ROBUSTNESS OF GRADIENT METHODS FOR DATA DRIVEN DECISION MAKING

by

Hesameddin Mohammadi

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

December 2022

Acknowledgments

I am grateful to all my colleagues, friends, and family who tremendously supported me during my PhD journey. I try to give a partial accounting of some of their aid and support.

I would like to begin by saying a special thank you to my great advisor Prof. Mihailo R. Jovanović who certainly falls into all the above three categories. Mihailo never hesitated to offer his strongest support, ranging from suggesting new research directions and ideas, helping with technical challenges, and providing invaluable feedback all along, to patiently going over our papers revising draft after draft till near perfection, to accommodating and aiding me during some of my hardest experiences in this journey. In addition, he generously dedicated funding resources at his disposal to my research throughout these years, which was one of the most important enablers of this work. Mihailo was always a great source of inspiration and I am truly thankful to have had him as my top personal and professional mentor, and academic father.

I would also like to express my sincere gratitude to my committee members Prof. Urbashi Mitra, Prof. Pierluigi Nuzzo, Prof. Mahdi Soltanolkotabi, and Prof. Meisam Razaviyayn. I benefited so much from their feedback and the discussions I had with them. Also, my collaboration with Mahdi and Meisam was extremely helpful to the development of this dissertation. Interacting with them was always intellectually inspiring and it significantly helped me broaden my perspectives.

I would also like to thank my former/current labmates Dr. Reza Kamyar, Dr. Morgan Jones, Dr. Marziye Rahimi, Prof. Armin Zare, Dr. Wei Ran, Dr. Sepideh Hassan Moghaddam, Dr. Dongsheng Ding, Dr. Anubhav Dwivedi, Evgeny Meyer, Samantha Samuelson, and Ibrahim Ozaslan. Interacting with each of them was always a pleasure and I learnt so much from all of them. The many hours I spent in the lab/office everyday would be unbearable if I did not have the greatest company of them.

I am thankful to all the great friends I have kept and made: Iman Bonakdar, Javad Abazari, Sina Tohidi, Dr. Mehdi Ataei, Zalan Fabian, Bowen Song, Mohammadmahdi Sajedi, Dr. Mo Hekmat, and Dr. Mohamadreza Ahmadi. The never-ending support I have received from them made it possible for me to continue.

Last but not least, I am always indebted to my parents Elaheh and Behrouz, my sister Noura, my brother Shahab, and my dearest friend Tara. You were always there for me and remained the main source of motivation for me to finish the PhD.

Table of Contents

List of Figures

Abstract

First-order optimization algorithms are increasingly used for data-driven control and many learning applications that often involve uncertain and noisy environments. In this thesis, we employ control-theoretic tools to study the stochastic performance of these algorithms in solving general (strongly) convex and some nonconvex optimization problems that arise in reinforcement learning and control theory.

In particular, we first study momentum-based accelerated optimization algorithms in which the iterations utilize information from the two previous steps and are subject to additive white noise. This class of algorithms includes Polyak's heavy-ball and Nesterov's accelerated methods as special cases and noise accounts for uncertainty in either gradient evaluation or iteration updates. For unconstrained, smooth, strongly convex optimization problems, we examine the mean-squared error in the optimization variable to quantify noise amplification. By leveraging the theory of Lyapunov and integral quadratic constraints, we establish an upper bound on the noise amplification of Nesterov's method with standard parameters that is tight up to a constant factor. We also use strongly convex quadratic problems to identify fundamental tradeoffs between noise amplification and convergence rate for the two-step momentum algorithms. For this class of problems, we explicitly evaluate the steady-state variance of the optimization variable in terms of the eigenvalues of the Hessian of the objective function. We also introduce a novel geometric characterization of conditions for linear convergence that clarifies the relation between the noise amplification and convergence rate as well as their dependence on the condition number and the constant algorithmic parameters. This geometric insight leads to simple alternative proofs of standard convergence results and allows us to establish analytical lower bounds on the product between the settling time and noise amplification that scale quadratically with the condition number. Our analysis also identifies a key difference between the gradient and iterate noise models: while the amplification of gradient noise can be made arbitrarily small by sufficiently decelerating the algorithm, the best achievable variance amplification for the iterate noise model increases linearly with the settling time in the decelerated regime. We also characterize the impact of condition number on worst-case transient responses of popular accelerated algorithms and examine the noise amplification of a class of primal-dual gradient flow dynamics based on the proximal augmented Lagrangian that can be used for non-smooth convex constrained optimization problems.

We next focus on model-free reinforcement learning which attempts to find an optimal control action for an unknown dynamical system by directly searching over the parameter space of controllers. The convergence behavior and statistical properties of these approaches are often poorly understood because of the nonconvex nature of the underlying optimization problems and the lack of exact gradient computation. In this thesis, we take a step towards

demystifying the performance and efficiency of such methods by focusing on the standard linear quadratic regulator (LQR) with unknown state-space parameters. For this problem, we establish exponential stability for the ordinary differential equation (ODE) that governs the gradient-flow dynamics over the set of stabilizing feedback gains and show that a similar result holds for the standard gradient descent. We also provide theoretical bounds on the convergence rate and sample complexity of the random search method with two-point gradient estimates. We prove that in the model-free setup, the required simulation time and the total number of function evaluations both scale with the logarithm of the inverse of the desired accuracy. The key enabler of our results is the PL condition that holds for the LQR problem both in continuous and discrete time. We finish the thesis by showing the absence of this condition for the linear quadratic Gaussian problem with incomplete state information.

Chapter 1

Introduction

First-order methods are well suited for solving a broad range of optimization problems that arise in statistics, signal and image processing, control, and machine learning [1]–[5]. Among these algorithms, accelerated methods enjoy the optimal rate of convergence and they are popular because of their low per-iteration complexity. There is a large body of literature dedicated to the convergence analysis of these methods under different stepsize selection rules [4]–[9]. In many applications, however, these algorithms are brought into uncertain and noisy environments and they may only be used with limited time budgets.

For example, the exact value of the gradient is often not fully available or noise may corrupt the iterates of the algorithm due to uncertain communication. This happens when the objective function is obtained via costly simulations (e.g., tuning of hyper-parameters in supervised/unsupervised learning [10]–[12] and model-free optimal control [13]–[15]), when evaluation of the objective function relies on noisy measurements (e.g., embedded and real-time applications), or when the noise is due to communication between different agents (e.g., distributed computation over networks). Another related application arises in the context of (batch) stochastic gradient, where at each iteration the gradient of the objective function is computed from a small batch of data points. Such a batch gradient is known to be a noisy unbiased estimator for the gradient of the training loss. Moreover, additive noise may be introduced deliberately in the context of nonconvex optimization to help the iterates escape saddle points and improve generalization [16], [17].

In addition to uncertainty, many emerging applications [18], [19] that arise in modern Reinforcement Learning (RL) involve optimization landscapes that lack convexity. In these applications control-oriented models are not readily available and classical approaches from optimal control may not be directly applicable. In spite of these challenges, model-free RL approaches that rely on first-order optimization algorithms and prescribe control actions using estimated values of a cost function achieve empirical success in a variety of domains [20], [21]. Unfortunately, however, our mathematical understanding of these algorithms is still in its infancy and there are many open questions surrounding convergence and sample complexity.

Motivated by these observations, in this dissertation, we first use control theoretic tools to analyze the stochastic performance and transient response of accelerated optimization algorithms for smooth strongly convex problems and identify fundamental tradeoffs between convergence rate and noise amplification. Then, we turn our attention to the performance of first-order methods in model-free RL and focus on the infinite-horizon Linear Quadratic

Regulator (LQR) problem. In spite of the lack of convexity, we establish linear convergence of gradient descent and examine the convergence and sample complexity of the random search method [22] that attempts to emulate the behavior of gradient descent via gradient approximations resulting from evaluating random estimates of the objective function.

1.1 Main topics

In this section, we discuss the main topics of the dissertation.

1.1.1 Noise amplification of accelerated optimization algorithms

There is a vast body of literature that considers the robustness of first-order accelerated optimization algorithms under different types of noisy/inexact gradient oracles [23]–[28]. For example, in a deterministic noise scenario, an upper bound on the error in iterates for accelerated proximal gradient methods was established in [29]. This study showed that both proximal gradient and its accelerated variant can maintain their convergence rates provided that the noise is bounded and that it vanishes fast enough. Moreover, it has been shown that in the presence of random noise, with the proper diminishing stepsize, acceleration can be achieved for general convex problems. However, in this case optimal rates are *sub-linear* [30].

In the context of stochastic approximation, while early results suggest to use a stepsize that is inversely proportional to the iteration number [24], a more robust behavior can be obtained by combining larger stepsizes with averaging [25], [31]–[33]. Utility of these averaging schemes and their modifications for solving quadratic optimization and manifold problems has been examined thoroughly in recent years [34]–[36]. Moreover, several studies have suggested that accelerated first-order algorithms are more susceptible to errors in the gradient compared to their non-accelerated counterparts [26], [27], [29], [37]–[39].

One of the basic sources of error that arises in computing the gradient can be modeled by additive white stochastic noise. This source of error is typical for problems in which the gradient is being sought through measurements of a real system [40] and it has a rich history in analysis of stochastic dynamical systems and control theory [41]. Moreover, in many applications including distributed computing over networks [42], [43], coordination in vehicular formations [44], [45], and control of power systems [46]–[48], additive white noise is a convenient abstraction for the robustness analysis of distributed control strategies [43] and of first-order optimization algorithms [49], [50]. Motivated by this observation, we consider the scenario in which a white stochastic noise with zero mean and identity covariance is added to the iterates of standard first-order algorithms: gradient descent, Polyak's heavy-ball method, and Nesterov's accelerated algorithm. By focusing on smooth strongly convex problems, we use control theoretic tools to provide a tight quantitative characterization for the mean-squared error of the optimization variable. Since this quantity provides a measure of how noise gets amplified by the dynamics resulting from optimization algorithms, we also refer to it as *noise* (or *variance*) *amplification.*

1.1.2 Tradeoffs between noise amplification and convergence rate

While convergence properties of accelerated algorithms have been carefully studied [6], [9], [51]–[56], their performance and fundamental limitations in the presence of noise has received less attention [10]–[12], [57], [58]. Prior studies indicate that inaccuracies in the computation of gradient values can adversely impact the convergence rate of accelerated methods and that gradient descent may have advantages relative to its accelerated variants in noisy environments [23]–[26], [28]. In this dissertation, we consider the class of first-order methods with constant parameters in which the iterations involve information from the two previous steps. This class includes heavy-ball and Nesterov's accelerated algorithms as special cases and we examine its stochastic performance in the presence of additive white noise.

For strongly convex quadratic problems, we establish analytical lower bounds on the product of the settling time and the steady-state variance of the error in the optimization variable that hold for any constant stabilizing parameters and for both gradient and iterate noise models. Our lower bounds reveal a fundamental limitation posed by the problem condition number for this class of algorithms. Our results build upon a simple, yet powerful geometric viewpoint, which clarifies the relation between condition number, convergence rate, and algorithmic parameters for strongly convex quadratic problems. This viewpoint allows us to present alternative proofs for the optimal convergence rate of the two-step momentum algorithm [59], [60] and that of the standard gradient descent, heavy-ball method, and Nesterov's accelerated algorithm [52]. In addition, this viewpoint enables a novel geometric characterization of noise amplification in terms of stability margins and it allows us to precisely quantify tradeoffs between convergence rate and robustness to noise.

1.1.3 Transient growth of accelerated optimization algorithms

In addition to deterioration of robustness in the face of uncertainty, asymptotically stable accelerated algorithms may also exhibit undesirable transient behavior [61]. This is in contrast to gradient descent which is a contraction for strongly convex problems with suitable stepsize [62]. In real-time optimization and in applications with limited time budgets, the transient growth can limit the appeal of accelerated methods. In addition, first-order methods are often used as a building block in multi-stage optimization including ADMM [63] and distributed optimization methods [64]. In these settings, at each stage we can perform only a few iterations of first-order updates on primal or dual variables and transient growth can have a detrimental impact on the performance of the entire algorithm. This motivates an in-depth study of the behavior of accelerated first-order methods in non-asymptotic regimes. It is widely recognized that large transients may arise from the presence of resonant modal interactions and non-normality of linear dynamical generators [65]. Even in the absence of unstable modes, these can induce large transient responses, significantly amplify exogenous disturbances, and trigger departure from nominal operating conditions. For example, in fluid dynamics, such mechanisms can initiate departure from stable laminar flows and trigger transition to turbulence [66], [67].

To quantify the transient behavior of accelerated algorithms, we examine the ratio of the largest error in the optimization variable to the initial error. For convex quadratic problems, these algorithms can be cast as a linear time-invariant (LTI) system and modal analysis of

the state-transition matrix can be performed. For both accelerated algorithms, we identify non-normal modes that create large transient growth, derive analytical expressions for the state-transition matrices, and establish bounds on the transient response in terms of the convergence rate and the iteration number. We show that both the peak value of the transient response and the rise time to this value increase with the square root of the condition number of the problem. Moreover, for general strongly convex problems, we combine a Lyapunov-based approach with the theory of Integral Quadratic Constraints (IQCs) to establish similar upper bound on the transient response of Nesterov's accelerated algorithm.

1.1.4 Noise amplification of primal-dual gradient flow dynamics based on proximal augmented Lagrangian

We consider a class of primal-dual gradient flow dynamics based on proximal augmented Lagrangian [68] that can be used for solving large-scale nonsmooth constrained optimization problems in continuous time. These problems arise in many areas e.g. signal processing [69], statistical estimation [70], and control [71]. In addition, primal-dual methods have received renewed attention due to their prevalent application in distributed optimization [72] and their convergence and stability properties have been greatly studied [73]–[79]. While gradient-based methods are not readily applicable to nonsmooth optimization, we can utilize their proximal variants to address such problems [80]. In the context of nonsmooth constrained optimization, proximal-based extensions of primal-dual methods can also be obtained using the augmented Lagrangian [68], which preserve structural separability and remain suitable for distributed optimization.

We extend our analysis of noise amplification to the primal-dual flow subject to additive white noise. We examine the mean-squared error of the primal optimization variable as a measure of how noise gets amplified by the dynamics. For convex quadratic optimization problems, the primal-dual flow becomes a linear time invariant system, for which the noise amplification can be characterized using Lyapunov equations. For non-quadratic problems, the flow is no longer linear, however, tools from robust control theory can be utilized to quantify upper bounds on the noise amplification. In particular, we use IQCs [81], [82] to characterize upper bounds on the noise amplification of the primal-dual flow based on proximal augmented Lagrangian using solutions to a certain linear matrix inequality. Our results establish tight upper-upper bounds on the noise amplification that are inversely proportional to the strong-convexity module of the corresponding objective function.

1.1.5 Gradient methods for model-free linear quadratic regulator

In many emerging applications, control-oriented models are not readily available and classical approaches from optimal control may not be directly applicable. This challenge has led to the emergence of Reinforcement Learning (RL) approaches that often perform well in practice. Examples include learning complex locomotion tasks via neural network dynamics [18] and playing Atari games based on images using deep-RL [19]. In spite of the empirical success of RL in a variety of domains, our mathematical understanding of it is still in its infancy and there are many open questions surrounding convergence and sample complexity. In this

dissertation, we take a step towards answering such questions with a focus on the infinite-horizon Linear Quadratic Regulator (LQR) for continuous-time systems.

The LQR problem is the cornerstone of control theory. The globally optimal solution can be obtained by solving the Riccati equation and efficient numerical schemes with provable convergence guarantees have been developed [83]. However, computing the optimal solution becomes challenging for large-scale problems, when prior knowledge is not available, or in the presence of structural constraints on the controller. This motivates the use of direct search methods for controller synthesis. Unfortunately, the nonconvex nature of this formulation complicates the analysis of first- and second-order optimization algorithms. To make matters worse, structural constraints on the feedback gain matrix may result in a disjoint search landscape limiting the utility of conventional descent-based methods [84]. Furthermore, in the model-free setting, the exact model (and hence the gradient of the objective function) is unknown so that only zeroth-order methods can be used.

We study the sample complexity and convergence of random search method for the infinite-horizon LQR problem. For the continuous-time LQR, we employ a standard convex reparameterization [85], [86] to establish exponential stability of the ODE that governs the gradient-flow dynamics over the set of stabilizing feedback gains, and linear convergence of the gradient descent algorithm with a suitable stepsize for the nonconvex formulation. In the model-free setting, we also examine convergence and sample complexity of the random search method [22] that attempts to emulate the behavior of gradient descent via gradient approximations resulting from objective function values. For the *discrete-time* LQR, global convergence guarantees were recently provided in [13] for gradient decent and the random search method with one-point gradient estimates. For the two-point gradient estimation setting, we prove linear convergence of the random search method and show that the total number of function evaluations and the simulation time required in our results scale with the logarithm of the inverse of the desired accuracy in both continuous and discrete time.

1.1.6 Optimization landscape of the linear Quadratic Gaussian

Among model-free RL approaches, simple random search achieves a logarithmic complexity if one can access the so-called *two-point* gradient estimates [14], [87]. These results build on the fact that the gradient descent itself achieves linear convergence for both discrete [13] and continuous-time LQR problems [88] despite lack of convexity. A key enabler for these results is the so-called gradient dominance property of the underlying optimization problem that can be used as a surrogate for strong convexity [89].

Motivated by this observation, we study the convergence of gradient descent for the Linear Quadratic Gaussian (LQG) problem with incomplete state information. The separation principle states that the solution to the LQG problem is given by an observer-based controller, which consists of a Kalman filter and the corresponding LQR solution. This problem is also closely related to the output-feedback problem for distributed control, which is known to be fundamentally more challenging than LQR. In particular, the output-feedback problem has been shown to involve an optimization domain with exponential number of connected components [84], [90]. In contrast, the standard LQG problem allows for dynamic controllers and do not impose structural constraints on the controller.

We reformulate the LQG problem as a joint optimization of the control and observer feedback gains whose domain, unlike the output feedback problem is connected. We derive analytical expressions for the gradient of the LQG cost function with respect to gain matrices and demonstrate through examples that LQG does not satisfy the gradient dominance property. In particular, we show that, in addition to the global solution, the gradient vanishes at the origin for open-loop stable systems. Our study disproves global exponential convergence of policy gradient methods for LQG.

1.2 Dissertation structure

This dissertation consists of two main parts that each focuses on a specific topic and includes individual chapters that study relevant subjects. Each chapter is self contained in that it provides introduction, preliminaries and background material, problem formulation, methodology, technical results, and concluding remarks. Proofs of technical results are relegated to the corresponding appendices.

Part I

We study the stochastic performance of two-step momentum algorithms with additive white noise that accounts for uncertainty in either gradient evaluation or iteration updates. For smooth, strongly convex optimization problems, we examine the mean-squared error in the optimization variable to quantify noise amplification. By leveraging the theory of Lyapunov and integral quadratic constraints, we establish an upper bound on the noise amplification of Nesterov's method with standard parameters that is tight up to a constant factor. We also use strongly convex quadratic problems to identify fundamental tradeoffs between noise amplification and convergence rate for the two-step momentum algorithms. We use modal decomposition to introduce a novel geometric characterization of conditions for linear convergence that clarifies the relation between the noise amplification and convergence rate as well as their dependence on the condition number and the constant algorithmic parameters. This geometric insight leads to simple alternative proofs of standard convergence results and allows us to establish analytical lower bounds on the product between the settling time and noise amplification that scale quadratically with the condition number. We also characterize the impact of condition number on worst-case transient responses of popular accelerated algorithms, and examine the noise amplification of a class of primal-dual gradient flow dynamics based on proximal augmented Lagrangian that can be used for non-smooth convex constrained optimization problems.

Part II

In the second part, we focus on model-free reinforcement learning which attempts to find an optimal control action for an unknown dynamical system by directly searching over the parameter space of controllers. The convergence behavior and statistical properties of these approaches are often poorly understood because of the nonconvex nature of the underlying optimization problems and the lack of exact gradient computation. In this thesis, we take

a step towards demystifying the performance and efficiency of such methods by focusing on the standard infinite-horizon linear quadratic regulator problem with unknown state-space parameters. We establish exponential stability for the ordinary differential equation (ODE) that governs the gradient-flow dynamics over the set of stabilizing feedback gains and show that a similar result holds for the standard gradient descent. We also provide theoretical bounds on the convergence rate and sample complexity of the random search method with two-point gradient estimates. We prove that in the model-free setup, the required simulation time and the total number of function evaluations both scale with the logarithm of the inverse of the desired accuracy. The key enabler of our results is the PL condition that holds for the LQR problem both in continuous and discrete time. We finish the thesis by showing the absence of this condition for the linear quadratic Gaussian problem with incomplete state information.

1.3 Contributions of the dissertation

In this section, we provide a summary of the main contributions of each part. The chapters presented here are a reproduction of the materials that have been (or are still under review to be) published in journals and conference proceedings. We have made only some minor changes that were necessary to meet the guidelines for this document.

Part I

Noise amplification of accelerated algorithms

We study the robustness of noisy heavy-ball and Nesterov's accelerated methods for smooth, strongly convex optimization problems. Even though the underlying dynamics of these algorithms are in general nonlinear, we establish upper bounds on noise amplification that are accurate up to constant factors. For quadratic objective functions, we provide analytical expressions that quantify the effect of all eigenvalues of the Hessian matrix on variance amplification. We use these expressions to establish lower bounds demonstrating that although the acceleration techniques improve the convergence rate they significantly amplify noise for problems with large condition numbers κ. In problems of size $n \ll \kappa$, the noise amplification increases from $O(\kappa)$ to $\Omega(\kappa^{3/2})$ when moving from standard gradient descent to accelerated algorithms. We specialize our results to the problem of distributed averaging over noisy undirected networks and also study the role of network size and topology on robustness of accelerated algorithms [91]–[93].

Tradeoffs between convergence rate and noise amplification

We examine the amplification of stochastic disturbances for a class of two-step momentum algorithms in which the iterates are perturbed by an additive white noise. This class of algorithms includes Polyak's heavy-ball and Nesterov's accelerated methods as special cases and noise arises from uncertainties in gradient evaluation or in computing the iterates. For both gradient and iterate noise models, we establish lower bounds on the product of the

settling time and the smallest/largest steady-state variance of the error in the optimization variable among the class of strongly convex quadratic optimization problems. Our bounds scale quadratically with the condition number for all stabilizing parameters, which reveals a fundamental limitation imposed by the condition number in designing algorithms that tradeoff noise amplification and convergence rate. In addition, we provide a novel geometric viewpoint of stability and linear convergence. This viewpoint brings insight into the relation between noise amplification, convergence rate, and algorithmic parameters. It also allows us to (i) take an alternative approach to optimizing convergence rates for standard algorithms; (ii) identify key similarities and differences between the iterate and gradient noise models; and (iii) introduce parameterized families of algorithms for which the parameters can be continuously adjusted to tradeoff noise amplification and settling time. By utilizing positive and negative momentum parameters in accelerated and decelerated regimes, respectively, we demonstrate that a parameterized family of heavy-ball-like algorithms can achieve order-wise Pareto optimality for all settling times and both noise models. We also extend our analysis to continuous-time dynamical systems that can be discretized via an implicit-explicit Euler scheme to obtain the two-step momentum algorithm. For such gradient flow dynamics, we show that similar fundamental stochastic performance limitations hold as in discrete time [94], [95].

Transient growth of accelerated algorithms

We examine the impact of acceleration on the transient responses of popular first-order optimization algorithms. Without imposing restrictions on initial conditions, we establish bounds on the largest value of the Euclidean distance between the optimization variable and the global minimizer. For convex quadratic problems, we utilize the tools from linear systems theory to fully capture transient responses and for general strongly convex problems, we employ the theory of integral quadratic constraints to establish an upper bound on transient growth. This upper bound is proportional to the square root of the condition number and we identify quadratic problem instances for which accelerated algorithms generate transient responses which are within a constant factor of this upper bound [96]–[98].

Noise amplification of primal-dual gradient flow dynamics based on proximal augmented Lagrangian

We examine the noise amplification of proximal primal-dual gradient flow dynamics that can be used to solve non-smooth composite optimization problems. For quadratic problems, we employ algebraic Lyapunov equations to establish analytical expressions for the noise amplification. We also utilize the theory of IQCs to characterize tight upper bounds in terms of a solution to a linear matrix inequality. Our results show that stochastic performance of the primal-dual dynamics is inversely proportional to the strong-convexity module of the smooth part of the objective function [99].

Part II

Random search for continuous-time LQR

We prove exponential/linear convergence of gradient flow/descent algorithms for solving the continuous-time Linear Quadratic Regulator problem based on a nonconvex formulation that directly searches for the controller. A salient feature of our analysis is that we relate the gradient-flow dynamics associated with this nonconvex formulation to that of a convex reparameterization. This allows us to deduce convergence of the nonconvex approach from its convex counterpart. We also establish a bound on the sample complexity of the random search method for solving the continuous-time LQR problem that does not require the knowledge of system parameters. We prove that in the model-free setup with two-point gradient estimates, the required simulation time and the total number of function evaluations both scale with the logarithm of the inverse of the desired accuracy [14], [15], [88], [100], [101].

Random search for discrete-time LQR

We study the convergence and sample complexity of the random search method with two-point gradient estimates for the discrete-time LQR problem. Despite nonconvexity, we establish that the random search method with a fixed number of roll-outs per iteration that is proportional to the problem size requires a simulation time and the total number of function evaluations that scale with the logarithm of the inverse of the desired accuracy [87].

Lack of gradient domination for LQG

Motivated by the recent results on the global exponential convergence of policy gradient algorithms for the model-free LQR problem that rely on the so-called gradient dominance property, we study the standard Linear Quadratic Gaussian problem as optimization over controller and observer feedback gains. We present an explicit formula for the gradient and demonstrate that for open-loop stable systems, in addition to the unique global minimizer, the origin is also a critical point for the LQG problem, thus disproving the gradient dominance property for this class of problems [102].

Part I

Robustness of accelerated first-order optimization algorithms for strongly convex optimization problems

Chapter 2

Noise amplification of accelerated algorithms

In this chapter, we study the robustness of accelerated first-order algorithms to stochastic uncertainties in gradient evaluation. Specifically, for unconstrained, smooth, strongly convex optimization problems, we examine the mean-squared error in the optimization variable when the iterates are perturbed by additive white noise. This type of uncertainty may arise in situations where an approximation of the gradient is sought through measurements of a real system or in a distributed computation over a network. Even though the underlying dynamics of first-order algorithms for this class of problems are nonlinear, we establish upper bounds on the mean-squared deviation from the optimal solution that are tight up to constant factors. Our analysis quantifies fundamental tradeoffs between noise amplification and convergence rates obtained via *any* acceleration scheme similar to Nesterov's or heavy-ball methods. To gain additional analytical insight, for strongly convex quadratic problems, we explicitly evaluate the steady-state variance of the optimization variable in terms of the eigenvalues of the Hessian of the objective function. We demonstrate that the entire spectrum of the Hessian, rather than just the extreme eigenvalues, influence noise amplification. We specialize this result to the problem of distributed averaging over undirected networks and examine the role of network size and topology on the robustness of noisy accelerated algorithms.

2.1 Introduction

First-order methods are well suited for solving a broad range of optimization problems that arise in statistics, signal and image processing, control, and machine learning [1]–[5]. Among these algorithms, accelerated methods enjoy the optimal rate of convergence and they are popular because of their low per-iteration complexity. There is a large body of literature dedicated to the convergence analysis of these methods under different stepsize selection rules [4]–[9]. In many applications, however, the exact value of the gradient is not fully available, e.g., when the objective function is obtained via costly simulations (e.g., tuning of hyper-parameters in supervised/unsupervised learning [10]–[12] and model-free optimal control [13]–[15]), when evaluation of the objective function relies on noisy measurements (e.g., real-time and embedded applications), or when the noise is introduced via communication between different agents (e.g., distributed computation over networks). Another related application arises in the context of (batch) stochastic gradient, where at each iteration the gradient of the objective function is computed from a small batch of data points. Such a batch gradient is known to be a noisy unbiased estimator for the gradient of the training

loss. Moreover, additive noise may be introduced deliberately in the context of nonconvex optimization to help the iterates escape saddle points and improve generalization [16], [17].

In all above situations, first-order algorithms only have access to noisy estimates of the gradient. This observation has motivated the robustness analysis of these algorithms under different types of noisy/inexact gradient oracles [23]–[28]. For example, in a deterministic noise scenario, an upper bound on the error in iterates for accelerated proximal gradient methods was established in [29]. This study showed that both proximal gradient and its accelerated variant can maintain their convergence rates provided that the noise is bounded and that it vanishes fast enough. Moreover, it has been shown that in the presence of random noise, with the proper diminishing stepsize, acceleration can be achieved for general convex problems. However, in this case optimal rates are *sub-linear* [30].

In the context of stochastic approximation, while early results suggest to use a stepsize that is inversely proportional to the iteration number [24], a more robust behavior can be obtained by combining larger stepsizes with averaging [25], [31]–[33]. Utility of these averaging schemes and their modifications for solving quadratic optimization and manifold problems has been examined thoroughly in recent years [34]–[36]. Moreover, several studies have suggested that accelerated first-order algorithms are more susceptible to errors in the gradient compared to their non-accelerated counterparts [26], [27], [29], [37]–[39].

One of the basic sources of error that arises in computing the gradient can be modeled by additive white stochastic noise. This source of error is typical for problems in which the gradient is being sought through measurements of a real system [40] and it has a rich history in analysis of stochastic dynamical systems and control theory [41]. Moreover, in many applications including distributed computing over networks [42], [43], coordination in vehicular formations [44], [45], and control of power systems [46]–[48], additive white noise is a convenient abstraction for the robustness analysis of distributed control strategies [43] and of first-order optimization algorithms [49], [50]. Motivated by this observation, in this chapter we consider the scenario in which a white stochastic noise with zero mean and identity covariance is added to the iterates of standard first-order algorithms: gradient descent, Polyak's heavy-ball method, and Nesterov's accelerated algorithm. By focusing on smooth strongly convex problems, we provide a tight quantitative characterization for the mean-squared error of the optimization variable. Since this quantity provides a measure of how noise gets amplified by the dynamics resulting from optimization algorithms, we also refer to it as *noise* (or *variance*) *amplification.* We demonstrate that our quantitative characterization allows us to identify fundamental tradeoffs between the noise amplification and the rate of convergence obtained via acceleration.

This work builds on our recent conference papers [91], [92]. In a concurrent work [103], a similar approach was taken to analyze the robustness of gradient descent and Nesterov's accelerated method. Therein, it was shown that for a given convergence rate, one can select the algorithmic parameters such that the steady-state mean-squared error in the *objective value* of a Nesterov-like method becomes smaller than that of gradient descent. This is not surprising because gradient descent can be viewed as a special case of Nesterov's method with a zero momentum parameter. Using this argument, similar assertions have been made about the variance amplification of the *iterates*. This observation has been used to design an optimal multi-stage algorithm that does not require any information about the variance of the noise [104]. On the contrary, we demonstrate that there are fundamental differences

between these two robustness measures, i.e., *objective values* and *iterates*, as the former does not capture the negative impact of acceleration in the presence of noise.

By confining our attention to the error in the iterates, we show that any choice of parameters for Nesterov's or heavy-ball methods that yields an accelerated convergence rate increases variance amplification relative to gradient descent. More precisely, *for the problem with the condition number κ, an algorithm with accelerated convergence rate of at least $1 - c/\sqrt{\kappa}$, where c is a positive constant, increases the variance amplification in the iterates by a factor of $\sqrt{\kappa}$.* The robustness problem was also studied in [58] where the authors show a similar behavior of Nesterov's method and gradient descent in an asymptotic regime in which the stepsize goes to zero. In contrast, we focus on the non-asymptotic stepsize regime and establish fundamental differences between gradient descent and its accelerated variants in terms of noise amplification.

More recently, the problem of finding upper bounds on the variance amplification was cast as a semidefinite program [105]. This formulation provided numerical results that are consistent with our theoretical upper bounds in terms of the condition number. In [105], structured objective functions (e.g., diagonal Hessians) that arise in distributed optimization were also studied and the problem of designing robust algorithms were formulated as a bilinear matrix inequality (which, in general, is not convex).

Contributions

The effect of imperfections on the performance and robustness of first-order algorithms has been studied in [27], [35] but the influence of acceleration on stochastic gradient perturbations has not been precisely characterized. We employ control-theoretic tools suitable for analyzing stochastic dynamical systems to quantify such influence and identify fundamental tradeoffs between acceleration and noise amplification. The main contributions of this chapter are:

1. We start our analysis by examining strongly convex quadratic optimization problems for which we can explicitly characterize variance amplification of first-order algorithms and obtain analytical insight. In contrast to convergence rates, which solely depend on the extreme eigenvalues of the Hessian matrix, we demonstrate that the *variance amplification is influenced by the entire spectrum.*

2. We establish the relation between the noise amplification of accelerated algorithms and gradient descent for parameters that provide the optimal convergence rate for strongly convex quadratic problems. We also explain how the distribution of the eigenvalues of the Hessian influences these relations and provide examples to show that *acceleration can significantly increase the noise amplification.*

3. We address the problem of tuning the algorithm parameters and demonstrate the existence of a fundamental tradeoff between convergence rate and noise amplification: for problems with condition number κ and bounded dimension n, we show that any choice of parameters in accelerated methods that yields the linear convergence rate of at least $1 - c/\sqrt{\kappa}$, where c is a positive constant, *increases noise amplification in the iterates relative to gradient descent* by a factor of at least $\sqrt{\kappa}$.

4. We extend our analysis from quadratic objective functions to general strongly convex problems. We borrow an approach based on linear matrix inequalities from control theory to establish upper bounds on the noise amplification of both gradient descent and Nesterov's accelerated algorithm. Furthermore, for any given condition number, we demonstrate that *these bounds are tight up to constant factors.*

5. We apply our results to distributed averaging over large-scale undirected networks. We examine the role of network size and topology on noise amplification and further illustrate the subtle influence of the entire spectrum of the Hessian matrix on the robustness of noisy optimization algorithms. In particular, *we identify a class of large-scale problems for which accelerated Nesterov's method achieves the same order-wise noise amplification* (in terms of condition number) *as gradient descent.*

Chapter structure

The rest of our presentation is organized as follows. In Section 2.2, we formulate the problem and provide background material. In Section 2.3, we explicitly evaluate the variance amplification (in terms of the algorithmic parameters and problem data) for strongly convex quadratic problems, derive lower and upper bounds, and provide a comparison between the accelerated methods and gradient descent. In Section 2.4, we extend our analysis to general strongly convex problems. In Section 2.5, we establish fundamental tradeoffs between the rate of convergence and noise amplification. In Section 2.6, we apply our results to the problem of distributed averaging over noisy undirected networks. We highlight the subtle influence of the distribution of the eigenvalues of the Laplacian matrix on variance amplification and discuss the roles of network size and topology. We provide concluding remarks in Section 2.7 and technical details in appendices.

2.2 Preliminaries and background

In this chapter, we quantify the effect of stochastic uncertainties in gradient evaluation on the performance of first-order algorithms for unconstrained optimization problems

$$\underset{x}{\text{minimize}} \quad f(x) \tag{2.1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is strongly convex with Lipschitz continuous gradient ∇f. Specifically, we examine how gradient descent,

$$x^{t+1} \;=\; x^t \;-\; \alpha \nabla f(x^t) \;+\; \sigma w^t \tag{2.2a}$$

Polyak's heavy-ball method,

$$x^{t+2} \;=\; x^{t+1} \;+\; \beta(x^{t+1} \;-\; x^t) \;-\; \alpha \nabla f(x^{t+1}) \;+\; \sigma w^t \tag{2.2b}$$

and Nesterov's accelerated method,

$$x^{t+2} = \; x^{t+1} \;+\; \beta(x^{t+1} \;-\; x^t) \;-\; \alpha \nabla f(x^{t+1} \;+\; \beta(x^{t+1} \;-\; x^t)) \;+\; \sigma w^t \tag{2.2c}$$

Method	Parameters	Linear rate
Gradient	$\alpha = \frac{1}{L}$	$\rho = \sqrt{1 - \frac{2}{\kappa+1}}$
Nesterov	$\alpha = \frac{1}{L}$, $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$	$\rho = \sqrt{1 - \frac{1}{\sqrt{\kappa}}}$

Table 2.1: Conventional values of parameters and the corresponding rates for $f \in \mathcal{F}_m^L$, $\|x^t - x^\star\| \leq c\,\rho^t\,\|x^0 - x^\star\|$, where $\kappa := L/m$ and $c > 0$ is a constant [9, Theorems 2.1.15, 2.2.1]. The heavy-ball method does not offer acceleration guarantees for all $f \in \mathcal{F}_m^L$.

amplify the additive white stochastic noise w^t with zero mean and identity covariance matrix, $\mathbb{E}\,[w^t] = 0$, $\mathbb{E}\left[w^t (w^\tau)^T\right] = I\,\delta(t-\tau)$. Here, t is the iteration index, x^t is the optimization variable, α is the stepsize, β is an extrapolation parameter used for acceleration, σ is the noise magnitude, δ is the Kronecker delta, and $\mathbb{E}$ is the expected value. When the only source of uncertainty is a noisy gradient, we set $\sigma = \alpha$ in (2.2).

The set of functions f that are m-strongly convex and L-smooth is denoted by $\mathcal{F}_m^L$; $f \in \mathcal{F}_m^L$ means that $f(x) - \frac{m}{2}\|x\|^2$ is convex and that the gradient ∇f is L-Lipschitz continuous. In particular, for a twice continuously differentiable function f with the Hessian matrix $\nabla^2 f$, we have

$$f \in \mathcal{F}_m^L \;\Leftrightarrow\; mI \preceq \nabla^2 f(x) \preceq LI, \quad \forall\, x \in \mathbb{R}^n.$$

In the absence of noise (i.e., for $\sigma = 0$), for $f \in \mathcal{F}_m^L$, the parameters α and β can be selected such that gradient descent and Nesterov's accelerated method converge to the global minimum $x^\star$ of (2.1) with a linear rate $\rho < 1$, i.e.,

$$\|x^t - x^\star\| \leq c\,\rho^t\,\|x^0 - x^\star\|$$

for all t and some $c > 0$. Table 2.1 provides the conventional values of these parameters and the corresponding guaranteed convergence rates [9]. Nesterov's method with the parameters provided in Table 2.1 enjoys the convergence rate $\rho_{\text{na}} = \sqrt{1 - 1/\sqrt{\kappa}} \leq 1 - 1/(2\sqrt{\kappa})$, where $\kappa := L/m$ is the condition number associated with $\mathcal{F}_m^L$. This rate is *orderwise optimal* in the sense that no first-order algorithm can optimize all $f \in \mathcal{F}_m^L$ with the rate $\rho_{\text{lb}} = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$ [9, Theorem 2.1.13]. Note that $1 - \rho_{\text{lb}} = O(1/\sqrt{\kappa})$ and $1 - \rho_{\text{na}} = \Omega(1/\sqrt{\kappa})$. In contrast to Nesterov's method, the heavy-ball method does not offer any acceleration guarantees for all $f \in \mathcal{F}_m^L$. However, for strongly convex quadratic f, parameters can be selected to guarantee linear convergence of the heavy-ball method with a rate that outperforms the one achieved by Nesterov's method [52]; see Table 2.2.

To provide a quantitative characterization for the robustness of algorithms (2.2) to the noise w^t, we examine the performance measure,

$$J \;:=\; \limsup_{t\,\to\,\infty}\; \frac{1}{t} \sum_{k\,=\,0}^{t} \mathbb{E}\left(\|x^k - x^\star\|^2\right). \tag{2.3}$$

For quadratic objective functions, algorithms (2.2) are linear dynamical systems. In this case, J quantifies the steady-state variance amplification and it can be computed from the

solution of the algebraic Lyapunov equation; see Section 2.3. For general strongly convex problems, there is no explicit characterization for J but techniques from control theory can be utilized to compute an upper bound; see Section 2.4.

Notation

We write $g = \Omega(h)$ (or, equivalently, $h = O(g)$) to denote the existence of positive constants c_i such that, for any $x > c_2$, the functions g and h: $\mathbb{R} \to \mathbb{R}$ satisfy $g(x) \geq c_1 h(x)$. We write $g = \Theta(h)$, or more informally $g \approx h$, if both $g = \Omega(h)$ and $g = O(h)$.

2.3 Strongly convex quadratic problems

Consider a strongly convex quadratic objective function,

$$f(x) \;=\; \tfrac{1}{2}\, x^T Q\, x \;-\; q^T x \tag{2.4}$$

where Q is a symmetric positive definite matrix and q is a vector. Let $f \in \mathcal{F}_m^L$ and let the eigenvalues λ_i of Q satisfy

$$L \;=\; \lambda_1 \;\geq\; \lambda_2 \;\geq\; \ldots \;\geq\; \lambda_n \;=\; m \;>\; 0.$$

In the absence of noise, the constant values of parameters α and β provided in Table 2.2 yield linear convergence (with optimal decay rates) to the globally optimal point $x^\star = Q^{-1}q$ for all three algorithms [52]. In the presence of additive white noise w^t, we derive analytical expressions for the variance amplification J of algorithms (2.2) and demonstrate that J depends not only on the algorithmic parameters α and β but also on all eigenvalues of the Hessian matrix Q. This should be compared and contrasted to the optimal rate of linear convergence which only depends on $\kappa := L/m$, i.e., the ratio of the largest and smallest eigenvalues of Q.

For constant α and β, algorithms (2.2) can be described by a linear time-invariant (LTI) first-order recursion

$$\begin{aligned} \psi^{t+1} &= A\,\psi^t \;+\; \sigma B\, w^t \\ z^t &= C\,\psi^t \end{aligned} \tag{2.5}$$

where ψ^t is the state, $z^t := x^t - x^\star$ is the performance output, and w^t is a white stochastic input. In particular, choosing $\psi^t := x^t - x^\star$ for gradient descent and $\psi^t := [\,(x^t - x^\star)^T\; (x^{t+1} - x^\star)^T\,]^T$ for accelerated algorithms yields state-space model (2.5) with

$$A \;=\; I \;-\; \alpha\, Q\,, \quad B \;=\; C \;=\; I$$

for gradient descent and

$$A \;=\; \begin{bmatrix} 0 & I \\ -\beta I & (1+\beta)I - \alpha Q \end{bmatrix}, \quad A \;=\; \begin{bmatrix} 0 & I \\ -\beta(I - \alpha Q) & (1+\beta)(I - \alpha Q) \end{bmatrix}$$

for the heavy-ball and Nesterov's methods, respectively, with

$$B^T \;=\; \begin{bmatrix} 0 & I \end{bmatrix}, \;\; C \;=\; \begin{bmatrix} I & 0 \end{bmatrix}.$$

Since w^t is zero mean, we have $\mathbb{E}\,(\psi^{t+1}) \;=\; A\,\mathbb{E}\,(\psi^t)$. Thus, $\mathbb{E}\,(\psi^t) \;=\; A^t\,\mathbb{E}\,(\psi^0)$ and, for any stabilizing parameters α and β, $\lim_{t\to\infty}\mathbb{E}\,(\psi^t) = 0$, with the same linear rate as in the absence of noise. Furthermore, it is well-known that the covariance matrix $P^t := \mathbb{E}\left(\psi^t(\psi^t)^T\right)$ of the state vector satisfies the linear recursion

$$P^{t+1} \;=\; A\,P^tA^T \;+\; \sigma^2 BB^T \tag{2.6a}$$

and that its steady-state limit

$$P \;:=\; \lim_{t\to\infty}\;\mathbb{E}\left(\psi^t(\psi^t)^T\right) \tag{2.6b}$$

is the unique solution to the algebraic Lyapunov equation [41]

$$P \;=\; A\,PA^T \;+\; \sigma^2 BB^T. \tag{2.6c}$$

For stable LTI systems, performance measure (2.3) simplifies to the steady-state variance of the error in the optimization variable $z^t := x^t - x^\star$,

$$J \;=\; \lim_{t\to\infty}\;\frac{1}{t}\sum_{k\,=\,0}^{t}\mathbb{E}\left(\|z^k\|^2\right) \;=\; \lim_{t\to\infty}\;\mathbb{E}\left(\|z^t\|^2\right) \tag{2.6d}$$

and it can be computed using either of the following two equivalent expressions

$$J \;=\; \lim_{t\to\infty}\;\frac{1}{t}\sum_{k\,=\,0}^{t}\mathrm{trace}\left(Z^k\right) \;=\; \mathrm{trace}\,(Z) \tag{2.6e}$$

where $Z = CPC^T$ is the steady-state limit of the covariance matrix $Z^t := \mathbb{E}\left(z^t(z^t)^T\right) = CP^tC^T$ of the output z^t.

We next provide analytical solution P to (2.6c) that depends on the parameters α and β as well as on the spectrum of the Hessian matrix Q. This allows us to explicitly characterize the variance amplification J and quantify the impact of additive white noise on the performance of first-order optimization algorithms.

2.3.1 Influence of the eigenvalues of the Hessian matrix

We use the modal decomposition of the symmetric matrix $Q = V\Lambda V^T$ to bring A, B, and C in (2.5) into a block diagonal form, $\hat{A} = \mathrm{diag}\,(\hat{A}_i)$, $\hat{B} = \mathrm{diag}\,(\hat{B}_i)$, $\hat{C} = \mathrm{diag}\,(\hat{C}_i)$, with $i = 1,\ldots,n$. Here, $\Lambda = \mathrm{diag}\,(\lambda_i)$ is the diagonal matrix of the eigenvalues and V is the orthogonal matrix of the eigenvectors of Q. More specifically, the unitary coordinate transformation

$$\hat{x}^t \;:=\; V^Tx^t, \;\; \hat{x}^\star \;:=\; V^Tx^\star, \;\; \hat{w}^t \;:=\; V^Tw^t \tag{2.7}$$

Method	Optimal parameters	Rate of linear convergence
Gradient	$\alpha = \dfrac{2}{L+m}$	$\rho = \dfrac{\kappa-1}{\kappa+1}$
Nesterov	$\alpha = \dfrac{4}{3L+m}$, $\beta = \dfrac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$	$\rho = \dfrac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}}$
Heavy-ball	$\alpha = \dfrac{4}{(\sqrt{L}+\sqrt{m})^2}$, $\beta = \dfrac{(\sqrt{\kappa}-1)^2}{(\sqrt{\kappa}+1)^2}$	$\rho = \dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

Table 2.2: Optimal parameters and the corresponding convergence rates for a strongly convex quadratic objective function $f \in \mathcal{F}_m^L$ with $\lambda_{\max}(\nabla^2 f) = L$ and $\lambda_{\min}(\nabla^2 f) = m$, and $\kappa := L/m$ [52, Proposition 1].

brings the state-space model of gradient descent into a diagonal form with

$$\hat{\psi}_i^t \;=\; \hat{x}_i^t \;-\; \hat{x}_i^\star, \quad \hat{A}_i \;=\; 1 \;-\; \alpha\lambda_i, \quad \hat{B}_i \;=\; \hat{C}_i \;=\; 1. \tag{2.8a}$$

Similarly, for Polyak's heavy-ball and Nesterov's accelerated methods, we can use the change of coordinates (2.7) in conjunction with a permutation of variables, $\hat{\psi}_i^t = [\,\hat{x}_i^t - \hat{x}_i^\star \;\; \hat{x}_i^{t+1} - \hat{x}_i^\star\,]^T$, respectively to obtain

$$\hat{A}_i \;=\; \begin{bmatrix} 0 & 1 \\ -\beta & 1+\beta-\alpha\lambda_i \end{bmatrix}, \;\; \hat{B}_i \;=\; \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \;\; \hat{C}_i \;=\; \begin{bmatrix} 1 & 0 \end{bmatrix} \tag{2.8b}$$

$$\hat{A}_i \;=\; \begin{bmatrix} 0 & 1 \\ -\beta(1-\alpha\lambda_i) & (1+\beta)(1-\alpha\lambda_i) \end{bmatrix}, \;\; \hat{B}_i \;=\; \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \;\; \hat{C}_i \;=\; \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{2.8c}$$

This block diagonal structure allows us to explicitly solve Lyapunov equation (2.6c) for P and derive an analytical expression for J in terms of the eigenvalues λ_i of the Hessian matrix Q and the algorithmic parameters α and β. Namely, under coordinate transformation (2.7) and a suitable permutation of variables, equation (2.6c) can be brought into an equivalent set of equations,

$$\hat{P}_i \;=\; \hat{A}_i\,\hat{P}_i\,\hat{A}_i^T \;+\; \sigma^2\hat{B}_i\hat{B}_i^T, \quad i \;=\; 1,\ldots,n \tag{2.9}$$

where $\hat{P}_i$ is a scalar for the gradient descent method and a 2×2 matrix for the accelerated algorithms. In Theorem 1, we use the solution to these decoupled Lyapunov equations to express the variance amplification as

$$J \;=\; \sum_{i=1}^{n} \hat{J}(\lambda_i) \;:=\; \sum_{i=1}^{n} \operatorname{trace}\left(\hat{C}_i\hat{P}_i\hat{C}_i^T\right)$$

where $\hat{J}(\lambda_i)$ determines the contribution of the eigenvalue λ_i of the matrix Q to the variance amplification. In what follows, we use subscripts gd, hb, and na (e.g., J_{gd}, J_{hb}, and J_{na}) to denote quantities that correspond to gradient descent (2.2a), heavy-ball method (2.2b), and Nesterov's accelerated method (2.2c).

Theorem 1 *For strongly convex quadratic problems, the variance amplification of noisy first-order algorithms* (2.2) *with any constant stabilizing parameters* α *and* β *is determined by* $J = \sum_{i=1}^{n} \hat{J}(\lambda_i)$*, where* λ_i *is the* i*th eigenvalue of* $Q = Q^T \succ 0$ *and the modal contribution to the variance amplification* $\hat{J}(\lambda)$ *is given by*

$$\begin{aligned} \textit{Gradient:} \quad \hat{J}_{\mathrm{gd}}(\lambda) &= \frac{\sigma^2}{\alpha\lambda\,(2 - \alpha\lambda)} \\ \textit{Polyak:} \quad \hat{J}_{\mathrm{hb}}(\lambda) &= \frac{\sigma^2(1 + \beta)}{\alpha\lambda\,(1 - \beta)\,(2(1 + \beta) - \alpha\lambda)} \\ \textit{Nesterov:} \quad \hat{J}_{\mathrm{na}}(\lambda) &= \frac{\sigma^2(1 + \beta(1 - \alpha\lambda))}{\alpha\lambda\,(1 - \beta(1 - \alpha\lambda))\,(2(1 + \beta) - (2\beta + 1)\alpha\lambda)}. \end{aligned}$$

Proof: See Appendix A.1. □

For strongly convex quadratic problems, Theorem 1 provides *exact expressions* for variance amplification of the first-order algorithms. In addition to quantifying the dependence of J on the algorithmic parameters α and β and the impact of the largest and smallest eigenvalues, these expressions capture the effect of all other eigenvalues of the Hessian matrix Q. We also observe that the variance amplification J is proportional to σ^2. Apart from Section 2.5, where we examine the role of parameters α and β on acceleration/robustness tradeoff and allow the dependence of σ on α, without loss of generality we choose $\sigma = 1$ in the rest of the chapter.

Remark 1 *The performance measure* J *in* (2.6d) *quantifies the steady-state variance of the iterates of first-order algorithms. Robustness of noisy algorithms can be also evaluated using alternative performance measures, e.g., the mean value of the error in the objective function [103],*

$$J' \;=\; \lim_{t\to\infty} \mathbb{E}\left((x^t - x^\star)^T Q\,(x^t - x^\star)\right). \tag{2.10}$$

This measure of variance amplification can be characterized using our approach by defining $C = Q^{1/2}$ *for gradient descent and* $C = [\,Q^{1/2}\;\; 0\,]$ *for accelerated algorithms in state-space model* (2.5)*. Furthermore, repeating the above procedure for the modified performance output* z^t *yields* $J' = \sum_{i=1}^{n} \lambda_i \hat{J}(\lambda_i)$*, where the respective expressions for* $\hat{J}(\lambda_i)$ *are given in Theorem 1.*

2.3.2 Comparison for parameters that optimize convergence rate

We next examine the robustness of first-order algorithms applied to strongly convex quadratic problems for the parameters that optimize the linear convergence rate; see Table 2.2. For these parameters, the eigenvalues of the matrix A are inside the open unit disk, which implies exponential stability of system (2.5). We first use the expressions presented in Theorem 1 to compare the variance amplification of the heavy-ball method to gradient descent.

Theorem 2 *Let the strongly convex quadratic objective function* f *in* (2.4) *satisfy* $\lambda_{\max}(Q) = L$*,* $\lambda_{\min}(Q) = m > 0$*, and let* $\kappa := L/m$ *be the condition number. For the optimal parameters*

provided in Table 2.2, the ratio between the variance amplification of the heavy-ball method and gradient descent with equal values of σ is given by

$$\frac{J_{\mathrm{hb}}}{J_{\mathrm{gd}}} \;=\; \frac{(\sqrt{\kappa}\,+\,1)^4}{8\sqrt{\kappa}\,(\kappa\,+\,1)}. \tag{2.11}$$

Proof: For the parameters provided in Table 2.2 we have $\alpha_{\mathrm{hb}} = (1+\beta)\alpha_{\mathrm{gd}}$, where $\beta = (\sqrt{\kappa}-1)^2/(\sqrt{\kappa}+1)^2$ is the momentum parameter for the heavy-ball method. It is now straightforward to show that the modal contributions $\hat{J}_{\mathrm{hb}}$ and $\hat{J}_{\mathrm{gd}}$ to the variance amplification of the iterates given in Theorem 1 satisfy

$$\frac{\hat{J}_{\mathrm{hb}}(\lambda)}{\hat{J}_{\mathrm{gd}}(\lambda)} \;=\; \frac{1}{1\,-\,\beta^2} \;=\; \frac{(\sqrt{\kappa}\,+\,1)^4}{8\sqrt{\kappa}\,(\kappa\,+\,1)}, \quad \forall\,\lambda\,\in\,[m,L]. \tag{2.12}$$

Thus, *the ratio $\hat{J}_{\mathrm{hb}}(\lambda)/\hat{J}_{\mathrm{gd}}(\lambda)$ does not depend on λ and is only a function of the condition number κ.* Substitution of (2.12) into $J = \sum_i \hat{J}(\lambda_i)$ yields relation (2.11). □

Theorem 2 establishes the linear relation between the variance amplification of the heavy-ball algorithm J_{hb} and the gradient descent J_{gd}. We observe that the ratio $J_{\mathrm{hb}}/J_{\mathrm{gd}}$ *only* depends on the condition number κ and that *acceleration increases variance amplification*: for $\kappa \gg 1$, J_{hb} is larger than J_{gd} by a factor of $\sqrt{\kappa}$. We next study the ratio between the variance amplification of Nesterov's accelerated method and gradient descent. In contrast to the heavy-ball method, this ratio depends on the entire spectrum of the Hessian matrix Q. The following proposition, which examines the modal contributions $\hat{J}_{\mathrm{na}}(\lambda)$ and $\hat{J}_{\mathrm{gd}}(\lambda)$ of Nesterov's accelerated method and gradient descent, is the key technical result that allows us to establish the largest and smallest values that the ratio $J_{\mathrm{na}}/J_{\mathrm{gd}}$ can take for a given pair of extreme eigenvalues m and L of Q in Theorem 3.

Proposition 1 *Let the strongly convex quadratic function f in (2.4) satisfy $\lambda_{\max}(Q) = L$, $\lambda_{\min}(Q) = m > 0$, and let $\kappa := L/m$ be the condition number. For the optimal parameters provided in Table 2.2, the ratio $\hat{J}_{\mathrm{na}}(\lambda)/\hat{J}_{\mathrm{gd}}(\lambda)$ of modal contributions to variance amplification of Nesterov's method and gradient descent is a decreasing function of $\lambda \in [m,L]$. Furthermore, for $\sigma = 1$, the function $\hat{J}_{\mathrm{gd}}(\lambda)$ satisfies*

$$\begin{aligned} \max_{\lambda\,\in\,[m,L]}\; \hat{J}_{\mathrm{gd}}(\lambda) \;&=\; \hat{J}_{\mathrm{gd}}(m) \;\;=\; \hat{J}_{\mathrm{gd}}(L) \;=\; \frac{(\kappa\,+\,1)^2}{4\kappa} \\ \min_{\lambda\,\in\,[m,L]}\; \hat{J}_{\mathrm{gd}}(\lambda) \;&=\; \hat{J}_{\mathrm{gd}}(1/\alpha) \;=\; 1 \end{aligned} \tag{2.13a}$$

and the function $\hat{J}_{\rm na}(\lambda)$ *satisfies*

$$\begin{aligned} \hat{J}_{\rm na}(L) &= \frac{9\,\bar{\kappa}^2\big(\bar{\kappa}\,+\,2\sqrt{\bar{\kappa}}\,-2\,\big)}{32\,(\bar{\kappa}\,-\,1)\big(\bar{\kappa}\,-\,\sqrt{\bar{\kappa}}\,+\,1\big)\big(2\sqrt{\bar{\kappa}}\,-\,1\big)} \\ \max_{\lambda\,\in\,[m,L]}\ \hat{J}_{\rm na}(\lambda) \,=\, \hat{J}_{\rm na}(m) &= \frac{\bar{\kappa}^2\big(\bar{\kappa}\,-\,2\sqrt{\bar{\kappa}}\,+\,2\big)}{32\big(\sqrt{\bar{\kappa}}\,-\,1\big)^3} \\ \min_{\lambda\,\in\,[m,L]}\ \hat{J}_{\rm na}(\lambda) \,=\, \hat{J}_{\rm na}(1/\alpha) &= 1 \end{aligned} \tag{2.13b}$$

where $\bar{\kappa} := 3\kappa + 1$.

Proof: See Appendix A.1. □

For all three algorithms, Proposition 1 and Theorem 2 demonstrate that the modal contribution to the variance amplification of the iterates at the extreme eigenvalues of the Hessian matrix m and L only depends on the condition number $\kappa := L/m$. For gradient descent and the heavy-ball method, $\hat{J}$ achieves its largest value at m and L, i.e.,

$$\begin{aligned} \max_{\lambda\,\in\,[m,L]}\ \hat{J}_{\rm gd}(\lambda) \,=\, \hat{J}_{\rm gd}(m) \,&=\, \hat{J}_{\rm gd}(L) \,=\, \Theta(\kappa) \\ \max_{\lambda\,\in\,[m,L]}\ \hat{J}_{\rm hb}(\lambda) \,=\, \hat{J}_{\rm hb}(m) \,&=\, \hat{J}_{\rm hb}(L) \,=\, \Theta(\kappa\sqrt{\kappa}). \end{aligned} \tag{2.14a}$$

On the other hand, for Nesterov's method, (2.13b) implies a gap of $\Theta(\kappa)$ between the boundary values

$$\max_{\lambda\,\in\,[m,L]}\ \hat{J}_{\rm na}(\lambda) \,=\, \hat{J}_{\rm na}(m) \,=\, \Theta(\kappa\sqrt{\kappa}),\quad \hat{J}_{\rm na}(L) \,=\, \Theta(\sqrt{\kappa}). \tag{2.14b}$$

Remark 2 *Theorem 1 provides explicit formulas for the variance amplification of noisy algorithms* (2.2) *in terms of the eigenvalues* λ_i *of the Hessian matrix* Q*. Similarly, we can represent the variance amplification in terms of the eigenvalues* $\hat{\lambda}_i$ *of the dynamic matrices* $\hat{A}_i$ *in* (2.8)*. For gradient descent,* $\hat{\lambda}_i = 1 - \alpha\lambda_i$ *and it is straightforward to verify that* $J_{\rm gd}$ *is determined by the sum of reciprocals of distances of these eigenvalues to the stability boundary,* $J_{\rm gd} = \sum_{i\,=\,1}^{n} \sigma^2/(1-\hat{\lambda}_i^2)$*. Similarly, for accelerated methods we have,*

$$J \,=\, \sum_{i\,=\,1}^{n} \frac{\sigma^2(1+\hat{\lambda}_i\hat{\lambda}_i')}{(1-\hat{\lambda}_i\hat{\lambda}_i')(1-\hat{\lambda}_i)(1-\hat{\lambda}_i')(1+\hat{\lambda}_i)(1+\hat{\lambda}_i')}$$

where $\hat{\lambda}_i$ *and* $\hat{\lambda}_i'$ *are the eigenvalues of* $\hat{A}_i$*. For Nesterov's method with the parameters provided in Table 2.2, the matrix* $\hat{A}_n$*, which corresponds to* $\lambda_n = m$*, admits a Jordan canonical form with repeated eigenvalues* $\hat{\lambda}_n = \hat{\lambda}_n' = 1 - 2/\sqrt{3\kappa+1}$*. In this case,* $\hat{J}_{\rm na}(m) = \sigma^2(1+\hat{\lambda}_n^2)/(1-\hat{\lambda}_n^2)^3$*, which should be compared and contrasted to the above expression for gradient descent. Furthermore, for both* $\lambda_1 = L$ *and* $\lambda_n = m$*, the matrices* $\hat{A}_1$ *and* $\hat{A}_n$ *for the heavy-ball method with the parameters provided in Table 2.2 have eigenvalues with algebraic multiplicity two and incomplete sets of eigenvectors.*

We next establish the range of values that the ratio $J_{\rm na}/J_{\rm gd}$ can take.

Theorem 3 *For the strongly convex quadratic objective function f in (2.4) with $x \in \mathbb{R}^n$, $\lambda_{\max}(Q) = L$, and $\lambda_{\min}(Q) = m > 0$, the ratio between the variance amplification of Nesterov's accelerated method and gradient descent, for the optimal parameters provided in Table 2.2 and equal values of σ satisfies*

$$\frac{\hat{J}_{\mathrm{na}}(m) + (n-1)\hat{J}_{\mathrm{na}}(L)}{\hat{J}_{\mathrm{gd}}(m) + (n-1)\hat{J}_{\mathrm{gd}}(L)} \leq \frac{J_{\mathrm{na}}}{J_{\mathrm{gd}}} \leq \frac{\hat{J}_{\mathrm{na}}(L) + (n-1)\hat{J}_{\mathrm{na}}(m)}{\hat{J}_{\mathrm{gd}}(L) + (n-1)\hat{J}_{\mathrm{gd}}(m)}. \tag{2.15}$$

Proof: See Appendix A.1. □

Theorem 3 provides tight upper and lower bounds on the ratio between J_{na} and J_{gd} for strongly convex quadratic problems. As shown in Appendix A.1, the lower bound is achieved for a quadratic function in which the Hessian matrix Q has one eigenvalue at m and $n-1$ eigenvalues at L, and the upper bound is achieved when Q has one eigenvalue at L and the remaining ones at m. Theorem 3 in conjunction with Proposition 1 demonstrate that *for a fixed problem dimension n, J_{na} is larger than J_{gd} by a factor of $\sqrt{\kappa}$ for $\kappa \gg 1$.*

This tradeoff is further highlighted in Theorem 4 which provides tight bounds on the variance amplification of iterates in terms of the problem dimension n and the condition number κ for all three algorithms. To simplify the presentation, we first use the explicit expressions for $\hat{J}_{\mathrm{na}}(m)$ and $\hat{J}_{\mathrm{na}}(L)$ in Proposition 1 to obtain the following upper and lower bounds on $\hat{J}_{\mathrm{na}}(m)$ and $\hat{J}_{\mathrm{na}}(L)$ (see Appendix A.1)

$$\frac{(3\kappa+1)^{\frac{3}{2}}}{32} \leq \hat{J}_{\mathrm{na}}(m) \leq \frac{(3\kappa+1)^{\frac{3}{2}}}{8}, \qquad \frac{9\sqrt{3\kappa+1}}{64} \leq \hat{J}_{\mathrm{na}}(L) \leq \frac{9\sqrt{3\kappa+1}}{8}. \tag{2.16}$$

Theorem 4 *For the strongly convex quadratic objective function f in (2.4) with $x \in \mathbb{R}^n$, $\lambda_{\max}(Q) = L$, $\lambda_{\min}(Q) = m > 0$, and $\kappa := L/m$, the variance amplification of the first-order optimization algorithms, with the parameters provided in Table 2.2 and $\sigma = 1$, is bounded by*

$$\begin{aligned}
\frac{(\kappa-1)^2}{2\kappa} + n \leq\ & J_{\mathrm{gd}} \leq \frac{n(\kappa+1)^2}{4\kappa} \\
\frac{(\sqrt{\kappa}+1)^4}{8\sqrt{\kappa}(\kappa+1)}\left(\frac{(\kappa-1)^2}{2\kappa}+n\right) \leq\ & J_{\mathrm{hb}} \leq \frac{n(\kappa+1)(\sqrt{\kappa}+1)^4}{32\kappa\sqrt{\kappa}} \\
\frac{(3\kappa+1)^{\frac{3}{2}}}{32} + \frac{9\sqrt{3\kappa+1}}{64} + n - 2 \leq\ & J_{\mathrm{na}} \leq \frac{(n-1)(3\kappa+1)^{\frac{3}{2}}}{8} + \frac{9\sqrt{3\kappa+1}}{8}.
\end{aligned}$$

Proof: As shown in Proposition 1, the functions $\hat{J}(\lambda)$ for gradient descent and Nesterov's algorithm attain their largest and smallest values over the interval $[m, L]$ at $\lambda = m$ and $\lambda = 1/\alpha$, respectively. Thus, fixing the smallest and largest eigenvalues, the variance amplification J is maximized when the other $n-2$ eigenvalues are all equal to m and is minimized when they are all equal to $1/\alpha$. This combined with the explicit expressions for $\hat{J}_{\mathrm{gd}}(m)$, $\hat{J}_{\mathrm{gd}}(L)$, and $\hat{J}_{\mathrm{gd}}(1/\alpha)$ in (2.13a) leads to the tight upper and lower bounds for gradient descent. For the heavy-ball method, the bounds follow from Theorem 2 and for Nesterov's algorithm, the bounds follow from (2.16). □

For problems with a fixed dimension n and a condition number $\kappa \gg n$, there is an $\Omega(\sqrt{\kappa})$ difference in both upper and lower bounds provided in Theorem 4 for the accelerated algorithms relative to gradient descent. Even though Theorem 4 considers only the values of α and β that optimize the convergence rate, in Section 2.5 we demonstrate that this gap is fundamental in that it holds for any parameters that yield an accelerated convergence rate. It is worth noting that both the lower and upper bounds are influenced by the problem dimension n and the condition number κ. For large-scale problems, there may be a subtle relation between n and κ and the established bounds may exhibit different scaling trends. In Section 2.6, we identify a class of quadratic optimization problems for which $J_{\rm na}$ scales in the same way as $J_{\rm gd}$ for $\kappa \gg 1$ and $n \gg 1$.

Before we elaborate further on these issues, we provide two illustrative examples that highlight the importance of the choice of the performance metric in the robustness analysis of noisy algorithms. It is worth noting that an $O(\kappa)$ upper bound for gradient descent and an $O(\kappa^2)$ upper bound for Nesterov's accelerated algorithm was established in [29]. Relative to this upper bound for Nesterov's method, the upper bound provided in Theorem 4 is tighter by a factor of $\sqrt{\kappa}$. Theorem 4 also provides lower bounds, reveals the influence of the problem dimension n, and identifies constants that multiply the leading terms in the condition number κ. Moreover, in Section 2.4 we demonstrate that similar upper bounds can be obtained for general strongly convex objective functions with Lipschitz continuous gradients.

2.3.3 Examples

We next provide illustrative examples to (i) demonstrate the agreement of our theoretical predictions with the results of stochastic simulations; and (ii) contrast two relevant measures of performance, namely the variance of the iterates J in (2.6d) and the mean objective error J' in (2.10), for assessing robustness of noisy optimization algorithms.

Example 1

Let us consider the quadratic objective function in (2.4) with

$$Q \;=\; \begin{bmatrix} L & 0 \\ 0 & m \end{bmatrix}, \quad q \;=\; \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{2.17}$$

For all three algorithms, the performance measures J and J' are given by

$$\begin{aligned} J \;&=\; \hat{J}(m) \;+\; \hat{J}(L) \\ J' \;&=\; m\hat{J}(m) \;+\; L\hat{J}(L) \;=\; L\left(\tfrac{1}{\kappa}\,\hat{J}(m) \;+\; \hat{J}(L)\right) \;=\; m\left(\hat{J}(m) \;+\; \kappa\,\hat{J}(L)\right). \end{aligned}$$

As shown in (2.14), $\hat{J}(m)$ and $\hat{J}(L)$ only depend on the condition number κ and the variance amplification of the iterates satisfies

$$J_{\rm gd} \;=\; \Theta(\kappa), \quad J_{\rm hb} \;=\; \Theta(\kappa\sqrt{\kappa}), \quad J_{\rm na} \;=\; \Theta(\kappa\sqrt{\kappa}). \tag{2.18a}$$

On the other hand, J' also depends on m and L. In particular, it is easy to verify the following relations for two scenarios that yield $\kappa \gg 1$:

- for $m \ll 1$ and $L = O(1)$

$$J'_{\rm gd} \;=\; \Theta(\kappa), \quad J'_{\rm hb} \;=\; \Theta(\kappa\sqrt{\kappa}), \quad J'_{\rm na} \;=\; \Theta(\sqrt{\kappa}). \tag{2.18b}$$

- for $L \gg 1$ and $m = O(1)$

$$J'_{\rm gd} \;=\; \Theta(\kappa^2), \quad J'_{\rm hb} \;=\; \Theta(\kappa^2\sqrt{\kappa}), \quad J'_{\rm na} \;=\; \Theta(\kappa\sqrt{\kappa}). \tag{2.18c}$$

Relation (2.18a) reveals the detrimental impact of acceleration on the variance of the optimization variable. On the other hand, (2.18b) and (2.18c) show that, relative to gradient descent, the heavy-ball method increases the mean error in the objective function while Nesterov's method reduces it. Thus, if the mean value of the error in the objective function is to be used to assess performance of noisy algorithms, one can conclude that Nesterov's method significantly outperforms gradient descent both in terms of convergence rate and robustness to noise. However, this performance metric fails to capture large variance of the mode associated with the smallest eigenvalue of the matrix Q in Nesterov's algorithm. Theorem 2 and Proposition 1 show that the modal contributions to the variance amplification of the iterates for gradient descent and the heavy-ball method are balanced at m and L, i.e., $\hat{J}_{\rm gd}(m) = \hat{J}_{\rm gd}(L) = \Theta(\kappa)$ and $\hat{J}_{\rm hb}(m) = \hat{J}_{\rm hb}(L) = \Theta(\kappa\sqrt{\kappa})$. On the other hand, for Nesterov's method there is a $\Theta(\kappa)$ gap between $\hat{J}_{\rm na}(m) = \Theta(\kappa\sqrt{\kappa})$ and $\hat{J}_{\rm na}(L) = \Theta(\sqrt{\kappa})$. While the performance measure J' reveals a superior performance of Nesterov's algorithm at large condition numbers, it fails to capture the negative impact of acceleration on the variance of the optimization variable; see Fig. 2.1 for an illustration.

