11-1-2018

A Mathematical Framework on Machine Learning: Theory and Application

Bin Shi

Florida International University, bshi001@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

A MATHEMATICAL FRAMEWORK ON MACHINE LEARNING: THEORY AND APPLICATION

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Bin Shi

2018

To: John L. Volakis
Dean of College of Engineering and Computing

This dissertation, written by Bin Shi, and entitled A Mathematical Framework on Machine Learning: Theory and Application, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Leonardo Bobadilla

Zhenmin Chen

Xudong He

Jason Liu

Sundaraja S. Iyengar, Major Professor

Date of Defense: September 11, 2018

The dissertation of Bin Shi is approved.

John L. Volakis
Dean of College of Engineering and Computing

Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2018

DEDICATION

I dedicate this dissertation work to my beloved family, especially my parents. Without their patience, understanding, support or love, the completion of this work would not have been possible.

ACKNOWLEDGMENTS

It is the support from many people that brings me to the completion of my dissertation and conclusion of my Ph.D. study.

First, I would like to express my sincerest thanks and appreciation to my advisors Dr.Sundaraja S. Iyengar and Dr.Tao Li, for introducing me to the field of Data Science, which is the rising interdiscipline of machine learning, optimization and statistics. With Dr.Tao Li's success in the techniques of deep learning in practice, he was strongly aware of that the theoretical aspect is not mature by his perceptive insight. By Identifying with my strengths, he guided me into the field of theoretical development of machine learning.

Second, I would like to extend my gratitude to my major advisor, Dr. Sundaraja S. Iyengar, who has not only given huge support and encouragement for my research, but also supplied constructive suggestion in developing my Ph.D. career.

Third, my thanks to all my dissertation committee members: Dr.Xudong He, Dr.Jason Liu, Dr.Leonardo Bobadilla and Dr.Zhenmin Chen, for their helpful advice, insightful comments on my dissertation research and future research career plans.

Finally, I would like to express my utmost gratitude to my parents and family, whose endless love and understanding are with me in whatever I pursue. Without the unlimited support from them, I would never be able to survive the tough times in my life.

ABSTRACT OF THE DISSERTATION

A MATHEMATICAL FRAMEWORK ON MACHINE LEARNING: THEORY AND APPLICATION

by

Bin Shi

Florida International University, 2018

Miami, Florida

Professor Sundaraja S. Iyengar, Major Professor

The dissertation addresses the research topics of machine learning outlined below. We developed the theory about traditional first-order algorithms from convex optimization and provide new insights in nonconvex objective functions from machine learning. Based on the theory analysis, we designed and developed new algorithms to overcome the difficulty of nonconvex objective and to accelerate the speed to obtain the desired result. In this thesis, we answer the two questions: (1) How to design a step size for gradient descent with random initialization? (2) Can we accelerate the current convex optimization algorithms and improve them into nonconvex objective? For application, we apply the optimization algorithms in sparse subspace clustering. A new algorithm, CoCoSSC, is proposed to improve the current sample complexity under the condition of the existence of noise and missing entries.

Gradient-based optimization methods have been increasingly modeled and interpreted by ordinary differential equations (ODEs). Existing ODEs in the literature are, however, inadequate to distinguish between two fundamentally different methods, Nesterov's acceleration gradient method for strongly convex functions (NAG-SC) and Polyak's heavy-ball method. In this paper, we derive high-resolution ODEs as more accurate surrogates for the two methods in addition to Nesterov's acceleration gradient method for general convex functions (NAG-C), respectively. These novel

ODEs can be integrated into a general framework that allows for a fine-grained analysis of the discrete optimization algorithms through translating properties of the amenable ODEs into those of their discrete counterparts. As a first application of this framework, we identify the effect of a term referred to as gradient correction in NAG-SC but not in the heavy-ball method, shedding deep insight into why the former achieves acceleration while the latter does not. Moreover, in this high-resolution ODE framework, NAG-C is shown to boost the squared gradient norm minimization at the inverse cubic rate, which is the sharpest known rate concerning NAG-C itself. Finally, by modifying the high-resolution ODE of NAG-C, we obtain a family of new optimization methods that are shown to maintain the accelerated convergence rates as NAG-C for minimizing convex functions.

Key Words. Convex optimization, first-order method, Polyak's heavy ball method, Nesterov's accelerated gradient methods, ordinary differential equation, Lyapunov function, gradient minimization, dimensional analysis, phase space representation, numerical stability

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

With the explosive growth of data nowadays, a young and interdisciplinary field, **Data Science**, has emerged, which uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured. This data science field is becoming popular and needs to be developed urgently so that it can serve and guide for the industry of the society. Rigorously, applied **Data Science** is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Within the field of data analytics, **Machine Learning** is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. The name **Machine Learning** was coined in 1959 by Arthur Samuel, evolved from the study of pattern recognition and computational learning theory in artificial intelligence. **Computational Statistics**, which also focuses on prediction-making through the use of computers, is a closely related field and often overlaps with **Machine Learning**.

The name, **Computational Statistics**, tells us that it is composed of two indispensable parts, statistics inference models as well as the corresponding algorithms implemented in computers. Based on the different kinds of hypotheses, statistics inference can be divided into two schools, frequentist inference school and Bayesian inference school. Now, we briefly describe them. Let $\mathcal{H}$ be a hypothesis and $\mathcal{D}$ be data which may give evidence for $\mathcal{H}$. The probabilities about the event are defined as below

- The priori $P(\mathcal{H})$ is the probability that $\mathcal{H}$ is true before the data is considered.
- The posterior $P(\mathcal{H}|\mathcal{D})$ is the probability that $\mathcal{H}$ is true after the data $\mathcal{D}$ is considered.
- The likelihood $P(\mathcal{D}|\mathcal{H})$ is the evidence about $\mathcal{H}$ provided by the data $\mathcal{D}$.
- $P(\mathcal{D})$ is the total probability, shown as below

$$P(\mathcal{D}) = \sum_{\mathcal{H}} P(\mathcal{D}|\mathcal{H})\, P(\mathcal{H})$$

Connecting the probabilities above is the significant Bayes' formula in the theory of probability

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})\, P(\mathcal{H})}{P(\mathcal{D})} \sim P(\mathcal{D}|\mathcal{H})\, P(\mathcal{H})\,. \tag{1.1}$$

where $P(\mathcal{D})$ can be calculated automatically if we have known the likelihood $P(\mathcal{D}|\mathcal{H})$ and $P(\mathcal{H})$. If we presume that some hypothesis (parameter specifying the conditional distribution of the data) is true and that the observed data is sampled from that distribution, that is,

$$P(\mathcal{H}) = 1,$$

only using conditional distributions of data given specific hypotheses is the view of the frequentist school. However, if there is no presumption that some hypothesis (parameter specifying the conditional distribution of the data) is true, that is, there is a prior probability for the hypothesis $\mathcal{H}$,

$$\mathcal{H} \sim P(\mathcal{H}),$$

summing up the information from the prior and likelihood is the view from the Bayesian school. Apparently, the view from the frequentist school is a special case of the view from the Bayesian school, but the view from the Bayesian school is more comprehensive and requires more information.

Take the Gaussian distribution with known variance for the likelihood as an example. Without loss of generality, we assume the variance $\sigma^2 = 1$. In other words, the data point is viewed as a random variable $\mathbf{X}$ following the rule below,

$$\mathbf{X} \sim P\left(x|\mathcal{H}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

where the hypothesis is $\mathcal{H} = \{\mu|\mu \in (-\infty, \infty) \text{ is some fixed real number}\}$. Let the data set be $\mathcal{D} = \{x_i\}_{i=1}^n$. The frequentist school requires to compute maximum likelihood or maximum log-likelihood, that is

$$\begin{aligned}
\underset{\mu\in(-\infty,\infty)}{\text{argmax}}\ f\left(\mu\right) &= \underset{\mu\in(-\infty,\infty)}{\text{argmax}}\ \log P\left(\mathcal{D}|\mathcal{H}\right) \\
&= \underset{\mu\in(-\infty,\infty)}{\text{argmax}} \left(\log \prod_{i=1}^{n} P\left(x_i \in \mathcal{D}|\mathcal{H}\right) \right) \\
&= \underset{\mu\in(-\infty,\infty)}{\text{argmax}}\ \log \left[\left(\frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}} \right] \\
&= -\underset{\mu\in(-\infty,\infty)}{\text{argmin}} \left[\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 + n \log \sqrt{2\pi} \right],
\end{aligned} \tag{1.2}$$

which has been shown in the classical textbooks, such as [RS15]; whereas the Bayesian school requires to compute maximum posterior estimate or maximum log-posterior estimate, that is, we need to assume reasonable prior distribution

- If the prior distribution is a Gauss distribution $\mu \sim \mathcal{N}(0, \sigma_0^2)$, we have

$$\begin{aligned}
\underset{\mu\in(-\infty,\infty)}{\text{argmax}}\ f\left(\mu\right) &= \underset{\mu\in(-\infty,\infty)}{\text{argmax}}\ \log P\left(\mathcal{D}|\mathcal{H}\right) P\left(\mathcal{H}\right) \\
&= \underset{\mu\in(-\infty,\infty)}{\text{argmax}}\ \log \left(\prod_{i=1}^{n} \log P\left(x_i \in \mathcal{D}|\mathcal{H}\right) \right) P\left(\mathcal{H}\right) \\
&= \underset{\mu\in(-\infty,\infty)}{\text{argmax}}\ \log \left\{ \left[\left(\frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}} \right] \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_0} \right) e^{-\frac{\mu^2}{2\sigma_0^2}} \right\} \\
&= -\underset{\mu\in(-\infty,\infty)}{\text{argmin}} \left[\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{2\sigma_0^2} \cdot \mu^2 + n \log \sqrt{2\pi} + \log \sqrt{2\pi}\sigma_0 \right]
\end{aligned} \tag{1.3}$$

- If the prior distribution is Laplace distribution $\mu \sim \mathcal{L}(0, \sigma_0^2)$, we have

$$\begin{aligned}
\max_{\mu \in (-\infty,\infty)} f(\mu) &= \operatorname*{argmax}_{\mu \in (-\infty,\infty)} \log P(\mathcal{D}|\mathcal{H})\, P(\mathcal{H}) \\
&= \operatorname*{argmax}_{\mu \in (-\infty,\infty)} \log \left(\prod_{i=1}^{n} \log P(x_i \in \mathcal{D}|\mathcal{H}) \right) P(\mathcal{H}) \\
&= \operatorname*{argmax}_{\mu \in (-\infty,\infty)} \log \left\{ \left[\left(\frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}} \right] \cdot \left(\frac{1}{2\sigma_0^2} \right) e^{-\frac{|\mu|}{\sigma_0^2}} \right\} \\
&= - \operatorname*{argmin}_{\mu \in (-\infty,\infty)} \left[\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{\sigma_0^2} \cdot |\mu| + n \log \sqrt{2\pi} + \log 2\sigma_0^2 \right]
\end{aligned} \tag{1.4}$$

- If the prior distribution is the mixed distribution combined with Laplace distribution and Gaussian distribution $\mu \sim \mathcal{M}(0, \sigma_{0,1}^2, \sigma_{0,2}^2)$, we have

$$\begin{aligned}
\operatorname*{argmax}_{\mu \in (-\infty,\infty)} f(\mu) &= \operatorname*{argmax}_{\mu \in (-\infty,\infty)} \log P(\mathcal{D}|\mathcal{H})\, P(\mathcal{H}) \\
&= \operatorname*{argmax}_{\mu \in (-\infty,\infty)} \log \left(\prod_{i=1}^{n} \log P(x_i \in \mathcal{D}|\mathcal{H}) \right) P(\mathcal{H}) \\
&= \operatorname*{argmax}_{\mu \in (-\infty,\infty)} \log \left\{ \left[\left(\frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}} \right] \cdot C(\sigma_{0,1}, \sigma_{0,2})^{-1} e^{-\frac{|\mu|}{\sigma_{0,1}^2} - \frac{\mu^2}{2\sigma_{0,2}^2}} \right\} \\
&= - \operatorname*{argmin}_{\mu \in (-\infty,\infty)} \left[\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{\sigma_0^2} \cdot |\mu| + \frac{1}{2\sigma_{0,2}^2} \cdot \mu^2 \right. \\
&\qquad \left. + n \log \sqrt{2\pi} + \log C(\sigma_{0,1}, \sigma_{0,2}) \right]
\end{aligned} \tag{1.5}$$

where $C = 2\sqrt{2\pi}\sigma_{0,1}^2 \sigma_{0,2}$.

In summary, based on the description above, to solve this statistic problem can be transformed into an optimization problem.

1.2 Problem Statement

Based on the description on the statistics model in the previous section, we state the problems that we need to solve from two aspects. One is from the field of optimization, the other is from samples of probability distribution. Practically, from the view of efficient algorithms in computers, the representation of the first one is the expectation-maximization (**EM**) algorithm. The EM algorithm is used to find (local) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either missing values exist among the data, or the model can be formulated more simply by assuming the existence of further unobserved data points. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component to which each data point belongs.

Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values, the parameters and the latent variables, and simultaneously solving the resulting equations. In statistical models with latent variables, this is usually impossible. Instead, the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation.

The EM algorithm proceeds from the observation that there is a way to solve these two sets of equations numerically. One can simply pick arbitrary values for one of the two sets of unknowns, use them to estimate the second set, then use these new values to find a better estimate of the first set, and then keep alternating between the two until the resulting values both converge to fixed points. It's not obvious that this will work, but it can be proven that in this context it does, and that the derivative of the likelihood is (arbitrarily close to) zero at that point, which in turn means that the point is either a maximum or a saddle point. In general, multiple maxima may occur, with no guarantee that the global maximum will be found. Some likelihoods also have singularities in them, i.e., nonsensical maxima. For example, one of the solutions that may be found by EM in a mixture model involves setting one of the components to have zero variance and the mean parameter for the same component to be equal to one of the data points.

The second one is the Markov chain Monte Carlo (**MCMC**) method. Markov chain Monte Carlo methods are primarily used for calculating numerical approximations of multi-dimensional integrals, for example in Bayesian statistics, computational physics, computational biology and computational linguistics.

In Bayesian statistics, the recent development of Markov chain Monte Carlo methods has been a key step in making it possible to compute large hierarchical models that require integrations over hundreds or even thousands of unknown parameters.

In rare event sampling, they are also used for generating samples that gradually populate the rare failure region.

1.2.1 Optimization

Recall the process of finding the maximum probability, which is equivalent to the maximum log-likelihood or the maximum log-posterior estimate in essential. We describe them rigorously in statistics language as below.

- Finding the maximum likelihood (1.2) is equivalent to the expression below

$$\underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} f(\mu) = -\underset{\mu\in(-\infty,\infty)}{\operatorname{argmin}} \left[\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right], \tag{1.6}$$

  which is named **linear regression** in statistics.

- Finding the maximum posterior estimate (1.3) is equivalent to the expression below

$$\underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} f(\mu) = -\underset{\mu\in(-\infty,\infty)}{\operatorname{argmin}} \left[\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{2\sigma_0^2}\cdot\mu^2\right], \tag{1.7}$$

which is named **ridge regression** in statistics.

- Finding the maximum posterior estimate (1.3) is equivalent to the expression below
$$\underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} f(\mu) = - \underset{\mu\in(-\infty,\infty)}{\operatorname{argmin}} \left[\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{\sigma_0^2}\cdot|\mu| \right], \qquad (1.8)$$
which is named **lasso** in statistics.

- Finding the maximum posterior estimate (1.3) is equivalent to the expression below
$$\underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} f(\mu) = - \underset{\mu\in(-\infty,\infty)}{\operatorname{argmin}} \left[\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{\sigma_{0,1}^2}\cdot|\mu| + \frac{1}{2\sigma_{0,2}^2}\cdot\mu^2 \right], \quad (1.9)$$
which is named **elastic-net** in statistics.

Linear regression (1.6) is considered as one of the standard models in statistics, the variants (1.7), (1.8) and (1.9) of which are viewed as linear regression with regularizers. Every regularizer has its own advantage, the advantage of ridge regression (1.7) is stability, that of lasso (1.8) is sparsity, and that of elastic-net (1.9) owns sparsity and group-effect. Especially, due to the sparse property, the lasso (1.8) become one of the most significant models in statistics.

The linear regression and its variants above can be reduced to finding a minimizer of the convex objective function without constraint:
$$\min_{x\in\mathbb{R}} f(x),$$
of which the corresponding high-dimension expression highly concerned in practice is
$$\min_{x\in\mathbb{R}^n} f(x).$$

All of descriptions above are from the simple likelihood. In biology, the models above are suitable to study for a single species. Take the tigers in China for example.

Figure 1.1: Left: Siberian Tiger; Right: South China tiger.

There are two kinds of tigers in China, Siberian tiger and South China tiger (Figure 1.1). If we only consider one kind of tigers, Siberian tiger or South China tiger, then we can assume the likelihood is a single Gaussian; but if we consider the total tigers in China, both Siberian tiger and South China tiger, then the likelihood is a superposition of two single Gaussian. The simple sketch in $\mathbb{R}$ is shown in Figure 1.2. Comparing the left two and the right one in Figure 1.2, there exists three stationary points, two local maximal points and one local minimal point. In other words, the objective function is nonconvex. The classical convex optimization algorithms, based on the principle that the local minimal point is the global minimal point, are not suitable for the original convex case. Furthermore, if the dimension of the objective

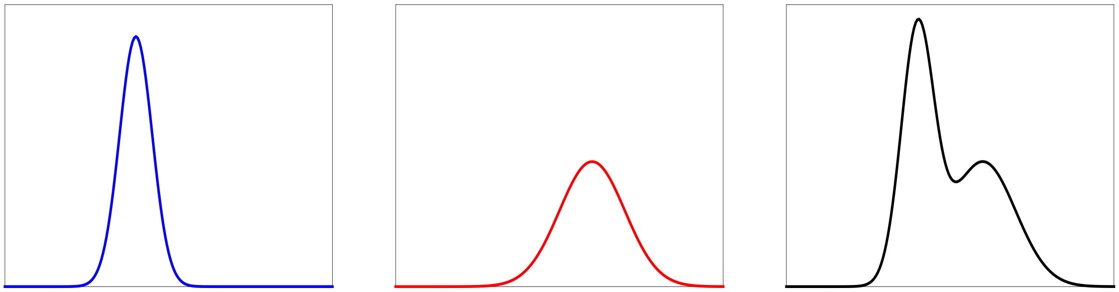


Figure 1.2: Left: Gaussian-1; Middle: Gaussian-2; Right: Mixed Gaussian: Gaussian-1+Gaussian-2.

function is greater than and equal 2, there exists another stationary point: saddle. We demonstrate the different stationary points in Figure 1.3.

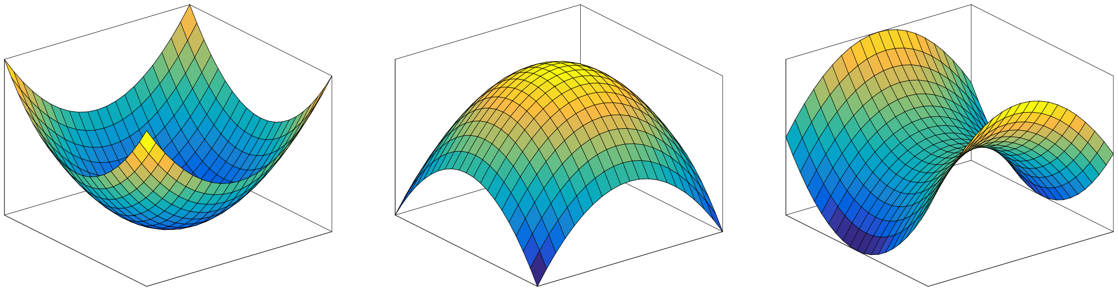


Figure 1.3: Left: Local Minimal Point; Middle: Local Maximal Point; Right: Saddle.

From the descriptions above, many statistics models are finally transformed to solve an optimization problem, not only simple convex optimization but also complex nonconvex optimization. What's more, the optimization algorithms are based on the information from the objective function. The classical oracle assumption for the smoothness is described in [Nes13] as below

- Zero-order oracle assumption: returns the value $f(x)$;
- First-order Oracle assumption: returns the value $f(x)$ and the gradient $\nabla f(x)$;
- Second-order oracle assumption: returns the value $f(x)$, the gradient $\nabla f(x)$ and the Hessian $\nabla^2 f(x)$.

To discriminate if an optimization algorithm is highly efficient in practice, based on the performance, the main characters are from oracle information and iteration complexity. Apparently, zero-order oracle algorithms are firstly considered. Currently, there are two main kinds of methods involved to implement: kernel-based bandit algorithms [BLE17] and algorithms of single-point gradient estimation [FKM05], [HL14]. Since the fewer oracle information leads to the higher iteration complexity, the zero-order oracle algorithms are not popular in practice.

Furthermore, developing zero-order oracle algorithms are still in the convex stage. Second-order oracle algorithms have been studied widespread for last four decades, which are essentially based on classical Newton iteration, such as modified Newton's method [MS79], modified Cholesky's method [GM74], Cubic-Regularization method [NP06] and Trust Region method [CRS14]. Currently, with the success of deep learning, some algorithms based on Hessian-prodcut in nonconvex objective have been proposed in [AAZB+17, CD16, CDHS16, LY17, RZS+17, RW17]. However, the difficulty of computing the Hessian information leads to infeasibility in current computers.

Now, we come to the first-order algorithms which have been widespread used. First-order algorithms only need to compute gradient which takes $O(d)$ time complexity, where the dimension d is large. Recall the statistics model (1.6), (1.7), (1.8) and (1.9), if we compute the full gradient $\nabla f(\mu)$, it leads to deterministic algorithms; if we compute one gradient $\nabla f_i(\mu)$, that is, $(x_i - \mu)$ for some $1 \leq i \leq n$, it leads to stochastic algorithms. In this thesis, we focus on deterministic algorithms.

1.2.1.1 Gradient Descent

Gradient descent (GD) and its variants provide the core optimization methodology in machine learning problems. Given a C^1 or C^2 function $f : \mathbb{R}^n \to \mathbb{R}$ with unconstrained variable $x \in \mathbb{R}^n$, GD uses the following update rule:

$$x_{k+1} = x_k - h_k \nabla f(x_k) \tag{1.10}$$

where h_k are step size, which may be either fixed or vary across iterations. When f is convex, $h_k < \frac{2}{L}$ is a necessary and sufficient condition to guarantee the (worst-case) convergence of GD, where L is the Lipschitz constant of the gradient of the function f. On the other hand, there is far less understanding of GD for non-convex problems. For general smooth non-convex problems, GD is only known to converge to a stationary point (i.e., a point with zero gradient) [Nes13].

Machine learning tasks often require finding a local minimizer instead of just a stationary point, which can also be a saddle point or a maximizer. In recent years, there has been an increasing focus on geometric conditions under which GD escapes saddle points and converges to a local minimizer. More specifically, if the objective function satisfies 1) all saddle point are strict and 2) all local minima are global minima then GD finds a global optimal solution. These two properties hold for a wide range of machine learning problems, such as matrix factorization [LWL+16], matrix completion [GLM16, GJZ17], matrix sensing [BNS16, PKCS17], tensor decomposition [GHJY15], dictionary learning [SQW17] and phase retrieval [SQW16].

Recent works showed when the objective function has the strict saddle property, then GD converges to a minimizer provided the initialization is randomized and the step sizes are fixed and smaller than $1/L$ [LSJR16, PP16]. While this was the first results establishing convergence of GD, there are still gaps toward fully understanding GD for strict saddle problems.

1.2.1.2 Accelerated Gradient Descent

Non-convex optimization is the dominating algorithmic technique behind many state-of-art results in machine learning, computer vision, natural language processing and reinforcement learning. Finding a global minimizer of a non-convex optimization problem is NP-hard. Instead, the local search method become increasingly important, which is based on the method from convex optimization problem. Formally, the problem of unconstrained optimization is stated in general terms as that of finding the minimum value that a function attains over Euclidean space, i.e.

$$\min_{x \in \mathbb{R}^n} f(x).$$

Numerous methods and algorithms have been proposed to solve the minimization problem, notably gradient methods, Newton's methods, trust-region method, ellipsoid method and interior-point method [Pol87b, Nes13, WN99, LY+84, BV04, B+15].

First-order optimization algorithms are the most popular algorithms to perform optimization and by far the most common way to optimize neural networks, since the second-order information obtained is supremely expensive. The simplest and earliest method for minimizing a convex function f is the gradient method, i.e.,

$$\begin{cases} x_{k+1} = x_k - h\nabla f(x_k) \\ \text{Any Initial Point}: \ x_0. \end{cases} \tag{1.11}$$

There are two significant improvements of the gradient method to speed up the convergence. One is the momentum method, named as Polyak heavy ball method, first proposed in [Pol64], i.e.,

$$\begin{cases} x_{k+1} = x_k - h\nabla f(x_k) + \gamma_k(x_k - x_{k-1}) \\ \text{Any Initial Point}: \ x_0. \end{cases} \tag{1.12}$$

Let κ be the condition number, which is the ratio of the smallest eigenvalue and the largest eigenvalue of Hessian at local minima. The momentum method speed up the local convergence rate from $1 - 2\kappa$ to $1 - 2\sqrt{\kappa}$. The other is the Notorious Nesterov's accelerated gradient method, first proposed in [Nes83] and an improved version [NN88, Nes13], i.e.

$$\begin{cases} y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} = x_k + \gamma_k(x_{k+1} - x_k) \\ \text{Any Initial Point}: \ x_0 = y_0 \end{cases} \tag{1.13}$$

where the parameter is set as

$$\gamma_k = \frac{\alpha_k(1 - \alpha_k)}{\alpha_k^2 + \alpha_{k+1}} \quad \text{and} \quad \alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + \alpha_{k+1}\kappa.$$

The scheme devised by Nesterov does not only own the property of the local convergence for strongly convex function, but also is the global convergence scheme, from $1-2\kappa$ to $1-\sqrt{\kappa}$ for strongly convex function and from $\mathcal{O}\left(\frac{1}{n}\right)$ to $\mathcal{O}\left(\frac{1}{n^2}\right)$ for non-strongly convex function.

Although there is the complex algebraic trick in Nesterov's accelerated gradient method, the three methods above can be considered from continuous-time limits [Pol64, SBC14, WWJ16, WRJ16] to obtain physical intuition. In other words, the three methods can be regarded as the discrete scheme for solving the ODE. The gradient method (1.11) is correspondent to

$$\begin{cases} \dot{x} = -\nabla f(x_k) \\ x(0) = x_0, \end{cases} \tag{1.14}$$

and the momentum method and Nesterov accelerated gradient method are correspondent to

$$\begin{cases} \ddot{x} + \gamma_t \dot{x} + \nabla f(x) = 0 \\ x(0) = x_0,\ \dot{x}(0) = 0, \end{cases} \tag{1.15}$$

the difference of which are the setting of the friction parameter γ_t. There are two significant intuitive physical meaning in the two ODEs (1.14) and (1.15). The ODE (1.14) is the governing equation for potential flow, a correspondent phenomena of waterfall from the height along the gradient direction. The infinitesimal generalization is correspondent to heat conduction in nature. Hence, the gradient method (1.11) is viewed as the implement in computer or optimization simulating the phenomena in the real nature. The ODE (1.15) is the governing equation for the heavy ball motion with friction. The infinitesimal generalization is correspondent to chord vibration in nature. Hence, the momentum method (1.12) and the Nesterov's accelerated gradient method (1.13) are viewed as the update version implement in computer or optimization by use of setting the friction force parameter γ_t.

Furthermore, we can view the three methods above as the thought for dissipating energy implemented in the computer. The unknown objective function in black box model can be viewed as the potential energy. Hence, the initial energy is from the potential function $f(x_0)$ at x_0 to the minimization value $f(x^\star)$ at $x^\star$. The total energy is combined with the kinetic energy and the potential energy. The key observation in this paper is that we find the kinetic energy, or the velocity, is observable and controllable variable in the optimization process. In other words, we can compare the velocities in every step to look for local minimum in the computational process or re-set them to zero to arrive to artificially dissipate energy.

Let us introduce firstly the governing motion equation in a conservation force field, that we use in this paper, for comparison as below,

$$\begin{cases} \ddot{x} = -\nabla f(x) \\ x(0) = x_0, \ \dot{x}(0) = 0. \end{cases} \tag{1.16}$$

The concept of phase space, developed in the late 19th century, usually consists of all possible values of position and momentum variables. The governing motion equation in a conservation force field (1.16) can be rewritten as

$$\begin{cases} \dot{x} = v \\ \dot{v} = -\nabla f(x) \\ x(0) = x_0, \ v(0) = 0. \end{cases} \tag{1.17}$$

1.2.1.3 Application to sparse subspace clustering

Subspace clustering is an important problem in machine learning, signal processing and computer vision research [Vid11]. Subspace clustering aims at grouping data points into disjoint *clusters* so that data points within each cluster lie near a *low-dimensional linear subspace*. It has found many successful applications in computer vision and machine learning, as many high dimensional data can be approximated by a union of low-dimensional subspaces. Example data include motion trajectories [CK98], face images [BJ03], network hop counts [EBN12], movie ratings [ZFIM12] and social graphs [JCSX11].

Mathematically, let $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ be an $n \times N$ data matrix, where n is the ambient dimension and N is the number of data points. We suppose there are L clusters $\mathcal{S}_1, \cdots, \mathcal{S}_L$, and each column (data point) of $\mathbf{X}$ belongs to exactly one cluster, and cluster $\mathcal{S}_\ell$ has $N_\ell \leq N$ points in $\mathbf{X}$. It is further assumed that data points within each subspace lie approximately on a low-dimensional linear subspace $\mathcal{U}_\ell \subseteq \mathbb{R}^n$ of dimension $d_\ell \ll n$. The question is to recover the clustering of all points in $\mathbf{X}$ without additional supervision.

In the case where data are noiseless (i.e., $\boldsymbol{x}_i \in \mathcal{U}_\ell$ if $\boldsymbol{x}_i$ belongs to cluster $\mathcal{S}_\ell$), the following *sparse subspace clustering* [EV13] approach can be used:

$$\text{SSC}: \qquad \boldsymbol{c}_i := \arg\min_{\boldsymbol{c}_i \in \mathbb{R}^{N-1}} \|\boldsymbol{c}_i\|_1 \quad s.t. \quad \boldsymbol{x}_i = \mathbf{X}_{-i}\boldsymbol{c}_i. \tag{1.18}$$

The vectors $\{\boldsymbol{c}_i\}_{i=1}^N$ are usually referred to as the *self-similarity matrix*, or simply *similarity matrix*, with the property that $|\boldsymbol{c}_{ij}|$ being large if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same cluster and vice versa. Afterwards, spectral clustering methods can be applied on $\{\boldsymbol{c}_i\}_{i=1}^N$ to produce the clustering [EV13].

While the noiseless subspace clustering model is ideal for simplified theoretical analysis, in practice data are almost always corrupted by additional noise. A general formulation for the noisy subspace clustering model is $\mathbf{X} = \mathbf{Y} + \mathbf{Z}$ where $\mathbf{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N)$ is an unknown noiseless data matrix (i.e., $\boldsymbol{y}_i \in \mathcal{U}_\ell$ if $\boldsymbol{y}_i$ belongs to $\mathcal{S}_\ell$) and $\mathbf{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N)$ is a noise matrix such that $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N$ are independent and $\mathbb{E}[\boldsymbol{z}_i|\mathbf{Y}] = \mathbf{0}$. Only the corrupted data matrix $\mathbf{X}$ is observed. Two important examples can be formulated under this framework:

- **Gaussian noise**: $\{\boldsymbol{z}_i\}$ are i.i.d. Gaussian random variables $\mathcal{N}(\mathbf{0}, \sigma^2/n \cdot \mathbf{I}_{n\times n})$.

- **Missing data**: let $R_{ij} \in \{0, 1\}$ be random variables indicating whether entry $\mathbf{Y}_{ij}$ is observed; that is, $\mathbf{X}_{ij} = R_{ij}\mathbf{Y}_{ij}/\rho$. The noise matrix $\mathbf{Z}$ can be taken as $\mathbf{Z}_{ij} = (1 - R_{ij}/\rho)\mathbf{Y}_{ij}$, where $\rho > 0$ is a parameter governing the probability of observing an entry; that is, $\Pr[R_{ij} = 1] = \rho$.

Many methods have been proposed to cluster noisy data with subspace clustering [SEC14, WX16, QX15, Sol14]. Existing work can be categorized primarily into two formulations: the Lasso SSC formulation

$$\textsc{Lasso SSC}: \qquad \boldsymbol{c}_i := \arg\min_{\boldsymbol{c}_i \in \mathbb{R}^{N-1}} \|\boldsymbol{c}_i\|_1 + \frac{\lambda}{2}\|\boldsymbol{x}_i - \mathbf{X}_{-i}\boldsymbol{c}_i\|_2^2, \tag{1.19}$$

which was analyzed in [SEC14, WX16, CJW17], and a de-biased Dantzig selector approach

$$\textsc{De-biased Dantzig Selector}: \qquad \boldsymbol{c}_i := \arg\min_{\boldsymbol{c}_i \in \mathbb{R}^{N-1}} \|\boldsymbol{c}_i\|_1 + \frac{\lambda}{2}\left\|\widetilde{\boldsymbol{\Sigma}}_{-i}\boldsymbol{c}_i - \widetilde{\boldsymbol{\gamma}}_i\right\|_\infty \tag{1.20}$$

which was proposed in [Sol14] and analyzed for an irrelevant feature setting in [QX15]. Here in Eq. (1.20) the terms $\widetilde{\boldsymbol{\Sigma}}_{-i}$ and $\widetilde{\boldsymbol{\gamma}}_i$ are de-biased second-order statistics, defined as $\widetilde{\boldsymbol{\Sigma}}_{-i} = \mathbf{X}_{-i}^\top\mathbf{X}_{-i} - \mathbf{D}$ and $\widetilde{\boldsymbol{\gamma}}_i = \boldsymbol{x}_i^\top\mathbf{X}_{-i}$, where $\mathbf{D} = \mathrm{diag}(\mathbb{E}[\boldsymbol{z}_1^\top\boldsymbol{z}_1], \cdots, \mathbb{E}[\boldsymbol{z}_N^\top\boldsymbol{z}_N])$ is a diagonal matrix that approximately de-biases the inner product and is assumed to be known. In particular, in the Gaussian noise model we have $\mathbf{D} = \sigma^2\mathbf{I}$ and in the missing data model we have $\mathbf{D} = (1-\rho)^2/\rho \cdot \mathrm{diag}(\|\boldsymbol{y}_1\|_2^2, \cdots, \|\boldsymbol{y}_N\|_2^2)$ which can be approximated by $\widehat{\mathbf{D}} = (1-\rho)^2\mathrm{diag}(\mathbf{X}^\top\mathbf{X})$ computable from corrupted data.

1.2.2 Online Algorithms: Sequential Updating

Based on the sampling methods, we here briefly introduce the principle behind the online time-varying algorithms. Let $t \in \{0, 1, 2, \ldots, N\}$ be a discrete finite time set. In every $t \in \{0, 1, 2, \ldots, N\}$, there are always new data being observed, noted as $\mathcal{D}_t$.

Recally the Bayesian formula (1.1), at time $t = 0$, with the prior $P(\mathcal{H})$ and likelihood $P(D_0|\mathcal{H})$, we have

$$P(\mathcal{H}|D_0) \sim P(D_0|\mathcal{H})P(\mathcal{H}).$$

At time $t = 1$, we take the posterior $P(\mathcal{H}|D_0)$ at time $t = 0$ as the prior at time $t = 1$ and the likelihood $P(D_1|\mathcal{H}, D_0)$, then the new posterior at $t = 1$ can be calculated as

$$P(\mathcal{H}|D_0, D_1) \sim P(D_1|\mathcal{H}, D_0)P(\mathcal{H}|D_0).$$

By analogy, at time $t = N$, we take the posterior $P(\mathcal{H}|D_0, \ldots, D_{N-1})$ at time $t = N-1$ as the prior at time $t = N$ and the likelihood $P(D_N|\mathcal{H}, D_0, \ldots, D_{N-1})$, then the new posterior at $t = 1$ can be calculated as

$$P(\mathcal{H}|D_0, \ldots, D_N) \sim P(D_N|\mathcal{H}, D_0, \ldots, D_{N-1})P(\mathcal{H}|D_0, \ldots, D_{N-1}).$$

With the description above, we actually implement $N+1$ times maximum posterior estimate, that is, maximum posterior estimate sequence as below,

$$P(\mathcal{H}|D_0), P(\mathcal{H}|D_0, D_1), \ldots, P(\mathcal{H}|D_0, D_1, D_N).$$

In other words, obtaining the distribution $P(\mathcal{H}|D_0, \ldots, D_k)$ $(k = 0, \ldots, N)$ is sequential updating. With the probability distribution $P(\mathcal{H}|D_0, \ldots, D_k)$ at time $t = k$, we can implement sampling process to generate data to observe the trend from time $t = 0$ to $t = N$ and to compare with the actual trend. Here, without any difficulty, we can find the core part of sequential updating is how to implement the likelihood sequence experimentally

$$P(D_0|\mathcal{H}), P(D_1|\mathcal{H}, D_0), \ldots, P(D_N|\mathcal{H}, D_0, \ldots, D_{N-1}).$$

A popular technique is named as particle learning, which assume actually the likelihood sequence following Gaussian random walk.

1.2.2.1 Application to multivariate time series

MTS analysis has been extensively employed across diverse application domains [BJRL15, Ham94], such as finance, social network, system management, weather forecast, etc. For example, it is well-known that there exists spatial and temporal correlations between air temperatures across certain regions [JHS+11, BBW+90]. Discovering and quantifying the hidden spatial-temporal dependences of the temperatures at different locations and time brings great benefits for weather forecast, especially in disaster prevention [LZZ+16].

Mining temporal dependency structure from MTS data is extensively studied across diverse domains. The Granger Causality framework is the most popular method. The intuition behind it is that if the time series A Granger causes the time series B, the future value prediction of B can be improved by giving the value of A. Regression model has evolved to be one of the principal approaches for Granger Causality. Specifically, to predict the future value of B, one regression model built only on the past values of B should be statistically significantly less accurate than the regression model inferred by giving the past values of both A and B. Regression model with L_1 regularizer [Tib96], named Lasso-Granger, is an advanced and effective approach for Granger causal relationship analysis. Lasso-Granger can effectively identify the sparse Granger Causality especially in high dimensions [BL13].

However, Lasso-Granger suffers some essential disadvantages. The number of non-zero coefficients chosen by Lasso is bounded by the number of training instances and also it tends to randomly select only one variable and ignore the others within a variable group which leads to instability. Moreover, all the work described above assumes a constant dependency structure among MTS. However, this assumption rarely holds in practice, since real-world problems often involve underlying processes that are dynamically evolving over time. Take a scenario in temperature forecast as an example.

Local temperature is usually impacted by its neighborhoods, but the dependency relationships dynamically change when monsoon comes from different directions. In order to capture the dynamic dependency typically happening in practice, a hidden Markov regression model [LKJ09] and a time-varying dynamic Bayesian network algorithm [ZWW$^+$16] have been proposed. However, both methods infer the underlying dependency structure based on the offline mode.

1.3 Contributions

In this section, we introduce our main contributions.

1.3.1 Gradient Descent

Question 1: Maximum Allowable Fixed Step Size. Recall that for convex optimization by gradient decent with fixed step-size rule $h_k \equiv h$, $h < 2/L$ is both a necessary and a sufficient condition for the convergence of GD. However, for non-convex optimization existing works all required the (fixed) step size to be smaller than $1/L$. Because larger step sizes lead to faster convergence, a nature question is to identify the maximum allowable step size such that GD escapes saddle points. The main technical difficulty to analyze larger step size is that the gradient map

$$g(x) = x - h\nabla f\left(x\right)$$

may *not* be a diffeomorphism when $h \geq 1/L$. Thus, techniques used in [LSJR16, PP16] are no longer sufficient.

Here, we take a finer look at the dynamics of GD. Our main observation is that the GD procedure escapes strict saddle points under much weaker conditions than g being a diffeomorphism everywhere. In particular, the probability of GD with random initialization converging to a strict saddle point is 0 provided that

$$g(x_k) = x_k - h_t \nabla f(x_k)$$

is a *local* diffeomorphism at every x_t. We further show that

$$\lambda\left(\{h \in [1/L, 2/L) : \exists t, g(x_k) \text{ is not a local diffeomorphism}\}\right) = 0$$

where $\lambda(\cdot)$ is the standard Lebesgue measure on $\mathbb{R}$, meaning that for almost every fixed step size choice in $[1/L, 2/L)$, $g(x_k)$ is a local diffeomorphism for every t. Therefore, if a step size h is chosen uniformly at random from $\left(\frac{2}{L} - \epsilon, \frac{2}{L}\right)$ for any $\epsilon > 0$, GD escapes all strict saddle points and converges to a local minimum. See Section 3.1 for the precise statement and Section 3.3 for the proof.

Question 2: Analysis of Adaptive Step Sizes. Another open question we consider in this paper is to analyze the convergence of GD for non-convex objectives when the step sizes $\{h_t\}$ vary as t evolves. In convex optimization, adaptive step size rules such as exact or backtracking line search [Nes13] are commonly used in practice to improve convergence, and convergence of GD is guaranteed provided that the adaptively tuned step sizes do not exceed twoce the inverse of local gradient Lipschitz constant. On the other hand, in non-convex optimization, whether gradient descent with varying step sizes can escape all strict saddle points is unknown.

Existing techniques [LSJR16, PP16, LPP$^+$17, OW17] cannot solve this question because they relied on the classical Stable Manifold Theorem [Shu13], which requires a fixed gradient map whereas when step sizes vary, the gradient maps also change across iterations. To deal with this issue, we adopt the powerful Hartman product map Theorem [Har71], which gives a finer characterization of local behavior of GD and allows the gradient map to change at every iteration. Based on Hartman product map Theorem, we show that as long as the step size at each iteration is proportional to the inverse of the *local* gradient Lipschitz constant, GD still escapes all strict saddle points. To our knowledge, this is the first result establishing convergence to local minima for non-convex gradient descent with varying step sizes.

1.3.2 Accelerated Gradient Descent

Here, we implement our discrete strategy into algorithms with the utility of the observability and controllability of the velocity, or the kinetic energy, as well as artificially dissipating energy for two directions as below,

- To look for local minima in non-convex function or global minima in convex function, the kinetic energy, or the norm of the velocity, is compared with that in the previous step, it will be re-set to zero until it becomes larger no longer.
- To look for global minima in non-convex function, an initial larger velocity $v(0) = v_0$ is implemented at the any initial position $x(0) = x_0$. A ball is implemented with (1.17), the local maximum of the kinetic energy is recorded to discern how many local minima exists along the trajectory. Then implementing the strategy above to find the minimum of all the local minima.

For implementing our thought in practice, we utilize the scheme in the numerical method for Hamiltonian system, the symplectic Euler method. We remark that a more accuracy version is the Störmer-Verlet method for practice.

1.3.3 The CoCoSSC Method

In this paper, we consider an alternative formulation CoCoSSC to solve the noisy subspace clustering problem, inspired by the CoCoLasso estimator for high-dimensional regression with measurement error [DZ17]. First, a pre-processing step is used that computes $\widetilde{\mathbf{\Sigma}} = \mathbf{X}^T\mathbf{X} - \widehat{\mathbf{D}}$ and then finds a matrix belonging to the following set:

$$S := \left\{\mathbf{A} \in \mathbb{R}^{N\times N} : \mathbf{A} \succeq \mathbf{0}\right\} \cap \left\{\mathbf{A} : \left|\mathbf{A}_{jk} - \widetilde{\mathbf{\Sigma}}_{jk}\right| \leq |\mathbf{\Delta}_{jk}|, \forall j, k \in [N]\right\}, \tag{1.21}$$

where $\mathbf{\Delta} \in \mathbb{R}^{N\times N}$ is an error tolerance matrix to be specified by the data analyst. For Gaussian random noise, all entries in $\mathbf{\Delta}$ can be set to a common parameter, while for the missing data model we recommend setting two different parameters for diagonal and off-diagonal elements in $\mathbf{\Delta}$, as estimation errors of these elements of $\mathbf{A}$ behave differently under the missing data model. We give theoretical guidelines on how to set the parameters in $\mathbf{\Delta}$ in our main theorems, while in practice we observe that setting the elements in $\mathbf{\Delta}$ to be sufficiently large would normally yield good results. Because S in Eq. (1.21) is a convex set, and we will later prove that $S \neq \emptyset$ with high probability, a matrix $\widetilde{\mathbf{\Sigma}}_+ \in S$ can be easily found by alternating projection from $\widetilde{\mathbf{\Sigma}}$.

For any $\widetilde{\mathbf{\Sigma}}_+ \in S$ and let $\widetilde{\mathbf{\Sigma}}_+ = \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$, where $\widetilde{\mathbf{X}} = (\widetilde{\boldsymbol{x}}_1, \cdots, \widetilde{\boldsymbol{x}}_N) \in \mathbb{R}^{N\times N}$. Such a decomposition exists because $\widetilde{\mathbf{\Sigma}}_+$ is positive semidefinite. The self-regression vector $\boldsymbol{c}_i$ is then obtained by solving the following (convex) optimization problem:

$$\textsc{CoCoSSC}: \qquad \boldsymbol{c}_i := \arg\min_{\boldsymbol{c}_i \in \mathbb{R}^{N-1}} \|\boldsymbol{c}_i\|_1 + \frac{\lambda}{2}\left\|\widetilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}_{-i}\boldsymbol{c}_i\right\|_2^2. \tag{1.22}$$

Eq. (1.22) is an ℓ_1-regularized least squares self regression problem, with the difference of using $\widetilde{\boldsymbol{x}}_i$ and $\widetilde{\mathbf{X}}_{-i}$ for self-regression instead of directly using the raw noise-corrupted observations $\boldsymbol{x}_i$ and $\mathbf{X}_{-i}$. This leads to improved sample complexity, as shown in Table 1.1 and our main theorems. On the other hand, CoCoSSC retains the nice structure of Lasso SSC, making it easier to optimize. We further discuss this aspect and other advantages of CoCoSSC in the next section.

1.3.3.1 Advantages of CoCoSSC

The CoCoSSC has the following advantages:

1. Eq. (1.22) is easier to optimize, especially compared to the de-biased Dantzig selector approach in Eq. (1.20), because it has a smoothly differentiable objective with an ℓ_1 regularization term. Many existing methods such as ADMM [BPC+11]

can be used to obtain fast convergence. We refer the readers to [WX16, Appendix B] for further details on efficient implementation of Eq. (1.22). The pre-processing step Eq. (1.21) can also be efficiently computed using alternating projection, as both sets in Eq. (1.21) are convex. On the other hand, the de-biased Dantzig selector formulation in Eq. (1.20) is usually solved using linear programming [CT05, CT07] and could be very slow as the number of variables is large. Indeed, our empirical results show that the debiased Dantzig selector is almost 5-10 times slower than both Lasso SSC and CoCoSSC.

2. Eq. (1.22) has improved or equal sample complexity in both the Gaussian noise model and the missing data model, compared to Lasso SSC and the de-biased Dantzig selector. This is because a "de-biasing" pre-processing step in Eq. (1.21) is used, and an error tolerance matrix $\mathbf{\Delta}$ with different diagonal and off-diagonal elements is considered to reflect the heterogeneous estimation error in $\mathbf{A}$. Table 1.1 gives an overview of our results and compare them with existing results.

Table 1.1: Summary of success conditions with normalized signals $\|y_i\|_2 = 1$. Polynomial dependency on d, $\overline{C}$, $\underline{C}$ and $\log N$ are omitted. In the last line χ is a subspace affinity quantity introduced in Definition 6.1.3 for the non-uniform semi-random model. χ is always upper bounded by $\sqrt{d}$.

	Gaussian model	Missing data (MD)	MD (random subspaces)
Lasso SSC [SEC14]	$\sigma = O(1)$	-	-
Lasso SSC [WX16]	$\sigma = O(n^{1/6})$	$\rho = \Omega(n^{-1/4})$	$\rho = \Omega(n^{-1/4})$
Lasso SSC [CJW17]	-	$\rho = \Omega(1)$	$\rho = \Omega(1)$
PZF-SSC[TV18]	-	$\rho = \Omega(1)$	$\rho = \Omega(1)$
Debiased Dantzig [QX15]	$\sigma = O(n^{1/4})$	$\rho = \Omega(n^{-1/3})$	$\rho = \Omega(n^{-1/3})$
CoCoSSC (this paper)	$\sigma = O(n^{1/4})$	$\rho = \Omega(\chi^{2/3}n^{-1/3} + n^{-2/5})^\dagger$	$\rho = \Omega(n^{-2/5})^\dagger$

† If $\|y_i\|_2$ is exactly known, the success condition can be improved to $\rho = \Omega(n^{-1/2})$. See Remark 6.1.3 for details.

1.3.4 Online Time-Varying Elastic-Net Algorithm

To overcome the deficiency of Lasso-Granger and capture the dynamical change of causal relationships among MTS, in this paper, we investigate the Granger Causality framework with Elasitc-Net [ZH05], which imposes a mixed L_1 and L_2 regularization penalty on the linear regression. The Elastic-Net cannot only obtain strongly stable coefficients [SHB16], but also capture grouped effects of variables [SHB16, ZH05]. Furthermore, our approach explicitly models the dynamical change behaviors of the dependency as a set of random walk particles, and utilizes particle learning [CJLP10, ZWW$^+$16] to provide a fully adaptive inference strategy which allows our model to effectively capture the varying dependency and learns the latent parameters simultaneously. Empirical studies on both synthetic and real dataset demonstrate the effectiveness of our proposed approach.

1.4 Organization

In Chapter **??**, we introduce preliminaries including notations, definitions as well as some related works. In Chapter 3, the maximum allowable step size and varying step size rules for gradient descent are shown. In Chapter 4, we show the conservation law algorithms based on accelerated gradient descent for nonconvex optimization. In Chapter 6, we introduce improved algorithm, CoCoSSC, and analyze sample complexity in sparse subspace Clustering with noisy and missing Entries. Finally, we show the online time-varying elastic-net algorithm to practically capture the dynamic group effect for MTS in Chapter 7.

CHAPTER 2

PRELIMINARIES AND NOTATIONS

2.1 Preliminaries and Notations

We define necessary notations and review important definitions that will be used later in our analysis. Let $C^2(\mathbb{R}^n)$ be the vector space of real-valued twice-continuously differentiable functions. Let ∇ be the gradient operator and ∇^2 be the Hessian operator. Let $\|\cdot\|_2$ be the Euclidean norm in $\mathbb{R}^n$. Let μ be the Lebesgue measure in $\mathbb{R}^n$.

Definition 2.1.1 (Global Gradient Lipschitz Continuity Condition)**.** $f \in C^2(\mathbb{R}^n)$ satisfies the global gradient Lipschitz continuity condition if there exists a constant $L > 0$ such that

$$\|\nabla f(x_1) - \nabla f(x_2)\|_2 \leq L \|x_1 - x_2\|_2 \qquad \forall x_1, x_2 \in \mathbb{R}^n. \tag{2.1}$$

Definition 2.1.2 (Global Hessian Lipschitz Continuity Condition)**.** $f \in C^2(\mathbb{R}^n)$ satisfies the global Hessian Lipschitz continuity condition if there exists a constant $K > 0$ such that

$$\left\|\nabla^2 f(x_1) - \nabla^2 f(x_2)\right\|_2 \leq K \|x_1 - x_2\|_2 \qquad \forall x_1, x_2 \in \mathbb{R}^n. \tag{2.2}$$

Intuitively, a twice-continuously differentiable function $f \in C^2(\mathbb{R}^n)$ satisfies the global gradient and Hessian Lipschitz continuity condition if its gradients and Hessians do not change dramatically for any two points in $\mathbb{R}^n$. However, the global Lipschitz constant L for many objective functions that arise in machine learning applications (e.g., $f(x) = x^4$) may be large or even non-existent. To handle such cases, one can use a finer definition of gradient continuity that characterizes the *local* behavior of gradients, especially for non-convex functions. This definition is adopted in many subjects of mathematics, such as in dynamical systems research.

Let $\delta > 0$ be some fixed constant. For every $x_0 \in \mathbb{R}^n$, its δ-closed neighborhood is defined as

$$V(x_0, \delta) = \{x \in \mathbb{R}^n | \, \|x - x_0\|_2 < \delta\} \, . \tag{2.3}$$

Definition 2.1.3 (Local Gradient Lipschitz Continuity Condition)**.** $f \in C^2(\mathbb{R}^n)$ satisfies the local gradient Lipschitz continuity condition at $x_0 \in \mathbb{R}^n$ with radius $\delta > 0$ if there exists a constant $L_{(x_0,\delta)} > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L_{(x_0,\delta)} \|x - y\|_2 \qquad \forall x, y \in V(x_0, \delta). \tag{2.4}$$

We next review the concepts of *stationary point*, *local minimizer* and *strict saddle point*, which are important in (non-convex) optimization.

Definition 2.1.4 (Stationary Point)**.** $x^* \in \mathbb{R}^n$ is a *stationary point* of $f \in C^2(\mathbb{R}^n)$ if $\nabla f\left(x^*\right) = 0$.

Definition 2.1.5 (Local Minimizer)**.** $x^* \in \mathbb{R}^n$ is a local minimum of f if there is a neighborhood U around x^* such that for all $x \in U$, $f(x^*) < f(x)$.

A stationary point can be a local minimizer, a saddle point or a maximizer. It is an standard fact that if a stationary point $x^\star \in \mathbb{R}^n$ is a local minimizer of $f \in C^2(\mathbb{R}^n)$, then $\nabla^2 f(x^\star)$ is positive semidefinite; on the other hand, if $x^* \in \mathbb{R}^n$ is a stationary point of $f \in C^2(\mathbb{R}^n)$ and $\nabla^2 f(x^\star)$ is positive definite, then x^* is also a local minimizer of f. It should also be noted that the stationary point $x^\star$ in the second case is isolated.

The following definition concerns "strict" saddle points, which was also analyzed in [GHJY15].

Definition 2.1.6 (Strict Saddle Points)**.** $x^* \in \mathbb{R}^n$ is a *strict saddle*[1] of $f \in C^2(\mathbb{R}^n)$ if x^* is a stationary point of f and furthermore $\lambda_{\min}\left(\nabla^2 f\left(x^*\right)\right) < 0$.

[1] For the purposes of this paper, strict saddle points include local maximizers.

We denote the set of all strict saddle points by $\mathcal{X}$. By definition, a strict saddle point must have an escaping direction so that the eigenvalue of the Hessian along that direction is trictly negative. For many non-convex problems studied in machine learning, all saddle points are strict.

We next review additional concepts in multivariate analysis and differential geometry/topology that will be used in our analysis.

Definition 2.1.7 (Gradient Map and Its Jacobian)**.** For any $f \in C^2(\mathbb{R}^n)$, the gradient map $g : \mathbb{R}^n \to \mathbb{R}^n$ with step size h is defined as

$$g(x) = x - h\nabla f(x). \tag{2.5}$$

The Jacobian $Dg : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ of the gradient map g is defined as

$$Dg(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x) & \cdots & \frac{\partial g_1}{\partial x_n}(x) \\ \cdots & \cdots & \cdots \\ \frac{\partial g_n}{\partial x_1}(x) & \cdots & \frac{\partial g_n}{\partial x_n}(x) \end{pmatrix}, \tag{2.6}$$

or equivalently, $Dg = I - h\nabla^2 f$.

We write $a_n \lesssim b_n$ if there exists an absolute constant $C > 0$ such that, for sufficiently large n, $|a_n| \le C|b_n|$. Similarly, $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ are true. We write $a_n \ll b_n$ if for a sufficiently small constant $c > 0$ and sufficiently large n, $|a_n| \le c|b_n|$. For any integer M, $[M]$ denotes the finite set $\{1, 2, \cdots, M\}$.

Definition 2.1.8 (Local Diffeomorphism)**.** Let M and N be two differentiable manifolds. A map $f : M \to N$ is a *local diffeomorphism* if for each point x in M, there exists an open set U containing x such that $f(U)$ is open in N and $f|_U : U \to f(U)$ is a diffeomorphism.

Definition 2.1.9 (Compact Set)**.** $S \subseteq \mathbb{R}^n$ is *compact* if every open cover of S has a finite sub-cover.

Definition 2.1.10 (Sublevel Set)**.** The α-sublevel set of $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$C_\alpha = \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\} .$$

2.2 Related Work

Over the past few years, there have been increasing interest in understanding the geometry of non-convex programs that naturally arise from machine learning problems. It is particularly interesting to study additional properties of the considered non-convex objective such that popular optimization methods (such as gradient descent) escape saddle points and converge to a local minimum. The strict saddle property (Definition 2.1.6) is one such property. which was also shown to hold in a broad range of applications.

Many existing works leveraged Hessian information in order to circumvent saddle points This includes a modified Newton's method [MS79], the modified Cholesky method [GM74], the cubic-regularization method [NP06] and trust region methods [CRS14]. The major drawback of such second-order methods is the requirement of access to the full Hessian, which could be computationally expensive. as the per-iteration computational complexity scales quadratically or even cubically in the problem dimension, unsuitable for optimization of high-dimensional functions. Some recent works [CDHS16, AAB+17, CD16] showed that the requirement of full Hessian can be relaxed to Hessian-vector products, which can be computed efficiently in certain machine learning applications. Several works [LY17, RZS+17, RW17] also presented algorithms that combine first-order methods with faster eigenvector algorithms to obtain lower per-iteration complexity.

Another line of works focus on noise-injected gradient methods whose per-iteration computational complexity scale linearly in the problem dimension. Earlier work have shown that first-order method with unbiased noise with sufficiently large variance can escape strict saddle points [Pem90]. [GHJY15] gave quantitative rates on the convergence. Recently, more refined algorithms and analyses [JGN+17, JNJ17] have been proposed to improve the convergence rate of such algorithms. Nevertheless, gradient methods with *deliberately injected* noise are almost never used in practical applications, limiting the applicability of the above-mentioned analysis.

Empirically, [SQW16] observed that gradient descent with 100 random initializations for the phase retrieval problem always converges to a local minimizer. Theoretically, the most important existing result is due to [LSJR16], who showed that gradient descent with fixed step size and any reasonable random initialization always escapes isolated strict saddle points. [PP16] later relaxed the requirement that strict saddle points are isolated. [OW17] extended the analysis to accelerated gradient descent and [LPP+17] generalized the result to a broader range of first-order methods, including proximal gradient descent and coordinate descent. However these works all require the step size to be significantly smaller than the inverse of Lipschitz constant of gradients, which has factor of 2 gap from results in the convex setting and do not allow the step size to vary across iterations. Our paper resolve both two problems.

The history of gradient method for convex optimization can be back to the time of Euler and Lagrange. However, since it is relatively cheaper to only calculation for first-order information, this simplest and earliest method is still active in machine learning and nonconvex optimization, such as the recent work [GHJY15, AG16, LSJR16, HMR16]. The natural speedup algorithms are the momentum method first proposed in [Pol64] and Nesterov accelerated gradient method first proposed in [Nes83] and an improved version [NN88]. A acceleration algorithm similar as Nesterov accelerated gradient method, named as FISTA, is designed to solve composition problems [BT09]. A related comprehensive work is proposed in [B+15].

The original momentum method, named as Polyak heavy ball method, is from the view of ODE in [Pol64], which contains extremely rich physical intuitive ideas and mathematical theory. An extremely important work in application on machine learning is the backpropagation learning with momentum [RHW+88]. Based on the thought of ODE, a lot of understanding and application on the momentum method and Nesterov accelerated gradient methods have been proposed. In [SMDH13], a well-designed random initialization with momentum parameter algorithm is proposed to train both DNNs and RNNs. A seminal deep insight from ODE to understand the intuition behind Nesterov scheme is proposed in [SBC14]. The understanding for momentum method based on the variation perspective is proposed on [WWJ16], and the understanding from Lyaponuv analysis is proposed in [WRJ16]. From the stability theorem of ODE, the gradient method always converges to local minima in the sense of almost everywhere is proposed in [LSJR16]. Analyzing and designing iterative optimization algorithms built on integral quadratic constraints from robust control theory is proposed in [LRP16].

Actually the "high momentum" phenomenon has been firstly observed in [OC15] for a restarting adaptive accelerating algorithm, and also the restarting scheme is proposed by [SBC14]. However, both works above utilize restarting scheme for an auxiliary tool to accelerate the algorithm based on friction. With the concept of phase space in mechanics, we observe that the kinetic energy, or velocity, is controllable and utilizable parameter to find the local minima. Without friction term, we can still find the local minima only by the velocity parameter. Based on this view, the algorithm is proposed very easy to practice. Meanwhile, the thought can be generalized to nonconvex optimization to detect local minima along the trajectory of the particle.

Sparse subspace clustering was proposed by [EV13] as an effective method for subspace clustering. [SC12] initiated the study of theoretical properties of sparse subspace clustering, which was later extended to noisy data [SEC14, WX16], dimensionality-reduced data [WWS15a, HTB17, TG17] and data consisting of sensitive private information [WWS15b]. [YRV15] considered some heuristics for subspace clustering with missing entries, and [TV18] considered a PZF-SSC approach and proved success conditions with $\rho = \Omega(1)$. [PCS14, HB15, LLY+13, TV17] proposed alternative approaches for subspace clustering. Some earlier references include k-plane [BM00], q-flat [Tse00], ALC [MDHW07], LSA [YP06] and GPCA [VMS05].

It is an important task to reveal the casual dependencies between historical and current observations in MTS analysis. Bayesian Network [JYG+03, Mur02] and Granger Causality [ALA07, ZF09] are two main frameworks for inference of temporal dependency. Comparing with Bayesian Network, Granger Causality is more straightforward, robust and extendable [ZF09].

Originally, Granger Causality is designed for a pair of time series. The appearance of pioneering work of combining the notion of Granger Causality with graphical model [Eic06] leads to the emergence of causal relationship analysis among MTS data. Two typical techniques, statistical significance test and Lasso-Granger [ALA07], are developed to inference the Granger Causality among MTS. Lasso-Granger gains more popularity due to its robust performance even in high dimensions [BL12]. However, Lasso-Granger suffers from instability and failure of group variable selection because of the high sensitivity of L_1 norm. To address this challenging, our method adopts Elastic-Net regularizer [ZH05] which is stable since it encourages a group variable selection (group effect) where strongly correlated predictors tend to be zero or non-zero simultaneously.

Particle learning [CJLP10] is a powerful tool to provide an online inference strategy for Bayesian models. It belongs to the Sequential Monte Carlo (SMC) methods consisting of a set of Monte Carlo methodologies to solve the filtering problem [DGA00].

Particle learning provides state filtering, sequential parameter learning and smoothing in a general class of state space models [CJLP10]. The central idea behind particle learning is the creation of a particle algorithm that directly samples from the particle approximation to the joint posterior distribution of states and conditional sufficient statistics for fixed parameters in a fully-adapted resample-propagate framework.

CHAPTER 3

GRADIENT DESCENT CONVERGES TO MINIMIZERS: OPTIMAL AND ADAPTIVE STEP SIZE RULES

In this chapter, we first introduce our main result formally with Theorems for maximum allowable step size in Section 3.1 and adaptive step size rules in Section 3.2. The both full proofs are shown in Section 3.3 and Section 3.4, respectively. Finally, we conclude the content of this chapter and introduce some future directions.

3.1 Maximum Allowable Step Size

We first consider gradient descent with a fixed step size. The following theorem provides a sufficient condition for escaping all strict saddle points.

Theorem 3.1.1. Suppose $f \in C^2(\mathbb{R}^n)$ satisfies the global gradient Lipschitz condition (Definition 2.1.1) with constant $L > 0$. Then there exists a zero-measure set $U \subset \left[\frac{1}{L}, \frac{2}{L}\right)$ such that if $h \in \left(0, \frac{2}{L}\right) \setminus U$ and $x_0 \in \mathbb{R}^n$ is randomly initialized with respect to an absolute continuous measure over $\mathbb{R}^n$, then

$$\mathbf{Pr}\left(\lim_k x_k \in \mathcal{X}\right) = 0,$$

where $\mathcal{X}$ denotes the set of all strict saddle points of f.

The complete proof of Theorem 3.1.1 is given in Sec. 3.3. Here we give a high-level sketch of our proof. Similar to [LSJR16], our proof relies on the seminal stable manifold theorem [Shu13]. For a fixed saddle point x^*, the stable manifold theorem asserts that locally, all points that converge to x^* lie in a manifold $W^{cs}_{loc}(x^*)$. Further, $W^{cs}_{loc}(x^*)$ has dimension at most $n-1$, thus $\mu\left(W^{cs}_{loc}(x^*)\right) = 0$. By Lindelöf's Lemma (Lemma 3.5.2), we can show that the union of these manifolds, $W^{cs}_{loc} = \bigcup_{x^* \in \mathcal{X}} W^{cs}_{loc}(x^*)$, also has Lebesgue measure 0. Next, we analyze what initialization points converge to W^{cs}_{loc}. Using the notion of the inverse gradient map, we can show the initialization points that converge W^{cs}_{loc} belongs to the set

$$\bigcup_{i=0}^{\infty} g^{-i}(W^{cs}_{loc}).$$

Thus, we only need to upper bound the Lebesgue measure of this set. If g is a local diffeomorphism, then by Lemma 3.3.2, we have $\mu\left(\bigcup_{i=0}^{\infty} g^{-i}(W^{cs}_{loc})\right) \leq \sum_{i=0}^{\infty} \mu\left(g^{-i}\left(W^{cs}_{loc}\right)\right) = 0$. Therefore, we only need to show g is a local diffeomorphism. Existing works require $\eta \leq 1/L$ to ensure g is a *global* diffeomorphism whereas a *local* diffeomorphism is already sufficient. Our main observation is that for h in $(1/L, 2/L)$, there is only a zero-measure set U such that g with respect to $h \in U$ is *not* a local diffeomorphism at some x_t. In other words, for almost every step size $h \in (1/L, 2/L)$, g is a local diffeomorphism at x_t for every t.

Theorem 3.1.1 shows that the step sizes in $[1/L, 2/L)$ that potentially leads to GD convergence towards a strict saddle point have measure zero. Comparing to recent results on gradient descent by [LSJR16, LPP+17, PP16], our theorem allows a maximum (fixed) step size of $2/L$ instead of $1/L$.

3.1.1 Consequences of Theorem 3.1.1

A direct corollary of Theorem 3.1.1 is that GD (with fixed step sizes $< 2/L$) can only converge to minimizers when the limit $\lim_k x_k$ exists.

Corollary 3.1.2 (GD Converges to Minimizers)**.** Under the conditions in Theorem 3.1.1 and the additional assumption that all saddle points of f are strict, if $\lim_k x_k$ exists then with probability 1 $\lim_k x_k$ is a local minimizer of f.

We now discuss when $\lim_k x_k$ exists. The following lemma gives a sufficient condition on its existence.

Lemma 3.1.3. Suppose $f \in C^2(\mathbb{R}^n)$ has global gradient Lipschitz constant L and owns compact sublevel sets. Further assume f only contains isolated stationary points. If $0 < h < 2/L$, $\lim_k x_k$ converges to a stationary point of f for any initialization x_0.

Theorem 3.1.1 and Lemma 3.1.3 together imply Corollary 3.1.2, which asserts that if the objective function has compact sub-level sets and the fixed step size h is smaller than $2/L$, GD converges to a minimizer. This result generalizes [LSJR16, PP16] where the fixed step sizes of GD cannot exceed $1/L$.

3.1.2 Optimality of Theorem 3.1.1

A natural question is whether the condition $h < 2/L$ in Theorem 3.1.1 can be further improved. The following proposition gives a negative answer, showing that GD with fixed step sizes $h \geq 2/L$ *diverges* on worst-case objective function f with probability 1. This shows that $h < 2/L$ is the optimal fixed step size rule one can hope for with which GD converges to a local minimum almost surely.

Proposition 3.1.1. There exists $f \in C^2(\mathbb{R}^n)$ with global gradient Lipschitz constant $L > 0$, compact sublevel sets and only isolated stationary points such that if $h \geq 2/L$ and x_0 is randomly initialized with respect to an absolutely continuous density on $\mathbb{R}^n$, then $\lim_k x_k$ does not exist with probability 1.

The proof of the proposition is simple by considering a quadratic function $f \in C^2(\mathbb{R}^n)$ that serves as a counter-example of GD with fixed step sizes larger than or equal to $h/2$. A complete proof of Proposition 3.1.1 is given in the appendix.

3.2 Adaptive Step Size Rules

In many machine learning applications, the global gradient Lipschitz constant L of the objective function f may be very large, but at most points the *local* gradient Lipschitz constant could be much smaller. It is thus desirable to consider varying step size rules that select step sizes h_t *adaptively* corresponding to the local gradient Lipschitz constant of f at x_t. When the objective function f is convex, the convergence of gradient descent with varying step sizes is well-understood [Nes13]. However, when f is non-convex, whether GD with varying step sizes can escape strict saddle points is still unknown. Existing works [LSJR16, LPP$^+$17, PP16] all require the step sizes to be fixed. Our following result closes this gap, showing that GD escapes strict saddle points if the step sizes chosen at each point x_t is proportional to the local gradient Lipschitz constant $L_{x_t,\delta}$.

