Omega}_{3,K}(s))ds)\Big)\\
=&-\frac{\epsilon^4}{L^4}\sum_{\substack{K_1-K_2+K_3-K_4+K_5=K\\ K_1\neq K_2,K_1\neq K}} \frac{d_{K_1}\overline{d_{K_2}}d_{K_3}\overline{d_{K_4}}d_{K_5}}{2\pi i\Omega_{3,K}(K_1,K_2,K_3-K_4+K_5)}\times\\
&\qquad\qquad\times e(\int_0^t \widetilde{\Omega}_{5,K}(K_1,K_2,K_3,K_4,K_5)(s)ds)
\end{aligned}
\tag{2.19}
$$

Do this to each term in the expansion of (2.17) and relabel the indices. We have

$$
\begin{aligned}
&-i\partial_t c^1_K = \frac{i\epsilon^4}{2\pi L^4}\sum_{\mathcal{S}_{5,K}=0} d_{K_1}\overline{d_{K_2}}d_{K_3}\overline{d_{K_4}}d_{K_5}e(\int_0^t \widetilde{\Omega}_{5,K}(s)ds)\Big[\frac{1_{K_1\neq K,K_5\neq K,K_2\neq K_3,K_3\neq K_4}}{\Omega_{3,K}(K_1,K_2-K_3+K_4,K_5)}\\
&-\frac{1_{K_1\neq K_2,K_2\neq K_3,K_4\neq K_5,K_5\neq K}}{\Omega_{3,K}(K_1-K_2+K_3,K_4,K_5)}-\frac{1_{K_1\neq K_2,K_1\neq K,K_3\neq K_4,K_4\neq K_5}}{\Omega_{3,K}(K_1,K_2,K_3-K_4+K_5)}\Big]\\
&+\frac{\epsilon^4}{L^4}\sum_{\substack{\mathcal{S}_{3,K}=0\\ \Omega_{3,K}\neq 0}}\frac{1}{2\pi\Omega_{3,K}}d_{K_1}\overline{d_{K_2}}d_{K_3}D_{3,K}(t)e(\int_0^t\widetilde{\Omega}_{3,K}(s)ds).
\end{aligned}
\tag{2.20}
$$

By doing normal form transformation once, we have transformed the cubic non-resonance interaction into a quintic form.

Let us introduce the following sets

$$
\begin{aligned}
B_1&=\{K_1=K_2=K_3,K_3\neq K_4,K_4\neq K_5\},\\
B_2&=\{K_1\neq K_2,K_2=K_3=K_4,K_4\neq K_5\},\\
B_3&=\{K_1\neq K_2,K_2\neq K_3,K_3=K_4=K_5\},\\
B_4&=\{K_1\neq K_2,K_2\neq K_3,K_4=K_5=K\}.
\end{aligned}
\tag{2.21}
$$

Then the second term of the left hand side may be reformulated as

$$
\begin{aligned}
&\frac{\epsilon^4}{L^4}\sum_{\substack{\mathcal{S}_{3,K}=0\\ \Omega_{3,K}\neq 0}}\frac{1}{2\pi\Omega_{3,K}}d_{K_1}\overline{d_{K_2}}d_{K_3}D_{3,K}(t)e(\int_0^t\widetilde{\Omega}_{3,K}(s)ds)\\
=&\frac{\epsilon^4}{L^4}\sum_{\mathcal{S}_{5,K}=0}\frac{1_{B_1}-1_{B_2}+1_{B_3}-1_{B_4}}{2\pi\Omega_{5,K}}d_{K_1}\overline{d_{K_2}}d_{K_3}\overline{d_{K_4}}d_{K_5}e(\int_0^t\widetilde{\Omega}_{5,K}(s)ds)
\end{aligned}
\tag{2.22}
$$

Denote $G^2_{K_1\dots K_5}=\frac{1_{B_1}-1_{B_2}+1_{B_3}-1_{B_4}}{2\pi\Omega_{5,K}}$ and

$$H^2_{K_1K_2\cdots K_5}=\frac{1_{A_2}}{2\pi\Omega_{3,K}(K_1,K_2-K_3+K_4,K_5)}$$
$$-\frac{1_{A_1}}{2\pi\Omega_{3,K}(K_1-K_2+K_3,K_4,K_5)}-\frac{1_{A_3}}{2\pi\Omega_{3,K}(K_1,K_2,K_3-K_4+K_5)},$$

where

$$\begin{aligned}A_1&=\{K_1\neq K_2,K_2\neq K_3,K_4\neq K_5,K_5\neq K\},\\ A_2&=\{K_1\neq K,K_2\neq K_3,K_3\neq K_4,K_5\neq K\},\\ A_3&=\{K_1\neq K_2,K_1\neq K,K_3\neq K_4,K_4\neq K_5\}.\end{aligned}\tag{2.23}$$

We rewrite (2.20) as

$$\begin{aligned}-i\partial_t c^1_K&=\frac{\epsilon^4}{L^4}\sum_{\mathcal{S}_{5,K}=0}H^2_{K_1K_2\cdots K_5}d_{K_1}\overline{d_{K_2}}d_{K_3}\overline{d_{K_4}}d_{K_5}e(\int_0^t\widetilde{\Omega}_{5,K}(s)ds)\\&+\frac{\epsilon^4}{L^4}\sum_{\mathcal{S}_{5,K}=0}G^2_{K_1\dots K_5}d_{K_1}\overline{d_{K_2}}d_{K_3}\overline{d_{K_4}}d_{K_5}e(\int_0^t\widetilde{\Omega}_{5,K}(s)ds)\\&=\frac{\epsilon^4}{L^4}\sum_{\mathcal{S}_{5,K}=0}H^{2,tot}_{K_1K_2\cdots K_5}d_{K_1}\overline{d_{K_2}}d_{K_3}\overline{d_{K_4}}d_{K_5}e\left(\int_0^t\widetilde{\Omega}_{5,K}(s)ds\right),\end{aligned}\tag{2.24}$$

where $H^{2,tot}_{K_1K_2\cdots K_5}=H^2_{K_1K_2\cdots K_5}+G^2_{K_1K_2\cdots K_5}$.

### 2.2.3 Higher order normal form transformations

In this subsection, we continue with normal form transformation up to $P$ times to remove the quintic form into higher order forms. Notice that in (2.24), we don't have the restriction that prevents $\Omega_{5,K}$ to vanish. So we shall do higher order normal form transformation by first splitting the sums in (2.24) into $\sum_{\mathcal{S}_{5,K}=0}=\sum_{\substack{\mathcal{S}_{5,K}=0\\ \Omega_{5,K}=0}}+\sum_{\substack{\mathcal{S}_{5,K}=0\\ \Omega_{5,K}\neq 0}}$, then always leaving the resonance interaction i.e. $\Omega_{5,K}=0$, and doing differentiation by parts on the nonresonance

interaction i.e. $\Omega_{5,K} \neq 0$. In doing the differentiation by parts we exploit

$$e(\int_0^t \widetilde{\Omega}_{5,K}(s)ds) = e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{5,K}(s)ds)\frac{\partial_t e(\Omega_{5,K}t)}{2\pi i\Omega_{5,K}}$$

and use differentiation by parts to move the $\partial_t$ to $d_{K_1}\overline{d_{K_2}}\cdots d_{K_5}$ or $e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{5,K}(s)ds)$. When the derivative hits the product of $d_{K_j}$, we substitute the equation for $\partial_t d_{K_j}$ to produce higher order terms and keep do differentiation by parts to these term. When the derivative hits $e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{5,K}(s)ds)$, we get an extra $\frac{\epsilon^2}{L^2}D_{5,K}(s)$ factor. We denote this term by $S_K^2(u)$ and leave it unchanged in following normal form transformations. Then repeat this process for 7 order terms and higher order terms. Let $c_K^d$ be the variable after $d$ times normal form transformation. By about analysis we may assume that the equations after $d$ normal form transformations writes

$$\begin{aligned}
-i\partial_t c_K^d =& \frac{\epsilon^4}{L^4}\sum_{\substack{\mathcal{S}_{5,K}=0\\ \Omega_{5,K}=0}} H^2_{K_1K_2\cdots K_5} d_{K_1}\overline{d_{K_2}}\cdots d_{K_5} e\left(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{5,K}(s)ds\right)\\
&+\frac{\epsilon^6}{L^6}\sum_{\substack{\mathcal{S}_{7,K}=0\\ \Omega_{7,K}=0}} H^{3,tot}_{K_1K_2\cdots K_7} d_{K_1}\overline{d_{K_2}}\cdots d_{K_7} e\left(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{7,K}(s)ds\right)+\cdots\\
&+\frac{\epsilon^{2d}}{L^{2d}}\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0}} H^{d,tot}_{K_1K_2\ldots K_{2d+1}} d_{K_1}\overline{d_{K_2}}\cdots d_{K_{2d+1}} e\left(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{2d+1,K}(s)ds\right)\\
&+\frac{\epsilon^{2d+2}}{L^{2d+2}}\sum_{\mathcal{S}_{2d+3,K}=0} H^{d+1,tot}_{K_1K_2\ldots K_{2d+3}} d_{K_1}\overline{d_{K_2}}\cdots d_{K_{2d+3}} e\left(\int_0^t \widetilde{\Omega}_{2d+3,K}(s)ds\right)\\
&+\sum_{k=3}^{d+1}\frac{\epsilon^{2k}}{L^{2k}}S_K^k(u),
\end{aligned}$$

where

$$\widetilde{\Omega}_{2d+3,K}(t) = \Omega_{2d+1,K} - \frac{\epsilon^2}{2\pi L^2}\sum_{j=1}^{2d+1}(-1)^{j-1}|d_{K_j}|^2(t),$$

$$D_{2d+1,K}=\sum_{j=1}^{2d+1}(-1)^{j-1}|d_{K_j}|^2(t).$$

Note that because $G^2_{K_1...K_5}=\frac{1_{B_1}-1_{B_2}+1_{B_3}-1_{B_4}}{2\pi\Omega_{5,K}}$ vanishes on the resonance surface $\{\mathcal{S}_{5,K}=0,\Omega_{5,K}=0\}$, we don't have $G^2_{K_1...K_5}$ in the first term on the right hand side. Do differentiation by parts in the fourth line and fifth line in above equation,

$$\begin{aligned}
&d_{K_1}\overline{d_{K_2}}\cdots d_{K_{2d+3}}e(\int_0^t\widetilde{\Omega}_{2d+3,K}(s)ds)\\
=&d_{K_1}\overline{d_{K_2}}\cdots d_{K_{2d+3}}e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{2d+3,K}(s)ds)\frac{\partial_t e(\Omega_{2d+3,K}t)}{2\pi i\Omega_{2d+3,K}}\\
=&i\partial_t\left(d_{K_1}\overline{d_{K_2}}\cdots d_{K_{2d+3}}\frac{e(\int_0^t\widetilde{\Omega}_{2d+3,K}(s)ds)}{2\pi\Omega_{2d+3,K}}\right)\\
&-\partial_t\left(d_{K_1}\overline{d_{K_2}}...d_{K_{2d+3}}e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{2d+3,K}(s)ds)\right)\frac{e(\Omega_{2d+3,K}t)}{2\pi i\Omega_{2d+3,K}}.
\end{aligned}\tag{2.25}$$

If the $\partial_t$ in the second term hits $d_{K_1}\overline{d_{K_2}}...d_{K_{2d+3}}$. Then by substitute the equation of $\partial_t d_K$ (2.10), we have

$$\begin{aligned}
&-\frac{\epsilon^{2d+2}}{L^{2d+2}}\sum_{\mathcal{S}_{2d+3,K}=0}H^{d+1,tot}_{K_1K_2...K_{2d+3}}\partial_t(d_{K_1}\overline{d_{K_2}}\cdots d_{K_{2d+3}})e\left(\int_0^t\widetilde{\Omega}_{2d+3,K}(s)ds\right)\\
=&-\frac{\epsilon^{2d+4}}{L^{2d+4}}\sum_{j=1}^{2d+3}\sum_{\substack{\mathcal{S}_{2d+3,K}=0\\ \Omega_{2d+3,K}\neq 0}}H^{d+1,tot}_{K_1K_2...K_{2d+3}}d_{K_1}...\\
&\times\left(\sum_{S_{3,K_j}(M,N,L)=0}d_M\overline{d_N}d_Le\left(\int_0^t\widetilde{\Omega}_{3,K_j}(M,N,L)(s)ds\right)\right)...d_{K_{2d+3}}\frac{e(\int_0^t\widetilde{\Omega}_{2d+3,K}(s)ds)}{2\pi\Omega_{2d+3,K}}\\
=&\frac{\epsilon^{2d+4}}{L^{2d+4}}\sum_{\mathcal{S}_{2d+5,K}=0}\Big(\sum_{j=1}^{2d+3}\frac{(-1)^j1_{\Omega_{2d+3,K}(K_1,...,K_j-K_{j+1}+K_{j+2},...,K_{2d+5})\neq 0}H^{d+1,tot}_{K_1...(K_j-K_{j+1}+K_{j+2})...K_{2d+5}}}{2\pi\Omega_{2d+3,K}(K_1,...,K_j-K_{j+1}+K_{j+2},...,K_{2d+5})}\Big)\\
&\times d_{K_1}\overline{d_{K_2}}...d_{K_{2d+5}}e(\int_0^t\widetilde{\Omega}_{2d+5,K}(s)ds).
\end{aligned}\tag{2.26}$$

If the $\partial_t$ in the second term hits $e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{2d+3,K}(s)ds)$, we get

$$\sum_{\mathcal{S}_{2d+3,K}=0} H^{d+1,tot}_{K_1K_2\dots K_{2d+3}} d_{K_1}\overline{d_{K_2}}\dots d_{K_{2d+3}}\frac{D_{2d+3,K}(t)}{2\pi\Omega_{2d+3,K}} e\left(\int_0^t \widetilde{\Omega}_{2d+3,K}(s)ds\right). \tag{2.27}$$

Combining all above calculations, we have

$$\begin{aligned}
&-i\partial_t\Big(c_K^d-\frac{\epsilon^{2d+2}}{L^{2d+2}}\sum_{\substack{\mathcal{S}_{2d+3,K}=0\\ \Omega_{2d+3,K}\neq 0}} H^{d+1,tot}_{K_1K_2\dots K_{2d+3}} d_{K_1}\overline{d_{K_2}}\dots d_{K_{2d+3}}\frac{e(\int_0^t\widetilde{\Omega}_{2d+3,K}(s)ds)}{2\pi\Omega_{2d+3,K}}\Big)\\
=&\frac{\epsilon^4}{L^4}\sum_{\substack{\mathcal{S}_{5,K}=0\\ \Omega_{5,K}=0}} H^2_{K_1K_2\cdots K_5} d_{K_1}\overline{d_{K_2}}\cdots d_{K_5}e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{5,K}(s)ds)\\
+&\frac{\epsilon^6}{L^6}\sum_{\substack{\mathcal{S}_{7,K}=0\\ \Omega_{7,K}=0}} H^{3,tot}_{K_1K_2\cdots K_7} d_{K_1}\overline{d_{K_2}}\cdots d_{K_7}e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{7,K}(s)ds)+\cdots\\
+&\frac{\epsilon^{2d+2}}{L^{2d+2}}\sum_{\substack{\mathcal{S}_{2d+3,K}=0\\ \Omega_{2d+3,K}=0}} H^{d+1,tot}_{K_1K_2\dots K_{2d+3}} d_{K_1}\overline{d_{K_2}}\cdots d_{K_{2d+3}}e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{2d+3,K}(s)ds)\\
+&\frac{\epsilon^{2d+4}}{L^{2d+4}}\sum_{\mathcal{S}_{2d+5,K}=0}\Big(\sum_{j=1}^{2d+3}\frac{(-1)^j 1_{\Omega_{2d+3,K}(K_1,\dots,K_j-K_{j+1}+K_{j+2},\dots,K_{2d+5})\neq 0} H^{d+1,tot}_{K_1\dots(K_j-K_{j+1}+K_{j+2})\dots K_{2d+5}}}{2\pi\Omega_{2d+3,K}(K_1,\dots,K_j-K_{j+1}+K_{j+2},\dots,K_{2d+5})}\Big)\\
&\times d_{K_1}\overline{d_{K_2}}\dots d_{K_{2d+5}}e(\int_0^t\widetilde{\Omega}_{2d+5,K}(s)ds)+\sum_{k=3}^{d+1}\frac{\epsilon^{2k}}{L^{2k}}S_K^k(u)\\
+&\frac{\epsilon^{2d+4}}{L^{2d+4}}\sum_{\mathcal{S}_{2d+3,K}=0} H^{d+1,tot}_{K_1K_2\dots K_{2d+3}} d_{K_1}\overline{d_{K_2}}\dots d_{K_{2d+3}}\frac{D_{2d+3,K}(t)}{2\pi\Omega_{2d+3,K}}e(\int_0^t\widetilde{\Omega}_{2d+3,K}(s)ds)
\end{aligned}$$

From above calculation, we get the recurrence formulas for $c_K^{d+1}$, $H^{d+2,tot}$ and $S_K^{d+2}(u)$.

$$c_K^{d+1}=c_K^d-\frac{\epsilon^{2d+2}}{L^{2d+2}}\sum_{\substack{\mathcal{S}_{2d+3,K}=0\\ \Omega_{2d+3,K}\neq 0}} H^{d+1,tot}_{K_1K_2\dots K_{2d+3}} d_{K_1}\overline{d_{K_2}}\cdots d_{K_{2d+3}}e(\int_0^t\widetilde{\Omega}_{2d+3,K}(s)ds), \tag{2.28}$$

$$S_K^{d+2}(u) =$$
$$\frac{\epsilon^{2d+4}}{L^{2d+4}} \sum_{\mathcal{S}_{2d+3,K}=0} H^{d+1,tot}_{K_1K_2...K_{2d+3}} d_{K_1} \overline{d_{K_2}} ... d_{K_{2d+3}} \frac{D_{2d+3,K}(t)}{2\pi\Omega_{2d+3,K}} e(\int_0^t \widetilde{\Omega}_{2d+3,K}(s)ds). \tag{2.29}$$

$$H^{d+2,tot}_{K_1...K_{2d+5}} =$$
$$\sum_{j=1}^{2d+3} \frac{(-1)^j 1_{K_j \neq K_{j+1}, K_{j+1} \neq K_{j+2}, \Omega_{2d+3,K}(K_1,...,K_j - K_{j+1} + K_{j+2},...,K_{2d+5}) \neq 0} H^{d+1,tot}_{K_1...(K_j - K_{j+1} + K_{j+2})...K_{2d+5}}}{2\pi\Omega_{2d+3,K}(K_1,...,K_j - K_{j+1} + K_{j+2},...,K_{2d+5})}, \tag{2.30}$$

Since the initial value of $H^{d+2,tot}$ has a decomposition $H^{2,tot}_{K_1K_2\cdots K_5} = H^2_{K_1K_2\cdots K_5} + G^2_{K_1K_2\cdots K_5}$, we also have a decomposition $H^{d+2,tot}_{K_1...K_{2d+5}} = H^{d+2}_{K_1...K_{2d+5}} + G^{d+2}_{K_1...K_{2d+5}}$ by the linearity of above recurrence formula. Here $H^{d+2}$, $G^{d+2}$ are obtained by running above recurrence formula using initial value $H^2_{K_1K_2\cdots K_5}$, $G^2_{K_1K_2\cdots K_5}$ respectively.

Since $H^{d+2}$, $G^{d+2}$ have different estimates and properties, we shall use them in the rest instead of $H^{d+2,tot}$. Their recurrence formulas are

$$H^{d+2}_{K_1...K_{2d+5}} =$$
$$\sum_{j=1}^{2d+3} \frac{(-1)^j 1_{K_j \neq K_{j+1}, K_{j+1} \neq K_{j+2}, \Omega_{2d+3,K}(K_1,...,K_j - K_{j+1} + K_{j+2},...,K_{2d+5}) \neq 0} H^{d+1}_{K_1...(K_j - K_{j+1} + K_{j+2})...K_{2d+5}}}{2\pi\Omega_{2d+3,K}(K_1,...,K_j - K_{j+1} + K_{j+2},...,K_{2d+5})}, \tag{2.31}$$

$$G^{d+2}_{K_1...K_{2d+5}} =$$
$$\sum_{j=1}^{2d+3} \frac{(-1)^j 1_{K_j \neq K_{j+1}, K_{j+1} \neq K_{j+2}, \Omega_{2d+3,K}(K_1,...,K_j - K_{j+1} + K_{j+2},...,K_{2d+5}) \neq 0} G^{d+1}_{K_1...(K_j - K_{j+1} + K_{j+2})...K_{2d+5}}}{2\pi\Omega_{2d+3,K}(K_1,...,K_j - K_{j+1} + K_{j+2},...,K_{2d+5})}. \tag{2.32}$$

Initially,

$$c_K^1 = d_K - \frac{\epsilon^2}{L^2} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} d_{K_1}\overline{d_{K_2}}d_{K_3}\frac{1}{2\pi\Omega_{3,K}} e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds), \tag{2.33}$$

$$\begin{aligned} H^2_{K_1K_2\cdots K_5} =& \frac{1_{A_2}}{\Omega_{3,K}(K_1, K_2 - K_3 + K_4, K_5)} \\ & - \frac{1_{A_1}}{\Omega_{3,K}(K_1 - K_2 + K_3, K_4, K_5)} - \frac{1_{A_3}}{\Omega_{3,K}(K_1, K_2, K_3 - K_4 + K_5)}, \end{aligned} \tag{2.34}$$

$$G^2_{K_1\ldots K_5} = \frac{1_{B_1} - 1_{B_2} + 1_{B_3} - 1_{B_4}}{2\pi\Omega_{5,K}} \tag{2.35}$$

with $A_1, A_2, A_3$ defined in (2.23) $B_1, B_2, B_3.B_4$ defined in (2.21).

Combining (2.28), (2.31), (2.32) (2.33), (2.34), (2.35). we can calculate all the quantities in the normal form transformations.

Assume that we have done normal form transformation for $P$ times. We also introduce the following notations

$$H^d_K(u) = \sum_{\substack{\mathcal{S}_{2d+1,K}=0 \\ \Omega_{2d+1,K}=0}} H^d_{K_1K_2\ldots K_{2d+1}} d_{K_1}\overline{d_{K_2}}\ldots d_{K_{2d+1}} e(\int_0^t D_{2d+1,K}(s)ds), \tag{2.36}$$

$$G^d_K(u) = \sum_{\substack{\mathcal{S}_{2d+1,K}=0 \\ \Omega_{2d+1,K}=0}} G^d_{K_1K_2\ldots K_{2d+1}} d_{K_1}\overline{d_{K_2}}\ldots d_{K_{2d+1}} e(\int_0^t D_{2d+1,K}(s)ds), \tag{2.37}$$

$$\widetilde{H}^{P+1}_K(u) = \sum_{\mathcal{S}_{2P+3,K}=0} H^{P+1}_{K_1K_2\ldots K_{2P+3}} d_{K_1}\overline{d_{K_2}}\ldots d_{K_{2P+3}} e(\int_0^t \widetilde{\Omega}_{2P+3,K}(s)ds), \tag{2.38}$$

$$\widetilde{G}^{P+1}_K(u) = \sum_{\mathcal{S}_{2P+3,K}=0} G^{P+1}_{K_1K_2\ldots K_{2P+3}} d_{K_1}\overline{d_{K_2}}\ldots d_{K_{2P+3}} e(\int_0^t \widetilde{\Omega}_{2P+3,K}(s)ds), \tag{2.39}$$

$$F_K^d(u) = \sum_{\substack{\mathcal{S}_{2d+1,K}=0 \\ \Omega_{2d+1,K}\neq 0}} H^d_{K_1K_2\dots K_{2d+1}} d_{K_1}\overline{d_{K_2}}\dots d_{K_{2d+1}} \frac{1}{2\pi\widetilde{\Omega}_{2d+1,K}} e(\int_0^t \Omega_{2d+1,K}(s)ds). \tag{2.40}$$

$$E_K^d(u) = \sum_{\substack{\mathcal{S}_{2d+1,K}=0 \\ \Omega_{2d+1,K}\neq 0}} G^d_{K_1K_2\dots K_{2d+1}} d_{K_1}\overline{d_{K_2}}\dots d_{K_{2d+1}} \frac{1}{2\pi\widetilde{\Omega}_{2d+1,K}} e(\int_0^t \Omega_{2d+1,K}(s)ds). \tag{2.41}$$

Denote $c_K^P$ by $c_K$, we have

$$c_K = d_K - \sum_{d=1}^{P} \frac{\epsilon^{2d}}{L^{2d}} F_K^d(u) - \sum_{d=1}^{P} \frac{\epsilon^{2d}}{L^{2d}} E_K^d(u), \tag{2.42}$$

And the transformed equation writes

$$-i\partial_t c_K = \frac{\epsilon^4}{L^4} H_K^2(u) + \sum_{d=3}^{P} \frac{\epsilon^{2d}}{L^{2d}} (H_K^d(u) + G_K^d(u)) + \sum_{d=3}^{P+1} \frac{\epsilon^{2d}}{L^{2d}} S_K^d(u) + \frac{\epsilon^{2P+2}}{L^{2P+2}} (\widetilde{H}_K^{P+1}(u) + \widetilde{G}_K^{P+1}(u)) \tag{2.43}$$

We shall extract our main term from $\frac{\epsilon^4}{L^4} H_K^2(u)$ by using vanishing property.

## 2.3 The Feynman diagrams and cancellations of resonances

In this section, we introduce Feynman diagrams and get explicit formulas for coefficients $H^d_{K_1K_2\dots K_{2d+1}}$ and $G^d_{K_1K_2\dots K_{2d+1}}$ of normal form transformation. This formula applies to all dispersive equations. Then we apply an induction argument on the level of trees in our Feynman diagram to prove a vanishing property of the coefficient of normal form transformation.

Feynman diagrams are often used in the study of NLS via normal form transformations, see for example [24] and [85]. A Feynman diagram is a forest diagram whose trees corresponds to terms in $H^d$, $\widetilde{H}^P$, $G^d$, or $\widetilde{G}^P$. Each node at the bottom of these trees correspond to a summation of indices in these terms. In this section we do not distinguish $H^d$ and $\widetilde{H}^P$, $G^d$ and $\widetilde{G}^P$ because they have the same coefficients. We will just explain the Feynman diagram technique for coefficients $H$. Coefficients $G$ can be treated in a very similar way.

### 2.3.1 The partition trees

We first introduce some notations in trees.

For a set of indices $\{K_1, ..., K_{2d+3}\}$, a $2k+3$-partition $\mathscr{P}$ is a collection of subsets whose union is $\{K_1, ..., K_{2d+3}\}$

$$\mathscr{P} = \{\{K_1, ..., K_{i_1}\}, \{K_{i_1+1}, ..., K_{i_2}\}, \cdots, \{K_{i_{2k+2}+1}, ..., K_{2d+3}\}\},$$

for all $0 = i_0 < i_1 < \cdots < i_{2k+2} < i_{2k+3} = 2d+3$. A $2k+3$ odd partition is a partition that consists of subsets of odd cardinality. In above notations, all of $\{K_{i_s+1}, ..., K_{i_{s+1}}\}$ have odd cardinality is equivalent to $i_s \equiv s \bmod(2)$. The collection of all $2k+3-$odd partitions of $\{K_1, ..., K_{2d+3}\}$ is denoted by $\mathbf{P}_{2k+3}$

We say $\mathscr{P}' \in \mathbf{P}_{2k+5}$ is a ternary refinement of $\mathscr{P} \in \mathbf{P}_{2k+3}$, denoted by $\mathscr{P} \leq \mathscr{P}'$, if $\mathscr{P}'$ is a $2k+5$ partition obtained by dividing a set in $\mathscr{P}$ into three new sets with odd cardinality. We will use a simpler term refinement to refer to ternary refinement. For example,

$$\{\{K_1, K_2, K_3, K_4, K_5\}, \{K_6\}, \{K_7\}\} \leq \{\{K_1\}, \{K_2\}, \{K_3, K_4, K_5\}, \{K_6\}, \{K_7\}\},$$

where the first set of the first partition is divided into three subsets.

$\mathscr{P}_3, \mathscr{P}_5, \cdots, \mathscr{P}_{2d+3}$ is called a full chain of refinements if $\mathscr{P}_{2j+3} \in \mathbf{P}_{2j+3}$ and $\mathscr{P}_3 \leq \mathscr{P}_5 \leq \cdots \leq \mathscr{P}_{2d+3}$. For example, the following sequence of partition is a full chain of refinements.

$$\begin{gathered}\{\{K_1, K_2, K_3, K_4, K_5\}, \{K_6\}, \{K_7\}\} \leq \\ \{\{K_1\}, \{K_2\}, \{K_3, K_4, K_5\}, \{K_6\}, \{K_7\}\} \leq \\ \{\{K_1\}, \{K_2\}, \{K_3\}, \{K_4\}, \{K_5\}, \{K_6\}, \{K_7\}\}\end{gathered} \tag{2.44}$$

One may define a tree diagram out of a chain of refinements, which is the partition tree. Recall that a tree diagram can be specified by collecting several nodes and assigning the relationship between them, i.e. assigning some nodes to be the parents/children/descendants/ancestors of other nodes. For a chain of refinements, one may define a tree whose nodes are sets in all partitions in this chain. One set $A$ in this tree is a descendant (resp. an ancestor) of another set $B$ if and only if $A \subseteq B$ (resp. $B \subseteq A$). Since the three sets in $\mathscr{P}_3$ do not have parent (or ancestor), we add a new node corresponding to the total set $\{K_1, ..., K_{2d+3}\}$. Then according to above rules, sets in $\mathscr{P}_3$ are children of it.

For example, (2.44) corresponding to the following tree. For simplicity, denote $K_{i,j} := \{K_i, , K_{i+1}, ..., K_j\}$.

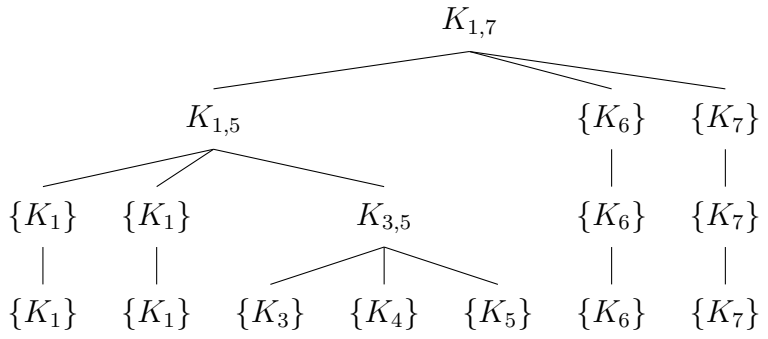


Figure 2.1: An Example of Trees Defined by a Chain of Refinements

In what follows, we omit labels for all nodes other than leaves (nodes without children), as the following picture demonstrated.

We may omit labels of nodes because these labels (or the subsets to which the nodes correspond) can be determined by a simple rule. For example, consider the node enclosed by

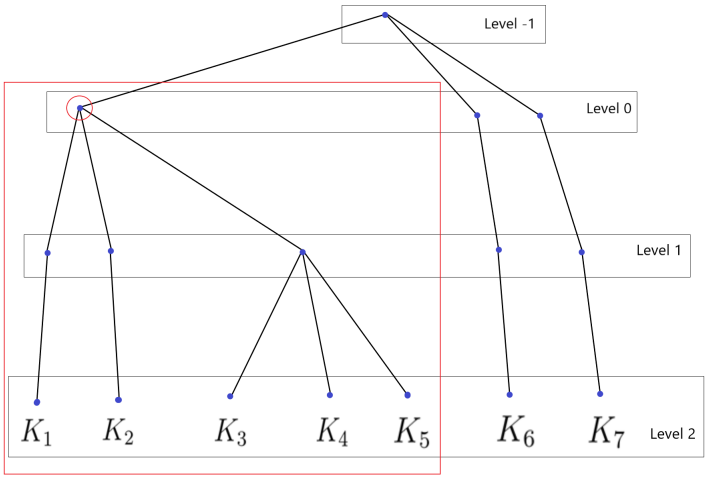


Figure 2.2: Trees with Labels Omitted

the red circle in Figure II (Let's call it $*$). The subtree rooted at $*$ is the subtree enclosed in the red rectangle. The bottom (collection of all leaf nodes) of this subtree are nodes $\{K_1\}$, $\{K_2\}$, ..., $\{K_5\}$. Then the set to which $*$ corresponds is the union of $\{K_1\}$, $\{K_2\}$, ..., $\{K_5\}$, which equals to $K_{1,5} = \{K_1, K_2, ..., K_5\}$. Hence the label of $*$ should be $K_{1,5}$. This label coincides with the label in Figure I.

Since the bottom of subtree rooted at $*$ are used very frequently in the rest, we denote it by $bot(*)$. We may also call it bottom of $*$ or bottom of subtree depending on the context. If we view nodes in a tree as sets, the bottom of a node is the set it corresponds to.

Each node in a tree is associated with a number, called level of these nodes. This is demonstrated in Figure II. For example, nodes at the bottom of this tree are enclosed in the black box labeled by level 2, so each node is of level 2.

### 2.3.2 Feynman diagrams and repeated normal form transformations

The Feynman diagram here will be a forest whose trees correspond to terms in the normal form coefficients $H^d$, $G^d$.

To see the connection between trees and normal forms, recall that in normal forms, we constantly have $\partial_t$ hitting on terms like $d_{K_1} \cdots d_{K_{2d+3}}$. When $d = 0$ and the derivative hits on $d_{K_1}$, we get $\partial_t d_{K_1} d_{K_2} d_{K_3}$ and we obtain $\sum d_{K_1'} d_{K_2'} d_{K_3'} d_{K_2} d_{K_3}$ by substituting the equation of $\partial_t d_{K_1}$. It becomes $\sum d_{K_1} d_{K_2} d_{K_3} d_{K_4} d_{K_5}$ by relabeling the indices $K_1' \to K_1$, $K_2' \to K_2$, $K_3' \to K_3$, $K_2 \to K_4$, $K_3 \to K_5$. Each time of differentiation by parts one $d_{K_j}$ becomes three $d_{K_l}$'s on the next level. Originally, we have $d_{K_1} d_{K_2} d_{K_3}$. After differentiation by parts $d$ times, we have $d_{K_1} \cdots d_{K_{2d+3}}$, so all $d_{K_j}$ in $d_{K_1} \cdots d_{K_{2d+3}}$ are children of the original $d_{K_j}$. The three sets of children of original $d_{K_j}$ constitute a 3-odd partition of $\{K_1, ..., K_{2d+3}\}$. The children sets of $d_{K_1}$, ..., $d_{K_5}$ in $d_{K_1} d_{K_2} \cdots d_{K_5}$ (terms from the first differentiation by parts) is a 5-odd partition. For the same reason, we also have a 7-odd partition, a 9-partition and so on. These partitions constitutes a full chain of refinement. We will see that it's convenient to use children sets to label all nodes.

Notice that the trees considered above (trees obtained from partitions or normal forms) are trees of a special type characterized by following properties.

*(1) The root node has three children.*

*(2) On any level d, exactly one node has three children and others have one child.* (2.45)

In the context of trees obtained from normal forms, the node $d_{K_j}$ hit by $\partial_t$ has three children. On level $d$ of a tree $T$ (resp. in $d$-th normal form), let $l_d^T$ denote the $l_d^T$-th node on level $d$ (from left to right) that has three children (resp. $d_{K_{l_d^T}}$ is hit by $\partial_t$). By (2.45), the number of nodes on level $d$ equals to $2d + 3$.

Given a tree $T$ satisfying (2.45), we want to determine the term to which it corresponds. In $d$-th differentiation by parts $\partial_t$ will hit on one of $d_{K_1}$,..., $d_{K_{2d+3}}$, and each choices will give a new term, so one $d$-th order term will split into $2d + 3$ new $d + 1$-th order terms. Thus all terms in $P$-th normal form are determined by a sequence of numbers $l_1^T, \cdots, l_P^T$ that specifies

which $d_{K_j}$ gets $\partial_t$ in each step. For a tree $T$ satisfying (2.45), we have also defined a sequence $l_1^T, \cdots, l_P^T$. So to obtain a term from a tree one can follow the map tree→sequence→term.

Coefficients of a term can be calculated from $T$. Each time of doing differentiation by parts gives a factor $1/\Omega_{2P+3}$ and a sign $\epsilon_{\mathcal{A}} \in \{-1, 0, 1\}$. Thus the coefficient $H^{P+1}_{K_1 \cdots K_{2P+3}}$ in front of $d_{K_1} \cdots d_{K_{2P+3}}$ should be of the form

$$\epsilon_{\mathcal{A}} / [\Omega_{3,K}(*,*,*)\Omega_{5,K}(*,...,*) \cdots \Omega_{2P+1,K}(*,...,*)]. \tag{2.46}$$

To see what the entries $*$ are, recall that there are $2d+3$ nodes $*_1, \cdots, *_{2d+3}$ on level $d$. The $2d+3$ entries $*$ in $\Omega_{2d+3,K}(*,...,*)$ are alternative sums of $bot(*_s)$. Recall that $bot(*_s)$ are bottoms of subtrees rooted at $*_s$ and alternative sums are defined as the following. Assume that

$$\{bot(*_1), \cdots, bot(*_{2d+3})\} = \{\{K_1, ..., K_{i_1}\}, \{K_{i_1+1}, ..., K_{i_2}\}, \cdots, \{K_{i_{2k+2}+1}, ..., K_{2d+3}\}\}, \tag{2.47}$$

where $i_s \equiv s \bmod(2)$. The alternative sum $S_{*_s}$ of $bot(*_s)$ is defined by

$$S_{*_s} = K_{i_s+1} - K_{i_s+2} + \cdots + K_{i_{s+1}}. \tag{2.48}$$

The alternative sum always ends with $+$ since $i_s \equiv s \bmod(2)$ implies that there are odd number of $K_j$'s in the sum. The sign $\epsilon_{\mathcal{A}}$ can be determined from trees but the rule is more complicated. We leave this task to Lemma 1. We will call alternative sums of $bot(*_s)$ alternative sums of $*_s$ depending on the context.

To see how the calculation works, we have the following example. When doing first differentiation by parts, we should get three terms depending on which $d_{K_j}$ in $d_{K_1}d_{K_2}d_{K_3}$ gets $\partial_t$. For example if $d_{K_1}$ gets $\partial_t$, we get a term,

$$-\frac{\epsilon^4}{2\pi L^4}\sum_{\substack{\mathcal{S}_{5,K}=0\\ \Omega_{5,K}=0}} d_{K_1}\overline{d_{K_2}}d_{K_3}\overline{d_{K_4}}d_{K_5}e(-\frac{\epsilon^2}{2\pi L^2}\int_0^t D_{5,K}(s)ds)\frac{1_{K_4\neq K_5,K_5\neq K}}{\Omega_{3,K}(K_1-K_2+K_3,K_4,K_5)}.$$

The normal form coefficients is

$$\frac{1_{K_4\neq K_5,K_5\neq K}}{\Omega_{3,K}(K_1-K_2+K_3,K_4,K_5)}. \tag{2.49}$$

The coefficient can be calculated by trees. Since $\partial_t$ hits $d_{K_1}$, the first node on level 0 have three children.

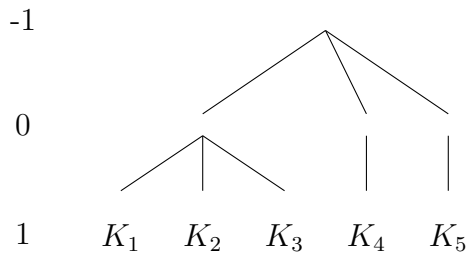


Figure 2.3: An Example of Feynman diagram

Notice that if $*_1$, $*_2$, $*_3$ are the three nodes of level 0 in Figure III, $S_{*_1}=K_1-K_2+K_3$, $S_{*_2}=K_4$, $S_{*_3}=K_5$. Taking $P=1$ in (2.46), up to a sign factor the coefficient is

$$\frac{1}{\Omega_{3,K}(K_1-K_2+K_3,K_4,K_5)}. \tag{2.50}$$

This coincides with (2.49).

To calculate the coefficients after doing differentiation by parts twice, assume that the first and the second $\partial_t$ both hit on $d_{K_1}$. Then the diagram is

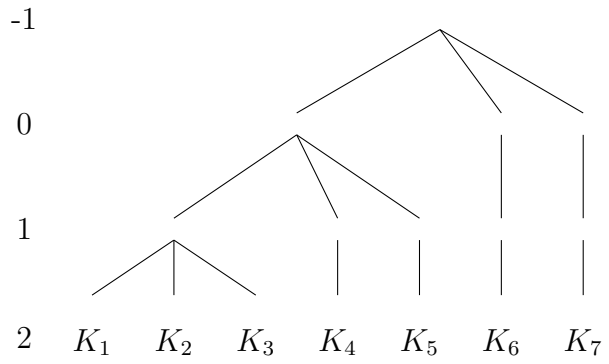


Figure 2.4: Another Example of Feynman diagram

Taking $P = 2$ in (2.46), the denominator of $H^3$ is of the form

$$\Omega_{3,K}(*,*,*)\Omega_{5,K}(*,*,*,*,*).$$

The alternative sums of nodes on level 0 are $K_1-K_2+K_3-K_4+K_5$, $K_6$, $K_7$. The alternative sums of nodes on level 1 are $K_1 - K_2 + K_3$, $K_4$, $K_5$, $K_6$, $K_7$. So the precise formula of the denominator is

$$\Omega_{3,K}(K_1 - K_2 + K_3 - K_4 + K_5, K_6, K_7)\Omega_{5,K}(K_1 - K_2 + K_3, K_4, K_5, K_6, K_7).$$

### 2.3.3 Correspondence of trees and terms

In this section, we summarize in Lemma 1 the correspondence between the trees and terms in the coefficients of $\widetilde{H}_K^{d+1}(u)$, $\widetilde{G}_K^{d+1}(u)$.

A distinguished feature of (2.54) is the appearance of the inequality constrains $1_{\mathcal{A}}$, $1_{\mathcal{B}}$. These inequality constrains will be crucial in the number theoretic estimate in section 2.4. Without them, some denominators $\Omega$ will vanish and the defining equation of (2.101) will have a large number of trivial solutions.

**Lemma 1.** *Denote by $\mathcal{G}_d$ the collection of all trees satisfying (2.45) of depth $d$ (leaf nodes of these trees are of level $d$). We have the following description of $H^{d+1}_{K_1\dots K_{2d+3}}$, $G^{d+1}_{K_1\dots K_{2d+3}}$.*

1. *Both* $H^{d+1}_{K_1...K_{2d+3}}$ *and* $G^{d+1}_{K_1...K_{2d+3}}$ *are sums of rational functions of* $K_1, ..., K_{2d+3}, K$ *multiplied by indicator functions. Any* $T \in \mathcal{G}_d$ *corresponds to one term in* $H^{d+1}_{K_1...K_{2d+3}}$ *or* $G^{d+1}_{K_1...K_{2d+3}}$.

2. *The corresponding term of* $T$ *in* $H^{d+1}_{K_1...K_{2d+3}}$ *is*

$$\frac{\epsilon_{\mathcal{A}}}{\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d+1,K}(T)}. \tag{2.51}$$

*The corresponding term of* $T$ *in* $G^{d+1}_{K_1...K_{2d+3}}$ *is*

$$\frac{\epsilon_{\mathcal{B}}}{\Omega_{5,K}(T)^2\Omega_{7,K}(T)\cdots\Omega_{2d+1,K}(T)}. \tag{2.52}$$

*Here*

$$\Omega_{2k+1,K}(T) = 2\pi\Omega_{2k+1,K}(S_{*_1}, S_{*_2}, \cdots, S_{*_{2k+1}}) \tag{2.53}$$

*for* $k = 1, 2, ..., d$, *where* $S_{*_j}$ *is the alternative sum of* $*_j$, *the* $j$*-th nodes on the level* $k-1$ *of* $T$. $\epsilon_{\mathcal{A}}$ *are functions of* $K_1$, ..., $K_{2d+3}$, $K$ *taking value in* $\{-1, 0, 1\}$ *and* $\epsilon_{\mathcal{B}}$ *is a sum of four such functions.*

3. *Let* $l^T_k$ *be the number such that* $*_{l^T_k}$ *is the only node on level* $k$ *that has three children. Assume* $*'$, $*''$ *and* $*'''$ *are three children of* $*_{l^T_k}$. *Let* $S^1_k = S_{*'}$, $S^2_k = S_{*''}$, $S^3_k = S_{*'''}$, $\mathcal{A}_k(T) = \mathcal{B}_k(T) = \{S^1_k \neq S^2_k, S^2_k \neq S^3_k, \Omega_{2k+3,K}(T) \neq 0\}$, $\mathcal{B}^{(s)}_0(T) = \{S^s_1 = S^{s+1}_1 = S^{s+2}_1\}$ $(1 \leq s \leq 3)$, $\mathcal{B}^{(4)}_0(T) = \{S^4_1 = S^5_1 = K\}$. *Then in above equations*

$$\epsilon_{\mathcal{A}} = (-1)^{\sum_{k=1}^{d-1} l^T_k} \prod_{k=1}^{d-1} 1_{\mathcal{A}_k(T)},$$

$$\epsilon_{\mathcal{B}} = \sum_{s=1}^{4} (-1)^{s+\sum_{k=2}^{d-1} l^T_k} 1_{\mathcal{B}^{(s)}_0(T)} \prod_{k=1}^{d-1} 1_{\mathcal{B}_k(T)}.$$

*Thus*

$$H^{d+1}_{K_1\dots K_{2d+3}} = \sum_{T\in\mathcal{G}_d} \frac{(-1)^{\sum_{k=0}^{d-1} l_k^T} \prod_{k=0}^{d-1} 1_{\mathcal{A}_k(T)}}{\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d+1,K}(T)}(K_1,...,K_{2d+3}). \tag{2.54}$$

$$G^{d+1}_{K_1\dots K_{2d+3}} = \sum_{s=1}^{4}\sum_{T\in\mathcal{G}_d} \frac{(-1)^{s+\sum_{k=2}^{d-1} l_k^T} 1_{\mathcal{B}_0^{(s)}(T)} \prod_{k=1}^{d-1} 1_{\mathcal{B}_k(T)}}{\Omega_{5,K}(T)^2\Omega_{7,K}(T)\cdots\Omega_{2d+1,K}(T)}(K_1,...,K_{2d+3})d_{K_1}\cdots d_{K_{2d+3}} \tag{2.55}$$

*Proof.* The result of $G$-coefficients can be proved in the same way as the that of $H$-coefficients. So we only consider the proof for $H$-coefficients.

Note that (1) follows easily from (2), (3). We shall prove (2), (3) by an induction on $d$. Assume that (2.54) holds for $1, 2, ..., d+1$, let's prove the case $d+2$, i.e.

$$H^{d+2}_{K_1\dots K_{2d+5}} = \sum_{T\in\mathcal{G}_{d+1}} \frac{(-1)^{\sum_{k=0}^{d} l_k^T} \prod_{k=0}^{d} 1_{\mathcal{A}_k}}{\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d+3,K}(T)}(K_1,...,K_{2d+5}) \tag{2.56}$$

By (2.31), we have

$$H^{d+2}_{K_1\dots K_{2d+5}} = \sum_{j=1}^{2d+3} \frac{(-1)^j 1_{K_j\neq K_{j+1}, K_{j+1}\neq K_{j+2}, \Omega_{2d+3}(\cdots)\neq 0} H^{d+1}_{K_1\dots(K_j-K_{j+1}+K_{j+2})\dots K_{2d+5}}}{2\pi\Omega_{2d+3,K}(K_1,...,K_j-K_{j+1}+K_{j+2},...,K_{2d+5})} \tag{2.57}$$

Now let's calculate the sum by substituting the induction assumption (2.56) into (2.57).

$$\begin{aligned} H^{d+2}_{K_1\dots K_{2d+5}} = \sum_{T\in\mathcal{G}_d}\sum_{j=1}^{2d+3} & \frac{(-1)^{\sum_{k=0}^{d-1} l_k^T} \prod_{k=0}^{d-1} 1_{\mathcal{A}_k} (-1)^j 1_{K_j\neq K_{j+1}, K_{j+1}\neq K_{j+2}, \Omega_{2d+3}(\cdots)\neq 0}}{\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d+1,K}(T)(K_1,...,K_j-K_{j+1}+K_{j+2},...,K_{2d+5})} \\ & \times \frac{1}{2\pi\Omega_{2d+3,K}(K_1,...,K_j-K_{j+1}+K_{j+2},...,K_{2d+5})} \end{aligned} \tag{2.58}$$

For any $T \in \mathcal{G}_d$, we can make a new tree $T_j \in \mathcal{G}_{d+1}$ by assigning three children to its $j$-th leaf node and one child to other leaf nodes. All trees $T' \in \mathcal{G}_{d+1}$ can be obtained in this way, i.e. there exist unique $T \in \mathcal{G}_d$ and $j$ such that $T' = T_j$. Then $\mathcal{G}_{d+1} = \cup_{T\in\mathcal{G}_d}\{T_1, \cdots, T_{2d+3}\}$. Thus we may reformulate (2.56) as

$$H^{d+2}_{K_1\dots K_{2d+5}} = \sum_{T\in\mathcal{G}_d}\sum_{j=1}^{2d+3} \frac{(-1)^{\sum_{k=0}^d l_k^T}\prod_{k=0}^d 1_{\mathcal{A}_k}}{\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d+3,K}(T)}(K_1, ..., K_{2d+5}) \tag{2.59}$$

Therefore, to prove (2.56), it suffices show that the summand in (2.58) equals to

$$\frac{(-1)^{\sum_{k=0}^d l_k^{T_j}}\prod_{k=0}^d 1_{\mathcal{A}_k}}{\Omega_{3,K}(T_j)\Omega_{5,K}(T_j)\cdots\Omega_{2d+3,K}(T_j)}(K_1, ..., K_{2d+5}), \tag{2.60}$$

the term corresponding to tree $T_j$. To do this, we only need to verify following equations.

$$\Omega_{2d+3,K}(T_j)(K_1, ..., K_{2d+5}) = 2\pi\Omega_{2d+3,K}(K_1, ..., K_j - K_{j+1} + K_{j+2}, ..., K_{2d+5}), \tag{2.61}$$

$$l^{T_j}_{d+1} = (-1)^j,\ \mathcal{A}_{d+1}(T_j) = \{K_j \neq K_{j+1}, K_{j+1} \neq K_{j+2}, \Omega_{2d+3}(\cdots) \neq 0\}, \tag{2.62}$$

$$\Omega_{2k+3,K}(T_j)(K_1, ..., K_{2d+5}) = \Omega_{2k+3,K}(T)(K_1, ..., K_j - K_{j+1} + K_{j+2}, ..., K_{2d+5}) \text{ for } k \leq d-1, \tag{2.63}$$

and

$$l^{T_j}_k = l^T_k,\ 1_{\mathcal{A}_k(T_j)}(K_1, ..., K_{2d+5}) = 1_{\mathcal{A}_k(T)}(K_1, ..., K_j - K_{j+1} + K_{j+2}, ..., K_{2d+5}) \text{ for } k \leq d-1. \tag{2.64}$$

Proof of (2.61) and (2.62): When generating $T_j$ from $T$, each node $K_i$ on the bottom of $T$ could have either one child labelled by $K_i$, one child labelled by $K_{i+2}$ or three child labelled by $K_j$, $K_{j+1}$, $K_{j+2}$, according to $i < j$, $i > j$ and $i = j$.

Thus in $\Omega_{2d+3,K}(T_j)$, all $S_{*_i} = K_i$, $K_{i+2}$ or $K_j - K_{j+1} + K_{j+2}$ according to $i < j$, $i > j$ or $i = j$. Thus by (2.53)

$$\Omega_{2d+3,K}(T_j)(K_1, ..., K_{2d+5}) = 2\pi\Omega_{2d+3,K}(K_1, ..., K_j - K_{j+1} + K_{j+2}, ..., K_{2d+5}) \tag{2.65}$$

Hence, we have proved (2.61). (2.62) can be proved similarly.

Proof of (2.63) and (2.64):

Because $\Omega_{2k+3,K}(T)$ and $1_{\mathcal{A}_k(T)}$ are functions of $S_*$ and by obvious reason $l_k^{T_j} = l_k^T$, we only need to verify that

$$S_{*_n}^{T_j}(K_1, ..., K_{2d+5}) = S_{*_n}^{T}(K_1, ..., K_j - K_{j+1} + K_{j+2}, ..., K_{2d+5}).$$

Here $S_{*_n}^{T_j}$ and $S_{*_n}^{T}$ are the alternative sums of bottom of $*_n$ in $T_j$ and $T$ respectively. Notice that nodes on the bottom of $T_j$ are children of nodes on the bottom of $T$. Denote by $*_1, ..., *_{2k+3}$ the $2k+3$ nodes of $T$ on level $k$. One nodes in $*_1, ..., *_{2k+3}$ is the ancestor of the $j$-th node $K_j$ on the bottom of $T$. Denote the bottom of $*_n$ In $T$ by $\{K_{i_n+1}, ..., K_{i_{n+1}}\}$ for $0 = i_1 < i_2 < \cdots < i_{2k+4} = 2d + 5$.

• For $n < n_0$ the bottom of $*_n$ in $T_j$ is also $\{K_{i_n+1}, ..., K_{i_{n+1}}\}$ and $i_{n+1} < j$. So

$$\begin{aligned} S_{*_n}^{T_j}(K_1, ..., K_{2d+5}) &= K_{i_n+1} - K_{i_n+2} + ... + K_{i_{n+1}} \\ =&S_{*_n}^{T}(K_1, ..., K_j - K_{j+1} + K_{j+2}, ..., K_{2d+5}). \end{aligned}$$

• For $n > n_0$, the bottom of $*_n$ in $T_j$ is $\{K_{i_n+3}, ..., K_{i_{n+1}+2}\}$ (the label of all nodes is shifted by 2 comparing to that of their parents). So

$$S^{T_j}_{*_n}(K_1, ..., K_{2d+5}) = K_{i_n+3} - K_{i_n+4} + ... + K_{i_{n+1}+2}$$
$$= S^{T}_{*_n}(K_1, ..., K_j - K_{j+1} + K_{j+2}, ..., K_{2d+5})$$

• For $n = n_0$, the bottom of $*_n$ is $\{K_{i_n+1}, ..., K_j, K_{j+1}, K_{j+2}, ..., K_{i_{n+1}+2}\}$ (the label of all nodes before $K_j$ is unchanged while the label of nodes after $K_j$ is shifted by 2, comparing to that of their parents). So

$$S^{T}_{*_{n_0}}(K_1, ..., K_j - K_{j+1} + K_{j+2}, ..., K_{2d+5}) = K_{i_{n_0}+1} - ... + (K_j - K_{j+1} + K_{j+2}) + ... + K_{i_{n_0+1}}$$
$$= K_{i_{n_0}+1} - ... + K_j - K_{j+1} + K_{j+2} + ... + K_{i_{n_0+1}}$$
$$= S^{T_j}_{*_{n_0}}(K_1, K_2, ..., K_{2d+5})$$

This completes the induction step. □

### 2.3.4 A vanishing property of normal form coefficients

In this section, we describe a vanishing property of the coefficients of normal form transformation, which states that $H^{d+1}_{K_1...K_{2d+3}}$ vanishes outside of a lower dimensional subvariety of the resonance surface $\mathscr{R} = \{(K_1, ..., K_{2d+3}) : \mathcal{S}_K(K_1, ..., K_{2d+3}) = \Omega_K(K_1, ..., K_{2d+3}) = 0\}$.

**Theorem 2.** *For nonlinear Schrodinger equation (2.2), the coefficients $H^{d+1}_{K_1...K_{2d+3}}$ of its normal form vanishes if $(K_1, ..., K_{2d+3}) \in \mathscr{R} \bigcap \left( \bigcap_{T \in \mathcal{G}_d} \bigcap_{k=0}^{d-1} \mathcal{A}_k(T) \right)$ for any $d \leq P$. Here $\mathcal{A}_k(T)$ is defined in Lemma 1 and*

$$\mathscr{R} = \{(K_1, ..., K_{2d+3}) : \mathcal{S}_K(K_1, ..., K_{2d+3}) = \Omega_K(K_1, ..., K_{2d+3}) = 0\} \tag{2.66}$$

**Remark 4.** *Because $\mathcal{A}_k(T)$ is the complement of hypersurfaces of the form $S_* = S_{*'}$ and $\Omega_{2d+1,K}(T)$, above theorem basically asserts that most of terms in $H^{d+1}_{K_1...K_{2d+3}}$ vanish. i.e. They vanish if $(K_1, \cdots, K_{2d+3})$ does not lie on a lower dimensional subvariety of the resonance surface $\mathscr{R}$.*

**Remark 5.** *To see this cancellation in the quintic form, recall that*

$$-i\partial_t c^1_K = \frac{i\epsilon^4}{2\pi L^4} \sum_{\mathcal{S}_{5,K}=0} d_{K_1}\overline{d_{K_2}} d_{K_3}\overline{d_{K_4}} d_{K_5} e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds)\Big[\frac{1_{K_1\neq K, K_5\neq K, K_2\neq K_3, K_3\neq K_4}}{\Omega_{3,K}(K_1, K_2 - K_3 + K_4, K_5)}$$
$$- \frac{1_{K_1\neq K_2, K_2\neq K_3, K_4\neq K_5, K_5\neq K}}{\Omega_{3,K}(K_1 - K_2 + K_3, K_4, K_5)} - \frac{1_{K_1\neq K_2, K_1\neq K, K_3\neq K_4, K_4\neq K_5}}{\Omega_{3,K}(K_1, K_2, K_3 - K_4 + K_5)}\Big]$$
$$+ \frac{\epsilon^4}{L^4} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}\neq 0}} \frac{1}{2\pi\Omega_{3,K}} d_{K_1}\overline{d_{K_2}} d_{K_3} D_{3,K}(t) e(\int_0^t \widetilde{\Omega}_{3,K}(s)ds).$$

*Apply this theorem, we have for any $(K_1, ..., K_5) \in \{\mathcal{S}_K(K_1, ..., K_5) = \Omega_K(K_1, ..., K_5) = 0\}$,*

$$\frac{1}{\Omega_{3,K}(K_1, K_2 - K_3 + K_4, K_5)}$$
$$- \frac{1}{\Omega_{3,K}(K_1 - K_2 + K_3, K_4, K_5)} - \frac{1}{\Omega_{3,K}(K_1, K_2, K_3 - K_4 + K_5)} = 0.$$

*Proof.* The proof of this theorem is based on induction on the level of trees in the Feynman diagram.

**Step 1 (Setting up the Notation):** To initiate the induction, let's first introduce some notation.

Chain of Refinements starting at $\mathscr{P}_{2k+3}$: If $\mathscr{P}_{2k+3} \in \mathbf{P}_{2k+3}$ are partitions and $\mathscr{P}_{2k+3} \leq \mathscr{P}_{2k+5} \leq \cdots \leq \mathscr{P}_{2d+3}$, then we call $\mathscr{P}_{2k+3}, \mathscr{P}_{2k+5}, \cdots, \mathscr{P}_{2d+3}$ a chain of refinements starting at $\mathscr{P}_{2k+3}$. In above definitions $\mathscr{P}_{2d+3} = \{\{K_1\}, \cdots, \{K_{2d+3}\}\}$.

Equivalence of Trees and Full Chains of Refinements: As explained before, for nodes on level $k$ in $T \in \mathcal{G}_d$, all their bottoms constitute a partition $\mathscr{P}_{2k+3}$ of $\{K_1, ..., K_{2d+3}\}$. $\{\mathscr{P}_{2k+3}\}_{k=0}^d$ constitute a full chain of refinements canonically associated with $T$. On the

other hand, given a full chain of refinements $\{\mathscr{P}_{2k+3}\}_{k=0}^{d}$, we may view all sets in $\mathscr{P}_{2k+3}$ as nodes of level $2k+3$. If a set $A \in \mathscr{P}_{2k+5}$ is a subset of a set $B \in \mathscr{P}_{2k+3}$, then $A$ is a child of $B$. We also introduce a virtual root node and let the three sets in $\mathscr{P}_3$ be its children. So we can also produce a canonical tree associated with a full chain of refinements. Thus these two objects are equivalent.

For a chain of refinements starting at $\mathscr{P}_{2k+3}$, we can also construct a tree $T$ in the way mentioned above. A virtual root is added so that the $2k+3$ sets in $\mathscr{P}_{2k+3}$ are its children. This tree almost satisfies (2.45) except at the root node which has $2k+3$ children. To define level in this tree, we assign the level of the root node to be $d-1$ so that nodes corresponding to a set in $\mathscr{P}_{2j+3}$ is on level $j$ in $T$. For any $\mathscr{P} \in \mathbf{P}_{2k+3}$ we define

$$\begin{aligned} \mathcal{G}_{d,k,\mathscr{P}} &= \{\textit{full chain of refinements starting at } \mathscr{P}\} \\ &= \{\textit{trees } T \textit{ constructed from chains of refinement starting at } \mathscr{P}\} \end{aligned} \tag{2.67}$$

$\Omega_{2k+3,K}(T)$ is a function of the alternating sums of nodes on level $k$. If we view trees as full chain of refinements and nodes as sets like $\{K_i, K_{i+1}, \cdots\}$, these alternative sums are $K_i - K_{i+1} + K_{i+2} + \cdots$.

Reformulation of (2.54): An important observation is

$$H^{d+1}_{K_1 \ldots K_{2d+3}} = \sum_{\mathscr{P} \in \mathbf{P}_3} \sum_{T \in \mathcal{G}_{d,1,\mathscr{P}}} \frac{(-1)^{l_0^T + \cdots + l_{d-1}^T}}{\Omega_{3,K}(T) \cdots \Omega_{2d+1,K}(T)} (K_1, ..., K_{2d+3}) \tag{2.68}$$

This follows easily from (2.54), if we are willing to view the sum over all trees as sum over all full chains of refinements, because summing over all chains clearly equals to summing over all chains with a fixed start followed by summing over all possible start.

Define sum over all chains of refinements starting at $\mathscr{P}$ as the following

$$\Sigma_{k,\mathscr{P}} = \sum_{T \in \mathcal{G}_{d,k,\mathscr{P}}} \frac{(-1)^{l_k^T + \cdots + l_{d-1}^T}}{\Omega_{2k+3,K}(T) \cdots \Omega_{2d+1,K}(T)} (K_1, ..., K_{2d+3}). \tag{2.69}$$

Here $l_j^T$ and $\Omega_{2k+3,K}(T)$ can be defined in the same way as before, since we have identified a chain of refinements with a tree $T$ and defined the notation of level in $T$.

Therefore

$$H^{d+1}_{K_1 \dots K_{2d+3}} = \sum_{\mathscr{P} \in \mathbf{P}_3} \Sigma_{1,\mathscr{P}}. \tag{2.70}$$

Basically $\Sigma_{k,\mathscr{P}}$ is what we called sum over all graph in which the partition consists of bottom of nodes of level $k$ equals to $\mathscr{P}$.

Define $l_1 = K_1 - K_2$, $l_2 = K_2 - K_3$, ...,$l_{2d+2} = K_{2d+2} - K_{2d+3}$.

We use an induction on $k$ to show that

$$\Sigma_{k,\mathscr{P}} = \frac{1}{2^{d-k} l_1 \cdots l_{i_1-1} \hat{l}_{i_1} l_{i_1+1} \cdots \hat{l}_{i_2} \cdots \hat{l}_{i_{2k+2}} \cdots l_{2d+2}} \tag{2.71}$$

Here $\hat{l}$ indicates the absence of this term.

We will first show that above conclusion holds when $k = d-1$, then assume the conclusion for $k+1$ and get the conclusion for $k$

**Step 2 (Verifying the Induction for $k = d-1$):** For $\mathscr{P} \in \mathbf{P}_{2d+1}$, $\mathscr{P}$ takes the form

$$\{\{K_1\}, \{K_2\}, \cdots, \{K_i, K_{i+1}, K_{i+2}\}, \cdots, \{K_{2d+3}\}\}.$$

By definition $\mathcal{G}_{d,d-1,\mathscr{P}}$ are all chain of refinements starting at $\mathscr{P}$, so there are only one element in $\mathcal{G}_{d,d-1,\mathscr{P}}$. Therefore

$$\Sigma_{d-1,\mathscr{P}} = \frac{(-1)^{l^T_{d-1}}}{\Omega_{2d+1,K}(T)(K_1, ..., K_{2d+3})}. \tag{2.72}$$

Since the $i$-th nodes on level $d-1$ which corresponds to $\{K_i, K_{i+1}, K_{i+2}\}$ have three children, by definition $l_j^T = i$ and its three children are $\{K_i\}$, $\{K_{i+1}\}$, $\{K_{i+2}\}$.

By definition of $\Omega_{2d+1,K}(T)$

$$\begin{aligned}\Omega_{2d+1,K}(T)(K_1,...,K_{2d+3}) &= \Omega_{2d+1,K}(K_1,K_2,\cdots,K_i-K_{i+1}+K_{i+2},\cdots,K_{2d+3})\\ &= K_1^2-K_2^2+...+(-1)^{i-1}(K_i-K_{i+1}+K_{i+2})^2+...+K_{2d+3}^2-K^2\\ &= (-1)^i(K_i^2-K_{i+1}^2+K_{i+2}^2-(K_i-K_{i+1}+K_{i+2})^2)\\ &= 2(-1)^i l_i l_{i+1}\end{aligned} \tag{2.73}$$

Here in the second equality, we apply the condition $\Omega_K(K_1,...,K_{2d+3})=0$ to cancel terms.

Thus

$$\Sigma_{d-1,\mathscr{P}} = \frac{(-1)^i}{2(-1)^i l_i l_{i+1}} = \frac{1}{2l_i l_{i+1}}. \tag{2.74}$$

This is exactly the equation (2.71) for $k=d-1$.