Figure 2.2 shows the performance outputs $z^t = x^t$ and $z^t = Q^{1/2}x^t$ resulting from 10^5 iterations of noisy first-order algorithms with the optimal parameters provided in Table 2.2 for the strongly convex objective function $f(x) = 0.5\,x_1^2 + 0.25 \times 10^{-4}\,x_2^2$ ($\kappa = 2 \times 10^4$). Although Nesterov's method exhibits good performance with respect to the error in the objective function (performance measure J'), the plots in the first row illustrate detrimental impact of noise on both accelerated algorithms with respect to the variance of the iterates (performance measure J). In particular, we observe that: (i) for gradient descent and the heavy-ball method, the iterates x^t are scattered uniformly along the eigen-directions of the Hessian matrix Q and acceleration increases variance equally along all directions; and (ii) relative to gradient descent, Nesterov's method exhibits larger variance in the iterates x^t along the direction that corresponds to the smallest eigenvalue $\lambda_{\min}(Q)$.

Example 2

Figure 2.3 compares the results of twenty stochastic simulations for a strongly convex quadratic objective function (2.4) with $q = 0$ and a Toeplitz matrix $Q \in \mathbb{R}^{50\times 50}$ with the first row $[\,2\;\;-1\;\;0\;\;\cdots\;\;0\;\;0\,]^T$. This figure illustrates the dependence of the variance of the performance outputs $z^t = x^t$ and $z^t = Q^{1/2}x^t$ on time t for the algorithms subject to additive white noise with zero initial conditions. The plots further demonstrate that the mean error in the objective function does not capture the detrimental impact of noise on the variance of the iterates for Nesterov's algorithm. The bottom row also compares variance obtained

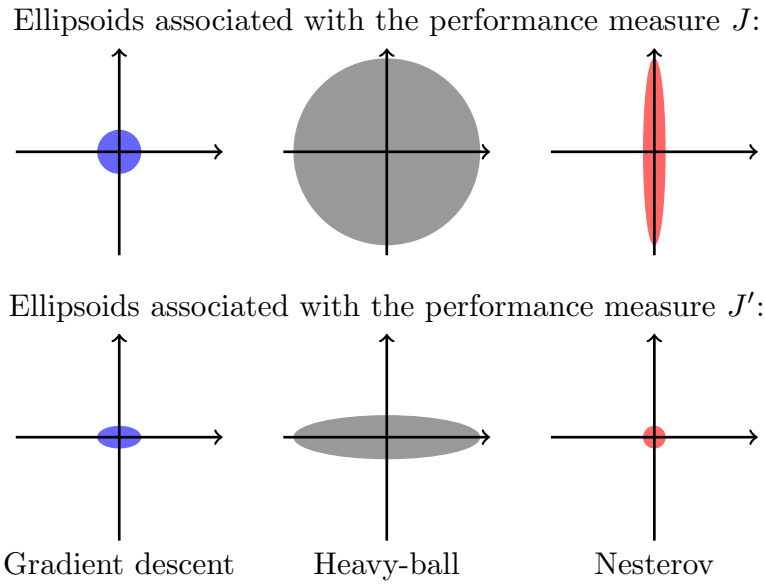


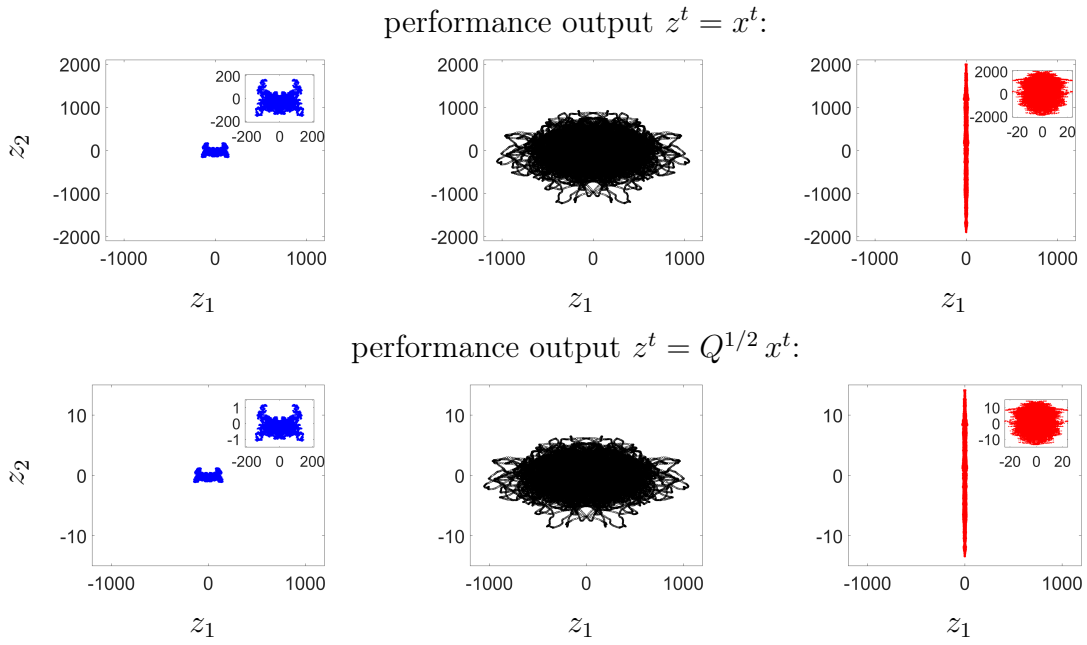
Figure 2.1: Ellipsoids $\{z \,|\, z^T Z^{-1} z \leq 1\}$ associated with the steady-state covariance matrices $Z = CPC^T$ of the performance outputs $z^t = x^t - x^\star$ (top row) and $z^t = Q^{1/2}(x^t - x^\star)$ (bottom row) for algorithms (2.2) with the parameters provided in Table 2.2 for the matrix Q given in (2.17) with $m \ll L = O(1)$. The horizontal and vertical axes show the eigenvectors $[\,1\;\;0\,]^T$ and $[\,0\;\;1\,]^T$ associated with the eigenvalues $\hat{J}(L)$ and $\hat{J}(m)$ (top row) and $\hat{J}'(L)$ and $\hat{J}'(m)$ (bottom row) of the respective output covariance matrices Z.

by averaging outcomes of twenty stochastic simulations with the corresponding theoretical values resulting from the Lyapunov equations.

2.4 General strongly convex problems

In this section, we extend our results to the class $\mathcal{F}_m^L$ of m-strongly convex objective functions with L-Lipschitz continuous gradients. While a precise characterization of noise amplification for general problems is challenging because of the nonlinear dynamics, we employ tools from robust control theory to obtain meaningful upper bounds. Our results utilize the theory of integral quadratic constraints [81], a convex control-theoretic framework that was recently used to analyze optimization algorithms [52] and study convergence and robustness of the first-order methods [53], [54], [56], [68]. We establish analytical upper bounds on the mean-squared error of the iterates (2.3) for gradient descent (2.2a) and Nesterov's accelerated (2.2c) methods. Since there are no known accelerated convergence guarantees for the heavy-ball method when applied to general strongly convex functions, we do not consider it in this section.

We first exploit structural properties of the gradient and employ quadratic Lyapunov functions to formulate a semidefinite programing problem (SDP) that provides upper bounds



(a) Gradient descent (b) Heavy-ball (c) Nesterov

Figure 2.2: Performance outputs $z^t = x^t$ (top row) and $z^t = Q^{1/2}x^t$ (bottom row) resulting from 10^5 iterations of noisy first-order algorithms (2.2) with the parameters provided in Table 2.2. Strongly convex problem with $f(x) = 0.5\,x_1^2 + 0.25 \times 10^{-4}\,x_2^2$ ($\kappa = 2 \times 10^4$) is solved using algorithms with additive white noise and zero initial conditions.

on J in (2.3). While quadratic Lyapunov functions yield tight upper bounds for gradient descent, they fail to provide any upper bound for Nesterov's method for large condition numbers ($\kappa > 100$). To overcome this challenge, we present a modified semidefinite program that uses more general Lyapunov functions which are obtained by augmenting standard quadratic terms with the objective function. This type of generalized Lyapunov functions has been introduced in [56], [106] and used to study convergence of optimization algorithms for non-strongly convex problems. We employ a modified SDP to derive meaningful upper bounds on J in (2.3) for Nesterov's method as well.

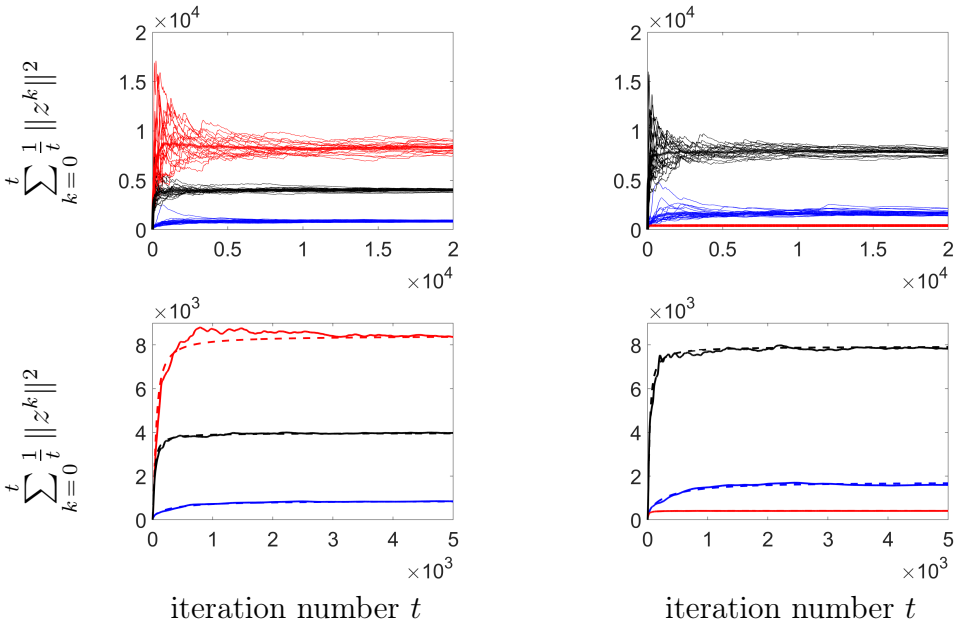
We note that algorithms (2.2) are invariant under translation, i.e., if we let $\tilde{x} := x - \bar{x}$ and $g(\tilde{x}) := f(\tilde{x} + \bar{x})$, then (2.2c), for example, satisfies

$$\tilde{x}^{t+2} = \tilde{x}^{t+1} + \beta(\tilde{x}^{t+1} - \tilde{x}^t) - \alpha\nabla g(\tilde{x}^{t+1} + \beta(\tilde{x}^{t+1} - \tilde{x}^t)) + \sigma w^t.$$

Thus, in what follows, without loss of generality, we assume that $x^\star = 0$ is the unique minimizer of (2.1).

2.4.1 An approach based on contraction mappings

Before we present our approach based on Linear Matrix Inequalities (LMIs), we provide a more intuitive approach that can be used to examine noise amplification of gradient descent. Let φ: $\mathbb{R}^n \to \mathbb{R}^n$ be a contraction mapping, i.e., there exists a positive scalar $\eta < 1$ such



(a) performance output $z^t = x^t$ (b) performance output $z^t = Q^{1/2}x^t$

Figure 2.3: $(1/t)\sum_{k=0}^{t}\|z^k\|^2$ for the performance output z^t in Example 2. Top row: the thick blue (gradient descent), black (heavy-ball), and red (Nesterov's method) lines mark variance obtained by averaging results of twenty stochastic simulations. Bottom row: comparison between results obtained by averaging outcomes of twenty stochastic simulations (thick lines) with the corresponding theoretical values $(1/t)\sum_{k=0}^{t}\operatorname{trace}(CP^kC^T)$ (dashed lines) resulting from the Lyapunov equation (2.6a).

that $\|\varphi(x)-\varphi(y)\| \le \eta\|x-y\|$ for all $x, y \in \mathbb{R}^n$, and let $x^\star = 0$ be the unique fixed point of φ, i.e, $\varphi(0) = 0$. For the noisy recursion $x^{t+1} = \varphi(x^t) + \sigma w^t$, where w^t is a zero-mean white noise with identity covariance and $\mathbb{E}((w^t)^T\varphi(x^t)) = 0$, the contractiveness of φ implies

$$\mathbb{E}(\|x^{t+1}\|^2) \;=\; \mathbb{E}(\|\varphi(x^t)+\sigma w^t\|^2) \;\le\; \eta^2\,\mathbb{E}(\|x^t\|^2) \;+\; n\sigma^2.$$

Since $\eta < 1$, this relation yields

$$\lim_{t\to\infty}\mathbb{E}(\|x^t\|^2) \;\le\; \frac{n\sigma^2}{1\;-\;\eta^2}.$$

If $\eta := \max\{|1-\alpha m|, |1-\alpha L|\} < 1$, the map $\varphi(x) := x - \alpha\nabla f(x)$ is a contraction [62]. Thus, for the conventional stepsize $\alpha = 1/L$ we have $\eta = 1 - 1/\kappa$, and the bound becomes

$$\lim_{t\to\infty}\mathbb{E}(\|x^t\|^2) \;\le\; \frac{n\sigma^2}{1\;-\;\eta^2} \;=\; \frac{n\sigma^2\kappa^2}{2\kappa\;-\;1} \;=\; n\Theta(\kappa).$$

In the next section, we show that this upper bound is indeed tight for the class of functions $\mathcal{F}_m^L$. While this approach yields a tight upper bound for gradient descent, it cannot be used for Nesterov's method (because it is not a contraction).

2.4.2 An approach based on linear matrix inequalities

For any function $f \in \mathcal{F}_m^L$, the nonlinear mapping $\Delta\colon \mathbb{R}^n \to \mathbb{R}^n$

$$\Delta(y) \;:=\; \nabla f(y) \;-\; m\,y$$

satisfies the quadratic inequality [52, Lemma 6]

$$\begin{bmatrix} y \,-\, y_0 \\ \Delta(y) \,-\, \Delta(y_0) \end{bmatrix}^T \Pi \begin{bmatrix} y \,-\, y_0 \\ \Delta(y) \,-\, \Delta(y_0) \end{bmatrix} \;\geq\; 0 \tag{2.19}$$

for all y, $y_0 \in \mathbb{R}^n$, where the matrix Π is given by

$$\Pi \;:=\; \begin{bmatrix} 0 & (L \,-\, m)I \\ (L \,-\, m)I & -2I \end{bmatrix}. \tag{2.20}$$

We can bring algorithms (2.2) with constant parameters into a time-invariant state-space form

$$\begin{aligned} \psi^{t+1} &= A\,\psi^t \;+\; \sigma B_w w^t \;+\; B_u u^t \\ \begin{bmatrix} z^t \\ y^t \end{bmatrix} &= \begin{bmatrix} C_z \\ C_y \end{bmatrix} \psi^t \\ u^t &= \Delta(y^t) \end{aligned} \tag{2.21a}$$

that contains a feedback interconnection of linear and nonlinear components. Figure 2.4 illustrates the block diagram of system (2.21a), where ψ^t is the state, w^t is a white stochastic noise, z^t is the performance output, and u^t is the output of the nonlinear term $\Delta(y^t)$. In particular, if we let

$$\psi^t \;:=\; \begin{bmatrix} x^t \\ x^{t+1} \end{bmatrix}, \quad z^t \;:=\; x^t, \quad y^t \;:=\; -\beta x^t + (1+\beta) x^{t+1}$$

and define the corresponding matrices as

$$\begin{aligned} A &= \begin{bmatrix} 0 & I \\ -\beta(1-\alpha\,m)I & (1+\beta)(1-\alpha\,m)I \end{bmatrix}, \quad B_w \;=\; \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad B_u \;=\; \begin{bmatrix} 0 \\ -\alpha\,I \end{bmatrix} \\ C_z &= \begin{bmatrix} I & 0 \end{bmatrix}, \quad C_y \;=\; \begin{bmatrix} -\beta\,I & (1+\beta)I \end{bmatrix} \end{aligned} \tag{2.21b}$$

then (2.21a) represents Nesterov's method (2.2c). For gradient descent (2.2a), we can alternatively use $\psi^t = z^t = y^t := x^t$ with the corresponding matrices

$$A \;=\; (1 \,-\, \alpha\,m)I, \quad B_w \;=\; I, \quad B_u \;=\; -\alpha I, \quad C_z \;=\; C_y \;=\; I. \tag{2.21c}$$

In what follows, we demonstrate how property (2.19) of the nonlinear mapping Δ allows us to obtain upper bounds on J when system (2.21a) is driven by the white stochastic input w^t with zero mean and identity covariance. Lemma 1 uses a quadratic Lyapunov function of the form $V(\psi) = \psi^T X \psi$ and provides upper bounds on the steady-state second-order

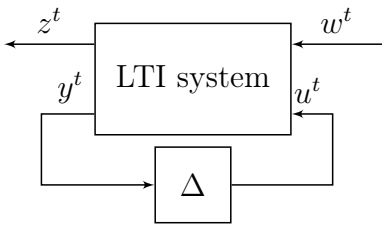


Figure 2.4: Block diagram of system (2.21a).

moment of the performance output z^t in terms of solutions to a certain LMI. This approach yields a tight upper bound for gradient descent.

Lemma 1 *Let the nonlinear function $u = \Delta(y)$ satisfy the quadratic inequality*

$$\begin{bmatrix} y \\ u \end{bmatrix}^T \Pi \begin{bmatrix} y \\ u \end{bmatrix} \ \geq \ 0 \tag{2.22}$$

for some matrix Π, let X be a positive semidefinite matrix, and let λ be a nonnegative scalar such that system (2.21a) *satisfies*

$$\begin{bmatrix} A^T X\, A - X + C_z^T\, C_z & A^T X\, B_u \\ B_u^T\, X\, A & B_u^T\, X\, B_u \end{bmatrix} \ + \ \lambda \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \preceq 0. \tag{2.23}$$

Then the steady-state second-order moment J of the performance output z^t in (2.21a) *is bounded by*

$$J \ \leq \ \sigma^2 \operatorname{trace}(B_w^T\, X\, B_w).$$

Proof: See Appendix A.2. □

For Nesterov's accelerated method with the parameters provided in Table 2.1, we have conducted computational experiments showing that LMI (2.23) becomes infeasible for large values of the condition number κ. Thus, Lemma 1 does not provide sensible upper bounds on J for Nesterov's algorithm. This observation is consistent with the results of [52], where it was suggested that analyzing the convergence rate requires the use of additional quadratic inequalities, apart from (2.19), to further tighten the constraints on the gradient ∇f and reduce conservativeness. In what follows, we build on the results of [56] and present an alternative LMI in Lemma 2 that is obtained using a Lyapunov function of the form $V(\psi) = \psi^T X \psi + f([\,0 \ \ I\,]\psi)$, where X is a positive semidefinite matrix and f is the objective function in (2.1). Such Lyapunov functions have been used to study convergence of optimization algorithms in [106]. The resulting approach allows us to establish an order-wise tight analytical upper bound on J for Nesterov's accelerated method.

Lemma 2 *Let the matrix $M(m, L; \alpha, \beta)$ be defined as*

$$M \ := \ N_1^T \begin{bmatrix} L\, I & I \\ I & 0 \end{bmatrix} N_1 \ + \ N_2^T \begin{bmatrix} -m\, I & I \\ I & 0 \end{bmatrix} N_2$$

where

$$N_1 := \begin{bmatrix} \alpha\, m\, \beta\, I & -\alpha\, m(1+\beta)\, I & -\alpha\, I \\ -m\, \beta\, I & m(1+\beta)\, I & I \end{bmatrix}, \quad N_2 := \begin{bmatrix} -\beta\, I & \beta\, I & 0 \\ -m\, \beta\, I & m(1+\beta)\, I & I \end{bmatrix}.$$

Consider the state-space model in (2.21a)-(2.21b) *for algorithm* (2.2c) *and let* Π *be given by* (2.20)*. Then, for any positive semidefinite matrix* X *and scalars* $\lambda_1 \geq 0$ *and* $\lambda_2 \geq 0$ *that satisfy*

$$\begin{bmatrix} A^T X\, A - X + C_z^T\, C_z & A^T X\, B_u \\ B_u^T\, X\, A & B_u^T\, X\, B_u \end{bmatrix} + \lambda_1 \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} + \lambda_2\, M \preceq 0 \quad (2.24)$$

the steady-state second-order moment J *of the performance output* z^t *in* (2.21a) *is bounded by*

$$J \;\leq\; \sigma^2 \left(n\, L\, \lambda_2 \;+\; \mathrm{trace}\,(B_w^T\, X\, B_w) \right). \quad (2.25)$$

Proof: See Appendix A.2. □

Remark 3 *Since LMI* (2.24) *simplifies to* (2.23) *by setting* $\lambda_2 = 0$*, Lemma 2 represents a relaxed version of Lemma 1. This modification is the key enabler to establishing tight upper bound on* J *for Nesterov's method.*

The upper bounds provided in Lemmas 1 and 2 are proportional to σ^2. In what follows, to make a connection between these bounds and our analytical expressions for the variance amplification in the quadratic case (Section 2.3), we again set $\sigma = 1$. The best upper bound on J that can be obtained using Lemma 2 is given by the optimal objective value of the semidefinite program

$$\begin{aligned} \underset{X,\, \lambda_1,\, \lambda_2}{\text{minimize}} \quad & n\, L\, \lambda_2 \;+\; \mathrm{trace}\,(B_w^T\, X\, B_w) \\ \text{subject to} \quad & \text{LMI (2.24)}, \;\; X \succeq 0, \;\; \lambda_1 \geq 0, \;\; \lambda_2 \geq 0. \end{aligned} \quad (2.26)$$

For system matrices (2.21b), LMI (2.24) is of size $3n \times 3n$ where $x^t \in \mathbb{R}^n$. However, if we impose the additional constraint that the matrix X has the same block structure as A,

$$X \;=\; \begin{bmatrix} x_1 I & x_0 I \\ x_0 I & x_2 I \end{bmatrix}$$

for some scalars x_1, x_2, and x_0, then using appropriate permutation matrices, we can simplify (2.23) into an LMI of size 3×3. Furthermore, imposing this constraint comes without loss of generality. In particular, the optimal objective value of problem (2.26) does not change if we require X to have this structure; see [52, Section 4.2] for a discussion of this lossless dimensionality reduction for LMI constraints with similar structure.

In Theorem 5, we use Lemmas 1 and 2 to establish tight upper bounds on J_{gd} and J_{na} for all $f \in \mathcal{F}_m^L$.

Theorem 5 *For gradient descent and Nesterov's accelerated method with the parameters provided in Table 2.1 and $\sigma = 1$, the performance measures $J_{\rm gd}$ and $J_{\rm na}$ of the error $x^t - x^\star \in \mathbb{R}^n$ satisfy*

$$\sup_{f \in \mathcal{F}_m^L} J_{\rm gd} \;=\; q_{\rm gd}, \qquad q_{\rm na} \;\leq\; \sup_{f \in \mathcal{F}_m^L} J_{\rm na} \;\leq\; 4.08\, q_{\rm na}$$

where

$$q_{\rm gd} \;=\; \frac{n\kappa^2}{2\kappa - 1} \;=\; n\,\Theta(\kappa), \quad q_{\rm na} \;=\; \frac{n\kappa^2(2\kappa - 2\sqrt{\kappa} + 1)}{(2\sqrt{\kappa} - 1)^3} \;=\; n\,\Theta(\kappa^{\frac{3}{2}})$$

and $\kappa := L/m$ is the condition number of the set $\mathcal{F}_m^L$.

Proof: See Appendix A.2. □

The variance amplification of gradient descent and Nesterov's method for $f(x) = \frac{m}{2}\, x^T x$ in $\mathcal{F}_m^L$ is determined by $q_{\rm gd}$ and $q_{\rm na}$, respectively, and these two quantities can be obtained using Theorem 1. In Theorem 5, we use this strongly convex quadratic objective function to certify the accuracy of the upper bounds on $\sup J$ for all $f \in \mathcal{F}_m^L$. In particular, we observe that the upper bound is exact for gradient descent and that it is within a 4.08 factor of the optimal for Nesterov's method.

For strongly convex objective functions with the condition number κ, Theorem 5 proves that gradient descent outperforms Nesterov's accelerated method in terms of the largest noise amplification by a factor of $\sqrt{\kappa}$. This uncovers the fundamental performance limitation of Nesterov's accelerated method when the gradient evaluation is subject to additive stochastic uncertainties.

2.5 Tuning of algorithmic parameters

The parameters provided in Table 2.2 yield the optimal convergence rate for strongly convex quadratic problems. For these specific values, Theorem 4 establishes upper and lower bounds on the variance amplification that reveal the negative impact of acceleration. However, it is relevant to examine whether the parameters can be designed to provide acceleration while reducing the variance amplification.

While the convergence rate solely depends on the extreme eigenvalues $m = \lambda_{\min}(Q)$ and $L = \lambda_{\max}(Q)$ of the Hessian matrix Q, variance amplification is influenced by the entire spectrum of Q and its minimization is challenging as it requires the use of all eigenvalues. In this section, we first consider the special case of eigenvalues being symmetrically distributed over the interval $[m, L]$ and demonstrate that for gradient descent and the heavy-ball method, the parameters provided in Table 2.2 yield a variance amplification that is within a constant factor of the optimal value. As we demonstrate in Section 2.6, symmetric distribution of the eigenvalues is encountered in distributed consensus over undirected torus networks. We also consider the problem of designing parameters for objective functions in which the problem size satisfies $n \ll \kappa$ and establish a tradeoff between convergence rate and variance amplification. More specifically, we show that for any accelerating pair of parameters α and β and

bounded problem dimension n, the variance amplification of accelerated methods is larger than that of gradient descent by a factor of $\Omega(\sqrt{\kappa})$.

2.5.1 Tuning of parameters using the whole spectrum

Let $L = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n = m > 0$ be the eigenvalues of the Hessian matrix Q of the strongly convex quadratic objective function in (2.4). Algorithms (2.2) converge linearly in expected value to the optimizer $x^\star$ with the rate

$$\rho \;:=\; \max_i \; \hat{\rho}(\lambda_i) \tag{2.27}$$

where $\hat{\rho}(\lambda_i)$ is the spectral radius of the matrix $\hat{A}_i$ given by (2.8). For any scalar $c > 0$ and fixed σ, let

$$\begin{aligned}(\alpha^\star_{\rm hb}(c), \beta^\star_{\rm hb}(c)) := \; & \underset{\alpha,\,\beta}{\rm argmin} \qquad J_{\rm hb}(\alpha, \beta) \\ & \text{subject to} \quad \rho_{\rm hb} \;\leq\; 1 - \frac{c}{\sqrt{\kappa}}\end{aligned} \tag{2.28a}$$

for the heavy-ball method, and

$$\begin{aligned}\alpha^\star_{\rm gd}(c) := \; & \underset{\alpha}{\rm argmin} \qquad J_{\rm gd}(\alpha) \\ & \text{subject to} \quad \rho_{\rm gd} \;\leq\; 1 - \frac{c}{\kappa}\end{aligned} \tag{2.28b}$$

for gradient descent, where the expression for the variance amplification J is provided in Theorem 1. Here, the constraints enforce a standard rate of linear convergence for gradient descent and an accelerated rate of linear convergence for the heavy-ball method parametrized with the constant c. Obtaining a closed form solution to (2.28) is challenging because J depends on all eigenvalues of the Hessian matrix Q. Herein, we focus on objective functions for which the spectrum of Q is symmetric, i.e., for any eigenvalue λ, the corresponding mirror image $\lambda' := L + m - \lambda$ with respect to $\frac{1}{2}(L + m)$ is also an eigenvalue with the same algebraic multiplicity. For this class of problems, Theorem 6 demonstrates that the parameters provided in Table 2.2 for gradient descent and the heavy-ball method yield variance amplification that is within a constant factor of the optimal.

Theorem 6 *For any scalar $c > 0$ and fixed σ, there exist constants $c_1 \geq 1$ and $c_2 > 0$ such that for any strongly convex quadratic objective function in which the spectrum of the Hessian matrix Q is symmetrically distributed over the interval $[m, L]$ with $\kappa := L/m > c_1$, we have*

$$J_{\rm gd}(\alpha^\star_{\rm gd}(c)) \;\geq\; \frac{1}{2}\, J_{\rm gd}(\alpha_{\rm gd}), \quad J_{\rm hb}(\alpha^\star_{\rm hb}(c), \beta^\star_{\rm hb}(c)) \;\geq\; c_2\, J_{\rm hb}(\alpha_{\rm hb}, \beta_{\rm hb})$$

where parameters $\alpha_{\rm gd}$ and $(\alpha_{\rm hb}, \beta_{\rm hb})$ are provided in Table 2.2, and $\alpha^\star_{\rm gd}(c)$ and $(\alpha^\star_{\rm hb}(c), \beta^\star_{\rm hb}(c))$ solve (2.28).

Proof: See Appendix A.2.4. □

For strongly convex quadratic objective functions with symmetric spectrum of the Hessian matrix over the interval $[m, L]$, Theorem 6 shows that the variance amplifications of gradient descent and the heavy-ball method with the parameters provided in Table 2.2 are within a constant factors of the optimal values. As we illustrate in Section 2.6, this class of problems is encountered in distributed averaging over noisy undirected networks. Combining this result with the lower bound on $J_{\rm hb}(\alpha_{\rm hb}, \beta_{\rm hb})$ and the upper bound on $J_{\rm gd}(\alpha_{\rm gd})$ established in Theorem 4, we see that regardless of the choice of parameters, there is a fundamental gap of $\Omega(\sqrt{\kappa})$ between $J_{\rm hb}$ and $J_{\rm gd}$ as long as we require an accelerated rate of convergence.

2.5.2 Fundamental lower bounds

We next establish lower bounds on the variance amplification of accelerated methods that hold for any pair of α and β for strongly convex quadratic problems with $\kappa \gg 1$. In particular, we show that the variance amplification of accelerated algorithms is lower bounded by $\Omega(\kappa^{3/2})$ irrespective of the choice of α and β.

The next theorem establishes a fundamental tradeoff between the convergence rate and variance amplification for the heavy-ball method.

Theorem 7 *For strongly convex quadratic problems with any stabilizing parameters $\alpha > 0$ and $0 < \beta < 1$ and with a fixed noise magnitude σ, the heavy-ball method with the linear convergence rate ρ satisfies*

$$\frac{J_{\rm hb}}{1-\rho} \;\geq\; \sigma^2 \left(\frac{\kappa+1}{8}\right)^2. \tag{2.29a}$$

Furthermore, if $\sigma = \alpha$, i.e., when the only source of uncertainty is a noisy gradient, we have

$$\frac{J_{\rm hb}}{1-\rho} \;\geq\; \left(\frac{\kappa}{8L}\right)^2. \tag{2.29b}$$

Proof: See Appendix A.3. □

To gain additional insight, let us consider two special cases: (i) for $\alpha = 1/L$ and $\beta \to 0^+$, we obtain gradient descent algorithm for which $1 - \rho = \Theta(1/\kappa)$ and $J = \Theta(\kappa)$; (ii) for the heavy-ball method with the parameters provided in Table 2.2, we have $1 - \rho = \Theta(1/\sqrt{\kappa})$ and $J = \Theta(\kappa\sqrt{\kappa})$. Thus, in both cases, $J_{\rm hb}/(1-\rho) = \Omega(\kappa^2)$. Theorem 7 shows that this lower bound is fundamental and it therefore quantifies the tradeoff between the convergence rate and the variance amplification of the heavy-ball method for any choice of parameters α and β. It is also worth noting that the lower bound for $\sigma = \alpha$ depends on the largest eigenvalue L of the Hessian matrix Q. Thus, this bound is meaningful when the value of L is uniformly upper bounded. This scenario occurs in many applications including consensus over undirected tori networks; see Section 2.6.

While we are not able to show a similar lower bound for Nesterov's method, in the next theorem, we establish an asymptotic lower bound on the variance amplification that holds for any pair of accelerating parameters (α, β) for both Nesterov's and heavy-ball methods.

Theorem 8 *For a strongly convex quadratic objective function with condition number κ, let $c > 0$ be a constant such that either Nesterov's algorithm or the heavy-ball method with some (possibly problem dependent) parameters $\alpha > 0$ and $0 < \beta < 1$ converges linearly with a rate $\rho \leq 1 - c/\sqrt{\kappa}$. Then, for any fixed noise magnitude σ, the variance amplification satisfies*

$$\frac{J}{\sigma^2} \;=\; \Omega(\kappa^{\frac{3}{2}}). \tag{2.30a}$$

Furthermore, if $\sigma = \alpha$, i.e., when the only source of uncertainty is a noisy gradient, we have

$$J \;=\; \Omega(\frac{\kappa^{\frac{3}{2}}}{L^2}). \tag{2.30b}$$

Proof: For the heavy-ball method, the result follows from combining Theorem 7 with the inequality $1 - \rho \geq c/\sqrt{\kappa}$. For Nesterov's method, the proof is provided in Appendix A.3. □

For problems with $n \ll \kappa$, we recall that the variance amplification of gradient descent with conventional values of parameters scales as $O(\kappa)$; see Theorem 5. Irrespective of the choice of parameters α and β, this result in conjunction with Theorem 8 demonstrates that acceleration cannot be achieved without increasing the variance amplification J by a factor of $\Omega(\sqrt{\kappa})$.

2.6 Application to distributed computation

Distributed computation over networks has received significant attention in optimization, control systems, signal processing, communications, and machine learning communities. In this problem, the goal is to optimize an objective function (e.g., for the purpose of training a model) using multiple processing units that are connected over a network. Clearly, the structure of the network (e.g., node dynamics and network topology) may impact the performance (e.g., convergence rate and noise amplification) of any optimization algorithm. As a first step toward understanding the impact of the network structure on performance of noisy first-order optimization algorithms, in this section, we examine the standard distributed consensus problem.

The consensus problem arises in applications ranging from social networks, to distributed computing networks, to cooperative control in multi-agent systems. In the simplest setup, each node updates a scalar value using the values of its neighbors such that they all agree on a single consensus value. Simple updating strategies of this kind can be obtained by applying a first-order algorithm to the convex quadratic problem

$$\underset{x}{\text{minimize}} \;\; \frac{1}{2}\, x^T \mathbf{L}\, x \tag{2.31}$$

where $\mathbf{L} = \mathbf{L}^T \in \mathbb{R}^{n \times n}$ is the Laplacian matrix of the graph associated with the underlying undirected network and $x \in \mathbb{R}^n$ is the vector of node values.

The graph Laplacian matrix $\mathbf{L} \succeq 0$ has a nontrivial null space that consists of the minimizers of problem (2.31). In the absence of noise, for gradient descent and both of its accelerated variants, it is straightforward to verify that the projections v^t of the iterates x^t onto the null space of $\mathbf{L}$ remain constant ($v^t = v^0$, for all t) and also that x^t converges linearly to v^0. In the presence of additive noise, however, v^t experiences a random walk which leads to an unbounded variance of x^t as $t \to \infty$. Instead, as described in [42], the performance of algorithms in this case can be quantified by examining $\bar{J} := \lim_{t\to\infty} \mathbb{E}\left(\|x^t - v^t\|^2\right)$. For connected networks, the null space of $\mathbf{L}$ is given by $\mathcal{N}(\mathbf{L}) = \{c\mathbb{1} \mid c \in \mathbb{R}\}$ and

$$\bar{J} \;=\; \lim_{t\to\infty} \mathbb{E}\left(\|x^t - (\mathbb{1}^T x^t/n)\mathbb{1}\|^2\right) \tag{2.32}$$

quantifies the mean-squared deviation from the network average, where $\mathbb{1}$ denotes the vector of all ones, i.e., $\mathbb{1} := [\,1 \;\cdots\; 1\,]^T$. Finally, it is straightforward to show that $\bar{J}$ can also be computed using the formulae in Theorem 1 by summing over the non-zero eigenvalues of $\mathbf{L}$.

In what follows, we consider a class of networks for which the structure allows for the explicit evaluation of the eigenvalues of the Laplacian matrix $\mathbf{L}$. For d-dimensional torus networks, fundamental performance limitations of standard consensus algorithms in continuous time were established in [43], but it remains an open question if gradient descent and its accelerated variants suffer from these limitations. We use such torus networks to show that standard gradient descent exhibits the same scaling trends as consensus algorithms studied in [43] and that, in lower spatial dimensions, acceleration always increases variance amplification.

2.6.1 Explicit formulae for d-dimensional torus networks

We next examine the asymptotic scaling trends of the performance metric $\bar{J}$ given by (2.32) for large problem dimensions $n \gg 1$ and highlight the subtle influence of the distribution of the eigenvalues of $\mathbf{L}$ on the variance amplification for d-dimensional torus networks. Tori with nearest neighbor interactions generalize one-dimensional rings to higher spatial dimensions. Let $\mathbb{Z}_{n_0}$ denote the group of integers modulo n_0. A d-dimensional torus $\mathbb{T}^d_{n_0}$ consists of $n := n_0^d$ nodes denoted by v_a where $a \in \mathbb{Z}^d_{n_0}$ and its set of edges is given by

$$\{\{v_a\, v_b\} \mid \|a \,-\, b\| \;=\; 1 \mod n_0\}$$

where the nodes v_a and v_b are neighbors if and only if a and b differ exactly at a single entry by one. For example, $\mathbb{T}^1_{n_0}$ denotes a ring with $n = n_0$ nodes and $\mathbb{T}^5_{n_0}$ denotes a five dimensional torus with $n = n_0^5$ nodes.

The multidimensional discrete Fourier transform can be used to determine the eigenvalues of the Laplacian matrix $\mathbf{L}$ of a d-dimensional torus $\mathbb{T}^d_{n_0}$,

$$\lambda_i \;=\; \sum_{l=1}^{d} 2\left(1 \,-\, \cos\tfrac{2\pi i_l}{n_0}\right), \quad i_l \,\in\, \mathbb{Z}_{n_0} \tag{2.33}$$

where $i := (i_1, \ldots, i_d) \in \mathbb{Z}_{n_0}^d$. We note that $\lambda_0 = 0$ is the only zero eigenvalue of $\mathbf{L}$ with the eigenvector $\mathbb{1}$ and that all other eigenvalues are positive. Let $\kappa := \lambda_{\max}/\lambda_{\min}$ be the ratio of the largest and smallest nonzero eigenvalues of $\mathbf{L}$. A key observation is that, for $n_0 \gg 1$,

$$\kappa \;=\; \Theta(\frac{2}{1 \,-\, \cos\frac{2\pi}{n_0}}) \;=\; \Theta(n_0^2) \;=\; \Theta(n^{2/d}). \tag{2.34}$$

This is because $\lambda_{\min} = 2d\,(1-\cos\,(2\pi/n_0))$ goes to zero as $n_0 \to \infty$, and the largest eigenvalue of $\mathbf{L}$, $\lambda_{\max} = 2d\,(1 - \cos\,(2\pi\lfloor\frac{n_0}{2}\rfloor/n_0))$, is equal to $4\,d$ for even n_0 and it approaches $4\,d$ from below for odd n_0.

As aforementioned, the performance metric $\bar{J}$ can be obtained by

$$\bar{J} \;=\; \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \hat{J}(\lambda_i)$$

where $\hat{J}(\lambda)$ for each algorithm is determined in Theorem 1 and λ_i are the non-zero eigenvalues of $\mathbf{L}$. The next theorem characterizes the asymptotic value of the network-size normalized mean-squared deviation from the network average, $\bar{J}/n$, for a fixed spatial dimension d and condition number $\kappa \gg 1$. This result is obtained using analytical expression (2.33) for the eigenvalues of the Laplacian matrix $\mathbf{L}$.

Theorem 9 *Let $\mathbf{L} \in \mathbb{R}^{n\times n}$ be the graph Laplacian of the d-dimensional undirected torus $\mathbb{T}_{n_0}^d$ with $n = n_0^d \gg 1$ nodes. For convex quadratic optimization problem* (2.31)*, the network-size normalized performance metric $\bar{J}/n$ of noisy first-order algorithms with the parameters provided in Table 2.2 and $\sigma = 1$, is determined by*

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
Gradient	$\Theta(\sqrt{\kappa})$	$\Theta(\log\,\kappa)$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
Nesterov	$\Theta(\kappa)$	$\Theta(\sqrt{\kappa}\log\,\kappa)$	$\Theta(\kappa^{\frac{1}{4}})$	$\Theta(\log\,\kappa)$	$\Theta(1)$
Polyak	$\Theta(\kappa)$	$\Theta(\sqrt{\kappa}\,\log\,\kappa)$	$\Theta(\sqrt{\kappa})$	$\Theta(\sqrt{\kappa})$	$\Theta(\sqrt{\kappa})$

where $\kappa = \Theta(n^{2/d})$ is the condition number of $\mathbf{L}$ given in (2.34)*.*

Proof: See Appendix A.4. □

Theorem 9 demonstrates that the variance amplification of gradient descent is equivalent to that of the standard consensus algorithm studied in [43] and that, in lower spatial dimensions, acceleration always negatively impacts the performance of noisy algorithms. Our results also highlight the subtle influence of the distribution of the eigenvalues of $\mathbf{L}$ on the variance amplification. For rings (i.e., $d = 1$), lower bounds provided in Theorem 4 capture the trends that our detailed analysis based on the distribution of the entire spectrum of $\mathbf{L}$ reveals. In higher spatial dimensions, however, the lower bounds that are obtained using only the extreme eigenvalues of $\mathbf{L}$ are conservative. Similar conclusion can be made about the upper bounds provided in Theorem 4. This observation demonstrates that the naïve bounds that result only from the use of the extreme eigenvalues can be overly conservative.

We also note that gradient descent significantly outperforms Nesterov's method in lower spatial dimensions. In particular, while $\bar{J}/n$ becomes network-size-independent for $d = 3$ for gradient descent, Nesterov's algorithm reaches "critical connectivity" only for $d = 5$. On the other hand, in any spatial dimension, there is no network-size independent upper bound on $\bar{J}/n$ for the heavy-ball method. These conclusions could not have been reached without performing an in-depth analysis of the impact of all eigenvalues on performance of noisy networks with $n \gg 1$ and $\kappa \gg 1$.

2.7 Concluding remarks

We study the robustness of noisy first-order algorithms for smooth, unconstrained, strongly convex optimization problems. Even though the underlying dynamics of these algorithms are in general nonlinear, we establish upper bounds on noise amplification that are accurate up to constant factors. For quadratic objective functions, we provide analytical expressions that quantify the effect of all eigenvalues of the Hessian matrix on variance amplification. We use these expressions to establish lower bounds demonstrating that although the acceleration techniques improve the convergence rate they significantly amplify noise for problems with large condition numbers. In problems of bounded dimension $n \ll \kappa$, the noise amplification increases from $O(\kappa)$ to $\Omega(\kappa^{3/2})$ when moving from standard gradient descent to accelerated algorithms. We specialize our results to the problem of distributed averaging over noisy undirected networks and also study the role of network size and topology on robustness of accelerated algorithms. Future research directions include (i) extension of our analysis to multiplicative and correlated noise; and (ii) robustness analysis of broader classes of optimization algorithms.

Chapter 3

Tradeoffs between convergence rate and noise amplification for accelerated algorithms

We study momentum-based first-order optimization algorithms in which the iterations utilize information from the two previous steps and are subject to an additive white noise. This class of algorithms includes Polyak's heavy-ball and Nesterov's accelerated methods as special cases and noise accounts for uncertainty in either gradient evaluation or iteration updates. For strongly convex quadratic problems, we use the steady-state variance of the error in the optimization variable to quantify noise amplification and identify fundamental stochastic performance tradeoffs. Our approach utilizes the Jury stability criterion to provide a novel geometric characterization of conditions for linear convergence, and it clarifies the relation between the noise amplification and convergence rate as well as their dependence on the condition number and the constant algorithmic parameters. This geometric insight leads to simple alternative proofs of standard convergence results and allows us to establish analytical lower bounds on the product between the settling time and noise amplification that scale quadratically with the condition number. Our analysis also identifies a key difference between the gradient and iterate noise models: while the amplification of gradient noise can be made arbitrarily small by sufficiently decelerating the algorithm, the best achievable variance amplification for the iterate noise model increases linearly with the settling time in decelerating regime. Furthermore, we introduce two parameterized families of algorithms that strike a balance between noise amplification and settling time while preserving order-wise Pareto optimality for both noise models. Finally, by analyzing a class of accelerated gradient flow dynamics, whose suitable discretization yields the two-step momentum algorithm, we establish that stochastic performance tradeoffs also extend to continuous time.

3.1 Introduction

Accelerated first-order algorithms [5], [7], [8] are often used in solving large-scale optimization problems [1], [2], [4] because of their scalability, fast convergence, and low per-iteration complexity. Convergence properties of these algorithms have been carefully studied [6], [9], [51]–[56], but their performance in the presence of noise has received less attention [10]–[12], [57], [58]. Prior studies indicate that inaccuracies in the computation of gradient values can adversely impact the convergence rate of accelerated methods and that gradient descent may

have advantages relative to its accelerated variants in noisy environments [23]–[26], [28]. In contrast to gradient descent, accelerated algorithms can also exhibit undesirable transient behavior [61], [98], [107]; for convex quadratic problems, the non-normal dynamic modes in accelerated algorithms induce large transient responses of the error in the optimization variable [98].

Analyzing the performance of accelerated algorithms with additive white noise that arises from uncertainty in gradient evaluation dates back to [57] where Polyak established the optimal linear convergence rate for strongly convex quadratic problems. In addition, he used time-varying parameters to obtain convergence in the error variance at a sub-linear rate and with an improved constant factor compared to gradient descent. Acceleration in a sub-linear regime can also be achieved for smooth strongly convex problems with properly diminishing stepsize [30] and averaging techniques can be used to prevent the accumulation of gradient noise by accelerated algorithms [108]. For standard accelerated methods with constant parameters, control-theoretic tools were utilized in [93] and [109] to study the steady-state variance of the error in optimization variable for smooth strongly convex problems. In particular, for the parameters that optimize convergence rates for quadratic problems, tight upper and lower bounds on the noise amplification of gradient descent, heavy-ball method, and Nesterov's accelerated algorithm were developed in [93]. These bounds are expressed in terms of the condition number κ and the problem dimension n, and they demonstrate opposite trends relative to the settling time: *for a fixed problem size n, accelerated algorithms increase noise amplification by a factor of $\Theta(\sqrt{\kappa})$ relative to gradient descent.* Similar result also holds for heavy-ball and Nesterov's algorithms with parameters that provide convergence rate $\rho \leq 1 - c/\sqrt{\kappa}$ with $c > 0$ [93]. Furthermore, for all strongly convex optimization problems with a condition number κ, tight and attainable upper bounds for noise amplification of gradient descent and Nesterov's accelerated method were provided in [93].

In this chapter, we extend the results of [93] to the class of first-order algorithms with three constant parameters in which the iterations involve information from the two previous steps. This class includes heavy-ball and Nesterov's accelerated algorithms as special cases and we examine its stochastic performance for strongly convex quadratic problems. Our results are complementary to [103], which evaluates stochastic performance in the objective error, and to a recent work [109] that studies the steady-state variance of the error associated with the point at which the gradient is evaluated. This reference combines theory with computational experiments to demonstrate that a parameterized family of heavy-ball-like methods with reduced stepsize provides Pareto-optimal algorithms for the simultaneous optimization of convergence rate and amplification of gradient noise. In contrast to [109], we establish analytical lower bounds on the product of the settling time and the steady-state variance of the error in the optimization variable that hold for any constant stabilizing parameters and for both gradient and iterate noise models. Our lower bounds scale with the square of the condition number and thus reveal a fundamental limitation of this class of algorithms.

In addition to considering noise arising from gradient evaluation, we study the stochastic performance of algorithms when noise is directly added to the iterates (rather than the gradient). For the iterate noise model, we establish an alternative lower bound on the noise amplification which scales linearly with the settling time and is order-wise tight for settling times that are larger than that of gradient descent with the standard stepsize. In this

decelerated regime, our results identify a key difference between the two noise models: while the impact of gradient uncertainties on variance amplification can be made arbitrarily small by decelerating the two-step momentum algorithm, the best achievable variance amplification for the iterate noise model increases linearly with the settling time in the decelerated regime.

Our results build upon a simple, yet powerful geometric viewpoint, which clarifies the relation between condition number, convergence rate, and algorithm parameters for strongly convex quadratic problems. This viewpoint allows us to present alternative proofs for (i) the optimal convergence rate of the two-step momentum algorithm, which recovers Nesterov's fundamental lower bound on the convergence rate [59] for finite dimensional problems [60]; and (ii) the optimal rates achieved by standard gradient descent, heavy-ball method, and Nesterov's accelerated algorithm [52]. In addition, it enables a novel geometric characterization of noise amplification in terms of stability margins and it allows us to precisely quantify tradeoffs between convergence rate and robustness to noise.

We also introduce two parameterized families of algorithms that are structurally similar to the heavy-ball and Nesterov's accelerated algorithms. These algorithms utilize continuous transformations from gradient descent to the corresponding accelerated algorithm (with the optimal convergence rate) via a homotopy path, and they can be used to provide additional insight into the tradeoff between convergence rate and noise amplification. We prove that these parameterized families are order-wise (in terms of the condition number) Pareto-optimal for simultaneous minimization of settling time and noise amplification. Another family of algorithms that facilitates similar tradeoff was proposed in [54], and it includes the fastest known algorithm for the class of smooth strongly convex problems. We also utilize negative momentum parameters to decelerate a heavy-ball-like family of algorithms relative to gradient descent with the optimal stepsize. For both noise models, our parameterized family yields order-wise optimal algorithms and it allows us to further highlight the key difference between them in the decelerated regime.

Finally, we examine the noise amplification of a class of stochastically-forced momentum-based accelerated gradient flow dynamics. Such dynamics were introduced in [110] as a continuous-time variant of Nesterov's accelerated algorithm and a Lyapunov-based method was used to establish their stability properties and infer the convergence rate. Inspired by this work, we examine the tradeoffs between the noise amplification and convergence rate of similar gradient flow dynamics for strongly convex quadratic problems. We introduce a geometric viewpoint analogous to the discrete-time setting to characterize the optimal convergence rate and identify the corresponding algorithmic parameters. We then examine the dependence of the noise amplification on the parameters and the spectrum of the Hessian matrix and demonstrate that our findings regarding the restrictions imposed by the condition number on the product of the settling time and noise amplification extend to the continuous-time case as well.

The rest of the chapter is organized as follows. In Section 3.2, we provide preliminaries and background material and, in Section 3.3, we summarize our key contributions. In Section 3.4, we introduce the tools and ideas that enable our analysis. In particular, we utilize the Jury stability criterion to provide a novel geometric characterization of stability and ρ-linear convergence and exploit this insight to derive simple alternative proofs of standard convergence results and quantify fundamental stochastic performance tradeoffs. In

Section 3.5, we introduce two parameterized families of algorithms that allow us to constructively tradeoff settling time and noise amplification. In Section 3.6, we extend our results to the continuous-time setting, in Section 3.7, we provide proofs of our main results, and in Section 3.8, we conclude the chapter.

3.2 Preliminaries and background

For the unconstrained optimization problem

$$\underset{x}{\text{minimize}} \;\; f(x) \tag{3.1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a strongly convex function with a Lipschitz continuous gradient ∇f, we consider noisy momentum-based first-order algorithms that use information from the two previous steps to update the optimization variable:

$$x^{t+2} = x^{t+1} + \beta(x^{t+1} - x^t) - \alpha\nabla f(x^{t+1} + \gamma(x^{t+1} - x^t)) + \sigma_w w^t. \tag{3.2}$$

Here, t is the iteration index, α is the stepsize, β and γ are momentum parameters, σ_w is the noise magnitude, and w^t is an additive white noise with zero mean and identity covariance matrix,

$$\mathbb{E}\left[w^t\right] = 0,\; \mathbb{E}\left[w^t(w^\tau)^T\right] = I\,\delta(t-\tau)$$

where δ is the Kronecker delta and $\mathbb{E}$ is the expected value operator. In this chapter, we consider two noise models.

1. Iterate noise ($\sigma_w = \sigma$): models uncertainty in computing the iterates of (3.2), where σ denotes the stepsize-independent noise magnitude.

2. Gradient noise ($\sigma_w = \alpha\sigma$): models uncertainty in the gradient evaluation. In this case, the stepsize α directly impacts magnitude of the additive noise.

Iterate noise models scenarios where uncertainties in optimization variables exist because of roundoff, quantization, and communication errors. This model has also been used to improve generalization and robustness in machine learning [111]. On the other hand, the second noise model accounts for gradient computation error or scenarios in which the gradient is estimated from noisy measurements [40]. Also, noise may be intentionally added to the gradient for privacy reasons [112].

Remark 1 *An alternative noise model with $\sigma_w = \sqrt{\alpha}\sigma$ has been used to escape local minima in stochastic gradient descent [113] and to provide non-asymptotic guarantees in nonconvex learning [114], [115]. This model arises from a discretization of the continuous-time Langevin diffusion dynamics [114] and, for strongly convex quadratic problems, our framework can be used to examine acceleration/robustness tradeoffs. For algorithms that are faster than the standard gradient descent, this model has order-wise identical performance bounds as the other two models and the only difference arises in decelerated regime. We omit details for brevity.*

Special cases of (3.2) include noisy gradient descent ($\beta = \gamma = 0$), Polyak's heavy-ball method ($\gamma = 0$), and Nesterov's accelerated algorithm ($\gamma = \beta$). In the absence of noise (i.e., for $\sigma = 0$), the parameters (α, β, γ) can be selected such that the iterates converge linearly to the globally optimal solution [9]. For the family of smooth strongly convex problems, the parameters that yield the fastest known linear convergence rate were provided in [55].

3.2.1 Linear dynamics for quadratic problems

Let $\mathcal{Q}_m^L$ denote the class of m-strongly convex L-smooth quadratic functions

$$f(x) \;=\; \tfrac{1}{2}\, x^T Q\, x \;-\; q^T x \tag{3.3}$$

with the condition number $\kappa := L/m$, where q is a vector and $Q = Q^T \succ 0$ is the Hessian matrix with eigenvalues

$$L \;=\; \lambda_1 \;\geq\; \lambda_2 \;\geq\; \ldots \;\geq\; \lambda_n \;=\; m \;>\; 0.$$

For the quadratic objective function in (3.3), we can use a linear time-invariant (LTI) state-space model to describe the *two-step momentum algorithm* (3.2) with constant parameters,

$$\begin{aligned} \psi^{t+1} &= A\,\psi^t \;+\; B\,w^t \\ z^t &= C\,\psi^t \end{aligned} \tag{3.4a}$$

where ψ^t is the state, $z^t := x^t - x^\star$ is the performance output, and w^t is the white stochastic input. In particular, choosing $\psi^t := [(x^t - x^\star)^T\ (x^{t+1} - x^\star)^T]^T$ yields

$$\begin{aligned} A &= \begin{bmatrix} 0 & I \\ -\beta I + \gamma\alpha Q & (1+\beta)I - (1+\gamma)\alpha Q \end{bmatrix} \\ B^T &= \begin{bmatrix} 0 & \sigma_w I \end{bmatrix}, \quad C \;=\; \begin{bmatrix} I & 0 \end{bmatrix}. \end{aligned} \tag{3.4b}$$

3.2.2 Convergence rates

An algorithm is stable if in the absence of noise (i.e., $\sigma_w = 0$), the state converges linearly with some rate $\rho < 1$,

$$\|\psi^t\|_2 \;\leq\; c\,\rho^t\,\|\psi^0\|_2 \quad \text{for all } t \;\geq\; 1 \tag{3.5}$$

for all $f \in \mathcal{Q}_m^L$ and all initial conditions ψ^0, where $c > 0$ is a constant. For LTI system (3.4a), the spectral radius $\rho(A)$ determines the best achievable convergence rate. In addition,

$$T_s \;:=\; \frac{1}{1 \,-\, \rho} \tag{3.6}$$

determines the *settling time*, i.e., the number of iterations required to reach a given desired accuracy; see Appendix B.1. For the class $\mathcal{Q}_m^L$ of high-dimensional functions (i.e., for $n \gtrsim T_s$),

method	fastest parameters (α, β, γ)	T_s	$J_{\min}/\sigma_w^2$	$J_{\max}/\sigma_w^2$
Gradient	$(2/(L+m),\ 0,\ 0)$	$(\kappa+1)/2$	$\Theta(\kappa)+n$	$n\Theta(\kappa)$
Heavy-ball	$(4/(\sqrt{L}+\sqrt{m})^2,\ (1-2/(\sqrt{\kappa}+1))^2,\ 0)$	$(\sqrt{\kappa}+1)/2$	$\Theta(\kappa\sqrt{\kappa})+n\Theta(\sqrt{\kappa})$	$n\Theta(\kappa\sqrt{\kappa})$
Nesterov	$(4/(3L+m),\ 1-4/(\sqrt{3\kappa+1}+2,\ \beta)$	$\sqrt{3\kappa+1}/2$	$\Theta(\kappa\sqrt{\kappa})+n$	$n\Theta(\kappa\sqrt{\kappa})$

Table 3.1: Settling times $T_s := 1/(1-\rho)$ [52, Proposition 1] along with the corresponding noise amplification bounds in (3.10) [93, Theorem 4] for the parameters that optimize the linear convergence rate ρ for strongly convex quadratic function $f \in \mathcal{Q}_m^L$ with the condition number $\kappa := L/m$, Here, n is the dimension of x and σ_w^2 is the variance of the white noise.

Nesterov established the fundamental lower bound on the settling time (convergence rate) of any first-order algorithm [9],

$$T_s \ \geq \ \frac{\sqrt{\kappa}\ +\ 1}{2}. \tag{3.7}$$

This lower bound is sharp and it is achieved by the heavy-ball method with the parameters provided in Table 3.1 [52].

3.2.3 Noise amplification

For LTI system (3.4a) driven by an additive white noise w^t, $\mathbb{E}\left(\psi^{t+1}\right) = A\,\mathbb{E}\left(\psi^t\right)$. Thus, $\mathbb{E}\left(\psi^t\right) = A^t\,\mathbb{E}\left(\psi^0\right)$ and, for any stabilizing parameters (α, β, γ), the iterates reach a statistical steady-state with $\lim_{t\to\infty}\mathbb{E}\left(\psi^t\right) = 0$ and a variance that can be computed from the solution of the algebraic Lyapunov equation [41], [93]. We call the steady-state variance of the error in the optimization variable noise (or variance) amplification,

$$J \ := \ \lim_{t\to\infty}\ \frac{1}{t}\sum_{k=0}^{t}\mathbb{E}\left(\|x^k\ -\ x^\star\|_2^2\right). \tag{3.8}$$

In addition to the algorithmic parameters (α, β, γ), the entire spectrum $\{\lambda_i \,|\, i = 1, \ldots, n\}$ of the Hessian matrix Q impacts the noise amplification J of algorithm (3.2) [93].

Remark 2 *An alternative performance metric that examines the steady-state variance of $y^t - x^\star$ was considered in [109], where $y^t := x^t + \gamma(x^t - x^{t-1})$ is the point at which the gradient is evaluated in* (3.2)*. For all $\gamma \geq 0$, we have $J_x \leq J_y \leq (1+2\gamma)^2 J_x$, where the subscripts x and y denote the noise amplification in terms of the error in x^t and y^t. Thus, these performance metrics are within a constant factor of each other for bounded values of $\gamma \geq 0$.*

3.2.4 Parameters that optimize convergence rate

For special instances of the two-step momentum algorithm (3.2) applied to strongly convex quadratic problems, namely gradient descent (gd), heavy-ball method (hb), and Nesterov's

accelerated algorithm (na), the parameters that yield the fastest convergence rates were established in [52], [57]. These parameters along with the corresponding rates and the noise amplification bounds are provided in Table 3.1. The convergence rates are determined by the spectral radius of the corresponding A-matrices and the noise amplification bounds are computed by examining the solution to the algebraic Lyapunov equation and determining the functions $f \in \mathcal{Q}_m^L$ for which the steady-state variance is maximized/minimized [93, Proposition 1]. Since the optimal convergence rate for the heavy-ball method meets the fundamental lower bound (3.7), this choice of parameters also optimizes the convergence rate of the two-step momentum algorithm (3.2) for $f \in \mathcal{Q}_m^L$.

For the optimal parameters provided in Table 3.1, there is a $\Theta(\sqrt{\kappa})$ improvement in settling times of the heavy-ball and Nesterov's accelerated algorithms relative to gradient descent,

$$T_s \;=\; \begin{cases} \Theta(\kappa) & \text{gd} \\ \Theta(\sqrt{\kappa}) & \text{hb, na} \end{cases} \tag{3.9}$$

where $a = \Theta(b)$ means that a lies within constant factors of b as $b \to \infty$. This improvement makes accelerated algorithms popular for problems with large condition number κ.

While convergence rate is only affected by the largest and smallest eigenvalues of Q, the entire spectrum of Q influences the noise amplification J. On the other hand, the largest and smallest values of J over the function class $\mathcal{Q}_m^L$,

$$J_{\max} \;:=\; \max_{f \,\in\, \mathcal{Q}_m^L} J, \qquad J_{\min} \;:=\; \min_{f \,\in\, \mathcal{Q}_m^L} J \tag{3.10}$$

depend only on the noise magnitude σ_w, the algorithmic parameters (α, β, γ), the problem dimension n, and the extreme eigenvalues m and L of Q.

For the parameters that optimize convergence rates, tight upper and lower bounds on the noise amplification were developed in [93, Theorem 4]. These bounds are expressed in terms of the condition number κ and the problem dimension n, and they demonstrate opposite trends relative to the settling time. In particular, for gradient descent,

$$J_{\max} \;=\; \sigma_w^2 n \Theta(\kappa), \qquad J_{\min} \;=\; \sigma_w^2 (\Theta(\kappa) + n) \tag{3.11a}$$

and for accelerated algorithms,

$$J_{\max} \;=\; \sigma_w^2 n \Theta(\kappa\sqrt{\kappa}), \qquad J_{\min} = \begin{cases} \sigma_w^2 (\Theta(\kappa\sqrt{\kappa}) \,+\, n\Theta(\sqrt{\kappa})) & \text{hb} \\ \sigma_w^2 (\Theta(\kappa\sqrt{\kappa}) \,+\, n) & \text{na.} \end{cases} \tag{3.11b}$$

We observe that for fixed problem dimension n and noise magnitude σ_w, the accelerated algorithms increase noise amplification by a factor of $\Theta(\sqrt{\kappa})$ relative to gradient descent for the parameters that optimize convergence rates. While similar result also holds for heavy-ball and Nesterov's algorithms with arbitrary values of parameters α and β that provide settling time $T_s \leq c\sqrt{\kappa}$ with $c > 0$ [93, Theorem 8], in this chapter we establish fundamental tradeoffs between noise amplification and settling time for the class of the two-step momentum algorithms (3.2) with arbitrary stabilizing values of constant parameters.

3.3 Summary of main results

In this section, we summarize our key contributions regarding tradeoffs between robustness and convergence of noisy two-step momentum algorithm (3.2). In addition, our geometric characterization of stability and ρ-linear convergence allows us to provide alternative proofs of standard convergence results and quantify fundamental performance tradeoffs. The proofs of results presented here can be found in Section 3.7.

3.3.1 Bounded noise amplification for stabilizing parameters

For a discrete-time LTI system with a convergence rate ρ, the distance of the eigenvalues to the unit circle is lower bounded by $1-\rho$. We use this stability margin to establish an upper bound on the noise amplification J of the two-step momentum method (3.2) for *any* stabilizing parameters (α, β, γ).

Theorem 1 *Let the parameters (α, β, γ) be such that the two-step momentum algorithm in (3.2) converges linearly with the rate $\rho = 1 - 1/T_s$ for all $f \in \mathcal{Q}_m^L$. Then,*

$$J \;\leq\; \frac{\sigma_w^2(1\,+\,\rho^2)}{(1\,+\,\rho)^3}\, n\, T_s^3 \tag{3.12a}$$

where n is the problem size. Furthermore, for the gradient noise model ($\sigma_w = \alpha\sigma$),

$$J \;\leq\; \frac{\sigma^2(1\,+\,\rho)(1\,+\,\rho^2)}{L^2}\, n\, T_s^3. \tag{3.12b}$$

For $\rho < 1$, both upper bounds in (3.12) scale with nT_s^3 and they are exact for the heavy-ball method with the parameters that optimize the convergence rate provided by Table 3.1. However, these bounds are not tight for all stabilizing parameters; e.g., applying (3.12a) to gradient descent with the optimal stepsize $\alpha = 2/(L+m)$ yields $J \leq \sigma_w^2 n \Theta(\kappa^3)$, which is off by a factor of κ^2; cf. Table 3.1. The bound in (3.12b) is obtained by combining (3.12a) with $\alpha L \leq (1+\rho)^2$, which follows from the conditions for ρ-linear convergence in Section 3.4.

3.3.2 Tradeoff between settling time and noise amplification

In this subsection, we establish lower bounds on the products $J_{\max} \times T_s$ and $J_{\min} \times T_s$ for any stabilizing constant parameters (α, β, γ) in the two-step momentum algorithm (3.2), where $J_{\max}$ and $J_{\min}$ defined in (3.10) are the largest and the smallest noise amplification for the class of functions $\mathcal{Q}_m^L$ and T_s is the settling time.

Theorem 2 *Let the parameters (α, β, γ) be such that the two-step momentum algorithm in* (3.2) *converges linearly with the rate $\rho = 1 - 1/T_s$ for all $f \in \mathcal{Q}_m^L$. Then, $J_{\max}$ and $J_{\min}$ in* (3.10) *satisfy,*

$$J_{\max} \times T_s \;\geq\; \sigma_w^2 \left((n \,-\, 1) \frac{\kappa^2}{64} \,+\, \frac{\sqrt{\kappa}+1}{2} \right) \tag{3.13a}$$

$$J_{\min} \times T_s \;\geq\; \sigma_w^2 \left(\frac{\kappa^2}{64} \,+\, (n \,-\, 1) \frac{\sqrt{\kappa}+1}{2} \right). \tag{3.13b}$$

Furthermore, for the gradient noise model ($\sigma_w = \alpha\sigma$), we have

$$J_{\max} \times T_s \;\geq\; \frac{\sigma^2}{L^2} \left((n \,-\, 1) \frac{\kappa^2}{4} \,+\, \max\left\{ \frac{\kappa^2}{T_s^3},\, \frac{1}{4} \right\} \right) \tag{3.13c}$$

$$J_{\min} \times T_s \;\geq\; \frac{\sigma^2}{L^2} \left(\frac{\kappa^2}{4} \,+\, (n \,-\, 1) \max\left\{ \frac{\kappa^2}{T_s^3},\, \frac{1}{4} \right\} \right). \tag{3.13d}$$

For both noise models, the condition number κ restricts the performance of the two-step momentum algorithm with constant parameters: *for a fixed problem size n, all four lower bounds in* (3.13) *scale with κ^2.* Relative to the dominant term in κ, the problem dimension n appears in a multiplicative fashion for the lower bounds on $J_{\max}$ and in an additive fashion for the lower bounds on $J_{\min}$. Next, by establishing upper bounds on $J_{\max} \times T_s$ and $J_{\min} \times T_s$ for a parameterized family of heavy-ball-like algorithms in Theorem 3, we prove that for any settling time T_s these bounds are *order-wise tight* (in κ) for the gradient noise model. On the other hand, for the iterate noise model, they are tight only if T_s is smaller than the best achievable settling time of gradient descent, $(\kappa + 1)/2$.

Theorem 3 *For the class of strongly convex quadratic functions $\mathcal{Q}_m^L$ with the condition number $\kappa = L/m$, let the scalar ρ be such that the fundamental lower bound $T_s = 1/(1 - \rho) \geq (\sqrt{\kappa} + 1)/2$ given by* (3.7) *holds. Then, the two-step momentum algorithm* (3.2) *with parameters*

$$\alpha \;=\; \frac{(1 \,+\, \rho)(1 \,+\, \beta/\rho)}{L}, \;\; \beta \;=\; \rho\, \frac{\kappa \,-\, (1 \,+\, \rho)/(1 \,-\, \rho)}{\kappa \,+\, (1 \,+\, \rho)/(1 \,-\, \rho)}, \;\; \gamma \;=\; 0 \tag{3.14}$$

converges linearly with the rate ρ and, for settling times $T_s \leq (\kappa + 1)/2$, $J_{\max}$ and $J_{\min}$ in (3.10) *satisfy*

$$J_{\max} \times T_s \;\leq\; \sigma_w^2\, n\kappa\, (\kappa \,+\, 1)/2 \tag{3.15a}$$

$$J_{\min} \times T_s \;\leq\; \sigma_w^2\, \kappa\, (\kappa \,+\, n \,-\, 1)\,. \tag{3.15b}$$

Furthermore, for the gradient noise model ($\sigma_w = \alpha\sigma$) and any settling time that satisfies the inequality in (3.7)*, parameters* (3.14) *lead to*

$$J_{\max} \times T_s \;\leq\; \sigma^2\, n\kappa\, (\kappa \,+\, 1)/L^2 \tag{3.15c}$$

$$J_{\min} \times T_s \;\leq\; \sigma^2\, 2\kappa\, (\kappa \,+\, 4n \,-\, 7)/L^2. \tag{3.15d}$$

Theorem 3 provides upper bounds on $J_{\max} \times T_s$ and $J_{\min} \times T_s$ for a family of heavy-ball-like algorithms ($\gamma = 0$) parameterized by the settling time T_s. For both noise models, the upper bounds in (3.15) scale with κ^2 which matches the corresponding lower bounds in (3.13). For the gradient noise model, the upper and lower bounds are order-wise tight (in κ) for any settling time. However, for the iterate noise model, the lower bounds in Theorem 2 can be improved when $T_s \geq (\kappa+1)/2$. In Theorem 4, we establish alternative lower bounds on $J_{\max}$ and $J_{\min}$ that scale with T_s for the two-step momentum algorithm (3.2) with any stabilizing parameters. We also utilize parameterized family (3.14) of heavy-ball-like algorithms with negative momentum parameter β to increase T_s beyond $(\kappa+1)/2$ and obtain upper bounds on $J_{\max}$ and $J_{\min}$ that scale linearly with T_s for the iterate noise model.

Theorem 4 *Let the parameters (α, β, γ) be such that the two-step momentum algorithm in (3.2) achieves the convergence rate $\rho = \mathsf{\rho}(A) = 1 - 1/T_s$, where the matrix A is given by (3.4). Then, $J_{\max}$ and $J_{\min}$ in (3.10) satisfy,*

$$J_{\max} \;\geq\; \sigma_w^2 \left((n \,-\, 1)\, \frac{T_s}{2(1 \,+\, \rho)^2} \,+\, 1 \right) \tag{3.16a}$$

$$J_{\min} \;\geq\; \sigma_w^2 \left(\frac{T_s}{2(1 \,+\, \rho)^2} \,+\, (n \,-\, 1) \right). \tag{3.16b}$$

Furthermore, for the parameterized family of heavy-ball-like algorithms (3.14) with $T_s \geq (\kappa+1)/2$,

$$J_{\max} \;\leq\; \sigma_w^2 n\, T_s \tag{3.17a}$$

$$J_{\min} \;\leq\; 2\sigma_w^2 (1 + (n-2)/\kappa)\, T_s. \tag{3.17b}$$

We note that the condition $T_s \geq (\kappa+1)/2$ in Theorem 4 corresponds to non-positive momentum parameter $\beta \leq 0$. We also observe that both upper and lower bounds on $J_{\max}$ and $J_{\min}$ in Theorem 4 grow linearly with T_s and that for the iterate noise model with $T_s \geq (\kappa+1)/2$ the lower bound is sharper than the one established in Theorem 2.