Theorem 3.2.1. Suppose $f \in C^2(\mathbb{R}^n)$ satisfies the global Hessian Lipschitz continuity condition (Definition 2.1.2) with parameter K and for every $x^* \in \mathcal{X}$, $\nabla^2 f(x^*)$ is non-singular. Fix $\epsilon_0 \in (0,1)$ and define $r = \max_{x^* \in \mathcal{X}} K^{-1} \epsilon_0 \|\nabla^2 f(x^*)\|_2$. Then there exists $U \subset \mathbb{R}^+$ with $\mu(U) = 0$ such that if the step size at the tth iteration satisfies $h_t \in \left[\frac{\epsilon_0}{L_{x_t,r}}, \frac{2-\epsilon_0}{L_{x_t,r}}\right] \setminus U$ for all $t = 0, 1, \ldots$ and x_0 is randomly initialized with respect to an absolutely continuous density on $\mathbb{R}^n$, then

$$\mathbf{Pr}\left(\lim_t x_t \in \mathcal{X}\right) = 0.$$

Theorem 3.2.1 shows that even though the step sizes vary across iterations, GD still escapes all strict saddle points provided that all step sizes are proportional to their *local* smoothness. To our knowledge, this is the first result showing GD with varying step size escapes all strict saddle points. Theorem 3.2.1 requires $h_t \in \left[\frac{\epsilon_0}{L_{x_t,\delta}}, \frac{2-\epsilon_0}{L_{x_t,\delta}}\right]$, which are the desired local step size.

The proof of Theorem 3.2.1 follows a similar path as that of Theorem 3.1.1. We first locally characterize Lebesgue measure of the set of points that converge to saddle points and then use Lemma 3.3.2 to relates this set to the initialization. The main technical difficulty is the inapplicability of the stable manifold theorem in this setting, as the gradient maps g are no longer *fixed* and change across iterations. Instead of using the stable manifold theorem, we adopt the more general Hartman's product map theorem [Har82] which gives a finer characterization of the local behavior of a series of gradient maps around a saddle point.

Different from Theorem 3.1.1, Theorem 3.2.1 has two additional assumptions. First, we require that the Hessian matrices $\nabla^2 f(x^*)$ at each saddle point x^* is non-singular (i.e., no zero eigenvalues). This is a technical regularity condition for using Hartman's product map theorem. To remove this assumption, we need to generalize Hartman's product map theorem which is a challenging problem in dynamical systems. Second, we require that Hessian matrices $\nabla^2 f(x)$ satisfies a global Lipschitz continuity condition (Definition 2.1.2). This is because the Hartman's product map theorem requires the step size to be proportional to the gradient Lipschitz constant in a neighborhood of each saddle point and the radius of the neighborhood needs to be carefully quantified. Under the Hessian Lipschitz continuity assumption, we can give an upper bound on this radius which is sufficient for applying Hartman's product map theorem. It is possible to give finer upper bounds on this radius based on other quantitative continuity assumptions on the Hessian. The complete proof of Theorem 3.2.1 is given in Section 3.4.

3.3 Proof of Theorem 3.1.1

To prove Theorem 3.1.1, similar to [LSJR16], we rely on the following seminal stable manifold theorem from dynamical systems research.

Theorem 3.3.1 (**Theorem III.7, pp.65**, [Shu13])**.** Let 0 be a fixed point for a C^r local differomorphism $f : U \to \mathbb{R}^n$, where U is a neighborhood of zero in $\mathbb{R}^n$ and $1 \leq r < \infty$. Let $E^s \bigoplus E^c \bigoplus E^u$ be the invariant splitting of $\mathbb{R}^n$ into the generalized eigenspaces of $Df(0)$ corresponding to eigenvalues of absolute value less than one, equal to one, and greater than one. To the $Df(0)$ invariant subspaces $E^s \bigoplus E^c$, E^c there is associated a local f invariant C^r embedded disc W^{cs}_{loc} tangent to the linear subspace at 0 and a ball B around zero in an adapted norm such that

$$f(W^{cs}_{loc}) \cap B \subset W^{cs}_{loc}.$$

In addition, for any x satisfying $f^n(x) \in B$ for all $n \geq 0$, [1] then $x \in W^{cs}_{loc}$.

For each saddle point x^*, Theorem 3.3.1 implies the existence of a ball B_{x^*} centered at x^* and an invariant manifold $W^{cs}_{loc}(x^*)$ whose dimension is at most $n-1$. Let $B = \bigcup_{x^* \in \mathcal{X}} B_{x^*}$. With Lindelöf's Lemma (Lemma 3.5.2), there exists a countable $\mathcal{X}' \subset \mathcal{X}$ such that

$$B = \bigcup_{x^\star \in \mathcal{X}'} B_{x^*}.$$

Recall the dimension of $W^{cs}_{loc}(x^*)$ is at most $n-1$. Therefore $\mu(W^{cs}_{loc}(x^*)) = 0$. The measure of W^{cs}_{loc} can be subsequently bounded as

$$\mu(W^{cs}_{loc}) = \mu\left(\bigcup_{x^\star \in S'} W^{cs}_{loc}(x^\star)\right) \leq \sum_{x^* \in \mathcal{X}'} \mu\left(W^{cs}_{loc}(x^*)\right) = 0$$

where the first inequality is from the semi-countable additivity of Lebesgue measure.

To relate the stable manifolds of these saddle points to the initialization, we need to analyze the gradient map. In contrast to previous analyses, we only show the gradient maps is a *local* diffeomophism instead of a global one, which is considerably weaker but sufficient for our purposes. This result is in the following lemma, which is proved in the appendix.

[1] $f^n(x)$ means the application of f on x repetitively for n times.

Lemma 3.3.2. If a smooth map $g : \mathbb{R}^n \to \mathbb{R}^n$ is a local diffeomorphism, then for every open set S with $\mu(S) = 0$, the inverse set $g^{-1}(S)$ is also a zero-measure set; that is, $\mu\left(g^{-1}(S)\right) = 0$.

Next, we show we can choose a step size in $(0, 2/L)$ to make g a local diffeomorphism except for a zero-measure set.

Lemma 3.3.3. The gradient map $g : \mathbb{R}^n \to \mathbb{R}^n$ in (2.5) is a local diffeomorphism in $\mathbb{R}^n$ for step sizes $h \in (0, 2/L)\backslash H$, where $H \subseteq [1/L, 2/L)$ has measure zero.

Given Lemma 3.3.2 and Lemma 3.3.3, the rest of the proof is fairly straightforward. With Lemma 3.3.3, we know that under the step size $h \in (0, 2/L) \setminus H$ and $\mu(H) = 0$, gradient descent is a local diffeomorphism. Furthermore, with Lemma 3.3.2, we have

$$\mu\left(\bigcup_{i=0}^{\infty} g^{-i}(W_{loc}^{cs})\right) \leq \sum_{i=0}^{\infty} \mu(g^{-i}(W_{loc}^{cs})) = 0.$$

Thus, as long as the random initialization scheme is absolutely continuous with respect to the Lebesgue measure, GD will not converge a saddle point.

3.4 Proof of Theorem 3.2.1

In this section we prove Theorem 3.2.1. First observe that if we can prove a local manifold that converges to the strict saddle point has Lebesgue measure 0, then we can re-use the arguments for proving Theorem 3.1.1. To characterize the local behavior of GD with varying step sizes, we resort to a generalization of the seminal Hartman product map theorem.

3.4.1 Hartman Product Map Theorem

Before describing the Theorem, we need to introduce some conditions and definitions.

Assumption A1 (Hypothesis (H_1) [Har71])**.** Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Banach spaces and $Z = X \times Y$ with norm $\|\cdot\|_Z = \max(\|\cdot\|_X, \|\cdot\|_Y)$. Define $Z_r(0) = \{z \in Z : \|z\|_Z < r\}$. Let $T_n(z) = (A_n x, B_n y) + (F_n(z), G_n(z))$ be a map from $Z_r(0)$ to Z with fixed point 0 and having a continuous Fréchet derivative. Let $A_n : X \to X$ and $B_n : Y \to Y$ be two linear operators and assume B_n is invertible. Suppose

$$\|A_n\|_X \leq a < 1, \quad \|B_n^{-1}\|_Y \leq 1/b \leq 1. \quad 0 < 4\delta < b - a, \quad 0 < a + 2\delta < 1, \tag{3.1}$$

$F_n(0) = 0$, $G_n(0) = 0$ and

$$\begin{cases} \|F_n(z_1) - F_n(z_2)\|_X \leq \delta \|z_1 - z_2\|_Z \\ \|G_n(z_1) - G_n(z_2)\|_Y \leq \delta \|z_1 - z_2\|_Z \end{cases}$$

Here A_n represents local linear operator that acts on the space that corresponds to positive eigenvalues of the Hessian of a saddle point and B_n is a local linear operator that acts on the remaining space. F_n and G_n are higher order functions which vanish at 0.

The main mathematical object we study in this section is the following invariant set.

Definition 3.4.1 (Invariant Set)**.** With the same notations in Assumption A1, let $T_1, \ldots, T_n$ be n maps from $Z_r(0)$ to Z and $S_n = T_n \circ T_{n-1} \circ \cdots \circ T_1$ be the product of the maps. Let $\mathcal{D}_n$ be the invariant set of the product operator S_n and $\mathcal{D} = \bigcap_{n=1}^{\infty} \mathcal{D}_n$.

This set corresponds the points that will converge to the strict saddle point. To study its property, we consider a particular subset.

Definition 3.4.2 ([Har71])**.**

$$\mathcal{D}^{a\delta} = \{z_0 = (x_0, y_0) \in \mathcal{D} : \ z_n \equiv S_n(z_0) \equiv (x_n, y_n) \ \text{ s.t. } \forall n, \|y_n\|_Y \leq \|x_n\|_X \leq (a + 2\delta)^n \|x_0\|_X\}.$$

Now we are ready to state Hartman product map Theorem.

Theorem 3.4.3 (Theorem 7.1, [Har71])**.** Under the Assumption A1, the set $\mathcal{D}^{a\delta}$ is a C^1-manifold and satisfies $D^{a\delta} = \{z = (x, y) \in D | y = y_0(x)\}$ for some function y_0 which is continuous and has continuous Frèchet derivative $D_x y_0$ on $X_r(0)$ and $z_n = S_n(x_0, y_0) \equiv (x_n(x_0), y_n(x_0))$. Further, we have

$$\|y_n(x^0) - y_n(x_0)\|_Y \leq \|x_n(x^0) - x_n(x_0)\|_X \leq (a + 2\delta)^n \|x^0 - x_0\|_X,$$
$$y_n(x_0) = y_0(x_n(x_0)),$$

for any $|x_0|$, $|x^0| < r$ and $n = 0, 1, \ldots$.

Remark 3.4.1. The C^1-manifold $y = y_0(x)$ is equivalent to $y - y_0(x) = 0$. The tangent manifold of y at the fixed point 0 is the intersection set $\bigcap_{i=1}^{dim(y)}\{(x, y)|\nabla_x y_i(x_0) \cdot x - y_i = 0\}$. In the $\mathbb{R}^n$ case, $\{(x, y)|\nabla_x y_i(x_0) \cdot x - y_i = 0\}$ is a subspace of $\mathbb{R}^n$ with dimension at most $n - 1$. Hence, its Lebesgue measure is 0.

Remark 3.4.2. Taking $x^0 = 0$ where 0 is a fixed point, we can rewrite the result of Theorem 3.4.3 as

$$\|y_n(x_0) - y_n(0)\|_Y \leq \|x_n(x_0) - x_n(0)\|_X \leq (a + 2\delta)^n \|x_0 - 0\|_X,$$
$$y_n(0) = y_0(x_n(0)) = 0, \quad x_n(0) = 0.$$

The following theorem from [Har71] implies $D^{a\delta}$ is actually D.

Theorem 3.4.4 (Proposition 7.1, [Har71])**.** Let $z_0 \in \mathcal{D}$; and $z_n = S_n(z_0)$ for $n = 0, 1, \ldots$.

1. If the inequality

$$\|y_m\|_Y \geq \|x_m\|_X$$

holds for some $m \in \mathbb{N}$, then for $n > m$, we have

$$\|y_m\|_Y \geq \|x_m\|_X \qquad \|y_n\|_Y \geq (b - 2\delta)^{n-m} \|y_m\|_Y$$

2. Otherwise, for every $n \in \mathbb{N}$, we have

$$\|y_n\|_Y \leq \|x_n\|_X \leq (a + 2\delta)^n \|x_0\|_X$$

Using Theorem 3.4.4 we have the following useful corollary.

Corollary 3.4.5. If $b - 2\delta > 1$, we have $D = D^{a\delta}$.

3.4.2 Complete Proof of Theorem 3.2.1

We first correspond the parameters of GD to the notations in Assumption A1. Let x^* be a strict saddle point. Since $\nabla^2 f\left(x^*\right)$ is non-singular, it only contain positive and negative eigenvalues. We let $\mathbb{R}^n = X \times Y$ where X corresponds to the space of positive eigenvalues of $\nabla^2 f\left(x^*\right)$ and Y corresponds to the space of negative eigenvalues of $\nabla^2 f\left(x^*\right)$. For any $z \in \mathbb{R}^n$, we write $z = (x, y)$ where x represents to the component in X and y represents the component in Y. Mappings $T_1, T_2, \ldots$ in Assumption A1 correspond to the gradient maps. A_n,B_n,F_n and G_n are thus defined accordingly. The next lemma shows under our assumption on the step size, GD dynamics satisfies Assumption A1.

Lemma 3.4.6. *Suppose $f \in C^2\left(\mathbb{R}^n\right)$ with Hessian Lipschitz constant K and x^* a strict saddle point with $L = \|\nabla^2 f(x^*)\|_2$ and $\mu = \|\left(\nabla^2 f(x^*)\right)^{-1}\|_2^{-1}$. For any fixed $\epsilon_0 \in (0,1)$, if the step size satisfies $h_t \in \left[\frac{\epsilon_0}{L}, \frac{2-\epsilon_0}{L}\right]$, we have*

$$\|A_t\|_2 \leq 1 - \epsilon_0 \qquad \textit{and} \qquad \|B_t^{-1}\|_2 \leq \frac{1}{1 + \frac{\epsilon_0 \mu}{L}}$$

and for any $z_1, z_2 \in \mathbb{R}^n$.

$$\max(\|F_t(z_1) - F_t(z_2)\|_2, \|G_t(z_1) - G_t(z_2)\|_2) \leq \delta \|z_1 - z_2\|_2.$$

where $\delta = \frac{\epsilon}{5}$ and $r = \frac{\epsilon L}{20K}$ (c.f. Assumption A1).

Let D be the invariant set defined in Definition 3.4.2. From Theorem 3.4.3, Remark 3.4.1 and Remark 3.4.2, we know the induced shrinking C^1 manifold $\mathcal{D}^{a\delta}$ defined in Definition 3.4.2 has dimension at most $n-1$. Furthermore, by Corollary 3.4.5 we know that $\mathcal{D} = \mathcal{D}^{a\delta}$. Therefore, the set of points converging to the strict saddle point has zero Lebesgue measure. Similar to the proof of Theorem 3.1.1, since the gradient map is a local diffeomorphism, we can see that with random initialization, GD will not converge to any saddle point. The proof is complete. $\square$

3.5 Appendix

3.5.1 Additional Theorems

Lemma 3.5.1 (The Inverse Function Theorem)**.** Let $f : M \to N$ be a smooth map, and $\dim(M) = \dim(N)$. Suppose that the Jacobian Df_p is nonsingular at some $p \in M$. Then f is a local diffeomorphism at p, i.e., there exists an open neighborhood U of p such that

1. f is one-to-one on U.
2. $f(U)$ is open in N.
3. $f^{-1} : f(U) \to U$ is smooth.

In particular, $D(f^{-1})_{f(p)} = (Df_p)^{-1}$.

Lemma 3.5.2 (Lindelöf's Lemma)**.** For every open cover there is a countable subcover.

3.5.2 Additional Techniques

Proof. [Proof of Lemma 3.3.2] With Theorem 3.5.1, we know that for every $x \in S$, there exists an open set $U_x \in \mathbb{R}^n$ such that g is non-singular. Let $W_x = S \cap U_x$, then we have

$$S \subseteq \bigcup_{x \in S} W_x$$

With Lindelöf's Lemma, there exists a set S' with countable elements x such that

$$S \subseteq \bigcup_{x \in S'} W_x$$

Since the Dg is non-singular on W_x , we know that g on W_x is one-one-onto. Hence, we have $\mu(g^{-1}(W_x)) = 0$. Hence, we have

$$\mu(g^{-1}(\bigcup_{x \in S} W_x)) = \mu(g^{-1}(\bigcup_{x \in S'} W_x)) \leq \mu(\bigcup_{x \in S'} g^{-1}(x)) \leq \sum_{x \in S'} \mu(g^{-1}(x)) = 0$$

where the second inequality is from monotony of Lebesgue measure and the third inequality is from semi-countable additivity of Lebesgue measure. □

Proof. Proof of Lemma 3.3.3 If the Jacobian of the gradient map Dg is non-singular at some point $x \in \mathbb{R}^n$, with the continuity of the Jacobian Dg, we know that Dg is non-singular at some open neighborhood $\mathcal{U}_x$ of the point x. Hence, we have

$$\mathbb{R}^n \subseteq \bigcup_{x \in \mathbb{R}^n} \mathcal{U}_x$$

With Lindelöf's Lemma, there exists a set $\mathcal{S}$ with countable number of $x \in \mathbb{R}^n$ such that

$$\mathbb{R}^n \subseteq \bigcup_{x \in \mathcal{S}} \mathcal{U}_x$$

Let $\mathcal{H}_x$ be the step size that Jacobian Dg s singular at the open set $\mathcal{U}_x$. With the definition of $\mathcal{U}_x$, we know that there are at most n elements in $\mathcal{H}_x$. Hence, we have

$$\mu(H_x) = 0 \qquad and \qquad H = \bigcup_{x \in \mathcal{S}} H_x,$$

where H satisfies that the Jacobian of the gradient map non-singular with step size $h \in \left(0, \frac{2}{L}\right) \setminus H$. With the semi-countable additivity of Lebesgue measure, we have

$$\mu(H) = \mu\left(\bigcup_{x \in \mathcal{S}} H_x\right) \leq \sum_{x \in \mathcal{S}} \mu(H_x) = 0.$$

□

Proof. Proof of Proposition 3.1.1 Consider the following quadratic function

$$f(x) = \frac{1}{2} x^T A x$$

where A is a diagonal matrix $A = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$ and satisfies $\lambda_1 > \lambda_2 > \ldots, \lambda_n > 0$. The global gradient Lipschitz constant L of f is λ_1. Now consider the gradient dynamics

$$x_{t+1} = x_t - hAx_t = (I - hA)\, x_t.$$

Since $h \geq \frac{2}{L}$, $\lambda_{\max}(I - hA) \geq 1$. Therefore, the sequence $\{x_0, x_1, \ldots\}$ does not converge. □

Proof. Proof of Lemma 3.4.6 If $x_t \in V_r(x^\star)$, the step size satisfies

$$h_t \in \left[\frac{\epsilon_0}{L_{(x_t,r)}}, \frac{2-\epsilon_0}{L_{(x_t,r)}}\right] \subseteq \left[\frac{\epsilon_0}{L - 2Kr}, \frac{2-\epsilon_0}{L}\right] \subseteq \left[\frac{\epsilon_0}{L(1-0.1\epsilon_0)}, \frac{2-\epsilon_0}{L}\right]$$

Therefore, we known

$$h_t \in \left[\frac{\epsilon_0'}{L}, \frac{2-\epsilon_0'}{L}\right].$$

where $\epsilon_0' = \frac{\epsilon_0}{1-0.1\epsilon_0'}$. Since both A_t and B_t are diagonal, then the 2-norm is equal to the maximum eigenvalue, that is,

$$\begin{aligned}
\|A_t\|_2 &= 1 - \max|\lambda(\nabla^2|f(x^\star)) \cdot h \leq 1 - \epsilon_0 \\
\|B_t^{-1}\|_2 &= \frac{1}{1 + \mu \min|\lambda(\nabla^2 f(x^*))|h} \leq \frac{1}{1 + \frac{\epsilon_0 \mu}{L}}.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\max(\|F_1(z_1) - F_1(z_2)\|_2, \|F_2(z_1) - F_2(z_2)\|_2) \\
&\leq h\|(\nabla f(x) - \nabla f(y)) + \nabla^2 f(x^\star)(x - y)\|_2 \\
&\leq hK\left(\|z_1\|_2 + \|z_2\|_2\right)\|z_1 - z_2\|_2
\end{aligned}$$

Plugging in our assumption on the step size we have the desired result. □

CHAPTER 4

A CONSERVATION LAW METHOD IN OPTIMIZATION

This chapter is organized as follows. In Section 4.1, we warm up with a analytical solution for simple 1-d quadratic function. In Section 4.3, we propose the artificially dissipating energy algorithm, energy conservation algorithm and the combined algorithm based on the symplectic Euler scheme, and remark a second-order scheme — the Störmer-Verlet scheme . In Section 4.4, we propose the locally theoretical analysis for High-Speed converegnce. Section 4.5 propose the experimental demonstration. In section 4.5, we propose the experimental result for the proposed algorithms on strongly convex function, non-strongly convex function and nonconvex function in high-dimension. Finally, we propose some perspective view for the proposed algorithms and two adventurous ideas based on the evolution of Newton Second Law — fluid and quantum.

4.1 Warm-up: An Analytical Demonstration for Intuition

For a simple 1-D function with ill-conditioned Hessian, $f(x) = \frac{1}{200}x^2$ with the initial position at $x_0 = 1000$. The solution and the function value along the solution for (1.14) are given by

$$\begin{cases} x(t) = x_0 e^{-\frac{1}{100}t} & (4.1) \\ f(x(t)) = \dfrac{1}{200}x_0^2 e^{-\frac{1}{50}t}. & (4.2) \end{cases}$$

The solution and the function value along the solution for (1.15) with the optimal friction parameter $\gamma_t = \frac{1}{5}$ are

$$\begin{cases} x(t) = x_0 \left(1 + \dfrac{1}{10}t\right) e^{-\frac{1}{10}t} & (4.3) \\ f(x(t)) = \dfrac{1}{200}x_0^2 \left(1 + \dfrac{1}{10}t\right)^2 e^{-\frac{1}{5}t}. & (4.4) \end{cases}$$

The solution and the function value along the solution for (1.17) are

$$\begin{cases} x(t) = x_0 \cos\left(\dfrac{1}{10}t\right) \quad \text{and} \quad v(t) = x_0 \sin\left(\dfrac{1}{10}t\right) & (4.5) \\ f(x(t)) = \dfrac{1}{200}x_0^2 \cos^2\left(\dfrac{1}{10}t\right) & (4.6) \end{cases}$$

stop at the point that $|v|$ arrive maximum. Combined with (4.2), (4.4) and (4.6) with stop at the point that $|v|$ arrive maximum, the function value approximating $f(x^\star)$ are shown as below,

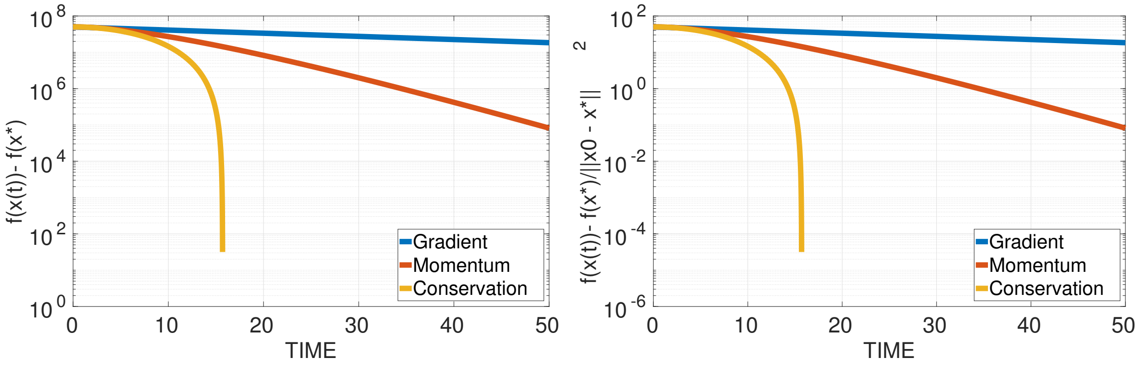


Figure 4.1: Minimizing $f(x) = \frac{1}{200}x^2$ by the analytical solution for (4.2), (4.4) and (4.6) with stop at the point that $|v|$ arrive maximum, starting from $x_0 = 1000$ and the numerical step size $\Delta t = 0.01$.

From the analytical solution for local convex quadratic function with maximum eigenvalue L and minimum eigenvalue μ, in general, the step size by $\frac{1}{\sqrt{L}}$ for momentum method and Nesterov accelerated gradient method, hence the simple estimate for iterative times is approximately

$$n \sim \frac{\pi}{2}\sqrt{\frac{L}{\mu}}.$$

hence, the iterative times n is proportional to the reciprocal of the square root of minimal eigenvalue $\sqrt{\mu}$, which is essentially different from the convergence rate of the gradient method and momentum method.

4.2 Related Work

The history of gradient method for convex optimization can be back to the time of Euler and Lagrange. However, since it is relatively cheaper to only calculation for first-order information, this simplest and earliest method is still active in machine learning and nonconvex optimization, such as the recent work [GHJY15, AG16, LSJR16, HMR16]. The natural speedup algorithms are the momentum method first proposed in [Pol64] and Nesterov accelerated gradient method first proposed in [Nes83] and an improved version [NN88]. A acceleration algorithm similar as Nesterov accelerated gradient method, named as FISTA, is designed to solve composition problems [BT09]. A related comprehensive work is proposed in [B+15].

The original momentum method, named as Polyak heavy ball method, is from the view of ODE in [Pol64], which contains extremely rich physical intuitive ideas and mathematical theory. An extremely important work in application on machine learning is the backpropagation learning with momentum [RHW+88]. Based on the thought of ODE, a lot of understanding and application on the momentum method and Nesterov accelerated gradient methods have been proposed. In [SMDH13], a well-designed random initialization with momentum parameter algorithm is proposed to train both DNNs and RNNs. A seminal deep insight from ODE to understand the intuition behind Nesterov scheme is proposed in [SBC14]. The understanding for momentum method based on the variation perspective is proposed on [WWJ16], and the understanding from Lyaponuv analysis is proposed in [WRJ16]. From the stability theorem of ODE, the gradient method always converges to local minima in the sense of almost everywhere is proposed in [LSJR16]. Analyzing and designing iterative optimization algorithms built on integral quadratic constraints from robust control theory is proposed in [LRP16].

Actually the "high momentum" phenomenon has been firstly observed in [OC15] for a restarting adaptive accelerating algorithm, and also the restarting scheme is proposed by [SBC14]. However, both works above utilize restarting scheme for an auxiliary tool to accelerate the algorithm based on friction. With the concept of phase space in mechanics, we observe that the kinetic energy, or velocity, is controllable and utilizable parameter to find the local minima. Without friction term, we can still find the local minima only by the velocity parameter. Based on this view, the algorithm is proposed very easy to practice. Meanwhile, the thought can be generalized to nonconvex optimization to detect local minima along the trajectory of the particle.

4.3 Symplectic Scheme and Algorithms

In this chapter, we utilize the first-order symplectic Euler scheme from numerically solving Hamiltonian system as below

$$\begin{cases} x_{k+1} = x_k + hv_{k+1} \\ v_{k+1} = v_k - h\nabla f(x_k) \end{cases} \tag{4.7}$$

to propose the corresponding artifically dissipating energy algorithm to find the global minima for convex function, or local minima in non-convex function. Then by the observability of the velocity, we propose the energy conservation algorithm for detecting local minima along the trajectory. Finally, we propose a combined algorithm to find better local minima between some local minima.

Remark 4.3.1. In all the algorithms below, the symplectic Euler scheme can be taken place by the Störmer-Verlet scheme, i.e.

$$\begin{cases} v_{k+1/2} = v_k - \dfrac{h}{2}\nabla f(x_k) \\ x_{k+1} = x_k + hv_{k+1/2} \\ v_{k+1} = v_{k+1/2} - \dfrac{h}{2}\nabla f(x_{k+1}) \end{cases} \tag{4.8}$$

which works perfectly better than the symplectic scheme even if doubling step size and keep the left-right symmetry of the Hamiltonian system. The Störmer-Verlet scheme is the natural discretization for 2nd-order ODE

$$x_{k+1} - 2x_k + x_{k-1} = -h^2 \nabla f(x_k) \tag{4.9}$$

which is named as leap-frog scheme in PDEs. We remark that the discrete scheme (4.9) is different from the finite difference approximation by the forward Euler method to analyze the stability of 2nd ODE in [SBC14], since the momentum term is biased.

4.3.1 Artifically Dissipating Energy Algorithm

Firstly, the artificially dissipating energy algorithm based on (4.7) is proposed as below.

Algorithm 1 Artifically Dissipating Energy Algorithm

```
 1: Given a starting point $x_0 \in \mathbf{dom}(f)$
 2: Initialize the step length $h$, maxiter, and the velocity variable $v_0 = 0$
 3: Initialize the iterative variable $v_{iter} = v_0$
 4: while $\|\nabla f(x)\| > \epsilon$ and  $k <$ maxiter do
 5:     Compute $v_{iter}$ from the below equation in (4.7)
 6:     if $\|v_{iter}\| \leq \|v\|$ then
 7:         $v = 0$
 8:     else
 9:         $v = v_{iter}$
10:     end if
11:     Compute $x$ from the above equation in (4.7)
12:     $x_k = x$;
13:     $f(x_k) = f(x)$;
14:     $k = k + 1$;
15: end while
```

Remark 4.3.2. In the actual algorithm 1, the codes in line 15 and 16 are not need in the while loop in order to speed up the computation.

4.3.1.1 A Simple Example For Illustration

Here, we use a simple convex quadratic function with ill-conditioned eigenvalue for illustration as below,

$$f(x_1, x_2) = \frac{1}{2}\left(x_1^2 + \alpha x_2^2\right), \tag{4.10}$$

of which the maximum eigenvalue is $L = 1$ and the minimum eigenvalue is $\mu = \alpha$. Hence the scale of the step size for (4.10) is

$$\frac{1}{L} = \sqrt{\frac{1}{L}} = 1.$$

In figure 4.2, we demonstrate the convergence rate of gradient method, momentum method, Nesterov accelerated gradient method and artifically dissipating energy method with the common step size $h = 0.1$ and $h = 0.5$, where the optimal friction parameter for momentum method $\gamma = \frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}}$ with $\alpha = 10^{-5}$. A further result for comparison with the optimal step size in gradient method $h = \frac{2}{1+\alpha}$, the momentum method $h = \frac{4}{(1+\sqrt{\alpha})^2}$, and Nesterov accelerated gradient method with $h = 1$ and the artifically disspating energy method with $h = 0.5$ shown in figure 4.3.

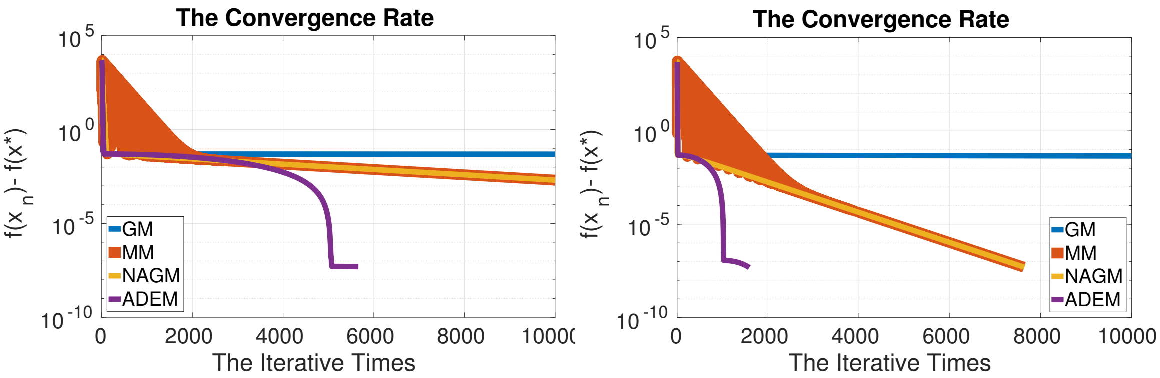


Figure 4.2: Mimimize the function in (4.10) for artificially dissipating energy algorithm comparing with gradient method, momentum method and Nesterov accelerated gradient method with stop criteria $\epsilon = 1e - 6$. The Step size: Left: $h = 0.1$; Right: $h = 0.5$.

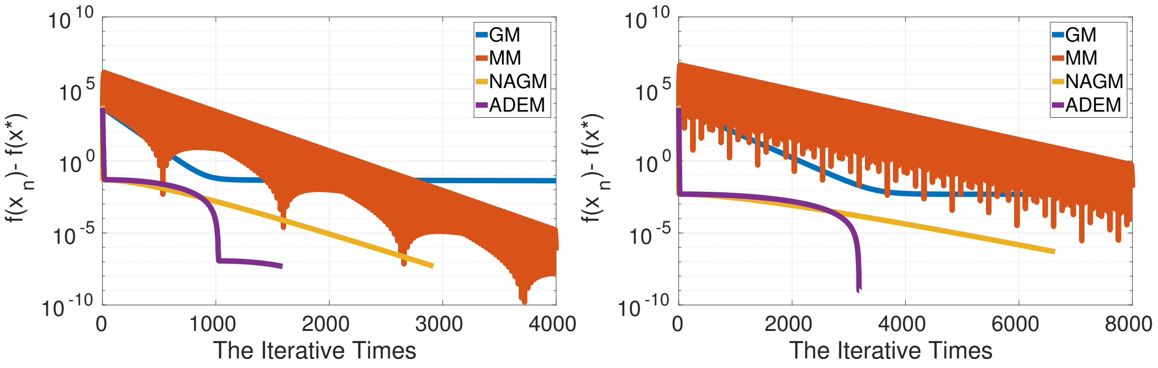


Figure 4.3: Mimimize the function in (4.10) for artificially dissipating energy algorithm comparing with gradient method, momentum method and Nesterov accelerated gradient method with stop criteria $\epsilon = 1e - 6$. The Coefficient α: Left: $\alpha = 10^{-5}$; Right: $\alpha = 10^{-6}$.

With the illustrative convergence rate, we need to learn the trajectory. Since the trajectories of all the four methods are so narrow in ill-condition function in (4.10), we use a relatively good-conditioned function to show it as $\alpha = \frac{1}{10}$ in figure 4.4.

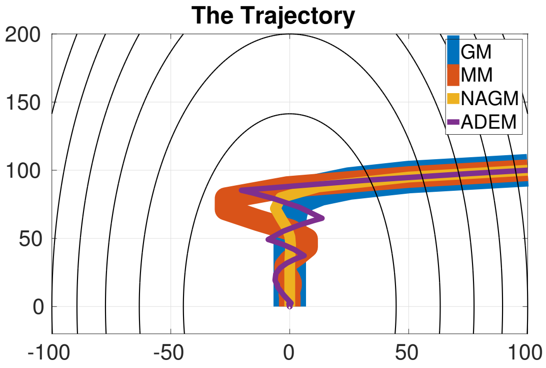


Figure 4.4: The trajectory for gradient method, momentum method, Nesterov accelerated method and artifically dissipating energy method for the function (4.10) with $\alpha = 0.1$.

A clear fact in figure 4.4 shows that the gradient correction decrease the oscillation to comparing with momentum method. A more clear observation is that artificially dissipating method owns the same property with the other three method by the law of nature, that is, if the trajectory come into the local minima in one dimension will not leave it very far. However, from figure 4.2 and figure 4.3, the more rapid convergence rate from artificially dissipating energy method has been shown.

4.3.2 Energy Conservation Algorithm for Detecting Local Minima

Here, the energy conservation algorithm based on (4.7) is proposed as below.

Algorithm 2 Energy Conservation Algorithm

```
Given a starting point $x_0 \in \mathbf{dom}(f)$
Initialize the step size $h$ and the maxiter
Initialize the velocity $v_0 > 0$ and compute $f(x_0)$
Compute the velocity $x_1$ and $v_1$ from the equation (4.7), and compute $f(x_1)$
for $k = 1 : n$ do
    Compute $x_{k+1}$ and $v_{k+1}$ from (4.7)
    Compute $f(x_{k+1})$
    if $\|v_k\| \geq \|v_{k+1}\|$ and $\|v_k\| \geq \|v_{k-1}\|$ then
        Record the position $x_k$
    end if
end for
```

Remark 4.3.3. In the algorithm 2, we can set $v_0 > 0$ such that the total energy large enough to climb up some high peak. Same as the algorithm 1, the function value $f(x)$ is not need in the while loop in order to speed up the computation.

4.3.2.1 The Simple Example For Illustration

Here, we use the non-convex function for illustration as below,

$$f(x) = \begin{cases} 2\cos(x), & x \in [0, 2\pi] \\ \cos(x) + 1, & x \in [2\pi, 4\pi] \\ 3\cos(x) - 1, & x \in [4\pi, 6\pi] \end{cases} \tag{4.11}$$

which is the 2nd-order smooth function but not 3rd-order smooth. The maximum eigenvalue can be calculated as below

$$\max_{x\in[0,6\pi]} |f''(x)| = 3.$$

then, the step length is set $h \sim \sqrt{\frac{1}{L}}$. We illustrate that the algorithm 2 simulate the trajectory and find the local minima in figure 4.5.

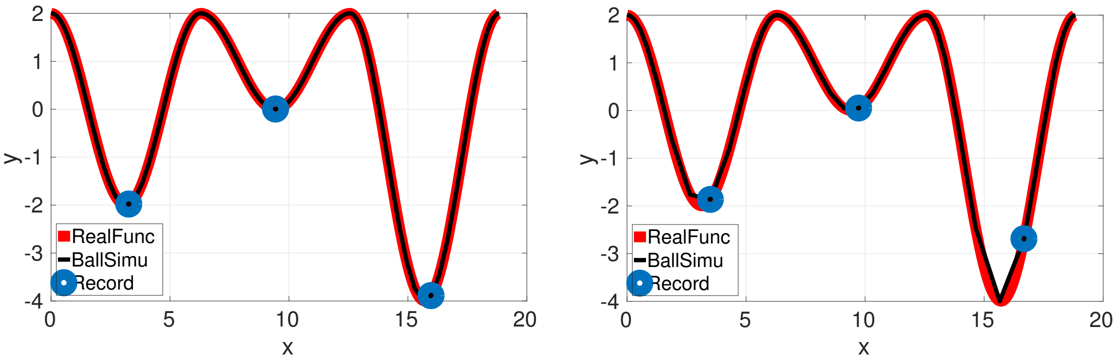


Figure 4.5: The Left: the step size $h = 0.1$ with 180 iterative times. The Right: the step size $h = 0.3$ with 61 iterative times.

Another 2D potential function is shown as below,

$$f(x_1, x_2) = \frac{1}{2}\left[(x_1 - 4)^2 + (x_2 - 4)^2 + 8\sin(x_1 + 2x_2)\right]. \tag{4.12}$$

which is the smooth function with domain in $(x_1, x_2) \in [0, 8] \times [0, 8]$. The maximum eigenvalue can be calculated as below

$$\max_{x\in[0,6\pi]} |\lambda(f''(x))| \geq 16.$$

then, the step length is set $h \sim \sqrt{\frac{1}{L}}$. We illustrate that the algorithm 2 simulate the trajectory and find the local minima in figure 4.6.

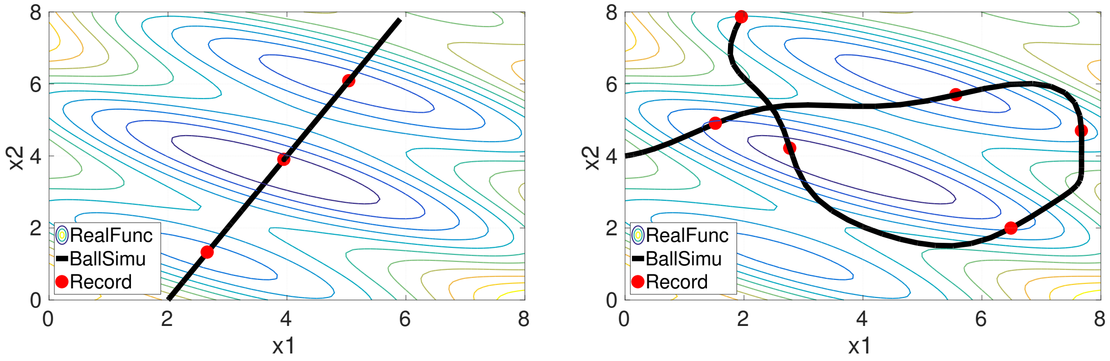


Figure 4.6: The common step size is set $h = 0.1$. The Left: the position at $(2, 0)$ with 23 iterative times. The Right: the position at $(0, 4)$ with 62 iterative times.

Remark 4.3.4. We point out that for the energy conservation algorithm for detecting local minima along the trajectory cannot detect saddle point in the sense of almost every, since the saddle point in original function $f(x)$ is also a saddle point for the energy function $H(x, v) = \frac{1}{2}\|v\|^2 + f(x)$. The proof process is fully the same in [LSJR16].

4.3.3 Combined Algorithm

Finally, we propose the comprehensive algorithm combining the artificially dissipating energy algorithm (algorithm 1) and the energy conservation algorithm (2) to find global minima.

Algorithm 3 Combined Algorithm

```
1: Given some starting points $x_{0,i} \in \mathbf{dom}(f)$ with $i = 1, \ldots, n$
2: Implement algorithm 2 detecting the position there exists local minima, noted as $x_j$ with $j = 1, \ldots, m$
3: Implement algorithm 1 from the result on line 2 finding the local minima, noted as $x_k$ with $k = 1, \ldots, l$
4: Comparison of $f(x_k)$ with $k = 1, \ldots, l$ to find global minima.
```

Remark 4.3.5. We remark that the combined algorithm (algorithm 3) cannot guarantee to find global minima if the initial position is not ergodic. The tracking local minima is dependent on the trajectory. However, the time of computation and precision based on the proposed algorithm is far better than the large sampled gradient method. Our proposed algorithm first makes the global minima found become possible.

4.4 An Asymptotic Analysis for the Phenomena of Local High-Speed Convergence

In this section, we analyze the phenomena of high-speed convergence shown in figure 4.1, figure 4.2 and figure 4.3. Without loss of generality, we use the translate transformation $y_k = x_k - x^\star$ ($x^\star$ is the point of local minima) and $v_k = v_k$ into (4.7), shown as below,

$$\begin{cases} y_{k+1} = y_k + hv_{k+1} \\ v_{k+1} = v_k - h\nabla f(x^\star + y_k), \end{cases} \tag{4.13}$$

the locally linearized scheme of which is given as below,

$$\begin{cases} y_{k+1} = y_k + hv_{k+1} \\ v_{k+1} = v_k - h\nabla^2 f(x^\star) y_k. \end{cases} \tag{4.14}$$

Remark 4.4.1. The local linearized analysis is based on the stability theorem in finite dimension, the invariant stable manifold theorem and Hartman-Grobman linearized map theorem [Har82]. The thought is firstly used in [Pol64] to estimate the local convergence of momentum method. And in the paper [LSJR16], the thought is used to exclude the possiblity to converegnce to saddle point. However, the two theorems above belong to the qualitative theorem of ODE. Hence, the linearized scheme (4.14) is only an approximate estimate for the original scheme (4.13) locally.

4.4.1 Some Lemmas for the Linearized Scheme

Let A be the positive-semidefinite and symmetric matrix to represent $\nabla^2 f(x^\star)$ in (4.14).

Lemma 4.4.2. The numerical scheme, shown as below

$$\begin{pmatrix} x_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} I - h^2 A & hI \\ -hA & I \end{pmatrix} \begin{pmatrix} x_k \\ v_k \end{pmatrix} \tag{4.15}$$

is equivalent to the linearized symplectic-Euler scheme (4.14), where we note that the linear transformation is

$$M = \begin{pmatrix} I - h^2 A & hI \\ -hA & I \end{pmatrix}. \tag{4.16}$$

Proof.

$$\begin{pmatrix} I & -hI \\ 0 & I \end{pmatrix} \begin{pmatrix} x_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ -hA & I \end{pmatrix} \begin{pmatrix} x_k \\ v_k \end{pmatrix} \Leftrightarrow \begin{pmatrix} x_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} I - h^2 A & hI \\ -hA & I \end{pmatrix} \begin{pmatrix} x_k \\ v_k \end{pmatrix}$$

□

Lemma 4.4.3. For every $2n \times 2n$ matrix M in (4.16), there exists the orthogonal transformation $U_{2n\times 2n}$ such that the matrix M is similar as below

$$U^T M U = \begin{pmatrix} T_1 & & & \\ & T_2 & & \\ & & \ddots & \\ & & & T_n \end{pmatrix} \tag{4.17}$$

where T_i $(i = 1, \ldots, n)$ is 2×2 matrix with the form

$$T_i = \begin{pmatrix} 1 - \omega_i^2 h^2 & h \\ -\omega_i^2 h & 1 \end{pmatrix} \tag{4.18}$$

where ω_i^2 is the eigenvalue of the matrix A.

Proof. Let Λ be the diagonal matrix with the eigenvalues of the matrix A as below

$$\Lambda = \begin{pmatrix} \omega_1^2 & & & \\ & \omega_2^2 & & \\ & & \ddots & \\ & & & \omega_n^2 \end{pmatrix}.$$

Since A is positive define and symmetric, there exists orthogonal matrix U_1 such that

$$U_1^T A U_1 = \Lambda$$

Let Π be the permuation matrix satisfying

$$\Pi_{i,j} = \begin{cases} 1, & j \text{ odd}, \; i = \dfrac{j+1}{2} \\ 1, & j \text{ even}, \; i = n + \dfrac{j}{2} \\ 0, & \text{otherwise} \end{cases}$$

where i is the row index and j is the column index. Then, let $U = \mathbf{diag}(U_1, U_1)\Pi$, we have by conjugation

$$\begin{aligned} U^T M U &= \Pi^T \begin{pmatrix} U_1^T & \\ & U_1^T \end{pmatrix} \begin{pmatrix} I - h^2 A & hI \\ -hA & I \end{pmatrix} \begin{pmatrix} U_1 & \\ & U_1 \end{pmatrix} \Pi \\ &= \Pi^T \begin{pmatrix} I - h^2 \Lambda & hI \\ -h\Lambda & I \end{pmatrix} \Pi \\ &= \begin{pmatrix} T_1 & & & \\ & T_2 & & \\ & & \ddots & \\ & & & T_n \end{pmatrix} \end{aligned}$$

□

From Lemma 4.4.3, we know that the equation (4.15) can be written as the equivalent form

$$\begin{pmatrix} (U_1^T x)_{k+1,i} \\ (U_1^T v)_{k+1,i} \end{pmatrix} = T_i \begin{pmatrix} (U_1^T x)_{k,i} \\ (U_1^T v)_{k,i} \end{pmatrix} = \begin{pmatrix} 1-\omega_i^2 h^2 & h \\ -\omega_i^2 h & 1 \end{pmatrix} \begin{pmatrix} (U_1^T x)_{k,i} \\ (U_1^T v)_{k,i} \end{pmatrix} \tag{4.19}$$

where $i = 1, \ldots, n$.

Lemma 4.4.4. For any step size h satisfying $0 < h\omega_i < 2$, the eigenvalues of the matrix T_i are complex with absolute value 1.

Proof. For $i = 1, \ldots, n$, we have

$$|\lambda I - T_i| = 0 \Leftrightarrow \lambda_{1,2} = 1 - \frac{h^2\omega_i^2}{2} \pm h\omega_i\sqrt{1 - \frac{h^2\omega_i^2}{4}}.$$

□

Let θ_i and ϕ_i for $i = 1, \ldots, n$ for the new coordinate variables as below

$$\begin{cases} \cos\theta_i = 1 - \dfrac{h^2\omega_i^2}{2} \\ \sin\theta_i = h\omega_i\sqrt{1 - \dfrac{h^2\omega_i^2}{4}} \end{cases}, \qquad \begin{cases} \cos\phi_i = \dfrac{h\omega_i}{2} \\ \sin\phi_i = \sqrt{1 - \dfrac{h^2\omega_i^2}{4}} \end{cases} \tag{4.20}$$

In order to make θ_i and ϕ_i located in $\left(0, \frac{\pi}{2}\right)$, we need to shrink to $0 < h\omega_i < \sqrt{2}$.

Lemma 4.4.5. With the new coordinate in (4.20) for $0 < h\omega_i < \sqrt{2}$, we have

$$2\phi_i + \theta_i = \pi \tag{4.21}$$

and

$$\begin{cases} \sin\theta_i = \sin(2\phi_i) = h\omega_i \sin\phi_i \\ \sin(3\phi_i) = -\left(1 - h^2\omega_i^2\right)\sin\phi_i \end{cases} \tag{4.22}$$

Proof. With Sum-Product identities of trigonometric function, we have

$$\begin{aligned}\sin(\theta_i+\phi_i) &= \sin\theta_i\cos\phi_i+\cos\theta_i\sin\phi_i\\ &= h\omega_i\sqrt{1-\frac{h^2\omega_i^2}{4}}\cdot\frac{h\omega_i}{2}+\left(1-\frac{h^2\omega_i^2}{2}\right)\sqrt{1-\frac{h^2\omega_i^2}{4}}\\ &= \sqrt{1-\frac{h^2\omega_i^2}{4}}\\ &= \sin\phi_i.\end{aligned}$$

Since $0<h\omega_i<2$, we have $\theta_i,\phi_i\in\left(0,\frac{\pi}{2}\right)$, we can obtain that

$$\theta_i+\phi_i=\pi-\phi_i \Leftrightarrow \theta_i=\pi-2\phi_i$$

and with the coordinate transfornation in (4.20), we have

$$\sin\theta_i=h\omega_i\sin\phi_i \Leftrightarrow \sin(2\phi_i)=h\omega_i\sin\phi_i.$$

Next, we use Sum-Product identities of trigonometric function furthermore,

$$\begin{aligned}\sin(\theta_i-\phi_i) &= \sin\theta_i\cos\phi_i-\cos\theta_i\sin\phi_i\\ &= h\omega_i\sqrt{1-\frac{h^2\omega_i^2}{4}}\cdot\frac{h\omega_i}{2}-\left(1-\frac{h^2\omega_i^2}{2}\right)\sqrt{1-\frac{h^2\omega_i^2}{4}}\\ &= \left(h^2\omega_i^2-1\right)\sqrt{1-\frac{h^2\omega_i^2}{4}}\\ &= -\left(1-h^2\omega_i^2\right)\sin\phi_i\end{aligned}$$

and with $\theta_i=\pi-2\phi_i$, we have

$$\sin(3\phi_i)=-\left(1-h^2\omega_i^2\right)\sin\phi_i$$

□

Lemma 4.4.6. With the new coordinate in (4.20), the matrix T_i $(i=1,\ldots,n)$ in (4.18) can expressed as below,

$$T_i=\frac{1}{\omega_i\left(e^{-i\phi_i}-e^{i\phi_i}\right)}\begin{pmatrix}1 & 1\\ \omega_i e^{i\phi_i} & \omega_i e^{-i\phi_i}\end{pmatrix}\begin{pmatrix}e^{i\theta_i} & 0\\ 0 & e^{-i\theta_i}\end{pmatrix}\begin{pmatrix}\omega_i e^{-i\phi_i} & -1\\ -\omega_i e^{i\phi_i} & 1\end{pmatrix} \tag{4.23}$$

Proof. For the coordinate transformation in (4.20), we have

$$T_i \begin{pmatrix} 1 \\ \omega_i e^{i\phi_i} \end{pmatrix} = \begin{pmatrix} 1 \\ \omega_i e^{i\phi_i} \end{pmatrix} e^{i\theta_i} \qquad \text{and} \qquad T_i \begin{pmatrix} 1 \\ \omega_i e^{-i\phi_i} \end{pmatrix} = \begin{pmatrix} 1 \\ \omega_i e^{-i\phi_i} \end{pmatrix} e^{-i\theta_i}$$

Hence, (4.23) is proved. □

4.4.2 The Asymptotic Analysis

Theorem 4.4.7. Let the initial value x_0 and v_0, after the first k steps without reseting the velocity, the iterative solution (4.14) with the equivalent form (4.19) has the form as below

$$\begin{pmatrix} (U_1^T x)_{k,i} \\ (U_1^T v)_{k,i} \end{pmatrix} = T_i^k \begin{pmatrix} (U_1^T x)_{0,i} \\ (U_1^T v)_{0,i} \end{pmatrix} = \begin{pmatrix} -\frac{\sin(k\theta_i - \phi_i)}{\sin\phi_i} & \frac{\sin(k\theta_i)}{\omega_i \sin\phi_i} \\ -\frac{\omega_i \sin(k\theta_i)}{\sin\phi_i} & \frac{\sin(k\theta_i + \phi_i)}{\sin\phi_i} \end{pmatrix} \begin{pmatrix} (U_1^T x)_{0,i} \\ (U_1^T v)_{0,i} \end{pmatrix} \tag{4.24}$$

Proof. With Lemma 4.4.6 and the coordinate transformation (4.20), we have

$$\begin{aligned} T_i^k &= \frac{1}{\omega_i \left(e^{-i\phi_i} - e^{i\phi_i}\right)} \begin{pmatrix} 1 & 1 \\ \omega_i e^{i\phi_i} & \omega_i e^{-i\phi_i} \end{pmatrix} \begin{pmatrix} e^{i\theta_i} & 0 \\ 0 & e^{-i\theta_i} \end{pmatrix}^k \begin{pmatrix} \omega_i e^{-i\phi_i} & -1 \\ -\omega_i e^{i\phi_i} & 1 \end{pmatrix} \\ &= \frac{1}{\omega_i \left(e^{-i\phi_i} - e^{i\phi_i}\right)} \begin{pmatrix} 1 & 1 \\ \omega_i e^{i\phi_i} & \omega_i e^{-i\phi_i} \end{pmatrix} \begin{pmatrix} \omega e^{i(k\theta_i - \phi_i)} & -e^{ik\theta_i} \\ -\omega e^{-i(k\theta_i - \phi_i)} & e^{-ik\theta_i} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{\sin(k\theta_i - \phi_i)}{\sin\phi_i} & \frac{\sin(k\theta_i)}{\omega_i \sin\phi_i} \\ -\frac{\omega_i \sin(k\theta_i)}{\sin\phi_i} & \frac{\sin(k\theta_i + \phi_i)}{\sin\phi_i} \end{pmatrix} \end{aligned}$$

The proof is complete. □

Comparing (4.24) and (4.19), we can obtain that

$$\frac{\sin(k\theta_i - \phi_i)}{\sin\phi_i} = 1 - h^2\omega_i^2.$$

With the intial value $(x_0, 0)^T$, then the initial value for (4.19) is $(U_1^T x_0, 0)$. In order to make sure the numerical solution, or the iterative solution owns the same behavior as the analytical solution, we need to set $0 < h\omega_i < 1$.

Remark 4.4.8. Here, the behavior is similar as the thought in [LSJR16]. The step size $0 < hL < 2$ make sure the global convergence of gradient method. And the step size $0 < hL < 1$ make the uniqueness of the trajectory along the gradient method, the thought of which is equivalent of the existencen and uniqueness of the solution for ODE. Actually, the step size $0 < hL < 1$ owns the property with the solution of ODE, the continous-limit version. A global existence of the solution for gradient system is proved in [Per13].

For the good-conditioned eigenvalue of the Hessian $\nabla^2 f(x^\star)$, every method such as gradient method, momentum method, Nesterov accelerated gradient method and artificially dissipating energy method has the good convergence rate shown by the experiment. However, for our artificially dissipating energy method, since there are trigonometric functions from (4.24), we cannot propose the rigorous mathematic proof for the convergence rate. If everybody can propose a theoretical proof, it is very beautiful. Here, we propose a theoretical approximation for ill-conditioned case, that is, the direction with small eigenvalue $\lambda(\nabla^2 f(x^\star)) \ll L$.

Assumption A2. If the step size $h = \frac{1}{\sqrt{L}}$ for (4.14), for the ill-conditioned eigenvalue $\omega_i \ll \sqrt{L}$, the coordinate variable can be approximated by the analytical solution as

$$\theta_i = h\omega_i, \qquad \text{and} \qquad \phi_i = \frac{\pi}{2}. \tag{4.25}$$

With Assumption A2, the iterative solution (4.24) can be rewritten as

$$\begin{pmatrix} (U_1^T x)_{k,i} \\ (U_1^T v)_{k,i} \end{pmatrix} = \begin{pmatrix} \cos(kh\omega_i) & \frac{\sin(kh\omega_i)}{\omega_i} \\ -\omega_i \sin(kh\omega_i) & -\cos(kh\omega_i) \end{pmatrix} \begin{pmatrix} (U_1^T x)_{0,i} \\ (U_1^T v)_{0,i} \end{pmatrix} \tag{4.26}$$

Theorem 4.4.9. For every ill-conditioned eigen-direction, with every initial condition $(x_0, 0)^T$, if the algorithm 1 is implemented at $\|v_{iter}\| \leq \|v\|$, then there exist an eigenvalue ω_i^2 such that

$$k\omega_i h \geq \frac{\pi}{2}.$$

Proof. When $\|v_{iter}\| \leq \|v\|$, then $\left\|U_1^T v_{iter}\right\| \leq \left\|U_1^T v\right\|$. While for the $\left\|U_1^T v\right\|$, we can write in the analytical form,

$$\left\|U_1^T v\right\| = \sqrt{\sum_{i=1}^{n} \omega_i^2 (U_1 x_0)_i^2 \sin^2(kh\omega_i)}$$

if there is no $k\omega_i h < \frac{\pi}{2}$, $\left\|U_1^T v\right\|$ increase with k increasing. □

For some i such that $k\omega_i h$ approximating $\frac{\pi}{2}$, we have

$$\begin{aligned} \frac{\left|(U_1^T x)_{k+1,i}\right|}{\left|(U_1^T x)_{k,i}\right|} &= \frac{\cos\left((k+1)h\omega_i\right)}{\cos\left(kh\omega_i\right)} \\ &= e^{\ln\cos((k+1)h\omega_i) - \ln\cos(kh\omega_i)} \\ &= e^{-\tan(\xi)h\omega_i} \end{aligned} \tag{4.27}$$

where $\xi \in (kh\omega_i, (k+1)h\omega_i)$. Hence, with ξ approximating $\frac{\pi}{2}$, $\left|(U_1^T x)_{k,i}\right|$ approximatie 0 with the linear convergence, but the coefficient will also decay with the rate $e^{-\tan(\xi)h\omega_i}$ with $\xi \to \frac{\pi}{2}$. With the Laurent expansion for $\tan\xi$ at $\frac{\pi}{2}$, i.e.,

$$\tan\xi = -\frac{1}{\xi - \frac{\pi}{2}} + \frac{1}{3}\left(\xi - \frac{\pi}{2}\right) + \frac{1}{45}\left(\xi - \frac{\pi}{2}\right)^3 + \mathcal{O}\left(\left(\xi - \frac{\pi}{2}\right)^5\right)$$

the coefficient has the approximating formula

$$e^{-\tan(\xi)h\omega_i} \approx e^{\frac{h\omega_i}{\xi - \frac{\pi}{2}}} \leq \left(\frac{\pi}{2} - \xi\right)^n.$$

where n is an arbitrary large real number in $\mathbb{R}^+$ for $\xi \to \frac{\pi}{2}$.

4.5 Experimental Demonstration

In this section, we implement the artificially dissipating energy algorithm (algorithm 1), energy conservation algorithm (algorithm 2) and the combined algorithm (algorithm 3) into high-dimension data for comparison with gradient method, momentum method and Nesterov accelerated gradient method.

4.5.1 Strongly Convex Function

Here, we investigate the artificially dissipating energy algorithm (algorithm 1) for the strongly convex function for comparison with gradient method, momentum method and Nesterov accelerated gradient method (strongly convex case) by the quadratic function as below.

$$f(x) = \frac{1}{2}x^T Ax + b^T x \tag{4.28}$$

where A is symmetric and positive-definite matrix. The two cases are shown as below:

(a) The generate matrix A is 500×500 random positive define matrix with eigenvalue from $1e-6$ to 1 with one defined eigenvalue $1e-6$. The generate vector b follows i.i.d. Gaussian distribution with mean 0 and variance 1.

(b) The generate matrix A is the notorious example in Nesterov's book [Nes13], i.e.,

$$A = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

the eigenvalues of the matrix are

$$\lambda_k = 2 - 2\cos\left(\frac{k\pi}{n+1}\right) = 4\sin^2\left(\frac{k\pi}{2(n+1)}\right)$$

and n is the dimension of the matrix A. The eigenvector can be solved by the second Chebyshev's polynomial. We implement $\dim(A) = 1000$ and b is zero vector. Hence, the smallest eigenvalue is approximating

$$\lambda_1 = 4\sin^2\left(\frac{\pi}{2(n+1)}\right) \approx \frac{\pi^2}{1001^2} \approx 10^{-5}$$

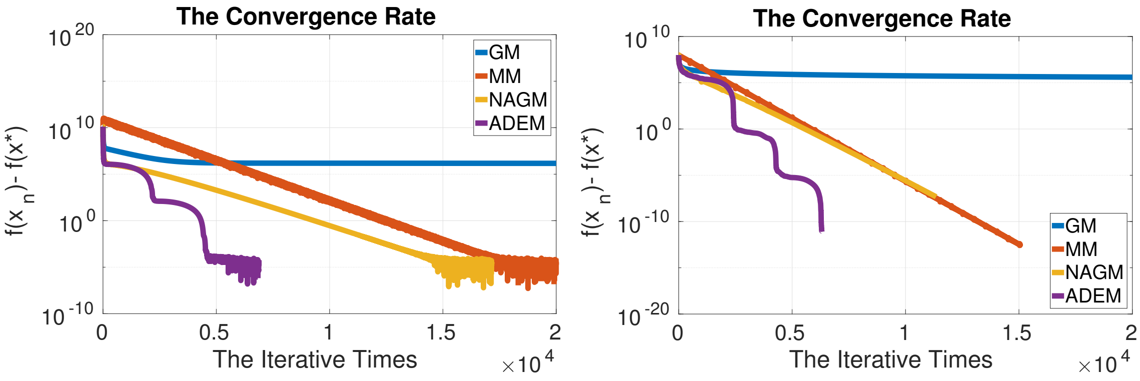


Figure 4.7: The Left: the case (**a**) with the initial point $x_0 = 0$. The Right: the case (**b**) with the initial point $x_0 = 1000$

4.5.2 Non-Strongly Convex Function

Here, we investigate the artificially dissipating energy algorithm (algorithm 1) for the non-strongly convex function for comparison with gradient method, Nesterov accelerated gradient method (non-strongly convex case) by the log-sum-exp function as below.

$$f(x) = \rho \log \left[\sum_{i=1}^{n} \exp\left(\frac{\langle a_i, x\rangle - b_i}{\rho} \right) \right] \tag{4.29}$$

where A is the $m \times n$ matrix with a_i, $(i = 1, \ldots, m)$ the column vector of A and b is the $n \times 1$ vector with component b_i. ρ is the parameter. We show the experiment in (4.29): the matrix $A = (a_{ij})_{m\times n}$ and the vector $b = (b_i)_{n\times 1}$ are set by the entry following i.i.d Gaussian distribution for the paramter $\rho = 5$ and $\rho = 10$.

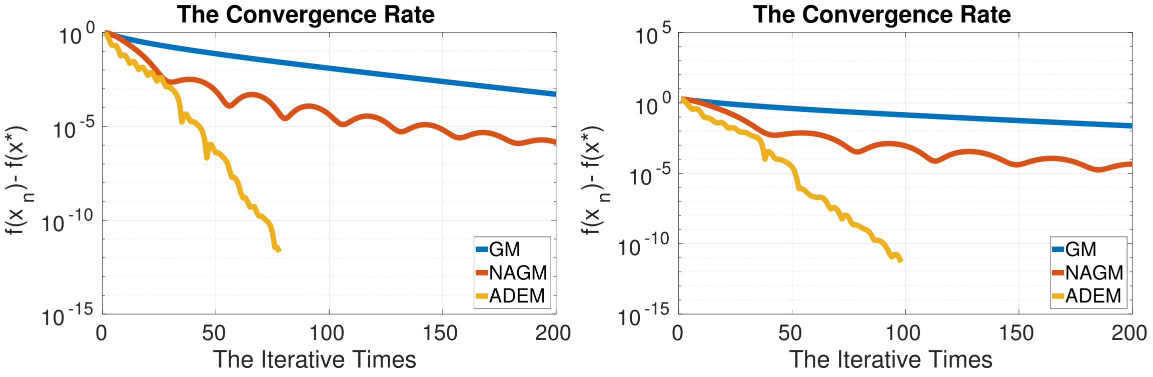


Figure 4.8: The convergence rate is shown from the initial point $x_0 = 0$. The Left: $\rho = 5$; The Right: $\rho = 10$.

4.5.3 Non-convex Function

For the nonconvex function, we exploit classical test function, known as artificial landscape, to evaluate characteristics of optimization algorithms from general performance and precision. In this paper, we show our algorithms implementing on the Styblinski-Tang function and Shekel function, which is recorded in the virtual library of simulation experiments[1]. Firstly, we investigate Styblinski-Tang function, i.e.

$$f(x) = \frac{1}{2}\sum_{i=1}^{d}\left(x_i^4 - 16x_i^2 + 5x_i\right) \tag{4.30}$$

to demonstrate the general performance of the algorithm 2 to track the number of local minima and then find the local minima by algorithm 3.

[1] https://www.sfu.ca/ ssurjano/index.html

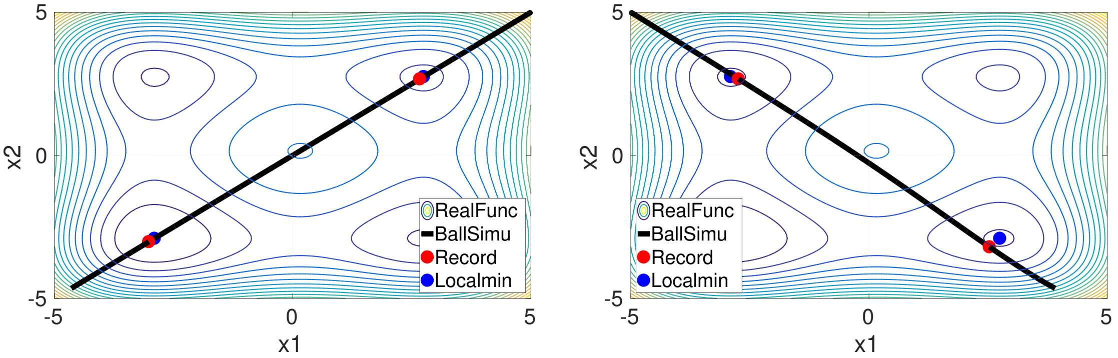


Figure 4.9: Detecting the number of the local minima of 2-D Styblinski-Tang function by algorithm 3 with step length $h = 0.01$. The red points are recorded by algorithm 2 and the blue point are the local minima by algorithm 1. The Left: The Initial Position $(5, 5)$; The Right: The Initial Position $(-5, 5)$.

To the essential 1-D nonconvex Styblinski-Tang function of high dimension, we implement the algorithm 3 to obtain the precision of the global minima as below.

	Local_min1	Local_min2	Local_min3	Local_min4
Initial Position	(5,5,...)	(5,5,...)	(5,-5,...)	(5,-5,...)
Position	(2.7486,2.7486,...)	(-2.9035,-2.9035,...)	(2.7486,-2.9035,...)	(-2.9035,2.7486,...)
Function Value	-250.2945	-391.6617	-320.9781	-320.9781

Table 4.1: The example for ten-dimensional Styblinski-Tang function from two initial positions.

The global minima calculated at the position $(-2.9035, -2.9035, \ldots)$ is -391.6617 shown on the Table 4.1. And the real global minima at $(-2.903534, -2.903534, \ldots)$ is $-39.16599 \times 10 = -391.6599$.

Furthermore, we demonstrate the numerical experiment from Styblinski-Tang function to more complex Shekel function

$$f(x) = -\sum_{i=1}^{m} \left(\sum_{j=1}^{4} (x_j - C_{ji})^2 + \beta_i \right)^{-1} \tag{4.31}$$

where

$$\beta = \frac{1}{10} (1, 2, 2, 4, 4, 6, 3, 7, 5, 5)^T$$

and

$$C = \begin{pmatrix} 4.0 & 1.0 & 8.0 & 6.0 & 3.0 & 2.0 & 5.0 & 8.0 & 6.0 & 7.0 \\ 4.0 & 1.0 & 8.0 & 6.0 & 7.0 & 9.0 & 3.0 & 1.0 & 2.0 & 3.6 \\ 4.0 & 1.0 & 8.0 & 6.0 & 3.0 & 2.0 & 5.0 & 8.0 & 6.0 & 7.0 \\ 4.0 & 1.0 & 8.0 & 6.0 & 7.0 & 9.0 & 3.0 & 1.0 & 2.0 & 3.6 \end{pmatrix}.$$

(1) Case $m = 5$, the global minima at $x^\star = (4, 4, 4, 4)$ is $f(x^\star) = -10.1532$.