**Step 3 (Applying the Induction Hypothesis):** Assume that in $T\in\mathcal{G}_{d,k,\mathscr{P}}$ is a chain of refinements $\{\mathscr{P},\mathscr{P}_{2k+5},\cdots,\mathscr{P}_{2d+3}\}$ starting at $\mathscr{P}$ with

$$\mathscr{P}=\{\{K_1,...,K_{i_1}\},\{K_{i_1+1},...,K_{i_2}\},\cdots,\{K_{i_{2k+2}+1},...,K_{2d+3}\}\}.$$

Then all possibility of $\mathscr{P}_{2k+5}\in\mathbf{P}_{2k+3}$ are listed below (depending on which set is divided into three or in other words, which nodes on level $k$ have children)

(1) For $1\le i<j\le i_1$, $i$ odd, $j$ even,

$$\begin{aligned}\mathscr{P}_{1,ij}=\{&\{K_1,...,K_i\},\{K_{i+1},...,K_j\},\{K_{j+1},...,K_{i_1}\},\{K_{i_1+1},...,K_{i_2}\},\\ &\cdots,\{K_{i_{2k+2}+1},...,K_{2d+3}\}\}.\end{aligned}$$

(2) For $i_1 + 1 \leq i < j \leq i_2$, $i$ even, $j$ odd,

$$\mathscr{P}_{2,ij} = \{\{K_1, ..., K_{i_1}\}, \{K_{i_1+1}, ..., K_i\}, \{K_{i+1}, ..., K_j\}, \{K_{j+1}, ..., K_{i_2}\},$$
$$\cdots, \{K_{i_{2k+2}+1}, ..., K_{2d+3}\}\}.$$

$$\cdots$$

(2k+3) For $i_{2k+2} + 1 \leq i < j \leq 2k + 3$, $i$ odd, $j$ even,

$$\mathscr{P}_{2k+3,ij} = \{\{K_1, ..., K_{i_1}\}, \{K_{i_1+1}, ..., K_{i_2}\},$$
$$\cdots, \{K_{i_{2k+2}+1}, ..., K_i\}, \{K_{i+1}, ..., K_j\}, \{K_{j+1}, ..., K_{2k+3}\}\}.$$

Denote by $\mathbf{Ref}_{\mathscr{P}}$ the set of all $\mathscr{P}_{s,ij}$. If $\mathscr{P}_{2k+5} = \mathscr{P}_{s,ij}$, then we know $s$-th node on level $k$ of $T$ has children. Thus $l_k^T = s$

We can split the sum in the following way, (summing over all chains start at $\mathscr{P}$ equals to summing over all chains with second step $\mathscr{P}_{2k+5}$ followed by summing over all possibility of $\mathscr{P}_{2k+5}$)

$$\sum_{T \in \mathcal{G}_{d,k,\mathscr{P}}} = \sum_{\mathscr{P}' \in \mathbf{Ref}_{\mathscr{P}}} \sum_{T \in \mathcal{G}_{d,k+1,\mathscr{P}'}} . \tag{2.75}$$

Now notice that $\Omega_{2k+3,K}(T)$ depends only on the alternative sum of the sets in $\mathscr{P}_{2k+5}$ which implies that $\Omega_{2k+3,K}(T)$ is independent of the inner sum. So we have

$$\begin{aligned} \Sigma_{k,\mathscr{P}} &= \sum_{T \in \mathcal{G}_{d,k,\mathscr{P}}} \frac{(-1)^{l_k^T + \cdots + l_{d-1}^T}}{\Omega_{2k+3,K}(T) \cdots \Omega_{2d+1,K}(T)}(K_1, ..., K_{2d+3}) \\ &= \sum_{\mathscr{P}' \in \mathbf{Ref}_{\mathscr{P}}} \frac{(-1)^{l_k^T}}{\Omega_{2k+3,K}(T)} \sum_{T \in \mathcal{G}_{d,k+1,\mathscr{P}'}} \frac{(-1)^{l_{k+1}^T + \cdots + l_{d-1}^T}}{\Omega_{2k+5,K}(T) \cdots \Omega_{2d+1,K}(T)}(K_1, ..., K_{2d+3}) \end{aligned} \tag{2.76}$$

Since $\mathbf{Ref}_{\mathscr{P}}$ equals to the set of all $\{\mathscr{P}_{s,ij}\}$, we get

$$
\begin{aligned}
\Sigma_{k,\mathscr{P}} &= \sum_{T\in\mathcal{G}_{d,k,\mathscr{P}}} \frac{(-1)^{l_k^T+\cdots+l_{d-1}^T}}{\Omega_{2k+3,K}(T)\cdots\Omega_{2d+1,K}(T)}(K_1,...,K_{2d+3}) \\
&= \frac{1}{\Omega_{2k+3,K}(T)}\Bigg(\sum_{\substack{1\le s\le 2k+3\\ s \text{ odd}}}(-1)^s \sum_{\substack{i_{s-1}+1\le i<j\le i_s\\ i \text{ odd } j \text{ even}}}\sum_{T\in\mathcal{G}_{d,k+1,\mathscr{P}_{2k+3,ij}}} \\
&\frac{(-1)^{l_{k+1}^T+\cdots+l_{d-1}^T}}{\Omega_{2k+5,K}(T)\cdots\Omega_{2d+1,K}(T)}(K_1,...,K_{2d+3}) \\
&+ \sum_{\substack{1\le s\le 2k+3\\ s \text{ even}}}(-1)^s \sum_{\substack{i_{s-1}+1\le i<j\le i_s\\ i \text{ even } j \text{ odd}}}\sum_{T\in\mathcal{G}_{d,k+1,\mathscr{P}_{2k+3,ij}}} \frac{(-1)^{2^{d-k-1}l_{k+1}^T+\cdots+l_{d-1}^T}}{\Omega_{2k+5,K}(T)\cdots\Omega_{2d+1,K}(T)}(K_1,...,K_{2d+3})\Bigg).
\end{aligned}
\tag{2.77}
$$

Substitute the definition of $\Sigma_{k+1,\mathscr{P}_{2k+3,ij}}$ (2.69),

$$
\begin{aligned}
\Sigma_{k,\mathscr{P}} = \frac{1}{\Omega_{2k+3,K}(T)}\Bigg(&\sum_{\substack{1\le s\le 2k+3\\ s \text{ odd}}}(-1)^s \sum_{\substack{i_{s-1}+1\le i<j\le i_s\\ i \text{ odd } j \text{ even}}}\Sigma_{k+1,\mathscr{P}_{2k+3,ij}} \\
&+ \sum_{\substack{1\le s\le 2k+3\\ s \text{ even}}}(-1)^s \sum_{\substack{i_{s-1}+1\le i<j\le i_s\\ i \text{ even } j \text{ odd}}}\Sigma_{k+1,\mathscr{P}_{2k+3,ij}}\Bigg).
\end{aligned}
\tag{2.78}
$$

Then apply the induction hypothesis (2.71),

$$\Sigma_{k,\mathscr{P}} = \frac{1}{\Omega_{2k+3,K}(T)} \Bigg( \sum_{\substack{1\le s\le 2k+3 \\ s \text{ odd}}} \sum_{\substack{i_{s-1}+1\le i<j\le i_s \\ i \text{ odd } j \text{ even}}} \frac{1}{2^{d-k-1} l_1 \cdots \hat{l}_{i_1} \cdots \hat{l}_{i_{s-1}} \cdots \hat{l}_i \cdots \hat{l}_j \cdots \hat{l}_{i_s} \cdots l_{2d+2}}$$

$$- \sum_{\substack{1\le s\le 2k+3 \\ s \text{ odd}}} \sum_{\substack{i_{s-1}+1\le i<j\le i_s \\ i \text{ even } j \text{ odd}}} \frac{1}{2^{d-k-1} l_1 \cdots \hat{l}_{i_1} \cdots \hat{l}_{i_{s-1}} \cdots \hat{l}_i \cdots \hat{l}_j \cdots \hat{l}_{i_s} \cdots l_{2d+2}} \Bigg)$$

$$= \frac{1}{\Omega_{2k+3,K}(T)} \left( \sum_{\substack{1\le s\le 2k+3 \\ s \text{ odd}}} \sum_{\substack{i_{s-1}+1\le i<j\le i_s \\ i \text{ odd } j \text{ even}}} \frac{l_i l_j}{2^{d-k-1} l_1 \cdots \hat{l}_{i_1} \cdots l_{2d+2}} - \sum_{\substack{1\le s\le 2k+3 \\ s \text{ even}}} \sum_{\substack{i_{s-1}+1\le i<j\le i_s \\ i \text{ even } j \text{ odd}}} \cdots \right)$$

$$= \frac{1}{2^{d-k-1}\Omega_{2k+3,K}(T) l_1 \cdots \hat{l}_{i_1} \cdots l_{2d+2}} \left( \sum_{\substack{1\le s\le 2k+3 \\ s \text{ odd}}} \sum_{\substack{i_{s-1}+1\le i<j\le i_s \\ i \text{ odd } j \text{ even}}} l_i l_j - \sum_{\substack{1\le s\le 2k+3 \\ s \text{ even}}} \sum_{\substack{i_{s-1}+1\le i<j\le i_s \\ i \text{ even } j \text{ odd}}} l_i l_j \right) \tag{2.79}$$

By $\Omega_K(K_1, ..., K_{2d+3}) = 0$, we have

$$\begin{aligned}
\Omega_{2k+3,K}(T) &= (K_1 - ... + K_{i_1})^2 - (K_{i_1+1} - ... + K_{i_2})^2 + \cdots + (K_{i_{2k+2}} - ... + K_{2d+3})^2 - K^2 \\
&= \big((K_1 - ... + K_{i_1})^2 - K_1^2 + ... - K_{i_1})^2\big) - \big((K_{i_1+1} - ... + K_{i_2})^2 - K_{i_1+1}^2 + ... - K_{i_2}\big) \\
&+ \cdots + \big((K_{i_{2k+2}} - ... + K_{2k+3})^2 - K_{i_{2k+2}}^2 + ... - K_{2d+3}^2\big) \\
&= 2 \sum_{\substack{1\le i<j\le i_1 \\ i \text{ odd, } j \text{ even}}} l_i l_j - 2 \sum_{\substack{i_1+1\le i<j\le i_2 \\ i \text{ even, } j \text{ odd}}} l_i l_j + 2 \sum_{\substack{i_2+1\le i<j\le i_3 \\ i \text{ odd, } j \text{ even}}} l_i l_j - \cdots + 2 \sum_{\substack{i_{2k+2}+1\le i<j\le 2d+3 \\ i \text{ even, } j \text{ odd}}} l_i l_j \\
&= 2 \sum_{\substack{1\le s\le 2k+3 \\ s \text{ odd}}} \sum_{\substack{i_{s-1}+1\le i<j\le i_s \\ i \text{ odd } j \text{ even}}} l_i l_j - 2 \sum_{\substack{1\le s\le 2k+3 \\ s \text{ even}}} \sum_{\substack{i_{s-1}+1\le i<j\le i_s \\ i \text{ even } j \text{ odd}}} l_i l_j
\end{aligned} \tag{2.80}$$

Here in the second equality, we applied the condition $\Omega_K(K_1, ..., K_{2d+3}) = 0$.

Substituting (2.80) in (2.79), we get

$$\Sigma_{k,\mathscr{P}} = \frac{1}{2^{d-k} l_1 \cdots \hat{l}_{i_1} \cdots \hat{l}_{i_2} \cdots \hat{l}_{i_{2k+2}} \cdots l_{2d+2}} \tag{2.81}$$

So we finish the proof of (2.71).

**Step 4 (Concluding the Proof):** Now we take $k = 1$ in (2.71). There are three nodes on level 0. So their associated bottoms are

$$\mathscr{P}_{i_1 i_2} = \{\{K_1, ..., K_{i_1}\}, \{K_{i_1+1}, ..., K_{i_2}\}, \{K_{i_2+1}, ..., K_{2d+3}\}\}.$$

So (2.71) reads

$$\Sigma_{1,\mathscr{P}} = \frac{1}{2^d l_1 \cdots \hat{l}_{i_1} \cdots \hat{l}_{i_2} \cdots l_{2d+2}} \tag{2.82}$$

By (2.68)

$$\begin{aligned} H^{d+1}_{K_1 \ldots K_{2d+3}} &= \sum_{\mathscr{P} \in \mathbf{P}_3} \Sigma_{1,\mathscr{P}} \\ &= \sum_{\substack{1 \le i_1 < i_2 \le 2d+3 \\ i_1 \text{ odd},\ i_2 \text{ even}}} \frac{1}{2^d l_1 \cdots \hat{l}_{i_1} \cdots \hat{l}_{i_2} \cdots l_{2d+2}} \\ &= \frac{1}{2^d l_1 l_2 \cdots l_{2d+1} l_{2d+2}} \sum_{\substack{1 \le i_1 < i_2 \le 2d+3 \\ i_1 \text{ odd},\ i_2 \text{ even}}} l_{i_1} l_{i_j} \\ &= 0 \end{aligned} \tag{2.83}$$

The last line follows from

$$2 \sum_{\substack{i<j \\ i \text{ odd},\ j \text{ even}}} l_i l_j = K_1^2 - K_2^2 + \cdots + K_{2d+3}^2 - (K_1 - K_2 + \cdots + K_{2d+3})^2 = 0. \tag{2.84}$$

This completes the proof. □

## 2.4 The long time dynamics of 1D cubic NLS

This section is devoted to the proof of Theorem 1. As an application of the general tools we developed in section 2.3, we shall prove a long time dynamics result of (2.2).

### 2.4.1 Ideas of the proof

The differentiation by parts in normal forms is similar to integration by parts in oscillatory integrals. In oscillatory integrals, we gain $t^{-1}$ each time when doing integration by parts. In this section, we will show that similarly in normal forms, either $L^{-1}$ or $\epsilon^2$ factors appear when doing normal form. In following explanation, we often abuse $L^+$ and $L^0$, although sometimes their difference is crucial.

Ideas from number theory: We first introduce some number theoretical prerequisities that can help us to estimate the magnitude of terms in normal forms. Consider a sum over lattice $\mathbb{Z}_L^k$,

$$\sum_{f_1=0,\cdots,f_k=0} a(K_1)\cdots a(K_\ell). \tag{2.85}$$

Let $f_1,\cdots,f_k$ be linear functions in $K_1,\cdots,K_\ell$ and $a(K_1),\cdots,a(K_\ell)$ be rapidly decaying functions, then this sum is of order $L^{\ell-k}$. Because $a_{K_j}$ is rapidly decaying, we think it as a indicator function on interval $1_{[-1,1]}(K_j)$. Let $K_j=\frac{k_j}{L}$ for $j=1,\cdots,\ell$, then (2.85) is equal to number of integer solutions of

$$\{f_1(k_1,\cdots,k_l)=0,\cdots,f_k(k_1,\cdots,k_l)=0,\quad |k_j|\le L\}. \tag{2.86}$$

Since $f_j$ are linear functions, this is the number of integer points on a $l-k$ linear subspace. There should be in total $L^{l-k}$ solutions, since on a $l-k$ linear subspace there are $l-k$ freedoms and each freedom takes $L$ possible values.

Consider $f_1, \cdots, f_k$ being polynomial functions with $2\sum_{j=1}^{k} degf_j < l$, number theorists believe that the number of integer points of (2.86) is $L^{l-\sum_{j=1}^{k} degf_j}$. This principle has been verified in may concrete situations, while itself is a major conjecture in number theory, see for example [20] and [53]. Intuitively, $f$ has $L^{degf}$ possible values since $|k_j| \leq L$, $|f| \leq L^{degf}$, if we impose an equidistribution assumption that each value of $f$ are equally likely to appear, then the probability that $f = 0$ equals to $|f| \leq L^{-degf}$, which implies that the number of solutions to $f = 0$ equals to $L^{dim-degf}$. Thus $f = 0$ reduces the freedom by $degf$.

The assumption $2\sum_{j=1}^{k} degf_j < l$ is crucial to obtain the equidistribution result. The assumption means that the number of nontrivial solutions (equidistributed solutions) should exceed the number of trivial solutions. For example consider

$$k_1^3 + k_2^3 = k_3^3, \quad |k_j| \leq L. \tag{2.87}$$

We have a set of trivial solution $k_1 = -k_2$ and $k_3 = 0$, hence the number of solutions is at least $L$, whereas the equidistribution result only gives a constant number of solutions. However, the number theorists believe that, after remove a lower dimensional subvariety of (2.86), the equidistribution result should still hold and solutions on this lower dimensional subvariety are the trivial solutions. In above example, the subvariety should be $\{k_3 = 0\}$. If $k_3 \neq 0$, by the special case of the Fermat's last theorem, (2.87) has no solution. Therefore, we know the number of solutions can be bounded by any positive constant. Typically the number of nontrivial solutions is $L^{l-\sum_{j=1}^{k} degf_j}$ and the number of trivial solutions is $L^{\frac{\ell}{2}}$. The total number of solutions is $L^{l-\sum_{j=1}^{k} degf_j} + L^{\frac{\ell}{2}}$. When $2\sum_{j=1}^{k} degf_j < \ell$, the equidistributed solutions dominates, and this is where the condition, $2\sum_{j=1}^{k} degf_j < \ell$, comes from.

The time scale: In [45] and [23], they considered the NLS equation with dimension $d \geq 2$

where for fixed $K \in \mathbb{Z}_L^d$, the resonance set

$$\begin{cases} K_1 - K_2 + K_3 = K, \\ |K_1|^2 - |K_2|^2 + |K_3|^2 = |K|^2, \\ K_1, K_2, K_3 \in \mathbb{Z}_L^d. \end{cases} \tag{2.88}$$

has $3d$ variables. In their case the number of trivial solutions can be bounded by $L^d$, and the number of nontrivial solutions is of order $L^{2d-2}$ ($2d-2 = 3d-d-2$, $d$ linear equations, 1 quadratic equation). When $d \geq 2$, we have $L^{2d-2} \geq L^d$, i.e. the number of nontrivial solutions dominates. Consider the cubic term

$$\frac{\epsilon^2}{L^{2d}} \sum_{\mathcal{S}_{3,K}=0} a_{K_1} a_{K_2} a_{K_3} e(\Omega_{3,K} t). \tag{2.89}$$

Separate this into resonance and nonresonance interaction and remove the nonresonance interaction to higher order term by doing normal form. We have the cubic resonance interaction

$$\frac{\epsilon^2}{L^{2d}} \sum_{\substack{\mathcal{S}_{3,K}=0 \\ \Omega_{3,K}=0}} a_{K_1} a_{K_2} a_{K_3} e(\Omega_{3,K} t). \tag{2.90}$$

This sum is of magnitude $L^{2d-2}$ from the number theory analysis (think summand as $O(1)$ and we are suming $L^{2d-2}$ many $O(1)$s), hence the cubic resonance is of order $\frac{\epsilon^2}{L^2}$.

In general for any term like

$$\frac{\epsilon^{2P}}{L^{2Pd}} \sum_{\substack{f_1=0 \\ \cdots \\ f_k=0}} a_{K_1} \cdots a_{K_{2P+1}} e(\Omega_{2P+1,K} t), \tag{2.91}$$

its magnitude equals to $\epsilon^{2P}/L^k$, i.e. one constrain equations gives one $L^{-1}$.

For higher order terms, after doing normal form transformation once, one gains $\epsilon^2$. Since there are always to defining equations, one does not gain $L^{-1}$. Combining this with the

assumption $\epsilon^2 L^+ \ll 1$, after doing normal form $P$ times,

$$-i\partial_t a_K = \frac{\epsilon^2}{L^2}\left(\frac{1}{L^{2d-2}}\sum_{\substack{\mathcal{S}_{3,K}=0\\ \Omega_{3,K}=0}} a_{K_1}a_{K_2}a_{K_3}\right) + O(\frac{\epsilon^4}{L^2}) + \cdots + O(\frac{\epsilon^{2P}}{L^2}) + \epsilon^{2P+2}.$$

Consider the time of existence up to $\frac{L^2}{\epsilon^2}$, the contribution of $O(\frac{\epsilon^4}{L^2})$ can be bounded by $\int_0^t O(\frac{\epsilon^4}{L^2}) \leq \epsilon^2$. Therefore, when $t \leq \frac{L^2}{\epsilon^2}$, the dynamics of $a_K$ can be approximately described by (rescale $t$ by $s = \frac{\epsilon^2}{L^2}t$)

$$-i\partial_s a_K = \frac{1}{L^{2d-2}}\sum_{\substack{\mathcal{S}_{3,K}=0\\ \Omega_{3,K}=0}} a_{K_1}a_{K_2}a_{K_3},$$

whose left hand side after applying the circle method, converges to a integral, see Theorem 5 in [23].

In our case, the cubic resonance interaction is

$$\begin{cases} K_1 - K_2 + K_3 = K, \\ K_1^2 - K_2^2 + K_3^2 = K^2, \\ K_1, K_2, K_3 \in \mathbb{Z}_L. \end{cases} \tag{2.92}$$

Due to lack of number of variables, the number of trivial solutions dominates, which is of order $L$. Combining with the factor $\frac{\epsilon^2}{L^2}$, our cubic resonance interaction is of order $\frac{\epsilon^2}{L}$, which means the time which we can describe the solution is only up to $\frac{L}{\epsilon^2}$. To go beyond this time scale, we change the phase of the solution by $d_K = a_K e(\cdots)$ as in subsection 2.2.1 to remove the cubic resonance interaction and transform the cubic nonlinearity into

$$\frac{\epsilon^2}{L^2}\sum_{\substack{\mathcal{S}_{3,K}=0\\ \Omega_{3,K}\neq 0}} d_{K_1}d_{K_2}d_{K_3}e\left(\int_0^t \widetilde{\Omega}_{3,K}\right),$$

which can be transformed into quintic form by normal form transformation. And the quintic resonance interaction will be our main term that describes the long time dynamics up to time $t < \frac{L^2}{\epsilon^4}$, which is the wave turbulence time scale.

Outline of the Proof: First, we shall further analyze the main term $H_K^2(u)$. In [45] and [23], they approximate their main term using the circle method. We may also follow their strategy but because of the integrable structure, we have the cancellation property (Theorem 2) that we proved in section 2.3, which states that outside of a submanifold of the resonance surface, the normal form coefficients vanish. Apply this cancellation to $H_K^2(u)$, $H_K^2(u) = 0$ if $(K_1, \cdots, K_5) \in A_1 \bigcap A_2 \bigcap A_3$. And this will significantly simplify the approximation of the main term. The first following section is devoted to the simplification of the main term $H_K^2(u)$. This step is a refinement of Theorem 2 for quintic term. Notice that the vanishing property implies that the normal form coefficient is zero for summation indices outside a subvariety of $\mathcal{S} = 0, \Omega = 0$, so it has the same effect as adding a new constrain equality in the summation, which may imply a gain of $L^{-3}$ instead of $L^{-2}$ in $H_K^d(u)$, $d \geq 3$. However, a gain of $L^{-3}$ requires that there are not so many trivial solutions i.e. solutions of $\mathcal{S} = 0, \Omega = 0$ lie on a subvariety. This is not true in our main term $H_K^2(u)$, although the number of nontrivial solutions with non-vanishing normal form coefficient are of order $L^{2+}/L$, there are $L^2$ trivial solutions.

The next step is to treat all other terms as errors. We apply the explicit formula of the normal form coefficients in Lemma 1 to reduce the problem into a number theoretical problem (2.102). Because in (2.102) there are 1 linear equation and $d$ quadratic equations, number of nontrivial solutions are expected to be $L^{2d+1-1-2\times d} = L^+$. Therefore, we just need to exclude the potential possibility of $L^d$ trivial solutions ($K_1 = K_2$, $K_3 = K_4$, $\cdots$). We have to keep track inequality constrains in the explicit formulas to eliminate this possibility. A useful tool of getting bound and taking advantage of these constrains is the Lemma 3.

Combining all above estimates, we are able to conclude Proposition 1 by boostrap argument, which describes the dynamics of $d_K$ that is $u_K$ with a phase correction $\int_0^t |d_K(s)|^2 ds$.

To obtain the dynamics of $u_K$, a refined estimate of $|d_K(s)|^2$ is needed so that we may replace it by some known quantities. We will do energy estimate on $c_K$ which give us an estimate of $|c_K(s)|^2$ and then we show that $|c_K|$ differs from $|d_K|$ by some lower order error. Finally, we obtain the desired estimate on $\int_0^t |b_K(s)|^2 ds$ in terms of $\int_0^t |c_K(s)|^2 ds$.

### 2.4.2 Identification of the main term

In this section, we shall calculate the coefficients of $H_K^2$ carefully. In following derivation of the dynamics of $d_K$, we will think $H_K^2$ as the main term. All other terms will be errors.

Recall that

$$\frac{\epsilon^4}{L^4} H_K^2(u) = \frac{\epsilon^4}{2\pi L^4} \sum_{\substack{\mathcal{S}_{5,K}=0 \\ \Omega_{5,K}=0}} \Big[ \frac{1_{A_2}}{\Omega_{3,K}(K_1, K_2 - K_3 + K_4, K_5)} - \frac{1_{A_1}}{\Omega_{3,K}(K_1 - K_2 + K_3, K_4, K_5)} - \frac{1_{A_3}}{\Omega_{3,K}(K_1, K_2, K_3 - K_4 + K_5)} \Big] d_{K_1} \overline{d_{K_2}} d_{K_3} \overline{d_{K_4}} d_{K_5} e\Big( \int_0^t \widetilde{\Omega}_{5,K}(s) ds \Big),$$

where

$$A_1 = \{K_1 \neq K_2, K_2 \neq K_3, K_4 \neq K_5, K_5 \neq K\},$$

$$A_2 = \{K_1 \neq K, K_2 \neq K_3, K_3 \neq K_4, K_5 \neq K\},$$

$$A_3 = \{K_1 \neq K_2, K_1 \neq K, K_3 \neq K_4, K_4 \neq K_5\}.$$

For simplicity, let $H = H^2_{K_1 \dots K_5}$ and we will fix $K_1$, $K_2$, ..., $K_5$ in most part of this section. We know that

$$H = \frac{1_{l_1 \neq 0, l_2 \neq 0, l_1 + l_3 \neq 0, l_4 \neq 0}}{l_1 l_2} + \frac{1_{l_1 + l_3 \neq 0, l_2 + l_4 \neq 0, l_2 \neq 0, l_3 \neq 0}}{l_2 l_3} + \frac{1_{l_1 \neq 0, l_2 + l_4 \neq 0, l_3 \neq 0, l_4 \neq 0}}{l_3 l_4},$$

if we use the change of variables that we introduced in the earlier section

$$\begin{cases} l_1 = K_1 - K_2 \\ l_2 = K_2 - K_3 \\ l_3 = K_3 - K_4 \\ l_4 = K_4 - K_5 \end{cases} \tag{2.93}$$

The resonance surface equation $\Omega_{5,K} = 0$ becomes $l_1 l_2 + l_3 l_4 + l_1 l_4 = 0$.

By the principle of inclusion and exclusion, we know that

$$H = H1_{l_1 \neq 0, \ldots, l_4 \neq 0} + \sum_{i=1}^{4} H1_{l_i = 0} - \sum_{i,j=1}^{4} H1_{l_i = l_j = 0} + \cdots \tag{2.94}$$

From Theorem 2, we know that the following rational function

$$\frac{1}{\Omega_{3,K}(K_1, K_2 - K_3 + K_4, K_5)} - \frac{1}{\Omega_{3,K}(K_1 - K_2 + K_3, K_4, K_5)} - \frac{1}{\Omega_{3,K}(K_1, K_2, K_3 - K_4 + K_5)}$$

vanishes on the resonance surface $\mathcal{S}_{5,K} = 0$, $\Omega_{5,K} = 0$, which implies $H1_{l_1 \neq 0, \ldots, l_4 \neq 0} = 0$

All other terms also vanish, except

(1) $1_{l_1 = 0} H = \frac{1_{l_2 \neq 0, l_3 \neq 0}}{l_2 l_3}$. In this case by $l_1 = 0$, $l_1 l_2 + l_3 l_4 + l_1 l_4 = 0$ is equivalent to $l_3 l_4 = 0$. Then by $l_3 \neq 0$, we have $l_4 = 0$. Thus in this case $1_{l_1 = 0} H = 1_{l_1 = 0, l_4 = 0} H$.

(2) $1_{l_2 = 0} H = \frac{1_{l_1 \neq 0, l_3 \neq 0, l_4 \neq 0}}{l_3 l_4}$. For similar reason as in (1), $1_{l_2 = 0} H = 1_{l_2 = 0, l_1 + l_3 = 0} H$.

(3) $1_{l_3 = 0} H = \frac{1_{l_1 \neq 0, l_2 \neq 0, l_4 \neq 0}}{l_1 l_2}$. For similar reason as in (1), $1_{l_3 = 0} H = 1_{l_3 = 0, l_2 + l_4 = 0} H$.

(4) $1_{l_4 = 0} H = \frac{1_{l_2 \neq 0, l_3 \neq 0}}{l_2 l_3}$. For the same reason as (1), $1_{l_4 = 0} H = 1_{l_1 = 0, l_4 = 0} H$.

(5) $1_{l_1 = 0, l_4 = 0} H = \frac{1_{l_2 \neq 0, l_3 \neq 0}}{l_2 l_3}$.

From (1), (4), (5), we have

$$\sum_{\substack{\mathcal{S}_{5,K} = 0 \\ \Omega_{5,K} = 0}} (H^2_{K_1 \ldots K_5} 1_{l_1 = 0} + H^2_{K_1 \ldots K_5} 1_{l_4 = 0} - H^2_{K_1 \ldots K_5} 1_{l_1 = 0, l_4 = 0}) d_{K_1} \overline{d_{K_2}} d_{K_3} \overline{d_{K_4}} d_{K_5} e\Big(\int_0^t \widetilde{\Omega}_{5,K}(s) ds\Big)$$

$$= (2 - 1) \sum_{\substack{K_1 \neq K \\ K_2 \neq K}} \frac{|d_{K_1}|^2}{(K_1 - K)} \frac{|d_{K_2}|^2}{(K - K_2)} d_K = -\Big( \sum_{K_1 \neq K} \frac{|d_{K_1}|^2}{(K_1 - K)} \Big)^2 d_K.$$

From (2), (3), we have

$$\begin{aligned}
&\sum_{\substack{\mathcal{S}_{5,K}=0\\ \Omega_{5,K}=0}} (H^2_{K_1\dots K_5}1_{l_2=0}+H^2_{K_1\dots K_5}1_{l_3=0})d_{K_1}\overline{d_{K_2}}d_{K_3}\overline{d_{K_4}}d_{K_5}e(\int_0^t \widetilde{\Omega}_{5,K}(s)ds)\\
=&2\sum_{\substack{K_1\neq K\\ K_2\neq K_1}} \frac{|d_{K_1}|^2}{(K_2-K_1)}\frac{|d_{K_2}|^2}{(K_1-K)}d_K\\
=&-2\sum_{K_1\neq K}\frac{|d_{K_1}|^4}{(K_1-K)^2}d_K+\sum_{\substack{K_1\neq K\\ K_2\neq K_1}}\frac{|d_{K_1}|^2|d_{K_2}|^2}{(K_2-K_1)}\Big(\frac{1}{K_1-K}-\frac{1}{K_2-K}\Big)d_K\\
=&-3\sum_{K_1\neq K}\frac{|d_{K_1}|^4}{(K_1-K)^2}d_K+\Big(\sum_{K_1\neq K}\frac{|d_{K_1}|^2}{(K_1-K)}\Big)^2 d_K.
\end{aligned}$$

Therefore,

$$\frac{\epsilon^4}{L^4}H^2_K(u)=\frac{3\epsilon^4}{L^4}\sum_{K_1\neq K}\frac{|d_{K_1}|^4}{(K_1-K)^2}d_K. \tag{2.95}$$

### 2.4.3 Estimate of errors in normal form transformation

In this section, we shall obtain estimates of $H^d_K$, $d\geq 2$, $F^d_K$, $d\geq 1$, $\widetilde{H}^{P+1}_K$, $G^d_K$, $d\geq 3$, $E^d_K$, $d\geq 2$, $\widetilde{G}^{P+1}_K$, $S^d_K$ $d\geq 3$. Note that the equations (2.10) of $d_K$ contains two inequalities. All terms produced by normal form transformations inherits inequality constrains from these two constrains. These inequalities are crucial in excluding the presence of trivial solutions $K_1=K_2$, $K_2=K_3$, ..., in Lemma 5.

Three expression appears frequently in the expression of normal form coefficient, so we first obtain an estimate for them.

**Lemma 2.** *Fix $d,P\in\mathbb{N}$, $T\in\mathcal{G}_{d-1}$ or $\mathcal{G}_P$, $\ell>1$. Let $\{d_K\}_{K\in\mathbb{Z}_L}\in X^\ell$ i.e. $||d_K||_{X^\ell}=\sup_{K\in\mathbb{Z}_L}|\langle K\rangle^\ell d_K|<+\infty$. Then for any $K\in\mathbb{Z}_L$, we have*

$$\langle K\rangle^{\ell}\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0}}\frac{\prod_{k=0}^{d-2}1_{\mathcal{A}_k(T)}}{|\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d-1,K}(T)|}(K_1,...,K_{2d+1})d_{K_1}\cdots d_{K_{2d+1}}\lesssim L^{2(d-1)+}||d_K||_{X^\ell}^{2d+1},$$

$$\langle K\rangle^{\ell}\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}\neq 0}}\frac{\prod_{k=0}^{d-2}1_{\mathcal{A}_k(T)}}{|\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d+1,K}(T)|}(K_1,...,K_{2d+1})d_{K_1}\cdots d_{K_{2d+1}}\lesssim L^{2d+}||d_K||_{X^\ell}^{2d+1},$$

*and*

$$\langle K\rangle^{\ell}\sum_{\mathcal{S}_{2P+3,K}=0}\frac{\prod_{k=0}^{P-1}1_{\mathcal{A}_k(T)}}{|\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2P+1,K}(T)|}(K_1,...,K_{2P+3})d_{K_1}...d_{K_{2P+3}}\leq L^{2(P+1)+}||d_K||_{X^\ell}^{2P+1}.$$

*Proof.* We only show how to prove the first inequality. The second and the third ones can be handled similarly. It suffices to show that

$$\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0}}\frac{\prod_{k=0}^{d-2}1_{\mathcal{A}_k(T)}}{|\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d-1,K}(T)|}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\lesssim L^{2d-2+}\langle K\rangle^{-\ell}. \tag{2.96}$$

Since $K_1,...,K_{2d+1}\in\mathbb{Z}_L$, $\Omega_{2k+1,K}(T)=\frac{\mu_k}{L^2}$, $\mu_k\in\mathbb{Z}$, $k=1,\cdots,d-1$. Split the summation into two parts,

$$\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0}} \frac{\prod_{k=0}^{d-2} 1_{\mathcal{A}_k(T)}}{|\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d-1,K}(T)|}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}$$

$$= \sum_{\mu_1\in\mathbb{N}_+}\cdots\sum_{\mu_{d-1}\in\mathbb{N}_+}\frac{L^{2d-2}}{\mu_1\cdots\mu_{d-1}} \sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ |\Omega_{3,K}(T)|=\frac{\mu_1}{L^2},\cdots,|\Omega_{2d-1,K}(T)|=\frac{\mu_{d-1}}{L^2},\Omega_{2d+1,K}=0}} \tag{2.97}$$

$$\Big(\prod_{k=0}^{d-2} 1_{\mathcal{A}_k(T)}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\Big)$$

$$= \sum_{0\le\mu_k\le L^{10d},\forall 1\le k\le d-1} + \sum_{\mu_k\ge L^{10d},\exists k\le d-1}$$

**Estimate First Sum in (2.97):** For the first sum, since $K_1 - \cdots + K_{2d+1} = K$, we have $\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell} \le \langle K\rangle^{-l}$ (consider largest $K_j$ and bound it by $\langle K\rangle^{-\ell}$, and bound all other $\langle K_j\rangle^{-\ell}$ by 1). Therefore,

$$\sum_{0\le\mu_k\le L^{10d},\forall 1\le k\le d-1}\frac{L^{2d-2}}{\mu_1\cdots\mu_{d-1}} \sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ |\Omega_{3,K}(T)|=\frac{\mu_1}{L^2},\cdots,|\Omega_{2d-1,K}(T)|=\frac{\mu_{d-1}}{L^2},\Omega_{2d+1,K}=0}} \tag{2.98}$$

$$\Big(\prod_{k=0}^{d-2} 1_{\mathcal{A}_k(T)}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\Big)$$

$$\le \frac{L^{2d-2}}{\langle K\rangle^{\ell}}\sum_{0\le\mu_k\le L^{10d},\forall 1\le k\le d-1}\frac{1}{\mu_1\cdots\mu_{d-1}} \sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ |\Omega_{3,K}(T)|=\frac{\mu_1}{L^2},\cdots,|\Omega_{2d-1,K}(T)|=\frac{\mu_{d-1}}{L^2},\Omega_{2d+1,K}=0}} \prod_{k=0}^{d-2} 1_{\mathcal{A}_k(T)}$$

$$= \frac{L^{2d-2}}{\langle K\rangle^{\ell}}\sum_{0\le\mu_k\le L^{10d},\forall 1\le k\le d-1}\frac{1}{\mu_1\cdots\mu_{d-1}}\#\{(K_1,...,K_{2d+1})\in\cap_{k=0}^{d-2}\mathcal{A}_k(T) : \mathcal{S}_{2d+1,K}=0, | \tag{2.99}$$

$$\Omega_{3,K}(T)| = \frac{\mu_1}{L^2},\cdots,|\Omega_{2d-1,K}(T)| = \frac{\mu_{d-1}}{L^2},\Omega_{2d+1,K}=0\}$$

$$= \frac{L^{2d-2}}{\langle K\rangle^{\ell}}\sum_{0\le|\mu_k|\le L^{10d},\forall 1\le k\le d-1}\frac{1}{\mu_1\cdots\mu_{d-1}}\#_{\mu_1,...,\mu_{d-1};K} \tag{2.100}$$

Here in the last step we introduce the notation

$$\#_{\mu_1,...,\mu_{d-1},K} = \#\{(K_1,...,K_{2d+1}) \in \cap_{k=0}^{d-2}\mathcal{A}_k(T) : \mathcal{S}_{2d+1,K} = 0, |\Omega_{3,K}(T)| = \frac{\mu_1}{L^2} \\ ,\cdots,|\Omega_{2d-1,K}(T)| = \frac{\mu_{d-1}}{L^2}, \Omega_{2d+1,K} = 0\}. \tag{2.101}$$

We want to show that

$$\#_{\mu_1,...,\mu_{d-1};K} \lesssim \mu_1^+ \cdots \mu_{d-1}^+. \tag{2.102}$$

Replacing $K_j$ by $K_j + K$ in the definition of $\#_{\mu_1,...,\mu_{d-1};K}$, we get $\#_{\mu_1,...,\mu_{d-1};K} = \#_{\mu_1,...,\mu_{d-1};0}$. Denote $\#_{\mu_1,...,\mu_{d-1}} = \#_{\mu_1,...,\mu_{d-1};0}$. We want to show that

$$\#_{\mu_1,...,\mu_{d-1}} \lesssim \mu_1^+ \cdots \mu_{d-1}^+. \tag{2.103}$$

To prove this, we need the following three lemmas.

**Lemma 3.** *For any $\mu,\ \mu \in \mathbb{Z}$,*

$$\#\{(K_1,K_2,K_3) \in \mathbb{Z}_L^3 : K_1 \neq K_2, K_2 \neq K_3, K_1 - K_2 + K_3 = \frac{\mu}{L^2}, K_1^2 - K_2^2 + K_3^2 = \frac{\mu'}{L^2}\} \lesssim |\mu' - \mu^2|^+. \tag{2.104}$$

*Proof.* Let $K_j = \frac{k_j}{L}$ with $k_j \in \mathbb{Z}$, $j = 1,2,3$, we have

$$\begin{aligned}
&\#\{(K_1,K_2,K_3) \in \mathbb{Z}_L^3 : K_1 \neq K_2, K_2 \neq K_3, K_1 - K_2 + K_3 = \frac{\mu}{L^2}, K_1^2 - K_2^2 + K_3^2 = \frac{\mu'}{L^2}\} \\
=&\#\{(k_1,k_2,k_3) \in \mathbb{Z}^3 : k_1 \neq k_2, k_2 \neq k_3, k_1 - k_2 + k_3 = \mu, k_1^2 - k_2^2 + k_3^2 = \mu'\} \\
=&\#\{(k_1,k_2,k_3) \in \mathbb{Z}^3 : k_1 \neq k_2, k_2 \neq k_3, k_1 - k_2 + k_3 = \mu, (k_1 - k_2)(k_2 - k_3) = \mu' - \mu^2\} \\
=&\sum_{ab=|\mu'-\mu^2|} \#\{(k_1,k_2,k_3) \in \mathbb{Z}^3 : k_1 \neq k_2, k_2 \neq k_3, k_1 - k_2 + k_3 = \mu, k_1 - k_2 = a, k_2 - k_3 = b\} \\
\leq&\sum_{ab=|\mu'-\mu^2|} 1 = \#\{(a,b) \in \mathbb{Z}^2 : ab = |\mu' - \mu^2|\}
\end{aligned} \tag{2.105}$$

Now we need the following lemma,

**Lemma 4.** *For any number $n \in \mathbb{Z}$, the number of its divisors is bounded by $n^+$.*

*Proof.* We follow proposition 1.15 in [96]. Let $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$, then

$$\#\{k : k|n\} = (\alpha_1 + 1) \cdots (\alpha_r + 1).$$

Notice that $ln(n) = \alpha_1 ln(p_1) + \cdots + \alpha_r ln(p_r)$, so we have $\alpha_i \leq \frac{ln(n)}{ln(2)}$. And we also have $ln(1 + \alpha_i) \leq \alpha_i$. We will apply the first inequality when $p_i$ is small and the second when $p_i$ is large. Set a threshold $c$ which will be given later. Define $m = \max\{i : ln(p_i) < c\}$ ($m = 0$ if no such $i$). Thus $ln(m) < ln(p_m) < c$.

$$\begin{aligned}
ln(\#\{k : k|n\}) &= ln(\alpha_1 + 1) + \cdots + ln(\alpha_r + 1) \\
&= \sum_{k=1}^{m} ln(\alpha_k + 1) + \sum_{k=m}^{r} ln(\alpha_k + 1) \\
&\leq m \cdot ln(\frac{ln(n)}{ln(2)} + 1) + \sum_{k=m}^{r} \alpha_k \\
&\leq e^c ln(\frac{ln(n)}{ln(2)} + 1) + \frac{1}{c} \sum_{k=m}^{r} \alpha_k ln(p_k) \\
&\leq e^c ln(\frac{ln(n)}{ln(2)} + 1) + \frac{1}{c} ln(n)
\end{aligned}$$

Here the fourth inequlity is because $ln(m) < c$ and $ln(n) = \alpha_1 ln(p_1) + \cdots + \alpha_r ln(p_r)$.

Take $c = (1 - \delta) lnln(n)$. We have

$$ln(\#\{k : k|n\}) \leq ln^{1-\delta}(n) ln(\frac{ln(n)}{ln(2)} + 1) + (1 - \delta)^{-1} \frac{ln(n)}{lnln(n)} \lesssim \frac{ln(n)}{lnln(n)} \lesssim n^+.$$

□

Recall $ab = |\mu' - \mu^2|$, by Lemma 4, there are $|\mu' - \mu^2|^+$ possible values of $a$ and $b$. Thus $\#\{(a, b) \in \mathbb{Z}^2 : ab = |\mu' - \mu^2|\} \leq \#\{\text{divisor of } |\mu' - \mu^2|\}^2 \leq |\mu' - \mu^2|^+$. This completes the proof of Lemma 3. □

Given Lemma 3, we can prove the following lemma.

**Lemma 5.** *Fix $T \in \mathcal{G}_{d-1}$. If $(K_1, ..., K_{2d+1}) \in \{(K_1, ..., K_{2d+1}) \in \cap_{k=0}^{d-2}\mathcal{A}_k(T) : \mathcal{S}_{2d+1,0} = 0, |\Omega_{3,0}(T)| = \frac{\mu_1}{L^2}, \cdots, |\Omega_{2d-1,0}(T)| = \frac{\mu_{d-1}}{L^2}, \Omega_{2d+1,0} = 0\}$, then there are at most $\mu_1^+ \cdots \mu_{d-1}^+$ possible values of $S_*$ (see Lemma 1 for its definition) for any node $* \in T$.*

*Proof.* We prove this lemma by induction on the level of $*$.

For the three nodes on level 0, $*_1$, $*_2$, $*_3$, we have $S_{*_1} - S_{*_2} + S_{*_3} = 0$ and $S_{*_1}^2 - S_{*_2}^2 + S_{*_3}^2 = \pm\frac{\mu_1}{L^2}$. Then by Lemma 3, there are at most $\mu_1^+ \cdots \mu_{d-1}^+$ possible values of them. Note that the inequality constrains in $\mathcal{A}_1(T)$ guarantees $S_{*_1} \neq S_{*_2}$ and $S_{*_2} \neq S_{*_3}$.

Thus this claim is true for all nodes of level 0. Assume that this lemma is true for nodes of level $\leq k$, let us consider nodes on $k+1$-th level.

Let $*_1$,..., $*_{l_{k-1}^T}$,..., $*_{2k+1}$ be all nodes of level $k$. $*_{l_{k-1}^T}$ has three children $*'_{l_{k-1}^T}$, $*'_{l_{k-1}^T+1}$, $*'_{l_{k-1}^T+2}$. $*_j$ has only one child $*'_j$ or one child $*'_{j+2}$ for $j < l_{k-1}^T$ or $j > l_{k-1}^T$ respectively. Then $*'_1$,..., $*'_{l_{k-1}^T}$, $*'_{l_{k-1}^T+1}$, $*'_{l_{k-1}^T+2}$,..., $*'_{2k+3}$ are all nodes on level $k+1$.

By induction assumption there are at most $\mu_1^+ \cdots \mu_{d-1}^+$ possible values of $S_{*_j}$ for any $j$.

If $j \neq l_{k-1}^T$, then $S_{*'_j} = S_{*_j}$ or $S_{*'_{j+2}} = S_{*_j}$ for $j < l_{k-1}^T$ or $j > l_{k-1}^T$ respectively. By induction assumption, there are at most $\mu_1^+ \cdots \mu_{d-1}^+$ possible values of them.

If $j = l_{k-1}^T$, $S_{*'_j} - S_{*'_{j+1}} + S_{*'_{j+1}} = S_{*_j}$ and $S_{*'_j}^2 - S_{*'_{j+1}}^2 + S_{*'_{j+1}}^2 = \pm\frac{\mu_k}{L^2} - S_{*_1}^2 + \cdots$. By induction assumption, the right hand side of previous equations have only $\mu_1^+ \cdots \mu_{d-1}^+$ possible values. By the inequality constrains in $\mathcal{A}_k$, $S_{*'_j} \neq S_{*'_{j+1}}$, $S_{*'_{j+1}} \neq S_{*'_{j+2}}$. Thus by Lemma 3, there are at most $\mu_1^+ \cdots \mu_{d-1}^+$ possible values of $S_{*'_j}$, $S_{*'_{j+1}}$, $S_{*'_{j+1}}$. So we complete the induction. □

If $*$ is the $j$-th nodes on the bottom of $T$, then $S_* = K_j$. Thus there are at most $\mu_1^+ \cdots \mu_{d-1}^+$ possible values of all $K_j$, $1 \leq j \leq 2d+1$. This fact implies $\#_{\mu_1,...,\mu_{d-1}} \lesssim \mu_1^+ \cdots \mu_{d-1}^+$. So from (2.100), we know that the first sum in (2.97) can be bounded by

$$\begin{aligned}&\frac{L^{2d-2}}{\langle K\rangle^{\ell}}\sum_{0\leq\mu_k\leq L^{10d},\forall 1\leq k\leq d-1}\frac{1}{\mu_1\cdots\mu_{d-1}}\mu_1^{+}\cdots\mu_{d-1}^{+}\\ \lesssim& L^{2d-2+}\langle K\rangle^{-\ell}\end{aligned}\tag{2.106}$$

So the first sum can be bounded by the right hand side of (2.96).

**Estimate Second Sum in (2.97):** Since $\mu_k \geq L^{10d} \Leftrightarrow |\Omega_{2k+1,K}| \geq L^{10d-2}$, by a minute of reflection, we know that the second sum equals to

$$\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0\\ \exists k,\ |\Omega_{2k+1,K}(T)|\geq L^{10d-2}}}\frac{\prod_{k=0}^{d-2}1_{\mathcal{A}_k(T)}}{|\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d-1,K}(T)|}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\tag{2.107}$$

By inequality constrains in $\mathcal{A}_j$, $\Omega_{2j+1,K}(T)\neq 0$. Thus $|\Omega_{2j+1,K}(T)|\geq\frac{1}{L^2}$ by the fact that $(K_1,...,K_{2d+1})\in\mathbb{Z}_L^{2d+1}$.

So above expression can be bounded by

$$
\begin{aligned}
&\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0\\ \exists k,\ |\Omega_{2k+1,K}(T)|\geq L^{10d-2}}} L^{2d-2}L^{-(10d-2)}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\\
\lesssim& L^{-8d}\sum_{\mathcal{S}_{2d+1,K}=0}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\\
\lesssim& L^{-8d}\sum_{j=1}^{2d+1}\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ |K_j|=\max_m|K_m|}}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\\
\lesssim& L^{-8d}\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ |K_1|=\max_m|K_m|}}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\\
\lesssim& L^{-8d}\langle K\rangle^{-\ell}\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ |K_1|=\max_m|K_m|}}\langle K_2\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\\
\lesssim& L^{-8d}\langle K\rangle^{-\ell}\sum_{(K_2,\ldots,K_{2d+1})\in\mathbb{Z}_L^{2d}}\langle K_2\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\\
=& L^{-8d}\langle K\rangle^{-\ell}\left(\sum_{K_2\in\mathbb{Z}_L}\langle K_2\rangle^{-\ell}\right)^{2d}\\
\lesssim& L^{-6d}\langle K\rangle^{-\ell}
\end{aligned}
\tag{2.108}
$$

Here the third inequality uses the symmetry between $K_j$, $1\leq j\leq 2d+1$. The last inequality follows from the elementry inequality $\sum_{K_2\in\mathbb{Z}_L}\langle K_2\rangle^{-\ell}\leq L$.

So the second sum can also be bounded by the right hand side of (2.96). Thus (2.96) is proved. We have finished the first conclusion of this lemma.

**Proof of other conclusions:** The proof of the second conclusion of this lemma is essentially the same as above. Now let's describe how to prove the third one. As same as (2.97), we split the sum into two parts. The second part can be bounded by the same strategy as described above. The first sum can be bounded by

$$\sum_{0\le\mu_k\le L^{10(P+1)},\forall 1\le k\le P}\frac{L^{2P}}{\mu_1\cdots\mu_P}\sum_{\substack{\mathcal{S}_{2P+3,K}=0\\ |\Omega_{3,K}(T)|=\frac{\mu_1}{L^2},\cdots,|\Omega_{2P+1,K}(T)|=\frac{\mu_P}{L^2}}}\Big(\prod_{k=0}^{P-1}1_{\mathcal{A}_k(T)}\langle K_1\rangle^{-\ell}\cdots\langle K_{2P+3}\rangle^{-\ell}\Big).$$

Replacing $K_j$ by $K_j+K$, we get

$$\sum_{0\le\mu_k\le L^{10(P+1)},\forall 1\le k\le P}\frac{L^{2P}}{\mu_1\cdots\mu_P}\sum_{\substack{\mathcal{S}_{2P+3,0}=0\\ |\Omega_{3,0}(T)|=\frac{\mu_1}{L^2},\cdots,|\Omega_{2P+1,0}(T)|=\frac{\mu_P}{L^2}}}\Big(\prod_{k=0}^{P-1}1_{\mathcal{A}_k(T)}\langle K_1+K\rangle^{-\ell}\cdots\langle K_{2P+3}+K\rangle^{-\ell}\Big). \tag{2.109}$$

Note we no longer have the restriction $\Omega_{2d+3,K}=0$. Consider

$$\{(K_1,...,K_{2d+1})\in\cap_{k=0}^{d-2}\mathcal{A}_k(T):\mathcal{S}_{2d+1,0}=0,|\Omega_{3,0}(T)|=\frac{\mu_1}{L^2},\cdots,|\Omega_{2d-1,0}(T)|=\frac{\mu_{d-1}}{L^2}\}.$$

A similar induction works as in Lemma 5, except for the last step, due to the absence of $\Omega_{2d+3,K}=0$. But we know that for nodes $*_1$, ...,$*_{2P+1}$ on level $P-1$, $S_{*_1}$, ..., $S_{*_{2P+1}}$ has only $\mu_1^+\cdots\mu_P^+$ possible values. Since $S_{*_j}=K_j$ or $K_{j+2}$ according to $j<l_{P-1}^T$ or $j>l_{P-1}^T$ and $S_{*_{l_{P-1}^T}}=K_{l_{P-1}^T}-K_{l_{P-1}^T+1}+K_{l_{P-1}^T+2}$, we know that $K_1$, ..., $K_{l_{P-1}^T}-K_{l_{P-1}^T+1}+K_{l_{P-1}^T+2}$, ..., $K_{2P+3}$ all have only $\mu_1^+\cdots\mu_P^+$ possible values.

Let $j_0$ satisfies $|K_{j_0}+K|=\max_j|K_j+K|$. If $j_0\notin\{l_{P-1}^T,l_{P-1}^T+1,l_{P-1}^T+2\}$, we bound $\langle K_{j_0}\rangle^{-\ell}$ by $\langle K\rangle^{-\ell}$ and other $\langle K_j\rangle^{-\ell}$ by 1. Thus the first sum can be further bounded by

$$\begin{aligned}
&\frac{L^{2P}}{\langle K\rangle^{\ell}}\sum_{0\le\mu_k\le L^{10(P+1)},\forall 1\le k\le d-1}\frac{\mu_1^+\cdots\mu_P^+}{\mu_1\cdots\mu_{d-1}}\sum_{K_{l_{P-1}^T}-K_{l_{P-1}^T+1}+K_{l_{P-1}^T+2}\in\mathscr{K}}\langle K_{l_{P-1}^T}+K\rangle^{-\ell}\\
&\times\langle K_{l_{P-1}^T+1}+K\rangle^{-\ell}\langle K_{l_{P-1}^T+2}+K\rangle^{-\ell}\\
=&\frac{L^{2P}}{\langle K\rangle^{\ell}}\sum_{0\le\mu_k\le L^{10(P+1)},\forall 1\le k\le d-1}\frac{\mu_1^+\cdots\mu_P^+}{\mu_1\cdots\mu_{d-1}}\sum_{a\in\mathscr{K}}\sum_{K_{l_{P-1}^T},K_{l_{P-1}^T+1}\in\mathbb{Z}}\langle K_{l_{P-1}^T}+K\rangle^{-\ell}\\
&\times\langle K_{l_{P-1}^T+1}+K\rangle^{-\ell}\langle a-K_{l_{P-1}^T}+K_{l_{P-1}^T+1}+K\rangle^{-\ell}\\
\lesssim&\frac{L^{2P}}{\langle K\rangle^{\ell}}\sum_{0\le\mu_k\le L^{10(P+1)},\forall 1\le k\le d-1}\frac{\mu_1^+\cdots\mu_P^+}{\mu_1\cdots\mu_{d-1}}\sum_{K_{l_{P-1}^T},K_{l_{P-1}^T+1}\in\mathbb{Z}}\langle K_{l_{P-1}^T}+K\rangle^{-\ell}\langle K_{l_{P-1}^T+1}+K\rangle^{-\ell}\\
\lesssim&\frac{L^{2P+}}{\langle K\rangle^{\ell}}(\sum_{K_{l_{P-1}^T}\in\mathbb{Z}}\langle K_{l_{P-1}^T}+K\rangle^{-\ell})^2\lesssim\frac{L^{2P+2+}}{\langle K\rangle^{\ell}},
\end{aligned}\tag{2.110}$$

where $\mathscr{K}$ is some set with cardinality $\le\mu_1^+\cdots\mu_P^+$. The last inequality follows from the elementry inequality $\sum_{K_2\in\mathbb{Z}_L}\langle K_{l_{P-1}^T}+K\rangle^{-\ell}\le L$.

If $j_0\in\{l_{P-1}^T,l_{P-1}^T+1,l_{P-1}^T+2\}$, without loss of generality assume that $j_0=l_{P-1}^T+2$, we have $\langle K_{l_{P-1}^T+2}+K\rangle\ge\langle K\rangle$. After bounding all $\langle K_j\rangle^{-\ell}$ by 1, the first sum can be further

bounded by

$$
\begin{aligned}
& L^{2P} \sum_{0\le\mu_k\le L^{10(P+1)},\forall 1\le k\le d-1} \frac{\mu_1^+\cdots\mu_P^+}{\mu_1\cdots\mu_{d-1}} \sum_{\substack{K_{l_{P-1}^T}-K_{l_{P-1}^T+1}+K_{l_{P-1}^T+2}\in\mathscr{K}\\ \langle K_{l_{P-1}^T+2}+K\rangle\ge\langle K\rangle}} \langle K_{l_{P-1}^T}+K\rangle^{-\ell} \\
& \times \langle K_{l_{P-1}^T+1}+K\rangle^{-\ell}\langle K_{l_{P-1}^T+2}+K\rangle^{-\ell} \\
=& \frac{L^{2P}}{\langle K\rangle^{\ell}} \sum_{0\le\mu_k\le L^{10(P+1)},\forall 1\le k\le d-1} \frac{\mu_1^+\cdots\mu_P^+}{\mu_1\cdots\mu_{d-1}} \sum_{K_{l_{P-1}^T},K_{l_{P-1}^T+1}\in\mathbb{Z}} \langle K_{l_{P-1}^T}+K\rangle^{-\ell}\langle K_{l_{P-1}^T+1}+K\rangle^{-\ell} \\
\lesssim& \frac{L^{2P}}{\langle K\rangle^{\ell}} \sum_{0\le\mu_k\le L^{10(P+1)},\forall 1\le k\le d-1} \frac{\mu_1^+\cdots\mu_P^+}{\mu_1\cdots\mu_{d-1}} \sum_{K_{l_{P-1}^T},K_{l_{P-1}^T+1}\in\mathbb{Z}} \langle K_{l_{P-1}^T}+K\rangle^{-\ell}\langle K_{l_{P-1}^T+1}+K\rangle^{-\ell} \\
\lesssim& \frac{L^{2P+}}{\langle K\rangle^{\ell}} \Big(\sum_{K_{l_{P-1}^T}\in\mathbb{Z}} \langle K_{l_{P-1}^T}+K\rangle^{-\ell}\Big)^2 \lesssim \frac{L^{2P+2+}}{\langle K\rangle^{\ell}}
\end{aligned}
\tag{2.111}
$$

where $\mathscr{K}$ is some set with cardinality $\le \mu_1^+\cdots\mu_P^+$. So we get a bound on the first sum, thus we finished the proof of the last conclusion. This completes the proof of Lemma 2. □

**Lemma 6.** *For any* $3\le d\le P$, $\ell>1$,

$$
\|H_K^d(u)\|_{X_K^\ell} \lesssim L^{2(d-1)+}\|u\|_{X^\ell}^{2d+1} .
$$

*Proof.* Recall that

$$
H_K^d(u) = \sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0}} H^d_{K_1K_2\ldots K_{2d+1}} d_{K_1}\overline{d_{K_2}}\ldots d_{K_{2d+1}} e\Big(\int_0^t D_{2d+1,K}(s)ds\Big), \tag{2.112}
$$

By (2.54),

$$
H^d_{K_1\ldots K_{2d+1}} = \sum_{T\in\mathcal{G}_{d-1}} \frac{(-1)^{\sum_{k=0}^{d-2} l_k^T}\prod_{k=0}^{d-2} 1_{\mathcal{A}_k(T)}}{\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d-1,K}(T)}(K_1,\ldots,K_{2d+1}). \tag{2.113}
$$

By (2.112),

$$|H_K^d(u)| \le \sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0}} |H^d_{K_1K_2...K_{2d+1}}||d_{K_1}\overline{d_{K_2}}...d_{K_{2d+1}}|$$
$$\le \sum_{T\in\mathcal{G}_{d-1}} \sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0}} \frac{\prod_{k=0}^{d-2} 1_{\mathcal{A}_k(T)}}{|\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d-1,K}(T)|}(K_1,...,K_{2d+1})|d_{K_1}\overline{d_{K_2}}...d_{K_{2d+1}}|. \tag{2.114}$$

By the first conclusion in Lemma 2,

$$|H_K^d(u)| \le |\mathcal{G}_{d-1}|||d_K||^{2d+1}_{X^\ell_K}\frac{L^{2(d-1)+}}{\langle K\rangle^\ell} \lesssim ||u||^{2d+1}_{X^\ell}\frac{L^{2(d-1)+}}{\langle K\rangle^\ell}, \tag{2.115}$$

where $|\mathcal{G}_{d-1}|$ is the number of trees in $\mathcal{G}_{d-1}$. Thus $||H_K^d(u)||_{X^\ell} \lesssim L^{2(d-1)+}||u||^{2d+1}_{X^\ell}$. This completes the proof. □

**Lemma 7.** *For any $1 \le d \le P$, $\ell > 1$,*

$$\|F_K^d(u)\|_{X^\ell_K} \lesssim L^{2d+}\|u\|^{2d+1}_{X^\ell} \ .$$

*Proof.* Recall that

$$F_K^d(u) = \sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}\neq 0}} H^d_{K_1K_2...K_{2d+1}} d_{K_1}\overline{d_{K_2}}...d_{K_{2d+1}}\frac{1}{2\pi\Omega_{2d+1,K}}e(\int_0^t \widetilde{\Omega}_{2d+1,K}(s)ds). \tag{2.116}$$

Denote

$$F^d_{K_1K_2...K_{2d+1}} = \frac{H^d_{K_1K_2...K_{2d+1}}1_{\Omega_{2d+1,K}\neq 0}}{2\pi\Omega_{2d+1,K}(K_1,...,K_{2d+1})}, \tag{2.117}$$

then

$$F_K^d(u) = \sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}\neq 0}} F^d_{K_1K_2...K_{2d+1}} b_{K_1}\overline{b_{K_2}}...b_{K_{2d+1}}e(\int_0^t \widetilde{\Omega}_{2d+1,K}(s)ds). \tag{2.118}$$

By Lemma 1,

$$F^d_{K_1...K_{2d+1}}=\sum_{T\in\mathcal{G}_{d-1}}\frac{(-1)^{\sum_{k=0}^{d-2}l_k^T}\prod_{k=0}^{d-2}1_{\mathcal{A}_k(T)}}{\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d+1,K}(T)}(K_1,...,K_{2d+1}). \tag{2.119}$$

By (2.118),

$$\begin{aligned}|F^d_K(u)|&\leq\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}\neq 0}}|F^d_{K_1K_2...K_{2d+1}}||b_{K_1}\overline{b_{K_2}}...b_{K_{2d+1}}|\\ &\leq\sum_{T\in\mathcal{G}_{d-1}}\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}\neq 0}}\frac{\prod_{k=0}^{d-2}1_{\mathcal{A}_k(T)}}{|\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d-1,K}(T)|}(K_1,...,K_{2d+1})|d_{K_1}\overline{d_{K_2}}...d_{K_{2d+1}}|.\end{aligned} \tag{2.120}$$

By the second conclusion in Lemma 2,

$$|F^d_K(u)|\leq|\mathcal{G}_{d-1}|||d_K||^{2d+1}_{X^\ell_K}\frac{L^{2d+}}{\langle K\rangle^\ell}\lesssim||u||^{2d+1}_{X^\ell}\frac{L^{2d+}}{\langle K\rangle^\ell}, \tag{2.121}$$

where $|\mathcal{G}_{d-1}|$ is the number of trees in $\mathcal{G}_{d-1}$. Thus $||F^d_K(u)||_{X^\ell}\lesssim L^{2(d-1)+}||u||^{2d+1}_{X^\ell}$. □

**Lemma 8.** *For any $\ell>1$*

$$\|\widetilde{H}^{P+1}_K\|_{X^\ell_K}\lesssim L^{2P+2+}\|u\|^{2P+3}_{X^\ell}\ .$$

*Proof.* Recall that

$$\widetilde{H}^{P+1}_K(u)=\sum_{\mathcal{S}_{2P+3,K}=0}H^{P+1}_{K_1K_2...K_{2P+3}}d_{K_1}\overline{d_{K_2}}...d_{K_{2P+3}}e(\int_0^t\widetilde{\Omega}_{2P+3,K}(s)ds), \tag{2.122}$$

By (2.54),

$$H^{P+1}_{K_1...K_{2P+3}}=\sum_{T\in\mathcal{G}_P}\frac{(-1)^{\sum_{k=0}^{P-1}l_k^T}\prod_{k=0}^{P-1}1_{\mathcal{A}_k(T)}}{\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2P+1,K}(T)}(K_1,...,K_{2P+3}). \tag{2.123}$$

By (2.122),

$$|\widetilde{H}_K^{P+1}(u)| \leq \sum_{\mathcal{S}_{2P+3,K}=0} |H^{P+1}_{K_1K_2...K_{2P+3}}||d_{K_1}\overline{d_{K_2}}...d_{K_{2P+3}}|$$
$$\leq \sum_{T\in\mathcal{G}_P}\sum_{\mathcal{S}_{2P+3,K}=0} \frac{\prod_{k=0}^{P-1}1_{\mathcal{A}_k(T)}}{|\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2P+1,K}(T)|}(K_1,...,K_{2P+3})|d_{K_1}\overline{d_{K_2}}...d_{K_{2P+3}}|. \tag{2.124}$$

By the last conclusion in Lemma 2,

$$|\widetilde{H}_K^{P+1}(u)| \leq |\mathcal{G}_P|||d_K||^{2P+3}_{X^\ell_K}\frac{L^{2P+2+}}{\langle K\rangle^\ell} \lesssim ||u||^{2P+3}_{X^\ell}\frac{L^{2P+2+}}{\langle K\rangle^\ell}, \tag{2.125}$$

where $|\mathcal{G}_P|$ is the number of trees in $\mathcal{G}_P$. Thus $||\widetilde{H}_K^{P+1}(u)||_{X^\ell} \lesssim L^{2P+2+}||u||^{2P+3}_{X^\ell}$. □

**Lemma 9.** *Fix $d, P \in \mathbb{N}$, $T \in \mathcal{G}_{d-1}$ or $\mathcal{G}_P$, $\ell > 1$, $s = 1,2,3,4$. Let $\{d_K\}_{K\in\mathbb{Z}_L} \in X^\ell$ i.e. $||d_K||_{X^\ell} = \sup_{K\in\mathbb{Z}_L}|\langle K\rangle^\ell d_K| < +\infty$. Then for any $K \in \mathbb{Z}_L$, we have*

$$\langle K\rangle^\ell \sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0}} \frac{1_{\mathcal{B}_0^{(s)}(T)}\prod_{k=1}^{d-2}1_{\mathcal{B}_k(T)}}{|\Omega_{5,K}(T)^2\Omega_{7,K}(T)\cdots\Omega_{2d-1,K}(T)|}(K_1,...,K_{2d+1})d_{K_1}\cdots d_{K_{2d+1}} \lesssim L^{2(d-1)+}||d_K||^{2d+1}_{X^\ell},$$

$$\langle K\rangle^\ell \sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}\neq 0}} \frac{1_{\mathcal{B}_0^{(s)}(T)}\prod_{k=1}^{d-2}1_{\mathcal{B}_k(T)}}{|\Omega_{5,K}(T)^2\Omega_{7,K}(T)\cdots\Omega_{2d+1,K}(T)|}(K_1,...,K_{2d+1})d_{K_1}\cdots d_{K_{2d+1}} \lesssim L^{2d+}||d_K||^{2d+1}_{X^\ell},$$

*and*

$$\langle K\rangle^\ell \sum_{\mathcal{S}_{2P+3,K}=0} \frac{1_{\mathcal{B}_0^{(s)}(T)}\prod_{k=1}^{P-1}1_{\mathcal{B}_k(T)}}{|\Omega_{5,K}(T)^2\Omega_{7,K}(T)\cdots\Omega_{2P+1,K}(T)|}(K_1,...,K_{2P+3})d_{K_1}...d_{K_{2P+3}} \leq L^{2(P+1)+}||d_K||^{2P+1}_{X^\ell}.$$

*Proof.* Because the idea of this proof is identical to Lemma 2, so the proof here will be sketchy. We first prove the first inequality. In the end of the proof, we explain how to prove the second and the third ones similarly. As same as before, it suffices to show that

$$\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0}} \frac{1_{\mathcal{B}_0^{(s)}(T)}\prod_{k=1}^{d-2}1_{\mathcal{B}_k(T)}}{|\Omega_{5,K}(T)^2\Omega_{7,K}(T)\cdots\Omega_{2d-1,K}(T)|}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\lesssim L^{2d-2+}\langle K\rangle^{-\ell}. \quad (2.126)$$

Since $K_1,...,K_{2d+1}\in\mathbb{Z}_L$, $\Omega_{2k+1,K}(T)=\frac{\mu_k}{L^2}$, $\mu_k\in\mathbb{Z}$. The same as before, split the summation into two parts,

$$\begin{aligned}
&\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0}} \frac{1_{\mathcal{B}_0^{(s)}(T)}\prod_{k=1}^{d-2}1_{\mathcal{B}_k(T)}}{|\Omega_{5,K}(T)^2\Omega_{7,K}(T)\cdots\Omega_{2d-1,K}(T)|}\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\\
&=\sum_{\mu_2\in\mathbb{N}_+}\cdots\sum_{\mu_{d-1}\in\mathbb{N}_+}\frac{L^{2d-2}}{\mu_2^2\cdots\mu_{d-1}}\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ |\Omega_{5,K}(T)|=\frac{\mu_2}{L^2},\cdots,|\Omega_{2d-1,K}(T)|=\frac{\mu_{d-1}}{L^2},\Omega_{2d+1,K}=0}}\left(1_{\mathcal{B}_0^{(s)}(T)}\prod_{k=1}^{d-2}1_{\mathcal{B}_k(T)}\langle K_1\rangle^{-\ell}\cdots\right)\\
&=\sum_{0\le\mu_k\le L^{10d},\forall 2\le k\le d-1}+\sum_{\mu_k\ge L^{10d},\exists k\le d-1}
\end{aligned} \quad (2.127)$$

**Estimate First Sum in (2.127):** For the first sum, using the same argument as before, since $K_1-\cdots+K_{2d+1}=K$, we have $\langle K_1\rangle^{-\ell}\cdots\langle K_{2d+1}\rangle^{-\ell}\le\langle K\rangle^{-\ell}$. (Simply consider largest $K_j$ and bound it by $\langle K\rangle^{-\ell}$. Bound all other $\langle K_j\rangle^{-\ell}$ by 1) So we have,

$$\sum_{0\le\mu_k\le L^{10d},\forall 2\le k\le d-1}\frac{L^{2d-2}}{\mu_2^2\cdots\mu_{d-1}}\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ |\Omega_{5,K}(T)|=\frac{\mu_2}{L^2},\cdots,|\Omega_{2d-1,K}(T)|=\frac{\mu_{d-1}}{L^2},\Omega_{2d+1,K}=0}}\Big(1_{\mathcal{B}_0^{(s)}(T)}\prod_{k=0}^{d-2}1_{\mathcal{B}_k(T)}\langle K_1\rangle^{-\ell}\cdots\Big)$$

$$\le\frac{L^{2d-2}}{\langle K\rangle^{\ell}}\sum_{0\le\mu_k\le L^{10d},\forall 2\le k\le d-1}\frac{1}{\mu_2^2\cdots\mu_{d-1}}\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ |\Omega_{5,K}(T)|=\frac{\mu_2}{L^2},\cdots,|\Omega_{2d-1,K}(T)|=\frac{\mu_{d-1}}{L^2},\Omega_{2d+1,K}=0}}1_{\mathcal{B}_0^{(s)}(T)}\prod_{k=0}^{d-2}1_{\mathcal{B}_k(T)}$$

$$=\frac{L^{2d-2}}{\langle K\rangle^{\ell}}\sum_{0\le\mu_k\le L^{10d},\forall 2\le k\le d-1}\frac{1}{\mu_2^2\cdots\mu_{d-1}}\#\{(K_1,...,K_{2d+1})\in\cap_{k=0}^{d-2}\mathcal{B}_k(T)\cap\mathcal{B}_0^{(s)}(T):\mathcal{S}_{2d+1,K}=0,$$

$$|\Omega_{5,K}(T)|=\frac{\mu_2}{L^2},\cdots,|\Omega_{2d-1,K}(T)|=\frac{\mu_{d-1}}{L^2},\Omega_{2d+1,K}=0\}$$

$$=\frac{L^{2d-2}}{\langle K\rangle^{\ell}}\sum_{0\le|\mu_k|\le L^{10d},\forall 2\le k\le d-1}\frac{1}{\mu_2^2\cdots\mu_{d-1}}\#_{\mu_2,...,\mu_{d-1};K}\tag{2.128}$$

Here in the last step we introduce the notation $\#_{\mu_2,...,\mu_{d-1},K}$. We want to show that

$$\#_{\mu_2,...,\mu_{d-1};K}\lesssim\mu_2^+\cdots\mu_{d-1}^+.\tag{2.129}$$

Replacing $K_j$ by $K_j+K$ in the definition of $\#_{\mu_2,...,\mu_{d-1};K}$, we get $\#_{\mu_1,...,\mu_{d-1};K}=\#_{\mu_1,...,\mu_{d-1};0}$. Denote $\#_{\mu_2,...,\mu_{d-1}}=\#_{\mu_2,...,\mu_{d-1};0}$. We want to show that

$$\#_{\mu_2,...,\mu_{d-1}}\lesssim\mu_2^+\cdots\mu_{d-1}^+.\tag{2.130}$$

We prove (2.130) inductively.

**Lemma 10.** *Fix $T\in\mathcal{G}_{d-1}$. If $(K_1,...,K_{2d+1})\in\{(K_1,...,K_{2d+1})\in\cap_{k=0}^{d-2}\mathcal{B}_k(T)\cap\mathcal{B}_0^{(s)}(T):\mathcal{S}_{2d+1,0}=0,|\Omega_{5,0}(T)|=\frac{\mu_2}{L^2},\cdots,|\Omega_{2d-1,0}(T)|=\frac{\mu_{d-1}}{L^2},\Omega_{2d+1,0}=0\}$, then there are at most $\mu_2^+\cdots\mu_{d-1}^+$ possible values of $S_*$ (see Lemma 1 for its definition) for any node $*\in T$.*

*Proof.* This lemma is proved by induction on the level of the nodes. In the proof, we only consider the case that $s=1$. In all other cases the proof has no essential changes.

Unlike the proof of Lemma 5, we don't have $\Omega$ on level 0. For the five nodes on level 1, $*_1$, $*_2$, $*_3$, $*_4$, $*_5$, we have $S_{*_1} - S_{*_2} + S_{*_3} - S_{*_4} + S_{*_5} = 0$ and $S_{*_1}^2 - S_{*_2}^2 + S_{*_3}^2 - S_{*_4}^2 + S_{*_5}^2 = \pm\frac{\mu_2}{L^2}$. Unlike the proof of Lemma 5, we have several equations in $\mathcal{B}_0^{(s)}(T)$, i.e. $S_{*_1} = S_{*_2} = S_{*_3}$.

Applying them, we get $S_{*_3} - S_{*_4} + S_{*_5} = 0$ and $S_{*_3}^2 - S_{*_4}^2 + S_{*_5}^2 = \pm\frac{\mu_2}{L^2}$. Then by Lemma 3, there are at most $\mu_1^+ \cdots \mu_{d-1}^+$ possible values of $S_{*_1}$, $S_{*_2}$, $S_{*_3}$, $S_{*_4}$, $S_{*_5} = 0$. Note that the inequality constrains in $\mathcal{B}_1(T)$ guarantees $S_{*_3} \neq S_{*_4}$ and $S_{*_4} \neq S_{*_5}$.