Remark 3 *Since $\mathcal{Q}_m^L$ is a subset of the class of m-strongly convex functions with L-Lipschitz continuous gradients, the fundamental lower bounds on $J_{\max} \times T_s$ established in Theorem 2 carry over to this broader class of problems. Thus, the restriction imposed by the condition number on the tradeoff between settling time and noise amplification goes beyond $\mathcal{Q}_m^L$ and holds for general strongly convex problems.*

Remark 4 *The upper bounds in Theorems 3 and 4 are obtained for a particular choice of constant parameters. Thus, they also provide upper bounds on the best achievable noise amplification bounds $J_{\max}^\star := \min_{\alpha,\beta,\gamma} J_{\max}$ and $J_{\min}^\star := \min_{\alpha,\beta,\gamma} J_{\min}$ for a settling time T_s; see Figure 3.1.*

Iterate noise $\sigma_w = 1$	$J \;\le\; \frac{1+\rho^2}{(1+\rho)^3}\, n\, T_s^3$	(3.12a) ⊗
	$J^\star_{\max} \;\le\; \begin{cases} \frac{1}{2}\, n\kappa\,(\kappa+1)\, T_s^{-1} & \text{if} \quad T_s \le (\kappa+1)/2 \\ n\, T_s & \text{if} \quad T_s \ge (\kappa+1)/2 \end{cases}$	(3.15a) ■ (3.17a)
	$J^\star_{\max} \;\ge\; \max\left\{\left(\frac{1}{64}(n-1)\kappa^2 + \frac{\sqrt{\kappa}+1}{2}\right) T_s^{-1}, \frac{1}{8}(n-1)T_s + 1\right\}$	(3.13a) ▲ (3.16a)
	$J^\star_{\min} \;\le\; \begin{cases} \kappa\,(\kappa+n-1)\, T_s^{-1} & \text{if} \quad T_s \le (\kappa+1)/2 \\ 2\,(1+\frac{n-2}{\kappa})\, T_s & \text{if} \quad T_s \ge (\kappa+1)/2 \end{cases}$	(3.15b) □ (3.17b)
	$J^\star_{\min} \;\ge\; \max\left\{\left(\frac{1}{64}\kappa^2 + (n-1)\frac{\sqrt{\kappa}+1}{2}\right) T_s^{-1}, \frac{1}{8}\, T_s + n - 1\right\}$	(3.13b) △ (3.16b)

Gradient noise $\sigma_w = \alpha$	$J \;\le\; \frac{(1+\rho)(1+\rho^2)}{L^2}\, n\, T_s^3$	(3.12b) ⊗
	$J^\star_{\max} \;\le\; \frac{1}{L^2}\, n\kappa\,(\kappa+1)\, T_s^{-1}$	(3.15c) ■
	$J^\star_{\max} \;\ge\; \frac{1}{L^2}\left(\frac{1}{4}(n-1)\kappa^2 + \max\{\kappa^2/T_s^3, 1/4\}\right) T_s^{-1}$	(3.13d) ▲
	$J^\star_{\min} \;\le\; \frac{2}{L^2}\,\kappa\,(\kappa+4n-7)\, T_s^{-1}$	(3.15d) □
	$J^\star_{\min} \;\ge\; \frac{1}{L^2}\left(\frac{1}{4}\kappa^2 + (n-1)\max\{\kappa^2/T_s^3, 1/4\}\right) T_s^{-1}$	(3.13c) △

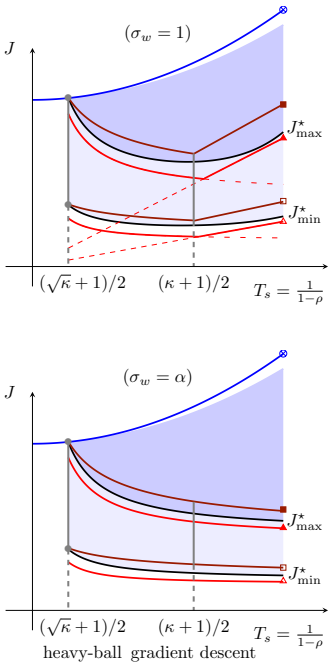


Figure 3.1: Summary of the results established in Theorems 1-4 for $\sigma^2 = 1$. The top and bottom rows correspond to the iterate and gradient noise models, respectively, and they illustrate (i) $J^\star_{\max} := \min_{\alpha,\beta,\gamma} \max_f J$ and $J^\star_{\min} := \min_{\alpha,\beta,\gamma} \min_f J$ subject to a settling time T_s for $f \in \mathcal{Q}_m^L$ (black curves); and (ii) their corresponding upper (maroon curves) and lower (red curves) bounds in terms of the condition number $\kappa = L/m$, problem size n, and settling time T_s. The upper bounds on J established in Theorem 1 are marked by blue curves. The dark shaded region and its union with the light shaded region respectively correspond to all possible pairs $(T_s, \max_f J)$ and $(T_s, \min_f J)$ for $f \in \mathcal{Q}_m^L$ and any stabilizing parameters (α, β, γ).

3.4 Geometric characterization

In this section, we examine the relation between the convergence rate and noise amplification of the two-step momentum algorithm (3.2) for strongly convex quadratic problems. In particular, the eigenvalue decomposition of the Hessian matrix Q allows us to bring the dynamics into n decoupled second-order systems parameterized by the eigenvalues of Q and the algorithmic parameters (α, β, γ). We utilize the Jury stability criterion to provide a novel geometric characterization of stability and ρ-linear convergence and exploit this insight to derive alternative proofs of standard convergence results and quantify fundamental performance tradeoffs.

3.4.1 Modal decomposition

We utilize the eigenvalue decomposition of the Hessian matrix $Q = Q^T \succ 0$, $Q = V\Lambda V^T$, where Λ is the diagonal matrix of the eigenvalues and V is the orthogonal matrix of the corresponding eigenvectors. The change of variables $\hat{x}^t := V^T(x^t - x^\star)$ and $\hat{w}^t := V^T w^t$ allows us to bring (3.4) into n decoupled second-order subsystems,

$$\begin{aligned} \hat{\psi}_i^{t+1} &= \hat{A}_i \hat{\psi}_i^t + \hat{B}_i \hat{w}_i^t \\ \hat{z}_i^t &= \hat{C}_i \hat{\psi}_i^t \end{aligned} \tag{3.18a}$$

where $\hat{w}_i^t$ is the ith component of the vector $\hat{w}^t \in \mathbb{R}^n$, $\hat{\psi}_i^t = \left[\begin{array}{cc} \hat{x}_i^t & \hat{x}_i^{t+1} \end{array}\right]^T$,

$$\hat{A}_i = \hat{A}(\lambda_i) := \left[\begin{array}{cc} 0 & 1 \\ -a(\lambda_i) & -b(\lambda_i) \end{array}\right], \quad \hat{B}_i = \left[\begin{array}{cc} 0 & \sigma_w \end{array}\right]^T, \quad \hat{C}_i = \left[\begin{array}{cc} 1 & 0 \end{array}\right] \tag{3.18b}$$

and

$$a(\lambda) := \beta - \gamma\alpha\lambda, \; b(\lambda) := (1+\gamma)\alpha\lambda - (1+\beta). \tag{3.18c}$$

3.4.2 Conditions for linear convergence

For the class of strongly convex quadratic functions $\mathcal{Q}_m^L$, the best convergence rate ρ is determined by the largest spectral radius of the matrices $\hat{A}(\lambda)$ in (3.18) for $\lambda \in [m, L]$,

$$\rho = \max_{\lambda \in [m,L]} \; \rho(\hat{A}(\lambda)). \tag{3.19}$$

For the heavy-ball and Nesterov's accelerated methods, analytical expressions for $\rho(\hat{A}(\lambda))$ were developed and algorithmic parameters that optimize convergences rate were obtained in [52]. Unfortunately, these expressions do not provide insight into the relation between convergence rates and noise amplification.

In this chapter, we ask the dual question:

- *For a fixed convergence rate ρ, what is the largest condition number κ that can be handled by the two-step momentum algorithm* (3.2) *with constant parameters* (α, β, γ)?

We note that the matrices $\hat{A}(\lambda)$ share the same structure as

$$M \;=\; \begin{bmatrix} 0 & 1 \\ -a & -b \end{bmatrix} \tag{3.20a}$$

with the real scalars a and b and that the characteristic polynomial associated with the matrix M is given by

$$F(z) \;:=\; \det\left(zI \,-\, M\right) \;=\; z^2 \,+\, b\,z \,+\, a. \tag{3.20b}$$

We next utilize the Jury stability criterion [116, Chap. 4-3] to provide conditions for stability of the matrix M given by (3.20a).

Lemma 1 *For the matrix $M \in \mathbb{R}^{2\times 2}$ given by* (3.20a)*,*

$$\rho(M) \,<\, 1 \;\;\Longleftrightarrow\;\; (b,a) \,\in\, \Delta \tag{3.21a}$$

where the stability set

$$\Delta \;:=\; \{(b,a) \mid |b| \,-\, 1 \,<\, a \,<\, 1\} \tag{3.21b}$$

is an open triangle in the (b,a)-plane with vertices

$$X \,=\, (-2,1), \quad Y \,=\, (2,1), \quad Z \,=\, (0,-1). \tag{3.21c}$$

Proof: The characteristic polynomial $F(z)$ associated with the matrix M is given by equation (3.20b) and the Jury stability criterion [116, Chap. 4-3] provides necessary and sufficient conditions for stability,

$$|a| \,<\, 1,\; F(\pm 1) \,=\, 1 \,\pm\, b \,+\, a \,>\, 0.$$

The condition $a > -1$ is ensured by the positivity of $F(\pm 1)$. □

For any $\rho > 0$, the spectral radius $\rho(M)$ of the matrix M is smaller than ρ if and only if $\rho(M/\rho)$ is smaller than 1. This observation in conjunction with Lemma 1 allow us to obtain necessary and sufficient conditions for stability with the linear convergence rate ρ of the two-step momentum algorithm (3.2).

Lemma 2 *For any positive scalar $\rho < 1$ and the matrix $M \in \mathbb{R}^{2\times 2}$ given by* (3.20a)*, we have*

$$\rho(M) \,\leq\, \rho \;\;\Longleftrightarrow\;\; (b,a) \,\in\, \Delta_\rho \tag{3.22a}$$

where the ρ-linear convergence set

$$\Delta_\rho \;:=\; \left\{(b,a) \mid \rho\left(|b| \,-\, \rho\right) \,\leq\, a \,\leq\, \rho^2\right\} \tag{3.22b}$$

is a closed triangle in the (b,a)-plane with vertices

$$X_\rho \,=\, (-2\rho, \rho^2),\; Y_\rho \,=\, (2\rho, \rho^2),\; Z_\rho \,=\, (0, -\rho^2). \tag{3.22c}$$

Proof: See Appendix B.3. □

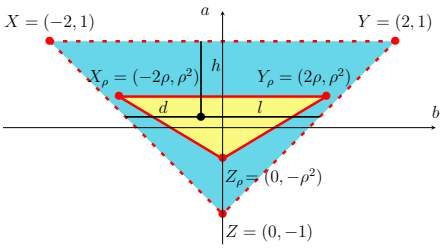


Figure 3.2: The stability set Δ (the open, cyan triangle) in (3.21b) and the ρ-linear convergence set Δ_ρ (the closed, yellow triangle) in (3.22b) along with the corresponding vertices. For the point (b, a) (black dot) associated with the matrix M in (3.20a), the corresponding distances (d, h, l) in (3.29) are marked by black lines.

Figure 3.2 illustrates the stability and the ρ-linear convergence sets Δ and Δ_ρ. We note that for any $\rho \in (0, 1)$, we have $\Delta_\rho \subset \Delta$. This can be verified by observing that the vertices (X_ρ, Y_ρ, Z_ρ) of Δ_ρ all lie in Δ.

For the two-step momentum algorithm (3.2), the functions $a(\lambda)$ and $b(\lambda)$ given by (3.18c) satisfy the affine relation,

$$(1 + \gamma)\, a(\lambda) \;+\; \gamma\, b(\lambda) \;=\; \beta \;-\; \gamma. \tag{3.23}$$

This fact in conjunction with Lemmas 1 and 2 allow us to derive conditions for stability and the convergence rate.

Lemma 3 *The two-step momentum algorithm* (3.2) *with constant parameters* (α, β, γ) *is stable for all functions* $f \in \mathcal{Q}_m^L$ *if and only if the following equivalent conditions hold:*

1. $(b(\lambda), a(\lambda)) \in \Delta$ *for all* $\lambda \in [m, L]$;
2. $(b(\lambda), a(\lambda)) \in \Delta$ *for* $\lambda \in \{m, L\}$.

Furthermore, the linear convergence rate $\rho < 1$ *is achieved for all functions* $f \in \mathcal{Q}_m^L$ *if and only if the following equivalent conditions hold:*

1. $(b(\lambda), a(\lambda)) \in \Delta_\rho$ *for all* $\lambda \in [m, L]$;
2. $(b(\lambda), a(\lambda)) \in \Delta_\rho$ *for* $\lambda \in \{m, L\}$.

Here, $(b(\lambda), a(\lambda))$ *is given by* (3.18c)*, and the stability and* ρ*-linear convergence triangles* Δ *and* Δ_ρ *are given by* (3.21b) *and* (3.22b)*, respectively.*

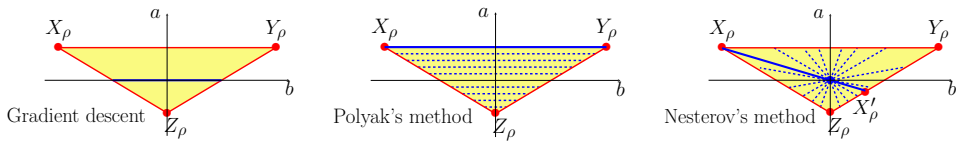


Figure 3.3: For a fixed ρ-linear convergence triangle Δ_ρ (yellow), dashed blue lines mark the line segments $(b(\lambda), a(\lambda))$ with $\lambda \in [m, L]$ for gradient descent, Polyak's heavy-ball, and Nesterov's accelerated methods as particular instances of the two-step momentum algorithm (3.2) with constant parameters. The solid blue line segments correspond to the parameters for which the algorithm achieves rate ρ for the largest possible condition number given by (3.28).

Proof: The conditions in 1) follow from combing (3.19) with Lemma 1 (for stability) and with Lemma 2 (for ρ-linear convergence). The conditions in 2) follow from the facts that Δ and Δ_ρ are convex sets and that $(b(\lambda), a(\lambda))$ is a line segment in the (b, a)-plane with end points corresponding to $\lambda = m$ and $\lambda = L$. □

Lemma 3 exploits the affine relation (3.23) between $a(\lambda)$ and $b(\lambda)$ and the convexity of the sets Δ and Δ_ρ to establish necessary and sufficient conditions for stability and ρ-linear convergence: *the inclusion of the end points of the line segment $(b(\lambda), a(\lambda))$ associated with the extreme eigenvalues m and L of the matrix Q in the corresponding triangle.* A similar approach was taken in [109, Appendix A.1], where the affine nature of the conditions resulting from the Jury stability criterion with respect to λ was used to conclude that $\rho(\hat{A}(\lambda))$ is a quasi-convex function of λ and show that the extreme points m and L determine $\rho(A)$. In contrast, we exploit the triangular shapes of the stability and ρ-linear convergence sets Δ and Δ_ρ and utilize this geometric insight to identify the parameters that optimize the convergence rate and to establish tradeoffs between noise amplification and convergence rate.

The following corollary is immediate.

Corollary 1 *Let the two-step momentum algorithm* (3.2) *with constant parameters* (α, β, γ) *minimize a function* $f \in \mathcal{Q}_m^L$ *with a linear rate* $\rho < 1$. *Then, the convergence rate* ρ *is achieved for all functions* $f \in \mathcal{Q}_m^L$.

Proof: Lemma 3 implies that only the extreme eigenvalues m and L of Q determine ρ. Since all functions $f \in \mathcal{Q}_m^L$ share the same extreme eigenvalues, this completes the proof. □

For the two-step momentum algorithm (3.2) with constant parameters, Lemma 3 leads to a simple alternative proof for the fundamental lower bound (3.7) on the settling time established by Nesterov. Our proof utilizes the fact that for any point $(b(\lambda), a(\lambda)) \in \Delta_\rho$, the horizontal signed distance to the edge XZ of the stability triangle Δ satisfies

$$d(\lambda) \;:=\; a(\lambda) + b(\lambda) + 1 \;=\; \alpha\lambda. \tag{3.24}$$

where a and b are given by (3.18c); see Figure 3.2 for an illustration.

Proposition 1 *Let the two-step momentum algorithm in* (3.2) *with constant parameters* (α, β, γ) *achieve the linear convergence rate* $\rho < 1$ *for all functions* $f \in \mathcal{Q}_m^L$. *Then, lower bound* (3.7) *on the settling time holds and it is achieved by the heavy-ball method with the parameters provided in Table 3.1.*

Proof: Let $d(m) = \alpha m$ and $d(L) = \alpha L$ denote the values of the function $d(\lambda)$ associated with the points $(b(m), a(m))$ and $(b(L), a(L))$, where (b, a) and d are given by (3.18c) and (3.24), respectively. Lemma 3 implies that $(b(L), a(L))$ and $(b(m), a(m))$ lie in the ρ-linear convergence triangle Δ_ρ. Thus,

$$d_{\max}/d_{\min} \;\geq\; d(L)/d(m) \;=\; \kappa \tag{3.25}$$

where $d_{\max}$ and $d_{\min}$ are the largest and smallest values of d among all points $(b, a) \in \Delta_\rho$. From the shape of Δ_ρ, we conclude that $d_{\max}$ and $d_{\min}$ correspond to the vertices Y_ρ and X_ρ of Δ_ρ given by (3.22c); see Figure 3.2. Thus,

$$d_{\max} \;=\; d_{Y_\rho} \;=\; 1 + \rho^2 + 2\,\rho \;=\; (1 + \rho)^2 \tag{3.26a}$$

$$d_{\min} \;=\; d_{X_\rho} \;=\; 1 + \rho^2 - 2\,\rho \;=\; (1 - \rho)^2. \tag{3.26b}$$

Combining (3.25) with (3.26) yields

$$\kappa \;=\; \frac{d(L)}{d(m)} \;\leq\; \frac{d_{\max}}{d_{\min}} \;=\; \frac{(1 + \rho)^2}{(1 - \rho)^2}. \tag{3.27}$$

Rearranging terms in (3.27) gives lower bound (3.7). □

To provide additional insight, we next examine the implications of Lemma 3 for gradient descent, Polyak's heavy-ball, and Nesterov's accelerated algorithms. In all three cases, our dual approach recovers the optimal convergence rates provided in Table 3.1. From the affine relation (3.23), it follows that $(b(\lambda), a(\lambda))$ with $\lambda \in [m, L]$ for,

- gradient descent ($\beta = \gamma = 0$), is a horizontal line segment parameterized by $a(\lambda) = 0$;
- heavy-ball method ($\gamma = 0$), is a horizontal line segment parameterized by $a(\lambda) = \beta$;
- Nesterov's accelerated method ($\beta = \gamma$), is a line segment parameterized by $a(\lambda) = -\beta b(\lambda)/(1 + \beta)$.

These observations are illustrated in Figure 3.3 and, as we show in the proof of Lemma 3, to obtain the largest possible condition number for which the convergence rate ρ is feasible for each algorithm, one needs to find the largest ratio $d(L)/d(m) = \kappa$ among all possible orientations for the line segment $(b(\lambda), a(\lambda))$ with $\lambda \in [m, L]$ to lie within Δ_ρ. This leads to the following conditions:

- For gradient descent, the largest ratio $d(L)/d(m)$ corresponds to the intersections of the horizontal axis and the edges $Y_\rho Z_\rho$ and $X_\rho Z_\rho$ of the triangle Δ_ρ, which are given by $(\rho, 0)$ and $(-\rho, 0)$, respectively. Thus, we have

$$\kappa \;=\; d(L)/d(m) \;\leq\; (1 + \rho)/(1 - \rho). \tag{3.28a}$$

  Rearranging terms in (3.28a) yields a lower bound on the settling time for gradient descent $1/(1-\rho) \geq (\kappa+1)/2$. This lower bound is tight as it can be achieved by choosing the parameters in Table 3.1, which place $(b(\lambda), a(\lambda))$ to $(\rho, 0)$ and $(-\rho, 0)$ for $\lambda = L$ and $\lambda = m$, respectively.

- For the heavy-ball method, the optimal rate is recovered by designing the parameters (α, β) such that the vertices X_ρ and Y_ρ belong to the line segment $(b(\lambda), a(\lambda))$,

$$\kappa \;=\; d(L)/d(m) \;\leq\; (1 + \rho)^2/(1 - \rho)^2. \tag{3.28b}$$

  By choosing $d(L) = d_{Y_\rho}$ and $d(m) = d_{X_\rho}$, we recover the optimal parameters provided in Table 3.1 and achieve the fundamental lower bound (3.7) on the convergence rate.

- For Nesterov's accelerated method, the largest ratio $d(L)/d(m)$ corresponds to the line segment $X_\rho X'_\rho$ that passes through the origin, where $X'_\rho = (2\rho/3, -\rho^2/3)$ lies on the edge $Y_\rho Z_\rho$; see Appendix B.3. This yields

$$\kappa \;=\; d(L)/d(m) \;\leq\; (1 + \rho)(3 - \rho)/(3(1 - \rho)^2). \tag{3.28c}$$

  Rearranging terms in this inequality provides a lower bound on the settling time $1/(1-\rho) \geq \sqrt{3\kappa+1}/2$. This lower bound is tight and it can be achieved with the parameters provided in Table 3.1, which place $(b(L), a(L))$ to X'_ρ and $(b(m), a(m))$ to X_ρ.

Figure 3.3 illustrates the optimal orientations discussed above.

3.4.3 Noise amplification

To quantify the noise amplification of the two-step momentum algorithm (3.2), we utilize an alternative characterization of the stability and ρ-linear convergence triangles Δ and Δ_ρ. As illustrated in Figure 3.2, let d and l denote the horizontal signed distances of the point (a, b) to the edges XZ and YZ of the stability triangle Δ,

$$\begin{aligned} d(\lambda) \;&:=\; a(\lambda) \,+\, b(\lambda) \,+\, 1 \\ l(\lambda) \;&:=\; a(\lambda) \,-\, b(\lambda) \,+\, 1. \end{aligned} \tag{3.29a}$$

and let h denote its vertical signed distance to the edge XY,

$$h(\lambda) \;:=\; 1 \,-\, a(\lambda). \tag{3.29b}$$

Then, the following equivalence conditions,

$$(b,a) \in \Delta \iff h,\, d,\, l > 0 \tag{3.30a}$$

$$(b,a) \in \Delta_\rho \iff \begin{cases} h \geq (1-\rho)(1+\rho) \\ d \geq (1-\rho)(1+\rho+b) \\ l \geq (1-\rho)(1+\rho-b) \end{cases} \tag{3.30b}$$

follow from the definition of the sets Δ in (3.21b), Δ_ρ in (3.22b), and (h,d,l) in (3.29).

In Theorem 5, we quantify the steady-state variance of the error in the optimization variable in terms of the spectrum of the Hessian matrix and the algorithmic parameters for noisy two-step momentum algorithm (3.2). Special cases of this result for gradient decent, heavy-ball, and Nesterov's accelerated algorithms were established in [93]. The proof of Theorem 5 follows from similar arguments and we omit it for brevity.

Theorem 5 *For a strongly convex quadratic objective function $f \in \mathcal{Q}_m^L$ with the Hessian matrix Q, the steady-state variance of $x^t - x^\star$ for the two-step momentum algorithm* (3.2) *with any stabilizing parameters (α,β,γ) is determined by*

$$J \;=\; \sum_{i=1}^{n} \frac{\sigma_w^2 (d(\lambda_i) + l(\lambda_i))}{2\, d(\lambda_i)\, h(\lambda_i)\, l(\lambda_i)} \;=:\; \sum_{i=1}^{n} \hat{J}(\lambda_i)$$

Here, $\hat{J}(\lambda_i)$ denotes the modal contribution of the ith eigenvalue λ_i of Q to the steady-state variance, (d,h,l) are defined in (3.29)*, and (a,b) are given by* (3.18c)*.*

In Appendix B.6, we describe how the algebraic Lyapunov equation for the steady-state covariance matrix of the error in the optimization variable can be used to compute the noise amplification J. Theorem 5 demonstrates that J depends on the entire spectrum of the Hessian matrix Q and not only on its extreme eigenvalues m and L, which determine the convergence rate. Since for any $f \in \mathcal{Q}_m^L$ the extreme eigenvalues of Q are fixed at m and L, we have

$$\begin{aligned} J_{\max} &:= \max_{f \in \mathcal{Q}_m^L} J \;=\; \hat{J}(m) \,+\, \hat{J}(L) \,+\, (n-2)\hat{J}_{\max} \\ J_{\min} &:= \min_{f \in \mathcal{Q}_m^L} J \;=\; \hat{J}(m) \,+\, \hat{J}(L) \,+\, (n-2)\hat{J}_{\min} \end{aligned} \tag{3.31a}$$

where

$$\hat{J}_{\max} := \max_{\lambda \in [m,L]} \hat{J}(\lambda), \qquad \hat{J}_{\min} := \min_{\lambda \in [m,L]} \hat{J}(\lambda). \tag{3.31b}$$

We use these expressions to determine explicit upper and lower bounds on $J_{\max}$ and $J_{\min}$ in terms of the condition number and the settling time.

3.5 Designing order-wise Pareto-optimal algorithms with adjustable parameters

We now utilize the geometric insight developed in Section 3.4 to design algorithm parameters that tradeoff settling time and noise amplification. In particular, we introduce two instances

of parameterized families of heavy-ball-like ($\gamma = 0$) and Nesterov-like ($\gamma = \beta$) algorithms that provide *continuous transformations* from gradient descent to the corresponding accelerated algorithm (with the optimal convergence rate) via a homotopy path parameterized by the settling time T_s. For both the iterate and gradient noise models, we establish an order-wise tight scaling $\Theta(\kappa^2)$ for $J_{\max} \times T_s$ and $J_{\min} \times T_s$ in accelerated regime (i.e., when T_s is smaller than the settling time of gradient descent with the optimal stepsize, $(\kappa + 1)/2$). This is a direct extension of [93, Theorem 4] which studied gradient descent and its accelerated variants for the parameters that optimize the corresponding convergence rates.

We also examine performance tradeoffs for the parameterized family of heavy-ball-like algorithms with negative momentum parameter $\beta < 0$. This decelerated regime corresponds to settling times larger than $(\kappa+1)/2$ and it captures a key difference between the two noise models: *for $T_s \geq (\kappa + 1)/2$, $J_{\max}$ and $J_{\min}$ grow linearly with the settling time T_s for the iterate noise model and they remain inversely proportional to T_s for the gradient noise model.* Comparison with the lower bounds in Theorems 2 and 4 shows that the parameterized family of heavy-ball-like methods yields order-wise optimal (in κ and T_s) $J_{\max}$ and $J_{\min}$ for both noise models. The results presented here prove all upper bounds in Theorems 3 and 4.

3.5.1 Parameterized family of heavy-ball-like methods

For the two-step momentum algorithm (3.2) with $\gamma = 0$, the line segment $(b(\lambda), a(\lambda))$ parameterized by $\lambda \in [m, L]$ is parallel to the b-axis in the (b, a)-plane and it satisfies $a(\lambda) = \beta$. As described in Section 3.4, gradient descent and heavy-ball methods with the optimal parameters provided in Table 3.1 are obtained for $\beta = 0$ and $\beta = \rho^2$, respectively, and the corresponding end points $(b(m), a(m))$ and $(b(L), a(L))$ lie at the edges $X_\rho Z_\rho$ and $Y_\rho Z_\rho$ of the ρ-linear convergence triangle Δ_ρ. Inspired by this observation, we propose a family of parameters for which $\beta = c\rho^2$, for some scalar $c \in [-1, 1]$, and determine the stepsize α such that the above end points lie at $X_\rho Z_\rho$ and $Y_\rho Z_\rho$,

$$\alpha \;=\; (1 \,+\, \rho)(1 \,+\, c\rho)/L, \;\; \beta \;=\; c\rho^2, \;\; \gamma \;=\; 0. \tag{3.32}$$

This yields a continuous transformation between the standard heavy-ball method ($c = 1$) and gradient descent ($c = 0$) for a fixed condition number κ. In addition, the momentum parameter β in (3.32) becomes negative for $c < 0$; see Figure 3.3 for an illustration. In Lemma 4, we provide expressions for the scalar c as well as for $\hat{J}_{\max}$ and $\hat{J}_{\min}$ defined in (3.31b) in terms of the condition number κ and the convergence rate ρ.

Lemma 4 *For the class of functions $\mathcal{Q}_m^L$ with the condition number $\kappa = L/m$, let the scalar ρ be such that*

$$T_s \;=\; 1/(1 \,-\, \rho) \;\geq\; (\sqrt{\kappa} \,+ 1\,)/2.$$

Then, the two-step momentum algorithm in (3.2) *with parameters in* (3.32) *achieves the convergence rate* ρ*, and the largest and smallest values* $\hat{J}_{\max}$ *and* $\hat{J}_{\min}$ *of* $\hat{J}(\lambda)$ *for* $\lambda \in [m, L]$ *satisfy*

$$\begin{aligned}\hat{J}_{\max} &= \hat{J}(m) = \hat{J}(L) = \frac{\sigma_w^2(\kappa+1)}{2(1-c\rho^2)(1+\rho)(1+c\rho)} \\ \hat{J}_{\min} &= \hat{J}(\hat{\lambda}) = \frac{\sigma_w^2}{(1+c\rho^2)(1-c\rho^2)}\end{aligned}$$

where $\hat{\lambda} := (m+L)/2$ *and the scalar* c *is given by*

$$c := \frac{\kappa - (1+\rho)/(1-\rho)}{\rho\,(\kappa + (1+\rho)/(1-\rho))} \in [-1,\, 1]. \tag{3.33}$$

Proof: See Appendix B.4. □

The parameters in (3.32) with c given by (3.33) are equivalent to the parameters presented in Theorem 3. Lemma 4 in conjunction with (3.31) allow us to derive analytical expressions for $J_{\max}$ and $J_{\min}$.

Corollary 2 *The parameterized family of heavy-ball-like methods* (3.32) *satisfies*

$$\begin{aligned}J_{\max} &= n\hat{J}(m) = n\hat{J}(L) \\ J_{\min} &= 2\hat{J}(m) + (n-2)\hat{J}(\hat{\lambda})\end{aligned}$$

where $\hat{J}(m)$ *and* $\hat{J}(\hat{\lambda})$ *are given in Lemma 4, and* $J_{\max}$ *and* $J_{\min}$ *defined in* (3.10) *are the largest and smallest values of* J *when the algorithm is applied to* $f \in \mathcal{Q}_m^L$ *with the condition number* $\kappa = L/m$*.*

The next proposition uses the analytical expressions in Corollary 2 to establish order-wise tight upper and lower bounds on $J_{\max}$ and $J_{\min}$ for the parameterized family of heavy-ball-like algorithms (3.32). Our upper and lower bounds are within constant factors of each other and they are expressed in terms of the problem size n, condition number κ, and settling time T_s.

Proposition 2 *For the parameterized family of heavy-ball-like algorithms in* (3.32)*,* $J_{\max}$ *and* $J_{\min}$ *in* (3.10) *satisfy,*

$$J_{\max} \times T_s = \sigma_w^2\, p_{1c}(\rho)\, n\, \kappa(\kappa+1) \tag{3.34a}$$

$$J_{\min} \times T_s = \sigma_w^2\, \kappa\,(2\, p_{1c}(\rho)\,(\kappa+1) + (n-2)\, p_{2c}(\rho))\,. \tag{3.34b}$$

Furthermore, for the gradient noise model ($\sigma_w = \alpha\sigma$)*,*

$$J_{\max} \times T_s = \sigma^2 p_{3c}(\rho)\, n\, \kappa(\kappa+1) \tag{3.35a}$$

$$J_{\min} \times T_s = \sigma^2 \kappa\,(2\, p_{3c}(\rho)\,(\kappa+1) + (n-2)\, p_{4c}(\rho)) \tag{3.35b}$$

where

$$\begin{aligned} p_{1c}(\rho) &:= q_c(\rho)/(2(1+\rho)^2(1+c\rho)^2), & p_{2c}(\rho) &:= q_c(\rho)/((1+\rho)(1+c\rho^2)(1+c\rho)) \\ p_{3c}(\rho) &:= q_c(\rho)/(2L^2), & p_{4c}(\rho) &:= q_c(\rho)q_{-c}(\rho)(1+\rho)/L^2 \end{aligned} \tag{3.36}$$

and $q_c(\rho) := (1-c\rho)/(1-c\rho^2)$. *In addition, for* $c \in [0,1]$, $p_{1c}(\rho) \in [1/64, 1/2]$ *and* $p_{2c}(\rho) \in [1/16, 1]$*; and for* $c \in [-1,1]$, $p_{3c}(\rho) \in [1/(4L^2), 1/L^2]$ *and* $p_{4c}(\rho) \in [1/(4L^2), 4/L^2]$.

Proof: See Appendix B.4. □

Proposition 3 *For the parameterized family of heavy-ball-like methods* (3.32) *with* $c \in [-1,0]$, $J_{\max}$ *and* $J_{\min}$ *in* (3.10) *satisfy,*

$$J_{\max} = \sigma_w^2\, p_{5c}(\rho)\, n\, (1 + 1/\kappa)\, T_s \tag{3.37a}$$

$$J_{\min} = \sigma_w^2 \big(2\, p_{5c}(\rho)\, (1 + 1/\kappa) + p_{6c}(\rho)(n-2)/\kappa\big)\, T_s \tag{3.37b}$$

where $p_{5c}(\rho) := 1/(2(1+|c|\rho)(1+|c|\rho^2)) \in [1/8, 1/2]$ *and* $p_{6c}(\rho) := 2(1+\rho)p_{5c}(\rho)q_{-c}(\rho) \in [1/8, 2]$.

Proof: See Appendix B.4. □

The upper bounds in Theorems 3 and 4 follow from Propositions 2 and 3, respectively. Since these upper bounds have the same scaling as the corresponding lower bounds in Theorems 2 and 4 that hold for all stabilizing parameters (α, β, γ), this demonstrates the tightness of lower bounds for all settling times and for both noise models.

3.5.2 Parameterized family of Nesterov-like methods

For the two-step momentum algorithm (3.2) with $\gamma = \beta$, the line segment $(b(\lambda), a(\lambda))$ parameterized by $\lambda \in [m, L]$ passes through the origin. As described in Section 3.4, gradient descent and Nesterov's method with the optimal parameters provided in Table 3.1 are obtained for $a = 0$ and $a = -(\rho/2)b$, respectively, and the corresponding end points $(b(m), a(m))$ and $(b(L), a(L))$ lie on the edges $X_\rho Z_\rho$ and $Y_\rho Z_\rho$ of the ρ-linear convergence triangle Δ_ρ. To provide a continuous transformation between these two standard algorithms, we introduce a parameter $c \in [0, 1/2]$, and let the line segment satisfy $a(\lambda) = -c\rho b(\lambda)$ and take its end points at the edges $X_\rho Z_\rho$ and $Y_\rho Z_\rho$; see Figure 3.3 for an illustration. This can be accomplished with the following choice of parameters,

$$\alpha = (1+\rho)(1+c-c\rho)/(L(1+c)), \qquad \gamma = \beta = c\rho^2/((\alpha L - 1)(1+c)). \tag{3.38}$$

For the parameterized family of Nesterov-like algorithms (3.38), Proposition 4 establishes the settling time and characterizes the dependence of $J_{\min} \times T_s$ and $J_{\max} \times T_s$ on the condition number κ and the problem size n.

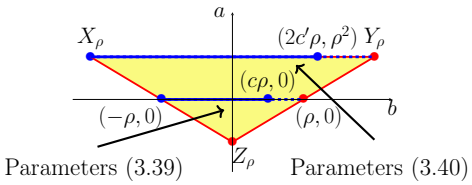


Figure 3.4: The triangle Δ_ρ (yellow) and the line segments $(b(\lambda), a(\lambda))$ with $\lambda \in [m, L]$ (blue) for gradient descent with reduced stepsize (3.39) and heavy-ball-like method (3.40), which place the end point $(b(m), a(m))$ at X_ρ and the end point $(b(L), a(L))$ at $(2c'\rho, \rho^2)$ on the edge $X_\rho Y_\rho$, where $c' := \kappa(1-\rho)^2/\rho - (1+\rho^2)/\rho$ ranges over the interval $[-1, 1]$.

Proposition 4 *For the class $\mathcal{Q}_m^L$ with the condition number $\kappa = L/m$, let the scalar ρ be such that*

$$T_s \;=\; 1/(1-\rho) \;\in\; [(\sqrt{3\kappa+1})/2, (\kappa+1)/2].$$

The two-step momentum algorithm (3.2) *with parameters* (3.38) *achieves the convergence rate ρ and satisfies*

$$\sigma_w^2\left((n-1)\kappa(\kappa+1)/32 + \sqrt{3\kappa+1}/2\right) \le J_{\max} \times T_s \le 6\,\sigma_w^2 n\,\kappa(3\kappa+1)$$
$$\sigma_w^2\left(\kappa(\kappa+1)/32 + (n-1)\sqrt{3\kappa+1}/2\right) \le J_{\min} \times T_s \le \sigma_w^2\left(6\,\kappa(3\kappa+1) + (n-1)(\kappa+1)/2\right)$$

where $J_{\max}$ and $J_{\min}$ are the largest and smallest values that J can take when the algorithm is applied to $f \in \mathcal{Q}_m^L$ with the condition number $\kappa = L/m$, and the scalar $c \in [0, 1/2]$ is the solution to the quadratic equation

$$\kappa(1-\rho)(1-c\rho-c^2(1+\rho)) = (1+\rho)(1-c\rho-c^2(1-\rho)).$$

Proof: See Appendix B.4. □

Since the stepsize in (3.38) satisfies $\alpha \in [1/L, 3/L]$, comparing the upper bounds in Proposition 4 with the lower bounds in Theorem 2 shows that, for settling times $T_s \le (\kappa+1)/2$, the parameters in (3.38) achieve order-wise optimal $J_{\max}$ and $J_{\min}$ for both the iterate ($\sigma_w = \sigma$) and gradient ($\sigma_w = \alpha\sigma$) noise models.

3.5.3 Impact of reducing the stepsize

When the only source of uncertainty is a noisy gradient, i.e., $\sigma_w = \alpha\sigma$, one can attempt to reduce the noise amplification J by decreasing the stepsize α at the expense of increasing the settling time $T_s = 1/(1-\rho)$ [56], [58], [103], [109]. In particular, for gradient descent, α can be reduced from its optimal value $2/(L+m)$ by keeping $(b(m), a(m))$ at $(-\rho, 0)$ and moving the point $(b(L), a(L))$ from $(\rho, 0)$ towards $(-\rho, 0)$ along the horizontal axis; see Figure 3.4. This can be accomplished with

$$\alpha \;=\; (1+c\rho)/L, \quad \gamma \;=\; \beta \;=\; 0 \tag{3.39}$$

for some $c \in [-1, 1]$ parameterizing $(b(L), a(L)) = (c\rho, 0)$. In this case, the settling time satisfies $T_s = (\kappa + c)/(c + 1) \in [(\kappa + 1/2), \infty)$ and similar arguments to those presented in the proof of Lemma 4 can be used to obtain

$$\hat{J}_{\max} \;=\; \hat{J}(m) \;=\; \sigma^2\kappa^2(1-\rho)/L^2$$
$$\hat{J}_{\min} \;=\; \begin{cases} \hat{J}(L) = \sigma^2\alpha^2/(1-c^2\rho^2) & c \;\leq\; 0 \\ \hat{J}(1/\alpha) = \sigma^2\alpha^2 & c \;\geq\; 0. \end{cases}$$

For a fixed n, the stepsize in (3.39) yields a $\Theta(\kappa^2)$ scaling for both $J_{\max} \times T_s$ and $J_{\min} \times T_s$ for all $c \in [-1, 1]$. Thus, gradient descent with reduced stepsize order-wise matches the lower bounds in Theorem 2. An IQC-based approach [93, Lemma 1] was utilized in [109, Theorem 13] to show that stepsize (3.39) also yields the above discussed convergence rate and worst-case noise amplification for one-point m-strongly convex L-smooth functions.

Remark 5 *Any desired settling time $T_s = 1/(1-\rho) \in [(\sqrt{\kappa}+1)/2, \infty)$ can be achieved by the heavy-ball-like method with reduced stepsize,*

$$\alpha \;=\; (1-\rho)^2/m, \quad \beta \;=\; \rho^2, \quad \gamma \;=\; 0. \tag{3.40}$$

This choice yields $J_{\max} = \sigma^2 n\kappa^2(1-\rho^4)/(L^2(1+\rho)^4)$ [109, Theorem 9]; see Figure 3.4. In addition, by considering the error in $y^t = x^t + \gamma(x^t - x^{t-1})$ as the performance metric, it was stated and numerically verified in [109] that the choice of parameters (3.40) *yields Pareto-optimal algorithms for simultaneously optimizing $J_{\max}$ and ρ. We note that the settling time $T_s = \Theta(\kappa)$ of gradient descent with standard stepsizes ($\alpha = 1/L$ or $2/(m+L)$) can be achieved via* (3.40) *by reducing α to $O(1/(\kappa L))$. In contrast, the parameterized family of heavy-ball-like methods* (3.32) *is order-wise Pareto-optimal (cf. Theorems 2-4) while maintaining $\alpha \in [1/L, 4/L]$.*

3.6 Continuous-time gradient flow dynamics

Noisy gradient descent can be viewed as the forward Euler discretization of gradient flow dynamics (gfd),

$$\dot{x} \;+\; \alpha\nabla f(x) \;=\; \sigma w \tag{3.41a}$$

where $\dot{x}$ denotes the derivative of x with respect to time τ and w is a white noise with zero mean and identity covariance matrix, $\mathbb{E}[w(\tau)] = 0$, $\mathbb{E}[w(\tau_1)w^T(\tau_2)] = I\delta(\tau_1 - \tau_2)$. Similarly, noisy two-step momentum algorithm (3.2) can be obtained by discretizing the accelerated gradient flow dynamics (agd),

$$\ddot{x} \;+\; \theta\dot{x} \;+\; \alpha\nabla f(x + \gamma\dot{x}) \;=\; \sigma w \tag{3.41b}$$

with $\theta := 1 - \beta$ by approximating x, $\dot{x}$, and $\ddot{x}$ using

$$x \;=\; x^{t+1}, \;\; \dot{x} \;\approx\; x^{t+1} - x^t, \;\; \ddot{x} \;\approx\; x^{t+2} - 2x^{t+1} + x^t.$$

System (3.41b) with $\beta = \gamma$ was introduced in [110] as a continuous-time analogue of Nesterov's accelerated algorithm and a Lyapunov-based method was employed to characterize its stability properties for smooth strongly convex problems.

For a time dilation $s = c\tau$, the solution to (3.41b) satisfies

$$x'' \;+\; \bar{\theta} x' \;+\; \bar{\alpha} \nabla f(x + \bar{\gamma} x') \;=\; \bar{\sigma} w$$

where $\dot{x} = \mathrm{d}x/\mathrm{d}\tau$, $x' = \mathrm{d}x/\mathrm{d}s$, and

$$\bar{\theta} \;=\; \theta/c, \quad \bar{\gamma} \;=\; c\gamma, \quad \bar{\alpha} \;=\; \alpha/c^2, \quad \bar{\sigma} \;=\; \sigma/(c\sqrt{c}).$$

This follows by combining $\dot{x} = cx'$ and $\ddot{x} = c^2 x''$ with the fact that the time dilation yields a $\sqrt{c}$ increase in the noise magnitude σ. Similar change of variables can be applied to gradient flow dynamics (3.41a) and to study stability and noise amplification of (3.41) we set $\alpha = 1/L$ and $\sigma = 1$ without loss of generality.

3.6.1 Modal-decomposition

For the quadratic problem (3.3) with $Q = Q^T \succ 0$, we follow the approach of Section 3.4.1 and utilize the eigenvalue decomposition of $Q = V \Lambda V^T$ and the change of variables, $\hat{x} := V^T(x - x^\star)$, $\hat{w} := V^T w$, to bring (3.41) to,

$$\begin{aligned} \dot{\hat{\psi}}_i \;&=\; \hat{A}_i \hat{\psi}_i \;+\; \hat{B}_i \hat{w}_i, \\ \hat{z}_i \;&=\; \hat{C}_i \hat{\psi}_i \end{aligned} \tag{3.42a}$$

where $\hat{w}_i$ is the ith component of the vector $\hat{w}$. For gradient flow dynamics (3.41a), we let $\hat{\psi}_i := \hat{x}_i$, which leads to

$$\hat{A}_i \;=\; -\alpha\lambda_i \;=:\; -a(\lambda_i), \;\; \hat{B}_i \;=\; 1, \;\; \hat{C}_i \;=\; 1. \tag{3.42b}$$

On the other hand, for accelerated gradient flow dynamics (3.41b), $\hat{\psi}_i := \begin{bmatrix} \hat{x}_i & \dot{\hat{x}}_i \end{bmatrix}^T$, and

$$\begin{aligned} \hat{A}_i \;&=\; \hat{A}(\lambda_i) \;:=\; \begin{bmatrix} 0 & 1 \\ -a(\lambda_i) & -b(\lambda_i) \end{bmatrix} \\ \hat{B}_i \;&=\; \begin{bmatrix} 0 & 1 \end{bmatrix}^T, \;\; \hat{C}_i \;=\; \begin{bmatrix} 1 & 0 \end{bmatrix} \\ a(\lambda) &:=\; \alpha\lambda, \;\; b(\lambda) \;:=\; \theta \;+\; \gamma\alpha\lambda. \end{aligned} \tag{3.42c}$$

Even though functions $a(\lambda)$ and $b(\lambda)$ take different forms in continuous time, matrices $\hat{A}_i$, $\hat{B}_i$, and $\hat{C}_i$ in (3.42c) have the same structure as their discrete-time counterparts in (3.18).

3.6.2 Optimal convergence rate

System (3.42) is stable if and only if the matrix $\hat{A}_i$ is Hurwitz (i.e., if all of its eigenvalues have negative real parts). Moreover, the system is exponentially stable with the rate ρ,

$$\|\hat{\psi}_i(\tau)\|_2 \;\leq\; c\,\mathrm{e}^{-\rho\tau}\,\|\hat{\psi}_i(0)\|_2$$

if and only if the real parts of all eigenvalues of $\hat{A}_i$ are less than or equal to $-\rho$. For gradient flow dynamics (3.41a) with $\alpha = 1/L$, $\hat{A}_i$'s are real scalars and ρ is determined by

$$\rho_{\mathrm{gfd}} \;:=\; \min_i\,|\alpha\lambda_i| \;=\; m/L \;=\; 1/\kappa. \tag{3.43}$$

Note that $\hat{A}_i$ in (3.42c) has the same structure as the matrix M in (3.20a). Lemma 5 is a continuous-time counterpart for Lemmas 1 and 2 and it provides conditions for (exponential) stability of matrices $\hat{A}_i$ for accelerated gradient flow dynamics (3.41b).

Lemma 5 *The real matrix M in* (3.20a) *satisfies*

$$M \text{ is Hurwitz} \iff a,\, b \,>\, 0.$$

In addition, for any $\rho > 0$, we have

$$\max\{\Re\,(\mathrm{eig}(M))\} \;\leq\; -\rho \iff \begin{cases} a \geq\; \rho(b \,-\, \rho) \\ b \geq\; 2\rho. \end{cases}$$

Proof: See Appendix B.5. □

Conditions for stability and ρ-exponential stability in Lemma 5 respectively require inclusion of the point (b, a) to the open positive orthant and the ρ-parameterized cone shown in Figure 3.5. Furthermore, the normalization of the parameter α to $\alpha = 1/L$ yields the extra condition $a \leq 1$. For $\rho < 1$, combining this inequality with the exponential stability conditions in Lemma 5 further restricts the ρ-exponential stability cone to the triangle in the (b, a)-plane,

$$\Delta_\rho \;:=\; \{(b,a) \mid b \,\geq\, 2\rho,\; \rho(b-\rho) \,\leq\, a \,\leq\, 1\} \tag{3.44a}$$

whose vertices are given by

$$X_\rho \;=\; (2\rho, \rho^2),\; Y_\rho \;=\; (2\rho, 1),\; Z_\rho \;=\; (\rho + 1/\rho, 1). \tag{3.44b}$$

For $\rho = 1$, the triangle Δ_ρ is a single point and, for $\rho > 1$, adding the normalization condition $a \leq 1$ makes the ρ-exponential stability conditions in Lemma 5 infeasible. Thus, in what follows, we confine our attention to $\rho < 1$.

Figure 3.5 illustrates the stability and ρ-exponential stability cones as well as the ρ-exponential stability triangle Δ_ρ. The geometry of Δ_ρ allows us to determine the largest condition number for which (3.41b) is ρ-exponentially stable.

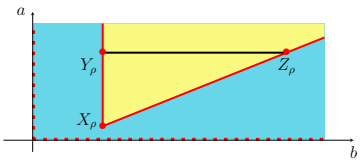


Figure 3.5: The open positive orthant (cyan) in the (b, a)-plane is the stability region for the matrix M in (3.20a). The intersections Y_ρ and Z_ρ of the stepsize normalization line $a = 1$ (black) and the boundary of the ρ-exponential stability cone (yellow) established in Lemma 5, along with the cone apex X_ρ determine the vertices of the ρ-exponential stability triangle Δ_ρ given by (3.44).

Proposition 5 *For a strongly convex quadratic objective function $f \in \mathcal{Q}_m^L$ with the condition number $\kappa = L/m$, the optimal convergence rate and the corresponding parameters (β, γ) of accelerated gradient flow dynamics* (3.41b) *with $\alpha = 1/L$ are*

$$\rho \;=\; 1/\sqrt{\kappa},\;\; \beta \;=\; 1 + (v-2)/\sqrt{\kappa},\;\; \gamma \;=\; v\sqrt{\kappa} \tag{3.45}$$

where $v \in [0, 1]$. This rate is achieved by the heavy-ball method ($\gamma = 0$) with $v = 0$ and, for $\kappa \geq 4$, by Nesterov's accelerated method ($\gamma = \beta$) with $v = (\sqrt{\kappa} - 2)/(\kappa - 1)$.

Proof: See Appendix B.5. □

Proposition 5 uses necessary and sufficient condition for ρ-exponential stability:

$$(b(\lambda), a(\lambda)) \;\in\; \Delta_\rho \text{ for all } \lambda \in [m, L].$$

Figure 3.6 illustrates the orientation of this line segment in Δ_ρ for the heavy-ball and Nesterov's algorithms. For the optimal values of parameters, Proposition 5 implies that accelerated gradient flow dynamics (3.41b) reduces the settling time $1/\rho$ relative to gradient flow dynamics (3.41a) by a factor of $\sqrt{\kappa}$, i.e.,

$$\rho_{\text{agd}}/\rho_{\text{gfd}} \;=\; \sqrt{\kappa}.$$

3.6.3 Noise amplification

Similar to the discrete-time setting, exponentially stable LTI systems in (3.42) driven by white noise reach a statistical steady-state with $\lim_{t\to\infty} \mathbb{E}\,(\hat{\psi}_i(t)) = 0$. Furthermore, the variance

$$J \;:=\; \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{E}\left(\|x(\tau) \,-\, x^\star\|_2^2\right) \mathrm{d}\tau \tag{3.46}$$

can be computed from the solution of the *continuous-time algebraic Lyapunov equation* [41]. The following theorem provides analytical expressions for the steady-state variance J.

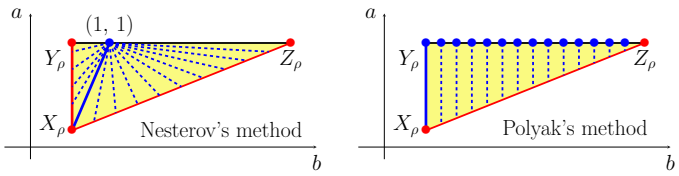


Figure 3.6: For a fixed ρ-exponential stability triangle Δ_ρ (yellow) in (3.44), the line segments $(b(\lambda), a(\lambda))$, $\lambda \in [m, L]$ for Nesterov's accelerated ($\gamma = \beta$) and the heavy-ball ($\gamma = 0$) dynamics, as special examples of accelerated dynamics (3.41b) with constant parameters γ, β, and $\alpha = 1/L$ are marked by dashed blue lines. The blue bullets correspond to the locus of the end point $(b(L), a(L))$, and the solid blue line segments correspond to the parameters for which the rate ρ is achieved for the largest possible condition number (3.45).

Theorem 6 *For a strongly convex quadratic objective function $f \in \mathcal{Q}_m^L$ with Hessian Q, the noise amplification J of (3.41) with any constant stabilizing parameters (α, β, γ) is determined by $J = \sum_{i=1}^n \hat{J}(\lambda_i)$. Here, $\hat{J}(\lambda_i)$ is the modal contribution of the ith eigenvalue λ_i of Q to the noise amplification*

$$\hat{J}_{\text{gfd}}(\lambda) \;=\; 1/(2a(\lambda)), \quad \hat{J}_{\text{agd}}(\lambda) \;=\; 1/(2a(\lambda)b(\lambda))$$

where the functions a and b are given by (3.42c).

We omit the proof of Theorem 6 as it uses similar arguments to those used in the proof of [93, Theorem 1].

For $\alpha = 1/L$ and the parameters that optimize the convergence rate in Proposition 5, we can use the explicit forms of $\hat{J}(\lambda)$ established in Theorem 6 to obtain

$$\begin{aligned} J_{\max} &= \begin{cases} ((n-1)\kappa + 1)/2 & \text{gfd} \\ ((n-1)\kappa\sqrt{\kappa} + \sqrt{\kappa})/4 & \text{agd (hb)} \\ ((n-1)\kappa\sqrt{\kappa} + 2)/4 & \text{agd (na)} \end{cases} \\ J_{\min} &= \begin{cases} (\kappa + (n-1))/2 & \text{gfd} \\ (\kappa\sqrt{\kappa} + (n-1)\sqrt{\kappa})/4 & \text{agd (hb)} \\ (\kappa\sqrt{\kappa} + (n-1)2)/4 & \text{agd (na)} \end{cases} \end{aligned} \tag{3.47}$$

For all three cases, the largest noise amplification $J_{\max}$ occurs when the Hessian matrix Q has $n-1$ eigenvalues at $\lambda = L$ and one at $\lambda = m$, and the smallest noise amplification $J_{\min}$ occurs when Q has $n-1$ eigenvalues at $\lambda = m$ and one at $\lambda = L$. Despite the $\sqrt{\kappa}$ improvement in the convergence rate achieved by the accelerated gradient flow dynamics, the corresponding $J_{\max}$ and $J_{\min}$ are larger than those of gradient flow dynamics by a factor of $\sqrt{\kappa}$. We next generalize this result to any stabilizing (β, γ) and establish similar trends for all $f \in \mathcal{Q}_m^L$.

3.6.4 Convergence and noise amplification tradeoffs

The next result is the continuous-time analogue of Theorem 2 and it establishes a lower bound on the product of the noise amplification and the settling time $T_s = 1/\rho$ of the accelerated gradient flow dynamics for any (β, γ).

Theorem 7 *Let the parameters (β, γ) be such that the accelerated gradient flow dynamics in (3.41b) with $\alpha = 1/L$ is exponentially stable with rate $\rho = 1/T_s$ for all $f \in \mathcal{Q}_m^L$. Then, $J_{\max}$ and $J_{\min}$ in (3.10) satisfy,*

$$J_{\max} \times T_s \;\geq\; (n \,-\, 1)\frac{\kappa^2}{4} \,+\, \frac{1}{2(1 \,+\, \rho^2)} \tag{3.48a}$$

$$J_{\min} \times T_s \;\geq\; \frac{\kappa^2}{4} \,+\, \frac{(n \,-\, 1)}{2(1 \,+\, \rho^2)}. \tag{3.48b}$$

Proof: See Appendix B.5. □

Theorem 7 demonstrates that the tradeoff between $J_{\max}$ and $J_{\min}$ and the settling time established in Theorem 2 for the two-step momentum algorithm extends to the continuous-time dynamics. For a fixed problem size n and the parameters that optimize the convergence rate provided in Lemma 5, we can use (3.47) to conclude that the bounds in Theorem 7 are order-wise tight for the parameters that achieve the optimal convergence rate.

3.7 Proofs of Theorems 1-4

3.7.1 Proof of Theorem 1

From Theorem 5 it follows that we can use upper bounds on $\hat{J}(\lambda)$ over $\lambda \in [m, L]$ to establish an upper bound on J. Since the algorithm achieves the convergence rate ρ, combining equation (3.19) and Lemma 2 yield $(b(\lambda), a(\lambda)) \in \Delta_\rho$ for all $\lambda \in [m, L]$. As we demonstrate in Appendix B.2, the function $\hat{J}$ is convex in (b, a) over the stability triangle Δ. In addition, $\Delta_\rho \subset \Delta$ is the convex hull of the points X_ρ, Y_ρ, Z_ρ in the (b, a)-plane. Since the maximum of a convex function over the convex hull of a finite set of points is attained at one of these points, $\hat{J}$ attains its maximum over Δ_ρ at X_ρ, Y_ρ, or Z_ρ.

Using the definition of X_ρ, Y_ρ, and Z_ρ in (3.22c), the affine relations (3.29), and the analytical expression for $\hat{J}$ in Theorem 5, it follows that the maximum occurs at vertices X_ρ and Y_ρ,

$$\hat{J}_{\max} \;:=\; \max_{\lambda \,\in\, [m,L]} \hat{J}(\lambda) \;=\; \frac{\sigma_w^2(1+\rho^2)}{(1-\rho)^3(1+\rho)^3}$$

where we use $d_{X_\rho} = l_{Y_\rho} = (1-\rho)^2$, $l_{X_\rho} = d_{Y_\rho} = (1+\rho)^2$, and $h_{X_\rho} = h_{Y_\rho} = 1-\rho^2$. Combining the above identity with Theorem 5 completes the proof of (3.12a).

We use an argument similar to the proof of Proposition 1 to prove (3.12b). In particular, since $(b(L), a(L)) \in \Delta_\rho$, we have

$$\alpha L \;=\; d(L) \;\leq\; d_{\max} \;=\; (1 \,+\, \rho)^2$$

where d given by (3.24) is the horizontal signed distance to the edge XZ of the stability triangle Δ. On the other hand, $d_{\max}$ is the largest value that d can take among all points $(b, a) \in \Delta_\rho$ and it corresponds to the vertex Y_ρ; see equation (3.26a). Combining this inequality with $\sigma_w = \alpha\sigma$ and (3.12a) completes the proof of Theorem 1.

3.7.2 Proof of Theorem 2

Using the expression $J = \sum_i \hat{J}(\lambda_i)$ established in Theorem 5, we have the decomposition

$$J \;=\; \hat{J}(m) \;+\; \sum_{i=1}^{n-1} \hat{J}(\lambda_i). \tag{3.49}$$

To prove the lower bounds (3.13b) and (3.13d) on $J_{\min} \times T_s$, we establish a lower bound on $\hat{J}(m) \times T_s$ that scales quadratically with κ, and a general lower bound on $\hat{J}(\lambda) \times T_s$.

Case $\sigma_w = \sigma$

The proof of (3.13b) utilizes the following inequalities

$$\frac{\hat{J}(m)}{1 \,-\, \rho} \;\geq\; \frac{\sigma_w^2 \kappa^2}{2(1 \,+\, \rho)^5} \tag{3.50a}$$

$$\frac{\hat{J}(\lambda)}{1 \,-\, \rho} \;\geq\; \frac{\sigma_w^2(\sqrt{\kappa} \,+\, 1)}{2}. \tag{3.50b}$$

We first prove (3.50a). Our approach builds on the proof of Proposition 1. In particular, $d(\lambda) = \alpha\lambda$ for the point $(b(\lambda), a(\lambda))$, where d and (b, a) are defined in (3.29) and (3.18c), respectively. Thus, $d(m) = d(L)/\kappa$. Furthermore, Lemma 3 implies $(b(\lambda), a(\lambda)) \in \Delta_\rho$ for $\lambda \in [m, L]$. Thus, the trivial inequality $d(L) \leq d_{\max}$ leads to

$$d(m) \;\leq\; d_{\max}/\kappa \;=\; (1 \,+\, \rho)^2/\kappa \tag{3.51}$$

where $d_{\max} = (1 + \rho)^2$ is the largest value that d can take among all points $(b, a) \in \Delta_\rho$; see equation (3.26a). We now use Theorem 5 to write

$$\frac{\hat{J}(\lambda)}{1 \,-\, \rho} \;=\; \frac{\sigma_w^2(d(\lambda) \,+\, l(\lambda))}{2d(\lambda)h(\lambda)l(\lambda)(1 \,-\, \rho)} \;\geq\; \frac{\sigma_w^2}{2d(\lambda)h(\lambda)(1 \,-\, \rho)}. \tag{3.52}$$

Next, we lower bound the right-hand side of (3.52). Let $\mathcal{L}$ be the line that passes through $(b(\lambda), a(\lambda))$ which is parallel to the edge XZ of the stability triangle Δ, and let G be the intersection of $\mathcal{L}$ and the edge $X_\rho Z_\rho$ of the ρ-stability triangle Δ_ρ; see Figure 3.7 for an illustration. It is easy to verify that

$$h_G \;\geq\; h(\lambda), \quad d_G \;=\; d(\lambda) \tag{3.53a}$$

where h_G and d_G correspond to the values of h and d associated with the point G. In addition, since G lies on the edge $X_\rho Z_\rho$, h_G and d_G satisfy the affine relation

$$h_G \;=\; 1 \,-\, \rho \,+\, d_G\rho/(1 \,-\, \rho). \tag{3.53b}$$

This follows from the equation of the line $X_\rho Z_\rho$ in the (b, a)-plane and from the definitions of d and h in (3.29). Furthermore, combining (3.53a) and (3.53b) implies

$$\frac{\sigma_w^2}{2\,d(\lambda)\,h(\lambda)(1\,-\,\rho)} \overset{(a)}{\geq} \frac{\sigma_w^2}{2\,d(\lambda)\,h_G(1\,-\,\rho)} \overset{(b)}{=} \frac{\sigma_w^2}{2\,d(\lambda)\,((1\,-\,\rho)^2\,+\,\rho\,d(\lambda))}. \tag{3.54a}$$

For $\lambda = m$, we can further write

$$\frac{\sigma_w^2}{2\,d(m)\,((1\,-\,\rho)^2\,+\,\rho\,d(m))} \;\geq\; \frac{\sigma_w^2}{2\,\frac{(1+\rho)^2}{\kappa}(\frac{(1+\rho)^2}{\kappa}\,+\,\rho\,\frac{(1+\rho)^2}{\kappa})} \;=\; \frac{\sigma_w^2\kappa^2}{2\,(1+\rho)^5} \tag{3.54b}$$

where the inequality is obtained from (3.27) and (3.51). Combining (3.52) with (3.54a) and (3.54b) completes the proof of (3.50a).

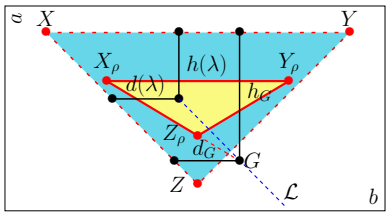


Figure 3.7: The line $\mathcal{L}$ (blue, dashed) and the intersection point G, along with the distances d_1, h_1, d_G, and h_G as introduced in the proof of Theorem 2.

Next, we prove the general lower bound in (3.50b). As we demonstrate in Appendix B.2, the modal contribution $\hat{J}$ to the noise amplification is a convex function of (b, a) which takes its minimum $\hat{J}_{\min} = \sigma_w^2$ over the stability triangle Δ at the origin $b = a = 0$. Combining this fact with the lower bound in (3.7) on ρ completes the proof of (3.50b). Finally, we can obtain the lower bound (3.13b) on $J_{\min} \times T_s$ by combining (3.49) and (3.50).

Case $\sigma_w = \alpha\sigma$

The proof of (3.13d) utilizes the following inequalities

$$\frac{\hat{J}(\lambda)}{1\,-\,\rho} \;\geq\; \frac{\sigma^2}{2\lambda^2(1\,+\,\rho)} \tag{3.55a}$$

$$\frac{\hat{J}(\lambda)}{1\,-\,\rho} \;\geq\; \frac{\sigma^2(1\,-\,\rho)^3\kappa^2}{L^2}. \tag{3.55b}$$

In particular, (3.13d) follows from using (3.55a) for $\lambda = m$ and taking the maximum of (3.55a) and (3.55b) for the other eigenvalues to bound the expression for J established by Theorem 5.

We first prove (3.55a). By combining (3.52) and (3.54a), we obtain

$$\frac{\hat{J}(\lambda)}{1 - \rho} \geq \frac{\alpha^2\sigma^2}{2\, d(\lambda)\, ((1 - \rho)^2 + \rho\, d(\lambda))}. \tag{3.56}$$

Since $d(\lambda) \geq d_{\min} := (1-\rho)^2$, where $d_{\min}$ is the smallest value of d over Δ_ρ, [cf. (3.26b)], we can write

$$\frac{\alpha^2\sigma^2}{2\, d(\lambda)\, ((1 - \rho)^2 + \rho d(\lambda))} \geq \frac{\alpha^2\sigma^2}{2\, d(\lambda)^2(1 + \rho)} = \frac{\sigma^2}{2\lambda^2(1 + \rho)}. \tag{3.57}$$

Combining (3.56) and (3.57) completes the proof of (3.55a).

To prove (3.55b) we use $d(\lambda) \geq d_{\min} := (1-\rho)^2$ and $d(m) = \alpha m$, to obtain $\alpha \geq (1-\rho)^2\kappa/L$. Combining this inequality with $\hat{J}_{\min} = \sigma_w^2 = \alpha^2\sigma^2$ yields (3.55b). Finally, we obtain the lower bound (3.13d) on $J_{\min} \times T_s$ by combining (3.49) and (3.55).

To obtain the lower bounds (3.13a) and (3.13c) on $J_{\max} \times T_s$, we consider a quadratic function for which the Hessian has $n-1$ eigenvalues at $\lambda = m$ and one eigenvalue at $\lambda = L$. For such a function, we can use Theorem 5 to write

$$J_{\max} \geq J = (n - 1)\hat{J}(m) + \hat{J}(L). \tag{3.58}$$

Case $\sigma_w = \sigma$

To prove (3.13a), we use inequalities in (3.50a) and (3.50b) to bound $\hat{J}(m)/(1-\rho)$ and $\hat{J}(L)/(1-\rho)$ in (3.58), respectively.

Case $\sigma_w = \alpha\sigma$

To prove (3.13c), we use inequality in (3.55a) with $\lambda = m$ to lower bound $\hat{J}(m)/(1-\rho)$, and combine (3.55a) and (3.55b) to lower bound $\hat{J}(L)/(1-\rho)$ in (3.58). This completes the proof.

3.7.3 Proof of Theorem 3

As described in Section 3.5, the parameters in Theorem 3 are obtained by placing the end points of the horizontal line segment $(b(\lambda), a(\lambda))$ parameterized by $\lambda \in [m, L]$ at the edges $X_\rho Z_\rho$ and $Y_\rho Z_\rho$ of the ρ-linear convergence triangle Δ_ρ. These parameters can be equivalently represented by (3.32) where the scalar c given in Lemma 4 satisfies $c \geq 0$ if and only if $T_s \leq (\kappa+1)/2$. The proof of Theorem 3 follows from combining Lemma 4 and Proposition 2.

3.7.4 Proof of Theorem 4

The following proposition allows us to prove the lower bounds in Theorem 4.

Proposition 6 *Let $\rho = \mathsf{\rho}(A) = 1 - 1/T_s$ be the convergence rate of the two-step momentum algorithm* (3.2)*. Then, the largest and smallest modal contributions to noise amplification given by* (3.31b) *satisfy*

$$\hat{J}_{\max} \;\geq\; \frac{\sigma_w^2}{2(1\,+\,\rho)^2}\,T_s, \qquad \hat{J}_{\min} \;\geq\; \sigma_w^2.$$

Proof: The inequality $\hat{J}_{\min} \geq \sigma_w^2$ follows from the fact that $\hat{J}$, as a function of (b, a), takes its minimum value at the origin; see Appendix B.2. The proof for $\hat{J}_{\max}$ utilizes the fact that for any constant parameters (α, β, γ) and fixed condition number, the spectral radius $\mathsf{\rho}(A)$ corresponds to the smallest ρ-linear convergence triangle Δ_ρ that contains the line segment $(b(\lambda), a(\lambda))$ for $\lambda \in [m, L]$. Thus, at least one of the end points $(b(m), a(m))$ or $(b(L), a(L))$ will be on the boundary of the triangle $\Delta_{\mathsf{\rho}(A)}$. Combining this with the fact that $d(m) \leq d(L)$, it follows that at least one of the following holds

$$\begin{aligned}(b(m), a(m)) \;&\in\; X_\rho Z_\rho \text{ or } X_\rho Y_\rho,\\ (b(L), a(L)) \;&\in\; Y_\rho Z_\rho \text{ or } X_\rho Y_\rho.\end{aligned}$$

Together with the concrete values of vertices (3.21c) in terms of ρ, this yields

$$1 \,-\, \rho \;\geq\; \min\{h(m),\, h(L),\, l(L)/(1\,+\,\rho),\, d(m)/(1\,+\,\rho)\} \tag{3.59}$$

Also, using Theorem 5 and noting that the maximum values that $h(\lambda)$, $d(\lambda)$, and $l(\lambda)$ can take among Δ_ρ are given by $1 + \rho^2$, $(1 + \rho)^2$, and $(1 + \rho)^2$, respectively, we can write

$$\begin{aligned}\hat{J}(m) \;&\geq\; \frac{\sigma_w^2}{2h(m)d(m)} \;\geq\; \max\left\{\frac{\sigma_w^2}{2h(m)(1+\rho)^2}, \frac{\sigma_w^2}{2d(m)(1+\rho^2)}\right\}\\ \hat{J}(L) \;&\geq\; \frac{\sigma_w^2}{2h(L)l(L)} \;\geq\; \max\left\{\frac{\sigma_w^2}{2h(L)(1+\rho)^2}, \frac{\sigma_w^2}{2l(L)(1+\rho^2)}\right\}.\end{aligned} \tag{3.60}$$

Finally, by the convexity of $\hat{J}$ (see Appendix B.2), we have $\hat{J}_{\max} \geq \max\{\hat{J}(m), \hat{J}(L)\}$. Combining this with (3.59) and (3.60) completes the proof. □

The lower bounds in Theorem 4 follow from combining Proposition 6 with the expression for J in Theorem 5. To obtain the upper bounds, we note that the parameter c in Lemma 4 satisfies $c \in [-1, 0]$ if and only if $T_s \geq (\kappa+1)/2$. Thus, we can use Proposition 3 to complete the proof.