(a) From the position $(10, 10, 10, 10)$, the experimental result with the step length $h = 0.01$ and the iterative times 3000 is shown as below

Detect Position (Algorithm 2)

$$\begin{pmatrix} 7.9879 & 6.0136 & 3.8525 & 6.2914 & 2.7818 \\ 7.9958 & 5.9553 & 3.9196 & 6.2432 & 6.7434 \\ 7.9879 & 6.0136 & 3.8525 & 6.2914 & 2.7818 \\ 7.9958 & 5.9553 & 3.9196 & 6.2432 & 6.7434 \end{pmatrix}$$

Detect value

$$\begin{pmatrix} -5.0932 & -2.6551 & -6.5387 & -1.6356 & -1.7262 \end{pmatrix}$$

Final position (Algorithm 1)

$$\begin{pmatrix} 7.9996 & 5.9987 & 4.0000 & 5.9987 & 3.0018 \\ 7.9996 & 6.0003 & 4.0001 & 6.0003 & 6.9983 \\ 7.9996 & 5.9987 & 4.0000 & 5.9987 & 3.0018 \\ 7.9996 & 6.0003 & 4.0001 & 6.0003 & 6.9983 \end{pmatrix}$$

Final value

$$\begin{pmatrix} -5.1008 & -2.6829 & -10.1532 & -2.6829 & -2.6305 \end{pmatrix}$$

(b) From the position $(3, 3, 3, 3)$, the experimental result with the step length $h = 0.01$ and the iterative times 1000 is shown as below

Detect Position (Algorithm 2)

$$\begin{pmatrix} 3.9957 & 6.0140 \\ 4.0052 & 6.0068 \\ 3.9957 & 6.0140 \\ 4.0052 & 6.0068 \end{pmatrix}$$

Detect value

$$\begin{pmatrix} -10.1443 & -2.6794 \end{pmatrix}$$

Final position (Algorithm 1)

$$\begin{pmatrix} 4.0000 & 5.9987 \\ 4.0001 & 6.0003 \\ 4.0000 & 5.9987 \\ 4.0001 & 6.0003 \end{pmatrix}$$

Final value

$$\begin{pmatrix} -10.1532 & -2.6829 \end{pmatrix}$$

(2) Case $m = 7$, the global minima at $x^\star = (4, 4, 4, 4)$ is $f(x^\star) = -10.4029$.

(a) From the position $(10, 10, 10, 10)$, the experimental result with the step length $h = 0.01$ and the iterative times 3000 is shown as below

Detect Position (Algorithm 2)

$$\begin{pmatrix} 7.9879 & 6.0372 & 3.1798 & 5.0430 & 6.2216 & 2.6956 \\ 8.0041 & 5.9065 & 3.8330 & 2.8743 & 6.2453 & 6.6837 \\ 7.9879 & 6.0372 & 3.1798 & 5.0430 & 6.2216 & 2.6956 \\ 8.0041 & 5.9065 & 3.8330 & 2.8743 & 6.2453 & 6.6837 \end{pmatrix}$$

Detect value

$$\begin{pmatrix} -5.1211 & -2.6312 & -0.9428 & -3.3093 & -1.8597 & -1.5108 \end{pmatrix}$$

Final position (Algorithm 1)

$$\begin{pmatrix} 7.9995 & 5.9981 & 4.0006 & 4.9945 & 5.9981 & 3.0006 \\ 7.9996 & 5.9993 & 3.9996 & 3.0064 & 5.9993 & 7.0008 \\ 7.9995 & 5.9981 & 4.0006 & 4.9945 & 5.9981 & 3.0006 \\ 7.9996 & 5.9993 & 3.9996 & 3.0064 & 5.9993 & 7.0008 \end{pmatrix}$$

Final value

$$\begin{pmatrix} -5.1288 & -2.7519 & -10.4029 & -3.7031 & -2.7519 & -2.7496 \end{pmatrix}$$

(b) From the position $(3, 3, 3, 3)$, the experimental result with the step length $h = 0.01$ and the iterative times 1000 is shown as below

Detect Position (Algorithm 2)

$$\begin{pmatrix} 4.0593 & 3.0228 \\ 3.9976 & 7.1782 \\ 4.0593 & 3.0228 \\ 3.9976 & 7.1782 \end{pmatrix}$$

Detect value

$$\begin{pmatrix} -9.7595 & -2.4073 \end{pmatrix}$$

Final position (Algorithm 1)

$$\begin{pmatrix} 4.0006 & 3.0006 \\ 3.9996 & 7.0008 \\ 4.0006 & 3.0006 \\ 3.9996 & 7.0008 \end{pmatrix}$$

Final value

$$\begin{pmatrix} -10.4029 & -2.7496 \end{pmatrix}$$

(3) Case $m = 10$, the global minima at $x^\star = (4, 4, 4, 4)$ is $f(x^\star) = -10.5364$.

(a) From the position $(10, 10, 10, 10)$, the experimental result with the step length $h = 0.01$ and the iterative times 3000 is shown as below

Detect Position (Algorithm 2)

$$\begin{pmatrix} 7.9977 & 5.9827 & 4.0225 & 2.7268 & 6.1849 & 6.2831 & 6.3929 \\ 7.9942 & 6.0007 & 3.8676 & 7.3588 & 6.0601 & 3.2421 & 1.9394 \\ 7.9977 & 5.9827 & 4.0225 & 2.7268 & 6.1849 & 6.2831 & 6.3929 \\ 7.9942 & 6.0007 & 3.8676 & 7.3588 & 6.0601 & 3.2421 & 1.9394 \end{pmatrix}$$

Detect value

$$\begin{pmatrix} -5.1741 & -2.8676 & -7.9230 & -1.5442 & -2.4650 & -1.3703 & -1.7895 \end{pmatrix}$$

Final position (Algorithm 1)

$$\begin{pmatrix} 7.9995 & 5.9990 & 4.0007 & 3.0009 & 5.9990 & 6.8999 & 5.9919 \\ 7.9994 & 5.9965 & 3.9995 & 7.0004 & 5.9965 & 3.4916 & 2.0224 \\ 7.9995 & 5.9990 & 4.0007 & 3.0009 & 5.9990 & 6.8999 & 5.9919 \\ 7.9994 & 5.9965 & 3.9995 & 7.0004 & 5.9965 & 3.4916 & 2.0224 \end{pmatrix}$$

Final value

$$\begin{pmatrix} -5.1756 & -2.8712 & -10.5364 & -2.7903 & -2.8712 & -2.3697 & -2.6085 \end{pmatrix}$$

(b) From the position $(3, 3, 3, 3)$, the experimental result with the step length $h = 0.01$ and the iterative times 1000 is shown as below

Detect Position (Algorithm 2)

$$\begin{pmatrix} 4.0812 & 3.0206 \\ 3.9794 & 7.0173 \\ 4.0812 & 3.0206 \\ 3.9794 & 7.0173 \end{pmatrix}$$

Detect value

$$\begin{pmatrix} -9.3348 & -2.7819 \end{pmatrix}$$

Final position (Algorithm 1)

$$\begin{pmatrix} 4.0007 & 3.0009 \\ 3.9995 & 7.0004 \\ 4.0007 & 3.0009 \\ 3.9995 & 7.0004 \end{pmatrix}$$

Final value

$$\begin{pmatrix} -10.5364 & -2.7903 \end{pmatrix}$$

CHAPTER 5

UNDERSTANDING THE ACCELERATION PHENOMENON VIA HIGH-RESOLUTION DIFFERENTIAL EQUATIONS

5.1 Introduction

Machine learning has become one of the major application areas for optimization algorithms during the past decade. While there have been many kinds of applications, to a wide variety of problems, the most prominent applications have involved large-scale problems in which the objective function is the sum over terms associated with individual data, such that stochastic gradients can be computed cheaply, while gradients are much more expensive and the computation (and/or storage) of Hessians is often infeasible. In this setting, simple first-order gradient descent algorithms have become dominant, and the effort to make these algorithms applicable to a broad range of machine learning problems has triggered a flurry of new research in optimization, both methodological and theoretical.

We will be considering unconstrained minimization problems,

$$\min_{x \in \mathbb{R}^n} \quad f(x), \tag{5.1}$$

where f is a smooth convex function. Perhaps the simplest first-order method for solving this problem is gradient descent. Taking a fixed step size s, gradient descent is implemented as the recursive rule

$$x_{k+1} = x_k - s\nabla f(x_k),$$

given an initial point x_0.

As has been known at least since the advent of conjugate gradient algorithms, improvements to gradient descent can be obtained within a first-order framework by using the history of past gradients. Modern research on such extended first-order methods arguably dates to Polyak [Pol64, Pol87a], whose *heavy-ball method* incorporates a momentum term into the gradient step. This approach allows past gradients to influence the current step, while avoiding the complexities of conjugate gradients and permitting a stronger theoretical analysis. Explicitly, starting from an initial point $x_0,\ x_1 \in \mathbb{R}^n$, the heavy-ball method updates the iterates according to

$$x_{k+1} = x_k + \alpha\left(x_k - x_{k-1}\right) - s\nabla f(x_k), \tag{5.2}$$

where $\alpha > 0$ is the momentum coefficient. While the heavy-ball method provably attains a faster rate of *local* convergence than gradient descent near a minimum of f, it does not come with *global* guarantees. Indeed, [LRP16] demonstrate that even for strongly convex functions the method can fail to converge for some choices of the step size.[1]

The next major development in first-order methodology was due to Nesterov, who discovered a class of *accelerated gradient methods* that have a faster global convergence rate than gradient descent [Nes83, Nes13]. For a μ-strongly convex objective f with L-Lipschitz gradients, Nesterov's accelerated gradient method (NAG-`SC`) involves the following pair of update equations:

$$\begin{aligned} y_{k+1} &= x_k - s\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\left(y_{k+1} - y_k\right), \end{aligned} \tag{5.3}$$

given an initial point $x_0 = y_0 \in \mathbb{R}^n$. Equivalently, NAG-`SC` can be written in a single-variable form that is similar to the heavy-ball method:

$$x_{k+1} = x_k + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\left(x_k - x_{k-1}\right) - s\nabla f(x_k) - \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\cdot s\left(\nabla f(x_k) - \nabla f(x_{k-1})\right), \tag{5.4}$$

[1] [Pol64] considers $s = 4/(\sqrt{L}+\sqrt{\mu})^2$ and $\alpha = (1-\sqrt{\mu s})^2$. This momentum coefficient is basically the same as the choice $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$ (adopted starting from Section 5.1.1) if s is small.

starting from x_0 and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1+\sqrt{\mu s}}$. Like the heavy-ball method, NAG-SC blends gradient and momentum contributions into its update direction, but defines a specific momentum coefficient $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$. Nesterov also developed the *estimate sequence technique* to prove that NAG-SC achieves an accelerated linear convergence rate:

$$f(x_k) - f(x^\star) \le O\left((1-\sqrt{s\mu})^k\right),$$

if the step size satisfies $0 < s \le 1/L$. Moreover, for a (weakly) convex objective f with L-Lipschitz gradients, Nesterov defined a related accelerated gradient method (NAG-C), that takes the following form:

$$\begin{aligned} y_{k+1} &= x_k - s\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{k}{k+3}(y_{k+1} - y_k), \end{aligned} \tag{5.5}$$

with $x_0 = y_0 \in \mathbb{R}^n$. The choice of momentum coefficient $\frac{k}{k+3}$, which tends to one, is fundamental to the estimate-sequence-based argument used by Nesterov to establish the following inverse quadratic convergence rate:

$$f(x_k) - f(x^\star) \le O\left(\frac{1}{sk^2}\right), \tag{5.6}$$

for any step size $s \le 1/L$. Under an oracle model of optimization complexity, the convergence rates achieved by NAG-SC and NAG-C are *optimal* for smooth strongly convex functions and smooth convex functions, respectively [NY83].

5.1.1 Gradient Correction: Small but Essential

Throughout the present paper, we let $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$ and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1+\sqrt{\mu s}}$ to define a specific implementation of the heavy-ball method in (5.2). This choice of the momentum coefficient and the second initial point renders the heavy-ball method and NAG-SC identical except for the last (small) term in (5.4). Despite their close resemblance, however, the two methods are in fact fundamentally different, with contrasting convergence results (see, for example, [B+15]). Notably, the former algorithm in general only achieves *local* acceleration, while the latter achieves acceleration method for all initial values of the iterate [LRP16]. As a numerical illustration, Figure 5.1 presents the trajectories that arise from the two methods when minimizing an ill-conditioned convex quadratic function. We see that the heavy-ball method exhibits pronounced oscillations throughout the iterations, whereas NAG-`SC` is monotone in the function value once the iteration counter exceeds 50.

This striking difference between the two methods can *only* be attributed to the last term in (5.4):

$$\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \cdot s\left(\nabla f(x_k) - \nabla f(x_{k-1})\right), \tag{5.7}$$

which we refer to henceforth as the *gradient correction*[2]. This term corrects the update direction in NAG-`SC` by contrasting the gradients at consecutive iterates. Although an essential ingredient in NAG-`SC`, the effect of the gradient correction is unclear from the vantage point of the estimate-sequence technique used in Nesterov's proof. Accordingly, while the estimate-sequence technique delivers a proof of acceleration for NAG-`SC`, it does not explain why the absence of the gradient correction prevents the heavy-ball method from achieving acceleration for strongly convex functions.

A recent line of research has taken a different point of view on the theoretical analysis of acceleration, formulating the problem in continuous time and obtaining algorithms via discretization [SBC14, KBB15, WWJ16]). This can be done by taking continuous-time limits of existing algorithms to obtain ordinary differential equations (ODEs) that can be analyzed using the rich toolbox associated with ODEs, including

[2] The gradient correction for NAG-`C` is $\frac{k}{k+3} \cdot s(\nabla f(x_k) - \nabla f(x_{k-1}))$, as seen from the single-variable form of NAG-`C`: $x_{k+1} = x_k + \frac{k}{k+3}(x_k - x_{k-1}) - s\nabla f(x_k) - \frac{k}{k+3} \cdot s(\nabla f(x_k) - \nabla f(x_{k-1}))$.

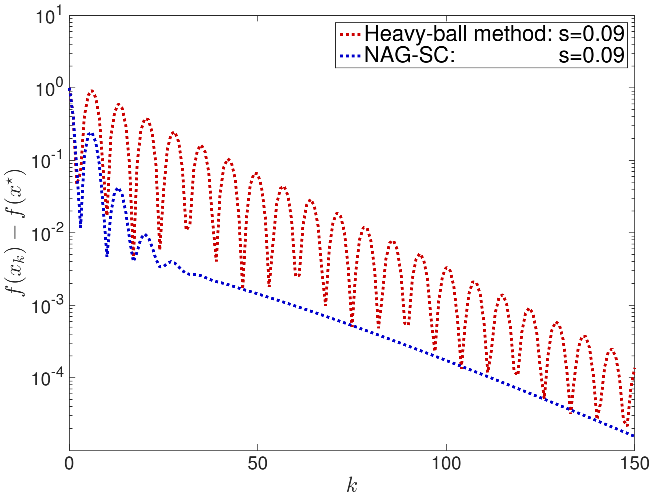


Figure 5.1: A numerical comparison between NAG-SC and heavy-ball method. The objective function (ill-conditioned $\mu/L \ll 1$) is $f(x_1, x_2) = 5 \times 10^{-3}x_1^2 + x_2^2$, with the initial iterate $(1, 1)$.

Lyapunov functions[3]. For instance, [SBC16] shows that

$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla f(X(t)) = 0, \tag{5.8}$$

with initial conditions $X(0) = x_0$ and $\dot{X}(0) = 0$, is the exact limit of NAG-C (5.5) by taking the step size $s \to 0$. Alternatively, the starting point may be a Lagrangian or Hamiltonian framework [WWJ16]. In either case, the continuous-time perspective not only provides analytical power and intuition, but it also provides design tools for new accelerated algorithms.

Unfortunately, existing continuous-time formulations of acceleration stop short of differentiating between the heavy-ball method and NAG-SC. In particular, these two methods have the *same* limiting ODE (see, for example, [WRJ16]):

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \nabla f(X(t)) = 0, \tag{5.9}$$

[3] One can think of the Lyapunov function as a generalization of the idea of the energy of a system. Then the method studies stability by looking at the rate of change of this measure of energy.

and, as a consequence, this ODE does not provide any insight into the stronger convergence results for NAG-`SC` as compared to the heavy-ball method. As will be shown in Section 5.2, this is because the gradient correction $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} s \left(\nabla f(x_k) - \nabla f(x_{k-1})\right) = O(s^{1.5})$ is an order-of-magnitude smaller than the other terms in (5.4) if $s = o(1)$. Consequently, the gradient correction is *not* reflected in the *low-resolution* ODE (5.9) associated with NAG-`SC`, which is derived by simply taking $s \to 0$ in both (5.2) and (5.4).

5.1.2 Overview of Contributions

Just as there is not a singled preferred way to discretize a differential equation, there is not a single preferred way to take a continuous-time limit of a difference equation. Inspired by dimensional-analysis strategies widely used in fluid mechanics in which physical phenomena are investigated at multiple scales via the inclusion of various orders of perturbations [Ped13], we propose to incorporate $O(\sqrt{s})$ terms into the limiting process for obtaining an ODE, including the (Hessian-driven) gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ in (5.7). This will yield *high-resolution ODEs* that differentiate between the NAG methods and the heavy-ball method.

We list the high-resolution ODEs that we derive in the paper here[4]:

(a) The high-resolution ODE for the heavy-ball method (5.2):

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0, \tag{5.10}$$

with $X(0) = x_0$ and $\dot{X}(0) = -\frac{2\sqrt{s}\nabla f(x_0)}{1+\sqrt{\mu s}}$.

[4] We note that the form of the initial conditions is fixed for each ODE throughout the paper. For example, while x_0 is arbitrary, $X(0)$ and $\dot{X}(0)$ must always be equal to x_0 and $-2\sqrt{s}f(x_0)/(1+\sqrt{\mu s})$ respectively in the high-resolution ODE of the heavy-ball method. This is in accordance with the choice of $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$ and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1+\sqrt{\mu s}}$.

(b) The high-resolution ODE for NAG-SC (5.3):

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s})\,\nabla f(X(t)) = 0, \tag{5.11}$$

with $X(0) = x_0$ and $\dot{X}(0) = -\frac{2\sqrt{s}\nabla f(x_0)}{1+\sqrt{\mu s}}$.

(c) The high-resolution ODE for NAG-C (5.5):

$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + \left(1 + \frac{3\sqrt{s}}{2t}\right)\nabla f(X(t)) = 0 \tag{5.12}$$

for $t \geq 3\sqrt{s}/2$, with $X(3\sqrt{s}/2) = x_0$ and $\dot{X}(3\sqrt{s}/2) = -\sqrt{s}\nabla f(x_0)$.

High-resolution ODEs are more accurate continuous-time counterparts for the corresponding discrete algorithms than low-resolution ODEs, thus allowing for a better characterization of the accelerated methods. This is illustrated in Figure 5.2, which presents trajectories and convergence of the discrete methods, and the low- and high-resolution ODEs. For both NAGs, the high-resolution ODEs are in much better agreement with the discrete methods than the low-resolution ODEs[5]. Moreover, for NAG-SC, its high-resolution ODE captures the non-oscillation pattern while the low-resolution ODE does not.

The three new ODEs include $O(\sqrt{s})$ terms that are not present in the corresponding low-resolution ODEs (compare, for example, (5.12) and (5.8)). Note also that if we let $s \to 0$, each high-resolution ODE reduces to its low-resolution counterpart. Thus, the difference between the heavy-ball method and NAG-SC is reflected only in their high-resolution ODEs: the gradient correction (5.7) of NAG-SC is preserved only in its high-resolution ODE in the form $\sqrt{s}\nabla^2 f(X(t))\dot{X}(t)$. This term, which we refer to as the (Hessian-driven) gradient correction, is connected with the discrete gradient correction by the approximate identity:

$$\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \cdot s\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) \approx s\nabla^2 f(x_k)(x_k - x_{k-1}) \approx s^{\frac{3}{2}}\nabla^2 f(X(t))\dot{X}(t)$$

[5] Note that for the heavy-ball method, the trajectories of the high-resolution ODE and the low-resolution ODE are almost identical.

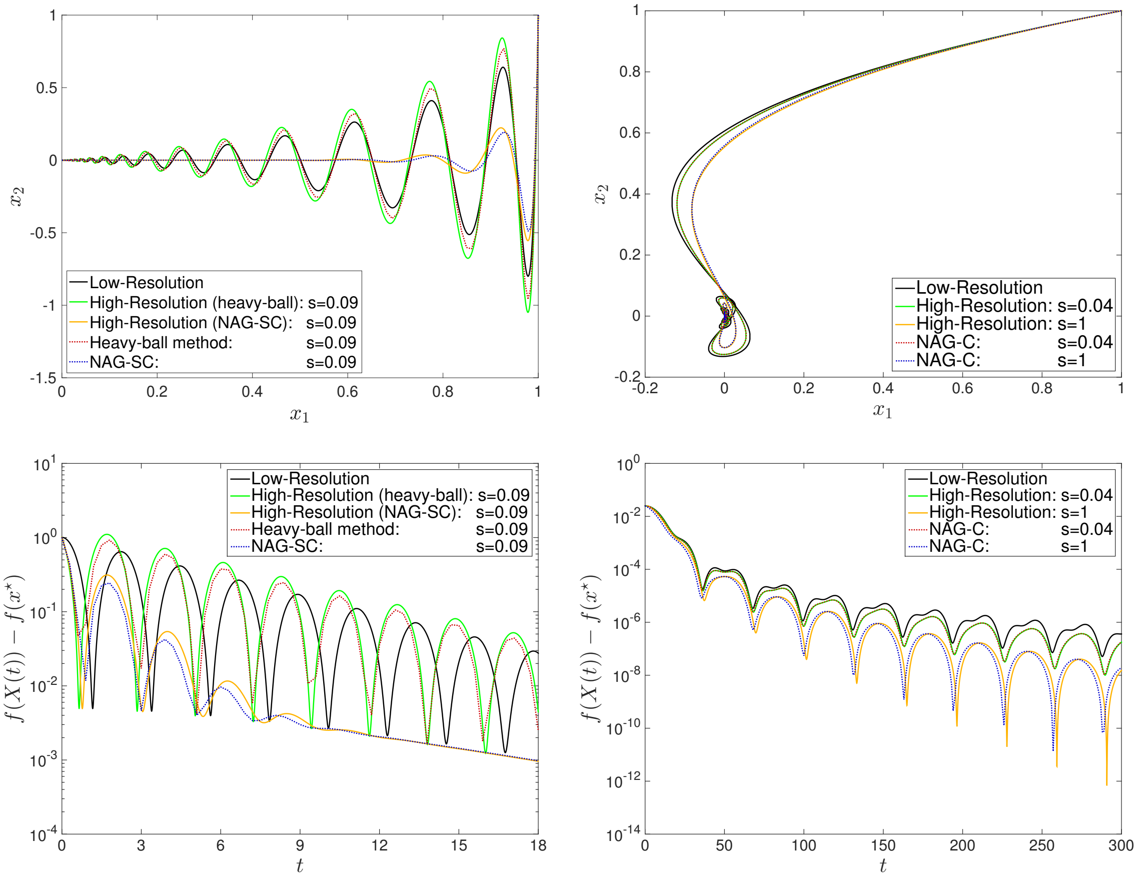


Figure 5.2: Top left and bottom left: trajectories and errors of NAG-SC and the heavy-ball method for minimizing $f(x_1, x_2) = 5 \times 10^{-3} x_1^2 + x_2^2$, from the initial value $(1, 1)$, the same setting as Figure 5.1. Top right and bottom right: trajectories and errors of NAG-C for minimizing $f(x_1, x_2) = 2 \times 10^{-2} x_1^2 + 5 \times 10^{-3} x_2^2$, from the initial value $(1, 1)$. For the two bottom plots, we use the identification $t = k\sqrt{s}$ between time and iterations for the x-axis.

for small s, with the identification $t = k\sqrt{s}$. The gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ in NAG-C arises in the same fashion[6]. Interestingly, although both NAGs are first-order methods, their gradient corrections brings in second-order information from the objective function.

[6] Henceforth, the dependence of X on t is suppressed when clear from the context.

Despite being small, the gradient correction has a fundamental effect on the behavior of both NAGs, and this effect is revealed by inspection of the high-resolution ODEs. We provide two illustrations of this.

- **Effect of the gradient correction in acceleration.** Viewing the coefficient of $\dot{X}$ as a damping ratio, the ratio $2\sqrt{\mu}+\sqrt{s}\nabla^2 f(X)$ of $\dot{X}$ in the high-resolution ODE (5.11) of NAG-`SC` is *adaptive* to the position X, in contrast to the *fixed* damping ratio $2\sqrt{\mu}$ in the ODE (5.10) for the heavy-ball method. To appreciate the effect of this adaptivity, imagine that the velocity $\dot{X}$ is highly correlated with an eigenvector of $\nabla^2 f(X)$ with a large eigenvalue, such that the large friction $(2\sqrt{\mu}+\sqrt{s}\nabla^2 f(X))\dot{X}$ effectively "decelerates" along the trajectory of the ODE (5.11) of NAG-`SC`. This feature of NAG-`SC` is appealing as taking a cautious step in the presence of high curvature generally helps avoid oscillations. Figure 5.1 and the left plot of Figure 5.2 confirm the superiority of NAG-`SC` over the heavy-ball method in this respect.

  If we can translate this argument to the discrete case we can understand why NAG-`SC` achieves acceleration globally for strongly convex functions but the heavy-ball method does not. We will be able to make this translation by leveraging the high-resolution ODEs to construct discrete-time Lyapunov functions that allow maximal step sizes to be characterized for the NAG-`SC` and the heavy-ball method. The detailed analyses is given in Section 5.3.

- **Effect of gradient correction in gradient norm minimization.** We will also show how to exploit the high-resolution ODE of NAG-`C` to construct a continuous-time Lyapunov function to analyze convergence in the setting of a smooth convex objective with L-Lipschitz gradients. Interestingly, the time derivative of the Lyapunov function is not only negative, but it is smaller than $-O(\sqrt{s}t^2\|\nabla f(X)\|^2)$. This bound arises from the gradient correction and, indeed, it cannot be obtained from the Lyapunov function studied in the low-resolution case by [SBC16]. This finer characterization in the high-resolution case allows us to establish a new phenomenon:

$$\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2 \leq O\left(\frac{L^2}{k^3}\right).$$

That is, we discover that NAG-`C` achieves an inverse *cubic* rate for minimizing the squared gradient norm. By comparison, from (5.6) and the L-Lipschitz continuity of ∇f we can only show that $\|\nabla f(x_k)\|^2 \leq O\left(L^2/k^2\right)$. See Section 5.4 for further elaboration on this cubic rate for NAG-`C`.

As we will see, the high-resolution ODEs are based on a phase-space representation that provides a systematic framework for translating from continuous-time Lyapunov functions to discrete-time Lyapunov functions. In sharp contrast, the process for obtaining a discrete-time Lyapunov function for low-resolution ODEs presented by [SBC16] relies on "algebraic tricks" (see, for example, Theorem 6 of [SBC16]). On a related note, a Hessian-driven damping term also appears in ODEs for modeling Newton's method [AABR02, AMR12, APR16].

5.1.3 Related Work

There is a long history of using ODEs to analyze optimization methods [HM12, Sch00, Fio05]. Recently, the work of [SBC14, SBC16] has sparked a renewed interest in leveraging continuous dynamical systems to understand and design first-order methods and to provide more intuitive proofs for the discrete methods. Below is a rather incomplete review of recent work that uses continuous-time dynamical systems to study accelerated methods.

In the work of [WWJ16, WRJ16, BJW18], Lagrangian and Hamiltonian frameworks are used to generate a large class of continuous-time ODEs for a unified treatment of accelerated gradient-based methods. Indeed, [WWJ16] extend NAG-C to non-Euclidean settings, mirror descent and accelerated higher-order gradient methods, all from a single "Bregman Lagrangian." In [WRJ16], the connection between ODEs and discrete algorithms is further strengthened by establishing an equivalence between the estimate sequence technique and Lyapunov function techniques, allowing for a principled analysis of the discretization of continuous-time ODEs. Recent papers have considered symplectic [BJW18] and Runge–Kutta [ZMSJ18] schemes for discretization of the low-resolution ODEs.

An ODE-based analysis of mirror descent has been pursued in another line of work by [KBB15, KBB16, KB17], delivering new connections between acceleration and constrained optimization, averaging and stochastic mirror descent.

In addition to the perspective of continuous-time dynamical systems, there has also been work on the acceleration from a control-theoretic point of view [LRP16, HL17, FRMP18] and from a geometric point of view [BLS15, CML17]. See also [OC15, FB15, GL16, DO17, LMH18, DFR18] for a number of other recent contributions to the study of the acceleration phenomenon.

5.1.4 Organization and Notation

The remainder of the paper is organized as follows. In Section 5.2, we briefly introduce our high-resolution ODE-based analysis framework. This framework is used in Section 5.3 to study the heavy-ball method and NAG-SC for smooth strongly convex functions. In Section 5.4, we turn our focus to NAG-C for a general smooth convex objective. In Section 5.5 we derive some extensions of NAG-C. We conclude the paper in Section 5.6 with a list of future research directions. Most technical proofs are deferred to the Appendix.

We mostly follow the notation of [Nes13], with slight modifications tailored to the present paper. Let $\mathcal{F}_L^1(\mathbb{R}^n)$ be the class of L-smooth convex functions defined on $\mathbb{R}^n$; that is, $f \in \mathcal{F}_L^1$ if $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ for all $x, y \in \mathbb{R}^n$ and its gradient is L-Lipschitz continuous in the sense that

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|,$$

where $\|\cdot\|$ denotes the standard Euclidean norm and $L > 0$ is the Lipschitz constant. (Note that this implies that ∇f is also L'-Lipschitz for any $L' \geq L$.) The function class $\mathcal{F}_L^2(\mathbb{R}^n)$ is the subclass of $\mathcal{F}_L^1(\mathbb{R}^n)$ such that each f has a Lipschitz-continuous Hessian. For $p = 1, 2$, let $\mathcal{S}_{\mu,L}^p(\mathbb{R}^n)$ denote the subclass of $\mathcal{F}_L^p(\mathbb{R}^n)$ such that each member f is μ-strongly convex for some $0 < \mu \leq L$. That is, $f \in \mathcal{S}_{\mu,L}^p(\mathbb{R}^n)$ if $f \in \mathcal{F}_L^p(\mathbb{R}^n)$ and

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2,$$

for all $x, y \in \mathbb{R}^n$. Note that this is equivalent to the convexity of $f(x) - \frac{\mu}{2}\|x - x^\star\|^2$, where $x^\star$ denotes a minimizer of the objective f.

5.2 The High-Resolution ODE Framework

This section introduces a high-resolution ODE framework for analyzing gradient-based methods, with NAG-SC being a guiding example. Given a (discrete) optimization algorithm, the first step in this framework is to derive a high-resolution ODE using dimensional analysis, the next step is to construct a continuous-time Lyapunov function to analyze properties of the ODE, the third step is to derive a discrete-time Lyapunov function from its continuous counterpart and the last step is to translate properties of the ODE into that of the original algorithm. The overall framework is illustrated in Figure 5.3.

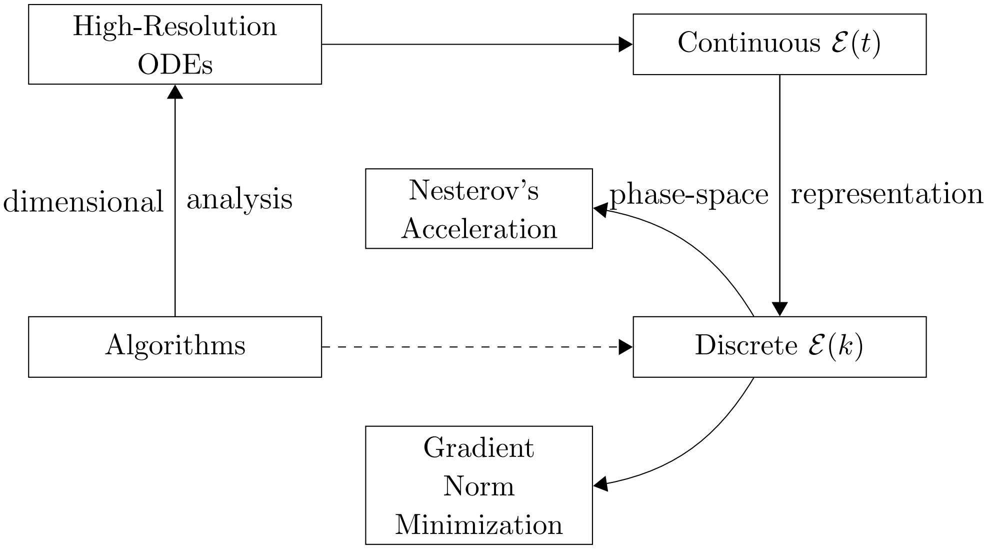


Figure 5.3: An illustration of our high-resolution ODE framework. The three solid straight lines represent Steps 1, 2 and 3, and the two curved lines denote Step 4. The dashed line is used to emphasize that it is difficult, if not impractical, to construct discrete Lyapunov functions directly from the algorithms.

Step 1: Deriving High-Resolution ODEs

Our focus is on the single-variable form (5.4) of NAG-SC. For any nonnegative integer k, let $t_k = k\sqrt{s}$ and assume $x_k = X(t_k)$ for some sufficiently smooth curve $X(t)$. Performing a Taylor expansion in powers of $\sqrt{s}$, we get

$$\begin{aligned} x_{k+1} = X(t_{k+1}) = X(t_k) + \dot{X}(t_k)\sqrt{s} + \frac{1}{2}\ddot{X}(t_k)\left(\sqrt{s}\right)^2 + \frac{1}{6}\dddot{X}(t_k)\left(\sqrt{s}\right)^3 + O\left(\left(\sqrt{s}\right)^4\right) \\ x_{k-1} = X(t_{k-1}) = X(t_k) - \dot{X}(t_k)\sqrt{s} + \frac{1}{2}\ddot{X}(t_k)\left(\sqrt{s}\right)^2 - \frac{1}{6}\dddot{X}(t_k)\left(\sqrt{s}\right)^3 + O\left(\left(\sqrt{s}\right)^4\right). \end{aligned} \tag{5.13}$$

We now use a Taylor expansion for the gradient correction, which gives

$$\nabla f(x_k) - \nabla f(x_{k-1}) = \nabla^2 f(X(t_k))\dot{X}(t_k)\sqrt{s} + O\left(\left(\sqrt{s}\right)^2\right). \tag{5.14}$$

Multiplying both sides of (5.4) by $\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{1}{s}$ and rearranging the equality, we can rewrite NAG-SC as

$$\frac{x_{k+1} + x_{k-1} - 2x_k}{s} + \frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{x_{k+1} - x_k}{s}$$

$$+ \nabla f(x_k) - \nabla f(x_{k-1}) + \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \nabla f(x_k) = 0. \quad (5.15)$$

Next, plugging (5.13) and (5.14) into (5.15), we have[7]

$$\ddot{X}(t_k) + O\left((\sqrt{s})^2 \right) + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}} \left[\dot{X}(t_k) + \frac{1}{2}\ddot{X}(t_k)\sqrt{s} + O\left((\sqrt{s})^2 \right) \right]$$
$$+ \nabla^2 f(X(t_k))\dot{X}(t_k)\sqrt{s} + O\left((\sqrt{s})^2 \right) + \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right) \nabla f(X(t_k)) = 0,$$

which can be rewritten as

$$\frac{\ddot{X}(t_k)}{1-\sqrt{\mu s}} + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}\dot{X}(t_k) + \sqrt{s}\nabla^2 f(X(t_k))\dot{X}(t_k) + \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \nabla f(X(t_k)) + O(s) = 0.$$

Multiplying both sides of the last display by $1-\sqrt{\mu s}$, we obtain the following high-resolution ODE of NAG-SC:

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + \sqrt{s}\nabla^2 f(X)\dot{X} + (1+\sqrt{\mu s})\nabla f(X) = 0,$$

where we ignore any $O(s)$ terms but retain the $O(\sqrt{s})$ terms (note that $(1-\sqrt{\mu s})\sqrt{s} = \sqrt{s} + O(s)$).

Our analysis is inspired by dimensional analysis [Ped13], a strategy widely used in physics to construct a series of differential equations that involve increasingly high-order terms corresponding to small perturbations. In more detail, taking a small s, one first derives a differential equation that consists only of $O(1)$ terms, then derives a differential equation consisting of both $O(1)$ and $O(\sqrt{s})$, and next, one proceeds to obtain a differential equation consisting of $O(1), O(\sqrt{s})$ and $O(s)$ terms. High-order terms in powers of $\sqrt{s}$ are introduced sequentially until the main characteristics of the original algorithms have been extracted from the resulting approximating differential

[7] Note that we use the approximation $\frac{x_{k+1}+x_{k-1}-2x_k}{s} = \ddot{X}(t_k) + O(s)$, whereas [SBC16] relies on the low-accuracy Taylor expansion $\frac{x_{k+1}+x_{k-1}-2x_k}{s} = \ddot{X}(t_k) + o(1)$ in the derivation of the low-resolution ODE of NAG-C. We illustrate this derivation of the three low-resolution ODEs in Appendix 5.7.1.2; they can be compared to the high-resolution ODEs that we derive here.

equation. Thus, we aim to understand Nesterov acceleration by incorporating $O(\sqrt{s})$ terms into the ODE, including the (Hessian-driven) gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ which results from the (discrete) gradient correction (5.7) in the single-variable form (5.4) of NAG-`SC`. We also show (see Appendix 5.7.1.1 for the detailed derivation) that this $O(\sqrt{s})$ term appears in the high-resolution ODE of NAG-`C`, but is not found in the high-resolution ODE of the heavy-ball method.

As shown below, each ODE admits a unique global solution under mild conditions on the objective, and this holds for an arbitrary step size $s > 0$. The solution is accurate in approximating its associated optimization method if s is small. To state the result, we use $C^2(I; \mathbb{R}^n)$ to denote the class of twice-continuously-differentiable maps from I to $\mathbb{R}^n$ for $I = [0, \infty)$ (the heavy-ball method and NAG-`SC`) and $I = [1.5\sqrt{s}, \infty)$ (NAG-`C`).

Proposition 5.2.1. For any $f \in \mathcal{S}^2_\mu(\mathbb{R}^n) := \cup_{L \geq \mu} \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$, each of the ODEs (5.10) and (5.11) with the specified initial conditions has a unique global solution $X \in C^2([0, \infty); \mathbb{R}^n)$. Moreover, the two methods converge to their high-resolution ODEs, respectively, in the sense that

$$\limsup_{s \to 0} \max_{0 \leq k \leq \frac{T}{\sqrt{s}}} \left\| x_k - X(k\sqrt{s}) \right\| = 0,$$

for any fixed $T > 0$.

In fact, Propositition 5.2.1 holds for $T = \infty$ because both the discrete iterates and the ODE trajectories converge to the unique minimizer when the objective is stongly convex.

Proposition 5.2.2. For any $f \in \mathcal{F}^2(\mathbb{R}^n) := \cup_{L > 0} \mathcal{F}^2_L(\mathbb{R}^n)$, the ODE (5.12) with the specified initial conditions has a unique global solution $X \in C^2([1.5\sqrt{s}, \infty); \mathbb{R}^n)$. Moreover, NAG-`C` converges to its high-resolution ODE in the sense that

$$\limsup_{s \to 0} \max_{0 \leq k \leq \frac{T}{\sqrt{s}}} \left\| x_k - X(k\sqrt{s} + 1.5\sqrt{s}) \right\| = 0,$$

for any fixed $T > 0$.

The proofs of these propositions are given in Appendix 5.7.1.3.1 and Appendix 5.7.1.3.2.

Step 2: Analyzing ODEs Using Lyapunov Functions

With these high-resolution ODEs in place, the next step is to construct Lyapunov functions for analyzing the dynamics of the corresponding ODEs, as is done in previous work [SBC16, WRJ16, LRP16]. For NAG-SC, we consider the Lyapunov function

$$\mathcal{E}(t) = (1+\sqrt{\mu s})\left(f(X) - f(x^\star)\right)+\frac{1}{4}\|\dot{X}\|^2+\frac{1}{4}\|\dot{X}+2\sqrt{\mu}(X-x^\star)+\sqrt{s}\nabla f(X)\|^2. \quad (5.16)$$

The first and second terms $(1 + \sqrt{\mu s})\left(f(X) - f(x^\star)\right)$ and $\frac{1}{4}\|\dot{X}\|^2$ can be regarded, respectively, as the potential energy and kinetic energy, and the last term is a mix. For the mixed term, it is interesting to note that the time derivative of $\dot{X} + 2\sqrt{\mu}(X - x^\star) + \sqrt{s}\nabla f(X)$ equals $-(1 + \sqrt{\mu s})\nabla f(X)$.

The differentiability of $\mathcal{E}(t)$ will allow us to investigate properties of the ODE (5.11) in a principled manner. For example, we will show that $\mathcal{E}(t)$ decreases exponentially along the trajectories of (5.11), recovering the accelerated linear convergence rate of NAG-SC. Furthermore, a comparison between the Lyapunov function of NAG-SC and that of the heavy-ball method will explain why the gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ yields acceleration in the former case. This is discussed in Section 5.3.1.

Step 3: Constructing Discrete Lyapunov Functions

Our framework make it possible to translate continuous Lyapunov functions into discrete Lyapunov functions via a phase-space representation (see, for example, [Arn13]). We illustrate the procedure in the case of NAG-SC. The first step is formulate explicit position and velocity updates:

$$\begin{aligned} &x_k - x_{k-1} = \sqrt{s} v_{k-1} \\ &v_k - v_{k-1} = -\frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}} v_k - \sqrt{s}(\nabla f(x_k) - \nabla f(x_{k-1})) - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \sqrt{s} \nabla f(x_k), \end{aligned} \tag{5.17}$$

where the velocity variable v_k is defined as:

$$v_k = \frac{x_{k+1} - x_k}{\sqrt{s}}.$$

The initial velocity is $v_0 = -\frac{2\sqrt{s}}{1+\sqrt{\mu s}} \nabla f(x_0)$. Interestingly, this phase-space representation has the flavor of symplectic discretization, in the sense that the update for $x_k - x_{k-1}$ is explicit (it only depends on the last iterate v_{k-1}) while the update for $v_k - v_{k-1}$ is implicit (it depends on the current iterates x_k and v_k)[8].

The representation (5.17) suggests translating the continuous-time Lyapunov function (5.16) into a discrete-time Lyapunov function of the following form:

$$\begin{aligned} \mathcal{E}(k) = &\underbrace{\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(f(x_k) - f(x^\star) \right)}_{\mathbf{I}} + \underbrace{\frac{1}{4} \left\| v_k \right\|^2}_{\mathbf{II}} \\ &+ \underbrace{\frac{1}{4} \left\| v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}} (x_{k+1} - x^\star) + \sqrt{s} \nabla f(x_k) \right\|^2}_{\mathbf{III}} \underbrace{- \frac{s \left\| \nabla f(x_k) \right\|^2}{2(1-\sqrt{\mu s})}}_{\textbf{a negative term}}, \end{aligned} \tag{5.18}$$

by replacing continuous terms (e.g., $\dot{X}$) by their discrete counterparts (e.g., v_k). Akin to the continuous (5.16), here **I**, **II**, and **III** correspond to potential energy, kinetic energy, and mixed energy, respectively, from a mechanical perspective. To better appreciate this translation, note that the factor $\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}$ in **I** results from the term $\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \sqrt{s} \nabla f(x_k)$ in (5.17). Likewise, $\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}$ in **III** is from the term $\frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}} v_k$ in (5.17). The need for the final (small) negative term is technical; we discuss it in Section 5.3.2.

[8] Although this suggestion is a heuristic one, it is also possible to rigorously derive a symplectic integrator of the high-resolution ODE of NAG-SC; this integrator has the form:

$$\begin{aligned} &x_k - x_{k-1} = \sqrt{s} v_{k-1} \\ &v_k - v_{k-1} = -2\sqrt{\mu s} v_k - s \nabla^2 f(x_k) v_k - (1 + \sqrt{\mu s}) \sqrt{s} \nabla f(x_k). \end{aligned}$$

Step 4: Analyzing Algorithms Using Discrete Lyapunov Functions

The last step is to map properties of high-resolution ODEs to corresponding properties of optimization methods. This step closely mimics Step 2 except that now the object is a discrete algorithm and the tool is a discrete Lyapunov function such as (5.18). Given that Step 2 has been performed, this translation is conceptually straightforward, albeit often calculation-intensive. For example, using the discrete Lyapunov function (5.18), we will recover the optimal linear rate of NAG-`SC` and gain insights into the fundamental effect of the gradient correction in accelerating NAG-`SC`. In addition, NAG-`C` is shown to minimize the squared gradient norm at an inverse cubic rate by a simple analysis of the decreasing rate of its discrete Lyapunov function.

5.3 Gradient Correction for Acceleration

In this section, we use our high-resolution ODE framework to analyze NAG-`SC` and the heavy-ball method. Section 5.3.1 focuses on the ODEs with an objective function $f \in \mathcal{S}_{\mu,L}^2(\mathbb{R}^n)$, and in Section 5.3.2 we extend the results to the discrete case for $f \in \mathcal{S}_{\mu,L}^1(\mathbb{R}^n)$. Finally, in Section 5.3.3 we offer a comparative study of NAG-`SC` and the heavy-ball method from a finite-difference viewpoint.

Throughout this section, the strategy is to analyze the two methods in parallel, thereby highlighting the differences between the two methods. In particular, the comparison will demonstrate the vital role of the gradient correction, namely $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \cdot s\left(\nabla f(x_k) - \nabla f(x_{k-1})\right)$ in the discrete case and $\sqrt{s}\nabla^2 f(X)\dot{X}$ in the ODE case, in making NAG-`SC` an accelerated method.

5.3.1 The ODE Case

The following theorem characterizes the convergence rate of the high-resolution ODE corresponding to NAG-SC.

Theorem 5.3.1 (Convergence of NAG-SC ODE)**.** Let $f \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$. For any step size $0 < s \leq 1/L$, the solution $X = X(t)$ of the high-resolution ODE (5.11) satisfies

$$f(X(t)) - f(x^\star) \leq \frac{2\left\|x_0 - x^\star\right\|^2}{s} \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}.$$

The theorem states that the functional value $f(X)$ tends to the minimum $f(x^\star)$ at a linear rate. By setting $s = 1/L$, we obtain $f(X) - f(x^\star) \leq 2L\left\|x_0 - x^\star\right\|^2 \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}$.

The proof of Theorem 5.3.1 is based on analyzing the Lyapunov function $\mathcal{E}(t)$ for the high-resolution ODE of NAG-SC. Recall that $\mathcal{E}(t)$ defined in (5.16) is

$$\mathcal{E}(t) = (1 + \sqrt{\mu s})\left(f(X) - f(x^\star)\right) + \frac{1}{4}\|\dot{X}\|^2 + \frac{1}{4}\|\dot{X} + 2\sqrt{\mu}(X - x^\star) + \sqrt{s}\nabla f(X)\|^2.$$

The next lemma states the key property we need from this Lyapunov function

Lemma 5.3.2 (Lyapunov function for NAG-SC ODE)**.** Let $f \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$. For any step size $s > 0$, and with $X = X(t)$ being the solution to the high-resolution ODE (5.11), the Lyapunov function (5.16) satisfies

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} \leq -\frac{\sqrt{\mu}}{4}\mathcal{E}(t) - \frac{\sqrt{s}}{2}\left[\|\nabla f(X(t))\|^2 + \dot{X}(t)^\top \nabla^2 f(X(t))\dot{X}(t)\right]. \tag{5.19}$$

The proof of this theorem relies on Lemma 5.3.2 through the inequality $\dot{\mathcal{E}}(t) \leq -\frac{\sqrt{\mu}}{4}\mathcal{E}(t)$. The term $\frac{\sqrt{s}}{2}(\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X)\dot{X}) \geq 0$ plays no role at the moment, but Section 5.3.2 will shed light on its profound effect in the discretization of the high-resolution ODE of NAG-SC.

Proof. [Proof of Theorem 5.3.1]

Lemma 5.3.2 implies $\dot{\mathcal{E}}(t) \leq -\frac{\sqrt{\mu}}{4}\mathcal{E}(t)$, which amounts to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{E}(t)\mathrm{e}^{\frac{\sqrt{\mu}t}{4}}\right) \leq 0.$$

By integrating out t, we get

$$\mathcal{E}(t) \le \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}\mathcal{E}(0). \tag{5.20}$$

Recognizing the initial conditions $X(0) = x_0$ and $\dot{X}(0) = -\frac{2\sqrt{s}\nabla f(x_0)}{1+\sqrt{\mu s}}$, we write (5.20) as

$$\begin{aligned} f(X) - f(x^\star) \le & \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}\left[f(x_0) - f(x^\star) + \frac{s}{\left(1+\sqrt{\mu s}\right)^3}\left\|\nabla f(x_0)\right\|^2 \right. \\ & \left. + \frac{1}{4(1+\sqrt{\mu s})}\left\| 2\sqrt{\mu}(x_0 - x^\star) - \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\cdot\sqrt{s}\nabla f(x_0)\right\|^2\right]. \end{aligned}$$

Since $f \in \mathcal{S}^2_{\mu,L}$, we have that $\|\nabla f(x_0)\| \le L\|x_0 - x^\star\|$ and $f(x_0) - f(x^\star) \le L\|x_0 - x^\star\|^2/2$. Together with the Cauchy–Schwarz inequality, the two inequalities yield

$$\begin{aligned} & f(X) - f(x^\star) \\ \le & \left[f(x_0) - f(x^\star) + \frac{2 + (1-\sqrt{\mu s})^2}{2(1+\sqrt{\mu s})^3}\cdot s\left\|\nabla f(x_0)\right\|^2 + \frac{2\mu}{1+\sqrt{\mu s}}\left\|x_0 - x^\star\right\|^2\right]\mathrm{e}^{-\frac{\sqrt{\mu}t}{4}} \\ \le & \left[\frac{L}{2} + \frac{3 - 2\sqrt{\mu s} + \mu s}{2(1+\sqrt{\mu s})^3}\cdot sL^2 + \frac{2\mu}{1+\sqrt{\mu s}}\right]\left\|x_0 - x^\star\right\|^2\mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}, \end{aligned}$$

which is valid for all $s > 0$. To simplify the coefficient of $\left\|x_0 - x^\star\right\|^2\mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}$, note that L can be replaced by $1/s$ in the analysis since $s \le 1/L$. It follows that

$$f(X(t)) - f(x^\star) \le \left[\frac{1}{2} + \frac{3 - 2\sqrt{\mu s} + \mu s}{2(1+\sqrt{\mu s})^3} + \frac{2\mu s}{1+\sqrt{\mu s}}\right]\frac{\left\|x_0 - x^\star\right\|^2\mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}}{s}.$$

Furthermore, a bit of analysis reveals that

$$\frac{1}{2} + \frac{3 - 2\sqrt{\mu s} + \mu s}{2(1+\sqrt{\mu s})^3} + \frac{2\mu s}{1+\sqrt{\mu s}} < 2,$$

since $\mu s \le \mu/L \le 1$, and this step completes the proof of Theorem 5.3.1. □

We now consider the heavy-ball method (5.2). Recall that the momentum coefficient α is set to $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$. The following theorem characterizes the rate of convergence of this method.

Theorem 5.3.3 (Convergence of heavy-ball ODE)**.** Let $f \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$. For any step size $0 < s \leq 1/L$, the solution $X = X(t)$ of the *high-resolution ODE* (5.10) satisfies

$$f(X(t)) - f(x^\star) \leq \frac{7\left\|x_0 - x^\star\right\|^2}{2s} \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}.$$

As in the case of NAG-SC, the proof of Theorem 5.3.3 is based on a Lyapunov function:

$$\mathcal{E}(t) = (1 + \sqrt{\mu s})\left(f(X) - f(x^\star)\right) + \frac{1}{4}\|\dot{X}\|^2 + \frac{1}{4}\|\dot{X} + 2\sqrt{\mu}(X - x^\star)\|^2, \tag{5.21}$$

which is the same as the Lyapunov function (5.16) for NAG-SC except for the lack of the $\sqrt{s}\nabla f(X)$ term. In particular, (5.16) and (5.21) are identical if $s = 0$. The following lemma considers the decay rate of (5.21).

Lemma 5.3.4 (Lyapunov function for the heavy-ball ODE)**.** Let $f \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$. For any step size $s > 0$, the Lyapunov function (5.21) for the high-resolution ODE (5.10) satisfies

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} \leq -\frac{\sqrt{\mu}}{4}\mathcal{E}(t).$$

The proof of Theorem 5.3.3 follows the same strategy as the proof of Theorem 5.3.1. In brief, Lemma 5.3.4 gives $\mathcal{E}(t) \leq \mathrm{e}^{-\sqrt{\mu}t/4}\mathcal{E}(0)$ by integrating over the time parameter t. Recognizing the initial conditions

$$X(0) = x_0, \quad \dot{X}(0) = -\frac{2\sqrt{s}\nabla f(x_0)}{1 + \sqrt{\mu s}}$$

in the high-resolution ODE of the heavy-ball method and using the L-smoothness of ∇f, Lemma 5.3.4 yields

$$f(X) - f(x^\star) \leq \left[\frac{1}{2} + \frac{3}{(1 + \sqrt{\mu s})^3} + \frac{2(\mu s)}{1 + \sqrt{\mu s}}\right] \frac{\left\|x_0 - x^\star\right\|^2 \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}}{s},$$

if the step size $s \leq 1/L$. Finally, since $0 < \mu s \leq \mu/L \leq 1$, the coefficient satisfies

$$\frac{1}{2} + \frac{3}{(1 + \sqrt{\mu s})^3} + \frac{2\mu s}{1 + \sqrt{\mu s}} < \frac{7}{2}.$$

The proofs of Lemma 5.3.2 and Lemma 5.3.4 share similar ideas. In view of this, we present only the proof of the former here, deferring the proof of Lemma 5.3.4 to Appendix 5.7.2.1.

Proof. [Proof of Lemma 5.3.2] Along trajectories of (5.11) the Lyapunov function (5.16) satisfies

$$\begin{aligned}\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} &= (1+\sqrt{\mu s})\langle \nabla f(X), \dot{X}\rangle + \frac{1}{2}\left\langle \dot{X}, -2\sqrt{\mu}\dot{X} - \sqrt{s}\nabla^2 f(X)\dot{X} - (1+\sqrt{\mu s})\nabla f(X)\right\rangle \\ &\quad + \frac{1}{2}\left\langle \dot{X} + 2\sqrt{\mu}\,(X - x^\star) + \sqrt{s}\nabla f(X), -(1+\sqrt{\mu s})\nabla f(X)\right\rangle \\ &= -\sqrt{\mu}\left(\|\dot{X}\|^2 + (1+\sqrt{\mu s})\left\langle \nabla f(X), X - x^\star\right\rangle + \frac{s}{2}\left\|\nabla f(X)\right\|^2\right) \\ &\quad - \frac{\sqrt{s}}{2}\left[\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X)\dot{X}\right] \\ &\le -\sqrt{\mu}\left(\|\dot{X}\|^2 + (1+\sqrt{\mu s})\left\langle \nabla f(X), X - x^\star\right\rangle + \frac{s}{2}\left\|\nabla f(X)\right\|^2\right).\end{aligned} \tag{5.22}$$

Furthermore, $\langle \nabla f(X), X - x^\star\rangle$ is greater than or equal to both $f(X) - f(x^\star) + \frac{\mu}{2}\|X - x^\star\|^2$ and $\mu\|X - x^\star\|^2$ due to the μ-strong convexity of f. This yields

$$\begin{aligned}&(1+\sqrt{\mu s})\left\langle \nabla f(X), X - x^\star\right\rangle \\ \ge& \frac{1+\sqrt{\mu s}}{2}\left\langle \nabla f(X), X - x^\star\right\rangle + \frac{1}{2}\left\langle \nabla f(X), X - x^\star\right\rangle \\ \ge& \frac{1+\sqrt{\mu s}}{2}\left[f(X) - f(x^\star) + \frac{\mu}{2}\|X - x^\star\|^2\right] + \frac{\mu}{2}\|X - x^\star\|^2 \\ \ge& \frac{1+\sqrt{\mu s}}{2}(f(X) - f(x^\star)) + \frac{3\mu}{4}\|X - x^\star\|^2,\end{aligned}$$

which together with (5.22) suggests that the time derivative of this Lyapunov function can be bounded as

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} \le -\sqrt{\mu}\left(\frac{1+\sqrt{\mu s}}{2}(f(X) - f(x^\star)) + \|\dot{X}\|^2 + \frac{3\mu}{4}\left\|X - x^\star\right\|^2 + \frac{s}{2}\left\|\nabla f(X)\right\|^2\right). \tag{5.23}$$

Next, the Cauchy–Schwarz inequality yields

$$\left\|2\sqrt{\mu}(X - x^\star) + \dot{X} + \sqrt{s}\nabla f(X)\right\|^2 \le 3\left(4\mu\left\|X - x^\star\right\|^2 + \|\dot{X}\|^2 + s\left\|\nabla f(X)\right\|^2\right),$$

from which it follows that

$$\mathcal{E}(t) \leq (1+\sqrt{\mu s})\left(f(X)-f(x^\star)\right)+\|\dot{X}\|^2+3\mu\left\|X-x^\star\right\|^2+\frac{3s}{4}\left\|\nabla f(X)\right\|^2. \quad (5.24)$$

Combining (5.23) and (5.24) completes the proof of the theorem.

□

Remark 5.3.5. The only inequality in (5.22) is due to the term $\frac{\sqrt{s}}{2}(\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X)\dot{X})$, which is discussed right after the statement of Lemma 5.3.2. This term results from the gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ in the NAG-$\texttt{SC}$ ODE. For comparison, this term does not appear in Lemma 5.3.4 in the case of the heavy-ball method as its ODE does not include the gradient correction and, accordingly, its Lyapunov function (5.21) is free of the $\sqrt{s}\nabla f(X)$ term.

5.3.2 The Discrete Case

This section carries over the results in Section 5.3.1 to the two discrete algorithms, namely NAG-$\texttt{SC}$ and the heavy-ball method. Here we consider an objective $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$ since second-order differentiability of f is not required in the two discrete methods. Recall that both methods start with an arbitrary x_0 and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1+\sqrt{\mu s}}$.

Theorem 5.3.6 (Convergence of NAG-$\texttt{SC}$)**.** Let $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. If the step size is set to $s = 1/(4L)$, the iterates $\{x_k\}_{k=0}^\infty$ generated by NAG-$\texttt{SC}$ (5.3) satisfy

$$f(x_k) - f(x^\star) \leq \frac{5L\left\|x_0 - x^\star\right\|^2}{\left(1+\frac{1}{12}\sqrt{\mu/L}\right)^k},$$

for all $k \geq 0$.

In brief, the theorem states that $\log(f(x_k) - f(x^\star)) \leq -O(k\sqrt{\mu/L})$, which matches the optimal rate for minimizing smooth strongly convex functions using only first-order information [Nes13]. More precisely, [Nes13] shows that $f(x_k) - f(x^\star) = O((1-$

$\sqrt{\mu/L})^k$) by taking $s = 1/L$ in NAG-$\texttt{SC}$. Although this optimal rate of NAG-$\texttt{SC}$ is well known in the litetature, this is the first Lyapunov-function-based proof of this result.

As indicated in Section 5.2, the proof of Theorem 5.3.6 rests on the discrete Lyapunov function (5.18):

$$\begin{aligned}\mathcal{E}(k) =& \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_k) - f(x^\star)\right) + \frac{1}{4}\left\|v_k\right\|^2 \\ &+ \frac{1}{4}\left\|v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^\star) + \sqrt{s}\nabla f(x_k)\right\|^2 - \frac{s\left\|\nabla f(x_k)\right\|^2}{2(1-\sqrt{\mu s})}.\end{aligned}$$

Recall that this functional is derived by writing NAG-$\texttt{SC}$ in the phase-space representation (5.17). Analogous to Lemma 5.3.2, the following lemma gives an upper bound on the difference $\mathcal{E}(k+1) - \mathcal{E}(k)$.

Lemma 5.3.7 (Lyapunov function for NAG-$\texttt{SC}$)**.** Let $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. Taking any step size $0 < s \leq 1/(4L)$, the discrete Lyapunov function (5.18) with $\{x_k\}_{k=0}^{\infty}$ generated by NAG-$\texttt{SC}$ satisfies

$$\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\frac{\sqrt{\mu s}}{6}\mathcal{E}(k+1).$$

The form of the inequality ensured by Lemma 5.3.7 is consistent with that of Lemma 5.3.2. Alternatively, it can be written as $\mathcal{E}(k+1) \leq \frac{1}{1+\frac{\sqrt{\mu s}}{6}}\mathcal{E}(k)$. With Lemma 5.3.7 in place, we give the proof of Theorem 5.3.6.

Proof. [Proof of Theorem 5.3.6] Given $s = 1/(4L)$, we have

$$f(x_k) - f(x^\star) \leq \frac{4(1-\sqrt{\mu/(4L)})}{3+4\sqrt{\mu/(4L)}}\mathcal{E}(k). \tag{5.25}$$

To see this, first note that

$$\mathcal{E}(k) \geq \frac{1+\sqrt{\mu/(4L)}}{1-\sqrt{\mu/(4L)}}\left(f(x_k) - f(x^\star)\right) - \frac{\left\|\nabla f(x_k)\right\|^2}{8L(1-\sqrt{\mu/(4L)})}$$

and

$$\frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 \leq f(x_k) - f(x^\star).$$

Combining these two inequalities, we get

$$\begin{aligned}\mathcal{E}(k) &\geq \frac{1+\sqrt{\mu/(4L)}}{1-\sqrt{\mu/(4L)}}\left(f(x_k)-f(x^\star)\right) - \frac{f(x_k)-f(x^\star)}{4(1-\sqrt{\mu/(4L)})}\\ &= \frac{3+4\sqrt{\mu/(4L)}}{4(1-\sqrt{\mu/(4L)})}(f(x_k)-f(x^\star)),\end{aligned}$$

which gives (5.25).

Next, we inductively apply Lemma 5.3.7, yielding

$$\mathcal{E}(k) \leq \frac{\mathcal{E}(0)}{\left(1+\frac{\sqrt{\mu s}}{6}\right)^k} = \frac{\mathcal{E}(0)}{\left(1+\frac{1}{12}\sqrt{\mu/L}\right)^k}. \tag{5.26}$$

Recognizing the initial velocity $v_0 = -\frac{2\sqrt{s}\nabla f(x_0)}{1+\sqrt{\mu s}}$ in NAG-$\texttt{SC}$, one can show that

$$\begin{aligned}\mathcal{E}(0) \leq & \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_0)-f(x^\star)\right) + \frac{s}{(1+\sqrt{\mu s})^2}\left\|\nabla f(x_0)\right\|^2\\ &+\frac{1}{4}\left\|\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_0-x^\star) - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\sqrt{s}\nabla f(x_0)\right\|^2\\ \leq & \left[\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right) + \frac{Ls}{(1+\sqrt{\mu s})^2} + \frac{2\mu/L}{(1-\sqrt{\mu s})^2} + \frac{Ls}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\right]\\ &\cdot L\left\|x_0-x^\star\right\|^2.\end{aligned} \tag{5.27}$$

Taking $s = 1/(4L)$ in (5.27), it follows from (5.25) and (5.26) that

$$f(x_k) - f(x^\star) \leq \frac{C_{\mu/L}\, L\left\|x_0-x^\star\right\|^2}{\left(1+\frac{1}{12}\sqrt{\mu/L}\right)^k}.$$

Here the constant factor $C_{\mu/L}$ is a short-hand for

$$\begin{aligned}\frac{4\left(1-\sqrt{\mu/(4L)}\right)}{3+4\sqrt{\mu/(4L)}}\cdot\Bigg[&\frac{1+\sqrt{\mu/(4L)}}{2-2\sqrt{\mu/(4L)}} + \frac{1}{4(1+\sqrt{\mu/(4L)})^2}\\ &+\frac{2\mu/L}{(1-\sqrt{\mu/(4L)})^2} + \frac{1}{8}\left(\frac{1+\sqrt{\mu/(4L)}}{1-\sqrt{\mu/(4L)}}\right)^2\Bigg],\end{aligned}$$

which is less than five by making use of the fact that $\mu/L \leq 1$. This completes the proof.

□

We now turn to the heavy-ball method (5.2). Recall that $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$ and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1+\sqrt{\mu s}}$.

Theorem 5.3.8 (Convergence of heavy-ball method)**.** Let $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. If the step size is set to $s = \mu/(16L^2)$, the iterates $\{x_k\}_{k=0}^{\infty}$ generated by the heavy-ball method satisfy

$$f(x_k) - f(x_0) \leq \frac{5L \left\| x_0 - x^\star \right\|^2}{\left(1 + \frac{\mu}{16L}\right)^k}$$

for all $k \geq 0$.

The heavy-ball method minimizes the objective at the rate $\log(f(x_k) - f(x^\star)) \leq -O(k\mu/L)$, as opposed to the optimal rate $-O(k\sqrt{\mu/L})$ obtained by NAG-SC. Thus, the acceleration phenomenon is not observed in the heavy-ball method for minimizing functions in the class $\mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. This difference is, on the surface, attributed to the much smaller step size $s = \mu/(16L^2)$ in Theorem 5.3.8 than the ($s = 1/(4L)$) in Theorem 5.3.6. Further discussion of this difference is given after Lemma 5.3.9 and in Section 5.3.3.

In addition to allowing us to complete the proof of Theorem 5.3.8, Lemma 5.3.9 will shed light on why the heavy-ball method needs a more conservative step size. To state this lemma, we consider the discrete Lyapunov function defined as

$$\mathcal{E}(k) = \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_k) - f(x^\star)\right) + \frac{1}{4}\left\|v_k\right\|^2 + \frac{1}{4}\left\| v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^\star) \right\|^2, \quad (5.28)$$

which is derived by discretizing the continuous Lyapunov function (5.21) using the phase-space representation of the heavy-ball method:

$$\begin{aligned} x_k - x_{k-1} &= \sqrt{s} v_{k-1} \\ v_k - v_{k-1} &= -\frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}} v_k - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \sqrt{s}\nabla f(x_k). \end{aligned} \quad (5.29)$$

Lemma 5.3.9 (Lyapunov function for the heavy-ball method)**.** Let $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. For any step size $s > 0$, the discrete Lyapunov function (5.28) with $\{x_k\}_{k=0}^{\infty}$ generated by the heavy-ball method satisfies

$$\begin{aligned}&\mathcal{E}(k+1)-\mathcal{E}(k)\leq-\sqrt{\mu s}\min\left\{\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}},\frac{1}{4}\right\}\mathcal{E}(k+1)\\&-\left[\frac{3\sqrt{\mu s}}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_{k+1})-f(x^\star)\right)-\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2\right].\end{aligned}\tag{5.30}$$

The proof of Lemma 5.3.9 can be found in Appendix 5.7.2.3. To apply this lemma to prove Theorem 5.3.8, we need to ensure

$$\frac{3\sqrt{\mu s}}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_{k+1})-f(x^\star)\right)-\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2\geq 0.\tag{5.31}$$

A sufficient and necessary condition for (5.31) is

$$\frac{3\sqrt{\mu s}}{4}\left(f(x_{k+1})-f(x^\star)\right)-\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)sL\left(f(x_{k+1})-f(x^\star)\right)\geq 0.\tag{5.32}$$

This is because $\left\|\nabla f(x_{k+1})\right\|^2\leq 2L\left(f(x_{k+1})-f(x^\star)\right)$, which can be further reduced to an equality (for example, $f(x)=\frac{L}{2}\|x\|^2$). Thus, the step size s must obey

$$s=O\left(\frac{\mu}{L^2}\right).$$

In particular, the choice of $s=\frac{\mu}{16L^2}$ fulfills (5.32) and, as a consequence, Lemma 5.3.9 implies

$$\mathcal{E}(k+1)-\mathcal{E}(k)\leq-\frac{\mu}{16L}\mathcal{E}(k+1).$$

The remainder of the proof of Theorem 5.3.8 is similar to that of Theorem 5.3.6 and is therefore omitted. As an aside, [Pol64] uses $s=4/(\sqrt{L}+\sqrt{\mu})^2$ for *local* accelerated convergence of the heavy-ball method. This choice of step size is larger than our step size $s=\frac{\mu}{16L^2}$, which yields a non-accelerated but global convergence rate.

The term $\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2$ in (5.30) that arises from finite differencing of (5.28) is a (small) term of order $O(s)$ and, as a consequence, this term is not reflected in Lemma 5.3.4. In relating to the case of NAG-SC, one would be tempted to ask why this term does not appear in Lemma 5.3.7. In fact, a similar term can be found in $\mathcal{E}(k+1) - \mathcal{E}(k)$ by taking a closer look at the proof of Lemma 5.3.7. However, this term is canceled out by the discrete version of the quadratic term $\frac{\sqrt{s}}{2}(\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X)\dot{X})$ in Lemma 5.3.2 and is, therefore, not present in the statement of Lemma 5.3.7. Note that this quadratic term results from the gradient correction (see Remark 5.3.5). In light of the above, the gradient correction is the key ingredient that allows for a larger step size in NAG-`SC`, which is necessary for achieving acceleration.

For completeness, we finish Section 5.3.2 by proving Lemma 5.3.7.