Assume that this lemma is true for nodes of level less than $k$, let us consider nodes on $k$-th level. The following argument is identically the same as that in the proof of Lemma 5.

Let $*_1$,..., $*_{l_{k-1}^T}$,..., $*_{2k+1}$ be all nodes of level $k$. $*_{l_{k-1}^T}$ has three children $*'_{l_{k-1}^T}$, $*'_{l_{k-1}^T+1}$, $*'_{l_{k-1}^T+2}$. $*_j$ has only one child $*'_j$ for $j < l_{k-1}^T$, one child $*'_{j+2}$ for $j > l_{k-1}^T$.

If $j \neq l_{k-1}^T$, $S_{*'_j} = S_{*_j}$ or $S_{*'_{j+2}} = S_{*_j}$ depending on $j < l_{k-1}^T$ or $j > l_{k-1}^T$. Thus by induction assumption, there are at most $\mu_1^+ \cdots \mu_{d-1}^+$ possible values of them.

If $j = l_{k-1}^T$, $S_{*'_j} - S_{*'_{j+1}} + S_{*'_{j+1}} = S_{*_j}$ and $S_{*'_j}^2 - S_{*'_{j+1}}^2 + S_{*'_{j+1}}^2 = \pm\frac{\mu_k}{L^2} - S_{*_1}^2 + \cdots$. By induction assumption, the right hand side of previous equations have only $\mu_1^+ \cdots \mu_{d-1}^+$ possible values. By the inequality constrains in $\mathcal{B}_k$, $S_{*'_j} \neq S_{*'_{j+1}}$, $S_{*'_{j+1}} \neq S_{*'_{j+2}}$. Thus by Lemma 3, there are at most $\mu_1^+ \cdots \mu_{d-1}^+$ possible values of $S_{*'_j}$, $S_{*'_{j+1}}$, $S_{*'_{j+1}}$. So we complete the induction. $\square$

If $*$ is the $j$-th nodes on the bottom of $T$, then $S_* = K_j$. Thus there are at most $\mu_1^+ \cdots \mu_{d-1}^+$ possible values of all $K_j$, $1 \leq j \leq 2d+1$. This fact implies $\#_{\mu_1,\ldots,\mu_{d-1}} \lesssim \mu_1^+ \cdots \mu_{d-1}^+$. So from (2.128), we know that the first sum in (2.127) can be bounded by

$$\begin{aligned} &\frac{L^{2d-2}}{\langle K\rangle^{\ell}} \sum_{0\leq\mu_k\leq L^{10d},\forall 1\leq k\leq d-1} \frac{1}{\mu_1\cdots\mu_{d-1}} \mu_1^+\cdots\mu_{d-1}^+ \\ \lesssim & L^{2d-2+}\langle K\rangle^{-\ell} \end{aligned} \tag{2.131}$$

So the first sum can be bounded by the right hand side of (2.126).

**Estimate Second Sum in (2.127):** Since $\mu_k \geq L^{10d} \Leftrightarrow |\Omega_{2k+1,K}| \geq L^{10d-2}$, by a minute of reflection, we know that the second sum equals to

$$\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0\\ \exists k,\ |\Omega_{2k+1,K}(T)|\geq L^{10d-2}}} \frac{1_{\mathcal{B}_0^{(s)}(T)} \prod_{k=1}^{d-2} 1_{\mathcal{B}_k(T)}}{|\Omega_{5,K}(T)^2 \cdots \Omega_{2d-1,K}(T)|} \langle K_1\rangle^{-l} \cdots \langle K_{2d+1}\rangle^{-l} \tag{2.132}$$

By inequality constrains in $\mathcal{B}_j$, $\Omega_{2j+1,K}(T) \neq 0$. Thus $|\Omega_{2j+1,K}(T)| \geq \frac{1}{L^2}$ since $(K_1, ..., K_{2d+1}) \in \mathbb{Z}_L^{2d+1}$. So above expression can be bounded by

$$\begin{aligned}
&\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0\\ \exists k,\ |\Omega_{2k+1,K}(T)|\geq L^{10d-2}}} L^{2d-2} L^{-(10d-2)} 1_{\mathcal{B}_0^{(s)}(T)} \langle K_1\rangle^{-\ell} \cdots \langle K_{2d+1}\rangle^{-\ell} \\
\lesssim &\sum_{\substack{\mathcal{S}_{2d+1,K}=0\\ \Omega_{2d+1,K}=0\\ \exists k,\ |\Omega_{2k+1,K}(T)|\geq L^{10d-2}}} L^{2d-2} L^{-(10d-2)} \langle K_1\rangle^{-\ell} \cdots \langle K_{2d+1}\rangle^{-\ell} \qquad (2.133)\\
\lesssim &L^{-6d} \langle K\rangle^{-\ell}
\end{aligned}$$

The last inequality follows from the same argument as in (2.108) in Lemma 2. Note that $\mathcal{B}_0^{(s)}(T)$ produces two equations. If we take advantage of it in the second step, we can get a gain of $L^{-2}$, which we don't need here.

The second sum can also be bounded by the right hand side of (2.126). Thus (2.126) is proved. We have finished the first conclusion of this lemma.

**Proof of Other Conclusions:** The proof of the second conclusion of this lemma is essentially the same as above. The proof of the third one follows from of the argument of the last part in Lemma 2, because the same induction argument works in Lemma 10 as in Lemma 5. So we complete the proof of Lemma 9. □

**Lemma 11.** *For any* $3 \leq d \leq P$, $\ell > 1$,

$$\|G_K^d(u)\|_{X_K^\ell} \lesssim L^{2(d-1)+} \|u\|_{X^\ell}^{2d+1} .$$

*Proof.* Replacing all the letter 'H' by 'G' and 'Lemma 2' by 'Lemma 9' in the proof of Lemma 6 we get a proof of this lemma. □

**Lemma 12.** *For any* $2 \le d \le P$, $\ell > 1$,

$$\|E_K^d(u)\|_{X_K^\ell} \lesssim L^{2d+}\|u\|_{X^\ell}^{2d+1} .$$

*Proof.* Replacing all the letter 'F' by 'E' and 'Lemma 2' by 'Lemma 9' in the proof of Lemma 7 we get a proof of this lemma. □

**Lemma 13.** *For any* $\ell > 1$

$$\|\widetilde{G}_K^{P+1}\|_{X_K^\ell} \lesssim L^{2d+2+}\|u\|_{X^\ell}^{2d+3} .$$

*Proof.* Replacing all the letter '$\widetilde{H}$' by '$\widetilde{G}$' and 'Lemma 2' by 'Lemma 9' in the proof of Lemma 8 we get a proof of this lemma. □

**Lemma 14.** *For any* $\ell > 1$, $3 \le d \le P$, *assume that* $\epsilon^2\|u\|_{X^\ell}^2 \le \frac{1}{10}$ *if* $d = 3$, *then we have*

$$\|S_K^d(u)\|_{X_K^\ell} \lesssim L^{2(d-1)+}\|u\|_{X^\ell}^{2d+1} .$$

*Proof.* Recall that

$$S_K^{d+2}(u) = \sum_{\mathcal{S}_{2d+3,K}=0} (H^{d+1}_{K_1K_2...K_{2d+3}} + G^{d+1}_{K_1K_2...K_{2d+3}}) d_{K_1}\overline{d_{K_2}}...d_{K_{2d+3}} \frac{D_{2d+3,K}(t)}{2\pi\Omega_{2d+3,K}} e(\int_0^t \widetilde{\Omega}_{2d+3,K}(s)ds). \tag{2.134}$$

Denote

$$S^{d+2}_{K_1K_2...K_{2d+3}} = \frac{(H^{d+1}_{K_1K_2...K_{2d+3}} + G^{d+1}_{K_1K_2...K_{2d+3}})1_{\Omega_{2d+3,K}\neq 0}}{2\pi\Omega_{2d+3,K}(K_1,...,K_{2d+3})}, \tag{2.135}$$

then

$$S_K^{d+2}(u)=\sum_{\mathcal{S}_{2d+3,K}=0}S^{d+2}_{K_1K_2...K_{2d+3}}b_{K_1}\overline{b_{K_2}}...b_{K_{2d+3}}D_{2d+3,K}(t)e(\int_0^t\widetilde{\Omega}_{2d+3,K}(s)ds). \tag{2.136}$$

By Lemma 1,

$$\begin{aligned}S^{d+2}_{K_1...K_{2d+3}}&=\sum_{T\in\mathcal{G}_d}\frac{(-1)^{\sum_{k=0}^{d-1}l_k^T}\prod_{k=0}^{d-1}1_{\mathcal{A}_k(T)}}{\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d+3,K}(T)}(K_1,...,K_{2d+3}).\\+\sum_{j=1}^{4}\sum_{T\in\mathcal{G}_d}&\frac{1^{s+\sum_{k=2}^{d-1}l_k^T}1_{\mathcal{B}_0^{(s)}(T)}\prod_{k=1}^{d-1}1_{\mathcal{B}_k(T)}}{\Omega_{5,K}(T)^2\Omega_{7,K}(T)\cdots\Omega_{2d+3,K}(T)}(K_1,...,K_{2d+3})d_{K_1}\cdots d_{K_{2d+3}}\end{aligned} \tag{2.137}$$

By (2.136),

$$\begin{aligned}|S_K^{d+2}(u)|&\leq\sum_{\mathcal{S}_{2d+3,K}=0}|F^{d+1}_{K_1K_2...K_{2d+3}}||b_{K_1}\overline{b_{K_2}}...b_{K_{2d+3}}||D_{2d+3,K}(t)|\\&\leq||u||_{X^\ell}^2\sum_{T\in\mathcal{G}_d}\sum_{\substack{\mathcal{S}_{2d+3,K}=0\\\Omega_{2d+3,K}\neq0}}\frac{\prod_{k=0}^{d-1}1_{\mathcal{A}_k(T)}}{|\Omega_{3,K}(T)\Omega_{5,K}(T)\cdots\Omega_{2d+3,K}(T)|}(K_1,...,K_{2d+3})|d_{K_1}\overline{d_{K_2}}...d_{K_{2d+3}}|\\&+||u||_{X^\ell}^2\sum_{T\in\mathcal{G}_d}\sum_{\substack{\mathcal{S}_{2d+3,K}=0\\\Omega_{2d+3,K}\neq0}}\frac{1_{\mathcal{B}_0^{(s)}(T)}\prod_{k=1}^{d-1}1_{\mathcal{B}_k(T)}}{|\Omega_{5,K}(T)^2\Omega_{7,K}(T)\cdots\Omega_{2d+3,K}(T)|}(K_1,...,K_{2d+3})|d_{K_1}\overline{d_{K_2}}\cdots d_{K_{2d+3}}|.\end{aligned} \tag{2.138}$$

By the second conclusion in Lemma 2 and Lemma 9,

$$|S_K^{d+2}(u)|\leq|\mathcal{G}_d|||d_K||_{X_K^\ell}^{2d+5}\frac{L^{2d+2+}}{\langle K\rangle^\ell}\lesssim||u||_{X^\ell}^{2d+5}\frac{L^{2d+2+}}{\langle K\rangle^\ell} \tag{2.139}$$

where $|\mathcal{G}_d|$ is the number of trees in $\mathcal{G}_d$. Thus $\|S_K^{d+2}\|_{X_K^\ell}\lesssim L^{2d+2+}\|u\|_{X^\ell}^{2d+5}$. Replacing $d$ by $d-2$, we complete the proof. □

### 2.4.4 Dynamics of $d_K$

Consider the continuous equation

$$-i\partial_t f(K) = \frac{\pi^2}{3}|f(K)|^4 f(K), \quad f(K)|_{t=0} = f_0(K), \quad K \in \mathbb{R}. \tag{2.140}$$

We can solve this ODE and solution is $f(K) = e(\frac{\pi}{6}|f_0(K)|^4 t)f_0(K)$. In this section, we compare $f(K)$ with $c_K$ (2.43) and obtain a long time dynamics description for $d_K$ in Proposition 1.

**Proposition 1.** *Fix $\ell > 1$, $0 < \gamma < 1$, and let $f_0 \in X^{\ell,2}(\mathbb{R})$. Let $f(t,\xi)$ be the solution of (2.140) over time interval $[0, M]$ with initial data $f_0$. Let $B \stackrel{def}{=} \|f(0)\|_{X^{\ell,2}(\mathbb{R})}$. Let $u$ be the solution of (2.2) with initial data $u_0 = \frac{1}{L}\sum_{\mathbb{Z}_L} f_0(K)e(Kx)$. Then for $L$ sufficiently large and $\epsilon^2 L^\gamma$ sufficiently small, depending on $M$, $B$, there exists a constant $C_\gamma$ such that or all $t \in [0, MT_R]$,*

$$\left\| d_K - f(\frac{t}{T_R}, K) \right\|_{X^\ell(\mathbb{Z}_L)} \lesssim C_\gamma\Big(\epsilon^2 L^\gamma + L^{-1}\Big),$$

*where $T_R = \frac{L^2}{\epsilon^4}$.*

*Proof.* We shall use bootstrap argument to show that $\sup_{0\le t\le T} ||u_K(t)||_{X^\ell} \le 2B$. To do this, it suffices to show that

$$\sup_{0\le t\le T} ||u_K(t)||_{X^\ell} \le 2B \Rightarrow \sup_{0\le t\le T} ||u_K(t)||_{X^\ell} < 2B \tag{2.141}$$

for any $T < MT_R$. To see why (2.141) implies $\sup_{0\le t\le T} ||u_K(t)||_{X^\ell} \le 2B$, notice that initially $||u(0)||_{X^\ell} = ||f(0)||_{X^\ell} \le B$, if at some time $t_0 \le MT_R$, $||u(t_0)||_{X^\ell} \ge 2B$, by continuity there exists $0 < T \le t_0$ s.t. $||u_K(T)||_{X^\ell} = 2B$ and $||u_K(T)||_{X^\ell} = \sup_{0\le t\le T} ||u_K(t)||_{X^\ell}$. But by (2.141), $||u_K(T)||_{X^\ell} < 2B$, which is a contradiction.

Notice that

$$\sup_{0\le t\le T}||u_K(t)||_{X^\ell}\le 2B \Rightarrow \sup_{0\le t\le T}\left\|d_K(t)-f\left(\frac{t}{T_R},K\right)\right\|_{X^\ell}\lesssim C_\gamma\Big(\epsilon^2L^\gamma+L^{-1}\Big). \tag{2.142}$$

implies (2.141). So we only need to prove (2.142).

In what follows, we sometimes denote $f(\frac{t}{T_R},K)$ by $f(K)$ for simplicity.

From Lemma 7 we have

$$\begin{aligned}\left\|d_K(t)-f\left(\frac{t}{T_R},K\right)\right\|_{X^\ell}&\le\left\|c_K(t)-f\left(\frac{t}{T_R},K\right)\right\|_{X^\ell}+\|d_K(t)-c_K(t)\|_{X^\ell}\\&\le\left\|c_K(t)-f\left(\frac{t}{T_R},K\right)\right\|_{X^\ell}+C_\gamma\left(\sum_{d=1}^P\epsilon^{2d}\|u\|_{X^\ell}^{2d+1}\right)L^\gamma\\&\le\left\|c_K(t)-f\left(\frac{t}{T_R},K\right)\right\|_{X^\ell}+C_\gamma\left(\sum_{d=1}^P\epsilon^{2d}B^{2d+1}\right)L^\gamma.\end{aligned}$$

To bound $w_K:=c_K(t)-f(\frac{t}{T_R},K)$, from (2.43), we have

$$\begin{aligned}-i\partial_t w_K&=\underbrace{\frac{\epsilon^4}{L^2}\Big(\frac{1}{L^2}H_K^2(u)-\frac{\pi^2}{3}|f(K)|^4f(K)\Big)}_{I}+\underbrace{\sum_{d=3}^P\frac{\epsilon^{2d}}{L^{2d}}(H_K^d(u)+G_K^d(u))}_{II}\\&\quad+\underbrace{\sum_{d=3}^{P+1}\frac{\epsilon^{2d}}{L^{2d}}S_K^d(u)}_{III}+\underbrace{\frac{\epsilon^{2P+2}}{L^{2P+2}}(\widetilde{H}_K^{P+1}(u)+\widetilde{G}_K^{P+1}(u))}_{IV}\\&=I+II+III+IV,\end{aligned} \tag{2.143}$$

where $I$ is the main term, while $II$, $III$, and $IV$ are errors.

**Estimate of Main Term $I$:** To estimate the main term, we use the method of symmetrization. Recall that, from (2.95), we have

$$\frac{\epsilon^4}{L^4}H_K^2(u) = \frac{3\epsilon^4}{L^4}\sum_{K_1\neq K}\frac{|d_{K_1}|^4}{(K_1-K)^2}d_K \tag{2.144}$$

$$= \frac{3\epsilon^4}{L^4}\sum_{K_1\neq 0}\frac{|d_{K_1+K}|^4+|d_{K-K_1}|^4-2|d_K|^4}{2K_1^2}d_K + \frac{3\epsilon^4}{L^4}\Big(\sum_{K_1\neq 0}\frac{1}{K_1^2}\Big)|d_K|^4 d_K. \tag{2.145}$$

Therefore,

$$\begin{aligned} I &= \frac{3\epsilon^4}{L^2}\left(L^{-2}\Big(\sum_{K_1\neq 0}\frac{1}{K_1^2}\Big)|d_K|^4 d_K - \frac{\pi^2}{3}|f(K)|^4 f(K)\right) \\ &\quad + \frac{3\epsilon^4}{L^4}\sum_{K_1\neq 0}\frac{|d_{K_1+K}|^4+|d_{K-K_1}|^4-2|d_K|^4}{2K_1^2}d_K \\ &=: I_1 + I_2. \end{aligned}$$

Notice that

$$I_1 = \frac{3\epsilon^4}{L^2}\left(\Big(\sum_{\substack{\mu\neq 0\\ \mu\in\mathbb{Z}\\ \mu\leq L}}\frac{1}{\mu^2}\Big)|d_K|^4 d_K - \frac{\pi^2}{3}|f(K)|^4 f(K)\right),$$

Since $\sum_{\mu\neq 0} 1/\mu^2 = \frac{\pi^2}{3}$, we have $\sum_{\substack{\mu\neq 0\\ \mu\in\mathbb{Z}\\ \mu\leq L}}\frac{1}{\mu^2} = O(L^{-1})$. We also have equation $|d_K|^4 d_K - \frac{\pi^2}{3}|f(K)|^4 f(K) = (d_K - f(K))O(|d_K|, |f(K)|)$. So using $|d_K| \leq B$, we get an bound for $I_1$

$$\|I_1\|_{X^\ell} \leq \frac{3\epsilon^4}{L^2}B^4\left\|d_K(t) - f\left(\frac{t}{T_R}, K\right)\right\|_{X^\ell}.$$

For $I_2$, recall that for $g \in X^{\ell,2}$

$$\begin{aligned}
&\sum_{x\in\mathbb{Z}_L\setminus 0} \frac{g(x)+g(-x)-2g(0)}{x^2} \\
=&\sum_{\substack{x\in\mathbb{Z}_L\setminus 0\\ |x|<1}} \frac{g(x)+g(-x)-2g(0)}{x^2} + \sum_{\substack{x\in\mathbb{Z}_L\setminus 0\\ |x|>1}} \frac{g(x)+g(-x)-2g(0)}{x^2} \\
\lesssim& L\left\|g''\right\|_{L^\infty} + \sum_{\substack{x\in\mathbb{Z}_L\setminus 0\\ |x|>1}} \frac{1}{x^2}\left\|g\right\|_{L^\infty} \lesssim L\left\|g\right\|_{X^{0,2}}.
\end{aligned} \tag{2.146}$$

Therefore,

$$\begin{aligned}
I_2 =& \frac{3\epsilon^4}{L^4}\sum_{K_1\neq 0}\Big(\frac{|d_{K_1+K}|^4+|d_{K-K_1}|^4-2|d_K|^4}{2K_1^2}d_K \\
&- \frac{|f_0(K_1+K)|^4+|f_0(K-K_1)|^4-2|f_0(K)|^4}{K_1^2}f(K)\Big) \\
&+ \frac{3\epsilon^4}{L^4}\sum_{K_1\neq 0}\frac{|f_0(K_1+K)|^4+|f_0(K-K_1)|^4-2|f_0(K)|^4}{K_1^2}f(K) \\
\leq& \frac{3\epsilon^4}{L^2}B^4\left\|d_K-f(K)\right\|_{X^\ell} + \frac{3\epsilon^4}{L^3}\left\|f\right\|_{X^{0,2}}^5.
\end{aligned}$$

Here in last step, we apply (2.146).

**Estimate of Errors:**

Bound on $II$

From Lemma 6 and Lemma 11,

$$\left\|II\right\|_{X^\ell} \leq \sum_{d=3}^{P}\frac{\epsilon^{2d}}{L^{2d}}L^{2(d-1)+}\left\|u\right\|_{X^\ell}^{2d+1} = \sum_{d=3}^{P}\frac{\epsilon^{2d}}{L^{2-}}B^{2d+1}.$$

Bound on $III$

From Lemma 14,

$$\left\|III\right\|_{X^\ell} \leq \sum_{d=3}^{P}\frac{\epsilon^{2d}}{L^{2d}}L^{2(d-1)+}\left\|u\right\|_{X^\ell}^{2d+1} = \sum_{d=3}^{P}\frac{\epsilon^{2d}}{L^{2-}}B^{2d+1}.$$

Bound on $IV$

From Lemma 8 and Lemma 13,

$$\|IV\|_{X^\ell} \leq \frac{\epsilon^{2P+2}}{L^{2P+2}} L^{2P+2+} B^{2P+3} = \epsilon^{2P+2} L^+ B^{2P+3}.$$

Integrating (2.143), we have

$$\begin{aligned}
&\left\| d_K(t) - f\left(\frac{t}{T_R}, K\right)\right\|_{X^\ell} - C_{\gamma,B}\epsilon^2 L^\gamma \\
\leq &\int_0^t \left( \frac{3\epsilon^4}{L^2} B^4 \left\| d_K(t) - f\left(\frac{t}{T_R}, K\right)\right\|_{X^\ell} + \frac{3\epsilon^4}{L^2} B^4 \left\| d_K - f(K)\right\|_{X^\ell} + C_{\gamma,B}\frac{3\epsilon^4}{L^3} \right. \\
&\left. + C_{\gamma,B}\frac{\epsilon^6 L^+}{L^2} + C_{\gamma,B}\frac{\epsilon^6 L^+}{L^2} + C_{\gamma,B}\epsilon^{2P+2} L^+ \right) ds
\end{aligned}$$

From Gronwall's inequality, and $0 \leq t \leq T_R M$, we obtain,

$$\begin{aligned}
&\left\| d_K(t) - f\left(\frac{t}{T_R}, K\right)\right\|_{X^\ell} \\
\leq &C_{\gamma,B}\left(\epsilon^2 L^\gamma + \frac{M}{L} + M\epsilon^2 L^\gamma + M\epsilon^{2P+2} L^{2+\gamma}\right) e^{CB^4 M} \qquad (2.147) \\
\lesssim &C_\gamma\left(\epsilon^2 L^\gamma + L^{-1}\right)
\end{aligned}$$

Therefore, $\|d_K(t)\|_{X^\ell} \leq 2B$ and this completes the proof.

□

### 2.4.5 A refined estimate of $|d_K|^2$

In previous subsection, we have described the dynamics of $d_K$, when $t < T_R$. We know that $d_K$ is different from $u_K$ by a phase factor that depends on the module of the Fourier coefficients $|d_K|$. From Proposition 1, we have $\||d_K| - |f(K)|\|_{X^\ell} \lesssim \epsilon^2 L^+ + L^{-1}$. Given this

estimate we know that

$$\frac{\epsilon^2}{2\pi L^2}\left|\int_0^t |d_K|^2(s) - |f(K)|^2(s)ds\right| \lesssim \frac{\epsilon^2}{2\pi L^2} t(\epsilon^2 L^+ + L^{-1}).$$

When $t \sim \frac{L^2}{\epsilon^4}$, this roughly gives a bound $L^+ + L^{-1}\epsilon^{-2}$. This bound do not allow us to replace $e\left(\frac{\epsilon^2}{2\pi L^2}\int_0^t |d_K|^2(s)ds\right)$ by $e\left(\frac{\epsilon^2}{2\pi L^2}\int_0^t |f(K)|^2(s)ds\right)$. Only a refined estimate like $\||d_K|^2 - (known)\|_{X^\ell} \leq \epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L}$ allows us to estimate $u_K$ in terms of $d_K$, which still requires some nontrivial work.

In this subsection, we prove the following proposition.

**Proposition 2.** *We have the following estimate for $|c_K|^2$ and $|d_K|^2$*

$$\left\||c_K|^2(t) - P_K(f)(t)\right\|_{X^\ell} \lesssim \epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L}. \tag{2.148}$$

*with*

$$\begin{aligned} P_K(f) =& -\frac{2\epsilon^6}{L^6}\int_0^t \Im\Big( \left(H_K^3(f) + G_K^3(f) + S_K^3(f)\right)\overline{f(K)} + H_K^2(f)F_K^1(\overline{f}) \\ &+ H_K^2(f)F_K^1(\overline{f}) + H_K^2(f)E_K^1(\overline{f})\Big)(s)ds \end{aligned} \tag{2.149}$$

*and*

$$\left\||d_K|^2(t) - Q_K(f)(t)\right\|_{X^\ell} \lesssim \epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L}. \tag{2.150}$$

*with*

$$\begin{aligned} Q_K(f) =& \frac{2\epsilon^2}{L^2}\Re\left(F_K^d(f)\overline{f(K)}\right) - \frac{2\epsilon^6}{L^6}\int_0^t \Im\Big( \left(H_K^3(f) + G_K^3(f) + S_K^3(f)\right)\overline{f(K)} \\ &+ H_K^2(f)F_K^1(\overline{f}) + H_K^2(f)F_K^1(\overline{f}) + H_K^2(f)E_K^1(\overline{f})\Big)(s)ds \end{aligned} \tag{2.151}$$

*Proof.* Our final goal is to to estimate $|d_K(t)|^2$ to obtain the dynamics of $u_K$. Our strategy here is to first estimate $|c_K(t)|^2$ using the equation for $c_K$. Then estimate the difference

between $|c_K(t)|^2$ and $|d_K(t)|^2$. For $c_K(t)$, recall we have

$$-i\partial_t c_K = \frac{\epsilon^4}{L^4}H_K^2(u)+\sum_{d=3}^{P}\frac{\epsilon^{2d}}{L^{2d}}(H_K^d(u)+G_K^d(u))+\sum_{d=3}^{P+1}\frac{\epsilon^{2d}}{L^{2d}}S_K^d(u)+\frac{\epsilon^{2P+2}}{L^{2P+2}}(\widetilde{H}_K^{P+1}(u)+\widetilde{G}_K^{P+1}(u)).$$

Therefore,

$$\begin{aligned}\partial_t|c_K|^2 &= 2\Re(\partial_t c_K\overline{c}_K)\\ &= -2\Im\Big(\frac{\epsilon^4}{L^4}H_K^2(u)\overline{c}_K+\sum_{d=3}^{P}\frac{\epsilon^{2d}}{L^{2d}}(H_K^d(u)+G_K^d(u))\overline{c}_K\\ &+\sum_{d=3}^{P+1}\frac{\epsilon^{2d}}{L^{2d}}S_K^d(u)\overline{c}_K+\frac{\epsilon^{2P+2}}{L^{2P+2}}(\widetilde{H}_K^{P+1}(u)+\widetilde{G}_K^{P+1}(u))\overline{c}_K\Big).\end{aligned}$$

Since

$$\overline{c}_K = \overline{d}_K-\sum_{d=1}^{P}\frac{\epsilon^{2d}}{L^{2d}}F_K^d(\overline{u})-\sum_{d=1}^{P}\frac{\epsilon^{2d}}{L^{2d}}E_K^d(\bar{u}),$$

we have

$$\begin{aligned}\partial_t|c_K|^2 &= -2\Im\bigg(\frac{\epsilon^4}{L^4}H_K^2(u)\Big(\overline{d}_K-\sum_{d=1}^{P}\frac{\epsilon^{2d}}{L^{2d}}F_K^d(\overline{u})-\sum_{d=1}^{P}\frac{\epsilon^{2d}}{L^{2d}}E_K^d(\bar{u})\Big)\\ &+\sum_{d=3}^{P}\frac{\epsilon^{2d}}{L^{2d}}(H_K^d(u)+G_K^d(u))\Big(\overline{d}_K-\sum_{d=1}^{P}\frac{\epsilon^{2d}}{L^{2d}}F_K^d(\overline{u})-\sum_{d=1}^{P}\frac{\epsilon^{2d}}{L^{2d}}E_K^d(\bar{u})\Big)\\ &+\sum_{d=3}^{P+1}\frac{\epsilon^{2d}}{L^{2d}}S_K^d(u)\Big(\overline{d}_K-\sum_{d=1}^{P}\frac{\epsilon^{2d}}{L^{2d}}F_K^d(\overline{u})-\sum_{d=1}^{P}\frac{\epsilon^{2d}}{L^{2d}}E_K^d(\bar{u})\Big)\\ &+\frac{\epsilon^{2P+2}}{L^{2P+2}}(\widetilde{H}_K^{P+1}(u)+\widetilde{G}_K^{P+1}(u))\Big(\overline{d}_K-\sum_{d=1}^{P}\frac{\epsilon^{2d}}{L^{2d}}F_K^d(\overline{u})-\sum_{d=1}^{P}\frac{\epsilon^{2d}}{L^{2d}}E_K^d(\bar{u})\Big)\bigg)\\ &= -2\Im\left(\frac{\epsilon^4}{L^4}H_K^2(u)\overline{d}_K\right)+2\Im\left(\frac{\epsilon^6}{L^6}\big(H_K^3(u)+G_K^3(u)+S_K^3(u)\big)\overline{d}_K\right)\\ &+2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(u)F_K^1(\overline{u})\right)+2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(u)E_K^1(\overline{u})\right)+\text{Higher Order Terms}\\ &=: \mathcal{N}_1+\mathcal{N}_2+\mathcal{N}_3+\mathcal{N}_4+\mathcal{N}_5.\end{aligned}\tag{2.152}$$

Here we collect all terms with the factor of order lower than $\frac{\epsilon^6}{L^6}$ on the right hand side and denote them by $\mathcal{N}_1$, $\mathcal{N}_2$, $\mathcal{N}_3$, $\mathcal{N}_4$. The collection of all other terms is called higher order terms and is denoted by $\mathcal{N}_5$. Note that our final goal is $\||c_K|^2 - (known)\|_{X^\ell} \lesssim \epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L}$ for $t < \frac{L^2}{\epsilon^4}$. All terms of order greater than $\frac{\epsilon^8}{L^8}$ still have $\epsilon^4$ in front after integration over $t$ and thus can be treated as error. So we collect all terms of order lower than $\frac{\epsilon^6}{L^6}$ for which further estimate should be done.

Notice that $\mathcal{N}_1 = 0$. This is because by (2.95), we know that

$$\frac{\epsilon^4}{L^4} H_K^2(u) = \frac{3\epsilon^4}{L^4} \sum_{K_1 \neq K} \frac{|d_{K_1}|^4}{(K_1 - K)^2} d_K. \tag{2.153}$$

Thus

$$-2\Im \frac{\epsilon^4}{L^4} H_K^2(u) \overline{d}_K = -2\Im \frac{3\epsilon^4}{L^4} \sum_{K_1 \neq K} \frac{|d_{K_1}|^4}{(K_1 - K)^2} |d_K|^2 = 0$$

From Proposition 1, we have

$$\sup_{0 \leq t \leq MT_R} \left| |d_K(t)| - |f(\frac{t}{T_R}, K)| \right| \lesssim \epsilon^2 L^\gamma + \frac{1}{L}.$$

To treat all other $\mathcal{N}_j$, $j = 2, 3, 4.$, we first introduce the following lemma.

**Lemma 15.** *For any multilinear form $F(x_1, \cdots, x_m) : X^\ell \times \cdots \times X^\ell \to X^\ell$ with norm $\|F\|$, i.e.*

$$\|F(x_1, \cdots, x_m)\|_{X^\ell} \leq \|F\| \ \|x_1\|_{X^\ell} \cdots \|x_m\|_{X^\ell}.$$

*If $\left\|d_{K_j}\right\|_{X^\ell} \leq B$ and $\|f(K)\|_{X^\ell} \leq B$, we have*

$$\begin{aligned} &\|F(d_{K_1}, \cdots, d_{K_m}) - F(f(K_1), \cdots, f(K_m))\|_{X^\ell} \\ &\leq CB^{m-1} \|F\| \|d_K - f(K)\|_{X^\ell} \leq CB^{m-1} \|F\| (\epsilon^2 L^+ + \frac{1}{L}). \end{aligned} \tag{2.154}$$

**Remark 6.** *By the polar identity,*

$$\begin{aligned}&\|F(x_1,\cdots,x_m)\|_{X^\ell}\le\|F\|\ \|x_1\|_{X^\ell}\cdots\|x_m\|_{X^\ell},\ \forall x_1,\cdots,x_m.\\&\Leftrightarrow\|F(x,\cdots,x)\|_{X^\ell}\le\|F\|\ \|x\|_{X^\ell}\cdots\|x\|_{X^\ell},\ \forall x.\end{aligned}\tag{2.155}$$

*Thus the results in Lemma 6, 7, 8, 11, 12, 13, 14 can be viewed as bounds on the norm of the multilinear forms $H^d_K(u)$, $F^d_K(u)$, $\widetilde{H}^d_K(u)$, $G^d_K(u)$ and $\widetilde{G}^d_K(u)$ and $S^d_K(u)$.*

*Proof.*

$$\begin{aligned}&\|F(d_{K_1},\cdots,d_{K_m})-F(f(K_1),\cdots,f(K_m))\|_{X^\ell}\\=&\left\|\int_0^1\frac{d}{dt}F(td_{K_1}+(1-t)f(K_1),\cdots,td_{K_m}+(1-t)f(K_m))dt\right\|_{X^\ell}\\=&\int_0^1\left\|\frac{d}{dt}F(td_{K_1}+(1-t)f(K_1),\cdots,td_{K_m}+(1-t)f(K_m))\right\|_{X^\ell}dt\\=&\int_0^1\sum_{j=1}^m\left\|\partial_{x_j}F(td_{K_1}+(1-t)f(K_1),\cdots,td_{K_m}+(1-t)f(K_m))(d_{K_j}-f(K_j)\right\|_{X^\ell}dt\\\lesssim&C(\|d_K\|^{m-1}_{X^\ell}+\|f(K)\|^{m-1}_{X^\ell})\,\|F\|\,\|d_K-f(K)\|_{X^\ell}\\\lesssim&CB^{m-1}\,\|F\|\,\|d_K-f(K)\|_{X^\ell}\end{aligned}$$

□

Basically this lemma tells us that we may replace $d_K$ by $f(K)$ up to errors. For sake of simplicity, we introduce $H^d_K(f)$, $G^d_K(f)$ and $S^d_K(f)$, which is obtained simply by replacing all $d_{K_j}$ in $H^d_K(u)$, $G^d_K(u)$ and $S^d_K(u)$ by $f(K)$.

Bound on $\mathcal{N}_2$: By Lemma 6, 11 and 14, we know that the norms of $H^3_K(u)$, $G^3_K(u)$ and $S^3_K(u)$ can be bounded by $L^{4+}$. Thus replace $d_K$ by $f(K)$ using Lemma 15 we get

$$\begin{aligned}\left\|\mathcal{N}_2-2\left(\Im\frac{\epsilon^6}{L^6}\big(H^3_K(f)+G^3_K(f)+S^3_K(f)\big)\right)\overline{f(K)}\right\|_{X^\ell}&\lesssim\frac{\epsilon^6}{L^6}L^{4+}(\epsilon^2L^++\frac{1}{L})\\&\lesssim\frac{\epsilon^4}{L^2}(\epsilon^4L^++\frac{\epsilon^2L^+}{L}).\end{aligned}$$

Bound on $\mathcal{N}_3$: By (2.95), Lemma 7, we know that the norms of $H_K^2(u)$, $F_K^1(u)$ can be bounded by $L^{2+}$. Thus replace $d_K$ by $f(K)$ using Lemma 15 we get

$$\begin{aligned}
&\left\|\mathcal{N}_4 - 2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(f)F_K^1(\overline{f})\right)\right\|_{X^\ell}\\
=&\left\|2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(u)F_K^1(\overline{u})\right) - 2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(f)F_K^1(\overline{f})\right)\right\|_{X^\ell}\\
=&\left\|2\Im\left(\frac{\epsilon^6}{L^6}(H_K^2(u)-H_K^2(f))F_K^1(\overline{u})\right) - 2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(f)(F_K^1(\overline{f}-F_K^1(\overline{u}))\right)\right\|_{X^\ell}\\
\lesssim&\frac{\epsilon^6}{L^6}L^{4+}(\epsilon^2L^+ + \frac{1}{L})\\
\lesssim&\frac{\epsilon^4}{L^2}(\epsilon^4L^+ + \frac{\epsilon^2L^+}{L}).
\end{aligned}$$

Bound on $\mathcal{N}_4$: By (2.95), Lemma 12, we know that the norms of $H_K^2(u)$, $E_K^1(u)$ can be bounded by $L^{2+}$. Thus replace $d_K$ by $f(K)$ using Lemma 15 we get

$$\begin{aligned}
&\left\|\mathcal{N}_3 - 2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(f)E_K^1(\overline{f})\right)\right\|_{X^\ell}\\
=&\left\|2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(u)E_K^1(\overline{u})\right) - 2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(f)E_K^1(\overline{f})\right)\right\|_{X^\ell}\\
=&\left\|2\Im\left(\frac{\epsilon^6}{L^6}(H_K^2(u)-H_K^2(f))E_K^1(\overline{u})\right) - 2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(f)(E_K^1(\overline{f}-E_K^1(\overline{u}))\right)\right\|_{X^\ell}\\
\lesssim&\frac{\epsilon^6}{L^6}L^{4+}(\epsilon^2L^+ + \frac{1}{L})\\
\lesssim&\frac{\epsilon^4}{L^2}(\epsilon^4L^+ + \frac{\epsilon^2L^+}{L}).
\end{aligned}$$

Bound on $\mathcal{N}_5$: From Lemma 6, Lemma 7, Lemma 8, Lemma 11, Lemma 12, Lemma 13, Lemma 14, it's not very difficult to prove that the $X^\ell$ norm of the sum of all other terms should be of the order $\frac{\epsilon^4L^+}{L^2}(\epsilon^4L^+)$.

From above calculations we know that

$$\begin{aligned}
\partial_t|c_K|^2 =& -2\left(\Im\frac{\epsilon^6}{L^6}\big(H_K^3(f)+G_K^3(f)+S_K^3(f)\big)\overline{f(K)}\right) - 2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(f)F_K^1(\overline{f})\right)\\
&-2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(f)F_K^1(\overline{f})\right) - 2\Im\left(\frac{\epsilon^6}{L^6}H_K^2(f)E_K^1(\overline{f})\right) + O(\frac{\epsilon^4}{L^2}(\epsilon^4L^+ + \frac{\epsilon^2L^+}{L}))
\end{aligned} \tag{2.156}$$

Integrate in $t$, and note that because $t \lesssim \frac{L^2}{\epsilon^4}$, $\int_0^t O(\frac{\epsilon^4}{L^2}(\epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L}))ds \leq O(\epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L})$. We have

$$\begin{aligned}|c_K|^2(t) = &-\frac{2\epsilon^6}{L^6}\int_0^t \Im\Big( \big(H_K^3(f) + G_K^3(f) + S_K^3(f)\big)\overline{f(K)} + H_K^2(f)F_K^1(\overline{f}) \\ &+H_K^2(f)F_K^1(\overline{f}) + H_K^2(f)E_K^1(\overline{f})\Big)(s)ds + O(\epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L})\end{aligned} \tag{2.157}$$

Define

$$\begin{aligned}P_K(f) = &-\frac{2\epsilon^6}{L^6}\int_0^t \Im\Big( \big(H_K^3(f) + G_K^3(f) + S_K^3(f)\big)\overline{f(K)} + H_K^2(f)F_K^1(\overline{f}) \\ &+H_K^2(f)F_K^1(\overline{f}) + H_K^2(f)E_K^1(\overline{f})\Big)(s)ds\end{aligned} \tag{2.158}$$

Then we know that

$$\big\||c_K|^2(t) - P_K(f)(t)\big\|_{X^\ell} \lesssim \epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L}. \tag{2.159}$$

We complete the prove of (2.148).

Now let us started to prove (2.150). Let us consider the difference between $|c_K|^2$ and $|d_K|^2$. By (2.28), we have

$$c_K = d_K - \sum_{d=1}^P \frac{\epsilon^{2d}}{L^{2d}}F_K^d(u) - \sum_{d=1}^P \frac{\epsilon^{2d}}{L^{2d}}E_K^d(u), \tag{2.160}$$

So

$$\begin{aligned}&|c_K|^2 - |d_K|^2 \\ =&-2\sum_{d=1}^P \frac{\epsilon^{2d}}{L^{2d}}\Re\left(F_K^d(u) + E_K^d(u)\right)\overline{d_K} + \left|\sum_{d=1}^P \frac{\epsilon^{2d}}{L^{2d}}(F_K^d(u) + F_K^d(u))\right|^2 \\ =&-\frac{2\epsilon^2}{L^2}\Re\left(F_K^d(u) + E_K^d(u)\right)\overline{d_K} + \text{Higher Order Term} \\ =&-\frac{2\epsilon^2}{L^2}\Re\left(F_K^d(u)\overline{d_K}\right) + \text{Higher Order Term}\end{aligned} \tag{2.161}$$

Here the last step follows from the fact that $G^1_{K_1K_2K_3} = 0$. Here we collect all terms with the factor of order lower than $\frac{\epsilon^4}{L^4}$ on the right hand side. The collection of all other terms is called higher order terms. Note that our final goal is $\||d_K|^2 - (known)\|_{X^\ell} \leq \epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L}$ for $t < \frac{L^2}{\epsilon^4}$. All terms of order greater than $\frac{\epsilon^4}{L^4}$ have $\epsilon^4$ in front and thus can be treated as error. So we collect all terms of order lower than $\frac{\epsilon^4}{L^4}$ for which further estimate should be done.

The first term can be treated in the same way as that of $\mathcal{N}_2$. We first notice that the by Lemma 7, we know that the norms of $F^1_K(u)$ can be bounded by $L^{2+}$. Thus replace $d_K$ by $f(K)$ using Lemma 15 we get

$$\begin{aligned}
&\left\| \text{first term} + \frac{2\epsilon^2}{L^2} \Re \left( F^d_K(f) \overline{f(K)} \right) \right\|_{X^\ell} \\
=& \frac{2\epsilon^2}{L^2} \left\| \Re \left( F^d_K(u) \overline{d_K} \right) - \Re \left( F^d_K(f) \overline{f(K)} \right) \right\|_{X^\ell} \\
\lesssim& \frac{\epsilon^2}{L^2} L^{2+} (\epsilon^2 L^+ + \frac{1}{L}) \\
\lesssim& \epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L}.
\end{aligned}$$

From Lemma 6, 7, 8, 11, 12, 13, 14, it's easy to verify that the $X^\ell$ norm of the sum of all higher order terms should be bounded by $\epsilon^4 L^+$.

Thus

$$\begin{aligned}
&\left\| |d_K|^2 - |c_K|^2 - \frac{2\epsilon^2}{L^2} \Re \left( F^d_K(f) \overline{f(K)} \right) \right\|_{X^\ell} \\
\lesssim& \epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L}.
\end{aligned}$$

Now apply the first conclusion and this completes the proof of the second conclusion.

□

### 2.4.6 Wrapping up

In this section, we shall prove Theorem 1 which states that

$$\left\|u_K(t) - f_0(K)e\left(\left(K^2 + \frac{\epsilon^2}{\pi L}\|u\|_{L^2}^2 + \frac{\pi\epsilon^2}{6L^2}|f_0(K)|^4\right)t - \frac{\epsilon^2}{2\pi L^2}\int_0^t Q_K(f)(s)ds\right)\right\|_{X^\ell(\mathbb{Z}_L^n)}$$
$$\lesssim C_\gamma(\epsilon^2 L^\gamma + \frac{1}{L^{1-\gamma}}). \tag{2.162}$$

From Proposition 1, we know that in $[0, MT_R]$,

$$\left\|u_K(t) - f_0(K)e\left(\left(K^2 + \frac{\epsilon^2}{\pi L}\|u\|_{L^2}^2 + \frac{\pi\epsilon^2}{6L^2}|f_0(K)|^4\right)t - \frac{\epsilon^2}{2\pi L^2}\int_0^t |d_K(s)|^2 ds\right)\right\|_{X^\ell(\mathbb{Z}_L^n)}$$
$$\lesssim C_\gamma(\epsilon^2 L^\gamma + \frac{1}{L^{1-\gamma}}) \tag{2.163}$$

and

$$\|u_K(t)\|_{X^\ell(\mathbb{Z}_L^n)} \leq B. \tag{2.164}$$

Thus by Proposition 2, we know that

$$\begin{aligned}
&\left\|u_K(t) - f_0(K)e\left(\left(K^2 + \frac{\epsilon^2}{\pi L}\|u\|_{L^2}^2 + \frac{\pi\epsilon^2}{6L^2}|f_0(K)|^4\right)t - \frac{\epsilon^2}{2\pi L^2}\int_0^t Q_K(f)(s)ds\right)\right\|_{X^\ell(\mathbb{Z}_L^n)}\\
\leq&\left\|f_0(K)e\left((\cdots)t - \frac{\epsilon^2}{2\pi L^2}\int_0^t |d_K(s)|^2 ds\right) - f_0(K)e\left((\cdots)t - \frac{\epsilon^2}{2\pi L^2}\int_0^t Q_K(f)(s)ds\right)\right\|_{X^\ell(\mathbb{Z}_L^n)}\\
+&\left\|u_K(t) - f_0(K)e\left((\cdots)t - \frac{\epsilon^2}{2\pi L^2}\int_0^t |d_K(s)|^2 ds\right)\right\|_{X^\ell(\mathbb{Z}_L^n)}\\
\lesssim&\left\|f_0(K)\left(e\left(\frac{\epsilon^2}{2\pi L^2}\int_0^t (Q_K(f)(s) - |d_K(s)|^2)ds\right) - 1\right)\right\|_{X^\ell(\mathbb{Z}_L^n)} + \epsilon^2 L^+ + L^{-1+}\\
\lesssim&\frac{\epsilon^2}{L^2}\int_0^t \left\|Q_K(f)(s) - |d_K(s)|^2\right\|_{X^\ell(\mathbb{Z}_L^n)} ds + \epsilon^2 L^+ + L^{-1+}\\
\lesssim&\frac{\epsilon^2}{L^2}T_R(\epsilon^4 L^+ + \frac{\epsilon^2 L^+}{L}) + \epsilon^2 L^+ + L^{-1+}\\
=&\epsilon^2 L^+ + L^{-1+}.
\end{aligned}$$

The second step follows from (2.163) and the third step follows from $|e(x) - 1| \leq |x|$. This completes the proof of Theorem 1.

# Chapter 3

# Continuous-time stochastic gradient descent: the stochastic gradient process

## 3.1 Introduction

The training of a machine learning model is often represented through an optimization problem, where the goal is to calibrate the model's parameters to optimize its goodness-of-fit with respect to training data. The goodness-of-fit is usually quantified through a loss function that sums up the losses from the misrepresentation of every single training data set, see, e.g., [48] for or, e.g., [52, 25] for similar optimization problems in imaging and statistics. In 'big data' settings, the sum of these loss functions consists of thousands or millions of terms, making classical optimization methods computationally infeasible. Thus, efficiently solving optimization problems of this form has been a focus of machine learning and optimization research in the past decades. Here, methods often build upon the popular stochastic gradient descent method.

Originally, stochastic gradient descent was proposed by [107] to optimize not only sums of loss functions, but also expectations of randomized functions.[1] Of course, a normalized sum is just a special case of an expected value, making stochastic gradient descent available for the kind of training problem described above. Based on stochastic gradient descent ideas, improved algorithms have been proposed for optimizing sums of loss functions, such as [27, 36, 40]. Unfortunately, these methods often specifically target sums of loss functions and can then be infeasible to optimize general expected values of loss functions.

The optimization of expected values of loss functions appears in the presence of countably infinite and continuous data in functional data analysis and non-parametric statistics (e.g., [119]), physics-informed deep learning (e.g., [105]), inverse problems (e.g., [19]), and continuous data augmentation/adversarial robustness (e.g., [32, 117, 100]). Some of these problems are usually studied after discretising the data. As hinted above, algorithms for discrete data sometimes deteriorate when approaching the continuum limit, i.e. as the number of data sets goes to infinity. Thus, we prefer studying the continuum case immediately. Finally, we note that 'continuous data' can also refer to optimization under general noise models. Here, expected values are minimized in robust optimization (e.g., [91]), variational Bayesian inference (e.g., [29]), and optimal control (e.g., [88]). Overall, the optimization of general expected values is a very important task in modern data science, machine learning, and related fields.

In this chapter, we study stochastic gradient descent for general expected values in a continuous-time framework. We now proceed with the formal introduction of the optimization problem, the stochastic gradient descent algorithm, its continuous-time limits, and current research in this area.

[1] Actually, the 'stochastic approximation method' of [107] aims at finding roots of functions that are given as expectations of randomized functions. The method they construct resembles stochastic gradient descent for a least squares loss function.

### 3.1.1 Problem setting and state of the art

We study optimization problems of the form

$$\min_{\theta \in X} \Phi(\theta) := \int_S f(\theta, y)\pi(\mathrm{d}y), \tag{3.1}$$

where $X := \mathbb{R}^K$, $S$ is a Polish space, $f : S \times X \to \mathbb{R}$ is a measurable function that is continuously differentiable in the first variable, and $\pi$ is a probability measure on $S$. Moreover, we assume that the integral above always exists. We refer to $X$ as *parameter space*, $S$ as *index set*, $\Phi$ as *full target function*, and $f$ as *subsampled target function*. In these optimization problems, it is usually impossible or intractable to evaluate the integral $\Phi$ or its gradient $\nabla\Phi$.Hence, traditional optimization algorithms, such as steepest gradient descent or Newton methods are not applicable.

As mentioned above, it is possible to employ stochastic optimization methods, such as the *stochastic gradient descent (SGD)* method, see [69, 107]. The stochastic gradient descent method for (3.1) proceeds through the following discrete-time dynamic that iterates over $n \in \mathbb{N} := \{1, 2, \ldots\}$:

$$\theta_n = \theta_{n-1} - \eta_n \nabla_\theta f(\theta_{n-1}, y_n), \tag{3.2}$$

where $y_1, y_2, \ldots \sim \pi$ independent and identically distributed (i.i.d.), $(\eta_n)_{n=1}^\infty \in (0, \infty)^{\mathbb{N}}$ is a non-increasing sequence of *learning rates*, and $\theta_0 \in X$ is an appropriate initial value. Hence, SGD is an iterative method that employs only the gradient of the integrand $f$, but not $\Phi$. SGD converges to the minimizer of $\Phi$, if $\eta_n \to 0$, as $n \to \infty$, sufficiently slowly, and $f(\cdot, y)$ is strongly convex and $\nabla f(\cdot, y)$ is bounded ($y \in S$); see, e.g., [21]. SGD is used in practice also for non-convex optimization problems and with constant learning rate. The constant learning rate setting is popular especially due to its regularizing properties; see [1, 122].

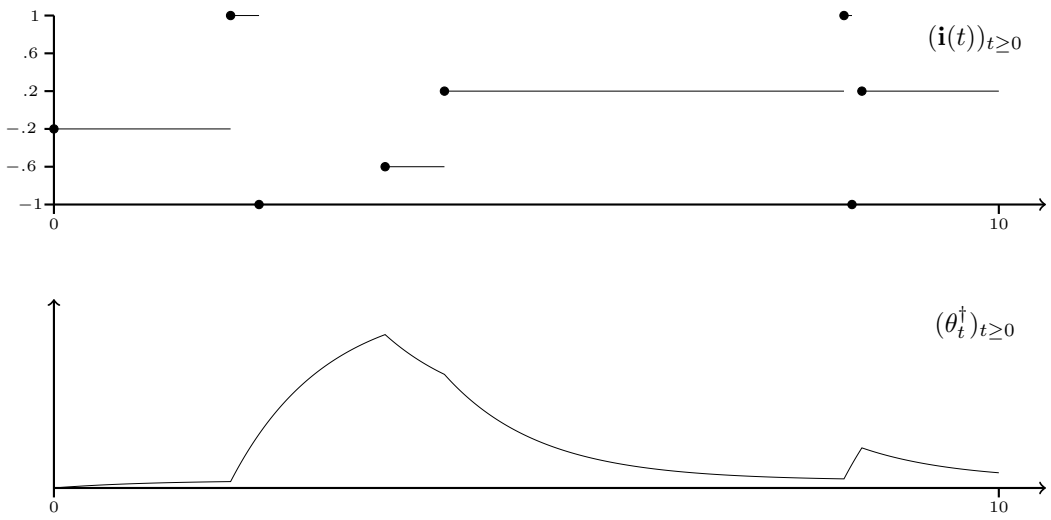


Figure 3.1: Cartoon of the stochastic gradient process $(\theta_t^\dagger)_{t\geq 0}$ with index process $(\mathbf{i}(t))_{t\geq 0}$ on the discrete index set $S := \{-1, -0.6, \ldots, 1\}$. The index process is a Markov pure jump process on $S$. The process $(\theta_t^\dagger)_{t\geq 0}$ aims at optimizing the $\mathrm{Unif}(S)$-integral of the subsampled target functional is $f(\theta, y) := \frac{1}{2}(\theta - y^2)^2$ $(\theta \in X := \mathbb{R}, y \in S)$.

To understand, improve, and study discrete-time dynamical systems, it is sometimes advantageous to represent them in continuous time, see, e.g. the works by [35, 65, 127]. Continuous-time models allow us to concentrate on the underlying dynamics and omit certain numerical considerations. Moreover, they give us natural ways to construct new, efficient algorithms.

The discrete-time dynamic in (3.2) is sometimes represented through a continuous-time diffusion process, see [1, 77, 76, 86, 87, 131]:

$$\mathrm{d}\theta_t = -\nabla\Phi(\theta_t)\mathrm{d}t + \sqrt{\eta(t)}\Sigma(\theta_t)^{1/2}\mathrm{d}W_t,$$

where $\Sigma(\theta) = \int(\nabla_\theta f(\theta; y) - \nabla_\theta\Phi(\theta)) \otimes (\nabla_\theta f(\theta; y) - \nabla_\theta\Phi(\theta))\pi(\mathrm{d}y)$, $(W_t)_{t\geq 0}$ is a $K$-dimensional Brownian motion, and $(\eta(t))_{t\geq 0}$ is an interpolation of the learning rate sequence. While this diffusion approach is suitable to describe the dynamic of the moments of SGD, it does not immediately allow us to construct new stochastic optimization algorithms, as the system depends on the inaccessible $\nabla\Phi$.

A continuous-time representation of stochastic gradient descent that does not depend on $\nabla\Phi$ has recently been proposed by [70]. This work only considers the discrete data case, i.e., $S$ is finite and $\pi := \mathrm{Unif}(S)$. SGD is represented by the *stochastic gradient process* $(\theta_t^\dagger)_{t\geq 0}$. It is defined through the coupled dynamical system

$$\mathrm{d}\theta_t^\dagger = -\nabla_\theta f(\theta_t^\dagger; \mathbf{i}(t))\mathrm{d}t, \tag{3.3}$$

where $(\mathbf{i}(t))_{t\geq 0}$ is a suitable continuous-time Markov process on $S$, which we call *index process*. Hence, the process $(\theta_t^\dagger)_{t\geq 0}$ represents gradient flows with respect to the subsampled target functions that are switched after random waiting times. The random waiting times are controlled by the continuous-time Markov process $(\mathbf{i}(t))_{t\geq 0}$. We show an example of the coupling of exemplary process $(\mathbf{i}(t))_{t\geq 0}$ and $(\theta_t^\dagger)_{t\geq 0}$ in Figure 3.1. The setting is $S := \{-1, -0.6, \ldots, 1\}$, $\pi := \mathrm{Unif}(S)$, $X := \mathbb{R}$, and $f(\theta, y) := \frac{1}{2}(\theta - y^2)^2$ $(\theta \in X, y \in S)$. There, we see that the sample path of $(\theta_t^\dagger)_{t\geq 0}$ is piecewise smooth, with non-smooth behavior at the jump times of $(\mathbf{i}(t))_{t\geq 0}$.

If the process $(\mathbf{i}(t))_{t\geq 0}$ is homogeneous-in-time, the dynamical system represents a constant learning rate. Inhomogeneous $(\mathbf{i}(t))_{t\geq 0}$ with decreasing mean waiting times, on the other hand, model a decreasing learning rate. Under certain assumptions, the process $(\theta_t^\dagger)_{t\geq 0}$ converges to a unique stationary measure when the learning rate is constant or to the minimizer of $\Phi$ when the learning rate decreases.

### 3.1.2 The stochastic gradient process

We now briefly introduce the continuous-time stochastic gradient descent methods that we study throughout this chapter. Then, we summarize our main contributions and give a short outline.

In chapter, we aim to generalize the dynamical system (3.3) to include more general spaces $S$ and probability measures $\pi$ – studying the more general optimization problems of

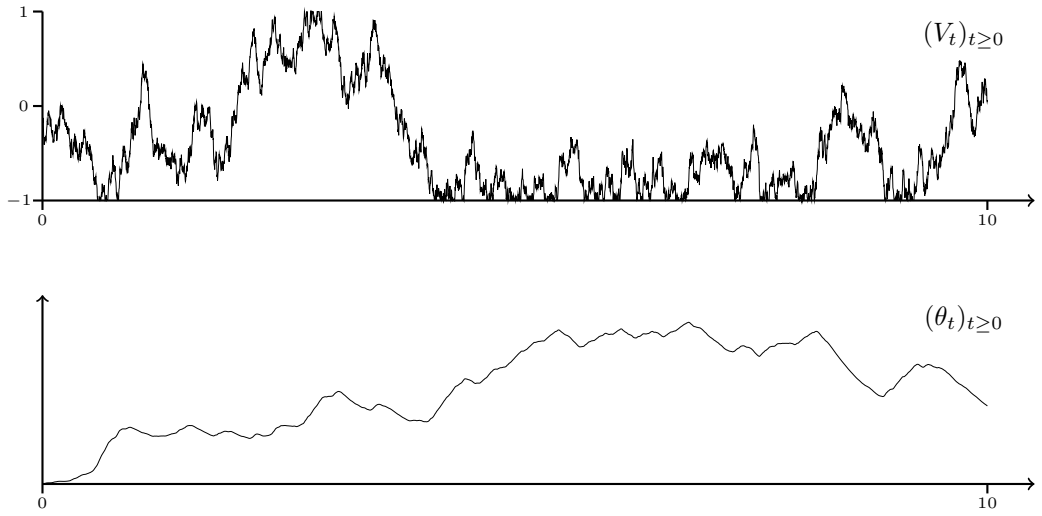


Figure 3.2: Cartoon of the stochastic gradient process $(\theta_t)_{t\geq 0}$ with index process $(V_t)_{t\geq 0}$, with continuous index set $S := [-1, 1]$. The index process is a reflected Brownian motion on $S$.

type (3.1). We proceed as follows: We define a stationary continuous-time Markov process $(V_t)_{t\geq 0}$ on $S$ that is geometrically ergodic and has $\pi$ as its stationary measure. This process $(V_t)_{t\geq 0}$ is now our *index process*. Similarly to (3.3), we then couple $(V_t)_{t\geq 0}$ with the following gradient flow:

$$\mathrm{d}\theta_t = -\nabla_\theta f(\theta_t, V_t)\mathrm{d}t. \tag{3.4}$$

Note that the index process $(V_t)_{t\geq 0}$ can be considerably more general than the Markov jump processes studied by [70]; we discuss examples below and in Section 3.2. As the dynamical system (3.4) contains the discrete version (3.3) as a special case, we refer to $(\theta_t)_{t\geq 0}$ also as *stochastic gradient process*.

We give an example for $(\theta_t, V_t)_{t\geq 0}$ in Figure 3.2. There, we consider $S := [-1, 1]$, $\pi := \mathrm{Unif}[-1, 1]$, $X := \mathbb{R}$, and $f(\theta, y) := \frac{1}{2}(\theta - y^2)^2$ $(\theta \in X, y \in S)$. A suitable choice for $(V_t)_{t\geq 0}$ is a reflected Brownian motion on $[-1, 1]$. Although it is coupled with $(V_t)_{t\geq 0}$, the process $(\theta_t)_{t\geq 0}$ appears to be relatively smooth. This may be due to the smoothness of the subsampled target function $f$. Moreover, we note that the example in Figure 3.1 is a discretized data version of the example here in Figure 3.2.

More similarly to the discrete data case (3.3), one could also choose $(V_t)_{t\geq 0}$ to be a Markov pure jump process on $S$ that has $\pi$ as a stationary measure. Indeed, the reflected Brownian motion was constructed rather artificially. Sampling from $\mathrm{Unif}[-1,1]$ is not actually difficult in practice and we just needed a way to find a continuous-time Markov process that is stationary with respect to $\mathrm{Unif}[-1,1]$. However, there are cases, where one may not be able to sample independently from $\pi$. For instance, $\pi$ could be the measure of interest in a statistical physics simulation or Bayesian inference. In those cases, Markov chain Monte Carlo methods are used to approximate $\pi$ through a Markov chain stationary with respect to it, see, e.g., [108]. In other cases, the data might be time series data that is streamed at runtime of the algorithm – a related problem has been studied by [120]. Hence, in this chapter, we also discuss stochastic optimization in those cases or – more generally – stochastic optimization with respect to data from arbitrary sources.

As the index process $(V_t)_{t\geq 0}$ is stationary, the stochastic gradient process as defined above would, again, represent the situation of a constant learning rate $(\eta_n)_{n=1}^{\infty}$. However, as before we usually cannot hope for convergence to a stationary point if there is not a sense of a decreasing learning rate. Hence, we need to introduce an inhomogeneous variant of $(V_t)_{t\geq 0}$ that represents a decreasing learning rate. We now introduce a way to obtain such a decreasing learning rate in our continuous setting.

We start with the stochastic process $(V_t^{\mathrm{dc}})_{t\geq 0}$ that represents the index process associated to the discrete-time stochastic gradient descent dynamic (3.2) with constant learning rate parameter $\eta = 1$. This index process is given by

$$V_t^{\mathrm{dc}} = \sum_{n=1}^{\infty} y_n \mathbf{1}[t \in [n-1, n)] = \sum_{n=1}^{\infty} y_n \mathbf{1}[t - n + 1 \in [0, 1)] \qquad (t \geq 0),$$

where $y_1, y_2, \ldots \sim \pi$ i.i.d. and $\mathbf{1}[\cdot]$ is the indicator function: $\mathbf{1}[\text{true}] = 1$ and $\mathbf{1}[\text{false}] = 0$. We now want to turn the process $(V_t^{\mathrm{dc}})_{t\geq 0}$ into the index process $(V_t^{\mathrm{dd}})_{t\geq 0}$ that represents a

decreasing learning rate $(\eta_n)_{n=1}^{\infty}$. It is defined through:

$$(V_t^{\mathrm{dd}})_{t\geq 0} = \sum_{n=1}^{\infty} y_n \mathbf{1}\left[t \in [H_{n-1}, H_n)\right] = \sum_{n=1}^{\infty} y_n \mathbf{1}\left[\frac{t - H_{n-1}}{\eta_n} \in [0,1)\right],$$

where we denote $H_n := \sum_{m=1}^{n} \eta_m$. Hence, we can represent $(V_t^{\mathrm{dd}})_{t\geq 0} := (V_{\beta(t)}^{\mathrm{dc}})_{t\geq 0}$, where $\beta : [0,\infty) \to [0,\infty)$ is given by

$$\beta(t) = \sum_{n=1}^{\infty} \frac{t + n - 1 - H_{n-1}}{\eta_n} \mathbf{1}\left[t \in [H_{n-1}, H_n)\right] \qquad (t \geq 0) \tag{3.5}$$

is a piecewise linear, non-decreasing function with $\beta(t) \to \infty$, as $t \to \infty$.

Following this idea, we turn our homogeneous index process $(V_t)_{t\geq 0}$ that represents a constant learning rate into an inhomogeneous process with decreasing learning rate using a suitable rescaling function $\beta$. In that case, we obtain a stochastic gradient process of type

$$\mathrm{d}\xi_t = -\nabla_{\xi} f(\xi_t, V_{\beta(t)})\mathrm{d}t,$$

which we will use to represent the stochastic gradient descent algorithm with decreasing learning rate. Note that while we require $\beta$ to satisfy certain conditions that ensure the well-definedness of the dynamical system, it is not strictly necessary for it to be of the form (3.5). Actually, we later assume that $\beta$ is smooth.

The main contributions of this chapter are the following:

- We study stochastic gradient processes for optimization problems of the form (3.1) with finite, countably infinite, and continuous index sets $S$.
- We give conditions under which the stochastic gradient process with constant learning rate is well-defined and that it can approximate the full gradient flow $\mathrm{d}\zeta_t = -\nabla\Phi(\zeta_t)\mathrm{d}t$ at any accuracy. In addition, we study the geometric ergodicity of the stochastic gradient process and properties of its stationary measure.

- We study the well-definedness of the stochastic gradient process with decreasing learning rate and give conditions under which the process converges to the minimizer of $\Phi$ in the optimization problem (3.1).

- In numerical experiments, we show the suitability of our stochastic gradient process for (convex) polynomial regression with continuous data and the (non-convex) training of physics-informed neural networks with continuous sampling of function-valued data.

This chapter is organized as follows. In Section 3.2, we study the index process $(V_t)_{t\geq 0}$ and give examples for various combinations of index spaces $S$ and probability measures $\pi$. Then, in Sections 3.3 and 3.4, we analyze the stochastic gradient process with constant and decreasing learning rate, respectively. In Section 3.5, we review discretization techniques that allow us to turn the continuous dynamical systems into practical optimization algorithms. We employ these techniques in Section 3.6, where we present numerical experiments regarding polynomial regression and the training of physics-informed neural networks. We end with conclusions and outlook in Section 3.7.

## 3.2 The index process: Feller processes and geometric ergodicity

Before we define the stochastic gradient flow, we introduce and study the class of stochastic processes $(V_t)_{t\geq 0}$ that can be used for the data switching in (3.4). Moreover, we give an overview of appropriate processes for various measures $\pi$. For more background material on (continuous-time) stochastic processes, we refer the reader to the book by [106], the book by [79], and other standard literature.

Let $\mathcal{S} = (S, m)$ be a compact Polish space and

$$\Omega = \{\omega : [0, \infty) \to S \mid \omega \text{ is right continuous with left limits.}\}.$$

We consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, (\mathbb{P}_x)_{x\in S})$, where $\mathcal{F}$ is the smallest $\sigma$-algebra on $\Omega$ such that the mapping $\omega \to \omega(t)$ is measurable for any $t \geq 0$ and the filtration $\mathcal{F}_t$ is right continuous. Let $(V_t)_{t\geq 0}$ be a $(\mathcal{F}_t)_{t\geq 0}$ adapted stochastic process from $\Omega$ to $S$. We assume that $(V_t)_{t\geq 0}$ is Feller with respect to $(\mathcal{F}_t)_{t\geq 0}$. $(\mathbb{P}_x)_{x\in S}$ is a collection of probability measures on $\Omega$ such that $\mathbb{P}_x(V_0 = x) = 1$. For any probability measure $\mu$ on $S$, we define

$$\mathbb{P}_\mu(\cdot) := \int_S \mathbb{P}_y(\cdot)\mu(\mathrm{d}y)$$

and denote expectations with respect to $\mathbb{P}_x$ and $\mathbb{P}_\mu$ by $\mathbb{E}_x$ and $\mathbb{E}_\mu$, respectively.

Below we give a set of assumptions on the process $(V_t)_{t\geq 0}$. We need those to ensure that a certain coupling property holds. We comment on these assumptions after stating them.

**Assumption 1.** *Let $(V_t)_{t\geq 0}$ be a Feller process on $(\Omega, \mathcal{F}, (\mathcal{F}_t))_{t\geq 0}, (\mathbb{P}_x)_{x\in S})$. We assume the following:*

*(i)* *$(V_t)_{t\geq 0}$ admits a unique invariant measure $\pi$.*

*(ii)* *For any $x \in S$, there exist a family $(V_t^x)_{t\geq 0}$ and a stationary version $(V_t^\pi)_{t\geq 0}$ defined on the same probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ such that, $(V_t^x)_{t\geq 0} \overset{d}{=} (V_t)_{t\geq 0}$ in $\mathbb{P}_x$ and $(V_t^\pi)_{t\geq 0} \overset{d}{=} (V_t)_{t\geq 0}$ in $\mathbb{P}_\pi$, i.e. for any $0 \leq t_1 < \cdots < t_n$,*

$$\tilde{\mathbb{P}}(V_{t_1}^x \in A_1, \cdots, V_{t_n}^x \in A_n) = \mathbb{P}_x(V_{t_1} \in A_1, \cdots, V_{t_n} \in A_n),$$

$$\tilde{\mathbb{P}}(V_{t_1}^\pi \in A_1, \cdots, V_{t_n}^\pi \in A_n) = \mathbb{P}_\pi(V_{t_1} \in A_1, \cdots, V_{t_n} \in A_n),$$

*where $A_1, \cdots, A_n \in \mathcal{B}(S)$.*

*(iii)* *Let $T^x := \inf\{t \geq 0 \mid V_t^x = V_t^\pi\}$ be a stopping time. There exist constants $C, \delta > 0$ such that for any $t \geq 0$,*

$$\sup_{x\in S} \tilde{\mathbb{P}}(T^x \geq t) \leq C\exp(-\delta t).$$

First, we assume that $(V_t)_{t\geq 0}$ has a stationary measure $\pi$. Second, we assume that for the process $(V_t)_{t\geq 0}$ that starts from $x$ with probability 1, we can find a coupled process $(V_t^x)_{t\geq 0}$. Also, given that the process $(V_t)_{t\geq 0}$ starts with its invariant measure $\pi$, we can find a stationary version $(V_t^\pi)_{t\geq 0}$ of $(V_t)_{t\geq 0}$. Here, the processes $(V_t^x)_{t\geq 0}$ and $(V_t^\pi)_{t\geq 0}$ are defined on the same probability space. Third, we assume that the processes $(V_t^x)_{t\geq 0}$ and $(V_t^\pi)_{t\geq 0}$ intersect exponentially fast. The exponential rate can be chosen uniformly in $x$ since $S$ is compact. With Assumption 1, we have the following lemma.