3.8 Concluding remarks

We examined the amplification of stochastic disturbances for a class of two-step momentum algorithms in which the iterates are perturbed by an additive white noise which arises from uncertainties in gradient evaluation or in computing the iterates. For both noise models, we establish lower bounds on the product of the settling time and the smallest/largest steady-state variance of the error in the optimization variable. These bounds scale with κ^2 for

all stabilizing parameters, which reveals a fundamental limitation imposed by the condition number κ in designing algorithms that tradeoff noise amplification and convergence rate. In addition, we provide a novel geometric viewpoint of stability and ρ-linear convergence. This viewpoint brings insight into the relation between noise amplification, convergence rate, and algorithmic parameters. It also allows us to (i) take an alternative approach to optimizing convergence rates for standard algorithms; (ii) identify key similarities and differences between the iterate and gradient noise models; and (iii) introduce parameterized families of algorithms for which the parameters can be continuously adjusted to tradeoff noise amplification and settling time. By utilizing positive and negative momentum parameters in accelerated and decelerated regimes, respectively, we demonstrate that a parameterized family of the heavy-ball-like algorithms can achieve order-wise Pareto optimality for all settling times and both noise models. We also extend our analysis to continuous-time dynamical systems that can be discretized via an implicit-explicit Euler scheme to obtain the two-step momentum algorithm. For such gradient flow dynamics, we show that similar fundamental stochastic performance limitations hold as in discrete time.

Our ongoing work focuses on extending these results to algorithms with more complex structures including update strategies that utilize information from more than the last two iterates and time-varying algorithmic parameters [117]. It is also of interest to identify fundamental performance limitations of stochastic gradient descent algorithms in which both additive and multiplicative stochastic disturbances exist [118], [119].

Chapter 4

Transient growth of accelerated algorithms

First-order optimization algorithms are increasingly being used in applications with limited time budgets. In many real-time and embedded scenarios, only a few iterations can be performed and traditional convergence metrics cannot be used to evaluate performance of the algorithms in these non-asymptotic regimes. In this chapter, we examine the transient behavior of accelerated first-order optimization algorithms. For convex quadratic problems, we employ tools from linear systems theory to show that transient growth arises from the presence of non-normal dynamics. We identify the existence of modes that yield an algebraic growth in early iterations and quantify the transient excursion from the optimal solution caused by these modes. For strongly convex smooth optimization problems, we utilize the theory of integral quadratic constraints (IQCs) to establish an upper bound on the magnitude of the transient response of Nesterov's accelerated algorithm. We show that both the Euclidean distance between the optimization variable and the global minimizer and the rise time to the transient peak are proportional to the square root of the condition number of the problem. Finally, for problems with large condition numbers, we demonstrate tightness of the bounds that we derive up to constant factors.

4.1 Introduction

First-order optimization algorithms are widely used in a variety of fields including statistics, signal/image processing, control, and machine learning [1]–[5], [71], [120], [121]. Acceleration is often utilized as a means to achieve a faster rate of convergence relative to gradient descent while maintaining low per-iteration complexity. There is a vast literature focusing on the convergence properties of accelerated algorithms for different stepsize rules and acceleration parameters, including [7]–[9], [122]. There is also a growing body of work which investigates robustness of accelerated algorithms to various types of uncertainty [27], [53], [91]–[93], [123], [124]. These studies demonstrate that acceleration increases sensitivity to uncertainty in gradient evaluation.

In addition to deterioration of robustness in the face of uncertainty, asymptotically stable accelerated algorithms may also exhibit undesirable transient behavior [61]. This is in contrast to gradient descent which is a contraction for strongly convex problems with suitable stepsize [62]. In real-time optimization and in applications with limited time budgets, the transient growth can limit the appeal of accelerated methods. In addition, first-order algorithms are often used as a building block in multi-stage optimization including ADMM [63]

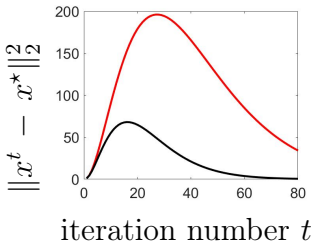


Figure 4.1: Error in the optimization variable for Polyak's heavy-ball (black) and Nesterov's (red) algorithms with the parameters that optimize the convergence rate for a strongly convex quadratic problem with the condition number 10^3 and a unit norm initial condition with $x^0 \neq x^\star$.

and distributed optimization methods [64]. In these settings, at each stage we can perform only a few iterations of first-order updates on primal or dual variables and transient growth can have a detrimental impact on the performance of the entire algorithm. This motivates an in-depth study of the behavior of accelerated first-order methods in non-asymptotic regimes.

It is widely recognized that large transients may arise from the presence of resonant modal interactions and non-normality of linear dynamical generators [65]. Even in the absence of unstable modes, these can induce large transient responses, significantly amplify exogenous disturbances, and trigger departure from nominal operating conditions. For example, in fluid dynamics, such mechanisms can initiate departure from stable laminar flows and trigger transition to turbulence [66], [67].

In this chapter, we consider the optimization problem

$$\underset{x}{\text{minimize}} \;\; f(x) \tag{4.1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a convex and smooth function, and we focus on a class of accelerated first-order algorithms

$$x^{t+2} \;=\; x^{t+1} \;+\; \beta(x^{t+1} - x^t) \;-\; \alpha\nabla f(x^{t+1} \;+\; \gamma(x^{t+1} - x^t)) \tag{4.2}$$

where t is the iteration index, α is the stepsize, and β is the momentum parameter. In particular, we are interested in Nesterov's accelerated and Polyak's heavy-ball methods that correspond to $\gamma = \beta$ and $\gamma = 0$, respectively. While these algorithms have faster convergence rates compared to the standard gradient descent ($\gamma = \beta = 0$), they may suffer from large transient responses; see Fig. 4.1 for an illustration. To quantify the transient behavior, we examine the ratio of the largest error in the optimization variable to the initial error.

For convex quadratic problems, the algorithm in (4.2) can be cast as a linear time-invariant (LTI) system and modal analysis of the state-transition matrix can be performed. For both accelerated algorithms, we identify non-normal modes that create large transient growth, derive analytical expressions for the state-transition matrices, and establish bounds on the transient response in terms of the convergence rate and the iteration number. We show that both the peak value of the transient response and the rise time to this value increase with the square root of the condition number of the problem. Moreover, for general

strongly convex problems, we combine a Lyapunov-based approach with the theory of IQCs to establish an upper bound on the transient response of Nesterov's accelerated algorithm. As for quadratic problems, we demonstrate that this bound scales with the square root of the condition number.

This work builds on our conference papers [96], [97]. In contrast to these preliminary results, we provide a comprehensive analysis of transient growth of accelerated algorithms for convex quadratic problems and address the important issue of eliminating transient growth of Nesterov's accelerated algorithm with the proper choice of initial conditions. Adaptive restarting, which was introduced in [61] to address the oscillatory behavior of Nesterov's accelerated method, provides heuristics for improving transient responses. In [107], the transient growth of second-order systems was studied and a framework for establishing upper bounds was introduced, with a focus on real eigenvalues. The result was applied to the heavy-ball method but was not applicable to quadratic problems in which the dynamical generator may have complex eigenvalues. We account for complex eigenvalues and conduct a thorough analysis for Nesterov's accelerated algorithm as well. Furthermore, for convex quadratic problems, we provide tight upper and lower bounds on transient responses in terms of the condition number and identify the initial condition that induces the largest transient response. Similar results with extensions to the Wasserstein distance have been recently reported in [125]. Previous work on non-asymptotic bounds for Nesterov's accelerated algorithm includes [126], where bounds on the objective error in terms of the condition number were provided. However, in contrast to our work, this result introduces a restriction on the initial conditions. Finally, while [56] presents computational bounds we develop analytical bounds on the non-asymptotic value of the estimated optimizer.

4.2 Convex quadratic problems

In this section, we examine transient responses of accelerated algorithms for convex quadratic objective functions,

$$f(x) \,=\, \frac{1}{2}\, x^T Q\, x \tag{4.3a}$$

where $Q = Q^T \succeq 0$ is a positive semi-definite matrix. In what follows, we first bring (4.2) into a standard LTI state-space form and then utilize appropriate coordinate transformation to decompose the dynamics into decoupled subsystems. Using this decomposition, we provide analytical expressions for the state-transition matrix and establish sharp bounds on the transient growth and the location of the transient peak for accelerated algorithms. We also examine the influence of initial conditions on transient responses and relegate the proofs to Appendix C.1.

4.2.1 LTI formulation

The matrix Q admits an eigenvalue decomposition, $Q = V\Lambda V^T$, where Λ is the diagonal matrix of eigenvalues with

$$\begin{aligned} &L \,:=\, \lambda_1 \,\geq\, \cdots \,\geq\, \lambda_r \,=:\, m \,>\, 0 \\ &\lambda_i \,=\, 0 \;\text{ for }\; i \,=\, r+1, \ldots, n \end{aligned} \tag{4.3b}$$

Method	Optimal parameters		Linear rate ρ
Nesterov	$\alpha = \frac{4}{3L+m}$	$\beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$	$1 - \frac{2}{\sqrt{3\kappa+1}}$
Polyak	$\alpha = \frac{4}{(\sqrt{L}+\sqrt{m})^2}$	$\beta = \frac{(\sqrt{\kappa}-1)^2}{(\sqrt{\kappa}+1)^2}$	$1 - \frac{2}{\sqrt{\kappa}+1}$

Table 4.1: Parameters that provide optimal convergence rates for a convex quadratic objective function (4.3) with $\kappa := L/m$.

and V is the unitary matrix of the corresponding eigenvectors. We define the condition number $\kappa := L/m$ as the ratio of the largest and smallest non-zero eigenvalues of the matrix Q. For f in (4.3a), we have $\nabla f(x) = Qx$, and the change of variables $\hat{x}^t := V^T x^t$ brings dynamics (4.2) to

$$\hat{x}^{t+2} \;=\; (I \;-\; \alpha\Lambda)\,\hat{x}^{t+1} \;+\; (\beta I \;-\; \gamma\alpha\Lambda)(\hat{x}^{t+1} \;-\; \hat{x}^t). \tag{4.4}$$

This system can be represented via n decoupled second-order subsystems of the form,

$$\hat{\psi}_i^{t+1} \;=\; A_i\hat{\psi}_i^t, \quad \hat{x}_i^t \;=\; C_i\hat{\psi}_i^t \tag{4.5a}$$

where $\hat{x}_i^t$ is the ith element of the vector $\hat{x}^t \in \mathbb{R}^n$, $\hat{\psi}_i^t := \begin{bmatrix} \hat{x}_i^t & \hat{x}_i^{t+1} \end{bmatrix}^T$, $C_i := \begin{bmatrix} 1 & 0 \end{bmatrix}$, and

$$A_i \;=\; \begin{bmatrix} 0 & 1 \\ -(\beta - \gamma\alpha\lambda_i) & 1 - \alpha\lambda_i + (\beta - \gamma\alpha\lambda_i) \end{bmatrix}. \tag{4.5b}$$

4.2.2 Linear convergence of accelerated algorithms

The minimizers of (4.3a) are determined by the null space of the matrix Q, $x^\star \in \mathcal{N}(Q)$. The constant parameters α and β can be selected to provide stability of subsystems in (4.5) for all $\lambda_i \in [m, L]$, and guarantee convergence of $\hat{x}_i^t$ to $\hat{x}_i^\star := 0$ with a linear rate determined by the spectral radius $\rho(A_i) < 1$. On the other hand, for $i = r+1, \ldots, n$ the eigenvalues of A_i are β and 1. In this case, the solution to (4.5) is given by

$$\hat{x}_i^t \;=\; \frac{1 \;-\; \beta^t}{1 \;-\; \beta}\,(\hat{x}_i^1 \;-\; \hat{x}_0^0) \;+\; \hat{x}_i^0 \tag{4.6a}$$

and the steady-state limit of $\hat{x}_i^t$,

$$\hat{x}_i^\star \;:=\; \frac{1}{1 \;-\; \beta}\,(\hat{x}_i^1 \;-\; \hat{x}_i^0) \;+\; \hat{x}_i^0 \tag{4.6b}$$

is achieved with a linear rate $\beta < 1$. Thus, the iterates of (4.2) converge to the optimal solution $x^\star = V\hat{x}^\star \in \mathcal{N}(Q)$ with a linear rate $\rho < 1$ and Table 4.1 provides the parameters α and β that optimize the convergence rate [52, Proposition 1].

4.2.3 Transient growth of accelerated algorithms

In spite of a significant improvement in the rate of convergence, acceleration may deteriorate performance on finite time intervals and lead to large transient responses. This is in contrast to gradient descent which is a contraction [62]. At any t, we are interested in the worst-case ratio of the two norm of the error of the optimization variable $z^t := x^t - x^\star$ to the two norm of the initial condition $\psi^0 - \psi^\star = \left[\begin{array}{cc} (z^0)^T & (z^1)^T \end{array}\right]^T$,

$$J^2(t) \;:=\; \sup_{\psi^0 \neq \psi^\star} \frac{\|x^t \,-\, x^\star\|_2^2}{\|\psi^0 \,-\, \psi^\star\|_2^2}. \tag{4.7}$$

Proposition 1 *For the accelerated algorithms applied to convex quadratic problems, $J(t)$ in (4.7) is determined by*

$$J^2(t) \;=\; \max\left\{ \max_{i \leq r} \|C_i A_i^t\|_2^2,\; \beta^{2t}/(1 \,+\, \beta^2) \right\}. \tag{4.8}$$

Proof: Since V is unitary and dynamics (4.5) that govern the evolution of each $\hat{x}_i^t$ are decoupled, $J(t)$ is determined by

$$J^2(t) \;=\; \max_i \sup_{\hat{\psi}_i^0 \neq \hat{\psi}_i^\star} \frac{(\hat{x}_i^t \,-\, \hat{x}_i^\star)^2}{\|\hat{\psi}_i^0 \,-\, \hat{\psi}_i^\star\|_2^2} \tag{4.9}$$

where $\hat{\psi}_i^\star := \left[\begin{array}{cc} \hat{x}_i^\star & \hat{x}_i^\star \end{array}\right]^T$. Furthermore, the mapping from $\hat{\psi}_i^0 - \hat{\psi}_i^\star$ to $\hat{x}_i^t - \hat{x}_i^\star$ is given by $\Phi_i(t) := C_i A_i^t$ where the state-transition matrix A_i^t is determined by the tth power of A_i,

$$\hat{x}_i^t \,-\, \hat{x}_i^\star \;=\; C_i A_i^t (\hat{\psi}_i^0 \,-\, \hat{\psi}_i^\star) \;=:\; \Phi_i(t)(\hat{\psi}_i^0 \,-\, \hat{\psi}_i^\star). \tag{4.10}$$

For $\lambda_i \neq 0$, $\hat{\psi}_i^0 - \hat{\psi}_i^\star = \hat{\psi}_i^0$ is an arbitrary vector in $\mathbb{R}^2$. Thus,

$$\sup_{\hat{\psi}_i^0 \neq \hat{\psi}_i^\star} \frac{(\hat{x}_i^t \,-\, \hat{x}_i^\star)^2}{\|\hat{\psi}_i^0 \,-\, \hat{\psi}_i^\star\|_2^2} \;=\; \|C_i A_i^t\|_2^2, \quad i \,=\, 1, \ldots, r. \tag{4.11}$$

This expression, however, does not hold when $\lambda_i = 0$ in (4.5) because $\psi_i^0 - \psi_i^\star$ is restricted to a line in $\mathbb{R}^2$. Namely, from (4.6),

$$\begin{aligned} \hat{x}_i^t \,-\, \hat{x}_i^\star &= \frac{-\beta^t}{1 \,-\, \beta}\,(\hat{x}_i^1 \,-\, \hat{x}_0^0) \\ \psi_i^0 \,-\, \psi_i^\star &= \left[\begin{array}{c} \hat{x}_i^0 \,-\, \hat{x}_i^\star \\ \hat{x}_i^1 \,-\, \hat{x}_i^\star \end{array}\right] \;=\; \frac{-(\hat{x}_i^1 \,-\, \hat{x}_i^0)}{1 \,-\, \beta} \left[\begin{array}{c} 1 \\ \beta \end{array}\right] \end{aligned} \tag{4.12}$$

which, for any initial condition with $\hat{x}_i^0 \neq \hat{x}_i^1$, leads to

$$\frac{(\hat{x}_i^t \,-\, \hat{x}_i^\star)^2}{\|\psi_i^0 \,-\, \psi_i^\star\|_2^2} \;=\; \frac{\beta^{2t}}{1 \,+\, \beta^2}, \quad i \,=\, r+1, \ldots, n. \tag{4.13}$$

Finally, substitution of (4.11) and (4.13) to (4.9) yields (4.8). □

4.2.4 Analytical expressions for transient response

We next derive analytical expressions for the state-transition matrix A_i^t and the response matrix $\Phi_i(t) = C_i A_i^t$ in (4.5).

Lemma 1 *Let μ_1 and μ_2 be the eigenvalues of the matrix*

$$M = \begin{bmatrix} 0 & 1 \\ a & b \end{bmatrix}$$

and let t be a positive integer. For $\mu_1 \neq \mu_2$,

$$M^t = \frac{1}{\mu_2 - \mu_1} \begin{bmatrix} \mu_1\mu_2(\mu_1^{t-1} - \mu_2^{t-1}) & \mu_2^t - \mu_1^t \\ \mu_1\mu_2(\mu_1^t - \mu_2^t) & \mu_2^{t+1} - \mu_1^{t+1} \end{bmatrix}.$$

Moreover, for $\mu := \mu_1 = \mu_2$, the matrix M^t is determined by

$$M^t = \begin{bmatrix} (1-t)\,\mu^t & t\,\mu^{t-1} \\ -t\,\mu^{t+1} & (t+1)\,\mu^t \end{bmatrix}. \tag{4.14}$$

Lemma 1 with $M = A_i$ determines explicit expressions for A_i^t. These expressions allow us to establish a bound on the norm of the response for each decoupled subsystem (4.5). In Lemma 2, we provide a tight upper bound on $\|C_i A_i^t\|_2^2$ for each t in terms of the spectral radius of the matrix A_i.

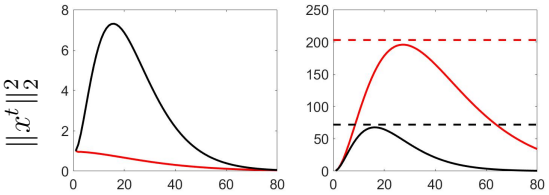
Lemma 2 *The matrix M in Lemma 1 satisfies*

$$\| \begin{bmatrix} 1 & 0 \end{bmatrix} M^t \|_2^2 \leq (t-1)^2\rho^{2t} + t^2\rho^{2t-2} \tag{4.15}$$

where ρ is the spectral radius of M. Moreover, (4.15) becomes equality if M has repeated eigenvalues.

Remark 1 *For Nesterov's accelerated algorithm with the parameters that optimize the rate of convergence (cf. Table 4.1), the matrix $\hat{A}_r$, which corresponds to the smallest non-zero eigenvalue of Q, $\lambda_r = m$, has an eigenvalue $1 - 2/\sqrt{3\kappa + 1}$ with algebraic multiplicity two and incomplete sets of eigenvectors. Similarly, for both $\lambda_1 = L$ and $\lambda_r = m$, $\hat{A}_1$ and $\hat{A}_r$ for the heavy-ball method with the parameters provided in Table 4.1 have repeated eigenvalues which are, respectively, given by $(1 - \sqrt{\kappa})/(1 + \sqrt{\kappa})$ and $-(1 - \sqrt{\kappa})/(1 + \sqrt{\kappa})$.*

We next use Lemma 2 with $M = A_i$ to establish an analytical expression for $J(t)$.



(a) $x^1 = x^0$ (b) $x^1 = -x^0$

Figure 4.2: Dependence of the error in the optimization variable on the iteration number for the heavy-ball (black) and Nesterov's methods (red), as well as the peak magnitudes (dashed lines) obtained in Proposition 2 for two different initial conditions with $\|x^1\|_2 = \|x^0\|_2 = 1$.

Theorem 1 *For accelerated algorithms applied to convex quadratic problems, $J(t)$ in (4.7) satisfies*

$$J^2(t) \;\leq\; \max\left\{(t-1)^2\rho^{2t} \;+\; t^2\rho^{2(t-1)},\, \beta^{2t}/(1 \;+\; \beta^2)\right\}$$

where $\rho := \max_{i\leq r} \rho(A_i)$. Moreover, for the parameters provided in Table 4.1

$$J^2(t) \;=\; (t-1)^2\rho^{2t} \;+\; t^2\rho^{2(t-1)}. \tag{4.16}$$

Theorem 1 highlights the source of disparity between the long and short term behavior of the response. While the geometric decay of ρ^t drives x^t to $x^\star$ as $t \to \infty$, early stages are dominated by the algebraic term which induces a transient growth. We next provide tight bounds on the time $t_{\max}$ at which the largest transient response takes place and the corresponding peak value $J(t_{\max})$. Even though we derive the explicit expressions for these two quantities, our tight upper and lower bounds are more informative and easier to interpret.

Theorem 2 *For accelerated algorithms with the parameters provided in Table 4.1, let $\rho \in [1/\mathrm{e}, 1)$. Then the rise time $t_{\max} := \operatorname{argmax}_t J(t)$ and the peak value $J(t_{\max})$ satisfy*

$$-1/\log(\rho) \;\leq\; t_{\max} \;\leq\; 1 \;-\; 1/\log(\rho)$$

$$-\frac{\sqrt{2}\rho}{\mathrm{e}\log(\rho)} \;\leq\; J(t_{\max}) \;\leq\; -\frac{\sqrt{2}}{\mathrm{e}\,\rho\log(\rho)}.$$

For accelerated algorithms with the parameters provided in Table 4.1, Theorem 2 can be used to determine the rise time to the peak in terms of condition number κ. We next establish that both $t_{\max}$ and $J(t_{\max})$ scale as $\sqrt{\kappa}$.

Proposition 2 *For accelerated algorithms with the parameters provided in Table 4.1, the rise time $t_{\max} := \operatorname{argmax}_t J(t)$ and the peak value $J(t_{\max})$ satisfy*

(i) Polyak's heavy-ball method with $\kappa \geq 4.69$

$$\begin{aligned}(\sqrt{\kappa}-1)/2 \;&\leq\; t_{\max} \;\leq\; (\sqrt{\kappa}+3)/2\\ \frac{(\sqrt{\kappa}-1)^2}{\sqrt{2}\,\mathrm{e}(\sqrt{\kappa}+1)} \;&\leq\; J(t_{\max}) \;\leq\; \frac{(\sqrt{\kappa}+1)^2}{\sqrt{2}\,\mathrm{e}(\sqrt{\kappa}-1)}\end{aligned}$$

(ii) Nesterov's accelerated method with $\kappa \geq 3.01$

$$\begin{aligned}(\sqrt{3\kappa+1}-2)/2 \;&\leq\; t_{\max} \;\leq\; (\sqrt{3\kappa+1}+2)/2\\ \frac{(\sqrt{3\kappa+1}-2)^2}{\sqrt{2}\,\mathrm{e}\sqrt{3\kappa+1}} \;&\leq\; J(t_{\max}) \;\leq\; \frac{3\kappa+1}{\sqrt{2}\,\mathrm{e}(\sqrt{3\kappa+1}-2)}.\end{aligned}$$

In Proposition 2, the lower-bounds on κ are only required to ensure that the convergence rate ρ satisfies $\rho \geq 1/\mathrm{e}$, which allows us to apply Theorem 2. We also note that the upper and lower bounds on $t_{\max}$ and $J(t_{\max})$ are *tight* in the sense that their ratio converges to 1 as $\kappa \to \infty$.

4.2.5 The role of initial conditions

The accelerated algorithms need to be initialized with x^0 and $x^1 \in \mathbb{R}^n$. This provides a degree of freedom that can be used to potentially improve their transient performance. To provide insight, let us consider the quadratic problem with $Q = \mathrm{diag}\,(\kappa, 1)$. Figure 4.2 shows the error in the optimization variable for Polyak's and Nesterov's algorithms as well as the peak magnitudes obtained in Proposition 2 for two different types of initial conditions with $x^1 = x^0$ and $x^1 = -x^0$, respectively. For $x^1 = -x^0$, both algorithms recover their worst-case transient responses. However, for $x^1 = x^0$, Nesterov's method shows no transient growth.

Our analysis shows that large transient responses arise from the existence of non-normal modes in the matrices A_i. However, such modes do not move the entries of the state transition matrix A_i^t in arbitrary directions. For example, using Lemma 1, it is easy to verify that A_r in (4.5b), associated with the smallest non-zero eigenvalue $\lambda_r = m$ of Q in Nesterov's algorithm with the parameters provided by Table 4.1 has the repeated eigenvalue $\mu = 1 - 2/\sqrt{3\kappa+1}$ and A_r^t is determined by (4.14) with $M = A_r$. Even though each entry of A_r^t experiences a transient growth, its row sum is determined by

$$A_r^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \,+\, 2t/(\sqrt{3\kappa+1}-2) \\ 1 \,+\, 2t/\sqrt{3\kappa+1} \end{bmatrix} (1 \,-\, 2/\sqrt{3\kappa+1})^t$$

and entries of this vector are monotonically decaying functions of t. Furthermore, for $i < r$, it can be shown that the entries of $A_i^t \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ remain smaller than 1 for all i and t. In Theorem 3, we provide a bound on the transient response of Nesterov's method for *balanced* initial conditions with $x^1 = x^0$.

Theorem 3 *For convex quadratic optimization problems, Nesterov's accelerated method with a balanced initial condition* $x^1 = x^0$ *and parameters provided in Table 4.1 satisfies*

$$\|x^t \,-\, x^\star\|_2 \;\leq\; \|x^0 \,-\, x^\star\|_2.$$

Proof: See Appendix C.2. □

It is worth mentioning that the transient growth of the heavy-ball method cannot be eliminated with the use of balanced initial conditions. To see this, we note that the matrices A_r^t and A_1^t for the heavy-ball method with parameters provided in Table 4.1 also take the form in (4.14) with $\mu = (1-\sqrt{\kappa})/(1+\sqrt{\kappa})$ and $\mu = -(1-\sqrt{\kappa})/(1+\sqrt{\kappa})$, respectively. In contrast to $A_r^t \begin{bmatrix} 1 & 1 \end{bmatrix}^T$, which decays monotonically,

$$A_1^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 + 2t\sqrt{\kappa}/(1-\sqrt{\kappa}) \\ 1 + 2t\sqrt{\kappa}/(1+\sqrt{\kappa}) \end{bmatrix} \frac{(1-\sqrt{\kappa})^t}{(1+\sqrt{\kappa})^t}$$

experiences transient growth. It was recently shown that an averaged version of the heavy-ball method experiences smaller peak deviation than the heavy-ball method [127]. We also note that adaptive restarting provides effective heuristics for reducing oscillatory behavior of accelerated algorithms [61].

Remark 2 *For accelerated algorithms with the parameters provided in Table 4.1, the initial condition that leads to the largest transient growth at any time τ is determined by*

$$\hat{\psi}_r^0 = c \begin{bmatrix} (1-\tau)\,\rho^\tau & \tau\rho^{\tau-1} \end{bmatrix}^T,\ \hat{\psi}_i^0 = 0 \text{ for } i \neq r$$

where $c \neq 0$ and $\hat{\psi}_r^0$ is the principal right singular vector of $C_r A_r^\tau$. Thus, the largest peak $J(t_{\max})$ occurs for $\{\hat{\psi}_i^0 = 0,\ i \neq r\}$ and $\hat{\psi}_r^0 = c \begin{bmatrix} (1-t_{\max})\,\rho^{t_{\max}} & t_{\max}\,\rho^{t_{\max}-1} \end{bmatrix}^T$, where tight bounds on $t_{\max}$ are established in Proposition 2.

Remark 3 *For $\lambda_i = 0$ in (4.5), $|\hat{x}_i^t - \hat{x}_i^\star|$ decays monotonically with a linear rate β and only non-zero eigenvalues of Q contribute to the transient growth. Furthermore, for the parameters provided in Table 4.1, our analysis shows that $J^2(t) = \max_{i \leq r} \|C_i A_i^t\|_2^2$. In what follows, we provide bounds on the largest deviation from the optimal solution for Nesterov's algorithm for general strongly convex problems.*

4.3 General strongly convex problems

In this section, we combine a Lyapunov-based approach with the theory of IQCs to provide bounds on the transient growth of Nesterov's accelerated algorithm for the class $\mathcal{F}_m^L$ of m-strongly convex and L-smooth functions. When f is not quadratic, first-order algorithms are no longer LTI systems and eigenvalue decomposition cannot be utilized to simplify analysis. Instead, to handle nonlinearity and obtain upper bounds on J in (4.7), we augment standard quadratic Lyapunov functions with the objective error.

For $f \in \mathcal{F}_m^L$, algorithm (4.2) is invariant under translation. Thus, without loss of generality, we assume that $x^\star = 0$ is the unique minimizer of (4.1) with $f(0) = 0$. In what follows,

we present a framework based on Linear Matrix Inequalities (LMIs) that allows us to obtain time-independent bounds on the error in the optimization variable. This framework combines certain IQCs [81] with Lyapunov functions of the form

$$V(\psi) \;=\; \psi^T X \psi \;+\; \theta f(C\psi) \tag{4.17}$$

which consist of the objective function evaluated at $C\psi$ and a quadratic function of ψ, where X is a positive definite matrix.

The IQC theory provides a convex control-theoretic approach to analyzing optimization algorithms [52] and it was recently employed to study convergence and robustness of the first-order methods [53], [54], [56], [68], [93], [128]. The type of Lyapunov functions in (4.17) was introduced in [56], [106] to study convergence for convex problems. For Nesterov's accelerated algorithm, we demonstrate that this approach provides *orderwise-tight* analytical upper bounds on $J(t)$.

Nesterov's accelerated algorithm can be viewed as a feedback interconnection of linear and nonlinear components

$$\begin{aligned} \psi^{t+1} \;&=\; A\,\psi^t \;+\; B\,u^t \\ y^t \;&=\; C_y\,\psi^t, \quad u^t \;=\; \Delta(y^t) \end{aligned} \tag{4.18a}$$

where the LTI part of the system is determined by

$$A = \begin{bmatrix} 0 & I \\ -\beta I & (1+\beta)I \end{bmatrix}, \;\; B \;=\; \begin{bmatrix} 0 \\ -\alpha I \end{bmatrix}, \;\; C_y \;=\; \begin{bmatrix} -\beta I & (1+\beta)I \end{bmatrix} \tag{4.18b}$$

and the nonlinear mapping $\Delta\colon \mathbb{R}^n \to \mathbb{R}^n$ is $\Delta(y) := \nabla f(y)$. Moreover, the state vector ψ^t and the input y^t to Δ are determined by

$$\psi^t \;:=\; \begin{bmatrix} x^t \\ x^{t+1} \end{bmatrix}, \;\; y^t \;:=\; (1+\beta)x^{t+1} \;-\; \beta x^t. \tag{4.18c}$$

For smooth and strongly convex functions $f \in \mathcal{F}_m^L$, Δ satisfies the quadratic inequality [52, Lemma 6]

$$\begin{bmatrix} y \;-\; y_0 \\ \Delta(y) \;-\; \Delta(y_0) \end{bmatrix}^T \Pi \begin{bmatrix} y \;-\; y_0 \\ \Delta(y) \;-\; \Delta(y_0) \end{bmatrix} \;\geq\; 0 \tag{4.19a}$$

for all y, $y_0 \in \mathbb{R}^n$, where the matrix Π is given by

$$\Pi \;:=\; \begin{bmatrix} -2mLI & (L+m)I \\ (L+m)I & -2I \end{bmatrix}. \tag{4.19b}$$

Using $u^t := \Delta(y^t)$ and $y^t := C_y\psi^t$ and evaluating (4.19a) at $y = y^t$ and $y_0 = 0$ leads to,

$$\begin{bmatrix} \psi^t \\ u^t \end{bmatrix}^T M_1 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} \;\geq\; 0 \tag{4.19c}$$

where

$$M_1 := \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} -2mLC_y^T C_y & (L+m)C_y^T \\ (L+m)C_y & -2I \end{bmatrix}. \tag{4.19d}$$

In Lemma 3, we provide an upper bound on the difference between the objective function at two consecutive iterations of Nesterov's algorithm. In combination with (4.19), this result allows us to utilize Lyapunov function of the form (4.17) to establish an upper bound on transient growth. We note that variations of this lemma have been presented in [56, Lemma 5.2] and in [93, Lemma 3].

Lemma 3 *Along the solution of Nesterov's accelerated algorithm* (4.18)*, the function* $f \in \mathcal{F}_m^L$ *with* $\kappa := L/m$ *satisfies*

$$f(x^{t+2}) \,-\, f(x^{t+1}) \,\leq\, \frac{1}{2} \begin{bmatrix} \psi^t \\ u^t \end{bmatrix}^T M_2 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} \tag{4.20a}$$

where the matrix M_2 *is given by*

$$M_2 := \begin{bmatrix} -mC_2^T C_2 & C_2^T \\ C_2 & -\alpha(2-\alpha L)I \end{bmatrix}, \quad C_2 := \begin{bmatrix} -\beta I & \beta I \end{bmatrix}. \tag{4.20b}$$

Using Lemma 3, we next demonstrate how a Lyapunov function of the form (4.17) with $\theta := 2\theta_2$ and $C := [\,0 \;\; I\,]$ in conjunction with property (4.19) of the nonlinear mapping Δ can be utilized to obtain an upper bound on $\|x^t\|_2^2$.

Lemma 4 *Let* M_1 *be given by* (4.19d) *and let* M_2 *be defined in Lemma 3. Then, for any positive semi-definite matrix* X *and nonnegative scalars* θ_1 *and* θ_2 *that satisfy*

$$W \,:=\, \begin{bmatrix} A^T X\, A - X & A^T X\, B \\ B^T\, X\, A & B^T\, X\, B \end{bmatrix} \,+\, \theta_1 M_1 \,+\, \theta_2 M_2 \,\preceq\, 0 \tag{4.21}$$

the transient growth of Nesterov's accelerated algorithm (4.18) *for all* $t \geq 1$ *is upper bounded by*

$$\|x^t\|_2^2 \,\leq\, \frac{\lambda_{\max}(X)\|x^0\|_2^2 \,+\, (\lambda_{\max}(X) + L\theta_2)\|x^1\|_2^2}{\lambda_{\min}(X) + m\theta_2}. \tag{4.22}$$

In Lemma 4, the Lyapunov function candidate $V(\psi) := \psi^T X \psi + 2\theta_2 f([\,0 \;\; I\,]\psi)$ is used to show that the state vector ψ^t is confined within the sublevel set $\{\psi \in \mathbb{R}^{2n} \,|\, V(\psi) \leq V(\psi^0)\}$ associated with $V(\psi^0)$. We next establish an *order-wise* tight upper bound on $\|x^t\|_2$ that scales linearly with $\sqrt{\kappa}$ by finding a feasible point to LMI (4.21) in Lemma 4.

Theorem 4 *For* $f \in \mathcal{F}_m^L$ *with the condition number* $\kappa := L/m$*, the iterates of Nesterov's accelerated algorithm* (4.18) *for any stabilizing parameters* $\alpha \leq 1/L$ *and* $\beta < 1$ *satisfy*

$$\|x^t\|_2^2 \,\leq\, \kappa \left(\frac{1+\beta^2}{\alpha\beta L} \,\|x^0\|_2^2 \,+\, (1 + \frac{1+\beta^2}{\alpha\beta L}) \,\|x^1\|_2^2 \right). \tag{4.23a}$$

Furthermore, for the conventional values of parameters

$$\alpha = 1/L,\ \beta = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1) \tag{4.23b}$$

the largest transient error, defined in (4.7)*, satisfies*

$$\frac{\sqrt{2}\,(\sqrt{\kappa}-1)^2}{\mathrm{e}\sqrt{\kappa}} \leq \sup_{\{t\in\mathbb{N},\, f\in\mathcal{F}_m^L\}} J(t) \leq \sqrt{3\kappa+\frac{4\kappa}{\kappa-1}}. \tag{4.23c}$$

For balanced initial conditions, i.e., $x^1 = x^0$, Nesterov established the upper bound $\sqrt{\kappa+1}$ on J in [9]. Theorem 4 shows that similar trends hold without restriction on initial conditions. Linear scaling of the upper and lower bounds with $\sqrt{\kappa}$ illustrates a potential drawback of using Nesterov's accelerated algorithm in applications with limited time budgets. As $\kappa \to \infty$, the ratio of these bounds converges to $\mathrm{e}\sqrt{3/2} \approx 3.33$, thereby demonstrating that the largest transient response for all $f \in \mathcal{F}_m^L$ is within the factor of 3.33 relative to the bounds established in Theorem 4.

4.4 Concluding remarks

We have examined the impact of acceleration on the transient responses of accelerated first-order optimization algorithms. Without imposing restrictions on initial conditions, we establish bounds on the largest value of the Euclidean distance between the optimization variable and the global minimizer. For convex quadratic problems, we utilize the tools from linear systems theory to fully capture transient responses and for general strongly convex problems, we employ the theory of integral quadratic constraints to establish an upper bound on transient growth. This upper bound is proportional to the square root of the condition number and we identify quadratic problem instances for which accelerated algorithms generate transient responses which are within a constant factor of this upper bound. Future directions include extending our analysis to nonsmooth optimization problems and devising algorithms that balance acceleration with quality of transient responses.

Chapter 5

Noise amplification of primal-dual gradient flow dynamics based on proximal augmented Lagrangian

In this chapter, we examine amplification of additive stochastic disturbances to primal-dual gradient flow dynamics based on proximal augmented Lagrangian. These dynamics can be used to solve a class of non-smooth composite optimization problems and are convenient for distributed implementation. We utilize the theory of integral quadratic constraints to show that the upper bound on noise amplification is inversely proportional to the strong-convexity module of the smooth part of the objective function. Furthermore, to demonstrate tightness of these upper bounds, we exploit the structure of quadratic optimization problems and derive analytical expressions in terms of the eigenvalues of the corresponding dynamical generators. We further specialize our results to a distributed optimization framework and discuss the impact of network topology on the noise amplification.

5.1 Introduction

We consider a class of primal-dual gradient flow dynamics based on proximal augmented Lagrangian [68] that can be used for solving large-scale non-smooth constrained optimization problems in continuous time. These problems arise in many areas e.g. signal processing [69], statistical estimation [70], and control [71]. In addition, primal-dual methods have received renewed attention due to their prevalent application in distributed optimization [72] and their convergence and stability properties have been greatly studied [73]–[79].

While gradient-based methods are not readily applicable to non-smooth optimization, we can utilize their proximal counterparts to address such problems [80]. In the context of non-smooth constrained optimization, proximal-based extensions of primal-dual methods can also be obtained using the augmented Lagrangian [68], which preserve structural separability and remain suitable for distributed optimization.

Using primal-dual algorithms in real-world distributed settings motivates the robustness analysis of such methods as uncertainty can potentially enter the dynamics due to noisy communication channels [129]. Moreover, uncertainties can also arise in applications where the exact value of the gradient is not fully available, e.g., when the objective function is obtained via costly simulations or its computation relies on noisy measurements e.g., real-time and embedded applications.

In this chapter, we consider the scenario in which the dynamics of the primal-dual flow are perturbed by additive white noise. We examine the mean-squared error of the primal optimization variable as a measure of how noise gets amplified by the dynamics – we refer to this quantity as *noise* (or *variance*) *amplification.* For convex quadratic optimization problems, the primal-dual flow becomes a linear time invariant system, for which the noise amplification can be characterized using Lyapunov equations. For non-quadratic problems, the flow is no longer linear, however, tools from robust control theory can be utilized to quantify upper bounds on the noise amplification. In particular, we use the theory of Integral Quadratic Constraints (IQC) [81], [82] to characterize upper bounds on the noise amplification of the primal-dual flow based on proximal augmented Lagrangian using solutions to a certain linear matrix inequality. Our results establish tight upper-upper bounds on the noise amplification that are inversely proportional to the strong-convexity module of the corresponding objective function.

The approach taken in this chapter is similar to those in [53], [56], [93], [98], [103], [105], wherein IQCs have been used to analyze convergence and robustness of first-order optimization algorithms and their accelerated variants. The noise amplification of primal-dual methods has also been studied in [129] where the authors have focused on quadratic problems and considered the average error in the objective function. In contrast, we consider the average error in the optimization variable and extend the noise amplification analysis to the case of strongly convex and non-smooth optimization problems. For smooth strongly convex problems, an input-output analysis with a focus on the induced $\mathcal{L}_2$ norm using the passivity theory has been provided in [49]. In contrast, we study stochastic performance of primal-dual algorithms that can be utilized to solve non-smooth composite optimization problems.

The rest of the chapter is structured as follows. We describe the proximal-augmented Lagrangian and the noisy primal-dual gradient flow dynamics in Section 5.2. We next study the variance amplification for quadratic problems in Section 5.3. We present our IQC-based approach for general strongly convex but non-smooth optimization problems in Section 5.4. We study the noise amplification in a distributed optimization setting in Section 5.5, and provide concluding remarks in Section 5.6.

5.2 Proximal Augmented Lagrangian

We study a nonsmooth composite optimization problem

$$\begin{array}{ll} \underset{x,z}{\text{minimize}} & f(x) \ + \ g(z) \\ \text{subject to} & Tx \ - \ z \ = \ 0 \end{array} \tag{5.1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a convex, continuously differentiable function, $g\colon \mathbb{R}^k \to \mathbb{R}$ is a convex, but possibly non-differentiable function, and $T \in \mathbb{R}^{k\times n}$ is a given matrix. The augmented Lagrangian associated with (5.1) is given by

$$\mathcal{L}_\mu(x, z; \nu) = f(x) + g(z) + \nu^T(Tx - z) + \tfrac{1}{2\mu}\|Tx - z\|_2^2$$

where $\mu > 0$ is a parameter and ν is the Lagrange multiplier. The infimum of the augmented Lagrangian $\mathcal{L}_\mu$ with respect to z is given by the proximal augmented Lagrangian [68]

$$\begin{aligned}\mathcal{L}_\mu(x;\nu) \;&:=\; \inf_z \;\mathcal{L}_\mu(x,z;\nu)\\ &=\; f(x) \;+\; \mathrm{M}_{\mu g}(Tx+\mu\nu) \;-\; \tfrac{\mu}{2}\|\nu\|_2^2\end{aligned} \tag{5.2}$$

where $\mathrm{M}_{\mu g}(\xi) := g(\mathbf{prox}_{\mu g}(\xi)) + \frac{1}{2\mu}\|\mathbf{prox}_{\mu g}(\xi) - \xi\|_2^2$ is the Moreau envelope of the function g and

$$\mathbf{prox}_{\mu g}(\xi) \;:=\; \operatorname*{argmin}_z \Big(g(z) \;+\; \tfrac{1}{2\mu}\|z \;-\; \xi\|^2\Big)$$

is the corresponding proximal operator. In addition, the Moreau envelope is continuously differentiable and its gradient is determined by $\mu\nabla\mathrm{M}_{\mu g}(\xi) = \xi - \mathbf{prox}_{\mu g}(\xi)$.

For convex problems, solving (5.1) amounts to finding the saddle points of $\mathcal{L}_\mu(x;\nu)$. To this end, continuous differentiability of $\mathcal{L}_\mu(x;\nu)$ was utilized in [68] to introduce associated Arrow-Hurwicz-Uzawa gradient flow dynamics

$$\begin{aligned}\dot{x} \;&=\; -\nabla_x\mathcal{L}_\mu(x;\nu)\\ \dot{\nu} \;&=\; \nabla_\nu\mathcal{L}_\mu(x;\nu)\end{aligned} \tag{5.3}$$

which is a continuous-time algorithm that performs gradient primal-descent and dual-ascent on the proximal augmented Lagrangian. For $\mathcal{L}_\mu(x;\nu)$ given by (5.2), gradient flow dynamics in (5.3) take the following form,

$$\begin{aligned}\dot{x} \;&=\; -\nabla f(x) - \tfrac{1}{\mu}\,T^T(Tx+\mu\nu - \mathbf{prox}_{\mu g}(Tx+\mu\nu))\\ \dot{\nu} \;&=\; Tx \;-\; \mathbf{prox}_{\mu g}(Tx+\mu\nu).\end{aligned} \tag{5.4}$$

5.2.1 Stability properties

When f is convex with a Lipschitz continuous gradient, and g is proper, closed, and convex, the set of equilibrium points of (5.4) is characterized by minimizers of problem (5.1) and is globally asymptotically stable [68, Theorem 2]. Furthermore, when f is strongly convex and T is full-row-rank, there is a unique equilibrium point $(x^\star, \nu^\star)$ which is globally exponentially stable and $(x^\star, z^\star = \mathbf{prox}_{\mu g}(Tx^\star + \mu\nu^\star))$ is the unique optimal solution of problem (5.1) [75, Theorem 6].

5.2.2 Noise amplification

We examine the impact of additive stochastic uncertainties on performance of the primal-dual gradient flow dynamics. In particular, we consider the noisy version of (5.4),

$$\begin{aligned}\mathrm{d}x \;&=\; -\big(\nabla f(x) \;+\; T^T\nabla\mathrm{M}_{\mu g}(Tx+\mu\nu)\big)\,\mathrm{d}t \;+\; \mathrm{d}w_1\\ \mathrm{d}\nu \;&=\; \big(Tx \;-\; \mathbf{prox}_{\mu g}(Tx+\mu\nu)\big)\,\mathrm{d}t \;+\; \mathrm{d}w_2\end{aligned} \tag{5.5}$$

where $\mathrm{d}w_i(t)$ are the increments of independent Wiener processes with covariance matrices $\mathbb{E}[w_i(t)w_i^T(t)] = s_i I t$ and $s_i > 0$ for $i \in \{1, 2\}$. We quantify the noise amplification using [82]

$$J \;=\; \limsup_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[\|x(t) \;-\; x^\star\|_2^2]\,\mathrm{d}t. \tag{5.6}$$

For quadratic objective functions $f(x) := \frac{1}{2}x^T Q x$, if we let g be the indicator function of the set $\{b\}$ with $b \in R^k$, (5.5) is a linear time-invariant system and J quantifies the steady-state variance of the error in the optimization variable $x(t) - x^\star$,

$$J \;=\; \lim_{t\to\infty} \mathbb{E}[\|x(t) \;-\; x^\star\|_2^2]. \tag{5.7}$$

In the next section, we examine this class of problems.

5.3 Quadratic optimization problems

To provide insight into the noise amplification of the primal-dual gradient flow dynamics, we first examine the special case in which the quadratic objective function $f(x) = \frac{1}{2}x^T Q x$ is strongly convex with $Q = Q^T \succ 0$ and $g(z) = \mathrm{I}_{\{b\}}(z)$, where $\mathrm{I}_{\mathcal{S}}$ is the indicator function of the set $\mathcal{S}$, i.e., $\mathrm{I}_{\mathcal{S}}(z) := 0$ for $z \in \mathcal{S}$ and $\mathrm{I}_{\mathcal{S}}(z) := \infty$ for $z \notin \mathcal{S}$. For this choice of g, optimization problem (5.1) simplifies to

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & Tx \;=\; b \end{array} \tag{5.8}$$

and the nonlinear terms in (5.5) are determined by

$$\nabla f(x) \;=\; Qx,\; \mathbf{prox}_{\mu g}(\xi) \;=\; b,\; \nabla \mathrm{M}_{\mu g}(\xi) \;=\; \tfrac{1}{\mu}(\xi \;-\; b).$$

Hence, (5.5) simplifies to

$$\begin{aligned} \mathrm{d}x &= \;-\Big((Q \;+\; \tfrac{1}{\mu}T^T T)x + T^T\nu - \tfrac{1}{\mu}Tb\Big)\,\mathrm{d}t \;+\; \mathrm{d}w_1 \\ \mathrm{d}\nu &= \;(Tx \;-b)\,\mathrm{d}t \;+\; \mathrm{d}w_2 \end{aligned} \tag{5.9}$$

In what follows, without loss of generality, we set $b = 0$. In this case, noisy dynamics (5.5) are described by an LTI system

$$\mathrm{d}\psi \;=\; A\,\psi\,\mathrm{d}t \;+\; \mathrm{d}w \tag{5.10}$$

where $w := \left[\begin{array}{cc} w_1^T & w_2^T \end{array}\right]^T$ and

$$\psi \;:=\; \left[\begin{array}{c} x \;-\; x^\star \\ \nu \;-\; \nu^\star \end{array}\right],\; A \;=\; \left[\begin{array}{cc} -(Q + \frac{1}{\mu}T^T T) & -T^T \\ T & 0 \end{array}\right].$$

For $Q \succ 0$ and a full-row-rank T, A is a Hurwitz matrix and LTI system (5.10) is stable. Moreover from linearity, it follows that the variance amplification can be computed as

$$J \;=\; \lim_{t\to\infty} \mathbb{E}[\|x(t) - x^\star\|_2^2] \;=\; \mathrm{trace}\,(XC^TC) \;=\; \mathrm{trace}\,(X_1) \tag{5.11}$$

where $X := \lim_{t\to\infty} \mathbb{E}[\psi(t)\psi^T(t)] = \begin{bmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{bmatrix}$ is the steady-state covariance matrix of the state $\psi(t)$ which can be obtained by solving the algebraic Lyapunov equation

$$AX \;+\; XA^T \;=\; -\mathrm{diag}\,(s_1 I, s_2 I) \tag{5.12}$$

and $C := \begin{bmatrix} I & 0 \end{bmatrix}$. Theorem 1 addresses the special case with $Q = mI$ and provides an analytical expression for the variance amplification of the corresponding primal-dual gradient flow dynamics. This result is obtained by computing the steady-state covariance matrix of the state ψ.

Theorem 1 *Let $f(x) = \frac{m}{2}\|x\|^2$, $g(z) = \mathrm{I}_{\{0\}}(z)$, and T be a full-row-rank matrix in* (5.1)*. Then, the steady-state variance of the primal optimization variable in* (5.5) *with $\mathrm{d}w_i(t)$ being the increments of independent Wiener processes with covariance $\mathbb{E}[w_i(t)w_i^T(t)] = s_i It$ is determined by*

$$J \;=\; \frac{(n-k)s_1}{2m} \;+\; \sum_{i=1}^{k} \frac{s_1 + s_2}{2(m + (1/\mu)\sigma_i^2(T))}$$

where $\sigma_i(T)$ is the ith singular values of the matrix T.

Proof: Let $T = U\Sigma V^T$ be the singular value decomposition with unitary matrices $U \in \mathbb{R}^{k\times k}$ and $V \in \mathbb{R}^{n\times n}$ and $\Sigma = \begin{bmatrix} \Sigma_0 & 0_{k\times(n-k)} \end{bmatrix} \in \mathbb{R}^{k\times n}$, with

$$\Sigma_0 \;:=\; \mathrm{diag}\,(\sigma_1, \ldots, \sigma_k) \;\in\; \mathbb{R}^{k\times k}.$$

Multiplication of the Lyapunov equation (5.12) by $M = \mathrm{diag}\,(V, U)$ and M^T from right and left, respectively, yields

$$\hat{A}\,\hat{X} \;+\; \hat{X}\,\hat{A}^T \;=\; -\mathrm{diag}\,(s_1 I, s_2 I) \tag{5.13}$$

where

$$\hat{A} \;=\; \begin{bmatrix} -mI - \frac{1}{\mu}\Sigma^T\Sigma & -\Sigma^T \\ \Sigma & 0 \end{bmatrix}, \qquad \hat{X} \;=\; \begin{bmatrix} \hat{X}_1 & \hat{X}_2 \\ \hat{X}_2^T & \hat{X}_3 \end{bmatrix} \;:=\; M^T X M.$$

Finally, it is straightforward to verify that

$$\hat{X}_1 \;=\; \begin{bmatrix} \frac{s_1+s_2}{2}\left(mI+\frac{1}{\mu}\Sigma_0\Sigma_0\right)^{-1} & 0 \\ 0 & \frac{s_1}{2m}I \end{bmatrix}$$

$$\hat{X}_2 \;=\; \begin{bmatrix} -\frac{s_2}{2}\Sigma_0^{-1} \\ 0_{(n-k)\times k} \end{bmatrix} \in \mathbb{R}^{n\times k}, \qquad \hat{X}_3 \;=\; \mathrm{diag}\,(a_1,\ldots,a_k) \;\in\; \mathbb{R}^{k\times k}$$

where $a_i = \dfrac{s_1+s_2}{2(m+\sigma_i^2/\mu)} + \dfrac{s_2(m+\sigma_i^2/\mu)}{2\sigma_i^2}$. The result follows from

$$J \;=\; \mathrm{trace}\,(X_1) \;=\; \mathrm{trace}\,(\hat{X}_1).$$

□

The following corollary is immediate from Theorem 1.

Corollary 1 *Under the conditions of Theorem 1, the steady-state variance of the primal optimization variable in* (5.5) *is upper bounded by* $J \leq (ns_1+ks_2)/(2m)$.

Corollary 1 establishes that, for $\mu > 0$ and a full-row-rank matrix T, the variance of the primal optimization variable in (5.5) satisfies an upper bound that is independent of T and μ. In addition, using the explicit expression for J provided in Theorem 1, it follows that for any fixed $\mu > 0$, in the limit of $\sigma_{\max}(T) \to 0$ and/or $n/k \to \infty$, the upper bound on the variance amplification J in Corollary 1 becomes exact.

It is also noteworthy that, as demonstrated in the proof of Theorem 1, the dual variable ν may experience an unbounded steady-state variance for $s_2 > 0$ if $\sigma_{\min}(T) \to 0$.

Even though it is challenging to derive an analytical expression for the covariance matrix X for a general strongly convex quadratic objective function f, we next demonstrate that the upper bound in Corollary 1 remains valid.

Theorem 2 *Let* $f(x) = \frac{1}{2}x^TQx$ *with* $Q \succeq mI$, $g(z) = \mathrm{I}_{\{0\}}(z)$, *and* T *be a full-row-rank matrix in* (5.1). *Then, the steady-state variance of the primal optimization variable in* (5.5) *with* $\mathrm{d}w_i(t)$ *being the increments of independent Wiener processes with covariance* $\mathbb{E}[w_i(t)w_i^T(t)] = s_iIt$ *satisfies*

$$J \;\leq\; \frac{ns_1+ks_2}{2m}. \tag{5.14}$$

Proof: To quantify J, an alternative method to using the state covariance matrix is to write $J = \mathrm{trace}(P\mathrm{diag}(s_1I, s_2I))$, where P is the observability gramian of system (5.10)

$$A^TP \,+\, PA \;=\; -C^TC \tag{5.15}$$

with $C = \begin{bmatrix} I & 0 \end{bmatrix}$. Thus, any matrix $P' \succeq P$ satisfies $J \leq \mathrm{trace}(P'\mathrm{diag}(s_1 I, s_2 I))$. To find such a P', we note that A satisfies

$$A^T I + I A \;=\; -2\mathrm{diag}\left(Q + \tfrac{1}{\mu} T^T T, 0\right) \;\preceq\; -2\lambda_{\min}(Q) C^T C.$$

Dividing this inequality by $2\lambda_{\min}(Q)$ and subtracting from (5.15) yields

$$A^T\left(\tfrac{1}{2\lambda_{\min}(Q)}\, I \;-\; P\right) \;+\; \left(\tfrac{1}{2\lambda_{\min}(Q)}\, I \;-\; P\right) A \;\preceq\; 0.$$

Since A is Hurwitz, it follows that $P \preceq \frac{1}{2\lambda_{\min}(Q)} I$, and hence

$$J \;=\; \mathrm{trace}\left(P\,\mathrm{diag}\left(s_1 I, s_2 I\right)\right) \;\leq\; \frac{1}{2\lambda_{\min}(Q)}\,\mathrm{trace}\left(\mathrm{diag}\left(s_1 I, s_2 I\right)\right) \;\leq\; \frac{n s_1 + k s_2}{2m}.$$

□

5.4 Beyond quadratic problems

In this section, we extend our upper bounds on the noise amplification of the primal-dual gradient flow dynamics to problems with a general strongly convex function f, a convex but possibly non-differentiable function g, and a matrix T of an arbitrary rank. Our approach is based on Integral Quadratic Constraints (IQCs) which provide a convex control-theoretic framework for stability and robustness analysis of systems with structured nonlinear components [81]. This framework has been recently used to analyze convergence and robustness of first-order optimization methods [53], [56], [93], [105]. In what follows, we first demonstrate how IQCs can be combined with quadratic storage functions to characterize upper bounds on the noise amplification of continuous-time dynamical systems via solutions to a certain linear matrix inequality (LMI). We then specialize this result to the primal-dual gradient flow dynamics and establish tight upper bounds on the noise amplification by finding feasible solutions to the associated LMI.

5.4.1 An IQC-based approach

As demonstrated in Section 5.4.2, noisy primal-dual gradient flow dynamics can be viewed as a feedback interconnection of an LTI system with a static nonlinear component

$$\begin{aligned} \mathrm{d}\psi \;&=\; A\,\psi\,\mathrm{d}t \;+\; B\,u\,\mathrm{d}t \;+\; \mathrm{d}w \\ \begin{bmatrix} z \\ y \end{bmatrix} \;&=\; \begin{bmatrix} C_z \\ C_y \end{bmatrix} \psi, \quad u(t) \;=\; \Delta(y(t)). \end{aligned} \tag{5.16}$$

Here, $\psi(t)$ is the state, $\mathrm{d}w(t)$ is the increment of a Wiener process with covariance

$$\mathbb{E}[w(t) w^T(t)] \;=\; W\,t$$

where W is a positive semidefinite matrix, $z(t)$ is the performance output, and $u(t)$ is the output of the nonlinear term Δ: $\mathbb{R}^n \to \mathbb{R}^n$ that satisfies the quadratic inequalities

$$\begin{bmatrix} y \\ \Delta(y) \end{bmatrix}^T \Pi_i \begin{bmatrix} y \\ \Delta(y) \end{bmatrix} \;\geq\; 0 \tag{5.17}$$

for some matrices Π_i and all $y \in \mathbb{R}^n$.

Lemma 1 utilizes property (5.17) of the nonlinear mapping Δ and provides an upper bound on the average energy [82]

$$J \;=\; \limsup_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}\left[\|z(t)\|\right]_2^2 \mathrm{d}t.$$

Lemma 1 *Let the nonlinear function $u = \Delta(y)$ satisfy*

$$\begin{bmatrix} y \\ u \end{bmatrix}^T \Pi_i \begin{bmatrix} y \\ u \end{bmatrix} \;\geq\; 0 \tag{5.18}$$

for some matrices Π_i, let P be a positive semidefinite matrix, and let λ_i be nonnegative scalars such that system (5.16) *satisfies*

$$\begin{bmatrix} A^TP + PA + C_z^T C_z & PB \\ B^T P & 0 \end{bmatrix} + \sum_i \lambda_i \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi_i \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \;\preceq\; 0. \tag{5.19}$$

Then the average energy of the performance output in statisticaly steady-state is bounded by $J \leq \mathrm{trace}(PW)$.

The proof of Lemma 1 follows from similar arguments as in [82, Theorem 7.2] and is omitted for brevity. Lemma 1 introduces a quadratic storage function, $\psi^T P \psi$, for continuous-time primal-dual gradient flow dynamics. We note that discrete-time variants of this result have been used to quantify noise amplification of accelerated optimization algorithms [93, Lemmas 1, 2], [103].

5.4.2 State-space representation

We next demonstrate how noisy primal-dual gradient flow dynamics (5.5) can be brought into the standard state-space form (5.16). In particular, choosing $\psi = \begin{bmatrix} x^T & \nu^T \end{bmatrix}^T$ as the state variable along with $z := x$ and

$$y \;=\; \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \;:=\; \begin{bmatrix} x \\ Tx + \mu\nu \end{bmatrix}, \qquad u \;=\; \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \;:=\; \begin{bmatrix} \nabla f(x) - mx \\ \mathbf{prox}_{\mu g}(Tx + \mu\nu) \end{bmatrix}$$

brings system (5.5) into the state-space form (5.16) with

$$A \;=\; \begin{bmatrix} -(mI + \frac{1}{\mu} T^T T) & -T^T \\ T & 0 \end{bmatrix}, \quad B \;=\; \begin{bmatrix} -I & \frac{1}{\mu} T^T \\ 0 & -I \end{bmatrix}, \quad C_y \;=\; \begin{bmatrix} I & 0 \\ T & \mu I \end{bmatrix}. \tag{5.20}$$

and $C_z = \begin{bmatrix} I & 0 \end{bmatrix}$, where m is the strong-convexity module of f. We note that the input-output pair (u, y) satisfies the pointwise nonlinear equation $u = \Delta(y)$ with $\Delta = \operatorname{diag}(\Delta_1, \Delta_2)$, where

$$u_1 = \Delta_1(y_1) := \nabla f(y_1) - m y_1, \qquad u_2 = \Delta_2(y_2) := \mathbf{prox}_{\mu g}(y_2).$$

It is worth mentioning that for the special case $g(z) = \mathrm{I}_{\{0\}}(z)$, which we considered in our analysis of quadratic problems in Section 5.3, the nonlinear term u_2 vanishes and the primal-dual gradient flow dynamics simplify to

$$\begin{aligned} \mathrm{d}x &= -\Big(\nabla f(x) + \tfrac{1}{\mu} T^T T x + T^T \nu\Big)\,\mathrm{d}t + \mathrm{d}w_1 \\ \mathrm{d}\nu &= Tx\,\mathrm{d}t + \mathrm{d}w_2. \end{aligned} \tag{5.21}$$

5.4.3 Characterizing the structural properties via IQCs

The input-output pairs (y_i, u_i) associated with nonlinear mappings Δ_i satisfy

$$\begin{bmatrix} y_i - y_i' \\ u_i - u_i' \end{bmatrix}^T \pi_i \begin{bmatrix} y_i - y_i' \\ u_i - u_i' \end{bmatrix} \geq 0 \tag{5.22}$$

where

$$\pi_1 := \begin{bmatrix} 0 & (L-m)I \\ (L-m)I & -2I \end{bmatrix}, \qquad \pi_2 := \begin{bmatrix} 0 & I \\ I & -2I \end{bmatrix}.$$

The above inequalities follow from the facts that Δ_1 is the gradient of the $(L-m)$-smooth convex function $f(\cdot) - (m/2)\|\cdot\|^2$ and that $\Delta_2 = \mathbf{prox}_{\mu g}$ is firmly non-expansive.

To make the above IQCs conform to the required format in Lemma 1, we can employ a suitable permutation combined with a change of variables that utilizes deviations from the optimal solution to obtain the inequalities in (5.18) with

$$\Pi_1 = \begin{bmatrix} 0 & 0 & (L-m)\,I & 0 \\ 0 & 0 & 0 & 0 \\ (L-m)\,I & 0 & -2I & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \Pi_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I \\ 0 & 0 & 0 & 0 \\ 0 & I & 0 & -2I \end{bmatrix}. \tag{5.23}$$

5.4.4 General convex g

The main result of the chapter is presented in Theorem 3. It demonstrates that proximal primal-dual gradient flow dynamics enjoys the same upper bound on noise amplification as the primal-dual gradient flow dynamics for smooth problems.

Theorem 3 *Let the function f be m-strongly convex and let g be closed, proper, convex. Then, the noise amplification of noisy primal-dual gradient flow dynamics satisfies* (5.14).

Proof: It is easy to verify that $P = pI$, $\lambda_1 = 1/(L-m)$, $\lambda_2 = 1/\mu$ with $p \geq 1/(2m)$ provides a feasible solution to the LMI in Lemma 1 for the system matrices in (5.20) and

matrices Π_1, Π_2 in (5.23). Thus, the result follows from Lemma 1. $\square$

For general strongly convex problems, Theorem 3 establishes the same upper bound on the noise amplification as what we obtained using Lyapunov equations for quadratic problems in Theorem 2. In addition, as we discussed in Section 5.3, this upper bound is tight in the sense that the noise amplification for the quadratic problem in Theorem 1 converges to this upper bound in the limit $\sigma_{\max}(T) \to 0$ and/or as $n/k \to \infty$. Another advantage of the IQC framework is that it does not require the matrix A to be Hurwitz. Therefore, the upper bound established in Theorem 3 holds for any matrix T independent of its rank.

5.5 Application to distributed optimization

The primal-dual gradient flow dynamics provide a distributed strategy for solving

$$\underset{\theta}{\text{minimize}} \ \sum_{i=1}^{n} f_i(\theta) \tag{5.24}$$

where f_i are convex functions [72]. Assuming without loss of generality that $\theta \in \mathbb{R}$, given a connected network with an incidence matrix $E = T^T$, we can assign a different scalar variable x_i to each agent and define the equivalent problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \displaystyle\sum_{i=1}^{n} f_i(x_i) \\ \text{subject to} & Tx \ = \ 0 \end{array} \tag{5.25}$$

where the constraint enforces that

$$x \ := \ \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T \ \in \mathcal{N}(T) \ = \ \{c\mathbb{1} \,|\, c \in \mathbb{R}\}$$

where $\mathbb{1} := [\, 1 \ \cdots \ 1 \,]^T$. Letting $f(x) := \sum_i f_i(x_i)$, the primal-dual gradient flow for solving problem (5.25) is determined by (5.21) and, in the absence of noise, it converges to $x = \theta^\star \mathbb{1}$, where $\theta^\star$ is an optimal solution of problem (5.24). In this formulation, the primal and dual variables x_i and ν_i correspond to the nodes and the edges of the network, respectively.

Theorem 3 provides an upper bound on noise amplification of a distributed primal-dual algorithm

$$J \ \leq \ \frac{ns_1 + ks_2}{2m}$$

for strongly convex problems. Here, k denotes the number of edges in the network and m is the strong convexity module of the function f. However, if f lacks strong convexity, then an additive white noise with a full-rank covariance matrix can result in unbounded variance of $x(t)$ as $t \to \infty$.

To see one such example, we can let f_i be constants, in which case the primal-dual gradient flow simplifies to a consensus-type algorithm. In this case, the average mode $a(t) := \frac{1}{n}(\mathbb{1}^T x(t))\mathbb{1}$ experiences a random walk, and its variance

$$J_a \;:=\; \lim_{t\to\infty} \mathbb{E}\left(\|a(t) - \theta^\star \mathbb{1}\|^2\right) \tag{5.26a}$$

is unbounded. However, the mean-square deviation from the network average

$$\bar{J} \;:=\; \lim_{t\to\infty} \mathbb{E}\left(\|x(t) - a(t)\|^2\right) \tag{5.26b}$$

becomes a relevant quantity and it can be used in lieu of J to quantify stochastic performance as it remains bounded [43].

Using the fact that $\langle x(t) - a(t), \mathbb{1}\rangle = 0$, this idea can be generalized to the distributed optimization framework by noting that the variance amplification can be split into two terms,

$$J \;=\; J_a \;+\; \bar{J}.$$

To provide insight, let us examine the special case with $f_i(\theta) = \frac{1}{2}m(\theta - c_i)^2$, where the agents aim to compute the average of c_i. Although the underlying dynamics are linear in this case, the results of Theorem 1 are not applicable because the matrix T is full row-rank only when the corresponding graph is a tree. However, by eliminating modes from the dual-variable that are not stable, a similar argument as in the proof of Theorem 1 can be used to establish an expression for the noise amplification in the distributed setting in terms of the non-zero eigenvalues λ_i of the Laplacian matrix $\mathbf{L} = T^T T$.

Proposition 1 *The noisy primal-dual gradient flow dynamics* (5.9) *for solving distributed optimization problem* (5.25) *with* $f_i(x_i) = \frac{1}{2}m(x_i - c_i)^2$ *satisfies* $J = J_a + \bar{J}$*, where*

$$J_a \;=\; \frac{s_1}{2m}, \quad \bar{J} \;=\; \sum_{i=1}^{n-1} \frac{s_1 + s_2}{2(m + \lambda_i(\mathbf{L})/\mu)}$$

and λ_i *are the non-zero eigenvalues of the Laplacian matrix* $\mathbf{L} = T^T T$ *of connected undirected network.*

Proof: Let us without loss of generality assume that $c_i = 0$; using the change of variables $y := T^T \nu$, we obtain that the noisy primal-dual flow satisfies

$$\begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \end{bmatrix} = \begin{bmatrix} -mI - \frac{1}{\mu}\mathbf{L} & -I \\ \mathbf{L} & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \mathrm{d}w_1 \\ T^T \mathrm{d}w_2 \end{bmatrix}.$$

Noting that $\mathbf{L}\mathbb{1} = 0$, we can let $\mathbf{L} = V\Lambda V^T$, where $\Lambda = \mathrm{diag}(0, \hat{\Lambda})$ is the diagonal matrix of eigenvalues and the columns of the unitary matrix $V = \begin{bmatrix} \mathbb{1}/\sqrt{n} & U \end{bmatrix}$ are the corresponding eigenvectors. Using the change of variables

$$\hat{x} \;:=\; U^T x, \quad \hat{y} \;:=\; U^T y, \quad \hat{\psi}^T = \begin{bmatrix} \hat{x}^T & \hat{y}^T \end{bmatrix}$$

it is easy to verify that

$$\mathrm{d}\hat{\psi} \;=\; \begin{bmatrix} -mI - \frac{1}{\mu}\hat{\Lambda} & -I \\ \hat{\Lambda} & 0 \end{bmatrix} \hat{\psi}\mathrm{d}t + \begin{bmatrix} \mathrm{d}\hat{w}_1 \\ \mathrm{d}\hat{w}_2 \end{bmatrix}$$

where $\mathrm{d}\hat{w}_1$ and $\mathrm{d}\hat{w}_2$ are the increments of independent Wiener process with covariance $s_1 It$ and $s_2\hat{\Lambda}t$, respectively. In addition, the average modes associated with the primal and dual variables $a = (x^T\mathbb{1})\mathbb{1}/n$ and $b = (y^T\mathbb{1})\mathbb{1}/n$ satisfy

$$\mathrm{d}a \;=\; -m\,a\,\mathrm{d}t + \mathrm{d}w_a, \quad b \;=\; 0$$

and the variance amplification is determined by

$$J \;=\; J_a \,+ \bar{J} \;=\; \lim_{t\to\infty} \mathbb{E}[\|\hat{x}\|^2] + \mathbb{E}[a^2] \;=\; \mathrm{trace}(X_1) + \frac{s_1}{2m}$$

where $X = \begin{bmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{bmatrix}$ is the corresponding state covariance matrix at the steady state

$$\begin{bmatrix} -mI - \frac{1}{\mu}\hat{\Lambda} & -I \\ \hat{\Lambda} & 0 \end{bmatrix} X \,+\, X \begin{bmatrix} -mI - \frac{1}{\mu}\hat{\Lambda} & \hat{\Lambda} \\ -I & 0 \end{bmatrix} \;=\; \begin{bmatrix} -s_1 I & 0 \\ 0 & -s_2\hat{\Lambda} \end{bmatrix}$$

The result follows from noting that X_1, X_2, and X_3 are all diagonal and

$$X_1 \;=\; \frac{s_1+s_2}{2}(mI+\hat{\Lambda})^{-1}, \qquad X_2 \;=\; \frac{-s_2}{2} I.$$

□

For quadratic optimization problems, Proposition 1 demonstrates that, in addition to the strong-convexity module of the function f, the topology of the network also impacts the variance amplification. In the limit as m goes to 0, while the variance of the average mode J_a becomes unbounded, the mean-square deviation from the average mode remains bounded and is captured by the sum of reciprocals of the eigenvalues of the graph Laplacian. This dependence of variance amplification on the spectral properties of $\mathbf{L}$ is identical to the one observed in standard consensus algorithms [43], [93].