Proof. [Proof of Lemma 5.3.7] Using the Cauchy–Schwarz inequality, we have[9]

$$\begin{aligned}\mathbf{III} =&\frac{1}{4}\left\|\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_k - x^\star) + \sqrt{s}\nabla f(x_k)\right\|^2 \\ \leq&\frac{3}{4}\left[\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 \|v_k\|^2 + \frac{4\mu}{(1-\sqrt{\mu s})^2}\|x_k - x^\star\|^2 + s\|\nabla f(x_k)\|^2\right],\end{aligned}$$

which, together with the inequality

$$\begin{aligned}\frac{3s}{4}\|\nabla f(x_k)\|^2 - \frac{s\|\nabla f(x_k)\|^2}{2(1-\sqrt{\mu s})} &= \frac{s}{4}\|\nabla f(x_k)\|^2 + \frac{s}{2}\|\nabla f(x_k)\|^2 - \frac{s\|\nabla f(x_k)\|^2}{2(1-\sqrt{\mu s})} \\ &\leq \frac{Ls}{2}\left(f(x_k) - f(x^\star)\right) - \frac{s\sqrt{\mu s}\|\nabla f(x_k)\|^2}{2(1-\sqrt{\mu s})},\end{aligned}$$

for $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$, shows that the Lyapunov function (5.18) satisfies

$$\begin{aligned}\mathcal{E}(k) \leq &\left(\frac{1}{1-\sqrt{\mu s}} + \frac{Ls}{2}\right)\left(f(x_k) - f(x^\star)\right) + \frac{1+\sqrt{\mu s}+\mu s}{(1-\sqrt{\mu s})^2}\|v_k\|^2 \\ &+ \frac{3\mu}{(1-\sqrt{\mu s})^2}\|x_k - x^\star\|^2 + \frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_k) - f(x^\star) - \frac{s}{2}\|\nabla f(x_k)\|^2\right).\end{aligned} \tag{5.33}$$

Next, as shown in Appendix 5.7.2.2, the inequality

$$\begin{aligned}\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\sqrt{\mu s}&\left[\frac{1-2Ls}{\left(1-\sqrt{\mu s}\right)^2}\left(f(x_{k+1}) - f(x^\star)\right) + \frac{1}{1-\sqrt{\mu s}}\|v_{k+1}\|^2\right. \\ &\left.+\frac{\mu}{2(1-\sqrt{\mu s})^2}\|x_{k+1} - x^\star\|^2 + \frac{\sqrt{\mu s}}{(1-\sqrt{\mu s})^2}\left(f(x_{k+1}) - f(x^\star) - \frac{s}{2}\|\nabla f(x_{k+1})\|^2\right)\right]\end{aligned} \tag{5.34}$$

[9] See the definition of **III** in (5.18).

holds for $s \le 1/(2L)$. Comparing the coefficients of the same terms in (5.33) for $\mathcal{E}(k+1)$ and (5.34), we conclude that the first difference of the discrete Lyapunov function (5.18) must satisfy

$$
\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
\le & -\sqrt{\mu s}\min\left\{\frac{1-2Ls}{1-\sqrt{\mu s}+\frac{Ls}{2}\left(1-\sqrt{\mu s}\right)^2}, \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}+\mu s}, \frac{1}{6}, \frac{1}{1-\sqrt{\mu s}}\right\}\mathcal{E}(k+1) \\
\le & -\sqrt{\mu s}\min\left\{\frac{1-2Ls}{1+\frac{Ls}{2}}, \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}+\mu s}, \frac{1}{6}, \frac{1}{1-\sqrt{\mu s}}\right\}\mathcal{E}(k+1) \\
= & -\frac{\sqrt{\mu s}}{6}\mathcal{E}(k+1),
\end{aligned}
$$

since $s \le 1/(4L)$. □

5.3.3 A Numerical Stability Perspective on Acceleration

As shown in Section 5.3.2, the gradient correction is the fundamental cause of the difference in convergence rates between the heavy-ball method and NAG-SC. This section aims to further elucidate this distinction from the viewpoint of numerical stability. A numerical scheme is said to be stable if, roughly speaking, this scheme does not magnify errors in the input data. Accordingly, we address the question of what values of the step size s are allowed for solving the high-resolution ODEs (5.10) and (5.11) in a stable fashion. While various discretization schemes on low-resolution ODEs have been explored in [WWJ16, WRJ16, ZMSJ18], we limit our attention to the forward Euler scheme to simplify the discussion (see [SB13] for an exposition on discretization schemes).

For the heavy-ball method, the forward Euler scheme applied to (5.10) is

$$
\frac{X(t+\sqrt{s}) - 2X(t) + X(t-\sqrt{s})}{s} + 2\sqrt{\mu}\cdot\frac{X(t) - X(t-\sqrt{s})}{\sqrt{s}} + (1+\sqrt{\mu s})\nabla f(X(t-\sqrt{s})) = 0. \quad (5.35)
$$

Using the approximation $\nabla f(X(t-\sqrt{s})+\epsilon) \approx \nabla f(X(t-\sqrt{s})) + \nabla^2 f(X(t-\sqrt{s}))\epsilon$ for a small perturbation ϵ, we get the characteristic equation of (5.35):

$$\det\left(\lambda^2 \boldsymbol{I} - (2-2\sqrt{\mu s})\lambda \boldsymbol{I} + (1-2\sqrt{\mu s})\boldsymbol{I} + (1+\sqrt{\mu s})s\nabla^2 f(X(t-\sqrt{s}))\right) = 0,$$

where $\boldsymbol{I}$ denotes the $n \times n$ identity matrix. The numerical stability of (5.35) requires the roots of the characteristic equation to be no larger than one in absolute value. Therefore, a necessary condition for the stability is that[10]

$$(1-2\sqrt{\mu s})\boldsymbol{I} + (1+\sqrt{\mu s})s\nabla^2 f(X(t-\sqrt{s})) \preceq \boldsymbol{I}. \tag{5.36}$$

By the L-smoothness of f, the largest singular value of $\nabla^2 f(X(t-\sqrt{s}))$ can be as large as L. Therefore, (5.36) is guaranteed in the worst case analysis only if

$$(1+\sqrt{\mu s})sL \le 2\sqrt{\mu s},$$

which shows that the step size must obey

$$s \le O\left(\frac{\mu}{L^2}\right). \tag{5.37}$$

Next, we turn to the high-resolution ODE (5.11) of NAG-SC, for which the forward Euler scheme reads

$$\begin{aligned}&\frac{X(t+\sqrt{s}) - 2X(t) + X(t-\sqrt{s})}{s}\\ &+ (2\sqrt{\mu} + \sqrt{s}\nabla^2 f(X(t-\sqrt{s}))) \cdot \frac{X(t) - X(t-\sqrt{s})}{\sqrt{s}} + (1+\sqrt{\mu s})\nabla f(X(t-\sqrt{s})) = 0.\end{aligned} \tag{5.38}$$

Its characteristic equation is

$$\begin{aligned}\det\big(&\lambda^2 \boldsymbol{I} - (2-2\sqrt{\mu s} - s\nabla^2 f(X(t-\sqrt{s})))\lambda \boldsymbol{I}\\ &+(1-2\sqrt{\mu s})\boldsymbol{I} + \sqrt{\mu s^3}\nabla^2 f(X(t-\sqrt{s}))\big) = 0,\end{aligned} \tag{5.39}$$

[10] The notation $A \preceq B$ indicates that $B - A$ is positive semidefinite for symmetric matrices A and B.

which, as earlier, suggests that the numerical stability condition of (5.38) is

$$(1 - 2\sqrt{\mu s})\boldsymbol{I} + \sqrt{\mu s^3}\nabla^2 f(X(t - \sqrt{s})) \preceq \boldsymbol{I}.$$

This inequality is ensured by setting the step size

$$s = O\left(\frac{1}{L}\right). \tag{5.40}$$

As constraints on the step sizes, both (5.37) and (5.40) are in agreement with the discussion in Section 5.3.2, albeit from a different perspective. In short, a comparison between (5.35) and (5.38) reveals that the Hessian $\sqrt{s}\nabla^2 f(X(t - \sqrt{s}))$ makes the forward Euler scheme for the NAG-SC ODE numerically stable with a larger step size, namely $s = O(1/L)$. This is yet another reflection of the vital importance of the gradient correction in yielding acceleration for NAG-SC.

5.4 Gradient Correction for Gradient Norm Minimization

In this section, we extend the use of the high-resolution ODE framework to NAG-C (5.5) in the setting of minimizing an L-smooth convex function f. The main result is an improved rate of NAG-SC for minimizing the squared gradient norm. Indeed, we show that NAG-C achieves the $O(L^2/k^3)$ rate of convergence for minimizing $\|\nabla f(x_k)\|^2$. To the best of our knowledge, this is the *sharpest* known bound for this problem using NAG-C *without* any modification. Moreover, we will show that the gradient correction in NAG-C is responsible for this rate and, as it is therefore unsurprising that this inverse cubic rate was not perceived within the low-resolution ODE frameworks such as that of [SBC16]. In Section 5.4.3, we propose a new accelerated method with the same rate $O(L^2/k^3)$ and briefly discuss the benefit of the phase-space representation in simplifying technical proofs.

5.4.1 The ODE Case

We begin by studying the high-resolution ODE (5.12) corresponding to NAG-`C` with an objective $f \in \mathcal{F}_L^2(\mathbb{R}^n)$ and an arbitrary step size $s > 0$. For convenience, let $t_0 = 1.5\sqrt{s}$.

Theorem 5.4.1. Assume $f \in \mathcal{F}_L^2(\mathbb{R}^n)$ and let $X = X(t)$ be the solution to the ODE (5.12). The squared gradient norm satisfies

$$\inf_{t_0 \le u \le t} \|\nabla f(X(u))\|^2 \le \frac{(12 + 9sL)\|x_0 - x^\star\|^2}{2\sqrt{s}(t^3 - t_0^3)},$$

for all $t > t_0$.

By taking the step size $s = 1/L$, this theorem shows that

$$\inf_{t_0 \le u \le t} \|\nabla f(X(u))\|^2 = O(\sqrt{L}/t^3),$$

where the infimum operator is necessary as the squared gradient norm is generally not decreasing in t. In contrast, directly combining the convergence rate of the function value (see Corollary 5.4.3) and inequality $\|\nabla f(X)\|^2 \le 2L(f(X) - f(x^\star))$ only gives a $O(L/t^2)$ rate for squared gradient norm minimization.

The proof of the theorem is based on the continuous Lyapunov function

$$\mathcal{E}(t) = t\left(t + \frac{\sqrt{s}}{2}\right)(f(X) - f(x^\star)) + \frac{1}{2}\|t\dot{X} + 2(X - x^\star) + t\sqrt{s}\nabla f(X)\|^2, \quad (5.41)$$

which reduces to the continuous Lyapunov function in [SBC16] when setting $s = 0$.

Lemma 5.4.2. Let $f \in \mathcal{F}_L^2(\mathbb{R}^n)$. The Lyapunov function defined in (5.41) with $X = X(t)$ being the solution to the ODE (5.12) satisfies

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} \le -\left[\sqrt{s}t^2 + \left(\frac{1}{L} + \frac{s}{2}\right)t + \frac{\sqrt{s}}{2L}\right]\|\nabla f(X)\|^2 \quad (5.42)$$

for all $t \ge t_0$.

The decreasing rate of $\mathcal{E}(t)$ as specified in the lemma is sufficient for the proof of Theorem 5.4.1. First, note that Lemma 5.4.2 readily gives

$$\begin{aligned}\int_{t_0}^{t}\left[\sqrt{s}u^2+\left(\frac{1}{L}+\frac{s}{2}\right)u+\frac{\sqrt{s}}{2L}\right]\|\nabla f(X(u))\|^2\,\mathrm{d}u &\le -\int_{t_0}^{t}\frac{\mathrm{d}\mathcal{E}(u)}{\mathrm{d}u}\mathrm{d}u\\ &= \mathcal{E}(t_0)-\mathcal{E}(t)\\ &\le \mathcal{E}(t_0),\end{aligned}$$

where the last step is due to the fact $\mathcal{E}(t)\ge 0$. Thus, it follows that

$$\begin{aligned}\inf_{t_0\le u\le t}\|\nabla f(X(u))\|^2 &\le \frac{\int_{t_0}^{t}\left[\sqrt{s}u^2+\left(\frac{1}{L}+\frac{s}{2}\right)u+\frac{\sqrt{s}}{2L}\right]\|\nabla f(X(u))\|^2\,\mathrm{d}u}{\int_{t_0}^{t}\sqrt{s}u^2+\left(\frac{1}{L}+\frac{s}{2}\right)u+\frac{\sqrt{s}}{2L}\mathrm{d}u}\\ &\le \frac{\mathcal{E}(t_0)}{\sqrt{s}(t^3-t_0^3)/3+\left(\frac{1}{L}+\frac{s}{2}\right)(t^2-t_0^2)/2+\frac{\sqrt{s}}{2L}(t-t_0)}.\end{aligned}\tag{5.43}$$

Recognizing the initial conditions of the ODE (5.12), we get

$$\begin{aligned}\mathcal{E}(t_0) &= t_0(t_0+\sqrt{s}/2)(f(x_0)-f(x^\star))\\ &\quad+\frac{1}{2}\left\|-t_0\sqrt{s}\nabla f(x_0)+2(x_0-x^\star)+t_0\sqrt{s}\nabla f(x_0)\right\|^2\\ &\le 3s\cdot\frac{L}{2}\|x_0-x^\star\|^2+2\left\|x_0-x^\star\right\|^2,\end{aligned}$$

which together with (5.43) gives

$$\inf_{t_0\le u\le t}\|\nabla f(X(u))\|^2 \le \frac{(2+1.5sL)\left\|x_0-x^\star\right\|^2}{\sqrt{s}(t^3-t_0^3)/3+\left(\frac{1}{L}+\frac{s}{2}\right)(t^2-t_0^2)/2+\frac{\sqrt{s}}{2L}(t-t_0)}.\tag{5.44}$$

This bound reduces to the one claimed by Theorem 5.4.1 by only keeping the first term $\sqrt{s}(t^3-t_0^3)/3$ in the denominator.

The gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ in the high-resolution ODE (5.12) plays a pivotal role in Lemma 5.4.2 and is, thus, key to Theorem 5.4.1. As will be seen in the proof of the lemma, the factor $\|\nabla f(X)\|^2$ in (5.42) results from the term $t\sqrt{s}\nabla f(X)$ in the Lyapunov function (5.41), which arises from the gradient correction in the ODE (5.12). In light of this, the low-resolution ODE (5.8) of NAG-C cannot yield a result similar to Lemma 5.4.2 and; furthermore, we conjecture that the $O(\sqrt{L}/t^3)$

rate does applies to this ODE. Section 5.4.2 will discuss this point further in the discrete case.

In passing, it is worth pointing out that the analysis above applies to the case of $s = 0$. In this case, we have $t_0 = 0$, and (5.44) turns out to be

$$\inf_{0 \le u \le t} \left\| \nabla f(X(u)) \right\|^2 \le \frac{4L \left\| x_0 - x^\star \right\|^2}{t^2}.$$

This result is similar to that of the low-resolution ODE in [SBC16][11].

This section is concluded with the proof of Lemma 5.4.2.

Proof. [Proof of Lemma 5.4.2] The time derivative of the Lyapunov function (5.41) obeys

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} &= \left(2t + \frac{\sqrt{s}}{2}\right)(f(X) - f(x^\star)) + t\left(t + \frac{\sqrt{s}}{2}\right)\left\langle \nabla f(X), \dot{X} \right\rangle \\
&\quad + \left\langle t\dot{X} + 2(X - x^\star) + t\sqrt{s}\nabla f(X), -\left(\frac{\sqrt{s}}{2} + t\right)\nabla f(X) \right\rangle \\
&= \left(2t + \frac{\sqrt{s}}{2}\right)(f(X) - f(x^\star)) - (\sqrt{s} + 2t)\left\langle X - x^\star, \nabla f(X) \right\rangle \\
&\quad - \sqrt{s}t\left(t + \frac{\sqrt{s}}{2}\right)\left\| \nabla f(X) \right\|^2.
\end{aligned}$$

Making use of the basic inequality $f(x^\star) \ge f(X) + \langle \nabla f(X), x^\star - X \rangle + \frac{1}{2L}\left\| \nabla f(X) \right\|^2$ for L-smooth f, the expression of $\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t}$ above satisfies

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} &\le -\frac{\sqrt{s}}{2}(f(X) - f(x^\star)) - \left(\sqrt{s}t + \frac{1}{L}\right)\left(t + \frac{\sqrt{s}}{2}\right)\left\| \nabla f(X) \right\|^2 \\
&\le -\left(\sqrt{s}t + \frac{1}{L}\right)\left(t + \frac{\sqrt{s}}{2}\right)\left\| \nabla f(X) \right\|^2 \\
&= -\left[\sqrt{s}t^2 + \left(\frac{1}{L} + \frac{s}{2}\right)t + \frac{\sqrt{s}}{2L}\right]\left\| \nabla f(X) \right\|^2.
\end{aligned}$$

□

[11] To see this, recall that [SBC16] shows that $f(X(t)) - f(x^\star) \le \frac{2\|x_0 - x^\star\|^2}{t^2}$, where $X = X(t)$ is the solution to (5.44) with $s = 0$. Using the L-smoothness of f, we get $\|\nabla f(X(t))\|^2 \le 2L(f(X(t)) - f(x^\star)) \le \frac{4L\|x_0 - x^\star\|^2}{t^2}$.

Note that Lemma 5.4.2 shows $\mathcal{E}(t)$ is a decreasing function, from which we get

$$f(X) - f(x^\star) \le \frac{\mathcal{E}(t_0)}{t\left(t + \frac{\sqrt{s}}{2}\right)} = \frac{3s(f(x_0) - f(x^\star)) + 2\left\|x_0 - x^\star\right\|^2}{t\left(t + \frac{\sqrt{s}}{2}\right)}$$

by recognizing the initial conditions of the high-resolution ODE (5.12). This gives the following corollary.

Corollary 5.4.3. Under the same assumptions as in Theorem 5.4.1, for any $t > t_0$, we have

$$f(X(t)) - f(x^\star) \le \frac{(4 + 3sL)\left\|x_0 - x^\star\right\|^2}{t\,(2t + \sqrt{s})}.$$

5.4.2 The Discrete Case

We now turn to the discrete NAG-$\texttt{C}$ (5.5) for minimizing an objective $f \in \mathcal{F}_L^1(\mathbb{R}^n)$. Recall that this algorithm starts from any x_0 and $y_0 = x_0$. The discrete counterpart of Theorem 5.4.1 is as follows.

Theorem 5.4.4. Let $f \in \mathcal{F}_L^1(\mathbb{R}^n)$. For any step size $0 < s \le 1/(3L)$, the iterates $\{x_k\}_{k=0}^\infty$ generated by NAG-$\texttt{C}$ obey

$$\min_{0 \le i \le k} \left\|\nabla f(x_i)\right\|^2 \le \frac{8568\left\|x_0 - x^\star\right\|^2}{s^2(k+1)^3},$$

for all $k \ge 0$. In additional, we have

$$f(x_k) - f(x^\star) \le \frac{119\left\|x_0 - x^\star\right\|^2}{s(k+1)^2},$$

for all $k \ge 0$.

Taking $s = 1/(3L)$, Theorem 5.4.4 shows that NAG-$\texttt{C}$ minimizes the squared gradient norm at the rate $O(L^2/k^3)$. This theoretical prediction is in agreement with two numerical examples illustrated in Figure 5.4. To our knowledge, the bound

$O(L^2/k^3)$ is sharper than any existing bounds in the literature for NAG-$\texttt{C}$ for squared gradient norm minimization. In fact, the convergence result $f(x_k) - f(x^\star) = O(L/k^2)$ for NAG-$\texttt{C}$ and the L-smoothness of the objective immediately give $\|\nabla f(x_k)\|^2 \leq O(L^2/k^2)$. This well-known but loose bound can be improved by using a recent result from [AP16], which shows that a slightly modified version NAG-$\texttt{C}$ satisfies $f(x_k) - f(x^\star) = o(L/k^2)$ (see Section 5.5.2 for more discussion of this improved rate). This reveals

$$\|\nabla f(x_k)\|^2 \leq o\left(\frac{L^2}{k^2}\right),$$

which, however, remains looser than that of Theorem 5.4.4. In addition, the rate $o(L^2/k^2)$ is not valid for $k \leq n/2$ and, as such, the bound $o(L^2/k^2)$ on the squared gradient norm is *dimension-dependent* [AP16]. For completeness, the rate $O(L^2/k^3)$ can be achieved by introducing an additional sequence of iterates and a more aggressive step size policy in a variant of NAG-$\texttt{C}$ [GL16]. In stark contrast, our result shows that no adjustments are needed for NAG-$\texttt{C}$ to yield an accelerated convergence rate for minimizing the gradient norm.

An $\Omega(L^2/k^4)$ lower bound has been established by [Nes12] as the optimal convergence rate for minimizing $\|\nabla f\|^2$ with access to only first-order information. (For completeness, Appendix 5.7.3.3 presents an exposition of this fundamental barrier.) In the same paper, a regularization technique is used in conjunction with NAG-$\texttt{SC}$ to obtain a matching upper bound (up to a logarithmic factor). This method, however, takes as input the distance between the initial point and the minimizer, which is not practical in general [KF18].

Returning to Theorem 5.4.4, we present a proof of this theorem using a Lyapunov function argument. By way of comparison, we remark that Nesterov's estimate sequence technique is unlikely to be useful for characterizing the convergence of the gradient norm as this technique is essentially based on local quadratic approxima-

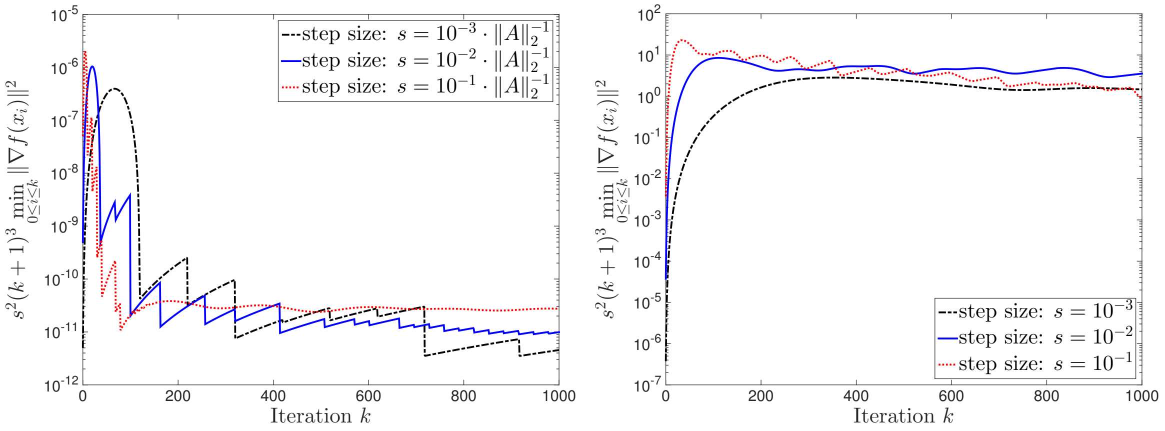


Figure 5.4: Scaled squared gradient norm $s^2(k+1)^3 \min_{0\le i\le k} \|\nabla f(x_i)\|^2$ of NAG-C. In both plots, the scaled squared gradient norm stays bounded as $k \to \infty$. Left: $f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle$, where $A = T'T$ is a 500×500 positive semidefinite matrix and b is 1×500. All entries of b, $T \in \mathbb{R}^{500\times500}$ are i.i.d. uniform random variables on $(0,1)$, and $\|\cdot\|_2$ denotes the matrix spectral norm. Right: $f(x) = \rho \log\left\{\sum_{i=1}^{200} \exp\left[\left(\langle a_i, x\rangle - b_i\right)/\rho\right]\right\}$, where $A = [a_1, \ldots, a_{200}]'$ is a 200×50 matrix and b is a 200×1 column vector. All entries of A and b are i.i.d.-sampled from $\mathcal{N}(0,1)$ and $\rho = 20$.

tions. The phase-space representation of NAG-C (5.5) takes the following form:

$$
\begin{aligned}
&x_k - x_{k-1} = \sqrt{s} v_{k-1} \\
&v_k - v_{k-1} = -\frac{3}{k} v_k - \sqrt{s}(\nabla f(x_k) - \nabla f(x_{k-1})) - \left(1 + \frac{3}{k}\right)\sqrt{s}\nabla f(x_k),
\end{aligned}
\tag{5.45}
$$

for any initial position x_0 and the initial velocity $v_0 = -\sqrt{s}\nabla f(x_0)$. This representation allows us to discretize the continuous Lyapunov function (5.41) into

$$
\begin{aligned}
\mathcal{E}(k) = s(k+3)(k+1)\left(f(x_k) - f(x^\star)\right)& \\
+ \frac{1}{2}\left\|(k+1)\sqrt{s}v_k + 2(x_{k+1} - x^\star) + (k+1)s\nabla f(x_k)\right\|^2.&
\end{aligned}
\tag{5.46}
$$

The following lemma characterizes the dynamics of this Lyapunov function.

Lemma 5.4.5. Under the assumptions of Theorem 5.4.4, we have

$$
\mathcal{E}(k+1) - \mathcal{E}(k) \le -\frac{s^2\left((k+3)(k-1) - Ls(k+3)(k+1)\right)}{2}\left\|\nabla f(x_{k+1})\right\|^2
$$

for all $k \ge 0$.

Next, we provide the proof of Theorem 5.4.4.

Proof. [Proof of Theorem 5.4.4] We start with the fact that

$$(k+3)(k-1) - Ls(k+3)(k+1) \geq 0, \tag{5.47}$$

for $k \geq 2$. To show this, note that it suffices to guarantee

$$s \leq \frac{1}{L} \cdot \frac{k-1}{k+1}, \tag{5.48}$$

which is self-evident since $s \leq 1/(3L)$ by assumption.

Next, by a telescoping-sum argument, Lemma 5.4.5 leads to the following inequalities for $k \geq 4$:

$$\begin{aligned}
\mathcal{E}(k) - \mathcal{E}(3) &= \sum_{i=3}^{k-1} \left(\mathcal{E}(i+1) - \mathcal{E}(i)\right) \\
&\leq \sum_{i=3}^{k-1} -\frac{s^2}{2}\left[(i+3)(i-1) - Ls(i+3)(i+1)\right] \left\|\nabla f(x_{i+1})\right\|^2 \\
&\leq -\frac{s^2}{2} \min_{4 \leq i \leq k} \left\|\nabla f(x_i)\right\|^2 \sum_{i=3}^{k-1} \left[(i+3)(i-1) - Ls(i+3)(i+1)\right] \\
&\leq -\frac{s^2}{2} \min_{4 \leq i \leq k} \left\|\nabla f(x_i)\right\|^2 \sum_{i=3}^{k-1} \left[(i+3)(i-1) - \frac{1}{3}(i+3)(i+1)\right],
\end{aligned} \tag{5.49}$$

where the second inequality is due to (5.47). To further simplify the bound, observe that

$$\sum_{i=3}^{k-1} \left[(i+3)(i-1) - \frac{1}{3}(i+3)(i+1)\right] = \frac{2k^3 - 38k + 60}{9} \geq \frac{(k+1)^3}{36},$$

for $k \geq 4$. Plugging this inequality into (5.49) yields

$$\mathcal{E}(k) - \mathcal{E}(3) \leq -\frac{s^2(k+1)^3}{72} \min_{4 \leq i \leq k} \left\|\nabla f(x_i)\right\|^2,$$

which gives

$$\min_{4 \leq i \leq k} \left\|\nabla f(x_i)\right\|^2 \leq \frac{72(\mathcal{E}(3) - \mathcal{E}(k))}{s^2(k+1)^3} \leq \frac{72\mathcal{E}(3)}{s^2(k+1)^3}. \tag{5.50}$$

It is shown in Appendix 5.7.3.1 that

$$\mathcal{E}(3) \le \mathcal{E}(2) \le 119 \left\|x_0 - x^\star\right\|^2,$$

for $s \le 1/(3L)$. As a consequence of this, (5.50) gives

$$\min_{4 \le i \le k} \left\|\nabla f(x_i)\right\|^2 \le \frac{8568 \left\|x_0 - x^\star\right\|^2}{s^2(k+1)^3}. \tag{5.51}$$

For completeness, Appendix 5.7.3.1 proves, via a brute-force calculation, that $\left\|\nabla f(x_0)\right\|^2$, $\left\|\nabla f(x_1)\right\|^2$, $\left\|\nabla f(x_2)\right\|^2$, and $\left\|\nabla f(x_3)\right\|^2$ are all bounded above by the right-hand side of (5.51). This completes the proof of the first inequality claimed by Theorem 5.4.4.

For the second claim in Theorem 5.4.4, the definition of the Lyapunov function and its decreasing property ensured by (5.47) implies

$$f(x_k) - f(x^\star) \le \frac{\mathcal{E}(k)}{s(k+3)(k+1)} \le \frac{\mathcal{E}(2)}{s(k+3)(k+1)} \le \frac{119 \left\|x_0 - x^\star\right\|^2}{s(k+1)^2}, \tag{5.52}$$

for all $k \ge 2$. Appendix 5.7.3.1 establishes that $f(x_0) - f(x^\star)$ and $f(x_1) - f(x^\star)$ are bounded by the right-hand side of (5.52). This completes the proof.

□

Now, we prove Lemma 5.4.5.

Proof. [Proof of Lemma 5.4.5]

The difference of the Lyapunov function (5.46) satisfies

$$\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
=& s(k+3)(k+1)\left(f(x_{k+1}) - f(x_k)\right) + s(2k+5)\left(f(x_{k+1}) - f(x^\star)\right) \\
&+ \big\langle 2(x_{k+2} - x_{k+1}) + \sqrt{s}(k+2)(v_{k+1} + \sqrt{s}\nabla f(x_{k+1})) - \sqrt{s}(k+1)(v_k + \sqrt{s}\nabla f(x_k)), \\
&2(x_{k+2} - x^\star) + (k+2)\sqrt{s}(v_{k+1} + \sqrt{s}\nabla f(x_{k+1}))\big\rangle \\
&- \frac{1}{2}\big\|2(x_{k+2} - x_{k+1}) + \sqrt{s}(k+2)(v_{k+1} + \sqrt{s}\nabla f(x_{k+1})) \\
&\qquad -(k+1)\sqrt{s}(v_k + \sqrt{s}\nabla f(x_k))\big\|^2
\end{aligned}$$

$$
\begin{aligned}
=& s(k+3)(k+1)\left(f(x_{k+1})-f(x_k)\right)+s(2k+5)\left(f(x_{k+1})-f(x^\star)\right)\\
&+\left\langle -s(k+3)\nabla f(x_{k+1}), 2(x_{k+2}-x^\star)+\sqrt{s}(k+2)(v_{k+1}+\sqrt{s}\nabla f(x_{k+1}))\right\rangle\\
&-\frac{1}{2}\left\|s(k+3)\nabla f(x_{k+1})\right\|^2\\
=& s(k+3)(k+1)\left(f(x_{k+1})-f(x_k)\right)+s(2k+5)\left(f(x_{k+1})-f(x^\star)\right)\\
&-s^{\frac{3}{2}}(k+3)(k+4)\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle-2s(k+3)\left\langle \nabla f(x_{k+1}), x_{k+1}-x^\star\right\rangle\\
&-s^2(k+3)(k+2)\left\|\nabla f(x_{k+1})\right\|^2-\frac{s^2}{2}(k+3)^2\left\|\nabla f(x_{k+1})\right\|^2,
\end{aligned}
$$

where the last two equalities are due to

$$
(k+3)\left(v_k+\sqrt{s}\nabla f(x_k)\right)-k\left(v_{k-1}+\sqrt{s}\nabla f(x_{k-1})\right)=-k\sqrt{s}\nabla f(x_k), \qquad (5.53)
$$

which follows from the phase-space representation (5.45). Rearranging the identity for $\mathcal{E}(k+1)-\mathcal{E}(k)$, we get

$$
\begin{aligned}
\mathcal{E}(k+1)-\mathcal{E}(k)=& s(k+3)(k+1)\left(f(x_{k+1})-f(x_k)\right)\\
&-s^{\frac{3}{2}}(k+3)(k+4)\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle\\
&+s(2k+5)\left(f(x_{k+1})-f(x^\star)\right)-s(2k+6)\left\langle \nabla f(x_{k+1}), x_{k+1}-x^\star\right\rangle\\
&-\frac{s^2(k+3)(3k+7)}{2}\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned}
\qquad (5.54)
$$

The next step is to recognize that the convexity and the L-smoothness of f gives

$$
\begin{aligned}
&f(x_{k+1})-f(x_k)\leq\left\langle \nabla f(x_{k+1}), x_{k+1}-x_k\right\rangle-\frac{1}{2L}\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2\\
&f(x_{k+1})-f(x^\star)\leq\left\langle \nabla f(x_{k+1}), x_{k+1}-x^\star\right\rangle.
\end{aligned}
$$

Plugging these two inequalities into (5.54), we have

$$
\begin{aligned}
\mathcal{E}(k+1)-\mathcal{E}(k)\leq& -s^{\frac{3}{2}}(k+3)\left\langle \nabla f(x_{k+1}), (k+4)v_{k+1}-(k+1)v_k\right\rangle\\
&-\frac{s}{2L}(k+3)(k+1)\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2\\
&-s\left\langle \nabla f(x_{k+1}), x_{k+1}-x^\star\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \frac{s^2(k+3)(3k+7)}{2}\left\|\nabla f(x_{k+1})\right\|^2 \\
&\leq -s^{\frac{3}{2}}(k+3)\left\langle \nabla f(x_{k+1}), (k+4)v_{k+1} - (k+1)v_k \right\rangle \\
&\quad - \frac{s}{2L}(k+3)(k+1)\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
&\quad - \frac{s^2(k+3)(3k+7)}{2}\left\|\nabla f(x_{k+1})\right\|^2,
\end{aligned}
$$

where the second inequality uses the fact that $\langle \nabla f(x_{k+1}), x_{k+1} - x^\star \rangle \geq 0$.

To further bound $\mathcal{E}(k+1) - \mathcal{E}(k)$, making use of (5.53) with $k+1$ in place of k, we get

$$
\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) &\leq s^2(k+3)(k+1)\left\langle \nabla f(x_{k+1}), \nabla f(x_{k+1}) - \nabla f(x_k) \right\rangle \\
&\quad - \frac{s}{2L}(k+3)(k+1)\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
&\quad - s^2\left(\frac{(k+3)(3k+7)}{2} - (k+3)(k+4)\right)\left\|\nabla f(x_{k+1})\right\|^2 \\
&= \frac{Ls^3(k+3)(k+1)}{2}\|\nabla f(x_{k+1})\|^2 \\
&\quad - \frac{s(k+3)(k+1)}{2L}\left\|(1-Ls)\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
&\quad - \frac{s^2(k+3)(k-1)}{2}\left\|\nabla f(x_{k+1})\right\|^2 \\
&\leq -\frac{s^2}{2}\left[(k+3)(k-1) - Ls(k+3)(k+1)\right]\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned}
$$

This completes the proof.

□

In passing, we remark that the gradient correction sheds light on the superiority of the high-resolution ODE over its low-resolution counterpart, just as in Section 5.3. Indeed, the absence of the gradient correction in the low-resolution ODE leads to the lack of the term $(k+1)s\nabla f(x_k)$ in the Lyapunov function (see Section 4 of [SBC16]), as opposed to the high-resolution Lyapunov function (5.46). Accordingly, it is unlikely to carry over the bound $\mathcal{E}(k+1) - \mathcal{E}(k) \leq -O(s^2k^2\|\nabla f(x_{k+1})\|^2)$ of Lemma 5.4.5 to the low-resolution case and, consequently, the low-resolution ODE approach pioneered by [SBC16] is insufficient to obtain the $O(L^2/k^3)$ rate for squared gradient norm minimization.

5.4.3 A Modified NAG-$\texttt{C}$ without a Phase-Space Representation

This section proposes a new accelerated method that also achieves the $O(L^2/k^3)$ rate for minimizing the squared gradient norm. This method takes the following form:

$$\begin{aligned} y_{k+1} &= x_k - s\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{k}{k+3}(y_{k+1} - y_k) - s\left(\frac{k}{k+3}\nabla f(y_{k+1}) - \frac{k-1}{k+3}\nabla f(y_k)\right), \end{aligned} \tag{5.55}$$

starting with x_0 and $y_0 = x_0$. As shown by the following theorem, this new method has the same convergence rates as NAG-$\texttt{C}$.

Theorem 5.4.6. Let $f \in \mathcal{F}_L^1(\mathbb{R}^n)$. Taking any step size $0 < s \leq 1/L$, the iterates $\{(x_k, y_k)\}_{k=0}^{\infty}$ generated by the modified NAG-$\texttt{C}$ (5.55) satisfy

$$\begin{aligned} \min_{0\leq i\leq k} \left\|\nabla f(x_i) + \nabla f(y_i)\right\|^2 &\leq \frac{882\left\|x_0 - x^\star\right\|^2}{s^2(k+1)^3} \\ f(y_k) - f(x^\star) &\leq \frac{21\left\|x_0 - x^\star\right\|^2}{s(k+1)^2}, \end{aligned}$$

for all $k \geq 0$.

We refer readers to Appendix 5.7.3.2 for the proof of Theorem 5.4.6, which is, as earlier, based on a Lyapunov function. However, since both $f(x_k)$ and $f(y_k)$ appear in the iteration, (5.55) does not admit a phase-space representation. As a consequence, the construction of the Lyapunov function is complex; we arrived at it via trial and error. Our initial aim was to seek possible improved rates of the original NAG-$\texttt{C}$ without using the phase-space representation, but the enormous challenges arising in this process motivated us to (1) modify NAG-$\texttt{C}$ to the current (5.55), and (2) to adopt the phase-space representation. Employing the phase-space representation yields a simple proof of the $O(L^2/k^3)$ rate for the original NAG-$\texttt{C}$ and this technique turned out to be useful for other accelerated methods.

5.5 Extensions

Motivated by the high-resolution ODE (5.12) of NAG-$\texttt{C}$, this section considers a family of generalized high-resolution ODEs that take the form

$$\ddot{X} + \frac{\alpha}{t}\dot{X} + \beta\sqrt{s}\nabla^2 f(X)\dot{X} + \left(1 + \frac{\alpha\sqrt{s}}{2t}\right)\nabla f(X) = 0, \tag{5.56}$$

for $t \geq \alpha\sqrt{s}/2$, with initial conditions $X(\alpha\sqrt{s}/2) = x_0$ and $\dot{X}(\alpha\sqrt{s}/2) = -\sqrt{s}\nabla f(x_0)$. As demonstrated in [SBC16, ACR17, VJFC18], the low-resolution counterpart (that is, set $s = 0$) of (5.56) achieves acceleration if and only if $\alpha \geq 3$. Accordingly, we focus on the case where the friction parameter $\alpha \geq 3$ and the gradient correction parameter $\beta > 0$. An investigation of the case of $\alpha < 3$ is left for future work.

By discretizing the ODE (5.56), we obtain a family of new accelerated methods for minimizing smooth convex functions:

$$\begin{aligned} y_{k+1} &= x_k - \beta s\nabla f(x_k) \\ x_{k+1} &= x_k - s\nabla f(x_k) + \frac{k}{k+\alpha}(y_{k+1} - y_k), \end{aligned} \tag{5.57}$$

starting with $x_0 = y_0$. The second line of the iteration is equivalent to

$$x_{k+1} = \left(1 - \frac{1}{\beta}\right)x_k + \frac{1}{\beta}y_{k+1} + \frac{k}{k+\alpha}(y_{k+1} - y_k).$$

In Section 5.5.1, we study the convergence rates of this family of generalized NAC-$\texttt{C}$ algorithms along the lines of Section 5.4. To further our understanding of (5.57), Section 5.5.2 shows that this method in the super-critical regime (that is, $\alpha > 3$) converges to the optimum actually faster than $O(1/(sk^2))$. As earlier, the proofs of all the results follow the high-resolution ODE framework introduced in Section 5.2. Proofs are deferred to Appendix 5.7.4. Finally, we note that Section 5.6 briefly sketches the extensions along this direction for NAG-`SC`.

5.5.1 Convergence Rates

The theorem below characterizes the convergence rates of the generalized NAG-`C` (5.57).

Theorem 5.5.1. Let $f \in \mathcal{F}_L^1(\mathbb{R}^n), \alpha \geq 3$, and $\beta > \frac{1}{2}$. There exists $c_{\alpha,\beta} > 0$ such that, taking any step size $0 < s \leq c_{\alpha,\beta}/L$, the iterates $\{x_k\}_{k=0}^{\infty}$ generated by the generalized NAG-`C` (5.57) obey

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\|^2 \leq \frac{C_{\alpha,\beta}\|x_0 - x^\star\|^2}{s^2(k+1)^3}, \tag{5.58}$$

for all $k \geq 0$. In addition, we have

$$f(x_k) - f(x^\star) \leq \frac{C_{\alpha,\beta}\|x_0 - x^\star\|^2}{s(k+1)^2},$$

for all $k \geq 0$. The constants $c_{\alpha,\beta}$ and $C_{\alpha,\beta}$ only depend on α and β.

The proof of Theorem 5.5.1 is given in Appendix 5.7.4.1 for $\alpha = 3$ and Appendix 5.7.4.2 for $\alpha > 3$. This theorem shows that the generalized NAG-`C` achieves the same rates as the original NAG-`C` in both squared gradient norm and function value minimization. The constraint $\beta > \frac{1}{2}$ reveals that further leveraging of the gradient correction does not hurt acceleration, but perhaps not the other way around (note that NAG-`C` in its original form corresponds to $\beta = 1$). It is an open question whether this constraint is a technical artifact or is fundamental to acceleration.

5.5.2 Faster Convergence in Super-Critical Regime

We turn to the case in which $\alpha > 3$, where we show that the generalized NAG-C in this regime attains a faster rate for minimizing the function value. The following proposition provides a technical inequality that motivates the derivation of the improved rate.

Proposition 5.5.2. Let $f \in \mathcal{F}_L^1(\mathbb{R}^n), \alpha > 3$, and $\beta > \frac{1}{2}$. There exists $c'_{\alpha,\beta} > 0$ such that, taking any step size $0 < s \leq c'_{\alpha,\beta}/L$, the iterates $\{x_k\}_{k=0}^{\infty}$ generated by the generalized NAG-C (5.57) obey

$$\sum_{k=0}^{\infty} \left[(k+1)\left(f(x_k) - f(x^\star)\right) + s(k+1)^2 \left\|\nabla f(x_k)\right\|^2\right] \leq \frac{C'_{\alpha,\beta}\left\|x_0 - x^\star\right\|^2}{s},$$

where the constants $c'_{\alpha,\beta}$ and $C'_{\alpha,\beta}$ only depend on α and β.

In relating to Theorem 5.5.1, one can show that Proposition 5.5.2 in fact implies (5.58) in Theorem 5.5.1. To see this, note that for $k \geq 1$, one has

$$\begin{aligned}\min_{0 \leq i \leq k} \left\|\nabla f(x_i)\right\|^2 &\leq \frac{\sum_{i=0}^{k} s(i+1)^2 \left\|\nabla f(x_i)\right\|^2}{\sum_{i=0}^{k} s(i+1)^2} \\ &\leq \frac{\frac{C'_{\alpha,\beta}\|x_0 - x^\star\|^2}{s}}{\frac{s}{6}(k+1)(k+2)(2k+1)} = O\left(\frac{\left\|x_0 - x^\star\right\|^2}{s^2 k^3}\right),\end{aligned}$$

where the second inequality follows from Proposition 5.5.2.

Proposition 5.5.2 can be thought of as a generalization of Theorem 6 of [SBC16]. In particular, this result implies an intriguing and important message. To see this, first note that, by taking $s = O(1/L)$, Proposition 5.5.2 gives

$$\sum_{k=0}^{\infty}(k+1)\left(f(x_k) - f(x^\star)\right) = O(L\left\|x_0 - x^\star\right\|^2), \tag{5.59}$$

which would not be valid if $f(x_k) - f(x^\star) \geq cL\left\|x_0 - x^\star\right\|^2/k^2$ for a constant $c > 0$. Thus, it is tempting to suggest that there might exist a faster convergence rate in the sense that

$$f(x_k) - f(x^\star) \leq o\left(\frac{L\left\|x_0 - x^\star\right\|^2}{k^2}\right). \tag{5.60}$$

This faster rate is indeed achievable as we show next, though there are examples where (5.59) and $f(x_k) - f(x^\star) = O(L\left\|x_0 - x^\star\right\|^2/k^2)$ are both satisfied but (5.60) does not hold (a counterexample is given in Appendx 5.7.4.3).

Theorem 5.5.3. Under the same assumptions as in Proposition 5.5.2, taking the step size $s = c'_{\alpha,\beta}/L$, the iterates $\{x_k\}_{k=0}^{\infty}$ generated by the generalized NAG-$\texttt{C}$ (5.57) starting from any $x_0 \neq x^\star$ satisfy

$$\lim_{k\to\infty} \frac{k^2(f(x_k) - f(x^\star))}{L\left\|x_0 - x^\star\right\|^2} = 0.$$

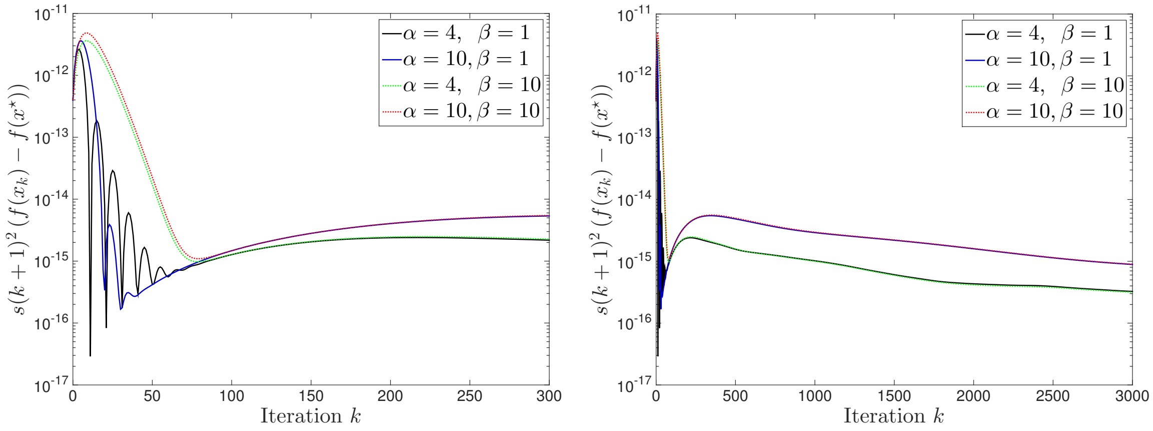


Figure 5.5: Scaled error $s(k+1)^2(f(x_k) - f(x^\star))$ of the generalized NAG-$\texttt{C}$ (5.57) with various (α, β). The setting is the same as the left plot of Figure 5.4, with the objective $f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle$. The step size is $s = 10^{-1}\|A\|_2^{-1}$. The left shows the short-time behaviors of the methods, while the right focuses on the long-time behaviors. The scaled error curves with the same β are very close to each other in the short-time regime, but in the long-time regime, the scaled error curves with the same α almost overlap. The four scaled error curves slowly tend to zero.

Figures 5.5 and 5.6 present several numerical studies concerning the prediction of Theorem 5.5.3. For a fixed dimension n, the convergence in Theorem 5.5.3 is uniform over functions in $\mathcal{F}^1 = \cup_{L>0}\mathcal{F}^1_L$ and, consequently, is independent of the Lipschitz constant L and the initial point x_0. In addition to following the high-resolution ODE framework, the proof of this theorem reposes on the finiteness of the series in

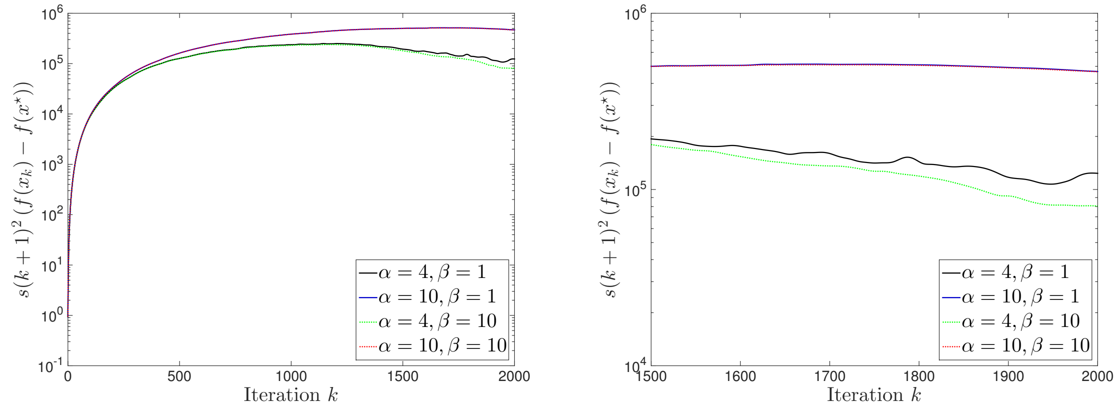


Figure 5.6: Scaled error $s(k+1)^2(f(x_k)-f(x^\star))$ of the generalized NAG-C (5.57) with various (α,β). The setting is the same as the right plot of Figure 5.4, with the objective $f(x)=\rho\log\left\{\sum_{i=1}^{200}\exp\left[\left(\langle a_i,x\rangle-b_i\right)/\rho\right]\right\}$. The step size is $s=0.1$. This set of simulation studies implies that the convergence in Theorem 5.5.3 is slow for some problems.

Proposition 5.5.2. See Appendix 5.7.4.2 and Appendix 5.7.4.4 for the full proofs of the proposition and the theorem, respectively.

In the literature, [AP16, May17, ACPR18] use low-resolution ODEs to establish the faster rate $o(1/k^2)$ for the generalized NAG-C (5.57) in the special case of $\beta=1$. In contrast, our proof of Theorem 5.5.3 is more general and applies to a broader class of methods.

In passing, we make the observation that Proposition 5.5.2 reveals that

$$\sum_{k=1}^{\infty} sk^2\left\|\nabla f(x_k)\right\|^2 \le \frac{C'_{\alpha,\beta}\left\|x_0-x^\star\right\|^2}{s},$$

which would not hold if $\min_{0\le i\le k}\|\nabla f(x_i)\|^2 \ge c\|x_0-x^\star\|^2/(s^2k^3)$ for all k and a constant $c>0$. In view of the above, it might be true that the rate of the generalized NAG-C for minimizing the squared gradient norm can be improved to

$$\min_{0\le i\le k}\|\nabla f(x_i)\|^2 = o\left(\frac{\|x_0-x^\star\|^2}{s^2k^3}\right).$$

We leave the confirmation or disconfirmation of this asymptotic result for future research.

5.6 Discussion

In this paper, we have proposed high-resolution ODEs for modeling three first-order optimization methods—the heavy-ball method, NAG-SC, and NAG-C. These new ODEs are more faithful surrogates for the corresponding discrete optimization methods than existing ODEs in the literature, thus serving as a more effective tool for understanding, analyzing, and generalizing first-order methods. Using this tool, we identified a term that we refer to as "gradient correction" in NAG-SC and in its high-resolution ODE, and we demonstrate its critical effect in making NAG-SC an accelerated method, as compared to the heavy-ball method. We also showed via the high-resolution ODE of NAG-C that this method minimizes the squared norm of the gradient at a faster rate than expected for smooth convex functions, and again the gradient correction is the key to this rate. Finally, the analysis of this tool suggested a new family of accelerated methods with the same optimal convergence rates as NAG-C.

The aforementioned results are obtained using the high-resolution ODEs in conjunction with a new framework for translating findings concerning the amenable ODEs into those of the less "user-friendly" discrete methods. This framework encodes an optimization property under investigation to a continuous-time Lyapunov function for an ODE and a discrete-time Lyapunov function for the discrete method. As an appealing feature of this framework, the transformation from the continuous Lyapunov function to its discrete version is through a phase-space representation. This representation links continuous objects such as position and velocity variables to their discrete counterparts in a faithful manner, permitting a transparent analysis of the three discrete methods that we studied.

There are a number of avenues open for future research using the high-resolution ODE framework. First, the discussion of Section 5.5 can carry over to the heavy-ball method and NAG-SC, which correspond to the high-resolution ODE

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \beta\sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s})\,\nabla f(X(t)) = 0$$

with $\beta = 0$ and $\beta = 1$, respectively. This ODE with a general $0 < \beta < 1$ corresponds to a new algorithm that can be thought of as an interpolation between the two methods. It is of interest to investigate the convergence properties of this class of algorithms. Second, we recognize that new optimization algorithms are obtained in [WWJ16, WRJ16] by using different discretization schemes on low-resolution ODE. Hence, a direction of interest is to apply the techniques therein to our high-resolution ODEs and to explore possible appealing properties of the new methods. Third, the technique of dimensional analysis, which we have used to derive high-resolution ODEs, can be further used to incorporate even higher-order powers of $\sqrt{s}$ into the ODEs. This might lead to further fine-grained findings concerning the discrete methods.

More broadly, we wish to remark on possible extensions of the high-resolution ODE framework beyond smooth convex optimization in the Euclidean setting. In the non-Euclidean case, it would be interesting to derive a high-resolution ODE for mirror descent [KBB15, WWJ16]. This framework might also admit extensions to non-smooth optimization and stochastic optimization, where the ODEs are replaced, respectively, by differential inclusions [ORX$^+$16, VJFC18] and stochastic differential equations [KB17, HLLL17, LTE17, LS17, XWG18, HMC$^+$18, GGZ18]. Finally, recognizing that the high-resolution ODEs are well-defined for non-convex functions, we believe that this framework will provide more accurate characterization of local behaviors of first-order algorithms near saddle points [JGN$^+$17, DJL$^+$17, HLS17]. On a related note, given the centrality of the problem of finding an approximate stationary point in the non-convex setting [CDHS17a, CDHS17b, AZ18], it is worth using the high-resolution ODE framework to explore possible applications of the faster rate for minimizing the squared gradient norm that we have uncovered.

5.7 Technical Details and Proofs

5.7.1 Technical Details in Section 5.2

5.7.1.1 Derivation of High-Resolution ODEs

In this section, we formally derive the high-resolution ODEs of the heavy-ball method and NAG-C. Let $t_k = k\sqrt{s}$. For the moment, let $X(t)$ be a sufficiently smooth map from $[0, \infty)$ (the heavy-ball method) or $[1.5\sqrt{s}, \infty)$ (NAG-C) to $\mathbb{R}^n$, with the correspondence $X(t_k) = X(k\sqrt{s}) = x_k$, where $\{x_k\}_{k=0}^{\infty}$ is the sequence of iterates generated by the heavy-ball method or NAG-C, depending on the context.

The heavy-ball method. For any function $f(x) \in \mathcal{S}_{\mu,L}^2(\mathbb{R}^n)$, setting $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$, multiplying both sides of (5.2) by $\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{1}{s}$ and rearranging the equality, we obtain

$$\frac{x_{k+1} + x_{k-1} - 2x_k}{s} + \frac{2\sqrt{\mu s}}{1 - \sqrt{\mu s}} \frac{x_{k+1} - x_k}{s} + \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \nabla f(x_k) = 0. \tag{5.61}$$

Plugging (5.13) into (5.61), we have

$$\ddot{X}(t_k) + O\left(\sqrt{s}\right) + \frac{2\sqrt{\mu}}{1 - \sqrt{\mu s}} \left[\dot{X}(t_k) + \frac{1}{2}\sqrt{s}\ddot{X}(t_k) + O\left(\left(\sqrt{s}\right)^2\right)\right] + \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \nabla f(X(t_k)) = 0.$$

By only ignoring the $O(s)$ term, we obtain the high-resolution ODE (5.10) for the heavy-ball method

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + (1 + \sqrt{\mu s})\, \nabla f(X) = 0.$$

NAG-C. For any function $f(x) \in \mathcal{F}_L^2(\mathbb{R}^n)$, multiplying both sides of (5.5) by $\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{1}{s}$ and rearranging the equality, we get

$$\frac{x_{k+1} + x_{k-1} - 2x_k}{s} + \frac{3}{k} \cdot \frac{x_{k+1} - x_k}{s} + \left(\nabla f(x_k) - \nabla f(x_{k-1})\right) + \left(1 + \frac{3}{k}\right) \nabla f(x_k) = 0. \tag{5.62}$$

For convenience, we slightly change the definition $t_k = k\sqrt{s} + (3/2)\sqrt{s}$ instead of $t_k = k\sqrt{s}$. Plugging (5.13) into (5.62), we have

$$\ddot{X}(t_k) + O\left(\left(\sqrt{s}\right)^2\right) + \frac{3}{t_k - (3/2)\sqrt{s}}\left[\dot{X}(t_k) + \frac{1}{2}\sqrt{s}\ddot{X}(t_k) + O\left(\left(\sqrt{s}\right)^2\right)\right]$$
$$+ \nabla^2 f(X(t_k))\dot{X}(t_k)\sqrt{s} + O\left(\left(\sqrt{s}\right)^2\right) + \frac{t_k + (3/2)\sqrt{s}}{t_k - (3/2)\sqrt{s}}\nabla f(X(t_k)) = 0.$$

Ignoring any $O(s)$ terms, we obtain the high-resolution ODE (5.12) for NAG-`C`

$$\ddot{X} + \frac{3}{t}\dot{X} + \sqrt{s}\nabla^2 f(X)\dot{X} + \left(1 + \frac{3\sqrt{s}}{2t}\right)\nabla f(X) = 0.$$

5.7.1.2 Derivation of Low-Resolution ODEs

In this section, we derive low-resolution ODEs of accelerated gradient methods for comparison. The results presented here are well-known in the literature and the purpose is for ease of reading. In [SBC16], the second-order Taylor expansions at both x_{k-1} and x_{k+1} with the step size $\sqrt{s}$ are,

$$\begin{cases} x_{k+1} = X\left((k+1)\sqrt{s}\right) = X(t_k) + \dot{X}(t_k)\sqrt{s} + \frac{1}{2}\ddot{X}(t_k)\left(\sqrt{s}\right)^2 + O\left(\left(\sqrt{s}\right)^3\right) \\ x_{k-1} = X\left((k-1)\sqrt{s}\right) = X(t_k) - \dot{X}(t_k)\sqrt{s} + \frac{1}{2}\ddot{X}(t_k)\left(\sqrt{s}\right)^2 + O\left(\left(\sqrt{s}\right)^3\right). \end{cases} \tag{5.63}$$

With the Taylor expansion (5.63), we obtain the gradient correction

$$\nabla f(x_k) - \nabla f(x_{k-1}) = \nabla^2 f(X(t_k))\dot{X}(t_k)\sqrt{s} + O\left(\left(\sqrt{s}\right)^2\right) = O\left(\sqrt{s}\right). \tag{5.64}$$

From (5.63) and (5.64), we can derive the following low-resolution ODEs.

(1) For any function $f(x) \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$.

(a) Recall the equivalent form (5.15) of NAG-`SC` (5.3) is

$$\frac{x_{k+1} + x_{k-1} - 2x_k}{s} + \frac{2\sqrt{\mu s}}{1 - \sqrt{\mu s}}\frac{x_{k+1} - x_k}{s}$$
$$+ (\nabla f(x_k) - \nabla f(x_{k-1})) + \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}}\nabla f(x_k) = 0.$$

Plugging (5.63) and (5.64) into (5.15), we have

$$\ddot{X}(t_k) + O\left(\sqrt{s}\right) + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}\left[\dot{X}(t_k) + \frac{1}{2}\ddot{X}\sqrt{s} + O\left(\left(\sqrt{s}\right)^2\right)\right]$$
$$+ O\left(\sqrt{s}\right) + \left(1 + O(\sqrt{s})\right)\nabla f(X(t_k)) = 0.$$

Hence, taking $s \to 0$, we obtain the low-resolution ODE (5.9) of NAG-SC

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X) = 0.$$

(b) Recall the equivalent form (5.61) of the heavy-ball method (5.2) is

$$\frac{x_{k+1} + x_{k-1} - 2x_k}{s} + \frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}}\frac{x_{k+1}-x_k}{s} + \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\nabla f(x_k) = 0.$$

Plugging (5.63) and (5.64) into (5.61), we have

$$\ddot{X}(t_k) + O\left(\sqrt{s}\right) + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}\left[\dot{X}(t_k) + \frac{1}{2}\sqrt{s}\ddot{X}(t_k) + O\left(\left(\sqrt{s}\right)^2\right)\right]$$
$$+ \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\nabla f(X(t_k)) = 0.$$

Hence, taking $s \to 0$, we obtain the low-resolution ODE (5.9) of the heavy-ball method

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X) = 0.$$

Notably, NAG-SC and the heavy-ball method share the same low-resolution ODE (5.9), which is almost consistent with (5.10). Thus the low-resolution ODE fails to capture the information from the "gradient correction" of NAG-SC.

(2) For any function $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$, recall the equivalent form (5.62) of NAG-C (5.5) is

$$\frac{x_{k+1} + x_{k-1} - 2x_k}{s} + \frac{3}{k}\cdot\frac{x_{k+1}-x_k}{s} + (\nabla f(x_k) - \nabla f(x_{k-1})) + \left(1 + \frac{3}{k}\right)\nabla f(x_k) = 0.$$

Plugging (5.63) and (5.64) into (5.62), we have

$$\ddot{X}(t_k) + O\left(\sqrt{s}\right) + \frac{3}{t_k} \cdot \left[\dot{X}(t_k) + \frac{1}{2}\ddot{X}(t_k)\sqrt{s} + O\left(\left(\sqrt{s}\right)^2\right)\right] \\ + O\left(\sqrt{s}\right) + \left(1 + \frac{3\sqrt{s}}{t_k}\right)\nabla f(X(t_k)) = 0.$$

Thus, by taking $s \to 0$, we obtain the low-resolution ODE (5.8) of NAG-C

$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X) = 0,$$

which is the same as [SBC16].

5.7.1.3 Solution Approximating Optimization Algorithms

To investigate the property about the high-resolution ODEs (5.10), (5.11) and (5.12), we need to state the relationship between them and their low-resolution corresponding ODEs. Here, we denote the solution to high-order ODE by $X_s = X_s(t)$. Actually, the low-resolution ODE is the special case of high-resolution ODE with $s = 0$. Take NAG-SC for example

$$\ddot{X}_s + \mu\dot{X}_s + \sqrt{s}\nabla f(X_s)\dot{X}_s + (1 + \sqrt{\mu s})\nabla f(X_s) = 0$$
$$X_s(0) = x_0, \quad \dot{X}_s(0) = -\frac{2\sqrt{s}\nabla f(x_0)}{1 + \sqrt{\mu s}}.$$

In other words, we consider a family of ODEs about the step size parameter s.

5.7.1.3.1 Proof of Proposition 5.2.1

Global Existence and Uniqueness To prove the global existence and uniqueness of solution to the high-resolution ODEs (5.10) and (5.11), we first emphasize a fact that if $X_s = X_s(t)$ is the solution of (5.10) or (5.11), there exists some constant $\mathcal{C}_1 > 0$ such that

$$\sup_{0 \le t < \infty} \left\| \dot{X}_s(t) \right\| \le \mathcal{C}_1, \tag{5.65}$$

which is only according to the following Lyapunov function

$$\mathcal{E}(t) = (1+\sqrt{\mu s})\left(f(X_s) - f(x^\star)\right) + \frac{1}{2}\|\dot{X}_s\|^2. \tag{5.66}$$

Now, we proceed to prove the global existence and uniqueness of solution to the high-resolution ODEs (5.10) and (5.11). Recall initial value problem (IVP) for first-order ODE system in $\mathbb{R}^m$ as

$$\dot{x} = b(x), \quad x(0) = x_0, \tag{5.67}$$

of which the classical theory about global existence and uniqueness of solution is shown as below.

Theorem 5.7.1 (Chillingworth, Chapter 3.1, Theorem 4 [Per13])**.** Let $M \in \mathbb{R}^m$ be a compact manifold and $b \in C^1(M)$. If the vector field b satisfies the global Lipschitz condition

$$\|b(x) - b(y)\| \leq \mathfrak{L}\,\|x-y\|$$

for all $x, y \in M$. Then for any $x_0 \in M$, the IVP (5.67) has a unique solution $x(t)$ defined for all $t \in \mathbb{R}$.

Apparently, the set $M_{\mathcal{C}_1} = \left\{(X_s, \dot{X}_s) \in \mathbb{R}^{2n} \middle| \|\dot{X}_s\| \leq \mathcal{C}_1\right\}$ is a compact manifold satisfying Theorem 5.7.1 with $m = 2n$.

- For the heavy-ball method, the phase-space representation of high-resolution ODE (5.10) is

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} X_s \\ \dot{X}_s \end{pmatrix} = \begin{pmatrix} \dot{X}_s \\ -\mu\dot{X}_s - (1+\sqrt{\mu s})\nabla f(X_s) \end{pmatrix}. \tag{5.68}$$

  For any $(X_s, \dot{X}_s)^\top, (Y_s, \dot{Y}_s)^\top \in M_{\mathcal{C}_1}$, we have

$$\left\| \begin{pmatrix} \dot{X}_s \\ -\mu\dot{X}_s - (1+\sqrt{\mu s})\nabla f(X_s) \end{pmatrix} - \begin{pmatrix} \dot{Y}_s \\ -\mu\dot{Y}_s - (1+\sqrt{\mu s})\nabla f(Y_s) \end{pmatrix} \right\|$$

$$
\begin{aligned}
&= \left\| \begin{pmatrix} \dot{X}_s - \dot{Y}_s \\ -\mu(\dot{X}_s - \dot{Y}_s) \end{pmatrix} \right\| + (1+\sqrt{\mu s}) \left\| \begin{pmatrix} 0 \\ \nabla f(X_s) - \nabla f(Y_s) \end{pmatrix} \right\| \\
&\leq \sqrt{1+\mu^2} \left\| \dot{X}_s - \dot{Y}_s \right\| + (1+\sqrt{\mu s}) L \left\| X_s - Y_s \right\| \\
&\leq 2 \max\left\{ \sqrt{1+\mu^2}, (1+\sqrt{\mu s})L \right\} \left\| \begin{pmatrix} X_s \\ \dot{X}_s \end{pmatrix} - \begin{pmatrix} Y_s \\ \dot{Y}_s \end{pmatrix} \right\|.
\end{aligned} \tag{5.69}
$$

- For NAG-SC, the phase-space representation of high-resolution ODE (5.11) is

$$
\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} X_s \\ \dot{X}_s \end{pmatrix} = \begin{pmatrix} \dot{X}_s \\ -\mu \dot{X}_s - \sqrt{s} \nabla^2 f(X_s) \dot{X}_s - (1+\sqrt{\mu s}) \nabla f(X_s) \end{pmatrix}. \tag{5.70}
$$

For any $(X_s, \dot{X}_s)^\top, (Y_s, \dot{Y}_s)^\top \in M_{\mathcal{C}_1}$, we have

$$
\left\| \begin{pmatrix} \dot{X}_s \\ -\mu \dot{X}_s - \sqrt{s} \nabla^2 f(X_s) \dot{X}_s - (1+\sqrt{\mu s}) \nabla f(X_s) \end{pmatrix} \right. \tag{5.71}
$$

$$
\begin{aligned}
&\left. - \begin{pmatrix} \dot{Y}_s \\ -\mu \dot{Y}_s - \sqrt{s} \nabla^2 f(Y_s) \dot{Y}_s - (1+\sqrt{\mu s}) \nabla f(Y_s) \end{pmatrix} \right\| \\
&\leq \left\| \begin{pmatrix} \dot{X}_s - \dot{Y}_s \\ -\left(\mu \boldsymbol{I} + \sqrt{s} \nabla^2 f(X_s)\right) (\dot{X}_s - \dot{Y}_s) \end{pmatrix} \right\| + \sqrt{s} \left\| \begin{pmatrix} 0 \\ \left(\nabla^2 f(X_s) - \nabla^2 f(Y_s)\right) \dot{Y}_s \end{pmatrix} \right\| \\
&\qquad + (1+\sqrt{\mu s}) \left\| \begin{pmatrix} 0 \\ \nabla f(X_s) - \nabla f(Y_s) \end{pmatrix} \right\| \\
&\leq \sqrt{1+2\mu^2+2sL^2} \left\| \dot{X}_s - \dot{Y}_s \right\| + \left[\sqrt{s} \mathcal{C}_1 L' + (1+\sqrt{\mu s}) L \right] \left\| X_s - Y_s \right\| \\
&\leq 2 \max\left\{ \sqrt{1+2\mu^2+2sL^2}, \sqrt{s} \mathcal{C}_1 L' + (1+\sqrt{\mu s}) L \right\} \left\| \begin{pmatrix} X_s \\ \dot{X}_s \end{pmatrix} - \begin{pmatrix} Y_s \\ \dot{Y}_s \end{pmatrix} \right\|.
\end{aligned} \tag{5.72}
$$

Based on the phase-space representation (5.68) and (5.70), together with the Lipschitz condition (5.69) and (5.71), Theorem 5.7.1 leads to the following Corollary.

Corollary 5.7.2. For any $f \in \mathcal{S}_{\mu}^{2}(\mathbb{R}^n) := \cup_{L\geq\mu}\mathcal{S}_{\mu,L}^{2}(\mathbb{R}^n)$, each of the two ODEs (5.10) and (5.11) with the specified initial conditions has a unique global solution $X \in C^2(I;\mathbb{R}^n)$

Approximation Based on the Lyapunov function (5.66), the gradient norm is bounded along the solution of (5.10) or (5.11), that is,

$$\sup_{0\leq t<\infty} \|\nabla f(X_s(t))\| \leq \mathcal{C}_2. \tag{5.73}$$

Recall the low-resolution ODE (5.9), the phase-space representation is proposed as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} X \\ \dot{X} \end{pmatrix} = \begin{pmatrix} \dot{X} \\ -\mu\dot{X} - \nabla f(X) \end{pmatrix}. \tag{5.74}$$

Similarly, using a Lyapunov function argument, we can show that if $X = X(t)$ is a solution of (5.9), we have

$$\sup_{0\leq t<\infty} \left\|\dot{X}(t)\right\| \leq \mathcal{C}_3. \tag{5.75}$$

Simple calculation tells us that there exists some constant $\mathcal{L}_1 > 0$ such that

$$\left\| \begin{pmatrix} \dot{X} \\ -\mu\dot{X} - \nabla f(X) \end{pmatrix} - \begin{pmatrix} \dot{Y} \\ -\mu\dot{Y} - \nabla f(Y) \end{pmatrix} \right\| \leq \mathcal{L}_1 \left\| \begin{pmatrix} X \\ \dot{X} \end{pmatrix} - \begin{pmatrix} Y \\ \dot{Y} \end{pmatrix} \right\|. \tag{5.76}$$

Now, we proceed to show the approximation.