**Lemma 16** (Geometric Ergodicity)**.** *Under Assumption 1, there exist constants $C, \delta > 0$ such that for any $x \in S$ and $t \geq 0$,*

$$\sup_{A\in\mathcal{B}(S)} |\mathbb{P}_x(V_t \in A) - \pi(A)| \leq C\exp(-\delta t),$$

*where $\mathcal{B}(S)$ is the set of all Borel measurable sets of $S$.*

*Proof.* For any given $x \in S$, we construct the following process by coupling $(V_t^x)_{t\geq 0}$ and $(V_t^\pi)_{t\geq 0}$:

$$\tilde{V}_t^x = \begin{cases} V_t^x, & 0 \leq t \leq T^x, \\ V_t^\pi, & t > T^x. \end{cases}$$

By the strong Markov property, $(\tilde{V}_t^x)_{t\geq 0} \stackrel{d}{=} (V_t^x)_{t\geq 0}$. For any $A \in \mathcal{B}(S)$, notice that

$$
\begin{aligned}
&|\mathbb{P}_x(V_t \in A) - \pi(A)| \\
&\qquad = |\tilde{\mathbb{P}}(V_t^x \in A) - \tilde{\mathbb{P}}(V_t^\pi \in A)| \\
&\qquad = |\tilde{\mathbb{P}}(\tilde{V}_t^x \in A) - \tilde{\mathbb{P}}(V_t^\pi \in A)| \\
&\qquad = |\tilde{\mathbb{P}}(\tilde{V}_t^x \in A, \tilde{V}_t^x \neq V_t^\pi) + \tilde{\mathbb{P}}(\tilde{V}_t^x \in A, \tilde{V}_t^x = V_t^\pi) \\
&\qquad\qquad - (\tilde{\mathbb{P}}(V_t^\pi \in A, \tilde{V}_t^x \neq V_t^\pi) + \tilde{\mathbb{P}}(V_t^\pi \in A, \tilde{V}_t^x = V_t^\pi))| \\
&\qquad = |\tilde{\mathbb{P}}(\tilde{V}_t^x \in A, \tilde{V}_t^x \neq V_t^\pi) - \tilde{\mathbb{P}}(V_t^\pi \in A, \tilde{V}_t^x \neq V_t^\pi)| \\
&\qquad \leq 2\tilde{\mathbb{P}}(\tilde{V}_t^x \neq V_t^\pi) \\
&\qquad \leq 2\tilde{\mathbb{P}}(T^x \geq t) \leq C\exp(-\delta t).
\end{aligned}
$$

From the third assumption in Assumption 1, $C$ and $\delta$ are independent of $x$ and this completes the proof. □

In the lemma above, we have shown geometric ergodicity of $(V_t)_{t\geq 0}$ in the total variation distance. Next, we show that the same rate of convergence of $(V_t)_{t\geq 0}$ holds in the weak topology.

**Corollary 1.** *Under Assumption 1, there exist constants $C, \delta > 0$ such that for any $h \in \mathcal{C}(S)$, i.e. the set of all continuous function on $S$, we have*

$$\sup_{x\in S}\left|\mathbb{E}_x[h(V_t)] - \int_S h(y)\pi(\mathrm{d}y)\right| \leq C\left\|h\right\|_\infty \exp(-\delta t)$$

*where $\left\|h\right\|_\infty := \sup_{x\in S}|h(x)|$*

*Proof.* Rewrite $\mathbb{E}_x[h(V_t)]$ as

$$\mathbb{E}_x[h(V_t)] = \int_S h(y)\mathbb{P}_x(V_t \in \mathrm{d}y).$$

Then, we have

$$\left|\mathbb{E}_x[h(V_t)] - \int_S h(y)\pi(\mathrm{d}y)\right| = \left|\int_S h(y)[\mathbb{P}_x(V_t \in \mathrm{d}y) - \pi(\mathrm{d}y)]\right|$$
$$\leq \|h\|_\infty \left|\mathbb{P}_x(V_t) - \pi\right|(S),$$

where $\frac{1}{2}\left|\mathbb{P}_x(V_t) - \pi\right|(S)$ is the total variation of the measure $\mathbb{P}_x(V_t) - \pi$. Notice that

$$\left|\mathbb{P}_x(V_t) - \pi\right|(S) = 2 \sup_{A \in \mathcal{B}(S)} |\mathbb{P}_x(V_t \in A) - \pi(A)|.$$

By Lemma 16, we have

$$\begin{aligned}\sup_{x\in S}\left|\mathbb{E}_x[h(V_t)] - \int_S h(y)\pi(\mathrm{d}y)\right| \leq& \|h\|_\infty \sup_{x\in S}\left|\mathbb{P}_x(V_t) - \pi\right|(S)\\ \leq& 2\|h\|_\infty \sup_{x\in S}\sup_{A\in\mathcal{B}(S)} |\mathbb{P}_x(V_t \in A) - \pi(A)|\\ \leq& C\|h\|_\infty \exp(-\delta t),\end{aligned}$$

which completes the proof. □

We now study four examples for processes that satisfy our assumptions: Lévy processes with two-sided reflections on a compact interval, continuous-time Markov processes on finite and countably infinite spaces, and processes on rectangular sets with independent coordinates.

### 3.2.1 Example 1: Lévy processes with two-sided reflection

For any $b > 0$, we say a triplet $((X_t)_{t\geq 0}, (L_t)_{t\geq 0}, (U_t)_{t\geq 0})$ is a solution to the Skorokhod problem of the Lévy process $(X_t)_{t\geq 0}$ on the space $S := [0, b]$ if for all $t \geq 0$,

$$V_t = X_t + L_t - U_t, \tag{3.6}$$

where $(L_t)_{t\geq 0},\ (U_t)_{t\geq 0}$ are non-decreasing right continuous processes such that

$$\int_0^\infty V_t \mathrm{d}L_t = \int_0^\infty (b - V_t)\mathrm{d}U_t = 0.$$

In other words, $(L_t)_{t\geq 0}$ and $(U_t)_{t\geq 0}$ can only increase when $(V_t)_{t\geq 0}$ is at the lower boundary 0 or the upper boundary $b$. From [2, Proposition 5.1], we immediately see that the process $(V_t)_{t\geq 0}$ in (3.6) satisfies Assumption 1. The geometric ergodicity follows from [2, Remark 5.3]. As an example, the standard Brownian Motion (BM) reflected at 0 and 1 can be written as

$$V_t = B_t + \tilde{L}_t^0 - \tilde{L}_t^1,$$

where $(B_t)_{t\geq 0}$ is a standard BM and $(\tilde{L}_t^a)_{t\geq 0}$ is the symmetric local time of $(V_t)_{t\geq 0}$ at $a \in \{0, 1\}$. Intuitively, a local time describes the time spent at a given point of a continuous stochastic process. The formal definition of symmetric local time of continuous semimartingales can be found, for example, in [106, Chapter VI]. For the optimization problem (3.1) with $S = [0, 1]$ and $\pi$ being the uniform measure on $S$, the corresponding stochastic process in (3.4) can be chosen to be this Brownian Motion with two-sided refection since its invariant measure is the uniform measure on $[0, 1]$. To see this, we employ [2, Theorem 5.4] and have

$$\pi([x, 1]) = \mathbb{P}(B_{\tau_x \wedge \tau_{x-1}} = x) = \mathbb{P}(\tau_x < \tau_{x-1}) = 1 - x \qquad (x \in [0, 1]),$$

where $\tau_a = \inf\{t \geq 0 | B_t = a\}$.

### 3.2.2 Example 2: Continuous-time Markov processes with finite states

We consider a continuous-time Markov process $V_t$ on state space $I = \{1, 2, \ldots, N\}$ with transition rate matrix

$$\mathbf{A}_N = \mathbf{\Lambda}_N - N\lambda \mathbf{I}_N,$$

where $\lambda > 0$, $\mathbf{\Lambda}_N$ is an $N \times N$ matrix the entries of which are all equal to $\lambda$, and $\mathbf{I}_N$ is the identity matrix. From [70], we know that the transition probability is given by

$$\mathbb{P}(V_{t+s} = i | V_s = j) = \frac{1 - \exp(-\lambda N t)}{N} + \exp(-\lambda N t)\mathbf{1}[i = j].$$

The invariant measure $\pi$ of $V_t$ is the uniform measure on $I$, i.e. $\pi(i) = 1/N$ for $i \in \{1, ..., N\}$. To see that $(V_t)_{t\geq 0}$ satisfies the rest of Assumption 1, consider a stationary version $(\hat{V}_t)_{t\geq 0}$ that is independent of $(V_t)_{t\geq 0}$. Let $V_0 = 1$. We define $T = \inf\{t \geq 0, V_t = \hat{V}_t\}$. Moreover, for $i,\ j \in \mathbb{N}$, we denote the $i$-th and $j$-th jump time of $(V_t)_{t\geq 0}$ and $(\hat{V}_t)_{t\geq 0}$ by $T_i$ and $\hat{T}_j$, respectively. Then we have

$$\mathbb{P}(T = 0) = \mathbb{P}(V_0 = 1, \hat{V}_0 = 1) = \mathbb{P}(V_0 = 1)\mathbb{P}(\hat{V}_0 = 1) = \frac{1}{N}.$$

For any $i,\ j \in \mathbb{N}$, since $T_i$ and $\hat{T}_j$ are independent, $\mathbb{P}(T_i = \hat{T}_j) = 0$. Let $(Y_t)_{t\geq 0} = ((V_t, \hat{V}_t))_{t\geq 0}$ be a Markov process on $I \times I$ with transition probability:

$$\begin{aligned}\mathbb{P}(Y_{t+s} = (i, j) | Y_s = (i_0, j_0)) = &\frac{1 - \exp(-2\lambda N t)}{2N}\Big(\mathbf{1}[i = i_0] + \mathbf{1}[j = j_0]\Big) \\ &+ \exp(-2\lambda N t)\mathbf{1}[(i, j) = (i_0, j_0)].\end{aligned}$$

Thus, $T$ is the first time when $(Y_t)_{t\geq 0}$ hits $\{(i, i) | i = 1, ..., N\}$. Let the $n$-th jump time of $(Y_t)_{t\geq 0}$ be $\tau_n$, for $t > 0$, we have

$$\begin{aligned}\mathbb{P}(T \geq t) =& \sum_{n\geq 1} \mathbb{P}(T = \tau_n, \tau_n \geq t) \\ =& \sum_{n\geq 1} \exp(-2(N-1)n\lambda t)\frac{N-1}{N}\frac{1}{N-1}\Big(\frac{N-2}{N-1}\Big)^{n-1} \\ \leq& C \exp(-2(N-1)\lambda t),\end{aligned}$$

where the second equality follows from

$$\mathbb{P}(T = \tau_n) = \frac{N-1}{N}\frac{1}{N-1}\Big(\frac{N-2}{N-1}\Big)^{n-1}$$

since there are $2N-4$ states available for the next jump. Thus, $(V_t)_{t\geq 0}$ satisfies Assumption 1.

### 3.2.3 Example 3: Continuous-time Markov processes with countable states

We consider a continuous-time Markov process $(V_t)_{t\geq 0}$ on state space $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ with exponential jump times. At time $t$, if $V_t \in \mathbb{N}$, it jumps to 0 with probability 1 at the next jump time. Otherwise, if $V_t = 0$, it jumps to $i$ with probability $1/2^i$. It is easy to verify that the invariant measure $\pi$ of $(V_t)_{t\geq 0}$ is $\pi(\{i\}) = 1/2^{i+1}$. One may consider $\mathbb{N}$ as one state and view $(V_t)_{t\geq 0}$ as a Markov process with two states.

To verify that $(V_t)_{t\geq 0}$ satisfies the rest of Assumption 1, similarly to the previous example, we consider a stationary version $(\hat{V}_t)_{t\geq 0}$ that is independent of $(V_t)_{t\geq 0}$. Let $V_0 = 0$ and $T = \inf\{t \geq 0, V_t = \hat{V}_t = 0\}$. For $i,\ j \in \mathbb{N}_0$, we denote the $i$-th and $j$-th jump time of $(V_t)_{t\geq 0}$ and $(\hat{V}_t)_{t\geq 0}$ by $T_i$ and $\hat{T}_j$, respectively. Then we have

$$\mathbb{P}(T = 0) = \mathbb{P}(V_0 = 0, \hat{V}_0 = 0) = \mathbb{P}(V_0 = 0)\mathbb{P}(\hat{V}_0 = 0) = \frac{1}{2}.$$

For any $i,\ j \in \mathbb{N}_0$, since $T_i$ and $\hat{T}_j$ are independent, $\mathbb{P}(T_i = \hat{T}_j) = 0$. Let $(Y_t)_{t\geq 0} = ((V_t, \hat{V}_t))_{t\geq 0}$ be a Markov process on $\mathbb{N}_0 \times \mathbb{N}_0$. Notice that $T$ is the first time when $(Y_t)_{t\geq 0}$ hits $(0, 0)$. Let

the $n$-th jump time of $(Y_t)_{t\geq 0}$ be $\tau_n$, for $t > 0$, we have

$$\begin{aligned}\mathbb{P}(T \geq t) &= \sum_{n\geq 1} \mathbb{P}(T = \tau_n, \tau_n \geq t) \\ &= \sum_{k\geq 0} \mathbb{P}(T = \tau_{2k+1}, \tau_{2k+1} \geq t) \\ &= \sum_{k\geq 0} \exp(-2(2k+1)t)\frac{1}{2^{k+2}} \leq \exp(-t),\end{aligned}$$

where the second and the third equality follows from $\mathbb{P}(T = \tau_{2k+1}) = 2^{-k-1}$ and $\mathbb{P}(T = \tau_{2k}) = 0$ for $k \geq 1$. Since $\inf\{t \geq 0, V_t = \hat{V}_t\}$ is upper bounded by $T$, $(V_t)_{t\geq 0}$ satisfies Assumption 1.

### 3.2.4 Example 4: Multidimensional processes

For multidimensional processes, Assumption 1 is satisfied if each component satisfies Assumption 1 and all components are mutually independent. We illustrate this by discussing the 2-dimensional case – higher-dimensional processes can be constructed inductively. Multidimensional processes arise, e.g., when the underlying space $S$ is multidimensional. They also arise when $S$ is one-dimensional, but we run multiple processes in parallel to obtain a mini-batch SGD instead of single-draw SGD.

Let $(S^1, m^1)$ and $(S^2, m^2)$ be two compact Polish spaces. We consider the probability triples $(\Omega^1, (\mathcal{F}_t^1))_{t\geq 0}, (\mathbb{P}_a^1)_{a\in S^1})$ and $(\Omega^2, (\mathcal{F}_t^2))_{t\geq 0}, (\mathbb{P}_b^2)_{b\in S^2})$ with $\mathbb{P}_a^1((V_0^1 = a) = \mathbb{P}_b^2(V_0^2 = b) = 1$. Let $(V_t^1)_{t\geq 0}$ and $(V_t^2)_{t\geq 0}$ be $(\mathcal{F}_t^1))_{t\geq 0}$ and $(\mathcal{F}_t^2))_{t\geq 0}$ adapted and from $\Omega^1$ to $S^1$ and $\Omega^2$ to $S^2$ respectively. In the following proposition, we construct a 2-dimensional process $(V_t^1, V_t^2)_{t\geq 0}$ from $\Omega^1 \times \Omega^2$ to $(S^1 \times S^2, m^1 + m^2)$ with a family of probability measures $(\mathbb{P}_{(a,b)})_{(a,b)\in S^1\times S^2}$ such that $\mathbb{P}_{(a,b)}(A \times B) = \mathbb{P}_a^1(A)\mathbb{P}_b^2(B)$ for $A \in \mathcal{F}^1$ and $B \in \mathcal{F}^2$.

We now show that the joint process $(V_t^1, V_t^2)_{t\geq 0}$ is Feller and satisfies Assumption 1, if the marginals do.

**Proposition 1.** *Let $(V_t^1)_{t\geq 0}$ and $(V_t^2)_{t\geq 0}$ be càdlàg and Feller with respect to $(\mathcal{F}_t^1)_{t\geq 0}$ and $(\mathcal{F}_t^2)_{t\geq 0}$, respectively, and satisfy Assumption 1 with probability $(\mathbb{P}_a^1)_{a\in S^2}$ and $(\mathbb{P}_b^2)_{b\in S^2}$, re-*

*spectively. Then* $(V_t^1, V_t^2)_{t\geq 0}$ *is also càdlàg and Feller with respect to* $\sigma(\mathcal{F}_t^1 \times \mathcal{F}_t^2)_{t\geq 0}$ *and satisfies Assumption 1 with* $(\mathbb{P}_{(a,b)})_{(a,b)\in S^1\times S^2}$.

*Proof.* It is obvious that the process $(V_t^1, V_t^2)_{t\geq 0}$ is càdlàg and Markovian. To verify the Feller property, we show that for any continuous function $F$ on $S^1 \times S^2$, $\mathbb{E}_{(x,y)}[F(V_t^1, V_t^2)]$ is continuous in $(x, y)$. We shall prove this by showing this property for separable $F$ and approximate general continuous functions using this special case. Let $f$ and $g$ be continuous functions on $S^1$ and $S^2$ respectively, then we have

$$\mathbb{E}_{(x,y)}[f(V_t^1)g(V_t^2)] = \mathbb{E}_x^1[f(V_t^1)]\mathbb{E}_y^2[g(V_t^2)], \tag{3.7}$$

which implies $\mathbb{E}_{(x,y)}[f(V_t^1)g(V_t^2)]$ is continuous in $(x, y)$ since $(V_t^1)_{t\geq 0}$ and $(V_t^2)_{t\geq 0}$ are Feller. By the Stone–Weierstrass theorem, for any $k \geq 1$, any continuous function $F$ on $S^1 \times S^2$ can be approximated as the following,

$$\sup_{(x,y)\in S^1\times S^2} \left| F(x, y) - \sum_{i=1}^{n_k} f_i^k(x) g_i^k(y) \right| \leq \frac{1}{k}$$

where $f_i^k$ and $g_i^k$ are continuous. From (3.7), this implies $\mathbb{E}_{(x,y)}[F(V_t^1, V_t^2)]$ is continuous on $S^1 \times S^2$.

Next, we prove that $(V_t^1, V_t^2)_{t\geq 0}$ satisfies Assumption 1. Let $\pi^1$ and $\pi^2$ be the invariant measures of $(V_t^1)_{t\geq 0}$ and $(V_t^2)_{t\geq 0}$, respectively. Then $\pi^1 \times \pi^2$ is the invariant measure of $(V_t^1, V_t^2)_{t\geq 0}$ since $(V_t^1)_{t\geq 0}$ and $(V_t^2)_{t\geq 0}$ are independent. From Assumption 1, we know there exist $(\tilde{\Omega}^1, \tilde{\mathcal{F}}^1, \tilde{\mathbb{P}}^1)$, $(\tilde{\Omega}^2, \tilde{\mathcal{F}}^2, \tilde{\mathbb{P}}^2)$, such that for any $a \in S^1$ and $b \in S^2$ , $(V_t^{1,a})_{t\geq 0} \overset{d}{=} (V_t^1)_{t\geq 0}$ in $\mathbb{P}_a^1$ and $(V_t^{2,b})_{t\geq 0} \overset{d}{=} (V_t^2)_{t\geq 0}$ in $\mathbb{P}_b^2$. We define $\tilde{\mathbb{P}}$ on $\tilde{\Omega}^1 \times \tilde{\Omega}^2$ such that

$$\tilde{\mathbb{P}}(A \times B) = \tilde{\mathbb{P}}^1(A)\tilde{\mathbb{P}}^2(B), \quad (A \in \tilde{\mathcal{F}}^1,\ B \in \tilde{\mathcal{F}}^2).$$

Then we have that $((V_t^{1,a})_{t\geq 0})_{a\in S^1}$ and $(V_t^{1,\pi^1})_{t\geq 0}$ are independent of $((V_t^{2,b})_{t\geq 0})_{b\in S^2}$ and $(V_t^{2,\pi^2})_{t\geq 0}$ under $\tilde{\mathbb{P}}$. Similar to the proof of Lemma 16, we construct the following processes

by the coupling method:

$$\tilde{V}_t^{1,a} = \begin{cases} V_t^{1,a}, & 0 \le t \le T^{1,a}, \\ V_t^{1,\pi^1}, & t > T^{1,a}, \end{cases}$$

and

$$\tilde{V}_t^{2,b} = \begin{cases} V_t^{2,b}, & 0 \le t \le T^{2,b}, \\ V_t^{2,\pi^2}, & t > T^{2,b}. \end{cases}$$

Then the distribution of $(\tilde{V}_t^{1,a}, \tilde{V}_t^{2,b})_{t\ge 0}$ under $\tilde{\mathbb{P}}$ is the same as the distribution of $(V_t^1, V_t^2)_{t\ge 0}$ under $\mathbb{P}_{(a,b)}$; the distribution of $(V_t^{1,\pi^1}, V_t^{2,\pi^2})_{t\ge 0}$ under $\tilde{\mathbb{P}}$ is the same as the distribution of $(V_t^1, V_t^2)_{t\ge 0}$ under $\mathbb{P}_{\pi^1\times\pi^2}$. Moreover, $(\tilde{V}_t^{1,a}, \tilde{V}_t^{2,b})_{t\ge 0}$ intersects the invariant state $(V_t^{1,\pi^1}, V_t^{2,\pi^2})_{t\ge 0}$ at time $T^{1,a} \vee T^{1,b}$. For any $(a,b) \in S^1 \times S^2$,

$$\tilde{\mathbb{P}}(T^{1,a} \vee T^{1,b} \ge t) \le \tilde{\mathbb{P}}(T^{1,a} \ge t) + \tilde{\mathbb{P}}(T^{1,b} \ge t) \le C\exp(-\delta t).$$

□

We have now discussed various index processes and their properties.We are ready to move on to study the stochastic gradient process.

## 3.3 Stochastic gradient processes with constant learning rate

We now define and study the stochastic gradient process with constant learning rate. Here, the switching between data sets is performed in a homogeneous-in-time way. Hence, it models the discrete-time stochastic gradient descent algorithm when employed with a constant

learning rate. Although, one can usually not hope to converge to the minimizer of the target functional in this case, this setting is popular in practice.

To obtain the stochastic gradient process with constant learning rate, we will couple the gradient flow (3.4) with the an appropriate process $(V_{t/\varepsilon})_{t\geq 0}$. Here, $(V_t)_{t\geq 0}$ is a Feller process introduced in Section 3.2 and $\varepsilon > 0$ is a scaling parameter that allows us to uniformly control a switching rate parameter. To define the stochastic process associated with this stochastic gradient descent problem, we first introduce the following assumptions that guarantee the existence and uniqueness of the solution of the associated dynamical system. After its formal definition and the proof of well-definedness, we move on to the analysis of the process. Indeed, we show that the process approximates the full gradient flow (3.9), as $\varepsilon \downarrow 0$. Moreover, we show that the process has a unique stationary measure to which it converges in the longtime limit at geometric speed.

We commence with regularity properties of the subsampled target function $f$ that are necessary to show the well-definedness of the stochastic gradient process.

**Assumption 2.** *Let $f(\theta, y) \in \mathcal{C}^2(\mathbb{R}^K \times S, \mathbb{R})$.*

1. *$\nabla_\theta f$, $H_\theta f$ are continuous,*
2. *$\nabla_\theta f(\theta, y)$ is Lipschitz in $x$ and the Lipschitz constant is uniform for $y \in S$, and*
3. *$f(\theta, \cdot)$ and $\nabla_\theta f$ are integrable w.r.t to the probability measure $\pi(\cdot)$, for $\theta \in \mathbb{R}^K$.*

Now, we move on to the formal definition of the stochastic gradient process.

**Definition 1.** *For $\varepsilon > 0$, the* stochastic gradient process with constant learning rate (SGPC) *is a solution of the following stochastic differential equation,*

$$\begin{cases} \mathrm{d}\theta_t^\varepsilon = -\nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon})\mathrm{d}t, \\ \theta_0^\varepsilon = \theta_0, \end{cases} \tag{3.8}$$

*where $f$ satisfies Assumption 2 and $(V_t)_{t\geq 0}$ is a Feller process that satisfies Assumption 1.*

Given these two assumptions, we can indeed show that SGPC is a well-defined Markov process.

**Proposition 2.** *Let Assumptions 1 and 2 hold. Then, equation (3.8) has a unique strong solution, i.e. the solution $(\theta_t^\varepsilon)_{t\geq 0}$ is measurable with respect to $\mathcal{F}_t^\varepsilon := \mathcal{F}_{t/\varepsilon}$ for any $t \geq 0$. For $y \in S$, $(\theta_t^\varepsilon, V_{t/\varepsilon})_{t\geq 0}$ is a Markov process under $\mathbb{P}_y$ with respect to $(\mathcal{F}_t^\varepsilon)_{t\geq 0}$.*

*Proof.* The existence and the uniqueness of the strong solution to the equation (3.8) can be found in [68, Chapter 2, Theorem 4.1]. To prove the Markov property, we define the operator $(Q_t^\varepsilon)_{t\geq 0}$ such that

$$Q_t^\varepsilon h(x,y) := \mathbb{E}_y[h(\theta_t^\varepsilon, V_{t/\varepsilon})|\theta_0^\varepsilon = x],$$

for any function $h$ bounded and measurable on $\mathbb{R}^K \times S$. For any $s, t \geq 0$, we want to show

$$\mathbb{E}[h(\theta_{t+s}^\varepsilon, V_{(t+s)/\varepsilon})|\mathcal{F}_s^\varepsilon] = Q_t^\varepsilon h(\theta_s^\varepsilon, V_{s/\varepsilon}).$$

We set $\hat{\theta}_t^\varepsilon := \theta_{t+s}^\varepsilon,\ \hat{\mathcal{F}}_t := \mathcal{F}_{t+s}^\varepsilon,\ \hat{V}_{t/\varepsilon} := V_{(t+s)/\varepsilon}$. Since

$$\theta_{t+s}^\varepsilon = \theta_s^\varepsilon - \int_s^{t+s} \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\mathrm{d}m,$$

we have

$$\hat{\theta}_t^\varepsilon = \hat{\theta}_0^\varepsilon - \int_0^t \nabla_\theta f(\hat{\theta}_m^\varepsilon, \hat{V}_{m/\varepsilon})\mathrm{d}m.$$

Hence $\hat{\theta}_t^\varepsilon$ is the solution of equation (3.8) with $\hat{\theta}_0^\varepsilon = \theta_s^\varepsilon$ and $\hat{V}_0 = V_{s/\varepsilon}$. Moreover,

$$\begin{aligned}
\mathbb{E}[h(\theta_{t+s}^\varepsilon, V_{(t+s)/\varepsilon})|\mathcal{F}_s^\varepsilon] &= \mathbb{E}[h(\hat{\theta}_t^\varepsilon, \hat{V}_{t/\varepsilon})|\hat{\theta}_0^\varepsilon = \theta_s^\varepsilon, \hat{V}_0 = V_{s/\varepsilon}] \\
&= \mathbb{E}_{\hat{V}_0}[h(\hat{\theta}_t^\varepsilon, \hat{V}_{t/\varepsilon})|\hat{\theta}_0^\varepsilon = \theta_s^\varepsilon] \\
&= Q_t^\varepsilon h(\theta_s^\varepsilon, V_{s/\varepsilon}),
\end{aligned}$$

where the second equality and third equality follow from the homogeneous Markov property of $(V_t^\varepsilon)_{t\geq 0}$. □

### 3.3.1 Approximation of the full gradient flow

We now let $\varepsilon \to 0$ and study the limiting behavior of SGPC. Indeed, we aim to show that here the SGPC converges to the *full gradient flow*

$$\mathrm{d}\zeta_t = -\Big[\int_S \nabla_\zeta f(\zeta_t, v)\pi(\mathrm{d}v)\Big]\mathrm{d}t. \tag{3.9}$$

We study this topic for two reasons: First, we aim to understand the interdependence of $(V_t)_{t\geq 0}$ and $(\theta_t^\varepsilon)_{t\geq 0}$. Second, we understand SGPC as an approximation to the full gradient flow (3.9), as motivated in the introduction. Hence, we should show that SGPC can approximate the full gradient flow at any accuracy.

We now denote $g(\cdot) := \int_S \nabla_\zeta f(\cdot, v)\pi(\mathrm{d}v) \in \mathcal{C}^1(\mathbb{R}^K, \mathbb{R}^K)$. Then, we can define $(\zeta_t)_{t\geq 0}$ through the dynamical system $\mathrm{d}\zeta_t = -g(\zeta_t)\mathrm{d}t$. Moreover, let $\mathcal{C}([0,\infty) : \mathbb{R}^K)$ be the space of continuous functions from $[0,\infty)$ to $\mathbb{R}^K$ equipped with the distance

$$\rho\Big((\varphi_t)_{t\geq 0}, (\varphi_t')_{t\geq 0}\Big) := \int_0^\infty \exp(-t)(1 \wedge \sup_{0\leq s\leq t} \|\varphi_s - \varphi_s'\|)\mathrm{d}t,$$

where $(\varphi_t)_{t\geq 0}, (\varphi_t')_{t\geq 0} \in \mathcal{C}([0,\infty) : \mathbb{R}^K)$. We study the weak limit of the system (3.8) as $\varepsilon \to 0$. Similar problems have been discussed in, for example, [68] and [67].

**Theorem 3.** *Let $\theta_0^\varepsilon = \theta_0$ and $\zeta_0 = \theta_0$. Moreover, let $(\theta_t^\varepsilon)_{t\geq 0}$ and $(\zeta_t)_{t\geq 0}$ solve (3.8) and (3.9), respectively. Then $(\theta_t^\varepsilon)_{t\geq 0}$ under $\mathbb{P}_\pi$ converges weakly to $(\zeta_t)_{t\geq 0}$ in $\mathcal{C}([0,\infty) : \mathbb{R}^K)$ as $\varepsilon \to 0$, i.e. for any bounded continuous function $F : \mathcal{C}([0,\infty) : \mathbb{R}^K) \to \mathbb{R}$, we have*

$$\mathbb{E}_\pi[F\big((\theta_t^\varepsilon)_{t\geq 0}\big)] \to \mathbb{E}_\pi[F\big((\zeta_t)_{t\geq 0}\big)] = F\big((\zeta_t)_{t\geq 0}\big).$$

*Proof.* We first verify that $(\theta_t^\varepsilon)_{t\geq 0}$ is tight by checking:

1. $\sup_{0<\varepsilon<1} \|\theta_0^\varepsilon\| < +\infty$;
2. *For any fixed* $T > 0$, $\lim_{\delta\to 0} \sup_{0<\varepsilon<1} \sup_{s,t\in[0,T],|s-t|\leq\delta} \|\theta_t^\varepsilon - \theta_s^\varepsilon\| \to 0$.

The first condition follows from $\theta_0^\varepsilon = \theta_0$. For the second condition, by Assumption 2, let $C_0 = \sup_{y\in S} \|\nabla_\theta f(0,y)\|$ and $L_f$ be the Lipschitz constant of $\nabla_\theta f(\cdot, y)$, we have

$$\begin{aligned}
\frac{\mathrm{d}\left\|\theta_t^\varepsilon\right\|^2}{\mathrm{d}t} &= -2\left\langle \theta_t^\varepsilon, \nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon})\right\rangle \\
&= -2\left\langle \theta_t^\varepsilon, \nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon}) - \nabla_\theta f(0, V_{t/\varepsilon})\right\rangle - 2\left\langle \theta_t^\varepsilon, \nabla_\theta f(0, V_{t/\varepsilon})\right\rangle \\
&\leq 2L_f \left\|\theta_t^\varepsilon\right\|^2 + 2C_0 \left\|\theta_t^\varepsilon\right\| \\
&\leq 2L_f \left\|\theta_t^\varepsilon\right\|^2 + \left\|\theta_t^\varepsilon\right\|^2 + C_0^2 \\
&= (2L_f + 1)\left\|\theta_t^\varepsilon\right\|^2 + C_0^2.
\end{aligned}$$

By Grönwall's inequality,

$$\left\|\theta_t^\varepsilon\right\|^2 \leq (\|\theta_0\|^2 + C_0^2)e^{(2L_f+1)t}. \tag{3.10}$$

Therefore, $\theta_t^\varepsilon$ is bounded on any finite time interval. For any fixed $T > 0$, let

$$C_{T,f,\theta_0} = \sup_{\|x\|\leq(\|\theta_0\|^2+C_0^2)e^{(2L_f+1)T},\ y\in S} \|\nabla_\theta f(x,y)\|.$$

Then for any $s, t \in [0,T]$,

$$\|\theta_t^\varepsilon - \theta_s^\varepsilon\| \leq \int_s^t \left\|\nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\right\| \mathrm{d}m \leq C_{T,f,\theta_0}|t-s|.$$

Hence, $(\theta_t^\varepsilon)_{t\geq 0}$ is tight in $\mathcal{C}([0,\infty) : \mathbb{R}^K)$. By Prokhorov's theorem, let $(\theta_t)_{t\geq 0}$ be a weak limit of $(\theta_t^\varepsilon)_{t\geq 0}$. We shall verify that $(\theta_t)_{t\geq 0}$ satisfies equation (3.9), which is equivalent to show

that for any bounded differentiable function $\varphi$, $h$

$$\mathbb{E}_\pi\Big[\Big(\varphi(\theta_t)-\varphi(\theta_s)+\int_s^t\langle\nabla_\theta\varphi(\theta_m),g(\theta_m)\rangle\,\mathrm{d}m\Big)h\Big((\theta_{t_i})_{i=1,\dots,n}\Big)\Big]=0,$$

$\forall\ 0\le t_1<\cdots<t_n\le s$. The case $t=0$ is obvious. Since $(\theta_t^\varepsilon)_{t\ge0}$ is a strong solution to equation (3.8), for any $0\le s<t$,

$$\varphi(\theta_t^\varepsilon)=\varphi(\theta_s^\varepsilon)-\int_s^t\big\langle\nabla_\theta\varphi(\theta_m^\varepsilon),\nabla_\theta f(\theta_m^\varepsilon,V_{m/\varepsilon})\big\rangle\,\mathrm{d}m. \tag{3.11}$$

Hence, we have

$$\mathbb{E}_\pi\Big[\Big(\varphi(\theta_t^\varepsilon)-\varphi(\theta_s^\varepsilon)+\int_s^t\big\langle\nabla_\theta\varphi(\theta_s^\varepsilon),\nabla_\theta f(\theta_m^\varepsilon,V_{m/\varepsilon})\big\rangle\,\mathrm{d}m\Big)h\Big((\theta_{t_i}^\varepsilon)_{i=1,\dots,n}\Big)\Big]=0,$$

Moreover, when $\varepsilon\to0$,

$$\mathbb{E}_\pi\Big[\Big(\varphi(\theta_t^\varepsilon)-\varphi(\theta_s^\varepsilon)\Big)h\Big((\theta_{t_i}^\varepsilon)_{i=1,\dots,n}\Big)\Big]\to\mathbb{E}_\pi\Big[\Big(\varphi(\theta_t)-\varphi(\theta_s)\Big)h\Big((\theta_{t_i})_{i=1,\dots,n}\Big)\Big].$$

Hence, all we need to show is the following

$$\mathbb{E}_\pi\Big[\Big(\int_s^t\big\langle\nabla_\theta\varphi(\theta_m^\varepsilon),\nabla_\theta f(\theta_m^\varepsilon,V_{m/\varepsilon})\big\rangle\,\mathrm{d}m-\int_s^t\langle\nabla_\theta\varphi(\theta_m^\varepsilon),g(\theta_m^\varepsilon)\rangle\,\mathrm{d}m\Big)h\Big((\theta_{t_i}^\varepsilon)_{i=1,\dots,n}\Big)\Big]\to0, \tag{3.12}$$

which is equivalent to prove that

$$\mathbb{E}_\pi\Big[\int_s^t\big\langle\nabla_\theta\varphi(\theta_m^\varepsilon),\nabla_\theta f(\theta_m^\varepsilon,V_{m/\varepsilon})\big\rangle\,\mathrm{d}m-\int_s^t\langle\nabla_\theta\varphi(\theta_m^\varepsilon),g(\theta_m^\varepsilon)\rangle\,\mathrm{d}m\Big|\mathcal{F}_s^\varepsilon\Big]\to0. \tag{3.13}$$

Let $\tilde{\varepsilon} := 1/[1/\sqrt{\varepsilon}]$, where $[x]$ is the greatest integer less than or equal to $x$. Then we have the following decomposition

$$\begin{aligned}
&\mathbb{E}_\pi\Big[\int_s^t \big\langle \nabla_\theta \varphi(\theta_m^\varepsilon), \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\big\rangle \,\mathrm{d}m - \int_s^t \langle \nabla_\theta \varphi(\theta_m^\varepsilon), g(\theta_m^\varepsilon)\rangle \,\mathrm{d}m \Big| \mathcal{F}_s^\varepsilon\Big] \\
=&\tilde{\varepsilon}\sum_{i=0}^{1/\tilde{\varepsilon}} \tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_{s+i(t-s)\tilde{\varepsilon}}^{s+(i+1)(t-s)\tilde{\varepsilon}} \big\langle \nabla_\theta \varphi(\theta_m^\varepsilon), \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon}) - g(\theta_m^\varepsilon)\big\rangle \,\mathrm{d}m \Big| \mathcal{F}_s^\varepsilon\Big] \\
=&\tilde{\varepsilon}\sum_{i=0}^{1/\tilde{\varepsilon}} \mathbb{E}_\pi\Big[\tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_{s+i(t-s)\tilde{\varepsilon}}^{s+(i+1)(t-s)\tilde{\varepsilon}} \big\langle \nabla_\theta \varphi(\theta_m^\varepsilon), \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon}) - g(\theta_m^\varepsilon)\big\rangle \,\mathrm{d}m \Big| \mathcal{F}_{s+i(t-s)\tilde{\varepsilon}}^\varepsilon\Big]\Big| \mathcal{F}_s^\varepsilon\Big].
\end{aligned}$$

We claim that as $\varepsilon \to 0$,

$$\sup_{0\le r<t} \tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(\theta_m^\varepsilon, V_{m/\varepsilon})\mathrm{d}m \Big| \mathcal{F}_r^\varepsilon\Big] \to 0, \tag{3.14}$$

where $G(x,y) := \langle \nabla_\theta \varphi(x), \nabla_\theta f(x,y) - g(x)\rangle$. Notice that for any fixed $t > 0$, $(\theta_s^\varepsilon)_{0\le s\le t}$ is uniformly equicontinuous. Hence, we have

$$\begin{aligned}
\sup_{0\le r\le t}\sup_{r\le m\le r+(t-s)\tilde{\varepsilon}} \|\theta_m^\varepsilon - \theta_r^\varepsilon\| &= \sup_{0\le r\le t}\sup_{r\le m\le r+(t-s)\tilde{\varepsilon}} \int_r^{r+(t-s)\tilde{\varepsilon}} \big\|\nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\big\| \,\mathrm{d}m \\
&\le \tilde{\varepsilon} \sup_{0\le m\le t} \big\|\nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\big\|.
\end{aligned}$$

Therefore, as $\varepsilon \to 0$,

$$\sup_{0\le r<t}\Big|\tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(\theta_m^\varepsilon, V_{m/\varepsilon})\mathrm{d}m \Big| \mathcal{F}_r^\varepsilon\Big] - \tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(\theta_r^\varepsilon, V_{m/\varepsilon})\mathrm{d}m \Big| \mathcal{F}_r^\varepsilon\Big]\Big| \to 0.$$

Hence, (3.14) is equivalent to

$$\sup_{0\le r<t}\Big|\tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(\theta_r^\varepsilon, V_{m/\varepsilon})\mathrm{d}m \Big| \mathcal{F}_r^\varepsilon\Big]\Big| \to 0. \tag{3.15}$$

By Corollary 1,

$$
\begin{aligned}
&\sup_{0\le r<t}\left|\tilde{\varepsilon}^{-1}\mathbb{E}_{\pi}\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(\theta_r^{\varepsilon},V_{m/\varepsilon})\mathrm{d}m\Big|\mathcal{F}_r^{\varepsilon}\Big]\right|\\
=&\sup_{0\le r<t}\left|\tilde{\varepsilon}^{-1}\mathbb{E}_{V_{r/\varepsilon},x=\theta_r^{\varepsilon}}\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(x,V_{(m-r)/\varepsilon})\mathrm{d}m\Big]\right|\\
=&\sup_{0\le r<t}\left|\tilde{\varepsilon}^{-1}\mathbb{E}_{V_{r/\varepsilon},x=\theta_r^{\varepsilon}}\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} \big\langle\nabla_\theta\varphi(x),\nabla_\theta f(x,V_{(m-r)/\varepsilon})-g(x)\big\rangle\,\mathrm{d}m\Big]\right|\\
=&\sup_{0\le r<t}\left|\tilde{\varepsilon}^{-1}\int_r^{r+(t-s)\tilde{\varepsilon}}\mathbb{E}_{V_{r/\varepsilon},x=\theta_r^{\varepsilon}}\Big[\big\langle\nabla_\theta\varphi(x),\nabla_\theta f(x,V_{(m-r)/\varepsilon})-g(x)\big\rangle\Big]\mathrm{d}m\right|\\
\le&\sup_{0\le r<t}\tilde{\varepsilon}^{-1}\int_r^{r+(t-s)\tilde{\varepsilon}}\left|\mathbb{E}_{V_{r/\varepsilon},x=\theta_r^{\varepsilon}}\Big[\big\langle\nabla_\theta\varphi(x),\nabla_\theta f(x,V_{(m-r)/\varepsilon})-g(x)\big\rangle\Big]\right|\mathrm{d}m\\
\le&\sup_{0\le r<t}\tilde{\varepsilon}^{-1}\left\|\langle\nabla_\theta\varphi(\theta_r^{\varepsilon}),\nabla_\theta f(\theta_r^{\varepsilon},\cdot)\rangle\right\|_\infty\int_r^{r+(t-s)\tilde{\varepsilon}}e^{-\delta(m-r)/\varepsilon}dm\\
=&\sup_{0\le r<t}\tilde{\varepsilon}^{-1}\left\|\langle\nabla_\theta\varphi(\theta_r^{\varepsilon}),\nabla_\theta f(\theta_r^{\varepsilon},\cdot)\rangle\right\|_\infty\int_0^{(t-s)\tilde{\varepsilon}}e^{-\delta k/\varepsilon}dk\\
\le&C_{t,\varphi,f}\frac{\varepsilon}{\delta\tilde{\varepsilon}}\le C_{t,\varphi,f}\frac{\sqrt{\varepsilon}}{2\delta}\to 0.
\end{aligned}
$$

This completes the proof of (3.13). Hence, any weak limit of $(\theta_t^{\varepsilon})_{t\ge0}$ is a martingale solution to equation (3.9). Since equation (3.9) is a deterministic ordinary differential equation and $\theta_0^{\varepsilon}=\theta_0$ is independent of $\varepsilon$, we have $(\theta_t^{\varepsilon})_{t\ge0}$ converges weakly to $(\zeta_t)_{t\ge0}$ as $\varepsilon\to0$. □

In general, the constant $C_{t,\varphi,f}$ cannot be controlled since we only assume $\nabla_\theta f$ is Lipschitz in $\theta$. In this case, $\theta_t$ is not bounded uniformly in time. We shall study the convergence rate under Wasserstein distance in Corollary 2, which is also not uniform in time, yet goes to 0 as $\varepsilon\to0$. In section 3.3.2, we shall show that with an additional assumption, this constant can be controlled and we obtain a longtime result as $t\to\infty$ with fixed $\varepsilon$. Moreover, the obtained upper bound of the Wasserstein distance between the invariant states goes to 0 as $\varepsilon\to0$. See Proposition 3 for more details.

Before stating Corollary 2, we need to introduce some notation. Let $\nu$ and $\nu'$ be two probability measures on $(\mathbb{R}^K,\mathcal{B}(\mathbb{R}^K))$. We define the Wasserstein distance between those

measures by

$$\mathcal{W}_d(\nu,\nu') = \inf_{\Gamma\in\mathcal{H}(\nu,\nu')}\int_{\mathbb{R}^K\times\mathbb{R}^K} d(y,y')\Gamma(\mathrm{d}y,\mathrm{d}y'),$$

where $d(y,y') := 1\wedge\|y-y'\|$ and $\mathcal{H}(\nu,\nu')$ is the set of coupling between $\nu$ and $\nu'$, i.e.

$$\mathcal{H}(\nu,\nu') = \{\Gamma\in\Pr(\mathbb{R}^K\times\mathbb{R}^K) : \Gamma(A\times\mathbb{R}^K)=\nu(A), \Gamma(\mathbb{R}^K\times B)=\nu'(B), \forall A,B\in\mathcal{B}(\mathbb{R}^K)\}.$$

To simplify the notation, for $B\in\mathcal{B}(\mathbb{R}^K)$, $\theta\in\mathbb{R}^K$, and $y\in S$, we denote

$$C_t^\varepsilon(B|\theta,y) := \mathbb{P}_y(\theta_t^\varepsilon\in B|\theta_0^\varepsilon=\theta),$$

$$C_t^\varepsilon(B|\theta,\pi) := \mathbb{P}_\pi(\theta_t^\varepsilon\in B|\theta_0^\varepsilon=\theta),$$

where $\pi$ is the invariant measure of $(V_t)_{t\geq 0}$.

Now we study the approximation property of SGPC in the Wasserstein distance. Indeed, the following corollary follows immediately from Theorem 3.

**Corollary 2.** *There exists a function $\alpha:(0,1)\to[0,1]$, such that*

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|\theta_0,\pi),\delta(\cdot-\zeta_t)) \leq (\exp(t)\alpha(\varepsilon))\wedge 1$$

*and* $\lim_{\varepsilon\to 0}\alpha(\varepsilon)=0$.

*Proof.* By Theorem 3, we have $(\theta_t^\varepsilon)_{t\geq 0}\Rightarrow(\zeta_t)_{t\geq 0}$. By Skorokhod's representation theorem, there exists a sequence $(\tilde{\theta}_t^\varepsilon)_{t\geq 0}$ such that

$$\begin{aligned}&(\tilde{\theta}_t^\varepsilon)_{t\geq 0}\overset{d}{=}(\theta_t^\varepsilon)_{t\geq 0}\text{ under }\mathbb{P}_\pi,\\ &\rho\Big((\tilde{\theta}_t^\varepsilon-\zeta_t)_{t\geq 0},0\Big)\to 0\text{ almost surely in }\mathbb{P}_\pi.\end{aligned}$$

This implies

$$\mathbb{E}_\pi[F((\tilde{\theta}_t^\varepsilon - \zeta_t)_{t\geq 0})] \to F(0),$$

for any bounded continuous function $F$ on $\mathcal{C}([0,\infty) : \mathbb{R}^K)$. By taking

$$F((\tilde{\theta}_t^\varepsilon - \zeta_t)_{t\geq 0})) = \sup_{t\geq 0} \exp(-t) \left(1 \wedge \sup_{0\leq s\leq t} \left\|\tilde{\theta}_t^\varepsilon - \zeta_t\right\|\right)$$

and

$$\alpha(\varepsilon) := \mathbb{E}_\pi\Big[\sup_{t\geq 0} \exp(-t) \left(1 \wedge \sup_{0\leq s\leq t} \left\|\tilde{\theta}_t^\varepsilon - \zeta_t\right\|\right)\Big] \to 0,$$

we have, for all $t \geq 0$,

$$\mathbb{E}_\pi\Big[1 \wedge \left\|\tilde{\theta}_t^\varepsilon - \zeta_t\right\|\Big] \leq \exp(t)\alpha(\varepsilon).$$

Since $1 \wedge \left\|\tilde{\theta}_t^\varepsilon - \zeta_t\right\| \leq 1$ and $\tilde{\theta}_t^\varepsilon \stackrel{d}{=} \theta_t^\varepsilon$, denoting the distribution of $\tilde{\theta}_t^\varepsilon$ as $F_{\tilde{\theta}_t^\varepsilon}$,

$$\begin{aligned}\mathcal{W}_d(C_t^\varepsilon(\cdot|\theta_0,\pi), \delta(\cdot - \zeta_t)) \leq& \mathcal{W}_d(F_{\tilde{\theta}_t^\varepsilon}, C_t^\varepsilon(\cdot|\theta_0,\pi)) + \mathbb{E}_\pi\Big[1 \wedge \left\|\tilde{\theta}_t^\varepsilon - \zeta_t\right\|\Big] \\ \leq& (\exp(t)\alpha(\varepsilon)) \wedge 1.\end{aligned}$$

□

Finally in this section, we look at a technical result concerning the asymptotic behavior of the full gradient flow $(\zeta_t)_{t\geq 0}$. First, we will additionally assume that the subsampled target function $f(\cdot, y)$ in the optimization problem is strongly convex, with a convexity parameter that does not depend on $y \in S$. We state this assumption below.

**Assumption 3** (Strong Convexity)**.** *For any* $x_1, x_2 \in \mathbb{R}^K$,

$$\langle x_1 - x_2, \nabla_\theta f(x_1, y) - \nabla_\theta f(x_2, y)\rangle \geq \kappa \left\|x_1 - x_2\right\|^2$$

*where* $\kappa > 0$ *and* $\kappa$ *is independent of* $y \in S$.

Strong convexity implies, of course, that the full target function $g := \int_S \nabla_\theta f(\cdot, y)\pi(\mathrm{d}y)$ has a unique minimizer $\theta^*$. It also implies that the associated full gradient flow $(\zeta_t)_{t\geq 0}$ converges at exponential speed to this unique minimizer. We give a short proof of this statement below.

**Lemma 17.** *Let* $(\zeta_t)_{t\geq 0}$ *be the process that solves (3.9) with initial data* $\theta_0$*. Under Assumption 3, we have*

$$\left\|\zeta_t - \theta_*\right\|^2 \leq \left\|\theta_0 - \theta_*\right\|^2 \exp(-\kappa t),$$

*where* $\theta_*$ *is a stationary solution of (3.9).*

*Proof.* Since $\theta_*$ is a stationary solution,

$$g(\theta_*) = 0 \quad \text{and} \quad \mathrm{d}(\zeta_t - \theta_*) = -(g(\zeta_t) - g(\theta_*))\mathrm{d}t.$$

Therefore,

$$\frac{\mathrm{d}\left\|\zeta_t - \theta_*\right\|^2}{\mathrm{d}t} = 2\left\langle \zeta_t - \theta_*, \frac{\mathrm{d}(\zeta_t - \theta_*)}{\mathrm{d}t}\right\rangle = -2\left\langle \zeta_t - \theta_*, g(\zeta_t) - g(\theta_*)\right\rangle \leq -2\kappa\left\|\zeta_t - \theta_*\right\|^2.$$

By Grönwall's inequality, we have $\left\|\zeta_t - \theta_*\right\|^2 \leq \left\|\theta_0 - \theta_*\right\|^2 \exp(-\kappa t)$. □

### 3.3.2 Longtime behavior and ergodicity

We now study the longtime behavior of SGPC, i.e. the behavior and distribution of $(\theta_t^\varepsilon, V_{t/\varepsilon})$ for $t \gg 0$ large. Indeed, the main result of this section will be the geometric ergodicity of

this coupled process and a study of its stationary measure. Initially, we study stability of the stochastic gradient process $(\theta_t^\varepsilon)_{t\geq 0}$.

**Lemma 18.** *Under Assumption 3, we have*

$$\|\theta_t^\varepsilon\|^2 \leq \|\theta_0^\varepsilon\|^2 \exp(-\kappa t) + \frac{8K_f^2}{\kappa^2},$$

*where* $K_f := \sup_{y\in S} \|\nabla_\theta f(0,y)\|$.

*Proof.* By Itô's formula, we have

$$\frac{\mathrm{d}\,\|\theta_t^\varepsilon\|^2}{\mathrm{d}t} = 2\left\langle \theta_t^\varepsilon, d\theta_t^\varepsilon/dt\right\rangle = -2\left\langle \theta_t^\varepsilon, \nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon})\right\rangle. \tag{3.16}$$

Assumption 3 implies,

$$\begin{aligned}
\left\langle \theta_t^\varepsilon, \nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon})\right\rangle =& \left\langle \theta_t^\varepsilon - 0, \nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon}) - \nabla_\theta f(0, V_{t/\varepsilon})\right\rangle + \left\langle \theta_t^\varepsilon, \nabla_\theta f(0, V_{t/\varepsilon})\right\rangle \\
\geq& \kappa\,\|\theta_t^\varepsilon\|^2 - \|\theta_t^\varepsilon\|\,\left\|\nabla_\theta f(0, V_{t/\varepsilon})\right\| \\
\geq& \frac{\kappa}{2}\,\|\theta_t^\varepsilon\|^2 - \frac{4}{\kappa}\left\|\nabla_\theta f(0, V_{t/\varepsilon})\right\|^2 \\
\geq& \frac{\kappa}{2}\,\|\theta_t^\varepsilon\|^2 - \frac{4K_f^2}{\kappa}.
\end{aligned}$$

Hence, (3.16) implies

$$\frac{\mathrm{d}\,\|\theta_t^\varepsilon\|^2}{\mathrm{d}t} \leq -\kappa\,\|\theta_t^\varepsilon\|^2 + \frac{8K_f^2}{\kappa}. \tag{3.17}$$

Multiplying $\exp(\kappa t)$ on both sides of (3.17), we get

$$\frac{\mathrm{d}(\|\theta_t^\varepsilon\|^2 \exp(\kappa t))}{\mathrm{d}t} \leq \frac{8K_f^2 \exp(\kappa t)}{\kappa},$$

that is

$$\|\theta_t^\varepsilon\|^2 \exp(\kappa t) - \|\theta_0^\varepsilon\|^2 \leq \frac{8K_f^2(\exp(\kappa t) - 1)}{\kappa^2} \leq \frac{8K_f^2 \exp(\kappa t)}{\kappa^2}$$

and therefore,

$$\|\theta_t^\varepsilon\|^2 \leq \|\theta_0^\varepsilon\|^2 \exp(-\kappa t) + \frac{8K_f^2}{\kappa^2}.$$

□

Using this lemma, we are now able to prove the first main result of this section, showing geometric ergodicity of $(\theta_t^\varepsilon, V_{t/\varepsilon})_{t\geq 0}$. First, we introduce a Wasserstein distance, on the space on which $(\theta_t^\varepsilon, V_{t/\varepsilon})_{t\geq 0}$ lives. Let $\Pi$ and $\Pi'$ be two probability measures on $(\mathbb{R}^K \times S, \mathcal{B}(\mathbb{R}^K \times S))$. We define the Wasserstein distance between those measures by

$$\widetilde{\mathcal{W}}_{\tilde{d}}(\Pi, \Pi') = \inf_{\widetilde{\Gamma} \in \mathcal{H}(\Pi, \Pi')} \int_{(\mathbb{R}^K \times S) \times (\mathbb{R}^K \times S)} \tilde{d}((u, v), (u', v')) \widetilde{\Gamma}(\mathrm{d}u\mathrm{d}v, \mathrm{d}u'\mathrm{d}v'),$$

where $\tilde{d}((u, v), (u', v')) := \mathbf{1}_{v \neq v'} + (1 \wedge \|u - u'\|)\mathbf{1}_{v=v'}$. Intuitively, when the indices $v$ and $v'$ are different, the distance should be large regardless of the distance between $u$ and $u'$; when $v = v'$, the distance coincides with $d(\cdot, \cdot)$. For $a \in S$ and $m \in \mathbb{R}^K$, let $H_t^\varepsilon(\cdot|m, a)$ be the distribution of $(\theta_t^\varepsilon, V_{t/\varepsilon})$ under $\mathbb{P}_a$ with $\theta_0^\varepsilon = m$. Moreover, recall that $(V_t)_{t\geq 0}$ is a Feller process that satisfies Assumption 1. More specifically, it satisfies Assumption 1 (iii) with a constant $\delta$:

$$\sup_{x \in S} \tilde{\mathbb{P}}(T^x \geq t) \leq C \exp(-\delta t),$$

where $T^x := \inf\{t \geq 0 \mid V_t^x = V_t^\pi\}$. With $\delta$ defined in this way, we have the following theorem.

**Theorem 4.** *Under Assumption 3, for any $0 < \varepsilon \leq 1 \wedge (\delta/2\kappa)$, the (coupled) process $(\theta_t^\varepsilon, V_{t/\varepsilon})_{t\geq 0}$ admits an unique stationary measure $\Pi^\varepsilon$ on $(\mathbb{R}^K \times S, \mathcal{B}(\mathbb{R}^K \times S))$. Moreover,*

$$\widetilde{\mathcal{W}}_{\tilde{d}}^2(H_t^\varepsilon(\cdot|m,a), \Pi^\varepsilon) \leq C_f \exp(-\kappa t) \int_{\mathbb{R}^K} (1 + \|x-m\|^2) \Pi^\varepsilon(\mathrm{d}x, S), \tag{3.18}$$

$$\widetilde{\mathcal{W}}_{\tilde{d}}^2(H_t^\varepsilon(\cdot|m,\pi), \Pi^\varepsilon) \leq C_f \exp(-\kappa t) \int_{\mathbb{R}^K} \|x-m\|^2 \Pi^\varepsilon(\mathrm{d}x, S), \tag{3.19}$$

*where the constant $C_f$ only depends on $f$.*

*Proof.* To obtain the existence of the invariant measure, we apply the weak form of Harris' Theorem in [31, Theorem 3.7] by verifying the Lyapunov condition, the $\tilde{d}$-contracting condition, and the $\tilde{d}$-small condition.

(i) **Lyapunov condition:** Let $V(x,y) = \|x\|^2$. To verify that it satisfies (2.1) in [31, Definition 2.1], we take $x = \theta_0^\varepsilon$ and $(P_t)_{t\geq 0}$ to be the semi-group associated with $(\theta_t^\varepsilon)_{t\geq 0}$. Then Lemma 18 yields that $V(x,y)$ is a Lyapunov function. The existence of the Lyapunov function especially prevents the coupled processes from going to infinity.

(ii) **$\tilde{d}$-contracting condition:** $\tilde{d}$-contracting states that there exists $t^* > 0$ such that for any $t > t^*$, there exists some $\alpha < 1$ such that

$$\widetilde{\mathcal{W}}_{\tilde{d}}(\delta_{(m,a)} P_t^\varepsilon, \delta_{(n,b)} P_t^\varepsilon) \leq \alpha \tilde{d}((m,a),(n,b))$$

for any $(m,a),(n,b) \in \mathbb{R}^K \times S$ such that $\tilde{d}((m,a),(n,b)) < 1$. Here, $(P_t^\varepsilon)_{t\geq 0}$ is the semi-group operator associated with (3.8). Notice that $\tilde{d}((m,a),(n,b)) < 1$ implies $a = b$. Let $(\theta_t^{(m,a)})_{t\geq 0}$, $(\theta_t^{(n,a)})_{t\geq 0}$ solve the following equations:

$$\begin{aligned}
\theta_t^{(m,a)} &= m - \int_0^t \nabla_\theta f(\theta_t^{(m,a)}, \tilde{V}_{s/\varepsilon}^a) \mathrm{d}s, \\
\theta_t^{(n,a)} &= n - \int_0^t \nabla_\theta f(\theta_t^{(n,a)}, \tilde{V}_{s/\varepsilon}^a) \mathrm{d}s,
\end{aligned}$$

where $\tilde{V}_t^a = V_t^a$ for $t \le T^a$ and $\tilde{V}_t^a = V_t^\pi$ for $t > T^a$. Then by Itô's formula and Assumption 3, we obtain

$$\begin{aligned}\mathrm{d}\left\|\theta_t^{(m,a)} - \theta_t^{(n,a)}\right\|^2 / \mathrm{d}t =& -2\left\langle \theta_t^{(m,a)} - \theta_t^{(n,a)}, \nabla_\theta f(\theta_t^{(m,a)}, \tilde{V}_{s/\varepsilon}^a) - \nabla_\theta f(\theta_t^{(n,a)}, \tilde{V}_{s/\varepsilon}^a)\right\rangle \\ \le& -\kappa\left\|\theta_t^{(m,a)} - \theta_t^{(n,a)}\right\|^2.\end{aligned}$$

By Grönwall's inequality, we have

$$\left\|\theta_t^{(m,a)} - \theta_t^{(n,a)}\right\|^2 \le \exp(-\kappa t)\left\|m-n\right\|^2. \tag{3.20}$$

Noticing that $\tilde{d}((m,a),(n,b)) < 1$ implies $\|m-n\|^2 < 1$, by choosing $t \ge \frac{1}{\kappa}$, we obtain

$$\begin{aligned}\left\|\theta_t^{(m,a)} - \theta_t^{(n,a)}\right\|^2 &\le \exp(-1)\left\|m-n\right\|^2 \\ &= \exp(-1)(\|m-n\|^2 \wedge 1) = \exp(-1)\tilde{d}^2((m,a),(n,a)).\end{aligned}$$

Therefore, with $t^* = 1/\kappa$,

$$\widetilde{\mathcal{W}}_{\tilde{d}}(H_t^\varepsilon(\cdot|m,a), H_t^\varepsilon(\cdot|n,b)) \le \tilde{\mathbb{E}}\Big[\left\|\theta_t^{(m,a)} - \theta_t^{(n,a)}\right\|\Big] \le \exp(-1)\tilde{d}((m,a),(n,b)),$$

where $\tilde{\mathbb{E}}$ is the expectation under $\tilde{\mathbb{P}}$ (see Assumption 1).

(iii) $\tilde{d}$**-small condition:** We shall verify that there exists $t_* > 0$ such that for any $t > t_*$, the sublevel set $\mathscr{V} := \{(x,y) \in \mathbb{R}^K \times S \mid V(x,y) \le 32K_f^2/\kappa^2\}$ is $\tilde{d}$-small for $(P_t^\varepsilon)_{t\ge 0}$, meaning that there exists a constant $\zeta$ such that

$$\widetilde{\mathcal{W}}_{\tilde{d}}(\delta_{(m,a)}P_t^\varepsilon, \delta_{(n,b)}P_t^\varepsilon) \le 1-\zeta,$$

for all $(m,a),(n,b)\in\mathscr{V}$. Let $(\theta_t^{(m,a)})_{t\geq 0},\ (\theta_t^{(n,b)})_{t\geq 0}$ solve the following equations:

$$\begin{aligned}\theta_t^{(m,a)} &= m-\int_0^t \nabla_\theta f(\theta_t^{(m,a)},\tilde{V}^a_{s/\varepsilon})\mathrm{d}s,\\ \theta_t^{(n,b)} &= n-\int_0^t \nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^b_{s/\varepsilon})\mathrm{d}s,\end{aligned}$$

where $\tilde{V}_t^a=V_t^a$ for $t\leq T^a$ and $\tilde{V}_t^a=V_t^\pi$ for $t>T^a$; $\tilde{V}_t^b=V_t^b$ for $t\leq T^b$ and $\tilde{V}_t^b=V_t^\pi$ for $t>T^b$. By Itô's formula, Assumption 3, and the $\varepsilon$-Young inequality,

$$\begin{aligned}&\mathrm{d}\left\|\theta_t^{(m,a)}-\theta_t^{(n,b)}\right\|^2/\mathrm{d}t\\ =&-2\left\langle\theta_t^{(m,a)}-\theta_t^{(n,b)},\nabla_\theta f(\theta_t^{(m,a)},\tilde{V}^a_{t/\varepsilon})-\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^b_{t/\varepsilon})\right\rangle\\ =&-2\left\langle\theta_t^{(m,a)}-\theta_t^{(n,b)},\nabla_\theta f(\theta_t^{(m,a)},\tilde{V}^a_{t/\varepsilon})-\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^a_{t/\varepsilon})\right\rangle\\ &-2\left\langle\theta_t^{(m,a)}-\theta_t^{(n,b)},\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^a_{t/\varepsilon})-\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^b_{t/\varepsilon})\right\rangle\\ \leq&-2\kappa\left\|\theta_t^{(m,a)}-\theta_t^{(n,b)}\right\|^2+2\left|\left\langle\theta_t^{(m,a)}-\theta_t^{(n,b)},\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^a_{t/\varepsilon})-\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^b_{t/\varepsilon})\right\rangle\right|\\ \leq&-\kappa\left\|\theta_t^{(m,a)}-\theta_t^{(n,b)}\right\|^2+\frac{4}{\kappa}\left\|\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^a_{t/\varepsilon})-\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^b_{t/\varepsilon})\right\|^2.\end{aligned}$$

Multiplying $\exp(\kappa t)$ on both sides, we obtain

$$\mathrm{d}\Big(\exp(\kappa t)\left\|\theta_t^{(m,a)}-\theta_t^{(n,b)}\right\|^2\Big)/\mathrm{d}t\leq\frac{4\exp(\kappa t)}{\kappa}\left\|\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^a_{t/\varepsilon})-\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}^b_{t/\varepsilon})\right\|^2,$$

that is

$$\begin{aligned}\exp(\kappa t)\left\|\theta_t^{(m,a)}-\theta_t^{(n,b)}\right\|^2\leq&\|m-n\|^2\\ &+\frac{4}{\kappa}\int_0^t\exp(\kappa s)\left\|\nabla_\theta f(\theta_s^{(n,b)},\tilde{V}^a_{t/\varepsilon})-\nabla_\theta f(\theta_s^{(n,b)},\tilde{V}^b_{t/\varepsilon})\right\|^2\mathrm{d}s.\end{aligned}$$

Notice that $\tilde{V}_t^a = \tilde{V}_t^b$ if $t > T^a \vee T^b$. Hence, we have

$$\begin{aligned}
&\exp(\kappa t) \left\| \theta_t^{(m,a)} - \theta_t^{(n,b)} \right\|^2 \\
&\quad \leq \|m-n\|^2 + \frac{4}{\kappa} \int_0^{t \wedge \varepsilon(T^a \vee T^b)} \exp(\kappa s) \left\| \nabla_\theta f(\theta_s^{(n,b)}, \tilde{V}_{t/\varepsilon}^a) - \nabla_\theta f(\theta_s^{(n,b)}, \tilde{V}_{t/\varepsilon}^b) \right\|^2 \mathrm{d}s.
\end{aligned}$$

By Lemma 18, we have, for $(m,a),(n,b) \in \mathscr{V} \times S$,

$$\left\| \nabla_\theta f(\theta_t^{(m,a)}, \tilde{V}_{t/\varepsilon}^a) \right\| \text{ and } \left\| \nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^b) \right\|$$

are bounded by some constant $C_f$. This implies

$$\left\| \theta_t^{(m,a)} - \theta_t^{(n,b)} \right\|^2 \leq \exp(-\kappa t) \|m-n\|^2 + 4C_f \exp(-\kappa t) \exp(\kappa \varepsilon (T^a \vee T^b)). \tag{3.21}$$

Recall $0 < \varepsilon \leq 1 \wedge \frac{\delta}{2\kappa}$. By (iii), Assumption 1,

$$\begin{aligned}
\tilde{\mathbb{E}}[\exp(\kappa\varepsilon(T^a \vee T^b))] =& \kappa\varepsilon \int_0^\infty \exp(\kappa\varepsilon x) \tilde{\mathbb{P}}(T^a \vee T^b \geq x) \mathrm{d}x \\
\leq& C\kappa\varepsilon \int_0^\infty \exp(\kappa\varepsilon x) \exp(-\delta x) \mathrm{d}x \\
\leq& \frac{C\delta}{2} \int_0^\infty \exp\left(-\frac{\delta x}{2}\right) \mathrm{d}x = C.
\end{aligned}$$

Therefore,

$$\tilde{\mathbb{E}}\left[ \left\| \theta_t^{(m,a)} - \theta_t^{(n,b)} \right\|^2 \right] \leq \exp(-\kappa t) \|m-n\|^2 + C_f \exp(-\kappa t).$$

Moreover,

$$\tilde{\mathbb{P}}(\tilde{V}_{t/\varepsilon}^a \neq \tilde{V}_{t/\varepsilon}^b) = \tilde{\mathbb{P}}(T^a \vee T^b \geq t/\varepsilon) \leq C \exp(-\delta t/\varepsilon) \leq C \exp(-2\kappa t).$$

Hence for any $(m,a),(n,b)\in\mathscr{V}\times S$, by taking $t\geq\frac{1}{\kappa}[\log(8C_f)+\log(512K_f^2/\kappa^2)+\frac{1}{2}\log(8C)]$, we have

$$\widetilde{\mathcal{W}}_{\tilde{d}}^2(H_t^\varepsilon(\cdot|m,a),H_t^\varepsilon(\cdot|n,b))\leq\tilde{\mathbb{P}}(\tilde{V}_{t/\varepsilon}^a\neq\tilde{V}_{t/\varepsilon}^b)+\tilde{\mathbb{E}}\Big[\Big\|\theta_t^{(m,a)}-\theta_t^{(n,b)}\Big\|^2\Big]\leq\frac{1}{2}.$$

We have verified all three conditions from [31, Theorem 3.7] and hence conclude the existence and uniqueness of the invariant measure of $(\theta_t^\varepsilon,V_{t/\varepsilon})_{t\geq0}$ denoted as $\Pi^\varepsilon$.

Next, we are going to prove (3.18) and (3.19). Define $\Theta^\varepsilon$ such that $(\Theta^\varepsilon,\tilde{V}_0^\pi)\sim\Pi^\varepsilon$ in $\tilde{\mathbb{P}}$. Let $\theta_t^{\Pi^\varepsilon}$ and $\theta_t^{(m,\pi)}$ solve following equations

$$\begin{aligned}\theta_t^{\Pi^\varepsilon}&=\Theta^\varepsilon-\int_0^t\nabla_\theta f(\theta_t^{\Pi^\varepsilon},\tilde{V}_{t/\varepsilon}^\pi)\mathrm{d}s,\\ \theta_t^{(m,\pi)}&=m-\int_0^t\nabla_\theta f(\theta_t^{(m,\pi)},\tilde{V}_{t/\varepsilon}^\pi)\mathrm{d}s.\end{aligned}$$

Recall that from (3.21) and (3.20), we have

$$\begin{aligned}\Big\|\theta_t^{(m,a)}-\theta_t^{\Pi^\varepsilon}\Big\|^2&\leq\exp(-\kappa t)\left\|m-\Theta^\varepsilon\right\|^2+4C_f\exp(-\kappa t)\exp(\kappa\varepsilon T^a),\\ \Big\|\theta_t^{(m,\pi)}-\theta_t^{\Pi^\varepsilon}\Big\|^2&\leq\exp(-\kappa t)\left\|m-\Theta^\varepsilon\right\|^2.\end{aligned}$$

Therefore, by (iii), Assumption 1 and since $0<\varepsilon\leq1\wedge(\delta/2\kappa)$, we have

$$\begin{aligned}\widetilde{\mathcal{W}}_{\tilde{d}}^2(H_t^\varepsilon(\cdot|m,a),\Pi^\varepsilon)\leq&\tilde{\mathbb{P}}(\tilde{V}_{t/\varepsilon}^a\neq\tilde{V}_{t/\varepsilon}^\pi)+\tilde{\mathbb{E}}\Big[\Big\|\theta_t^{(m,a)}-\theta_t^{\Pi^\varepsilon}\Big\|^2\Big]\\ \leq&\tilde{\mathbb{P}}(T^a\geq\frac{t}{\varepsilon})+\exp(-\kappa t)\tilde{\mathbb{E}}[\|m-\Theta^\varepsilon\|^2]+4C_f\exp(-\kappa t)\tilde{\mathbb{E}}[\exp(\kappa\varepsilon T^a)]\\ \leq&C_f\exp(-\kappa t)\int_{\mathbb{R}^K}(1+\|x-m\|^2)\Pi^\varepsilon(\mathrm{d}x,S),\end{aligned}$$

$$\begin{aligned}\widetilde{\mathcal{W}}_{\tilde{d}}^2(H_t^\varepsilon(\cdot|m,\pi),\Pi^\varepsilon) \leq& \tilde{\mathbb{E}}\Big[\left\|\theta_t^{(m,\pi)} - \theta_t^{\Pi^\varepsilon}\right\|^2\Big] \\ \leq& \exp(-\kappa t)\tilde{\mathbb{E}}[\|m-\Theta^\varepsilon\|^2] \\ \leq& C_f \exp(-\kappa t)\int_{\mathbb{R}^K}\|x-m\|^2\,\Pi^\varepsilon(\mathrm{d}x,S).\end{aligned}$$

□

In the following corollaries, we study the integrals on the right-hand side of the inequalities in Theorem 4. Moreover, we show that the result above immediately implies not only geometric ergodicity of the coupled process $(\theta_t^\varepsilon, V_{t/\varepsilon})_{t\geq 0}$, but also of its marginal, the stochastic gradient process $(\theta_t^\varepsilon)_{t\geq 0}$.