5.6 Concluding remarks

We have examined the noise amplification of proximal primal-dual gradient flow dynamics that can be used to solve non-smooth composite optimization problems. For quadratic problems, we have employed algebraic Lyapunov equations to establish analytical expressions for the noise amplification. We have also utilized the theory of IQCs to characterize tight upper bounds in terms of a solution to an LMI. Our results show that stochastic performance of the primal-dual dynamics is inversely proportional to the strong-convexity module of the smooth part of the objective function. The ongoing work focuses on examining the impact

of network topology on the noise amplification in distributed settings and on extension of our results to discrete-time versions of primal-dual algorithms.

Part II

Convergence and sample complexity of gradient methods for the data-driven control

Chapter 6

Random search for continuous-time LQR

Model-free reinforcement learning attempts to find optimal control actions for an unknown dynamical system by directly searching over the parameter space of controllers. However, the statistical properties and convergence behavior of these approaches are often poorly understood because of the nonconvex nature of the underlying optimization problems and the lack of exact gradient computation. In this chapter, we take a step towards demystifying the performance and efficiency of such methods by focusing on the standard infinite-horizon linear quadratic regulator problem for continuous-time systems with unknown state-space parameters. We establish exponential stability for the ordinary differential equation (ODE) that governs the gradient-flow dynamics over the set of stabilizing feedback gains and show that a similar result holds for the gradient descent method that arises from the forward Euler discretization of the corresponding ODE. We also provide theoretical bounds on the convergence rate and sample complexity of the random search method with two-point gradient estimates. We prove that the required simulation time for achieving ϵ-accuracy in the model-free setup and the total number of function evaluations both scale as $\log(1/\epsilon)$.

6.1 Introduction

In many emerging applications, control-oriented models are not readily available and classical approaches from optimal control may not be directly applicable. This challenge has led to the emergence of Reinforcement Learning (RL) approaches that often perform well in practice. Examples include learning complex locomotion tasks via neural network dynamics [18] and playing Atari games based on images using deep-RL [19].

RL approaches can be broadly divided into model-based [130], [131] and model-free [20], [21]. While model-based RL uses data to obtain approximations of the underlying dynamics, its model-free counterpart prescribes control actions based on estimated values of a cost function without attempting to form a model. In spite of the empirical success of RL in a variety of domains, our mathematical understanding of it is still in its infancy and there are many open questions surrounding convergence and sample complexity. In this chapter, we take a step towards answering such questions with a focus on the infinite-horizon Linear Quadratic Regulator (LQR) for continuous-time systems.

The LQR problem is the cornerstone of control theory. The globally optimal solution can be obtained by solving the Riccati equation and efficient numerical schemes with provable convergence guarantees have been developed [83]. However, computing the optimal solution

becomes challenging for large-scale problems, when prior knowledge is not available, or in the presence of structural constraints on the controller. This motivates the use of direct search methods for controller synthesis. Unfortunately, the nonconvex nature of this formulation complicates the analysis of first- and second-order optimization algorithms. To make matters worse, structural constraints on the feedback gain matrix may result in a disjoint search landscape limiting the utility of conventional descent-based methods [84]. Furthermore, in the model-free setting, the exact model (and hence the gradient of the objective function) is unknown so that only zeroth-order methods can be used.

In this chapter, we study convergence properties of gradient-based methods for the continuous-time LQR problem. In spite of the lack of convexity, we establish (a) *exponential stability* of the ODE that governs the gradient-flow dynamics over the set of stabilizing feedback gains; and (b) *linear convergence* of the gradient descent algorithm with a suitable stepsize. We employ a standard convex reparameterization for the LQR problem [85], [86] to establish the convergence properties of gradient-based methods for the nonconvex formulation. In the model-free setting, we also examine convergence and sample complexity of the random search method [22] that attempts to emulate the behavior of gradient descent via gradient approximations resulting from objective function values. For the two-point gradient estimation setting, we prove linear convergence of the random search method and show that the total number of function evaluations and the simulation time required in our results to achieve ϵ-accuracy are proportional to $\log(1/\epsilon)$.

For the *discrete-time* LQR, global convergence guarantees were recently provided in [13] for gradient decent and the random search method with one-point gradient estimates. The authors established a bound on the sample complexity for reaching the error tolerance ϵ that requires a number of function evaluations that is at least proportional to $(1/\epsilon^4)\log(1/\epsilon)$. If one has access to the infinite-horizon cost values, the number of function evaluations for the random search method with one-point gradient estimates can be improved to $1/\epsilon^2$ [132]. In contrast, we focus on the *continuous-time* LQR and examine the two-point gradient estimation setting. The use of two-point gradient estimates reduces the required number of function evaluations to $1/\epsilon$ [132]. We significantly improve this result by showing that the required number of function evaluations is proportional to $\log(1/\epsilon)$. Similarly, the simulation time required in our results is proportional to $\log(1/\epsilon)$; this is in contrast to [13] that requires $\text{poly}(1/\epsilon)$ simulation time and [132] that assumes an infinite simulation time. Furthermore, our convergence results hold both in terms of the error in the objective value and the optimization variable (i.e., the feedback gain matrix) whereas [13] and [132] only prove convergence in the objective value. We note that the literature on model-free RL is rapidly expanding and recent extensions to Markovian jump linear systems [133], $\mathcal{H}_\infty$ robustness analysis through implicit regularization [134], learning distributed LQ problems [135], and output-feedback LQR [136] have been made.

Our presentation is structured as follows. In Section 6.2, we revisit the LQR problem and present gradient-flow dynamics, gradient descent, and the random search algorithm. In Section 6.3, we highlight the main results of the chapter. In Section 6.4, we utilize convex reparameterization of the LQR problem and establish exponential stability of the resulting gradient-flow dynamics and gradient descent method. In Section 6.5, we extend our analysis to the nonconvex landscape of feedback gains. In Section 6.6, we quantify the accuracy of two-point gradient estimates and, in Section 6.7, we discuss convergence and

sample complexity of the random search method. In Section 6.8, we provide an example to illustrate our theoretical developments and, in Section 6.9, we offer concluding remarks. Most technical details are relegated to the appendices.

Notation

We use $\mathrm{vec}(M) \in \mathbb{R}^{mn}$ to denote the vectorized form of the matrix $M \in \mathbb{R}^{m\times n}$ obtained by concatenating the columns on top of each other. We use $\|M\|_F^2 = \langle M, M\rangle$ to denote the Frobenius norm, where $\langle X, Y\rangle := \mathrm{trace}\,(X^T Y)$ is the standard matricial inner product. We denote the largest singular value of linear operators and matrices by $\|\cdot\|_2$ and the spectral induced norm of linear operators by $\|\cdot\|_S$

$$\|\mathcal{M}\|_2 \;:=\; \sup_M \frac{\|\mathcal{M}(M)\|_F}{\|M\|_F}, \quad \|\mathcal{M}\|_S \;:=\; \sup_M \frac{\|\mathcal{M}(M)\|_2}{\|M\|_2}.$$

We denote by $\mathbb{S}^n \subset \mathbb{R}^{n\times n}$ the set of symmetric matrices. For $M \in \mathbb{S}^n$, $M \succ 0$ means M is positive definite and $\lambda_{\min}(M)$ is the smallest eigenvalue. We use $S^{d-1} \subset \mathbb{R}^d$ to denote the unit sphere of dimension $d-1$. We denote the expected value by $\mathbb{E}[\cdot]$ and probability by $\mathbb{P}(\cdot)$. To compare the asymptotic behavior of $f(\epsilon)$ and $g(\epsilon)$ as ϵ goes to 0, we use $f = O(g)$ (or, equivalently, $g = \Omega(f)$) to denote $\limsup_{\epsilon\to 0} f(\epsilon)/g(\epsilon) < \infty$; $f = \tilde{O}(g)$ to denote $f = O(g\log^k g)$ for some integer k; and $f = o(\epsilon)$ to signify $\lim_{\epsilon\to 0} f(\epsilon)/\epsilon = 0$.

6.2 Problem formulation

The infinite-horizon LQR problem for continuous-time LTI systems is given by

$$\underset{x,\,u}{\text{minimize}} \quad \mathbb{E}\left[\int_0^\infty (x^T(t)Qx(t) \,+\, u^T(t)Ru(t))\,\mathrm{d}t\right] \tag{6.1a}$$

$$\text{subject to} \quad \dot{x} \,=\, Ax \,+\, Bu, \quad x(0) \,\sim\, \mathcal{D} \tag{6.1b}$$

where $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in \mathbb{R}^m$ is the control input, A and B are constant matrices of appropriate dimensions, Q and R are positive definite matrices, and the expectation is taken over a random initial condition $x(0)$ with distribution $\mathcal{D}$. For a controllable pair (A, B), the solution to (6.1) is given by

$$u(t) \;=\; -K^\star x(t) \;=\; -R^{-1}B^T P^\star x(t) \tag{6.2a}$$

where $P^\star$ is the unique positive definite solution to the Algebraic Riccati Equation (ARE)

$$A^T P^\star \,+\, P^\star A \,+\, Q \,-\, P^\star B R^{-1} B^T P^\star \;=\; 0. \tag{6.2b}$$

When the model is known, the LQR problem and the corresponding ARE can be solved efficiently via a variety of techniques [137]–[140]. However, these methods are not directly applicable in the model-free setting, i.e., when the matrices A and B are unknown. Exploiting

the linearity of the optimal controller, we can alternatively formulate the LQR problem as a direct search for the optimal linear feedback gain, namely

$$\underset{K}{\text{minimize}}\ f(K) \tag{6.3a}$$

where

$$f(K) := \begin{cases} \text{trace}\left((Q + K^T R\, K)X(K)\right), & K \in \mathcal{S}_K \\ \infty, & \text{otherwise.} \end{cases} \tag{6.3b}$$

Here, the function $f(K)$ determines the LQR cost in (6.1a) associated with the linear state-feedback law $u = -Kx$,

$$\mathcal{S}_K := \{K \in \mathbb{R}^{m\times n} \,|\, A - BK \text{ is Hurwitz}\} \tag{6.3c}$$

is the set of stabilizing feedback gains and, for any $K \in \mathcal{S}_K$,

$$X(K) := \int_0^\infty \mathbb{E}\left[x(t)x^T(t)\right] = \int_0^\infty \mathrm{e}^{(A-BK)t}\,\Omega\,\mathrm{e}^{(A-BK)^T t}\,\mathrm{d}t \tag{6.4a}$$

is the unique solution to the Lyapunov equation

$$(A - BK)X + X(A - BK)^T + \Omega = 0 \tag{6.4b}$$

and $\Omega := \mathbb{E}\,[x(0)x^T(0)]$. To ensure $f(K) = \infty$ for $K \notin \mathcal{S}_K$, we assume $\Omega \succ 0$. This assumption also guarantees $K \in \mathcal{S}_K$ if and only if the solution X to (6.4b) is positive definite.

In problem (6.3), the matrix K is the optimization variable, and (A, B, $Q \succ 0$, $R \succ 0$, $\Omega \succ 0$) are the problem parameters. This alternative formulation of the LQR problem has been studied for both continuous-time [83] and discrete-time systems [13], [141] and it serves as a building block for several important control problems including optimal static-output feedback design [142], optimal design of sparse feedback gain matrices [71], [143]–[147], and optimal sensor/actuator selection [121], [148]–[150].

For all stabilizing feedback gains $K \in \mathcal{S}_K$, the gradient of the objective function is determined by [142], [143]

$$\nabla f(K) = 2(R\,K - B^T P(K))X(K). \tag{6.5}$$

Here, $X(K)$ is given by (6.4a) and

$$P(K) = \int_0^\infty \mathrm{e}^{(A-BK)^T t}\,(Q + K^T R\,K)\,\mathrm{e}^{(A-BK)t}\,\mathrm{d}t \tag{6.6a}$$

is the unique positive definite solution of

$$(A - BK)^T P + P(A - BK) = -Q - K^T R\,K. \tag{6.6b}$$

To simplify our presentation, for any $K \in \mathbb{R}^{m\times n}$, we define the closed-loop Lyapunov operator $\mathcal{A}_K\colon \mathbb{S}^n \to \mathbb{S}^n$ as

$$\mathcal{A}_K(X) \;:=\; (A \,-\, BK)X \,+\, X(A \,-\, BK)^T. \tag{6.7a}$$

For $K \in \mathcal{S}_K$, both $\mathcal{A}_K$ and its adjoint

$$\mathcal{A}_K^*(P) \;=\; (A \,-\, BK)^T P \,+\, P(A \,-\, BK) \tag{6.7b}$$

are invertible and $X(K)$, $P(K)$ are determined by

$$X(K) \;=\; -\mathcal{A}_K^{-1}(\Omega), \quad P(K) \;=\; -(\mathcal{A}_K^*)^{-1}(Q \,+\, K^T R K).$$

In this chapter, we first examine the global stability properties of the gradient-flow dynamics

$$\dot{K} \;=\; -\nabla f(K), \;\; K(0) \,\in\, \mathcal{S}_K \tag{GF}$$

associated with problem (6.3) and its discretized variant,

$$K^{k+1} \;:=\; K^k \,-\, \alpha\, \nabla f(K^k), \;\; K^0 \,\in\, \mathcal{S}_K \tag{GD}$$

where $\alpha > 0$ is the stepsize. Next, we build on this analysis to study the convergence of a search method based on random sampling [22], [151] for solving problem (6.3). As described in Algorithm 1, at each iteration we form an empirical approximation $\overline{\nabla} f(K)$ to the gradient of the objective function via simulation of system (6.1b) for randomly perturbed feedback gains $K \pm U_i$, $i = 1, \ldots, N$, and update K via,

$$K^{k+1} \;:=\; K^k \,-\, \alpha\, \overline{\nabla} f(K^k), \quad K^0 \,\in\, \mathcal{S}_K. \tag{RS}$$

We note that the gradient estimation scheme in Algorithm 1 does not require knowledge of system matrices A and B in (6.1b) but only access to a simulation engine.

6.3 Main results

Optimization problem (6.3) is not convex [84]; see Appendix D.1 for an example. The function $f(K)$, however, has two important properties: *uniqueness of the critical points and the compactness of sublevel sets* [152], [153]. Based on these, the LQR objective error $f(K) - f(K^\star)$ can be used as a maximal Lyapunov function (see [154] for a definition and [155], [156] as examples) to prove asymptotic stability of gradient-flow dynamics (GF) over the set of stabilizing feedback gains $\mathcal{S}_K$. However, this approach does not provide any guarantee on the rate of convergence and additional analysis is necessary to establish exponential stability; see Section 6.5 for details.

6.3.1 Known model

We first summarize our results for the case when the model is known. In spite of the nonconvex optimization landscape, we establish the exponential stability of gradient-flow

Algorithm 1 Two-point gradient estimation

Require: Feedback gain $K \in \mathbb{R}^{m\times n}$, state and control weight matrices Q and R, distribution $\mathcal{D}$, smoothing constant r, simulation time τ, number of random samples N.

for $i = 1, \dots, N$ **do**

– Define perturbed feedback gains $K_{i,1} := K + rU_i$ and $K_{i,2} := K - rU_i$, where $\operatorname{vec}(U_i)$ is a random vector uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$.

– Sample an initial condition x_i from distribution $\mathcal{D}$.

– For $j \in \{1, 2\}$, simulate system (6.1b) up to time τ with the feedback gain $K_{i,j}$ and initial condition x_i to form

$$\hat{f}_{i,j} \;=\; \int_0^\tau (x^T(t)Qx(t) \;+\; u^T(t)Ru(t))\,\mathrm{d}t.$$

end for

Ensure: The gradient estimate

$$\overline{\nabla} f(K) \;=\; \frac{1}{2rN} \sum_{i=1}^{N} \left(\hat{f}_{i,1} \;-\; \hat{f}_{i,2}\right) U_i.$$

dynamics (GF) for any stabilizing initial feedback gain $K(0)$. This result also provides an explicit bound on the rate of convergence to the LQR solution $K^\star$.

Theorem 1 *For any initial stabilizing feedback gain $K(0) \in \mathcal{S}_K$, the solution $K(t)$ to gradient-flow dynamics* (GF) *satisfies*

$$\begin{aligned} f(K(t)) \;-\; f(K^\star) \;&\leq\; \mathrm{e}^{-\rho\, t}\,(f(K(0)) \;-\; f(K^\star)) \\ \|K(t) \;-\; K^\star\|_F^2 \;&\leq\; b\,\mathrm{e}^{-\rho\, t}\,\|K(0) \;-\; K^\star\|_F^2 \end{aligned}$$

where the convergence rate ρ and constant b depend on $K(0)$ and the parameters of the LQR problem (6.3).

The proof of Theorem 1 along with explicit expressions for the convergence rate ρ and constant b are provided in Section 6.5.1. Moreover, for a sufficiently small stepsize α, we show that gradient descent method (GD) also converges over $\mathcal{S}_K$ at a linear rate.

Theorem 2 *For any initial stabilizing feedback gain $K^0 \in \mathcal{S}_K$, the iterates of gradient descent* (GD) *satisfy*

$$\begin{aligned} f(K^k) \;-\; f(K^\star) \;&\leq\; \gamma^k \left(f(K^0) \;-\; f(K^\star)\right) \\ \|K^k \;-\; K^\star\|_F^2 \;&\leq\; b\,\gamma^k\,\|K^0 \;-\; K^\star\|_F^2 \end{aligned}$$

where the rate of convergence γ, stepsize α, and constant b depend on K^0 and the parameters of the LQR problem (6.3).

6.3.2 Unknown model

We now turn our attention to the model-free setting. We use Theorem 2 to carry out the convergence analysis of the random search method (RS) under the following assumption on the distribution of initial condition.

Assumption 1 *Let the distribution $\mathcal{D}$ of the initial conditions have i.i.d. zero-mean unit-variance entries with bounded sub-Gaussian norm, i.e., for a random vector $v \in \mathbb{R}^n$ that is distributed according to $\mathcal{D}$, $\mathbb{E}[v_i] = 0$ and $\|v_i\|_{\psi_2} \leq \kappa$, for some constant κ and $i = 1, \ldots, n$; see Appendix D.10 for the definition of $\|\cdot\|_{\psi_2}$.*

Our main convergence result holds under Assumption 1. Specifically, for a desired accuracy level $\epsilon > 0$, in Theorem 3 we establish that iterates of (RS) with constant stepsize (that does not depend on ϵ) reach accuracy level ϵ at a linear rate (i.e., in at most $O(\log(1/\epsilon))$ iterations) with high probability. Furthermore, the total number of function evaluations and the simulation time required to achieve an accuracy level ϵ are proportional to $\log(1/\epsilon)$. This significantly improves the existing results for discrete-time LQR [13], [132] that require $O(1/\epsilon)$ function evaluations and poly$(1/\epsilon)$ simulation time.

Theorem 3 (Informal) *Let the initial condition $x_0 \sim \mathcal{D}$ of the LTI system in* (6.1b) *obey Assumption 1. Also let the simulation time τ and the number of samples N used by Algorithm 1 satisfy*

$$\tau \;\geq\; \theta_1 \log(1/\epsilon) \;\text{ and }\; N \;\geq\; c\left(1 \,+\, \beta^4 \kappa^4\, \theta_1 \log^6 n\right) n$$

for some $\beta > 0$ and desired accuracy $\epsilon > 0$. Then, we can choose a smoothing parameter $r < \theta_3\sqrt{\epsilon}$ in Algorithm 1 and the constant stepsize α such that the random search method (RS) *that starts from any initial stabilizing feedback gain $K^0 \in \mathcal{S}_K$ achieves $f(K^k) - f(K^\star) \leq \epsilon$ in at most*

$$k \;\leq\; \theta_4 \log\left((f(K^0) \,-\, f(K^\star))/\epsilon\right)$$

iterations with probability not smaller than $1 - c'k(n^{-\beta} + N^{-\beta} + N\mathrm{e}^{-\frac{n}{8}} + \mathrm{e}^{-c'N})$. Here, the positive scalars c and c' are absolute constants and $\theta_1, \ldots, \theta_4 > 0$ depend on K^0 and the parameters of the LQR problem (6.3).

The formal version of Theorem 3 along with a discussion of parameters θ_i and stepsize α is presented in Section 6.7.

6.4 Convex reparameterization

The main challenge in establishing the exponential stability of (GF) arises from nonconvexity of problem (6.3). Herein, we use a standard change of variables to reparameterize (6.3) into a convex problem, for which we can provide exponential stability guarantees for gradient-flow dynamics. We then connect the gradient flow on this convex reparameterization to its nonconvex counterpart and establish the exponential stability of (GF).

6.4.1 Change of variables

The stability of the closed-loop system with the feedback gain $K \in \mathcal{S}_K$ in problem (6.3) is equivalent to the positive definiteness of the matrix $X(K)$ given by (6.4a). This condition allows for a standard change of variables $K = YX^{-1}$, for some $Y \in \mathbb{R}^{m\times n}$, to reformulate the LQR design as a convex optimization problem [85], [86]. In particular, for any $K \in \mathcal{S}_K$ and the corresponding matrix X, we have

$$f(K) \;=\; h(X,Y) \;:=\; \mathrm{trace}\,(Q\,X + Y^T R\,Y X^{-1})$$

where $h(X,Y)$ is a jointly convex function of (X,Y) for $X \succ 0$. In the new variables, Lyapunov equation (6.4b) takes the affine form

$$\mathcal{A}(X) \;-\; \mathcal{B}(Y) \;+\; \Omega \;=\; 0 \tag{6.8a}$$

where $\mathcal{A}$ and $\mathcal{B}$ are the linear maps

$$\mathcal{A}(X) \;:=\; A\,X \;+\; X A^T, \quad \mathcal{B}(Y) \;:=\; B\,Y \;+\; Y^T B^T. \tag{6.8b}$$

For an invertible map $\mathcal{A}$, we can express the matrix X as an affine function of Y

$$X(Y) \;=\; \mathcal{A}^{-1}(\mathcal{B}(Y) \;-\; \Omega) \tag{6.8c}$$

and bring the LQR problem into the convex form

$$\underset{Y}{\text{minimize}} \quad h(Y) \tag{6.9}$$

where

$$h(Y) \;:=\; \begin{cases} h(X(Y),Y), & Y \;\in\; \mathcal{S}_Y \\ \infty, & \text{otherwise} \end{cases}$$

and $\mathcal{S}_Y := \{Y \in \mathbb{R}^{m\times n} \,|\, X(Y) \succ 0\}$ is the set of matrices Y that correspond to stabilizing feedback gains $K = YX^{-1}$. The set $\mathcal{S}_Y$ is open and convex because it is defined via a positive definite condition imposed on the affine map $X(Y)$ in (6.8c). This positive definite condition in $\mathcal{S}_Y$ is equivalent to the closed-loop matrix $A - B\,Y(X(Y))^{-1}$ being Hurwitz.

Remark 1 *Although our presentation assumes invertibility of $\mathcal{A}$, this assumption comes without loss of generality. As shown in Appendix D.2, all results carry over to noninvertible $\mathcal{A}$ with an alternative change of variables $A = \hat{A} + BK^0$, $K = \hat{K} + K^0$, and $\hat{K} = \hat{Y}X^{-1}$, for some $K^0 \in \mathcal{S}_K$.*

6.4.2 Smoothness and strong convexity of $h(Y)$

Our convergence analysis of gradient methods for problem (6.4.1) relies on the L-smoothness and μ-strong convexity of the function $h(Y)$ over its sublevel sets $\mathcal{S}_Y(a) := \{Y \in \mathcal{S}_Y \,|\, h(Y) \leq a\}$. These two properties were recently established in [121] where it was shown that over any sublevel set $\mathcal{S}_Y(a)$, the second-order term $\left\langle \tilde{Y}, \nabla^2 h(Y;\tilde{Y}) \right\rangle$ in the Taylor series expansion of

$h(Y+\tilde{Y})$ around $Y \in \mathcal{S}_Y(a)$ can be upper and lower bounded by quadratic forms $L\|\tilde{Y}\|_F^2$ and $\mu\|\tilde{Y}\|_F^2$ for some positive scalars L and μ. While an explicit form for the smoothness parameter L along with an existence proof for the strong convexity modulus μ were presented in [121], in Proposition 1 we establish an explicit expression for μ in terms of a and parameters of the LQR problem. This allows us to provide bounds on the convergence rate for gradient methods.

Proposition 1 *Over any non-empty sublevel set $\mathcal{S}_Y(a)$, the function $h(Y)$ is L-smooth and μ-strongly convex with*

$$L \;=\; \frac{2a\|R\|_2}{\nu}\left(1 \,+\, \frac{a\|\mathcal{A}^{-1}\mathcal{B}\|_2}{\sqrt{\nu\lambda_{\min}(R)}}\right)^2 \tag{6.10a}$$

$$\mu \;=\; \frac{2\lambda_{\min}(R)\lambda_{\min}(Q)}{a\,(1\,+\,a^2\eta)^2} \tag{6.10b}$$

where the constants

$$\eta \;:=\; \frac{\|\mathcal{B}\|_2}{\lambda_{\min}(Q)\lambda_{\min}(\Omega)\sqrt{\nu\,\lambda_{\min}(R)}} \tag{6.10c}$$

$$\nu \;:=\; \frac{\lambda_{\min}^2(\Omega)}{4}\left(\frac{\|A\|_2}{\sqrt{\lambda_{\min}(Q)}} \,+\, \frac{\|B\|_2}{\sqrt{\lambda_{\min}(R)}}\right)^{-2} \tag{6.10d}$$

only depend on the problem parameters.

Proof: See Appendix D.3. □

6.4.3 Gradient methods over $\mathcal{S}_Y$

The LQR problem can be solved by minimizing the convex function $h(Y)$ whose gradient is given by [121, Appendix C]

$$\nabla h(Y) \;=\; 2\,R\,Y(X(Y))^{-1} \,-\, 2\,B^T W(Y) \tag{6.11a}$$

where $W(Y)$ is the solution to

$$A^T\,W \,+\, W\,A \;=\; (X(Y))^{-1}\,Y^T R\,Y(X(Y))^{-1} \,-\, Q. \tag{6.11b}$$

Using the strong convexity and smoothness properties of $h(Y)$ established in Proposition 1, we next show that the unique minimizer $Y^\star$ of the function $h(Y)$ is the exponentially stable equilibrium point of the gradient-flow dynamics over $\mathcal{S}_Y$,

$$\dot{Y} \;=\; -\nabla h(Y), \quad Y(0) \,\in\, \mathcal{S}_Y. \tag{GFY}$$

Proposition 2 *For any $Y(0) \in \mathcal{S}_Y$, the gradient-flow dynamics* (GFY) *are exponentially stable, i.e.,*

$$\|Y(t) \,-\, Y^\star\|_F^2 \;\leq\; (L/\mu)\,\mathrm{e}^{-2\,\mu\,t}\,\|Y(0) \,-\, Y^\star\|_F^2$$

where μ and L are the strong convexity and smoothness parameters of the function $h(Y)$ over the sublevel set $\mathcal{S}_Y(h(Y(0)))$.

Proof: The derivative of the Lyapunov function candidate $V(Y) := h(Y) - h(Y^\star)$ along the flow in (GFY) satisfies

$$\dot{V} \;=\; \left\langle \nabla h(Y), \dot{Y} \right\rangle \;=\; -\|\nabla h(Y)\|_F^2 \;\leq\; -2\,\mu V. \tag{6.12}$$

Inequality (6.12) is a consequence of the strong convexity of the function $h(Y)$ and it yields [157, Lemma 3.4]

$$V(Y(t)) \;\leq\; \mathrm{e}^{-2\,\mu\,t}\,V(Y(0)). \tag{6.13}$$

Thus, for any $Y(0) \in \mathcal{S}_Y$, $h(Y(t))$ converges exponentially to $h(Y^\star)$. Moreover, since $h(Y)$ is μ-strongly convex and L-smooth, $V(Y)$ can be upper and lower bounded by quadratic functions and the exponential stability of (GFY) over $\mathcal{S}_Y$ follows from Lyapunov theory [157, Theorem 4.10]. □

In Section 6.5, we use the above result to prove exponential/linear convergence of gradient flow/descent for the nonconvex optimization problem (6.3). Before we proceed, we note that similar convergence guarantees can be established for the gradient descent method with a sufficiently small stepsize α,

$$Y^{k+1} \;:=\; Y^k \;-\; \alpha\,\nabla h(Y^k), \quad Y^0 \;\in\; \mathcal{S}_Y \tag{GY}$$

Since the function $h(Y)$ is L-smooth over the sublevel set $\mathcal{S}_Y(h(Y^0))$, for any $\alpha \in [0, 1/L]$ the iterates Y^k remain within $\mathcal{S}_Y(h(Y^0))$. This property in conjunction with the μ-strong convexity of $h(Y)$ imply that Y^k converges to the optimal solution $Y^\star$ at a linear rate of $\gamma = 1 - \alpha\mu$.

6.5 Control design with a known model

The asymptotic stability of (GF) is a consequence of the following properties of the LQR objective function [152], [153]:

1. The function $f(K)$ is twice continuously differentiable over its open domain $\mathcal{S}_K$ and $f(K) \to \infty$ as $K \to \infty$ and/or $K \to \partial\mathcal{S}_K$.

2. The optimal solution $K^\star$ is the unique equilibrium point over $\mathcal{S}_K$, i.e., $\nabla f(K) = 0$ if and only if $K = K^\star$.

In particular, the derivative of the maximal Lyapunov function candidate $V(K) := f(K) - f(K^\star)$ along the trajectories of (GF) satisfies

$$\dot{V} \;=\; \Big\langle \nabla f(K), \dot{K} \Big\rangle \;=\; -\|\nabla f(K)\|_F^2 \;\leq\; 0$$

where the inequality is strict for all $K \neq K^\star$. Thus, Lyapunov theory [154] implies that, starting from any stabilizing initial condition $K(0)$, the trajectories of (GF) remain within the sublevel set $\mathcal{S}_K(f(K(0)))$ and asymptotically converge to $K^\star$.

Similar arguments were employed for the convergence analysis of the Anderson-Moore algorithm for output-feedback synthesis [152]. While [152] shows global asymptotic stability, it does not provide any information on the rate of convergence. In this section, we first demonstrate exponential stability of (GF) and prove Theorem 1. Then, we establish linear convergence of the gradient descent method (GD) and prove Theorem 2.

6.5.1 Gradient-flow dynamics: proof of Theorem 1

We start our proof of Theorem 1 by relating the convex and nonconvex formulations of the LQR objective function. Specifically, in Lemma 1, we establish a relation between the gradients $\nabla f(K)$ and $\nabla h(Y)$ over the sublevel sets of the objective function $\mathcal{S}_K(a) := \{K \in \mathcal{S}_K \,|\, f(K) \leq a\}$.

Lemma 1 *For any stabilizing feedback gain $K \in \mathcal{S}_K(a)$ and $Y := KX(K)$, we have*

$$\|\nabla f(K)\|_F \;\geq\; c\,\|\nabla h(Y)\|_F \tag{6.14a}$$

where $X(K)$ is given by (6.4a)*, the constant c is determined by*

$$c \;=\; \frac{\nu\sqrt{\nu\,\lambda_{\min}(R)}}{2\,a^2\,\|\mathcal{A}^{-1}\|_2\,\|B\|_2 \;+\; a\,\sqrt{\nu\,\lambda_{\min}(R)}} \tag{6.14b}$$

and the scalar ν given by Eq. (6.10d) *depends on the problem parameters.*

Proof: See Appendix D.4. □

Using Lemma 1 and the exponential stability of gradient-flow dynamics (GFY) over $\mathcal{S}_Y$, established in Proposition 2, we next show that (GF) is also exponentially stable. In particular, for any stabilizing $K \in \mathcal{S}_K(a)$, the derivative of $V(K) := f(K) - f(K^\star)$ along the gradient flow in (GF) satisfies

$$\dot{V} \;=\; -\|\nabla f(K)\|_F^2 \;\leq\; -c^2\,\|\nabla h(Y)\|_F^2 \;\leq\; -2\,\mu\,c^2\,V \tag{6.15}$$

where $Y = KX(K)$ and the constants c and μ are provided in Lemma 1 and Proposition 1, respectively. The first inequality in (6.15) follows from (6.14a) and the second follows from $f(K) = h(Y)$ combined with $\|\nabla h(Y)\|_F^2 \geq 2\mu V$ (which in turn is a consequence of the strong convexity of $h(Y)$ established in Proposition 1).

Now, since the sublevel set $\mathcal{S}_K(a)$ is invariant with respect to (GF), following [157, Lemma 3.4], inequality (6.15) guarantees that system (GF) converges exponentially in the

objective value with rate $\rho = 2\mu c^2$. This concludes the proof of part (a) in Theorem 1. In order to prove part (b), we use the following lemma which connects the errors in the objective value and the optimization variable.

Lemma 2 *For any stabilizing feedback gain K, the objective function $f(K)$ in problem* (6.3) *satisfies*

$$f(K) \;-\; f(K^\star) \;=\; \mathrm{trace}\left((K-K^\star)^T R\,(K-K^\star)\,X(K)\right)$$

where $K^\star$ is the optimal solution and $X(K)$ is given by (6.4a)*.*

Proof: See Appendix D.4. □

From Lemma 2 and part (a) of Theorem 1, we have

$$\begin{aligned}\|K(t) \;-\; K^\star\|_F^2 \;&\leq\; \frac{f(K(t)) \;-\; f(K^\star)}{\lambda_{\min}(R)\,\lambda_{\min}(X(K(t)))} \\ &\leq\; \mathrm{e}^{-\rho\,t}\,\frac{f(K(0)) \;-\; f(K^\star)}{\lambda_{\min}(R)\,\lambda_{\min}(X(K(t)))} \;\leq\; b'\,\mathrm{e}^{-\rho\,t}\,\|K(0) \;-\; K^\star\|_F^2 \quad (6.16)\end{aligned}$$

where $b' := \|R\|_2\|X(K(0))\|_2/(\lambda_{\min}(R)\lambda_{\min}(X(K(t))))$. Here, the first and third inequalities follow form basic properties of the matrix trace combined with Lemma 2 applied with $K = K(t)$ and $K = K(0)$, respectively. The second inequality follows from part (a) of Theorem 1.

Finally, to upper bound parameter b', we use Lemma 15 presented in Appendix D.11 that provides the lower and upper bounds $\nu/a \leq \lambda_{\min}(X(K))$ and $\|X(K)\|_2 \leq a/\lambda_{\min}(Q)$ on the matrix $X(K)$ for any $K \in \mathcal{S}_K(a)$, where the constant ν is given by (6.10d). Using these bounds and the invariance of $\mathcal{S}_K(a)$ with respect to (GF), we obtain

$$b' \;\leq\; b \;:=\; \frac{a^2\,\|R\|_2}{\nu\,\lambda_{\min}(R)\,\lambda_{\min}(Q)} \quad (6.17)$$

which completes the proof of part (b).

Remark 2 (Gradient domination) *Expression* (6.15) *implies that the objective function $f(K)$ over any given sublevel set $\mathcal{S}_K(a)$ satisfies the Polyak-Łojasiewicz (PL) condition [89]*

$$\|\nabla f(K)\|_F^2 \;\geq\; 2\,\mu_f\,(f(K) \;-\; f(K^\star)) \quad (6.18)$$

with parameter $\mu_f := \mu\,c^2$, where μ and c are functions of a that are given by (6.10b) *and* (6.14b)*, respectively. This condition is also known as gradient dominance and it was recently used to show convergence of gradient descent for* discrete-time *LQR problem [13].*

6.5.2 Geometric interpretation

The solution $Y(t)$ to gradient-flow dynamics (GFY) over the set $\mathcal{S}_Y$ induces the trajectory

$$K_{\mathrm{ind}}(t) \;:=\; Y(t)(X(Y(t)))^{-1} \quad (6.19)$$

over the set of stabilizing feedback gains $\mathcal{S}_K$, where the affine function $X(Y)$ is given by (6.8c). The induced trajectory $K_{\text{ind}}(t)$ can be viewed as the solution to the differential equation

$$\dot{K} \;=\; g(K) \tag{6.20a}$$

where $g\colon \mathcal{S}_K \to \mathbb{R}^{m\times n}$ is given by

$$g(K) \;:=\; \big(K\mathcal{A}^{-1}(\mathcal{B}(\nabla h(Y(K)))) - \nabla h(Y(K))\big)\,(X(K))^{-1}. \tag{6.20b}$$

Here, the matrix $X = X(K)$ is given by (6.4a) and $Y(K) = KX(K)$. System (6.20) is obtained by differentiating both sides of Eq. (6.19) with respect to time t and applying the chain rule. Figure 6.1 illustrates an induced trajectory $K_{\text{ind}}(t)$ and a trajectory $K(t)$ resulting from gradient-flow dynamics (GF) that starts from the same initial condition.

Moreover, using the definition of $h(Y)$, we have

$$h(Y(t)) \;=\; f(K_{\text{ind}}(t)). \tag{6.21}$$

Thus, the exponential decay of $h(Y(t))$ established in Proposition 2 implies that f decays exponentially along the vector field g, i.e., for $K_{\text{ind}}(0) \neq K^\star$, we have

$$\frac{f(K_{\text{ind}}(t)) \;-\; f(K^\star)}{f(K_{\text{ind}}(0)) \;-\; f(K^\star)} \;=\; \frac{h(Y(t)) \;-\; h(Y^\star)}{h(Y(0)) \;-\; h(Y^\star)} \;\leq\; \mathrm{e}^{-2\,\mu\,t}.$$

This inequality follows from inequality (6.13), where μ denotes the strong-convexity modulus of the function $h(Y)$ over the sublevel set $\mathcal{S}_Y(h(Y(0)))$; see Proposition 1. Herein, we provide a geometric interpretation of the exponential decay of f under the trajectories of (GF) that is based on the relation between the vector fields g and $-\nabla f$.

Differentiating both sides of Eq. (6.21) with respect to t yields

$$\|\nabla h(Y)\|^2 = \langle -\nabla f(K), g(K)\rangle\,. \tag{6.22}$$

Thus, for each $K \in \mathcal{S}_K$, the inner product between the vector fields $-\nabla f(K)$ and $g(K)$ is nonnegative. However, this is not sufficient to ensure exponential decay of f along (GF). To address this challenge, our proof utilizes inequality (6.14a) in Lemma 1. Based on the equation in (6.22), we observe that (6.14a) can be equivalently restated as

$$\frac{\|-\nabla f(K)\|_F}{\|\Pi_{-\nabla f(K)}(g(K))\|_F} \;=\; \frac{\|\nabla f(K)\|_F^2}{\langle -\nabla f(K), g(K)\rangle} \;\geq\; c^2$$

where $\Pi_b(a)$ denotes the projection of a onto b. Thus, Lemma 1 ensures that the ratio between the norm of the vector field $-\nabla f(K)$ associated with gradient-flow dynamics (GF) and the norm of the projection of $g(K)$ onto $-\nabla f(K)$ is uniformly lower bounded by a positive constant. This lower bound is the key geometric feature that allows us to deduce exponential decay of f along the vector field $-\nabla f$ from the exponential decay of the vector field g.

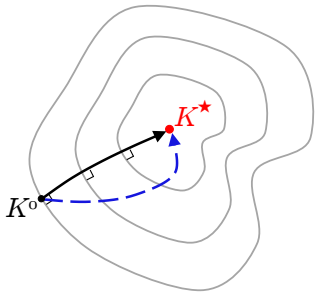


Figure 6.1: Trajectories $K(t)$ of (GF) (solid black) and $K_{\text{ind}}(t)$ resulting from Eq. (6.19) (dashed blue) along with the level sets of the function $f(K)$.

6.5.3 Gradient descent: proof of Theorem 2

Given the exponential stability of gradient-flow dynamics (GF) established in Theorem 1, the convergence analysis of gradient descent (GD) amounts to finding a suitable stepsize α. Lemma 3 provides a Lipschitz continuity parameter for $\nabla f(K)$, which facilitates finding such a stepsize.

Lemma 3 *Over any non-empty sublevel set $\mathcal{S}_K(a)$, the function gradient $\nabla f(K)$ is Lipschitz continuous with parameter*

$$L_f \;:=\; \frac{2a\|R\|_2}{\lambda_{\min}(Q)} \;+\; \frac{8a^3\|B\|_2}{\lambda^2_{\min}(Q)\lambda_{\min}(\Omega)}\Big(\frac{\|B\|_2}{\lambda_{\min}(\Omega)} \;+\; \frac{\|R\|_2}{\sqrt{\nu\lambda_{\min}(R)}}\Big)$$

where ν given by (6.10d) *depends on the problem parameters.*

Proof: See Appendix D.4. □

Let $K_\alpha := K - \alpha\,\nabla f(K)$, $\alpha \geq 0$ parameterize the half-line starting from $K \in \mathcal{S}_K(a)$ with $K \neq K^\star$ along $-\nabla f(K)$ and let us define the scalar $\beta_m := \max \beta$ such that $K_\alpha \in \mathcal{S}_K(a)$, for all $\alpha \in [0, \beta]$. The existence of β_m follows from the compactness of $\mathcal{S}_K(a)$ [152]. We next show that $\beta_m \geq 2/L_f$.

For the sake of contradiction, suppose $\beta_m < 2/L_f$. From the continuity of $f(K_\alpha)$ with respect to α, it follows that $f(K_{\beta_s}) = a$. Moreover, since $-\nabla f(K)$ is a descent direction of the function $f(K)$, we have $\beta_m > 0$. Thus, for $\alpha \in (0, \beta_m]$,

$$f(K_\alpha) \;-\; f(K) \;\leq\; -\frac{\alpha(2 \;-\; L_f\alpha)}{2}\,\|\nabla f(K)\|_F^2 \;<\; 0.$$

Here, the first inequality follows from the L_f-smoothness of $f(K)$ over $\mathcal{S}_K(a)$ (Descent Lemma [158, Eq. (9.17)]) and the second inequality follows from $\nabla f(K) \neq 0$ in conjunction with $\beta_m \in (0, 2/L_f)$. This implies $f(K_{\beta_m}) < f(K) \leq a$, which contradicts $f(K_{\beta_m}) = a$. Thus, $\beta_m \geq 2/L_f$.

We can now use induction on k to show that, for any stabilizing initial condition $K^0 \in \mathcal{S}_K(a)$, the iterates of (GD) with $\alpha \in [0, 2/L_f]$ remain in $\mathcal{S}_K(a)$ and satisfy

$$f(K^{k+1}) \; - \; f(K^k) \; \leq \; -\frac{\alpha(2 \; - \; L_f\alpha)}{2} \, \|\nabla f(K^k)\|_F^2. \tag{6.23}$$

Inequality (6.23) in conjunctions with the PL condition (6.18) evaluated at K^k guarantee linear convergence for gradient descent (GD) with the rate $\gamma \leq 1 - \alpha\,\mu_f$ for all $\alpha \in (0, 1/L_f]$, where μ_f is the PL parameter of the function $f(K)$. This completes the proof of part (a) of Theorem 2.

Using part (a) and Lemma 2, we can make a similar argument to what we used for the proof of Theorem 1 to establish part (b) with constant b in (6.17). We omit the details for brevity.

Remark 3 *Using our results, it is straightforward to show linear convergence of* $K^{k+1} = K^k - \alpha H_1^k \nabla f(K^k) H_2^k$ *with* $K^0 \in \mathcal{S}_K$ *and small enough stepsize, where* H_1^k *and* H_2^k *are uniformly upper and lower bounded positive definite matrices. In particular, the Kleinman iteration [137] is recovered for* $\alpha = 0.5$, $H_1^k = R^{-1}$, *and* $H_2^k = (X(K^k))^{-1}$. *Similarly, convergence of gradient descent may be improved by choosing* $H_1^k = I$ *and* $H_2^k = (X(K^k))^{-1}$. *In this case, the corresponding update direction provides the continuous-time variant of the so-called* natural gradient *for discrete-time systems [159].*

6.6 Bias and correlation in gradient estimation

In the model-free setting, we do not have access to the gradient $\nabla f(K)$ and the random search method (RS) relies on the gradient estimate $\overline{\nabla} f(K)$ resulting from Algorithm 1. According to [13], achieving $\|\overline{\nabla} f(K) - \nabla f(K)\|_F \leq \epsilon$ may take $N = \Omega(1/\epsilon^4)$ samples using one-point gradient estimates. Our computational experiments (not included in this chapter) also suggest that to achieve $\|\overline{\nabla} f(K) - \nabla f(K)\|_F \leq \epsilon$, N must scale as poly $(1/\epsilon)$ even when a two-point gradient estimate is used. To avoid this poor sample complexity, in our proof we take an alternative route and give up on the objective of controlling the gradient estimation error. By exploiting the problem structure, we show that with a linear number of samples $N = \tilde{O}(n)$, where n is the number of states, the estimate $\overline{\nabla} f(K)$ concentrates with *high probability* when projected to the direction of $\nabla f(K)$.

Our proof strategy allows us to significantly improve upon the existing literature both in terms of the required function evaluations and simulation time. Specifically, using the random search method (RS), the total number of function evaluations required in our results to achieve an accuracy level ϵ is proportional to $\log(1/\epsilon)$ compared to at least $(1/\epsilon^4)\log(1/\epsilon)$ in [13] and $1/\epsilon$ in [132]. Similarly, the simulation time that we require to achieve an accuracy level ϵ is proportional to $\log(1/\epsilon)$; this is in contrast to poly $(1/\epsilon)$ simulation times in [13] and infinite simulation time in [132].

Algorithm 1 produces a biased estimate $\overline{\nabla} f(K)$ of the gradient $\nabla f(K)$. Herein, we first introduce an unbiased estimate $\widehat{\nabla} f(K)$ of $\nabla f(K)$ and establish that the distance $\|\widehat{\nabla} f(K) - \overline{\nabla} f(K)\|_F$ can be readily controlled by choosing a large simulation time τ and an appropriate smoothing parameter r in Algorithm 1; we call this distance the estimation bias. Next, we

show that with $N = \tilde{O}(n)$ samples, the unbiased estimate $\widehat{\nabla} f(K)$ becomes highly correlated with $\nabla f(K)$. We exploit this fact in our convergence analysis.

6.6.1 Bias in gradient estimation due to finite simulation time

We first introduce an unbiased estimate of the gradient that is used to quantify the bias. For any $\tau \geq 0$ and $x_0 \in \mathbb{R}^n$, let

$$f_{x_0,\tau}(K) \;:=\; \int_0^\tau \left(x^T(t)Qx(t) \,+\, u^T(t)Ru(t)\right) \mathrm{d}t$$

denote the τ-truncated version of the LQR objective function associated with system (6.1b) with the initial condition $x(0) = x_0$ and feedback law $u = -Kx$ for all $K \in \mathbb{R}^{m\times n}$. Note that for any $K \in \mathcal{S}_K$ and $x(0) = x_0 \in \mathbb{R}^n$, the infinite-horizon cost

$$f_{x_0}(K) \;:=\; f_{x_0,\infty}(K) \tag{6.24a}$$

exists and it satisfies $f(K) = \mathbb{E}_{x_0}[f_{x_0}(K)]$. Furthermore, the gradient of $f_{x_0}(K)$ is given by (cf. (6.5))

$$\nabla f_{x_0}(K) \;=\; 2\,(RK \,-\, B^T P(K))X_{x_0}(K) \tag{6.24b}$$

where $X_{x_0}(K) = -\mathcal{A}_K^{-1}(x_0 x_0^T)$ is determined by the closed-loop Lyapunov operator in (6.7) and $P(K) = -(\mathcal{A}_K^*)^{-1}(Q + K^T R K)$. Note that the gradients $\nabla f(K)$ and $\nabla f_{x_0}(K)$ are linear in $X(K) = -\mathcal{A}_K^{-1}(\Omega)$ and $X_{x_0}(K)$, respectively. Thus, for any zero-mean random initial condition $x(0) = x_0$ with covariance $\mathbb{E}[x_0 x_0^T] = \Omega$, the linearity of the closed-loop Lyapunov operator $\mathcal{A}_K$ implies

$$\mathbb{E}_{x_0}[X_{x_0}(K)] \;=\; X(K),\; \mathbb{E}_{x_0}[\nabla f_{x_0}(K)] \;=\; \nabla f(K).$$

Let us define the following three estimates of the gradient

$$\begin{aligned}
\overline{\nabla} f(K) &:= \frac{1}{2rN} \sum_{i=1}^{N} \left(f_{x_i,\tau}(K + rU_i) - f_{x_i,\tau}(K - rU_i)\right) U_i \\
\widetilde{\nabla} f(K) &:= \frac{1}{2rN} \sum_{i=1}^{N} \left(f_{x_i}(K + rU_i) - f_{x_i}(K - rU_i)\right) U_i \\
\widehat{\nabla} f(K) &:= \frac{1}{N} \sum_{i=1}^{N} \langle \nabla f_{x_i}(K), U_i \rangle\, U_i
\end{aligned} \tag{6.25}$$

where $U_i \in \mathbb{R}^{m\times n}$ are i.i.d. random matrices with $\mathrm{vec}(U_i)$ uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$ and $x_i \in \mathbb{R}^n$ are i.i.d. initial conditions sampled from distribution $\mathcal{D}$. Here, $\widetilde{\nabla} f(K)$ is the infinite-horizon version of the output $\overline{\nabla} f(K)$ of Algorithm 1 and $\widehat{\nabla} f(K)$

provides an unbiased estimate of $\nabla f(K)$. To see this, note that by the independence of U_i and x_i we have

$$\begin{aligned}\mathbb{E}_{x_i,U_i}\Big[\mathrm{vec}(\widehat{\nabla} f(K))\Big] &= \mathbb{E}_{U_1}[\langle \nabla f(K), U_1\rangle \,\mathrm{vec}(U_1)] \\ &= \mathbb{E}_{U_1}[\mathrm{vec}(U_1)\mathrm{vec}(U_1)^T]\mathrm{vec}(\nabla f(K)) \;=\; \mathrm{vec}(\nabla f(K))\end{aligned}$$

and thus $\mathbb{E}[\widehat{\nabla} f(K)] = \nabla f(K)$. Here, we have utilized the fact that for the uniformly distributed random variable $\mathrm{vec}(U_1)$ over the sphere $\sqrt{mn}\, S^{mn-1}$, $\mathbb{E}_{U_1}[\mathrm{vec}(U_1)\mathrm{vec}(U_1)^T] = I$.

6.6.1.1 Local boundedness of the function $f(K)$

An important requirement for the gradient estimation scheme in Algorithm 1 is the stability of the perturbed closed-loop systems, i.e., $K \pm rU_i \in \mathcal{S}_K$; violating this condition leads to an exponential growth of the state and control signals. Moreover, this condition is necessary and sufficient for $\widetilde{\nabla} f(K)$ to be well defined. In Proposition 3, we establish a radius within which any perturbation of $K \in \mathcal{S}_K$ remains stabilizing.

Proposition 3 *For any stabilizing feedback gain $K \in \mathcal{S}_K$, we have $\{\hat{K} \in \mathbb{R}^{m\times n} \mid \|\hat{K}-K\|_2 < \zeta\} \subset \mathcal{S}_K$ where*

$$\zeta \;:=\; \lambda_{\min}(\Omega)/(2\,\|B\|_2\,\|X(K)\|_2)$$

and $X(K)$ is given by (6.4a).

Proof: See Appendix D.5. □

If we choose the parameter r in Algorithm 1 to be smaller than ζ, then the sample feedback gains $K \pm rU_i$ are all stabilizing. In this chapter, we further require that the parameter r is small enough so that $K \pm rU_i \in \mathcal{S}_K(2a)$ for all $K \in \mathcal{S}_K(a)$. Such upper bound on r is provided in the next lemma.

Lemma 4 *For any $U \in \mathbb{R}^{m\times n}$ with $\|U\|_F \le \sqrt{mn}$ and $K \in \mathcal{S}_K(a)$, $K + r(a)U \in \mathcal{S}_K(2a)$ where $r(a) := \tilde{c}/a$ for some positive constant $\tilde{c}$ that depends on the problem data.*

Proof: See Appendix D.5. □

Note that for any $K \in \mathcal{S}_K(a)$, and $r \le r(a)$ in Lemma 4, $\widetilde{\nabla} f(K)$ is well defined because $K + rU_i \in \mathcal{S}_K(2a)$ for all i.

6.6.1.2 Bounding the bias

Herein, we establish an upper bound on the difference between the output $\overline{\nabla} f(K)$ generated by Algorithm 1 and the unbiased estimate $\widehat{\nabla} f(K)$ of the gradient $\nabla f(K)$. We accomplish this by bounding the difference between these two quantities and $\widetilde{\nabla} f(K)$ through the use of the triangle inequality

$$\|\widehat{\nabla} f(K) - \overline{\nabla} f(K)\|_F \;\le\; \|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F \;+\; \|\widehat{\nabla} f(K) - \widetilde{\nabla} f(K)\|_F. \tag{6.26}$$

The first term on the right-hand side of (6.26) arises from a bias caused by the finite simulation time in Algorithm 1. The next proposition quantifies an upper bound on this term.

Proposition 4 *For any $K \in \mathcal{S}_K(a)$, the output of Algorithm 1 with parameter $r \leq r(a)$ (given by Lemma 4) satisfies*

$$\|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F \;\leq\; \frac{\sqrt{mn}\max_i \|x_i\|^2}{r}\,\kappa_1(2a)\,\mathrm{e}^{-\kappa_2(2a)\tau}$$

where $\kappa_1(a) > 0$ is a degree 5 polynomial and $\kappa_2(a) > 0$ is inversely proportional to a and they are given by (D.17).

Proof: See Appendix D.6. □

Although small values of r may result in a large error $\|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F$, the exponential dependence of the upper bound in Proposition 4 on the simulation time τ implies that this error can be readily controlled by increasing τ. In the next proposition, we handle the second term in (6.26).

Proposition 5 *For any $K \in \mathcal{S}_K(a)$ and $r \leq r(a)$ (given by Lemma 4), we have*

$$\|\widehat{\nabla} f(K) \;-\; \widetilde{\nabla} f(K)\|_F \;\leq\; \frac{(rmn)^2}{2}\,\ell(2a)\,\max_i \|x_i\|^2$$

where the function $\ell(a) > 0$ is a degree 4 polynomial and it is given by (D.21).

Proof: See Appendix D.7. □

The third-derivatives of the functions $f_{x_i}(K)$ are utilized in the proof of Proposition 5. It is also worth noting that unlike $\overline{\nabla} f(k)$ and $\widetilde{\nabla} f(K)$, the unbiased gradient estimate $\widehat{\nabla} f(K)$ is independent of the parameter r. Thus, Proposition 5 provides a quadratic upper bound on the estimation error in terms of r.

6.6.2 Correlation between gradient and gradient estimate

As mentioned earlier, one approach to analyzing convergence for the random search method in (RS) is to control the gradient estimation error $\overline{\nabla} f(K) - \nabla f(K)$ by choosing a large number of samples N. For the one-point gradient estimation setting, this approach was taken in [13] for the discrete-time LQR (and in [15] for the continuous-time LQR) and has led to an upper bound on the required number of samples for reaching ϵ-accuracy that grows at least proportionally to $1/\epsilon^4$. Alternatively, our proof exploits the problem structure and shows that with a linear number of samples $N = \tilde{O}(n)$, where n is the number of states, the gradient estimate $\widehat{\nabla} f(K)$ concentrates with *high probability* when projected to the direction

of $\nabla f(K)$. In particular, in Propositions 7 and 8 we show that the following events occur with high probability for some positive scalars μ_1, μ_2,

$$\mathsf{M}_1 \;:=\; \left\{ \left\langle \widehat{\nabla} f(K), \nabla f(K) \right\rangle \;\geq\; \mu_1 \|\nabla f(K)\|_F^2 \right\} \tag{6.27a}$$

$$\mathsf{M}_2 \;:=\; \left\{ \|\widehat{\nabla} f(K)\|_F^2 \;\leq\; \mu_2 \|\nabla f(K)\|_F^2 \right\}. \tag{6.27b}$$

To justify the definitions of these events, we first show that if they both take place then the unbiased estimate $\widehat{\nabla} f(K)$ can be used to decrease the objective error by a geometric factor.

Proposition 6 *[Approximate GD] If the matrix $G \in \mathbb{R}^{m\times n}$ and the feedback gain $K \in \mathcal{S}_K(a)$ are such that*

$$\langle G, \nabla f(K) \rangle \;\geq\; \mu_1 \|\nabla f(K)\|_F^2 \tag{6.28a}$$

$$\|G\|_F^2 \;\leq\; \mu_2 \|\nabla f(K)\|_F^2 \tag{6.28b}$$

for some positive scalars μ_1 and μ_2, then $K - \alpha G \in \mathcal{S}_K(a)$ for all $\alpha \in [0, \mu_1/(\mu_2 L_f)]$, and

$$f(K - \alpha G) \;-\; f(K^\star) \;\leq\; \gamma \left(f(K) \;-\; f(K^\star) \right)$$

with $\gamma = 1 - \mu_f \mu_1 \alpha$. Here, L_f and μ_f are the smoothness and the PL parameters of the function f over $\mathcal{S}_K(a)$.

Proof: See Appendix D.8. □

Remark 4 *The fastest convergence rate guaranteed by Proposition 6, $\gamma = 1 - \mu_f \mu_1^2/(L_f \mu_2)$, is achieved with the stepsize $\alpha = \mu_1/(\mu_2 L_f)$. This rate bound is tight in the sense that if $G = c\nabla f(K)$, for some $c > 0$, we recover the standard convergence rate $\gamma = 1 - \mu_f/L_f$ of gradient descent.*

We next quantify the probability of the events M_1 and M_2. In our proofs, we exploit modern non-asymptotic statistical analysis of the concentration of random variables around their average. While in Appendix D.10 we set notation and provide basic definitions of key concepts, we refer the reader to a recent book [160] for a comprehensive discussion. Herein, we use c, c', c'', etc. to denote positive absolute constants.

6.6.2.1 Handling M_1

We first exploit the problem structure to confine the dependence of $\widehat{\nabla} f(K)$ on the random initial conditions x_i into a zero-mean random vector. In particular, for any $K \in \mathcal{S}_K$ and $x_0 \in \mathbb{R}^n$,

$$\nabla f(K) \;=\; EX, \quad \nabla f_{x_0}(K) \;=\; EX_{x_0}$$

where $E := 2(RK - B^T P(K)) \in \mathbb{R}^{m\times n}$ is a fixed matrix, $X = -\mathcal{A}_K^{-1}(\Omega)$, and $X_{x_0} = -\mathcal{A}_K^{-1}(x_0 x_0^T)$. This allows us to represent the unbiased estimate $\widehat{\nabla} f(K)$ of the gradient as

$$\widehat{\nabla} f(K) \;=\; \frac{1}{N}\sum_{i=1}^{N} \langle EX_{x_i}, U_i\rangle\, U_i \;=\; \widehat{\nabla}_1 \;+\; \widehat{\nabla}_2 \tag{6.29a}$$

$$\widehat{\nabla}_1 \;=\; \frac{1}{N}\sum_{i=1}^{N} \langle E(X_{x_i} - X), U_i\rangle\, U_i \tag{6.29b}$$

$$\widehat{\nabla}_2 \;=\; \frac{1}{N}\sum_{i=1}^{N} \langle \nabla f(K), U_i\rangle\, U_i. \tag{6.29c}$$

Note that $\widehat{\nabla}_2$ does not depend on the initial conditions x_i. Moreover, from $\mathbb{E}[X_{x_i}] = X$ and the independence of X_{x_i} and U_i, we have $\mathbb{E}[\widehat{\nabla}_1] = 0$ and $\mathbb{E}[\widehat{\nabla}_2] = \nabla f(K)$.

In Lemma 5, we show that $\left\langle \widehat{\nabla}_1, \nabla f(K)\right\rangle$ can be made arbitrary small with a large number of samples N. This allows us to analyze the probability of the event M_1 in (6.27).

Lemma 5 *Let $U_1, \dots, U_N \in \mathbb{R}^{m\times n}$ be i.i.d. random matrices with each $\mathrm{vec}(U_i)$ uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$ and let $X_1, \dots, X_N \in \mathbb{R}^{n\times n}$ be i.i.d. random matrices distributed according to $\mathcal{M}(xx^T)$. Here, $\mathcal{M}$ is a linear operator and $x \in \mathbb{R}^n$ is a random vector whose entries are i.i.d., zero-mean, unit-variance, sub-Gaussian random variables with sub-Gaussian norm less than κ. For any fixed matrix $E \in \mathbb{R}^{m\times n}$ and positive scalars δ and β, if*

$$N \;\geq\; C\,(\beta^2\kappa^2/\delta)^2\,(\|\mathcal{M}^*\|_2 \;+\; \|\mathcal{M}^*\|_S)^2\, n\log^6 n \tag{6.30}$$

then, with probability not smaller than $1 - C'N^{-\beta} - 4N\mathrm{e}^{-\frac{n}{8}}$,

$$\left|\frac{1}{N}\sum_{i=1}^{N} \langle E\,(X_i - X)\,, U_i\rangle\,\langle EX, U_i\rangle\right| \;\leq\; \delta\|EX\|_F\|E\|_F$$

where $X := \mathbb{E}[X_1] = \mathcal{M}(I)$.

Proof: See Appendix D.9. □

In Lemma 6, we show that $\left\langle \widehat{\nabla}_2, \nabla f(K)\right\rangle$ concentrates with high probability around its average $\|\nabla f(K)\|_F^2$.

Lemma 6 *Let $U_1, \dots, U_N \in \mathbb{R}^{m\times n}$ be i.i.d. random matrices with each $\mathrm{vec}(U_i)$ uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$. Then, for any $W \in \mathbb{R}^{m\times n}$ and $t \in (0, 1]$,*

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} \langle W, U_i\rangle^2 \;<\; (1 \;-\; t)\|W\|_F^2\right\} \;\leq\; 2\,\mathrm{e}^{-cNt^2}.$$

Proof: See Appendix D.9. □

In Proposition 7, we use Lemmas 5 and 6 to address M_1.

Proposition 7 *Under Assumption 1, for any stabilizing feedback gain $K \in \mathcal{S}_K$ and positive scalar β, if*

$$N \geq C_1 \frac{\beta^4 \kappa^4}{\lambda_{\min}^2(X)} \left(\|(\mathcal{A}_K^*)^{-1}\|_2 + \|(\mathcal{A}_K^*)^{-1}\|_S \right)^2 n \log^6 n$$

then the event M_1 in (6.27) with $\mu_1 := 1/4$ satisfies $\mathbb{P}(\mathsf{M}_1) \geq 1 - C_2 N^{-\beta} - 4N\mathrm{e}^{-\frac{n}{8}} - 2\mathrm{e}^{-C_3 N}$.

Proof: We use Lemma 5 with $\delta := \lambda_{\min}(X)/4$ to show that

$$\left| \left\langle \widehat{\nabla}_1, \nabla f(K) \right\rangle \right| \leq \delta \, \|EX\|_F \, \|E\|_F \leq \frac{1}{4} \|EX\|_F^2 = \frac{1}{4} \|\nabla f(K)\|_F^2. \tag{6.31a}$$

holds with probability not smaller than $1 - C' N^{-\beta} - 4N\mathrm{e}^{-\frac{n}{8}}$. Furthermore, Lemma 6 with $t := 1/2$ implies that

$$\left\langle \widehat{\nabla}_2, \nabla f(K) \right\rangle \geq \frac{1}{2} \|\nabla f(K)\|_F^2 \tag{6.31b}$$

holds with probability not smaller than $1 - 2\mathrm{e}^{-cN}$. Since $\widehat{\nabla} f(K) = \widehat{\nabla}_1 + \widehat{\nabla}_2$, we can use a union bound to combine (6.31a) and (6.31b). This together with a triangle inequality completes the proof. □

6.6.2.2 Handling M_2

In Lemma 7, we quantify a high probability upper bound on $\|\widehat{\nabla}_1\|_F / \|\nabla f(K)\|$. This lemma is analogous to Lemma 5 and it allows us to analyze the probability of the event M_2 in (6.27).

Lemma 7 *Let X_i and U_i with $i = 1, \ldots, N$ be random matrices defined in Lemma 5, $X := \mathbb{E}[X_1]$, and let $N \geq c_0 n$. Then, for any $E \in \mathbb{R}^{m \times n}$ and positive scalar β,*

$$\frac{1}{N} \| \sum_{i=1}^{N} \langle E\left(X_i - X\right), U_i \rangle \, U_i \|_F \leq c_1 \beta \, \kappa^2 (\|\mathcal{M}^*\|_2 + \|\mathcal{M}^*\|_S) \|E\|_F \sqrt{mn} \, \log n$$

with probability not smaller than $1 - c_2(n^{-\beta} + N\mathrm{e}^{-\frac{n}{8}})$.

Proof: See Appendix D.10. □

In Lemma 8, we quantify a high probability upper bound on $\|\widehat{\nabla}_2\|_F / \|\nabla f(K)\|$.

Lemma 8 *Let $U_1, \dots, U_N \in \mathbb{R}^{m\times n}$ be i.i.d. random matrices with $\mathrm{vec}(U_i)$ being uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$ and let $N \geq Cn$. Then, for any $W \in \mathbb{R}^{m\times n}$,*

$$\mathbb{P}\left\{ \frac{1}{N} \| \sum_{j=1}^{N} \langle W, U_j \rangle\, U_j \|_F \;>\; C'\sqrt{m}\|W\|_F \right\} \;\leq\; 2N\mathrm{e}^{-\frac{mn}{8}} \;+\; 2\mathrm{e}^{-\hat{c}N}.$$

Proof: See Appendix D.10. □

In Proposition 8, we use Lemmas 7 and 8 to address M_2.

Proposition 8 *Let Assumption 1 hold. Then, for any $K \in \mathcal{S}_K$, scalar $\beta > 0$, and $N \geq C_4 n$, the event M_2 in (6.27) with $\mu_2 := C_5\big(\beta\kappa^2 \frac{\|(\mathcal{A}_K^*)^{-1}\|_2 + \|(\mathcal{A}_K^*)^{-1}\|_S}{\lambda_{\min}(X)}\sqrt{mn}\log n + \sqrt{m}\big)^2$ satisfies*

$$\mathbb{P}(\mathsf{M}_2) \;\geq\; 1 - C_6(n^{-\beta} + N\mathrm{e}^{-\frac{n}{8}} + \mathrm{e}^{-C_7 N}).$$

Proof: We use Lemma 7 to show that, with probability at least $1 - c_2(n^{-\beta} + N\mathrm{e}^{-\frac{n}{8}})$, $\widehat{\nabla}_1$ satisfies

$$\|\widehat{\nabla}_1\|_F \;\leq\; c_1\beta\kappa^2(\|(\mathcal{A}_K^*)^{-1}\|_2 + \|(\mathcal{A}_K^*)^{-1}\|_S)\|E\|_F\sqrt{mn}\,\log n \;\leq\; c_1\beta\kappa^2\,\frac{\|(\mathcal{A}_K^*)^{-1}\|_2 + \|(\mathcal{A}_K^*)^{-1}\|_S}{\lambda_{\min}(X)}\,\|\nabla f(K)\|_F\sqrt{mn}\,\log n.$$

Furthermore, we can use Lemma 8 to show that, with probability not smaller than $1 - 2N\mathrm{e}^{-\frac{mn}{8}} - 2\mathrm{e}^{-\hat{c}N}$, $\widehat{\nabla}_2$ satisfies

$$\|\widehat{\nabla}_2\|_F \;\leq\; C'\sqrt{m}\|\nabla f(K)\|_F.$$

Now, since $\widehat{\nabla} f(K) = \widehat{\nabla}_1 + \widehat{\nabla}_2$, we can use a union bound to combine the last two inequalities. This together with a triangle inequality completes the proof. □

6.7 Model-free control design

In this section, we prove a more formal version of Theorem 3.

Theorem 4 *Consider the random search method* (RS) *that uses the gradient estimates of Algorithm 1 for finding the optimal solution $K^\star$ of LQR problem* (6.3)*. Let the initial condition x_0 obey Assumption 1 and let the simulation time τ, the smoothing constant r, and the number of samples N satisfy*

$$\tau \;\geq\; \theta'(a)\log\frac{1}{r\epsilon}, \quad r < \min\{r(a), \theta''(a)\sqrt{\epsilon}\}, \quad N \;\geq\; c_1(1 + \beta^4\kappa^4\,\theta(a)\log^6 n)\,n \qquad (6.32)$$

for some $\beta > 0$ and a desired accuracy $\epsilon > 0$. Then, for any initial condition $K^0 \in \mathcal{S}_K(a)$, (RS) *with the constant stepsize $\alpha \leq 1/(32\mu_2(a)L_f)$ achieves $f(K^k) - f(K^\star) \leq \epsilon$ with probability not smaller than $1 - kp - 2kN\mathrm{e}^{-n}$ in at most*

$$k \;\leq\; \left(\log \frac{f(K^0) \;-\; f(K^\star)}{\epsilon}\right) \Bigg/ \left(\log \frac{1}{1 \;-\; \mu_f(a)\alpha/8}\right)$$

iterations. Here, $p := c_2(n^{-\beta} + N^{-\beta} + N\mathrm{e}^{-\frac{n}{8}} + \mathrm{e}^{-c_3 N})$, $\mu_2 := c_4\left(\sqrt{m} + \beta\kappa^2\theta(a)\sqrt{mn}\log n\right)^2$, $c_1, \ldots, c_4$ are positive absolute constants, μ_f and L_f are the PL and smoothness parameters of the function f over the sublevel set $\mathcal{S}_K(a)$, θ, θ', θ'' are positive functions that depend only on the parameters of the LQR problem, and $r(a)$ is given by Lemma 4.

Proof: The proof combines Propositions 4, 5, 6, 7, and 8. We first show that for any $r \leq r(a)$ and $\tau > 0$,

$$\|\overline{\nabla} f(K) \;-\; \widehat{\nabla} f(K)\|_F \;\leq\; \sigma \tag{6.33}$$

with probability not smaller than $1 - 2N\mathrm{e}^{-n}$, where

$$\sigma := c_5(\kappa^2 + 1)\left(\frac{n\sqrt{m}}{r}\kappa_1(2a)\mathrm{e}^{-\kappa_2(2a)\tau} + \frac{r^2 m^2 n^{\frac{5}{2}}}{2}\ell(2a)\right).$$

Here, $r(a)$, $\kappa_i(a)$, and $\ell(a)$ are positive functions that are given by Lemma 4, Eq. (D.17), and Eq. (D.21), respectively.