Lemma 5.7.3. Let the solution to high-resolution ODEs (5.10) and (5.11) as $X = X_s(t)$ and that of (5.9) as $X = X(t)$, then we have

$$\lim_{s\to 0} \max_{0\leq t\leq T} \|X_s(t) - X(t)\| = 0 \tag{5.77}$$

for any fixed $T > 0$

In order to prove (5.77), we prove a stronger result as

$$\lim_{s\to 0}\max_{0\le t\le T}\left(\|X_s(t)-X(t)\|^2+\|\dot{X}_s(t)-\dot{X}(t)\|^2\right)=0. \tag{5.78}$$

Before we start to prove (5.78), we first describe the standard Gronwall-inequality as below.

Lemma 5.7.4. Let $m(t)$, $t\in[0,T]$, be a nonnegative function satisfying the relation

$$m(t)\le C+\alpha\int_0^t m(s)\mathrm{d}s,\quad t\in[0,T],$$

with $C,\alpha>0$. Then

$$m(t)\le C\mathrm{e}^{\alpha t}$$

for any $t\in[0,T]$.

The proof is only according to simple calculus, here we omit it.

Proof. [Proof of Lemma 5.7.3] We separate it into two parts.

- For the heavy-ball method, the phase-space representations (5.68) and (5.74) tell us that

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}X_s-X\\ \dot{X}_s-\dot{X}\end{pmatrix}=\begin{pmatrix}\dot{X}_s-\dot{X}\\ -\mu\left(\dot{X}_s-\dot{X}\right)-\left(\nabla f(X_s)-\nabla f(X)\right)\end{pmatrix}-\sqrt{\mu s}\begin{pmatrix}0\\ \nabla f(X_s)\end{pmatrix}$$

  By the boundedness (5.73), (5.65) and (5.75) and the inequality (5.76), we have

$$\begin{aligned}&\|X_s(t)-X(t)\|^2+\|\dot{X}_s(t)-\dot{X}(t)\|^2\\ =&2\int_0^t\left\langle\begin{pmatrix}X_s(u)-X(u)\\ \dot{X}_s(u)-\dot{X}(u)\end{pmatrix},\frac{\mathrm{d}}{\mathrm{d}u}\begin{pmatrix}X_s(u)-X(u)\\ \dot{X}_s(u)-\dot{X}(u)\end{pmatrix}\right\rangle\mathrm{d}u\\ &+\|X_s(0)-X(0)\|^2+\|\dot{X}_s(0)-\dot{X}(0)\|^2\\ \le&2\mathcal{L}_1\int_0^t\|X_s(u)-X(u)\|^2+\|\dot{X}_s(u)-\dot{X}(u)\|^2\mathrm{d}u\end{aligned}$$

$$
\begin{aligned}
&+\left[\left(\mathcal{C}_1+\mathcal{C}_3\right)\mathcal{C}_2\sqrt{\mu}t+\frac{4\sqrt{s}}{(1+\sqrt{\mu s})^2}\left\|\nabla f(x_0)\right\|^2\right]\sqrt{s}\\
\leq& 2\mathcal{L}_1\int_0^t\left\|X_s(u)-X(u)\right\|^2+\|\dot{X}_s(u)-\dot{X}(u)\|^2\mathrm{d}u+\mathcal{C}_4\sqrt{s}.
\end{aligned}
$$

According to Lemma 5.7.4, we have

$$
\left\|X_s(t)-X(t)\right\|^2+\|\dot{X}_s(t)-\dot{X}(t)\|^2\leq\mathcal{C}_4\sqrt{s}\mathrm{e}^{2\mathcal{L}_1 t}.
$$

- For NAG-SC, the phase-space representations (5.70) and (5.74) tell us that

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}X_s-X\\ \dot{X}_s-\dot{X}\end{pmatrix}=&\begin{pmatrix}\dot{X}_s-\dot{X}\\ -\mu\left(\dot{X}_s-\dot{X}\right)-\left(\nabla f(X_s)-\nabla f(X)\right)\end{pmatrix}\\
&-\sqrt{s}\begin{pmatrix}0\\ \nabla^2 f(X_s)\dot{X}_s+\sqrt{\mu}\nabla f(X_s)\end{pmatrix}
\end{aligned}
$$

Similarly, by the boundedness (5.73), (5.65) and (5.75) and the inequality (5.76), we have

$$
\begin{aligned}
&\left\|X_s(t)-X(t)\right\|^2+\|\dot{X}_s(t)-\dot{X}(t)\|^2\\
=&2\int_0^t\left\langle\begin{pmatrix}X_s(u)-X(u)\\ \dot{X}_s(u)-\dot{X}(u)\end{pmatrix},\frac{\mathrm{d}}{\mathrm{d}u}\begin{pmatrix}X_s(u)-X(u)\\ \dot{X}_s(u)-\dot{X}(u)\end{pmatrix}\right\rangle\mathrm{d}u\\
&+\left\|X_s(0)-X(0)\right\|^2+\|\dot{X}_s(0)-\dot{X}(0)\|^2\\
\leq&2\mathcal{L}_1\int_0^t\left\|X_s(u)-X(u)\right\|^2+\|\dot{X}_s(u)-\dot{X}(u)\|^2\mathrm{d}u\\
&+\left[\left(\mathcal{C}_1+\mathcal{C}_3\right)\left(L\mathcal{C}_1+\mathcal{C}_2\sqrt{\mu}\right)t+\frac{4\sqrt{s}}{(1+\sqrt{\mu s})^2}\left\|\nabla f(x_0)\right\|^2\right]\sqrt{s}\\
\leq&2\mathcal{L}_1\int_0^t\left\|X_s(u)-X(u)\right\|^2+\|\dot{X}_s(u)-\dot{X}(u)\|^2\mathrm{d}u+\mathcal{C}_5\sqrt{s}
\end{aligned}
$$

According to Lemma 5.7.4, we have

$$
\left\|X_s(t)-X(t)\right\|^2+\|\dot{X}_s(t)-\dot{X}(t)\|^2\leq\mathcal{C}_5\sqrt{s}\mathrm{e}^{2\mathcal{L}_1 t}
$$

The proof is complete. □

Lemma 5.7.5. The two methods, heavy-ball method and NAG-`SC`, converge to their low-resolution ODE (5.9) in the sense that

$$\lim_{s\to 0}\max_{0\le k\le T/\sqrt{s}}\left\|x_k - X(k\sqrt{s})\right\| = 0$$

for any fixed $T > 0$.

This result has bee studied in [WRJ16] and the method for proof refer to [SBC16, Appendix 2]. Combined with Corollary 5.7.2, Lemma 5.7.3 and Lemma 5.7.5, we complete the proof of Proposition 5.2.1.

5.7.1.3.2 Proof of Proposition 5.2.2

Global Existence and Uniqueness Similar as Appendix 5.7.1.3.1, we first emphasize the fact that if $X_s = X_s(t)$ is the solution of high-resolution ODE (5.12), there exists some constant $\mathcal{C}_6$ such that

$$\sup_{\frac{3\sqrt{s}}{2}\le t<\infty}\left\|\dot{X}_s(t)\right\| \le \mathcal{C}_6, \tag{5.79}$$

which is only according to the following Lyapunov function

$$\mathcal{E}(t) = \left(1 + \frac{3\sqrt{s}}{2t}\right)(f(X_s) - f(x^\star)) + \frac{1}{2}\left\|\dot{X}_s\right\|^2. \tag{5.80}$$

Now, we proceed to prove the global existence and uniqueness of solution to the high-resolution ODEs (5.12). Recall initial value problem (IVP) for first-order nonautonomous system in $\mathbb{R}^m$ as

$$\dot{x} = b(x,t), \quad x(0) = x_0, \tag{5.81}$$

of which the classical theory about global existence and uniqueness of solution is shown as below.

Theorem 5.7.6. Let $M \in \mathbb{R}^m$ be a compact manifold and $b \in C^1(M \times I)$, where $I = [t_0, \infty)$. If the vector field b satisfies the global Lipschitz condition

$$\|b(x,t) - b(y,t)\| \leq \mathfrak{L}\, \|x - y\|$$

for all $(x,t), (y,t) \in M \times I$. Then for any $x_0 \in M$, the IVP (5.81) has a unique solution $x(t)$ defined for all $t \in I$.

The proof is consistent with Theorem 3 and Theorem 4 of Chapter 3.1 in [Per13] except the Lipschitz condition for the vector field

$$\|b(x,t) - b(y,t)\| \leq \mathfrak{L}\, \|x - y\|$$

instead of

$$\|b(x) - b(y)\| \leq \mathfrak{L}\, \|x - y\|$$

for any $x, y \in M$. The readers can also refer to [GH13]. Similarly, the set

$$M_{\mathcal{C}_6} = \left\{ (X_s, \dot{X}_s) \in \mathbb{R}^{2n} \middle| \, \|\dot{X}_s\| \leq \mathcal{C}_6 \right\}$$

is a compact manifold satisfying Theorem 5.7.6 with $m = 2n$.

For NAG-$\mathtt{C}$, the phase-space representation of high-resolution ODE (5.11) is

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} X_s \\ \dot{X}_s \end{pmatrix} = \begin{pmatrix} \dot{X}_s \\ -\dfrac{3}{t}\cdot \dot{X}_s - \sqrt{s}\nabla^2 f(X_s)\dot{X}_s - \left(1 + \dfrac{3\sqrt{s}}{2t}\right)\nabla f(X_s) \end{pmatrix}. \tag{5.82}$$

For any $(X_s, \dot{X}_s, t), (Y_s, \dot{Y}_s, t) \in M_{\mathcal{C}_6} \times [(3/2)\sqrt{s}, \infty)$, we have

$$\left\| \begin{pmatrix} \dot{X}_s \\ -\dfrac{3}{t}\cdot \dot{X}_s - \sqrt{s}\nabla^2 f(X_s)\dot{X}_s - \left(1 + \dfrac{3\sqrt{s}}{2t}\right)\nabla f(X_s) \end{pmatrix} \right. \tag{5.83}$$
$$\left. - \begin{pmatrix} \dot{Y}_s \\ -\dfrac{3}{t}\cdot \dot{Y}_s - \sqrt{s}\nabla^2 f(Y_s)\dot{Y}_s - \left(1 + \dfrac{3\sqrt{s}}{2t}\right)\nabla f(Y_s) \end{pmatrix} \right\|$$

$$
\begin{aligned}
=&\left\|\begin{pmatrix} \dot{X}_s - \dot{Y}_s \\ -\left(\frac{3}{t}\cdot \boldsymbol{I} + \sqrt{s}\nabla^2 f(X_s)\right)(\dot{X}_s - \dot{Y}_s)\end{pmatrix}\right\| + \sqrt{s}\left\|\begin{pmatrix} 0 \\ \left(\nabla^2 f(X_s) - \nabla^2 f(Y_s)\right)\dot{Y}_s\end{pmatrix}\right\| \\
&\quad + \left(1 + \frac{3\sqrt{s}}{2t}\right)\left\|\begin{pmatrix} 0 \\ \nabla f(X_s) - \nabla f(Y_s)\end{pmatrix}\right\| \\
\leq&\sqrt{1 + \frac{18}{t_0^2} + 2sL^2}\left\|\dot{X}_s - \dot{Y}_s\right\| + \left[\sqrt{s}\mathcal{C}_6 L' + \left(1 + \frac{3\sqrt{s}}{2t_0}\right)L\right]\|X_s - Y_s\| \\
\leq&2\max\left\{\sqrt{1 + \frac{8}{s} + 2sL^2}, \sqrt{s}\mathcal{C}_6 L' + 2L\right\}\left\|\begin{pmatrix} X_s \\ \dot{X}_s\end{pmatrix} - \begin{pmatrix} Y_s \\ \dot{Y}_s\end{pmatrix}\right\|. \qquad (5.84)
\end{aligned}
$$

Based on the phase-space representation (5.82), together with (5.83), Theorem 5.7.6 leads the following Corollary.

Corollary 5.7.7. For any $f \in \mathcal{F}^2(\mathbb{R}^n) := \cup_{L>0}\mathcal{F}_L^2(\mathbb{R}^n)$, the ODE (5.12) with the specified initial conditions has a unique global solution $X \in C^2(I;\mathbb{R}^n)$.

Approximation Using a linear transformation $t + (3/2)\sqrt{s}$ instead of t, we can rewrite high-resolution ODE (5.12) as

$$
\ddot{X}_s(t) + \frac{3}{t + 3\sqrt{s}/2}\dot{X}_s(t) + \sqrt{s}\nabla^2 f(X_s(t))\dot{X}_s(t) + \left(1 + \frac{3\sqrt{s}}{2t + 3\sqrt{s}}\right)\nabla f(X_s(t)) = 0 \qquad (5.85)
$$

for $t \geq 0$, with initial $X_s(0) = x_0$ and $\dot{X}_s(0) = -\sqrt{s}\nabla f(x_0)$, of which the phase-space representation is

$$
\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} X_s \\ \dot{X}_s\end{pmatrix} = \begin{pmatrix} \dot{X}_s \\ -\frac{3}{t + 3\sqrt{s}/2}\cdot\dot{X}_s - \sqrt{s}\nabla^2 f(X_s)\dot{X}_s - \left(1 + \frac{3\sqrt{s}}{2t + 3\sqrt{s}}\right)\nabla f(X_s)\end{pmatrix}. \qquad (5.86)
$$

Here, we adopt the technique $\max\{\delta, t\}$ instead of t for any $\delta > 0$ to overcome the singular point $t = 0$, which is used firstly in [SBC16]. Then (5.86) is replaced into

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} X_s^\delta \\ \dot{X}_s^\delta \end{pmatrix}$$
$$= \begin{pmatrix} \dot{X}_s^\delta \\ -\dfrac{3}{\max\{\delta,t\} + 3\sqrt{s}/2}\cdot\dot{X}_s^\delta - \sqrt{s}\nabla^2 f(X_s)\dot{X}_s^\delta - \left(1 + \dfrac{3\sqrt{s}}{2\max\{\delta,t\} + 3\sqrt{s}}\right)\nabla f(X_s^\delta) \end{pmatrix}, \tag{5.87}$$

with the initial $X_s^\delta(0) = x_0$ and $\dot{X}_s^\delta(0) = -\sqrt{s}\nabla f(x_0)$. Recall the low-resolution ODE (5.8), with the above technique, the phase-space representation is proposed as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} X^\delta \\ \dot{X}^\delta \end{pmatrix} = \begin{pmatrix} \dot{X}^\delta \\ -\dfrac{3}{\max\{t,\delta\}}\cdot\dot{X}^\delta - \nabla f(X^\delta) \end{pmatrix}, \tag{5.88}$$

with the initial $X_s^\delta(0) = x_0$ and $\dot{X}_s^\delta(0) = 0$. Then according to (5.87) and (5.88), if we can prove for any $\delta > 0$ and any $t \in [0, T]$, the following equality holds

$$\lim_{s\to 0}\|X_s^\delta(t) - X^\delta(t)\| = 0.$$

Then, we can obtain the desired result as

$$\lim_{s\to 0}\|X_s(t) - X(t)\| = \lim_{s\to 0}\lim_{\delta\to 0}\|X_s^\delta(t) - X^\delta(t)\| = \lim_{\delta\to 0}\lim_{s\to 0}\|X_s^\delta(t) - X^\delta(t)\| = 0.$$

Similarly, using Lyapunov function argument, we can show that the solutions X_s^δ and X^δ satisfy

$$\sup_{0\le t<\infty}\left\|\dot{X}_s^\delta(t)\right\| \le \mathcal{C}_7 \quad \text{and} \quad \sup_{0\le t<\infty}\left\|\nabla f(X_s^\delta(t))\right\| \le \mathcal{C}_8, \tag{5.89}$$

and

$$\sup_{0\le t<\infty}\left\|\dot{X}^\delta(t)\right\| \le \mathcal{C}_9 \quad \text{and} \quad \sup_{0\le t<\infty}\left\|\nabla f(X^\delta(t))\right\| \le \mathcal{C}_{10}. \tag{5.90}$$

Simple calculation tells us that for any $(X, \dot{X}), (Y, \dot{Y}) \in \mathbb{R}^{2n}$, there exists some constant $\mathcal{L}_2 > 0$ such that

$$\left\| \begin{pmatrix} \dot{X} \\ -\frac{3}{\max\{t, \delta\} + (3/2)\sqrt{s}} \cdot \dot{X} - \nabla f(X) \end{pmatrix} \right. \tag{5.91}$$

$$\left. - \begin{pmatrix} \dot{Y} \\ -\frac{3}{\max\{t, \delta\} + (3/2)\sqrt{s}} \cdot \dot{Y} - \nabla f(Y) \end{pmatrix} \right\|$$

$$\leq \mathcal{L}_2 \left\| \begin{pmatrix} X \\ \dot{X} \end{pmatrix} - \begin{pmatrix} Y \\ \dot{Y} \end{pmatrix} \right\|. \tag{5.92}$$

for all $t \geq 0$. Now, we proceed to show the approximation.

Lemma 5.7.8. Denote the solution to high-resolution ODE (5.12) as $X = X_s(t)$ and that to (5.8) as $X = X(t)$. We have

$$\lim_{s \to 0} \max_{0 \leq t \leq T} \|X_s(t) - X(t)\| = 0 \tag{5.93}$$

for any fixed $T > 0$

In order to prove (5.93), we prove a stronger result

$$\lim_{s \to 0} \max_{0 \leq t \leq T} \left(\|X_s(t) - X(t)\|^2 + \|\dot{X}_s(t) - \dot{X}(t)\|^2 \right) = 0. \tag{5.94}$$

Proof. [Proof of Lemma 5.7.8] The phase-space representation (5.87) and (5.88) tell us that

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} X_s^\delta - X^\delta \\ \dot{X}_s^\delta - \dot{X}^\delta \end{pmatrix} = \begin{pmatrix} \dot{X}_s^\delta - \dot{X}^\delta \\ -\frac{3}{\max\{t, \delta\} + (3/2)\sqrt{s}} \cdot \left(\dot{X}_s^\delta - \dot{X}^\delta \right) - \left(\nabla f(X_s^\delta) - \nabla f(X^\delta) \right) \end{pmatrix}$$

$$- \sqrt{s} \begin{pmatrix} 0 \\ \nabla^2 f(X_s^\delta) \dot{X}_s^\delta + \frac{3 \nabla f(X_s)}{2 \max\{t, \delta\} + 3\sqrt{s}} - \frac{9 \nabla f(X)}{\max\{t, \delta\} \left(2 \max\{t, \delta\} + 3\sqrt{s} \right)} \end{pmatrix}$$

By the boundedness (5.89) and (5.90) and the Lipschitz inequality (5.91), we have

$$
\begin{aligned}
&\left\|X_s^\delta(t) - X^\delta(t)\right\|^2 + \left\|\dot{X}_s^\delta(t) - \dot{X}^\delta(t)\right\|^2 \\
=&2\int_0^t \left\langle \begin{pmatrix} X_s^\delta(u) - X^\delta(u) \\ \dot{X}_s^\delta(u) - \dot{X}^\delta(u) \end{pmatrix}, \frac{\mathrm{d}}{\mathrm{d}u}\begin{pmatrix} X_s^\delta(u) - X^\delta(u) \\ \dot{X}_s^\delta(u) - \dot{X}^\delta(u) \end{pmatrix} \right\rangle \mathrm{d}u \\
&+ \left\|X_s^\delta(0) - X^\delta(0)\right\|^2 + \left\|\dot{X}_s^\delta(0) - \dot{X}^\delta(0)\right\|^2 \\
\leq&2\mathcal{L}_2 \int_0^t \left\|X_s^\delta(u) - X^\delta(u)\right\|^2 + \left\|\dot{X}_s^\delta(u) - \dot{X}^\delta(u)\right\|^2 \mathrm{d}u \\
&+ \left[(\mathcal{C}_7 + \mathcal{C}_9)\left(L\mathcal{C}_7 + \frac{3\mathcal{C}_8}{2\delta} + \frac{9\mathcal{C}_{10}}{2\delta^2} \right) t + \sqrt{s}\left\|\nabla f(x_0)\right\|^2 \right]\sqrt{s} \\
\leq&2\mathcal{L}_2 \int_0^t \left\|X_s(u) - X(u)\right\|^2 + \|\dot{X}_s(u) - \dot{X}(u)\|^2 \mathrm{d}u + \mathcal{C}_{11}\sqrt{s}
\end{aligned}
$$

According to Lemma 5.7.4, we obtain the result as (5.94)

$$
\left\|X_s^\delta(t) - X^\delta(t)\right\|^2 + \|\dot{X}_s^\delta(t) - \dot{X}^\delta(t)\|^2 \leq \mathcal{C}_{11}\sqrt{s}\mathrm{e}^{2\mathcal{L}_2 t}
$$

The proof is complete. □

Lemma 5.7.9 (Theorem 2 [SBC16]). NAG-C converges to its low-resolution ODE in the sense that

$$
\lim_{s\to 0} \max_{0\leq k\leq T/\sqrt{s}} \left\|x_k - X(k\sqrt{s})\right\| = 0
$$

for any fixed $T > 0$.

Combined with Corollary 5.7.7, Lemma 5.7.8 and Lemma 5.7.9, we complete the proof of Proposition 5.2.2.

5.7.1.4 Closed-Form Solutions for Quadratic Functions

In this section, we propose the closed-form solutions to the three high-resolution ODEs for the quadratic objective function

$$
f(x) = \frac{1}{2}\theta x^2. \tag{5.95}
$$

where θ is the parameter suitable for the function in $\mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$ and $\mathcal{F}^2_L(\mathbb{R}^n)$. We compare them with the corresponding low-resolution ODEs and show the key difference. Throughout this section, both c_1 and c_2 are arbitrary real constants.

5.7.1.4.1 Oscillations and Non-Oscillations For any function $f(x) \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$, the parameter θ is set in $[\mu, L]$. First, plugging the quadratic objective (5.95) into the low-resolution ODE (5.9) of both NAG-`SC` and heavy-ball method, we have

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + \theta X = 0. \tag{5.96}$$

The closed-form solution of (5.96) can be shown from the theory of ODE, as below.

- When $\theta > \mu$, that is, $4\mu - 4\theta < 0$, the closed-form solution is the superimposition of two independent oscillation solutions

  $$X(t) = c_1 \mathrm{e}^{-\sqrt{\mu}t} \cos\left(\sqrt{\theta - \mu} \cdot t\right) + c_2 \mathrm{e}^{-\sqrt{\mu}t} \sin\left(\sqrt{\theta - \mu} \cdot t\right),$$

  of which the asymptotic estimate is

  $$\|X(t)\| = \Theta\left(\mathrm{e}^{-\sqrt{\mu}t}\right).$$

- When $\theta = \mu$, that is, $4\mu - 4\theta = 0$, the closed-form solution is the superimposition of two independent non-oscillation solutions

  $$X(t) = (c_1 + c_2 t)\, \mathrm{e}^{-\sqrt{\mu}t},$$

  of which the asymptotic estimate is

  $$\|X(t)\| = \Theta\left(t\mathrm{e}^{-\sqrt{\mu}t}\right).$$

Second, plugging the quadratic objective (5.95) into the high-resolution ODE (5.11) of NAG-`SC`, we have

$$\ddot{X} + (2\sqrt{\mu} + \sqrt{s}\theta)\dot{X} + (1 + \sqrt{\mu s})\theta X = 0. \tag{5.97}$$

The closed-form solutions to (5.97) are shown as below.

- When $s < \frac{4(\theta-\mu)}{\theta^2}$, that is, $4(\mu-\theta)+s\theta^2 < 0$, the closed-form solution is the superimposition of two independent oscillation solutions
$$X(t) = \mathrm{e}^{-\left(\sqrt{\mu}+\frac{\sqrt{s}\theta}{2}\right)t}\left[c_1\cos\left(\sqrt{(\theta-\mu)-\frac{1}{4}s\theta^2}\cdot t\right)+c_2\sin\left(\sqrt{(\theta-\mu)-\frac{1}{4}s\theta^2}\cdot t\right)\right],$$
the asymptotic estimate of which is
$$\|X(t)\| = \Theta\left(\mathrm{e}^{-\left(\sqrt{\mu}+\frac{\sqrt{s}\theta}{2}\right)t}\right) \le o\left(\mathrm{e}^{-\sqrt{\mu}t}\right).$$

- When $s = \frac{4(\theta-\mu)}{\theta^2}$, that is, $4(\mu-\theta)+s\theta^2 = 0$, the closed-form solution is the superimposition of two independent non-oscillation solutions
$$X(t) = (c_1+c_2t)\,\mathrm{e}^{-\left(\sqrt{\mu}+\frac{\sqrt{s}\theta}{2}\right)t},$$
the asymptotic estimate of which is
$$\|X(t)\| \le O\left(t\mathrm{e}^{-\left(\sqrt{\mu}+\frac{\sqrt{s}\theta}{2}\right)t}\right) \le o\left(\mathrm{e}^{-\sqrt{\mu}t}\right).$$

- When $s > \frac{4(\theta-\mu)}{\theta^2}$, that is, $4(\mu-\theta)+s\theta^2 > 0$, the closed-form solution is also the superimposition of two independent non-oscillation solutions
$$X(t) = c_1\mathrm{e}^{-\left(\sqrt{\mu}+\frac{\sqrt{s}\theta}{2}+\sqrt{(\mu-\theta)+\frac{s\theta^2}{4}}\right)t}+c_2\mathrm{e}^{-\left(\sqrt{\mu}+\frac{\sqrt{s}\theta}{2}-\sqrt{(\mu-\theta)+\frac{s\theta^2}{4}}\right)t},$$
the asymptotic estimate of which is
$$\|X(t)\| \le O\left(\mathrm{e}^{-\left(\sqrt{\mu}+\frac{\sqrt{s}\theta}{2}-\sqrt{(\mu-\theta)+\frac{s\theta^2}{4}}\right)t}\right) \le o\left(\mathrm{e}^{-\sqrt{\mu}t}\right).$$

Note that a simple calculation shows
$$\frac{4(\theta-\mu)}{\theta^2} = \frac{4}{\theta-\mu+\frac{\mu^2}{\theta-\mu}+2} \le \frac{2}{1+\mu}, \qquad \text{for } \theta \ge \mu.$$

Hence, when the step size satisfies $s \ge 2$, there is always no oscillation in the closed-form solution of (5.97).

Finally, plugging the quadratic objective (5.95) into the high-resolution ODE (5.10) of the heavy-ball method, we have

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + (1 + \sqrt{\mu s})\theta X = 0. \tag{5.98}$$

Since $4\mu - 4(1 + \sqrt{\mu s})\theta < 0$ is well established, the closed-form solution of (5.98) is the superimposition of two independent oscillation solutions

$$X(t) = c_1 e^{-\sqrt{\mu}t} \cos\left(\sqrt{(1 + \sqrt{\mu s})\theta - \mu} \cdot t\right) + c_2 e^{-\sqrt{\mu}t} \sin\left(\sqrt{(1 + \sqrt{\mu s})\theta - \mu} \cdot t\right),$$

the asymptotic estimate is

$$\|X(t)\| = \Theta\left(e^{-\sqrt{\mu}t}\right).$$

In summary, both the closed-form solutions to (5.96) and (5.98) are oscillated except the fragile condition $\theta = \mu$ and the speed of linear convergence is $\Theta\left(e^{-\sqrt{\mu}t}\right)$. However, the rate of convergence in the closed-form solution to the high-resolution ODE (5.97) is always faster than $\Theta\left(e^{-\sqrt{\mu}t}\right)$. Additionally, when the step size $s \geq 2$, there is always no oscillation in the closed-form solution of the high-resolution ODE (5.97).

5.7.1.4.2 Kummer's Equation and Confluent Hypergeometric Function

For any function $f(x) \in \mathcal{F}_L^2(\mathbb{R}^n)$, the parameter θ is required to located in $(0, L]$. Plugging the quadratic objective (5.95) into the low-resolution ODE (5.8) of NAG-C, we have

$$\ddot{X} + \frac{3}{t}\dot{X} + \theta X = 0,$$

the closed-form solution of which has been proposed in [SBC16]

$$X(t) = \frac{1}{\sqrt{\theta}t} \cdot \left[c_1 J_1\left(\sqrt{\theta}t\right) + c_2 Y_1\left(\sqrt{\theta}t\right)\right],$$

where $J_1(\cdot)$ and $Y_1(\cdot)$ are the Bessel function of the first kind and the second kind, respectively. According to the asymptotic property of Bessel functions,

$$J_1(\sqrt{\theta}t) \sim \frac{1}{\sqrt{t}} \quad \text{and} \quad Y_1(\sqrt{\theta}t) \sim \frac{1}{\sqrt{t}},$$

we obtain the following estimate

$$\|X(t)\| = \Theta\left(\frac{1}{t^{\frac{3}{2}}}\right).$$

Now, we plug the quadratic objective (5.95) into the high-resolution ODE (5.12) of NAG-`C` and obtain

$$\ddot{X} + \left(\frac{3}{t} + \theta\sqrt{s}\right)\dot{X} + \left(1 + \frac{3\sqrt{s}}{2t}\right)\theta X = 0. \tag{5.99}$$

For convenience, we define two new parameters as

$$\xi = \sqrt{s\theta^2 - 4\theta} \quad \text{and} \quad \rho = \frac{\theta\sqrt{s} + \sqrt{s\theta^2 - 4\theta}}{2}.$$

Let $Y = X\mathrm{e}^{\rho t}$ and $t' = \xi t$, the high-resolution ODE (5.99) can be rewritten as

$$t'\ddot{Y}(t') + (3 - t')\dot{Y}(t') - (3/2)Y(t') = 0,$$

which actually corresponds to the Kummer's equation. According to the closed-form solution to Kummer's equation, the high-resolution ODE (5.99) for quadratic function can be solved analytically as

$$X(t) = \mathrm{e}^{-\rho t}\left[c_1 M\left(\frac{3}{2}, 3, \xi t\right) + c_2 U\left(\frac{3}{2}, 3, \xi t\right)\right] \tag{5.100}$$

where $M(\cdot,\cdot,\cdot)$ and $U(\cdot,\cdot,\cdot)$ are the confluent hypergeometric functions of the first kind and the second kind. The integral expressions of $M(\cdot,\cdot,\cdot)$ and $U(\cdot,\cdot,\cdot)$ are given as

$$\begin{cases} M\left(\frac{3}{2}, 3, \xi t\right) = \frac{\Gamma(3)}{\Gamma\left(\frac{3}{2}\right)^2}\int_0^1 e^{\xi t u} u^{\frac{1}{2}}(1-u)^{\frac{1}{2}} du \\ U\left(\frac{3}{2}, 3, \xi t\right) = \frac{1}{\Gamma\left(\frac{3}{2}\right)}\int_0^1 e^{\xi t u} u^{\frac{1}{2}}(1-u)^{\frac{1}{2}} du. \end{cases}$$

Since the possible value of $\arg(\xi t)$ either 0 or $\pi/2$, we have

$$\begin{cases} M\left(\frac{3}{2}, 3, \xi t\right) \sim \Gamma(3)\left(\frac{e^{\xi t}(\xi t)^{-\frac{3}{2}}}{\Gamma\left(\frac{3}{2}\right)} + \frac{(-\xi t)^{-\frac{3}{2}}}{\Gamma\left(\frac{3}{2}\right)}\right) \\ U\left(\frac{3}{2}, 3, \xi t\right) \sim (-\xi t)^{-\frac{3}{2}}. \end{cases} \tag{5.101}$$

Apparently, from the asymptotic estimate of (5.101), we have

- When $s < 4/\theta$, that is, $s\theta^2 - 4\theta < 0$, the closed-form solution (5.100) is estimated as

  $$\|X(t)\| \leq \Theta\left(t^{-\frac{3}{2}} \mathrm{e}^{-\frac{\sqrt{s}\theta t}{2}}\right).$$

  Hence, when the step size satisfies $s < 4/L$, the above upper bound always holds.

- When $s \geq 4/\theta$, that is, $s\theta^2 - 4\theta \geq 0$, the closed-form solution (5.100) is estimated as

  $$\|X(t)\| \sim \mathrm{e}^{-\frac{\sqrt{s}\theta - \sqrt{s\theta^2 - 4\theta}}{2} \cdot t} t^{-\frac{3}{2}}.$$

  Apparently, we can bound

  $$\|X(t)\| \leq O\left(e^{-\frac{t}{\sqrt{s}}} t^{-\frac{3}{2}}\right) = O\left(e^{-\frac{t}{\sqrt{s}} - \frac{3\log t}{2}}\right)$$

  and

  $$\|X(t)\| \geq \Omega\left(e^{-\frac{2t}{\sqrt{s}}} t^{-\frac{3}{2}}\right) = \Omega\left(e^{-\frac{2t}{\sqrt{s}} - \frac{3\log t}{2}}\right).$$

5.7.2 Technical Details in Section 5.3

5.7.2.1 Proof of Lemma 5.3.4

With Cauchy-Schwarz inequality

$$\|\dot{X} + 2\sqrt{\mu}(X - x^\star)\|^2 \leq 2\left(\|\dot{X}\|^2 + 4\mu\left\|X - x^\star\right\|_2^2\right),$$

the Lyapunov function (5.21) can be estimated as

$$\mathcal{E} \leq (1 + \sqrt{\mu s})\left(f(X) - f(x^\star)\right) + \frac{3}{4}\|\dot{X}\|^2 + 2\mu\left\|X - x^\star\right\|^2. \tag{5.102}$$

Along the solution to the high-resolution ODE (5.10), the time derivative of the Lyapunov function (5.21) is

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} = (1 + \sqrt{\mu s})\left\langle \nabla f(X), \dot{X}\right\rangle + \frac{1}{2}\left\langle \dot{X}, -2\sqrt{\mu}\dot{X} - (1 + \sqrt{\mu s})\nabla f(X)\right\rangle$$

$$
\begin{aligned}
&+ \frac{1}{2}\left\langle \dot{X} + 2\sqrt{\mu}\,(X - x^\star)\,, -(1+\sqrt{\mu s})\nabla f(X)\right\rangle \\
&= -\sqrt{\mu}\left[\|\dot{X}\|_2^2 + (1+\sqrt{\mu s})\left\langle \nabla f(X), X - x^\star\right\rangle\right].
\end{aligned}
$$

With (5.102) and the inequality for any function $f(x) \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$

$$
f(x^\star) \geq f(X) + \langle \nabla f(X), x^\star - X\rangle + \frac{\mu}{2}\|X - x^\star\|_2^2,
$$

the time derivative of the Lyapunov function can be estimated as

$$
\begin{aligned}
\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} &\leq -\sqrt{\mu}\left[(1+\sqrt{\mu s})(f(X) - f(x^\star)) + \|\dot{X}\|_2^2 + \frac{\mu}{2}\|X - x^\star\|_2^2\right] \\
&\leq -\frac{\sqrt{\mu}}{4}\mathcal{E}
\end{aligned}
$$

Hence, the proof is complete.

5.7.2.2 Completing the Proof of Lemma 5.3.7

5.7.2.2.1 Derivation of (5.34) Here, we first point out that

$$
\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
&\leq -\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left[\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\rangle - s\|\nabla f(x_{k+1})\|^2\right) + \|v_{k+1}\|^2\right] \\
&\quad - \frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\right)\left(\frac{1}{L} - s\right)\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \qquad (5.103)
\end{aligned}
$$

implies (5.34) with $s \leq 1/L$. With (5.103), noting the basic inequality for $f(x) \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$ as

$$
\begin{cases}
f(x^\star) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x^\star - x_{k+1}\rangle + \dfrac{1}{2L}\|\nabla f(x_{k+1})\|_2^2 \\
f(x^\star) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x^\star - x_{k+1}\rangle + \dfrac{\mu}{2}\|x_{k+1} - x^\star\|_2^2,
\end{cases}
$$

when the step size satisfies $s \leq 1/(2L) \leq 1/L$, we have

$$
\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
&\leq -\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left[\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)(f(x_{k+1}) - f(x^\star)) + \frac{1}{2L}\left(\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\|\nabla f(x_{k+1})\|^2\right.
\end{aligned}
$$

$$
\begin{aligned}
&\qquad + \frac{\mu}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)\|x_{k+1}-x^\star\|^2 - \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right) s\left\|\nabla f(x_{k+1})\right\|^2 + \|v_{k+1}\|^2\Bigg]\\
&\leq -\sqrt{\mu s}\Bigg[\left(\frac{1}{1-\sqrt{\mu s}}\right)^2\left(f(x_{k+1})-f(x^\star)-s\left\|\nabla f(x_{k+1})\right\|^2\right)\\
&\qquad\qquad + \frac{\sqrt{\mu s}}{(1-\sqrt{\mu s})^2}\left(f(x_{k+1})-f(x^\star)-\frac{s}{2}\left\|\nabla f(x_{k+1})\right\|^2\right)\\
&\qquad\qquad + \frac{\mu}{2(1-\sqrt{\mu s})^2}\|x_{k+1}-x^\star\|^2 + \frac{1}{1-\sqrt{\mu s}}\|v_{k+1}\|^2\Bigg]\\
&\leq -\sqrt{\mu s}\Bigg[\frac{1-2Ls}{\left(1-\sqrt{\mu s}\right)^2}\left(f(x_{k+1})-f(x^\star)\right) + \frac{1}{1-\sqrt{\mu s}}\|v_{k+1}\|^2\\
&\qquad\qquad + \frac{\mu}{2(1-\sqrt{\mu s})^2}\|x_{k+1}-x^\star\|^2\\
&\qquad\qquad + \frac{\sqrt{\mu s}}{(1-\sqrt{\mu s})^2}\left(f(x_{k+1})-f(x^\star)-\frac{s}{2}\left\|\nabla f(x_{k+1})\right\|^2\right)\Bigg].
\end{aligned}
$$

5.7.2.2.2 Derivation of (5.103) Now, we show the derivation of (5.103). Recall the discrete Lyapunov function (5.18),

$$
\begin{aligned}
\mathcal{E}(k) = &\underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_k)-f(x^\star)\right)}_{\mathbf{I}} + \underbrace{\frac{1}{4}\|v_k\|^2}_{\mathbf{II}}\\
&+\underbrace{\frac{1}{4}\left\|v_k+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1}-x^\star)+\sqrt{s}\nabla f(x_k)\right\|^2}_{\mathbf{III}}\underbrace{-\frac{s}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)\left\|\nabla f(x_k)\right\|^2}_{\textbf{additional term}}.
\end{aligned}
$$

For convenience, we calculate the difference between $\mathcal{E}(k)$ and $\mathcal{E}(k+1)$ by the three parts, **I**, **II** and **III** respectively.

- For the part **I**, potential, with the convexity, we have

$$
\begin{aligned}
&\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_{k+1})-f(x^\star)\right)-\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_k)-f(x^\star)\right)\\
\leq&\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left[\langle\nabla f(x_{k+1}),x_{k+1}-x_k\rangle-\frac{1}{2L}\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2\right]\\
\leq&\underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\langle\nabla f(x_{k+1}),v_k\rangle}_{\mathbf{I}_1}\underbrace{-\frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2}_{\mathbf{I}_2}.
\end{aligned}
$$

- For the part **II**, kinetic energy, with the phase representation of NAG-`SC` (5.17), we have

$$
\begin{aligned}
&\frac{1}{4}\left\|v_{k+1}\right\|^2 - \frac{1}{4}\left\|v_k\right\|^2 \\
=&\frac{1}{2}\left\langle v_{k+1} - v_k, v_{k+1}\right\rangle - \frac{1}{4}\left\|v_{k+1} - v_k\right\|^2 \\
&-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\|v_{k+1}\right\|^2 - \frac{\sqrt{s}}{2}\left\langle \nabla f(x_{k+1}) - \nabla f(x_k), v_{k+1}\right\rangle \\
&-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{\sqrt{s}}{2}\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle - \frac{1}{4}\left\|v_{k+1} - v_k\right\|^2 \\
=&\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\|v_{k+1}\right\|^2}_{\mathbf{II}_1}\underbrace{-\frac{\sqrt{s}}{2}\cdot\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\left\langle \nabla f(x_{k+1}) - \nabla f(x_k), v_k\right\rangle}_{\mathbf{II}_2} \\
&+\underbrace{\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\cdot\frac{s}{2}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2}_{\mathbf{II}_3} \\
&+\underbrace{\frac{s}{2}\left\langle \nabla f(x_{k+1}) - \nabla f(x_k), \nabla f(x_{k+1})\right\rangle}_{\mathbf{II}_4} \\
&\underbrace{-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{\sqrt{s}}{2}\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle}_{\mathbf{II}_5}\underbrace{-\frac{1}{4}\left\|v_{k+1} - v_k\right\|^2}_{\mathbf{II}_6}.
\end{aligned}
$$

- For the part **III**, mixed energy, with the phase representation of NAG-`SC` (5.17), we have

$$
\begin{aligned}
&\frac{1}{4}\left\|v_{k+1} + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+2} - x^{\star}) + \sqrt{s}\nabla f(x_{k+1})\right\|^2 \\
&-\frac{1}{4}\left\|v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^{\star}) + \sqrt{s}\nabla f(x_k)\right\|^2 \\
=&\frac{1}{2}\left\langle -\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\sqrt{s}\nabla f(x_{k+1}), \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}v_{k+1} + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^{\star}) + \sqrt{s}\nabla f(x_{k+1})\right\rangle \\
&-\frac{1}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 s\left\|\nabla f(x_{k+1})\right\|^2 \\
=&\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle \nabla f(x_{k+1}), x_{k+1} - x^{\star}\right\rangle}_{\mathbf{III}_1}\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\sqrt{s}\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle}_{\mathbf{III}_2}
\end{aligned}
$$

$$\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)s\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_3}\underbrace{-\frac{1}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 s\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_4}.$$

Both $\mathbf{II}_2$ and $\mathbf{III}_3$ above are the discrete correspondence of the terms $-\frac{\sqrt{s}}{2}\left\|\nabla f(X(t))\right\|^2$ and $-\frac{\sqrt{s}}{2}\dot{X}(t)^\top\nabla^2 f(X(t))\dot{X}(t)$ in (5.19). The impact can be found in the calculation. Now, we calculate the difference of discrete Lyapunov function (5.18) at k-th iteration by the simple operation

$$\begin{aligned}
&\mathcal{E}(k+1)-\mathcal{E}(k)\\
\le\;&\underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\left\langle\nabla f(x_{k+1}),v_k\right\rangle}_{\mathbf{I}_1}\underbrace{-\frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2}_{\mathbf{I}_2}\\
&\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\|v_{k+1}\right\|^2}_{\mathbf{II}_1}\underbrace{-\frac{\sqrt{s}}{2}\cdot\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\left\langle\nabla f(x_{k+1})-\nabla f(x_k),v_k\right\rangle}_{\mathbf{II}_2}\\
&+\underbrace{\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\cdot\frac{s}{2}\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2}_{\mathbf{II}_3}\\
&+\underbrace{\frac{s}{2}\left\langle\nabla f(x_{k+1})-\nabla f(x_k),\nabla f(x_{k+1})\right\rangle}_{\mathbf{II}_4}\underbrace{-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{\sqrt{s}}{2}\left\langle\nabla f(x_{k+1}),v_{k+1}\right\rangle}_{\mathbf{II}_5}\\
&\underbrace{-\frac{1}{4}\left\|v_{k+1}-v_k\right\|^2}_{\mathbf{II}_6}\\
&\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle\nabla f(x_{k+1}),x_{k+1}-x^\star\right\rangle}_{\mathbf{III}_1}\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\sqrt{s}\left\langle\nabla f(x_{k+1}),v_{k+1}\right\rangle}_{\mathbf{III}_2}\\
&\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)s\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_3}\underbrace{-\frac{1}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 s\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_4}\\
&\underbrace{-\frac{s}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)\left(\left\|\nabla f(x_{k+1})\right\|^2-\left\|\nabla f(x_k)\right\|^2\right)}_{\textbf{additional term}}\\
\le\;&\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle\nabla f(x_{k+1}),x_{k+1}-x^\star\right\rangle+\left\|v_{k+1}\right\|^2\right)}_{\mathbf{II}_1+\mathbf{III}_1}
\end{aligned}$$

$$\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left[\sqrt{s}\left\langle \nabla f(x_{k+1}), \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)v_{k+1}-v_k\right\rangle + s\left\|\nabla f(x_{k+1})\right\|^2\right]}_{\frac{1}{2}\mathbf{I}_1+\mathbf{III}_2+\mathbf{III}_3}$$

$$\underbrace{-\frac{\sqrt{s}}{2}\cdot\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\left\langle \nabla f(x_{k+1})-\nabla f(x_k), v_k\right\rangle}_{\mathbf{II}_2} + \underbrace{\frac{s}{2}\left\langle\nabla f(x_{k+1})-\nabla f(x_k),\nabla f(x_{k+1})\right\rangle}_{\mathbf{II}_4}$$

$$\underbrace{-\frac{1}{4}\left[\left\|v_{k+1}-v_k\right\|^2 + 2\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\left\langle\nabla f(x_{k+1}), v_{k+1}-v_k\right\rangle\right.}_{\frac{1}{2}\mathbf{I}_1+\mathbf{II}_5+\mathbf{II}_6+\mathbf{III}_4}$$

$$\underbrace{\left.+\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 s\left\|\nabla f(x_{k+1})\right\|^2\right]}_{\frac{1}{2}\mathbf{I}_1+\mathbf{II}_5+\mathbf{II}_6+\mathbf{III}_4}$$

$$\underbrace{-\frac{1}{2}\left[\frac{1}{L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right) - s\left(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\right)\right]\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2}_{\mathbf{I}_2+\mathbf{II}_3}$$

$$\underbrace{-\frac{1}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)s\left(\left\|\nabla f(x_{k+1})\right\|^2-\left\|\nabla f(x_k)\right\|^2\right)}_{\textbf{additional term}}$$

Now, the term, $(1/2)\mathbf{I}_1+\mathbf{II}_5+\mathbf{II}_6+\mathbf{III}_4$, can be calculated as

$$\begin{aligned}\frac{1}{2}\mathbf{I}_1+\mathbf{II}_5+\mathbf{II}_6+\mathbf{III}_4 &= -\frac{1}{4}\left[\left\|v_{k+1}-v_k\right\|^2 + 2\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\left\langle\nabla f(x_{k+1}), v_{k+1}-v_k\right\rangle\right.\\ &\qquad\qquad \left.+\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 s\left\|\nabla f(x_{k+1})\right\|^2\right]\\ &= -\frac{1}{4}\left\|v_{k+1}-v_k+\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\nabla f(x_k)\right\|^2\\ &\leq 0.\end{aligned}$$

With phase representation of NAG-`SC` (5.17), we have

$$\begin{aligned}&\frac{1}{2}\mathbf{I}_1+\mathbf{III}_2+\mathbf{III}_3\\ =&-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left[\sqrt{s}\left\langle \nabla f(x_{k+1}), \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)v_{k+1}-v_k\right\rangle + s\left\|\nabla f(x_{k+1})\right\|^2\right]\\ =&\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)s\left(\left\langle\nabla f(x_{k+1})-\nabla f(x_k),\nabla f(x_{k+1})\right\rangle + \frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\|\nabla f(x_{k+1})\right\|^2\right)\end{aligned}$$

$$
= \underbrace{\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\cdot s\cdot\left\langle \nabla f(x_{k+1})-\nabla f(x_k), \nabla f(x_{k+1})\right\rangle}_{\mathbf{IV}_1}
+ \underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\cdot\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot s\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{IV}_2}
$$

For convenience, we note the term $\mathbf{IV} = (1/2)\mathbf{I}_1 + \mathbf{III}_2 + \mathbf{III}_3$. Then, with phase representation of NAG-$\texttt{SC}$ (5.17), the difference of Lyapunov function (5.18) is

$$
\begin{aligned}
&\mathcal{E}(k+1)-\mathcal{E}(k)\\
\leq & \underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\left\langle \nabla f(x_{k+1}), x_{k+1}-x^\star\right\rangle - s\left\|\nabla f(x_{k+1})\right\|^2\right)+\left\|v_{k+1}\right\|^2\right)}_{\mathbf{II}_1+\mathbf{III}_1+\mathbf{IV}_2}\\
& \underbrace{-\frac{1}{2}\cdot\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\left\langle \nabla f(x_{k+1})-\nabla f(x_k), x_{k+1}-x_k\right\rangle}_{\mathbf{II}_2}\\
& + \underbrace{\left(\frac{1}{1-\sqrt{\mu s}}\right)s\left\langle \nabla f(x_{k+1})-\nabla f(x_k), \nabla f(x_{k+1})\right\rangle}_{\mathbf{II}_4+\mathbf{IV}_1}\\
& \underbrace{-\frac{1}{2}\left[\frac{1}{L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)-s\left(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\right)\right]\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2}_{\mathbf{I}_2+\mathbf{II}_3}\\
& \underbrace{-\frac{1}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)s\left(\left\|\nabla f(x_{k+1})\right\|^2-\left\|\nabla f(x_k)\right\|^2\right)}_{\textbf{additional term}}
\end{aligned}
$$

Now, we can find the impact of additional term in the Lyapunov function (5.18). In other words, the $\mathbf{II}_4 + \mathbf{IV}_1$ term added the additional term is a perfect square, as below

$$
\begin{aligned}
\mathbf{II}_4+\mathbf{IV}_1+\textbf{additional term} &= \left(\frac{1}{1-\sqrt{\mu s}}\right)s\left\langle \nabla f(x_{k+1})-\nabla f(x_k), \nabla f(x_{k+1})\right\rangle\\
&\quad -\frac{1}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)s\left(\left\|\nabla f(x_{k+1})\right\|^2-\left\|\nabla f(x_k)\right\|^2\right)\\
&= \frac{1}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)s\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2
\end{aligned}
$$

Merging all the similar items, $\mathbf{II}_4 + \mathbf{IV}_1 + \textbf{additional term}$, $\mathbf{I}_2 + \mathbf{II}_3$, we have

$$\begin{aligned}
& (\mathbf{II}_4 + \mathbf{IV}_1 + \textbf{additional term}) + (\mathbf{I}_2 + \mathbf{II}_3) \\
= \quad & \frac{1}{2}\left(\frac{1}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{1}{Ls} \right) s \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 \\
\leq \quad & \frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{1}{Ls} \right) s \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2
\end{aligned}$$

Now, we obtain that the difference of Lyapunov function (5.18) is

$$\begin{aligned}
& \mathcal{E}(k+1) - \mathcal{E}(k) \\
& \leq - \frac{\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(\langle \nabla f(x_{k+1}), x_{k+1} - x^\star \rangle - s \left\| \nabla f(x_{k+1}) \right\|^2 \right) + \left\| v_{k+1} \right\|^2 \right) \\
& \quad - \frac{1}{2} \cdot \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \langle \nabla f(x_{k+1}) - \nabla f(x_k), x_{k+1} - x_k \rangle \\
& \quad + \frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{1}{Ls} \right) s \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2
\end{aligned}$$

With the inequality for any function $f(x) \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$

$$\left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 \leq L \left\langle \nabla f(x_{k+1}) - \nabla f(x_k), x_{k+1} - x_k \right\rangle,$$

we have

$$\begin{aligned}
& \mathcal{E}(k+1) - \mathcal{E}(k) \\
& \leq - \frac{\sqrt{\mu s}}{1-\sqrt{\mu s}} \left[\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(\langle \nabla f(x_{k+1}), x_{k+1} - x^\star \rangle - s \left\| \nabla f(x_{k+1}) \right\|^2 \right) + \left\| v_{k+1} \right\|^2 \right] \\
& \quad - \frac{1}{2} \cdot \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \cdot \frac{1}{L} \cdot \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 \\
& \quad + \frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{1}{Ls} \right) s \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 \\
& \leq - \frac{\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(\langle \nabla f(x_{k+1}), x_{k+1} - x^\star \rangle - s \left\| \nabla f(x_{k+1}) \right\|^2 \right) + \left\| v_{k+1} \right\|^2 \right) \\
& \quad - \frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \right) \left(\frac{1}{L} - s \right) \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 .
\end{aligned}$$

5.7.2.3 Proof of Lemma 5.3.9

With the phase representation of the heavy-ball method (5.29) and Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left\| v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^\star) \right\|_2^2 &= \left\| \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_k - x^\star) \right\|_2^2 \\ &\leq 2\left[\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 \|v_k\|_2^2 + \frac{4\mu}{(1-\sqrt{\mu s})^2} \|x_k - x^\star\|_2^2 \right].\end{aligned}$$

The discrete Lyapunov function (5.28) can be estimated as

$$\mathcal{E}(k) \leq \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(f(x_k) - f(x^\star)\right) + \frac{1+\mu s}{(1-\sqrt{\mu s})^2} \|v_k\|_2^2 + \frac{2\mu}{(1-\sqrt{\mu s})^2} \|x_k - x^\star\|_2^2 . \tag{5.104}$$

For convenience, we also split the discrete Lyapunov function (5.28) into three parts and mark them as below

$$\mathcal{E}(k) = \underbrace{\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(f(x_k) - f(x^\star)\right)}_{\mathbf{I}} + \underbrace{\frac{1}{4}\|v_k\|^2}_{\mathbf{II}} + \underbrace{\frac{1}{4}\left\| v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^\star)\right\|^2}_{\mathbf{III}},$$

where the three parts **I**, **II** and **III** are corresponding to potential, kinetic energy and mixed energy in classical mechanics, respectively.

- For the part **I**, potential, with the basic convex of $f(x) \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$

  $$f(x_k) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + \frac{1}{2L} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|_2^2 ,$$

  we have

  $$\begin{aligned}&\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_{k+1}) - f(x^\star)\right) - \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_k) - f(x^\star)\right) \\ \leq &\underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\,\langle \nabla f(x_{k+1}), v_k \rangle}_{\mathbf{I}_1} \underbrace{- \frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2}_{\mathbf{I}_2}.\end{aligned}$$

- For the part **II**, kinetic energy, with the phase representation of the heavy-ball method (5.29), we have

$$
\begin{aligned}
\frac{1}{4}\left\|v_{k+1}\right\|^2 - \frac{1}{4}\left\|v_k\right\|^2 &= \frac{1}{2}\left\langle v_{k+1} - v_k, v_{k+1}\right\rangle - \frac{1}{4}\left\|v_{k+1} - v_k\right\|^2 \\
&= \underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\|v_{k+1}\right\|^2}_{\mathbf{II}_1} \underbrace{-\frac{1}{2}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\sqrt{s}\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle}_{\mathbf{II}_2} \\
&\quad \underbrace{-\frac{1}{4}\left\|v_{k+1} - v_k\right\|^2}_{\mathbf{II}_3}
\end{aligned}
$$

- For the part **III**, mixed energy, with the phase representation of the heavy-ball method (5.29), we have

$$
\begin{aligned}
&\frac{1}{4}\left\|v_{k+1} + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+2} - x^\star)\right\|^2 - \frac{1}{4}\left\|v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^\star)\right\|^2 \\
=&\frac{1}{4}\left\langle v_{k+1} - v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+2} - x_{k+1}), v_{k+1} + v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+2} + x_{k+1} - 2x^\star)\right\rangle \\
=&-\frac{1}{2}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\sqrt{s}\left\langle \nabla f(x_{k+1}), v_{k+1} + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+2} - x^\star)\right\rangle \\
&-\frac{s}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2 \\
=&\underbrace{-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\right\rangle}_{\mathbf{III}_1} \underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\sqrt{s}\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle}_{\mathbf{III}_2} \\
&\underbrace{-\frac{s}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_3}
\end{aligned}
$$

Now, we calculate the difference of discrete Lyapunov function (5.18) at the k-th iteration by the simple operation as

$$
\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
\leq&\underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\left\langle \nabla f(x_{k+1}), v_k\right\rangle}_{\mathbf{I}_1} \underbrace{-\frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2}_{\mathbf{I}_2}
\end{aligned}
$$

$$
\begin{aligned}
&\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\|v_{k+1}\right\|^2}_{\mathbf{II}_1}\underbrace{-\frac{1}{2}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\sqrt{s}\left\langle\nabla f(x_{k+1}),v_{k+1}\right\rangle}_{\mathbf{II}_2}\underbrace{-\frac{1}{4}\left\|v_{k+1}-v_k\right\|^2}_{\mathbf{II}_3}\\
&\underbrace{-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle\nabla f(x_{k+1}),x_{k+1}-x^\star\right\rangle}_{\mathbf{III}_1}\\
&\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\sqrt{s}\left\langle\nabla f(x_{k+1}),v_{k+1}\right\rangle}_{\mathbf{III}_2}\\
&\underbrace{-\frac{s}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_3}\\
=&\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle\nabla f(x_{k+1}),x_{k+1}-x^\star\right\rangle+\left\|v_{k+1}\right\|^2\right)}_{\mathbf{II}_1+\mathbf{III}_1}\\
&\underbrace{-\frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2}_{\mathbf{I}_2}\\
&\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\left\langle\nabla f(x_{k+1}),\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)v_{k+1}-v_k\right\rangle}_{\frac{1}{2}\mathbf{I}_1+\mathbf{III}_2}\\
&\underbrace{-\frac{1}{4}\left(\left\|v_{k+1}-v_k\right\|^2+2\sqrt{s}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle\nabla f(x_{k+1}),v_{k+1}-v_k\right\rangle\right.}_{\frac{1}{2}\mathbf{I}_1+\mathbf{II}_2+\mathbf{II}_3+\mathbf{III}_3}\\
&\underbrace{\left.+s\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2\right)}_{\frac{1}{2}\mathbf{I}_1+\mathbf{II}_2+\mathbf{II}_3+\mathbf{III}_3}
\end{aligned}
$$

With the phase representation of the heavy-ball method (5.29), we have

$$
\begin{aligned}
\frac{1}{2}\mathbf{I}_1+\mathbf{III}_2&=-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\left\langle\nabla f(x_{k+1}),\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)v_{k+1}-v_k\right\rangle\\
&=\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2;
\end{aligned}
$$

and

$$
\frac{1}{2}\mathbf{I}_1+\mathbf{II}_2+\mathbf{II}_3+\mathbf{III}_3
$$

$$
\begin{aligned}
=& -\frac{1}{4}\left[\left\|v_{k+1}-v_k\right\|^2+2\sqrt{s}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle\nabla f(x_{k+1}),v_{k+1}-v_k\right\rangle+s\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2\right]\\
=& -\frac{1}{4}\left\|v_{k+1}-v_k+\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\sqrt{s}\nabla f(x_{k+1})\right\|^2\\
\leq& 0.
\end{aligned}
$$

Now, the difference of discrete Lyapunov function (5.28) can be rewritten as

$$
\begin{aligned}
\mathcal{E}(k+1)-\mathcal{E}(k)\leq& -\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle\nabla f(x_{k+1}),x_{k+1}-x^\star\right\rangle+\left\|v_{k+1}\right\|^2\right)\\
&-\frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2\\
&+\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned}
$$

With the inequality for any function $f(x)\in\mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$

$$
f(x^\star)\geq f(x_{k+1})+\left\langle\nabla f(x_{k+1}),x^\star-x_{k+1}\right\rangle+\frac{\mu}{2}\left\|x_{k+1}-x^\star\right\|^2,
$$

we have

$$
\begin{aligned}
&\mathcal{E}(k+1)-\mathcal{E}(k)\\
\leq& -\sqrt{\mu s}\left[\frac{1+\sqrt{\mu s}}{(1-\sqrt{\mu s})^2}\left(f(x_{k+1})-f(x^\star)\right)+\frac{\mu}{2}\cdot\frac{1+\sqrt{\mu s}}{(1-\sqrt{\mu s})^2}\left\|x_{k+1}-x^\star\right\|^2\right.\\
&\left.+\frac{1}{1-\sqrt{\mu s}}\left\|v_{k+1}\right\|^2\right]+\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2\\
\leq& -\sqrt{\mu s}\left[\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_{k+1})-f(x^\star)\right)+\frac{\mu}{2}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\|x_{k+1}-x^\star\right\|^2\right.\\
&\left.+\frac{1}{1-\sqrt{\mu s}}\left\|v_{k+1}\right\|^2\right]+\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2\\
\leq& -\sqrt{\mu s}\left[\frac{1}{4}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_{k+1})-f(x^\star)\right)+\frac{1}{1-\sqrt{\mu s}}\left\|v_{k+1}\right\|^2\right.\\
&\left.+\frac{\mu}{2}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\|x_{k+1}-x^\star\right\|^2\right]\\
&-\left[\frac{3}{4}\sqrt{\mu s}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_{k+1})-f(x^\star)\right)-\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\left\|\nabla f(x_{k+1})\right\|^2\right].
\end{aligned}
$$

Comparing the coefficient of the estimate of Lyapunov function (5.104), we have

$$
\begin{aligned}
&\mathcal{E}(k+1)-\mathcal{E}(k)\\
\leq &-\sqrt{\mu s}\min\left\{\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}, \frac{1}{4}\right\}\mathcal{E}(k+1)\\
&-\left[\frac{3}{4}\sqrt{\mu s}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)(f(x_{k+1})-f(x^\star))-\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\|\nabla f(x_{k+1})\|^2\right].
\end{aligned}
$$

The proof is complete.