**Corollary 3.** *Under the same assumptions as Theorem 4, there exists a constant $C_{f,m}$ that depends only on $f$ and the initial value $m = \theta_0^\varepsilon$, such that*

$$\widetilde{\mathcal{W}}_{\tilde{d}}(H_t^\varepsilon(\cdot|m,a),\Pi^\varepsilon) \leq C_{f,m}\exp\left(-\frac{\kappa t}{2}\right), \tag{3.22}$$

$$\widetilde{\mathcal{W}}_{\tilde{d}}(H_t^\varepsilon(\cdot|m,\pi),\Pi^\varepsilon) \leq C_{f,m}\exp\left(-\frac{\kappa t}{2}\right), \tag{3.23}$$

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|m,a),\Pi^\varepsilon(\cdot,S)) \leq C_{f,m}\exp\left(-\frac{\kappa t}{2}\right), \tag{3.24}$$

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|m,\pi),\Pi^\varepsilon(\cdot,S)) \leq C_{f,m}\exp\left(-\frac{\kappa t}{2}\right). \tag{3.25}$$

*Proof.* By Lemma 18,

$$\int_{\|x\|^2\geq\frac{8K^2}{\kappa^2}+\|m\|^2+1}\Pi^\varepsilon(\mathrm{d}x,S) = \lim_{t\to\infty}\tilde{\mathbb{P}}(\left\|\theta_t^{(m,a)}\right\| \geq \frac{8K_f^2}{\kappa^2}+\|m\|^2+1) = 0.$$

Let $C_{f,m} = C_f(2\|m\| + \frac{4K_f}{\kappa} + 2)$. From (3.18), we have

$$\begin{aligned}\tilde{\mathcal{W}}_{\tilde{d}}(H_t^\varepsilon(\cdot|m,a),\Pi^\varepsilon) \leq & C_f \exp\left(\frac{-\kappa t}{2}\right)\Big(\int_{\mathbb{R}^K}(1+\|x-m\|^2)\Pi^\varepsilon(\mathrm{d}x,S)\Big)^{1/2} \\ = & C_f \exp\left(\frac{-\kappa t}{2}\right)\Big(\int_{\|x\|^2\leq\frac{8K_f^2}{\kappa^2}+\|m\|^2+1}(1+\|x-m\|^2)\Pi^\varepsilon(\mathrm{d}x,S)\Big)^{1/2} \\ \leq & C_{f,m}\exp\left(\frac{-\kappa t}{2}\right).\end{aligned}$$

(3.23) can be derived similarly from (3.19). Moreover, notice that

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|m,a),\Pi^\varepsilon(\cdot,S)) \leq \tilde{\mathbb{E}}\Big[\Big\|\theta_t^{(m,a)} - \theta_t^{\Pi^\varepsilon}\Big\|\Big], \tag{3.26}$$

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|m,\pi),\Pi^\varepsilon(\cdot,S)) \leq \tilde{\mathbb{E}}\Big[\Big\|\theta_t^{(m,\pi)} - \theta_t^{\Pi^\varepsilon}\Big\|\Big]. \tag{3.27}$$

(3.24) and (3.25) can be derived similarly from (3.26) and (3.27). □

Combining (3.24) and (3.25), we immediately have:

**Corollary 4.** *Under the same assumptions as Theorem 4,*

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|m,a), C_t^\varepsilon(\cdot|m,\pi)) \leq C_{f,m}\exp\left(-\frac{\kappa t}{2}\right).$$

So far, we have shown that the stochastic gradient process $(\theta_t^\varepsilon)_{t\geq 0}$ converges to a unique stationary measure $\Pi^\varepsilon(\cdot,S)$. It is often not possible to determine this stationary measure. However, we can comment on its asymptotic behavior as $\varepsilon \to 0$. Indeed, we will show that $\Pi^\varepsilon(\cdot,S)$ concentrates around the minimizer $\theta_*$ of the full target function.

**Proposition 3.** *Under Assumption 3, the measure* $\Pi^\varepsilon(\cdot,S)$ *on* $(\mathbb{R}^K,\mathcal{B}(\mathbb{R}^K))$ *approximates* $\delta(\cdot-\theta_*)$. *In other words, we have*

$$\mathcal{W}_d(\Pi^\varepsilon(\cdot,S),\delta(\cdot-\theta_*)) \leq \rho(\varepsilon)$$

*where* $\rho : (0,1) \to [0,1]$ *and* $\lim_{\varepsilon \to 0} \rho(\varepsilon) = 0$.

*Proof.* By the triangle inequality,

$$\begin{aligned}
&\mathcal{W}_d(\Pi^\varepsilon(\cdot, S), \delta(\cdot - \theta_*)) \\
\leq &\mathcal{W}_d(\Pi^\varepsilon(\cdot, S), C_t^\varepsilon(\cdot|m,\pi)) + \mathcal{W}_d(C_t^\varepsilon(\cdot|m,\pi), \delta(\cdot - \zeta_t)) + \mathcal{W}_d(\delta(\cdot - \zeta_t), \delta(\cdot - \theta_*)).
\end{aligned}$$

Let $\theta_0 = \theta_0^\varepsilon = \theta_*$. Then by Lemma 17, we have the last term

$$\mathcal{W}_d(\delta(\cdot - \zeta_t), \delta(\cdot - \theta_*)) \leq \|\theta_* - \theta_*\| \exp(-\kappa t) = 0.$$

By (3.22) and Corollary 2, for any $t \geq 0$,

$$\mathcal{W}_d(\Pi^\varepsilon(\cdot, S), C_t^\varepsilon(\cdot|m,\pi)) + \mathcal{W}_d(C_t^\varepsilon(\cdot|m,\pi), \delta(\cdot - \zeta_t)) \leq C_{f,\theta_*} \exp\left(-\frac{\kappa t}{2}\right) + (\exp(t)\alpha(\varepsilon)) \wedge 1.$$

By choosing $t = -\log(1 \wedge \alpha(\varepsilon))/2$, we get

$$\begin{aligned}
\mathcal{W}_d(\Pi^\varepsilon(\cdot, S), \delta(\cdot - \theta_*)) \leq &C_{f,\theta_*} \exp\left(-\frac{\kappa t}{2}\right) + (\exp(t)\alpha(\varepsilon)) \wedge 1 \\
\leq &C_{f,\theta_*} (\alpha(\varepsilon))^{\frac{\kappa}{4}} + (\alpha(\varepsilon))^{1/2}.
\end{aligned}$$

Taking $\rho(\varepsilon) := \left(C_{f,\theta_*} (\alpha(\varepsilon))^{\frac{\kappa}{4}} + (\alpha(\varepsilon))^{1/2}\right) \wedge 1$ completes the proof. □

## 3.4 Stochastic gradient processes with decreasing learning rate

Constant learning rates are popular in some practical situations, but the associated stochastic gradient process usually does not converge to the minimizer of $\Phi$. This is also true for the discrete-time stochastic gradient descent algorithm. However, SGD can converge to the minimizer if the learning rate is decreased over time. In the following, we discuss a

decreasing learning version of the stochastic gradient process and show that this dynamical system indeed converges to the minimizer of $\Phi$.

As discussed in Section 3.1.2, we obtain the stochastic gradient process with decreasing learning rate by non-linearly rescaling the time in the constant-learning-rate index process $(V_t)_{t\geq 0}$. Indeed, we choose a function $\beta : [0,\infty) \to [0,\infty)$ and then define the decreasing learning rate index process by $(V_{\beta(t)})_{t\geq 0}$. We have discussed an intuitive way to construct a rescaling function $\beta$ also in Section 3.1.2.

In the following, we define $\beta$ through an integral $\beta(t) = \int_0^t \mu(s)\mathrm{d}s$, $t \geq 0$. We commence this section with necessary growth conditions on $\mu$ which allow us to then give the formal definition of the stochastic gradient process with decreasing learning rate. Then, we study the longtime behavior of this process.

**Assumption 4.** *Let $\mu : [0,\infty) \to (0,\infty)$ be a non-decreasing continuously differentiable function with $\lim_{t\to\infty} \mu(t) = \infty$ and*

$$\lim_{t\to\infty} \frac{\mu'(t)t}{\mu(t)} = 0.$$

Assumption 4 implies that $\mu$ goes to infinity, but at a very slow pace. Indeed it says that $\lim_{t\to\infty} \mu(t)/t^\gamma = 0$, $\gamma > 0$, that is $\mu$ grows slower than any polynomial.

**Definition 2.** *The* stochastic gradient process with decreasing learning rate (SGPD) *is a solution of the following stochastic differential equation,*

$$\begin{cases} \mathrm{d}\xi_t = -\nabla_\xi f(\xi_t, V_{\beta(t)})\mathrm{d}t, \\ \xi_0 = \theta_0, \end{cases} \tag{3.28}$$

*where $f$ satisfies Assumption 2, $(V_t)_{t\geq 0}$ is a Feller process that satisfies Assumption 1, and $\beta(t) = \int_0^t \mu(s)\mathrm{d}s$ with $\mu$ satisfying Assumption 4.*

To see that $(\xi_t)_{t\geq 0}$ is well-defined, consider the following: $t \mapsto \beta(t)$ is an increasing continuous function. Thus, $(V_{\beta(t)})_{t\geq 0}$ is càdlàg and Feller with respect to $(\mathcal{F}_{\beta(t)})_{t\geq 0}$. We

then obtain well-definedness of $(\xi_t)_{t\geq 0}$ by replacing $(V_{t/\varepsilon})_{t\geq 0}$ by $(V_{\beta(t)})_{t\geq 0}$ in the proof of Proposition 2.

We now move on to studying the longtime behavior of the SGPD $(\xi_t)_{t\geq 0}$. In a first technical result, we establish a connection between SGPD $(\xi_t)_{t\geq 0}$ and a time-rescaled version of SGPC $(\theta^\varepsilon_t)_{t\geq 0}$. To this end, note that $\dot\beta(t) = \mu(t) > 0$, $\beta(t)$ is strictly increasing. Hence, the inverse of $\beta(t)$ exists and

$$\beta^{-1}(t) = \int_0^t \frac{1}{\mu(\beta^{-1}(s))}\mathrm{d}s.$$

This gives us the following inequality.

**Proposition 4.** *For any $0 < \varepsilon < 1$,*

$$\left\|\xi_t - \theta^\varepsilon_{\varepsilon\beta(t)}\right\|^2 \leq C_{f,\theta_0,\mu}\left[\frac{\exp(-2\varepsilon\kappa(\beta(t)-\beta(\frac{t}{2})))}{\varepsilon} + \frac{1}{\varepsilon}\Big(\left|\frac{1}{\mu(t)} - \varepsilon\right| + \left|\frac{1}{\mu(\frac{t}{2})} - \varepsilon\right|\Big)\right]$$

*almost surely, where the constant $C_{f,\theta_0,\mu}$ depends only on $f$, the initial data $\theta_0$, and $\mu$.*

*Proof.* From (3.28) and (3.8), we have

$$\begin{aligned}\xi_t =& \theta_0 - \int_0^{\beta(t)} \nabla_\xi f(\xi_{\beta^{-1}(s)}, V_s)\mathrm{d}\beta^{-1}(s),\\ \theta^\varepsilon_t =& \theta_0 - \varepsilon\int_0^{\frac{t}{\varepsilon}} \nabla_\theta f(\theta^\varepsilon_{\varepsilon s}, V_s)\mathrm{d}s.\end{aligned}$$

Let $b_t := \mathrm{d}\beta^{-1}(t)/\mathrm{d}t = 1/\mu(\beta^{-1}(t)) > 0$, we have

$$\begin{aligned}\xi_{\beta^{-1}(t)} =& \theta_0 - \int_0^t \nabla_\xi f(\xi_{\beta^{-1}(s)}, V_s)b_s\mathrm{d}s\\ \theta^\varepsilon_{\varepsilon t} =& \theta_0 - \varepsilon\int_0^t \nabla_\theta f(\theta^\varepsilon_{\varepsilon s}, V_s)\mathrm{d}s\end{aligned}$$

Therefore, by Itô's formula and Assumption 3,

$$\begin{aligned}\mathrm{d}\left\|\theta^{\varepsilon}_{\varepsilon t}-\xi_{\beta^{-1}(t)}\right\|^2/\mathrm{d}t =& -2\left\langle\theta^{\varepsilon}_{\varepsilon t}-\xi_{\beta^{-1}(t)},\varepsilon\nabla_\theta f(\theta^{\varepsilon}_{\varepsilon t},V_t)-\varepsilon\nabla_\xi f(\xi_{\beta^{-1}(t)},V_t)\right\rangle\\ &-2(\varepsilon-b_t)\left\langle\theta^{\varepsilon}_{\varepsilon t}-\xi_{\beta^{-1}(t)},\nabla_\xi f(\xi_{\beta^{-1}(t)},V_t)\right\rangle\\ \leq& -2\varepsilon\kappa\left\|\theta^{\varepsilon}_{\varepsilon t}-\xi_{\beta^{-1}(t)}\right\|^2+C_{f,\theta_0}\left|b_t-\varepsilon\right|,\end{aligned}$$

where the last step follows from the boundedness of $\theta^{\varepsilon}_{\varepsilon t}$, $\xi_{\beta^{-1}(t)}$, and $\nabla_\xi f(\xi_{\beta^{-1}(t)},V_t)$. $\xi_{\beta^{-1}(t)}$ is bounded can be showed similarly to Lemma 18. Multiplying $\exp(2\varepsilon\kappa t)$ on both sides, we obtain

$$\mathrm{d}\Big(\exp(2\varepsilon\kappa t)\left\|\theta^{\varepsilon}_{\varepsilon t}-\xi_{\beta^{-1}(t)}\right\|^2\Big)/\mathrm{d}t\leq C_{f,\theta_0}\left|b_t-\varepsilon\right|\exp(2\varepsilon\kappa t),$$

which implies

$$\left\|\theta^{\varepsilon}_{\varepsilon t}-\xi_{\beta^{-1}(t)}\right\|^2\leq C_{f,\theta_0}\exp(-2\varepsilon\kappa t)\int_0^t\left|b_s-\varepsilon\right|\exp(2\varepsilon\kappa s)\mathrm{d}s.$$

Notice that $b_s$ is bounded and non-increasing, hence we have

$$\begin{aligned}\left\|\xi_t-\theta^{\varepsilon}_{\varepsilon\beta(t)}\right\|^2\leq& C_{f,\theta_0}\exp(-2\varepsilon\kappa\beta(t))\int_0^{\beta(t)}\left|b_s-\varepsilon\right|\exp(2\varepsilon\kappa s)\mathrm{d}s\\ =& C_{f,\theta_0}\exp(-2\varepsilon\kappa\beta(t))\Big(\int_0^{\beta(\frac{t}{2})}+\int_{\beta(\frac{t}{2})}^{\beta(t)}\Big)\left|b_s-\varepsilon\right|\exp(2\varepsilon\kappa s)\mathrm{d}s\\ \leq& C_{f,\theta_0,\mu}\frac{\exp(-2\varepsilon\kappa(\beta(t)-\beta(\frac{t}{2})))}{\varepsilon}\\ &+C_{f,\theta_0}\exp(-2\varepsilon\kappa\beta(t))\int_{\beta(\frac{t}{2})}^{\beta(t)}\left|b_s-\varepsilon\right|\exp(2\varepsilon\kappa s)\mathrm{d}s\\ \leq& C_{f,\theta_0,\mu}\left[\frac{\exp(-2\varepsilon\kappa(\beta(t)-\beta(\frac{t}{2})))}{\varepsilon}+\frac{1}{\varepsilon}\Big(\left|\frac{1}{\mu(t)}-\varepsilon\right|+\left|\frac{1}{\mu(\frac{t}{2})}-\varepsilon\right|\Big)\right].\end{aligned}$$

□

Now, we get to the main result of this section, where we show the convergence of $(\xi_t)_{t\geq 0}$ to the minimizer $\theta_*$ of $\Phi$. In the following, we denote

$$D_t(B|\theta_0, a) := \mathbb{P}_a(\xi_t \in B|\xi_0 = \theta_0),$$
$$D_t(B|\theta_0, \pi) := \mathbb{P}_\pi(\xi_t \in B|\xi_0 = \theta_0) \qquad (B \in \mathcal{B}(\mathbb{R}^K), \theta_0 \in \mathbb{R}^K),$$

where $a \in S$ and $\pi$ is the invariant measure of $(V_t)_{t\geq 0}$, respectively.

**Theorem 5.** *Under Assumption 3, given $\theta_0 \in \mathbb{R}^K$ and $a \in S$, there exists $T > 0$ such that for any $t > T$,*

$$\mathcal{W}_d(D_t(\cdot|\theta_0, \pi), \delta(\cdot - \theta_*)) \leq C_{f,\theta_0,\mu} A(t), \tag{3.29}$$

$$\mathcal{W}_d(D_t(\cdot|\theta_0, a), \delta(\cdot - \theta_*)) \leq C_{f,\theta_0,\mu} A(t), \tag{3.30}$$

*where*

$$A(t) := \exp\left(\frac{-\kappa t}{8}\right) + \left[\frac{\mu(t) - \mu(\frac{t}{2})}{\mu(t)}\right]^{1/2} + \rho\left(\frac{1}{\mu(\frac{t}{2})}\right)$$

*and* $\lim_{t\to\infty} A(t) = 0$.

*Proof.* To prove (3.29), by the triangle inequality,

$$\begin{aligned}&\mathcal{W}_d(D_t(\cdot|\theta_0, \pi), \delta(\cdot - \theta_*))\\ \leq&\mathcal{W}_d(D_t(\cdot|\theta_0, \pi), C^\varepsilon_{\varepsilon\beta(t)}(\cdot|\theta_0, \pi)) + \mathcal{W}_d(C^\varepsilon_{\varepsilon\beta(t)}(\cdot|\theta_0, \pi), \Pi^\varepsilon(\cdot, S)) + \mathcal{W}_d(\Pi^\varepsilon(\cdot, S), \delta(\cdot - \theta_*)).\end{aligned}$$

For the last two terms, by (3.25) and Proposition 3,

$$\mathcal{W}_d(C^\varepsilon_{\varepsilon\beta(t)}(\cdot|\theta_0, \pi), \Pi^\varepsilon(\cdot, S)) + \mathcal{W}_d(\Pi^\varepsilon(\cdot, S), \delta(\cdot - \theta_*)) \leq C_{f,m} \exp(-\kappa\varepsilon\beta(t)/2) + \rho(\varepsilon).$$

For the first term, by Proposition 4,

$$\begin{aligned}&\mathcal{W}_d(D_t(\cdot|\theta_0,\pi), C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,\pi))\\ &\quad\le C_{f,\theta_0,\mu}\left[\frac{\exp(-2\varepsilon\kappa(\beta(t)-\beta(\frac{t}{2})))}{\varepsilon}+\frac{1}{\varepsilon}\Big(\Big|\frac{1}{\mu(t)}-\varepsilon\Big|+\Big|\frac{1}{\mu(\frac{t}{2})}-\varepsilon\Big|\Big)\right]^{1/2}.\end{aligned}$$

Since $\lim_{t\to\infty}\mu(t)=\infty$, there exists $T>0$ such that $1/\mu(\frac{T}{2})<\frac{\delta}{2\kappa}$. Let $\varepsilon=1/\mu(\frac{t}{2})$, $t>T$, we have

$$\exp\left(\frac{-\kappa\varepsilon\beta(t)}{2}\right)=\exp\left(\frac{-\kappa\int_0^t\mu(s)\mathrm{d}s}{2\mu(\frac{t}{2})}\right)\le\exp\left(\frac{-\kappa t}{8}\right)$$

and

$$\begin{aligned}\frac{\exp\left(-2\varepsilon\kappa(\beta(t)-\beta(\frac{t}{2}))\right)}{\varepsilon}&=\mu\left(\frac{t}{2}\right)\exp\left(\frac{-\kappa\int_{\frac{t}{2}}^t\mu(s)\mathrm{d}s}{\mu(\frac{t}{2})}\right)\\&\le\mu\left(\frac{t}{2}\right)\exp\left(\frac{-\kappa t}{2}\right)\le C\exp\left(\frac{-\kappa t}{8}\right).\end{aligned}$$

Therefore,

$$\mathcal{W}_d(D_t(\cdot|\theta_0,\pi),\delta(\cdot-\theta_*))\le C_{f,\theta_0,\mu}\Big[\exp\left(\frac{-\kappa t}{8}\right)+\left(\frac{\mu(t)-\mu(\frac{t}{2})}{\mu(t)}\right)^{1/2}+\rho\left(\frac{1}{\mu\left(\frac{t}{2}\right)}\right)\Big]$$

From Assumption 4, by the mean value theorem,

$$\frac{\mu(t)-\mu(\frac{t}{2})}{\mu(t)}=\frac{t\mu'(\tau_t)}{2\mu(t)}=\frac{\tau_t\mu'(\tau_t)}{\mu(\tau_t)}\frac{t}{2\tau_t}\frac{\mu(\tau_t)}{\mu(t)}\le\frac{\tau_t\mu'(\tau_t)}{\mu(\tau_t)}\to 0$$

where $\tau_t\in[\frac{t}{2},t]$. Thus, (3.29) is obtained by taking

$$A(t):=\exp\left(\frac{-\kappa t}{8}\right)+\left[\frac{\mu(t)-\mu(\frac{t}{2})}{\mu(t)}\right]^{1/2}+\rho\left(\frac{1}{\mu(\frac{t}{2})}\right).$$

To prove (3.30), we have

$$\begin{aligned}\mathcal{W}_d(D_t(\cdot|\theta_0,a),\delta(\cdot-\theta_*)) \le\ & \mathcal{W}_d(D_t(\cdot|\theta_0,a),C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,a))\\ &+\mathcal{W}_d(C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,a),C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,\pi))\\ &+\mathcal{W}_d(C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,\pi),\Pi^{\varepsilon}(\cdot,S))\\ &+\mathcal{W}_d(\Pi^{\varepsilon}(\cdot,S),\delta(\cdot-\theta_*)),\end{aligned}$$

by the triangle inequality. By Corollary 4, we have

$$\mathcal{W}_d(C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,a),C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,\pi)) \le C_{f,m}\exp\left(-\frac{\kappa\varepsilon\beta(t)}{2}\right).$$

Notice that $C_{f,m}\exp(-\frac{\kappa t}{4}) \le C_{f,m}A(t)$ when $\varepsilon = 1/\mu(\frac{t}{2})$. Similar to the proof of (3.29), we have

$$\begin{aligned}\mathcal{W}_d(D_t(\cdot|\theta_0,a),C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,a)) + \mathcal{W}_d(C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,\pi),\Pi^{\varepsilon}(\cdot,S)) + \mathcal{W}_d(\Pi^{\varepsilon}(\cdot,S),\delta(\cdot-\theta_*))&\\ \le C_{f,\theta_0,\mu}A(t),&\end{aligned}$$

which completes the proof. □

Thus, we have shown that the distribution of $(\xi_t)_{t\ge0}$ converges in Wasserstein distance to the Dirac measure concentrated in the minimizer $\theta_*$ of $\Phi$. This result is independent of whether we initialize the index process $(V_{\beta(t)})_{t\ge0}$ with its stationary measure or with any deterministic value.

## 3.5 From continuous dynamics to practical optimization.

So far, we have discussed the stochastic gradient process as a continuous-time coupling of an ODE and a stochastic process. In order to apply the stochastic gradient process in practice, we need to discretize ODE and stochastic process with appropriate time-stepping schemes. That means, for a given increasing sequence $(t(k))_{k=0}^{\infty}$, with $t(0) := 0$ and $\lim_{k\to\infty} t(k) = \infty$, we seek discrete-time stochastic processes $(\widehat{V}_k, \widehat{\theta}_k)_{k=0}^{\infty}$, such that $(\widehat{V}_k, \widehat{\theta}_k)_{k=0}^{\infty} \approx (V_{t(k)}, \theta^{\varepsilon}_{t(k)})_{k=0}^{\infty}$ and analogous discretizations for $(V_{\beta(t)}, \xi_t)_{t\geq 0}$.

In the following, we propose and discuss time stepping strategies and the algorithms arising from them. We discuss the index process and gradient flow separately, which we consider sufficient as the coupling is only one-sided.

### 3.5.1 Discretization of the index process

We have defined the stochastic gradient process for a huge range of potential index processes $(V_t)_{t\geq 0}$. The discretization of such processes has been the topic of several works, see, e.g., [47, 83]. In the following, we focus on one case and refer to those previous works for other settings and details.

Indeed, we study the setting $S := [-1, 1]$ and $\pi := \mathrm{Unif}[-1, 1]$ and discuss the discretization of $(V_t)_{t\geq 0}$ as a Markov pure jump process and as a reflected Brownian motion.

**Markov pure jump process.**

A suitable Markov pure jump process is a piecewise constant càdlàg process $(V_t)_{t\geq 0}$ with Markov transition kernel

$$\mathbb{P}_x(V_t \in \cdot) = \exp(-\lambda t)\delta(\cdot - x) + (1 - \exp(-\lambda t))\mathrm{Unif}[-1, 1] \qquad (t \geq 0),$$

where $\lambda > 0$ is a rate parameter. We can now discretize the process $(V_t)_{t\geq 0}$ just through sampling from this Markov kernel for our discrete time points. We describe this in Algorithm 1.

**Algorithm 1** Discretized Markov pure jump process

initialize $\widehat{V}_0$, $\lambda > 0$, and a sequence of points $(t(k))_{k=0}^{\infty}$
**for** $k = 1, 2, \ldots$ **do**
  sample $U \sim \mathrm{Unif}[0, 1]$
  **if** $U \leq \exp(-\lambda(t(k) - t(k-1)))$ **then**
    $\widehat{V}_k \leftarrow \widehat{V}_{k-1}$ {process stays at its current position}
  **else**
    sample $\widehat{V}_k \sim \mathrm{Unif}[-1, 1]$ {process jumps to a new position}
  **end if**
**end for**
**return** $(\widehat{V}_k)_{k=0}^{\infty}$

**Reflected Brownian motion.**

We have defined the reflected Brownian Motion on a non-empty compact interval through the Skorohod problem in Subsection 3.2.1.

Let $\sigma > 0$ and $(W_t)_{t\geq 0}$ be a standard Brownian motion. Probably the easiest way to sample a reflected Brownian motion is by discretizing the rescaled Brownian motion $(\sigma \cdot W_t)_{t\geq 0}$ using the Euler–Maruyama scheme and projecting back to $S$, whenever the sequence leaves $S$. This scheme has been studied by [99]. We describe the full scheme in Algorithm 2.

[99] shows that this scheme converges at a rather slow rate. As we usually assume that the domain on which we move is rather low-dimensional and the sampling is rather cheap, we can afford small discretization stepsizes $t(k) - t(k-1)$, for $k \in \mathbb{N}$. Thus, the slow rate of convergence is manageable. Other schemes for the discretization of reflected Brownian motions have been discussed by, e.g., [14, 81].

**Algorithm 2** Discretized Reflected Brownian motion on $S$

initialize $\widehat{V}_0$, $\sigma > 0$, a sequence of points $(t(k))_{k=0}^{\infty}$, and the projection operator $\mathrm{proj}_S$ mapping onto $S$
**for** $k = 1, 2, \ldots$ **do**
  $V' \leftarrow V_{k-1} + \sigma\sqrt{t(k) - t(k-1)}\psi, \quad \psi \sim \mathrm{N}(0, 1^2)$ {Euler-Maruyama update}
  **if** $V' \notin S$ **then**
    $\widehat{V}_k \leftarrow \mathrm{proj}_S V'$ {project back}
  **else**
    $\widehat{V}_k \leftarrow V'$ {accept Euler–Maruyama update}
  **end if**
**end for**
**return** $(\widehat{V}_k)_{k=0}^{\infty}$

### 3.5.2 Discretization of the gradient flow

We now briefly discuss the discretization of the gradient flow in the stochastic gradient process. Based on these ideas, we will conduct numerical experiments in Section 3.6.

#### Stochastic gradient descent

In stochastic gradient descent, the gradient flow is discretized with a forward Euler method. This method leads to an accurate discretization of the respective gradient flow if the stepsize/learning rates are sufficiently small. In the presence of rather large stepsizes and stiff vector fields, however, the forward Euler method may be inaccurate and unstable, see, e.g., [104].

#### Stability

Several ideas have been proposed to mitigate this problem. The stochastic proximal point method, for instance, uses the backward Euler method to discretize the gradient flow; see [13]. Unfortunately, such implicit ODE integrators require us to invert a possibly highly non-linear and complex vector field. In convex stochastic optimization this inversion can be replaced by evaluating a proximal operator. For strongly convex optimization, on the other hand, [43] proposes stable explicit methods.

**Efficient optimizers**

Plenty of highly efficient methods for stochastic optimization methods are nowadays available, especially in machine learning. Those have often been proposed without necessarily thinking of the stable and accurate discretization of a gradient flow: such are adaptive methods ([64]), variance reduced methods (e.g., [36]), or momentum methods (e.g., [65] for an overview), which have been shown in multiple works to be highly efficient; partially also in non-convex optimization. We could understand those methods also as certain discretizations of the gradient flow. Thus, we may also consider the combination of a feasible index process $(V_t)_{t\geq 0}$ with the discrete dynamical system in, e.g., the Adam method ([64]).

## 3.6 Applications

We now study two fields of application of the stochastic gradient process for continous data. In the first example, we consider regularized polynomial regression with noisy functional data. In this case, we can easily show that the necessary assumptions for our analysis hold. Thus, we use it to illustrate our analytical results and especially to learn about the implicit regularization that is put in place due to different index proccesses.

In the second example, we study so-called physics-informed neural networks. In these continuous-data machine learning problems, a deep neural network is used to approximate the solution of a partial differential equation. The associated optimization problem is usually non-convex. Our analysis does not hold in this case: We study it to get more insights in the behavior of the stochastic gradient process in state-of-the-art deep learning problems.

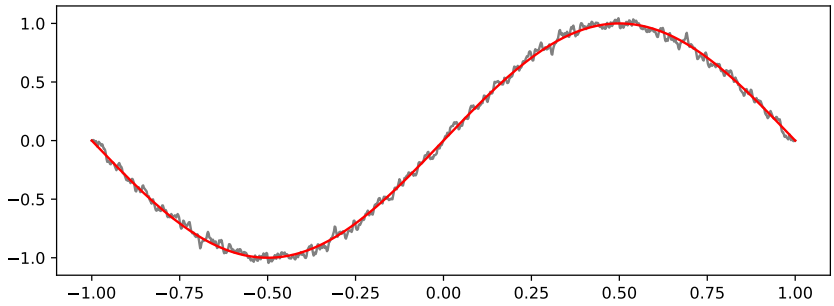


Figure 3.3: True function $\Theta$ (red) and noisy observation $g$ (grey) in the polynomial regression example.

### 3.6.1 Polynomial regression with functional data

We begin with a simple polynomial regression problem with noisy functional data. We observe the function $g : [-1, 1] \to \mathbb{R}$ which is given through

$$g(y) := \Theta(y) + \Xi(y) \qquad (y \in [-1, 1]),$$

where $\Theta : [-1, 1] \to \mathbb{R}$ is a smooth function and $\Xi$ is a Gaussian process with highly oscillating, continuous realizations. We aim at identifying the unknown function $\Theta$ subject to the observational noise $\Xi$. Here, we represent the function $\Theta$ on a basis consisting of a finite number of Legendre polynomials on $[-1, 1]$. We denote this basis of Legendre polynomials by $(\ell_k)_{k=1}^K$. To estimate the prefactors of the polynomials, we minimize the potential

$$\Phi(\theta) := \frac{1}{2} \int_{[-1,1]} \left( g(y) - \sum_{k=1}^{K} \theta_k \ell_k(y) \right)^2 \mathrm{d}y + \frac{\alpha}{2} \|\theta\|_2^2 \qquad (\theta \in X), \tag{3.31}$$

where $\alpha > 0$ is a regularization parameter. This can be understood as a maximum-a-posteriori estimation of the unknown $\theta$ with Gaussian prior under the (misspecified) assumption that the data is perturbed with Gaussian white noise. We employ the following

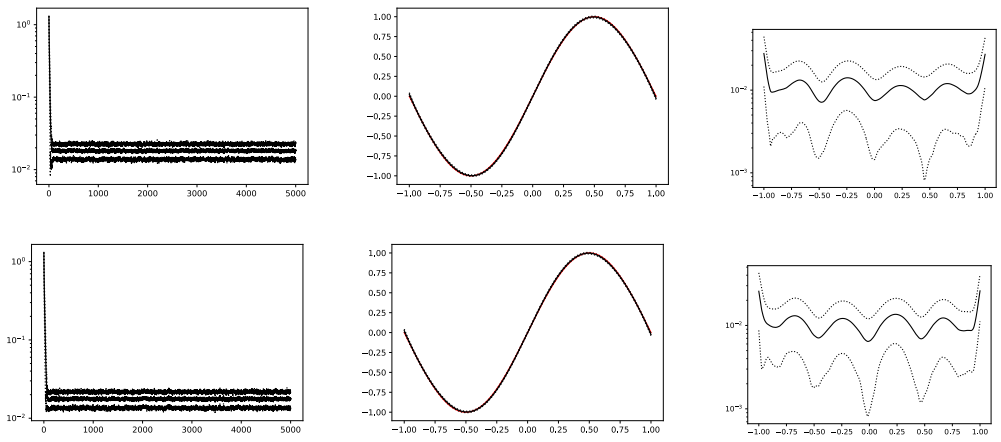

Figure 3.4: Estimation results of the polynomial regression problem using stochastic gradient descent with constant learning rate $\eta = 0.1$ (top row) and a version of stochastic gradient descent that uses the implicit midpoint rule (bottom row). The figures depict the mean over 100 runs (black solid line), mean $\pm$ standard deviation (black dotted line). Left column: trajectory of the rel_err over time; centre column: comparison of $\Theta$ (solid red line) and estimated polynomial; right column: estimation error in terms of abs_err.

associated subsampled potentials:

$$f(\theta, y) := \frac{1}{2}\left(g(y) - \sum_{k=1}^{K} \theta_k \ell_k(y)\right)^2 + \frac{\alpha}{2}\|\theta\|_2^2 \qquad (\theta \in X, y \in [-1,1]). \tag{3.32}$$

Those subsampled potentials satisfy the strong convexity assumption, i.e., Assumption 3.

**Setup**

In particular, we have produced artificial data $g$, by setting $\Theta := \sin(\pi\cdot)$ and choosing

$$\Xi(x) = \sum_{j=1}^{200} \frac{10}{1000 + (\pi j)^{3/2}} \sin(2\pi j(x - 0.5))\Xi_j \qquad (x \in [-1,1])$$

and i.i.d. random variables $\Xi_1, \ldots, \Xi_{200} \sim \mathrm{N}(0, 1^2)$. Note that $\Xi$ is a Gaussian random field given through the truncated Karhunen-Loève expansion of a covariance operator that is related to the Matérn family, see, e.g., [80].

We show $\Theta$ and $g$ in Figure 3.3. For our estimation, we set $\alpha := 10^{-4}$ and use the $K = 9$ Legendre polynomials with degrees $0, \ldots, 8$. We employ the stochastic gradient process with

constant learning rate, using either a reflected diffusion process or a pure Markov jump process for the index process $(V_t)_{t\geq 0}$. We discretize the gradient flow using the implicit midpoint rule: an ODE $z' = q(z), z(0) = z_0$ is then discretized with stepsize $h > 0$ by successively solving the implicit formula

$$z_k = z_{k-1} + \frac{h}{2}q(z_k) + \frac{h}{2}q(z_{k-1}) \qquad (k \in \mathbb{N}).$$

In our experiments, we choose $h = 0.1$. We use Algorithms 1 and 2 to discretize the index processes with constant stepsize $t(\cdot) - t(\cdot - 1) = 10^{-2}$. We perform $J := 100$ repeated runs for each of the considered settings for $N := 5 \cdot 10^4$ time steps and thus, obtain a family of trajectories $(\theta^{(j,n)})_{n=1,\dots,N,j=1,\dots,J}$. In each case, we choose the initial values $V(0) := 0$ and the $\theta^{(j,0)} := (0.5, \dots, 0.5)$.

We study the distance of the estimated polynomial to the true function $\Theta$ by the relative error:

$$\text{rel_err}_{n,j} := \frac{\sum_{l=1}^{L}\left(\Theta(x_l) - \sum_{k=1}^{K}\theta_k^{(j,n)}\ell_k(x_l)\right)^2}{\sum_{l'=1}^{L}\Theta(x_{l'})^2},$$

for trajectory $j \in \{1, \dots, J\}$ and time step $n \in \{1, \dots, N\}$. Here $(x_l)_{l=1}^{L}$ are $L := 10^3$ equispaced points in $[-1, 1]$. Moreover, we compare the estimated polynomial to the true function $\Theta$ by

$$\text{abs_err}_{j,x} := \left|\Theta(x) - \sum_{k=1}^{K}\theta_k^{(j,N)}\ell_k(x)\right|$$

for trajectory $j \in \{1, \dots, J\}$ at position $x \in [-1, 1]$. In each case, we study mean and standard deviation (StD) computed over the 100 runs.

**Results and discussion**

For the polynomial regression problem we now study:

- stochastic gradient descent, as given in (3.2), with constant learning rate $\eta_{(\cdot)} = h = 0.1$ (Figure 3.4 top row),

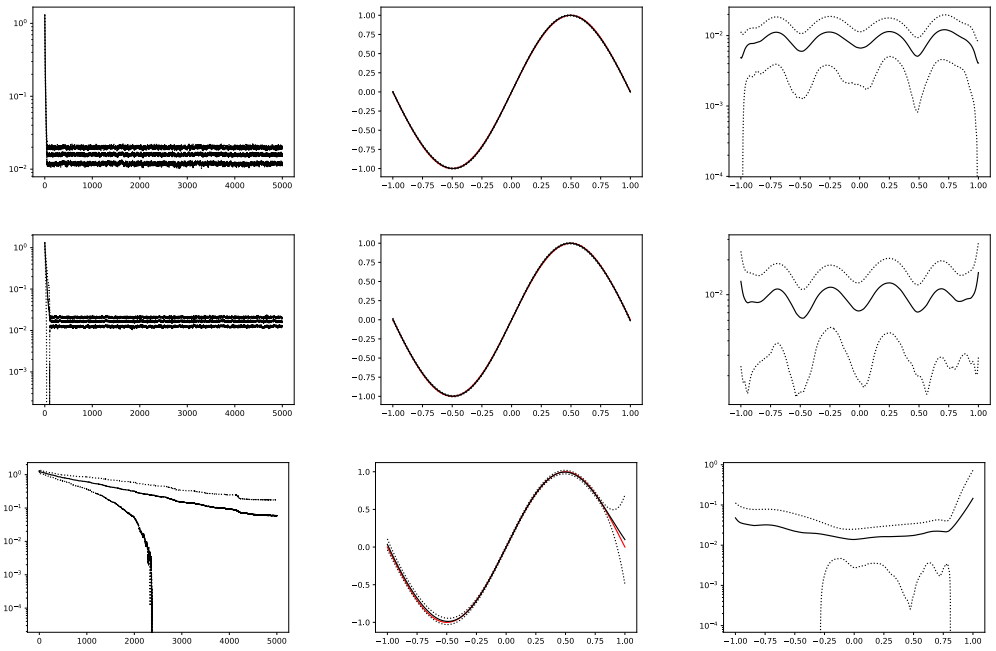

Figure 3.5: Estimation results of the polynomial regression problem using the stochastic gradient process with reflected Brownian motion process with $\sigma = 5$ (top row), $\sigma = 0.5$ (centre row), and $\sigma = 0.05$ (bottom row). The figures depict the mean over 100 runs (black solid line), mean $\pm$ standard deviation (black dotted line). Left column: trajectory of the rel_err over time; centre column: comparison of $\Theta$ (solid red line) and estimated polynomial; right column: estimation error in terms of abs_err.

- stochastic gradient descent algorithm, for which the forward Euler update is replaced by an implicit midpoint rule update, with constant learning rate $\eta_{(\cdot)} = h = 0.1$ (Figure 3.4 bottom row),

- the stochastic gradient process with reflected Brownian motion as an index process with standard deviation $\sigma \in \{5, 0.5, 0.05\}$ (Figure 3.5),

- the stochastic gradient process with Markov pure jump process as an index process with rate parameter $\lambda \in \{10, 1, 0.1, 0.01\}$ (Figure 3.6), and

- the quality of discretisation of the index processes $(V_t)_{t \geq 0}$ and their convergence (Figure 3.7).

In addition to those plots, we give means and standard deviations of the relative errors at the terminal state of the iterations in Table 3.1. To compare the convergence behavior of the different methods, we plot the rel_err within the first 2000 discrete time steps in Figure 3.8.

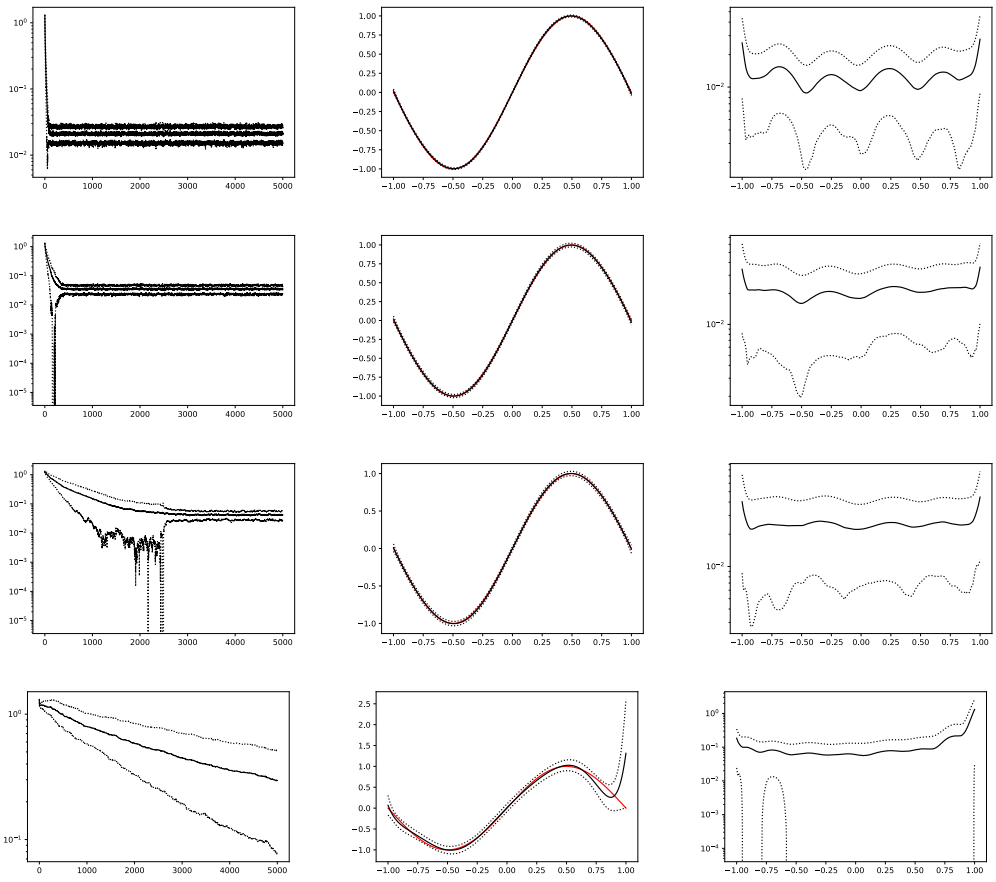

Figure 3.6: Estimation results of the polynomial regression problem using the stochastic gradient process with pure jump index process with $\lambda = 10$ (first row), $\lambda = 1$ (second row), $\lambda = 0.1$ (third row), and $\lambda = 0.01$ (fourth row). The figures depict the mean over 100 runs (black solid line), mean $\pm$ standard deviation (black dotted line). Left column: trajectory of the rel_err over time; centre column: comparison of $\Theta$ (solid red line) and estimated polynomial; right column: estimation error in terms of abs_err.

We learn several things from these results. Unsurprisingly, the index processes with a strong autocorrelation ($\lambda = 0.01, \sigma = 0.05$) lead to larger errors in the reconstruction: the processes move too slowly to capture the index spaces appropriately, see also Figure 3.7. In the other cases, we can assume that the processes have reached their stationary regime. Thus, in the figures and table, we should learn about the implicit regularization that is

| **Method** | **Parameters** | **Mean of** $\text{rel_err}_{N,(\cdot)}$ | $\pm$ **StD** |
|---|---|---|---|
| SGD | $\eta_{(\cdot)} = 0.1$ | $1.849 \cdot 10^{-2}$ | $\pm 5.114 \cdot 10^{-3}$ |
| SGD implicit | $\eta_{(\cdot)} = 0.1$ | $1.772 \cdot 10^{-2}$ | $\pm 4.483 \cdot 10^{-3}$ |
| SGPC with reflected diffusion index process | $\sigma = 5$ | $1.502 \cdot 10^{-2}$ | $\pm 3.339 \cdot 10^{-3}$ |
| | $\sigma = 0.5$ | $1.627 \cdot 10^{-2}$ | $\pm 3.729 \cdot 10^{-3}$ |
| | $\sigma = 0.05$ | $5.933 \cdot 10^{-2}$ | $\pm 1.153 \cdot 10^{-1}$ |
| SGPC with Markov pure jump index process | $\lambda = 10$ | $2.127 \cdot 10^{-2}$ | $\pm 5.494 \cdot 10^{-3}$ |
| | $\lambda = 1$ | $3.542 \cdot 10^{-2}$ | $\pm 1.102 \cdot 10^{-2}$ |
| | $\lambda = 0.1$ | $4.193 \cdot 10^{-2}$ | $\pm 1.351 \cdot 10^{-2}$ |
| | $\lambda = 0.01$ | $2.959 \cdot 10^{-1}$ | $\pm 2.183 \cdot 10^{-1}$ |

Table 3.1: Accuracy of the estimation in the polynomial regression model. Mean and standard deviation of the relative error of the methods at the final point of their trajectory. In particular, sample mean and sample standard deviation of $j \mapsto \text{rel_err}_{N,j}$, with $N = 5 \cdot 10^4$, computed over 100 independent runs.

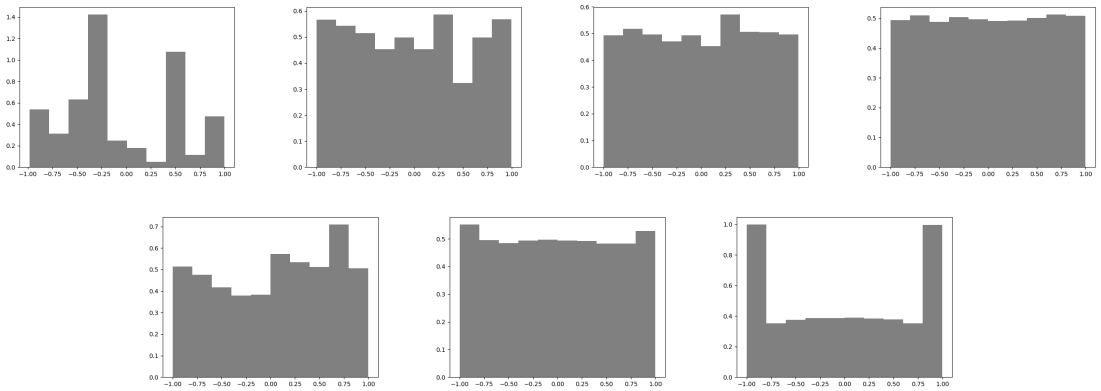

Figure 3.7: Histograms of single simulations of the index process $(V_t)_{t \geq 0}$, given by the pure jump process (top row) with $\lambda \in \{0.01, 0.1, 1, 10\}$ and the reflected diffusion process (bottom row) with $\sigma \in \{0.05, 0.5, 5\}$, increasing from left to right, respectively. In each case, we simulated $5 \cdot 10^4$ time steps.

implicated by the different subsampling schemes, see [1, 122]. We especially see that the mean errors are reduced as $\sigma$ respectively $\lambda$ increases, which illustrates the approximation of the full gradient flow as shown in Theorem 3. Although, we should note that we compute the error to the truth $\Theta$, which is likely not the true minimizer of the full optimization problem 3.31.

It appears that the stochastic gradient processes with reflected diffusion index process $\sigma \in \{0.5, 5\}$ returns the best results. Looking at the error plots in the right column of

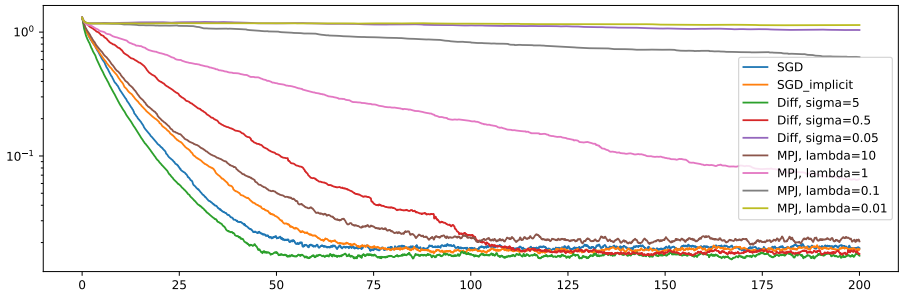


Figure 3.8: Comparison of the mean rel_err of the stochastic methods for time $t \leq 200$.

Figure 3.5, we see that SGPC especially outperforms the other algorithms close to the boundary. For $\sigma = 5$ this could be seen as a numerical artefact due to the time step $t - t(\cdot - 1)$ being too large. This though is likely not the case for $\sigma = 0.5$, where we see a similar effect, albeit a bit weaker. Indeed, considering Figure 3.7, we see that the discretised processes oversample the boundary considerably.

In the convergence plot, Figure 3.8, we see for different methods different speeds of convergence to their respective stationary regime. Those speeds again depend on the auto-correlation of the processes. Interestingly, the SGPC with reflected diffusion index process and $\sigma = 5$ appears to be the best of the algorithms.

### 3.6.2 Solving partial differential equations using neural networks (NNs)

Partial differential equations (PDEs) are used in science and engineering to model systems and processes, such as: turbulent flow, biological growth, or elasticity. Due to the implicit nature of a PDE and its complexity, the model they represent usually needs to be approximated ('solved') numerically. Finite differences, elements, and volumes have been the state of the art for solving PDEs for the last decades. Recently, deep learning approaches have gained popularity for the approximation of PDE solutions. Here, deep learning is particularly successful in high-dimensional settings, where classical methods suffer from the curse

of dimensionality. See for example [105, 84] for physics-informed neural networks (PINNs). Integrated PyTorch-based packages are available for example see [28, 98]. More recently, see [78] for state-of-the-art performance results based on the Fourier neural operator.

Physics-informed neural networks are a very natural field of application of deep learning with continuous data. Below we introduce PINNs, the associated continuous-data optimization problem, and the state-of-the-art in the training of PINNs. Then we consider a particular PDE, showcase the applicability of SGP, and compare its performance with the standard SGD-type algorithm.

The basic idea of PINNs consists in representing the PDE solution by a deep neural network where the parameters of the network are chosen such that the PDE is optimally satisfied. Thus, the problem is reduced to an optimization problem with the loss function formulated from differential equations, boundary conditions, and initial conditions. More precisely, for PDE problems of Dirichlet type, we aim to solve a system of equations of type

$$\begin{cases} \mathcal{L}(u(t,x)) &= s(t,x) & (t \in [0,\infty),\ x \in D) \\ u(0,x) &= u_0(x) & (x \in D) \\ u(t,x) &= b(t,x) & (t \in [0,\infty),\ x \in D) \end{cases} \tag{3.33}$$

where $D \subset \mathbb{R}^d$ is an open, connected, and bounded set and $\mathcal{L}$ is a differential operator defined on a function space $V$ (e.g. $H^1(D)$). The unknown is $u : \bar{D} \to \mathbb{R}^n$. Functions $s(t,x)$, $b(t,x)$, and $u_0(x)$ are given. In numerical practice, we need to replace the infinite-dimensional space $V$ by a – in some sense – discrete representation. Traditionally, one employs a finite-dimensional subspace of $V$, say $\text{span}\{\psi_1, \ldots, \psi_K\}$, where $\psi_1, \ldots, \psi_K$ are basis functions in a finite element method. To take advantage of the recent development of machine learning,

one could solve the problem on a set of deep neural networks contained in $V$, say

$$\Big\{\psi(\cdot;\theta) : \psi(x;\theta) = (W^{(K)}\sigma(\cdot) + b^{(K)}) \circ \cdots \circ (W^{(1)}\sigma(x) + b^{(1)}), x \in [0,\infty) \times D,$$
$$\theta = \big((W^{(K)}, b^{(K)}), \ldots, (W^{(1)}, b^{(1)})\big) \in \prod_{k=1}^{K} \left(\mathbb{R}^{n_k \times n_{k-1}} \times \mathbb{R}^{n_k}\right) =: X\Big\},$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function, applied component-wise, $n_0 = d + 1$ and $n_K = 1$ to match input and output of the PDE's solution space, and $n_1, \ldots, n_{K-1}$ determine the network's architecture.

In simpler terms, let $u(\cdot;\theta) \in V$ be the output of a feedforward neural network (FNN) with parameters (biases/weights) denoted by $\theta \in X$. The parameters can be learned by minimizing the mean squared error (MSE) loss

$$\begin{aligned}\Phi(\theta; \mathcal{L}, s, u_0, b) := &\int_0^\infty w(t) \int_D \left(\mathcal{L}(u(t,x;\theta)) - s(x)\right)^2 \mathrm{d}x\mathrm{d}t + \int_{\partial D} \left(u(0,x;\theta) - u_0(x)\right)^2 \mathrm{d}x \\ &+ \int_0^\infty w(t) \int_{\partial D} \left(u(t,x;\theta) - b(t,x)\right)^2 \mathrm{d}x\mathrm{d}t,\end{aligned}$$

where the first term is the $L^2$ norm of the PDE residual, the second term is the $L^2$ norm of the residual for the initial condition, the third term is the $L^2$ norm of the residual for the boundary conditions, and $w : [0,\infty) \to [0,\infty)$ is an appropriate weight function. The FNN then represents the solution via solving the following minimization problem

$$\min_{\theta \in X} \Phi(\theta; \mathcal{L}, s, u_0, b). \tag{3.34}$$

Note that in physics-informed neural networks, differential operators w.r.t. the input $x$ and the gradient w.r.t the parameter $\theta$ are both obtained using automatic differentiation.

**Training of physics-informed neural networks**

In practice, the optimization problem (3.34) is often replaced by an optimization problem with discrete potential

$$\widehat{\Phi}(\theta; \mathcal{L}, s, u_0, b) := \sum_{k=1}^{K} \left(\mathcal{L}(u(t_k, x_k; \theta)) - s(x_k)\right)^2 + \sum_{k'=1}^{K'} \left(u(0, x'_{k'}; \theta) - u_0(x'_{k'})\right)^2 + \sum_{k''=1}^{K''} \left(u(t''_{k''}, x''_{k''}; \theta) - b(t''_{k''}, x''_{k''})\right)^2,$$

for appropriate continuous indices

$$(x_k, t_k)_{k=1}^{K} \in [0, \infty)^K \times D^K, (x'_{k'})_{k'=1}^{K'} \in \partial D^K, (x''_{k''}, t''_{k''})_{k''=1}^{K''} \in [0, \infty)^K \times \partial D^K$$

that may be chosen deterministically or randomly, see for example [98, 84].

Focusing the training on a fixed set of samples can be problematic: fixing a set of random samples might be unreliable; a reliable cover of the domain will likely only be reached through tight meshing, which scales badly. [121] propose to use SGD on the continuous data space. They employ the discrete dynamic in (3.2). Naturally, we would like to follow [121] and employ the SGP dynamic on the continuous index set.

To train the PINNs with SGP, we again choose the reflected Brownian motion as an index process, which we discretize with the Euler–Maruyama scheme in Algorithm 2. In addition, we employ mini-batching to reduce the variance in the estimator: We sample $M \in \mathbb{N}$ independent index processes $(V_t^{(1)})_{t\geq 0}, \ldots, (V_t^{(M)})_{t\geq 0}$ and then employ the dynamical system

$$\mathrm{d}\theta_t = -\frac{1}{M} \sum_{m=1}^{M} \nabla_\theta f(\theta_t, V_t^{(m)}) \mathrm{d}t.$$

Hence, rather than optimizing with respect to a single data set, we optimize with respect to $M$ different data sets in each iteration. While we only briefly mention the mini-batching throughout our analysis, one can easily see that it is fully contained in our framework.

In preliminary experiments, we noticed that the Brownian motion for the sampling on the boundary is not very effective: possibly due to its localizing effect. Hence, we obtain training data on the boundary by sampling uniformly, which we consider justified as a mesh on the boundary scales more slowly as a mesh in the interior and as the boundary behavior of the considered PDE is rather predictable.

### PDE and results

We now describe the partial differential equation that we aim to solve with our PINN model. After introducing the PDE we immediately outline the PINN's architectures and show our estimation results. We the train networks on Google Colab Pro using GPUs (often T4 and P100, sometimes K80). We are certain that a more efficient PDE solution could be obtained by classical methods, e.g., the finite element method. We do not compare the deep learning methods with classical methods, as we are mainly interested in SGP and SGD in non-convex continuous-data settings. Other methods that could approximate the PDE solution are not our focus.

The PDE we study is a transport equation; which is a linear first order, time-dependent model. One of the main advantages of studying this particular model is that we know an analytical solution that allows us to compute a precise test error.

**Example 1** (1D Transport equation)**.** *We solve the one-dimensional transport equation on the space* $[0,1]$ *with periodic boundary conditions:*

$$\begin{cases} u_t + u_x &= 0 & (t \in [0,\infty), \quad x \in [0,1]) \\ u(0,x) &= \sin(2\pi x) & (x \in [0,1]) \\ u(t,0) &= u(t,1) & (t \in [0,\infty)). \end{cases} \tag{3.35}$$

The neural network approximation of this PDE has already been studied by [98], our experiments partially use the code associated to this work. The network architecture is defined by a three-layer deep neural network with 128 neurons per layer and a Rectified

Linear Unit (ReLu) activation function. While theoretically the solution exists globally in time, we restrict $t$ to a compact domain and w.l.o.g, we assume $t \in [0, 1]$. From the interior of the domain of time and space variables, i.e. $(0, 1) \times (0, 1)$, we use Algorithm 2 with $\sigma = 0.5$ to sample the train set of size $3 \cdot 10^4$ for SGPC and SGPD and we uniformly sample 600 points for the train set of SGD. In addition, as a part of the train set for all three methods, we sample uniformly 20 and 60 points for the initial condition and periodic boundary condition, respectively.

The learning rate for SGD and SGPC is 0.01. The learning rate for SGPD is defined as

$$\eta(t) = \frac{0.01}{\log(t+2)^{0.3}},$$

which is chosen such that the associated $\mu := 1/\eta$ satisfies Assumption 4. For all three methods, we use Adam [see 64] as the optimizer to speed up the convergence; we use an $L^2$ regularizer with weight 0.1 to avoid overfitting. Each model is trained over 600 iterations with batch size 50. The training process for SGPC and SGPD contains only one epoch, while we train 50 epochs in the SGD case. We evaluate the models by testing on a uniformly sampled test set of size $2 \cdot 10^3$ and compare the predicted values with the theoretical solution

$$u(t, x) = \sin(2\pi(x - t)).$$

We obtain the losses, the predicted solutions, and the test errors by averaging over 30 random experiments, i.e. 30 independent runs of SGD, SGPC, and SGPD, respectively. We give the results in Figures 3.9, 3.10, and 3.11. Note that the timings are very similar for each of the algorithms, the fact that SGPC and SGPD require us to first sample reflected Brownian motions is negligible.

From Figure 3.9, we notice that while SGD and SGPC behave similarly, SGPD does converge faster. Here, Assumption 4 provides a way of designing a non-constant learning rate in practice. On the test set, the mean squared errors for SGD, SGPC, and SGPD are

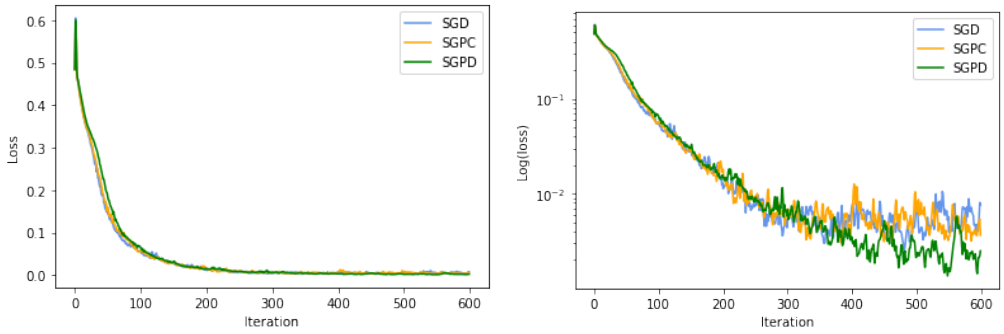


Figure 3.9: The plots of the loss vs iteration and its log scale for SGD, SGPC, and SGPD. The losses are obtained by averaging over 30 random experiments.

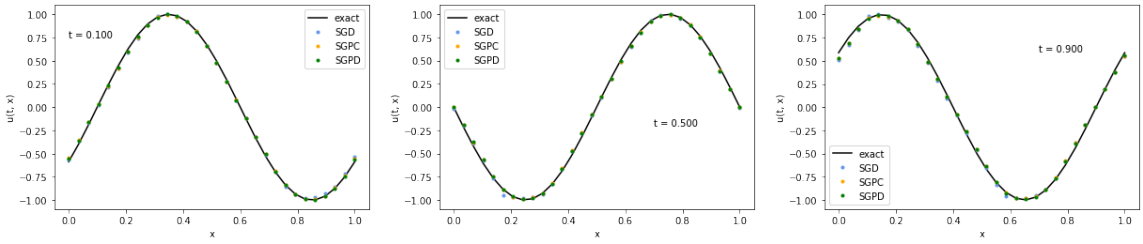


Figure 3.10: The plots of the solutions at time $t = 0.1, 0.5, 0.9$. We evaluate the models at 30 uniformly sampled points. For each method, the predicted values are taken by averaging over the predicted values from the best models (the model that achieves the lowest training loss within the 600 iteration steps) in 30 random experiments. The black curve is the theoretical solution.

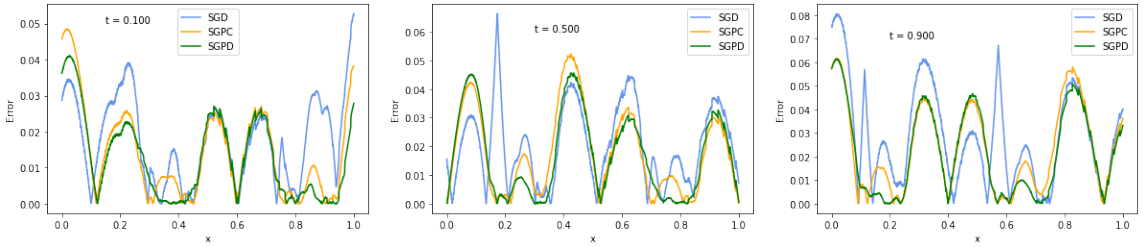


Figure 3.11: The plots of the test error at time $t = 0.1, 0.5, 0.9$. We evaluate the models at 2000 uniformly sampled points. For each method, the predicted values are taken by averaging over the predicted values from the best models (the model that achieves the lowest training loss within the 600 iteration steps) in 30 random experiments. At each point $x$, the error calculated by taking the absolute value of the difference between the predicted value and the true solution.

$4.5 \cdot 10^{-4}$, $3.5 \cdot 10^{-4}$, and $2.8 \cdot 10^{-4}$. These test errors refer to the averaged model output of the 30 models from independent experiments. Combined with Figure 3.10 and Figure 3.11, we observe that SGPC and SGPD generalize at least slightly better on the test set.

This improved generalization error might be due to the additional test data generated by the Brownian motion, as compared to the fixed training set used in PINNs. The combination with the reduction of the learning rate in SGPD, appears to be especially effective.

## 3.7 Conclusions and outlook

In this chapter we have proposed and analyzed a continuous-time stochastic gradient descent method for optimization with respect to continuous data. Our framework is very flexible: it allows for a whole range of random sampling patterns on the continuous data space, which is particularly useful when the data is streamed or simulated. Our analysis shows ergodicity of the dynamical system under convexity assumptions – converging to a stationary measure when the learning rate is constant and to the minimizer when the learning rate decreases. In experiments we see the suitability of the method and the effect of different sampling patterns on its implicit regularization.

We end this chapter by now briefly listing some interesting problems for future research in this area. First, we would like to learn how the SGP sampling patterns perform in large-scale (adversarially-)robust machine learning and in other applications we have mentioned but not studied here. Moreover, from a both practical and analytical perspective, it would be interesting to also consider non-compact index spaces $S$. Those appear especially in robust optimal control and variational Bayes. Finally, we consider the following generalization of the optimization problem (3.1) to be of high interest:

$$\min_{\theta \in X} \int_S f(\theta, y) \Pi(\mathrm{d}y|\theta),$$

where $\Pi$ is now a Markov kernel from $X$ to $S$. Hence, in this case the probability distribution and the sampling pattern itself depend on the parameter $\theta$. Optimization problems of this form appear in the optimal control of random systems (e.g., [38]) and empirical Bayes (e.g., [26]) but also in reinforcement learning (e.g., [125]).

# Chapter 4

# The stochastic gradient-momentum process

## 4.1 Introduction

In this chapter, we investigate momentum-based stochastic optimization methods in a continuous-time framework. We now introduce the optimization problem, motivate stochastic and momentum-based optimization, review past work, and summarize our contributions and main results.

### 4.1.1 Optimization and continuous dynamics

We study optimization problems of the form

$$\min_{\theta \in X} \bar{\Phi}(\theta), \tag{4.1}$$

on the space $X := \mathbb{R}^K$, where

$$\bar{\Phi}(\theta) := \frac{1}{N} \sum_{i=1}^{N} \Phi_i(\theta).$$

We denote the number of terms by $N \in \mathbb{N} := \{1, 2, \ldots\}$ and assume that the functions $\Phi_i$, with $i \in I := \{1, \ldots, N\}$, are continuously differentiable with Lipschitz gradients. We assume that (4.1) is well-defined and denote $\theta^* :\in \operatorname{argmin} \bar{\Phi}$. In this chapter, we study continuous-time dynamical systems that can be used to optimize functions of type $\bar{\Phi}$ and to represent the dynamics of optimization algorithms. The most basic of these dynamical systems is the *gradient flow*

$$\frac{\mathrm{d}\zeta_t}{\mathrm{dt}} = -\nabla\bar{\Phi}(\zeta_t), \qquad \zeta_0 \in X, \tag{4.2}$$

a continuous-time version of the *gradient descent method*. The ODE solution $(\zeta(t))_{t\geq 0}$ converges to a stationary point of $\bar{\Phi}$ under, e.g. certain convexity conditions. In practice, especially in modern machine learning, we often encounter optimization problems that are non-convex, where the solution $(\zeta(t))_{t\geq 0}$ may converge to a saddle point or a local minimizer. When discretizing (4.2), the iterates actually only converge to (local) minimizers, see, e.g. [75, 97]. However, as observed in [39], the algorithm could take a lot of iterations to escape a saddle point.

The *stochastic gradient descent (SGD)* method going back to [107] has been a popular alternative to the gradient descent method. In SGD, we randomly select one of the $\Phi_i$ ($i \in I$) and optimize with respect to that function before switching to another $\Phi_j$ ($j \in I\backslash\{i\}$). This process is repeated until a stationary state is reached. A continuous-time model for stochastic gradient descent has recently proposed by [71] and extended by [58] and [72]. There, stochastic gradient descent is represented through the dynamical system

$$\frac{\mathrm{d}\theta_t}{\mathrm{dt}} = -\nabla\Phi_{i(t)}(\theta_t), \qquad \theta_0 \in X, \tag{4.3}$$

where the *index process* $(i(t))_{t\geq 0}$ is a suitable homogeneous continuous-time Markov process on $I$. The processes $(i(t))_{t\geq 0}$, $(\theta_t)_{t\geq 0}$ and any other stochastic processes and random variables throughout this chapter are defined on the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The

process $(\theta_t)_{t\geq 0}$ represents a randomized gradient flow whose potential is randomly replaced by another one when $(i(t))_{t\geq 0}$ jumps from one state to another. Both $(\theta_t)_{t\geq 0}$ and $(i(t), \theta_t)_{t\geq 0}$ are called *stochastic gradient process*. Since $\bar{\Phi}$ often represents the averaging over data subsets, we refer to the process of replacing $\bar{\Phi}$ by a $\Phi_i$ as *subsampling*.