Under Assumption 1, the vector $v \sim \mathcal{D}$ satisfies [160, Eq. (3.3)],

$$\mathbb{P}\left\{\|v\| \;\leq\; c_5(\kappa^2 + 1)\sqrt{n}\right\} \geq 1 - 2\mathrm{e}^{-n}.$$

Thus, for the random initial conditions $x_1, \ldots, x_N \sim \mathcal{D}$, we can apply the union bound (Boole's inequality) to obtain

$$\mathbb{P}\left\{\max_i \|x_i\| \;\leq\; c_5(\kappa^2 + 1)\sqrt{n}\right\} \;\geq\; 1 \;-\; 2N\mathrm{e}^{-n}. \tag{6.34}$$

Now, we combine Propositions 4 and 5 to write

$$\|\overline{\nabla} f(K) - \widehat{\nabla} f(K)\|_F \;\leq\; \left(\frac{\sqrt{mn}}{r}\,\kappa_1(2a)\mathrm{e}^{-\kappa_2(2a)\tau} \;+\; \frac{(rmn)^2}{2}\,\ell(2a)\right)\max_i \|x_i\|^2 \;\leq\; \sigma.$$

The first inequality is obtained by combining Propositions 4 and 5 through the use of the triangle inequality, and the second inequality follows from (6.34). This completes the proof of (6.33).

Let $\theta(a)$ be a uniform upper bound on

$$\frac{\|(\mathcal{A}_K^*)^{-1}\|_2 + \|(\mathcal{A}_K^*)^{-1}\|_S}{\lambda_{\min}(X)} \;\leq\; \theta(a)$$

for all $K \in \mathcal{S}_K(a)$; see Appendix D.12 for a discussion on $\theta(a)$. Since, the number of samples satisfies (6.32), for any given $K \in \mathcal{S}_K(a)$, we can combine Propositions 7 and 8 with a union bound to show that

$$\left\langle \widehat{\nabla} f(K), \nabla f(K) \right\rangle \;\geq\; \mu_1 \|\nabla f(K)\|_F^2 \tag{6.35a}$$

$$\|\widehat{\nabla} f(K)\|_F^2 \;\leq\; \mu_2 \|\nabla f(K)\|_F^2 \tag{6.35b}$$

holds with probability not smaller than $1-p$, where $\mu_1 = 1/4$, and μ_2 and p are determined in the statement of the theorem.

Without loss of generality, let us assume that the initial error satisfies $f(K^0) - f(K^\star) > \epsilon$. We next show that

$$\left\langle \overline{\nabla} f(K^0), \nabla f(K^0) \right\rangle \;\geq\; \frac{\mu_1}{2} \|\nabla f(K^0)\|_F^2 \tag{6.36a}$$

$$\|\overline{\nabla} f(K^0)\|_F^2 \;\leq\; 4\mu_2 \|\nabla f(K^0)\|_F^2 \tag{6.36b}$$

holds with probability not smaller than $1 - p - 2N\mathrm{e}^{-n}$.

Since the function f is gradient dominant over the sublevel set $\mathcal{S}_K(a)$ with parameter μ_f, combining $f(K^0) - f(K^\star) > \epsilon$ and (6.18) yields $\|\nabla f(K^0)\|_F \geq \sqrt{2\mu_f \epsilon}$. Also, let the positive scalars $\theta'(a)$ and $\theta''(a)$ be such that for any pair of τ and r satisfying $\tau \geq \theta'(a)\log(1/(r\epsilon))$ and $r < \min\{r(a), \theta''(a)\sqrt{\epsilon}\}$, the upper bound σ in (6.33) becomes smaller than $\sigma \leq \sqrt{2\mu_f\epsilon}\,\min\{\mu_1/2, \sqrt{\mu_2}\}$. The choice of θ' and θ'' with the above property is straightforward using the definition of σ. Combining $\|\nabla f(K^0)\|_F \geq \sqrt{2\mu_f\epsilon}$ and $\sigma \leq \sqrt{2\mu_f\epsilon}\,\min\{\mu_1/2, \sqrt{\mu_2}\}$ yields

$$\sigma \;\leq\; \|\nabla f(K^0)\|_F \,\min\{\mu_1/2, \sqrt{\mu_2}\}. \tag{6.37}$$

Using the union bound, we have

$$\begin{aligned}
\left\langle \overline{\nabla} f(K^0), \nabla f(K^0) \right\rangle \;&=\; \left\langle \widehat{\nabla} f(K^0), \nabla f(K^0) \right\rangle \;+\; \left\langle \overline{\nabla} f(K^0) - \widehat{\nabla} f(K^0), \nabla f(K^0) \right\rangle \\
&\overset{(a)}{\geq}\; \mu_1 \|\nabla f(K^0)\|_F^2 \;-\; \|\overline{\nabla} f(K^0) - \widehat{\nabla} f(K^0)\|_F \|\nabla f(K^0)\|_F \\
&\overset{(b)}{\geq}\; \mu_1 \|\nabla f(K^0)\|_F^2 \;-\; \sigma \|\nabla f(K^0)\|_F \;\overset{(c)}{\geq}\; \frac{\mu_1}{2} \|\nabla f(K^0)\|_F^2
\end{aligned}$$

with probability not smaller than $1 - p - 2N\mathrm{e}^{-n}$. Here, (a) follows from combining (6.35a) and the Cauchy-Schwartz inequality, (b) follows from (6.33), and (c) follows from (6.37). Moreover,

$$\begin{aligned}
\|\overline{\nabla} f(K^0)\|_F \;&\overset{(a)}{\leq}\; \|\widehat{\nabla} f(K^0)\|_F + \|\overline{\nabla} f(K^0) - \widehat{\nabla} f(K^0)\|_F \\
&\overset{(b)}{\leq}\; \sqrt{\mu_2} \|\nabla f(K^0)\|_F \;+\; \sigma \;\overset{(c)}{\leq}\; 2\sqrt{\mu_2} \|\nabla f(K^0)\|_F
\end{aligned}$$

where (a) follows from the triangle inequality, (b) from (6.33), and (c) from (6.37). This completes the proof of (6.36).

Inequality (6.36) allows us to apply Proposition 6 and obtain with probability not smaller than $1 - p - 2N\mathrm{e}^{-n}$ that for the stepsize $\alpha \leq \mu_1/(8\mu_2 L_f)$, we have $K^1 \in \mathcal{S}_K(a)$ and also $f(K^1) - f(K^\star) \leq \gamma \left(f(K^0) - f(K^\star)\right)$, with $\gamma = 1 - \mu_f \mu_1 \alpha/2$, where L_f is the smoothness parameter of the function f over $\mathcal{S}_K(a)$. Finally, using the union bound, we can repeat this procedure via induction to obtain that for some

$$k \;\leq\; \left(\log \frac{f(K^0) - f(K^\star)}{\epsilon}\right) \Big/ \left(\log \frac{1}{\gamma}\right)$$

the error satisfies

$$f(K^k) - f(K^\star) \;\leq\; \gamma^k \left(f(K^0) - f(K^\star)\right) \;\leq\; \epsilon$$

with probability not smaller than $1 - kp - 2kN\mathrm{e}^{-n}$. □

Remark 5 *For the failure probability in Theorem 4 to be negligible, the problem dimension n needs to be large. Moreover, to account for the conflicting term $N\mathrm{e}^{-n/8}$ in the failure probability, we can require a crude exponential bound $N \leq \mathrm{e}^{n/16}$ on the sample size. We also note that although Theorem 4 only guarantees convergence in the objective value, similar to the proof of Theorem 1, we can use Lemma 2 that relates the error in optimization variable, K, and the error in the objective function, $f(K)$, to obtain convergence guarantees in the optimization variable as well.*

Remark 6 *Theorem 4 requires the lower bound on the simulation time τ in (6.32) to ensure that, for any desired accuracy ϵ, the smoothing constant r satisfies $r \geq (1/\epsilon)\,\mathrm{e}^{-\tau/\theta'(a)}$. As we demonstrate in the proof, this requirement accounts for the bias that arises from a finite value of τ. Since this form of bias can be readily controlled by increasing τ, the above lower bound on r does not contradict the upper bound $r = O(\sqrt{\epsilon})$ required by Theorem 4. Finally, we note that letting $r \to 0$ can cause large bias in the presence of other sources of inaccuracy in the function approximation process.*

6.8 Computational experiments

We consider a mass-spring-damper system with s masses, where we set all mass, spring, and damping constants to unity. In state-space representation (6.1b), the state $x = [\,p^T\, v^T\,]^T$ contains the position and velocity vectors and the dynamic and input matrices are given by

$$A \;=\; \begin{bmatrix} 0 & I \\ -T & -T \end{bmatrix},\; B \;=\; \begin{bmatrix} 0 \\ I \end{bmatrix}$$

where 0 and I are $s \times s$ zero and identity matrices, and T is a Toeplitz matrix with 2 on the main diagonal and -1 on the first super and sub-diagonals.

6.8.1 Known model

To compare the performance of gradient descent methods (GD) and (GY) on K and Y, we solve the LQR problem with $Q = I + 100\,\mathrm{e}_1\mathrm{e}_1^T$, $R = I + 1000\,\mathrm{e}_4\mathrm{e}_4^T$, and $\Omega = I$ for $s \in \{10,\, 20\}$

masses (i.e., $n = 2s$ state variables), where e_i is the ith unit vector in the standard basis of $\mathbb{R}^n$.

Figure 6.2 illustrates the convergence curves for both algorithms with a stepsize selected using a backtracking procedure that guarantees stability of the closed-loop system. Both algorithms were initialized with $Y^0 = K^0 = 0$. Even though Fig. 6.2 suggests that gradient decent/flow on $\mathcal{S}_K$ converges faster than that on $\mathcal{S}_Y$, this observation does not hold in general.

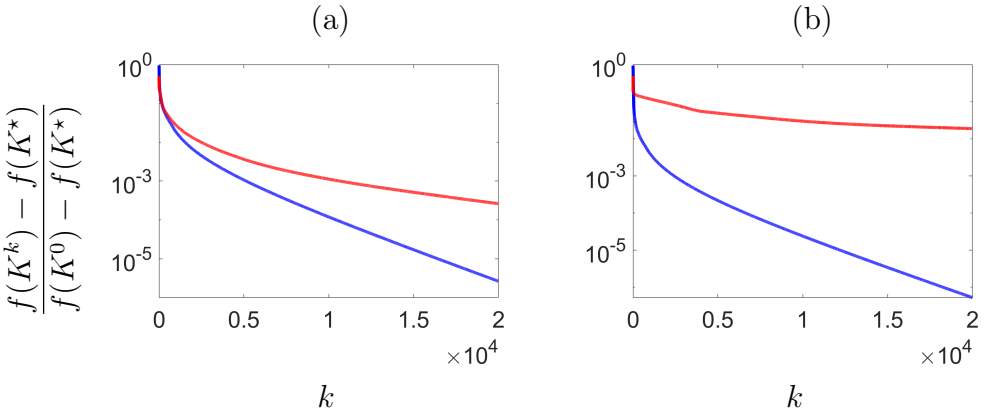


Figure 6.2: Convergence curves for gradient descent (blue) over the set $\mathcal{S}_K$, and gradient descent (red) over the set $\mathcal{S}_Y$ with (a) $s = 10$ and (b) $s = 20$ masses.

6.8.2 Unknown model

To illustrate our results on the accuracy of the gradient estimation in Algorithm 1 and the efficiency of our random search method, we consider the LQR problem with Q and R equal to identity for $s = 10$ masses (i.e., $n = 20$ state variables). We also let the initial conditions x_i in Algorithm 1 be standard normal and use $N = n = 2s$ samples.

Figure 6.3 (a) illustrates the dependence of $\|\widehat{\nabla} f(K) - \overline{\nabla} f(K)\|_F / \|\widehat{\nabla} f(K)\|_F$ on the simulation time τ for $K = 0$ and two values of the smoothing parameter $r = 10^{-4}$ (blue) and $r = 10^{-5}$ (red). We observe an exponential decrease in error for small values of τ. In addition, the error does not pass a saturation level which is determined by r. We also see that, as r decreases, this saturation level becomes smaller. These observations are in harmony with our theoretical developments; in particular, combining Propositions 4 and 5 through the use of the triangle inequality yields

$$\|\widehat{\nabla} f(K) - \overline{\nabla} f(K)\|_F \leq \left(\frac{\sqrt{mn}}{r} \kappa_1(2a)\, \mathrm{e}^{-\kappa_2(2a)\tau} + \frac{r^2 m^2 n^2}{2} \ell(2a) \right) \max_i \|x_i\|^2.$$

This upper bound clearly captures the exponential dependence of the bias on the simulation time τ as well as the saturation level that depends quadratically on the smoothing parameter r.

In Fig. 6.3 (b), we demonstrate the dependence of the total relative error $\|\nabla f(K) - \overline{\nabla} f(K)\|_F / \|\nabla f(K)\|_F$ on the simulation time τ for two values of the smoothing parameter $r = 10^{-4}$ (blue) and $r = 10^{-5}$ (red), resulting from the use of $N = n$ samples. We observe that the distance between the approximate gradient and the true gradient is rather large.

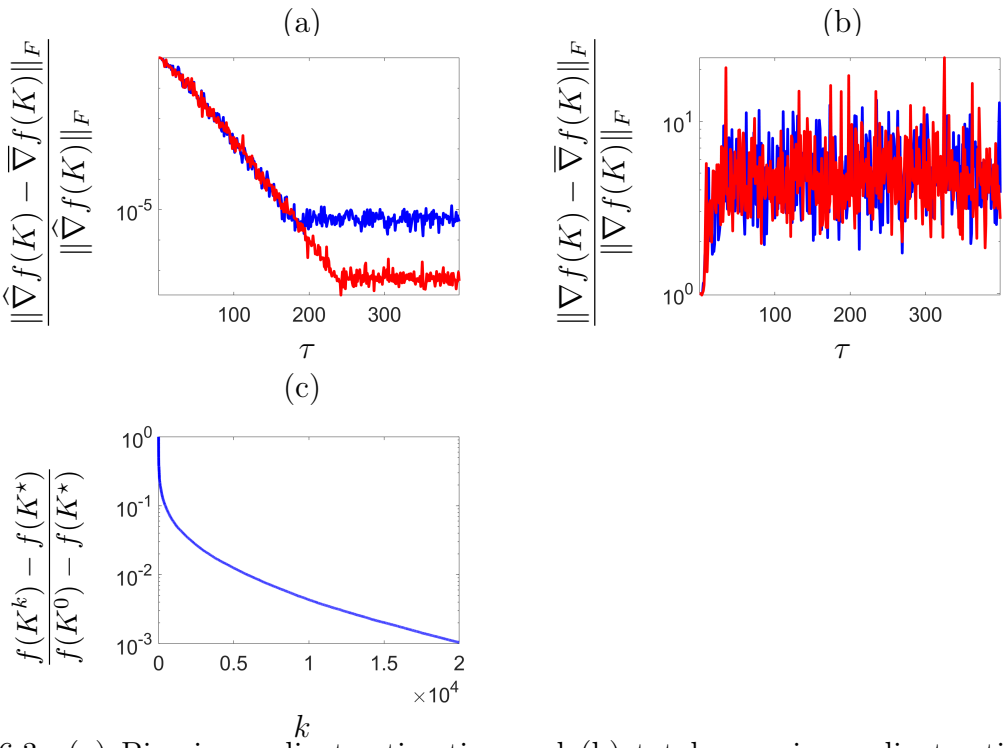


Figure 6.3: (a) Bias in gradient estimation and (b) total error in gradient estimation as functions of the simulation time τ. The blue and red curves correspond to two values of the smoothing parameter $r = 10^{-4}$ and $r = 10^{-5}$, respectively. (c) Convergence curve of the random search method (RS).

This is exactly why prior analysis of sample complexity and simulation time is subpar to our results. In contrast to the existing results which rely on the use of the estimation error shown in Fig. 6.3 (b), our analysis shows that the simulated gradient $\overline{\nabla} f(K)$ is close to the gradient estimate $\widehat{\nabla} f(K)$. While $\widehat{\nabla} f(K)$ is not close to the true gradient $\nabla f(K)$, it is highly correlated with it. This is sufficient for establishing convergence guarantees and it allows us to significantly improve upon existing results [13], [132] in terms of sample complexity and simulation time reducing both to $O(\log(1/\epsilon))$.

Finally, Fig. 6.3 (c) demonstrates linear convergence of the random search method (RS) with stepsize $\alpha = 10^{-4}$, $r = 10^{-5}$, and $\tau = 200$ in Algorithm 1, as established in Theorem 4. In this experiment, we implemented Algorithm 1 using the ode45 and trapz subroutines in MATLAB to numerically integrate the state/input penalties with the corresponding weight matrices Q and R. However, our theoretical results only account for an approximation error that arises from a finite simulation horizon. Clearly, employing empirical ODE solvers and numerical integration may introduce additional errors in our gradient approximation that require further scrutiny.

6.9 Concluding remarks

We prove exponential/linear convergence of gradient flow/descent algorithms for solving the continuous-time LQR problem based on a nonconvex formulation that directly searches for

the controller. A salient feature of our analysis is that we relate the gradient-flow dynamics associated with this nonconvex formulation to that of a convex reparameterization. This allows us to deduce convergence of the nonconvex approach from its convex counterpart. We also establish a bound on the sample complexity of the random search method for solving the continuous-time LQR problem that does not require the knowledge of system parameters. We have recently proved similar result for the discrete-time LQR problem [87].

Our ongoing research directions include: (i) providing theoretical guarantees for the convergence of gradient-based methods for sparsity-promoting as well as structured control synthesis; and (ii) extension to nonlinear systems via successive linearization techniques.

Chapter 7

Random search for discrete-time LQR

Model-free reinforcement learning techniques directly search over the parameter space of controllers. Although this often amounts to solving a nonconvex optimization problem, for benchmark control problems simple local search methods exhibit competitive performance. To understand this phenomenon, we study the discrete-time Linear Quadratic Regulator (LQR) problem with unknown state-space parameters. In spite of the lack of convexity, we establish that the random search method with two-point gradient estimates and a fixed number of roll-outs achieves ϵ-accuracy in $O(\log(1/\epsilon))$ iterations. This significantly improves existing results on the model-free LQR problem which require $O(1/\epsilon)$ total roll-outs.

7.1 Introduction

We study the sample complexity and convergence of random search method for the infinite-horizon discrete-time LQR problem. Random search method is a derivative-free optimization algorithm that directly searches over the parameter space of controllers using approximations of the gradient obtained through simulation data. Despite its simplicity, this approach has been used to solve benchmark control problems with state-of-the-art sample efficiency [22], [151]. However, even for the standard LQR problem, many open theoretical questions surround convergence properties and sample complexity of this method mainly because of the lack of convexity.

For *discrete-time* LQR problem, global convergence guarantees were recently provided for gradient descent and the random search method with one-point gradient estimates [13]. The key observation was that the LQR cost satisfies the Polyak-Łojasiewicz (PL) condition which can ensure convergence of gradient descent at a linear rate even for nonconvex problems. This reference also established a bound on the sample complexity of random search for reaching the error tolerance ϵ that requires a number of function evaluations that is proportional to $(1/\epsilon^4)\log(1/\epsilon)$. Extensions to the *continuous-time* LQR [15], [88], the $\mathcal{H}_\infty$ regularized LQR [134], and Markovian jump linear systems [133] have also been made.

Assuming access to the infinite horizon cost, the number of function evaluations for the random search method with one-point estimates was improved to $1/\epsilon^2$ in [132]. Moreover, this work showed that the use of two-point estimates reduces the number of function evaluations to $1/\epsilon$. Apart from the PL property, these results do not exploit structure of the LQR problem. Our recent work [14] focused on the *continuous-time* LQR problem, and established that the random search method with two-point gradient estimates converges to

the optimal solution at a linear rate with high probability. In this chapter, we extend the results of [14] to the *discrete-time* case. Relative to the existing literature, our results offer a significant improvement both in terms of the required number of function evaluations and simulation time. Specifically, the total number of function evaluations to achieve an ϵ-accuracy is proportional to $\log(1/\epsilon)$ compared to at least $(1/\epsilon^4)\log(1/\epsilon)$ in [13] and $1/\epsilon$ in [132]. Similarly, the required simulation time is proportional to $\log(1/\epsilon)$; this is in contrast to [13] which requires $\text{poly}(1/\epsilon)$ simulation time.

7.2 State-feedback characterization

Consider the LTI system

$$x^{t+1} \;=\; Ax^t \;+\; Bu^t, \quad x^0 \;=\; \zeta \tag{7.1a}$$

where $x^t \in \mathbb{R}^n$ is the state, $u^t \in \mathbb{R}^m$ is the control input, A and B are constant matrices, and $x^0 = \zeta$ is a zero-mean random initial condition with distribution $\mathcal{D}$. The LQR problem associated with system (7.1a) is given by

$$\underset{x,\,u}{\text{minimize}} \;\; \mathbb{E}\left[\sum_{t=0}^{\infty}(x^t)^T Q\, x^t \;+\; (u^t)^T R\, u^t\right] \tag{7.1b}$$

where Q and R are positive definite matrices and the expectation is taken over $\zeta \sim \mathcal{D}$. For a controllable pair (A, B), the solution to (7.1) takes a state-feedback form,

$$u^t \;=\; -K^\star x^t \;=\; -(R \;+\; B^T P^\star B)^{-1} B^T P^\star A\, x^t$$

where $P^\star$ is the unique positive definite solution to the Algebraic Riccati Equation (ARE) ,

$$A^T P^\star A \;+\; Q \;-\; A^T P^\star B(R \;+\; B^T P^\star B)^{-1} B^T P^\star A \;=\; P^\star.$$

When the model parameters A and B are known, the ARE can be solved efficiently via a variety of techniques [138], [161]. However, these techniques are not directly applicable when the matrices A and B are not known. One approach to dealing with the model-free scenario is to use the linearity of the optimal controller and reformulate the LQR problem as an optimization over state-feedback gains,

$$\underset{K}{\text{minimize}} \;\; f(K) \;:=\; \mathbb{E}\left[f_\zeta(K)\right] \tag{7.2}$$

where $f_\zeta(K) := \left\langle Q + K^T R\, K, X_\zeta(K)\right\rangle = \zeta^T P(K)\zeta$ and the matrices $P(K)$ and $X_\zeta(K)$ are given by

$$\begin{aligned} P(K) \;&:=\; \sum_{t=0}^{\infty}((A - BK)^T)^t (Q + K^T R\, K)(A - BK)^t \\ X_\zeta(K) \;&:=\; \sum_{t=0}^{\infty}(A - BK)^t\, \zeta\zeta^T((A - BK)^T)^t. \end{aligned} \tag{7.3}$$

Here, $f_\zeta(K)$ determines the LQR cost in (7.1b) associated with the feedback law $u = -Kx$ and the initial condition $x^0 = \zeta$. A necessary and sufficient condition for the boundedness of $f_\zeta(K)$ for all $\zeta \in \mathbb{R}^n$ is closed-loop stability,

$$K \in \mathcal{S}_K := \{K \in \mathbb{R}^{m\times n} \mid \rho(A - BK) < 1\} \tag{7.4}$$

where $\rho(\cdot)$ is the spectral radius.

For any $K \in \mathcal{S}_K$, the matrices $P(K)$ and $X_\zeta(K)$ are well-defined and are, respectively, determined by the unique solutions to the Lyapunov equations

$$\mathcal{A}_K^*(P) = -Q - K^T R K, \quad \mathcal{A}_K(X_\zeta) = -\zeta\zeta^T. \tag{7.5}$$

Here, $\mathcal{A}_K$, $\mathcal{A}_K^*$: $\mathbb{S}_n \to \mathbb{S}_n$

$$\mathcal{A}_K(X) = (A - BK)X(A - BK)^T - X \tag{7.6a}$$
$$\mathcal{A}_K^*(P) = (A - BK)^T P(A - BK) - P \tag{7.6b}$$

determine the adjoint pairs of invertible closed-loop Lyapunov operators acting on the set of symmetric matrices $\mathbb{S}_n \subset \mathbb{R}^{n\times n}$.

The invertibility of $\mathcal{A}_K$ and $\mathcal{A}_K^*$ for $K \in \mathcal{S}_K$ allows us to express the LQR objective function in (7.2) as

$$f(K) = \begin{cases} \left\langle Q + K^T R\, K, X(K)\right\rangle = \langle \Omega, P(K)\rangle\,, K \in \mathcal{S}_K \\ \infty, \qquad\qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$$

where

$$X(K) := \mathbb{E}\left[X_\zeta(K)\right] = -\mathcal{A}_K^{-1}(\Omega) \tag{7.7}$$

and $\Omega := \mathbb{E}[\zeta\zeta^T]$ is the covariance matrix of the initial condition. We assume $\Omega \succ 0$ to ensure that the random vector $\zeta \sim \mathcal{D}$ has energy in all directions. This condition guarantees $f(K) = \infty$ for all $K \notin \mathcal{S}_K$. Finally, it is well known that for any $K \in \mathcal{S}_K$, the cone of positive definite matrices is closed under the action of $-\mathcal{A}_K^{-1}$ and $-(\mathcal{A}_K^*)^{-1}$. Thus, from the positive definiteness of the matrices $Q + K^T R K$ and Ω, it follows that $P(K)$, $X(K) \succ 0$ for all $K \in \mathcal{S}_K$. In (7.2), K is the optimization variable, and (A, B, $Q \succ 0$, $R \succ 0$, $\Omega \succ 0$) are the problem parameters.

For any feedback gain $K \in \mathcal{S}_K$, it can be shown that [162]

$$\nabla f_\zeta(K) = E(K)X_\zeta(K),\ \nabla f(K) = E(K)X(K) \tag{7.8a}$$

where

$$E(K) := 2\left((R + B^T P(K)B)K - B^T P(K)A\right) \tag{7.8b}$$

is a fixed matrix that does not depend on the random initial condition ζ. Thus, the randomness of the gradient $\nabla f_\zeta(K)$ arises from the random matrix $X_\zeta(K)$.

Remark 1 *The LQR problem for continuous-time systems can be treated in a similar way. In this case, although the Lyapunov operator $\mathcal{A}_K$ has a different definition, the form of the objective function in terms of the matrices $X(K)$ and $P(K)$ and also the form of the gradient in terms of $X(K)$ and $E(K)$ remain unchanged. While this similarity allows for our results to hold for both continuous and discrete-time systems, in this chapter we only focus on the latter and refer to [14] for a treatment of continuous-time systems.*

7.3 Random search

The formulation of the LQR problem given by (7.2) has been studied for both continuous-time [83], [88] and discrete-time systems [13], [141]. In this chapter, we analyze the sample complexity and convergence properties of the random search method for solving problem (7.2) with unknown model parameters. At each iteration $k \in \mathbb{N}$, the random search method calls Algorithm 2 that forms an empirical approximation $\overline{\nabla} f(K^k)$ to the gradient of the objective function via finite-time simulation of system (7.1a) for randomly perturbed feedback gains $K^k \pm U_i$, $i = 1, \ldots, N$.

Algorithm 2 does not require knowledge of matrices A and B but only access to a *two-point* simulation engine. The two-point setting means that for any pair of points K and K', the simulation engine can return the random values $f_{\zeta,\tau}(K)$ and $f_{\zeta,\tau}(K')$ for some random initial condition $x^0 = \zeta$, where

$$f_{\zeta,\tau}(K) \;:=\; \sum_{t=0}^{\tau} (x^t)^T Q\, x^t \;+\; (u^t)^T R\, u^t \tag{7.9}$$

is a finite-time random function approximation associated with system (7.1a), starting from a random initial condition $x^0 = \zeta$, with the state feedback $u = -Kx$ running up to time τ. This is in contrast to the *one-point* setting in which, at each query, the simulation engine can receive only one specified point K and return the random value $f_{\zeta,\tau}(K)$.

Starting from an initial feedback gain $K^0 \in \mathcal{S}_K$, the random search method uses the gradient estimates obtained via Algorithm 2 to update the iterates according to

$$K^{k+1} \;:=\; K^k \;-\; \alpha\, \overline{\nabla} f(K^k), \quad K^0 \in \mathcal{S}_K \tag{RS}$$

for some stepsize $\alpha > 0$. The stabilizing assumption on the initial iterate $K^0 \in \mathcal{S}_K$ is required in our analysis as we select the input parameters of Algorithm 2 and the stepsize so that all iterates satisfy $K^k \in \mathcal{S}_K$.

For convex problems, the gradient estimates obtained in the two-point setting are known to yield faster convergence rates than the one-point setting [163]. However, the two-point setting requires simulations of the system for two different feedback gain matrices under the same initial condition.

Algorithm 2 Gradient estimation

Require: Feedback gain $K \in \mathbb{R}^{m\times n}$, state and control weight matrices Q and R, distribution $\mathcal{D}$, smoothing constant r, simulation time τ, number of random samples N.

for $i = 1$ to N **do**

– Define two perturbed feedback gains $K_{i,1} := K + rU_i$ and $K_{i,2} := K - rU_i$, where $\mathrm{vec}(U_i)$ is a random vector uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$.

– Sample an initial condition ζ^i from distribution $\mathcal{D}$.

– For $j \in \{1, 2\}$, simulate system (7.1a) up to time τ with the feedback gain $K_{i,j}$ and initial condition ζ_i to form $f_{\zeta^i,\tau}(K_{i,j})$ as in Eq. (7.9).

end for

Ensure: The two-point gradient estimate

$$\overline{\nabla} f(K) := \frac{1}{2rN} \sum_{i=1}^{N} \left(f_{\zeta^i,\tau}(K_{i,1}) \; - \; f_{\zeta^i,\tau}(K_{i,2}) \right) U_i.$$

7.4 Main result

We analyze the sample complexity and convergence of the random search method (RS) for the model-free setting. Our main convergence result exploits two key properties of the LQR objective function f, namely smoothness and the Polyak-Łojasiewicz (PL) condition over its sublevel sets $\mathcal{S}_K(a) := \{K \in \mathcal{S}_K \mid f(K) \leq a\}$ where a is a positive scalar. In particular, it can be shown that, restricted to any sublevel set $\mathcal{S}_K(a)$, the function f is $L_f(a)$-smooth and satisfies the PL condition with parameter $\mu_f(a)$, i.e.,

$$f(K^{'}) \; - \; f(K) \; \leq \; \langle \nabla f(K), K' - K \rangle \; + \; \frac{L_f(a)}{2} \, \|K - K'\|_F^2$$

$$f(K) \; - \; f(K^\star) \leq \frac{1}{2\mu_f(a)} \, \|\nabla f(K)\|_F^2$$

for all K and K' such that the line segment between them belongs to $\mathcal{S}_K(a)$, where $L_f(a)$ and $\mu_f(a)$ are positive rational functions of a. This result has been established for both continuous-time [88] and discrete-time [13], [141] LQR problems. We also make the following assumption on the statistical properties of the initial condition.

Assumption 1 (Initial distribution) *Let the distribution $\mathcal{D}$ of the initial condition have i.i.d. zero-mean unit-variance entries with bounded sub-Gaussian norm. For a random vector $\zeta \in \mathbb{R}^n$ distributed according to $\mathcal{D}$, this implies $\mathbb{E}[\zeta] = 0$, $\mathbb{E}[\zeta\zeta^T] = I$, and $\|\zeta_i\|_{\psi_2} \leq \kappa$, for some constant κ and $i = 1, \ldots, n$, where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm [160].*

We now state our main theoretical result.

Theorem 1 *Consider the random search method* (RS) *that uses the gradient estimates of Algorithm 2 for finding the optimal solution $K^\star$ of problem* (7.2)*. Let the initial condition*

$x^0 \sim \mathcal{D}$ obey Assumption 1 and let the simulation time τ and the number of samples N in Algorithm 2 satisfy

$$\tau \;\geq\; \theta'(a)\log(1/\epsilon), \quad N \;\geq\; c\left(1+\beta^4\kappa^4\,\theta(a)\log^6 n\right)n,$$

for some $\beta > 0$ and a desired accuracy $\epsilon > 0$. Then, we can choose a smoothing parameter $r < \theta''(a)\sqrt{\epsilon}$ in Algorithm 2 such that, for any initial condition $K^0 \in \mathcal{S}_K(a)$, method (RS) *with the constant stepsize $\alpha = 1/(\omega(a)L_f(a))$ achieves $f(K^k) - f(K^\star) \leq \epsilon$ in at most*

$$k \;\leq\; -\log\left(\epsilon^{-1}\left(f(K^0) - f(K^\star)\right)\right)/\log\left(1-\mu_f(a)\alpha/8\right)$$

iterations. This holds with probability not smaller than

$$1 - c'k(n^{-\beta} + N^{-\beta} + N\mathrm{e}^{-\frac{n}{8}} + \mathrm{e}^{-c'N}).$$

Here, $\omega(a) := c''(\sqrt{m} + \beta\kappa^2\theta(a)\sqrt{mn}\log n)^2$, the positive scalars c, c', and c'' are absolute constants, $\mu_f(a)$ and $L_f(a)$ are the PL and smoothness parameters of f over the sublevel set $\mathcal{S}_K(a)$, and θ, θ', and θ'' are positive polynomials that depend only on the parameters of the LQR problem.

For a desired accuracy level $\epsilon > 0$, Theorem 1 shows that the random search iterates (RS) with constant stepsize (that does not depend on ϵ) reach an accuracy level ϵ at a linear rate (i.e., in at most $O(\log(1/\epsilon))$ iterations) with high probability. Furthermore, the total number of function evaluations and the simulation time required to achieve an accuracy level ϵ are proportional to $\log(1/\epsilon)$. As stated earlier, this significantly improves the existing results for discrete-time LQR [13], [132] that require $O(1/\epsilon)$ function evaluations and poly$(1/\epsilon)$ simulation time.

7.5 Proof sketch

In this section, we present a sketch of our proof strategy for the main result of the chapter. The smoothness of the objective function along with the PL condition are sufficient for the gradient descent method with a suitable stepsize α,

$$K^{k+1} \;:=\; K^k \;-\; \alpha\,\nabla f(K^k), \quad K^0 \;\in\; \mathcal{S}_K \tag{GD}$$

to achieve linear convergence even for nonconvex problems [89]. These properties were recently used to show convergence of gradient descent for both discrete-time [13] as well as continuous-time [88] LQR problems. In the model-free setting, the gradient descent method is not directly implementable because computing the gradient $\nabla f(K)$ requires knowledge of system parameters A and B. The random search method (RS) resolves this issue by using the *gradient estimate* $\overline{\nabla} f(K)$ obtained via Algorithm 2. One approach to the convergence analysis of random search is to first use a large number of samples N in order to make the estimation error small, and then relate the iterates of (RS) to that of gradient descent. It has been shown that achieving $\|\overline{\nabla} f(K) - \nabla f(K)\|_F \leq \epsilon$ takes $N = O(1/\epsilon^4)$ samples [13]; see also [15, Theorem 3] for the continuous-time LQR. This upper bound unfortunately leads to a sample complexity bound that grows polynomially with $1/\epsilon$. To improve this result, we

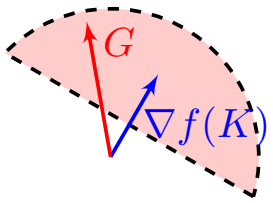


Figure 7.1: The intersection of the half-space and the ball parameterized by μ_1 and μ_2, respectively, in Proposition 1. If an update direction G lies within this region, then taking one step along $-G$ with a constant stepsize α yields a geometric decrease in the objective value.

take an alternative route and give up on the objective of controlling the gradient estimation error. In particular, by exploiting the problem structure, we show that with a fixed number of samples $N = \tilde{O}(n)$, where n denotes the number of states, the estimate $\overline{\nabla} f(K)$ concentrates with *high probability* when projected to the direction of $\nabla f(K)$.

In what follows, we first establish that for any $\epsilon > 0$, using a simulation time $\tau = O(\log(1/\epsilon))$ and an appropriate smoothing parameter r in Algorithm 2, the estimate $\overline{\nabla} f(K)$ can be made ϵ-close to an unbiased estimate $\widehat{\nabla} f(K)$ of the gradient with high probability, $\|\overline{\nabla} f(K) - \widehat{\nabla} f(K)\|_F \leq \epsilon$, where the definition of $\widehat{\nabla} f(K)$ is given in Eq. (7.12). We call this distance the *estimation bias*. We then show that, for a large number of samples N, our unbiased estimate $\widehat{\nabla} f(K)$ becomes highly correlated with the gradient. In particular, we establish that the following two events

$$\mathsf{M}_1 := \left\{ \left\langle \widehat{\nabla} f(K), \nabla f(K) \right\rangle \geq \mu_1 \|\nabla f(K)\|_F^2 \right\} \tag{7.10a}$$

$$\mathsf{M}_2 := \left\{ \|\widehat{\nabla} f(K)\|_F^2 \leq \mu_2 \|\nabla f(K)\|_F^2 \right\} \tag{7.10b}$$

occur with high probability for some positive scalars μ_1 and μ_2. To justify the definition of these events, let us first demonstrate that the gradient estimate $\widehat{\nabla} f(K)$ can be used to decrease the objective error by a geometric factor if both M_1 and M_2 occur.

Proposition 1 *If $G \in \mathbb{R}^{m \times n}$ and $K \in \mathcal{S}_K(a)$ are such that $\langle G, \nabla f(K) \rangle \geq \mu_1 \|\nabla f(K)\|_F^2$ and $\|G\|_F^2 \leq \mu_2 \|\nabla f(K)\|_F^2$ for some scalars $\mu_1, \mu_2 > 0$, then $K - \alpha G \in \mathcal{S}_K(a)$ for all $\alpha \in [0, \mu_1/(\mu_2 L_f(a))]$, and $f(K - \alpha G) - f(K^\star) \leq \big(1 - \mu_f(a)\mu_1 \alpha\big)\big(f(K) - f(K^\star)\big)$, where $L_f(a)$ and $\mu_f(a)$ are the smoothness and PL parameters of f over $\mathcal{S}_K(a)$.*

Proposition 1 demonstrates that, conditioned on the events M_1 and M_2, the unbiased estimate $\widehat{\nabla} f(K)$ yields a simple descent-based algorithm that has linear convergence. Fig. 7.1 illustrates the region parameterized by μ_1 and μ_2 in Proposition 1. This region has a different geometry than ϵ-neighborhoods of the gradient. A gradient estimate G can have an accuracy of $O(\nabla f(K))$ and still belong to this region. We leverage this fact in our convergence analysis which only requires the gradient estimate $\widehat{\nabla} f(K)$ to be in such a region for certain parameters μ_1 and μ_2 and not necessarily within an ϵ-neighborhood of the gradient.

7.5.1 Controlling the bias

Herein, we define the unbiased estimate $\widehat{\nabla} f(K)$ of the gradient and establish an upper bound on its distance to the output $\overline{\nabla} f(K)$ of Algorithm 2

$$\begin{aligned}
\overline{\nabla} f(K) &:= \frac{1}{2rN} \sum_{i=1}^{N} \left(f_{\zeta^i,\tau}(K + rU_i) - f_{\zeta^i,\tau}(K - rU_i) \right) U_i \\
\widetilde{\nabla} f(K) &:= \frac{1}{2rN} \sum_{i=1}^{N} \left(f_{\zeta^i}(K + rU_i) - f_{\zeta^i}(K - rU_i) \right) U_i \\
\widehat{\nabla} f(K) &:= \frac{1}{N} \sum_{i=1}^{N} \left\langle \nabla f_{\zeta^i}(K), U_i \right\rangle U_i
\end{aligned} \tag{7.12}$$

Here, $U_i \in \mathbb{R}^{m \times n}$ are i.i.d. random matrices whose vectorized form $\mathrm{vec}(U_i)$ are uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$ and $\zeta^i \in \mathbb{R}^n$ are i.i.d. random initial conditions sampled from distribution $\mathcal{D}$. Note that $\widetilde{\nabla} f(K)$ is the infinite horizon version of $\overline{\nabla} f(K)$ and $\widehat{\nabla} f(K)$ is an unbiased estimate of $\nabla f(K)$. The fact that $\mathbb{E}[\widehat{\nabla} f(K)] = \nabla f(K)$ follows from

$$\begin{aligned}
\mathbb{E}_{\zeta^i, U_i} \Big[\mathrm{vec}(\widehat{\nabla} f(K)) \Big] &= \mathbb{E}_{U_1} [\langle \nabla f(K), U_1 \rangle \, \mathrm{vec}(U_1)] \\
&= \mathbb{E}_{U_1} [\mathrm{vec}(U_1) \mathrm{vec}(U_1)^T] \mathrm{vec}(\nabla f(K)) \;=\; \mathrm{vec}(\nabla f(K)).
\end{aligned}$$

Local boundedness of the function $f(K)$: An important requirement for the gradient estimation scheme in Algorithm 2 is the stability of the perturbed closed-loop systems, i.e., $K \pm rU_i \in \mathcal{S}_K$; violating this condition leads to an exponential growth of the state and control signals. Moreover, this condition is necessary and sufficient for $\widetilde{\nabla} f(K)$ to be well defined. It can be shown that for any sublevel set $\mathcal{S}_K(a)$, there exists a positive radius r such that $K + rU \in \mathcal{S}_K$ for all $K \in \mathcal{S}_K(a)$ and $U \in \mathbb{R}^{m \times n}$ with $\|U\|_F \leq \sqrt{mn}$. In this chapter, we further require that r is small enough so that $K \pm rU_i \in \mathcal{S}_K(2a)$ for all $K \in \mathcal{S}_K(a)$. Such upper bound on r can be provided using the upper bound on the cost difference established in [13, Lemma 24]. A similar result has been established for the continuous-time LQR problem using the small-gain theorem and the KYP lemma [14].

Lemma 1 *For any $K \in \mathcal{S}_K(a)$ and $U \in \mathbb{R}^{m \times n}$ with $\|U\|_F \leq \sqrt{mn}$, $K + r(a)U \in \mathcal{S}_K(2a)$, where $r(a) := \tilde{c}/a$ for some constant $\tilde{c} > 0$ that depends on the problem data.*

Note that for any $K \in \mathcal{S}_K(a)$ and $r \leq r(a)$ in Lemma 1, $\widetilde{\nabla} f(K)$ is well defined since the feedback gains $K \pm rU_i$ are all stabilizing. We next establish an upper bound on the difference between the output $\overline{\nabla} f(K)$ of Algorithm 2 and the unbiased estimate $\widehat{\nabla} f(K)$ of the gradient $\nabla f(K)$. We accomplish this by bounding the difference between these two quantities and $\widetilde{\nabla} f(K)$ using the triangle inequality

$$\|\widehat{\nabla} f(K) - \overline{\nabla} f(K)\|_F \;\leq\; \|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F + \|\widehat{\nabla} f(K) - \widetilde{\nabla} f(K)\|_F. \tag{7.13}$$

Proposition 2 provides an upper bound on each term on the right-hand side of the above inequality.

Proposition 2 *For any $K \in \mathcal{S}_K(a)$ and $r \leq r(a)$, where $r(a)$ is given by Lemma 1,*

$$\begin{aligned}\|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F &\leq \frac{\sqrt{mn}\eta}{r}\kappa_1(2a)\,(1-\kappa_2(2a))^\tau \\ \|\widehat{\nabla} f(K) - \widetilde{\nabla} f(K)\|_F &\leq \frac{(rmn)^2\eta}{2}\,\ell(2a)\end{aligned}$$

where $\eta := \max_i \|\zeta^i\|^2$, and $\ell(a) > 0$, $\kappa_1(a) > 0$, and $1 > \kappa_2(a) > 0$ are rational functions that depend on the problem data.

The first term on the right-hand side of (7.13) corresponds to a bias arising from the finite-time simulation. Proposition 2 shows that although small values of r may result in a large $\|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F$, because of the exponential dependence of the upper bound on the simulation time τ, this error can be controlled by increasing τ. In addition, since $\widehat{\nabla} f(K)$ is independent of the parameter r, this result provides a quadratic bound on the estimation error in terms of r. It is also worth mentioning that the third derivative of the function $f_\zeta(K)$ is utilized in obtaining the second inequality.

7.5.2 Correlation of $\widehat{\nabla} f(K)$ and $\nabla f(K)$

We establish that under Assumption 1 on the initial distribution, with large enough number of samples $N = \tilde{O}(n)$, the events M_1 and M_2 with $\mu_1 := 1/4$ and

$$\mu_2 := Cm\left(\beta\kappa^2\,\frac{\|(\mathcal{A}_K^*)^{-1}\|_2 + \|(\mathcal{A}_K^*)^{-1}\|_S}{\lambda_{\min}(X(K))}\sqrt{n}\log n + 1\right)^2 \tag{7.14}$$

occur with high probability, where κ is an upper bound on the ψ_2-norm of the entries of ζ^i, $\beta > 0$ is a parameter that determines the failure probability, C is a positive absolute constant, and for an operator $\mathcal{M}$,

$$\|\mathcal{M}\|_2 := \sup_M \frac{\|\mathcal{M}(M)\|_F}{\|M\|_F},\ \|\mathcal{M}\|_S := \sup_M \frac{\|\mathcal{M}(M)\|_2}{\|M\|_2}.$$

We note that these parameters do not depend on the desired accuracy-level ϵ. Moreover, since the sub-level sets of the function $f(K)$ are compact [141], $\|(\mathcal{A}_K^*)^{-1}\|$ is a continuous function of K, and $X(K) \succeq \Omega$, we can uniformly upper bound μ_2 over any sublevel set $\mathcal{S}_K(a)$. Such bound has also been discussed and analytically quantified for the continuous-time LQR problem [14].

Our approach to accomplishing the above task exploits the problem structure, which allows for confining the dependence of $\widehat{\nabla} f(K)$ on the random initial conditions ζ^i into the zero-mean random matrices $X_{\zeta^i} - X$, where $X_{\zeta^i} := X_{\zeta^i}(K)$ and $X := X(K)$ are given

by (7.3) and (7.7), respectively. In particular, for any given feedback gain $K \in \mathcal{S}_K$, we can use the form of gradient (7.8) to write

$$\widehat{\nabla} f(K) \;=\; \frac{1}{N}\sum_{i=1}^{N} \big\langle E\, X_{\zeta^i}, U_i \big\rangle\, U_i \;=\; \widehat{\nabla}_1 \;+\; \widehat{\nabla}_2$$

where $\widehat{\nabla}_1 := (1/N)\sum_{i=1}^{N} \big\langle E\,(X_{\zeta^i} - X), U_i\big\rangle\, U_i$, $\widehat{\nabla}_2 := (1/N)\sum_{i=1}^{N} \langle \nabla f(K), U_i\rangle\, U_i$, and the matrix $E := E(K)$ is given by (7.8b). It is now easy to verify that $\mathbb{E}[\widehat{\nabla}_1] = 0$ and $\mathbb{E}[\widehat{\nabla}_2] = \nabla f(K)$. Furthermore, only the term $\widehat{\nabla}_1$ depends on the initial conditions ζ^i.

7.5.2.1 Quantifying the probability of M_1

We exploit results from modern high-dimensional statistics on the non-asymptotic analysis of the concentration of random quantities around their mean [160]. Our approach to analyzing the event M_1 consists of two steps. First, we establish that the zero-mean random variable $\Big\langle \widehat{\nabla}_1, \nabla f(K) \Big\rangle$ *highly* concentrates around zero with a large enough number of samples $N = \tilde{O}(n)$. Our proof technique relies on the Hanson-Wright inequality [164, Theorem 1.1]. Next, we study the concentration of the random variable $\Big\langle \widehat{\nabla}_2, \nabla f(K) \Big\rangle$ around its mean $\|\nabla f(K)\|_F^2$. The key enabler here is the Bernstein inequality [160, Corollary 2.8.3]. This leads to the next proposition.

Proposition 3 *Under Assumption 1, for any stabilizing feedback gain $K \in \mathcal{S}_K$ and positive scalar β, if*

$$N \;\geq\; C_1\, \frac{\beta^4 \kappa^4}{\lambda_{\min}^2(X)} \left(\|(\mathcal{A}_K^*)^{-1}\|_2 + \|(\mathcal{A}_K^*)^{-1}\|_S\right)^2 n \log^6 n$$

then the event M_1 in (7.10) with $\mu_1 := 1/4$ satisfies

$$\mathbb{P}(\mathsf{M}_1) \;\geq\; 1 - C_2 N^{-\beta} - 4N\mathrm{e}^{-\frac{n}{8}} - 2\mathrm{e}^{-C_3 N}.$$

7.5.2.2 Quantifying the probability of M_2

Similarly, we analyze the event M_2 in two steps. We establish upper bounds on the ratio $\|\widehat{\nabla}_i\|_F / \|\nabla f(K)\|_F$, for $i = \{1, 2\}$, that hold with high probability, and use the triangle inequality

$$\frac{\|\widehat{\nabla}_1\|_F}{\|\nabla f(K)\|_F} \;+\; \frac{\|\widehat{\nabla}_i\|_F}{\|\nabla f(K)\|_F} \;\geq\; \frac{\|\widehat{\nabla} f(K)\|_F}{\|\nabla f(K)\|_F}.$$

Our results are summarized in the next proposition.

Proposition 4 *Under Assumption 1, for any $K \in \mathcal{S}_K$, scalar $\beta > 0$, and $N \geq C_4 n$, the event M_2 in (7.10) with μ_2 given by (7.14) satisfies $\mathbb{P}(\mathsf{M}_2) \geq 1 - C_6(n^{-\beta} + N\mathrm{e}^{-\frac{n}{8}} + \mathrm{e}^{-C_7 N})$.*

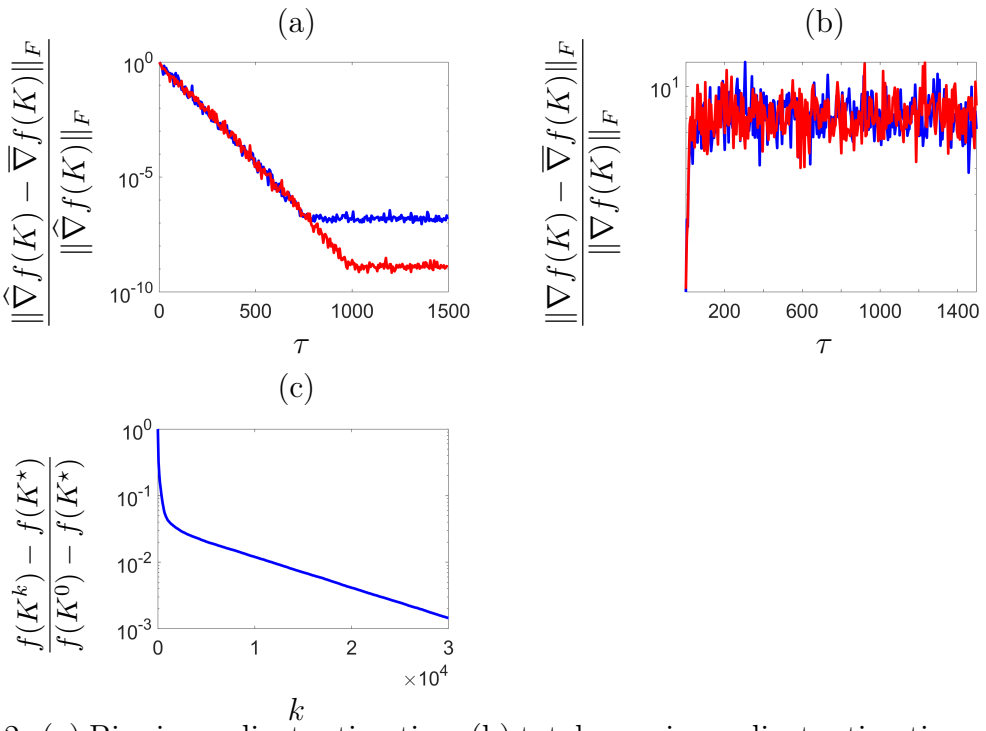


Figure 7.2: (a) Bias in gradient estimation; (b) total error in gradient estimation as functions of the simulation time τ. The blue and red curves correspond to two values of the smoothing parameter $r = 10^{-4}$ and $r = 10^{-6}$, respectively. (c) Convergence curve of the random search method (RS).

7.6 Computational experiments

We consider a system with $s = 10$ inverted pendula on force-controlled carts that are connected by springs and dampers; see Fig. 7.3. We set all masses, pendula lengths, spring and damping constants to unity and let the state vector $x := [\theta^T \omega^T p^T v^T]^T$ contain the angle and angular velocity of pendula as well as position and velocity of masses. Linearizing around the equilibrium point yields the continuous-time system $\dot{x} = A_c x + B_c u$, where

$$A_c = \begin{bmatrix} 0 & I & 0 & 0 \\ 20I & 0 & T & T \\ 0 & 0 & 0 & I \\ -10I & 0 & -T & -T \end{bmatrix}, \quad B_c = \begin{bmatrix} 0 \\ -I \\ 0 \\ I \end{bmatrix}.$$

Here, 0 and I are $s \times s$ zero and identity matrices, and T is a Toeplitz matrix with 2 on the main diagonal, -1 on the first upper and lower sub-diagonals, and zero elsewhere. We discretize this system with sampling time $t_s = 0.1$, which yields Eq. (7.1a) with $A = \mathrm{e}^{A_c t_s}$ and $B = \int_0^{t_s} \mathrm{e}^{A_c t} B_c \mathrm{d}t$. Since the open-loop system is unstable, we use a stabilizing feedback gain $K^0 = [-50I \ -10I \ -5I \ -5I]$ as a starting point for the random search method and choose $Q = \mathrm{blkdiag}(10I, I, I, I)$ and $R = I$ in the LQR cost. We also let the initial conditions ζ^i in Algorithm 2 be standard normal and use $N = n = 2s$ samples.

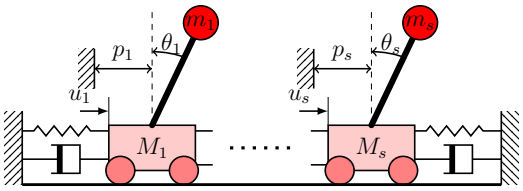


Figure 7.3: An interconnected system of inverted pendula on carts.

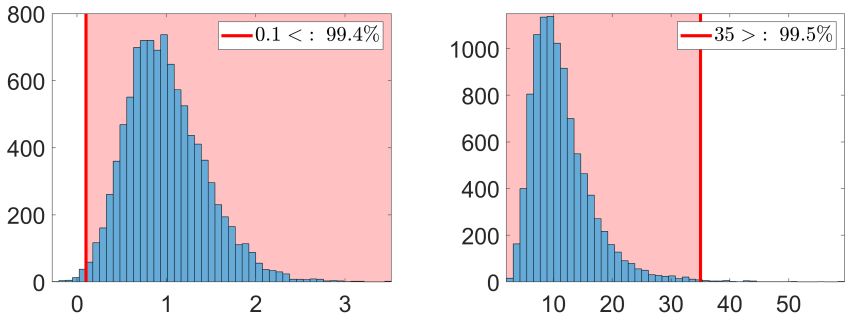


Figure 7.4: Histograms of two algorithmic quantities associated with the events M_1 and M_2 given by (7.10). The red lines demonstrate that M_1 with $\mu_1 = 0.1$ and M_2 with $\mu_2 = 35$ occur in more than 99% of trials.

Figure 7.2 (a) illustrates the dependence of the relative error

$$\|\widehat{\nabla} f(K) - \overline{\nabla} f(K)\|_F / \|\widehat{\nabla} f(K)\|_F$$

on the simulation time τ for $K = K^0 = [-50I \ -10I \ -5I \ -5I]$ and two values of smoothing parameter $r = 10^{-4}$ (blue) and $r = 10^{-6}$ (red). We see an exponential decrease in error for small values of τ and note that the error does not pass a saturation level determined by the smoothing parameter $r > 0$. We also observe that as r decreases, this saturation level becomes smaller. These observations are in harmony with the results established in Proposition 2. This should be compared and contrasted with Fig. 7.2 (b), which demonstrates that the relative error with respect to the true gradient does not vanish with increase in the simulation time τ.

In spite of this significant error, the key observation that allows us to establish the linear convergence of random search method in Theorem 1 is that the gradient estimate has high correlation with the true gradient. Figure 7.4 shows histograms of two algorithmic quantities associated with the events M_1 and M_2 given by (7.10). The red lines demonstrate that M_1 with $\mu_1 = 0.1$ and M_2 with $\mu_2 = 35$ occur in more than 99% of trials; cf. Propositions 3 and 4.

Figure 7.2 (c) illustrates the convergence curve of the random search method (RS) with stepsize $\alpha = 10^{-5}$, $r = 10^{-5}$, and $\tau = 1000$ in Algorithm 2. This figure confirms linear convergence of (RS) established in Theorem 1.

7.7 Concluding remarks

In this chapter, we studied the convergence and sample complexity of the random search method with two-point gradient estimates for the discrete-time LQR problem. Despite non-convexity, we established that the random search method with a fixed number of roll-outs $N = \tilde{O}(n)$ per iteration achieves ϵ-accuracy in $O(\log(1/\epsilon))$ iterations. This significantly improves existing results on the model-free LQR which require $O(1/\epsilon)$ total roll-outs. Our ongoing research directions include: (i) providing theoretical guarantees for the convergence of gradient-based methods for sparsity-promoting and structured control synthesis [71]; and (ii) extension to nonlinear systems via successive linearization techniques.

Chapter 8

Lack of gradient domination for linear quadratic Gaussian problems with incomplete state information

Policy gradient algorithms in model-free reinforcement learning have been shown to achieve global exponential convergence for the Linear Quadratic Regulator problem despite the lack of convexity. However, extending such guarantees beyond the scope of standard LQR and full-state feedback has remained open. A key enabler for existing results on LQR is the so-called gradient dominance property of the underlying optimization problem that can be used as a surrogate for strong convexity. In this chapter, we take a step further by studying the convergence of gradient descent for the Linear Quadratic Gaussian problem and demonstrate through examples that LQG does not satisfy the gradient dominance property. Our study shows the non-uniqueness of equilibrium points and thus disproves the global convergence of policy gradient methods for LQG.

8.1 Introduction

Modern reinforcement learning algorithms have shown great empirical performance in solving continuous control problems [18] with unknown dynamics. However, despite the recent surge in research, convergence and sample complexity of these methods are not yet fully understood. This has recently motivated a significant body of literature on data-driven control to focus on the Linear Quadratic Regulator (LQR) problem with unknown model parameters with the primary purpose of providing insight into the behavior and performance of RL algorithms in more challenging settings.

The LQR problem is the cornerstone of control theory. The globally optimal solution to LQR is given by a static linear feedback and, for problems with known models, the solution can be obtained by solving the celebrated Riccati equation using efficient numerical schemes with provable convergence guarantees [83]. In the data-driven setting, existing techniques are mainly divided into two categories, model-based [130] and model-free [21]. While model-based techniques use data to obtain approximations of the underlying dynamics, model-free methods directly search over the parameter space of controllers using the reward/cost values without attempting to form a model.

Among model-free approaches, simple random search, which emulates the behavior of gradient descent by forming estimates of the gradient via cost evaluations, has been shown to achieve sub-linear sample complexity for LQR [132]. This can be even further improved to

a logarithmic complexity if one can access the so-called *two-point* gradient estimates [14], [87]. These results build on the fact that the gradient descent itself achieves linear convergence for both discrete [13] and continuous-time LQR problems [88] despite lack of convexity. A key enabler for these results is the so-called gradient dominance property of the underlying optimization problem that can be used as a surrogate for strong convexity [89].

In this chapter, we take a step further by studying the convergence of gradient descent for the Linear Quadratic Gaussian (LQG) problem with incomplete state information. The separation principle states that the solution to the LQG problem is given by an observer-based controller, which consists of a Kalman filter and the corresponding LQR solution. This problem is also closely related to the output-feedback problem for distributed control, which is known to be fundamentally more challenging than LQR. In particular, the output-feedback problem has been shown to involve an optimization domain with exponential number of connected components [84], [90]. In contrast, the standard LQG problem allows for dynamic controllers and do not impose structural constraints on the controller.

Motivated by the convergence properties of gradient descent on LQR, we reformulate the LQG problem as a joint optimization of the control and observer feedback gains whose domain, unlike the output feedback problem is connected. We derive analytical expressions for the gradient of the LQG cost function with respect to gain matrices and demonstrate through examples that LQG does not satisfy the gradient dominance property. In particular, we show that, in addition to the global solution, the gradient vanishes at the origin for open-loop stable systems. Our study disproves global exponential convergence of policy gradient methods for LQG. The analysis of the optimization landscape of the LQG problem with unknown system parameters has also been recently provided in [165], where the authors relate the existence of multiple equilibrium points to the non-minimality of the controller transfer function.

The rest of the chapter is structured as follows. In Section 8.2, we formulate the LQG problem and provide background information. In Section 8.3, we derive an analytical expression for the gradient. In Section 8.4, we discuss the lack of gradient domination and non-uniqueness of equilibrium points. We present numerical experiments in Section 8.5 and finally provide concluding remarks in Section 8.6.

8.2 Linear Quadratic Gaussian

Consider the stochastic LTI system

$$\dot{x} = Ax + Bu + w, \quad y = Cx + v \tag{8.1a}$$

where $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in \mathbb{R}^m$ is the control input, $y(t) \in \mathbb{R}^p$ is the measured output, A, B, and C are constant matrices, and $w(t)$ and $v(t)$ are independent zero-mean Gaussian white noise processes with covariance functions $\mathbb{E}[w(t)w^T(\tau)] = \delta(t-\tau)\Sigma_w$ and $\mathbb{E}[v(t)v^T(\tau)] = \delta(t-\tau)\Sigma_v$. Here, δ is the Dirac delta (impulse) function and we assume

$\Sigma_w, \Sigma_v \succ 0$ are positive definite matrices. The Linear Quadratic Gaussian (LQG) problem associated with system (8.1a) is given by

$$\underset{u(t)\in\mathcal{Y}(t)}{\text{minimize}} \lim_{t\to\infty} \mathbb{E}\left[x^T(t)Q\,x(t) + u^T(t)R\,u(t)\right] \tag{8.1b}$$

where Q and R are positive definite matrices and $\mathcal{Y}(t)$ is the set of functions that depend only on the available information up to time t, i.e., the measured outputs $y(s)$ with $s \leq t$.

8.2.1 Separation principle

It is well-known that if the pair (A, B) is controllable and (A, C) is observable, the solution to (8.1) is given by an observer-based controller of the form

$$\begin{aligned} \dot{\hat{x}} &= A\,\hat{x} + B\,u - L(\hat{y} - y) \\ \hat{y} &= C\,\hat{x}, \quad u = -K\hat{x} \end{aligned} \tag{8.2}$$

where $\hat{x}(t) \in \mathbb{R}^n$ is the state estimate, and $L \in \mathbb{R}^{n\times p}$ and $K \in \mathbb{R}^{m\times n}$ are the observer and controller feedback gain matrices, respectively [83], [166]. The separation principle states that the optimal gains $K^\star$ and $L^\star$ correspond to solutions to two decoupled problems associated with (8.1), namely the linear quadratic regulator

$$\underset{K}{\text{minimize}} \lim_{t\to\infty} \mathbb{E}\left[x^T(t)Qx(t) + u^T(t)Ru(t)\right] \tag{8.3}$$

subject to (8.1a) with the full-state feedback $u = -Kx$, and the Kalman filter, which seeks to

$$\underset{L}{\text{minimize}} \lim_{t\to\infty} \mathbb{E}\left[\|e(t)\|^2\right] \tag{8.4a}$$

subject to the error dynamics

$$\dot{e} = (A - LC)\,e - L\,v + w \tag{8.4b}$$

where $e := x - \hat{x}$ is the state estimation error. The solutions to these two problems (and also to the original LQG problem) are given by

$$K^\star = R^{-1}B^T P_c^\star, \quad L^{\star T} = \Sigma_v^{-1} C X_o^\star \tag{8.5}$$

where $P_c^\star$ and $X_o^\star$ are the unique solutions to the decoupled pair of Algebraic Riccati Equations (ARE)

$$\begin{aligned} A^T P_c^\star + P_c^\star A + Q - P_c^\star B R^{-1} B^T P_c^\star &= 0 \\ A X_o^\star + X_o^\star A^T + \Sigma_w - X_o^\star C^T \Sigma_v^{-1} C X_o^\star &= 0. \end{aligned}$$

8.2.2 Characterization based on gain matrices

In this chapter, we analyze the LQG problem as optimization of feedback gain matrices K and L. In particular, the closed-loop dynamics in (8.1a) and (8.2) can be jointly described by

$$\dot{\xi} \;=\; A_{\mathcal{L}}\,\xi \,+\, \mu \tag{8.6}$$

where $\xi := \begin{bmatrix} x^T & e^T \end{bmatrix}^T \in \mathbb{R}^{2n}$ consists of the state and error signals,

$$\mu \;:=\; \begin{bmatrix} w^T & w^T - v^T L^T \end{bmatrix}^T$$

is white noise, and the closed-loop matrix $A_{\mathcal{L}}$ is given by

$$A_{\mathcal{L}} \;:=\; \begin{bmatrix} A - BK & BK \\ 0 & A - LC \end{bmatrix}. \tag{8.7}$$

The closed-loop representation given by (8.6) allows us to reformulate the LQG problem as an optimization over the set $\mathcal{S}_c \times \mathcal{S}_o$ of stabilizing gain matrices, where

$$\begin{aligned} \mathcal{S}_c &:= \{K \in \mathbb{R}^{m\times n} \,|\, A \,-\, BK \text{ is Hurwitz}\} \\ \mathcal{S}_o &:= \{L \in \mathbb{R}^{n\times p} \,|\, A \,-\, LC \text{ is Hurwitz}\}. \end{aligned} \tag{8.8}$$

In particular, the LQG problem in (8.1b) amounts to

$$\underset{K,L}{\text{minimize}} \;\; f(K,L) \;:=\; \langle \Omega, X \rangle \tag{8.9}$$

where $X = \lim\limits_{t\to\infty} \mathbb{E}\left[\xi(t)\xi^T(t)\right]$ is the steady-state covariance matrix associated with closed-loop system (8.6) and it can be determined by solving the algebraic Lyapunov equation

$$A_{\mathcal{L}}\,X \;+\; X A_{\mathcal{L}}^T \;+\; \Sigma \;=\; 0. \tag{8.10}$$

Here, the positive semi-definite matrices $\Omega,\ \Sigma$ are given by

$$\Omega \;:=\; \begin{bmatrix} Q + K^T R K & -K^T R K \\ -K^T R K & K^T R K \end{bmatrix} \tag{8.11a}$$

$$\Sigma \;:=\; \begin{bmatrix} \Sigma_w & \Sigma_w \\ \Sigma_w & \Sigma_w + L\Sigma_v L^T \end{bmatrix}. \tag{8.11b}$$

The matrix Ω accounts for the weight matrices in the cost function (8.1b) and the matrix Σ determines the covariance function $\Sigma\delta(t-\tau)$ of μ.

8.3 Gradient method

In this section, we introduce the gradient method on the LQG objective function over the set of stabilizing gain matrices $\mathcal{S}_c \times \mathcal{S}_o$ and discuss its convergence properties.

Lemma 1 *For any stabilizing pair of gain matrices $(K, L) \in \mathcal{S}_c \times \mathcal{S}_o$, the gradient of the LQG objective function f in (8.9) is given by*

$$\begin{aligned}\nabla_K f(K,L) &= 2(RK - B^T\hat{P}_1)\hat{X}_1 - 2B^T\hat{P}_2\hat{X}_2^T \\ \nabla_L f(K,L) &= 2P_3(L\Sigma_v - X_3C^T) - 2P_2^T X_2 C^T\end{aligned}$$

where the matrices

$$\begin{aligned}X = \begin{bmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{bmatrix}, \quad \hat{X} = \begin{bmatrix} \hat{X}_1 & \hat{X}_2 \\ \hat{X}_2^T & \hat{X}_3 \end{bmatrix} \\ P = \begin{bmatrix} P_1 & P_2 \\ P_2^T & P_3 \end{bmatrix}, \quad \hat{P} = \begin{bmatrix} \hat{P}_1 & \hat{P}_2 \\ \hat{P}_2^T & \hat{P}_3 \end{bmatrix}\end{aligned} \tag{8.12}$$

are the unique solutions to the Lyapunov equations

$$A_{\mathcal{L}} X + X A_{\mathcal{L}}^T + \Sigma = 0 \tag{8.13a}$$

$$\hat{A}_{\mathcal{L}} \hat{X} + \hat{X} \hat{A}_{\mathcal{L}}^T + \hat{\Sigma} = 0 \tag{8.13b}$$

$$A_{\mathcal{L}}^T P + P A_{\mathcal{L}} + \Omega = 0 \tag{8.13c}$$

$$\hat{A}_{\mathcal{L}}^T \hat{P} + \hat{P} \hat{A}_{\mathcal{L}} + \hat{\Omega} = 0. \tag{8.13d}$$

Here, the matrices $A_{\mathcal{L}}$, and Ω and Σ are given by (8.7) and (8.11), respectively, and

$$\hat{A}_{\mathcal{L}} := \begin{bmatrix} A - BK & LC \\ 0 & A - LC \end{bmatrix} \tag{8.14a}$$

$$\hat{\Omega} := \begin{bmatrix} Q + K^T R K & Q \\ Q & Q \end{bmatrix} \tag{8.14b}$$

$$\hat{\Sigma} := \begin{bmatrix} L\Sigma_v L^T & -L\Sigma_v L^T \\ -L\Sigma_v L^T & \Sigma_w + L\Sigma_v L^T \end{bmatrix}. \tag{8.14c}$$

Proof: To obtain $\nabla_L f(K, L)$, we use the Taylor series expansion of $f(K, L + \tilde{L})$ around (K, L) and collect first-order terms. From (8.9), we have

$$\begin{aligned}f(K, L+\tilde{L}) - f(K,L) &\approx \left\langle \nabla_L f(K,L), \tilde{L} \right\rangle \\ &= \left\langle \Omega, \tilde{X} \right\rangle\end{aligned} \tag{8.15a}$$

where $\tilde{X}$ is the unique solution to

$$A_{\mathcal{L}}\tilde{X} + \tilde{X}A_{\mathcal{L}}^T = -\tilde{A}_{\mathcal{L}}X - X\tilde{A}_{\mathcal{L}}^T - \tilde{\Sigma}$$
$$= \begin{bmatrix} 0 & X_2C^T\tilde{L}^T \\ \tilde{L}CX_2^T & \tilde{L}CX_3 + X_3C^T\tilde{L}^T \end{bmatrix} - \tilde{\Sigma} =: \Phi \quad (8.15b)$$

Here, the first equality is obtained by differentiating Lyapunov equation (8.10), and the second follows by noting that

$$\tilde{A}_{\mathcal{L}} = \begin{bmatrix} 0 & 0 \\ 0 & -\tilde{L}C \end{bmatrix}, \quad \tilde{\Sigma} = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{L}\Sigma_v L^T + L\Sigma_v\tilde{L}^T \end{bmatrix}.$$

Using the adjoint identity and (8.15), we obtain that

$$\left\langle \nabla_L f(K,L), \tilde{L} \right\rangle = \langle -\Phi, P \rangle$$

where P is given by (8.13c). Rearranging terms completes the proof for $\nabla_L f(K,L)$.