5.7.3 Technical Details in Section 5.4

5.7.3.1 Technical Details in Proof of Theorem 5.4.4

5.7.3.1.1 Iterates (x_k, y_k) at $k = 1, 2, 3$ The iterate (x_k, y_k) at $k = 1$ is

$$x_1 = y_1 = x_0 - s\nabla f(x_0). \tag{5.105}$$

When $k = 2$, the iterate (x_k, y_k) is

$$\begin{cases} y_2 = x_0 - s\nabla f(x_0) - s\nabla f(x_0 - s\nabla f(x_0)) \\ x_2 = x_0 - s\nabla f(x_0) - \dfrac{5}{4}s\nabla f(x_0 - s\nabla f(x_0)). \end{cases} \tag{5.106}$$

When $k = 3$, the iterate (x_k, y_k) is

$$\begin{cases} y_3 = x_0 - s\nabla f(x_0) - \dfrac{5}{4}s\nabla f(x_0 - s\nabla f(x_0)) \\ \qquad\qquad - s\nabla f\left(x_0 - s\nabla f(x_0) - \dfrac{5}{4}s\nabla f(x_0 - s\nabla f(x_0))\right) \\ x_3 = x_0 - s\nabla f(x_0) - \dfrac{27}{20}s\nabla f(x_0 - s\nabla f(x_0)) \\ \qquad\qquad - \dfrac{7}{5}s\nabla f\left(x_0 - s\nabla f(x_0) - \dfrac{5}{4}s\nabla f(x_0 - s\nabla f(x_0))\right). \end{cases} \tag{5.107}$$

5.7.3.1.2 Estimate For $\|\nabla f(x_k)\|^2$ at $k = 0, 1, 2, 3$ According to (5.105), we have

$$\|\nabla f(x_1)\|^2 = \|\nabla f(x_0 - s\nabla f(x_0))\|^2 \quad\leq\quad L^2\|x_0 - x^\star - s\nabla f(x_0)\|^2$$

$$
\begin{aligned}
&\leq 2L^2\left(\|x_0 - x^\star\|^2 + s^2\|\nabla f(x_0)\|^2\right) \\
&\leq 2L^2(1 + L^2s^2)\|x_0 - x^\star\|^2. \qquad (5.108)
\end{aligned}
$$

According to (5.106), we have

$$
\begin{aligned}
\|\nabla f(x_2)\|^2 &= \left\|\nabla f\left(x_0 - s\nabla f(x_0) - \frac{5}{4}s\nabla f\left(x_0 - s\nabla f(x_0)\right)\right)\right\|^2 \\
&\leq L^2\left\|x_0 - x^\star - s\nabla f(x_0) - \frac{5}{4}s\nabla f\left(x_0 - s\nabla f(x_0)\right)\right\|^2 \\
&\leq 3L^2\left(\|x_0 - x^\star\|^2 + s^2\|\nabla f(x_0)\|^2 + \frac{25}{16}s^2\|\nabla f(x_0 - s\nabla f(x_0))\|^2\right) \\
&\leq 3L^2\left[(1 + L^2s^2)\|x_0 - x^\star\|^2 + \frac{25}{16}L^2s^2\|x_0 - x^\star - s\nabla f(x_0)\|^2\right] \\
&\leq 3L^2\left[(1 + L^2s^2)\|x_0 - x^\star\|^2 + \frac{25}{8}L^2s^2\left(\|x_0 - x^\star\|^2 + s^2\|\nabla f(x_0)\|^2\right)\right] \\
&\leq 3L^2\left(1 + \frac{33}{8}L^2s^2 + \frac{25}{8}L^4s^4\right)\|x_0 - x^\star\|^2. \qquad (5.109)
\end{aligned}
$$

With (5.105)-(5.107), we have

$$
\begin{aligned}
&\|\nabla f(x_3)\|^2 \qquad (5.110) \\
\leq& L^2\|x_3 - x^\star\|^2 \\
\leq& L^2\left\|x_0 - x^\star - s\nabla f(x_0) - \frac{27}{20}s\nabla f(x_1) - \frac{7}{5}s\nabla f\left(x_2\right)\right\|^2 \\
=& 4L^2\left(\|x_0 - x^\star\|^2 + s^2\|\nabla f(x_0)\|^2 + \frac{729}{400}s^2\|\nabla f(x_1)\|^2 + \frac{49}{25}s^2\|\nabla f(x_2)\|^2\right) \\
=& 4L^2\left[1 + L^2s^2 + \frac{729}{200}L^2s^2(1 + L^2s^2) + \frac{147}{25}L^2s^2\left(1 + \frac{33}{8}L^2s^2 + \frac{25}{8}L^4s^4\right)\right]\|x_0 - x^\star\|^2 \\
=& \frac{L^2(40 + 381L^2s^2 + 1156L^4s^4 + 735L^6s^6)}{10}\|x_0 - x^\star\|^2. \qquad (5.111)
\end{aligned}
$$

Taking $s \leq 1/(3L)$ and using (5.108), (5.109) and (5.110), we have

$$
\begin{aligned}
\|\nabla f(x_0)\|^2 &\leq \frac{\|x_0 - x^\star\|^2}{9s^2}, & \|\nabla f(x_1)\|^2 &\leq \frac{20\|x_0 - x^\star\|^2}{81s^2}, \\
\|\nabla f(x_2)\|^2 &\leq \frac{485\|x_0 - x^\star\|^2}{972s^2}, & \|\nabla f(x_3)\|^2 &\leq \frac{2372\|x_0 - x^\star\|^2}{2187s^2}.
\end{aligned}
$$

5.7.3.1.3 Estimate For $f(x_k) - f(x^\star)$ at $k = 0, 1$

According to (5.105), we have

$$\begin{aligned} f(x_1) - f(x^\star) &\leq \frac{L}{2}\left\|x_1 - x^\star\right\|^2 \\ &\leq \frac{L}{2}\left\|x_0 - s\nabla f(x_0) - x^\star\right\|^2 \\ &\leq L\left(\left\|x_0 - x^\star\right\|^2 + s^2\left\|\nabla f(x_0)\right\|^2\right) \\ &\leq L(1 + L^2s^2)\left\|x_0 - x^\star\right\|^2. \end{aligned} \tag{5.112}$$

Taking $s \leq 1/(3L)$, (5.112) tells us that

$$f(x_0) - f(x^\star) \leq \frac{\left\|x_0 - x^\star\right\|^2}{6s}, \qquad f(x_1) - f(x^\star) \leq \frac{10\left\|x_0 - x^\star\right\|^2}{27s}.$$

5.7.3.1.4 Estimate for Lyapunov function $\mathcal{E}(2)$ and $\mathcal{E}(3)$

With the phase-space representation form (5.45), we have

$$v_2 = \frac{x_3 - x_2}{\sqrt{s}} = \frac{1}{10}\nabla f(x_1) + \frac{7}{5}\nabla f(x_2). \tag{5.113}$$

According to (5.46), the Lyapunov function $\mathcal{E}(2)$ can be written as

$$\mathcal{E}(2) = 15s\left(f(x_2) - f(x^\star)\right) + \frac{1}{2}\left\|2(x_2 - x^\star) + 5\sqrt{s}v_2 + 3s\nabla f(x_2)\right\|^2.$$

With (5.113) and Cauchy-Schwarz inequality, we have

$$\begin{aligned} \mathcal{E}(2) &\leq \frac{15Ls}{2}\left\|x_2 - x^\star\right\|^2 + \frac{3}{2}\left(4\left\|x_2 - x^\star\right\|^2 + 25s\left\|v_2\right\|^2 + 9s^2\left\|\nabla f(x_2)\right\|^2\right) \\ &\leq \left(\frac{15Ls}{2} + 6\right)\left\|x_2 - x^\star\right\|^2 + \frac{27}{2}s^2\left\|\nabla f(x_2)\right\|^2 + \frac{75}{2}s^2\left\|\frac{1}{10}\nabla f(x_1) + \frac{7}{5}\nabla f(x_2)\right\|^2 \\ &\leq \left(\frac{15Ls}{2} + 6\right)\left\|x_2 - x^\star\right\|^2 + \frac{27}{2}s^2\left\|\nabla f(x_2)\right\|^2 + \frac{3}{4}s^2\left\|\nabla f(x_1)\right\|^2 + 147s^2\left\|\nabla f(x_2)\right\|^2 \\ &= \left(\frac{15Ls}{2} + 6\right)\left\|x_2 - x^\star\right\|^2 + \frac{321}{2}s^2\left\|\nabla f(x_2)\right\|^2 + \frac{3}{4}s^2\left\|\nabla f(x_1)\right\|^2. \end{aligned}$$

Furthermore, with (5.106), we have

$$\begin{aligned} \mathcal{E}(2) \leq &\left(\frac{15Ls}{2} + 6\right)\left\|x_0 - x^\star - s\nabla f(x_0) - \frac{5}{4}s\nabla f(x_0 - s\nabla f(x_0))\right\|^2 \\ &+ \frac{321}{2}s^2\left\|\nabla f(x_2)\right\|^2 + \frac{3}{4}s^2\left\|\nabla f(x_1)\right\|^2. \end{aligned}$$

Finally, with (5.108)-(5.109), Cauchy-Schwarz inequality tells

$$\begin{aligned}
&\mathcal{E}(2) && (5.114)\\
\leq &\left\{ \left[\frac{3}{16}\left(12+15Ls\right) + \frac{963}{16}L^2s^2 \right] \left(8+33L^2s^2+25L^4s^4\right) + \frac{3}{2}L^2s^2(1+L^2s^2) \right\}\\
&\cdot \left\|x_0 - x^\star\right\|^2\\
= &\frac{288+360Ls+8916L^2s^2+1485L^3s^3+32703L^4s^4+1125L^5s^5+24075L^6s^6}{16}\\
&\cdot \left\|x_0 - x^\star\right\|^2. && (5.115)
\end{aligned}$$

By Lemma 5.4.5, when the step size $s \leq 1/(3L)$, (5.114) tells us

$$\mathcal{E}(3) \leq \mathcal{E}(2) \leq 119 \left\|x_0 - x^\star\right\|^2.$$

5.7.3.2 Proof of Theorem 5.4.6

Let $w_k = (1/2)\left[(k+2)x_k - ky_k + (k-1)s\nabla f(y_k)\right]$ for convenience. Using the dynamics of $\{(x_k, y_k)\}_{k=0}^{\infty}$ generated by the modified NAG-$\mathtt{C}$ (5.55), we have

$$\begin{aligned}
w_{k+1} &= \frac{1}{2}\left[(k+3)x_{k+1} - (k+1)y_{k+1} + sk\nabla f(y_{k+1})\right]\\
&= \frac{1}{2}\left[(k+3)\left(y_{k+1} + \frac{k}{k+3}(y_{k+1}-y_k) - \frac{sk}{k+3}\nabla f(y_{k+1})\right.\right.\\
&\qquad\qquad \left.\left. + \frac{s(k-1)}{k+3}\nabla f(y_k)\right) - (k+1)y_{k+1} + sk\nabla f(y_{k+1})\right]\\
&= \frac{1}{2}\left[(k+2)y_{k+1} - ky_k + s(k-1)\nabla f(y_{k-1})\right]\\
&= w_k - \frac{s(k+2)}{2}\nabla f(x_k).
\end{aligned}$$

Hence, the difference between $\|w_{k+1} - x^\star\|^2$ and $\|w_k - x^\star\|^2$ is

$$\begin{aligned}
&\frac{1}{2}\|w_{k+1} - x^\star\|^2 - \frac{1}{2}\|w_k - x^\star\|^2 \\
=&\left\langle w_{k+1} - w_k, \frac{w_{k+1} + w_k}{2} - x^\star \right\rangle \\
=&\frac{s^2(k+2)^2}{8}\|\nabla f(x_k)\|^2 - \frac{s(k+2)}{2}\langle \nabla f(x_k), w_k - x^\star\rangle \\
=&\frac{s^2(k+2)^2}{8}\|\nabla f(x_k)\|^2 - \frac{s^2(k-1)(k+2)}{4}\langle \nabla f(x_k), \nabla f(y_k)\rangle \\
&\quad - \frac{s(k+2)}{4}\langle \nabla f(x_k), (k+2)x_k - ky_k - 2x^\star\rangle .
\end{aligned}$$

If the step size satisfies $s \leq 1/L$, there exists a tighter basic inequality than [SBC16, Equation (22)] and [B+15, Lemma 3.6] for any function $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$

$$f(x - s\nabla f(x)) \leq f(y) + \langle \nabla f(x), x - y\rangle - \frac{s}{2}\|\nabla f(x)\|^2 - \frac{s}{2}\|\nabla f(x) - \nabla f(y)\|^2 . \tag{5.116}$$

With (5.116), we can obtain that

$$\begin{aligned}
(k+2)\left(f(y_{k+1}) - f(x^\star)\right) - k\left(f(y_k) - f(x^\star)\right) \leq& \langle \nabla f(x_k), (k+2)x_k - ky_k - 2x^\star\rangle \\
&- \frac{s(k+2)}{2}\|\nabla f(x_k)\|^2 - \frac{sk}{2}\|\nabla f(x_k) - \nabla f(y_k)\|^2 .
\end{aligned}$$

Consider the discrete Lyapunov function

$$\mathcal{E}(k) = \frac{s(k+1)^2}{4}\left(f(y_k) - f(x^\star)\right) + \frac{1}{2}\|w_k - x^\star\|^2 . \tag{5.117}$$

Hence, the difference between $\mathcal{E}(k+1)$ and $\mathcal{E}(k)$ in (5.117) is

$$\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) =& -\frac{1}{4}\left(f(y_k) - f(x^\star)\right) - \frac{s^2(k-1)(k+2)}{2}\langle \nabla f(x_k), \nabla f(y_k)\rangle \\
&- \frac{s^2k(k+2)}{8}\|\nabla f(x_k) - \nabla f(y_k)\|^2 \\
\leq& -\frac{1}{4}\left(f(y_k) - f(x^\star)\right) - \frac{s^2(k-1)(k+2)}{8}\|\nabla f(x_k) + \nabla f(y_k)\|^2 .
\end{aligned} \tag{5.118}$$

When $k \geq 2$, we have

$$\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(2) &= \sum_{i=2}^{k} \left(\mathcal{E}(i+1) - \mathcal{E}(i)\right) \\
&\leq -\sum_{i=2}^{k} \frac{s^2(i-1)(i+2)}{8} \left\|\nabla f(x_i) + \nabla f(y_i)\right\|^2 \\
&\leq -\frac{s^2}{8} \min_{2\leq i\leq k} \left\|\nabla f(x_i) + \nabla f(y_i)\right\|^2 \sum_{i=2}^{k}(i-1)(i+2) \\
&\leq -\frac{s^2}{24} \min_{2\leq i\leq k} \left\|\nabla f(x_i) + \nabla f(y_i)\right\|^2 \cdot k(k^2+3k-4) \\
&\leq -\frac{s^2}{24} \min_{2\leq i\leq k} \left\|\nabla f(x_i) + \nabla f(y_i)\right\|^2 \cdot \frac{(k+1)^3}{7} \\
&= -\frac{s^2(k+1)^3}{168} \min_{2\leq i\leq k} \left\|\nabla f(x_i) + \nabla f(y_i)\right\|^2.
\end{aligned}$$

Furthermore, we have

$$\min_{2\leq i\leq k} \left\|\nabla f(x_i) + \nabla f(y_i)\right\|^2 \leq \frac{168\left[\mathcal{E}(2) - \mathcal{E}(k+1)\right]}{s^2(k+1)^3} \leq \frac{168\mathcal{E}(2)}{s^2(k+1)^3}.$$

Combining with (5.118), we obtain that

$$\begin{aligned}
&\min_{2\leq i\leq k} \left\|\nabla f(x_i) + \nabla f(y_i)\right\|^2 \\
\leq& \frac{168\mathcal{E}(1)}{s^2(k+1)^3} \\
\leq& \frac{168}{s^2(k+1)^3}\left[s\left(f(y_1) - f(x^\star)\right) + \frac{1}{2}\left\|w_1 - x^\star\right\|^2\right] \\
\leq& \frac{168}{s^2(k+1)^3}\left(\frac{Ls}{2}\left\|y_1 - x^\star\right\|^2 + \frac{1}{2}\left\|w_0 - s\nabla f(x_0) - x^\star\right\|^2\right) \\
=& \frac{168}{s^2(k+1)^3}\left(\frac{Ls}{2}\left\|x_0 - s\nabla f(x_0) - x^\star\right\|^2 + \frac{1}{2}\left\|x_0 - \frac{3s}{2}\nabla f(x_0) - x^\star\right\|^2\right) \\
\leq& \frac{882\left\|x_0 - x^\star\right\|^2}{s^2(k+1)^3}.
\end{aligned}$$

Similarly, when $s \leq 1/L$, for $k = 0$, we have

$$\left\|\nabla f(x_0) + \nabla f(y_0)\right\|^2 = 4\left\|\nabla f(x_0)\right\|^2 \leq \frac{4\left\|x_0 - x^\star\right\|^2}{s^2};$$

for $k = 1$, following the modified NAG-$\texttt{C}$ (5.55), we obtain (x_1, y_1) as

$$y_1 = x_0 - s\nabla f(x_0), \quad x_1 = x_0 - \frac{4}{3}s\nabla f(x_0),$$

furthermore we have

$$\begin{aligned}
\|\nabla f(x_1) + \nabla f(y_1)\|^2 &\leq 2\left(\|\nabla f(x_1)\|^2 + \|\nabla f(y_1)\|^2\right) \\
&\leq \frac{2}{s^2}\left(\|x_1 - x^\star\|^2 + \|y_1 - x^\star\|^2\right) \\
&\leq \frac{4}{s^2}\left[\left(1 + L^2s^2\right)\|x_0 - x^\star\|^2 + \left(1 + (16/9)L^2s^2\right)\|x_0 - x^\star\|^2\right] \\
&\leq \frac{172s^2\|x_0 - x^\star\|^2}{9}.
\end{aligned}$$

For function value, (5.118) tells

$$f(y_k) - f(x^\star) \leq \frac{4\mathcal{E}(1)}{s(k+1)^2} \leq \frac{21\|x_0 - x^\star\|^2}{s(k+1)^2}$$

for all $k \geq 1$. Together with

$$f(y_0) - f(x^\star) \leq \frac{\|x_0 - x^\star\|^2}{s},$$

we complete the proof.

5.7.3.3 Nesterov's Lower Bound

Recall [Nes13, Theorem 2.1.7], for any k, $1 \leq k \leq (1/2)(n-1)$, and any $x_0 \in \mathbb{R}^n$, there exists a function $f \in \mathcal{F}_L^1(\mathbb{R}^n)$ such that any first-order method obeys

$$f(x_k) - f(x^\star) \geq \frac{3L\|x_0 - x^\star\|^2}{32(k+1)^2}.$$

Using the basic inequality for $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$,

$$\|\nabla f(x_k)\|\|x_k - x^\star\| \geq \langle \nabla f(x_k), x_k - x^\star \rangle \geq f(x_k) - f(x^\star),$$

we have

$$\|\nabla f(x_k)\| \geq \frac{3L\|x_0 - x^\star\|^2}{32(k+1)^2 \max\limits_{1 \leq k \leq \frac{n-1}{2}} \|x_k - x^\star\|}$$

for $1 \leq k \leq (1/2)(n-1)$.

5.7.4 Technical Details in Section 5.5

5.7.4.1 Proof of Theorem 5.5.1: Case $\alpha = 3$

Before starting to prove Theorem 5.5.1, we first look back our high-resolution ODE framework in Section 5.2.

- **Step** 1, the generalized high-resolution ODE has been given in (5.56).
- **Step** 2, the continuous Lyapunov function is constructed as

  $$\begin{aligned}\mathcal{E}(t) = t\left[t + \left(\frac{3}{2} - \beta\right)\sqrt{s}\right]\left(f(X(t)) - f(x^\star)\right) \\ + \frac{1}{2}\left\|2(X(t) - x^\star) + t\left(\dot{X}(t) + \beta\sqrt{s}\nabla f(X(t))\right)\right\|^2. \quad (5.119)\end{aligned}$$

  Following this Lyapunov function (5.119), we can definitely obtain similar results as Theorem 5.4.1 and Corollary 5.4.3. The detailed calculation, about the estimate of the optimal constant β and how the constant β influence the initial point, is left for readers.

- **Step** 3, before constructing discrete Lyapunov functions, we show the phase-space representation (5.57) as

  $$\begin{aligned}&x_k - x_{k-1} = \sqrt{s}v_{k-1} \\ &v_k - v_{k-1} = -\frac{\alpha}{k}v_k - \beta\sqrt{s}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \left(1 + \frac{\alpha}{k}\right)\sqrt{s}\nabla f(x_k).\end{aligned} \quad (5.120)$$

Now, we show how to construct the discrete Lyapunov function and analyze the algorithms (5.57) with $\alpha = 3$ in order to prove Theorem 5.5.1.

5.7.4.1.1 Case: $\beta < 1$ When $\beta < 1$, we know that the function

$$g(k) = \frac{k+3}{k+3-\beta}$$

decreases monotonically. Hence we can construct the discrete Lyapunov function as

$$\begin{aligned}\mathcal{E}(k) = s(k+4)(k+1)\left(f(x_k) - f(x^\star)\right)& \\ + \frac{k+3}{2(k+3-\beta)}\left\|2(x_{k+1} - x^\star) + \sqrt{s}(k+1)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right)\right\|^2&, \quad (5.121)\end{aligned}$$

which is slightly different from the discrete Lyapunov function (5.46) for NAG-C. When $\beta \to 1$, the discrete Lyapunov function (5.121) approximate to (5.46) as $k \to \infty$.

With the phase-space representation (5.120) for $\alpha = 3$, we can obtain

$$(k+3)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right) - k\left(v_{k-1} + \beta\sqrt{s}\nabla f(x_{k-1})\right) = -\sqrt{s}\left(k+3-3\beta\right)\nabla f(x_k). \quad (5.122)$$

The difference of the discrete Lyapunov function (5.121) of the k-th iteration is

$$\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
=&s(k+5)(k+2)\left(f(x_{k+1}) - f(x^\star)\right) - s(k+4)(k+1)\left(f(x_k) - f(x^\star)\right) \\
&+ \frac{k+4}{2(k+4-\beta)}\left\|2(x_{k+2} - x^\star) + \sqrt{s}(k+2)\left(v_{k+1} + \beta\sqrt{s}\nabla f(x_{k+1})\right)\right\|^2 \\
&- \frac{k+3}{2(k+3-\beta)}\left\|2(x_{k+1} - x^\star) + \sqrt{s}(k+1)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right)\right\|^2 \\
\leq&s\left(k+4\right)\left(k+1\right)\left(f(x_{k+1}) - f(x_k)\right) + s(2k+6)\left(f(x_{k+1}) - f(x^\star)\right) \\
&+ \frac{k+4}{k+4-\beta}\Big[\Big\langle 2(x_{k+2} - x_{k+1}) + \sqrt{s}(k+2)\left(v_{k+1} + \beta\sqrt{s}\nabla f(x_{k+1})\right) \\
&\qquad\qquad - \sqrt{s}(k+1)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right), \\
&\qquad\qquad 2(x_{k+2} - x^\star) + \sqrt{s}(k+2)\left(v_{k+1} + \beta\sqrt{s}\nabla f(x_{k+1})\right)\Big\rangle \\
&\qquad\qquad -\frac{1}{2}\Big\|2(x_{k+2} - x_{k+1}) + \sqrt{s}(k+2)\left(v_{k+1} + \beta\sqrt{s}\nabla f(x_{k+1})\right) \\
&\qquad\qquad -\sqrt{s}(k+1)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right)\Big\|^2\Big] \\
=&s\left(k+4\right)\left(k+1\right)\left(f(x_{k+1}) - f(x_k)\right) + s(2k+6)\left(f(x_{k+1}) - f(x^\star)\right) \\
&- \left\langle s(k+4)\nabla f(x_{k+1}), 2(x_{k+2} - x^\star) + \sqrt{s}(k+2)\left(v_{k+1} + \beta\sqrt{s}\nabla f(x_{k+1})\right)\right\rangle \\
&- \frac{1}{2}s^2(k+4)\left(k+4-\beta\right)\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned}$$

With the basic inequality of any function $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$

$$\begin{cases} f(x_k) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + \dfrac{1}{2L} \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 \\ f(x^\star) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x^\star - x_{k+1} \rangle , \end{cases}$$

and the phase-space representation (5.120)

$$x_{k+2} = x_{k+1} + \sqrt{s} v_{k+1},$$

the difference of the discrete Lyapunov function (5.121) can be estimated as

$$\begin{aligned} & \mathcal{E}(k+1) - \mathcal{E}(k) \\ \leq & s(k+4)(k+1) \left(\langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle - \frac{1}{2L} \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 \right) \\ & + s(2k+6) \left(f(x_{k+1}) - f(x^\star) \right) - s(2k+8) \langle \nabla f(x_{k+1}), x_{k+1} - x^\star \rangle \\ & - s^{\frac{3}{2}} (k+4)^2 \langle \nabla f(x_{k+1}), v_{k+1} \rangle - \beta s^2 (k+2)(k+4) \left\| \nabla f(x_{k+1}) \right\|^2 \\ & - \frac{1}{2} s^2 (k+4) \left(k+4-\beta \right) \left\| \nabla f(x_{k+1}) \right\|^2 \\ \leq & - s^{\frac{3}{2}} (k+4) \langle \nabla f(x_{k+1}), (k+4) v_{k+1} - (k+1) v_k \rangle \\ & - \frac{s(k+4)(k+1)}{2L} \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 \\ & - 2s \left(f(x_{k+1}) - f(x^\star) \right) \\ & - s^2 \left[\beta (k+4)(k+2) + \frac{1}{2} (k+4) \left(k+4-\beta \right) \right] \left\| \nabla f(x_{k+1}) \right\|^2 . \end{aligned}$$

Utilizing the phase-space representation (5.120) again, we calculate the difference of the discrete Lyapunov function (5.121) as

$$\begin{aligned} & \mathcal{E}(k+1) - \mathcal{E}(k) \\ \leq & s^{\frac{3}{2}} (k+4) \left\langle \nabla f(x_{k+1}), \beta \sqrt{s} (k+1) \left(\nabla f(x_{k+1}) - \nabla f(x_k) \right) + \sqrt{s} (k+4) \nabla f(x_{k+1}) \right\rangle \\ & - \frac{s(k+4)(k+1)}{2L} \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 \\ & - s^2 \left[\beta (k+4)(k+2) + \frac{1}{2} (k+4) \left(k+4-\beta \right) \right] \left\| \nabla f(x_{k+1}) \right\|^2 \end{aligned}$$

$$
\begin{aligned}
\leq & \beta s^2(k+4)(k+1)\left\langle \nabla f(x_{k+1}), \nabla f(x_{k+1}) - \nabla f(x_k)\right\rangle \\
& - \frac{s(k+4)(k+1)}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
& - \left[(k+2)(k+4)\beta - \frac{1}{2}\left(k+4+\beta\right)(k+4)\right] s^2 \left\|\nabla f(x_{k+1})\right\|^2 \\
\leq & \frac{L\beta^2 s^3}{2}(k+4)(k+1)\left\|\nabla f(x_{k+1})\right\|^2 \\
& - \left[(k+2)(k+4)\beta - \frac{1}{2}\left(k+4+\beta\right)(k+4)\right] s^2 \left\|\nabla f(x_{k+1})\right\|^2 \\
= & - \left[\beta(k+2) - \frac{1}{2}\left(k+4+\beta\right) - \frac{L\beta^2 s}{2}(k+1)\right](k+4)s^2\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned}
$$

To guarantee that the Lyapunov function $\mathcal{E}(k)$ is decreasing, a sufficient condition is

$$
\beta(k+2) - \frac{1}{2}\left(k+4+\beta\right) - \frac{L\beta^2 s}{2}(k+1) \geq 0. \tag{5.123}
$$

Simple calculation tells us that (5.123) can be rewritten as

$$
s \leq \frac{(2\beta-1)k + 3\beta - 4}{(k+1)L\beta^2} = \frac{1}{L\beta^2}\left(2\beta - 1 + \frac{\beta-3}{k+1}\right). \tag{5.124}
$$

Apparently, when $\beta \to 1$, the step size satisfies

$$
0 < s \leq \frac{k-1}{k+1}\cdot\frac{1}{L}
$$

which is consistent with (5.48). Now, we turn to discuss the parameter $0 \leq \beta < 1$ case by case.

- When the parameter $\beta \leq 1/2$, the sufficient condition (5.123) for the Lyapunov function $\mathcal{E}(k)$ decreasing cannot be satisfied for sufficiently large k.
- When the parameter $1/2 < \beta < 1$, since the function $h(k) = \frac{1}{L\beta^2}\left(2\beta - 1 + \frac{\beta-3}{k+1}\right)$ increases monotonically for $k \geq 0$, there exists $k_{3,\beta} = \left\lfloor \frac{4-3\beta}{2\beta-1}\right\rfloor + 1$ such that the step size

$$
s \leq \frac{(2\beta-1)k_{3,\beta} + 3\beta - 4}{(k_{3,\beta}+1)L\beta^2}
$$

works for any $k \geq k_{3,\beta}$ ($k_{3,\beta} \to 2$ with $\beta \to 1$). Then, the difference of the discrete Lyapunov function (5.121) can be estimated as

$$\mathcal{E}(k+1) - \mathcal{E}(k) \leq -s^2 \left(\frac{2\beta - 1 - L\beta^2 s}{2} \right) (k - k_{3,\beta})^2 \left\| \nabla f(x_{k+1}) \right\|^2 .$$

Here, the proof is actually complete. Without loss of generality, we briefly show the expression is consistent with Theorem 5.5.1 and omit the proofs for the following facts. When $k \geq k_{3,\beta} + 1$, there exists some constant $\mathfrak{C}_{3,\beta}^0 > 0$ such that

$$\mathcal{E}(k+1) - \mathcal{E}(k) \leq -s^2 \mathfrak{C}_{3,\beta}^0 (k+1)^2 \left\| \nabla f(x_{k+1}) \right\|^2 .$$

For $k \leq k_{3,\beta}$, using mathematic induction, there also exists some constant $\mathfrak{C}_{3,\beta}^1 > 0$ such that for $s = O(1/L)$, we have

$$\left\| \nabla f(x_{k+1}) \right\|^2 \leq \frac{\mathfrak{C}_{3,\beta}^1 \left\| x_0 - x^\star \right\|^2}{s^2} \quad \text{and} \quad f(x_k) - f(x^\star) \leq \frac{\mathcal{E}(k)}{4s} \leq \frac{\mathfrak{C}_{3,\beta}^1 \left\| x_0 - x^\star \right\|^2}{s}.$$

5.7.4.1.2 Case: $\beta \geq 1$ When $\beta \geq 1$, we know that the function

$$g(k) = \frac{k+2}{k+3-\beta}$$

decreases monotonically. Hence we can construct the discrete Lyapunov function as

$$\begin{aligned} \mathcal{E}(k) = {} & s(k+3)(k+1) \left(f(x_k) - f(x^\star) \right) \\ & + \frac{k+2}{2(k+3-\beta)} \left\| 2(x_{k+1} - x^\star) + \sqrt{s}(k+1) \left(v_k + \beta \sqrt{s} \nabla f(x_k) \right) \right\|^2 . \end{aligned} \quad (5.125)$$

which for $\beta = 1$ is consistent with the discrete Lyapunov function (5.46) for NAG-C.

With the expression (5.122)

$$(k+3) \left(v_k + \beta \sqrt{s} \nabla f(x_k) \right) - k \left(v_{k-1} + \beta \sqrt{s} \nabla f(x_{k-1}) \right) = -\sqrt{s} \left(k + 3 - 3\beta \right) \nabla f(x_k),$$

the difference of the discrete Lyapunov function (5.125) of the k-th iteration is

$$\mathcal{E}(k+1) - \mathcal{E}(k)$$

$$
\begin{aligned}
=&s(k+4)(k+2)\left(f(x_{k+1}) - f(x^\star)\right) - s(k+3)(k+1)\left(f(x_k) - f(x^\star)\right) \\
&+ \frac{k+3}{2(k+4-\beta)}\left\|2(x_{k+2}-x^\star) + \sqrt{s}(k+2)\left(v_{k+1} + \beta\sqrt{s}\nabla f(x_{k+1})\right)\right\|^2 \\
&- \frac{k+2}{2(k+3-\beta)}\left\|2(x_{k+1}-x^\star) + \sqrt{s}(k+1)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right)\right\|^2 \\
\leq&s\left(k+3\right)\left(k+1\right)\left(f(x_{k+1}) - f(x_k)\right) + s(2k+5)\left(f(x_{k+1}) - f(x^\star)\right) \\
&+ \frac{k+3}{k+4-\beta}\Big[\Big\langle 2(x_{k+2}-x_{k+1}) + \sqrt{s}(k+2)\left(v_{k+1} + \beta\sqrt{s}\nabla f(x_{k+1})\right) \\
&\qquad\qquad - \sqrt{s}(k+1)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right), \\
&\qquad 2(x_{k+2}-x^\star) + \sqrt{s}(k+2)\left(v_{k+1} + \beta\sqrt{s}\nabla f(x_{k+1})\right)\Big\rangle \\
&\qquad -\frac{1}{2}\Big\|2(x_{k+2}-x_{k+1}) + \sqrt{s}(k+2)\left(v_{k+1} + \beta\sqrt{s}\nabla f(x_{k+1})\right) \\
&\qquad\qquad -\sqrt{s}(k+1)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right)\Big\|^2\Big] \\
=&s\left(k+3\right)\left(k+1\right)\left(f(x_{k+1}) - f(x_k)\right) + s(2k+5)\left(f(x_{k+1}) - f(x^\star)\right) \\
&- \left\langle s(k+3)\nabla f(x_{k+1}), 2(x_{k+2}-x^\star) + \sqrt{s}(k+2)\left(v_{k+1} + \beta\sqrt{s}\nabla f(x_{k+1})\right)\right\rangle \\
&- \frac{1}{2}s^2(k+3)\left(k+4-\beta\right)\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned}
$$

With the basic inequality of any function $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$

$$
\begin{cases}
f(x_k) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle + \dfrac{1}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
f(x^\star) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x^\star - x_{k+1}\rangle,
\end{cases}
$$

and the phase-space representation (5.120)

$$
x_{k+2} = x_{k+1} + \sqrt{s}v_{k+1},
$$

the difference of the discrete Lyapunov function (5.125) can be estimated as

$$
\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
\leq&s(k+3)(k+1)\left(\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle - \frac{1}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2\right) \\
&+ s(2k+5)\left(f(x_{k+1}) - f(x^\star)\right) - s(2k+6)\left\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&- s^{\frac{3}{2}}(k+3)(k+4)\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle - \beta s^2(k+2)(k+3)\left\|\nabla f(x_{k+1})\right\|^2 \\
&- \frac{1}{2}s^2(k+3)\left(k+4-\beta\right)\left\|\nabla f(x_{k+1})\right\|^2 \\
\leq &- s^{\frac{3}{2}}(k+3)\left\langle \nabla f(x_{k+1}), (k+4)v_{k+1} - (k+1)v_k\right\rangle \\
&- \frac{s(k+3)(k+1)}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
&- 2s\left(f(x_{k+1}) - f(x^\star)\right) \\
&- s^2\left[\beta(k+3)(k+2) + \frac{1}{2}(k+3)\left(k+4-\beta\right)\right]\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned}
$$

Utilize the phase-space representation (5.120) again, we calculate the difference of the discrete Lyapunov function (5.125) as

$$
\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
\leq & s^{\frac{3}{2}}(k+3)\left\langle \nabla f(x_{k+1}), \beta\sqrt{s}(k+1)\left(\nabla f(x_{k+1}) - \nabla f(x_k)\right) + \sqrt{s}(k+4)\nabla f(x_{k+1})\right\rangle \\
&- \frac{s(k+3)(k+1)}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
&- s^2\left[\beta(k+3)(k+2) + \frac{1}{2}(k+3)\left(k+4-\beta\right)\right]\left\|\nabla f(x_{k+1})\right\|^2 \\
\leq & \beta s^2(k+3)(k+1)\left\langle \nabla f(x_{k+1}), \nabla f(x_{k+1}) - \nabla f(x_k)\right\rangle \\
&- \frac{s(k+3)(k+1)}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
&- \left[(k+2)(k+3)\beta - \frac{1}{2}\left(k+4+\beta\right)(k+3)\right]s^2\left\|\nabla f(x_{k+1})\right\|^2 \\
\leq & \frac{L\beta^2 s^3}{2}(k+3)(k+1)\left\|\nabla f(x_{k+1})\right\|^2 \\
&- \left[(k+2)(k+3)\beta - \frac{1}{2}\left(k+4+\beta\right)(k+3)\right]s^2\left\|\nabla f(x_{k+1})\right\|^2 \\
= &- \left[\beta(k+2) - \frac{1}{2}\left(k+4+\beta\right) - \frac{L\beta^2 s}{2}(k+1)\right](k+3)s^2\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned}
$$

Consistently, we can obtain the sufficient condition for the Lyapunov function $\mathcal{E}(k)$ decreasing (5.123) and the sufficient condition for step size (5.124).

Now, we turn to discuss the parameter $\beta \geq 1$ case by case.

- When the parameter $\beta \geq 3$, since the function $h(k) = \frac{1}{L\beta^2}\left(2\beta - 1 + \frac{\beta-3}{k+1}\right)$ decreases monotonically for $k \geq 0$, then the condition of the step size

  $$s \leq \frac{2\beta - 1}{(1+\epsilon)L\beta^2} < \frac{2\beta - 1}{L\beta^2}$$

  holds for (5.123), where $\epsilon > 0$ is a real number. Hence, when $k \geq k_{3,\beta} + 1$, where

  $$k_{3,\beta} = \max\left\{0, \lfloor \beta - 3 \rfloor + 1, \left\lfloor \frac{4 - 3\beta + L\beta^2 s}{2\beta - 1 - L\beta^2 s} \right\rfloor + 1\right\},$$

  the difference of the discrete Lyapunov function (5.125) can be estimated as

  $$\mathcal{E}(k+1) - \mathcal{E}(k) \leq -s^2\left(\frac{2\beta - 1 - L\beta^2 s}{2}\right)(k - k_{3,\beta})^2 \left\|\nabla f(x_{k+1})\right\|^2.$$

- When the parameter $1 \leq \beta < 3$, since the function $h(k) = \frac{1}{L\beta^2}\left(2\beta - 1 + \frac{\beta-3}{k+1}\right)$ increases monotonically for $k \geq 0$, there exists $k_{3,\beta} = \max\Big\{0, \lfloor \beta - 3 \rfloor + 1, \left\lfloor \frac{4-3\beta}{2\beta-1} \right\rfloor + 1\Big\}$ such that the step size

  $$s \leq \frac{(2\beta - 1)k_{3,\beta} + 3\beta - 4}{(k_{3,\beta} + 1)L\beta^2}$$

  works for any $k \geq k_{3,\beta}$. When $\beta = 1$, the step size satisfies

  $$0 < s \leq \frac{k-1}{k+1} \cdot \frac{1}{L}$$

  which is consistent with (5.48) and $k_{3,\beta} = 2$. Then, the difference of the discrete Lyapunov function (5.121) can be estimated as

  $$\mathcal{E}(k+1) - \mathcal{E}(k) \leq -s^2\left(\frac{2\beta - 1 - L\beta^2 s}{2}\right)(k - k_{3,\beta})^2 \left\|\nabla f(x_{k+1})\right\|^2.$$

  for all $k \geq k_{3,\beta} + 1$.

By simple calculation, we complete the proof.

5.7.4.2 Proof of Theorem 5.5.1: Case $\alpha > 3$

Before starting to prove Theorem 5.5.1: Case $\alpha > 3$, we first also look back our high-resolution ODE framework in Section 5.2.

- **Step** 1, the generalized high-resolution ODE has been given in (5.56).
- **Step** 2, the continuous Lyapunov function is constructed as

$$\begin{aligned}\mathcal{E}(t) = t\left[t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s}\right](f(X(t)) - f(x^\star)) \\ + \frac{1}{2}\left\|(\alpha-1)(X(t) - x^\star) + t\left(\dot{X}(t) + \beta\sqrt{s}\nabla f(X(t))\right)\right\|^2, \quad (5.126)\end{aligned}$$

which is consistent with (5.119) for $\alpha \to 3$. Following this Lyapunov function (5.126), we can obtain

$$\begin{aligned}& f(X(t)) - f(x^\star) \le O\left(\frac{\|X(t_0) - x^\star\|^2}{(t-t_0)^2}\right) \\ & \int_{t_0}^{t} u\left(f(X(u)) - f(x^\star)\right) + \sqrt{s}u^2\left\|\nabla f(X(u))\right\|^2 du \le O\left(\|X(t_0) - x^\star\|^2\right)\end{aligned} \quad (5.127)$$

for any $t > t_0 = \max\{\sqrt{s}(\alpha/2 - \beta)(\alpha-2)/(\alpha-3), \sqrt{s}(\alpha/2)\}$. The two inequalities of (5.127) for the convergence rate of function value is stronger than Corollary 5.4.3. The detailed calculation, about the estimate of the optimal constant β and how the constant β influences the initial point, is left for readers.

- **Step** 3, before constructing discrete Lyapunov functions, we look back the phase-space representation (5.120)

$$\begin{aligned}& x_k - x_{k-1} = \sqrt{s}v_{k-1} \\ & v_k - v_{k-1} = -\frac{\alpha}{k}v_k - \beta\sqrt{s}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \left(1 + \frac{\alpha}{k}\right)\sqrt{s}\nabla f(x_k).\end{aligned}$$

The discrete functional is constructed as

$$\begin{aligned}\mathcal{E}(k) = s(k+1)(k+\alpha-\beta+1)\left(f(x_k)-f(x^\star)\right)&\\ +\frac{1}{2}\left\|(\alpha-1)(x_{k+1}-x^\star)+\sqrt{s}(k+1)\left(v_k+\beta\sqrt{s}\nabla f(x_k)\right)\right\|^2.&\end{aligned} \quad (5.128)$$

When $\beta = 1$, with $\alpha \to 3$, the discrete Lyapunov function $\mathcal{E}(k)$ degenerates to (5.46).

Now, we procced to **Step** 4 to analyze the algorithms (5.57) with $\alpha > 3$ in order to prove Theorem 5.5.2. The simple transformation of (5.120) for $\alpha > 3$ is

$$(k+\alpha)\left(v_k+\beta\sqrt{s}\nabla f(x_k)\right)-k\left(v_{k-1}+\beta\sqrt{s}\nabla f(x_{k-1})\right) = -\sqrt{s}\left(k+\gamma-\gamma\beta\right)\nabla f(x_k). \quad (5.129)$$

Thus, the difference of the Lyapunov function (5.128) on the k-th iteration is

$$\begin{aligned}&\mathcal{E}(k+1)-\mathcal{E}(k)\\ =&s(k+2)(k+\alpha-\beta+2)\left(f(x_k)-f(x^\star)\right)\\ &+\frac{1}{2}\left\|(\alpha-1)(x_{k+2}-x^\star)+\sqrt{s}(k+2)\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right)\right\|^2\\ &-s(k+1)(k+\alpha-\beta+1)\left(f(x_k)-f(x^\star)\right)\\ &-\frac{1}{2}\left\|(\alpha-1)(x_{k+1}-x^\star)+\sqrt{s}(k+1)\left(v_k+\beta\sqrt{s}\nabla f(x_k)\right)\right\|^2\\ =&s(k+1)\left(k+\alpha-\beta+1\right)\left(f(x_{k+1})-f(x_k)\right)+s\left(2k+\alpha-\beta+3\right)\left(f(x_{k+1})-f(x^\star)\right)\\ &+\Big\langle(\alpha-1)(x_{k+2}-x_{k+1})+\sqrt{s}(k+2)\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right)\\ &\qquad\qquad-\sqrt{s}(k+1)\left(v_k+\beta\sqrt{s}\nabla f(x_k)\right),\\ &\quad(\alpha-1)(x_{k+2}-x^\star)+\sqrt{s}(k+2)\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right)\Big\rangle\\ &-\frac{1}{2}\Big\|(\alpha-1)(x_{k+2}-x_{k+1})+(k+2)\sqrt{s}\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right)\\ &\qquad\qquad-(k+1)\sqrt{s}\left(v_k+\beta\sqrt{s}\nabla f(x_k)\right)\Big\|^2\\ =&s(k+1)\left(k+\alpha-\beta+1\right)\left(f(x_{k+1})-f(x_k)\right)\\ &+s\left(2k+\alpha-\beta+3\right)\left(f(x_{k+1})-f(x^\star)\right)\\ &-\Big\langle s\left(k+\alpha-\beta+1\right)\nabla f(x_{k+1}),(\alpha-1)(x_{k+1}-x^\star)+\sqrt{s}(k+\alpha+1)v_{k+1}\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad +\beta s(k+2)\nabla f(x_{k+1})\rangle \\
&- \frac{1}{2}s^2(k+\alpha-\beta+1)^2\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned}
$$

With the basic inequality of convex function $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$,

$$
\begin{cases}
f(x_k) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle + \dfrac{1}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
f(x^\star) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x^\star - x_{k+1}\rangle
\end{cases}
$$

and the phase-space representation (5.120)

$$
x_{k+2} = x_{k+1} + \sqrt{s}v_{k+1},
$$

the difference of the discrete Lyapunov function (5.128) can be estimated as

$$
\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
=&s(k+1)\left(k+\alpha-\beta+1\right)\left(\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle - \frac{1}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2\right) \\
&+ s\left(2k+\alpha-\beta+3\right)\left(f(x_{k+1}) - f(x^\star)\right) \\
&- s(\alpha-1)\left(k+\alpha-\beta+1\right)\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\rangle \\
&- \left\langle s\left(k+\alpha-\beta+1\right)\nabla f(x_{k+1}), \sqrt{s}(k+\alpha+1)v_{k+1}\right\rangle \\
&- \frac{1}{2}s^2(k+\alpha-\beta+1)\left[(2\beta+1)k+\alpha+3\beta+1\right]\left\|\nabla f(x_{k+1})\right\|^2 \\
\leq& - s^{\frac{3}{2}}\left(k+\alpha-\beta+1\right)\langle \nabla f(x_{k+1}), (k+\alpha+1)v_{k+1} - (k+1)v_k\rangle \\
&- \frac{s(k+1)\left(k+\alpha-\beta+1\right)}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|_2^2 \\
&- s\left[(\alpha-3)k + \left(\alpha-2\right)\left(\alpha-\beta+1\right) - 2\right]\left(f(x_{k+1}) - f(x^\star)\right) \\
&- \frac{1}{2}s^2(k+\alpha-\beta+1)\left[(2\beta+1)k+\alpha+3\beta+1\right]\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned}
$$

Utilizing the phase-space representation (5.120) again, we calculate the difference of the discrete Lyapunov function (5.128) as

$$
\mathcal{E}(k+1) - \mathcal{E}(k)
$$

$$
\begin{aligned}
=&\beta s^2(k+1)\left(k+\alpha-\beta+1\right)\langle\nabla f(x_{k+1}),\nabla f(x_{k+1})-\nabla f(x_k)\rangle\\
&+s^2(k+\alpha+1)\left(k+\alpha-\beta+1\right)\left\|\nabla f(x_{k+1})\right\|^2\\
&-\frac{s(k+1)\left(k+\alpha-\beta+1\right)}{2L}\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|_2^2\\
&-s\left[(\alpha-3)k+(\alpha-2)\left(\alpha-\beta+1\right)-2\right]\left(f(x_{k+1})-f(x^\star)\right)\\
&-\frac{1}{2}s^2(k+\alpha-\beta+1)\left[(2\beta+1)k+\alpha+3\beta+1\right]\left\|\nabla f(x_{k+1})\right\|^2\\
=&\beta s^2(k+1)\left(k+\alpha-\beta+1\right)\langle\nabla f(x_{k+1}),\nabla f(x_{k+1})-\nabla f(x_k)\rangle\\
&-\frac{s(k+1)\left(k+\alpha-\beta+1\right)}{2L}\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|_2^2\\
&-s\left[(\alpha-3)k+(\alpha-2)\left(\alpha-\beta+1\right)-2\right]\left(f(x_{k+1})-f(x^\star)\right)\\
&-\frac{1}{2}s^2(k+\alpha-\beta+1)\left[(2\beta-1)k-\alpha+3\beta-1\right]\left\|\nabla f(x_{k+1})\right\|^2\\
\leq&\frac{L\beta^2s^3}{2}(k+1)\left(k+\alpha-\beta+1\right)\left\|\nabla f(x_{k+1})\right\|^2\\
&-s\left[(\alpha-3)k+(\alpha-2)\left(\alpha-\beta+1\right)-2\right]\left(f(x_{k+1})-f(x^\star)\right)\\
&-\frac{1}{2}s^2(k+\alpha-\beta+1)\left[(2\beta-1)k-\alpha+3\beta-1\right]\left\|\nabla f(x_{k+1})\right\|^2\\
=&-s\left[(\alpha-3)k+(\alpha-2)\left(\alpha-\beta+1\right)-2\right]\left(f(x_{k+1})-f(x^\star)\right)\\
&-\frac{1}{2}s^2(k+\alpha-\beta+1)\left[(2\beta-1)k-\alpha+3\beta-1-L\beta^2s(k+1)\right]\left\|\nabla f(x_{k+1})\right\|^2
\end{aligned}
$$

To guarantee the Lyapunov function $\mathcal{E}(k)$ decreasing, a sufficient condition is

$$
(2\beta-1)k-\alpha+3\beta-1-L\beta^2s(k+1)\geq 0. \tag{5.130}
$$

With the inequality (5.130), the step size can be estimated as

$$
s\leq\frac{2\beta-1}{L\beta^2}-\frac{\alpha-\beta}{(k+1)L\beta^2}.
$$

- When the parameter $\beta>1/2$ and $\alpha<\beta$, since the function $h(k)=\frac{2\beta-1}{L\beta^2}-\frac{\alpha-\beta}{(k+1)L\beta^2}$ decreases monotonically for $k\geq 0$, thus the step size

$$
s\leq\frac{2\beta-1}{(1+\epsilon)L\beta^2}<\frac{2\beta-1}{L\beta^2}
$$

holds for (5.130), where $\epsilon > 0$ is a real number. Hence, when $k \geq k_{\alpha,\beta} + 1$, where

$$\begin{aligned} k_{\alpha,\beta} = &\max\left\{0, \left\lfloor \frac{2-(\alpha-2)(\alpha-\beta+1)}{\alpha-3} \right\rfloor + 1, \right. \\ &\left. \left\lfloor \frac{4-3\beta+L\beta^2 s}{-1+2\beta-L\beta^2 s} \right\rfloor + 1, \lfloor \beta-\alpha-1 \rfloor + 1 \right\}, \end{aligned}$$

the difference of the discrete Lyapunov function (5.128) can be estimated as

$$\begin{aligned} \mathcal{E}(k+1) - \mathcal{E}(k) \leq &- s(\alpha-3)\left(k-k_{\alpha,\beta}\right)\left(f(x_{k+1}) - f(x^\star)\right) \\ &- s^2 \left(\frac{2\beta-1-L\beta^2 s}{2}\right)\left(k-k_{\alpha,\beta}\right)^2 \left\|\nabla f(x_{k+1})\right\|^2. \end{aligned}$$

- When the parameter $\beta > 1/2$ and $\alpha \geq \beta$, since the function $h(k) = \frac{2\beta-1}{L\beta^2} - \frac{\alpha-\beta}{(k+1)L\beta^2}$ increases monotonically for $k \geq 0$, there exists

$$k_{\alpha,\beta} = \max\left\{0, \left\lfloor \frac{2-(\alpha-2)(\alpha-\beta+1)}{\alpha-3} \right\rfloor + 1, \lfloor \beta-\alpha-1 \rfloor + 1, \left\lfloor \frac{1+\alpha-3\beta}{2\beta-1} \right\rfloor + 1 \right\}$$

such that the step size satisfies

$$s \leq \frac{(2\beta-1)k_{\alpha,\beta} - \alpha + 3\beta - 1}{L\beta^2(k_{\alpha,\beta}+1)}.$$

When $\beta = 1$, the step size satisfies

$$s \leq \frac{1}{L} \cdot \frac{k_{\alpha,\beta} - \alpha + 2}{(k_{\alpha,\beta}+1)} \to \frac{1}{L} \cdot \frac{k_{\alpha,\beta} - 1}{k_{\alpha,\beta}+1} \quad \text{with} \quad \alpha \to 3,$$

which is consistent with (5.48). Then, the difference of the discrete Lyapunov function (5.128) can be estimated as

$$\begin{aligned} \mathcal{E}(k+1) - \mathcal{E}(k) \leq &- s(\alpha-3)\left(k-k_{\alpha,\beta}\right)\left(f(x_{k+1}) - f(x^\star)\right) \\ &- s^2 \left(\frac{2\beta-1-L\beta^2 s}{2}\right)\left(k-k_{\alpha,\beta}\right)^2 \left\|\nabla f(x_{k+1})\right\|^2. \end{aligned}$$

5.7.4.3 A Simple Counterexample

The simple counterexample is constructed as

$$f(x_k) - f(x^\star) = \begin{cases} \dfrac{L\left\|x_0 - x^\star\right\|^2}{(k+1)^2}, & k = j^2 \\ 0, & k \neq j^2 \end{cases}$$

where $j \in \mathbb{N}$. Plugging it into (5.59), we have

$$\sum_{k=0}^{\infty}(k+1)\left(f(x_k) - f(x^\star)\right) = L\left\|x_0 - x^\star\right\|^2 \cdot \sum_{j=0}^{\infty}\left(\frac{1}{j^2+1}\right) < \infty.$$

Hence, Proposition 5.5.2 cannot guarantee the faster convergence rate.

5.7.4.4 Super-Critical Regime: Sharper Convergence Rate $o(1/t^2)$ and $o(L/k^2)$

5.7.4.4.1 The ODE Case Here, we still turn back to our high-resolution ODE framework in Section 5.2. The generalized high-resolution ODE has been still shown in (5.56). A more general Lyapunov function is constructed as

$$\begin{aligned} \mathcal{E}_\nu(t) = t\left[t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s} + (\alpha - \nu - 1)\beta\sqrt{s}\right]\left(f(X(t)) - f(x^\star)\right) \\ + \frac{\nu(\alpha - \nu - 1)}{2}\left\|X(t) - x^\star\right\|^2 + \frac{1}{2}\left\|\nu(X(t) - x^\star) + t\left(\dot{X}(t) + \beta\sqrt{s}\nabla f(X(t))\right)\right\|^2 \end{aligned} \tag{5.131}$$

where $2 < \nu \leq \alpha - 1$. When $\nu = \alpha - 1$, the Lyapunov function (5.131) degenerates to (5.126). Furthermore, when $\nu = \alpha - 1 \to 2$, the Lyapunov function (5.131) degenerates to (5.119). Finally, when $2 = \nu = \alpha - 1$ and $\beta = 1$, the Lyapunov function (5.131) is consistent with (5.41). We assume that initial time is

$$t_{\alpha,\beta,\nu} = \max\left\{\sqrt{s}\left(\beta - \frac{\alpha}{2}\right), \sqrt{s}\left(\frac{\beta(\alpha-2)}{\nu-2} - \frac{\alpha(\nu-1)}{2(\nu-2)}\right), \frac{\sqrt{s}\alpha}{2}\right\}.$$

Based on the Lyapunov function (5.131), we have the following results.

Theorem 5.7.10. Let $f(x) \in \mathcal{F}_L^2(\mathbb{R}^n)$ and $X = X(t)$ be the solution of the ODE (5.56) with $\alpha > 3$ and $\beta > 0$. Then, there exists $t_{\alpha,\beta,\nu} > 0$ such that

$$\begin{cases} \lim\limits_{t\to\infty} t^2 \left((f(X(t)) - f(x^\star)) + \left\| \dot{X}(t) + \beta\sqrt{s}\nabla f(X(t)) \right\|^2 \right) = \mathfrak{C}^2_{\alpha,\beta,\nu} \left\| x_0 - x^\star \right\|^2 \\ \displaystyle\int_{t_0}^{t} \left[u\left(f(X(u)) - f(x^\star)\right) + u \left\| \dot{X}(u) + \beta\sqrt{s}\nabla f(X(u)) \right\|^2 \right] \mathrm{d}u < \infty \end{cases} \tag{5.132}$$

for all $t \geq t_{\alpha,\beta,\nu}$, where the positive constant $\mathfrak{C}^2_{\alpha,\beta,\nu}$ and the integer $t_{\alpha,\beta,\nu}$ depend only on α, β and ν. In other words, the equivalent expression of (5.132) is

$$f(X(t)) - f(x^\star) + \left\| \dot{X}(t) + \beta\sqrt{s}\nabla f(X(t)) \right\|^2 \leq o\left(\frac{\|x_0 - x^\star\|^2}{t^2} \right).$$

Now, we start to show the proof. Since $X = X(t)$ is the solution of the ODE (5.56) with $\alpha > 3$ and $\beta > 0$, when $t > t_{\alpha,\beta,\nu}$, the time derivative of Lyapunov function (5.131) is

$$\begin{aligned} \frac{\mathrm{d}\mathcal{E}_\nu(t)}{\mathrm{d}t} &= \left[2t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s} + (\alpha - \nu - 1)\beta\sqrt{s} \right] (f(X(t)) - f(x^\star)) \\ &\quad + t\left[t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s} + (\alpha - \nu - 1)\beta\sqrt{s} \right] \left\langle \nabla f(X(t)), \dot{X}(t) \right\rangle \\ &\quad + \nu(\alpha - \nu - 1)\left\langle X(t) - x^\star, \dot{X}(t) \right\rangle \\ &\quad - \left\langle (\alpha - 1 - \nu)\dot{X}(t) + \left[t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s} \right] \nabla f(X(t)), \right. \\ &\quad \left. \nu(X(t) - x^\star) + t\left(\dot{X}(t) + \beta\sqrt{s}\nabla f(X(t)) \right) \right\rangle \\ &= \left[2t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s} + (\alpha - \nu - 1)\beta\sqrt{s} \right] (f(X(t)) - f(x^\star)) \qquad (5.133) \\ &\quad - (\alpha - 1 - \nu)t \left\| \dot{X}(t) \right\|^2 \\ &\quad - \nu\left[t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s} \right] \langle \nabla f(X(t)), X(t) - x^\star \rangle \qquad (5.134) \\ &\quad - \beta t\sqrt{s}\left[t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s} \right] \|\nabla f(X(t))\|^2. \end{aligned}$$

With the basic inequality for any $f(x) \in \mathcal{F}_L^2(\mathbb{R}^n)$

$$f(x^\star) \geq f(X(t)) + \langle \nabla f(X(t)), x^\star - X(t) \rangle,$$

the time derivative of Lyapunov function (5.133) can be estimated as

$$\begin{aligned}\frac{\mathrm{d}\mathcal{E}_\nu(t)}{\mathrm{d}t} \leq & -\left\{(\nu-2)t+\sqrt{s}\left[\frac{\alpha(\nu-1)}{2}-(\alpha-2)\beta\right]\right\}(f(X(t))-f(x^\star)) \\ & -(\alpha-1-\nu)t\left\|\dot{X}(t)\right\|^2-\beta t\sqrt{s}\left[t+\left(\frac{\alpha}{2}-\beta\right)\sqrt{s}\right]\|\nabla f(X(t))\|^2.\end{aligned}$$

With the Lyapunov function $\mathcal{E}_\nu(t) \geq 0$ and the technique for integral, for any $t > t_0$ we have

$$\begin{aligned}\int_{t_0}^{t} u(f(X(u))-f(x^\star))du \leq & \int_{t_0}^{t_0+\delta} u(f(X(u))-f(x^\star))du \\ & +\left(1+\frac{t_0}{\delta}\right)\int_{t_0+\delta}^{t}(u-t_0)(f(X(u))-f(x^\star))du,\end{aligned}$$

where $\delta < t - t_0$. Thus, we can obtain the following Lemma.

Lemma 5.7.11. Under the same assumption of Theorem 5.7.10, the following limits exist

$$\lim_{t\to\infty}\mathcal{E}_\nu(t),\ \lim_{t\to\infty}\int_{t_0}^{t}u(f(X(u))-f(x^\star))\mathrm{d}u,\ \lim_{t\to\infty}\int_{t_0}^{t}u\left\|\dot{X}(u)\right\|^2\mathrm{d}u,\ \lim_{t\to\infty}\int_{t_0}^{t}u^2\|\nabla f(X(u))\|^2 du.$$

With (5.133) and Lemma 5.7.11, the following Lemma holds.

Lemma 5.7.12. Under the same assumption of Theorem 5.7.10, the following limit exists

$$\lim_{t\to\infty}\int_{t_0}^{t}u\left\langle\nabla f(X(u)),X(u)-x^\star\right\rangle du.$$

Lemma 5.7.13. Under the same assumption of Theorem 5.7.10, the following limits exist

$$\lim_{t\to\infty}\|X(t)-x^\star\| \quad \text{and} \quad \lim_{t\to\infty}t\left\langle X(t)-x^\star,\dot{X}(t)+\beta\sqrt{s}\nabla f(X(t))\right\rangle.$$

Proof. [Proof of Lemma 5.7.13] Taking $\nu \neq \nu' \in [2, \gamma-1]$, we have

$$\mathcal{E}_\nu(t)-\mathcal{E}_{\nu'}(t)=(\nu-\nu')\left[-\beta\sqrt{s}t\left(f(X(t))-f(x^\star)\right)\right.$$

$$+t\left\langle X(t)-x^{\star},\dot{X}(t)+\beta\sqrt{s}\nabla f(X(t))\right\rangle+\frac{\alpha-1}{2}\left\|X(t)-x^{\star}\right\|^{2}\Bigg]$$

With Lemma 5.7.11 and (5.127), the following limit exists

$$\lim_{t\to\infty}\left[t\left\langle X(t)-x^{\star},\dot{X}(t)+\beta\sqrt{s}\nabla f(X(t))\right\rangle+\frac{\alpha-1}{2}\left\|X(t)-x^{\star}\right\|^{2}\right]. \tag{5.135}$$

Define a new function about time variable t:

$$\pi(t):=\frac{1}{2}\left\|X(t)-x^{\star}\right\|^{2}+\beta\sqrt{s}\int_{t_0}^{t}\left\langle\nabla f(X(u)),X(u)-x^{\star}\right\rangle du.$$

If we can prove the existence of the limit $\pi(t)$ with $t\to\infty$, we can guarantee $\lim\limits_{t\to\infty}\|X(t)-x^{\star}\|$ exists with Lemma 5.7.12. We observe the following equality

$$\begin{aligned}&t\dot{\pi}(t)+(\alpha-1)\pi(t)\\=&\beta(\alpha-1)\sqrt{s}\int_{t_0}^{t}\left\langle\nabla f(X(u)),X(u)-x^{\star}\right\rangle du\\&+t\left\langle X(t)-x^{\star},\dot{X}(t)+\beta\sqrt{s}\nabla f(X(t))\right\rangle+\frac{\alpha-1}{2}\left\|X(t)-x^{\star}\right\|^{2}.\end{aligned}$$

With (5.135) and Lemma 5.7.12, we obtain that the following limit exists

$$\lim_{t\to\infty}\left[t\dot{\pi}(t)+(\alpha-1)\pi(t)\right],$$

that is, there exists some constant $\mathfrak{C}^3$ such that the following equality holds,

$$\lim_{t\to\infty}\frac{\frac{\mathrm{d}(t^{\alpha-1}\pi(t))}{\mathrm{d}t}}{t^{\alpha-2}}=\lim_{t\to\infty}\left[t\dot{\pi}(t)+(\alpha-1)\pi(t)\right]=\mathfrak{C}^3.$$

For any $\epsilon>0$, there exists $t_0>0$ such that when $t\geq t_0$, we have

$$t^{\alpha-1}\left(\pi(t)-\frac{\mathfrak{C}^3}{\alpha-1}\right)-t_0^{\alpha-1}\left(\pi(t_0)-\frac{\mathfrak{C}^3}{\alpha-1}\right)\leq\frac{\epsilon}{\alpha-1}\cdot\left(t^{\alpha-1}-t_0^{\alpha-1}\right)$$

that is,

$$\left|\pi(t)-\frac{\mathfrak{C}^3}{\alpha-1}\right|\leq\left|\pi(t_0)-\frac{\mathfrak{C}^3}{\alpha-1}\right|\left(\frac{t_0}{t}\right)^{\alpha-1}+\frac{\epsilon}{\alpha-1}.$$

The proof is complete. □

Finally, we finish the proof for Theorem 5.7.10.

Proof. [Proof of Theorem 5.7.10] When $t > t_{\alpha,\beta,\nu}$, we expand the Lyapunov function (5.131) as

$$\begin{aligned}\mathcal{E}_\nu(t) =& t\left[t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s} + (\alpha - \nu - 1)\beta\sqrt{s}\right](f(X(t)) - f(x^\star))\\ &+ \frac{\nu(\alpha-1)}{2}\left\|X(t) - x^\star\right\|^2\\ &+ \frac{t^2}{2}\left\|\dot{X}(t) + \beta\sqrt{s}\nabla f(X(t))\right\|^2\\ &+ t\left\langle X(t) - x^\star, \dot{X}(t) + \beta\sqrt{s}\nabla f(X(t))\right\rangle.\end{aligned}$$

With Lemma 5.7.11 and Lemma 5.7.13, we obtain the first equation of (5.132). Furthermore, Cauchy-Scharwz inequality tells that

$$\begin{aligned}&\left[t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s} + (\alpha - \nu - 1)\beta\sqrt{s}\right](f(X(t)) - f(x^\star))\\ &+ \frac{t}{2}\left\|\dot{X}(t) + \beta\sqrt{s}\nabla f(X(t))\right\|^2\\ \leq& \left[t + \left(\frac{\alpha}{2} - \beta\right)\sqrt{s} + (\alpha - \nu - 1)\beta\sqrt{s}\right](f(X(t)) - f(x^\star))\\ &+ t\left\|\dot{X}(t)\right\|^2 + \beta^2 st\left\|\nabla f(X(t))\right\|^2.\end{aligned}$$

With Lemma 5.7.11, we obtain the second equation of (5.132). With basic calculation, we complete the proof. □

5.7.4.4.2 Proof of Theorem 5.5.3 Similarly, under the assumption of Theorem 5.5.3, if we can show a discrete version of (5.132), that is, there exists some constant $\mathfrak{C}^4_{\alpha,\beta,\nu} > 0$ and $\mathfrak{c}_{\alpha,\beta,\nu} > 0$ such that when the step size satisfies $0 < s \leq \mathfrak{c}_{\alpha,\beta,\nu}/L$, the following relationship holds

$$\begin{cases}\displaystyle\lim_{k\to\infty}(k+1)^2\left(f(x_k) - f(x^\star) + \left\|v_k + \beta\sqrt{s}\nabla f(x_k)\right\|^2\right) = \frac{\mathfrak{C}^4_{\alpha,\beta,\nu}\left\|x_0 - x^\star\right\|^2}{s}\\ \displaystyle\sum_{k=0}^{\infty}(k+1)\left((f(x_k) - f(x^\star)) + \left\|v_k + \beta\sqrt{s}\nabla f(x_k)\right\|^2\right) < \infty.\end{cases} \tag{5.136}$$

Thus, we obtain the sharper convergence rate as

$$f(x_k) - f(x^\star) + \left\| v_k + \beta\sqrt{s}\nabla f(x_k) \right\|^2 \leq o\left(\frac{\|x_0 - x^\star\|^2}{sk^2} \right).$$

Now we show the derivation of the inequality (5.136). The discrete Lyapunov function is constructed as

$$\begin{aligned} \mathcal{E}(k) = & \underbrace{s(k+1)\left[k + \alpha + 1 - \beta + \frac{(k+2)(\alpha - 1 - \nu)\beta}{k+\alpha+1}\right](f(x_k) - f(x^\star))}_{\mathbf{I}} \\ & + \underbrace{\frac{\nu(\alpha - \nu - 1)}{2}\|x_{k+1} - x^\star\|^2}_{\mathbf{II}} + \underbrace{\frac{1}{2}\left\|\nu(x_{k+1} - x^\star) + (k+1)\sqrt{s}\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right)\right\|^2}_{\mathbf{III}}, \end{aligned} \tag{5.137}$$

where $2 \leq \nu < \alpha - 1$ and parts **I**, **II** and **III** are potential, Euclidean distance and mixed energy respectively. Apparently, when $\nu = \alpha - 1$, the discrete Lyapunov function (5.137) is consistent with (5.128). When $\beta = 1$ and $\nu = \alpha - 1 \to 2$, the discrete Lyapunov function (5.137) degenerates to (5.46), Now, we turn to estimate the difference of Lyapunov function (5.137).

- For the part **I**, potential, we have

$$\begin{aligned} & s(k+2)\left[k + \alpha + 2 - \beta + \frac{(k+3)(\alpha - 1 - \nu)\beta}{k+\alpha+2}\right](f(x_{k+1}) - f(x^\star)) \\ & - s(k+1)\left[k + \alpha + 1 - \beta + \frac{(k+2)(\alpha - 1 - \nu)\beta}{k+\alpha+1}\right](f(x_k) - f(x^\star)) \\ = & s(k+1)\left[k + \alpha + 1 - \beta + \frac{(k+2)(\alpha - 1 - \nu)\beta}{k+\alpha+1}\right](f(x_{k+1}) - f(x_k)) \\ & + s\left(2k + \alpha + 3 - \beta\right)(f(x_{k+1}) - f(x^\star)) \\ & + s(k+2)(\alpha - 1 - \nu)\beta\left[\frac{k+3}{k+\alpha+2} - \frac{k+1}{k+\alpha+1}\right](f(x_{k+1}) - f(x^\star)) \\ \leq & \underbrace{s(k+1)\left[k + \alpha + 1 - \beta + \frac{(k+2)(\alpha - 1 - \nu)\beta}{k+\alpha+1}\right](f(x_{k+1}) - f(x_k))}_{\mathbf{I_1}} \\ & + \underbrace{s\left[2k + \alpha + 3 + (2\alpha - 3 - 2\nu)\beta\right](f(x_{k+1}) - f(x^\star))}_{\mathbf{I_2}}, \end{aligned}$$

where the last inequality follows $k+\alpha+2 > k+\alpha+1 > k+2$.

- For the part **II**, Euclidean distance, we have

$$
\begin{aligned}
&\frac{\nu(\alpha-\nu-1)}{2}\left\|x_{k+2}-x^{\star}\right\|^{2}-\frac{\nu(\alpha-\nu-1)}{2}\left\|x_{k+1}-x^{\star}\right\|^{2} \\
=&\underbrace{\nu(\alpha-\nu-1)\left\langle x_{k+2}-x_{k+1}, x_{k+2}-x^{\star}\right\rangle}_{\mathbf{II}_{1}}\underbrace{-\frac{\nu(\alpha-\nu-1)}{2}\left\|x_{k+2}-x_{k+1}\right\|^{2}}_{\mathbf{II}_{2}}.
\end{aligned}
$$

- For the part **III**, mixed energy, with the simple transformation (5.129) for $\alpha > 3$

$$
(k+\alpha)\left(v_{k}+\beta\sqrt{s}\nabla f(x_{k})\right)-k\left(v_{k-1}+\beta\sqrt{s}\nabla f(x_{k-1})\right)=-\sqrt{s}\left(k+\gamma-\gamma\beta\right)\nabla f(x_{k}),
$$

we have

$$
\begin{aligned}
&\frac{1}{2}\left\|\nu(x_{k+2}-x^{\star})+(k+2)\sqrt{s}\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right)\right\|^{2} \\
&\qquad\qquad-\frac{1}{2}\left\|\nu(x_{k+1}-x^{\star})+(k+1)\sqrt{s}\left(v_{k}+\beta\sqrt{s}\nabla f(x_{k})\right)\right\|^{2} \\
=&\Big\langle\nu(x_{k+2}-x_{k+1})+(k+2)\sqrt{s}\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right) \\
&\qquad\qquad\qquad-(k+1)\sqrt{s}\left(v_{k}+\beta\sqrt{s}\nabla f(x_{k})\right), \\
&\qquad\nu(x_{k+2}-x^{\star})+(k+2)\sqrt{s}\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right)\Big\rangle \\
&\quad-\frac{1}{2}\Big\|\nu(x_{k+2}-x_{k+1})+(k+2)\sqrt{s}\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right) \\
&\qquad\qquad-(k+1)\sqrt{s}\left(v_{k}+\beta\sqrt{s}\nabla f(x_{k})\right)\Big\|^{2} \\
=&-\left\langle s\left(k+\alpha+1-\beta\right)\nabla f(x_{k+1})+(\alpha-1-\nu)(x_{k+2}-x_{k+1}),\right. \\
&\qquad\qquad\left.\nu(x_{k+2}-x^{\star})+(k+2)(x_{k+2}-x_{k+1})+\beta s(k+2)\nabla f(x_{k+1})\right\rangle \\
&\quad-\frac{1}{2}\left\|s\left(k+\alpha+1-\beta\right)\nabla f(x_{k+1})+(\alpha-1-\nu)(x_{k+2}-x_{k+1})\right\|^{2} \\
=&-\nu(\alpha-\nu-1)\left\langle x_{k+2}-x_{k+1}, x_{k+2}-x^{\star}\right\rangle-(k+2)(\alpha-\nu-1)\left\|x_{k+2}-x_{k+1}\right\|^{2} \\
&\quad-\beta s(k+2)(\alpha-1-\nu)\left\langle\nabla f(x_{k+1}), x_{k+2}-x_{k+1}\right\rangle \\
&\quad-\left\langle s\left(k+\alpha+1-\beta\right)\nabla f(x_{k+1}),\right. \\
&\qquad\qquad\left.\nu(x_{k+1}-x^{\star})+(k+2+\nu)(x_{k+2}-x_{k+1})+\beta s(k+2)\nabla f(x_{k+1})\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{1}{2}\left\|s\left(k+\alpha+1-\beta\right)\nabla f(x_{k+1})\right\|^2 \\
&- \left\langle s\left(k+\alpha+1-\beta\right)\nabla f(x_{k+1}), (\alpha-1-\nu)(x_{k+2}-x_{k+1})\right\rangle \\
&- \frac{(\alpha-1-\nu)^2}{2}\left\|x_{k+2}-x_{k+1}\right\|^2 \\
= &\underbrace{-\nu(\alpha-\nu-1)\left\langle x_{k+2}-x_{k+1}, x_{k+2}-x^\star\right\rangle}_{\mathbf{III}_1} \\
&\underbrace{-\frac{(2k+\alpha+3-\nu)(\alpha-\nu-1)}{2}\left\|x_{k+2}-x_{k+1}\right\|^2}_{\mathbf{III}_2} \\
&\underbrace{-s(k+\alpha+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\langle\nabla f(x_{k+1}), x_{k+2}-x_{k+1}\right\rangle}_{\mathbf{III}_3} \\
&\underbrace{-s\nu\left(k+\alpha+1-\beta\right)\left\langle\nabla f(x_{k+1}), x_{k+1}-x^\star\right\rangle}_{\mathbf{III}_4} \\
&\underbrace{-\frac{1}{2}s^2\left[k+\alpha+1-\beta+2(k+2)\beta\right]\left(k+\alpha+1-\beta\right)\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_5}.
\end{aligned}
$$

Apparently, we can observe that

$$\mathbf{II}_1+\mathbf{III}_1=0,$$

and

$$\mathbf{II}_2+\mathbf{III}_2=-\frac{s(2k+\alpha+3)(\alpha-\nu-1)}{2}\left\|v_{k+1}\right\|^2.$$

Using the basic inequality for $f(x)\in\mathcal{F}_L^1(\mathbb{R}^n)$

$$f(x_k)\geq f(x_{k+1})+\left\langle\nabla f(x_{k+1}), x_k-x_{k+1}\right\rangle+\frac{1}{2L}\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2,$$

we have

$$
\begin{aligned}
&\mathbf{I}_1+\mathbf{III}_3+\mathbf{III}_5 \\
=&s(k+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left(f(x_{k+1})-f(x_k)\right) \\
&- s(k+\alpha+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\langle\nabla f(x_{k+1}), x_{k+2}-x_{k+1}\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \frac{1}{2}s^2\left[k+\alpha+1-\beta+2(k+2)\beta\right](k+\alpha+1-\beta)\left\|\nabla f(x_{k+1})\right\|^2 \\
&\leq - s^{\frac{3}{2}}\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\langle \nabla f(x_{k+1}), (k+\alpha+1)v_{k+1}-(k+1)v_k\right\rangle \\
&\quad - \frac{s(k+1)}{2L}\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2 \\
&\quad - \frac{1}{2}s^2\left[k+\alpha+1-\beta+2(k+2)\beta\right](k+\alpha+1-\beta)\left\|\nabla f(x_{k+1})\right\|^2 .
\end{aligned}
$$

Utilizing (5.129) again, we have

$$
\begin{aligned}
&\mathbf{I}_1+\mathbf{III}_3+\mathbf{III}_5 \\
\leq&\beta s^2(k+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\langle \nabla f(x_{k+1}), \nabla f(x_{k+1})-\nabla f(x_k)\right\rangle \\
&\quad + s^2(k+\alpha+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\|\nabla f(x_{k+1})\right\|^2 \\
&\quad - \frac{s(k+1)}{2L}\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2 \\
&\quad - \frac{1}{2}s^2\left[k+\alpha+1-\beta+2(k+2)\beta\right](k+\alpha+1-\beta)\left\|\nabla f(x_{k+1})\right\|^2 \\
\leq&\frac{L\beta^2s^2}{2}(k+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\|\nabla f(x_{k+1})\right\|^2 \\
&\quad + s^2(k+\alpha+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\|\nabla f(x_{k+1})\right\|^2 \\
&\quad - \frac{1}{2}s^2\left[k+\alpha+1-\beta+2(k+2)\beta\right](k+\alpha+1-\beta)\left\|\nabla f(x_{k+1})\right\|^2 \\
=&s^2\left[\frac{L\beta^2s}{2}(k+1)+(k+\alpha+1)\right]\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\|\nabla f(x_{k+1})\right\|^2 \\
&\quad - \frac{1}{2}s^2\left[(2\beta+1)k+\alpha+1+3\beta\right](k+\alpha+1-\beta)\left\|\nabla f(x_{k+1})\right\|^2 \\
&\leq s^2\left[\frac{L\beta^2s}{2}(k+1)+(k+\alpha+1)\right]\left[k+\alpha+1-\beta+(\alpha-1-\nu)\beta\right]\left\|\nabla f(x_{k+1})\right\|^2 \\
&\quad - \frac{1}{2}s^2\left[(2\beta+1)k+\alpha+1+3\beta\right](k+\alpha+1-\beta)\left\|\nabla f(x_{k+1})\right\|^2
\end{aligned}
$$

Since $\beta > 1/2$, let $n \in \mathbb{N}^+$ satisfy

$$
n = \left\lfloor \frac{2}{2\beta-1} \right\rfloor + 1.
$$

When $k \geq n(\alpha-1-\nu)\beta-(\alpha+1-\beta)$, we have

$$
\mathbf{I}_1+\mathbf{III}_3+\mathbf{III}_5
$$

$$
\begin{aligned}
\leq & s^2 \left[\frac{L\beta^2 s}{2}(k+1) + (k+\alpha+1) \right] [k+\alpha+1-\beta+(\alpha-1-\nu)\beta] \|\nabla f(x_{k+1})\|^2 \\
& - \frac{s^2 n}{2(n+1)} \cdot [(2\beta+1)k+\alpha+1+3\beta] [k+\alpha+1-\beta+(\alpha-1-\nu)\beta] \|\nabla f(x_{k+1})\|^2
\end{aligned}
$$

With the monotonicity of the following function about k

$$
\begin{aligned}
h(k) = & \frac{\left(\frac{n(2\beta+1)}{2(n+1)} - 1\right) k + \frac{n}{2(n+1)} \cdot (\alpha+1+3\beta) - \alpha - 1}{\frac{L\beta^2(k+1)}{2}} \\
= & \frac{(2\beta n - n - 2)(k+1) + (\beta-\alpha)n - 2\alpha}{L\beta^2(n+1)(k+1)},
\end{aligned}
$$

we know there exists some constant $\mathfrak{c}_{\alpha,\beta,\nu}$ and $k_{1,\alpha,\beta,\nu}$ such that the step size satisfies $0 < s \leq \mathfrak{c}_{\alpha,\beta,\nu}/L$. When $k \geq k_{1,\alpha,\beta,\nu}$, the following inequality holds

$$
\mathbf{I}_1 + \mathbf{III}_3 + \mathbf{III}_5 \leq -\frac{s^2}{2} \left(\frac{2\beta n}{n+1} - \frac{n+2}{n+1} - L\beta^2 s \right) (k - k_{1,\alpha,\beta,\nu})^2 \|\nabla f(x_{k+1})\|^2 .
$$

With the basic inequality for $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$,

$$
f(x^\star) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x^\star - x_{k+1} \rangle ,
$$

we know that there exists $k_{2,\alpha,\beta,\nu}$ such that when $k \geq k_{2,\alpha,\beta,\nu}$,

$$
\mathbf{I}_2 + \mathbf{III}_4 \leq -s(\nu-2)(k - k_{2,\alpha,\beta,\nu}) \langle \nabla f(x_{k+1}), x_{k+1} - x^\star \rangle .
$$

Let $k_{\alpha,\beta,\nu} = \max\{k_{1,\alpha,\beta,\nu}, k_{2,\alpha,\beta,\nu}\} + 1$. Summing up all the estimates above, when $\beta > 1/2$, the difference of discrete Lyapunov function, for any $k \geq k_{\alpha,\beta,\nu}$,

$$
\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) \leq & -\frac{s^2}{2} \left(\frac{2\beta n}{n+1} - \frac{n+2}{n+1} - L\beta^2 s \right) (k - k_{\alpha,\beta,\nu})^2 \|\nabla f(x_{k+1})\|^2 \\
& - s(\nu-2)(k - k_{\alpha,\beta,\nu}) \langle \nabla f(x_{k+1}), x_{k+1} - x^\star \rangle \\
& - \frac{s(2k+\alpha+3)(\alpha-\nu-1)}{2} \|v_{k+1}\|^2 .
\end{aligned}
$$

With the basic inequality for any function $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$

$$
\langle \nabla f(x_{k+1}), x_{k+1} - x^\star \rangle \geq f(x_{k+1}) - f(x^\star),
$$

we can obtain the following lemma.