SGD has two main advantages: Since we consider only one of the $\nabla\Phi_i$ at a time, we significantly reduce the computational cost when discretizing the dynamical system. Moreover, the perturbation through the randomized sampling in stochastic gradient descent can help to overcome saddle points, see [34, 57]. However, as the method is purely gradient-based, we would expect that the escaping saddle points is still slow. Saddle points and local minimizers can sometimes be escaped when using momentum-based optimization methods. Here, we replace the gradient flow (4.2), by the *underdamped gradient flow* given through

$$m\frac{\mathrm{d}^2 q_t}{\mathrm{d}t^2} - \alpha\frac{\mathrm{d}q_t}{\mathrm{d}t} + \nabla\bar{\Phi}(q_t) = 0, \quad \frac{\mathrm{d}q_0}{\mathrm{d}t} \in X, \quad q_0 \in X. \tag{4.4}$$

This dynamical system describes the motion of a ball with mass $m > 0$ constrained to slide on the graph $\bar{\Phi}$ subject to a constant gravitational field and viscosity friction $\alpha > 0$, see [9] for details. Intuitively, momentum is used to vault the optimizer out of saddle points and local minimizers.

### 4.1.2 Momentum

We now give a motivating example where the relevance of momentum in non-convex optimization can be observed and then summarize relevant literature on momentum-based optimization. The example is based on a simple one-dimensional non-smooth non-convex optimization problem inspired by the training of neural networks. As we shall see, the introduction of momentum can prevent the convergence of the gradient flow to a stationary point that is not the global minimizer. We later work in a framework that does not actually cover this introductory example, we present it only to give an intuition for the methodology.

Let us consider the function $\Phi : \mathbb{R} \to \mathbb{R}$ defined as

$$\Phi(x) := (\mathrm{ReLU}(x) - 1)^2 + x^2 = \begin{cases} x^2 + 1 & \text{if } x \leq 0 \\ 2x^2 - 2x + 1 & \text{if } x > 0 \end{cases} \qquad (x \in \mathbb{R}) \tag{4.5}$$

where $\mathrm{ReLU}(\cdot) := \max\{0, \cdot\}$. As shown in Figure 4.1, the function $\Phi$ attains its global

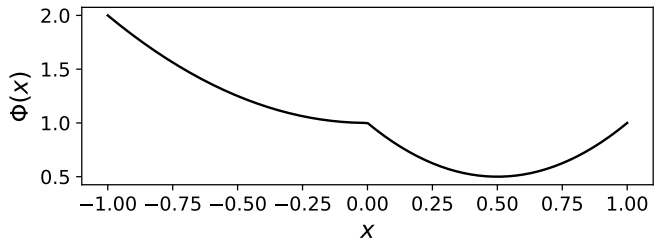


Figure 4.1: Plot of the potential $\Phi$.

minimum at the point $x^* = \frac{1}{2}$. However, $\tilde{x} = 0$ is the minimizer of $\Phi$ restricted to the non-positive half-line $(-\infty, 0]$. Therefore, any solution of the gradient flow equation (4.2) with potential $\bar{\Phi} := \Phi$ and initial value $\zeta_0 < 0$ converges to the saddle point $\tilde{x}$. In order to avoid the convergence to $\tilde{x}$, we can replace the gradient flow by the associated underdamped gradient flow (4.4) with mass $m > 0$, viscosity friction $\alpha > 0$, and potential $\bar{\Phi}$. If the initial value $q_0$ is negative, then the equation of the motion reduces to

$$m\frac{\mathrm{d}^2 q_t}{\mathrm{d}t^2} + 2q_t = -\alpha\frac{\mathrm{d}q_t}{\mathrm{d}t}, \tag{4.6}$$

whenever the dynamical system stays in the negative half-line. At this point the natural question is whether there exist combinations of $m$ and $\alpha$ such that, assuming that $q_0 < 0$, the particle is not confined in the negative half-line for every time $t > 0$. Using elementary theory of linear second-order differential equations, it turns out that the particle manages to overcome the "false" minimizer $\tilde{x} = 0$ if the following relation is satisfied:

$$\alpha^2 - 8m < 0, \tag{4.7}$$

i.e. the friction $\alpha$ is sufficiently small or the mass $m$ sufficiently large. In summary, momentum can be used to overcome stationary points of the target function.

The use of momentum in optimization dates back to the 1960s, when Polyak formulated the *heavy ball method* in convex optimization, see [101, 102]. In this context, the aim was to accelerate the convergence of the gradient descent method. Later, Nesterov proposed a momentum-based class of discrete accelerated methods for convex problems in his seminal paper [92]. Momentum-based optimization methods have been investigated thoroughly from the perspective of continuous-time dynamical systems. In-depth studies about the mechanical systems underlying Polyak's method can be found in [9] – a work which has sparked considerable interest in the community. We recall, for instance, [3] which have studied constrained convex optimization. In [123], the authors have formulated a dissipative second order ODE that provides a continuous-time counterpart to the Nesterov method. Here, the viscosity friction parameter $\alpha$ is time-dependent and vanishing as $t \to \infty$. This aspect was later further investigated in [6, 116, 115, 110]. In [8, 7], the authors combine vanishing viscosity with time-scaling of trajectories, leading to a dynamics similar to (4.4), but with an extra time-dependent coefficient $\beta(t)$ appearing in front of the term $\nabla\bar{\Phi}$, i.e.

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta(t) - \alpha(t)\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) + \beta(t)\nabla\bar{\Phi}(\theta(t)) = 0. \tag{4.8}$$

The parameter $\beta$ satisfies $\beta(t) \to \infty$ as $t \to \infty$.

Finally, we mention more recent advances on second-order dynamics with time scaling that include Hessian-driven damping [4] and averaging [5]. Next, we see how momentum is used within stochastic optimization practice.

### 4.1.3 Momentum-based stochastic optimization and the Adam algorithm

The most basic momentum-based stochastic optimization method combines a symplectic Euler discretization of the underdamped gradient flow (4.4) with subsampling. This *stochastic gradient descent method with classical momentum* is given by

$$\begin{cases} v_n = \alpha v_{n-1} - \eta \nabla \Phi_{i_n}(\theta_{n-1}), \\ \theta_n = \theta_{n-1} + v_n, \end{cases} \tag{4.9}$$

where $\alpha, \eta > 0$ are hyperparameters and $i_1, i_2, \ldots \sim \text{Unif}(I)$ are independent and identically distributed (i.i.d.) random variables. We refer to $\eta$ as learning rate and refer to (4.9) just as *classical momentum*. The connection to the underdamped gradient flow is more obvious when considering the timestep size to be equal to 1, $\eta$ to be a scaling factor of the potential, and $\alpha$ to be a parameter that incorporates the velocity at the previous timestep and the friction. Classical momentum has been studied by [109, 124]. In deep learning, more advanced momentum-based descent methods have been widely in use, including Adagrad [41]; Adadelta [135]; RMSprop [54]; Adam [64], etc. Adam is one of the most commonly used optimizers when training neural networks. The updating rule is the following,

$$\begin{cases} u_n = \beta_1 u_{n-1} + (1-\beta_1)\nabla \Phi_{i_n}(\theta_{n-1}), \\ v_n = \beta_2 v_{n-1} + (1-\beta_2)\big[\nabla \Phi_{i_n}(\theta_{n-1})\big]^2, \\ \theta_n = \theta_{n-1} - \alpha \frac{u_n/(1-\beta_1^n)}{\sqrt{v_n/(1-\beta_2^n)}+\beta_3}, \end{cases} \tag{4.10}$$

where $\beta_1$, $\beta_2$, and $\alpha$ are hyperparameters, $\beta_3$ is a small constant, and $i_1, i_2, \ldots \sim \text{Unif}(I)$ (i.i.d.). Comparing to the classical momentum method, Adam uses not only an estimate of the first moment of the gradient but also an estimate of the second moment as the part of the adaptive learning rate. The variable $u_n$ is the biased first moment, similar to the

one used in the classical momentum method. The variable $v_n$ is the biased second moment estimated using the gradient squared. When updating $\theta_n$, Adam normalises $u_n$ and $v_n$ so that they become unbiased estimators. Moreover, it was shown in [37] that Adam converges in the sense of [139] with speed $O(\log(t)/\sqrt{t}; t \to \infty)$.

While momentum-based stochastic optimization methods are popular in machine learning practice, they are overall rather badly understood; see the discussion in [82]. Here, we analyze a continuous-time stochastic gradient process with momentum and propose an efficient discretization technique for this dynamical system. Thus, we improve the understanding of momentum-based stochastic optimization in a theoretical framework and machine learning practice. Our analysis complements the Langevin-based analyzes by [77, 114], where the authors represent SGD in continuous time as a diffusion process. Before we summarize our contributions more precisely, we introduce our continuous-time stochastic gradient-momentum process.

### 4.1.4 The stochastic gradient-momentum process

Throughout this chapter, we discuss a continuous-time dynamical system that represents stochastic gradient descent with momentum – the stochastic gradient-momentum process. Due to the continuous-time nature of the system, it can equally represent classical momentum (4.9) and Adam (4.10) (although not the adaptive nature of the latter). We obtain this system by combining the underdamped dynamical system (4.4) with the stochastic gradient process (4.3). We now introduce the *index process*, the process that controls the subsampling.

**Definition 1** (Index process)**.** *Let $(i(t))_{t\geq 0}$ be a continuous-time Markov process on $I$ with transition rate matrix*

$$\mathbf{A}_N = \mathbf{\Gamma}_N - N\gamma \mathbf{I}_N.$$

*Here, $\gamma > 0$ denotes the jump rate, $\mathbf{\Gamma}_N$ is an $N \times N$ matrix whose entries are all equal to $\gamma$, and $\mathbf{I}_N$ is the identity matrix. We assume that the initial distribution $\mathbb{P}(i(0) \in \cdot)$ is the*

*uniform distribution on $I$. Moreover, when given a value $\nu > 0$ or an appropriate, strictly increasing function $\beta : [0, \infty) \to [0, \infty)$, we define the rescaled index processes $i^\nu(t) := i(t/\nu)$ and $i^\beta(t) := i(\beta^{-1}(t))$, for $t \geq 0$. We refer to each of $(i(t))_{t\geq 0}$, $(i^\nu(t))_{t\geq 0}$, and $(i^\beta(t))_{t\geq 0}$ as* index process.

The function $\beta^{-1}$ and the scalar $1/\nu$ play the role of a time-dependent and constant learning rate, respectively; see [58]. Thus, we sometimes refer to them as learning rate. Next, we make the following regularity assumptions.

**Assumption 5.** *Let $\Phi_i \in \mathcal{C}^1(X, \mathbb{R})$, i.e. it is continuously differentiable, and $\nabla_x \Phi_i$ be Lipschitz with Lipschitz constant $L > 0$, for $i \in I$.*

Throughout this chapter, we study multiple versions of the stochastic gradient-momentum process and introduce them in Definitions 2, 3, and 4. We now introduce the prototype of the dynamical systems studied in this chapter. Note that we often choose the more compact form $\mathrm{d}z_t = f(z_t)\mathrm{d}t$ to represent $\frac{\mathrm{d}z_t}{\mathrm{d}t} = f(z_t)$. The general *stochastic gradient-momentum process* (SGMP) is given by $(p_t^\dagger, q_t^\dagger, i^\beta(t))_{t\geq 0}$ satisfying

$$\begin{cases} \mathrm{d}q_t^\dagger &= p_t^\dagger \mathrm{d}t, \\ \varepsilon(t)\mathrm{d}p_t^\dagger &= -\nabla \Phi_{i^\beta(t)}(q_t^\dagger)\mathrm{d}t - \alpha p_t^\dagger \mathrm{d}t, \\ p^\dagger(t=0) &= p_0, \\ q^\dagger(t=0) &= q_0, \end{cases} \tag{4.11}$$

where $(p_t^\dagger)_{t\geq 0}$ describes the velocity of the particle, $(q_t^\dagger)_{t\geq 0}$ its position, $(\varepsilon(t))_{t\geq 0}$ its mass (that may depend on time), and $\alpha > 0$ the viscocity friction. The function $\beta$ controls the switching of the index process. In the following, we explain the importance behind the functions $m$ and $\beta$.

**Remark 7** (Mass $m$)**.** *The introduction of momentum allows the particle to be vaulted out of stationary points. An effect that is particularly pronounced if $m$ is large or $\alpha$ is small. While this behaviour is convenient when escaping local minimizers and saddle points, it can also*

*occur when approaching global minimizers. We give a very simple example in Figure 4.2, where we see that the method shows oscillatory behaviour and very slow convergence when the mass $m$ is large and constant. As a tuning of the mass $m$ is difficult in practice, we suggest an alternative strategy: reducing the mass over time. Rationale: a large mass in the beginning leads to a fast escaping of local minimizers, while a small mass later leads to fast convergence once the global minimizer is reached. This intuition is confirmed in Figure 4.2.*

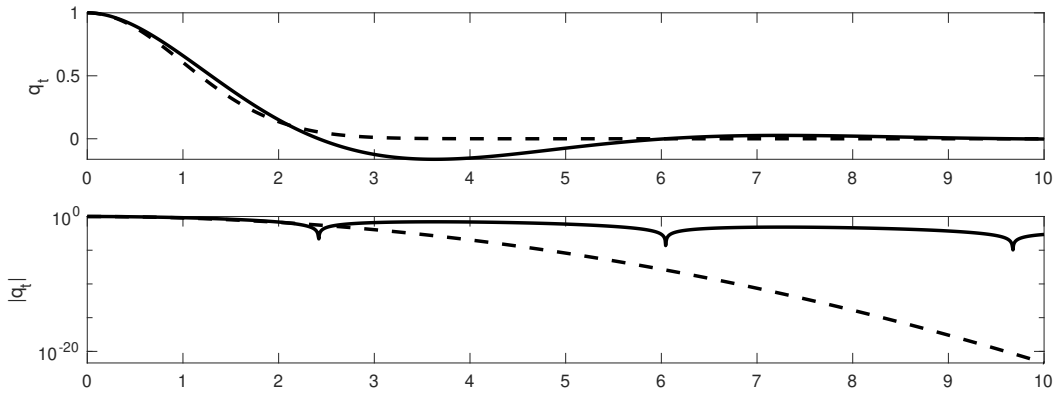


Figure 4.2: We consider the minimization of $\bar{\Phi}(\theta) := \theta^2/2$. Let $\alpha = 1, \varepsilon = 1$, $p_0 = 1$, $q_0 = 0$. Then, $q_t = \exp(-t/2)\left(\sin(\sqrt{3}t/2)/\sqrt{3} + \cos(\sqrt{3}t/2)\right)$ $(t \geq 0)$ (solid line) oscillates around the solution and converges ultimately to $\theta^* = 0$. If we choose instead $m(t) = (1+t)^{-1}$ $(t \geq 0)$ (dashed lines), we have $q_t = \exp(-t^2/2)$, which does not oscillate, but converges very quickly to $\theta^*$.

In [9], the authors have considered the singular perturbation of (4.4), corresponding to letting the mass $m \to 0$, and they established convergence of these dynamics to the gradient flow trajectories. Here, we also consider a vanishing mass and carefully bound the Euclidean distance between underdamped gradient flow and gradient flow. Moreover, we let $m$ depend on the time, and investigate the case $m(t) \to 0$ as $t \to \infty$. This setting is conceptually similar to the dynamics studied in [8, 7] that we give in (4.8). However, $\beta(t) \to \infty$ as $t \to \infty$ is not equivalent to the mass going to zero. We can rewrite the underdamped gradient flow to

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta(t) - \frac{\alpha}{m(t)}\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) + \underbrace{\frac{1}{m(t)}}_{=\beta(t)}\nabla\Phi(\theta(t)) = 0, \tag{4.12}$$

where the factor multiplied by the velocity explodes as well, as $m(t) \to 0$. Thus, we cannot rely on the results obtained in [8, 7] for the dynamics (4.8). Moreover, we shall not actually focus our attention on the deterministic evolution, but we rather investigate a stochastic process related to the ODE (4.12). However, an interesting similarity between our approach and [8, 7] arises when discretizing the respective dynamics. Indeed, the fact that $\beta(t) \to \infty$ in (4.8) and $m(t) \to 0$ in (4.12) make explicit schemes not suitable for deriving discrete approximations. In [8, 7] this issue is addressed by considering a *proximal* discretization, while in Section 4.4 we opt for a symplectic (semi-implicit) scheme.

**Remark 8** (Learning rate and $\beta$)**.** *The switching between the potentials leads to a similar effect as the momentum. As the flow approaches the minimizers of different target functions in each step, we are unlikely to converge to a single point. While this helps to escape local minimizers, it prohibits the convergence to a global minimizer. In several stochastic optimization methods, we need to reduce the learning rate or step-size throughout the algorithm to reduce the time in-between switches of data sets. We can attain this reduction of learning rate by rescaling the time in the index process $(i(t))_{t\geq 0}$ through an appropriate function $\beta$. The rescaling will lead to the waiting times between two switches to decrease over time.*

We now list and explain more assumptions that are required throughout this chapter. Indeed, some results of this chapter require the $(\Phi_i)_{i\in I}$ to be *($\mu$-)strongly convex*, i.e.

$$(x - x') \cdot (\nabla\Phi_i(x) - \nabla\Phi_i(x')) \geq \mu \left\|x - x'\right\|^2 \qquad (x, x' \in X, i \in I),$$

for some strong convexity parameter $\mu > 0$. Others do not rely on strong convexity, but still need some restriction on the potential function, as given in the following assumption.

**Assumption 6.** *Let $\Phi_i \in \mathcal{C}^1(X, \mathbb{R})$ admit a global minimizer $\theta_*^i$, for $i \in I$, and let there be constants $\lambda \in (0, 1/4]$ and $\bar{\alpha} > 0$ such that*

$$(x - \theta_*^i) \cdot \nabla\Phi_i(x)/2 \geq \lambda(\Phi_i(x) - \Phi_i(\theta_*^i) + \bar{\alpha}^2 \left\|x - \theta_*^i\right\|^2/4), \qquad (x \in X, i \in I). \tag{4.13}$$

We use $\mathcal{A}^{\lambda,\bar{\alpha}}$ as a short hand to refer to (4.13) with parameters $\lambda$ and $\bar{\alpha}$. We note that Assumption 6 is also known as the quasi-strong convexity, see, e.g. [10, 42, 90]. Assumption 6 is implied by strong convexity. We prove this assertion in Lemma 25 in the appendix. In the appendix, we also give a counterexample showing that the two assumptions are not equivalent, see Example 3.

In stochastic optimization, it is often necessary to decrease the learning rate over time to obtain a method that converges to the minimizer. In practice, however, this is sometimes purposefully disregarded – convergence to, e.g. a probability distribution is preferred. In stochastic optimization this leads to an *implicit regularization* of the optimization problem. Particularly in machine learning, this implicit regularization can be necessary to get good generalisation results, see, for example, [112, 63, 93, 136, 140]. Thus, we also study this constant learning rate case. Similarly, whilst we have seen that a decreasing mass can be beneficial, we also study the case of a constant mass. Indeed, when the mass $m$ does not depend on $t$ or the function $\beta$ is linear, i.e. $\beta(u) = \nu u$, we refer to mass or learning rate as *homogeneous*, respectively. Since we are otherwise always interested in reducing them over time, we speak of *decreasing* mass or learning rate, respectively. When decreasing mass and learning rate, an additional assumption – Assumption 8 – is needed. Since this assumption is rather technical, we delay its statement and discuss it instead along with the associated convergence result in Subsection 4.3.2.

In order to establish well-posedness of the system, the mass $(m(t))_{t\geq 0}$ must not decrease too quickly. Here, we have the following assumption.

**Assumption 7.** *There exists a constant $\lambda \in (0, 1]$, such that $|\varepsilon'(t)| \leq \lambda\varepsilon(t)$, $t > 0$. Moreover, we assume that $\varepsilon(0) =: \varepsilon_0 \leq 1$.*

It is easy to verify that $\varepsilon(t) = \frac{\varepsilon_0}{\lambda^{-1}+t}$ and $\varepsilon(t) = \varepsilon_0 e^{-\lambda t}$ satisfy Assumption 7. Setting $m_0 = \lambda = 1$, we retrieve the example in Figure 4.2. More generally, Assumption 7 implies $\varepsilon(t) \geq \varepsilon_0 e^{-\lambda t}$. Hence, the mass decreases at most exponentially with a rate $\geq -1$.

### 4.1.5 Contributions and main results

We have introduced the stochastic gradient-momentum process above. In this chapter, we analyze this dynamical system, discuss its discretization, and employ it in numerical experiments. We are interested in both, convergence to global minimizers and implicit regularization. Thus, we study three different stochastic gradient-momentum processes: (1) homogeneous mass and homogeneous learning rate, (2) decreasing mass and homogeneous learning rate, and (3) decreasing mass and decreasing learning rate, respectively. We now summarize our contributions:

- We study the connection between SGMP and the gradient flow, the underdamped gradient flow, as well as the stochastic gradient process. These connections allow us to investigate the longtime behavior of the stochastic gradient-momentum process.

We obtain those connections by considering momentum-to-no-momentum asymptotics (letting a homogeneous mass $m$ approach 0), and random-to-deterministic asymptotics (for a homogeneous $\beta(u) = \nu u$, letting $\nu$ approach 0). Understanding the longtime behavior eventually allows us to study the convergence of the underlying algorithm. We summarize our three main results below and also give informal statements of the corresponding theorems – the complete versions are given later in the text.

- In the fully homogeneous case, we show that the stochastic gradient-momentum process is close to the stochastic gradient process when the mass is small. This can be understood as the stochastic version of the deterministic result established in [9].

  **Theorem 7 (informal)** *Let $(q_t^\varepsilon)_{t\geq 0}$ be the stochastic gradient-momentum process (4.15) with constant mass $m$ and $(\theta_t^\varepsilon)_{t\geq 0}$ be the stochastic gradient process (4.16). Under Assumptions 5 and 6, we have*

$$\mathbb{E}\Big[\sup_{0\leq t\leq T}\|q_t^\varepsilon - \theta_t^\varepsilon\|\Big] \to 0 \quad (m \downarrow 0).$$

- If the mass decreases over time, the stochastic gradient-momentum process converges to the stochastic gradient process in the longtime limit.

  **Theorem 8 (informal)** *Let $(\tilde{q}_t)_{t\geq 0}$ be the decreasing-mass stochastic gradient-momentum process (4.19) with time-dependent momentum $(m(t))_{t\geq 0}$ and $(\theta_t)_{t\geq 0}$ be the stochastic gradient process (4.3). Under strong convexity and Assumptions 5 and 7, we have, almost surely and in expectation,*

$$\|\tilde{q}_t - \theta_t\| \to 0 \quad (t \to \infty).$$

- If mass and learning rate decrease over time, the stochastic gradient-momentum process converges to the minimizer of the full target function in the longtime limit.

  **Corollary 6 (informal)** *Let $(\hat{q}_t)_{t\geq 0}$ be the decreasing-mass, decreasing-learning-rate stochastic gradient-momentum process with time-dependent mass $(m(t))_{t\geq 0}$ and learning rate $(\beta(t))_{t\geq 0}$. Let $\theta_*$ be the minimizer of the full target function. Under strong convexity and Assumptions 5, 7, and 8, we have*

$$\|\hat{q}_t - \theta_*\| \xrightarrow{\textit{in probability}} 0 \quad (t \to \infty).$$

The connection to the stochastic gradient process is especially interesting, as its longtime behaviour in convex settings is well-understood, see [58, 71]. Thus, we can argue that SGMP leads to a similar implicit regularization when the mass is small or converging to zero. In a wider sense, the connections to the other methods allow us to interpret the stochastic gradient-momentum process as a method that can freely interpolate between gradient flow, underdamped gradient flow, and stochastic gradient process – through adjusting learning rate $\nu^{-1}$ and mass $m$. We depict this relationship in Figure 4.3.

Surprisingly, the continuous relationship shown in Figure 4.3 is lost when, e.g. comparing the discrete-time algorithms SGD and classical momentum. This is mainly due to the sym-

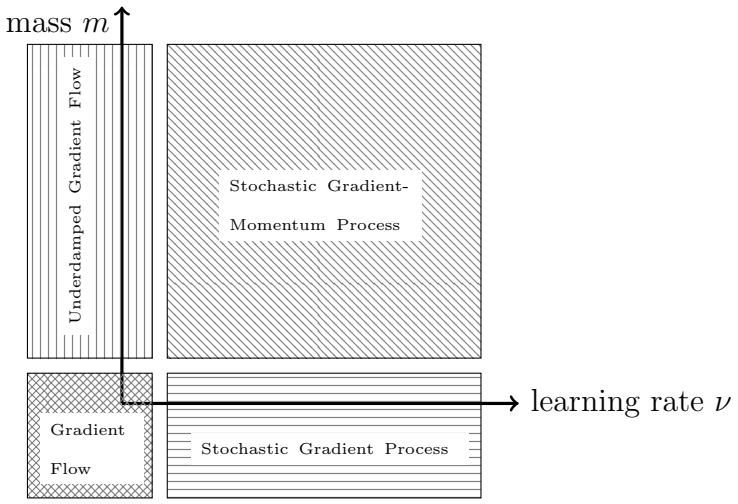


Figure 4.3: Schematic comparing gradient flow, underdamped gradient flow, stochastic gradient process, and stochastic gradient-momentum process in terms of the particle mass $m$ and learning rate $\nu$.

plectic Euler discretization that is employed in classical momentum. Especially problematic is the instability of classical momentum when the mass is small. The step-size needs to be chosen as $O(m; m \downarrow 0)$. From this computational perspective, our contributions are the following:

- We propose a discretization strategy for the stochastic gradient-momentum process. The strategy is a semi-implicit method that is explicit in the position $(q_t)_{t\geq 0}$ and implicit in the velocity $(p_t)_{t\geq 0}$. This discretization technique allows us to choose the step-size independently of the mass $m$. Thus, the method is stable for small masses as well as masses converging to zero. Similarly to the symplectic Euler method that is conventionally used in momentum-based optimization, our scheme is symplectic. Thus, it respects (local) conservation laws in the underdamped dynamics.

- We test our discretized algorithm in numerical experiments. We start with academic convex and non-convex optimization problems in low and high dimensions. Then we study the training of a convolutional neural network (CNN), which we use to classify images from the CIFAR-10 dataset. The achieved train and test accuracy with our stable discretization is comparable to stochastic gradient descent with classical mo-

mentum. Some of these experiments go beyond the convex setting that we study in our analysis.

This chapter is organized as follows. We introduce and investigate the homogeneous-in-time and heterogeneous-in-time stochastic gradient-momentum processes in Sections 4.2 and 4.3, respectively. In Section 4.4 we propose discretization techniques which we then employ in the mentioned academic and deep learning problems in Section 4.5. A conclusion is given in Section 4.6. Throughout the main part of the chapter, we require certain results about underdamped gradient flows, which we discuss in Appedices A.1 and A.2. Other auxiliary results needed throughout this chapter, are presented in Appendix A.3. Appendix A.4 recalls and summarizes a result by [68].

## 4.2 Homogeneous-in-time

In this section, we study the stochastic gradient-momentum process with homogeneous momentum and learning rate. In the first part, Subsection 4.2.1, we investigate the interplay between the underdamped gradient flow and the stochastic gradient-momentum process. In the second part, Subsection 4.2.2, we compare the stochastic gradient-momentum process and the stochastic gradient process. We first introduce the dynamical system central to this section.

**Definition 2.** *The* homogeneous stochastic gradient-momentum process (hSGMP) *is a solution of the following stochastic differential equation,*

$$\begin{cases} \mathrm{d}q_t^{\nu,m} &= p_t^{\nu,m}\mathrm{d}t, \\ m\mathrm{d}p_t^{\nu,m} &= -\nabla\Phi_{i^\nu(t)}(q_t^{\nu,m})\mathrm{d}t - \alpha p_t^{\nu,m}\mathrm{d}t, \\ p_0^{\nu,m} &= p_0 \in X, \\ q_0^{\nu,m} &= q_0 \in X, \end{cases} \tag{4.14}$$

*where $\alpha, m, \nu > 0$ are the viscosity friction parameter, the mass, and the learning rate, respectively – all kept constant. Finally, $\Phi_j$ satisfies Assumption 5, for every $j \in I$ and the stochastic process $i^\nu(t) = i(t/\nu)$ is given as in Definition 1. In the following, we denote*

*(i)* $(p_t^\nu, q_t^\nu)_{t\geq 0} := (p_t^{\nu,1}, q_t^{\nu,1})_{t\geq 0}$*, i.e. choosing (without loss of generality) a constant unit mass* $m := 1$ *and varying the learning rate* $\nu$*.*

*(ii)* $(p_t^m, q_t^m)_{t\geq 0} := (p_t^{m^\delta,m}, q_t^{m^\delta,m})_{t\geq 0}$*, for some* $\delta \in [0, 1)$*, i.e. varying the mass, and choosing a constant or a mass-depending learning rate* $\nu := m^\delta$*; in this case, we additionally choose (without loss of generality) a unit viscosity friction* $\alpha := 1$*.*

The Lipschitz condition in Assumption 5 guarantees the well-posedness of this system, which can be shown similarly as in [58, 71]. In Subsections 4.2.1 and 4.2.2, we consider the hSGMPs $(p_t^\nu, q_t^\nu)_{t\geq 0}$ and $(p_t^m, q_t^m)_{t\geq 0}$, respectively.

### 4.2.1 Momentum with and without subsampling

First, we study the interplay between subsampling and not subsampling in the momentum dynamics. Indeed, we show that the stochastic gradient-momentum process can approximate the underdamped gradient flow at any accuracy. More precisely, we are interested in the limiting behavior of hSGMP $(p_t^\nu, q_t^\nu)_{t\geq 0}$ as the learning rate approaches zero uniformly, i.e. we take $\nu \to 0$ in (4.14). We prove that the hSGMP converges to the solution to the underdamped gradient flow, which is given by

$$\begin{cases} \mathrm{d}q_t &= p_t \mathrm{d}t, \\ \mathrm{d}p_t &= -\nabla\bar{\Phi}(q_t)\mathrm{d}t - \alpha p_t \mathrm{d}t, \\ p_0, q_0 &\in X, \end{cases} \tag{4.15}$$

where we implicitly set the mass $m := 1$. In this case, we show that hSGMP is a stochastic approximation to the underdamped gradient flow.

We now formulate the statement about the convergence of hSGMP to the underdamped gradient flow more particularly. In principle, we show convergence of a random path to a deterministic path – both are contained in the space of continuous functions from $[0,\infty)$ to $X^2$ which we denote by $\mathcal{C}([0,\infty), X^2)$ and equip with the metric

$$\rho\Big((\zeta_t)_{t\geq 0}, (\xi_t)_{t\geq 0}\Big) := \int_0^\infty e^{-t}\left(1 \wedge \sup_{0\leq s\leq t} \|\zeta_s - \xi_s\|\right) \mathrm{d}t,$$

where as usual $a \wedge b := \min\{a, b\}$ for $a, b \in \mathbb{R}$. Probabilistically, we show convergence in the weak sense. We state and prove the convergence result below.

**Theorem 6.** *Let $(q_t^\nu, p_t^\nu)_{t\geq 0}$ and $(q_t, p_t)_{t\geq 0}$ solve (4.14) and (4.15). Then $(q_t^\nu, p_t^\nu)_{t\geq 0}$ converges weakly to $(q_t, p_t)_{t\geq 0}$ in $\mathcal{C}([0,\infty), X^2)$ as $\nu \to 0$, i.e. for any bounded continuous function $F : \mathcal{C}([0,\infty), X^2) \to \mathbb{R}$,*

$$\mathbb{E}[F\big((p_t^\nu, q_t^\nu)_{t\geq 0}\big)] \to \mathbb{E}[F\big((p_t, q_t)_{t\geq 0}\big)] = F\big((p_t, q_t)_{t\geq 0}\big).$$

*Proof.* We apply Theorem 4.3 in [68, Chapter 7], which studies tightness and weak convergence properties of solutions of a certain type of stochastic differential equations. The form of the differential equation is given in (2.2) [68, Chapter 7], we also recall this result in Appendix A.4. This corresponds to equation (4.15) with $\xi_t = i(t)$, $\bar{G}(x,y) = (y, -\nabla_x\bar{\Phi}(x) - \alpha y)$, $\tilde{G}(x,y,i) = (0, -\nabla_x\Phi_i(x) + \nabla_x\bar{\Phi}(x))$, and $F = 0$. In [68, Chapter 7, Theorem 4.3], the differential operator $A$ in our case is defined in (4.14), i.e. $Ah(x,y) = -\nabla h(x,y)\cdot(y, -\nabla_x\bar{\Phi}(x) - \alpha y)$ for any $h$ twice differentiable. If Assumptions (A4.2) to (A4.6) from [68] were satisfied, applying Theorem 4.3 in [68, Chapter 7] would imply that $(q_t^\nu, p_t^\nu)_{t\geq 0}$ is tight and the limit of any weakly convergent subsequence solves (4.14). Since $q_0^\nu = q_0$ and $p_0^\nu = p_0$, we have $(q_t^\nu, p_t^\nu)_{t\geq 0}$ converges weakly to $(q_t, p_t)_{t\geq 0}$. Therefore, we just need to verify Assumptions (A4.2) to (A4.6) from [68]. Assumption (A4.2) follows directly from Assumption 5. Assumption (A4.3) holds since $(i(t))_{t\geq 0}$ is càdlàg and bounded. Assumptions (A4.4) and (A4.6) trivially hold since $F = 0$. The only non-trivial part is to verify Assumption (A4.5), that is

as $t, \tau \to \infty$,

$$\frac{1}{\tau}\int_t^{t+\tau} \mathbb{E}[-\nabla_x \Phi_{i(s)}(x) + \nabla_x \bar{\Phi}(x) | (i(s'))_{s' \leq t}] \mathrm{d}s \to 0,$$

almost surely. By the Markov property of $(i(t))_{t\geq 0}$ and since $i(0)$ is stationary,

$$\begin{aligned}
&\Big|\frac{1}{\tau}\int_t^{t+\tau} \mathbb{E}[-\nabla_x \Phi_{i(s)}(x) + \nabla_x \bar{\Phi}(x) | \{i(s')\}_{s' \leq t}] \mathrm{d}s\Big| \\
&\qquad = \Big|\frac{1}{\tau}\int_0^{\tau} \mathbb{E}_{i(t)}[-\nabla_x \Phi_{i(s)}(x) + \nabla_x \bar{\Phi}(x)] \mathrm{d}s\Big| \\
&\qquad = \Big|\frac{1}{\tau}\int_0^{\tau} \mathbb{E}_{i(t)}[\nabla_x \Phi_{i(s)}(x) - \nabla_x \bar{\Phi}(x)] \mathrm{d}s\Big| \\
&\qquad \leq \sum_{k=1}^N \Big|\frac{1}{\tau}\int_0^{\tau} \mathbb{E}_k[\nabla_x \Phi_{i(s)}(x) - \nabla_x \bar{\Phi}(x)] \mathrm{d}s\Big| \\
&\qquad = \frac{1}{\tau}\sum_{k=1}^N \Big| \int_0^{\tau} \Big( \sum_{j=1}^N \frac{1-\exp(-Ns)}{N} \nabla_x \Phi_j(x)\Big) + \exp(-Ns)\Phi_k(x) - \nabla_x \bar{\Phi}(x) \mathrm{d}s\Big| \\
&\qquad = \frac{1}{\tau}\sum_{k=1}^N \Big| \int_0^{\tau} \exp(-Ns)\Big(\nabla_x \Phi_k(x) - \nabla_x \bar{\Phi}(x)\Big) \mathrm{d}s\Big| \\
&\qquad \leq \frac{1}{N\tau}\sum_{k=1}^N \big|\nabla_x \Phi_k(x) - \nabla_x \bar{\Phi}(x)\big| \to 0.
\end{aligned}$$

as $\tau \to \infty$. □

Hence, hSGMP $(q_t^\nu, p_t^\nu)_{t\geq 0}$ is a stochastic approximation of the underdamped gradient flow $(q_t, p_t)_{t\geq 0}$. Assumption 6 on the other hand implies that the position of the underdamped gradient flow $q_t$ converges to the minimizer of $\bar{\Phi}$ as $t \to \infty$. This already gives us reason to believe that an SGMP is able to identify the minimizer of $\bar{\Phi}$. We discuss the longtime behaviour of the underdamped gradient flow in detail in Appendix A.1.

### 4.2.2 Subsampling with and without momentum

In a similar way to the last section, we now study the relation between the homogeneous stochastic gradient-momentum process and the stochastic gradient process. Indeed, we let

the momentum in hSGMP go to zero by letting the mass $\varepsilon$ go to zero. We show that in the limit, we converge to the stochastic gradient process. We do this while either keeping a constant learning rate or while allowing the learning rate to depend on the mass and thus go to zero as well. Specifically, we consider the process $(p_t^m, q_t^m)_{t\geq 0}$ introduced in Definition 2(ii). Namely, we set $\delta \in [0,1)$, learning rate $\nu := m^\delta$, and $\alpha := 1$. We then aim to show that $(q_t^{i,m})_{t\geq 0}$ in (4.15) converges uniformly to some stochastic gradient process as $\varepsilon \to 0$. The associated stochastic gradient process is the following:

$$\begin{cases} \mathrm{d}\theta_t^\varepsilon &= -\nabla\Phi_{i(t/m^\delta)}(\theta_t^\varepsilon)\mathrm{d}t, \\ \theta_0^\varepsilon &= \theta_0. \end{cases} \tag{4.16}$$

Throughout our discussion, we need several auxiliary results regarding momentum-to-no-momentum limit in the deterministic case. We have collected these results, that may be of independent interest, in Appendix A.2. We establish the connection between $(q_t^\varepsilon)_{t\geq 0}$ and $(\theta_t^\varepsilon)_{t\geq 0}$ in the following theorem by iterating over jump times of the index process. In each of the iteration steps, we employ a result regarding the underdamped gradient flow that we summarize and prove in Proposition 6 in Appendix A.2.

**Theorem 7.** *Let $\Phi_i$ be convex and satisfy Assumptions 5 and 6 with $\mathcal{A}^{\lambda_i,1}$ and critical point $\theta_*^i$ for $i \in I$. Let $(p_t^\varepsilon, q_t^\varepsilon)_{t\geq 0}$ and $(\theta_t^\varepsilon)_{t\geq 0}$ solve (4.15) and (4.16), respectively. For $0 \leq \delta < 1$ and $T > 0$, we have*

$$\begin{aligned} \mathbb{E}\Big[\sup_{0\leq t\leq T} \|q_t^\varepsilon - \theta_t^\varepsilon\|\Big] \leq \Big(1 + C_L'\varepsilon(1+T)\Big)e^{\varepsilon^{1-\delta}T(1+T)\gamma N}\Big(\|q_0-\theta_0\| + \varepsilon\|p_0\|\Big) \\ + \Big(e^{\varepsilon^{1-\delta}T(1+T)\gamma N} - 1 + C_L'\varepsilon(T+1)\Big)K_{\Phi,T,\theta_0}, \end{aligned}$$

*where $C_L', K_{\Phi,T,\theta_0} > 0$ are constants.*

*Proof.* We now denote by $(\tau_n)_{n\geq 0}$ and $(\tau_n^{\varepsilon,\delta})_{n\geq 0}$ the sequences of the jump times of processes $(i(t))_{t\geq 0}$ and $(i^{\varepsilon,\delta}(t))_{t\geq 0}$, respectively. We set $\tau_0 = 0$. From the definition, we know that the

last jump time before $t$ is $\tau_{N_t}$, where $(N_t)_{t\geq 0}$ is a Poisson process with rate $\gamma N$. And since $i^{\varepsilon,\delta}(t) = i(t/\varepsilon^\delta)$, we have $\tau_n^{\varepsilon,\delta} = \varepsilon^\delta \tau_n$. By Proposition 6, for $\tau_n^{\varepsilon,\delta} < t \leq \tau_{n+1}^{\varepsilon,\delta}$, we have

$$\begin{aligned}\|q_t^\varepsilon - \theta_t^\varepsilon\| \leq & \left\| q^\varepsilon_{\tau_n^{\varepsilon,\delta}} - \theta^\varepsilon_{\tau_n^{\varepsilon,\delta}} \right\| \\ & + C_L^{(0)} \varepsilon (1 + t - \tau_n^{\varepsilon,\delta}) \Big( \left\| q^\varepsilon_{\tau_n^{\varepsilon,\delta}} - \theta^\varepsilon_{\tau_n^{\varepsilon,\delta}} \right\| + \left\| p^\varepsilon_{\tau_n^{\varepsilon,\delta}} \right\| + \left\| \theta^\varepsilon_{\tau_n^{\varepsilon,\delta}} - \theta_*^{i^{\varepsilon,\delta}(\tau_n^{\varepsilon,\delta})} \right\| \Big)\end{aligned}$$

and from the first inequality in Lemma 23, we obtain

$$\|p_t^\varepsilon\| \leq (1 + C_L \varepsilon) \left\| p^\varepsilon_{\tau_n^{\varepsilon,\delta}} \right\| + C_L \left\| q^\varepsilon_{\tau_n^{\varepsilon,\delta}} - \theta^\varepsilon_{\tau_n^{\varepsilon,\delta}} \right\| + \left\| \theta^\varepsilon_{\tau_n^{\varepsilon,\delta}} - \theta_*^{i^{\varepsilon,\delta}(\tau_n^{\varepsilon,\delta})} \right\|.$$

From Lemma 26, we know

$$\left\| \theta^\varepsilon_{\tau_n^{\varepsilon,\delta}} - \theta_*^{i^{\varepsilon,\delta}(\tau_n^{\varepsilon,\delta})} \right\| \leq \left\| \theta^\varepsilon_{\tau_n^{\varepsilon,\delta}} \right\| + \left\| \theta_*^{i^{\varepsilon,\delta}(\tau_n^{\varepsilon,\delta})} \right\| \leq \|\theta_0\| + C_\Phi T + K_{\Theta_*} := K_{\Phi,T,\theta_0}.$$

So we have

$$\begin{aligned}&\sup_{0\leq t\leq \tau_{n+1}^{\varepsilon,\delta}} \|q_t^\varepsilon - \theta_t^\varepsilon\| \leq \max\Big\{ \sup_{0\leq t\leq \tau_n^{\varepsilon,\delta}} \|q_t^\varepsilon - \theta_t^\varepsilon\|, \sup_{\tau_n^{\varepsilon,\delta}\leq t\leq \tau_{n+1}^{\varepsilon,\delta}} \|q_t^\varepsilon - \theta_t^\varepsilon\| \Big\} \\ &\leq \sup_{0\leq t\leq \tau_n^{\varepsilon,\delta}} \|q_t^\varepsilon - \theta_t^\varepsilon\| + C_L^{(0)} \varepsilon (1 + \tau_{n+1}^{\varepsilon,\delta} - \tau_n^{\varepsilon,\delta}) \Big( \left\| q^\varepsilon_{\tau_n^{\varepsilon,\delta}} - \theta^\varepsilon_{\tau_n^{\varepsilon,\delta}} \right\| + \left\| p^\varepsilon_{\tau_n^{\varepsilon,\delta}} \right\| + K_{\Phi,T,\theta_0} \Big).\end{aligned}$$

For $T > 0$, if we assume $\tau_{n+1}^{\varepsilon,\delta} \leq T$, we have

$$\begin{aligned}\sup_{0\leq t\leq \tau_{n+1}^{\varepsilon,\delta}} \|q_t^\varepsilon - \theta_t^\varepsilon\| \leq & \sup_{0\leq t\leq \tau_n^{\varepsilon,\delta}} \|q_t^\varepsilon - \theta_t^\varepsilon\| \\ & + C_L^{(0)} \varepsilon (1 + T) \Big( \sup_{0\leq t\leq \tau_n^{\varepsilon,\delta}} \|q_t^\varepsilon - \theta_t^\varepsilon\| + \sup_{0\leq t\leq \tau_n^{\varepsilon,\delta}} \|p_t^\varepsilon\| + K_{\Phi,T,\theta_0} \Big)\end{aligned}$$

and

$$\sup_{0\leq t\leq \tau_{n+1}^{\varepsilon,\delta}} \|p_t^\varepsilon\| \leq (1 + C_L \varepsilon) \sup_{0\leq t\leq \tau_n^{\varepsilon,\delta}} \|p_t^\varepsilon\| + C_L \Big( \sup_{0\leq t\leq \tau_n^{\varepsilon,\delta}} \|q_t^\varepsilon - \theta_t^\varepsilon\| + K_{\Phi,T,\theta_0} \Big).$$

We denote

$$A_n^{\varepsilon,\delta} := \sup_{0\le t\le \tau_n^{\varepsilon,\delta}} \|q_t^\varepsilon - \theta_t^\varepsilon\|\,,\ B_n^{\varepsilon,\delta} := \sup_{0\le t\le \tau_n^{\varepsilon,\delta}} \|p_t^\varepsilon\|\,,\ D_n^{\varepsilon,\delta} = A_n^{\varepsilon,\delta} + \varepsilon B_n^{\varepsilon,\delta}.$$

and then rewrite the inequalities

$$\begin{aligned} A_{n+1}^{\varepsilon,\delta} &\le A_n^{\varepsilon,\delta} + \varepsilon C_L^{(0)}(1+T)\Big(A_n^{\varepsilon,\delta} + B_n^{\varepsilon,\delta} + K_{\Phi,T,\theta_0}\Big),\\ B_{n+1}^{\varepsilon,\delta} &\le (1+C_L\varepsilon)B_n^{\varepsilon,\delta} + C_L\Big(A_n^{\varepsilon,\delta} + K_{\Phi,T,\theta_0}\Big). \end{aligned}$$

Since $\varepsilon < 1$, there exists a constant $C_L'$ such that

$$D_{n+1}^{\varepsilon,\delta} \le \Big(1 + C_L'\varepsilon(1+T)\Big)D_n^{\varepsilon,\delta} + C_L'\varepsilon(1+T)K_{\Phi,T,\theta_0}. \tag{4.17}$$

Hence, for $n \ge 0$, we have

$$D_n^{\varepsilon,\delta} \le \Big(1 + C_L'\varepsilon(1+T)\Big)^n\Big(\,\|q_0-\theta_0\| + \|p_0\|\,\Big) + \Big(\Big(1 + C_L'\varepsilon(1+T)\Big)^n - 1\Big)K_{\Phi,T,\theta_0}. \tag{4.18}$$

Let $n = N_{T/\varepsilon^\delta}$. From (4.18) and by using the moment-generating function of the Poisson distribution, we conclude:

$$\begin{aligned} \mathbb{E}[D_{N_{T/\varepsilon^\delta}}^{\varepsilon,\delta}] &\le \Big(\,\|q_0-\theta_0\| + m\,\|p_0\|\,\Big)\mathbb{E}\Big[\Big(1 + C_L'\varepsilon(1+T)\Big)^{N_{T/\varepsilon^\delta}}\Big]\\ &\qquad + K_{\Phi,T,\theta_0}\mathbb{E}\Big[\Big(1 + C_L'\varepsilon(1+T)\Big)^{N_{T/\varepsilon^\delta}} - 1\Big]\\ &= e^{\varepsilon^{1-\delta}T(1+T)\gamma N}\Big(\,\|q_0-\theta_0\| + C_L m\,\|p_0\|\,\Big) + \Big(e^{\varepsilon^{1-\delta}T(1+T)\gamma N} - 1\Big)K_{\Phi,T,\theta_0}. \end{aligned}$$

For any $T \geq 0$, there exists an integer $n \geq 0$, such that $\tau_n^{\varepsilon,\delta} \leq T \leq \tau_{n+1}^{\varepsilon,\delta}$ (i.e. $n = N_{T/\varepsilon^\delta}$). Hence,

$$\begin{aligned}
\mathbb{E}\Big[ \sup_{0\leq t\leq T} \|q_t^\varepsilon - \theta_t^\varepsilon\| \Big] &\leq \mathbb{E}\Big[ \sup_{0\leq t\leq \tau_{n+1}^{\varepsilon,\delta}} \|q_t^\varepsilon - \theta_t^\varepsilon\| \Big] \leq \mathbb{E}[D^{\varepsilon,\delta}_{N_{T/\varepsilon^\delta}+1}] \\
&\overset{(4.17)}{\leq} \Big(1 + C_L'\varepsilon(1+T)\Big)\mathbb{E}[D^{\varepsilon,\delta}_{N_{T/\varepsilon^\delta}}] + C_L'\varepsilon(1+T)K_{\Phi,T,\theta_0} \\
&\leq \Big(1 + C_L'\varepsilon(1+T)\Big)e^{\varepsilon^{1-\delta}T(1+T)\gamma N}\Big( \|q_0 - \theta_0\| + \varepsilon\|p_0\| \Big) \\
&\quad + \Big(e^{\varepsilon^{1-\delta}T(1+T)\gamma N} - 1 + C_L'\varepsilon(T+1)\Big)K_{\Phi,T,\theta_0}.
\end{aligned}$$

□

Thus, if we let the mass $\varepsilon$ go down to zero and the learning rate parameter $\nu$ is independent of $m$ or $\nu$ goes to zero at a certain rate that is slower than $m$, the stochastic gradient-momentum process converges to the stochastic gradient process. The speed of convergence is linear in $m$, which is also the case in the deterministic setting. Note that we show convergence in a possibly weaker sense as compared with the learning rate result in Theorem 6.

We finish our discussion of the homogeneous-in-time setting with a final remark regarding reducing mass and learning rate at the same time.

**Remark 9.** *From [71, Theorem 1] and [73], we already know that*

$$(\theta_t^m)_{t\geq 0} \to (\zeta_t)_{t\geq 0}$$

*if $m \to 0$, i.e. the stochastic gradient process converges to the gradient flow as the learning rate approaches zero. Convergence is in the weak sense probabilistically and a weighted $\infty$-norm*

$$\rho'\Big((\zeta_t)_{t\geq 0}, (\xi_t)_{t\geq 0}\Big) := \sum_{n=1}^{\infty} 2^{-n}\left(1 \wedge \sup_{0\leq s\leq n} \|\zeta_s - \xi_s\|\right),$$

*in space. Combined with Theorem 7, we can easily see that also* $(q_t^m)_{t\geq 0} \to (\zeta_t)_{t\geq 0}$ *in the case* $\delta \in (0,1)$*, where the convergence is in the sense of Theorem 7.*

## 4.3 Heterogeneous-in-time

In the previous section, we have studied the effect of losing momentum and reducing the learning rate in a uniform fashion. Having a homogeneous mass and learning rate is unquestionably very popular in practice. However, while we could show convergence to the minimizer of the deterministic dynamics, this is not clear for the stochastic dynamics. To obtain convergence in SGMP, we need to reduce both, learning rate and momentum, over time. Hence, we need to discuss a heterogeneous version of the SGMP. Indeed, we now allow the mass $m$ to be a non-constant function of time, especially to be reduced to zero over time, and the index process $(i^\beta(t))_{t\geq 0}$ to have waiting times that get smaller as time progresses. Thus we study the case where the stochastic gradient-momentum process converges to the stochastic gradient process and the gradient flow, respectively, at runtime. We have already seen the advantage of reducing mass over time in the deterministic setting in Figure 4.2.

We split this section into two parts. In Subsection 4.3.1, we let only the momentum depend on and decrease over time. We introduce the dynamical system, show well-posedness, and discuss the longtime behaviour of the new dynamical system. We see that asymptotically, the dynamical system behaves like the stochastic gradient process. Then, in Subsection 4.3.2, we decrease mass and learning rate over time. In this case, we discuss a setting, in which we see convergence of the stochastic dynamical system to the minimizer $\theta_*$ of the target function $\bar{\Phi}$.

### 4.3.1 Losing momentum over time

We consider the stochastic gradient-momentum process with a time-dependent, decreasing mass.

**Definition 3.** *The* decreasing-mass stochastic gradient-momentum process *(dmSGMP) is a solution of the following stochastic differential equation,*

$$\begin{cases} \mathrm{d}\tilde{q}_t &= \tilde{p}_t \mathrm{d}t, \\ \varepsilon(t)\mathrm{d}\tilde{p}_t &= -\nabla\Phi_{i(t)}(\tilde{q}_t)\mathrm{d}t - \tilde{p}_t \mathrm{d}t, \\ \tilde{p}_0, \tilde{q}_0 &\in X, \end{cases} \tag{4.19}$$

*where $\Phi_j$ satisfies Assumption 5, $j = 1, \cdots, N$. The stochastic process $(i(t))_{t\geq 0}$ is defined in Definition 1. The mass $\varepsilon(t) > 0$ is strictly decreasing and differentiable with $\lim_{t\to\infty} \varepsilon(t) = 0$.*

The formal limit of (4.19) is the stochastic gradient process without momentum, as given in (4.3). As in the previous section, we start our discussion considering the decreasing mass case with a fixed index, i.e. the following dynamical system:

$$\begin{cases} \mathrm{d}q_t^i &= p_t^i \mathrm{d}t, \\ \varepsilon(t)\mathrm{d}p_t^i &= -\nabla\Phi_i(q_t^i)\mathrm{d}t - p_t^i \mathrm{d}t, \\ p_0^i, q_0^i &\in X, \end{cases} \tag{4.20}$$

where the mass $(\varepsilon(t))_{t\geq 0}$ is chosen as before. In addition, we denote $\mathcal{E}(t) := \int_0^t 1/\varepsilon(s)\mathrm{d}s$. The associated limiting equation is given by (3.4). Next, we discuss the well-posedness of the system (4.20).

### Well-posedness

We study the well-posedness of the deterministic dynamical system (4.20). Well-posedness of the stochastic dynamical system (4.19) then follows from, e.g. Proposition 1 in [71].

**Proposition 3.** *For any fixed $i \in I$, let $\nabla\Phi_i$ satisfy Assumption 5 with constant $L$ and $\varepsilon(t)$ satisfy Assumption 7. Then* (4.20) *admits a unique solution.*

*Proof.* We first rewrite equation (4.20) as

$$\begin{cases} \mathrm{d}q_t^i &= p_t^i \mathrm{d}t, \\ \mathrm{d}p_t^i &= -\nabla\Phi_i(q_t^i)/\varepsilon(t)\mathrm{d}t - p_t^i/\varepsilon(t)\mathrm{d}t, \\ p_0^i, q_0^i &\in X. \end{cases}$$

We define a map $\mathcal{T}_0 : \mathcal{C}([0,t_1],\mathbb{R}) \times \mathcal{C}([0,t_1],\mathbb{R}) \to \mathcal{C}([0,t_1],\mathbb{R}) \times \mathcal{C}([0,t_1],\mathbb{R})$, as

$$\mathcal{T}_0(x,y) = \Big(q_0 + \int_0^t y_s \mathrm{d}s,\ p_0 - \int_0^t \nabla\Phi_i(x_s)/\varepsilon(s)\mathrm{d}s - \int_0^t y_s/\varepsilon(s)\mathrm{d}s\Big)$$

where $0 \le t \le t_1$. Next, we show that $\mathcal{T}_0$ is a contraction for $t_1$ small enough. Since $\varepsilon(t)$ is strictly decreasing and positive and using Assumption 7, we can conclude that

$$(e^{\lambda t}\varepsilon(t))' \ge 0 \quad \text{and} \quad \varepsilon(t) \ge \varepsilon_0 e^{-\lambda t}.$$

We then conclude

$$\begin{aligned} \|\mathcal{T}_0(q,p) - \mathcal{T}_0(\hat{q},\hat{p})\|_\infty &\le t_1 \|p-\hat{p}\|_\infty + L\|q-\hat{q}\|_\infty \int_0^{t_1} \frac{1}{\varepsilon(s)}\mathrm{d}s + \|p-\hat{p}\|_\infty \int_0^{t_1} \frac{1}{\varepsilon(s)}\mathrm{d}s \\ &\le (\|q-\hat{q}\|_\infty + \|p-\hat{p}\|_\infty)\frac{L+1}{\lambda\varepsilon_0}(e^{\lambda t_1} - 1 + t_1). \end{aligned}$$

Set $t_1 = \frac{\lambda\varepsilon_0}{2(L+1)(2\lambda+1)} =: C(\lambda,\varepsilon_0,L)$, and since $e^{\lambda x} - 1 \le 2\lambda x$ when $\lambda x \le 1$, we have

$$\frac{L+1}{\lambda\varepsilon_0}(e^{\lambda t_1} - 1 + t_1) \le \frac{L+1}{\lambda\varepsilon_0}(2\lambda+1)t_1 \le \frac{1}{2}.$$

Then $\mathcal{T}_0$ is contracting with constant smaller than $1/2$. By the Banach Fixed Point Theorem, we conclude that (4.20) admits a unique solution on the interval $[0,t_1]$ with initial value $(q_0,p_0)$. Let $\mathcal{T}_n : \mathcal{C}([t_n,t_{n+1}],\mathbb{R}) \times \mathcal{C}([t_n,t_{n+1}],\mathbb{R}) \to \mathcal{C}([t_n,t_{n+1}],\mathbb{R}) \times \mathcal{C}([t_n,t_{n+1}],\mathbb{R})$, with

$$\mathcal{T}_n(x,y) = \Big(q_{t_n} + \int_{t_n}^t y_s/\varepsilon(s)\mathrm{d}s,\ p_{t_n} - \int_{t_n}^t \nabla\Phi_i(x_s)/\varepsilon(s)\mathrm{d}s - \int_{t_n}^t y_s/\varepsilon(s)\mathrm{d}s\Big).$$

Then we have

$$\begin{aligned}\left\|\mathcal{T}_n(q,p)-\mathcal{T}_n(\hat{q},\hat{p})\right\|_\infty &\leq (t_{n+1}-t_n)\left\|p-\hat{p}\right\|_\infty + L\left\|q-\hat{q}\right\|_\infty \int_{t_n}^{t_{n+1}} \frac{1}{\varepsilon(s)}\mathrm{d}s \\ &\quad + \left\|p-\hat{p}\right\|_\infty \int_{t_n}^{t_{n+1}} \frac{1}{\varepsilon(s)}\mathrm{d}s \\ &\leq (\left\|q-\hat{q}\right\| + \left\|p-\hat{p}\right\|_\infty)\frac{L+1}{\lambda\varepsilon_0}e^{\lambda t_n}(e^{\lambda h_n}-1+h_n),\end{aligned}$$

where $h_n = t_{n+1}-t_n$. Set $h_n = C(\lambda,\varepsilon_0,L)e^{-\lambda t_n} = \frac{\lambda\varepsilon_0 e^{-\lambda t_n}}{2(L+1)(2\lambda+1)}$, then

$$\frac{L+1}{\lambda\varepsilon_0}e^{\lambda t_n}(e^{\lambda h_n}-1+h_n) \leq \frac{L+1}{\lambda\varepsilon_0}e^{\lambda t_n}(2\lambda+1)h_n \leq \frac{1}{2}.$$

This implies $\mathcal{T}_n$ is a contraction, and by the Banach Fixed Point Theorem, we conclude that (4.20) admits a unique solution on the interval $[t_n, t_{n+1}]$ with initial value $(\tilde{q}_{t_n}, \tilde{p}_{t_n})$. Therefore, by Lemma 28, we have $t_n \to \infty$, and, thus, (4.20) admits a unique solution globally. □

As the unique solution to the dynamical system (4.20) is a strong solution, the solution is continuously differentiable. Thus, the map $(p_0^i, p_0^i) \mapsto (p_t^i, p_t^i)$ is continuous for any $t \geq 0$. Hence, the differential equations (4.19) and (4.20) are well-posed.

**Longtime behavior**

Having understood the well-posedness of dmSGMP, we are now interested in its longtime behaviour. By reducing the mass over time, we aim to reduce the momentum and prove convergence to the stochastic gradient process (4.3).

To show this convergence result, we now collect a number of auxiliary results that mainly concern the deterministic system (4.20) that considers a single, fixed sample. Some of these results may remind the reader of similar results shown in Appendix A.2 regarding the fixed sample, homogeneous momentum dynamics; we make these connections clear in the titles of the following results. The intermediate goal is to show a bound between the system (4.20)

and its limiting equation (3.4). From there we then go back to the randomized setting and discuss the longtime behaviour and convergence of dmSGMP. We first show that the deterministic system converges at exponential speed to its unique stationary point.

**Lemma 19** (cf. Lemma 22)**.** *Let $(p_t^i, q_t^i)_{t\geq 0}$ solve (4.20). Let $\theta_*^i$ be the critical point of $\Phi_i$ for $i \in I$. Under the Assumptions 5, 6 with $\mathcal{A}^{\lambda_i,2}$, and 7 with $\lambda = \min_{i\in I} \lambda_i$, we have*

$$\left\|q_t^i - \theta_*^i\right\|^2 \leq 8e^{-\lambda t} V^i(0, q_0^i, p_0^i) \qquad (t \geq 0), \tag{4.21}$$

*where $V^i(t,x,y) = \varepsilon(t)(\Phi_i(x) - \Phi_i(\theta_*^i)) + \frac{1}{4}\Big( \|x - \theta_*^i + \varepsilon(t)y\|^2 + \|\varepsilon(t)y\|^2 \Big)$.*

*Proof.* By Lemma 29, we have $\|q_t^i - \theta_*^i\|^2 \leq 8V^i(t, q_t^i, p_t^i) \leq 8e^{-\lambda t} V^i(0, q_0^i, p_0^i)$. □

In the next auxiliary result, we show Lipschitz continuity of the position $(q_t^i)_{t\geq 0}$ and boundedness of the velocity $(p_t^i)_{t\geq 0}$.

**Lemma 20** (cf. Lemma 23)**.** *Let $(p_t^i, q_t^i)_{t\geq 0}$ solve (4.20). Let $\theta_*^i$ be the critical point of $\Phi_i$ for $i \in I$. Under the Assumptions 5, 6 with $\mathcal{A}^{\lambda_i,2}$, and 7 with $\lambda = \min_{i\in I} \lambda_i$, for any $0 \leq s \leq t$, we have*

$$\begin{aligned} \left\|q_t^i - q_s^i\right\| &\leq C_L^{(1)}(t-s)\Big( \left\|q_0^i - \theta_*^i\right\| + (\varepsilon_0 + e^{-\mathcal{E}(s)}) \left\|p_0^i\right\| \Big), \\ \left\|p_t^i\right\| &\leq e^{-\mathcal{E}(t)} \left\|p_0^i\right\| + C_L^{(1)}(\left\|q_0^i - \theta_*^i\right\| + \varepsilon_0 \left\|p_0^i\right\|) \qquad (t \geq 0), \end{aligned}$$

*where $C_L^{(1)} = 1 + 8L^2 + 8L$.*

*Proof.* From (4.20), we have

$$\frac{\mathrm{d}(e^{\mathcal{E}(t)} p_t^i)}{\mathrm{d}t} = -e^{\mathcal{E}(t)} \nabla \Phi_i(q_t^i)/\varepsilon(t),$$

which implies

$$p_t^i = e^{-\mathcal{E}(t)} p_0^i - e^{-\mathcal{E}(t)} \int_0^t e^{\mathcal{E}(s)} \nabla \Phi_i(q_s^i)/\varepsilon(s) \mathrm{d}s. \tag{4.22}$$

Hence,

$$\begin{aligned}
\left\|p_t^i\right\| \leq & e^{-\mathcal{E}(t)}\left\|p_0^i\right\| + e^{-\mathcal{E}(t)}\int_0^t \frac{e^{\mathcal{E}(s)}}{\varepsilon(s)}\left\|\nabla\Phi_i(q_s^i)\right\|\mathrm{d}s \\
= & e^{-\mathcal{E}(t)}\left\|p_0^i\right\| + e^{-\mathcal{E}(t)}\int_0^t \frac{e^{\mathcal{E}(s)}}{\varepsilon(s)}\left\|\nabla\Phi_i(q_s^i) - \nabla\Phi_i(\theta_*^i)\right\|\mathrm{d}s \\
\leq & e^{-\mathcal{E}(t)}\left\|p_0^i\right\| + Le^{-\mathcal{E}(t)}\int_0^t \frac{e^{\mathcal{E}(s)}}{\varepsilon(s)}\left\|q_s^i - \theta_*^i\right\|\mathrm{d}s \\
\overset{(4.21)}{\leq} & e^{-\mathcal{E}(t)}\left\|p_0^i\right\| + 8Le^{-\mathcal{E}(t)}(V^i(0,q_0^i,p_0^i))^{\frac{1}{2}}\int_0^t \frac{e^{\mathcal{E}(s)}}{\varepsilon(s)}e^{\frac{-\lambda s}{2}}\mathrm{d}s \\
\leq & e^{-\mathcal{E}(t)}\left\|p_0^i\right\| + 8L(V^i(0,q_0^i,p_0^i))^{\frac{1}{2}}.
\end{aligned}$$

The last inequality holds since $\int_0^t \frac{e^{\mathcal{E}(s)}}{\varepsilon(s)}e^{\frac{-\lambda s}{2}}\mathrm{d}s \leq e^{\mathcal{E}(t)}$, which can be seen by taking the derivative on both sides and comparing the initial values. And since $V^i(0,q_0^i,p_0^i) \leq (1+L)(\|q_0^i-\theta_*^i\|^2 + \varepsilon_0^2\|p_0^i\|^2)$, we have

$$\left\|p_t^i\right\| \leq e^{-\mathcal{E}(t)}\left\|p_0^i\right\| + (1+8L^2+8L)(\left\|q_0^i-\theta_*^i\right\| + \varepsilon_0\left\|p_0^i\right\|).$$

We set $C_L^{(1)} = (1+8L^2+8L)$ and have

$$\begin{aligned}
\left\|q_t^i - q_s^i\right\| \leq & \int_s^t \left\|p_m^i\right\|\mathrm{d}m \leq C_L^{(1)}(t-s)(\left\|q_0^i-\theta_*^i\right\| + \varepsilon_0\left\|p_0^i\right\|) + \left\|p_0^i\right\|\int_s^t e^{-\mathcal{E}(x)}\mathrm{d}x \\
\leq & C_L^{(1)}(t-s)\left(\left\|q_0^i-\theta_*^i\right\| + (\varepsilon_0 + e^{-\mathcal{E}(s)})\left\|p_0^i\right\|\right).
\end{aligned}$$

□

In the third lemma, we give a bound on the system's acceleration $(\mathrm{d}p_t^i/\mathrm{d}t)_{t\geq 0}$.

**Lemma 21** (cf. Lemma 24)**.** *Let $(p_t^i, q_t^i)_{t\geq 0}$ solve (4.20). Let $\theta_*^i$ be the critical point of $\Phi_i$ for $i \in I$. Under the Assumptions 5, 6 with $\mathcal{A}^{\lambda_i,2}$, and 7 with $\lambda = \min_{i\in I}\lambda_i$, we have*

$$\left\|\frac{\mathrm{d}p_t^i}{\mathrm{d}t}\right\| \leq C_L^{(2)}\left(\frac{e^{-\mathcal{E}(t)}}{\varepsilon(t)} + 2\right)(\left\|p_0^i\right\| + \left\|q_0^i-\theta_*^i\right\|) \qquad (t\geq 0), \tag{4.23}$$

*where* $C_L^{(2)} = 8(C_L^{(1)})^2 = 8(1+8L^2+8L)^2$.

*Proof.* From (4.22), we have

$$\frac{\mathrm{d}p_t^i}{\mathrm{d}t} = \frac{-e^{-\mathcal{E}(t)}p_0^i}{\varepsilon(t)} - \frac{\nabla\Phi_i(q_t^i)}{\varepsilon(t)} + \frac{e^{-\mathcal{E}(t)}}{\varepsilon(t)}\int_0^t e^{\mathcal{E}(s)}\nabla\Phi_i(q_s^i)/\varepsilon(s)\mathrm{d}s.$$

Notice that

$$\frac{\nabla\Phi_i(q_t^i)}{\varepsilon(t)} = \frac{e^{-\mathcal{E}(t)}}{\varepsilon(t)}\int_0^t e^{\mathcal{E}(s)}\nabla\Phi_i(q_t^i)/\varepsilon(s)\mathrm{d}s + \frac{e^{-\mathcal{E}(t)}\nabla\Phi_i(q_t^i)}{\varepsilon(t)}.$$

Then, we can rewrite $\frac{\mathrm{d}p_t^i}{\mathrm{d}t}$ as

$$\frac{\mathrm{d}p_t^i}{\mathrm{d}t} = \frac{-e^{-\mathcal{E}(t)}(p_0^i + \nabla\Phi_i(q_t^i))}{\varepsilon(t)} + \frac{e^{-\mathcal{E}(t)}}{\varepsilon(t)}\int_0^t e^{\mathcal{E}(s)}(\nabla\Phi_i(q_s^i) - \nabla\Phi_i(q_t^i))/\varepsilon(s)\mathrm{d}s.$$

Hence,

$$\begin{aligned}
\left\|\frac{\mathrm{d}p_t^i}{\mathrm{d}t}\right\| &\le \frac{e^{-\mathcal{E}(t)}}{\varepsilon(t)}\Big(\left\|p_0^i + \nabla\Phi_i(q_t^i) - \nabla\Phi_i(\theta_*^i)\right\| + \int_0^t e^{\mathcal{E}(s)}\left\|\nabla\Phi_i(q_s^i) - \nabla\Phi_i(q_t^i)\right\|/\varepsilon(s)\mathrm{d}s\Big)\\
&\le \frac{(1+L)e^{-\mathcal{E}(t)}}{\varepsilon(t)}\Big(\left\|p_0^i\right\| + \left\|q_t^i - \theta_*^i\right\| + \int_0^t e^{\mathcal{E}(s)}\left\|q_s^i - q_t^i\right\|/\varepsilon(s)\mathrm{d}s\Big)\\
&\underbrace{\le}_{(a_2)} \frac{4C_L^{(1)}e^{-\mathcal{E}(t)}}{\varepsilon(t)}\Big(\left\|q_0^i - \theta_*^i\right\| + \left\|p_0^i\right\|\\
&\qquad + C_L^{(1)}\int_0^t (t-s)e^{\mathcal{E}(s)}/\varepsilon(s)\mathrm{d}s\Big(\left\|q_0^i - \theta_*^i\right\| + (\varepsilon_0 + e^{-\mathcal{E}(s)})\left\|p_0^i\right\|\Big)\Big)\\
&= \frac{4C_L^{(1)}e^{-\mathcal{E}(t)}}{\varepsilon(t)}\left\|q_0^i - \theta_*^i\right\|\Big(1 + C_L^{(1)}\int_0^t (t-s)e^{\mathcal{E}(s)}/\varepsilon(s)\mathrm{d}s\Big)\\
&+ \frac{4C_L^{(1)}e^{-\mathcal{E}(t)}}{\varepsilon(t)}\left\|p_0^i\right\|\Big(1 + C_L^{(1)}\int_0^t (t-s)e^{\mathcal{E}(s)}(\varepsilon_0 + e^{-\mathcal{E}(s)})/\varepsilon(s)\mathrm{d}s\Big)\\
&\le 4C_L^{(1)}\Big(\frac{e^{-\mathcal{E}(t)}}{\varepsilon(t)} + 2C_L^{(1)}\Big)\left\|q_0^i - \theta_*^i\right\| + 4C_L^{(1)}\Big(\frac{e^{-\mathcal{E}(t)}}{\varepsilon(t)} + 4C_L^{(1)}\Big)\left\|p_0^i\right\|,
\end{aligned}$$

where the last step follows since

$$\int_0^t (t-s)e^{\mathcal{E}(s)}/\varepsilon(s)\mathrm{d}s \leq 2\varepsilon(t)e^{\mathcal{E}(t)}$$

and the inequality $(a_2)$ follows from Lemmas 19 and 20 and the fact that $V^i(0, q_0^i, p_0^i) \leq (1+L)(\|q_0^i - \theta_*^i\|^2 + \varepsilon_0^2 \|p_0^i\|^2)$. Hence,

$$\begin{aligned}\left\|\frac{\mathrm{d}p_t^i}{\mathrm{d}t}\right\| &\leq 4C_L^{(1)}\Big(\frac{e^{-\mathcal{E}(t)}}{\varepsilon(t)} + 4C_L^{(1)}\Big)(\|p_0^i\| + \|q_0^i - \theta_*^i\|) \\ &\leq 8(C_L^{(1)})^2\Big(\frac{e^{-\mathcal{E}(t)}}{\varepsilon(t)} + 2\Big)(\|p_0^i\| + \|q_0^i - \theta_*^i\|).\end{aligned}$$

We set $C_L^{(2)} = 8(C_L^{(1)})^2 = 8(1+8L^2+8L)^2$, which completes the proof. □

We have now attained the aforementioned intermediate goal where we show a bound between the system (4.20) and its limiting equation (3.4). We need to additionally ask for some underlying convexity; strong convexity in this case: Indeed, we aim to show that $q_{\tau_n} - \theta_{\tau_n}$ is contractive. Assumption 6 would only lead to contractivity of $q_{\tau_n} - \theta_*$.