In order to obtain $\nabla_K f(K,L)$, we use a slightly different representation of the objective function. In particular, if we let $\hat{\xi} := \begin{bmatrix} \hat{x}^T & e^T \end{bmatrix}^T$, it is easy to verify that the closed-loop system satisfies

$$\dot{\hat{\xi}} = \hat{A}_{\mathcal{L}}\hat{\xi} + \hat{\mu}$$

where the closed-loop matrix $\hat{A}_{\mathcal{L}}$ is given by (8.14a) and $\hat{\mu} = \begin{bmatrix} v^TL^T & w^T - v^TL^T \end{bmatrix}^T$. Furthermore, it is straightforward to verify that for any stabilizing gain matrices $K \in \mathcal{S}_c$ and $L \in \mathcal{S}_o$, the LQG cost in (8.1b) is given by

$$f(K,L) := \left\langle \hat{\Omega}, \hat{X} \right\rangle \quad (8.16)$$

where $\hat{X} = \lim_{t\to\infty} \mathbb{E}\left[\hat{\xi}(t)\hat{\xi}^T(t)\right]$ is the unique solution to the algebraic Lyapunov equation given by (8.13b) and and the matrices $\hat{\Omega}$ and $\hat{\Sigma}$ are given by (8.14). Now, using this representation, the same technique as in the first part of the proof can be used to obtain $\nabla_L f(K,L)$. This completes the proof. □

Using the explicit formula of the gradient in Lemma 1, the gradient descent method over the set of stabilizing gain matrices $\mathcal{S}_c \times \mathcal{S}_o$ follows the update rule

$$\begin{aligned} K^{k+1} &:= K^k - \alpha\,\nabla_K f(K^k,L^k), \quad K^0 \in \mathcal{S}_c \\ L^{k+1} &:= L^k - \alpha\,\nabla_L f(K^k,L^k), \quad L^0 \in \mathcal{S}_o \end{aligned} \quad \text{(GD)}$$

where $\alpha > 0$ is the stepsize.

8.3.1 Non-separability of gradients

For the LQG problem, unlike the optimal solution that satisfies the separation principle, we observe from Lemma 1 that the gradient is not separable as $\nabla_K f$ and $\nabla_L f$ depend on both L and K. To provide more insight, let us examine the value of gradient over two special subsets of the domain $\mathcal{S}_c \times \mathcal{S}_o$, namely $\mathcal{S}_c \times \{L^\star\}$, where $L^\star$ is the optimal Kalman gain, and $\{K^\star\} \times \mathcal{S}_o$, where $K^\star$ is the optimal control feedback gain in (8.5).

8.3.1.1 Optimal observer gain $L = L^\star$

In this case, from (8.5) and the corresponding Riccati equation, it follows that

$$L\Sigma_v \;=\; X_o^\star C^T \tag{8.17}$$

where $X_o^\star$ is the unique positive definite solution to the Lyapunov equation

$$(A \,-\, LC)X_o^\star \,+\, X_o^\star(A \,-\, LC)^T \;=\; -\Sigma_w - L\Sigma_v L^T.$$

Expanding (8.13a) and (8.13b), we observe that X_3 and $\hat{X}_3$ also satisfy the above Lyapunov equation. Thus, since $A - LC$ is Hurwitz, it follows that

$$X_o^\star \;=\; X_3 \;=\; \hat{X}_3. \tag{8.18}$$

In addition, combining equations (8.13b), (8.17), and (8.18) yields

$$(A - BK)\hat{X}_2 \,+\, \hat{X}_2(A - LC)^T \;=\; 0. \tag{8.19}$$

Now, since $K \in \mathcal{S}_c$ and $L \in \mathcal{S}_o$, we obtain that $\hat{X}_2 = 0$. Form this equation in conjunction with (8.17) and (8.18), we obtain that the following terms in the gradient vanish

$$B^T\hat{P}_2\hat{X}_2^T \;=\; 0, \quad P_3(L\Sigma_v \,-\, X_3C^T) \;=\; 0 \tag{8.20a}$$

and thus the gradient simplifies to

$$\begin{aligned}\nabla_K f(K, L^\star) \;&= 2(RK - B^T\hat{P}_1)\hat{X}_1\\ \nabla_L f(K, L^\star) \;&=\; -\,2P_2^T X_2 C^T.\end{aligned}$$

Remark 1 *As we demonstrate in the proof of Lemma 1, for any stabilizing gains L and K, the matrix $\hat{X}_2$ is given by*

$$\hat{X}_2 \;=\; \lim_{t\to\infty} \mathbb{E}\left[e(t)\hat{x}^T(t)\right].$$

Thus, the equality $\hat{X}_2 = 0$ can be directly established using the orthogonality principle which states that the optimal estimator is orthogonal to the estimation error.

8.3.1.2 Optimal control gain $K = K^\star$

Similar to the previous case, from (8.5) and the corresponding Riccati equation, it follows that

$$RK \;=\; B^T P_c^\star$$

where $P_c^\star$ is the unique positive definite solution to the Lyapunov equation

$$(A \,-\, BK)P_c^\star \;+\; P_c^\star(A \,-\, BK)^T \;=\; -Q \,-\, K^T RK.$$

Combining this equations with (8.13c) and (8.13d) yields $\hat{P}_1 = P_c^\star$ and $P_2 = 0$. Thus, we have

$$(RK - B^T\hat{P}_1)\hat{X}_1 \;=\; 0, \quad P_2^T X_2 C^T \;=\; 0 \tag{8.20b}$$

which yields

$$\begin{aligned}\nabla_K f(K^\star, L) \;&=\; -\,2B^T\hat{P}_2\hat{X}_2^T \\ \nabla_L f(K^\star, L) \;&= 2P_3(L\Sigma_v \,-\, X_3C^T).\end{aligned}$$

We observe that $\nabla_K f(K^\star, L)$ and $\nabla_L f(K, L^\star)$ do not vanish and thus the sets $\mathcal{S}_c \times \{L^\star\}$ and $\{K^\star\} \times \mathcal{S}_o$ are not invariant with respect to gradient descent. Therefore, unlike the optimal solutions, the gradient of the LQG objective function may not be decoupled.

8.4 Lack of gradient domination

Recently, it has been shown that the gradient descent method achieves linear convergence for the LQR problem with full-state feedback in both discrete [13] and continuous-time [88] settings. These results build on the key observation that the full-state feedback LQR cost in (8.3) as a function of the feedback gains, denoted by $g(K)$, satisfies the Polyak-Łojasiewicz (PL) condition over its sub-levelsets, i.e.

$$\|\nabla g(K)\|_F^2 \;\geq\; \mu_g\,(g(K) \,-\, g(K^\star)) \tag{8.21}$$

for some constant $\mu_g > 0$. The PL condition, also known as gradient dominance, can be used as a surrogate to strong convexity to ensure convergence of gradient descent at a linear rate even for nonconvex problems. This observation raises the question of whether the LQG problem is also gradient dominant.

In addition, it has been recently shown that the set of stabilizing gains for the case of static output feedback, i.e. $u = -Ky,\ y = Cx$ consists of multiple connected components and local minima [90], which hinders the convergence of local search algorithms. However, in contrast to the static output feedback problem, the joint optimization of the controller and observer feedback gains for the LQG, as studied in this chapter, involves the connected domain $\mathcal{S}_c \times \mathcal{S}_o$.

We now demonstrate that despite the connectivity of the optimization domain, this formulation yet suffers from the existence of non-optimal equilibrium points and thus lack of gradient domination.

8.4.1 Non-uniqueness of critical points

The nonconvexity of the function f suggests the possibility of having multiple critical points $\nabla f(K, L) = 0$. In this section, we demonstrate that this is in fact the case by providing two of such points for the LQG problem in the general form. This should be compared and contrasted to the full-state feedback LQR problem which, despite nonconvexity, has been shown to have a unique critical point.

Global minimizer

The most obvious critical point is the unique global minimizer of f, which is given by (8.5). To verify this, note that for the optimal gains $L^\star$ and $K^\star$, we have equations (8.20a) and (8.20b), respectively. Using these equations, and the form of gradient in Lemma 1, it immediately follows that $\nabla f(K^\star, L^\star) = 0$.

The origin for stable systems

To find another critical point, let us assume for simplicity that the system is open-loop stable. We next show that the origin $(K, L) = (0, 0)$ is also a critical point, i.e., $\nabla f(0, 0) = 0$.

For $(K, L) = (0, 0)$, from (8.13b) it follows that $\hat{X}_1 = \hat{X}_2 = 0$. In addition, from (8.13c), it follows that $P_2 = P_3 = 0$. Combining these equalities and the form of gradient in Lemma 1 ensures $\nabla f(0, 0) = 0$.

The existence of the sub-optimal critical point $(K, L) = (0, 0)$ also implies that gradient domination may not hold for the LQG problem.

8.5 An example

We consider the mass-spring-damper system in Figure 8.1 with s masses to demonstrate the performance of gradient descent given by (GD) on the LQG problem over the set $\mathcal{S}_c \times \mathcal{S}_o$ of stabilizing gains. We set all spring and damping constants as well as masses to unity. In state-space representation (8.1a), the state vector $x = [\, p^T \;\; v^T \,]^T$ contains the position and velocity of masses and the measured output $y = p$ is the position only. In this example, the dynamic, input, and output matrices are given by

$$A \;=\; \begin{bmatrix} 0 & I \\ -T & -T \end{bmatrix}, \quad B \;=\; \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad C \;=\; \begin{bmatrix} I & 0 \end{bmatrix}$$

where 0 and I are zero and identity matrices of suitable size, and T is a Toeplitz matrix with 2 on the main diagonal, -1 on the first super and sub-diagonals, and 0 elsewhere.

We solve the LQG problem with $Q = \Sigma_w = I$, $R = \Sigma_v = I$ for $s = 50$ masses, i.e., $n = 2s$ state variables. The algorithm was initialized with scaled matrices of all ones $K^0 = (L^0)^T =$

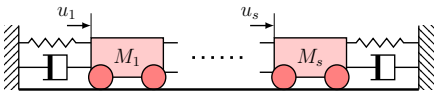


Figure 8.1: Mass-spring-damper system.

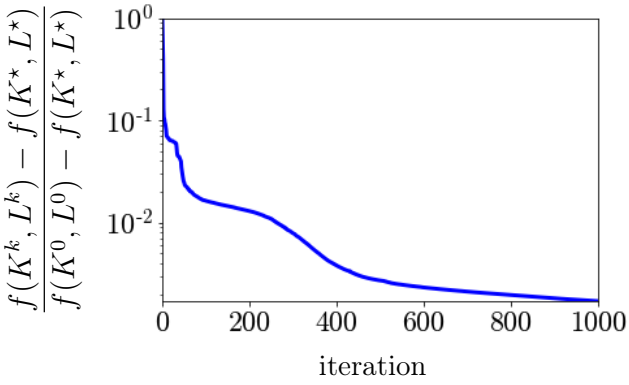


Figure 8.2: Convergence curve of gradient descent for $s = 50$.

$10^{-5}\mathbb{1}$. Figure 8.2 illustrates the convergence curves of gradient descent with a stepsize selected using a backtracking-based procedure initialized with $\alpha_0 = 10^{-3}$ that guarantees stability of the feedback loop and ensures descent. The optimal solution $K^\star$, $L^\star$ is obtained using (8.5) and the corresponding Riccati equations.

8.6 Concluding remarks

Motivated by the recent results on the global exponential convergence of policy gradient algorithms for the model-free LQR problem, in this chapter we studied the standard LQG problem as optimization over controller and observer feedback gains. We present an explicit formulae for the gradient and demonstrate that for open-loop stable systems, in addition to the unique global minimizer, the origin is also a critical point for the LQG problem, thus disproving the gradient dominance property. Numerical experiments for the convergence of gradient descent are also provided. Our work is ongoing to identify conditions under which gradient decent can solve the LQG problem at a linear rate.

Bibliography

[1] L. Bottou and Y. Le Cun, "On-line learning for very large data sets", *Appl. Stoch. Models Bus. Ind.*, vol. 21, no. 2, pp. 137–151, 2005.

[2] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing", *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, 2016.

[3] L. Bottou, F. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning", *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

[4] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems", *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[5] Y. Nesterov, "Gradient methods for minimizing composite objective functions", *Math. Program.*, vol. 140, no. 1, pp. 125–161, 2013.

[6] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning", in *Proc. International Conference on Machine Learning*, 2013, pp. 1139–1147.

[7] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods", *USSR Comput. Math. & Math. Phys.*, vol. 4, no. 5, pp. 1–17, 1964.

[8] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$", in *Dokl. Akad. Nauk SSSR*, vol. 27, 1983, pp. 543–547.

[9] ——, *Lectures on convex optimization*. Springer Optimization and Its Applications, 2018, vol. 137.

[10] D. Maclaurin, D. Duvenaud, and R. Adams, "Gradient-based hyperparameter optimization through reversible learning", in *Proc. International Conference on Machine Learning*, 2015, pp. 2113–2122.

[11] Y. Bengio, "Gradient-based optimization of hyperparameters", *Neural Comput.*, vol. 12, no. 8, pp. 1889–1900, 2000.

[12] A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh, "On optimal generalizability in parametric learning", in *Proc. Neural Information Processing (NIPS)*, 2017, pp. 3458–3468.

[13] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator", in *Proc. International Conference on Machine Learning*, 2018, pp. 1467–1476.

[14] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Convergence and sample complexity of gradient methods for the model-free linear-quadratic regulator problem", *IEEE Trans. Automat. Control*, vol. 67, no. 5, pp. 2435–2450, 2022.

[15] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "Random search for learning the linear quadratic regulator", in *Proc. the 2020 American Control Conference*, 2020, pp. 4798–4803.

[16] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points – online stochastic gradient for tensor decomposition", in *Proc. The 28th Conference on Learning Theory*, vol. 40, 2015, pp. 797–842.

[17] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently", in *Proc. International Conference on Machine Learning*, vol. 70, 2017, pp. 1724–1732.

[18] A. Nagabandi, G. Kahn, R. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning", in *IEEE Int Conf. Robot. Autom.*, 2018, pp. 7559–7566.

[19] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning", 2013, arXiv:1312.5602.

[20] D. Bertsekas, "Approximate policy iteration: A survey and some new methods", *J. Control Theory Appl.*, vol. 9, no. 3, pp. 310–335, 2011.

[21] Y. Abbasi-Yadkori, N. Lazic, and C. Szepesvári, "Model-free linear quadratic control via reduction to expert prediction", in *Proc. Mach. Learn. Res.*, vol. 89, 2019, pp. 3108–3117.

[22] H. Mania, A. Guy, and B. Recht, "Simple random search of static linear policies is competitive for reinforcement learning", in *NeurIPS*, vol. 31, 2018.

[23] Z.-Q. Luo and P. Tseng, "Error bounds and convergence analysis of feasible descent methods: A general approach", *Ann. Oper. Res.*, vol. 46, no. 1, pp. 157–178, 1993.

[24] H. Robbins and S. Monro, "A stochastic approximation method", *Ann. Math. Statist.*, pp. 400–407, 1951.

[25] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming", *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009.

[26] O. Devolder, "Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization", PhD thesis, Louvain-la-Neuve, 2013.

[27] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle", *Math. Program.*, vol. 146, no. 1-2, pp. 37–75, 2014.

[28] P. Dvurechensky and A. Gasnikov, "Stochastic intermediate gradient method for convex problems with stochastic inexact oracle", *J. Optimiz. Theory App.*, vol. 171, no. 1, pp. 121–145, 2016.

[29] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization", in *Proc. Neural Information Processing (NIPS)*, 2011, pp. 1458–1466.

[30] O. Devolder, "Stochastic first order methods in smooth convex optimization", Catholic Univ. Louvain, Louvain-la-Neuve, Tech. Rep., 2011.

[31] F. Bach, "Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression.", *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 595–627, 2014.

[32] B. T. Polyak, "New stochastic approximation type procedures", *Automat. i Telemekh*, vol. 7, no. 98-107, p. 2, 1990.

[33] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging", *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 838–855, 1992.

[34] A. Dieuleveut, N. Flammarion, and F. Bach, "Harder, better, faster, stronger convergence rates for least-squares regression", *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 3520–3570, 2017.

[35] E. Moulines and F. Bach, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning", in *Proc. Neural Information Processing (NIPS)*, 2011, pp. 451–459.

[36] N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan, "Averaging stochastic gradient descent on Riemannian manifolds", in *Proc. The 31st Conference On Learning Theory*, 2018, pp. 650–687.

[37] M. Baes, "Estimate sequence methods: Extensions and approximations", *IFOR Internal report, ETH, Zürich, Switzerland*, 2009.

[38] A. d'Aspremont, "Smooth optimization with approximate gradient", *SIAM J. Optim.*, vol. 19, no. 3, pp. 1171–1183, 2008.

[39] J.-F. Aujol and C. Dossal, "Stability of over-relaxations for the forward-backward algorithm, application to FISTA", *SIAM J. Optim.*, vol. 25, no. 4, pp. 2408–2433, 2015.

[40] B. T. Polyak, "Introduction to optimization. optimization software", *Inc., Publications Division, New York*, vol. 1, 1987.

[41] H. Kwakernaak and R. Sivan, *Linear optimal control systems*. Wiley-Interscience, 1972.

[42] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation", *J. Parallel Distrib. Comput.*, vol. 67, no. 1, pp. 33–46, 2007.

[43] B. Bamieh, M. R. Jovanović, P. Mitra, and S. Patterson, "Coherence in large-scale networks: Dimension dependent limitations of local feedback", *IEEE Trans. Automat. Control*, vol. 57, no. 9, pp. 2235–2249, 2012.

[44] F. Lin, M. Fardad, and M. R. Jovanović, "Optimal control of vehicular formations with nearest neighbor interactions", *IEEE Trans. Automat. Control*, vol. 57, no. 9, pp. 2203–2218, 2012.

[45] M. R. Jovanović and B. Bamieh, "On the ill-posedness of certain vehicular platoon control problems", *IEEE Trans. Automat. Control*, vol. 50, no. 9, pp. 1307–1321, 2005.

[46] F. Dörfler, M. R. Jovanović, M. Chertkov, and F. Bullo, "Sparsity-promoting optimal wide-area control of power networks", *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2281–2291, 2014.

[47] ——, "Sparse and optimal wide-area damping control in power networks", in *Proceedings of the 2013 American Control Conference*, Washington, DC, 2013, pp. 4295–4300.

[48] X. Wu, F. Dörfler, and M. R. Jovanović, "Input-output analysis and decentralized optimal control of inter-area oscillations in power systems", *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 2434–2444, 2016.

[49] J. W. Simpson-Porco, "Input/output analysis of primal-dual gradient algorithms", in *in Proc. 54th Annual Allerton Conference on Communication, Control, and Computing*, 2016, pp. 219–224.

[50] J. W. Simpson-Porco, B. K. Poolla, N. Monshizadeh, and F. Dörfler, "Quadratic performance of primal-dual methods with application to secondary frequency control of power systems", *in Proc. 55th IEEE Conf. Decision Control*, pp. 1840–1845, 2016.

[51] A. Badithela and P. Seiler, "Analysis of the heavy-ball algorithm using integral quadratic constraints", in *Proc. the 2019 American Control Conference*, IEEE, 2019, pp. 4081–4085.

[52] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints", *SIAM J. Optim.*, vol. 26, no. 1, pp. 57–95, 2016.

[53] B. Hu and L. Lessard, "Dissipativity theory for Nesterov's accelerated method", in *Proc. the 34th International Conference on Machine Learning*, ser. Proc. Mach. Learn. Res. 2017, pp. 1549–1557.

[54] S. Cyrus, B. Hu, B. Van Scoy, and L. Lessard, "A robust accelerated optimization algorithm for strongly convex functions", in *Proc. the 2018 American Control Conference*, 2018, pp. 1376–1381.

[55] B. Van Scoy, R. A. Freeman, and K. M. Lynch, "The fastest known globally convergent first-order method for minimizing strongly convex functions", *IEEE Control Syst. Lett.*, vol. 2, no. 1, pp. 49–54, 2018.

[56] M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado, "Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems", *SIAM J. Optim.*, vol. 28, no. 3, pp. 2654–2689, 2018.

[57] B. T. Polyak, "Comparison of the convergence rates for single-step and multi-step optimization algorithms in the presence of noise", *Engrg.Cybern.*, vol. 15, no. 1, pp. 6–10, 1977.

[58] K. Yuan, B. Ying, and A. H. Sayed, "On the influence of momentum acceleration on online learning", *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 6602–6667, 2016.

[59] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2004, vol. 87.

[60] Y. Arjevani, S. Shalev-Shwartz, and O. Shamir, "On lower and upper bounds in smooth and strongly convex optimization", *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 4303–4353, 2016.

[61] B. O'Donoghue and E. Candes, "Adaptive restart for accelerated gradient schemes", *Found. Comput. Math.*, vol. 15, pp. 715–732, 2015.

[62] D. P. Bertsekas, *Convex optimization algorithms*. Athena Scientific, 2015.

[63] Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao, "An accelerated linearized alternating direction method of multipliers", *SIAM J. Imaging Sci.*, vol. 8, no. 1, pp. 644–681, 2015.

[64] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization", *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, 2014.

[65] L. N. Trefethen and M. Embree, *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press, 2005.

[66] M. R. Jovanović and B. Bamieh, "Componentwise energy amplification in channel flows", *J. Fluid Mech.*, vol. 534, pp. 145–183, 2005.

[67] M. R. Jovanović, "From bypass transition to flow control and data-driven turbulence modeling: An input-output viewpoint", *Annu. Rev. Fluid Mech.*, vol. 53, no. 1, pp. 311–345, 2021.

[68] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented Lagrangian method for nonsmooth composite optimization", *IEEE Trans. Automat. Control*, vol. 64, no. 7, pp. 2861–2868, 2019.

[69] E. J. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements", *Comm. Pure and Applied Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[70] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices", *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[71] F. Lin, M. Fardad, and M. R. Jovanović, "Design of optimal sparse feedback gains via the alternating direction method of multipliers", *IEEE Trans. Automat. Control*, vol. 58, no. 9, pp. 2426–2431, 2013.

[72] J. Wang and N. Elia, "A control perspective for centralized and distributed convex optimization", in *in Proc. 50th IEEE Conf. Decision Control*, 2011, pp. 3800–3805.

[73] G. Qu and N. Li, "On the exponential stability of primal-dual gradient dynamics", *IEEE Control Syst. Lett.*, vol. 3, no. 1, pp. 43–48, 2018.

[74] H. D. Nguyen, T. L. Vu, K. Turitsyn, and J. Slotine, "Contraction and robustness of continuous time primal-dual dynamics", *IEEE Control Syst. Lett.*, vol. 2, no. 4, pp. 755–760, 2018.

[75] D. Ding and M. R. Jovanović, "Global exponential stability of primal-dual gradient flow dynamics based on the proximal augmented Lagrangian", in *Proc. the 2019 American Control Conference*, 2019, pp. 3414–3419.

[76] Y. Tang, G. Qu, and N. Li, "Semi-global exponential stability of augmented primal–dual gradient dynamics for constrained convex optimization", *Systems & Control Letters*, vol. 144, p. 104 754, 2020.

[77] D. Ding and M. R. Jovanović, "Global exponential stability of primal-dual gradient flow dynamics based on the proximal augmented Lagrangian: A Lyapunov-based approach", in *Proc. the 59th IEEE Conf. Decision Control*, 2020, pp. 4836–4841.

[78] D. Jakovetić, D. Bajović, J. Xavier, and J. M. Moura, "Primal–dual methods for large-scale and distributed convex optimization and data analytics", *Proc. the IEEE*, vol. 108, no. 11, pp. 1923–1938, 2020.

[79] P. You and E. Mallada, "Saddle flow dynamics: Observable certificates and separable regularization", in *Proc. the 2021 American Control Conference*, 2021, pp. 4817–4823.

[80] N. Parikh and S. Boyd, "Proximal algorithms", *Found. Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2013.

[81] A. Megretski and A. Rantzer, "System analysis via integral quadratic constraints", *IEEE Trans. Autom. Control*, vol. 42, no. 6, pp. 819–830, 1997.

[82] F. Paganini and E. Feron, "Linear matrix inequality methods for robust H_2 analysis: A survey with comparisons", in *Advances in linear matrix inequality methods in control*, SIAM, 2000, pp. 129–151.

[83] B. Anderson and J. Moore, *Optimal Control; Linear Quadratic Methods*. New York, NY: Prentice Hall, 1990.

[84] J. Ackermann, "Parameter space design of robust control systems", *IEEE Trans. Automat. Control*, vol. 25, no. 6, pp. 1058–1072, 1980.

[85] E. Feron, V. Balakrishnan, S. Boyd, and L. El Ghaoui, "Numerical methods for H_2 related problems", in *Proc. the 1992 American Control Conference*, 1992, pp. 2921–2922.

[86] G. E. Dullerud and F. Paganini, *A course in robust control theory: a convex approach*. New York: Springer-Verlag, 2000.

[87] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "On the linear convergence of random search for discrete-time LQR", *IEEE Control Syst. Lett.*, vol. 5, no. 3, pp. 989–994, 2021.

[88] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator", in *Proc. the 58th IEEE Conf. Decision Control*, 2019, pp. 7474–7479.

[89] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition", in *In European Conference on Machine Learning*, 2016, pp. 795–811.

[90] H. Feng and J. Lavaei, "On the exponential number of connected components for the feasible set of optimal decentralized control problems", in *Proc. the 2019 American Control Conference*, 2019, pp. 1430–1437.

[91] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Variance amplification of accelerated first-order algorithms for strongly convex quadratic optimization problems", in *Proc. the 57th IEEE Conf. Decision Control*, 2018, pp. 5753–5758.

[92] ——, "Performance of noisy Nesterov's accelerated method for strongly convex optimization problems", in *Proc. the 2019 American Control Conference*, 2019, pp. 3426–3431.

[93] ——, "Robustness of accelerated first-order algorithms for strongly convex optimization problems", *IEEE Trans. Automat. Control*, vol. 66, no. 6, pp. 2480–2495, 2021.

[94] ——, "Noise amplification of momentum-based optimization algorithms", in *Proc. the 2023 American Control Conference*, submitted, Sand Diego, CA, 2023.

[95] ——, "Tradeoffs between convergence rate and noise amplification for momentum-based accelerated optimization algorithms", 2022, arXiv:2209.11920.

[96] S. Samuelson, H. Mohammadi, and M. R. Jovanović, "Transient growth of accelerated first-order methods", in *Proc. the 2020 American Control Conference*, 2020, pp. 2858–2863.

[97] ——, "On the transient growth of Nesterov's accelerated method for strongly convex optimization problems", in *Proc. the 59th IEEE Conf. Decision Control*, 2020, pp. 5911–5916.

[98] H. Mohammadi, S. Samuelson, and M. R. Jovanović, "Transient growth of accelerated optimization algorithms", *IEEE Trans. Automat. Control*, 2022, doi:10.1109/TAC.2022.3162154.

[99] H. Mohammadi and M. R. Jovanović, "On the noise amplification of primal-dual gradient flow dynamics based on proximal augmented Lagrangian", in *Proc. the 2022 American Control Conference*, Atlanta, GA, 2022, pp. 926–931.

[100] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "Learning the model-free linear quadratic regulator via random search", in *Proc. Machine Learning Research, 2nd Annual Conference on Learning for Dynamics and Control*, vol. 120, Berkeley, CA, 2020, pp. 1–9.

[101] ——, "Model-free linear quadratic regulator", in *Handbook of Reinforcement Learning and Control*, K. G. Vamvoudakis, Y. Wan, F. Lewis, and D. Cansever, Eds., doi:10.1007/978-3-030-60990-0, Springer International Publishing, 2021.

[102] ——, "On the lack of gradient domination for linear quadratic Gaussian problems with incomplete state information", in *Proc. the 60th IEEE Conf. Decision Control*, Austin, TX, 2021, pp. 1120–1124.

[103] N. S. Aybat, A. Fallah, M. M. Gürbüzbalaban, and A. Ozdaglar, "Robust accelerated gradient methods for smooth strongly convex functions", *SIAM J. Opt.*, vol. 30, no. 1, pp. 717–751, 2020.

[104] N. S. Aybat, A. Fallah, M. Gürbüzbalaban, and A. Ozdaglar, "A universally optimal multistage accelerated stochastic gradient method", in *Proc. Neural Information Processing (NIPS)*, 2019.

[105] S. Michalowsky, C. Scherer, and C. Ebenbauer, "Robust and structure exploiting optimization algorithms: An integral quadratic constraint approach", *Int. J. Control*, vol. 94, no. 11, pp. 2956–2979, 2021.

[106] B. T. Polyak and P. Shcherbakov, "Lyapunov functions: An optimization theory perspective", *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 7456–7461, 2017.

[107] B. T. Polyak and G. V. Smirnov, "Transient response in matrix discrete-time linear systems", *Autom. Remote Control*, vol. 80, no. 9, pp. 1645–1652, 2019.

[108] M. B. Cohen, J. Diakonikolas, and L. Orecchia, "On acceleration with noise-corrupted gradients", in *Proc. the 35th International Conference on Machine Learning*, ser. Proc. Mach. Learn. Res. Vol. 80, 2018, pp. 1019–1028.

[109] B. Van Scoy and L. Lessard, "The speed-robustness trade-off for first-order methods with additive gradient noise", 2021, arXiv:2109.05059.

[110] M. Muehlebach and M. Jordan, "A dynamical systems perspective on Nesterov acceleration", in *International Conference on Machine Learning*, PMLR, 2019, pp. 4656–4662.

[111] Z. He, A. S. Rakin, and D. Fan, "Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack", in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 588–597.

[112] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds", in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 2014, pp. 464–473.

[113] S. B. Gelfand and S. K. Mitter, "Recursive stochastic algorithms for global optimization in R^d", *SIAM J. Control Optim.*, vol. 29, no. 5, pp. 999–1018, 1991.

[114] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis", in *Proc. the Conference on Learning Theory*, ser. Proc.Mach. Learn. Res. Vol. 65, 2017, pp. 1674–1703.

[115] Y. Zhang, P. Liang, and M. Charikar, "A hitting time analysis of stochastic gradient Langevin dynamics", in *Proc. the 2017 Conference on Learning Theory*, ser. Proc. Mach. Learn. Res. Vol. 65, 2017, pp. 1980–2022.

[116] K. Ogata, *Discrete-time control systems*. New Jersey: Prentice-Hall, 1994.

[117] R. Padmanabhan and P. Seiler, "Analysis of gradient descent with varying step sizes using integral quadratic constraints", 2022, arXiv:2210.00644.

[118] B. Hu, P. Seiler, and A. Rantzer, "A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints", in *Proc. the 2017 Conference on Learning Theory*, ser. Proc. Mach. Learn. Res. 2017, pp. 1157–1189.

[119] B. Hu, P. Seiler, and L. Lessard, "Analysis of biased stochastic gradient descent using sequential semidefinite programs", *Math. Program.*, vol. 187, no. 1, pp. 383–408, 2021.

[120] S. Hassan-Moghaddam and M. R. Jovanović, "Topology design for stochastically-forced consensus networks", *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1075–1086, 2018.

[121] A. Zare, H. Mohammadi, N. K. Dhingra, T. T. Georgiou, and M. R. Jovanović, "Proximal algorithms for large-scale statistical modeling and sensor/actuator selection", *IEEE Trans. Automat. Control*, vol. 65, no. 8, pp. 3441–3456, 2020.

[122] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning", in *Proc. International Conference on Machine Learning*, 2013, pp. 1139–1147.

[123] S. Michalowsky, C. Scherer, and C. Ebenbauer, "Robust and structure exploiting optimisation algorithms: An integral quadratic constraint approach", *Int. J. Control*, pp. 1–24, 2020.

[124] J. I. Poveda and N. Li, "Robust hybrid zero-order optimization algorithms with acceleration via averaging in time", *Automatica*, p. 109 361, 2021.

[125] B Can, M. Gurbuzbalaban, and L. Zhu, "Accelerated linear convergence of stochastic momentum methods in Wasserstein distances", in *International Conference on Machine Learning*, PMLR, 2019, pp. 891–901.

[126] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004, vol. 87.

[127] M. Danilova and G. Malinovsky, "Averaged heavy-ball method", 2021, arXiv:2111.05430.

[128] S. Hassan-Moghaddam and M. R. Jovanović, "Proximal gradient flow and Douglas-Rachford splitting dynamics: Global exponential stability via integral quadratic constraints", *Automatica*, vol. 123, 109311 (7 pages), 2021.

[129] J. W. Simpson-Porco, B. K. Poolla, N. Monshizadeh, and F. Dörfler, "Input-output performance of linear–quadratic saddle-point algorithms with application to distributed resource allocation problems", *IEEE Trans. Automat. Control*, vol. 65, no. 5, pp. 2032–2045, 2019.

[130] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator", *Found. Comput. Math.*, pp. 1–47, 2017.

[131] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification", in *Proc. Mach. Learn. Res.*, 2018, 439—473.

[132] D. Malik, A. Panajady, K. Bhatia, K. Khamaru, P. L. Bartlett, and M. J. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear-quadratic systems", *J. Mach. Learn. Res.*, vol. 51, 1–51, 2019.

[133] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud, "Convergence guarantees of policy optimization methods for Markovian jump linear systems", in *Proc. the American Control Conference*, 2020.

[134] K. Zhang, B. Hu, and T. Basar, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence", *SIAM J. Control Optim.*, vol. 59, no. 6, pp. 4081–4109, 2021.

[135] L. Furieri, Y. Zheng, and M. Kamgarpour, "Learning the globally optimal distributed LQ regulator", in *Learning for Dynamics and Control*, 2020, pp. 287–297.

[136] I. Fatkhullin and B. T. Polyak, "Optimizing static linear feedback: Gradient method", *SIAM J. Control Optim.*, vol. 59, no. 5, pp. 3887–3911, 2021.

[137] D Kleinman, "On an iterative technique for Riccati equation computations", *IEEE Trans. Automat. Control*, vol. 13, no. 1, pp. 114–115, 1968.

[138] S. Bittanti, A. J. Laub, and J. C. Willems, *The Riccati Equation*. Berlin, Germany: Springer-Verlag, 2012.

[139] P. L. D. Peres and J. C. Geromel, "An alternate numerical solution to the linear quadratic problem", *IEEE Trans. Automat. Control*, vol. 39, no. 1, pp. 198–202, 1994.

[140] V. Balakrishnan and L. Vandenberghe, "Semidefinite programming duality and linear time-invariant systems", *IEEE Trans. Automat. Control*, vol. 48, no. 1, pp. 30–41, 2003.

[141] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case", 2019, arXiv:1907.08921.

[142] W. S. Levine and M. Athans, "On the determination of the optimal constant output feedback gains for linear multivariable systems", *IEEE Trans. Automat. Control*, vol. 15, no. 1, pp. 44–48, 1970.

[143] F. Lin, M. Fardad, and M. R. Jovanović, "Augmented Lagrangian approach to design of structured optimal state feedback gains", *IEEE Trans. Automat. Control*, vol. 56, no. 12, pp. 2923–2929, 2011.

[144] M. Fardad, F. Lin, and M. R. Jovanović, "Sparsity-promoting optimal control for a class of distributed systems", in *Proc. the 2011 American Control Conference*, San Francisco, CA, 2011, pp. 2050–2055.

[145] M. R. Jovanović and N. K. Dhingra, "Controller architectures: Tradeoffs between performance and structure", *Eur. J. Control*, vol. 30, pp. 76–91, 2016.

[146] F. Lin, M. Fardad, and M. R. Jovanović, "Sparse feedback synthesis via the alternating direction method of multipliers", in *Proceedings of the 2012 American Control Conference*, Montréal, Canada, 2012, pp. 4765–4770.

[147] X. Wu and M. R. Jovanović, "Sparsity-promoting optimal control of systems with symmetries, consensus and synchronization networks", *Syst. Control Lett.*, vol. 103, pp. 1–8, 2017.

[148] B. T. Polyak, M. Khlebnikov, and P. Shcherbakov, "An LMI approach to structured sparse feedback design in linear control systems", in *Proc. the 2013 European Control Conference*, 2013, pp. 833–838.

[149] N. K. Dhingra, M. R. Jovanović, and Z. Q. Luo, "An ADMM algorithm for optimal sensor and actuator selection", in *Proc. the 53rd IEEE Conf. Decision Control*, 2014, pp. 4039–4044.

[150] A. Zare, T. T. Georgiou, and M. R. Jovanović, "Stochastic dynamical modeling of turbulent flows", *Annu. Rev. Control Robot. Auton. Syst.*, vol. 3, pp. 195–219, 2020.

[151] B. Recht, "A tour of reinforcement learning: The view from continuous control", *Annu. Rev. Control Robot. Auton. Syst.*, vol. 2, pp. 253–279, 2019.

[152] H. T. Toivonen, "A globally convergent algorithm for the optimal constant output feedback problem", *Int. J. Control*, vol. 41, no. 6, pp. 1589–1599, 1985.

[153] T. Rautert and E. W. Sachs, "Computational design of optimal output feedback controllers", *SIAM J. Optim*, vol. 7, no. 3, pp. 837–852, 1997.

[154] A. Vannelli and M. Vidyasagar, "Maximal Lyapunov functions and domains of attraction for autonomous nonlinear systems", *Automatica*, vol. 21, no. 1, pp. 69 –80, 1985.

[155] M. Jones, H. Mohammadi, and M. M. Peet, "Estimating the region of attraction using polynomial optimization: A converse lyapunov result", in *in Proc. 56th IEEE Conf. on Decision and Control*, 2017, pp. 1796–1802.

[156] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "On the stability of gradient flow dynamics for a rank-one matrix approximation problem", in *Proc. the 2018 American Control Conference*, Milwaukee, WI, 2018, pp. 4533–4538.

[157] H. K. Khalil, *Nonlinear Systems*. New York: Prentice Hall, 1996.

[158] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[159] S.-I Amari, "Natural gradient works efficiently in learning", *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.

[160] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

[161] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator", *IEEE Trans. Automat. Control*, vol. 16, no. 4, pp. 382–384, 1971.

[162] K. Mårtensson, "Gradient methods for large-scale and distributed linear quadratic control", PhD thesis, Lund University, 2012.

[163] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations", *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[164] M. Rudelson and R. Vershynin, "Hanson-Wright inequality and sub-Gaussian concentration", *Electron. Commun. Probab.*, vol. 18, 2013.

[165] Y. Zheng, Y. Tang, and N. Li, "Analysis of the optimization landscape of linear quadratic gaussian (lqg) control", 2021, arXiv:2102.04393.

[166] K. J. Åström, *Introduction to stochastic control theory*. Academic Press, New York, 1970.

[167] F. Topsok, "Some bounds for the logarithmic function", *Inequal. Theory Appl.*, vol. 4, p. 137, 2006.

[168] H. T. Toivonen and P. M. Mäkilä, "Newton's method for solving parametric linear quadratic control problems", *Int. J. Control*, vol. 46, no. 3, pp. 897–911, 1987.

[169] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks", *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 742–769, 2019.

[170] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[171] D. Pollard, "Mini empirical", 2015. [Online]. Available: `http://www.stat.yale.edu/~pollard/Books/Mini/`.

[172] A. W. Vaart and J. A. Wellner, *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996.

[173] T. Ma and A. Wigderson, "Sum-of-Squares lower bounds for sparse PCA", in *Advances in Neural Information Processing Systems*, 2015, pp. 1612–1620.

Appendices

Appendix A

Supporting proofs for Chapter 2

A.1 Quadratic problems

A.1.1 Proof of Theorem 1

For gradient descent, $\hat{A}_i = 1 - \alpha\lambda_i$ and $\hat{B}_i = 1$ are scalars and the solution to (2.9) is given by

$$\hat{P}_i \;:=\; \sigma^2 p_i \;=\; \frac{\sigma^2}{1 \,-\, (1 \,-\, \alpha\lambda_i)^2} \;=\; \frac{\sigma^2}{\alpha\lambda_i(2 \,-\, \alpha\lambda_i)}.$$

For the accelerated methods, we note that for any $\hat{A}_i$ and $\hat{B}_i$ of the form

$$\hat{A}_i \;=\; \begin{bmatrix} 0 & 1 \\ a_i & b_i \end{bmatrix},\; \hat{B}_i \;=\; \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

the solution $\hat{P}_i$ to Lyapunov equation (2.9) is given by

$$\hat{P}_i \;=\; \sigma^2 \begin{bmatrix} p_i & b_i p_i/(1-a_i) \\ b_i p_i/(1-a_i) & p_i \end{bmatrix}$$

where

$$p_i \;:=\; \frac{a_i \,-\, 1}{(a_i \,+\, 1)(b_i \,+\, a_i \,-\, 1)(b_i \,-\, a_i \,+\, 1)}. \tag{A.1}$$

The parameters a_i and b_i for Nesterov's algorithm are $\{a_i = -\beta(1-\alpha\lambda_i);\, b_i = (1+\beta)(1-\alpha\lambda_i)\}$ and for the heavy-ball method we have $\{a_i = -\beta;\, b_i = 1 + \beta - \alpha\lambda_i\}$. Now, since $\hat{C}_i = 1$ for gradient descent and $\hat{C}_i = [\,1\;\;0\,]$ for the accelerated algorithms, it follows that for all three algorithms we have $\hat{J}(\lambda_i) := \mathrm{trace}\,(\hat{C}_i \hat{P}_i \hat{C}_i^T) = \sigma^2 p_i$. Finally, if we use the expression for p_i for gradient descent and substitute for a_i and b_i in (A.1) for the accelerated algorithms, we obtain the expressions for $\hat{J}$ in the statement of the theorem.

A.1.2 Proof of Proposition 1

To show that $\hat{J}_{\mathrm{na}}(\lambda)/\hat{J}_{\mathrm{gd}}(\lambda)$ is a decreasing function of $\lambda \in [m, L]$, we split this ratio into the sum of two homographic functions $\hat{J}_{\mathrm{na}}(\lambda)/\hat{J}_{\mathrm{gd}}(\lambda) = \sigma_1(\lambda) + \sigma_2(\lambda)$, where

$$\sigma_1(\lambda) \;:=\; \frac{4\alpha_{\mathrm{gd}}\beta}{\alpha_{\mathrm{na}}(3\beta+1)(1-\beta)}\frac{1-\frac{\alpha_{\mathrm{gd}}}{2}\lambda}{1+\frac{\alpha_{\mathrm{na}}\beta}{1-\beta}\lambda}, \quad \sigma_2(\lambda) \;:=\; \frac{\alpha_{\mathrm{gd}}}{\alpha_{\mathrm{na}}(3\beta+1)}\frac{1-\frac{\alpha_{\mathrm{gd}}}{2}\lambda}{1-\frac{\alpha_{\mathrm{na}}(2\beta+1)}{2+2\beta}\lambda}. \tag{A.2}$$

Now, if we substitute the parameters provided in Table 2.2 into (A.2), it follows that the signs of the derivatives $\mathrm{d}\sigma_1/\mathrm{d}\lambda$ and $\mathrm{d}\sigma_2/\mathrm{d}\lambda$ satisfy

$$\begin{aligned}
\operatorname{sign}\left(\frac{\mathrm{d}\,\sigma_1}{\mathrm{d}\lambda}\right) \;&=\; \operatorname{sign}\left(-\tfrac{\alpha_{\mathrm{na}}\beta}{1-\beta}-\tfrac{\alpha_{\mathrm{gd}}}{2}\right) \;=\; \operatorname{sign}\left(-\frac{\kappa+\kappa\sqrt{3\,\kappa+1}+\sqrt{3\,\kappa+1}-1}{m\,(3\,\kappa+1)\,(\kappa+1)}\right) \;<\; 0, \quad \forall\,\kappa>1 \\
\operatorname{sign}\left(\frac{\mathrm{d}\,\sigma_2}{\mathrm{d}\lambda}\right) \;&=\; \operatorname{sign}\left(\tfrac{\alpha_{\mathrm{na}}(2\beta+1)}{2+2\beta}-\tfrac{\alpha_{\mathrm{gd}}}{2}\right) \;=\; \operatorname{sign}\left(-\frac{2\left(\kappa-\sqrt{3\,\kappa+1}+1\right)}{m\,(3\,\kappa+1)^{3/2}\,(\kappa+1)}\right) \;<\; 0, \quad \forall\,\kappa>1.
\end{aligned}$$

Furthermore, since the critical points of the functions $\sigma_1(\lambda)$ and $\sigma_2(\lambda)$ are not in $[m, L]$,

$$\lambda_{\mathrm{crt1}} \;=\; -\frac{m(3\kappa+1)}{\sqrt{3\kappa+1}-2} \;<\; 0 \;<\; m, \quad \lambda_{\mathrm{crt2}} \;=\; \frac{m\,(3\kappa+1)\,\sqrt{3\kappa+1}}{3\,\sqrt{3\kappa+1}-2} \;>\; m\,\kappa \;=\; L$$

we conclude that both σ_1 and σ_2 are decreasing functions over the interval $[m, L]$. We next prove (2.13a) and (2.13b).

It is straightforward to verify that both $\hat{J}_{\mathrm{gd}}(\lambda)$ and $\hat{J}_{\mathrm{na}}(\lambda)$ are quasi-convex functions over the interval $[m, L]$ and that the respective minima are attained at the critical point $\lambda = 1/\alpha$. Quasi-convexity also implies

$$\max_{\lambda\,\in\,[m,L]} \hat{J}(\lambda) \;=\; \max\{\hat{J}(m), \hat{J}(L)\}. \tag{A.3}$$

Now, letting $\alpha = 2/(L+m)$ in the expression for $\hat{J}_{\mathrm{gd}}$ gives $\hat{J}_{\mathrm{gd}}(m) = \hat{J}_{\mathrm{gd}}(L) = (\kappa+1)^2/(4\kappa)$ which in conjunction with (A.3) complete the proof for (2.13a). Finally, since the ratio $\hat{J}_{\mathrm{na}}(\lambda)/\hat{J}_{\mathrm{gd}}(\lambda)$ is decreasing, we have $\hat{J}_{\mathrm{na}}(L)/\hat{J}_{\mathrm{gd}}(L) \leq \hat{J}_{\mathrm{na}}(m)/\hat{J}_{\mathrm{gd}}(m)$. Combining this inequality with $\hat{J}_{\mathrm{gd}}(m) = \hat{J}_{\mathrm{gd}}(L)$ and (A.3) completes the proof of (2.13b).

A.1.3 Proof of Theorem 3

From Proposition 1, it follows that

$$\frac{\hat{J}_{\mathrm{na}}(L)}{\hat{J}_{\mathrm{gd}}(L)} \;\leq\; \frac{\hat{J}_{\mathrm{na}}(\lambda_i)}{\hat{J}_{\mathrm{gd}}(\lambda_i)} \;\leq\; \frac{\hat{J}_{\mathrm{na}}(m)}{\hat{J}_{\mathrm{gd}}(m)} \tag{A.4a}$$

for all λ_i and

$$\sum_{i=1}^{n-1} \hat{J}_{\mathrm{gd}}(\lambda_i) \;\leq\; (n-1)\hat{J}_{\mathrm{gd}}(m) = (n-1)\hat{J}_{\mathrm{gd}}(L). \tag{A.4b}$$

For the upper bound, we have

$$\frac{J_{\mathrm{na}}}{J_{\mathrm{gd}}} = \frac{\sum_{i=1}^{n} \hat{J}_{\mathrm{na}}(\lambda_i)}{\sum_{i=1}^{n} \hat{J}_{\mathrm{gd}}(\lambda_i)} \leq \frac{\hat{J}_{\mathrm{na}}(L) \ + \ \frac{\hat{J}_{\mathrm{na}}(m)}{\hat{J}_{\mathrm{gd}}(m)} \sum_{i=1}^{n-1} \hat{J}_{\mathrm{gd}}(\lambda_i)}{\hat{J}_{\mathrm{gd}}(L) \ + \ \sum_{i=1}^{n-1} \hat{J}_{\mathrm{gd}}(\lambda_i)} \leq \frac{\hat{J}_{\mathrm{na}}(L) \ + \ (n \ - \ 1)\hat{J}_{\mathrm{na}}(m)}{\hat{J}_{\mathrm{gd}}(L) \ + \ (n \ - \ 1)\hat{J}_{\mathrm{gd}}(m)}$$

where the first inequality follows from (A.4a). The second inequality can be verified by multiplying both sides with the product of the denominators and using $\hat{J}_{\mathrm{gd}}(m) = \hat{J}_{\mathrm{gd}}(L)$, $\hat{J}_{\mathrm{na}}(m) \geq \hat{J}_{\mathrm{na}}(L)$, and (A.4b). Similarly, for the lower bound we can write

$$\frac{J_{\mathrm{na}}}{J_{\mathrm{gd}}} = \frac{\sum_{i=1}^{n} \hat{J}_{\mathrm{na}}(\lambda_i)}{\sum_{i=1}^{n} \hat{J}_{\mathrm{gd}}(\lambda_i)} \geq \frac{\hat{J}_{\mathrm{na}}(m) \ + \ \frac{\hat{J}_{\mathrm{na}}(L)}{\hat{J}_{\mathrm{gd}}(L)} \sum_{i=2}^{n} \hat{J}_{\mathrm{gd}}(\lambda_i)}{\hat{J}_{\mathrm{gd}}(m) \ + \ \sum_{i=2}^{n} \hat{J}_{\mathrm{gd}}(\lambda_i)} \geq \frac{\hat{J}_{\mathrm{na}}(m) \ + \ (n \ - \ 1)\hat{J}_{\mathrm{na}}(L)}{\hat{J}_{\mathrm{gd}}(m) \ + \ (n \ - \ 1)\hat{J}_{\mathrm{gd}}(L)}.$$

Again, the first inequality follows from (A.4a) and the second inequality can be verified by multiplying both sides with the product of the denominators and using $\hat{J}_{\mathrm{gd}}(m) = \hat{J}_{\mathrm{gd}}(L)$, $\hat{J}_{\mathrm{na}}(m) \geq \hat{J}_{\mathrm{na}}(L)$, and (A.4b).

A.1.4 Proof of the bounds in (2.16)

From Proposition 1, we have

$$\hat{J}_{\mathrm{na}}(m) \ = \ \frac{b^4(b^2 - 2\,b + 2)}{32\,(b-1)^3}, \quad \hat{J}_{\mathrm{na}}(L) \ = \ \frac{9\,b^4(b^2 + 2\,b - 2)}{32\ (b^2-1)\,(2\,b-1)\,(b^2 - b + 1)}$$

where $b := \sqrt{3\,\kappa + 1} > 2$. The upper and lower bounds on $\hat{J}_{\mathrm{na}}(m)$ are obtained as follows

$$\frac{b^3}{32} \ \leq \ \frac{b^4((b-1)^2+1)}{32\,(b-1)^3} \ = \ \hat{J}_{\mathrm{na}}(m) \ \leq \ \frac{b^3(b + c_1(b))(b^2 - 2\,b + 2 + c_2(b))}{32\,(b-1)^3} = \frac{b^3}{8}$$

where the positive quantities $c_1(b) := b - 2$ and $c_2(b) := b^2 - 2b$ are added to yield a simple upper bound. Similarly, for $\hat{J}_{\mathrm{na}}(L)$ we have

$$\begin{aligned}
\frac{9b}{64} \ &= \ \frac{(9/32)\,b^4(b^2 + 2\,b - 2)}{((b^2-1)+1)\,((2\,b-1)+1)\,(b^2 - b + 1 + c_3(b))} \ \leq \ \hat{J}_{\mathrm{na}}(L) \\
\frac{9b}{8} \ &= \ \frac{(9/32)\,b^4(b^2 + 2\,b - 2 + c_4(b))}{(b^2-1)(2\,b-1-c_5(b))(b^2 - b + 1 - c_6(b))} \ \geq \ \hat{J}_{\mathrm{na}}(L)
\end{aligned}$$

where the positive quantities $c_3(b) := 3b - 3$, $c_4(b) := b^2 - 2b$, $c_5(b) := b - 1$, and $c_6(b) := (1/2)b^2 - b + 1$ are introduced to obtain tractable bounds.

A.2 General strongly convex problems

A.2.1 Proof of Lemma 1

Let us define the positive semidefinite function $V(\psi) := \psi^T X \psi$ and let $\eta := [\,\psi^T \; u^T\,]^T$. Using LMI (2.23) and (2.22), we can write

$$\begin{aligned}
\|z^t\|^2 \;&=\; (\eta^t)^T \begin{bmatrix} C_z^T C_z & 0 \\ 0 & 0 \end{bmatrix} \eta^t \\
&\leq\; -(\eta^t)^T \begin{bmatrix} A^T X\, A - X & A^T X\, B_u \\ B_u^T\, X\, A & B_u^T\, X\, B_u \end{bmatrix} \eta^t \;-\; \lambda\, (\eta^t)^T \begin{bmatrix} C_y^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C_y & 0 \\ 0 & I \end{bmatrix} \eta^t \\
&=\; (\eta^t)^T \Bigg(\begin{bmatrix} X & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} A^T \\ B_u^T \end{bmatrix} X \begin{bmatrix} A^T \\ B_u^T \end{bmatrix}^T \Bigg) \eta^t \;-\; \lambda \begin{bmatrix} y^t \\ u^t \end{bmatrix}^T \Pi \begin{bmatrix} y^t \\ u^t \end{bmatrix} \\
&\leq\; V(\psi^t) \;-\; V(\psi^{t+1}) \;+\; 2\sigma (\psi^t)^T A^T X\, B_w\, w^t \\
&\quad +\; \sigma^2 (w^t)^T B_w^T\, X\, B_w\, w^t \;+\; 2\sigma (u^t)^T B_u^T\, X\, B_w\, w^t.
\end{aligned}$$

Since w^t is a zero-mean white input with identity covariance which is independent of u^t and x^t, if we take the average of the above inequality over t and expectation over different realizations of w^t, we obtain

$$\frac{1}{\bar{T}} \sum_{t=1}^{\bar{T}} \mathbb{E}\left(\|z^t\|^2\right) \leq\; \frac{1}{\bar{T}}\, \mathbb{E}\left(V(\psi^1) \;-\; V(\psi^{\bar{T}+1})\right) \;+\; \sigma^2 \mathrm{trace}\,(B_w^T X B_w)$$

Therefore, letting $\bar{T} \to \infty$ and using $X \succeq 0$ lead to $J \leq \sigma^2 \mathrm{trace}\,(B_w^T\, X\, B_w)$, which completes the proof.

A.2.2 Proof of Lemma 2

In order to prove Lemma 2, we present a technical lemma which along the lines of results of [56] provides us with an upper bound on the difference between the objective value at two consecutive iterations.

Lemma 1 *Let $f \in \mathcal{F}_m^L$ and $\kappa := L/m$. Then, Nesterov's accelerated method, with the notation introduced in Section 2.4, satisfies*

$$\begin{aligned}
f(x^{t+2}) \;-\; f(x^{t+1}) \;\leq\; &\frac{1}{2} \left(N_1 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} + \begin{bmatrix} \sigma w^t \\ 0 \end{bmatrix} \right)^T \begin{bmatrix} L\, I & I \\ I & 0 \end{bmatrix} \left(N_1 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} + \begin{bmatrix} \sigma w^t \\ 0 \end{bmatrix} \right) + \\
&\frac{1}{2} \left(N_2 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} \right)^T \begin{bmatrix} -m\, I & I \\ I & 0 \end{bmatrix} \left(N_2 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} \right)
\end{aligned}$$

where N_1 and N_2 are defined in Lemma 2.

Proof: For any $f \in \mathcal{F}_m^L$, the Lipschitz continuity of ∇f implies

$$f(x^{t+2}) \,-\, f(y^t) \,\leq\, \frac{1}{2}\begin{bmatrix} x^{t+2} - y^t \\ \nabla f(y^t) \end{bmatrix}^T \begin{bmatrix} L\,I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} x^{t+2} - y^t \\ \nabla f(y^t) \end{bmatrix} \tag{A.5}$$

and the strong convexity of f yields

$$f(y^t) \,-\, f(x^{t+1}) \,\leq\, \frac{1}{2}\begin{bmatrix} y^t - x^{t+1} \\ \nabla f(y^t) \end{bmatrix}^T \begin{bmatrix} -m\,I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} y^t - x^{t+1} \\ \nabla f(y^t) \end{bmatrix}. \tag{A.6}$$

Moreover, the state and output equations in (2.5) lead to

$$\begin{bmatrix} x^{t+2} - y^t \\ \nabla f(y^t) \end{bmatrix} \,=\, N_1 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix} + \begin{bmatrix} \sigma w^t \\ 0 \end{bmatrix}, \quad \begin{bmatrix} y^t - x^{t+1} \\ \nabla f(y^t) \end{bmatrix} \,=\, N_2 \begin{bmatrix} \psi^t \\ u^t \end{bmatrix}. \tag{A.7}$$

Summing up (A.5) and (A.6) and substituting for the terms $\begin{bmatrix} x^{t+2} - y^t \\ \nabla f(y^t) \end{bmatrix}$ and $\begin{bmatrix} x^{t+2} - y^t \\ \nabla f(y^t) \end{bmatrix}$ from (A.7) completes the proof. □

Let us define the positive semidefinite function $V(\psi) := \psi^T X \psi$ and let $\eta := [\,\psi^T \;\; u^T\,]^T$. Similar to the first part of the proof of Lemma 1, we can use LMI (2.24) and inequality (2.19) to write

$$\begin{aligned} \|z^t\|^2 \,\leq\, V(\psi^t) \,-\, V(\psi^{t+1}) \,+\, 2\sigma(\psi^t)^T A^T X\, B_w\, w^t \\ +\, \sigma^2 (w^t)^T B_w^T X\, B_w\, w^t \,+\, 2\sigma (u^t)^T B_u^T X\, B_w\, w^t \,-\, (\eta^t)^T M\, \eta^t. \end{aligned} \tag{A.8}$$

From Lemma 1, it follows that

$$(\eta^t)^T M\, \eta^t \,\geq\, 2\left(f(x^{t+2}) \,-\, f(x^{t+1})\right) \,-\, \sigma^2 L\, \|w^t\|^2 \,-\, 2\begin{bmatrix} \sigma w^t \\ 0 \end{bmatrix}^T \begin{bmatrix} L\,I & I \\ I & 0 \end{bmatrix} N_1 \eta^t. \tag{A.9}$$

Now, combining inequalities (A.8) and (A.9) yields

$$\begin{aligned} \|z^t\|^2 \,\leq\, V(\psi^t) \,-\, V(\psi^{t+1}) \,+\, 2\sigma(\psi^t)^T A^T X\, B_w\, w^t \,+\, \sigma^2 (w^t)^T B_w^T X\, B_w\, w^t \\ +\, 2\sigma (u^t)^T B_u^T X\, B_w\, w^t \,-\, 2\,\lambda_2 \left(f(x^{t+2}) \,-\, f(x^{t+1})\right) \\ +\, \lambda_2 \sigma^2 L \|w^t\|^2 \,+\, 2\lambda_2 \begin{bmatrix} \sigma w^t \\ 0 \end{bmatrix}^T \begin{bmatrix} L\,I & I \\ I & 0 \end{bmatrix} N_1 \eta^t. \end{aligned} \tag{A.10}$$

Since w^t is a zero-mean white input with identity covariance which is independent of u^t and x^t, taking the expectation of the last inequality yields

$$\begin{aligned} \mathbb{E}\left(\|z^t\|^2\right) \,\leq\, \mathbb{E}\left(V(\psi^t) - V(\psi^{t+1})\right) + \sigma^2 \mathrm{trace}\,(B_w^T X\, B_w) \\ +\, 2\,\lambda_2\, \mathbb{E}\left(f(x^{t+1}) \,-\, f(x^{t+2})\right) + n\,\sigma^2 L\,\lambda_2 \end{aligned}$$

and taking the average over the first $\bar{T}$ iterations results in

$$\frac{1}{\bar{T}}\sum_{t=1}^{\bar{T}}\mathbb{E}\left(\|z^t\|^2\right) \;\leq\; \frac{1}{\bar{T}}\,\mathbb{E}\Big(V(\psi^1) \;-\; V(\psi^{\bar{T}+1})\Big) \\ +\; \sigma^2\mathrm{trace}\,(B_w^T\,X\,B_w) \;+\; \frac{2\,\lambda_2}{\bar{T}}\,\mathbb{E}\left(f(x^2) \;-\; f(x^{\bar{T}+2})\right) \;+\; n\,\sigma^2 L\,\lambda_2.$$

Finally, using positive definiteness of the function V, strong convexity of the function f, and letting $\bar{T}\to\infty$, it follows that $J\leq\sigma^2(nL\lambda_2+\mathrm{trace}\,(B_w^T X\,B_w))$ as required.

A.2.3 Proof of Theorem 5

Using Theorem (1), it is straightforward to show that for gradient descent and Nesterov's method with the parameters provided in Table 2.1, the function $f(x):=\frac{m}{2}\|x\|^2$ leads to the largest variance amplification J among the quadratic objective functions within $\mathcal{F}_m^L$. This yields the lower bounds

$$q_{\mathrm{gd}} \;=\; J_{\mathrm{gd}} \;\leq\; J_{\mathrm{gd}}^\star, \quad q_{\mathrm{na}} \;=\; J_{\mathrm{na}} \;\leq\; J_{\mathrm{na}}^\star$$

with J_{gd} and J_{na} corresponding to $f(x)=\frac{m}{2}\|x\|^2$. We next show that $J_{\mathrm{gd}}\leq q_{\mathrm{gd}}$.

To obtain the best upper bound on J_{gd} using Lemma 1, we minimize $\mathrm{trace}\,(B_w^T X B_w)$ subject to LMI (2.23), $X\succeq 0$, and $\lambda\geq 0$. For gradient descent, if we use the representation in (2.21c), then the negative definiteness of the $(1,1)$-block of LMI (2.23) implies that

$$X \;\succeq\; \frac{1}{\alpha\,m(2\;-\;\alpha\,m)}\,I \;=\; \frac{\kappa^2}{2\kappa-1}\,I. \tag{A.11}$$

It is straightforward to show that the pair

$$X \;=\; \frac{\kappa^2}{2\kappa-1}\,I, \quad \lambda \;=\; \frac{1-\alpha\,m}{m(2-\alpha\,m)(L-m)} \tag{A.12}$$

is feasible as the LMI (2.23) becomes

$$\begin{bmatrix} 0 & 0 \\ 0 & \frac{-1}{m^2(2\kappa-1)}\,I \end{bmatrix} \;\preceq\; 0.$$

Thus, X and λ given by (A.12) provide a solution to LMI (2.23). Therefore, inequality (A.11) is tight and it provides the best achievable upper bound

$$J_{\mathrm{gd}} \;\leq\; \mathrm{trace}\,(B_w^T\,X\,B_w) \;=\; \frac{n\kappa^2}{2\,\kappa\;-\;1}.$$

Finally, we show $J_{\rm na} \leq 4.08 q_{\rm na}$ by finding a sub-optimal feasible point for (2.26). Let $X := \left[\begin{array}{cc} x_1 I & x_0 I \\ x_0 I & x_2 I \end{array} \right]$ with

$$\begin{aligned} x_1 &:= \frac{1}{s(\kappa)} \left(2\,\kappa^{3.5} - 8\,\kappa^3 + 11\,\kappa^{2.5} + 5\,\kappa^2 - 14\,\kappa^{1.5} + 8\,\kappa - 2\,\kappa^{0.5} \right) \\ x_0 &:= \frac{-1}{s(\kappa)} \left(2\,\kappa^{1.5} \left(\kappa^{0.5} - 1 \right)^3 \left(\kappa^{0.5} + 1 \right) \right) \\ x_2 &:= \frac{\kappa^{1.5}}{s(\kappa)} \left(2\,\kappa^2 - 3\,\kappa + 5\,\kappa^{0.5} - 2 \right), \quad s(\kappa) := 8\,\kappa^2 - 6\,\kappa^{1.5} - 2\,\kappa + 3\,\kappa^{0.5} - 1 \end{aligned}$$

and let $\lambda_1 := (\kappa/L)^2/(2\kappa - 1)$ and $\lambda_2 := -x_0/(Ls(\kappa))$. We first show that $(\lambda_1, \lambda_2, X)$ is feasible for problem (2.26). It is straightforward to verify that $s(\kappa)$, $x_1 s(\kappa)$, $x_2 s(\kappa)$, and $-x_0 s(\kappa)$ (which are polynomials of degree less than 7 in $\sqrt{\kappa}$) are all positive for any $\kappa \geq 1$. Hence, $x_1 > 0$, $x_2 > 0$ and $\lambda_2 > 0$. It is also easy to see that $\lambda_1 > 0$ and that the determinant of X satisfies

$$\det(X) \;=\; \frac{\kappa^{2n}}{s^{2n}(\kappa)} \left(28\,\kappa^{3.5} - 65\,\kappa^3 + 56\,\kappa^{2.5} + 25\,\kappa^2 - 88\,\kappa^{1.5} + 70\,\kappa - 26\,\kappa^{0.5} + 4 \right)^n > 0, \; \forall\, \kappa \;\geq\; 1$$

which yields $X \succeq 0$. Moreover, it can be shown that the left-hand-side of LMI (2.24) becomes

$$\left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\lambda_1 I \end{array} \right] \;\preceq\; 0.$$

Therefore, the point $(\lambda_1, \lambda_2, X)$ is feasible to problem (2.26) and

$$J_{\rm na} \;\leq\; p(\kappa) \;:=\; n\,L\lambda_2 + n\,x_2 \;=\; \frac{n}{s(\kappa)} \left(4\,\kappa^{3.5} - 4\,\kappa^3 - 3\,\kappa^{2.5} + 9\,\kappa^2 - 4\,\kappa^{1.5} \right).$$

Comparing p with $q_{\rm na}$, it can be verified that, for all $\kappa \geq 1$, $4.08 q_{\rm na}(\kappa) \geq p(\kappa)$, which completes the proof.

A.2.4 Proof of Theorem 6

Without loss of generality, let $\sigma = 1$ and

$$G \;:=\; \sum_{i=1}^{n} \max\{ \hat{J}(\lambda_i), \hat{J}(\lambda_i') \} \tag{A.13}$$

where λ_i are the eigenvalues of the Hessian of the objective function f and $\lambda_i' = m + L - \lambda_i$ is the mirror image of λ_i with respect to $(m+L)/2$. Since $J = \sum_i \hat{J}(\lambda_i)$, if λ_i are symmetrically

distributed over the interval $[m, L]$ i.e., $(\lambda_1, \cdots, \lambda_n) = (\lambda'_n, \cdots, \lambda'_1)$, then for any parameters α and β we have

$$J \;\leq\; G \;\leq\; 2J. \tag{A.14}$$

Equation (A.14) implies that any bound on G simply carries over to J within an accuracy of constant factors. Thus, we focus on G and establish one of its useful properties in the next lemma that allows us to prove Theorem 6.

Lemma 2 *The heavy-ball method with any stabilizing parameter β satisfies*

$$\frac{2(1+\beta)}{L+m} \;=\; \operatorname*{argmin}_{\alpha}\, \rho(\alpha, \beta) \tag{A.15}$$

where ρ is the rate of linear convergence. Furthermore, if the Hessian of the quadratic objective function f has a symmetric spectrum over the interval $[\lambda_1, \lambda_n] = [m, L]$, then

$$\frac{2(1+\beta)}{L+m} \;=\; \operatorname*{argmin}_{\alpha}\, G(\alpha,\, \beta).$$

Proof: The linear convergence rate ρ is given by $\rho = \max_{1 \leq i \leq n} \hat{\rho}(\lambda_i)$, where $\hat{\rho}(\lambda)$ is the largest absolute value of the roots of the characteristic polynomial

$$\det(zI - \hat{A}) = z^2 + (\alpha\lambda - 1 - \beta)z + \beta$$

associated with the heavy-ball method and the eigenvalue λ of the Hessian of the objective function f; See (2.8) for the form of $\hat{A}$. Thus, we have

$$\hat{\rho}(\lambda) \;=\; \begin{cases} \sqrt{\beta} & \text{if } \Delta < 0 \\ \frac{1}{2}|1 + \beta - \alpha\lambda| + \frac{1}{2}\sqrt{\Delta} & \text{otherwise} \end{cases}$$

where $\Delta := (1 + \beta - \alpha\lambda)^2 - 4\beta$. This can be simplified to

$$\hat{\rho} \;=\; \begin{cases} \sqrt{\beta} & \text{if } (1 - \sqrt{\beta})^2 \leq \alpha\lambda \leq (1 + \sqrt{\beta})^2 \\ \frac{1}{2}|1 + \beta - \alpha\lambda| + \frac{1}{2}\sqrt{\Delta} & \text{otherwise.} \end{cases}$$

It is straightforward to show that $\hat{\rho}$ and $\hat{J}$ with $\sigma = 1$ are explicit quasi-convex functions of $\mu := \alpha\lambda$ which are symmetric with respect to $\mu = 1 + \beta$. Quasi-convexity of $\hat{\rho}$ yields

$$\rho \;=\; \max\{\hat{\rho}(\lambda_1), \hat{\rho}(\lambda_n)\} \;=\; \max\{\hat{\rho}(\lambda_1), \hat{\rho}(\lambda'_1)\}.$$

Let $\alpha^\sharp(\beta) = 2(1 + \beta)/(L + m)$. For any eigenvalue λ_i, from the symmetry of the spectrum, we have

$$\alpha^\sharp(\beta)\lambda_i \;-\; (1 + \beta) \;=\; (1 + \beta) \;-\; \alpha^\sharp(\beta)\lambda'_i$$

meaning that $\alpha^\sharp(\beta)\lambda_i$ and $\alpha^\sharp(\beta)\lambda'_i$ are the mirror images with respect to the middle point $1 + \beta$. Thus, from the quasi-convexity and symmetry of the functions $\hat{\rho}$ and $\hat{J}$, it follows

that $\alpha^\sharp(\beta)$ minimizes ρ as well as $\max\{\hat{J}(\lambda_i), \hat{J}(\lambda_i')\}$ for all i, which completes the proof. $\square$

Since gradient descent is obtained from the heavy-ball method by letting $\beta = 0$, from Lemma 2 it immediately follows that $\alpha_{\rm gd} = 2/(L+m)$ given in Table 2.2 optimizes both $G_{\rm gd}$ and the convergence rate $\rho_{\rm gd}$. This fact combined with (A.14) yields

$$2\, J_{\rm gd}(\alpha_{\rm gd}^\star(c)) \;\geq\; G_{\rm gd}(\alpha_{\rm gd}^\star(c)) \;\geq\; G_{\rm gd}(\alpha_{\rm gd}) \;\geq\; J_{\rm gd}(\alpha_{\rm gd}) \tag{A.16}$$

where $\alpha_{\rm gd}^\star(c)$ is given by (2.28b). This completes the proof for gradient descent.