Lemma 5.7.14. Under the same assumption of Theorem 5.5.3, the following limit exists

$$\lim_{k\to\infty} \mathcal{E}(k)$$

and the summation of the following series exist

$$\sum_{k=0}^{\infty}(k+1)^2 \left\|\nabla f(x_{k+1})\right\|^2, \qquad \sum_{k=0}^{\infty}(k+1)\left\langle \nabla f(x_{k+1}), x_{k+1}-x^\star\right\rangle,$$
$$\sum_{k=0}^{\infty}(k+1)(f(x_{k+1})-f(x^\star)), \qquad \sum_{k=0}^{\infty}(k+1)\left\|v_{k+1}\right\|^2.$$

Lemma 5.7.15. Under the same assumption of Theorem 5.5.3, the following limits exist

$$\lim_{k\to\infty}\|x_k-x^\star\| \quad \textbf{and} \quad \lim_{k\to\infty}(k+1)\left\langle x_{k+1}-x^\star, v_k+\beta\sqrt{s}\nabla f(x_k)\right\rangle.$$

Proof. [Proof of Lemma 5.7.15] Taking $\nu \neq \nu' \in (2,\gamma-1]$, we have

$$\begin{aligned}\mathcal{E}_\nu(k)-\mathcal{E}_{\nu'}(k) = (\nu-\nu')\Bigg[&-s\beta\cdot\frac{(k+1)(k+2)}{k+\alpha+1}\left(f(x_k)-f(x^\star)\right)\\ &+(k+1)\sqrt{s}\left\langle x_{k+1}-x^\star, v_k+\beta\sqrt{s}\nabla f(x_k)\right\rangle+\frac{(\alpha-1)}{2}\left\|x_{k+1}-x^\star\right\|^2\Bigg]\end{aligned}$$

With Lemma 5.7.14, the following limit exists

$$\lim_{k\to\infty}\left[(k+1)\sqrt{s}\left\langle x_{k+1}-x^\star, v_k+\beta\sqrt{s}\nabla f(x_k)\right\rangle+\frac{\alpha-1}{2}\left\|x_{k+1}-x^\star\right\|^2\right]. \tag{5.138}$$

Define a new function about k:

$$\pi(k) := \frac{1}{2}\left\|x_k-x^\star\right\|^2+\beta s\sum_{i=k_0}^{k-1}\left\langle\nabla f(x_i), x_{i+1}-x^\star\right\rangle.$$

If we can show the existence of the limit $\pi(k)$ with $k\to\infty$, we can guarantee $\lim\limits_{k\to\infty}\|x_{k+1}-x^\star\|$ exists with Lemma 5.7.14. We observe the following equality

$$(k+1)(\pi(k+1)-\pi(k))+(\alpha-1)\pi(k+1)-s\,(\alpha-1)\,\beta\sum_{i=0}^{k}\left\langle\nabla f(x_i), x_{i+1}-x^\star\right\rangle$$

$$
\begin{aligned}
=&(k+1)\left\langle x_{k+1}-x_k, x_{k+1}-x^\star\right\rangle - \frac{(k+1)s}{2}\left\|v_k\right\|^2\\
&+\frac{\alpha-1}{2}\left\|x_{k+1}-x^\star\right\|^2 + s(k+1)\beta\left\langle\nabla f(x_k), x_{k+1}-x^\star\right\rangle\\
=&(k+1)\sqrt{s}\left\langle x_{k+1}-x^\star, v_k+\beta\sqrt{s}\nabla f(x_k)\right\rangle - \frac{(k+1)s}{2}\left\|v_k\right\|^2 + \frac{\alpha-1}{2}\left\|x_{k+1}-x^\star\right\|^2.
\end{aligned}
$$

Lemma 5.7.14 and (5.138) tell us there exists some constant $\mathfrak{C}^5$ such that

$$
\lim_{k\to\infty}\left[(k+\alpha)\pi(k+1)-(k+1)\pi(k)\right]=\mathfrak{C}^5,
$$

that is, taking a simple translation $\pi'(k)=\pi(k)-\mathfrak{C}^5/(\gamma-1)$, we have

$$
\lim_{k\to\infty}\left[(k+\alpha)\pi'(k+1)-(k+1)\pi'(k)\right]=0.
$$

Since $\mathcal{E}(k)$ decreases for $k\geq k_{\alpha,\beta,\nu}$, thus, $\left\|x_k-x^\star\right\|^2$ is bounded. With Lemma 5.7.14, we obtain that $\pi(k)$ is bounded, that is, $\pi'(k)$ is bounded. Then we have

$$
\lim_{k\to\infty}\frac{(k+2)^{\alpha-1}\pi'(k+1)-(k+1)^{\alpha-1}\pi'(k)}{(k+1)^{\alpha-2}}=0,
$$

that is, for any $\epsilon>0$, there exists $k_0'>0$ such that

$$
|\pi'(k)|\leq\left(\frac{k_0'+1}{k+1}\right)^{\alpha-1}|\pi'(k_0')|+\frac{\epsilon\sum\limits_{i=k_0'}^{k-1}(i+1)^{\alpha-2}}{(k+1)^{\alpha-1}}.
$$

With arbitrary $\epsilon>0$, we complete the proof of Lemma 5.7.15. □

Proof. [Proof of (5.136)] When $k\geq k_{\alpha,\beta,\nu}$, we expand the discrete Lyapunov function (5.137) as

$$
\begin{aligned}
\mathcal{E}(k)=s(k+1)&\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right](f(x_k)-f(x^\star))\\
&+\sqrt{s}(k+1)\nu\left\langle x_{k+1}-x^\star, v_k+\beta\sqrt{s}\nabla f(x_k)\right\rangle\\
&\quad+\frac{\nu(\alpha-1)}{2}\left\|x_{k+1}-x^\star\right\|^2+\frac{s(k+1)^2}{2}\left\|v_k+\beta\sqrt{s}\nabla f(x_k)\right\|^2.
\end{aligned}
$$

With Lemma 5.7.14 and Lemma 5.7.15, we obtain the first equation of (5.136). Additionally, we have

$$\begin{aligned} & s\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left(f(x_k)-f(x^\star)\right) \\ & +\frac{(k+1)s}{2}\left\|v_k+\beta\sqrt{s}\nabla f(x_k)\right\|^2 \\ \leq & s\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left(f(x_k)-f(x^\star)\right) \\ & +(k+1)s\left\|v_k\right\|^2+(k+1)\beta^2 s^2\left\|\nabla f(x_k)\right\|^2. \end{aligned}$$

With Lemma 5.7.14, we obtain the second equation of (5.136). □

CHAPTER 6

IMPROVED SAMPLE COMPLEXITY IN SPARSE SUBSPACE CLUSTERING WITH NOISY AND MISSING ENTRIES

In this chapter, we show the results about the new CoCoSSC algorithm. The content is organized as follows. The main results about CoCoSSC algorithm are shown in Section 6.1. Following Section 6.1, we show the full proofs in Section 6.2. In Section 6.3, we show the performance for CoCoSSC algorithm and some related algorithms numerically. Finally, we conclude this work and some future directions.

6.1 Main Results about CoCoSSC Algorithm

We introduce our main results analyzing the performance of CoCoSSC under both the Gaussian noise model and the missing data model. Similar to [WX16], the quality of the computed self-similarity matrix $\{\boldsymbol{c}_i\}_{i=1}^N$ is assessed using a *subspace detection property (SDP)*:

Definition 6.1.1 (Subspace detection property (SDP), [WX16])**.** The self-similarity matrix $\{\boldsymbol{c}_i\}_{i=1}^N$ satisfies the *subspace detection property* if 1) for every $i \in [N]$, $\boldsymbol{c}_i$ is a non-zero vector; and 2) for every $i, j \in [N]$, $\boldsymbol{c}_{ij} \neq 0$ implies that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same cluster.

Intuitively, the subspace detection property asserts that the self-similarity matrix $\{\boldsymbol{c}_i\}_{i=1}^N$ has *no false positives*, where every non-zero entry in $\{\boldsymbol{c}_i\}_{i=1}^n$ links two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ to the same cluster. The first property in Definition 6.1.1 further rules out the trivial solution of $\boldsymbol{c}_i \equiv \mathbf{0}$.

The SDP stated in Definition 6.1.1 is, however, *not* sufficient for the success of a follow-up spectral clustering algorithm, or any clustering algorithm, as the "similarity graph" constructed by connecting every pairs of (i, j) with $\boldsymbol{c}_{ij} \neq 0$ might be poorly connected. Such "graph connectivity" is a well-known open problem in sparse subspace clustering [NH11] and remains largely unsolved except under strong assumptions [WWS16]. Nevertheless, in practical scenarios the SDP criterion correlates reasonably well with clustering performance [WX16, WWS15a] and therefore we choose to focus on the SDP success condition only.

6.1.1 The Non-Uniform Semi-Random Model

We adopt the following non-uniform semi-random model throughout the paper:

Definition 6.1.2 (Non-uniform semi-random model)**.** Suppose $\boldsymbol{y}_i$ belongs to cluster $\mathcal{S}_\ell$ and let $\boldsymbol{y}_i = \mathbf{U}_\ell \boldsymbol{\alpha}_i$, where $\mathbf{U}_\ell \in \mathbb{R}^{n \times d_\ell}$ is an orthonormal basis of $\mathcal{U}_\ell$ and $\boldsymbol{\alpha}_i$ is a d_ℓ-dimensional vector with $\|\boldsymbol{\alpha}_i\|_2 = 1$. We assume that $\boldsymbol{\alpha}_i$ are i.i.d. distributed according to an unknown underlying distribution P_ℓ, and that the density p_ℓ associated with P_ℓ satisfies

$$0 < \underline{C} \cdot p_0 \le p_\ell(\boldsymbol{\alpha}) \le \overline{C} \cdot p_0 < \infty \qquad \forall \boldsymbol{\alpha} \in \mathbb{R}^{d_\ell},\ \ \|\boldsymbol{\alpha}\|_2 = 1$$

for some constants $\underline{C}, \overline{C}$, where p_0 is the density of the uniform measure on $\{\boldsymbol{u} \in \mathbb{R}^{d_\ell} : \|\boldsymbol{u}\|_2 = 1\}$.

Remark 6.1.1. Our non-uniform semi-random model ensures that $\|\boldsymbol{y}_i\|_2 = 1$ for all $i \in [N]$, a common normalizing assumption made in previous works on sparse subspace clustering [SC12, SEC14, WX16]. However, such a property is only used in our theoretical analysis, and in our CoCoLasso algorithm the norms of $\{y_i\}_{i=1}^N$ are assumed unknown. Indeed, if the exact norms of $\|\boldsymbol{y}_i\|_2$ are known to the data analyst the sample complexity in our analysis can be further improved, as we remark in Remark 6.1.3.

The non-uniform semi-random model considers fixed (deterministic) subspaces $\{\mathcal{S}_\ell\}$, but assumes that data points within each low-dimensional subspace are independently generated from an unknown distribution P_ℓ with densities bounded away and above from below. This helps simplifying the "inter-subspace incoherence" (Definition 6.2.5) in our proof and yields interpretable results.

Compared with existing definitions of semi-random models [SC12, WX16, HB15, PCS14], the key difference is that in our model data are *not* uniformly distributed on each low-dimensional subspace. Instead, it is assumed that the data points are i.i.d., and that the data density is bounded away from both above and below. Such non-uniformity rules out algorithms that exploit the $\mathbb{E}[\boldsymbol{y}_i] = \mathbf{0}$ property in traditional semi-random models which is too strong and rarely holds true in practice.

Because the underlying subspaces are fixed, quantities that characterize the "affinity" between these subspace are needed because closer subspaces are harder to distinguish from each other. We adopt the following affinity measure, which was commonly used in previous works on sparse subspace clustering [WX16, WWS15a, CJW17]:

Definition 6.1.3 (Subspace affinity)**.** Let $\mathcal{U}_j$ and $\mathcal{U}_k$ be two linear subspaces of $\mathbb{R}^n$ of dimension d_j and d_k. The *affinity* between $\mathcal{U}_j$ and $\mathcal{U}_k$ is defined as $\chi^2_{j,k} := \cos^2\theta^{(1)}_{jk} + \cdots + \cos^2\theta^{(\min(d_j,d_k))}_{jk}$, where $\theta^{(\ell)}_{jk}$ is the ℓth canonical angle between $\mathcal{U}_j$ and $\mathcal{U}_k$.

Remark 6.1.2. $\chi_{jk} = \|\mathbf{U}_j^\top \mathbf{U}_k\|_F$ where $\mathbf{U}_j \in \mathbb{R}^{n\times d_j}, \mathbf{U}_k \in \mathbb{R}^{n\times d_k}$ are orthonormal basis of $\mathcal{U}_j, \mathcal{U}_k$.

Throughout the paper we also write $\chi := \max_{j\neq k}\chi_{j,k}$.

For the missing data model, we need the following additional "inner-subspace" incoherence of the subspaces to ensure that the observed data entries contain sufficient amount of information. Such incoherence assumptions were widely adopted in the matrix completion community [CR09, KMO10, Rec11].

Definition 6.1.4 (Inner-subspace incoherence)**.** Fix $\ell \in [L]$ and let $\mathbf{U}_\ell \in \mathbb{R}^{n\times d_\ell}$ be an orthonormal basis of subspace $\mathcal{U}_\ell$. The *subspace incoherence* of $\mathcal{U}_\ell$ is the smallest

μ_ℓ such that

$$\max_{1\leq i\leq n} \|\boldsymbol{e}_i^\top \mathbf{U}_\ell\|_2^2 \leq \mu_\ell d_\ell/n.$$

With the above definitions, we are now ready to state the following two theorems which give sufficient success conditions for the self-similarity matrix $\{\boldsymbol{c}_i\}_{i=1}^n$ produced by CoCoLasso.

Theorem 6.1.5 (The Gaussian noise model)**.** Suppose $\lambda \asymp 1/\sqrt{d}$ and $\boldsymbol{\Delta}_{jk} \asymp \sigma^2\sqrt{\frac{\log N}{n}}$ for all $j,k \in [N]$. Suppose also that $N_\ell \geq 2\overline{C}d_\ell/\underline{C}$. There exists a constant $K_0 > 0$ such that, if

$$\sigma < K_0\left(n/d^3\log^2(\overline{C}N/\underline{C})\right)^{1/4},$$

then the optimal solution $\{\boldsymbol{c}_i\}_{i=1}^N$ of the CoCoSSC estimator satisfies the subspace detection property (SDP) with probability $1-O(N^{-10})$.

Theorem 6.1.6 (The missing data model)**.** Suppose $\lambda \asymp 1/\sqrt{d}$, $\boldsymbol{\Delta}_{jk} \asymp \frac{\mu d\log N}{\rho\sqrt{n}}$ for $j \neq k$ and $\boldsymbol{\Delta}_{jk} \asymp \frac{\mu d\log N}{\rho^{3/2}\sqrt{n}}$ for $j = k$. Suppose also that $N_\ell \geq 2\overline{C}d_\ell/\underline{C}$. There exists a constant $K_1 > 0$ such that, if

$$\rho > K_1\max\left\{(\mu\chi d^{5/2}\log^2 N)^{2/3}\cdot n^{-1/3}, (\mu^2 d^{7/2}\log^2 N)^{2/5}\cdot n^{-2/5}\right\},$$

then the optimal solution $\{\boldsymbol{c}_i\}_{i=1}^N$ of the CoCoSSC estimator satisfies the subspace detection property (SDP) with probability $1-O(N^{-10})$.

Remark 6.1.3. If the norms of the data points $\|\boldsymbol{y}_i\|_2$ are exactly known and can be explicitly used in algorithm design, the diagonal terms of $\mathbf{A}$ in Eq. (1.21) can be directly set to $\mathbf{A}_{ii} = \|\boldsymbol{y}_i\|_2^2$ in order to avoid the ψ_2 concentration term in our proof (Definition 6.2.1). This would improve the sample complexity in the success condition to $\rho > \Omega(n^{-1/2})$, matching the sample complexity in linear regression problems with missing design entries [WWBS17].

Theorems 6.1.5 and 6.1.6 show that when the noise magnitude (σ in the Gaussian noise model and ρ^{-1} in the missing data model) is sufficiently small, a careful choice of tuning parameter λ results in a self-similarity matrix $\{\boldsymbol{c}_i\}$ satisfying the subspace detection property. Furthermore, the maximum amount of noise our method can tolerate is $\sigma = O(n^{1/4})$ and $\rho = \Omega(\chi^{2/3}n^{-1/3} + n^{-2/5})$, which improves over the sample complexity of existing methods (see Table 1.1).

6.1.2 The Fully Random Model

When the underlying subspaces $\mathcal{U}_1, \cdots, \mathcal{U}_L$ are independently uniformly sampled, a model referred to as the *fully random* model in the literature [SC12, SEC14, WX16], the success condition in Theorem 6.1.6 can be further simplified:

Corollary 6.1.7. Suppose subspaces $\mathcal{U}_1, \cdots, \mathcal{U}_L$ have the same intrinsic dimension d and are uniformly sampled, the condition in Theorem 6.1.6 can be simplified to

$$\rho > \widetilde{K}_1(\mu^2 d^{7/2}\log^2 N)^{2/5} \cdot n^{-2/5},$$

where $\widetilde{K}_1 > 0$ is a new universal constant.

Corollary 6.1.7 shows that in the fully-random model, the $\chi^{2/3}n^{-1/3}$ term in Theorem 6.1.6 is negligible and the success condition becomes $\rho = \Omega(n^{-2/5})$, strictly improving existing results (see Table 1.1).

6.2 Proofs

In this section we give proofs of our main results. Due to space constraints, we only give a proof framework and leave the complete proofs of all technical lemmas to the appendix.

6.2.1 Noise Characterization and Feasibility of Pre-Processing

Definition 6.2.1 (Characterization of noise variables)**.** $\{\boldsymbol{z}_i\}$ are independent random variables and $\mathbb{E}[\boldsymbol{z}_i] = \mathbf{0}$. Furthermore, there exist parameters $\psi_1, \psi_2 > 0$ such that with probability $1 - O(N^{-10})$ the following holds uniformly for all $i, j \in [N]$:

$$\left|\boldsymbol{z}_i^\top \boldsymbol{y}_j\right| \leq \psi_1 \sqrt{\frac{\log N}{n}}; \qquad \left|\boldsymbol{z}_i^\top \boldsymbol{z}_j - \mathbb{E}[\boldsymbol{z}_i^\top \boldsymbol{z}_j]\right| \leq \begin{cases} \psi_1 \sqrt{\frac{\log N}{n}} & i \neq j; \\ \psi_2 \sqrt{\frac{\log N}{n}} & i = j. \end{cases}$$

Proposition 6.2.1. Suppose $\boldsymbol{\Delta}$ are set as $\boldsymbol{\Delta}_{jk} \geq 3\psi_1 \sqrt{\frac{\log N}{n}}$ for $j \neq k$ and $\boldsymbol{\Delta}_{jk} \geq 3\psi_2 \sqrt{\frac{\log N}{n}}$ for $j = k$. Then with probability $1-O(N^{-10})$ the set S defined in Eq. (1.21) is not empty.

The following two lemmas derive explicit bounds on ψ_1 and ψ_2 for the two noise models.

Lemma 6.2.2. The Gaussian noise model satisfies Definition 6.2.1 with $\psi_1 \lesssim \sigma^2$ and $\psi_2 \lesssim \sigma^2$.

Lemma 6.2.3. Suppose $\rho = \Omega(n^{-1/2})$. The missing data model satisfies Definition 6.2.1 with $\psi_1 \lesssim \rho^{-1}\mu d\sqrt{\log N}$ and $\psi_2 \lesssim \rho^{-3/2}\mu d\sqrt{\log N}$, where $d = \max_{\ell \in [L]} d_\ell$ and $\mu = \max_{\ell \in [L]} \mu_\ell$.

6.2.2 Optimality Condition and Dual Certificates

We first write down the dual problem of CoCoSSC:

$$\text{Dual CoCoSSC}: \qquad \boldsymbol{\nu}_i = \arg\max_{\boldsymbol{\nu}_i \in \mathbb{R}^N} \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\nu}_i - \frac{1}{2\lambda}\|\boldsymbol{\nu}_i\|_2^2 \quad s.t. \quad \left\|\widetilde{\mathbf{X}}_{-i}^\top \boldsymbol{\nu}_i\right\|_\infty \leq 1. \quad (6.1)$$

Lemma 6.2.4 (Dual certificate, Lemma 12 of [WX16])**.** Suppose there exists triplet $(\boldsymbol{c}, \boldsymbol{e}, \boldsymbol{\nu})$ such that $\widetilde{\boldsymbol{x}}_i = \widetilde{\mathbf{X}}_{-i}\boldsymbol{c} + \boldsymbol{e}$, $\boldsymbol{c}$ has support $S \subseteq T \subseteq [N]$, and that $\boldsymbol{\nu}$ satisfies

$$[\widetilde{\mathbf{X}}_{-i}]_S^\top \boldsymbol{\nu} = \mathrm{sgn}(\boldsymbol{c}_S), \quad \boldsymbol{\nu} = \lambda \boldsymbol{e}, \quad \left\|[\widetilde{\mathbf{X}}_{-i}]_{T\cap S^c}^\top \boldsymbol{\nu}\right\|_\infty \leq 1, \quad \left\|[\widetilde{\mathbf{X}}_{-i}]_{T^c}^\top \boldsymbol{\nu}\right\|_\infty < 1,$$

then any optimal solution $\boldsymbol{c}_i$ to Eq. (1.22) satisfies $[\boldsymbol{c}_i]_{T^c} = \mathbf{0}$.

To construct such a dual certificate and to de-couple potential statistical dependency, we follow [WX16] to consider a constrained version of the optimization problem. Let $\widetilde{\mathbf{X}}_{-i}^{(\ell)}$ denote the data matrix of all but $\widetilde{\boldsymbol{x}}_i$ in cluster $\mathcal{S}_\ell$. The constrained problems are defined as follows:

$$\text{Constrained Primal :} \quad \widetilde{\boldsymbol{c}}_i = \arg \min_{\boldsymbol{c}_i \in \mathbb{R}^{N_\ell - 1}} \|\boldsymbol{c}_i\|_1 + \lambda/2 \cdot \|\widetilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}_{-i}^{(\ell)} \boldsymbol{c}_i\|_2^2; \tag{6.2}$$

$$\text{Constrained Dual :} \quad \widetilde{\boldsymbol{\nu}}_i = \arg \max_{\boldsymbol{\nu}_i \in \mathbb{R}^{N_\ell - 1}} \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\nu}_i - 1/(2\lambda) \cdot \|\boldsymbol{\nu}_i\|_2^2 \quad s.t. \quad \|(\widetilde{\mathbf{X}}_{-i}^{(\ell)})^\top \boldsymbol{\nu}_i\|_\infty \leq 1. \tag{6.3}$$

With $\boldsymbol{c} = [\widetilde{\boldsymbol{c}}_i, \mathbf{0}_{\mathcal{S}_{-\ell}}]$, $\boldsymbol{\nu} = [\widetilde{\boldsymbol{\nu}}_i, \mathbf{0}_{\mathcal{S}_{-\ell}}]$ and $\boldsymbol{e} = \widetilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}_{-i}^{(\ell)} \widetilde{\boldsymbol{c}}_i$, the certificate satisfies the first three conditions in Lemma 6.2.4 with $T = \mathcal{S}_\ell$ and $S = \operatorname{supp}(\widetilde{\boldsymbol{c}}_i)$. Therefore, we only need to establish that $|\langle \widetilde{\boldsymbol{x}}_j, \widetilde{\boldsymbol{\nu}}_i \rangle| < 1$ for all $\widetilde{\boldsymbol{x}}_j \notin \mathcal{S}_\ell$ to show no false discoveries, which we prove in the next section.

6.2.3 Deterministic Success Conditions

Define the following deterministic quantities as *inter-subspace incoherence* and *in-radius*, which are important quantities in deterministic analysis of sparse subspace clustering methods [SC12, WX16, SEC14].

Definition 6.2.5 (Inter-subspace incoherence)**.** The inter-subspace incoherence $\widetilde{\mu}$ is defined as $\widetilde{\mu} := \max_{\ell \in [L]} \max_{\boldsymbol{y}_i \in \mathcal{S}_\ell} \max_{\boldsymbol{y}_j \notin \mathcal{S}_\ell} \left|\langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle\right|$.

Definition 6.2.6 (In-radius)**.** Define r_i as the radius of the largest ball inscribed in the convex body of $\{\pm \mathbf{Y}_{-j}^{(\ell)}\}$. Also define that $r := \min_{1 \leq i \leq N} r_i$.

The following lemma derives an upper bound on $|\langle \widetilde{\boldsymbol{x}}_j, \widetilde{\boldsymbol{\nu}}_i \rangle|$, which is proved in the appendix.

Lemma 6.2.7. For every (i,j) belonging to different clusters, $|\langle \widetilde{\boldsymbol{x}}_j, \widetilde{\boldsymbol{\nu}}_i\rangle| \lesssim \lambda(1 + \|\widetilde{\boldsymbol{c}}_i\|_1)(\widetilde{\mu} + \psi_1\sqrt{\log N/n})$, where $\|\widetilde{\boldsymbol{c}}_i\|_1 \lesssim r^{-1}(1 + r^{-1}\lambda(\psi_1 + \psi_2)\sqrt{\log N/n})$.

Lemmas 6.2.4 and 6.2.7 immediately yield the following theorem:

Theorem 6.2.8 (no false discoveries)**.** There exists an absolute constant $\kappa_1 > 0$ such that if

$$\tfrac{\lambda}{r}\left(1 + \tfrac{\lambda}{r}(\psi_1 + \psi_2)\sqrt{\tfrac{\log N}{n}}\right) \cdot \left(\widetilde{\mu} + \psi_1\sqrt{\tfrac{\log N}{n}}\right) < \kappa_1, \tag{6.4}$$

then the optimal solution $\boldsymbol{c}_i$ of the CoCoSSC estimator in Eq. (1.22) has no false discoveries; that is, $\boldsymbol{c}_{ij} = 0$ for all $\boldsymbol{x}_j$ that belongs to a different cluster of $\boldsymbol{x}_i$.

The following theorem shows conditions under which $\boldsymbol{c}_i$ is not the trivial solution $\boldsymbol{c}_i = \boldsymbol{0}$.

Theorem 6.2.9 (Avoiding trivial solutions)**.** There exists an absolute constant $\kappa_2 > 0$ such that, if

$$\lambda\left(r - \psi_1\sqrt{\tfrac{\log N}{n}}\right) > \kappa_2, \tag{6.5}$$

then the optimal solution $\boldsymbol{c}_i$ of the CoCoSSC estimator in Eq. (1.22) is non-trivial; that is, $\boldsymbol{c}_i \neq \boldsymbol{0}$.

Finally, we remark that choosing $r = c/\lambda$ for some small constant $c > 0$ (depending only on κ_1 and κ_2), the choice of λ satisfies both Theorems 6.2.8 and 6.2.9 provided that

$$\max\left\{\tfrac{\psi_1}{r}\sqrt{\tfrac{\log N}{n}}, \tfrac{\widetilde{\mu}}{r^2}, \tfrac{\widetilde{\mu}(\psi_1+\psi_2)}{r^3}\sqrt{\tfrac{\log N}{n}}, \tfrac{\psi_1(\psi_1+\psi_2)}{r^3}\tfrac{\log N}{n}\right\} < \kappa_3 \tag{6.6}$$

for some sufficiently small absolute constant $\kappa_3 > 0$ that depends on κ_1, κ_2 and c.

6.2.4 Bounding $\widetilde{\mu}$ and r in Randomized Models

Lemma 6.2.10. Suppose $N_\ell = \Omega(\overline{C}d_\ell/\underline{C}_\ell)$ Under the non-uniform semi-random model, with probability $1 - O(N^{-10})$ it holds that $\widetilde{\mu} \lesssim \chi\sqrt{\log(\overline{C}N/\underline{C})}$ and $r \gtrsim 1/\sqrt{d}$.

Lemma 6.2.11. Suppose $\mathcal{U}_1, \cdots, \mathcal{U}_L$ are independently uniformly sampled linear subspaces of dimension d in $\mathbb{R}^n$. Then with probability $1 - O(N^{-10})$ we have that $\chi \lesssim d\sqrt{\log N/n}$ and $\mu \lesssim \sqrt{\log N}$.

6.3 Numerical results

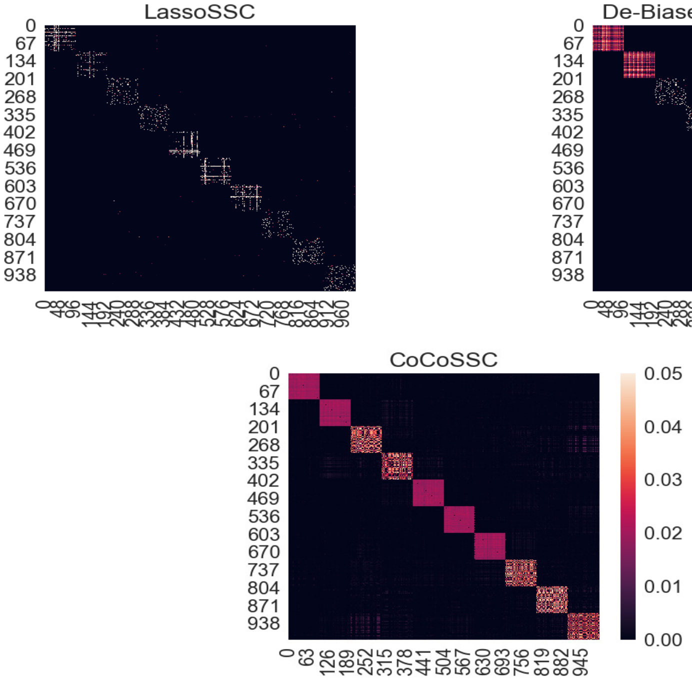


Figure 6.1: Heatmaps of similarity matrices $\{\boldsymbol{c}_i\}_{i=1}^N$, with brighter colors indicating larger absolute values of matrix entries. Left: LassoSSC; Middle: De-Biased Dantzig Selector; Right: CoCoSSC.

Experimental settings and methods We conduct numerical experiments based on synthetic generated data, using a computer with Intel Core i7 CPU (4 GHz) and 16GB memory. Each synthetic data set has ambient dimension $n = 100$, intrinsic dimension $d = 4$, number of underlying subspaces $L = 10$, and a total number of $N = 1000$ unlabeled data points. The observation rate ρ and Gaussian noise magnitude σ vary in our simulations. Underlying subspaces are generated uniformly at random, corresponding to our fully-random model. Each data point has an equal probability of being assigned to any cluster, and is generated uniformly at random on its corresponding low-dimensional subspace.

We compare the performance (explained later) of our proposed CoCoSSC approach, and two popular existing methods Lasso SSC and the de-biased Dantzig selector. The ℓ_1 regularized self-regression steps in both CoCoSSC and Lasso SSC are implemented using ADMM. The pre-processing step of CoCoSSC is implemented using alternating projections initialized at $\widetilde{\mathbf{\Sigma}} = \mathbf{X}^\top\mathbf{X} - \mathbf{D}$. Unlike the theoretical recommendations, we choose $\mathbf{\Delta}$ in Eq. (1.21) to be very large (3×10^3 for diagonal entries and 10^3 for off-diagonal entries) for fast convergence. The de-biased Dantzig selector is implemented using linear programming.

Evaluation measure We consider two measure to evaluate the performance of algorithms being compared. The first one evaluates the quality of the similarity matrix $\{\boldsymbol{c}_i\}_{i=1}^N$ by measuring how far (relatively) it deviates from having the subspace detection property. In particular, we consider the RelViolation metric proposed in [WX16] defined as

$$\text{RelViolation}(C, \mathcal{M}) = (\textstyle\sum_{(i,j)\notin\mathcal{M}} |C|_{i,j}) / (\textstyle\sum_{(i,j)\in\mathcal{M}} |C|_{i,j}). \tag{6.7}$$

where $\mathcal{M}$ is the mask of ground truth with all (i,j) satisfying $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{S}^{(\ell)}$ for some ℓ. A high RelViolation indicates frequent deviation from the subspace detection property and therefore poorer quality of $\{\boldsymbol{c}_i\}_{i=1}^N$.

For clustering results, we use the Fowlkes-Mallows index [FM83] to evaluate their quality. Suppose $\mathcal{A} \subseteq \{(i,j) \in [N]\times[N]\}$ consists of pairs of data points that are clustered together by a clustering algorithm, and $\mathcal{A}_0$ is the ground truth clustering. Define $TP = |\mathcal{A}\cap\mathcal{A}_0|$, $FP = |\mathcal{A}\cap\mathcal{A}_0^c|$, $FN = |\mathcal{A}^c\cap\mathcal{A}_0|$, $TN = |\mathcal{A}^c\cap\mathcal{A}_0^c|$. The Fowlkes-Mallows (FM) index is then expressed as

$$FM = \sqrt{TP^2/(TP+FP)(TP+FN)}.$$

The FM index of any two clusterings $\mathcal{A}$ and $\mathcal{A}_0$ is always between 0 and 1, with an FM index of one indicating perfectly identical clusterings and an FM index close to zero otherwise.

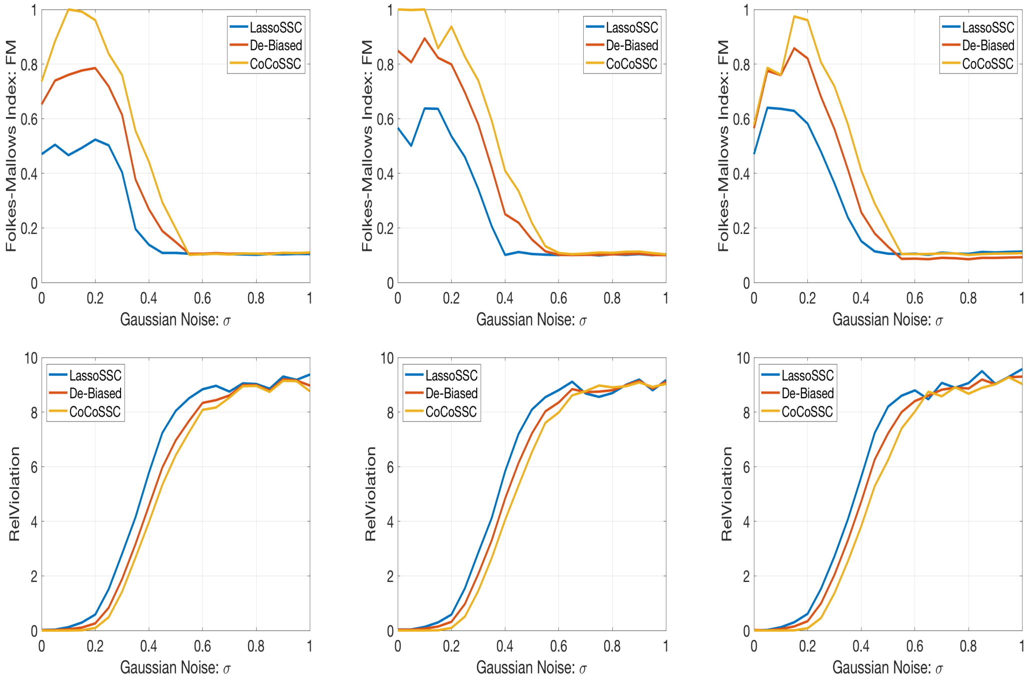


Figure 6.2: The Fowlkes-Mallows (FM) index of clustering results (top row) and RelViolation scores (bottom row) of the three methods, with noise of magnitude σ varying from 0 to 1. Left column: missing rate $1-\rho = 0.03$, middle column: $1-\rho = 0.25$, right column: $1-\rho = 0.9$.

Results We first give a qualitative illustration of similarity matrices $\{\boldsymbol{c}_i\}_{i=1}^N$ produced by the three algorithms of Lasso SSC, de-biased Dantzig selector and CoCoSSC in Fig. 6.1. We observe that the similarity matrix of Lasso SSC has several spurious connections, and both Lasso SSC and the de-biased Dantzig selector suffer from graph connectivity issues as signals within each block (cluster) are not very strong. On the other hand, the similarity matrix of CoCoSSC produces convincing signals within each block (cluster). This shows that our proposed CoCoSSC approach not only has few false discoveries as predicted by our theoretical results, but also has much better graph connectivity which our theory did not attempt to cover.

In Fig. 6.2 we report the Fowlkes-Mallows (FM) index for clustering results and RelViolation scores of similarity matrices $\{\boldsymbol{c}_i\}_{i=1}^N$ under various noise magnitude (σ) and observation rates (ρ) settings. A grid of tuning parameter values λ are attempted and the one leading to the best performance is reported. It is observed that our proposed CoCoLasso consistently outperforms its competitors Lasso SSC and de-biased Dantzig selector. Furthermore, CoCoLasso is very computationally efficient and converges in 8-15 seconds on each synthetic data set. On the other hand, de-biased Dantzig selector is computationally very expensive and typically takes over 100 seconds to converge.

6.4 Technical Details

Proof. [Proof of Proposition 6.2.1] By Definition 6.2.1 we know that $|\widetilde{\boldsymbol{\Sigma}}_{-i}-\mathbf{Y}_{-i}^T\mathbf{Y}_{-i}| \leq |\boldsymbol{\Delta}|$ in an element-wise sense. Also note that $\mathbf{Y}^\top\mathbf{Y}$ is positive semi-definite. Thus, $\mathbf{Y}^\top\mathbf{Y} \in S$. □

Proof. [Proofs of Lemmas 6.2.2 and 6.2.3] Lemma 6.2.2 is proved in [WX16]. See Lemmas 17 and 18 of [WX16] and note that $\mathbb{E}[\boldsymbol{z}_i^\top\boldsymbol{z}_i] = \sigma^2$.

We next prove Lemma 6.2.3. We first consider $|\boldsymbol{z}_i^\top\boldsymbol{y}_j|$. Let $\boldsymbol{z} = \boldsymbol{z}_i$, $\boldsymbol{y} = \boldsymbol{y}_i$, $\widetilde{\boldsymbol{y}} = \boldsymbol{y}_j$ and $\boldsymbol{r} = R_{j\cdot}$. Define $T_i := \boldsymbol{z}_i\boldsymbol{y}_i = (1-\boldsymbol{r}_i/\rho)\boldsymbol{y}_i\widetilde{\boldsymbol{y}}_j$. Because $\boldsymbol{r}$ is independent of $\boldsymbol{y}$ and $\widetilde{\boldsymbol{y}}$, we have that $\mathbb{E}[T_i] = 0$, $\mathbb{E}[T_i^2] \leq \boldsymbol{y}_i^2\widetilde{\boldsymbol{y}}_i{}^2/\rho \leq \mu^2 d^2/\rho n^2$ and $|T_i| \leq \mu d/\rho n =: M$ almost surely. Using Bernstein's inequality, we know that with probability $1 - O(N^{-10})$

$$|\boldsymbol{z}_i^\top\boldsymbol{y}_j| = \left|\sum_{i=1}^{T} T_i\right| \lesssim \sqrt{\sum_{i=1}^{n}\mathbb{E}[T_i^2]\cdot\log N + M\log N} \lesssim \mu d\sqrt{\frac{\log^2 N}{\rho n}}.$$

We next consider $|\boldsymbol{z}_i^\top\boldsymbol{z}_j|$ and the $i \neq j$ case. Let $\boldsymbol{y} = \boldsymbol{y}_i$, $\widetilde{\boldsymbol{y}} = \boldsymbol{y}_j$, $\boldsymbol{r} = R_{i\cdot}$ and $\widetilde{\boldsymbol{r}} = R_{j\cdot}$. By definition of μ, we have that $\|\boldsymbol{y}\|_\infty^2 \leq \mu d_i/n$ and $\|\widetilde{\boldsymbol{y}}\|_\infty^2 \leq \mu d_j/n$.

Define $T_i := \boldsymbol{z}_i\widetilde{\boldsymbol{z}}_i = (1-\boldsymbol{r}_i/\rho)(1-\widetilde{\boldsymbol{r}}_i/\rho)\cdot\boldsymbol{y}_i\widetilde{\boldsymbol{y}}_i$. Because $\boldsymbol{r}$ and $\widetilde{\boldsymbol{r}}$ are independent, $\mathbb{E}[T_i]=0$, $\mathbb{E}[T_i^2]\le\boldsymbol{y}_i^2\widetilde{\boldsymbol{y}}_i^2/\rho^2\le\mu^2d^2/\rho^2n^2$ and $|T_i|\le\mu d/\rho^2n =: M$ almost surely. Using Bernstein's inequality, we know that with probability $1-O(N^{-10})$

$$\left|\sum_{i=1}^n T_i\right| \lesssim \sqrt{\sum_{i=1}^n \mathbb{E}[T_i^2]\cdot\log N + M\log N} \lesssim \frac{\mu d}{\rho}\sqrt{\frac{\log^2 N}{n}},$$

where the last inequality holds because $\rho = O(n^{-1/2})$.

Finally is the case of $|\boldsymbol{z}_i^\top\boldsymbol{z}_j|$ and $i=j$. Let again $\boldsymbol{z} := \boldsymbol{z}_i = \boldsymbol{z}_j$. Define $T_i := \boldsymbol{z}_i^2 - \mathbb{E}[\boldsymbol{z}_i^2] = (1-\boldsymbol{r}_i/\rho)^2\boldsymbol{y}_i^2 - (1-\rho)^2/\rho\cdot\boldsymbol{y}_i^2$. It is easy to verify that $\mathbb{E}[T_i]=0$, $\mathbb{E}[T_i^2]\lesssim\boldsymbol{y}_i^4/\rho^3\le\mu^2d^2/\rho^3n^2$ and $|T_i|\lesssim\boldsymbol{y}_i^2/\rho^2\le\mu d/\rho^2n$. Subsequently, with probability $1-O(N^{-10})$ we have

$$\left|\sum_{i=1}^n T_i\right| \lesssim \frac{\mu d}{\rho^{3/2}}\sqrt{\frac{\log^2 N}{n}}.$$

The estimation error of $(1-\rho)(\mathbf{X}^\top\mathbf{X})_{ii}$ for $(1-\rho)/\rho\cdot\|\boldsymbol{y}_i\|_2^2 = (1-\rho)/\rho$ can be upper bounded similarly. □

Proof. [Proof of Lemma 6.2.7] Take $\boldsymbol{\Delta}_{jk} = 3\psi_1\sqrt{\frac{\log N}{n}}$ for $j\neq k$ and $\boldsymbol{\Delta}_{jk} = 3\psi_2\sqrt{\frac{\log N}{n}}$. Fix arbitrary $\widetilde{\boldsymbol{x}}_j\notin\mathcal{S}_\ell$ and $\widetilde{\boldsymbol{x}}_i\in\mathcal{S}_\ell$. Because $\widetilde{\boldsymbol{\nu}}_i = \lambda(\widetilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}_{-i}^{(\ell)}\widetilde{\boldsymbol{c}}_i)$, we have that

$$\begin{aligned}
\left|\langle\widetilde{\boldsymbol{x}}_j,\widetilde{\boldsymbol{\nu}}_i\rangle\right| = \lambda\left|\widetilde{\boldsymbol{x}}_j^\top(\widetilde{\boldsymbol{x}}_i + \widetilde{\mathbf{X}}_{-i}^{(\ell)}\widetilde{\boldsymbol{c}}_i)\right| &\le \lambda(1+\|\widetilde{\boldsymbol{c}}_i\|_1)\cdot\sup_{\widetilde{\boldsymbol{x}}_i\in\mathcal{S}_\ell}\left|\langle\widetilde{\boldsymbol{x}}_j,\widetilde{\boldsymbol{x}}_i\rangle\right| \\
&\le \lambda(1+\|\widetilde{\boldsymbol{c}}_i\|_1)\cdot\left(\widetilde{\mu} + \sup_{\widetilde{\boldsymbol{x}}_i\notin\mathcal{S}_\ell}\left|\langle\widetilde{\boldsymbol{x}}_j,\widetilde{\boldsymbol{x}}_i\rangle - \langle\boldsymbol{y}_j,\boldsymbol{y}_i\rangle\right|\right) \\
&\lesssim \lambda(1+\|\widetilde{\boldsymbol{c}}_i\|_1)\cdot\left(\widetilde{\mu} + \psi_1\sqrt{\frac{\log N}{n}}\right), \qquad (6.8)
\end{aligned}$$

where the last inequality holds by applying Definition 6.2.1 and the fact that

$$\begin{aligned}
\left|\langle\widetilde{\boldsymbol{x}}_i,\widetilde{\boldsymbol{x}}_j\rangle - \langle\widetilde{\boldsymbol{y}}_i,\widetilde{\boldsymbol{y}}_j\rangle\right| &\le \left|(\widetilde{\boldsymbol{\Sigma}}_+)_{ij} - (\widetilde{\boldsymbol{\Sigma}})_{ij}\right| + \left|(\widetilde{\boldsymbol{\Sigma}})_{ij} - \langle\widetilde{\boldsymbol{y}}_i,\widetilde{\boldsymbol{y}}_j\rangle\right| \\
&\le \left|\boldsymbol{\Delta}_{ij}\right| + \left|\langle\widetilde{\boldsymbol{x}}_i,\widetilde{\boldsymbol{x}}_j\rangle - \langle\widetilde{\boldsymbol{y}}_i,\widetilde{\boldsymbol{y}}_j\rangle\right| \\
&\le \left|\boldsymbol{\Delta}_{ij}\right| + \left|\langle\widetilde{\boldsymbol{z}}_i,\widetilde{\boldsymbol{y}}_j\rangle\right| + \left|\langle\widetilde{\boldsymbol{y}}_j,\widetilde{\boldsymbol{z}}_i\rangle\right| + \left|\langle\widetilde{\boldsymbol{z}}_j,\widetilde{\boldsymbol{z}}_i\rangle\right|
\end{aligned}$$

$$\lesssim \psi_1\sqrt{\frac{\log N}{n}} \quad \text{for } \; i \neq j.$$

To bound $\|\widetilde{\boldsymbol{c}}_i\|_1$, consider an auxiliary noiseless problem:

$$\widehat{\boldsymbol{c}}_i := \arg\min_{\boldsymbol{c}_i} \|\boldsymbol{c}_i\|_1 \quad s.t. \;\; \boldsymbol{y}_i = \mathbf{Y}_{-i}^{(\ell)}\boldsymbol{c}_i. \tag{6.9}$$

Note that when $r > 0$ Eq. (6.9) is always feasible. Following standard analysis (e.g., Lemma 15 and Eq. (5.15) of [WX16]), it can be established that $\|\widehat{\boldsymbol{c}}_i\|_1 \leq 1/r_i \leq 1/r$. On the other hand, by optimality we have $\|\widetilde{\boldsymbol{c}}_i\|_1 + \frac{\lambda}{2}\|\widetilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}_{-i}^{(\ell)}\widetilde{\boldsymbol{c}}_i\|_2^2 \leq \|\widehat{\boldsymbol{c}}_i\|_1 + \frac{\lambda}{2}\|\widetilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}_{-i}^{(\ell)}\widehat{\boldsymbol{c}}_i\|_2^2$. Therefore,

$$\begin{aligned}
\|\widetilde{\boldsymbol{c}}_i\|_1 &\leq \|\widehat{\boldsymbol{c}}_i\|_1 + \frac{\lambda}{2}\left\|\widetilde{\boldsymbol{x}}_i - \widetilde{\mathbf{X}}_{-i}^{(\ell)}\widehat{\boldsymbol{c}}_i\right\|_2^2 \\
&\lesssim \|\widehat{\boldsymbol{c}}_i\|_1 + \frac{\lambda}{2}\left\|\boldsymbol{y}_i - \mathbf{Y}_{-i}^{(\ell)}\widehat{\boldsymbol{c}}_i\right\|_2^2 + (1+\|\widehat{\boldsymbol{c}}_i\|_1)^2 \cdot \frac{\lambda}{2}\sup_{\boldsymbol{y}_i,\boldsymbol{y}_j\in\mathcal{S}_\ell}\left|\langle\widetilde{\boldsymbol{x}}_i,\widetilde{\boldsymbol{x}}_j\rangle - \langle\boldsymbol{y}_i,\boldsymbol{y}_j\rangle\right| \\
&= \|\widehat{\boldsymbol{c}}_i\|_1 + (1+\|\widehat{\boldsymbol{c}}_i\|_1)^2 \cdot \frac{\lambda}{2}\sup_{\boldsymbol{y}_i,\boldsymbol{y}_j\in\mathcal{S}_\ell}\left|\langle\widetilde{\boldsymbol{x}}_i,\widetilde{\boldsymbol{x}}_j\rangle - \langle\boldsymbol{y}_i,\boldsymbol{y}_j\rangle\right| \\
&\lesssim \|\widehat{\boldsymbol{c}}_i\|_1 + (1+\|\widehat{\boldsymbol{c}}_i\|_1)^2 \cdot (\psi_1+\psi_2)\sqrt{\frac{\log N}{n}} \\
&\lesssim \frac{1}{r}\left(1 + \frac{\lambda}{r}(\psi_1+\psi_2)\sqrt{\frac{\log N}{n}}\right).
\end{aligned} \tag{6.10}$$

□

Proof. [Proof of Theorem 6.2.9] Following the analysis of Lasso SSC solution path in [WX16], it suffices to show that $\lambda > 1/\|\widetilde{\boldsymbol{x}}_i^\top\widetilde{\mathbf{X}}_{-i}\|_\infty$. On the other hand, note that $\|\boldsymbol{y}_i^\top\mathbf{Y}_{-i}\|_\infty \geq \|\boldsymbol{y}_i^\top\mathbf{Y}_{-i}^{(\ell)}\|_\infty \geq r_i \geq r$ (see, for example, Eq. (5.19) of [WX16]). Subsequently,

$$\left\|\widetilde{\boldsymbol{x}}_i^\top\widetilde{\mathbf{X}}_{-i}\right\|_\infty \geq \left\|\boldsymbol{y}_i^\top\mathbf{Y}_{-i}\right\|_\infty - \sup_{j\neq i}\left|\langle\widetilde{\boldsymbol{x}}_i,\widetilde{\boldsymbol{x}}_j\rangle - \langle\boldsymbol{y}_i,\boldsymbol{y}_j\rangle\right| \gtrsim r - \psi_1\sqrt{\frac{\log N}{n}}.$$

□

Proof. [Proof of Lemma 6.2.10] We first prove

$$\max_{\boldsymbol{y}_i\in\mathcal{S}_k}\max_{\boldsymbol{y}_j\in\mathcal{S}_\ell}\left|\langle\boldsymbol{y}_i,\boldsymbol{y}_j\rangle\right| \lesssim \chi_{k\ell}\cdot\frac{\log(\overline{C}N/\underline{C})}{\sqrt{d_k d_\ell}} \qquad \forall j \neq k \in [L]. \tag{6.11}$$

Let N_k and N_ℓ be the total number of data points in $\mathcal{S}_k$ and $\mathcal{S}_\ell$, and let P_k and P_ℓ be the corresponding densities which are bounded from both above and below by $\overline{C}p_0$ and $\underline{C}p_0$. Consider a rejection sampling procedure: first sample $\boldsymbol{\alpha}$ randomly from the uniform measure over $\{\boldsymbol{\alpha} \in \mathbb{R}^{d_k} : \|\boldsymbol{\alpha}\|_2 = 1\}$, and then reject the sample if $u > p_k(\boldsymbol{\alpha})/\overline{C}p_0$, where $u \sim U(0,1)$. Repeat the procedure until N_k samples are obtained. This procedure is sound because $p_k/p_0 \leq \overline{C}$, and the resulting (accepted) samples are i.i.d. distributed according to P_k. On the other hand, for any $\boldsymbol{\alpha}$ the probability of acceptance is lower bounded by $\underline{C}/\overline{C}$. Therefore, the procedure terminates by producing a total of $O(\overline{C}N_k/\underline{C})$ samples (both accepted and rejected). Thus, without loss of generality we can assume both P_k and P_ℓ are uniform measures on the corresponding spheres, by paying the cost of adding $\widetilde{N}_k = O(\overline{C}N_k/\underline{C})$ and $\widetilde{N}_\ell = O(\overline{C}N_\ell/\underline{C})$ points to each subspace.

Now fix $\boldsymbol{y}_i = \mathbf{U}_k\boldsymbol{\alpha}_i$ and $\boldsymbol{y}_j = \mathbf{U}_\ell\boldsymbol{\alpha}_j$, where $\boldsymbol{\alpha}_i \in \mathbb{R}^{d_k}$, $\boldsymbol{\alpha}_j \in \mathbb{R}^{d_\ell}$ and $\|\boldsymbol{\alpha}_i\|_2 = \|\boldsymbol{\alpha}_j\|_2 = 1$. Then both $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ are uniformly distributed on the low-dimensional spheres, and that $|\langle \boldsymbol{y}_i, \boldsymbol{y}_j\rangle| = |\boldsymbol{\alpha}_i^\top(\mathbf{U}_k^\top\mathbf{U}_\ell)\boldsymbol{\alpha}_j|$. Applying Lemma 7.5 of [SC12] and note that $\chi_{k\ell} = \|\mathbf{U}_k^\top\mathbf{U}_\ell\|_F$ we complete the proof of Eq. (6.11).

We next prove

$$r_i \gtrsim \sqrt{\frac{\log(\underline{C}N_\ell/\overline{C}d_\ell)}{d_\ell}} \qquad \forall i \in [N], \ell \in [L], \boldsymbol{x}_i \in \mathcal{S}_\ell. \tag{6.12}$$

Let P_ℓ be the underlying measure of subspace $\mathcal{S}_\ell$. Consider the decomposition $P_\ell = \underline{C}/\overline{C}\cdot P_0 + (1-\underline{C}/\overline{C})\cdot P_\ell'$, where P_0 is the uniform measure. Such a decomposition and the corresponding density P_ℓ' exists because $\underline{C}P_0 \leq P_\ell \leq \overline{C}P_0$. This shows that the distribution of points in subspace $\mathcal{S}_\ell$ can be expressed as a mixture distribution, with a uniform density mixture with weight probability $\underline{C}/\overline{C}$. Because r_i decreases with smaller data set, it suffices to consider only the uniform mixture. Thus, we can assume P_ℓ is the uniform measure at the cost of considering only $\widetilde{N}_\ell = \Omega(\underline{C}N_\ell/\overline{C})$

points in subspace $\mathcal{S}_\ell$. Applying Lemma 21 of [WX16] and replacing N_ℓ with $\widetilde{N}_\ell$ we complete the proof of Eq. (6.12).

Finally Lemma 6.2.10 is an easy corollary of Eqs. (6.11) and (6.12). □

Proof. [Proof of Lemma 6.2.11] Fix $k,\ell \in [L]$ and let $\mathbf{U}_k = (\boldsymbol{u}_{k1},\cdots,\boldsymbol{u}_{kd})$, $\mathbf{U}_\ell = (\boldsymbol{u}_{\ell 1},\cdots,\boldsymbol{u}_{\ell d})$ be orthonormal basis of $\mathcal{U}_k$ and $\mathcal{U}_\ell$. Then $\chi_{k\ell} = \|\mathbf{U}_k^\top \mathbf{U}_\ell\|_F \leq d\|\mathbf{U}_k^\top \mathbf{U}_\ell\|_{\max} = d\cdot\sup_{1\leq i,j\leq d}|\langle \boldsymbol{u}_{ki}, \boldsymbol{u}_{\ell j}\rangle|$. Because $\mathcal{U}_k$ and $\mathcal{U}_\ell$ are random subspaces, $\boldsymbol{u}_{ki}$ and $\boldsymbol{u}_{\ell j}$ are independent vectors distributed uniformly on the d-dimensional unit sphere. Applying Lemma 17 of [WX16] and a union bound over all i,j,k,ℓ we prove the upper bound on χ. For the upper bound on μ, simply note that $\|\boldsymbol{u}_{jk}\|_\infty \lesssim \sqrt{\frac{\log N}{n}}$ with probability $1-O(N^{-10})$ by standard concentration result for Gaussian suprema. □

CHAPTER 7

ONLINE DISCOVERY FOR STABLE AND GROUPING CAUSALITIES IN MULTI-VARIATE TIME SERIES

The content of this chapter is organized as follows. The problem formulation is presented in Section 7.2. Section 7.3 introduces the details about our proposed approach and its equivalent Bayesian model. A solution capable of online inference with particle learning is given in Section 7.4. Extensive empirical evaluation are demonstrated in Section 7.5. Finally, we concludes our work and discusses the future work.

7.1 Related work

It is an important task to reveal the casual dependencies between historical and current observations in MTS analysis. Bayesian Network [JYG$^+$03, Mur02] and Granger Causality [ALA07, ZF09] are two main frameworks for inference of temporal dependency. Comparing with Bayesian Network, Granger Causality is more straightforward, robust and extendable [ZF09].

Originally, Granger Causality is designed for a pair of time series. The appearance of pioneering work of combining the notion of Granger Causality with graphical model [Eic06] leads to the emergence of causal relationship analysis among MTS data. Two typical techniques, statistical significance test and Lasso-Granger [ALA07], are developed to inference the Granger Causality among MTS. Lasso-Granger gains more popularity due to its robust performance even in high dimensions [BL12]. However, Lasso-Granger suffers from instability and failure of group variable selection because of the high sensitivity of L_1 norm. To address this challenging, our method adopts Elastic-Net regularizer [ZH05] which is stable since it encourages a group variable selection (group effect) where strongly correlated predictors tend to be zero or non-zero simultaneously.

Particle learning [CJLP10] is a powerful tool to provide an online inference strategy for Bayesian models. It belongs to the Sequential Monte Carlo (SMC) methods consisting of a set of Monte Carlo methodologies to solve the filtering problem [DGA00]. Particle learning provides state filtering, sequential parameter learning and smoothing in a general class of state space models [CJLP10]. The central idea behind particle learning is the creation of a particle algorithm that directly samples from the particle approximation to the joint posterior distribution of states and conditional sufficient statistics for fixed parameters in a fully-adapted resample-propagate framework.

7.2 Problem Formulation

In this section, we formally define the Granger Causality by the VAR model. Given a set of time series $\mathbf{Y}$ defined on $\mathbb{R}^n$ in the time interval $[0, T]$, that is,

$$\mathbf{Y} = \{\mathbf{y}_t : \ \mathbf{y}_t \in \mathbb{R}^n, \ t \in [0, T]\},$$

where $\mathbf{y}_t = (y_{t,1}, y_{t,2}, \ldots, y_{t,n})^T$. The inference of Granger Causality is usually achieved by fitting the time series data $\mathbf{Y}$ with a VAR model. Given the maximum time lag L, the VAR model is expressed as follows:

$$\mathbf{y}_t = \sum_{l=1}^{L} W_l^T \mathbf{y}_{t-l} + \boldsymbol{\epsilon}, \tag{7.1}$$

where $\boldsymbol{\epsilon}$ is the standard Gaussian noise and the vector-value $\mathbf{y}_t$ $(1 \leq t \leq T)$ only depends on the past vector-value $\mathbf{y}_{t-l}$ $(1 \leq l \leq L)$. The Granger causal relationship between $\mathbf{y}_t$ and $\mathbf{y}_{t-l}$ is formulated as the matrix W_l in the following:

$$W_l = (w_{l,ji})_{n \times n}$$

where the entry $w_{l,ji}$ express how large the component $y_{t-l,i}$ influence the component $y_{t,j}$, noted as $y_{t-l,i} \rightarrow_g y_{t,j}$.

To induce sparsity in the matrix W_l $(l = 1, 2, \ldots, L)$, the prior work [ZWW+16] proposed a VAR-Lasso model as follows:

$$\min_{W_l} \sum_{t=L+1}^{T} \left\| \mathbf{y}_t - \sum_{l=1}^{L} W_l^T \mathbf{y}_{t-l} \right\|_2^2 + \lambda_1 \sum_{l=1}^{L} \|W_l\|_1, \tag{7.2}$$

and an online time-varying method based on Bayesian update. However, it suffers instability and fails to select a group of variables that are highly correlated. To address these problems, we propose a method with Elastic-Net regularization and equivalent online inference strategy is given in following sections.

7.3 Elastic-net Regularizer

In this section, we describe the VAR-Elastic-Net model and its equivalent form in the perspective of Bayesian modeling.

7.3.1 Basic Optimization Model

Elastic-Net regularization [ZH05] is a combination of L_1 and L_2 norm and has the following objective function for MTS data:

$$\sum_{t=L+1}^{T} \left\| \mathbf{y}_t - \sum_{l=1}^{L} W_l^T \mathbf{y}_{t-l} \right\|_2^2 + \lambda_1 \sum_{l=1}^{L} \|W_l\|_1 + \lambda_2 \sum_{l=1}^{L} \|W_l\|_2^2, \tag{7.3}$$

where $\|\cdot\|_1$ is the entrywise norm and $\|\cdot\|_2$ is the Frobenius norm (or Hilbert-Schmidt norm).

In order to change Equation 7.3 into the standard form of the linear regression model, we define $\boldsymbol{\beta}$, a $nL \times n$ matrix, as follows:

$$\boldsymbol{\beta} = (W_1^T, W_2^T, \ldots, W_L^T)^T, \tag{7.4}$$

and $\mathbf{x}_t$ be a nL column vector as

$$\mathbf{x}_t = [\mathbf{y}_{t-1}^T, \mathbf{y}_{t-2}^T, \ldots, \mathbf{y}_{t-L}^T]^T. \tag{7.5}$$

Then, Equation 7.3 can be reformulate as

$$\sum_{t=L+1}^{T} \left(\mathbf{y}_t - \boldsymbol{\beta}^T \mathbf{x}_t\right)^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2. \tag{7.6}$$

The coefficient matrix $\boldsymbol{\beta}$ can be expressed as

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_n^T)^T, \tag{7.7}$$

where $\boldsymbol{\beta}_i$ $(i = 1, 2, \ldots, n)$ is a row vector of size nL. Based on Equation 7.7, the equivalent form of the Equation 7.6 is

$$\sum_{t=L+1}^{T} \left(y_{t,i} - \boldsymbol{\beta}_i^T \mathbf{x}_t\right)^2 + \lambda_1 \|\boldsymbol{\beta}_i\|_1 + \lambda_2 \|\boldsymbol{\beta}_i\|_2^2, \tag{7.8}$$

where $i = 1, 2, \ldots, n$. Thus, the original optimization problem of Equation 7.3 can be addressed as the optimization problem of n independent standard linear regression problem of Equation 7.8.

7.3.2 The Corresponding Bayesian Model

From Bayesian perspective, $y_{t,i}$ $(i = 1, 2, \ldots, n)$ follows a Gaussian distribution, given the coefficient vector $\boldsymbol{\beta}_i$ and the variance of random observation noise σ_i^2, as follows:

$$y_{t,i} | \boldsymbol{\beta}_i, \sigma_i^2 \sim \mathcal{N}\left(\boldsymbol{\beta}_i^T \mathbf{x}_t, \sigma_i^2\right). \tag{7.9}$$

The coefficient vector $\boldsymbol{\beta}_i$ is viewed as a random variable which follows the mixed Gaussian and Laplace distribution, as below [LL+10, Mur12]:

$$p(\boldsymbol{\beta}_i | \sigma_i^2) \propto \exp\left(-\lambda_1 \sigma_i^{-1} \sum_{j=1}^{nL} |\boldsymbol{\beta}_{ij}| - \lambda_2 \sigma_i^{-2} \sum_{j=1}^{nL} |\boldsymbol{\beta}_{ij}^2|\right). \tag{7.10}$$

Equation 7.10 represents a scale mixture of normal distributions and exponential distributions and equals a hierarchical form as below:

$$\begin{aligned}
&\tau_j^2|\lambda_1 \sim \sqrt{\exp\left(\lambda_1^2\right)}, \\
&\boldsymbol{\beta}_i|\sigma_i^2, \tau_1^2, \ldots, \tau_{nL}^2 \sim \mathcal{N}\left(0, \sigma_i^2 \mathbf{M}_{\boldsymbol{\beta}_i}\right), \\
&\mathbf{M}_{\boldsymbol{\beta}_i} = diag\left(\left(\lambda_2 + \tau_1^{-2}\right)^{-1}, \ldots, \left(\lambda_2 + \tau_{nL}^{-2}\right)^{-1}\right).
\end{aligned} \tag{7.11}$$

The variance σ_i^2 is a random variable following inverse Gamma distribution [Mur12] as follows:

$$\sigma_i^2 \sim \mathcal{IG}(\alpha_1, \alpha_2), \tag{7.12}$$

where α_1 and α_2 are hyper parameters.

Equation 7.10 can be obtained from integrating out the hyper parameters α_1 and α_2 in Equation 7.12 and it reduces to the regular Lasso when $\lambda_2 = 0$.

7.3.3 Time-Varying Causal Relationship Model

The aforementioned model is the traditional static regression model, based on the assumption that the coefficient $\boldsymbol{\beta}_i (i = 1, 2, \ldots, n)$ is unknown but fixed, which rarely holds in practice. To model dynamic dependencies, it is reasonable to view the coefficient vector $\boldsymbol{\beta}_{t,i} (i = 1, 2, \ldots, n)$ as a function of time t. Specifically, we propose a method modeling the coefficient vector as two parts including the stationary part and the drift part. The latter is to account for tracking the time-varying temporal dependency among the time series instantly.