**Proposition 4** (cf. Proposition 6)**.** *Let $(p_t^i, q_t^i)_{t\geq 0}$ solve (4.20). Let $\theta_*^i$ be the critical point of $\Phi_i$ for $i \in I$. We assume $\Phi_i$ is strongly convex with constant $\mu$ and let $\Phi_i$ satisfy Assumption 5 and $(\varepsilon(t))_{t\geq 0}$ satisfy Assumption 7 with $\lambda = (\mu/4) \wedge (1/4)$. Then we have*

$$\|q_t^i - \theta_t^i\| \leq e^{-\mu t/2}\|q_0^i - \theta_0^i\| + C_{L,\mu}^{(1)}\varepsilon_0^{\frac{1}{2}}\Big(\|p_0^i\| + \|q_0^i - \theta_*^i\|\Big) \qquad (t \geq 0),$$

*where $C_{L,\mu}^{(1)} = 2C_L^{(2)}\mu^{-1/2}(1+\mu^{-1/2}+\mu^{1/2}) + \mu^{1/2}$, which is larger than $\mu^{1/2}$.*

*Proof.* By Lemma 21 and Lemma 25, we have

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}\left\|q_t^i-\theta_t^i\right\|^2}{\mathrm{d}t} &= \left\langle q_t^i-\theta_t^i\right\rangle \frac{\mathrm{d}q_t^i}{\mathrm{d}t}-\frac{\mathrm{d}\theta_t^i}{\mathrm{d}t} \\
&= -\left\langle q_t^i-\theta_t^i\right\rangle \nabla\Phi_i(q_t^i)-\nabla\Phi_i(\theta_t^i)-\varepsilon(t)\left\langle q_t^i-\theta_t^i\right\rangle \frac{\mathrm{d}p_t^i}{\mathrm{d}t} \\
&\le -\mu\left\|q_t^i-\theta_t^i\right\|^2+\frac{\mu}{2}\left\|q_t^i-\theta_t^i\right\|^2+\frac{\varepsilon^2(t)}{2\mu}\left\|\frac{\mathrm{d}p_t^i}{\mathrm{d}t}\right\|^2 \\
&\overset{(4.23)}{\le} -\frac{\mu}{2}\left\|q_t^i-\theta_t^i\right\|^2+(C_L^{(2)})^2(\mu)^{-1}\varepsilon^2(t)\Big(\frac{e^{-2\mathcal{E}(t)}}{\varepsilon^2(t)}+4\Big)\Big(\left\|p_0^i\right\|^2+\left\|q_0^i-\theta_*^i\right\|^2\Big) \\
&= -\frac{\mu}{2}\left\|q_t^i-\theta_t^i\right\|^2+(2C_L^{(2)})^2(\mu)^{-1}\Big(e^{-2\mathcal{E}(t)}+\varepsilon^2(t)\Big)\Big(\left\|p_0^i\right\|^2+\left\|q_0^i-\theta_*^i\right\|^2\Big),
\end{aligned}$$

which implies

$$\begin{aligned}
\left\|q_t^i-\theta_t^i\right\|^2 \le\ & e^{-\mu t}\left\|q_0^i-\theta_0^i\right\|^2 \\
&+(C_L^{(2)})^2(\mu)^{-1}e^{-\mu t}\int_0^t\Big(e^{-2\mathcal{E}(s)}+\varepsilon^2(s)\Big)e^{\mu s}\mathrm{d}s\Big(\left\|p_0^i\right\|^2+\left\|q_0^i-\theta_*^i\right\|^2\Big).
\end{aligned}$$

From Lemma 30, we know that $e^{-\mu t}\int_0^t\varepsilon^2(s)e^{\mu s}\mathrm{d}s \le 2\mu^{-1}\varepsilon^2(t) \le 2\mu^{-1}\varepsilon_0^2 \le 2\mu^{-1}\varepsilon_0$, and $e^{-\mu t}\int_0^t e^{-2\mathcal{E}(s)}e^{\mu s}\mathrm{d}s \le (1+\mu)\varepsilon_0$, which implies

$$\left\|q_t^i-\theta_t^i\right\|^2 \le e^{-\mu t}\left\|q_0^i-\theta_0^i\right\|^2+(C_{L,\mu}^{(1)})^2\varepsilon_0\Big(\left\|p_0^i\right\|^2+\left\|q_0^i-\theta_*^i\right\|^2\Big).$$

□

From Lemma 20 and Proposition 4, we immediately have the following corollary.

**Corollary 5.** *Under the same condition as Proposition 4. For $0<\varepsilon_0\le 1$ and any fixed $i\in I$, we have*

$$\begin{aligned}
\left\|q_t^i-\theta_t^i\right\| &\le e^{-\mu t/2}\left\|q_0^i-\theta_0^i\right\|+C_{L,\mu}^{(1)}\varepsilon_0\Big(\left\|p_0^i\right\|+\left\|q_0^i-\theta_0^i\right\|+\left\|\theta_0^i-\theta_*^i\right\|\Big), \\
\left\|p_t^i\right\| &\le (e^{-\mathcal{E}(t)}+C_L^{(1)}\varepsilon_0)\left\|p_0^i\right\|+C_L^{(1)}\Big(\left\|q_0^i-\theta_0^i\right\|+\left\|\theta_0^i-\theta_*^i\right\|\Big) \qquad (t\ge 0).
\end{aligned}$$

We can now show the main statement of this section. The convergence of dmSGMP to the stochastic gradient process in the longtime limit. Importantly, we assume a coupling between the processes through the index process $(i(t))_{t\geq 0}$ that is identical in both dynamical systems. Throughout this section, we collected evidence for a contractive behaviour in between the deterministic processes (4.20) and (3.4). We now use these results to study the randomized version piece-by-piece. This, of course, is a very similar strategy to that used in Theorem 7.

**Theorem 8.** *Let $(\tilde{p}_t, \tilde{q}_t)_{t\geq 0}$ and $(\theta_t)_{t\geq 0}$ solve (4.19) and (4.3), respectively. Let $\theta_*^i$ be the critical point of $\Phi_i$ for $i \in I$. We assume that $\Phi_i$ is strongly convex with constant $\mu$ and let $\Phi_i$ satisfy Assumption 5 and $\varepsilon(t)$ satisfy Assumption 7 with $\lambda = (\mu/4) \wedge (1/4)$ for $i \in I$. Let $\{\tau_n\}_{n\geq 1}$ be the sequences of the jump times of the process $(i(t))_{t\geq 0}$. If we assume that $\varepsilon_0 \leq \frac{(C_{L,\mu}^{(2)})^{-2}\mu^2}{4(2\gamma N+\mu)^2}$, where $C_{L,\mu}^{(2)} = C_{L,\mu}^{(1)} + C_L^{(1)}$, then*

$$\mathbb{E}\Big[\, \|\tilde{q}_{\tau_n} - \theta_{\tau_n}\| \,\Big] \to 0 \qquad (n \to \infty).$$

*Furthermore, if we additionally assume that $\varepsilon$ decays at least exponentially, i.e. there exist $C, c > 0$ such that $\varepsilon(t) \leq Ce^{-ct}$, we have*

$$\|\tilde{q}_t - \theta_t\| \to 0 \qquad (t \to \infty)$$

*almost surely and in expectation.*

**Remark 10.** *After the statement of Assumption 7, we have concluded that $m(t) \geq m_0 e^{-\lambda t}$. Thus, the constant $c$ mentioned above needs to be smaller than $\lambda$.*

*Proof.* For any $n \geq 0$, by Corollary 5, and since the initial value of $\varepsilon(t)$ in the interval $[\tau_n, \tau_{n+1})$ is exactly $\varepsilon(\tau_n)$, we have

$$\begin{aligned}\left\|\tilde{q}_{\tau_{n+1}} - \theta_{\tau_{n+1}}\right\| &\leq e^{-\mu(\tau_{n+1}-\tau_n)/2} \left\|\tilde{q}_{\tau_n} - \theta_{\tau_n}\right\| \\ &\quad + C_{L,\mu}^{(1)}\varepsilon(\tau_n)\Big( \|\tilde{q}_{\tau_n} - \theta_{\tau_n}\| + \|\tilde{p}_{\tau_n}\| + \left\|\theta_{\tau_n} - \theta_*^{i(\tau_n)}\right\| \Big)\end{aligned}$$

and

$$\left\|\tilde{p}_{\tau_{n+1}}\right\| \leq \left(e^{-\mathcal{E}(\tau_{n+1}-\tau_n)} + C_L^{(1)}\varepsilon(\tau_n)\right)\left\|\tilde{p}^{\varepsilon}_{\tau_n}\right\| + C_L^{(1)}\Big(\left\|\tilde{q}_{\tau_n} - \theta_{\tau_n}\right\| + \left\|\theta_{\tau_n} - \theta_*^{i(\tau_n)}\right\|\Big).$$

Lemma 27 implies that

$$\left\|\theta_{\tau_n} - \theta_*^{i(\tau_n)}\right\| \leq \|\theta_0\| + K_\Phi = K_{\Phi,\theta_0}.$$

We denote $a_n := e^{-\mu(\tau_{n+1}-\tau_n)/2}$ and obtain

$$\left\|\tilde{q}_{\tau_{n+1}} - \theta_{\tau_{n+1}}\right\| \leq a_n \left\|\tilde{q}_{\tau_n} - \theta_{\tau_n}\right\| + C^{(1)}_{L,\mu}\varepsilon(\tau_n)\Big(\left\|\tilde{q}_{\tau_n} - \theta_{\tau_n}\right\| + \left\|\tilde{p}_{\tau_n}\right\| + K_{\Phi,\theta_0}\Big)$$

as well as

$$e^{-\mathcal{E}(\tau_{n+1}-\tau_n)} \leq e^{-(\tau_{n+1}-\tau_n)/\varepsilon_0} \leq e^{-(\tau_{n+1}-\tau_n)(C^{(2)}_{L,\mu})^2} \leq e^{-\mu(\tau_{n+1}-\tau_n)/2} := a_n.$$

Hence, we also have

$$\left\|\tilde{p}_{\tau_{n+1}}\right\| \leq (a_n + C_L^{(1)}\varepsilon(\tau_n))\left\|\tilde{p}_{\tau_n}\right\| + C_L^{(1)}\Big(\left\|\tilde{q}_{\tau_n} - \theta_{\tau_n}\right\| + K_{\Phi,\theta_0}\Big).$$

We denote $A_n := \left\|\tilde{q}_{\tau_n} - \theta_{\tau_n}\right\|,\ B_n := \left\|\tilde{p}_{\tau_n}\right\|$ and then have the following iteration inequalities:

$$\begin{aligned} A_{n+1} &\leq a_n A_n + C^{(1)}_{L,\mu}\varepsilon(\tau_n)\Big(A_n + B_n + K_{\Phi,\theta_0}\Big), \\ B_{n+1} &\leq (a_n + C_L^{(1)}\varepsilon(\tau_n))B_n + C_L^{(1)}\Big(A_n + K_{\Phi,\theta_0}\Big). \end{aligned}$$

Moreover, denoting $D_n = A_n + \varepsilon^{\frac{1}{2}}(\tau_n)B_n$, $C^{(2)}_{L,\mu} = C^{(1)}_{L,\mu} + C^{(1)}_L$, we have

$$\begin{aligned} D_{n+1} &\leq (a_n + C^{(1)}_{L,\mu}\varepsilon(\tau_n))A_n + \varepsilon^{\frac{1}{2}}(\tau_{n+1})(a_n + C^{(1)}_L\varepsilon(\tau_n))B_n + C^{(1)}_{L,\mu}\varepsilon(\tau_n)B_n \qquad (4.24)\\ &\quad + \varepsilon^{\frac{1}{2}}(\tau_{n+1})C^{(1)}_L A_n + C^{(1)}_{L,\mu}\varepsilon(\tau_n)K_{\Phi,\theta_0} + C^{(1)}_L\varepsilon^{\frac{1}{2}}(\tau_{n+1})K_{\Phi,\theta_0}\\ &\leq \Big(a_n + C^{(2)}_{L,\mu}\varepsilon^{\frac{1}{2}}(\tau_n)\Big)D_n + C^{(2)}_{L,\mu}\varepsilon^{\frac{1}{2}}(\tau_n)K_{\Phi,\theta_0}\\ &\leq \Big(a_n + C^{(2)}_{L,\mu}\varepsilon_0^{\frac{1}{2}}\Big)D_n + C^{(2)}_{L,\mu}\varepsilon^{\frac{1}{2}}(\tau_n)K_{\Phi,\theta_0} \qquad (4.25)\end{aligned}$$

and since $\tau_{n+1} - \tau_n$ are exponentially distributed with parameter $\gamma N$ is independent of $\mathcal{F}_{\tau_n}$, we have

$$\begin{aligned} \mathbb{E}[D_{n+1}] &\leq \mathbb{E}[e^{-\mu(\tau_{n+1}-\tau_n)/2} + C^{(2)}_{L,\mu}\sqrt{\varepsilon_0}]\mathbb{E}[D_n] + C^{(2)}_{L,\mu}K_{\Phi,\theta_0}\mathbb{E}[\varepsilon^{\frac{1}{2}}(\tau_n)], \qquad (4.26)\\ \mathbb{E}[D^2_{n+1}] &\leq \mathbb{E}[(e^{-\mu(\tau_{n+1}-\tau_n)/2} + C^{(2)}_{L,\mu}\sqrt{\varepsilon_0})^2]\mathbb{E}[D^2_n] + (C^{(2)}_{L,\mu}K_{\Phi,\theta_0})^2\mathbb{E}[\varepsilon(\tau_n)] \qquad (4.27)\\ &\quad + C^{(3)}_{L,\mu}\sqrt{\varepsilon_0}K_{\Phi,\theta_0}\mathbb{E}[D_n],\end{aligned}$$

where $C^{(3)}_{L,\mu} = 2C^{(2)}_{L,\mu}(1 + C^{(2)}_{L,\mu})$. Since $\varepsilon_0 \leq \frac{(C^{(2)}_{L,\mu})^{-2}\mu^2}{4(2\gamma N+\mu)^2}$, we have

$$\mathbb{E}[e^{-\mu(\tau_{n+1}-\tau_n)/2} + C^{(2)}_{L,\mu}\sqrt{\varepsilon_0}] \leq c_1,$$

where $c_1 = \frac{\mu+4\gamma N}{2(2\gamma N+\mu)} < 1$. From the iteration inequality and Lemma 31, we know that $\lim_{n\to\infty}\mathbb{E}[D_n] = 0$, which finish the proof of the first part. If we assume $\varepsilon(t) \leq Ce^{-ct}$, we then have

$$\mathbb{E}[\varepsilon^{\frac{1}{2}}(\tau_n)] \leq C\mathbb{E}[e^{-\frac{c\tau_n}{2}}] = C\mathbb{E}\Big[\prod_{i=1}^{n} e^{-\frac{c(\tau_i-\tau_{i-1})}{2}}\Big] = C\prod_{i=1}^{n}\mathbb{E}\Big[e^{-\frac{c(\tau_i-\tau_{i-1})}{2}}\Big] = C(c_2)^n$$

where $c_2 := \mathbb{E}\Big[e^{-\frac{c(\tau_i-\tau_{i-1})}{2}}\Big] = \frac{2\gamma N}{c+2\gamma N}$ which is a constant does not depend on $i$ and that is smaller than 1. We then rewrite (4.26) as

$$\mathbb{E}[D_{n+1}] \leq c_1\mathbb{E}[D_n] + CC^{(3)}_{L,\mu}\sqrt{\varepsilon_0}K_{\Phi,\theta_0}(c_2)^n,$$

which follows from the second part of Lemma 31. We then have $\mathbb{E}[D_n] \leq \tilde{B}e^{-\tilde{b}n}$. This also implies

$$\mathbb{E}[D^2_{n+1}] \leq c_3\mathbb{E}[D^2_n] + CC^{(3)}_{L,\mu}\sqrt{\varepsilon_0}K_{\Phi,\theta_0}(c_4)^n.$$

Furthermore, by using Lemma 31, we have $\mathbb{E}[D^2_n] \leq \tilde{B}e^{-\tilde{b}n}$ for some $\tilde{B}, \tilde{b} > 0$. The Markov inequality implies that we have

$$\sum_{n=1}^{+\infty}\mathbb{P}(|D_n| \geq \delta) \leq \delta^{-1}\sum_{n=1}^{+\infty}\mathbb{E}[D_n] \leq \delta^{-1}\sum_{n=1}^{+\infty}\mathbb{E}[D_n] \leq \delta^{-1}\tilde{B}\sum_{n=1}^{+\infty}e^{-\tilde{b}n} < +\infty,$$

for any $\delta > 0$. This implies that $D_n \to 0$ almost surely as $n \to \infty$, since $N_t \to \infty$ almost surely as $t \to \infty$. We then have $D_{N_t} \to 0$ almost surely as $t \to \infty$. By Corollary 5, for any $t \geq 0$, we have

$$\begin{aligned}
\|\tilde{q}_t - \theta_t\| &\leq e^{-\mu(t-\tau_{N_t})/2}\left\|\tilde{q}_{\tau_{N_t}} - \theta_{\tau_{N_t}}\right\| \\
&\quad + C^{(1)}_{L,\mu}\varepsilon(\tau_{N_t})\Big(\left\|\tilde{q}_{\tau_{N_t}} - \theta_{\tau_{N_t}}\right\| + \left\|\tilde{p}_{\tau_{N_t}}\right\| + \left\|\theta_{\tau_{N_t}} - \theta_*^{i(\tau_{N_t})}\right\|\Big) \\
&\leq (1 + C^{(3)}_{L,\mu})D_{N_t} + C^{(1)}_{L,\mu}\varepsilon(\tau_{N_t})K_{\Phi,\theta_0}.
\end{aligned}$$

This implies that $\|q_t - \theta_t\| \to 0$ almost surely as $t \to \infty$ since $D_{N_t} \to 0$ and $\varepsilon(\tau_{N_t}) \to 0$ almost surely.

Also, in order to prove $\mathbb{E}[\|\tilde{q}_t - \theta_t\|] \to 0$, it is sufficient to show $\mathbb{E}[D_{N_t}] + \mathbb{E}[\varepsilon(\tau_{N_t})] \to 0$ as $t \to +\infty$. Since

$$\mathbb{E}[D_{N_t}^2] = \sum_{k=0} \mathbb{E}[D_k^2 \mathbb{1}_{N_t=k}]^2 \leq \sum_{k=0} \mathbb{E}[D_k^2] < +\infty,$$

which implies that $\{D_{N_t}\}_{t\geq 0}$ is uniformly integrable. Then, $\mathbb{E}[D_{N_t}] \to 0$ since $D_{N_t} \to 0$ almost surely. Note that $\lim_{t\to+\infty} \varepsilon(\tau_{N_t}) = 0$. By the Dominated Convergence Theorem, we get

$$\lim_{t\to+\infty} \mathbb{E}[\varepsilon(\tau_{N_t})] = 0.$$

This completes the proof. □

### 4.3.2 Losing, both, momentum and randomness over time

In the previous subsection, we have seen that the stochastic gradient process with decreasing momentum approaches the stochastic gradient process. Asymptotically, the stochastic gradient process converges to a stationary distribution, not necessarily to a single point. As discussed before, to attain convergence to a single point, we usually need to decrease the learning rate over time, see [71]. We now study the following model.

**Definition 4.** *The* decreasing-mass, decreasing-learning-rate stochastic gradient-momentum process (ddSGMP) *is a solution of the following stochastic differential equation,*

$$\begin{cases} \mathrm{d}\hat{q}_t &= \hat{p}_t \mathrm{d}t, \\ \varepsilon(t)\mathrm{d}\hat{p}_t &= -\nabla\Phi_{i^\beta(t)}(\hat{q}_t)\mathrm{d}t - \hat{p}_t \mathrm{d}t, \\ \hat{p}_0, \hat{q}_0 &\in X, \end{cases} \tag{4.28}$$

*where $\Phi_j$ satisfies Assumption 5, $j = 1, \ldots, N$. The mass $\varepsilon(t) > 0$ is strictly decreasing and differentiable with $\lim_{t\to\infty} \varepsilon(t) = 0$. The stochastic process $(i(t))_{t\geq 0}$ is defined in Definition*

*1. For the re-scaled index process* $(i^\beta(t))_{t\geq 0} = (i(\beta(t)))_{t\geq 0}$*, we assume that* $\beta(t) = \int_0^t \eta(s)\mathrm{d}s$*,* $t \geq 0$ *with* $\eta : [0,\infty) \to (0,\infty)$ *being a non-decreasing continuously differentiable function.*

As opposed to the previous section, we do not consider the stochastic gradient process with constant learning rate as the limiting system, but actually the associated stochastic gradient process with *decreasing* learning rate, denoted by

$$\begin{cases} \mathrm{d}\xi_t &= -\nabla\Phi_{i^\beta(t)}(\xi_t)\mathrm{d}t, \\ \xi_0 &\in X. \end{cases} \tag{4.29}$$

Next, we define the sequence $\{\tau_n\}_{n\geq 1}$ to be the jump times of process $(i(t))_{t\geq 0}$ and $\tau_n^\beta = \beta^{-1}(\tau_n)$, to be the jump times of $(i^\beta(t))_{t\geq 0}$. We denote by

$$\Omega_n^\alpha := \left\{ \frac{\mu}{2\eta(\tau_{n+1}^\beta)} \geq \frac{\alpha_1}{\sqrt{n}} \text{ and } \varepsilon(\tau_n^\beta) \leq \alpha_2 e^{-\alpha_3\sqrt{n}} \right\},$$

for $n \in \mathbb{N}$ an event that is used to impose a growth condition on $\beta$ and $m$. Regarding this event, we now denote the following assumption.

**Assumption 8.** *For* $n \geq k$*, let* $W_k^{\alpha,n} = \bigcap_{i=k}^n \Omega_i^\alpha$*. There exist* $\alpha_1, \alpha_2, \alpha_3 > 0$ *such that* $\lim_{k\to+\infty} \mathbb{P}(W_k^{\alpha,\infty}) = 1$.

The event $W_k^{\alpha,n}$ is increasing in $k$ and decreasing in $n$. Assumption 8 implies that the complement of $\Omega_n^\alpha$ is eventually small. We describe a setting for which this assumption holds in Example 2 below, after stating and proving the main results of this section: the convergence of ddSGMP to the stochastic gradient process with decreasing learning rate and the convergence of ddSGMP to the minimizer of the target function. In neither of these cases, we obtain a convergence rate. We later study the speed of convergence through numerical experiments in Section 4.5. But now we start with the first of the two aforementioned results.

**Theorem 9.** *Let* $(\hat{p}_t, \hat{q}_t)_{t\geq 0}$ *and* $(\xi_t)_{t\geq 0}$ *solve (4.28) and (4.29), respectively. For* $i \in I$*, we assume that* $\Phi_i$ *is strongly convex with constant* $\mu$ *and let* $\Phi_i$ *satisfy Assumption 5. Let*

$(\varepsilon(t))_{t\geq 0}$ *satisfy Assumption 7 with* $\lambda = (\mu/4) \wedge (1/4)$. *In addition, we assume that* $(\varepsilon(t))_{t\geq 0}$ *and* $(\beta(t))_{t\geq 0}$ *satisfy Assumption 8. Then, we have*

$$\|\hat{q}_t - \xi_t\| \to 0$$

*almost surely, as* $t \to \infty$.

*Proof.* For any $n \geq 0$, by Corollary 5, we have the following iteration inequality,

$$\begin{aligned}\left\|\hat{q}_{\tau^\beta_{n+1}} - \xi_{\tau^\beta_{n+1}}\right\| \leq\ & e^{-\mu(\tau^\beta_{n+1}-\tau^\beta_n)/2}\left\|\hat{q}_{\tau^\beta_n} - \xi_{\tau^\beta_n}\right\| \\ & + C^{(1)}_{L,\mu}\varepsilon(\tau^\beta_n)\Big(\left\|\hat{q}_{\tau^\beta_n} - \xi_{\tau^\beta_n}\right\| + \left\|\hat{p}_{\tau^\beta_n}\right\| + \left\|\xi_{\tau^\beta_n} - \theta^{i(\tau^\beta_n)}_*\right\|\Big)\end{aligned}$$

and

$$\left\|\hat{p}_{\tau^\beta_{n+1}}\right\| \leq \left(e^{-\mu(\tau^\beta_{n+1}-\tau^\beta_n)/2} + C^{(1)}_L\varepsilon(\tau^\beta_n)\right)\left\|\hat{p}_{\tau^\beta_n}\right\| + C^{(1)}_{L,\mu}\Big(\left\|\hat{q}_{\tau^\beta_n} - \xi_{\tau^\beta_n}\right\| + \left\|\xi_{\tau^\beta_n} - \theta^{i(\tau^\beta_n)}_*\right\|\Big).$$

Since $\frac{\mathrm{d}\beta^{-1}}{\mathrm{d}t}(t) = \frac{1}{\eta(\beta^{-1}(t))}$, we have

$$\tau^\beta_{n+1} - \tau^\beta_n = \beta^{-1}(\tau_{n+1}) - \beta^{-1}(\tau_n) \geq (\tau_{n+1} - \tau_n)/\eta(\beta^{-1}(\tau_{n+1})).$$

Hence, under the event $W^{\alpha,n}_k$, for $n \geq k$, we have

$$\begin{aligned}&\left\|\hat{q}_{\tau^\beta_{n+1}} - \xi_{\tau^\beta_{n+1}}\right\| \\ &\leq e^{-\mu(\tau^\beta_{n+1}-\tau^\beta_n)/2}\left\|\hat{q}_{\tau^\beta_n} - \xi_{\tau^\beta_n}\right\| + C^{(1)}_{L,\mu}\varepsilon(\tau^\beta_n)\Big(\left\|\hat{q}_{\tau^\beta_n} - \xi_{\tau^\beta_n}\right\| + \left\|\hat{p}_{\tau^\beta_n}\right\| + \left\|\xi_{\tau^\beta_n} - \theta^{i(\tau^\beta_n)}_*\right\|\Big)\Big] \\ &\leq \left(e^{-\frac{\alpha_1(\tau_{n+1}-\tau_n)}{\sqrt{n}}} + C^{(1)}_{L,\mu}\alpha_2 e^{-\alpha_3\sqrt{n}}\right)\left\|\hat{q}_{\tau^\beta_n} - \xi_{\tau^\beta_n}\right\| + C^{(1)}_{L,\mu}\alpha_2 e^{-\alpha_3\sqrt{n}}\left\|\hat{p}_{\tau^\beta_n}\right\| + K_{\Phi,\xi_0}\alpha_2 e^{-\alpha_3\sqrt{n}}\end{aligned}$$

and

$$\left\|\hat{p}_{\tau^\beta_n}\right\| \leq \left(e^{-\frac{\alpha_1(\tau_{n+1}-\tau_n)}{\sqrt{n}}} + C^{(1)}_{L,\mu}\alpha_2 e^{-\alpha_3\sqrt{n}}\right)\left\|\hat{p}_{\tau^\beta_n}\right\| + C^{(1)}_{L,\mu}\Big(\left\|\hat{q}_{\tau^\beta_n} - \xi_{\tau^\beta_n}\right\| + \left\|\xi_{\tau^\beta_n} - \theta^{i(\tau^\beta_n)}_*\right\|\Big).$$

Since $\tau_{n+1}-\tau_n$ is independent of $\left\|\hat{q}_{\tau_n^\beta}-\xi_{\tau_n^\beta}\right\|$ and $W_k^{\alpha,n-1}$, for $n\geq k$, we have

$$\mathbb{E}\Big[\left\|\hat{q}_{\tau_{n+1}^\beta}-\xi_{\tau_{n+1}^\beta}\right\|\mathbb{1}_{W_k^{\alpha,n}}\Big]\leq\Big(\frac{\gamma N}{\gamma N+\frac{\alpha_1}{\sqrt{n}}}+C_{L,\mu}^{(1)}\alpha_2e^{-\alpha_3\sqrt{n}}\Big)\mathbb{E}\Big[\left\|\hat{q}_{\tau_n^\beta}-\xi_{\tau_n^\beta}\right\|\mathbb{1}_{W_k^{\alpha,n-1}}\Big]$$
$$+C_{L,\mu}^{(1)}\alpha_2e^{-\alpha_3\sqrt{n}}\mathbb{E}\Big[\left\|\hat{p}_{\tau_n^\beta}\right\|\mathbb{1}_{W_k^{\alpha,n-1}}\Big]+K_{\Phi,\xi_0}\alpha_2e^{-\alpha_3\sqrt{n}}$$

and

$$\mathbb{E}\Big[\left\|\hat{p}_{\tau_{n+1}^\beta}\right\|\mathbb{1}_{W_k^{\alpha,n}}\Big]\leq\Big(\frac{\gamma N}{\gamma N+\frac{\alpha_1}{\sqrt{n}}}+C_{L,\mu}^{(1)}\alpha_2e^{-\alpha_3\sqrt{n}}\Big)\mathbb{E}\Big[\left\|\hat{p}_{\tau_n^\beta}\right\|\mathbb{1}_{W_k^{\alpha,n-1}}\Big]$$
$$+C_{L,\mu}^{(1)}\mathbb{E}\Big[\left\|\hat{q}_{\tau_n^\beta}-\xi_{\tau_n^\beta}\right\|\mathbb{1}_{W_k^{\alpha,n-1}}\Big]+K_{\Phi,\xi_0}.$$

We denote

$$A_n^\beta:=\mathbb{E}\Big[\left\|\hat{q}_{\tau_n^\beta}-\xi_{\tau_n^\beta}\right\|\mathbb{1}_{W_k^{\alpha,n}}\Big],\ B_n^\beta:=\mathbb{E}\Big[\left\|\hat{p}_{\tau_n^\beta}\right\|\mathbb{1}_{W_k^{\alpha,n}}\Big].$$

Let $D_n^\beta=A_n^\beta+\alpha_2e^{-\alpha_3\sqrt{n}}B_n^\beta$. We can find some constant $0<C_{\gamma,N,L,\mu}<1$ such that for $n\geq k$ and $k$ large enough

$$D_{n+1}^\beta\leq\Big(1-\frac{C_{\gamma,N,L,\mu}}{\sqrt{n}}\Big)D_n^\beta+C_{\Phi,\xi_0,L}e^{-\alpha_3\sqrt{n}}.\tag{4.30}$$

Let $G_k^n:=\prod_{i=k}^n\Big(1-\frac{C_{\gamma,N,L,\mu}}{\sqrt{i}}\Big)$ with $G_n^n=1-\frac{C_{\gamma,N,L,\mu}}{\sqrt{n}}\geq\frac{1}{2}$. From (4.30), we have

$$D_{n+1}^\beta\leq G_k^nD_k^\beta+2C_{\Phi,\xi_0,L}\sum_{i=k}^n e^{-\alpha_3\sqrt{i}}G_i^n.$$

From Lemma 32, we know that there exist some constant $c > 0$ such that $G_i^n \leq e^{-c(\sqrt{n}-\sqrt{i})}$. This implies

$$\begin{aligned} D_{n+1}^{\beta} &\leq Ce^{-c\sqrt{n}} D_k^{\beta} + 2C_{\Phi,\xi_0,L} \sum_{i=k}^{n} e^{-\alpha_3\sqrt{i}} e^{-c(\sqrt{n}-\sqrt{i})} \\ &\leq Ce^{-c\sqrt{n}} D_k^{\beta} + 2C_{\Phi,\xi_0,L} \sum_{i=k}^{n} e^{-c\wedge\alpha_3\sqrt{i}} e^{-c\wedge\alpha_3(\sqrt{n}-\sqrt{i})} \\ &\leq Ce^{-c\sqrt{n}} D_k^{\beta} + 2C_{\Phi,\xi_0,L} n e^{-c\wedge\alpha_3\sqrt{n}}. \end{aligned}$$

Hence,

$$\sum_{n=k}^{\infty} \mathbb{P}\Big( \Big\| \hat{q}_{\tau_n^{\beta}} - \xi_{\tau_n^{\beta}} \Big\| \mathbb{1}_{W_k^{\alpha,n-1}} \geq \varepsilon \Big) \leq \sum_{n=k}^{\infty} \frac{D_n^{\beta}}{\varepsilon} \leq \frac{C}{\varepsilon} \sum_{n=k}^{\infty} n e^{-c\wedge\alpha_3\sqrt{n}} < +\infty.$$

Then, the Borel–Cantelli Lemma implies that

$$\lim_{n\to\infty} \Big\| \hat{q}_{\tau_n^{\beta}} - \xi_{\tau_n^{\beta}} \Big\| \mathbb{1}_{W_k^{\alpha,n-1}} = 0,$$

almost surely. Under the Assumption 8, we have

$$\lim_{k\to\infty} \mathbb{1}_{(W_k^{\alpha,\infty})^c} = 0,$$

almost surely. Since

$$\begin{aligned} \Big\| \hat{q}_{\tau_n^{\beta}} - \xi_{\tau_n^{\beta}} \Big\| &= \Big\| \hat{q}_{\tau_n^{\beta}} - \xi_{\tau_n^{\beta}} \Big\| \mathbb{1}_{W_k^{\alpha,n-1}} + \Big\| \hat{q}_{\tau_n^{\beta}} - \xi_{\tau_n^{\beta}} \Big\| \mathbb{1}_{(W_k^{\alpha,n})^c} \\ &\leq \Big\| \hat{q}_{\tau_n^{\beta}} - \xi_{\tau_n^{\beta}} \Big\| \mathbb{1}_{W_k^{\alpha,n-1}} + \Big\| \hat{q}_{\tau_n^{\beta}} - \xi_{\tau_n^{\beta}} \Big\| \mathbb{1}_{(W_k^{\alpha,\infty})^c}, \end{aligned}$$

for $k$ large enough, we have

$$\lim_{n\to\infty} \Big\| \hat{q}_{\tau_n^{\beta}} - \xi_{\tau_n^{\beta}} \Big\| = 0,$$

almost surely. Similarly, we have

$$\lim_{n\to\infty} e^{-\alpha_3\sqrt{n}} \left\|\hat{p}_{\tau_n^\beta}\right\| = 0$$

almost surely. Under the event $W_k^{\alpha,n}$, $\varepsilon(\tau_n^\beta)$ is smaller than $\alpha_2 e^{-\alpha_3\sqrt{n}}$. Hence, we have

$$\lim_{n\to\infty} \varepsilon(\tau_n^\beta) \left\|\hat{p}_{\tau_n^\beta}\right\| = 0,$$

almost surely. Since $\lim_{n\to\infty} \tau_n^\beta = +\infty$ and

$$\begin{aligned}\|\hat{q}_t - \xi_t\| \leq e^{-\mu(t-\tau_{N_t}^\beta)/2} \left\|\hat{q}_{\tau_{N_t}^\beta} - \xi_{\tau_{N_t}^\beta}\right\| \\ + C_{L,\mu}^{(1)}\varepsilon(\tau_{N_t}^\beta)\Big( \left\|\hat{q}_{\tau_{N_t}^\beta} - \xi_{\tau_{N_t}^\beta}\right\| + \left\|\hat{p}_{\tau_{N_t}^\beta}\right\| + \left\|\xi_{\tau_{N_t}^\beta} - \theta_*^{i(\tau_{N_t}^\beta)}\right\| \Big),\end{aligned}$$

where $\tau_{N_t}^\beta \leq t \leq \tau_{N_t+1}^\beta$, we have

$$\|\hat{q}_t - \xi_t\| \leq (1 + C_{L,\mu}^{(1)}\varepsilon_0) \left\|\hat{q}_{\tau_{N_t}^\beta} - \xi_{\tau_{N_t}^\beta}\right\| + \varepsilon(\tau_{N_t}^\beta)\Big( \left\|\hat{p}_{\tau_{N_t}^\beta}\right\| + \left\|\xi_{\tau_{N_t}^\beta} - \theta_*^{i(\tau_{N_t}^\beta)}\right\| \Big).$$

This implies that $\|\hat{q}_t - \xi_t\| \to 0$, almost surely, as $t \to \infty$. □

Using this result and prior knowledge about the stochastic gradient process, we can now finally show convergence to the minimizer of $\bar{\Phi}$.

**Corollary 6.** *Under the same conditions as Theorem 9, we have*

$$\|\hat{q}_t - \theta_*\| \to 0$$

*in probability, as $t \to \infty$, where $\theta_*$ is the unique minimizer of the function $\bar{\Phi}$.*

*Proof.* From [71, Theorem 4], we know that $\xi_t$ converges to $\theta_*$ weakly, as $t \to \infty$. This is equivalent to $\|\xi_t - \theta_*\| \to 0$ in probability since $\theta_*$ is deterministic. Then the result follows by applying Theorem 9. □

Hence, we have shown that when reducing – indeed, losing – momentum and decreasing the learning rate over time, we have convergence of the stochastic gradient-momentum process to the minimizer $\theta_*$.

We finish this section, by discussing the non-trivial Assumption 8. Indeed, we give an example below for how $(\beta(t))_{t\geq 0}$ and $(m(t))_{t\geq 0}$ can be chosen to satisfy the assumption.

**Example 2.** *Let $\beta(t) = t^2$ and $\varepsilon(t) = \varepsilon_0 e^{-\lambda t}$, where $0 < m_0 < 1$ is a constant. In this case $\eta(t) = 2t$ and $\beta^{-1}(t) = \sqrt{t}$. Hence,*

$$\begin{aligned}\Omega_n^\alpha &= \Big\{\frac{\mu}{2\eta(\beta^{-1}(\tau_{n+1}))} \geq \frac{\alpha_1}{\sqrt{n}} \text{ and } \varepsilon(\tau_n^\beta) \leq \alpha_2 e^{-\alpha_3\sqrt{n}}\Big\} \\ &= \Big\{\frac{\mu}{4\sqrt{\tau_{n+1}}} \geq \frac{\alpha_1}{\sqrt{n}} \text{ and } \varepsilon_0 e^{-\lambda\sqrt{\tau_n}} \leq \alpha_2 e^{-\alpha_3\sqrt{n}}\Big\}.\end{aligned}$$

*By a concentration inequality, we have*

$$\mathbb{P}\Big(\tau_n \leq \frac{n}{2\gamma N}\Big) = \mathbb{P}\Big(e^{-2\gamma N\tau_n} \geq e^{-n}\Big) \leq e^n\mathbb{E}[e^{-2\gamma N\tau_n}] = \frac{e^n}{3^n} \tag{4.31}$$

*and*

$$\mathbb{P}\Big(\tau_n \geq \frac{2n}{\gamma N}\Big) = \mathbb{P}\Big(e^{\gamma N\tau_n/2} \geq e^n\Big) \leq e^{-n}\mathbb{E}[e^{\gamma N\tau_n/2}] = \frac{2^n}{e^n}. \tag{4.32}$$

*When $n$ is sufficiently large, $\frac{n}{2\gamma N} \leq \tau_n \leq \frac{2n}{\gamma N}$ already implies that the event $\Omega_n^\alpha$ occurs for $\alpha_1 = \frac{2}{\mu\sqrt{2\gamma N}}, \alpha_2 = m_0, \alpha_3 = \frac{\lambda}{\sqrt{2\gamma N}}$. Hence, when $n$ is sufficiently large, $(\Omega_n^\alpha)^c \subset \{\tau_n \leq \frac{n}{2\gamma N}\} \cup \{\tau_n \geq \frac{2n}{\gamma N}\}$. Therefore, from (4.31) and (4.32), we have*

$$\sum_{n=1}^{\infty}\mathbb{P}(\Omega_n^c) \leq \sum_{n=1}^{\infty}\Big(\frac{e^n}{3^n} + \frac{2^n}{e^n}\Big) < +\infty.$$

*The Borel-Cantelli Lemma then implies that $\lim_{k\to+\infty}\mathbb{P}(\bigcup_{i=k}^{\infty}(\Omega_i^\alpha)^c) = 0$ which is equivalent to Assumption 8.*

## 4.4 discretization of the continuous-time system

After the previous theoretical study of stochastic gradient-momentum processes, we now propose a numerical scheme to discretise those dynamical systems. This is crucial to turn the continuous-time dynamics into practical algorithms, see the discussion in [110, 116, 128, 129, 130]: the continuous-time analysis given in Sections 4.2 and 4.3 is only worthwhile, when the discrete algorithm behaves similarly to the continuous-time dynamics. The most usual discretization of SGMP with an explicit symplectic Euler method would yield an algorithm very similar to stochastic gradient descent with classical momentum. However, this is hardly appropriate as explicit schemes are unstable when the mass is small or when it decreases over time. In the following, we aim to obtain a stable, symplectic, and efficient discretization strategy that retains the correct longtime behaviour for the deterministic parts of SGMP. We note that even in the decreasing mass case, we are committed to using a symplectic solver to retain local conversation laws in the dynamical system, see [89].

First, we discuss the discretization of the underdamped gradient flow with constant or decreasing mass. For convenience, we briefly recall the ODE of interest:

$$\begin{cases} \mathrm{d}q_t &= p_t \mathrm{d}t, \\ m(t)\mathrm{d}p_t &= -\nabla\bar{\Phi}(q_t)\mathrm{d}t - \alpha p_t \mathrm{d}t, \\ p_0, q_0 &\in X. \end{cases} \tag{4.33}$$

Since our theoretical analysis investigates the case $m(t) \to 0$ as $t \to \infty$, we need the discretization scheme to especially be stable for small values of $m(t)$. In this framework, a fully explicit method, such as the forward or symplectic Euler method, would not be suitable for the purpose. In particular, since the evolution of $t \mapsto p_t$ is directly affected by the value of the mass $m(t)$, it is natural to consider an implicit discretization for this variable:

$$m_n \frac{p_{n+1} - p_n}{h} = -(\nabla\bar{\Phi}(q_n) + \alpha p_{n+1}),$$

for every $n \geq 0$, yielding

$$p_{n+1} = \frac{1}{m_n/h + \alpha} \left( \frac{m_n}{h} p_n - \nabla \bar{\Phi}(q_n) \right), \tag{4.34}$$

where $h > 0$ denotes the discretization step-size and $m_n := m(nh)$. For the $q$ variable in (4.33) we use the scheme

$$q_{n+1} = q_n + h p_{n+1} \tag{4.35}$$

for every $n \geq 0$. Finally, combining (4.34) and (4.35), we obtain the following update rule for the discretized dynamical model:

$$\begin{cases} p_{n+1} = \frac{1}{m_n/h+\alpha} \left( \frac{m_n}{h} p_n - \nabla \bar{\Phi}(q_n) \right), \\ q_{n+1} = q_n + h p_{n+1}, \end{cases} \tag{4.36}$$

for every $n \geq 0$. As we discretise $q$ explicitly and $p$ implicity, this discretization scheme is a semi-implicit method. Moreover, the discretization scheme employed here is symplectic [51, Theorem VI.3.3], so that the discretized dynamics satisfy (local) conservation laws. For more details, we refer the reader to the discussion in [51, Section VI.7].

**Remark 11.** *As suggested in the previous sections and depicted in Figure 4.3, the parameter $m$ can be interpreted as an interpolation variable between the gradient flow and the underdamped gradient flow. We observe that this is still the case for the discretized dynamical system* (4.36)*. Indeed, if we set $\alpha = 1$ and let $m \to 0$,* (4.36) *turns into*

$$\begin{cases} p_{n+1} = -\nabla \bar{\Phi}(q_n), \\ q_{n+1} = q_n + h p_{n+1}, \end{cases}$$

*which is exactly the gradient descent method.*

We finish this section with a comparison of the semi-implicit discretization of the underdamped gradient flow with other momentum-based optimization methods. First, we note that we can rephrase the update rule (4.36) to obtain a recursive definition of $q_n$ that does not involve the velocity variable $p_n$. Namely, from (4.36) we deduce that

$$q_{n+1} = q_n + \frac{m_n}{m_n/h + \alpha} p_n - \frac{h}{m_n/h + \alpha} \nabla \bar{\Phi}(q_n).$$

Then, recalling that $hp_n = q_n - q_{n-1}$, we obtain that

$$q_{n+1} = q_n + \frac{m_n}{m_n + \alpha h}(q_n - q_{n-1}) - \frac{h}{m_n/h + \alpha} \nabla \bar{\Phi}(q_n). \tag{4.37}$$

This allows us to easily compare our scheme with other momentum-based optimization schemes. In particular, we recall that Polyak's Heavy Ball method has constant coefficients in front of the momentum term $q_n - q_{n-1}$ and in front of $\nabla \bar{\Phi}(q_n)$. Moreover, when writing Nesterov's update rule in single-variable form, we obtain:

$$q_{n+1} = q_n + \xi_n(q_n - q_{n-1}) - h\nabla \bar{\Phi}(q_n) - \xi_n h(\nabla \bar{\Phi}(q_n) - \nabla \bar{\Phi}(q_{n-1})),$$

with $\xi_n \equiv \frac{1-\sqrt{\mu h}}{1+\sqrt{\mu h}}$ when $\bar{\Phi}$ is $\mu$-strongly convex or $\xi_n = \frac{n}{n+3}$ when $\bar{\Phi}$ is just convex. Our scheme in (4.37) has a structure that is reminiscent of the one considered in [8, 7], where a time rescaling was introduced in the continuous-time dynamics, yielding a non-constant (and diverging) multiplicative factor in front of the gradient of the objective. However, it is important to note that in [8, 7] the gradient is evaluated at $q_{n+1}$, while in (4.37) we have $\nabla \bar{\Phi}(q_n)$.

### 4.4.1 Step-size choice

The first natural question about the discrete-time optimization method (4.36) concerns the choice of the discretization step-size $h > 0$. We derive a heuristic rule, assuming for simplicity

that $m_n = m$ for every $n \geq 1$. From (4.36), we obtain

$$\begin{aligned} q_{n+1} &= q_n + \frac{h^2}{m+\alpha h}\left(\frac{m}{h}p_n - \nabla\bar{\Phi}(q_n)\right) \\ &= q_n - \frac{h^2}{m+\alpha h}\left(\frac{m}{h}\frac{h}{m+\alpha h}\left(\frac{m}{h}p_{n-1} - \nabla\bar{\Phi}(q_{n-1})\right) - \nabla\bar{\Phi}(q_n)\right) \end{aligned}$$

for every $n \geq 0$. With a backward induction argument, assuming that $p_0 = 0$, we deduce that

$$q_{n+1} = q_n - \frac{h^2}{m+\alpha h}\sum_{j=0}^{n}\left(\frac{m}{m+\alpha h}\right)^j \nabla\bar{\Phi}(q_{n-j}).$$

The previous identity suggests that the position at the step $k+1$ is obtained through $k+1$ evaluations of the gradient at the previous points of the discrete trajectory. Moreover, each evaluation is weighted by a power of the *forgetting coefficient* $\frac{m}{m+\alpha h} < 1$. Therefore, it is reasonable to ask that the sum of the weights is of order $\frac{1}{L}$, where $L > 0$ is the Lipschitz constant of $\nabla\bar{\Phi}$, see Assumption 5. In particular, we require

$$\frac{h^2}{m+\alpha h}\sum_{j=0}^{n}\left(\frac{m}{m+\alpha h}\right)^j \simeq \frac{1}{L}.$$

Taking the limit as $n \to \infty$ in the previous sum, we obtain

$$\frac{1}{L} \simeq \frac{h^2}{m+\alpha h}\sum_{j=0}^{\infty}\left(\frac{m}{m+\alpha h}\right)^j = \frac{h}{\alpha},$$

or equivalently

$$h \simeq \frac{\alpha}{L}. \tag{4.38}$$

We first note that the expression at the right-hand side of (4.38) does not depend on the parameter $m$. Thus, according to the empirical argument presented above, there is no need to adjust the magnitude of the step-size according to $m$, also not if $m$ changes over time. This fact deviates from the usual discretization of the $t \mapsto p_t$ variable in (4.33). The second observation is that, we not only see convergence of our discrete method to gradient descent

as $m \downarrow 0$, see Remark 11. We also recover the correct step-size for that gradient descent scheme we converge to.

We end this section with a remark on a rescaling of the potential $\Phi$.

**Remark 12.** *Let us assume that the potential $\Phi$ is multiplied by a positive constant $\rho > 0$, i.e. $\Phi' = \rho\bar{\Phi}$. A natural question is how we should rescale the parameters of the method* (4.36) *such that the sequence of positions $(q'_n)_{n\geq 0}$ coincides with the sequence $(q_n)_{n\geq 0}$ corresponding to the original objective. It turns out that it is sufficient to set $h' = h/\rho$ and $m'_n = m_n/\rho$. Indeed, if we we consider the sequences $(q'_n)_{n\geq 0}$ and $(p'_n)_{n\geq 0}$ obtained with*

$$\begin{cases} p'_{n+1} = \frac{1}{\frac{m'_n}{h'}+\alpha}\left(\frac{m'_n}{h'}p'_n - \nabla\Phi'(q'_n)\right), & p'_0 = 0 \\ q'_{n+1} = q'_n + h'p'_{n+1}, & q'_0 = q_0, \end{cases}$$

*then a direct computation yields*

$$q'_n = q_n, \qquad p'_n = \rho p_n$$

*for every $k \geq 1$, where the sequences $(q_n)_{n\geq 0}$ and $(p_n)_{n\geq 0}$ are obtained using* (4.36) *with the original objective $\bar{\Phi}$ and the parameters $h$ and $m_n$.*

### 4.4.2 randomized version of the method

The stochastic dynamical systems discussed throughout this chapter consist of piecewise deterministic ODEs where the pieces are determined by random waiting times and a subsampling process. After having discussed the ODE discretization in the previous subsections, we now incorporate the stochasticity into the discrete-time setting. In this regard, the random waiting time for the switching of the objective — despite being a cornerstone of the theoretical analysis — does not seem natural in view of practical implementations. The use of random waiting times in practice within this framework has been discussed by [58, 71].

Here, we replace those random waiting times by deterministic waiting times that coincide with the time steps of the algorithm. This is exactly the well-established paradigm of the classical stochastic gradient descent method and of the classical momentum method as well. The waiting times may either be constant-in-time (homogeneous) or decreasing-in-time (heterogeneous).

Throughout this chapter, we have (implicitly) considered the case in which we choose one potential $\Phi_{i(t)}$ at any time $t \geq 0$. Of course, we can also employ *batch subsampling*. There, we have

$$\begin{cases} p_{n+1} = \frac{1}{\frac{m_n}{h}+\alpha} \left( \frac{m_n}{h} p_n - \mathbf{v}_n \right), \\ q_{n+1} = q_n + h p_{n+1}, \end{cases} \tag{4.39}$$

where

$$\mathbf{v}_n := \frac{1}{\ell} \sum_{r=1}^{\ell} \nabla \Phi_{i_r}(q_n), \tag{4.40}$$

and $\{i_1, \dots, i_\ell\}$ is a subset of $I$ that is sampled uniformly without replacement, i.e. we do not choose a single but rather a set of $\ell$ potentials at once and optimize with respect to their sample mean. This setting is very useful in practice, as it reduces the subsampling error. This setting is also already fully contained in our theory. To see this, we can just define a new set of potentials $(\Phi'_K)_{K \in I'}$, where $I' := \{K \subseteq I : \#K = \ell\}$ and $\Phi'_K = \frac{1}{\ell} \sum_{i \in K} \Phi_i$ for $K \in I'$. Then, the mean of the $(\Phi'_K)_{K \in I'}$ is $\bar{\Phi}$. What is not contained in our theory, but still helpful in practice, is *batch subsampling in epochs*, where at any time, we do not replace elements from which we sample in $I$ before we have picked each index once. This setting is not trivially contained in our theory due to its non-Markovian nature.

## 4.5 Numerical Experiments

We now present numerical experiments in which we test the discretization strategy proposed in Section 4.4. We start with academic convex and non-convex examples. Then, we employ SGMP for the training of a convolutional neural network regarding the classification of the

CIFAR-10 data set. We aim to show how the method compares with the classical momentum method and standard stochastic gradient descent.

### 4.5.1 One-dimensional non-smooth stationary point

The first numerical experiment involves the one-dimensional example discussed in Subsection 4.1.2. We recall that we want to minimise the function

$$\Phi(x) := (\mathrm{ReLU}(x) - 1)^2 + x^2 = \begin{cases} x^2 + 1 & \text{if } x \leq 0 \\ 2x^2 - 2x + 1 & \text{if } x > 0 \end{cases} \qquad (x \in \mathbb{R}),$$

which attains the global minimum at the point $x^* = \frac{1}{2}$. The point $\tilde{x} = 0$ is a non-smooth stationary point, since $\Phi(x) \geq \Phi(\tilde{x})$ for every $x \leq \tilde{x}$. In Subsection 4.1.2, we are able to show that the underdamped gradient flow can overcome the local minimizer $\tilde{x}$ if $\alpha^2 - 8m < 0$.

We now test this property using our discrete-time method (4.36) with the step-size constant along the iterations. The discretization scheme asks us to choose the learning rate $h$ according to the Lipschitz constant of $\Phi'$. Actually, the derivative of $\Phi$ is not continuous. However, in this framework, by "Lipschitz constant" of $\Phi'$ we mean $L_0 = \max\{\mathrm{Lip}(\Phi'|_{x\leq 0}), \mathrm{Lip}(\Phi'|_{x\geq 0})\}$. We test the scheme in this example using the precise Lipschitz constant and overestimated Lipschitz constants. The results are presented in Figure 4.4. We see that the condition $\alpha^2 - 8m < 0$ is also relevant for the discrete dynamical system. The inequality appears to be sharp whenever we overestimate the Lipschitz constant – unsurprisingly as a smaller step-size leads to a more accurate discretization of the underdamped gradient flow. Using the correct Lipschitz constant lets $\alpha^2 - 8m < 0$ appear quite conservative. Indeed, a much larger range of $\alpha, m$ allow for convergence to the global minimizer. A larger step-size makes the method more robust.

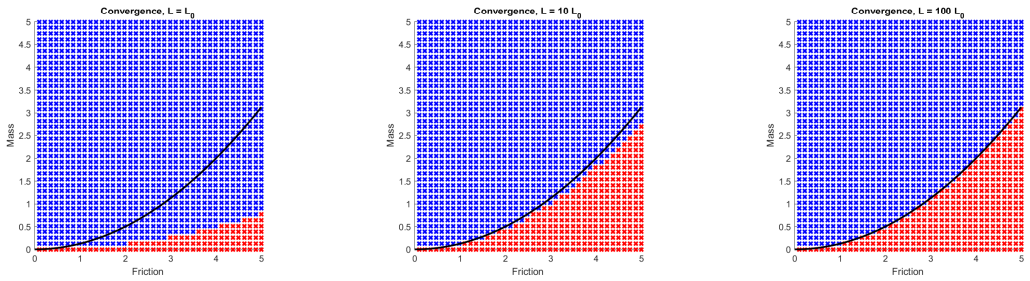


Figure 4.4: The plots depict for which combinations of $m, \alpha$ the discrete-time version of (4.4) derived in (4.36) manages to overcome the "false minimizer". The blue crosses represent convergence to the global minimizer of (4.5), the red ones to the origin. The black curve divides the $m, \alpha$ that do and do not satisfy (4.7). As the step-size $h = \frac{1}{L}$ gets smaller, the theoretical prediction (4.7) becomes more accurate. Finally, we observe that the gradient method (that corresponds to $m = 0$) never converges to the global minimizer.

### 4.5.2 Strongly convex example: quadratic objective

We now consider a target function of the form $\bar{\Phi} = \frac{1}{N}\sum_{i=1}^{N} \Phi_i$, where

$$\Phi_i(x) = \frac{N}{2} x^T A_i x + N b_i^T x,$$

and where $A_i \in \mathbb{R}^{K\times K}$ is a symmetric and positive definite matrix and $b_i \in \mathbb{R}^K$, for $i = 1, \dots, N$. Importantly, all of the $(\Phi_i)_{i\in I}$ are strongly convex. We consider $K = 500$ (dimension of the domain), and $N = 100$. For every $i = 1, \dots, N$, we have sampled the eigenvalues of the matrix $A_i$ using a uniform distribution in $[0.05, 15]$, and we have obtained $b_i$ using a normal distribution centered at the origin and with standard deviation $\sigma = 2$. We look at a total of 100 different randomly generated problems and later average over the results. We compare SGD and our method (4.36). At each iteration we use a mini-batch of $\ell = 10$ elements of $\{\Phi_1, \dots, \Phi_N\}$ to compute the stochastic approximation of the gradient of $\bar{\Phi}$. Finally, the discretization step-size has been chosen polynomially decreasing, namely at the $n$-th iteration we set $h_n = h_0/n$. This is to provide a numerical simulation for the theoretical framework analyzed in Section 4.3.2, where the re-scaling of the process $(i(t))_{t\geq 0}$ has the effect of progressively decreasing the learning rate. The results are reported in Figure 4.5. As we can see, SGMP shows a faster convergence than the classical SGD scheme.

In this case, decreasing the mass parameter $m$ leads to a deterioration of the performances. Indeed, consistently with the theoretical predictions, the behaviour of SGMP gets closer to the stochastic gradient process as $m$ diminishes. We also study the case where the mass is non-constant and decreases over time as $m_k = m_0(0.995)^k$. In that situation, the decrement of the mass leads to a slower convergence as well.

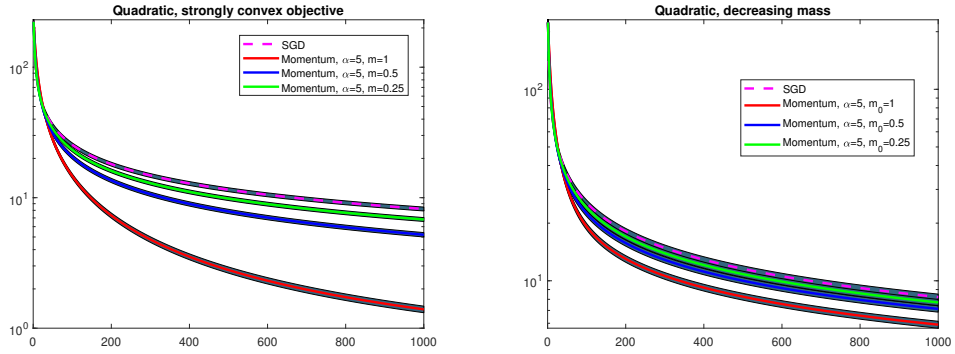


Figure 4.5: Convergence rate comparison. The plot represents the decreasing distance from the true minimizer achieved by the SGD and SGMP introduced in (4.39). We consider the case where the mass is constant, but the step-size $(h_n)_{n=0}^{\infty}$ is decreasing (left) and the case where the mass is reduced as $m_k = m_0(0.995)^k$ and the step-size decreases as before (right, see Subsection 4.3.1). The experiments are repeated 100 times (always resampling the potentials), and we report the mean distance achieved by each method, and the corresponding standard deviation.

### 4.5.3 Non-convex example: polynomial function

We studied the behaviour of our SGMP in the case of a polynomial non-convex function given as the sum $\bar{\Phi} = \frac{1}{N}\sum_{i=1}^{N}\Phi_i$, where for every $i = 1, \dots, N$

$$\Phi_i(x) = -\frac{N}{2}x^T A_i x + N b_i x + \frac{1}{4}\sum_{j=1}^{n} x_j^4$$

and where $A_i \in \mathbb{R}^{K\times K}$ is a symmetric and positive definite matrix and $b_i \in \mathbb{R}^K$. The problem is non-convex, since the Hessian of $\bar{\Phi}$ (as well as the one of each $\Phi_1, \dots, \Phi_N$) is negative definite at the origin $x = 0$. On the other hand, outside a large enough compact set, the objective function $\Phi$ is locally convex. We considered $K = 500$ (dimension of the domain),

and $N = 100$. For every $i = 1, \dots, N$, we have sampled the eigenvalues of the matrix $A_i$ using a uniform distribution in $[0.05, 15]$, and we obtained $b_i$ using a normal distribution centered at the origin and with standard deviation 2. At each iteration we use a mini-batch of $\ell = 5$ elements of $\{\Phi_1, \dots, \Phi_N\}$ to compute the stochastic approximation of the gradient of $\Phi$, and we keep the learning rate constant along the iterations. We have compared the SGMP with SGD. In this case, the learning rate is kept constant during the iterations. The results are reported in Figure 4.6. In this case it seems that the stochastic momentum method tends to stabilise in correspondence of lower values of the objective function. Interestingly, we observe that the implementations with smaller $m$ have a faster decay in the initial iterations – as opposed to the results obtained in the previous subsection.

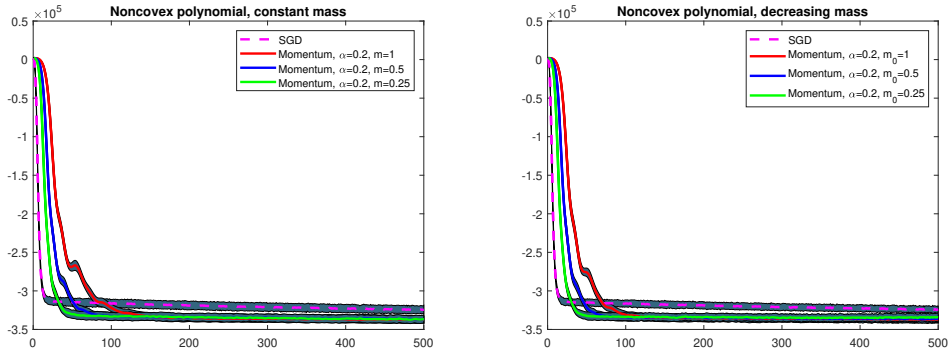


Figure 4.6: Decay of objective comparison. The plot represents the decreasing objective function achieved by the SGD and SGMP introduced in (4.39). We consider the constant-mass regime (left) and the exponentially decreasing case $m_k = m_0(0.995)^k$ (right). Experiments are run 100 times, and we reported the mean objective decrease achieved by the methods, and the respective standard deviations.

### 4.5.4 Convolutional Neural Network (CNN)

We now employ the discrete SGMP method (4.36) to solve the CIFAR-10 [66], image classification task with a convolutional neural network (CNN). The CIFAR-10 data set consists of $6 \cdot 10^4$ colour images ($32 \times 32$ pixels) which are split into $5 \cdot 10^4$ training images and $10^4$ test images. CIFAR-10 has 10 classes (e.g. airplane, dog, frog,...) with 6000 images per class. In the classification task, images with known class are used to train the CNN to automatically

recognise the class of any image. We use a VGG-like CNN architecture [118]. More precisely, we use $3 \times 3$ kernels with depth 32, 64, and 128. The network contains 6 convolutional layers, each of them followed by the ReLU activation, batch normalisation, max-pooling, and drop-out layers. The train data is augmented as in [74], that is a random horizontal flip and a $32 \times 32$ random crop after a $4 \times 4$ padding. The training is done with Google Colab using GPUs (often Tesla V100, sometime Tesla A100). We compare SGD, classical momentum, and SGMP (4.36) with different parameters, constant mass, and decreasing mass. The experiments are set in the following way. We train for 800 epochs with batch size $\ell = 100$ and no weight decay. We use constant learning rate $\eta = 0.01$ for SGD and the classical momentum. In classical momentum, we set the momentum hyperparameter $\rho = 0.9$. See Figure 4.7 for the plots of train loss for constant mass. In Figure 4.7, note that with fixed $\alpha$ and $h$, hSGMP performs better with small mass $m$. We have discussed the stability of (4.36) with small $m$ and (relatively) large stepsize in 4.4.1. In addition, when $m \downarrow 0$, the red line (hSGMP) converges to the blue line (SGD), which aligns with Theorem 7. Furthermore, we observe that with fixed step-size $h$, a small ratio of $\alpha$ and $m$ gives better results suggesting a possible rule for choosing hyperparameter in practice. For decreasing mass, see Figure 4.8 for the plots of train loss for dmSGMP. The decreasing rate is set to be $m = m_0 0.995^k$, where $k$ is the number of iterations and $m_0 = 0.1$ is the initial mass. The train loss for each epoch is calculated by averaging over batches. In Figure 4.8, dmSGMP achieves a similar train loss as the classical momentum method. They might arrive at different minimizers, but this shows that dmSGMP is a practical algorithm for machine learning optimization that we are able to lay a solid analytical foundation for. For this particular task, dmSGMP has attained a desirable result without decreasing learning rate – from Corollary 6 we would expect a behaviour similar to SGD. For larger datasets and more complicated tasks that we are not able to experiment due to lack of resource, we conjecture that ddSGMP may be an effective optimization method with proper choice of hyperparameters. See Table 4.1 for the train and test accuracy. Accuracy is measured using the model obtained after 800 epochs

of training. Overall, we conclude that hSGMP achieves competitive test accuracy compared with classical momentum. When decreasing the mass, dmSGMP achieves competitive train loss compared to classical momentum.

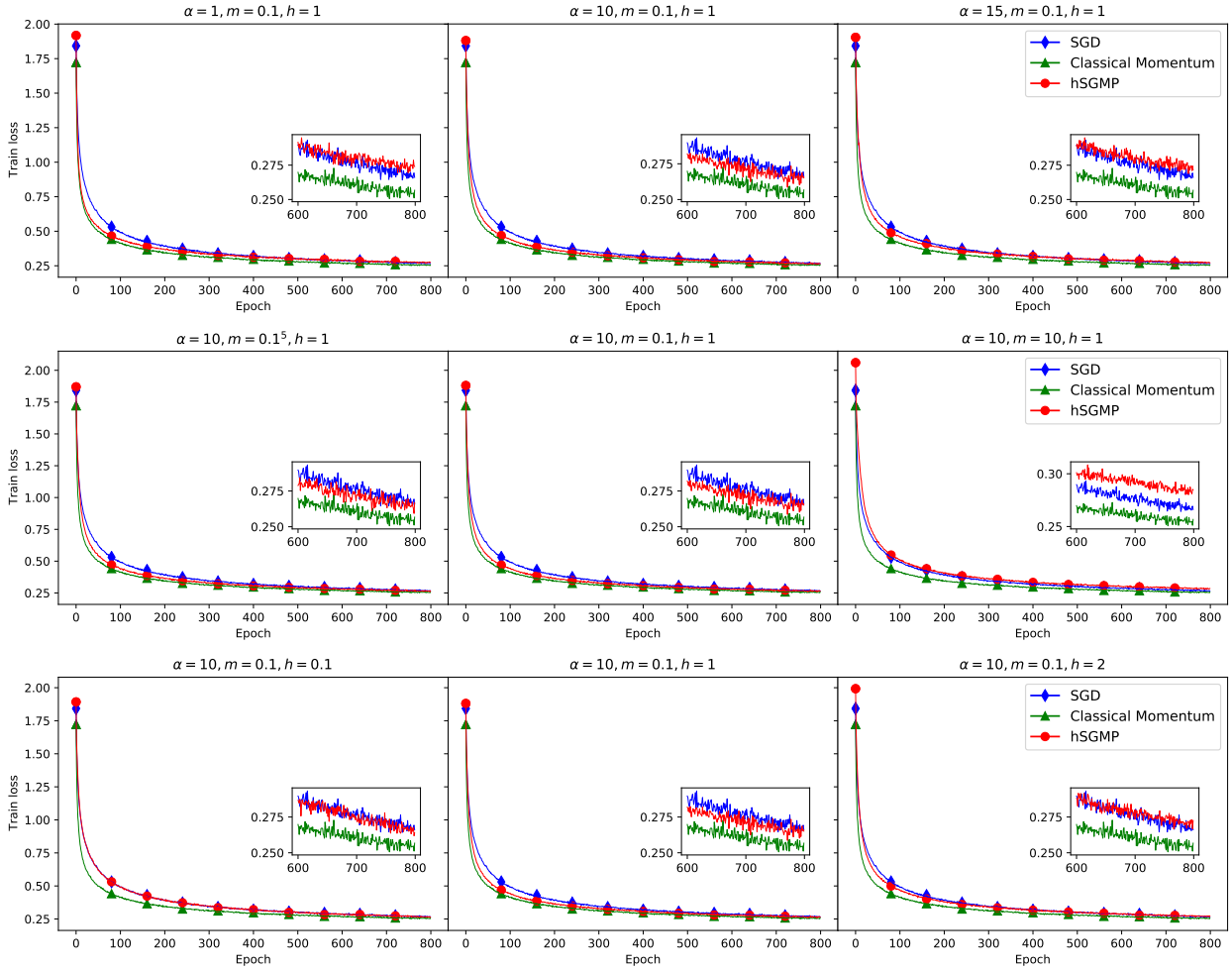


Figure 4.7: Train loss comparison for SGD ($\eta = 0.01$), classical momentum ($\eta = 0.01, \rho = 0.9$), and hSGMP for CNN on CIFAR-10. We vary $\alpha$, $\varepsilon$ and $h$ in each experiment, which are specified in the title of each plot.

## 4.6 Conclusions

In this chapter, we have proposed and analyzed the stochastic gradient-momentum process, a continuous-time dynamics representing momentum-based stochastic optimization. We have especially analyzed limiting behavior when reducing learning rate and/or particle mass.