Let the operation $\circ$ be the Hadamard product (entrywise-product). The form of the dynamic coefficient vector $\boldsymbol{\beta}_{t,i}$ $(i = 1, 2, \ldots, n)$ is constructed as

$$\boldsymbol{\beta}_{t,i} = \boldsymbol{\beta}_{t,i,1} + \boldsymbol{\beta}_{t,i,2} \circ \boldsymbol{\eta}_{t,i}, \tag{7.13}$$

where both $\boldsymbol{\beta}_{t,i,1}$ and $\boldsymbol{\beta}_{t,i,2}$ are stationary part and $\boldsymbol{\eta}_{t,i}$ is the drift part. The drift part at time t is caused by the standard Gaussian random walk from the information at time $t-1$, i.e., $\boldsymbol{\eta}_{t,i} = \boldsymbol{\eta}_{t-1,i} + \mathbf{v}$ and $\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_{nL})$. Thus $\boldsymbol{\eta}_{t,i}$ follows the Gaussian distribution

$$\boldsymbol{\eta}_{t,i} \sim \mathcal{N}(\boldsymbol{\eta}_{t-1,i}, \mathbf{I}_{nL}). \tag{7.14}$$

Combined with Equation 7.13, the equivalent time-varying Bayesian Elastic-Net model in Equation 7.8 becomes

$$\begin{aligned} &\sum_{t=L+1}^{T} \left(y_{t,i} - \boldsymbol{\beta}_{t,i}^T \mathbf{x}_t\right)^2 + \lambda_{1,1} \|\boldsymbol{\beta}_{t,i,1}\|_1 \\ &+ \lambda_{2,1} \|\boldsymbol{\beta}_{t,i,1}\|_2^2 + \lambda_{1,2} \|\boldsymbol{\beta}_{t,i,2}\|_1 + \lambda_{2,2} \|\boldsymbol{\beta}_{t,i,2}\|_2^2. \end{aligned} \tag{7.15}$$

Furthermore, the priors of the equivalent Bayesian model is given as below:

$$\begin{aligned} &\boldsymbol{\beta}_{i,1} | \sigma_i^2, \tau_{1,1}^2, \ldots, \tau_{1,nL}^2 \sim \mathcal{N}\left(0, \sigma_i^2 \mathbf{M}_{\boldsymbol{\beta}_{i,1}}\right), \\ &\boldsymbol{\beta}_{i,2} | \sigma_i^2, \tau_{2,1}^2, \ldots, \tau_{2,nL}^2 \sim \mathcal{N}\left(0, \sigma_i^2 \mathbf{M}_{\boldsymbol{\beta}_{i,2}}\right), \\ &\tau_{1,j}^2 | \lambda_{1,1} \sim \sqrt{\exp\left(\lambda_{1,1}^2\right)}, \\ &\tau_{2,j}^2 | \lambda_{1,2} \sim \sqrt{\exp\left(\lambda_{1,2}^2\right)}, \\ &\sigma_i^2 \sim \mathcal{IG}(\alpha_1, \alpha_2), \\ &\mathbf{M}_{\boldsymbol{\beta}_{i,1}} = diag\left((\lambda_{2,1} + \tau_{1,1}^{-2})^{-1}, \ldots, (\lambda_{2,1} + \tau_{1,nL}^{-2})^{-1}\right), \\ &\mathbf{M}_{\boldsymbol{\beta}_{i,2}} = diag\left((\lambda_{2,2} + \tau_{2,1}^{-2})^{-1}, \ldots, (\lambda_{2,2} + \tau_{2,nL}^{-2})^{-1}\right). \end{aligned} \tag{7.16}$$

It is difficult to solve straightforward the above regression model by traditional optimization method. We develop our solution to infer VAR-Elastic-Net model from a Bayesian perspective utilizing particle learning, which is presented in the following section.

7.4 Online Inference

In this section, we describe the online inference process to update the parameters from time $t-1$ to time t based on particle learning. At last, we give the pseudocode of algorithm to summarize the whole process.

Our goal is to infer both latent parameters and state variables in our Bayesian model. However, since the inference partially depends on the random walk which generates the latent state variables, we use particle learning strategy [CJLP10] to learn the distribution of both parameters and state variables. The definition of a particle is given as below.

Definition 7.4.1 (Particle). A particle used to predict $y_{t,i}$ $(i = 1, 2, \ldots, n)$ is a container which maintains the current status information for value prediction. The status information comprises of random variables, their distributions with corresponding hyperparameters.

Assume the number of particles is B. Let $\mathcal{P}_{t,i}^{(k)}$ be the k^{th} particle for predicting the value y_i at time t with particle weight $\rho_{t,i}^{(k)}$.

We define a new variable $\boldsymbol{\beta}_{t,i}^{\prime(k)} = \left(\boldsymbol{\beta}_{t,i,1}^{(k),T}, \boldsymbol{\beta}_{t,i,2}^{(k),T}\right)^T$ for concisely expressing the stationary parts given in Equation 7.13. At time $t-1$, the information of particle $\mathcal{P}_{t-1,i}^{(k)}$ includes the following variables and hyperparameters for corresponding distributions:

$$\begin{aligned}
&\boldsymbol{\beta}_{t-1,i}^{\prime(k)} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}_{t-1,i}^{\prime(k)}}, \sigma_i^2 \mathbf{M}_{\boldsymbol{\beta}_{t-1,i}^{\prime(k)}}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{t-1,i}^{\prime(k)}} \mathbf{M}_{\boldsymbol{\beta}_{t-1,i}^{\prime(k)}}^{\frac{1}{2}}\right), \\
&\boldsymbol{\eta}_{t-1,i}^{(k)} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\eta}_{t-1,i}^{(k)}}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{t-1,i}^{(k)}}\right), \\
&{\sigma^2}_{t-1,i}^{(k)} \sim \mathcal{IG}\left(\alpha_{t-1,1}^{(k)}, \alpha_{t-1,2}^{(k)}\right).
\end{aligned} \tag{7.17}$$

7.4.1 Particle Learning

The core idea for particle learning is iterated in following steps:

(1) At time $t-1$, there are B particles and each contains information in Equation 7.17. The coefficients at $t - 1$ is given as

$$\boldsymbol{\beta}_{t-1,i}^{(k)} = \boldsymbol{\beta}_{t-1,i,1}^{(k)} + \boldsymbol{\beta}_{t-1,i,2}^{(k)} \circ \boldsymbol{\eta}_{t-1,i}^{(k)}.$$

(2) At time t, sample the drift part $\boldsymbol{\eta}_{t,i}^{(k)}$ from Equation 7.14, and update parameters of all priors and sample the new values for $\boldsymbol{\beta}_{t,i,1}^{(k)}$, $\boldsymbol{\beta}_{t,i,2}^{(k)}$ (details is given in Section 7.4.2) for each particle.

(3) Finally, gain new feedback $y_{t,i}$ and resample B particles based on the recalcuated particle weights (details is given in Section 7.4.2). The value of $\boldsymbol{\beta}_{t,i}$ for prediction at time t is averaged as bellow:

$$\boldsymbol{\beta}_{t,i} = \frac{1}{B}\sum_{k=1}^{B}\left(\boldsymbol{\beta}_{t,i,1}^{(k)} + \boldsymbol{\beta}_{t,i,2}^{(k)} \circ \boldsymbol{\eta}_{t,i}^{(k)}\right). \tag{7.18}$$

7.4.2 Update Process

In the process of particle learning, the key step is to update all the parameters from time $t-1$ to time t and recalculate particle weights mentioned above. In this section, we describe the update process of particle weights and all the parameters in details.

7.4.2.1 Particle Weights Update

Each particle $\mathcal{P}_{t,i}^{(k)}$ has a weight, denoted as $\rho_{t,i}^{(k)}$, indicating its fitness for the new observed data at time t. Note that $\sum_{k=1}^{B}\rho_{t,i}^{(k)} = 1$. The fitness of each particle $\mathcal{P}_{t,i}^{(k)}$ is defined as likelihood of the observed data $\mathbf{x}_t$ and $y_{t,i}$. Therefore,

$$\rho_{t,i}^{(k)} \propto P(\mathbf{x}_t, y_{t,i}|\mathcal{P}_{t-1,i}^{(k)}).$$

Combined with Equation 7.14 for $\boldsymbol{\eta}_{t,i}^{(k)}$ and Equation 7.17 for $\boldsymbol{\eta}_{t-1,i}^{(k)}$, the particle weights $\rho_i^{(k)}$ at time t is proportional to the value as follows:

$$\begin{aligned}\rho_{t,i}^{(k)} \propto \iint & \mathcal{N}(y_{t,i}|\boldsymbol{\beta}_{t,i}^{(k),T}\mathbf{x}_t, \sigma_{t-1,i}^{2(k)})\mathcal{N}(\boldsymbol{\eta}_{t,i}^{(k)}|\boldsymbol{\eta}_{t-1,i}^{(k)}, I_{nL}) \\ & \mathcal{N}(\boldsymbol{\eta}_{t-1,i}^{(k)}|\boldsymbol{\mu}_{\boldsymbol{\eta}_{t-1,i}^{(k)}}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{t-1,i}^{(k)}})d\boldsymbol{\eta}_{t-1,i}^{(k)}d\boldsymbol{\eta}_{t,i}^{(k)}.\end{aligned} \tag{7.19}$$

Integrating out the variables of $\boldsymbol{\eta}_{t-1,i}^{(k)}$ and $\boldsymbol{\eta}_{t,i}^{(k)}$, we can obtain that the particle weights $\rho_i^{(k)}$ at time t follows Gaussian distribution as follows:

$$\rho_{t,i}^{(k)} \propto \mathcal{N}(y_{t,i}|m_{t,i}^{(k)}, Q_{t,i}^{(k)}), \tag{7.20}$$

where the mean value and the variance are respectively,

$$\begin{aligned} m_{t,i}^{(k)} &= (\boldsymbol{\beta}_{t,i,1}^{(k)} + \boldsymbol{\beta}_{t,i,2}^{(k)} \circ \boldsymbol{\mu}_{\boldsymbol{\eta}_{t-1,i}^{(k)}})^T \mathbf{x}_t, \\ Q_{t,i}^{(k)} &= \sigma_{t-1,i}^{2(k)} + (\mathbf{x}_t \circ \boldsymbol{\beta}_{t,i,2}^{(k)})^T \\ &\qquad (\mathbf{I}_{nL} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{t-1,i}^{(k)}})(\mathbf{x}_t \circ \boldsymbol{\beta}_{t,i,2}^{(k)}). \end{aligned} \tag{7.21}$$

Furthermore, the final k^{th} particle weights at time t can be obtained from normalization, as follows:

$$\rho_{t,i}^{(k)} = \frac{\mathcal{N}(y_{t,i}|m_{t,i}^{(k)}, Q_{t,i}^{(k)})}{\sum\limits_{k=1}^{B} \mathcal{N}(y_{t,i}|m_{t,i}^{(k)}, Q_{t,i}^{(k)})}. \tag{7.22}$$

With the particle weights $\rho_{t,i}^{(k)}$ $(k = 1, 2, \ldots, B)$ at time t obtained, the B particles are resampled at time t.

7.4.2.2 Latent State Update

Having the new observation $\mathbf{x}_t$ and $y_{t,i}$ at time t, both the mean $\boldsymbol{\mu}_{\boldsymbol{\eta}_{t-1,i}^{(k)}}$ and the variance $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{t-1,i}^{(k)}}$ need to update from time $t-1$ to time t. Here, we apply the Kalman filter method [Har90] to recursively update to the mean and the variance at time t as follows:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\eta}_{t,i}^{(k)}} &= \boldsymbol{\mu}_{\boldsymbol{\eta}_{t-1,i}^{(k)}} + \mathbf{G}_{t-1,i}^{(k)} \\ &\qquad \left(y_{t,i} - \left(\boldsymbol{\beta}_{t,i,1}^{(k)} + \boldsymbol{\beta}_{t,i,2}^{(k)} \circ \boldsymbol{\mu}_{\boldsymbol{\eta}_{t-1,i}^{(k)}} \right)^T \mathbf{x}_t \right), \\ \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{t,i}^{(k)}} &= \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{t-1,i}^{(k)}} + I_{nL} - \mathbf{G}_{t,i}^{(k)} Q_{t,i}^{(k)} \mathbf{G}_{t,i}^{(k),T}. \end{aligned} \tag{7.23}$$

where $\mathbf{G}_{t,i}^{(k)}$ is the Kalman gain defined as [Har90]

$$\mathbf{G}_{t,i}^{(k)} = \left(\mathbf{I}_{nL} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{t-1,i}^{(k)}} \right) \left(\mathbf{x}_t \circ \boldsymbol{\beta}_{t-1,i,2}^{\prime(k)} \right) {Q_{t,i}^{(k)}}^{-1}. \tag{7.24}$$

Then, we can sample the drift part at time t from Gaussian distribution as follows:

$$\boldsymbol{\eta}_{t,i}^{(k)} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\eta}_{t,i}^{(k)}}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{t,i}^{(k)}} \right). \tag{7.25}$$

Before updating parameters, a resampling process is conducted. We replace the particle set $\mathcal{P}^{(k)}_{t-1,i}$ with a new set $\mathcal{P}^{(k)}_{t,i}$, where $\mathcal{P}^{(k)}_{t,i}$ is generated from $\mathcal{P}^{(k)}_{t-1,i}$ using sampling with replacement based on the new particle weights.

7.4.2.3 Parameter update

Having sampled the drift part $\boldsymbol{\eta}^{(k)}_{t,i}$, the parameter update for the covariant matrix, mean value and the hyperparameters from time $t-1$ to time t are as follows:

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\beta}'^{(k)}_{t,i}} &= \left(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} + \mathbf{M}^{\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \mathbf{z}^{(k)}_{t,i} \mathbf{z}^{(k)^T}_{t,i} \mathbf{M}^{\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}}\right)^{-1}, \\
\boldsymbol{\mu}_{\boldsymbol{\beta}'^{(k)}_{t,i}} &= \mathbf{M}^{\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \boldsymbol{\Sigma}_{\boldsymbol{\beta}'^{(k)}_{t,i}} \mathbf{M}^{\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \mathbf{z}^{(k)}_{t,i} y_{t,i} \\
&\quad + \mathbf{M}^{\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \boldsymbol{\Sigma}_{\boldsymbol{\beta}'^{(k)}_{t,i}} \boldsymbol{\Sigma}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \mathbf{M}^{\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \boldsymbol{\beta}'^{(k)}_{t-1,i}, \\
\alpha^{(k)}_{t,1} &= \alpha^{(k)}_{t-1,1} + \frac{1}{2}, \\
\alpha^{(k)}_{t,2} &= \alpha^{(k)}_{t-1,2} + \frac{1}{2} y^2_{t,i} \\
&\quad + \frac{1}{2} \boldsymbol{\mu}^T_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \mathbf{M}^{-\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \boldsymbol{\Sigma}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \mathbf{M}^{-\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \boldsymbol{\mu}_{\boldsymbol{\beta}'^{(k)}_{t-1,i}} \\
&\quad - \frac{1}{2} \boldsymbol{\mu}^T_{\boldsymbol{\beta}'^{(k)}_{t,i}} \mathbf{M}^{-\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t,i}} \boldsymbol{\Sigma}_{\boldsymbol{\beta}'^{(k)}_{t,i}} \mathbf{M}^{-\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t,i}} \boldsymbol{\mu}_{\boldsymbol{\beta}'^{(k)}_{t,i}},
\end{aligned} \tag{7.26}$$

where $\mathbf{z}^{(k)}_{t,i} = (\mathbf{x}^T_t, (\boldsymbol{\eta}^{(k)}_{t,i} \circ \mathbf{x}_t)^T)^T$ be a $2n$ column vector. After parameters update in Equation 7.26 at time t, we can sample $\sigma^{2}{}^{(k)}_{t,i}$ and the stationary part of coefficients $\boldsymbol{\beta}'^{(k)}_{t,i}$ as follows:

$$\begin{aligned}
\sigma^{2}{}^{(k)}_{t,i} &\sim \mathcal{IG}(\alpha^{(k)}_{t,1}, \alpha^{(k)}_{t,2}), \\
\boldsymbol{\beta}'^{(k)}_{t,i} &\sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}'^{(k)}_{t,i}}, \sigma^{2}{}^{(k)}_{t,i} \mathbf{M}^{\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t,i}} \boldsymbol{\Sigma}_{\boldsymbol{\beta}'^{(k)}_{t,i}} \mathbf{M}^{\frac{1}{2}}_{\boldsymbol{\beta}'^{(k)}_{t,i}}\right).
\end{aligned} \tag{7.27}$$

7.4.3 Algorithm

Putting all the aforementioned descriptions together, an algorithm for Var-Elastic-Net by Bayesian Update is provided below.

Online inference for time-varying Bayesian Var-Elastic-Net model starts with MAIN procedure, as presented in Algorithm 1. The parameters $B, L, \alpha_1, \alpha_2, \lambda_{11}, \lambda_{12}, \lambda_{21}$ and λ_{22} are given as the input of MAIN procedure. The initialization is executed from line 2 to line 7. As new observation $\mathbf{y}_t$ arrives at time t, $\mathbf{x}_t$ is built using the time lag, then $\boldsymbol{\beta}_t$ is inferred by calling UPDATE procedure. Especially in the UPDATE procedure, we use the resample-propagate strategy in particle learning [CJLP10] rather than the resample-propagate strategy in particle filtering [DKZ+03]. With the resample-propagate strategy, the particles are resampled by taking $\rho_{t,i}^{(k)}$ as the k^{th} particle's weight, where the $\rho_{t,i}^{(k)}$ indicates the occurring probability of the observation at time t given the particle at time $t-1$. The resample-propagate strategy is considered as an optimal and fully adapted strategy, avoiding an importance sampling step.

7.5 Empirical Study

To demonstrate the efficiency of our proposed algorithm, we conduct experiments over both synthetic and real world climate change data set. In this section, we first outline the baseline algorithms for comparison and the evaluation metrics. Second, we present our approach to generate the synthetic data and then illustrate corresponding experimental results in details. Finally, a case study on real world climate change data set is given.

7.5.1 Baseline Algorithms

In our experiments, we demonstrate the performance of our method by comparing with the following baseline algorithms:

- $BL(\gamma)$: VAR by Bayesian prior Gaussian distribution $\mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{I}_d)$.

- $BLasso(\lambda_1)$: VAR-Lasso by Baysian prior Laplacian distribution $\mathcal{L}(\mathbf{0}, \lambda_1\mathbf{I})$.
- $TVLR(\gamma)$: VAR by Bayesian prior Gaussian distribution $\mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{I}_d)$ and online update with both stationary and drift components of the coefficient [ZWW$^+$16].
- $TVLasso(\lambda_1, \lambda_2)$: VAR-Lasso by Baysian prior Laplacian distribution $\mathcal{L}(\mathbf{0}, diag(\lambda_1\mathbf{I}, \lambda_2\mathbf{I}))$ and online update with both stationary and drift components of the coefficient [ZWW$^+$16].

Our proposed algorithm, VAR-Elastic-Net, is denoted as $TVEN(\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22})$. The penalty parameters λ_{ij} $(i = 1, 2; j = 1, 2)$ are presented in Equation 7.15, determining the L_1 and L_2 norm of both stationary component and drift component, respectively. During our experiments, we extract small subset of data with early time stamps and employ grid search to find the optimal parameters for all the algorithm. The parameter settings are verified by cross validation in terms of the prediction errors over the extracted data subset.

7.5.2 Evaluation Metrics

- **AUC Score:** At each time t, the AUC score is obtained by comparing its inferred temporal dependency structure with the ground truth. Nonzero value of $W_{l,ji}$ indicates $y_{t-l,i} \rightarrow_g y_{t,j}$ and the higher absolute value of $W_{l,ji}$ implies a larger likelihood of existing a temporal dependency $y_{t-l,i} \rightarrow_g y_{t,j}$.
- **Prediction Error:** At each time t, the true coefficient matrix is W_t and the estimated one is $\widehat{W}_t$. Hence, the prediction error ϵ_t defined by the Frobenius norm [CDG00] is $\epsilon_t = \|\widehat{W}_t - W_t\|_F$. A smaller prediction error ϵ_t indicates a more accurate inference of the temporal structure.

7.5.3 Synthetic Data and Experiments

In this section, we first present our approach to generate the synthetic data and then illustrate corresponding experimental results.

Table 7.1: **Parameters for Synthetic Data Generation**

Name	Description
K	The number of MTS
T	The total length of MTS with time line
L	The maximum time lag for VAR model
n	The number of different value in the case of piecewise constant
S	The sparsity of the spatial-temporal dependency, denoted as the ratio of zero-value coefficients in dependency matrix W
μ	The mean of the noise
σ^2	The variance of the noise

7.5.3.1 Synthetic Data Generation

By generating synthetic MTS with all types of dependency structures, we are able to comprehensively and systematically evaluate the performance of our proposed method in every scenario. Table 7.1 summarizes the parameters used to generate the Synthetic Data.

The dependency structure is shown by the coefficient matrix $W_{l,ji}$, which have beed constructed by five ways in [ZWW+16], such as **Zero value**, **Constant value**, **Piecewise constant**, **Periodic value** and **Random walk**. To show the efficiency our proposed algorithm, we add a new constructure by **Grouped value**. The variables are categorized into several groups. Each group first appoints a representative variable whose coefficient is sampled at time t. Meanwhile, the coefficients for other variables at the group is assigned with the same value adding a small Gaussian noise, that is, $\epsilon = 0.1\epsilon^*$, $\epsilon^* \sim \mathcal{N}(0, 1)$.

7.5.3.2 Overall Evaluation

We first conduct an overall evaluation in terms of AUC and prediction error over synthetic data generated by setting parameter $S = 0.85$, $T = 5000$, $n = 10$, $L = 2$, $\mu = 0$, $\sigma^2 = 1$ and $K = (30, 40, 50)$. From the experimental results shown in Figure 7.1, we conclude that our proposed method has the best performance in both evaluation metrics AUC and prediction error which indicates the superiority of our algorithm in dependency discovery for time series data.

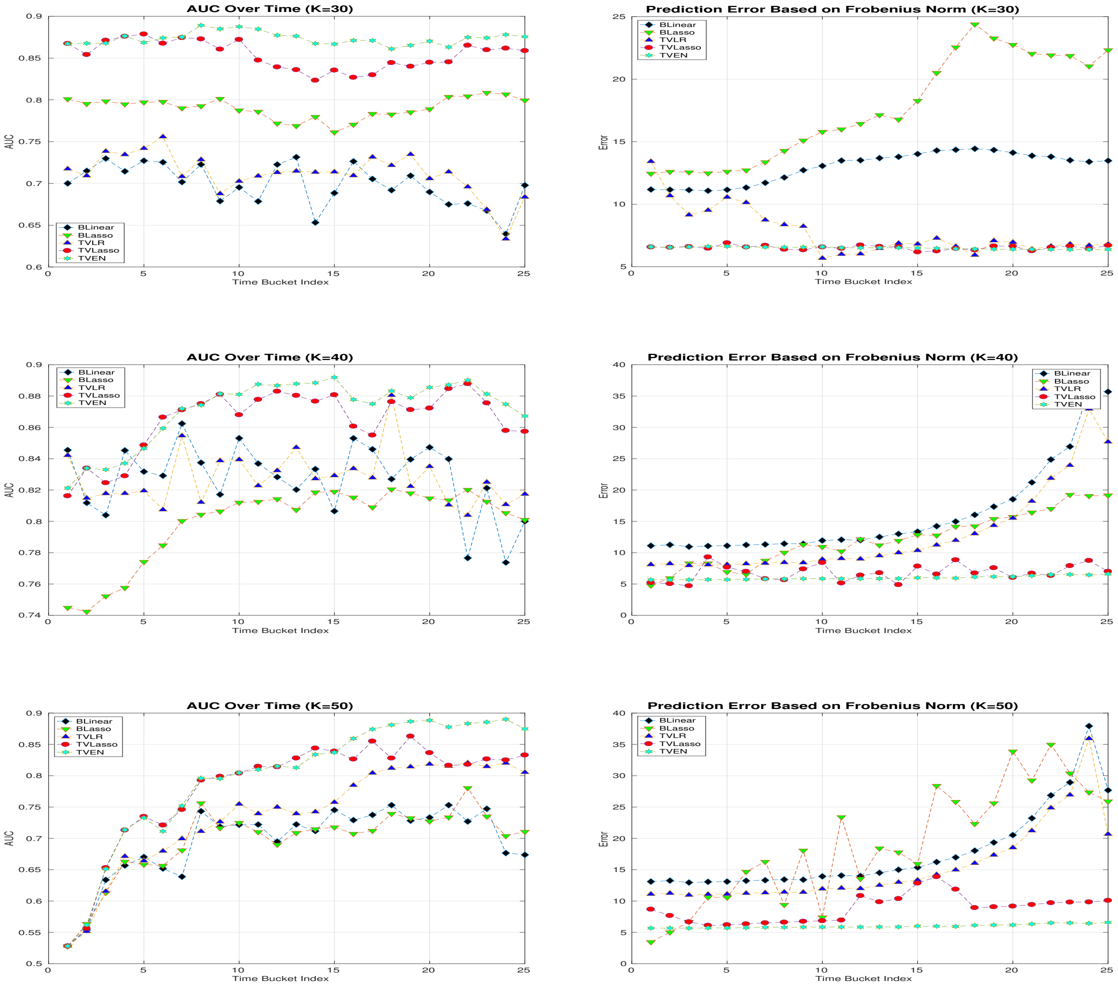


Figure 7.1: The temporal dependency identification performance is evaluated in terms of AUC and prediction error for algorithms `BLR`(1.0) `BLasso`(1k), `TVLR`(1.0), `TVLasso`(1k, 2k), `TVElastic-Net`(1k, 2k, 1m, 2m). The bucket size is 200.

To better show the capability of our algorithm in capturing the dynamic dependencies, we visualize and compare the ground truth coefficients and the estimated ones by different algorithms over synthetic data with all aforementioned dependency structures. The experiments start with simulations where $S = 0.87$, $T = 3000$, $L = 2$, $n = 10$, $\mu = 0$ and $\sigma^2 = 1$. In order to guarantee consistent comparison with the result in work [ZWW+16], we set parameter $K = 20$. The result shown on Figure 7.2 indicates our proposed approach is able to better track the dynamic temporal dependencies in all cases.

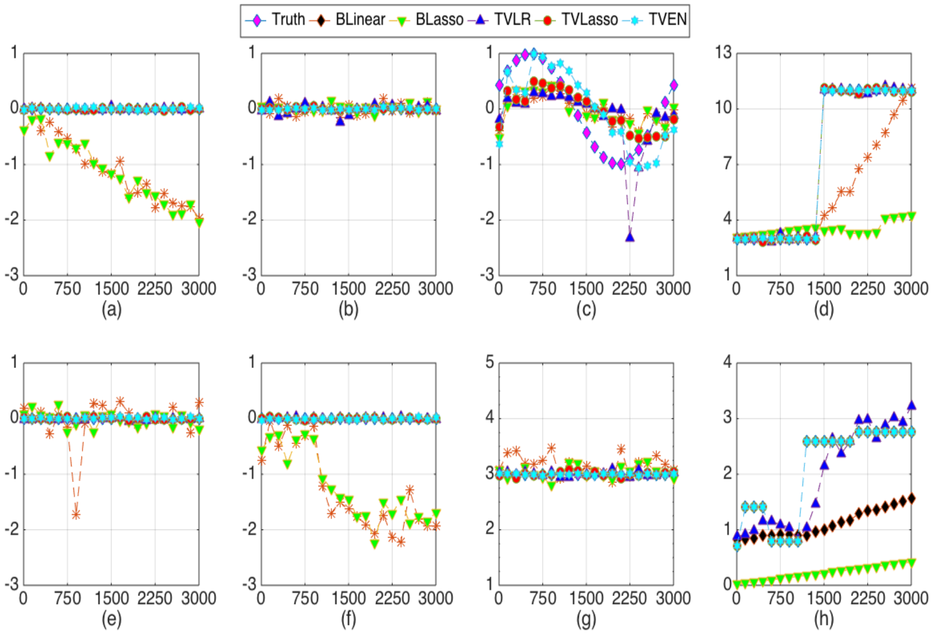


Figure 7.2: The temporal dependencies from 20 time series are learned and eight coefficients among all are selected for demonstration. Coefficients with zero values are displayed in (a), (b), (e) and (f). The coefficients with periodic change, piecewise constant, constant, and random walk are shown in (c), (d), (g) and (h), respectively.

7.5.3.3 Group Effect

To present the ability of our algorithm in better stability and group variable selection, we highlight our experiments on synthetic data with a *Group Value* dependency structure, where $T = 3000$, $n = 10$, $L = 1$, $\mu = 0$, $\sigma^2 = 1$ and $K = 20$. Among the dependency matrix sampled in this experiment, only 6 coefficients are non-zero and we equally categorize them into two groups. When sampling the coefficient values for each group, we first sample a value x and every member in this group is assigned with x adding a small Gaussian noise, that is, $\epsilon = 0.1\epsilon^*$, $\epsilon^* \sim \mathcal{N}(0, 1)$, such that the synthetic data will have group effect.

We make use of the tuning parameter *shrinkage ratio* [ZH05] s defined as following:

$$s = \|\beta\|_1 / \max(\|\beta\|_1),$$

where s is a value in $[0, 1]$. A smaller value s indicates a stronger penalty on the L_1 norm of coefficient β thus a smaller ratio of non-zero coefficient. We also have following definition:

Definition 7.5.1 (**Zero Point**)**.** A zero point for a variable α in our model equals to the value of *shrinkage ratio* s, which makes the coefficient of the variable α happen to change from zero to non-zero or vise versa.

From the definition of *shrinkage ratio* s, (1) a small zero point for variable α indicates a strong correlation between the variable α and the dependent variable and (2) group variables have closer zero points. However, it is static result on [ZH05]. Here, we show the dynamic change of zero point with time.

Figure 7.3 records the zero points for all variables with non-zero coefficients calculated by algorithm TVLasso and TVEN. From the result, it is safe to claim that Lasso regularization alone fail to identity group variables meanwhile our proposed method with Elastic-Net regularization succeeds.

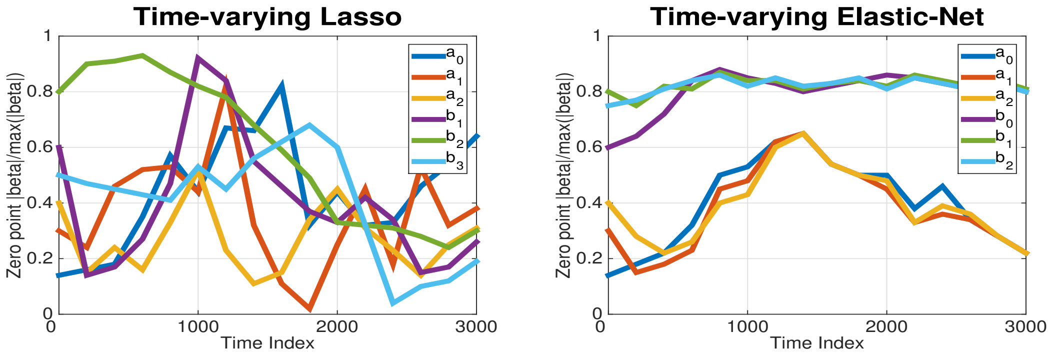


Figure 7.3: The zero point s change with time between TVLasso and TVEN. The penalty parameters are $\lambda_1 = \lambda_2 = 1000$ for TVLasso and $\lambda_{11} = \lambda_{12} = \lambda_{21} = \lambda_{22} = 1000$ for TVEN.

7.5.4 Climate Data and Experiments

In this section, we conduct experiments on real world climate data and display corresponding experimental results and analysis.

7.5.4.1 Data Source and Pre-Processing

The MTS data[1] records monthly global surface air temperature from 1948 to 2017 for each grid. The whole global surface is equally segmented into 360 × 720 grids (0.5 degree latitude × 0.5 degree longitude global grid for each).

In this paper, we only focus on East Florida area in U.S.A and is able to extract totally 46 contiguous land grids with complete monthly air temperature observations from Jan. 1948 to Dec. 2016. Each grid data is considered as one time series $\mathbf{y}$, so the number of multi-variate time series K is 46 and the total length of the time series T is 828. Normalization is applied to the data set for all grids.

[1] https://www.esrl.noaa.gov/psd/data/gridded/data.ghcncams.
-html

7.5.4.2 Spatial-temporal Overall Evaluation

To illustrate the efficacy of our algorithm on the real world climate data, we conduct experiments to inspect the prediction performance of our algorithm in the perspective of both space and time.

Figure 7.4 shows the average predicted value of the normalized air temperatures on the total 46 grids of East Florida, where the basic parameters are set to $K = 46$, $T = 828$, $L = 1$, $s = 0.85$, $\mu = 0$ and $\sigma^2 = 1$ for all the algorithms. As illustrated in Figure 7.4, our algorithm outperforms other baseline algorithms in predicting ability.

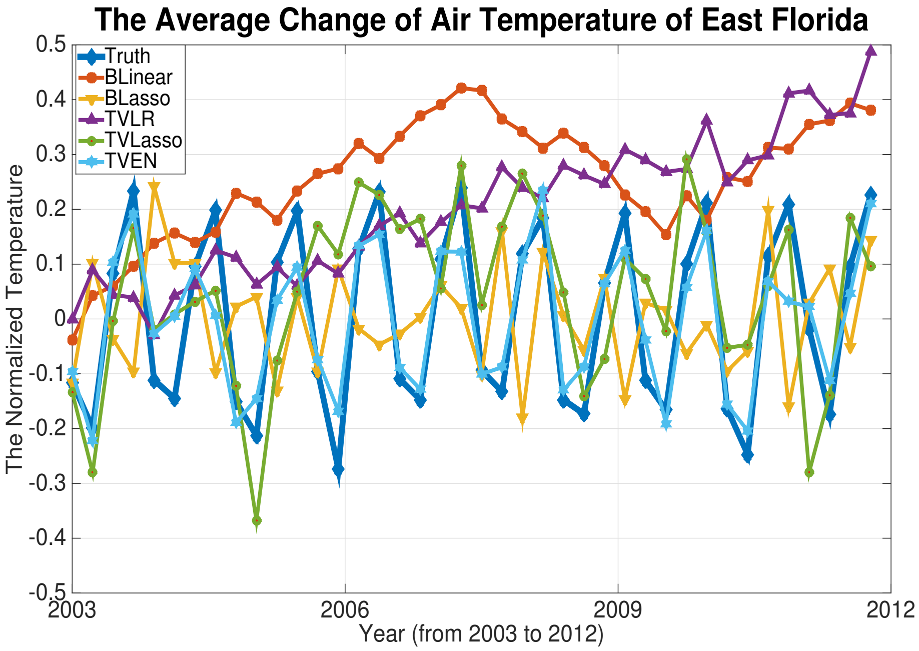


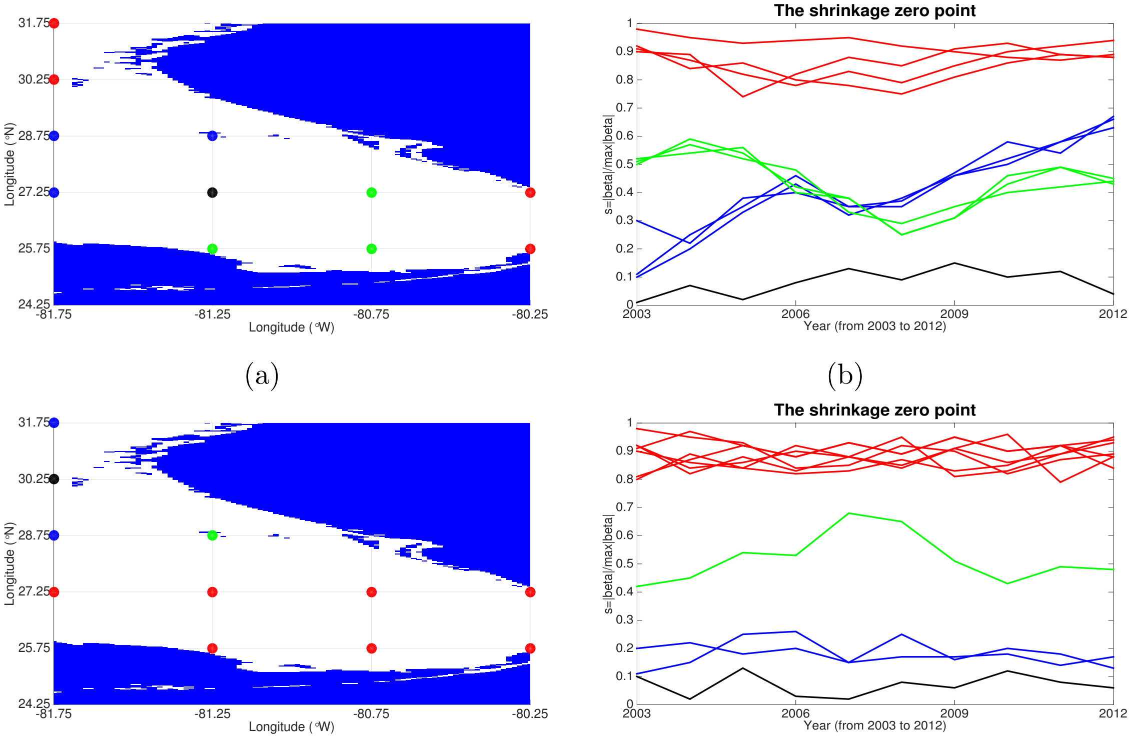
Figure 7.4: Average predicted value of the normalized air temperature on the total 46 grids of East Florida.

7.5.4.3 Group Effect

In this section, we further conduct experiments on data of contiguous 11 grids, a subset of aforementioned 46 points, to illustrate the ability of our algorithm in identifying group locations having similar dependencies towards one another location. Unlike the dependency analysis on the data of the globe or the entire USA [LLNM+09, Cas16], we ignore the influence of the weather in other far regions in our experiments since they are considered insignificant to the 11 grids [KKR+13], a relatively small area. We analyze the dependency matrices of the 11 locations towards two locations $(81.25°W, 27.25°N)$ and $(81.75°W, 30.25°N)$ to show the group effects of the air temperature among those locations.

Figure 7.5 shows the experimental results for the two target locations (black points) respectively. 4 groups are identified among the 11 locations by adjusting the *shrinkage ratio* s and locations in the same group are displayed with same colors. As shown in Figure 7.5, the black location, i.e., itself, has the most significant correlation for estimating the target air temperatures from time $t-1$ to time t for both two target locations. The relative close locations in color green and blue have larger power for predicting the target air temperatures from time $t-1$ to time t for the two locations than the red location.

The spatial-temporal dependency structures learned in our experiments are quite consistent with domain expertise which indicates our model is able to provide significant insights in MTS data



(a) (b)

(c) (d)

Figure 7.5: Group dependencies of air temperatures over time for two target locations. Subfigures (a) and (c) show the geographical locations and target locations are in black. Subfigures (b) and (d) show the zero points graph for the two target locations respectively.

Algorithm 4 The algorithm for VAR-Elastic-Net by Bayesian Update

```
procedure MAIN($B, L, \alpha_1, \alpha_2, \lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}, \boldsymbol{\beta}_t$)
    for $i = 1 : K$ do
        Initialize $y_{0,i}$ with $B$ particles;
        for $k = 1 : B$ do
            Initialize $\boldsymbol{\mu}_{\boldsymbol{\beta}'^{(k)}_{0,i}} = \mathbf{0}$;
            Initialize $\boldsymbol{\Sigma}_{\boldsymbol{\beta}'^{(k)}_{0,i}} = \mathbf{I}$;
        end for
    end for
    for $t = 1 : T$ do
        Obtain $\mathbf{x}_t$ using time lag $L$;
        for $i = 1 : K$ do
            UPDATE($\mathbf{x}_t, y_{t,i}, \boldsymbol{\beta}'_{t,i}, \boldsymbol{\eta}_{t,i}$);
            Output $\boldsymbol{\beta}_t$ according to Eq. 7.18;
        end for
    end for
end procedure

procedure UPDATE($\mathbf{x}_t, y_{t,i}, \boldsymbol{\beta}'_{t,i}, \boldsymbol{\eta}_{t,i}$)
    for $k = 1 : B$ do
        Compute particle weights $\rho^{(k)}_{t,i}$ by Eq. 7.22;
    end for
    Resample $\mathcal{P}^{(k)}_{t,i}$ from $\mathcal{P}^{(k)}_{t-1,i}$ according to $\rho^{(k)}_{t,i}$;
    for $i = 1 : B$ do
        Update $\boldsymbol{\mu}_{\boldsymbol{\eta}^{(k)}_{t,i}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}^{(k)}_{t,i}}$ by Eq. 7.23;
        Sample $\boldsymbol{\eta}^{(k)}_{t,i}$ according to Eq. 7.25;
        Update the parameters $\boldsymbol{\beta}'^{(k)}_{t,i}$, $\boldsymbol{\beta}'^{(k)}_{t,i}$, $\alpha^{(k)}_{t,1}$ and $\alpha^{(k)}_{t,2}$ by Eq. 7.26;
        Sample ${\sigma^2}^{(k)}_{t,i}$ and $\boldsymbol{\beta}'^{(k)}_{t,i}$ by Eq. 7.27;
    end for
end procedure
```

CHAPTER 8

CONCLUSION

The contribution of the thesis are as follows,

1. Here, we considered an optimal and adaptive step size rule for gradient descent (GD) applied to *non-convex* optimization problems. We finish to prove that GD with fixed step sizes not exceeding $2/L$, where L is the Lipschtiz constant, will not converge to strict saddle points almost surely, generalizing previous works of [LSJR16, PP16] that require step sizes to not exceed $1/L$. We also establish escaping strict saddle point properties of GD under varying/adaptive step sizes under additional conditions.

   We address an important open question and is to derive an explicit *rate of convergence* for the GD algorithm with different step size rules for non-convex objective functions. It is particularly interesting to study non-convex problems for which GD converges to local minima with number of iterations *polynomial* in problem dimension d. While the work of [DJL+17] rules out such possibility for general smooth f, polynomial iteration complexity of GD might still be possiblbe for non-convex objectives under additional assumptions.

2. Based on the idea of understanding arithmetical complexity from analytical complexity in the seminal book by Nesterov [Nes13] and the idea for viewing optimization from differential equation in the novel blog[1] , we propose some novel algorithms based on Newton Second Law with the kinetic energy observable and controllable in the computational process firstly. Although our algorithm cannot fully solve the global optimization problem, or it is dependent on the trajectory path more unpositively, this work introduce Hamilton system essentially to optimization such that it is possible that the global minima

[1] http://www.offconvex.org/2015/12/11/mission-statement/

can be obtained. Our algorithms are easy to implement and own more rapid convergence rate.

From the theoretical view, the Hamilton system is closer to nature and a lot of fundamental work have appeared on the last decades, such as KAM theory, Nekhoroshev estimate, operator spectral theory and so on [Arn13, Arn12]. Are these beautiful and essentially original work used to understand and improve the algorithm for optimization and machine learning? Furthermore, to estimate the convergence rate, the matrix containing the trigonometric function is hard to estimate. Some estimate for the trigonometric matrix based on spectral theory are proposed in [JL17, LY15]. For the numerical scheme, we only exploit the simple first-order symplectic Euler method. A lot of more efficient schemes, such as Störmer-Verlet scheme, Symplectic Runge-Kutta scheme, order condition method and so on, are proposed on [HLW06]. These schemes can make the algorithms in this paper more efficient and accurate. For the optimization, the method we proposed is only about unconstrained problem. In the nature, the classical Newton Second law, or the equivalent expression — Lagrange mechanics and Hamilton mechanics, is implemented on the manifold in the almost real physical world. In other word, a natural generalization is from unconstrained problem to constrained problem for our proposed algorithms. A more natural implementation is the geodesic descent in [LY$^+$84]. Similar as the development of the gradient method from smooth condition to nonsmooth condition, our algorithms can be generalized to nonsmooth condition by the subgradient. For application, we will implement our algorithms to Non-negative Matrix Factorization, Matrix Completion and Deep Neural Network and speed up the training of the objective function. Meanwhile, we apply the algorithms proposed in this paper to the maximum likelihood estimator and maximum a posteriori estimator in statistics.

Starting from Newton Second Law, we implement only a simple particle in classical mechanics, or macroscopic world. A natural generalization is from the macroscopic world to the microscopic world. In the field of fluid dynamics, the Newton second Law is expressed by Euler equation, or more complex Navier-Stokes equation. An important topic from fluid dynamics is geophysical fluid dynamics [Ped13, CRB11] , containing atmospheric science and oceanography. Especially, a key feature in the oceanography different from atmospheric science is the topography, which influence mainly vector field of the fluid. Many results have been demonstrated based on many numerical modeling , such as the classical POM[2], HYCOM[3], ROMS[4] and FVCOM[5]. A reverse idea is that if we view the potential function in black box is the topography, we observe the changing of the fluid vector field to find the number of local minima in order to obtain the global minima with a suitable initial vector field. A more adventurous idea is to generalize the classical particle to the quantum particle. For quantum particle, the Newton second law is expressed by the energy form, that is from the view of Hamilton mechanics, which is the starting point for the proposed algorithm in this paper. The particle appears in wave form in microscopic world. When the wave meet the potential barrier, the tunneling phenomena will appear. The tunneling phenomena still appear in high dimension [NM13]. It is very easy to observe the tunneling phenomena in the physical world. If the computer can be very easy to simulate the quantum world, we can find the global minima by binary section search. That is, if there exist tunneling phenomena in the

[2] http://ofs.dmcr.go.th/thailand/model.html

[3] https://hycom.org/

[4] https://www.myroms.org/

[5] http://fvcom.smast.umassd.edu/

upper level, continue to detect the upper level in the upper level, otherwise to go the lower level. In quantum world, it need only $\mathcal{O}(\log n)$ times to find global minima other than NP-hard.

3. In this paper we propose a general ODE-based approach to analyze accelerated gradient methods. Our approach explains why NAGs provide acceleration and imply new quantitative results on the convergence of gradient norms. In the sequel, we list some future directions.

   The variational perspective from Lagrangian mechanics is adopted in [WRJ16, WWJ16]. In contrast we draw ideas from Hamiltonian mechanics. While Lagrangian mechanics and Hamiltonian mechanics are equivalent under the Legendre transformation, the symplectic structure of Hamiltonian systems and the its view of energy [LL60] give us intuition on using phase space representation and how to construct energy functionals. Roughly speaking, the phase representation is a bridge between continuous energy functionals and discrete energy functionals.

8.0.0.0.1 Generating new algorithms through discretizing ODEs

[WWJ16, WRJ16, BJW18] considered different discretization schemes on *low-resolution ODE*s to generate new optimization algorithms. As we discussed in Section 5.1.3, only the implicit scheme can obtain the optimal convergence rate but it is not practical. Since our proposed ODEs are different from their *low-resolution ODE*s, applying discretization schemes on our proposed ODEs naturally generate another class of optimization algorithms and these new algorithms may admit competitive convergence rates as NAGs.

8.0.0.0.2 Generalization to Primal-dual Approach In many machine learning problems, one uses the primal-dual (accelerated) method to speed up convergence [DH18, DCL$^+$17, CP11, LYW$^+$17, WX17]. Similar to NAGs, existing analyses of the accelerated algorithms also require estaimted sequence based approach. It would be interesting to generalize ODE approach to characterize the convergence rates of accelerated primal-dual gradient methods and give simpler proofs.

8.0.0.0.3 Generalization to Non-convex Functions In deriving our *high-resolution ODE*s, we do not rely on the convexity of the objective function. Thus, even if the objective function is non-convex, we can still use these ODEs for analysis. A concrete problem is to analyze the convergence rate for the general smooth but non-convex problem [Nes13]. Another interesting direction is to use our framework to study the local behavior of the optimization algorithms around a strict saddle point [JGN$^+$17, DJL$^+$17].

8.0.0.0.4 Understanding optimization algorithms by higher order ODEs In this paper, we only consider gradient-Lipschitz objective functions. When the function has higher order smoothness, e.g., Hessian is Lipschitz, it is possible to obtain faster optimization algorithms. [Pol64] initiated the study in this direction. In his seminal paper, he proposed to study the following ODE

$$X^{(n)}(t) + \binom{n}{1} a^1 X^{(n-1)}(t) + \ldots + \binom{n}{n-1} a^{n-1} X^{(1)}(t) + \nabla f(X(t)) = 0.$$

where a is some problem dependent constants. Notice that this is essentially "low-resolution" because there is no step size parameter. An interesting direction is to derive high-resolution or higher resolution ODEs and study the convergence rates.

5. In this paper we propose a CoCoSSC formulation for the subspace clustering problem with data subjecting to stochastic Gaussian noise or missing entries. Our proposed method enjoys improved sample complexity and practical performance, and is also computationally efficient.

   An interesting future direction is to further improve the sample complexity to $\rho = \Omega(n^{-1/2})$ without knowing the norms $\|\boldsymbol{y}_i\|_2$, Such sample complexity is likely to be optimal because it is the smallest observation rate under which off-diagonal elements of sample covariance $\mathbf{X}^\top\mathbf{X}$ can be consistently estimated in max norm, which is also shown to be optimal for related regression problems [WWBS17].

6. In this paper, we proposed a novel VAR-Elastic-Net model with online Bayesian update allowing for both stable-sparsity and group-selection among MTS, which implements adaptive inference strategy of particle learning. Extensive empirical studies on both the synthetic and real MTS data demonstrate the effectiveness and the efficiency of the proposed method. In the process of time-varying temporal dependency discovery from MTS, the choice of regularizer is essential. One possible future work is to automate the identification of the proper regularizer for different MTS in an online setting.

BIBLIOGRAPHY

[AAB$^+$17] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *STOC*, 2017. Full version available at http://arxiv.org/abs/1611.01146.

[AABR02] Felipe Alvarez, Hedy Attouch, Jérôme Bolte, and P Redont. A second-order gradient-like dissipative dynamical system with hessian-driven damping.-application to optimization and mechanics. *Journal de mathématiques pures et appliquées*, 81(8):747–780, 2002.

[AAZB$^+$17] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199. ACM, 2017.

[ACPR18] Hedy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1-2):123–175, 2018.

[ACR17] Hedy Attouch, Zaki Chbani, and Hassan Riahi. Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. *arXiv preprint arXiv:1706.05671*, 2017.

[AG16] Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *arXiv preprint arXiv:1602.05908*, 2016.

[ALA07] Andrew Arnold, Yan Liu, and Naoki Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 66–75. ACM, 2007.

[AMR12] Hedy Attouch, Paul-Emile Maingé, and Patrick Redont. A second-order differential system with hessian-driven damping; application to non-elastic shock laws. *Differential Equations & Applications*, 4(1):27–65, 2012.

[AP16] Hedy Attouch and Juan Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

[APR16] Hedy Attouch, Juan Peypouquet, and Patrick Redont. Fast convex optimization via inertial dynamics with hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783, 2016.

[Arn12] Vladimir Igorevich Arnol'd. *Geometrical methods in the theory of ordinary differential equations*, volume 250. Springer Science & Business Media, 2012.

[Arn13] Vladimir Igorevich Arnol'd. *Mathematical methods of classical mechanics*, volume 60. Springer Science & Business Media, 2013.

[AZ18] Zeyuan Allen-Zhu. How to make the gradients small stochastically. *arXiv preprint arXiv:1801.02982*, 2018.

[B$^+$15] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[BBW$^+$90] FP Bretherton, K Bryan, JD Woods, et al. Time-dependent greenhouse-gas-induced climate change, 1990.

[BJ03] Ronen Basri and David Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

[BJRL15] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[BJW18] Michael Betancourt, Michael I Jordan, and Ashia C Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

[BL12] Mohammad Taha Bahadori and Yan Liu. On causality inference in time series. In *AAAI Fall Symposium: Discovery Informatics*, 2012.

[BL13] Mohammad Taha Bahadori and Yan Liu. An examination of practical granger causality inference. In *Proceedings of the 2013 SIAM International Conference on data Mining*, pages 467–475. SIAM, 2013.

[BLE17] Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85. ACM, 2017.

[BLS15] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[BM00] P.S. Bradley and O.L. Mangasarian. K-plane clustering. *Journal of Global Optimization*, 16(1):23–32, 2000.

[BNS16] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[BPC$^+$11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, January 2011.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[Cas16] Stefano Castruccio. Assessing the spatio-temporal structure of annual and seasonal surface temperature for cmip5 and reanalysis. *Spatial Statistics*, 18:179–193, 2016.

[CD16] Yair Carmon and John C Duchi. Gradient descent efficiently finds the cubic-regularized non-convex Newton step. *arXiv preprint arXiv:1612.00547*, 2016.

[CDG00] Bruno Carpentieri, Iain S Duff, and Luc Giraud. Sparse pattern selection strategies for robust frobenius-norm minimization preconditioners in electromagnetism. *Numerical linear algebra with applications*, 7(7-8):667–685, 2000.

[CDHS16] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for non-convex optimization. *arXiv preprint arXiv:1611.00756*, 2016.

[CDHS17a] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *arXiv preprint arXiv:1710.11606*, 2017.

[CDHS17b] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points II: First-order methods. *arXiv preprint arXiv:1711.00841*, 2017.

[CJLP10] Carlos M. Carvalho, Michael S. Johannes, Hedibert F. Lopes, and Nicholas G. Polson. Particle Learning and Smoothing. *Statistical Science*, 25:88–106, 2010.

[CJW17] Zachary Charles, Amin Jalali, and Rebecca Willett. Sparse subspace clustering with missing and corrupted data. *arXiv preprint: arXiv:1707.02461*, 2017.

[CK98] Joao Costeira and Takeo Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

[CML17] Shixiang Chen, Shiqian Ma, and Wei Liu. Geometric descent method for convex composite minimization. In *Advances in Neural Information Processing Systems*, pages 636–644, 2017.

[CP11] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[CR09] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[CRB11] Benoit Cushman-Roisin and Jean-Marie Beckers. *Introduction to geophysical fluid dynamics: physical and numerical aspects*, volume 101. Academic Press, 2011.

[CRS14] Frank E Curtis, Daniel P Robinson, and Mohammadreza Samadi. A trust region algorithm with a worst-case iteration complexity of $O(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, pages 1–32, 2014.

[CT05] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on Information Theory*, 51(12):4203–4215, 2005.

[CT07] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[DCL$^+$17] Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. *arXiv preprint arXiv:1702.07944*, 2017.

[DFR18] Dmitriy Drusvyatskiy, Maryam Fazel, and Scott Roy. An optimal first order method based on optimal quadratic averaging. *SIAM Journal on Optimization*, 28(1):251–271, 2018.

[DGA00] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.

[DH18] Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. *arXiv preprint arXiv:1802.01504*, 2018.

[DJL$^+$17] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Barnabas Poczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2017.

[DKZ$^+$03] Petar M Djuric, Jayesh H Kotecha, Jianqui Zhang, Yufei Huang, Tadesse Ghirmai, Mónica F Bugallo, and Joaquin Miguez. Particle filtering. *IEEE signal processing magazine*, 20(5):19–38, 2003.

[DO17] Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *arXiv preprint arXiv:1712.02485*, 2017.

[DZ17] Abhirup Datta and Hui Zou. Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45(6):2400–2426, 2017.

[EBN12] Brian Eriksson, Laura Balzano, and Robert Nowak. High rank matrix completion. In *AISTATS*, 2012.

[Eic06] Michael Eichler. Graphical modelling of multivariate time series with latent variables. *Preprint, Universiteit Maastricht*, 2006.

[EV13] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[FB15] Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695, 2015.

[Fio05] Simone Fiori. Quasi-geodesic neural learning algorithms over the orthogonal group: A tutorial. *Journal of Machine Learning Research*, 6(May):743–781, 2005.

[FKM05] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

[FM83] Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569, 1983.

[FRMP18] Mahyar Fazlyab, Alejandro Ribeiro, Manfred Morari, and Victor M Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018.

[GGZ18] Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient hamiltonian monte carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv preprint arXiv:1809.04618*, 2018.

[GH13] John Guckenheimer and Philip Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, volume 42. Springer Science & Business Media, 2013.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[GJZ17] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1233–1242, 2017.

[GL16] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

[GLM16] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[GM74] Philip E Gill and Walter Murray. Newton-type methods for unconstrained and linearly constrained optimization. *Mathematical Programming*, 7(1):311–350, 1974.

[Ham94] James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

[Har71] Philip Hartman. The stable manifold of a point of a hyperbolic map of a banach space. *Journal of Differential Equations*, 9(2):360–379, 1971.

[Har82] Philip Hartman. Ordinary differential equations, classics in applied mathematics, vol. 38, society for industrial and applied mathematics (siam), philadelphia, pa, 2002, corrected reprint of the second (1982) edition, 1982.

[Har90] Andrew C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, March 1990.

[HB15] Reinhard Heckel and Helmut Bölcskei. Robust subspace clustering via thresholding. *IEEE Transactions on Information Theory*, 61(11):6320–6342, 2015.

[HL14] Elad Hazan and Kfir Levy. Bandit convex optimization: Towards tight bounds. In *Advances in Neural Information Processing Systems*, pages 784–792, 2014.

[HL17] Bin Hu and Laurent Lessard. Dissipativity theory for nesterov's accelerated method. *arXiv preprint arXiv:1706.04381*, 2017.

[HLLL17] Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

[HLS17] Wenqing Hu, Chris Junchi Li, and Weijie Su. On the global convergence of a randomly perturbed dissipative nonlinear oscillator. *arXiv preprint arXiv:1712.05733*, 2017.

[HLW06] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

[HM12] Uwe Helmke and John B Moore. *Optimization and Dynamical Systems*. Springer Science & Business Media, 2012.

[HMC$^+$18] Li He, Qi Meng, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Differential equations for modeling asynchronous algorithms. *arXiv preprint arXiv:1805.02991*, 2018.

[HMR16] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

[HTB17] Reinhard Heckel, Michael Tschannen, and Helmut Bölcskei. Dimensionality-reduced subspace clustering. *Information and Inference: A Journal of the IMA*, 6(3):246–283, 2017.

[JCSX11] Ali Jalali, Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. In *ICML*, 2011.

[JGN$^+$17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1724–1732, 2017.

[JHS$^+$11] Manoj Joshi, Ed Hawkins, Rowan Sutton, Jason Lowe, and David Frame. Projections of when temperature change will exceed 2 [deg] c above preindustrial levels. *Nature Climate Change*, 1(8):407–412, 2011.

[JL17] Svetlana Jitomirskaya and Wencai Liu. Arithmetic spectral transitions for the maryland model. *Communications on Pure and Applied Mathematics*, 70(6):1025–1051, 2017.

[JNJ17] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*, 2017.

[JYG+03] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *science*, 302(5644):449–453, 2003.

[KB17] Walid Krichene and Peter L Bartlett. Acceleration and averaging in stochastic descent dynamics. In *Advances in Neural Information Processing Systems*, pages 6796–6806, 2017.

[KBB15] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in neural information processing systems*, pages 2845–2853, 2015.

[KBB16] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Adaptive averaging in accelerated descent dynamics. In *Advances in Neural Information Processing Systems*, pages 2991–2999, 2016.

[KF18] Donghwan Kim and Jeffrey A Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *arXiv preprint arXiv:1803.06600*, 2018.

[KKR+13] William Kleiber, Richard W Katz, Balaji Rajagopalan, et al. Daily minimum and maximum temperature simulation over complex terrain. *The Annals of Applied Statistics*, 7(1):588–612, 2013.

[KMO10] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.

[LKJ09] Yan Liu, Jayant R Kalagnanam, and Oivind Johnsen. Learning dynamic temporal graphs for oil-production equipment monitoring system. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234. ACM, 2009.

[LL60] LD Landau and EM Lifshitz. Classical mechanics. *ed: Pergamon Press, Oxford*, 1960.

[LL$^+$10] Qing Li, Nan Lin, et al. The bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 2010.

[LLNM$^+$09] Aurelie C Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan Hosking, and Naoki Abe. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 587–596. ACM, 2009.

[LLY$^+$13] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

[LMH18] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18(212):1–54, 2018.

[LPP$^+$17] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

[LRP16] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[LS17] Tengyuan Liang and Weijie Su. Statistical inference for the population landscape via moment adjusted stochastic gradients. *arXiv preprint arXiv:1712.07519*, 2017.

[LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

[LTE17] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110, 2017.

[LWL$^+$16] Xingguo Li, Zhaoran Wang, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, and Tuo Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*, 2016.

[LY$^+$84] David G Luenberger, Yinyu Ye, et al. *Linear and nonlinear programming*, volume 2. Springer, 1984.

[LY15] Wencai Liu and Xiaoping Yuan. Anderson localization for the completely resonant phases. *Journal of Functional Analysis*, 268(3):732–747, 2015.

[LY17] Mingrui Liu and Tianbao Yang. On noisy negative curvature descent: Competing with gradient descent for faster non-convex optimization. *arXiv preprint arXiv:1709.08571*, 2017.

[LYW$^+$17] Qi Lei, Ian En-Hsu Yen, Chao-yuan Wu, Inderjit S Dhillon, and Pradeep Ravikumar. Doubly greedy primal-dual coordinate descent for sparse empirical risk minimization. In *International Conference on Machine Learning*, pages 2034–2042, 2017.

[LZZ$^+$16] Tao Li, Wubai Zhou, Chunqiu Zeng, Qing Wang, Qifeng Zhou, Dingding Wang, Jia Xu, Yue Huang, Wentao Wang, Minjing Zhang, et al. Didap: An efficient disaster information delivery and analysis platform in disaster management. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1593–1602. ACM, 2016.

[May17] Ramzi May. Asymptotic for a second-order evolution equation with convex potential and vanishing damping term. *Turkish Journal of Mathematics*, 41(3):681–685, 2017.

[MDHW07] Yi Ma, Harm Derksen, Wei Hong, and John Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1546–1562, 2007.

[MS79] Jorge J Moré and Danny C Sorensen. On the use of directions of negative curvature in a modified newton method. *Mathematical Programming*, 16(1):1–20, 1979.

[Mur02] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning.* PhD thesis, University of California, Berkeley, 2002.

[Mur12] Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[Nes83] Yurii Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[Nes12] Yurii Nesterov. How to make the gradients small. *Optima*, 88:10–11, 2012.

[Nes13] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

[NH11] Behrooz Nasihatkon and Richard Hartley. Graph connectivity in sparse subspace clustering. In *CVPR*, 2011.

[NM13] Hiroki Nakamura and Gennady Mil'nikov. *Quantum mechanical tunneling in chemical physics.* CRC Press, 2013.

[NN88] Yurii Nesterov and A Nemirovsky. A general approach to polynomial-time algorithms design for convex programming. Technical report, Technical report, Centr. Econ. & Math. Inst., USSR Acad. Sci., Moscow, USSR, 1988.

[NP06] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[NY83] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[OC15] Brendan Odonoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[ORX$^+$16] Stanley Osher, Feng Ruan, Jiechao Xiong, Yuan Yao, and Wotao Yin. Sparse recovery via differential inclusions. *Applied and Computational Harmonic Analysis*, 41(2):436–469, 2016.

[OW17] Michael O'Neill and Stephen J Wright. Behavior of accelerated gradient methods near critical points of nonconvex problems. *arXiv preprint arXiv:1706.07993*, 2017.

[PCS14] Dohyung Park, Constantine Caramanis, and Sujay Sanghavi. Greedy subspace clustering. In *NIPS*, 2014.

[Ped13] Joseph Pedlosky. *Geophysical fluid dynamics*. Springer Science & Business Media, 2013.

[Pem90] Robin Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, pages 698–712, 1990.

[Per13] Lawrence Perko. *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media, 2013.

[PKCS17] Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

[Pol64] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[Pol87a] Boris T Polyak. Introduction to optimization. *Optimization Software, Inc, New York*, 1987.

[Pol87b] Boris T Polyak. Introduction to optimization. translations series in mathematics and engineering. *Optimization Software*, 1987.

[PP16] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*, 2016.

[QX15] Chao Qu and Huan Xu. Subspace clustering with irrelevant features via robust dantzig selector. In *NIPS*, 2015.

[Rec11] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[RHW+88] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[RS15] Vijay K Rohatgi and AK Md Ehsanes Saleh. *An introduction to probability and statistics*. John Wiley & Sons, 2015.

[RW17] Clément W Royer and Stephen J Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *arXiv preprint arXiv:1706.03131*, 2017.

[RZS+17] Sashank J Reddi, Manzil Zaheer, Suvrit Sra, Barnabas Poczos, Francis Bach, Ruslan Salakhutdinov, and Alexander J Smola. A generic approach for escaping saddle points. *arXiv preprint arXiv:1709.01434*, 2017.

[SB13] Josef Stoer and Roland Bulirsch. *Introduction to Numerical Analysis*, volume 12. Springer Science & Business Media, 2013.

[SBC14] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterovs accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[SBC16] Weijie Su, Stephen Boyd, and Emmanuel J Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[SC12] Mahdi Soltanolkotabi and Emmanuel J Candes. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[Sch00] Johannes Schropp. A dynamical systems approach to constrained minimization. *Numerical Functional Analysis and Optimization*, 21(3-4):537–551, 2000.

[SEC14] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J Candes. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

[SHB16] Yi Shen, Bin Han, and Elena Braverman. Stability of the elastic net estimator. *Journal of Complexity*, 32(1):20–39, 2016.

[Shu13] Michael Shub. *Global stability of dynamical systems.* Springer Science & Business Media, 2013.

[SMDH13] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[Sol14] Mahdi Soltanolkotabi. *Algorithms and theory for clustering and nonconvex quadratic programming.* PhD thesis, Stanford University, 2014.

[SQW16] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 2379–2383. IEEE, 2016.

[SQW17] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

[TG17] Panagiotis A Traganitis and Georgios B Giannakis. Sketched subspace clustering. *IEEE Transactions on Signal Processing*, 66(7):1663–1675, 2017.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Tse00] P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Application*, 105(1):249–252, 2000.

[TV17] Manolis Tsakiris and Rene Vidal. Algebraic clustering of affine subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[TV18] Manolis C Tsakiris and Rene Vidal. Theoretical analysis of sparse subspace clustering with missing entries. *arXiv preprint arXiv:1801.00393*, 2018.

[Vid11] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

[VJFC18] Apidopoulos Vassilis, Aujol Jean-François, and Dossal Charles. The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case $b < 3$. *SIAM Journal on Optimization*, 28(1):551–574, 2018.

[VMS05] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

[WN99] Stephen Wright and Jorge Nocedal. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

[WRJ16] Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

[WWBS17] Yining Wang, Jialei Wang, Sivaraman Balakrishnan, and Aarti Singh. Rate optimal estimation and confidence intervals for high-dimensional regression with missing covariates. *arXiv preprint arXiv:1702.02686*, 2017.

[WWJ16] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, page 201614734, 2016.

[WWS15a] Yining Wang, Yu-Xiang Wang, and Aarti Singh. A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data. In *ICML*, 2015.

[WWS15b] Yining Wang, Yu-Xiang Wang, and Aarti Singh. Differentially private subspace clustering. In *NIPS*, 2015.

[WWS16] Yining Wang, Yu-Xiang Wang, and Aarti Singh. Graph connectivity in noisy sparse subspace clustering. In *AISTATS*, 2016.

[WX16] Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. *Journal of Machine Learning Research*, 17(12):1–41, 2016.

[WX17] Jialei Wang and Lin Xiao. Exploiting strong convexity from data with primal-dual first-order algorithms. *arXiv preprint arXiv:1703.02624*, 2017.

[XWG18] Pan Xu, Tianhao Wang, and Quanquan Gu. Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions.

In *International Conference on Machine Learning*, pages 5488–5497, 2018.

[YP06] Jingyu Yan and Marc Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *ECCV*, 2006.

[YRV15] Congyuan Yang, Daniel Robinson, and Rene Vidal. Sparse subspace clustering with missing entries. In *ICML*, 2015.

[ZF09] Cunlu Zou and Jianfeng Feng. Granger causality vs. dynamic bayesian network inference: a comparative study. *BMC bioinformatics*, 10(1):122, 2009.

[ZFIM12] Amy Zhang, Nadia Fawaz, Stratis Ioannidis, and Andrea Montanari. Guess who rated this movie: Identifying users through subspace clustering. In *UAI*, 2012.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[ZMSJ18] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. *arXiv preprint arXiv:1805.00521*, 2018.

[ZWW$^+$16] Chunqiu Zeng, Qing Wang, Wentao Wang, Tao Li, and Larisa Shwartz. Online inference for time-varying temporal dependency discovery from time series. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 1281–1290. IEEE, 2016.

VITA

BIN SHI

September 13, 1983	Born, Qingdao, China
2006	B.S., Mathematics and Applied Mathematics Ocean University of China Qingdao, China
2011	M.S., Pure Mathematics Fudan University Shanghai, China
2015	M.S., Theoretical Physics University of Massachusetts Dartmouth Dartmouth, Massachusetts
2015 - Present	Ph.D, Computer Science Florida International Univerisity Miami, Florida

PUBLICATIONS AND PRESENTATIONS

[1] Bin Shi, Tao Li, Sundaraja S. Iyengar (2017). *A Conservation Law Method in Optimization.* 10th NIPS Workshop on Optimization for Machine Learning

[2] Bin Shi, Simon S. Du, Yining Wang, Jason Lee (2018).*Gradient Descent Converges to Minimizers: Optimal and Adaptive Step Size Rules.* INFORMS Journal on Optimization (Accepted with minor revision)

[3] Yining Wang, Bin Shi, Yuxiang Wang, Yudong Tao, Sundaraja S. Iyengar (2018). *Improved Sample Complexity in Sparse Subspace Clustering with Noisy and Missing Observations.* (Submit to AISTATS 2018, Joint work with CMU)

[4] Wentao Wang, Bin Shi, Tao Li, Sundaraja S. Iyengar (2018). *Online Discovery For Stable and Grouping Causalities in Multivariate Time Series.* (Submit to TKDD)

[5] Bin Shi, Simon S. Du, Michael I. Jordan, Weijie J. Su (2018) *Understanding the Acceleration Phenomenon via High-Resolution Differential Equations* (To be Submitted to Mathematic Programming)