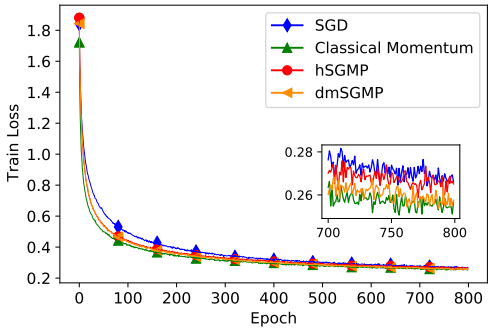


Figure 4.8: Train loss comparison for SGD ($\eta = 0.01$), classical momentum ($\eta = 0.01, \rho = 0.9$), hSGMP ($\alpha = 10, \varepsilon = 0.1, h = 1$), and dmSGMP ($\alpha = 10, \varepsilon_0 = 0.1, m = m_0 0.995^k, h = 1$) for CNN on CIFAR-10.

| **Method** | **Parameters** | **Train Acc.** | **Test Acc.** |
|---|---|---|---|
| SGD | $\eta = 0.01$ | 96.902 | 90.98 |
| Classical momentum | $\eta = 0.01,\ \rho = 0.9$ | 97.236 | 91.33 |
| hSGMP | $\alpha = 1,\ \varepsilon = 0.1,\ h = 1$ | 97.096 | 91.09 |
| | $\alpha = 10,\ \varepsilon = 0.1,\ h = 1$ | 96.986 | 91.25 |
| | $\alpha = 15,\ \varepsilon = 0.1,\ h = 1$ | 96.822 | 91.09 |
| | $\alpha = 10,\ \varepsilon = 10,\ h = 1$ | 96.556 | 90.75 |
| | $\alpha = 10,\ \varepsilon = 0.1^5,\ h = 1$ | 97.35 | 91.35 |
| | $\alpha = 10,\ \varepsilon = 0.1,\ h = 2$ | 96.984 | 90.87 |
| | $\alpha = 10,\ \varepsilon = 0.1,\ h = 0.1$ | 97.108 | 91.26 |
| dmSGMP | $\alpha = 10,\ \varepsilon_0 = 0.1,\ h = 1$ | 97.288 | 91.32 |

Table 4.1: Comparison of train and test accuracy (%) with different hyperparameters over the CIFAR-10 dataset.

In this context, learning rate and particle mass can either be reduced homogeneously or decrease over time. We have shown pathwise or longtime convergence to the underlying gradient flow or the stochastic gradient process, respectively. We have then proposed a stable discretization strategy for the stochastic gradient-momentum process and tested the strategy in several numerical examples. In those, we especially saw that the stable discretization of the stochastic gradient-momentum process can achieve a similar accuracy in a CNN training compared with (the possibly unstable) stochastic gradient descent with classical momentum algorithm.

Most of the theoretical results we have obtained throughout this chapter refer to the setting of convex optimization. Convex optimization is vital in, e.g. image reconstruction. The

training of neural network usually requires non-convex optimization and momentum-based methods are especially popular in non-convex settings. Hence, a natural future research direction are non-convex optimization problems. The stochastic gradient-momentum process does not represent the adaptivity in the Adam algorithm. To represent the adaptivity, we would need to study a two-sided dependence between $(i(t))_{t\geq 0}$ and $(p(t), q(t))_{t\geq 0}$ and a non-linear weighting in front of the gradient. Both these additions to the stochastic gradient-momentum process are a very interesting and challenging direction for future research.

# Appendix A

# Chapter 4 Appendix

## A.1 Longtime behavior of the underdamped gradient flow

We recall that the underdamped gradient flow (4.15) is defined as the following dynamical system

$$\begin{cases} \mathrm{d}q_t &= p_t \mathrm{d}t, \\ \mathrm{d}p_t &= -\nabla\bar{\Phi}(q_t)\mathrm{d}t - \alpha p_t \mathrm{d}t, \\ p_0, q_0 &\in X. \end{cases}$$

We now show that the position in the underdamped gradient flow converges to the global minimiser of the objective function $\bar{\Phi}$ and the velocity converges to 0. We shall assume that $\bar{\Phi}$ satisfies Assumption 3 with $\mathcal{A}^{\lambda,\alpha}$ for some $\lambda$ and $\alpha$ defined in (4.15). First, we note that since $\bar{\Phi}$ achieves a global minimum at $\theta_*$, $\theta_*$ is a critical point of $\bar{\Phi}$. Then, (4.13) implies that $\theta_*$ is the unique critical point of $\bar{\Phi}$. We use a technique similar to that of, e.g. [10, 11, 42, 111]. Indeed, we introduce the Lyapunov function $V : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$V(x,y) := \bar{\Phi}(x) - \bar{\Phi}(\theta_*) + \frac{\alpha^2}{4}\Big( \left\|x - \theta_* + \alpha^{-1}y\right\|^2 + \left\|\alpha^{-1}y\right\|^2 - \lambda \left\|x - \theta_*\right\|^2 \Big) \tag{A.1}$$

which we employ below. Using $V$, we can show Lyapunov exponential global stability for the underdamped gradient flow (4.15).

**Proposition 5.** *Assume that $\bar{\Phi}$ satisfies Assumption 2 and 3 with $\mathcal{A}^{\lambda,\alpha}$. Let $(p_t, q_t)_{t\geq 0}$ be the solution to (4.15) and $V$ be the Lyapunov function defined in* (A.1)*. Then we have the following inequality*

$$V(q_t, p_t) \leq V(q_0, p_0)e^{-\alpha\lambda t} \qquad (t \geq 0).$$

*Proof.* Notice that (4.13) implies that $\frac{\mathrm{d}V(q_t,p_t)}{\mathrm{d}t} \leq -\alpha\lambda V(q_t, p_t)$. By Grönwall's inequality, we obtain $V(q_t, p_t) \leq V(q_0, p_0)e^{-\alpha\lambda t}$. □

In the previous proposition, we show that the Lyapunov function $V(q_t, p_t)$ is bounded above by a number that exponentially decreases in $t > 0$. $V(q_t, p_t)$ is zero, if $q_t = \theta_*$ and $p_t = 0$, i.e. the particle is positioned at the unique minimiser of $\bar{\Phi}$ and the particle velocity is 0: the dynamical system has reached a stationary state. However, $V$ can have multiple zeros. Thus, we need some more work to show exponential convergence of the dynamical system.

**Corollary 7.** *Under the conditions of Proposition 5, we have*

$$\|p_t\|^2 + \|q_t - \theta_*\|^2 \leq C_{\alpha,\lambda}e^{-\alpha\lambda t}V(q_0, p_0) \qquad (t \geq 0).$$

*Proof.* Since $\lambda \leq \frac{1}{4}$, we have, by the $\varepsilon$-Young inequality:

$$\begin{aligned} V(x, y) &\geq \frac{\alpha^2}{4}\Big((1-\lambda)\|x-\theta_*\|^2 + 2\left\|\alpha^{-1}y\right\|^2 + 2(x-\theta_*)\cdot(\alpha^{-1}y)\Big) \\ &\geq \frac{\alpha^2}{4}\Big((1-\lambda)\|x-\theta_*\|^2 + 2\left\|\alpha^{-1}y\right\|^2 - 2\left\|\alpha^{-1}y\right\|^2 - \frac{1}{2}\|x-\theta_*\|^2\Big) \\ &\geq \frac{\alpha^2}{4}\left(\frac{1}{2}-\lambda\right)\|x-\theta_*\|^2, \end{aligned}$$

which implies that

$$\|q_t - \theta_*\|^2 \leq \frac{8}{\alpha^2(1-2\lambda)} V(q_t, p_t) \leq \frac{8}{\alpha^2(1-2\lambda)} V(q_0, p_0) e^{-\alpha\lambda t}.$$

Similarly, again by the $\varepsilon$-Young inequality, we have for $\|p_t\|$ that

$$\begin{aligned} V(x,y) &\geq \frac{\alpha^2}{4}\Big((1-\lambda)\|x-\theta_*\|^2 + 2\left\|\alpha^{-1}y\right\|^2 + 2(x-\theta_*)\cdot(\alpha^{-1}y)\Big) \\ &\geq \frac{\alpha^2}{4}\Big((1-\lambda)\|x-\theta_*\|^2 + 2\left\|\alpha^{-1}y\right\|^2 - \frac{1}{1-\lambda}\left\|\alpha^{-1}y\right\|^2 - (1-\lambda)\|x-\theta_*\|^2\Big) \\ &\geq \frac{\alpha^2}{4}\left(\frac{1-2\lambda}{1-\lambda}\left\|\alpha^{-1}y\right\|^2\right), \end{aligned}$$

which implies

$$\|p_t\|^2 \leq \frac{4(1-\lambda)}{1-2\lambda} e^{-\alpha\lambda t} V(q_0, p_0) \leq e^{-\alpha\lambda t} V(q_0, p_0).$$

Therefore, we conclude that

$$\|q_t - \theta_*\|^2 + \|p_t\|^2 \leq C_{\alpha,\lambda} e^{-\alpha\lambda t} V(q_0, p_0),$$

where $C_{\alpha,\lambda} = 4 + \frac{8}{\alpha^2(1-2\lambda)}$. □

## A.2 Decreasing momentum: the fixed-sample case

We now consider the SGMP dynamics subject to a fixed sample $i \in I$. More specifically, we define

$$\begin{cases} \mathrm{d}q_t^{i,\varepsilon} &= p_t^{i,\varepsilon}\mathrm{d}t, \\ \varepsilon\mathrm{d}p_t^{i,\varepsilon} &= -\nabla\Phi_i(q_t^{i,\varepsilon})\mathrm{d}t - p_t^{i,\varepsilon}\mathrm{d}t, \\ p_0^{i,\varepsilon} &= p_0^i \in X, \\ q_0^{i,\varepsilon} &= q_0^i \in X, \end{cases} \tag{A.2}$$

with mass $\varepsilon > 0$. When $\varepsilon \to 0$, we have the following formal limiting equation

$$\begin{cases} \mathrm{d}\theta_t^i &= -\nabla\Phi_i(\theta_t^i)\mathrm{d}t, \\ \theta_0^i &\in X. \end{cases} \tag{A.3}$$

This is the gradient flow with respect to the potential $\Phi_i$ for some $i \in I$. To prove this described limiting behaviour, we require three auxiliary results. In the first one we study, similarly to Corollary 7, the longtime behaviour of the deterministic underdamped dynamical system (A.2). This time with emphasis on the influence of the mass $\varepsilon$. Interestingly, we can see that the convergence rate is independent of the mass $\varepsilon$, if it is sufficiently small. We note that these results are similar to previous results by [9].

**Lemma 22.** *Let $(p_t^{i,\varepsilon}, q_t^{i,\varepsilon})_{t\geq 0}$ be the solution to (A.2). Let $\Phi_i$ satisfy Assumption 2 and 3 with $\mathcal{A}^{\lambda_i,1}$ and critical point $\theta_*^i$. We set $\lambda := \min_{i\in I} \lambda_i$. Then for $0 < \varepsilon \leq 1$, we have the following inequality*

$$\left\|q_t^{i,\varepsilon} - \theta_*^i\right\|^2 \leq 16(L+2)e^{-\lambda t}(\left\|q_0^i - \theta_*^i\right\|^2 + m^2\left\|p_0^i\right\|^2) \qquad (t \geq 0), \tag{A.4}$$

*where $L$ is the Lipschitz constant of the $\nabla\Phi_i$.*

*Proof.* We recall Assumption 3 for $\Phi_i$ with $\alpha = 1$:

$$(x - \theta_*^i) \cdot \nabla\Phi_i(x)/2 \geq \lambda(\Phi_i(x) - \Phi_i(\theta_*^i) + \left\|x - \theta_*^i\right\|^2/4),$$

which implies that

$$(x - \theta_*^i) \cdot \nabla\Phi_i(x)/(2\varepsilon) \geq \lambda(\varepsilon^{-1}\Phi_i(x) - \varepsilon^{-1}\Phi_i(\theta_*^i) + \varepsilon^{-1}\left\|x - \theta_*^i\right\|^2/4).$$

And since $0 < \varepsilon \leq 1$, we have

$$\begin{aligned}\lambda(\varepsilon^{-1}\Phi_i(x) - \varepsilon^{-1}\Phi_i(\theta_*^i) + \varepsilon^{-1}\left\|x-\theta_*^i\right\|^2/4) \\ &= \lambda\varepsilon(\varepsilon^{-2}\Phi_i(x) - \varepsilon^{-2}\Phi_i(\theta_*^i) + \varepsilon^{-2}\left\|x-\theta_*^i\right\|^2/4) \\ &\geq \lambda\varepsilon(\varepsilon^{-1}\Phi_i(x) - \varepsilon^{-1}\Phi_i(\theta_*^i) + \varepsilon^{-2}\left\|x-\theta_*^i\right\|^2/4).\end{aligned}$$

So Assumption 3 implies that

$$(x-\theta_*^i)\cdot\nabla\Phi_i(x)/(2\varepsilon) \geq \lambda\varepsilon(\varepsilon^{-1}\Phi_i(x) - \varepsilon^{-1}\Phi_i(\theta_*^i) + \varepsilon^{-2}\left\|x-\theta_*^i\right\|^2/4)$$

which exactly means that $\Phi_i(x)/\varepsilon$ satisfies Assumption 3 with $\mathcal{A}^{\lambda m, m^{-1}}$. We define

$$V^{i,\varepsilon}(x,y) = \frac{\Phi_i(x)-\Phi_i(\theta_*^i)}{\varepsilon} + \frac{1}{4\varepsilon^2}\Big(\left\|x-\theta_*^i+\varepsilon y\right\|^2 + \|\varepsilon y\|^2 - \varepsilon\lambda\left\|x-\theta_*^i\right\|^2\Big).$$

From Proposition 5, we have

$$V^{i,\varepsilon}(q_t^{i,\varepsilon}, q_t^{i,\varepsilon}) \leq V^{i,\varepsilon}(q_0^i, p_0^i)e^{-\lambda t}.$$

Next, we are going to show that

$$c\varepsilon^{-2}\left\|x-\theta_*^i\right\|^2 \leq V^{i,\varepsilon}(x,y) \leq (L+2)\Big(\varepsilon^{-2}\left\|x-\theta_*^i\right\|^2 + \|y\|^2\Big),$$

where $c = \frac{1}{16}$.

Since $\Phi_i(x) - \Phi_i(\theta_*^i) \geq 0$ and $\|x-\theta_*^i+\varepsilon y\|^2 + \|\varepsilon y\|^2 \geq \frac{\left\|x-\theta_*^i\right\|^2}{2}$, we have

$$\begin{aligned}V^{i,\varepsilon}(x,y) &\geq \frac{1}{4\varepsilon^2}\Big(\left\|x-\theta_*^i+\varepsilon y\right\|^2 + \|\varepsilon y\|^2 - \varepsilon\lambda\left\|x-\theta_*^i\right\|^2\Big) \\ &\geq \frac{1}{4\varepsilon^2}\Big(\frac{1}{2}-\varepsilon\lambda\Big)\left\|x-\theta_*^i\right\|^2 \geq \frac{\left\|x-\theta_*^i\right\|^2}{16\varepsilon^2}.\end{aligned}$$

Notice that

$$\Phi_i(x) - \Phi_i(\theta_*^i) = \int_0^1 (x-\theta_*^i)\Big[\nabla\Phi_i\Big((x-\theta_*^i)s+\theta_*^i\Big) - \nabla\Phi_i(\theta_*)\Big]ds \le \frac{L\left\|x-\theta_*^i\right\|^2}{2}.$$

This implies that

$$\begin{aligned} V^{i,\varepsilon}(x,y) &\le \frac{L\left\|x-\theta_*^i\right\|^2}{2\varepsilon} + \frac{1}{2\varepsilon^2}\Big(\left\|x-\theta_*^i\right\|^2 + \|\varepsilon y\|^2\Big) \\ &\le (L+2)\Big(\varepsilon^{-2}\left\|x-\theta_*^i\right\|^2 + \|y\|^2\Big). \end{aligned}$$

Hence, we immediately get

$$\varepsilon^{-2}\left\|q_t^{i,\varepsilon}-\theta_*^i\right\|^2 \le 16(L+2)e^{-\lambda t}(\varepsilon^{-2}\left\|q_0^i-\theta_*^i\right\|^2 + \left\|p_0^i\right\|^2),$$

which implies that

$$\left\|q_t^{i,\varepsilon}-\theta_*^i\right\|^2 \le 16(L+2)e^{-\lambda t}(\left\|q_0^i-\theta_*^i\right\|^2 + \varepsilon^2\left\|p_0^i\right\|^2).$$

□

In the next auxiliary result, we show boundedness of the velocity, i.e. $(p_t^{i,\varepsilon})_{t\ge0}$ that depends on the mass $\varepsilon$. Moreover, we show Lipschitz continuity of the particle position with respect to time. Note that the Lipschitz constant can be chosen independently of $\varepsilon$.

**Lemma 23.** *Let $(p_t^{i,\varepsilon}, q_t^{i,\varepsilon})_{t\ge0}$ be the solution to (A.2). Let $\Phi_i$ satisfy Assumption 2 and 3 with $\mathcal{A}^{\lambda_i,1}$ and critical point $\theta_*^i$. Then for $0<\varepsilon\le1$, we have for any $0\le s\le t$,*

$$\begin{aligned} \left\|p_t^{i,\varepsilon}\right\| &\le \Big(e^{\frac{-t}{\varepsilon}} + C_L\varepsilon\Big)\left\|p_0^i\right\| + C_L\left\|q_0^i-\theta_*^i\right\|, \\ \left\|q_t^{i,\varepsilon}-q_s^{i,\varepsilon}\right\| &\le C_L(t-s)\Big(\left\|q_0^i-\theta_*^i\right\| + \left\|p_0^i\right\|\Big) \qquad (t\ge0), \end{aligned}$$

*where $C_L = 16(L+2)L+1$ and $L$ is the Lipschitz constant of $\nabla\Phi_i$.*

*Proof.* We take the derivative of $e^{\frac{t}{\varepsilon}}p_t^{i,\varepsilon}$, from (A.2), we get

$$p_t^{i,\varepsilon} = e^{\frac{-t}{\varepsilon}}p_0^i - \varepsilon^{-1}e^{\frac{-t}{\varepsilon}}\int_0^t e^{\frac{s}{\varepsilon}}\nabla\Phi_i(q_s^{i,\varepsilon})\mathrm{d}s. \tag{A.5}$$

Hence, we have

$$\begin{aligned}
\left\|p_t^{i,\varepsilon}\right\| &\leq \left\|e^{\frac{-t}{\varepsilon}}p_0^i\right\| + \varepsilon^{-1}e^{\frac{-t}{\varepsilon}}\int_0^t e^{\frac{s}{\varepsilon}}\left\|\nabla\Phi_i(q_s^{i,\varepsilon}) - \nabla\Phi_i(\theta_*^i)\right\|\mathrm{d}s \\
&\leq e^{\frac{-t}{\varepsilon}}\left\|p_0^i\right\| + \varepsilon^{-1}e^{\frac{-t}{\varepsilon}}L\int_0^t e^{\frac{s}{\varepsilon}}\left\|q_s^{i,\varepsilon} - \theta_*^i\right\|ds \\
&\overset{(A.4)}{\leq} e^{\frac{-t}{\varepsilon}}\left\|p_0^i\right\| + 16(L+2)L\varepsilon^{-1}e^{\frac{-t}{\varepsilon}}(\left\|q_0^i - \theta_*^i\right\| + \varepsilon\left\|p_0^i\right\|)\int_0^t e^{\frac{s}{\varepsilon}}\mathrm{d}s \\
&\leq \left(e^{\frac{-t}{\varepsilon}} + 16(L+2)L\varepsilon\right)\left\|p_0^i\right\| + 16(L+2)L\left\|q_0^i - \theta_*^i\right\|.
\end{aligned}$$

Let $C_L := 16(L+2)L + 1$. Then we have $\left\|p_t^{i,\varepsilon}\right\| \leq C_L(\|p_0^i\| + \|q_0^i - \theta_*^i\|)$. From (A.2), we immediately get

$$\left\|q_t^{i,\varepsilon} - q_s^{i,\varepsilon}\right\| = \left\|\int_s^t p_m^{i,\varepsilon}\mathrm{d}m\right\| \leq \int_s^t \left\|p_m^{i,\varepsilon}\right\|\mathrm{d}m \leq C_L(t-s)\Big(\left\|q_0^i - \theta_*^i\right\| + \left\|p_0^i\right\|\Big).$$

□

The third auxiliary result is a bound on the time derivative of the velocity $(p_t^{i,m})_{t\geq 0}$, i.e. a bound on the particle's acceleration.

**Lemma 24.** *Let $(p_t^{i,\varepsilon}, q_t^{i,\varepsilon})_{t\geq 0}$ be the solution to (A.2). Let $\Phi_i$ satisfy Assumption 2 and 3 with $\mathcal{A}^{\lambda_i,1}$ and critical point $\theta_*^i$. Then for any $0 < \varepsilon \leq 1$, we have*

$$\left\|\frac{\mathrm{d}p_t^{i,\varepsilon}}{\mathrm{d}t}\right\| \leq C_L^{(0)}(1 + \varepsilon^{-1}e^{-t/\varepsilon})\Big(\left\|q_0^i - \theta_*^i\right\| + \left\|p_0^i\right\|\Big) \quad (t \geq 0), \tag{A.6}$$

*where $C_L^{(0)} = 2 + 4LC_L$ and $L$ is the Lipschitz constant of $\nabla\Phi_i$.*

*Proof.* We first recall (A.5), where we have

$$\frac{\mathrm{d}p_t^{i,\varepsilon}}{\mathrm{d}t} = \frac{-e^{\frac{-t}{\varepsilon}}p_0^i}{\varepsilon} - \frac{\nabla\Phi_i(q_t^{i,\varepsilon})}{\varepsilon} + \frac{e^{\frac{-t}{\varepsilon}}}{\varepsilon^2}\int_0^t e^{\frac{s}{\varepsilon}}\nabla\Phi_i(q_s^i)\mathrm{d}s. \tag{A.7}$$

and notice that

$$\frac{\nabla\Phi_i(q_t^{i,\varepsilon})}{\varepsilon} = \frac{e^{-\frac{t}{\varepsilon}}}{\varepsilon^2}\int_0^t e^{\frac{s}{\varepsilon}}\nabla\Phi_i(q_t^{i,\varepsilon})\mathrm{d}s + \frac{e^{-\frac{t}{\varepsilon}}\nabla\Phi_i(q_t^{i,\varepsilon})}{\varepsilon}. \tag{A.8}$$

Combining (A.7) and (A.8), we get

$$\begin{aligned}
\left\|\frac{\mathrm{d}p_t^{i,\varepsilon}}{\mathrm{d}t}\right\| &= e^{\frac{-t}{\varepsilon}}\left\|\frac{-p_0^i}{\varepsilon} - \frac{\nabla\Phi_i(q_t^{i,\varepsilon}) - \nabla\Phi_i(\theta_*^i)}{\varepsilon} + \frac{1}{\varepsilon^2}\int_0^t e^{\frac{s}{\varepsilon}}(\nabla\Phi_i(q_s^i) - \nabla\Phi_i(q_t^{i,\varepsilon}))\mathrm{d}s\right\| \\
&\leq e^{\frac{-t}{\varepsilon}}\Big[\frac{\|p_0^i\|}{\varepsilon} + \frac{L\left\|q_t^{i,\varepsilon} - \theta_*^i\right\|}{\varepsilon} + \frac{L}{\varepsilon^2}\int_0^t e^{\frac{s}{\varepsilon}}\left\|q_s^{i,\varepsilon} - q_t^{i,\varepsilon}\right\|\mathrm{d}s\Big] \\
&= \varepsilon^{-1}e^{\frac{-t}{\varepsilon}}\Big(\left\|p_0^i\right\| + L\left\|q_t^{i,\varepsilon} - \theta_*^i\right\|\Big) + \frac{L}{\varepsilon^2}\int_0^t e^{\frac{-(t-s)}{\varepsilon}}\left\|q_s^{i,\varepsilon} - q_t^{i,\varepsilon}\right\|\mathrm{d}s \\
&\leq \varepsilon^{-1}e^{\frac{-t}{\varepsilon}}\Big(\left\|p_0^i\right\| + L\left\|q_t^{i,\varepsilon} - \theta_*^i\right\|\Big) + \frac{LC_L}{\varepsilon^2}\int_0^t e^{\frac{-(t-s)}{\varepsilon}}(t-s)\mathrm{d}s\Big(\left\|q_0^i - \theta_*^i\right\| + \left\|p_0^i\right\|\Big) \\
&\underbrace{\leq}_{(a_1)} \varepsilon^{-1}e^{\frac{-t}{\varepsilon}}\Big(\left\|p_0^i\right\| + 4L(L+2)(\left\|q_0^i - \theta_*^i\right\| + \left\|p_0^i\right\|)\Big) + LC_L\Big(\left\|q_0^i - \theta_*^i\right\| + \left\|p_0^i\right\|\Big) \\
&\leq C_L^{(0)}(1 + \varepsilon^{-1}e^{\frac{-t}{\varepsilon}})\Big(\left\|q_0^i - \theta_*^i\right\| + \left\|p_0^i\right\|\Big),
\end{aligned}$$

where $(a_1)$ is correct due to $(A.4)$ and the fact that $\int_0^{\frac{t}{\varepsilon}} e^{-s}s\mathrm{d}s < 1$. □

The previous lemmas provide the boundedness results needed for proving the following proposition. We show that the difference between the fixed sample SGMP (A.2) and the fixed sample stochastic gradient process (A.3) can be bounded by the sum of the distance between their initial values and a term that is linear in $\varepsilon$. Hence, in the fixed subsample case, we show that the solution of (A.2) converges to the solution of the limiting equation (A.3). In addition to the previous assumptions, we now also need to assume $\Phi_i$ to be convex.

**Proposition 6.** *Let $\Phi_i$ be convex and satisfy Assumption 2 and 3 with $\mathcal{A}^{\lambda_i,1}$ and critical point $\theta_*^i$. Let $(p_t^{i,\varepsilon}, q_t^{i,\varepsilon})_{t\geq 0}$ be the solution to (A.2). Then, for $0 < \varepsilon \leq 1$, we have*

$$\left\|q_t^{i,\varepsilon} - \theta_t^i\right\| \leq \left\|q_0^i - \theta_0^i\right\| + C_L^{(0)}\varepsilon(1+t)\Big( \left\|p_0^i\right\| + \left\|q_0^i - \theta_0^i\right\| + \left\|\theta_0^i - \theta_*^i\right\| \Big) \qquad (t \geq 0),$$

*where $C_L^{(0)} = 2 + 4LC_L$ and $L$ is the Lipschitz constant of $\nabla\Phi_i$.*

*Proof.* We take the derivative of $\left\|q_t^{i,\varepsilon} - \theta_t^i\right\|^2$ and obtain

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}\left\|q_t^{i,\varepsilon} - \theta_t^i\right\|^2}{\mathrm{d}t} &= \left\langle q_t^{i,\varepsilon} - \theta_t^i \right\rangle \frac{\mathrm{d}q_t^{i,\varepsilon}}{\mathrm{d}t} - \frac{\mathrm{d}\theta_t^i}{\mathrm{d}t} \\
&= -\left\langle q_t^{i,\varepsilon} - \theta_t^i \right\rangle \nabla\Phi_i(q_t^{i,\varepsilon}) - \nabla\Phi_i(\theta_t^i) - \varepsilon\left\langle q_t^{i,\varepsilon} - \theta_t^i \right\rangle \frac{\mathrm{d}p_t^{i,\varepsilon}}{\mathrm{d}t} \\
&\leq -\varepsilon\left\langle q_t^{i,\varepsilon} - \theta_t^i \right\rangle \frac{\mathrm{d}p_t^{i,\varepsilon}}{\mathrm{d}t} \leq \varepsilon\left\|q_t^{i,\varepsilon} - \theta_t^i\right\| \left\|\frac{\mathrm{d}p_t^{i,\varepsilon}}{\mathrm{d}t}\right\| \\
&\overset{(A.6)}{\leq} C_L^{(0)}\varepsilon(1+\varepsilon^{-1}e^{\frac{-t}{\varepsilon}})\Big( \left\|q_0^i - \theta_*^i\right\| + \left\|p_0^i\right\| \Big) \left\|q_t^{i,\varepsilon} - \theta_t^i\right\|,
\end{aligned}$$

which implies

$$\begin{aligned}
\frac{\mathrm{d}\left\|q_t^{i,\varepsilon} - \theta_t^i\right\|}{\mathrm{d}t} &= \Big(2\left\|q_t^{i,\varepsilon} - \theta_t^i\right\|\Big)^{-1}\frac{\mathrm{d}\left\|q_t^{i,\varepsilon} - \theta_t^i\right\|^2}{\mathrm{d}t} \\
&\leq C_L^{(0)}\varepsilon(1+\varepsilon^{-1}e^{\frac{-t}{\varepsilon}})\Big( \left\|q_0^i - \theta_*^i\right\| + \left\|p_0^i\right\| \Big).
\end{aligned}$$

Integrating both sides, we get

$$\begin{aligned}
\left\|q_t^{i,\varepsilon} - \theta_t^i\right\| &\leq \left\|q_0^i - \theta_0^i\right\| + C_L^{(0)}\int_0^t (\varepsilon + e^{\frac{-s}{\varepsilon}})\mathrm{d}s\Big( \left\|p_0^i\right\| + \left\|q_0^i - \theta_*^i\right\| \Big) \\
&= \left\|q_0^i - \theta_0^i\right\| + C_L^{(0)}\Big(\varepsilon t + \varepsilon(1 - e^{\frac{-t}{\varepsilon}})\Big)\Big( \left\|p_0^i\right\| + \left\|q_0^i - \theta_*^i\right\| \Big) \\
&\leq \left\|q_0^i - \theta_0^i\right\| + C_L^{(0)}\varepsilon(1+t)\Big( \left\|p_0^i\right\| + \left\|q_0^i - \theta_*^i\right\| \Big) \\
&\leq \left\|q_0^i - \theta_0^i\right\| + C_L^{(0)}\varepsilon(1+t)\Big( \left\|p_0^i\right\| + \left\|q_0^i - \theta_0^i\right\| + \left\|\theta_0^i - \theta_*^i\right\| \Big).
\end{aligned}$$

□

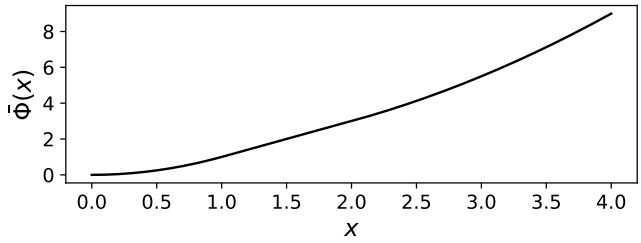


Figure A.1: Plot of function $\bar{\Phi}$ in Example 3.

## A.3 Other auxiliary results

**Example 3.** *Let $\bar{\Phi}$ be a function on $\mathbb{R}$ satisfying*

$$\bar{\Phi}(x) = \begin{cases} x^2, & x \in (-\infty, 1], \\ 2x - 1, & x \in (1, 2], \\ \frac{1}{2}x^2 + 1, & x \in (2, +\infty). \end{cases}$$

*$\bar{\Phi}$ satisfies Assumption 3 with $\lambda \in (0, 2/(8 + \bar{\alpha}^2)]$. As shown in Figure A.1, $\bar{\Phi}$ is convex but not strongly convex since it is linear on $(1, 2]$.*

**Lemma 25.** *Let $\bar{\Phi} \in \mathcal{C}^1(X, \mathbb{R})$ be strongly convex with constant $\mu$. Then $\bar{\Phi}$ satisfies Assumption 3 with $\mathcal{A}^{(\mu/\alpha^2)\wedge(1/4),\alpha}$.*

*Proof.* Since $\bar{\Phi}$ is strongly convex, we have

$$(x - \theta_*) \cdot (\nabla\bar{\Phi}(x) - \nabla\bar{\Phi}(\theta_*)) \geq \mu \|x - \theta_*\|^2 .$$

By Lagrange's mean value theorem, there exist some $\xi$ lying on the line between $x$ and $\theta_*$, such that

$$\bar{\Phi}(x) - \bar{\Phi}(\theta_*) = (x - \theta_*) \cdot \nabla\bar{\Phi}(\xi).$$

Let $\xi = \theta_* + t(x - \theta_*)$ for some $0 \leq t \leq 1$. By strong convexity, we have

$$(x - \xi) \cdot (\nabla\bar{\Phi}(x) - \nabla\bar{\Phi}(\xi)) \geq \mu \|x - \xi\|^2 .$$

Replacing $\xi$ by $\theta_* + t(x - \theta_*)$, we get

$$(x - \theta_*) \cdot (\nabla\bar{\Phi}(x) - \nabla\bar{\Phi}(\xi)) \geq (1-t)(x - \theta_*) \cdot (\nabla\bar{\Phi}(x) - \nabla\bar{\Phi}(\xi)) \geq \mu \left\| x - \xi \right\|^2 .$$

This implies

$$(x - \theta_*) \cdot \nabla\bar{\Phi}(x) \geq (x - \theta_*) \cdot \nabla\bar{\Phi}(\xi) + \mu \left\| x - \xi \right\|^2 \geq \bar{\Phi}(x) - \bar{\Phi}(\theta_*).$$

Hence,

$$(x - \theta_*) \cdot \nabla\Phi(x)/2 \geq (\bar{\Phi}(x) - \bar{\Phi}(\theta_*))/4 + \mu \left\| x - \theta_* \right\|^2 /4).$$

Then, we choose $\lambda = (\mu/\alpha^2) \wedge (1/4)$ and the proof is completed. □

**Lemma 26.** *We assume that $\Phi_i$ is convex and satisfies Assumption 2 for all $i \in I$. Let $(\theta_t^\varepsilon)_{t\geq 0}$ be the solution to system (4.16). Then we have*

$$\|\theta_t^\varepsilon\| \leq \|\theta_0\| + C_\Phi t \qquad (t \geq 0).$$

*Proof.* We differentiate $\|\theta_t^\varepsilon\|^2$ and obtain

$$\begin{aligned}
\frac{\mathrm{d}\left\|\theta_t^\varepsilon\right\|^2}{\mathrm{d}t} &= 2\left\langle \theta_t^\varepsilon \right\rangle \frac{\mathrm{d}\theta_t^\varepsilon}{\mathrm{d}t} = -2\left\langle \theta_t^\varepsilon \right\rangle \nabla\Phi_{i^{\varepsilon,\delta}(t)}(\theta_t^\varepsilon) \\
&= -2\left\langle \theta_t^\varepsilon, \nabla\Phi_{i^{\varepsilon,\delta}(t)}(\theta_t^\varepsilon) - \nabla\Phi_{i^{\varepsilon,\delta}(t)}(0)\right\rangle - 2\left\langle \theta_t^\varepsilon \right\rangle \nabla\Phi(0) \\
&\leq 2\left\|\theta_t^\varepsilon\right\| \left\|\nabla\Phi_{i^{\varepsilon,\delta}(t)}(0)\right\| \leq 2C_\Phi \left\|\theta_t^\varepsilon\right\|,
\end{aligned}$$

which implies $\frac{\mathrm{d}\|\theta_t^\varepsilon\|}{\mathrm{d}t} \leq C_\Phi$ and, hence, $\|\theta_t^\varepsilon\| \leq \|\theta_0\| + C_\Phi t$. □

**Lemma 27.** *We assume that $\Phi_i$ is strongly convex with $\mu > 0$ for all $i \in I$. Let $(\theta_t)_{t\geq 0}$ be the solution to (4.3). Then, we have*

$$\|\theta_t\| \leq \|\theta_0\| e^{-\mu t} + C_\Phi \qquad (t \geq 0).$$

*Proof.* We differentiate $\|\theta_t\|^2$,

$$\frac{\mathrm{d}\|\theta_t\|^2}{\mathrm{d}t} \leq -2\mu \|\theta_t\|^2 + 2\|\theta_t\| \left\|\nabla\Phi_{i(t)}(0)\right\| \leq -\mu\|\theta_t\|^2 + \frac{1}{\mu}\left\|\nabla\Phi_{i(t)}(0)\right\|^2,$$

By Grönwall's inequality, we have

$$\|\theta_t\|^2 \leq e^{-\mu t}\|\theta_0\|^2 + \frac{1}{\mu^2}\left\|\nabla\Phi_{i(t)}(0)\right\|^2$$

and choose $C_\Phi = \frac{1}{\mu^2}\sup_{i=1,\ldots,N}\|\nabla\Phi_i(0)\|^2$. □

**Lemma 28.** *Let $C$ and $\lambda$ be two positive constants and let $(t_n)_{n\geq 0}$ be a sequence with $t_0 \geq 0$ that satisfies the following recurrence relation*

$$t_{n+1} = t_n + Ce^{-\lambda t_n}.$$

*Then* $\lim_{n\to+\infty} t_n = +\infty$.

*Proof.* It is obvious that $\{t_n\}_{n\geq 0}$ is a strictly increasing sequence. We now assume that $\lim_{n\to+\infty} t_n < +\infty$. let $A$ be the smallest upper bound of $(t_n)_{n\geq 0}$. Then for any $\varepsilon > 0$, there exist $n_0$ such that, for any $n \geq n_0$, $t_n \geq A - \varepsilon$. We have

$$t_{n+1} = t_n + Ce^{-\lambda t_n} \geq A - \varepsilon + Ce^{-A}.$$

Set $\varepsilon \leq Ce^{-A}/2$, then $t_{n+1} > A$, contradicting the assumption that $t_n \leq A$ ($n \in \mathbb{N}$). Hence $\lim_{n\to+\infty} t_n = +\infty$. □

**Lemma 29.** *Let $(p_t^i, q_t^i)_{t\geq 0}$ solve (4.20). Let $\theta_*^i$ be the critical point of $\Phi_i$ for $i \in I$. Under Assumption 2, 6 with $\mathcal{A}^{\lambda_i,2}$, and 7 with $\lambda = \min_{i\in I} \lambda_i$, we have*

$$V^i(t, q_t^i, p_t^i) \leq e^{-\lambda t} V^i(0, q_0^i, p_0^i) \qquad (t \geq 0),$$

*where $V^i(t, x, y) = \varepsilon(t)[\Phi_i(x) - \Phi_i(\theta_*^i)] + \frac{1}{4}\Big( \|x - \theta_*^i + \varepsilon(t)y\|^2 + \|\varepsilon(t)y\|^2 \Big)$ is a Lyapunov function.*

*Proof.* First, we differentiate $V^i(t, q_t^i, p_t^i)$ and obtain by (4.20),

$$\begin{aligned}
\mathrm{d}V^i(t, q_t^i, p_t^i)/\mathrm{d}t &= \mathrm{d}(\varepsilon(t)(\Phi_i(q_t^i) - \Phi_i(\theta_*)))/\mathrm{d}t + \frac{1}{4}\mathrm{d}\Big( \big\|q_t^i - \theta_*^i + \varepsilon(t)p_t^i\big\|^2 + \big\|\varepsilon(t)p_t^i\big\|^2 \Big)/\mathrm{d}t \\
&= \varepsilon'(t)(\Phi_i(q_t^i) - \Phi_i(\theta_*^i)) + \varepsilon(t)\left\langle \nabla\Phi_i(q_t^i)\right\rangle \mathrm{d}q_t^i/\mathrm{d}t \\
&\quad + \frac{1}{2}\Big( \left\langle q_t^i - \theta_*^i + \varepsilon(t)p_t^i\right\rangle \mathrm{d}(q_t^i - \theta_*^i + \varepsilon(t)p_t^i)/\mathrm{d}t + \left\langle \varepsilon(t)p_t^i\right\rangle \mathrm{d}(\varepsilon(t)p_t^i)/\mathrm{d}t\Big) \\
&= \varepsilon'(t)(\Phi_i(q_t^i) - \Phi_i(\theta_*^i)) + \varepsilon(t)\left\langle \nabla\Phi_i(q_t^i)\right\rangle p_t^i \\
&\quad + \frac{1}{2}\Big( \left\langle q_t^i - \theta_*^i + \varepsilon(t)p_t^i\right\rangle p_t^i + \varepsilon'(t)p_t^i + \varepsilon(t)\mathrm{d}p_t^i/\mathrm{d}t \\
&\qquad + \left\langle \varepsilon(t)p_t^i\right\rangle \varepsilon'(t)p_t^i + \varepsilon(t)\mathrm{d}p_t^i/\mathrm{d}t\Big) \\
&\overset{(4.20)}{=} \varepsilon'(t)(\Phi_i(q_t^i) - \Phi_i(\theta_*^i)) + \varepsilon(t)\left\langle \nabla\Phi_i(q_t)\right\rangle p_t^i \\
&\quad + \frac{1}{2}\Big( \left\langle q_t^i - \theta_*^i + \varepsilon(t)p_t^i\right\rangle \varepsilon'(t)p_t^i - \nabla\Phi_i(q_t^i) \\
&\qquad + \left\langle \varepsilon(t)p_t^i\right\rangle \varepsilon'(t)p_t^i - \nabla\Phi_i(q_t^i) - p_t^i\Big) \\
&= \underbrace{\varepsilon'(t)(\Phi_i(q_t^i) - \Phi_i(\theta_*^i))}_{(\alpha_{11})} - \frac{1}{2}\left\langle q_t^i - \theta_*^i\right\rangle \nabla\Phi_i(q_t^i) \\
&\quad + \varepsilon(t)(\varepsilon'(t) - \tfrac{1}{2})\big\|p_t^i\big\|^2 + \underbrace{\frac{\varepsilon'(t)}{2}\left\langle q_t^i - \theta_*^i\right\rangle p_t^i}_{(\alpha_{12})}.
\end{aligned} \tag{A.9}$$

Note that $(\alpha_{11}) + m(t)\varepsilon'(t)\left\|p_t^i\right\|^2 \leq 0$ since $\varepsilon'(t) \leq 0$ and $\Phi_i(q_t^i) - \Phi_i(\theta_*^i) \geq 0$. By the $\varepsilon$-Young's inequality, we have $(\alpha_{12}) \leq \frac{\lambda}{4}\left\|q_t^i - \theta_*^i\right\|^2 + \frac{(\varepsilon'(t))^2}{4\lambda}\left\|p_t^i\right\|^2$. Then we have

$$
\begin{aligned}
\text{(A.9)} \leq & -\frac{1}{2}\left\langle q_t^i - \theta_*^i \right\rangle \nabla\Phi_i(q_t^i) - \frac{\varepsilon(t)}{2}\left\|p_t^i\right\|^2 + \frac{\lambda}{4}\left\|q_t^i - \theta_*^i\right\|^2 + \frac{(\varepsilon'(t))^2}{4\lambda}\left\|p_t^i\right\|^2 \\
\underbrace{\leq}_{(\beta_1)} & -\lambda\Big(\Phi_i(q_t^i) - \Phi_i(\theta_*^i) + \left\|q_t^i - \theta_*^i\right\|^2\Big) - \frac{\varepsilon(t)}{4}\left\|p_t^i\right\|^2 + \frac{\lambda}{4}\left\|q_t^i - \theta_*^i\right\|^2 \\
\underbrace{\leq}_{(\beta_2)} & -\lambda\Big(\Phi_i(q_t^i) - \Phi_i(\theta_*^i) + \frac{\left\|q_t^i - \theta_*^i + \varepsilon(t)p_t^i\right\|^2}{4} + \frac{\left\|\varepsilon(t)p_t^i\right\|^2}{4}\Big) + \frac{3\lambda\varepsilon^2(t) - \varepsilon(t)}{4}\left\|p_t^i\right\|^2 \\
\leq & -\lambda V^i(t, q_t^i, p_t^i).
\end{aligned}
$$

$(\beta_1)$ holds due to Assumption 6 and 7. In particular, Assumption 7 implies $\frac{(\varepsilon'(t))^2}{4\lambda} - \frac{\varepsilon(t)}{2} \leq \frac{\lambda^2(\varepsilon(t))^2}{4\lambda} - \frac{\varepsilon(t)}{2} \leq \frac{\lambda\varepsilon(t)}{4} - \frac{\varepsilon(t)}{2} \leq -\frac{\varepsilon(t)}{4}$. Moreover, $(\beta_2)$ follows from $\frac{\left\|q_t^i - \theta_*^i + \varepsilon(t)p_t^i\right\|^2}{4} \leq \frac{\left\|q_t^i - \theta_*^i\right\|^2 + \left\|\varepsilon(t)p_t^i\right\|^2}{2}$. Finally, by Grönwall's inequality we have

$$
V^i(t, q_t^i, p_t^i) \leq e^{-\lambda t}V^i(0, q_0^i, p_0^i).
$$

□

**Lemma 30.** *Let $\varepsilon(t) > 0$ be a strictly decreasing differentiable function and satisfy Assumption 7 with $\lambda = \mu/4$, where $\mu > 0$. Then we have the following inequalities,*

$$
\int_0^t \varepsilon^2(s)e^{\mu s}\mathrm{d}s \leq 2\mu^{-1}e^{\mu t}\varepsilon^2(t), \qquad \int_0^t e^{-2\mathcal{E}(s)}e^{\mu s}\mathrm{d}s \leq (1+\mu)\varepsilon_0 e^{\mu t}.
$$

*Proof.* For the first inequality, we have

$$
\int_0^t \varepsilon^2(s)e^{\mu s}\mathrm{d}s = \mu^{-1}\int_0^t \varepsilon^2(s)\mathrm{d}e^{\mu s} = \mu^{-1}\Big[\varepsilon^2(t)e^{\mu t} - \varepsilon_0 - 2\int_0^t \varepsilon(s)\varepsilon'(s)e^{\mu s}\mathrm{d}s\Big],
$$

using integration by parts. Under the Assumption 7,

$$-2\int_0^t \varepsilon(s)\varepsilon'(s)e^{\mu s}\mathrm{d}s \leq \frac{\mu}{2}\int_0^t \varepsilon^2(s)e^{\mu s}\mathrm{d}s.$$

Hence,

$$\int_0^t \varepsilon^2(s)e^{\mu s}\mathrm{d}s \leq \mu^{-1}e^{\mu t}\varepsilon^2(t) + \frac{1}{2}\int_0^t \varepsilon^2(s)e^{\mu s}\mathrm{d}s.$$

This implies that

$$\int_0^t \varepsilon^2(s)e^{\mu s}\mathrm{d}s \leq 2\mu^{-1}e^{\mu t}\varepsilon^2(t).$$

For the second inequality, we use integration by parts and notice that

$$\begin{aligned}
\int_0^t e^{-2\mathcal{E}(s)}e^{\mu s}\mathrm{d}s &= -\frac{1}{2}\int_0^t \varepsilon(s)e^{\mu s}\mathrm{d}e^{-2\mathcal{E}(s)} \\
&= \frac{1}{2}\Big(\varepsilon_0 - \varepsilon(t)e^{\mu t}e^{-2\mathcal{E}(t)}\Big) + \frac{1}{2}\int_0^t e^{-2\mathcal{E}(s)}e^{\mu s}(\mu\varepsilon(s) + \varepsilon'(s))\mathrm{d}s \\
&\leq \frac{\varepsilon_0}{2} + \frac{3\mu\varepsilon_0 e^{\mu t}}{8}\int_0^t e^{-2\mathcal{E}(s)}\mathrm{d}s \\
&\leq \frac{\varepsilon_0}{2} + \frac{3\mu\varepsilon_0 e^{\mu t}}{8}\int_0^t e^{\frac{-2s}{\varepsilon_0}}\mathrm{d}s \leq (1+\mu)\varepsilon_0 e^{\mu t}.
\end{aligned}$$

This completes the proof. □

**Lemma 31.** *Let $(\varepsilon_n)_{n\geq 0}$ be a non-negative decreasing sequence with $\lim_{n\to\infty}\varepsilon_n = 0$. Let $(d_n)_{n\geq 0}$ be a positive sequence satisfying $d_{n+1} \leq cd_n + \varepsilon_n$, where $0 < c < 1$. Then we have $\lim_{n\to\infty} d_n = 0$. Furthermore, if $\varepsilon_n \leq Be^{-bn}$ for some constant $\tilde{B}, \tilde{b} > 0$ we have $d_n \leq \tilde{B}e^{-\tilde{b}n}$.*

*Proof.* It is obvious that $(\varepsilon_n)_{n\geq 0}$ is bounded and we denote the upper bound by $C$. Hence $(d_n)_{n\geq 0}$ can be bounded by $D := d_0 + C/(1-c)$. Since $\lim_{n\to+\infty}\varepsilon_n = 0$, for any $\varepsilon > 0$, there exist a $N_0$, for any $N \geq N_0$, $\varepsilon_n \leq \varepsilon$. Hence, for any $n \geq N_0$, we have

$$d_{n+1} \leq cd_n + \varepsilon_n \leq cd_n + \varepsilon.$$

which implies that

$$d_n \le c^{n-N_0} d_{N_0} + \frac{\varepsilon}{1-c}.$$

Since $\varepsilon$ is arbitrary, we get $\lim_{n\to\infty} d_n = 0$, which finishes the proof of the first part.

Now, assuming that $\varepsilon_n \le Be^{-bn}$, the relation $d_{n+1} \le cd_n + \varepsilon_n$ implies that

$$d_{n+1} \le \sum_{i=0}^{n} c^{n-i}\varepsilon_i \le B\sum_{i=0}^{n} c^{n-i}e^{-bi} = Bc^n \sum_{i=0}^{n} e^{-(\log c+b)i}.$$

If $b + \log c \ge 0$, $d_{n+1} \le Bnc^n$ the result is obvious. For $b + \log c < 0$, we have

$$d_{n+1} \le Bc^n \frac{e^{-(\log c+b)(n+1)} - 1}{e^{-(\log c+b)} - 1} \le Bc^n \frac{e^{-(\log c+b)(n+1)}}{e^{-(\log c+b)} - 1} \le \tilde{B}e^{-b(n+1)}.$$

□

**Lemma 32.** *Let $G_k^n = \prod_{i=k}^{n} \left(1 - \frac{b}{\sqrt{i}}\right)$, where $0 < b < 1$. Then there exist $c > 0$ such that for any $0 \le k < n$, we have $G_k^n \le e^{-c(\sqrt{n}-\sqrt{k})}$.*

*Proof.* For any $0 \le x < 1$ we have $\log(1-x) \le -x$. This implies that

$$\begin{aligned}
\log G_k^n =& \log\Big(\prod_{i=k}^{n}\Big(1 - \frac{b}{\sqrt{i}}\Big)\Big) = \sum_{i=k}^{n} \log\Big(1 - \frac{b}{\sqrt{i}}\Big) \le -b\sum_{i=k}^{n} \frac{1}{\sqrt{i}} \\
=& -b\sum_{i=k}^{n} \frac{1}{\sqrt{i}} \int_i^{i+1} \mathrm{d}x \le -b\sum_{i=k}^{n} \int_i^{i+1} \frac{1}{\sqrt{x}}\mathrm{d}x = -b\int_k^{n+1} \frac{1}{\sqrt{x}}\mathrm{d}x \\
=& \frac{-b(\sqrt{n+1} - \sqrt{k})}{2} \le \frac{-b(\sqrt{n} - \sqrt{k})}{2}.
\end{aligned}$$

Finally, by taking exponential both side and let $c = -b/2$, we finish the proof. □

## A.4 Kushner's perturbed test function theory

In [68, Chapter 7], the author considers a dynamical system with state space $\mathbb{R}^d$ of the form

$$\frac{\mathrm{d}x^\gamma}{\mathrm{d}t} = \bar{G}(x^\gamma) + \tilde{G}(x^\gamma, \xi^\gamma) + F(x^\gamma, \xi^\gamma)/\gamma, \tag{A.10}$$

where $\gamma > 0$ is a parameter and $(\xi_t^\gamma)_{t\geq 0}$ is a stochastic process. The author then studies the limiting behaviour of this dynamical system to, e.g. ODEs of type $\frac{\mathrm{d}x}{\mathrm{d}t} = \bar{G}(x)$. We now cite their main result.

**Theorem 10** ([68])**.** *We assume that assumptions (A4.2) to (A4.6) (see below) hold and let* $(x^\gamma(0))_{\gamma>0}$ *be tight. Then* $(x^\gamma(t))_{t\geq 0}$ *is tight with respect to* $\gamma > 0$ *and the limit of any weakly convergent subsequence solves the martingale problem with the associate operator* $A$ *defined by*

$$Af(x) = \nabla_x f \cdot \bar{G}(x) + B_0 f(x),$$

*for appropriate test functions* $f : \mathbb{R}^d \to \mathbb{R}$*.*

We apply this theorem only in case $F = 0$, which simplifies the discussion in the following. For instance, we skip the definition of $B_0$, as $B_0 = 0$. Then, the operator $A$ is the infinitesimal generator of $\frac{\mathrm{d}x}{\mathrm{d}t} = \bar{G}(x)$ and we obtain convergence. We now list Assumptions (A4.2), (A4.3), and (A4.5). We skip Assumptions (A4.4) and (A4.6) as they are redundant if $F = 0$.

**(A4.2):** $\bar{G},\ \tilde{G},\ F,\ \nabla_x F$ are continuous.

**(A4.3):** The process $(\xi_t^\gamma)_{t\geq 0}$ satisfies $\xi_t^\gamma = \xi_{t/\gamma^2}$ $(t \geq 0)$ for a bounded and right continuous $(\xi_t)_{t\geq 0}$.

**(A4.5):** For each $x \in \mathbb{R}^d$, as $t, \tau \to \infty$,

$$\frac{1}{\tau}\int_t^{t+\tau} \mathbb{E}[\tilde{G}(x, \xi(s))|(\xi(s'))_{s'\leq t}]\mathrm{d}s \to 0,$$

almost surely.

We employ Theorem 10 in the proof of Theorem 6. There, we have $\xi_t = i(t)$, $\gamma = \nu^{1/2}$, $\bar{G}(x, y) = (y, -\nabla_x \bar{\Phi}(x) - \alpha y)$, $\tilde{G}(x, y, i) = (0, -\nabla_x \Phi_i(x) + \nabla_x \bar{\Phi}(x))$ and $F = 0$.

# Bibliography


[1] A. Ali, E. Dobriban, and R. Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pages 233–244, 2020.

[2] L. N. Andersen, S. Asmussen, P. W. Glynn, and M. Pihlsgård. Lévy processes with two-sided reflection. In *Lévy Matters V: Functionals of Lévy Processes*, pages 67–182. Springer International Publishing, 2015.

[3] H. Attouch and F. Alvarez. The heavy ball with friction dynamical system for convex constrained minimization problems. In *Optimization: Proceedings of the 9th Belgian-French-German Conference on Optimization Namur, September 7–11, 1998*, pages 25–35. Springer, 2000.

[4] H. Attouch, A. Balhag, Z. Chbani, and H. Riahi. Fast convex optimization via inertial dynamics combining viscous and Hessian-driven damping with time rescaling. *arXiv preprint arXiv:2009.07620*, 2020.

[5] H. Attouch, R. I. Bot, and D.-K. Nguyen. Fast convex optimization via time scale and averaging of the steepest descent. *arXiv preprint arXiv:2208.08260*, 2022.

[6] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168:123–175, 2018.

[7] H. Attouch, Z. Chbani, and H. Riahi. Fast convex optimization via time scaling of damped inertial gradient dynamics. *Pure and Applied Functional Analysis*, 2019.

[8] H. Attouch, Z. Chbani, and H. Riahi. Fast proximal methods via time scaling of damped inertial dynamics. *SIAM Journal on Optimization*, 29(3):2227–2256, 2019.

[9] H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method, I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(01):1–34, 2000.

[10] J.-F. Aujol, C. Dossal, and A. Rondepierre. Convergence Rates of the Heavy Ball Method for Quasi-strongly Convex Optimization. *SIAM Journal on Optimization*, 32(3):1817–1842, 2022.

[11] J.-F. Aujol, C. Dossal, and A. Rondepierre. Convergence rates of the heavy-ball method with Lojasiewicz property. *Mathematical Programming*, 198:195–254, 2023.

[12] A. V. Babin, A. A. Ilyin, and E. S. Titi. On the regularization mechanism for the periodic Korteweg–de Vries equation. *Communications on Pure and Applied Mathematics*, 64(5):591–648, 2011.

[13] P. Bianchi. A stochastic proximal point algorithm: convergence and application to convex optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–4, 2015.

[14] J. Blanchet and K. Murthy. Exact simulation of multidimensional reflected brownian motion. *Journal of Applied Probability*, 55(1):137–156, 2018.

[15] J. Bourgain. Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations. I. Schrödinger. *Geometric and Functional Analysis*, 1993.

[16] J. Bourgain. Approximation of solutions of the cubic nonlinear Schrödinger equations by finite-dimensional equations and nonsqueezing properties. *International Mathematics Research Notices*, 1994(2):79–90, 01 1994.

[17] J. Bourgain. Periodic nonlinear Schrödinger equation and invariant measures. *Communications in Mathematical Physics*, 166(1):1 – 26, 1994.

[18] J. Bourgain. A remark on normal forms and the "I-method" for periodic NLS. *Journal d'Analyse Mathematique*, 94, 2004.

[19] K. Bredies and D. Lorenz. *Variational Methods*, pages 251–443. Springer International Publishing, 2018.

[20] J. Brüdern and O. Robert. Rational points on linear slices of diagonal hypersurfaces. *Nagoya Mathematical Journal*, 218:51 – 100, 2015.

[21] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8:231–357, 2015.

[22] T. Buckmaster, P. Germain, Z. Hani, and J. Shatah. Analysis of (CR) in higher dimension. *International Mathematics Research Notices*, 2019(4):1265–1280, 07 2017.

[23] T. Buckmaster, P. Germain, Z. Hani, and J. Shatah. Effective dynamics of the nonlinear Schrödinger equation on large domains. *Communications on Pure and Applied Mathematics*, 71(7):1407–1460, 2018.

[24] T. Buckmaster, P. Germain, Z. Hani, and J. Shatah. Onset of the wave turbulence description of the longtime behavior of the nonlinear Schrödinger equation. *Inventiones mathematicae*, 225:787–855, 2021.

[25] L. L. Cam. Maximum likelihood: An introduction. *International Statistical Review*, 58(2):153–171, 1990.

[26] G. Casella. Empirical Bayes Gibbs sampling. *Biostatistics*, 2(4):485–500, 12 2001.

[27] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb. Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

[28] F. Chen, D. Sondak, P. Protopapas, M. Mattheakis, S. Liu, D. Agarwal, and M. D. Giovanni. Neurodiffeq: A python package for solving differential equations with neural networks. *Journal of Open Source Software*, 5(46), 2020.

[29] B.-E. Cherief-Abdellatif. Consistency of elbo maximization for model selection. In *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, volume 96, pages 11–31. PMLR, 2019.

[30] M. Christ. Power series solution of a nonlinear Schrödinger equation. *Annals of Mathematics Studies*, 163:131–155, 2007.

[31] B. Cloez and M. Hairer. Exponential ergodicity for Markov processes with random switching. *Bernoulli*, 21(1):505–536, 2015.

[32] J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pages 1310–1320, 2019.

[33] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao. Global well-posedness and scattering for the energy-critical nonlinear Schrödinger equation in $\mathbb{R}^3$. *Annals of Mathematics*, 167:767–865, 2008.

[34] H. Daneshmand, J. Kohler, A. Lucchi, and T. Hofmann. Escaping Saddles with Stochastic Gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1155–1164. PMLR, 10–15 Jul 2018.

[35] J. de Wiljes, S. Reich, and W. Stannat. Long-time stability and accuracy of the ensemble kalman–bucy filter for fully observed processes and small measurement noise. *SIAM Journal on Applied Dynamical Systems*, 17(2):1152–1181, 2018.

[36] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[37] A. Défossez, L. Bottou, F. Bach, and N. Usunier. A Simple Convergence Proof of Adam and Adagrad. *Transactions on Machine Learning Research*, 2022.

[38] H. Deqing and C. Sergei. Long-time average cost control of stochastic systems using sum of squares of polynomials. In *2015 34th Chinese Control Conference (CCC)*, pages 2344–2349, 2015.

[39] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos. Gradient Descent Can Take Exponential Time to Escape Saddle Points. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[40] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[41] J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

[42] A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47, 03 2017.

[43] A. Eftekhari, B. Vandereycken, G. Vilmart, and K. C. Zygalakis. Explicit stabilised gradient descent for faster strongly convex optimisation. *Bit Numer Math*, 61:119–139, 2021.

[44] M. Erdoğan and N. Tzirakis. Talbot effect for the cubic non-linear Schröedinger equation on the torus. *Mathematical Research Letters*, 20, 03 2013.

[45] E. Faou, P. Germain, and Z. Hani. The weakly nonlinear large box limit of the 2D cubic nonlinear Schrödinger equation. *J. Amer. Math. Soc.*, 29, 2015.

[46] P. Germain, N. Masmoudi, and J. Shatah. Global solutions for 3D quadratic Schrödinger equations. *International Mathematics Research Notices - INT MATH RES NOTICES*, 3:414–432, 2009.

[47] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[48] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[49] B. Grébert and T. Kappeler. *The Defocusing NLS Equation and Its Normal Form*. EMS Series of Lectures in Mathematics. European Mathematical Society Publishing House, 2014.

[50] Z. Guo, S. Kwon, and T. Oh. Poincaré-Dulac normal form reduction for unconditional well-posedness of the periodic cubic NLS. *Communications in Mathematical Physics*, 322:19–48, 2013.

[51] E. Hairer, C. Lubich, and G. Wanner. Geometric numerical integration illustrated by the Strömer-Verlet method. *Acta Numerica*, pages 1–51, 05 2003.

[52] P. C. Hansen. *Discrete Inverse Problems*. Society for Industrial and Applied Mathematics, 2010.

[53] D. Heath-Brown. A new form of the circle method, and its application to quadratic forms. *Journal für die reine und angewandte Mathematik*, 481:149–206, 1996.

[54] G. Hinton. Online course in Neural Networks for Machine Learning, 2012.

[55] A. D. Ionescu and F. Pusateri. Nonlinear fractional Schrödinger equations in one dimension. *Journal of Functional Analysis*, 266(1):139–176, 2014.

[56] A. D. Ionescu and F. Pusateri. Global regularity for 2D water waves with surface tension. *Memoirs of the American Mathematical Society*, 2018.

[57] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. On Nonconvex Optimization for Machine Learning: Gradients, Stochasticity, and Saddle Points. *J. ACM*, 68(2), feb 2021.

[58] K. Jin, J. Latz, C. Liu, and C.-B. Schönlieb. A Continuous-time Stochastic Gradient Descent Method for Continuous Data. *Journal of Machine Learning Research*, 24:1–48, 2023.

[59] T. Kappeler, P. Lohrmann, P. Topalov, and T. Z. Nguyen. Birkhoff coordinates for the focusing NLS equation. *Communications in Mathematical Physics*, 285:1087–1107, 02 2009.

[60] T. Kappeler and M. Makarov. On Birkhoff coordinates for KdV. *Annales Henri Poincaré*, 2:807–856, 2001.

[61] T. Kappeler, B. Schaad, and P. Topalov. Scattering-like phenomena of the periodic defocusing NLS equation. *Mathematical Research Letters*, 24, 05 2015.

[62] J. Kato and F. Pusateri. A new proof of long-range scattering for critical nonlinear Schrödinger equations. *Differential and Integral Equations*, 24:923–940, 2011.

[63] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations*, 2017.

[64] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 12 2014.

[65] N. B. Kovachki and A. M. Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021.

[66] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images, 2009.

[67] H. Kushner. *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, volume 6 of *MIT Press Series in Signal Processing, Optimization, and Control*. MIT Press, Cambridge, 1984.

[68] H. Kushner. *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*. Birkhäuser Basel, 1990.

[69] H. Kushner and G. Yin. *Stochastic Approximation Algorithms and Recursive Algorithms and Applications*. Springer, New York, NY, 2003.

[70] J. Latz. Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 31(39), 2021.

[71] J. Latz. Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 31, 07 2021.

[72] J. Latz. Gradient flows and randomised thresholding: sparse inversion and classification. *Inverse Problems*, 38(12), nov 2022.

[73] J. Latz. Correction to: analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 34(5):146, 2024.

[74] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-Supervised Nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 562–570. PMLR, 09–12 May 2015.

[75] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient Descent Only Converges to Minimizers. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257. PMLR, 23–26 Jun 2016.

[76] Q. Li, C. Tai, and W. E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 2101–2110, 2017.

[77] Q. Li, C. Tai, and W. E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

[78] Z. Li, N. B. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations, ICLR*, 2021.

[79] T. Liggett. *Continuous Time Markov Processes: An Introduction*. American Mathematical Soc., 2010.

[80] F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[81] Y. Liu. Discretization of a class of reflected diffusion processes. *Mathematics and Computers in Simulation*, 38(1):103–108, 1995.

[82] Y. Liu, Y. Gao, and W. Yin. An Improved Analysis of Stochastic Gradient Descent with Momentum. In *Advances in Neural Information Processing Systems*, volume 34, 2020.

[83] G. J. Lord, C. E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDEs*. Cambridge University Press, 2014.

[84] L. Lu, X. Meng, Z. Mao, and G. E. Karniadakis. Deepxde: A deep learning library for solving differential equations. *SIAM Review*, 63(1):208–228, 2021.

[85] J. Lukkarinen and H. Spohn. Weakly nonlinear Schrödinger equation with random initial data. *Inventiones mathematicae*, 183:79–188, 01 2011.

[86] S. Mandt, M. D. Hoffman, and D. M. Blei. A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML 2016)*, pages 354–363, 2016.

[87] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. *Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

[88] S. May, R. Rannacher, and B. Vexler. Error analysis for a finite element approximation of elliptic dirichlet boundary control problems. *SIAM Journal on Control and Optimization*, 51(3):2585–2611, 2013.

[89] R. I. McLachlan and A. Stern. Functional Equivariance and Conservation Laws in Numerical Integration. *Foundations of Computational Mathematics*, 24(1):149–177, 2024.

[90] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

[91] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[92] Y. Nesterov. A method for solving the convex programming problem with convergence rate $o(\frac{1}{k^2})$. *Soviet Mathematics Doklady*, pages 372–367, 1983.

[93] B. Neyshabur, R. Tomioka, and N. Srebro. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning, 2015.

[94] S. Novikov, S. V. Manakov, L. Pitaevskii, and V. Zakharov. *Theory of Solitons: The Inverse Scattering Method*. Monographs in Contemporary Mathematics. Springer-Verlag, 1984.

[95] T. Oh and Y. Wang. Normal form approach to the one-dimensional periodic cubic nonlinear Schrödinger equation in almost critical Fourier-Lebesgue spaces. *Journal d'Analyse Mathématique*, 143:1–40, 06 2021.

[96] M. Overholt. *A Course in Analytic Number Theory*, volume 160 of *Graduate Studies in Mathematics*. American Mathematical Society, 2014.

[97] I. Panageas and G. Piliouras. Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions. In *8th Innovations in Theoretical Computer Science Conference*. ITCS, 2017.

[98] J. B. Pedro, J. Maronas, and R. Paredes. Solving partial differential equations with neural networks. *arXiv:1912.04737*, 2019.

[99] R. Pettersson. Approximations for stochastic differential equations with reflecting convex boundaries. *Stochastic Processes and their Applications*, 59:295–308, 1995.

[100] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, page 2817–2826, 2017.

[101] B. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[102] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[103] M. Procesi and C. Procesi. A KAM algorithm for the resonant non–linear Schrödinger equation. *Advances in Mathematics*, 272:399–470, 2015.

[104] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Solution of Ordinary Differential Equations*, pages 479–538. Springer Berlin Heidelberg, 2007.

[105] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[106] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, Heidelberg, 2013.

[107] H. Robbins and S. Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[108] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer New York, 2004.

[109] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[110] A. Scagliotti and P. Colli Franzone. A Piecewise Conservative Method for Unconstrained Convex Optimization. *Comput. Optim. Appl.*, 81(1):251–288, 2022.

[111] O. Sebbouh, C. Dossal, and A. Rondepierre. Convergence Rates of Damped Inertial Dynamics under Geometric Conditions and Perturbations. *SIAM Journal on Optimization*, 30(3):1850–1877, 2020.

[112] A. Sekhari, K. Sridharan, and S. Kale. SGD: The Role of Implicit Regularization, Batch-size and Multiple-epochs. In *Advances in Neural Information Processing Systems*, 2021.

[113] J. Shatah. Normal forms and quadratic nonlinear Klein-Gordon equations. *Communications on Pure and Applied Mathematics*, 38(5):685–696, 1985.

[114] B. Shi. On the Hyperparameters in Stochastic Gradient Descent with Momentum, 2022.

[115] B. Shi, S. Du, M. Jordan, and W. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195:79–148, 01 2021.

[116] B. Shi, S. S. Du, W. Su, and M. I. Jordan. Acceleration via Symplectic Discretization of High-Resolution Differential Equations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[117] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, Jul 2019.

[118] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015.

[119] B. Sinova, G. González-Rodríguez, and S. V. Aelst. M-estimators of location for functional data. *Bernoulli*, 24(3):2328–2357, 2018.

[120] J. Sirignano and K. Spiliopoulos. Stochastic gradient descent in continuous time. *SIAM Journal on Financial Mathematics*, 8(1):933–961, 2017.

[121] J. Sirignano and K. Spiliopoulos. Dgm: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018.

[122] S. L. Smith, B. Dherin, D. G. T. Barrett, and S. De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations, ICLR*, 2021.

[123] W. Su, S. Boyd, and E. Candès. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[124] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[125] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[126] T. Tao. *Nonlinear dispersive equations*, volume 106 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI, 2006.

[127] N. G. Trillos and D. Sanz-Alonso. The Bayesian Update: Variational Formulations and Gradient Flows. *Bayesian Analysis*, 15(1):29–56, 2020.

[128] Z. Wang and J. Sirignano. Global Convergence of the ODE Limit for Online Actor-Critic Algorithms in Reinforcement Learning, 2021.

[129] Z. Wang and J. Sirignano. A Forward Propagation Algorithm for Online Optimization of Nonlinear Stochastic Differential Equations, 2022.

[130] Z. Wang and J. Sirignano. Continuous-time stochastic gradient descent for optimizing over the stationary distribution of stochastic differential equations. *Mathematical Finance*, 34(2):348–424, 2024.

[131] S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part II: Continuous time analysis. *arXiv:2106.02588*, 2021.

[132] V. Zakharov, A. Odesski, M. Cisternino, and M. Onorato. Five-wave classical scattering matrix and integrable equations. *Theoretical and Mathematical Physics*, 180(1):759–764, 2014.

[133] V. Zakharov and E. Schulman. On additional motion invariants of classical Hamiltonian wave systems. *Physica D: Nonlinear Phenomena*, 29(3):283–320, 1988.

[134] V. E. Zakharov and E. I. Schulman. *Integrability of Nonlinear Systems and Perturbation Theory*, pages 185–250. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991.

[135] M. D. Zeiler. ADADELTA: An Adaptive Learning Rate Method, 2012.

[136] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[137] F. Zheng. Long-term regularity of the periodic Euler–Poisson system for electrons in 2D. *Communications in Mathematical Physics*, 366(3):1135 – 1172, 2019.

[138] F. Zheng. Long-term regularity of 3D gravity water waves. *Communications on Pure and Applied Mathematics*, 75(5):1074–1180, 2022.

[139] M. Zinkevich. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, page 928–935, 2003.

[140] D. Zou, J. Wu, V. Braverman, Q. Gu, D. Foster, and S. M. Kakade. The Benefits of Implicit Regularization from SGD in Least Squares Problems. In *Advances in Neural Information Processing Systems*, 2021.