We next use Lemma 2 to establish a bound on the parameter $\beta_{\rm hb}^\star(c)$ that allows us to prove the result for the heavy-ball method as well.

Lemma 3 *There exists a positive constant a such that*

$$\beta_{\rm hb}^\star(c) \;\geq\; 1 \,-\, \frac{a}{\sqrt{\kappa}} \tag{A.17}$$

where $\beta_{\rm hb}^\star(c)$ is given by (2.28a).

Proof: We first show that for any parameters α and β, the convergence rate ρ of the heavy-ball method given by (2.27) is lower bounded by

$$\rho \;\geq\; \begin{cases} \sqrt{\beta} & \text{if } \beta \geq (\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^2 \\ \frac{(1+\beta)(L-m)+\sqrt{(1+\beta)^2(L-m)^2-4\beta(L+m)^2}}{2(L+m)} & \text{otherwise.} \end{cases} \tag{A.18}$$

The convergence rate satisfies

$$\rho \;=\; \max_{1\,\leq\, i\,\leq\, n} \hat{\rho}(\lambda_i) \;=\; \max_{\lambda\,\in\,\{m,L\}} \hat{\rho}(\lambda)$$

where the function $\hat{\rho}(\lambda)$ is given by (see proof of Lemma 2 for the proof of this statement)

$$\hat{\rho}(\lambda) = \begin{cases} \sqrt{\beta} & \text{if } (1-\sqrt{\beta})^2 \leq \alpha\lambda \leq (1+\sqrt{\beta})^2 \\ \frac{1}{2}|1+\beta-\alpha\lambda| + \frac{1}{2}\sqrt{\Delta} & \text{otherwise} \end{cases}$$

and $\Delta := (1+\beta-\alpha\lambda)^2 - 4\beta$. According to Lemma 2, $\alpha = 2(1+\beta)/(L+m)$ optimizes the rate ρ. This value of α yields

$$\hat{\rho}(m) \;=\; \hat{\rho}(L) \;=\; \begin{cases} \sqrt{\beta} & \text{if } \kappa \leq \frac{(1+\sqrt{\beta})^2}{(1-\sqrt{\beta})^2} \\ \frac{1}{2}|1+\beta-\alpha^\star\lambda| + \frac{1}{2}\sqrt{\Delta}\Big|_{\lambda=m} & \text{otherwise} \end{cases}$$

or equivalently

$$\hat{\rho}(m) \;=\; \hat{\rho}(L) \;=\; \begin{cases} \sqrt{\beta} & \text{if } \beta \geq (\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^2 \\ \frac{(1+\beta)(L-m)+\sqrt{(1+\beta)^2(L-m)^2-4\beta(L+m)^2}}{2(L+m)} & \text{otherwise} \end{cases} \tag{A.19}$$

which completes the proof of inequality (A.18). Now, if $\beta \geq (\sqrt{\kappa}-1)^2/(\sqrt{\kappa}+1)^2$, then (A.17) with $a = 2$ follows immediately. Otherwise, from (A.18) we obtain

$$\rho \geq \frac{(1+\beta)(L-m) + \sqrt{(1+\beta)^2(L-m)^2 - 4\beta(L+m)^2}}{2(L+m)}$$

which yields

$$\beta \;\geq\; v(\rho) \;:=\; \rho\,(\tfrac{L-m}{L+m} \;-\; \rho)/(1 \;-\; \tfrac{L-m}{L+m}\,\rho). \tag{A.20}$$

The convergence rate ρ satisfies $(\sqrt{\kappa}-1)^2/(\sqrt{\kappa}+1)^2 \leq \rho \leq 1-c/\sqrt{\kappa}$, where the lower bound follows from the optimal rate provided in Table 2.2 and the upper bound follows from the definition in (2.28a). Moreover, the derivative $\frac{\mathrm{d}v}{\mathrm{d}\rho} = 0$ vanishes only at $\rho = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$. Thus, we obtain a lower bound on β as

$$\beta \;\geq\; v(\rho) \;\geq\; \min\{v((\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^2),\; v(1-c/\sqrt{\kappa}),\; v(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})\}. \tag{A.21}$$

A simple manipulation of (A.21) allows us to find a constant a that satisfies (A.17), which completes the proof. □

Let $(\hat{\alpha}, \hat{\beta})$ be the optimal solution of the optimization problem

$$\begin{array}{ll} \underset{\alpha,\,\beta}{\text{minimize}} & G(\alpha,\beta) \\ \text{subject to} & \rho \;\leq\; 1 \;-\; c/\sqrt{\kappa} \end{array}$$

where G is defined in (A.13). We next show that there exists a scalar $c' > 0$ such that

$$G(\hat{\alpha}, \hat{\beta}) \;\geq\; c' J(\alpha_{\mathrm{hb}}, \beta_{\mathrm{hb}}) \tag{A.22}$$

where α_{hb} and β_{hb} are provided in Table 2.2. Let $\hat{\alpha}(\beta) := 2(1+\beta)/(L+m)$. It is straightforward to verify that

$$J(\hat{\alpha}(\beta), \beta) \;=\; \frac{1-\beta_{\mathrm{hb}}^2}{1-\beta^2}\, J(\alpha_{\mathrm{hb}}, \beta_{\mathrm{hb}}) \tag{A.23}$$

which allows us to write

$$
\begin{aligned}
G(\hat{\alpha},\hat{\beta}) &\overset{\text{(i)}}{=} \min_{\beta} \;\; G(\hat{\alpha}(\beta),\beta) \\
&\quad \text{subject to} \quad \rho \leq 1 - c/\sqrt{\kappa} \\
&\overset{\text{(ii)}}{\geq} \min_{\beta} \;\; J(\hat{\alpha}(\beta),\beta) \\
&\quad \text{subject to} \quad \rho \leq 1 - c/\sqrt{\kappa} \\
&\overset{\text{(iii)}}{=} \min_{\beta} \;\; \frac{1-\beta_{\text{hb}}^2}{1-\beta^2} J(\alpha_{\text{hb}},\beta_{\text{hb}}) \\
&\quad \text{subject to} \quad \rho \leq 1 - c/\sqrt{\kappa} \\
&\overset{\text{(iv)}}{\geq} \frac{1-\beta_{\text{hb}}^2}{1-(1-\frac{a}{\sqrt{\kappa}})^2} \, J(\alpha_{\text{hb}},\beta_{\text{hb}}).
\end{aligned}
\tag{A.24}
$$

Here, (i) determines partial minimization with respect to α which follows from Lemma 2; (ii) follows from (A.14); (iii) follows from (A.23), and (iv) follows from Lemma 3. Furthermore, it is easy to show the existence of a constant scalar c' such that

$$
\frac{1-\beta_{\text{hb}}^2}{1-(1-\frac{a}{\sqrt{\kappa}})^2} \;\geq\; c'. \tag{A.25}
$$

Inequality (A.22) follows from combining (A.25) and (A.24). Finally, we obtain that

$$
J(\alpha_{\text{gd}}^\star,\beta_{\text{gd}}^\star) \;\geq\; \frac{1}{2} G(\alpha_{\text{gd}}^\star,\beta_{\text{gd}}^\star) \;\geq\; \frac{1}{2} G(\hat{\alpha},\hat{\beta}) \;\geq\; \frac{c'}{2} J(\alpha_{\text{gd}},\beta_{\text{gd}})
$$

where the first inequality follows from (A.14), the second follows from the definition of $(\hat{\alpha},\hat{\beta})$, and the last inequality is given by (A.22). This completes the proof for the heavy-ball method in Theorem 6.

A.3 Fundamental lower bounds

A.3.1 Proof of Theorem 7

We first prove (2.29a). Without loss of generality, let the noise magnitude $\sigma = 1$. We define the trivial lower bound

$$
J \;\geq\; \hat{J}^\star \;:=\; \max\{\hat{J}(m),\hat{J}(L)\} \tag{A.26}
$$

and show that $\dfrac{\hat{J}^\star}{1-\rho} \;\geq\; (\dfrac{\kappa+1}{8})^2$. Let $\tilde{f}(x_1,x_2) := \frac{1}{2}(m\,x_1^2 + L\,x_2^2)$. The eigenvalues of the Hessian matrix $\nabla^2\tilde{f}$ are given by m and L which are clearly symmetric over the interval

$[m, L]$. Thus, for any given value of β, m, and L, we can use Lemma 2 with the objective function $\tilde{f}$ to obtain

$$\hat{\alpha}(\beta) \;:=\; \frac{2(1+\beta)}{L+m} \;=\; \operatorname*{argmin}_{\alpha}\, \hat{J}^{\star}(\alpha,\, \beta) = \operatorname*{argmin}_{\alpha}\, \rho(\alpha,\, \beta).$$

For the stepsize $\hat{\alpha}(\beta)$, the rate of convergence ρ is given by (A.19), i.e.,

$$\rho \;=\; \begin{cases} \sqrt{\beta} & \text{if } \beta \geq (\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^2 \\ \frac{(1+\beta)(L-m)+\sqrt{(1+\beta)^2(L-m)^2-4\beta(L+m)^2}}{2(L+m)} & \text{otherwise} \end{cases} \tag{A.27}$$

and the lower bound $\hat{J}^{\star}$ is given by

$$\hat{J}^{\star} = \hat{J}(m) = \hat{J}(L) = \frac{(L+m)^2}{4\,L\,m(1-\beta^2)}. \tag{A.28}$$

Therefore, we obtain a lower bound on $\hat{J}^{\star}/(1-\rho)$ as

$$\begin{aligned} \frac{\hat{J}^{\star}(\alpha,\beta)}{1-\rho(\alpha,\beta)} \;\geq\; \nu(\beta) \;&:=\; \frac{\hat{J}^{\star}(\hat{\alpha}(\beta),\beta)}{1-\rho(\hat{\alpha}(\beta),\beta)} \\ &= \begin{cases} \frac{(L+m)^2}{4\,L\,m(1-\beta^2)(1-\sqrt{\beta})} & \text{if } \beta \geq (\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^2 \\ \frac{(L+m)^3}{2\,L\,m(1-\beta^2)\left((1-\beta)L+(3+\beta)m-\sqrt{(1+\beta)^2(L-m)^2-4\beta(L+m)^2}\right)} & \text{otherwise} \end{cases} \end{aligned} \tag{A.29}$$

where the last equality follows from (A.27) and (A.28). It can be shown that $v(\beta)$ attains its minimum at $\beta = (\sqrt{\kappa}-1)^2/(\sqrt{\kappa}+1)^2$; see Figure A.1 for an illustration. Therefore,

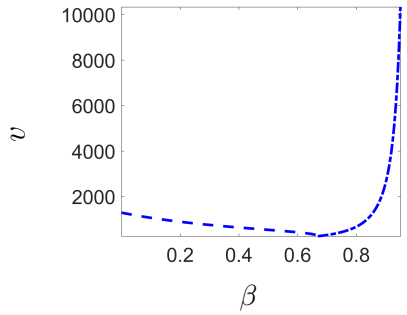


Figure A.1: The β-dependence of the function v in (A.29) for $L = 100$ and $m = 1$.

$$\begin{aligned} \nu(\beta) \;&\geq\; \frac{(L+m)^2}{4\,L\,m(1-\beta^2)(1-\sqrt{\beta})}\bigg|_{\beta=(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^2} \;=\; \frac{(L+m)^2}{4\,L\,m(1+\beta)(1+\sqrt{\beta})(1-\sqrt{\beta})^2}\bigg|_{\beta=(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^2} \\ &\geq\; \frac{(L+m)^2}{16\,L\,m(1-\sqrt{\beta})^2}\bigg|_{\beta=(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^2} \;=\; \frac{(\kappa+1)^2(\sqrt{\kappa}+1)^2}{64\kappa} \;\geq\; \left(\frac{\kappa+1}{8}\right)^2 \end{aligned}$$

which completes the proof of (2.29a). We next prove (2.29b) for $\sigma = \alpha$.

We analyze the two cases $\alpha > 1/L$ and $\alpha \leq 1/L$ separately. If $\alpha > 1/L$, inequality (2.29b) directly follows from inequality (2.29a)

$$\frac{J_{\text{hb}}}{1 - \rho} \;\geq\; \sigma^2 \left(\frac{\kappa+1}{8}\right)^2 \;=\; \alpha^2 \left(\frac{\kappa+1}{8}\right)^2 \;\geq\; \left(\frac{\kappa}{8L}\right)^2.$$

Here, the first inequality is given by (2.29a) and the second inequality holds since $\alpha > 1/L$.

Now suppose $\alpha \leq 1/L$. The convergence rate of Polyak's method is given by $\max_i \hat{\rho}(\lambda_i)$, where

$$\hat{\rho}(\lambda) \;=\; \begin{cases} \sqrt{\beta} & \text{if } (1-\sqrt{\beta})^2 \leq \alpha\lambda \leq (1+\sqrt{\beta})^2 \\ \frac{1}{2}|1+\beta-\alpha\lambda| + \frac{1}{2}\sqrt{\Delta} & \text{otherwise} \end{cases}$$

and $\Delta := (1+\beta-\alpha\lambda)^2 - 4\beta$ (see the proof of Lemma 2). Thus, for $\sigma = \alpha$, we have the trivial lower bound

$$\begin{aligned} \frac{J}{1-\rho} \;&\geq\; \frac{\hat{J}(m)}{1-\hat{\rho}(m)} \;=\; \frac{\alpha(1+\beta)}{m\,(1-\beta)\,(2(1+\beta)-\alpha m)\,(1-\hat{\rho}(m))} \\ &\geq\; p(\alpha,\beta) \;:=\; \frac{\alpha}{2m\,(1-\beta)\,(1-\hat{\rho}(m))} \\ &=\; \begin{cases} \dfrac{\alpha}{2m\,(1-\beta)\,\left(1-\sqrt{\beta}\right)}, & \beta \in [(1-\sqrt{\alpha m})^2,\; 1) \\ \dfrac{\alpha}{m\,(1-\beta)\left(1-\beta+\alpha m-\sqrt{\Delta}\right)}, & \beta \in [0,\; (1-\sqrt{\alpha m})^2). \end{cases} \end{aligned}$$

Here, the first inequality follows from combining $J = \sum_i \hat{J}(\lambda_i)$ and $\max_i \hat{\rho}(\lambda_i)$, and the second inequality follows from $\alpha m \leq \alpha L \leq 1$. We next show that for any fixed α, the function $p(\alpha,\cdot)$ attains its minimum at $\beta = (1-\sqrt{\alpha m})^2$. Before we do so, note that this fact allows us to use partial minimization with respect to β and obtain

$$p(\alpha,\beta) \;\geq\; p(\alpha,(1-\sqrt{\alpha m})^2) \;=\; \frac{1}{2m^2\,(2-\sqrt{\alpha m})} \;\geq\; \frac{1}{4m^2} \;\geq\; (\frac{\kappa}{2L})^2$$

which completes the proof of (2.29b).

For any fixed α, it is straightforward to verify that $p(\alpha,\beta)$ is increasing with respect to β over $[(1-\sqrt{\alpha m})^2,\, 1)$. Thus, it suffices to show that $p(\alpha,\beta)$ is decreasing with respect to β over $[0, (1-\sqrt{\alpha m})^2)$. To simplify the presentation, let us define the new set of parameters

$$\begin{aligned} q \;&:=\; s\,(s+x-\delta)\,, \quad s \;:=\; 1-\beta, \quad x \;:=\; \alpha m \\ \delta \;&:=\; \sqrt{\Delta} \;=\; \sqrt{(1+\beta-\alpha m)^2 - 4\beta} \;=\; \sqrt{(s+x)^2 - 4x}. \end{aligned}$$

It is now straightforward to verify that $p(\alpha,\beta) = \alpha/(mq)$ for $\beta \in [(1-\sqrt{\alpha m})^2,\ 1)$. It thus follows that $p(\alpha,\beta)$ is decreasing with respect to β over $[0,(1-\sqrt{\alpha m})^2)$ if and only if $q' = \mathrm{d}q/\mathrm{d}s \leq 0$ for $s \in (\sqrt{x}(2-\sqrt{x}),1]$. The derivative is given by

$$q' \ = \ \frac{1}{\delta}\left((2s+x)\delta - 2s^2 - 3sx - x^2 + 4x\right).$$

Thus, we have

$$q' \ \leq \ 0 \iff (2s+x)\delta \ \leq \ 2s^2 + 3sx + x^2 - 4x. \tag{A.30}$$

It is easy to verify that both sides of the inequality in (A.30), namely, $(2s+x)\delta$ and $2s^2 + 3sx + x^2 - 4x$ are positive for the specified range of $s \in (\sqrt{x}(2-\sqrt{x}),1]$. Thus, we can square both sides and obtain that

$$\begin{aligned} q' \ \leq \ 0 &\iff (2s+x)^2\delta^2 \ \leq \ (2s^2+3sx+x^2-4x)^2 \\ &\overset{(i)}{\iff} (2s+x)^2\left((s+x)^2-4x\right) \ \leq \ (2s^2+3sx+x^2-4x)^2 \\ &\overset{(ii)}{\iff} 8sx^2 + 4x^3 \ \leq \ 16x^2 \iff 8s + 4x \ \leq \ 16. \end{aligned}$$

where (i) follows from the definition of δ and (ii) is obtained by expanding both sides and rearranging the terms. Finally, the inequality $8s + 4x \ \leq \ 16$ clearly holds since $s \leq 1$ and $x \leq 1$. This proves that $p(\alpha,\cdot)$ attains its minimum at $\beta = (1-\sqrt{\alpha m})^2$.

A.3.2 Proof of Theorem 8

For the heavy-ball method, the result follows from combining Theorem 7 and the inequality $1-\rho > c/\sqrt{\kappa}$. Next, we present three additional lemmas that allow us to prove the result for Nesterov's method.

The following lemma provides a lower bound on the function $\hat{J}(m)$ associated with Nesterov's method which depends on κ and β.

Lemma 4 *For any strongly convex quadratic problem with condition number $\kappa > 2$ and the smallest eigenvalue of the Hessian m, the function $\hat{J}$ associated with Nesterov's accelerated method with any stabilizing pair of parameters $0 < \alpha$, $0 < \beta < 1$, and $\sigma = 1$ satisfies*

$$\hat{J}(m) \geq \frac{\kappa^2}{24(1-\beta)\kappa + 32\beta}. \tag{A.31}$$

Proof: We first show that Nesterov's method with $0 < \alpha$ and $0 < \beta < 1$ is stable if and only if

$$m < \frac{2\beta+2}{\alpha\,\kappa\,(2\beta+1)}. \tag{A.32}$$

The rate of linear convergence is given by $\rho = \max_{1\leq i\leq n}\ \hat{\rho}(\lambda_i)$, where $\hat{\rho}(\lambda)$ is the largest absolute value of the roots of the characteristic polynomial

$$\det(zI - \hat{A}) = z^2 - (1+\beta)(1-\alpha\lambda)z + \beta(1-\alpha\lambda)$$

associated with Nesterov's method and the eigenvalue λ of the Hessian of the objective function f; See (2.8) for the form of $\hat{A}$. For $\alpha > 0$ and $0 < \beta < 1$, it can be shown that

$$\hat{\rho}(\lambda) = \begin{cases} \sqrt{\beta(1-\alpha\lambda)} & \text{if} \quad \alpha\lambda \in ((\frac{1-\beta}{1+\beta})^2, 1) \\ \frac{1}{2}|(1+\beta)(1-\alpha\lambda)| + \frac{1}{2}\sqrt{(1+\beta)^2(1-\alpha\lambda)^2 - 4\beta(1-\alpha\lambda)} & \text{otherwise.} \end{cases} \tag{A.33}$$

The stability of the algorithm is equivalent to $\hat{\rho}(\lambda_i) < 1$ for all eigenvalues λ_i. For any positive stepsize α and parameter $\beta \in (0,1)$, it can be shown that the function $\hat{\rho}(\lambda)$ is quasi-convex and $\hat{\rho}(\lambda) = 1$ if and only if $\lambda \in \{0, \frac{2\beta+2}{\alpha(2\beta+1)}\}$. This fact along with $0 < m \leq \lambda_i \leq L = \kappa\, m$ imply that $\hat{\rho}(\lambda_i) < 1$ for all $\lambda_i \in [m, L]$ if and only if $\kappa\, m \leq \frac{2\beta+2}{\alpha(2\beta+1)}$ which completes the proof of (A.32).

For Nesterov's method, it is straightforward to show that the function $\hat{J}(\lambda)$ is quasi-convex over the interval $[0, \frac{2\beta+2}{\alpha(2\beta+1)}]$ and that it attains its minimum at $\lambda = 1/\alpha$. Also, from (A.32), for $\kappa > 2$ we obtain

$$m \ \leq\ \frac{2\beta+2}{\alpha\kappa(2\beta+1)} \ \leq\ \frac{1}{\alpha}$$

and thus,

$$\begin{aligned} \hat{J}(m) \ &\geq\ \hat{J}(\frac{2\beta+2}{\alpha\kappa(2\beta+1)}) \\ &=\ \frac{(2\,\beta+1)\,\kappa^2\ (\kappa - 2\,\beta + 2\,\beta\,\kappa)}{4\ (\beta+1)\ (\kappa-1)\ (2\,\beta+\kappa+\beta\,\kappa-2\,\beta^2\,\kappa+2\,\beta^2)} \ \geq\ \frac{\kappa^2}{24\,(1-\beta)\kappa + 32\beta} \end{aligned}$$

where the last inequality follows from the fact that $\beta \in (0,1)$. □

The next lemma presents a lower bound on any accelerating parameter β for Nesterov's method.

Lemma 5 *For Nesterov's method, under the conditions of Theorem 8, there exist positive constants c_3 and c_4 such that for any $\kappa > c_3$,*

$$\beta\ >\ 1\ -\ \frac{c_4}{\sqrt{\kappa}}. \tag{A.34}$$

Proof: For any $\alpha > 0$ and $\beta \in (0, 1)$, Nesterov's method converges with the rate $\rho = \max_{1 \leq i \leq n} \hat{\rho}(\lambda_i)$, where $\hat{\rho}(\lambda)$ is given by (A.33). We treat the two cases $(1-\beta)/(1+\beta)^2 < \alpha m$ and $(1-\beta)/(1+\beta)^2 \geq \alpha m$ separately. For $(1-\beta)/(1+\beta)^2 < \alpha m$, we have

$$(1-\beta)^2 \;\leq\; 4(\frac{1-\beta}{1+\beta})^2 \;<\; 4\alpha m \;=\; 4\frac{\alpha L}{\kappa} \;\leq\; \frac{8}{\kappa} \tag{A.35}$$

where the last inequality follows from (A.32). Therefore, we obtain $\beta \geq 1 - \sqrt{8}/\sqrt{\kappa}$ as required. Now, suppose $(1-\beta)/(1+\beta)^2 \geq \alpha m$. The convergence rate ρ satisfies

$$\rho \;\geq\; \frac{1}{2}(1+\beta)(1-\alpha m) \;+\; \frac{1}{2}\sqrt{(1+\beta)^2(1-\alpha m)^2 - 4\beta(1-\alpha m)}.$$

Thus,

$$\rho^2 - \rho(1+\beta)(1-\alpha m) + \beta(1-\alpha m) > 0$$

which yields a lower bound on β,

$$\beta \;\geq\; \nu(\rho, \alpha m) \;:=\; \frac{\rho(1-\alpha m-\rho)}{(1-\rho)(1-\alpha m)}. \tag{A.36}$$

In what follows, we establish a lower bound for ν. For a fixed αm, the critical point of $\nu(\rho)$ is given by $\rho_1 := 1 - \sqrt{\alpha m}$, i.e., $\partial\nu/\partial\rho = 0$ for $\rho = \rho_1$. Furthermore, the optimal rate from Table 2.2 and the condition on convergence rate in Theorem 8 for any $\kappa > c_1$ yield upper and lower bounds $\rho_3 < \rho < \rho_2$, where $\rho_2 := 1 - c_2/\sqrt{\kappa}$ and $\rho_3 := 1 - 2/\sqrt{3\kappa+1}$. Thus, the lower bound on ν is given by

$$\beta \;\geq\; \nu(\rho, \alpha m) \;\geq\; \min\{\nu(\rho_1, \alpha m), \nu(\rho_2, \alpha m), \nu(\rho_3, \alpha m)\}. \tag{A.37}$$

From the stability condition (A.32), we have

$$\alpha m \;<\; 2/\kappa \tag{A.38}$$

Furthermore, it can be shown that for any given $\rho \in (0, 1)$ the function $\nu(\rho, \alpha m)$ is decreasing with respect to αm. This fact combined with (A.37) and (A.38) yield

$$\beta \;\geq\; \min\{\nu(\rho_1, \alpha m), \nu(\rho_2, 2/\kappa), \nu(\rho_3, 2/\kappa)\}. \tag{A.39}$$

If we substitute for ρ_1. ρ_2, and ρ_3 their values as functions of κ and use $\alpha m < 2/\kappa$, then the result follows immediately. In particular,

$$\begin{aligned}
\nu(\rho_1, \alpha m) &= \frac{1-\sqrt{\alpha m}}{1+\sqrt{\alpha m}} \geq \frac{1-\sqrt{2/\kappa}}{1+\sqrt{2/\kappa}} = \frac{\sqrt{\kappa}-\sqrt{2}}{\sqrt{\kappa}+\sqrt{2}} \geq 1 - \frac{2\sqrt{2}}{\sqrt{\kappa}} \\
\nu(\rho_2, 2/\kappa) &= 1 - \frac{(\frac{2}{c_2}+c_2)\sqrt{\kappa}-4}{\kappa-2} \geq 1 - \frac{(\frac{2}{c_2}+c_2)}{\sqrt{\kappa}}, \quad \forall \kappa \geq (\frac{1}{c_2}+\frac{c_2}{2})^2 \\
\nu(\rho_3, 2/\kappa) &= 1 - \frac{5\kappa - 4\sqrt{3\kappa+1}+1}{(\kappa-2)\sqrt{3\kappa+1}} \geq 1 - \frac{5}{\sqrt{\kappa}}, \quad \forall \kappa \geq 9
\end{aligned}$$

which completes the proof. □

The next lemma provides a lower bound on $J_{\rm na}/(1-\rho)$ for Nesterov's method with $\sigma = \alpha \leq 1/L$.

Lemma 6 *Nesterov's accelerated method with any stabilizing pair of parameters $0 < \alpha \leq 1/L$ and $0 < \beta < 1$, and $\sigma = \alpha$ satisfies*

$$\frac{J_{\rm na}}{1-\rho} \geq \frac{1}{8}(\frac{\kappa}{L})^2.$$

Proof: The convergence rate of Nesterov's method is given by $\max_i \hat{\rho}(\lambda_i)$, where

$$\hat{\rho}(\lambda) = \begin{cases} \sqrt{\beta(1-\alpha\lambda)} & \text{if} \quad \alpha\lambda \in ((\frac{1-\beta}{1+\beta})^2, 1) \\ \frac{1}{2}|(1+\beta)(1-\alpha\lambda)| + \frac{1}{2}\sqrt{\Delta} & \text{otherwise} \end{cases}$$

and $\Delta := (1+\beta)^2(1-\alpha\lambda)^2 - 4\beta(1-\alpha\lambda)$; see equation (A.33). Thus, we have the trivial lower bound

$$\begin{aligned}
\frac{J}{1-\rho} &\geq \frac{\hat{J}(m)}{1-\hat{\rho}(m)} = \frac{\alpha\,(1+\beta(1-\alpha m))}{m\,(1-\beta(1-\alpha m))\,(2(1+\beta)-(2\beta+1)\alpha m)\,(1-\hat{\rho}(m))} \\
&\geq p(\alpha,\beta) := \frac{\alpha}{4m\,(1-\beta(1-\alpha m))\,(1-\hat{\rho}(m))} \\
&= \begin{cases} \dfrac{\alpha}{4m\,(1-\beta(1-\alpha m))\left(1-\sqrt{\beta(1-\alpha m)}\right)}, & \beta \in [\gamma,\ 1) \\ \dfrac{\alpha}{2m\,(1-\beta(1-\alpha m))\left(2-(1+\beta)(1-\alpha m)-\sqrt{\Delta}\right)}, & \beta \in [0,\ \gamma) \end{cases}
\end{aligned} \tag{A.40}$$

where $\gamma := \dfrac{1-\sqrt{\alpha m}}{1+\sqrt{\alpha m}}$. Here, the first inequality can be obtained by combining $J = \sum_i \hat{J}(\lambda_i)$ and $\max_i \hat{\rho}(\lambda_i)$, and the second inequality follows from the fact that $0 < \alpha m \leq 1$ and $0 \leq \beta < 1$. We next show that for any fixed α, the function $p(\alpha, \cdot)$ attains its minimum at

$\beta = \gamma$. Before we do so, note that this fact allows us to do partial minimization with respect to β and obtain

$$p(\alpha, \beta) \;\geq\; p(\alpha, \gamma) \;=\; \frac{1}{4m^2\,(2 - \sqrt{\alpha m})} \;\geq\; \frac{1}{8m^2} \;\geq\; \frac{1}{8}(\frac{\kappa}{L})^2.$$

For any fixed α, it is straightforward to verify that $p(\alpha, \beta)$ is increasing with respect to β over $[\gamma,\ 1)$. Thus, it suffices to show that $p(\alpha, \beta)$ is decreasing with respect to β over $[0, \gamma)$. To simplify the presentation, let us define

$$\begin{aligned} q \;&:=\; (1-s)(2-x-s-\delta), \quad x \;:=\; 1-\alpha m, \quad s \;:=\; \beta x \\ \delta \;&:=\; \sqrt{\Delta} \;=\; \sqrt{(1+\beta)^2(1-\alpha m)^2 - 4\beta(1-\alpha m)} \;=\; \sqrt{(x+s)^2 - 4s}. \end{aligned}$$

It is now straightforward to verify that $p(\alpha, \beta) = \alpha/(2mq)$ for $\beta \in [0, \gamma)$. It thus follows that $p(\alpha, \beta)$ is decreasing with respect to β over $[0, \gamma)$ if and only if $q' = \mathrm{d}q/\mathrm{d}s \geq 0$ for $s \in [0, (1 - \sqrt{1-x})^2)$. The derivative is given by

$$q' \;=\; \frac{1}{\delta}\left((x + 2s - 3)\delta + (1-s)(2-x-s) + \delta^2\right).$$

Thus, we have

$$q' \;\geq\; 0 \;\iff\; (1-s)(2-x-s) + \delta^2 \;\geq\; (3-x-2s)\delta. \tag{A.41}$$

It is easy to verify that both sides of the inequality in (A.41), namely, $(1-s)(2-x-s)+\delta^2$ and $(3-x-2s)\delta$ are positive for the specified range of $s \in [0, (1 - \sqrt{1-x})^2)$. Thus, we can square both sides and obtain that

$$\begin{aligned} q' \;\geq\; 0 \;&\iff\; \left((1-s)(2-x-s) + \delta^2\right)^2 \;\geq\; (3-x-2s)^2\delta^2 \\ &\overset{(i)}{\iff}\; \left((1-s)(2-x-s) + (x+s)^2 - 4s\right)^2 \;\geq\; (3-x-2s)^2\left((x+s)^2 - 4s\right) \\ &\overset{(ii)}{\iff}\; 4(x-1)^2(2s+x+1) \;\geq\; 0. \end{aligned}$$

where (i) follows from the definition of δ and (ii) is obtained by expanding both sides and rearranging the terms. Finally, the inequality $4(x-1)^2(2s+x+1) \;\geq\; 0$ trivially holds which completes the proof. □

We are now ready to prove Theorem 8 for Nesterov's method. The inequality in (2.30a) directly follows from combining (A.31) in Lemma 4 and (A.34) in Lemma 5. To show inequality (2.30b), we treat the two cases $\alpha > 1/L$ and $\alpha \leq 1/L$ separately. If $\alpha > 1/L$, then (2.30b) directly follows from (2.30a)

$$J_{\mathrm{na}} \;=\; \alpha^2 \frac{J_{\mathrm{na}}}{\sigma^2} \;=\; \Omega(\frac{\kappa^{\frac{3}{2}}}{L^2}).$$

Now suppose $\alpha \leq 1/L$. We can use Lemma 6 to obtain

$$J_{\text{na}} \;\geq\; (1-\rho)\frac{k^2}{8L^2} \;\geq\; \frac{c}{\sqrt{\kappa}}\frac{k^2}{8L^2} \;=\; \Omega(\frac{\kappa^{\frac{3}{2}}}{L^2}).$$

Here, the first inequality follows from Lemma 6 and the second inequality follows from the acceleration assumption $\rho \leq 1 - c/\sqrt{\kappa}$. This completes the proof.

A.4 Consensus over d-dimensional torus networks

The proof of Theorem 9 uses the explicit expression for the eigenvalues of torus in (2.33) to compute the variance amplification $\bar{J} = \sum_{i \neq 0} \hat{J}(\lambda_i)$ for all three algorithms. Several technical results that we use in the proof are presented next.

We borrow the following lemma, which provides tight bounds on the sum of reciprocals of the eigenvalues of a d-dimensional torus network, from [43, Appendix B].

Lemma 7 *The eigenvalues λ_i of the graph Laplacian of the d-dimensional torus $\mathbb{T}^d_{n_0}$ with $n_0 \gg 1$ satisfy*

$$\sum_{0 \neq i \in \mathbb{Z}^d_{n_0}} \frac{1}{\lambda_i} \;=\; \Theta(B(n_0))$$

where the function B is given by

$$B(n_0) \;=\; \begin{cases} \dfrac{1}{d-2}\,(n_0^d - n_0^2), & d \neq 2 \\ n_0^d \log\, n_0, & d = 2. \end{cases}$$

We next use Lemma 7 to establish an asymptotic expression for the variance amplification of the gradient descent algorithm for a d-dimensional torus.

Lemma 8 *For the consensus problem over a d-dimensional torus $\mathbb{T}^d_{n_0}$ with $n_0 \gg 1$, the performance metric $\bar{J}_{\text{gd}}$ corresponding to gradient decent with the stepsize $\alpha = 2/(L+m)$ satisfies*

$$\bar{J}_{\text{gd}} \;=\; \Theta(B(n_0))$$

where the function B is given in Lemma 7.

Proof: Using the expression for the noise amplification of gradient descent from Theorem 1, we have

$$\begin{aligned}
\bar{J}_{\mathrm{gd}} &= \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\alpha\lambda_i(2 - \alpha\lambda_i)} \\
&= \frac{1}{2\alpha} \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i} + \frac{1}{\frac{2}{\alpha} - \lambda_i} \\
&= \frac{1}{2\alpha} \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i} + \frac{1}{\lambda_{\max} + \lambda_{\min} - \lambda_i} \\
&\approx \frac{1}{\alpha} \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i} \approx 2d \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i}.
\end{aligned}$$

The first approximation follows from the facts that the eigenvalues satisfy

$$0 < \lambda_i \leq \lambda_{\max} + \lambda_{\min} \approx 4d$$

and that their distribution is asymptotically symmetric with respect to $\lambda = 2d$. The second approximation follows from

$$\alpha = \frac{2}{L + m} = \frac{2}{\lambda_{\max} + \lambda_{\min}} \approx \frac{1}{2d}.$$

The bounds for the sum of reciprocals of λ_i provided in Lemma 7 can now be used to complete the proof. □

The following lemma establishes a relationship between the variance amplifications of Nesterov's method and gradient descent. This relationship allows us to compute tight bounds on J_{na} by splitting it into the sum of two terms. The first term depends linearly on J_{gd} which is already computed in Lemma 8 and the second term can be evaluated separately using integral approximations for consensus problem on torus networks. This result holds in general for the scenarios in which the largest eigenvalue $L = \Theta(1)$ is bounded and the smallest eigenvalue m goes to zero causing the condition number κ to go to infinity.

Lemma 9 *For a strongly convex quadratic problem with $mI \preceq Q \preceq LI$ and condition number $\kappa := L/m \geq \kappa_0$, the ratio between variance amplifications of Nesterov's algorithm and gradient descent with the parameters given in Table 2.2 satisfies the asymptotic bounds*

$$\frac{c_1}{\sqrt{\kappa}} \leq \frac{J_{\mathrm{na}} - D}{J_{\mathrm{gd}}} \leq c_2, \qquad D := \frac{2}{(3\beta + 1)\,\alpha_{\mathrm{na}}^2} \sum_{i=1}^{n} \frac{1}{\lambda_i^2 + \frac{1-\beta}{\alpha_{\mathrm{na}}\beta}\lambda_i}$$

where κ_0, c_1, and c_2 are positive constants. Furthermore, depending on the distribution of the eigenvalues of the Laplacian matrix, D can take values between

$$\frac{c_3}{\kappa} \;\leq\; \frac{D}{J_{\rm gd}} \;\leq\; c_4\sqrt{\kappa} \tag{A.42}$$

where c_3 and c_4 are positive constants.

Proof: We can split $\hat{J}_{\rm na}(\lambda)/\hat{J}_{\rm gd}(\lambda)$ into the sum of two decreasing homographic functions $\sigma_1(\lambda) + \sigma_2(\lambda)$, where σ_1 and σ_2 are defined in (A.2); see the proof of Proposition 1. Furthermore, for $\kappa \gg 1$, these functions attain their extrema over the interval $[m, L]$ at

$$\sigma_1(L) \;\approx\; \frac{9}{8\kappa}\,,\; \sigma_1(m) \;\approx\; \frac{3\sqrt{3\kappa}}{8}\,,\quad \sigma_2(L) \;\approx\; \frac{9\sqrt{3}}{16\sqrt{\kappa}}\,,\; \sigma_2(m) \;\approx\; \frac{3}{8} \tag{A.43}$$

where we have kept the leading terms. It is straightforward to verify that

$$\sum_{i=1}^{n} \sigma_1(\lambda_i)\hat{J}_{\rm gd}(\lambda_i) \;=\; \tfrac{2}{(3\beta+1)\alpha_{\rm na}^2}\sum_{i=1}^{n}\tfrac{1}{\lambda_i^2 + \frac{1-\beta}{\alpha_{\rm na}\beta}\lambda_i} \;=\; D.$$

This equation in conjunction with (A.43), yield inequalities in (A.42). Moreover, we obtain that

$$\frac{J_{\rm na} - D}{J_{\rm gd}} \;=\; \frac{\sum_{i=1}^n \sigma_2(\lambda_i)\hat{J}_{\rm gd}(\lambda_i)}{\sum_{i=1}^n \hat{J}_{\rm gd}(\lambda_i)}.$$

This also implies that, asymptotically,

$$\frac{J_{\rm na} - D}{J_{\rm gd}} \;=\; O\left(\max_{\lambda\in[m,L]} \sigma_2(\lambda)\right) \;=\; O(1)$$
$$\frac{J_{\rm na} - D}{J_{\rm gd}} \;=\; \Omega\left(\min_{\lambda\in[m,L]} \sigma_2(\lambda)\right) \;=\; \Omega(\frac{1}{\sqrt{\kappa}})$$

which completes the proof. □

The next two lemmas provide us with asymptotic bounds on summations of the form $\sum_i 1/(\lambda_i^2 + \mu\lambda_i)$, where λ_i are the eigenvalues of the graph Laplacian matrix of a torus network. These bounds allow us to combine Lemma 8 and Lemma 9 to evaluate the variance amplification of Nesterov's accelerated algorithm.

Lemma 10 *For an integer $q \gg 1$ and any positive $a = O(q^3)$, we have*

$$\sum_{0\neq i\in\mathbb{Z}_q^d} \frac{1}{\|i\|^4 + a\|i\|^2} \;\approx\; q^{d-4}\int_{1/q}^{1} \frac{r^{d-1}}{r^4 + wr^2}\,{\rm d}r$$

where $\omega = a/q^2$.

Proof: The function $h(x) := \|x\|^4 + \omega\|x\|^2$ is strictly increasing over the positive orthant ($x \succ 0$) and $h((1/q)\mathbb{1})$ goes to 0 as q goes to infinity where $\mathbb{1} \in \mathbb{R}^d$ is the vector of all ones. Therefore, using the lower and upper Riemann sum approximations, it is straightforward to show that

$$\int \cdots \int_{\Delta \le \|x\| \le 1} \frac{1}{h(x)} \mathrm{d}x_1 \cdots \mathrm{d}x_d \approx \Delta^d \sum_{0 \neq i \in \mathbb{Z}_q^d} \frac{1}{\left(\sum_{l=1}^d (\Delta i_l)^2\right)^2 + \omega \sum_{l=1}^d (\Delta i_l)^2}$$

where $\Delta = 1/q$ is the incremental step in the Riemann approximation. Therefore, since $\omega = a\Delta^2$, we can write

$$\sum_{0 \neq i \in \mathbb{Z}_q^d} \frac{1}{\|i\|^4 + a\|i\|^2} \approx \Delta^{4-d} \int \cdots \int_{\Delta \le \|x\| \le 1} \frac{1}{h(x)} \mathrm{d}x_1 \cdots \mathrm{d}x_d.$$

Finally, we obtain the result by transforming the integral into a d-dimensional system with polar coordinates, i.e.,

$$\int \cdots \int_{\Delta \le \|x\| \le 1} \frac{1}{h(x)} \mathrm{d}x_1 \cdots \mathrm{d}x_d \;\approx\; \int_{\Delta}^1 \frac{r^{d-1}}{r^4 + \omega r^2} \, \mathrm{d}r.$$

□

Lemma 11 *Let λ_i be the eigenvalues of the Laplacian matrix for the d-dimensional torus $\mathbb{T}_{n_0}^d$. In the limit of large n_0, for any $\mu = O(n_0)$, we have*

$$\sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i^2 + \mu\lambda_i} \;=\; \Theta\left(n_0^d \int_{\frac{1}{n_0}}^1 \frac{r^{d-1}}{r^4 + \omega r^2} \, \mathrm{d}r \right) \tag{A.44}$$

where $\omega = \Theta(\mu)$.

Proof: Let $\zeta := \sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i^2 + \mu\lambda_i}$, where λ_i are the eigenvalues of the Laplacian matrix,

$$\lambda_i \;=\; 2 \sum_{l=1}^d \left(1 - \cos(i_l \frac{2\pi}{n_0}) \right).$$

Since $1 - \cos(\cdot - \pi)$ is an even function, for large n_0,

$$\zeta \;\approx\; 2^d \sum_{0 \neq i \in \mathbb{Z}_q^d} \frac{1}{\lambda_i^2 + \mu\lambda_i}$$

where $q = \lfloor n_0/2 \rfloor$. It is well-known that the function $1-\cos(x)$ can be bounded by quadratic functions as $x^2/\pi^2 \leq 1-\cos(x) \leq x^2$ for any $x \in [-\pi, \pi]$. Now, since for any $i \in \mathbb{Z}_q^d$, $i_l \frac{2\pi}{n_0} \in [0, \pi]$ for all l, we can use these quadratic bounds to obtain

$$\zeta \approx n_0^4 \sum_{0 \neq i \in \mathbb{Z}_q^d} \frac{1}{\|i\|^4 + c\mu n_0^2 \|i\|^2} \tag{A.45}$$

where c is a bounded constant. Finally, equation (A.44) follows from Lemma 10 where we let $a = c\mu n_0^2$ and $q \approx n_0/2$. □

The following proposition characterizes the network-size-normalized asymptotic variance amplification of noisy consensus algorithms for d-dimensional torus networks. This result is used to prove Theorem 9.

Proposition 1 *Let $\mathbf{L} \in \mathbb{R}^{n \times n}$ be the graph Laplacian of the d-dimensional undirected torus $\mathbb{T}_{n_0}^d$ with $n = n_0^d \gg 1$ nodes. For convex quadratic optimization problem (2.31), the network-size-normalized asymptotic variance amplification $\bar{J}/n$ of the first-order algorithms on the subspace $\mathbb{1}^\perp$ is determined by*

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
Gradient	$\Theta(n)$	$\Theta(\log\, n)$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
Nesterov	$\Theta(n^2)$	$\Theta(\sqrt{n}\log\, n)$	$\Theta(n^{1/6})$	$\Theta(\log\, n)$	$\Theta(1)$
Polyak	$\Theta(n^2)$	$\Theta(\sqrt{n}\, \log\, n)$	$\Theta(n^{1/3})$	$\Theta(n^{1/4})$	$\Theta(n^{1/5})$.

Proof: We prove the result for the three algorithms separately.

1. For gradient descent, the result follows from dividing the asymptotic bounds established in Lemma 8 with the total number of nodes $n = n_0^d$.

2. For Nesterov's algorithm, we use the relation established in Lemma 9 to write

$$\bar{J}_{\mathrm{na}}/n - \frac{c}{n}\sum_i \frac{1}{\lambda_i^2 + \mu\lambda_i} = O\left(\bar{J}_{\mathrm{gd}}/n\right) \tag{A.46a}$$

$$\bar{J}_{\mathrm{na}}/n - \frac{c}{n}\sum_i \frac{1}{\lambda_i^2 + \mu\lambda_i} = \Omega\left(\bar{J}_{\mathrm{gd}}/(n\sqrt{\kappa})\right) \tag{A.46b}$$

where $c = 2/\left((3\beta+1)\alpha_{\mathrm{na}}^2\right) \approx 9d^2/2$ and $\mu = (1-\beta)/(\alpha_{\mathrm{na}}\beta) = \Theta(1/\sqrt{\kappa}) = \Theta(n_0^{-1})$; see equation (2.34). We can use Lemma 11 to compute the second term

$$\frac{1}{n}\sum_{0 \neq i \in \mathbb{Z}_{n_0}^d} \frac{1}{\lambda_i^2 + \mu\lambda_i} = \Theta\left(\int_{\frac{1}{n_0}}^1 \frac{r^{d-1}}{r^4 + \omega r^2}\, \mathrm{d}r\right) \tag{A.47}$$

where $\omega = \Theta(\mu) = \Theta(n_0^{-1})$. Evaluating the above integral for different values of $d \in \mathbb{N}$ and letting $\omega = \Theta(n_0^{-1})$, it is straightforward to show that

$$\int_{\frac{1}{n_0}}^{1} \frac{r^{d-1}}{r^4 + \omega r^2} \, \mathrm{d}r = \begin{cases} \Theta(n_0^2) & d = 1 \\ \Theta(n_0 \log n_0) & d = 2 \\ \Theta(\sqrt{n_0}) & d = 3 \\ \Theta(\log n_0) & d = 4 \\ \Theta(1) & d = 5. \end{cases}$$

Finally, the result follows from the asymptotic values for $\bar{J}_{\mathrm{gd}}/n$ (shown in Part 1) and substituting for the second term on the left-hand-side of equation (A.46) from the above asymptotic values and using $n = n_0^d$. We note that we used the following integrals to evaluate $\bar{J}_{\mathrm{na}}$,

$$\begin{aligned} \int \frac{1}{r^4 + \omega r^2} \, \mathrm{d}r &= -\frac{\tan^{-1}(\frac{r}{\sqrt{\omega}})}{\omega^{3/2}} - \frac{1}{r\omega} \\ \int \frac{r}{r^4 + \omega r^2} \, \mathrm{d}r &= -\frac{\log(r^2 + \omega) - 2\log(r)}{2\omega} \\ \int \frac{r^2}{r^4 + \omega r^2} \, \mathrm{d}r &= \frac{\tan^{-1}(\frac{r}{\sqrt{\omega}})}{\sqrt{\omega}} \\ \int \frac{r^3}{r^4 + \omega r^2} \, \mathrm{d}r &= \tfrac{1}{2}\log(r^2 + \omega) \\ \int \frac{r^4}{r^4 + \omega r^2} \, \mathrm{d}r &= r - \sqrt{\omega}\tan^{-1}(\frac{r}{\sqrt{\omega}}). \end{aligned}$$

3. The result for the heavy-ball method directly follows from the first part of the proof, the relationship between variance amplifications of gradient descent and the heavy-ball method in Theorem 2, and equation (2.34).

□

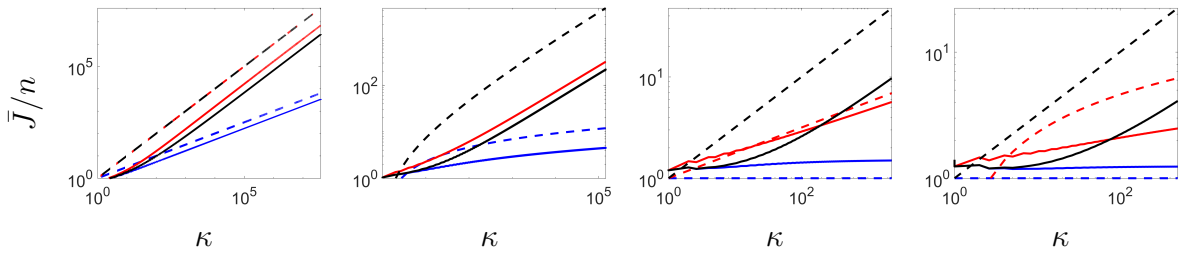
We now use Proposition 1 to proof Theorem 9 as follows.

A.4.1 Proof of Theorem 9

As stated in (2.34), the condition number satisfies $\kappa = \Theta(n^{2/d})$ and the result follows from combining this asymptotic relation with those provided in Proposition 1.

A.4.2 Computational experiments

To complement our asymptotic theoretical results, we compute the performance measure $\bar{J}$ in (2.32) for the consensus problem over d-dimensional torus $\mathbb{T}^d_{n_0}$ with $n = n_0^d$ nodes for different values of n_0 and d. We use expression (2.33) for the eigenvalues of the graph Laplacian



(a) $d = 1$ (b) $d = 2$ (c) $d = 3$ (d) $d = 4$

Figure A.2: The dependence of the network-size normalized performance measure $\bar{J}/n$ of the first-order algorithms for d-dimensional torus $\mathbb{T}^d_{n_0}$ with $n = n_0^d$ nodes on condition number κ. The blue, red, and black curves correspond to the gradient descent, Nesterov's method, and the heavy-ball method, respectively. Solid curves mark the actual values of $\bar{J}/n$ obtained using the expressions in Theorem 1 and the dashed curves mark the trends established in Theorem 9.

L to evaluate the formulae provided in Theorem 1 for each algorithm. Figure A.2 illustrates network-size normalized variance amplification $\bar{J}/n$ vs. condition number κ and verifies the asymptotic relations provided in Theorem 9. It is noteworthy that, *even though our analysis is asymptotic in the condition number (i.e., it assumes that $\kappa \gg 1$), our computational experiments exhibit similar scaling trends for small values of κ as well.*

Appendix B

Supporting proofs for Chapter 3

B.1 Settling time

If ρ denotes the linear convergence rate, $T_s = 1/(1-\rho)$ quantifies the *settling time*. The inequality in (3.5) shows that $c\rho^t \leq \epsilon$ provides a sufficient condition for reaching the accuracy level ϵ with $\|\psi^t\|_2/\|\psi^0\|_2 \leq \epsilon$. Taking the logarithm of $c\rho^t \leq \epsilon$ and using the first-order Taylor series approximation $\log(1-x) \approx -x$ around $x = 0$ yields a sufficient condition on the number of iterations t for an algorithm to reach ϵ-accuracy,

$$t \;\geq\; \log(\epsilon/c)/\log(1-1/T_s) \;\approx\; T_s \log(c/\epsilon).$$

In continuous time, the sufficient condition for reaching ϵ-accuracy $c\mathrm{e}^{-\rho t} \leq \epsilon$ yields $t \geq \log(c/\epsilon)/\rho$, and $T_s = 1/\rho$ can be used to asses the settling time.

B.2 Convexity of modal contribution $\hat{J}$

To show the convexity of $\hat{J}$, we use the fact that the function $g(x) = \prod_{i=1}^{d} x_i^{-1}$ is convex over the positive orthant $\mathbb{R}^d_{++}$. This can be verified by noting that its Hessian satisfies

$$\nabla^2 g(x) \;=\; g(x)\left(\mathrm{diag}(x) \,+\, xx^T\right) \succ 0$$

where $\mathrm{diag}(\cdot)$ is the diagonal matrix. By Theorem 5, we have

$$\frac{\hat{J}}{\sigma_w^2} \;=\; \frac{d\,+\,l}{2\,d\,h\,l} \;=\; \frac{1}{2\,h\,d} + \frac{1}{2\,h\,l}$$

where we have dropped the dependence on λ for simplicity. The functions $1/(2hd)$ and $1/(2hl)$ are both convex over the positive orthant $d, h, l > 0$. Thus, $\hat{J}$ is convex with respect to (d, h, l). In addition, since d, h, and l are all affine functions of a and b, we can use the equivalence relation in (3.30a) to conclude that $\hat{J}$ is also convex in (b, a) over the stability triangle Δ. Finally, since $b(\lambda)$ and $a(\lambda)$ are affine in λ, it follows that for any stabilizing parameters, $\hat{J}$ is also convex with respect to λ over the interval $[m, L]$.

Convexity of $\hat{J}$ allows us to use first-order conditions to find its minimizer. In particular, since for $\sigma_w = 1$

$$\frac{\partial \hat{J}}{\partial d} = -\frac{1}{2hd^2}, \quad \frac{\partial \hat{J}}{\partial l} = -\frac{1}{2hl^2}, \quad \frac{\partial \hat{J}}{\partial h} = -\frac{l+d}{2h^2dl}$$
$$\frac{\partial d}{\partial a} = \frac{\partial l}{\partial a} = -\frac{\partial h}{\partial a} = \frac{\partial d}{\partial b} = -\frac{\partial l}{\partial b} = 1, \quad \frac{\partial h}{\partial b} = 0$$

it is easy to verify that $\partial\hat{J}/\partial a = \partial\hat{J}/\partial b = 0$ at $a = b = 0$. Thus, $\hat{J}$ takes its minimum $\hat{J}_{\min} = \sigma_w^2$ over the stability triangle Δ at $a = b = 0$, which corresponds to $d = h = l = 1$.

B.3 Proofs of Section 3.4

B.3.1 Proof of Lemma 2

We start by noting that $\rho(M) \leq \rho$ if and only if $\rho(M') \leq 1$ where $M' := M/\rho$. The characteristic polynomial associated with M', $F_\rho(z) = z^2 + (b/\rho)z + a/\rho^2$, allows us to use similar arguments to those presented in the proof of Lemma 1 to show that

$$\rho(M') \leq 1 \quad \iff \quad (b/\rho,\, a/\rho^2) \in \Delta_1 \tag{B.1}$$

where $\Delta_1 := \{(b,a) \mid |b| - 1 \leq a \leq 1\}$ is the closure of the set Δ in (3.21b). Finally, the condition on the right-hand side of (B.1) is equivalent to $(b,a) \in \Delta_\rho$, where Δ_ρ is given by (3.22b).

Remark 1 *The eigenvalues of the matrix M in* (3.20a) *are given by $(-b \pm \sqrt{b^2-4a})/2$, and the sign of $b^2 - 4a$ determines if the eigenvalues are real or complex. The condition $a = b^2/4$ defines a parabola that passes through the vertices $X_\rho = (-2\rho, \rho^2)$ and $Y_\rho = (2\rho, \rho^2)$ of the triangle Δ_ρ and is tangent to the edges $X_\rho Z_\rho$ and $Y_\rho Z_\rho$ for all $\rho < 1$; see Figure B.1. For the optimal values of parameters provided in Table 3.1, we can combine this observation and the information in Figure 3.3 to conclude that while all eigenvalues of the matrix A in* (3.4a) *are real for gradient descent, they can be both real and complex for Nesterov's accelerated algorithm, and they come in complex-conjugate pairs for heavy-ball method.*

B.3.2 Proof of Equation (3.28c)

According to Figure 3.3, in order to find the largest ratio $d(L)/d(m)$ over the ρ-linear convergence set Δ_ρ for Nesterov's accelerated method, we need to check the pairs of points $\{E, E'\}$ that lie on the boundary of the triangle Δ_ρ, whose line segment EE' passes through the origin O. If one of the end points E lies on the edge $X_\rho Y_\rho$, then depending on whether the other end point E' lies on the edge $X_\rho Z_\rho$ or $Y_\rho Z_\rho$, we can continuously increase the ratio $d_E/d_{E'}$ by moving E toward the vertices Y_ρ or X_ρ, respectively. Thus, this case reduces to checking only the ratio $d_E/d_{E'}$ for the line segments $X_\rho X'_\rho$ and $Y_\rho Y'_\rho$, where

$$X'_\rho = (2\rho/3, -\rho^2/3), \quad Y'_\rho = (-2\rho/3, -\rho^2/3) \tag{B.2}$$

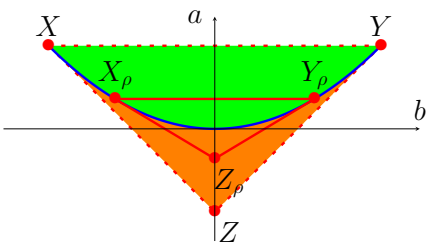


Figure B.1: The green and orange subsets of the stability triangle Δ (dashed-red) correspond to complex conjugate and real eigenvalues for the matrix M in (3.20a), respectively. The blue parabola $a = b^2/4$ corresponds to the matrix M with repeated eigenvalues and it is tangent to the edges $X_\rho Z_\rho$ and $Y_\rho Z_\rho$ of the ρ-linear convergence triangle Δ_ρ (solid red).

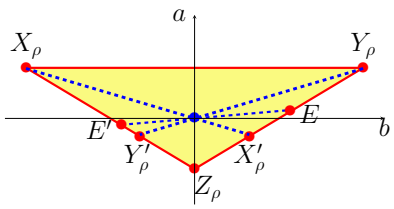


Figure B.2: The points X'_ρ and Y'_ρ as defined in (B.2) along with an arbitrary line segment EE' passing through the origin in the (b, a)-plane.

are the intersections of OX_ρ with $Y_\rho Z_\rho$, and OY_ρ with $X_\rho Z_\rho$; see Figure B.2. Regarding the case when neither E nor E' lies on the edge $X_\rho Y_\rho$, let us assume without loss of generality that E and E' lie on $Y_\rho Z_\rho$ and $X_\rho Z_\rho$, respectively. In this case, we can parameterize the ratio using

$$\frac{d_E}{d_{E'}} = \frac{(1+c)(1/(1-\rho)-c)}{(1-c)(1/(1+\rho)+c)}, \; c \in [-1/2, 1/2] \tag{B.3}$$

where $c\rho$ determines the slope of EE'. The general shape of this function is provided in Figure B.3. It is easy to verify that $d_E/d_{E'}$ takes its maximum over $c \in [-1/2, 1/2]$ at one of the boundaries. Thus, this case also reduces to checking only the ratio $d_E/d_{E'}$ for the line segments $X_\rho X'_\rho$ and $Y_\rho Y'_\rho$. We complete the proof by noting that

$$\frac{d_{X'_\rho}}{d_{X_\rho}} = \frac{(1+\rho)(3-\rho)}{3(1-\rho)^2}, \quad \frac{d_{Y_\rho}}{d_{Y'_\rho}} = \frac{3(1+\rho)^2}{(3+\rho)(1-\rho)}$$

satisfy $d_{X'_\rho}/d_{X_\rho} > d_{Y_\rho}/d_{Y'_\rho}$.

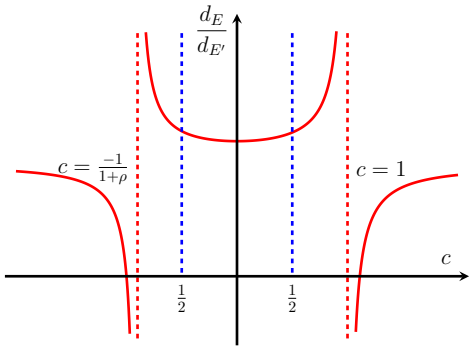


Figure B.3: The ratio $d_E/d_{E'}$ in (B.3) for Nesterov's method, where E and E' lie on the edges $Y_\rho Z_\rho$ and $X_\rho Z_\rho$ of the ρ-linear convergence triangle Δ_ρ, and $c\,\rho$ determines the slope of EE' which passes through the origin.

B.4 Proofs of Section 3.5

B.4.1 Proof of Lemma 4

We show that the parameters (α, β, γ) in (3.32) place the points $(b(m), a(m))$ and $(b(L), a(L))$ on the edges $X_\rho Z_\rho$ and $Y_\rho Z_\rho$ of the ρ-linear convergence triangle Δ_ρ, respectively. In particular, we can use a scalar $c \in [-1, 1]$ to parameterize the end points as

$$(b(m), a(m)) \;=\; (-(1+c)\rho,\; c\rho^2), \qquad (b(L), a(L)) \;=\; ((1+c)\rho,\; c\rho^2).$$

Using the definition of a and b in (3.18c), we can solve the above equations for (α, β, γ) to verify the desired parameters. Thus, the algorithm achieves the convergence rate ρ. In addition the points $c = 0$ and $c = 1$ recover gradient descent and heavy-ball method with the parameters that optimize the convergence rate; see Table 3.1.

Furthermore, h, d, and l in (3.29) are given by

$$\begin{aligned} h(m) \;&=\; h(L) \;=\; 1 \;-\; c\rho^2 \\ d(m) \;&=\; l(L) \;\;=\; (1 \;-\; \rho)(1 \;-\; c\rho) \\ l(m) \;&=\; d(L) \;=\; (1 \;+\; \rho)(1 \;+\; c\rho) \end{aligned} \tag{B.4a}$$

and the condition number is determined by

$$\kappa \;=\; \frac{\alpha L}{\alpha m} \;=\; \frac{d(L)}{d(m)} \;=\; \frac{l(m)}{d(m)}. \tag{B.4b}$$

Combining (B.4b) with (B.4a), and rearranging terms yields the desired expression for c in terms of ρ and κ.

The analytical expressions in Theorem 5 imply that for the parameters in (3.32), the function $\hat{J}(\lambda)$ is symmetric over $[m, L]$, i.e., $\hat{J}(\lambda) = \hat{J}(m + L - \lambda)$ for all $\lambda \in [m, L]$. In addition, as we demonstrate in Appendix B.2, $\hat{J}(\lambda)$ is convex. Thus, $\hat{J}(\lambda)$ attains its

maximum at $\lambda = m$ and $\lambda = L$ and we can use the expression for $\hat{J}(\lambda)$ in Theorem 5 to obtain the maximum value,

$$\hat{J}(m) \;=\; \frac{\sigma_w^2(d(m)+(m))}{2h(m)d(m)l(m)} \;=\; \frac{\sigma_w^2(\kappa\,+\,1)}{2h(m)l(m)} \tag{B.4c}$$

where the second equality follows from (B.4b). Combining (B.4a) and (B.4c) yields the expression for $\hat{J}(m)$.

Also, symmetry and convexity imply that $\hat{J}(\lambda)$ attains its minimum at the midpoint $\lambda = \hat{\lambda} := (m+L)/2 = (1+\beta)/\alpha$. This point corresponds to $(b(\hat{\lambda}), a(\hat{\lambda})) = (0, c\rho^2)$ in the (b,a)-plane and it thus satisfies

$$h(\hat{\lambda}) \;=\; 1 \,-\, c\rho^2, \;\; d(\hat{\lambda}) \;=\; l(\hat{\lambda}) \;=\; 1 \,+\, c\rho^2. \tag{B.4d}$$

Using (B.4d) to evaluate the expression for $\hat{J}(\lambda)$ at the point $\lambda = \hat{\lambda}$ yields the desired minimum value.

B.4.2 Proof of Proposition 2

Using the expressions established in Lemma 4, it is straightforward to verify that

$$\begin{aligned}\hat{J}(m) \times T_s \;&=\; \sigma_w^2\, p_{1c}(\rho)\kappa(\kappa\,+\,1)\\ \hat{J}(\hat{\lambda}) \times T_s \;&=\; \sigma_w^2\, \kappa\, p_{2c}(\rho)\end{aligned}$$

and that, for the gradient noise model ($\sigma_w = \alpha\sigma$), we have

$$\begin{aligned}\hat{J}(m) \times T_s \;&=\; \sigma^2\, p_{3c}(\rho)\kappa(\kappa\,+\,1)\\ \hat{J}(\hat{\lambda}) \;&=\; \sigma^2\, \kappa\, p_{2c}(\rho)\end{aligned}$$

where the functions $p_{1c}(\rho)$-$p_{4c}(\rho)$ are given by (3.36). Thus, the expressions for $J_{\max}$ and $J_{\min}$ follow from Corollary 2. The bounds on $p_{1c}(\rho)$-$p_{4c}(\rho)$ follow from the fact that, for $\rho \in (0,1)$, we have

$$q_c(\rho) \;=\; \frac{1-c\rho}{1-c\rho^2} \;\in\; \begin{cases} [1/(1+c\rho), 1] & c \,\in\, [0,1] \\ [1/2, 2] & c \,\in\, [-1,0]. \end{cases}$$

This completes the proof.

B.4.3 Proof of Proposition 3

Using the expressions established in Lemma 4, it is straightforward to verify that

$$\begin{aligned}\hat{J}(m) \;&=\; \sigma_w^2\, p_{5c}(\rho)\,(1\,+\,1/\kappa)\,T_s\\ \hat{J}(\hat{\lambda}) \;&=\; \sigma_w^2 p_{6c}(\rho) T_s/\kappa\end{aligned}$$

where p_{5c} and p_{6c} are given by Proposition 3. Thus, the expressions for $J_{\max}$ and $J_{\min}$ follow from Corollary 2. The bounds on p_{5c} and p_{6c} also follow from $c \in [-1,0]$ and $\rho \in (0,1)$.

B.4.4 Proof of Proposition 4

We show that (α, β, γ) correspond to the parameterized family of Nesterov-like algorithms in which the end points of the line segment $(b(\lambda), a(\lambda))$, $\lambda \in [m, L]$, lie on the edges $X_\rho Z_\rho$ and $Y_\rho Z_\rho$ of the ρ-linear convergence triangle Δ_ρ. In particular, we can use a scalar $c \in [0, 1/2]$ to parameterize the lines passing through the origin via $a = -c\rho b$. This yields

$$\begin{aligned}(b(m), a(m)) &= (-\rho/(1-c), c\rho^2/(1-c)) \\ (b(L), a(L)) &= (\rho/(1+c), -c\rho^2/(1+c)).\end{aligned}$$

Using the definitions of a and b in (3.18c), we can solve the above equations for (α, β, γ) to verify the desired parameters. Thus, the algorithm achieves the convergence rate ρ and the extreme points $c = 0$ and $c = 1/2$ recover gradient descent and Nesterov's method with the parameters provided in Table 3.1 that optimize settling times.

In Lemma 1, we establish expressions for the convergence rate and largest/smallest modal contributions to noise amplification in terms of the condition number for this family of parameters.

Lemma 1 *For the class of functions $\mathcal{Q}_m^L$ with condition number $\kappa = L/m$, the extreme values $\hat{J}_{\max}$ and $\hat{J}_{\min}$ of $\hat{J}(\lambda)$ over $[m, L]$ associated with the two-step momentum algorithm in* (3.2) *with parameters* (3.38) *satisfy*

$$\begin{aligned}\hat{J}_{\max} &= \hat{J}(m) = \frac{\sigma_w^2(1-c)^2(r\kappa+1)}{2(1-c-c\rho^2)(1+\rho)(1-c+c\rho)} \\ &\geq \hat{J}(L) = \frac{\sigma_w^2(1+c)^2(1+c-c\rho^2)}{(1-\rho^2)(1+c-c\rho)(1+c+c\rho)(1+c+c\rho^2)}\end{aligned}$$

and $\hat{J}_{\min} = \hat{J}(1/\alpha) = \sigma_w^2$, where the scalar $r \in [1, 3]$ is given by

$$r := (1+c)(1-c+c\rho)/\left((1-c)(1+c-c\rho)\right)$$

and the scalar $c \in [0, 1/2]$ is given by Proposition 4.

Proof: The values of h, d, and l in (3.29) are given by

$$\begin{aligned}h(m) &= (1-c-c\rho^2)/(1-c) & h(L) &= (1+c+c\rho^2)/(1+c) \\ d(m) &= (1-\rho)(1-c-c\rho)/(1-c) & d(L) &= (1+\rho)(1+c-c\rho)/(1+c) \\ l(m) &= (1+\rho)(1-c+c\rho)/(1-c) & l(L) &= (1-\rho)(1+c+c\rho)/(1+c)\end{aligned} \tag{B.5a}$$

and the condition number is determined by

$$\kappa = \frac{\alpha L}{\alpha m} = \frac{d(L)}{d(m)} = \frac{l(m)}{r d(m)} \tag{B.5b}$$

where we let $r := l(m)/d(L)$. By combining this identity with the expressions in (B.5a), and rearranging terms, we can obtain the desired quadratic equation for c in terms of ρ and κ. To

see that $r \in [1,3]$, from Figure 3.3 we observe that as we change the orientation from gradient descent ($c = 0$) to Nesterov's method with parameters that optimize the convergence rate ($c = 1/2$), $l(m)$ and $1/d(L)$ monotonically increase. Thus, r is also increasing in c, and its smallest and largest values are obtained for $c = 0$ and $c = 1/2$, respectively, which yields $1 \leq r \leq 3(1+\rho)/(3-\rho) \leq 3$.

As we demonstrate in Appendix B.2, $\hat{J}$ as a function of (b, a) takes its minimum $\hat{J}_{\min} = \sigma_w^2$ at the origin. In addition, for each $c \in [0, 1/2]$, the line segment $(b(\lambda), a(\lambda))$, $\lambda \in [m, L]$, passes through the origin at $\lambda = 1/\alpha$. Thus, the minimum of $\hat{J}(\lambda)$ occurs at $\lambda = 1/\alpha$ and is given by $\hat{J}_{\min} = \sigma_w^2$.

We next show that $\hat{J}(m)$ is the larges value of $\hat{J}(\lambda)$ over $[m, L]$. Since $\hat{J}(\lambda)$ is a convex function of λ (see Appendix B.2), it attains its maximum at one of the boundary points $\lambda = m$ and $\lambda = L$. To show $\hat{J}(m) > \hat{J}(L)$, we first obtain expressions for $\hat{J}(m)$ and $\hat{J}(L)$ in terms of ρ and c by combining (B.5a) with the analytical expression for $\hat{J}$ in Theorem 5. By properly rearranging terms and simplifying fractions, we can obtain the equivalence

$$\hat{J}(m) \;\geq\; \hat{J}(L) \iff c^4\rho^4 - c^4\rho^2 - c^2\rho^2 - c^2 + 1 \;\geq\; 0.$$

For $\rho \in [0, 1]$ and $c \in [0, 1/2]$, it is easy to verify that the inequality on the right-hand holds.

To obtain the maximum value, we use Theorem 5 to write

$$\frac{\hat{J}(m)}{\sigma_w^2} \;=\; \frac{d(m) + (m)}{2h(m)d(m)l(m)} \;=\; \frac{r\kappa \,+\, 1}{2h(m)l(m)} \tag{B.5c}$$

Combining (B.5a) with (B.5c) yields the desired value for $\hat{J}(m)$. □

Lemma 1 allows us to derive analytical expressions for the largest and smallest values that J takes over $f \in \mathcal{Q}_m^L$.

Corollary 1 *The parameterized family of Nesterov-like methods* (3.38) *satisfies*

$$\begin{aligned} J_{\max} &= (n \,-\, 1)\hat{J}(m) \,+\, \hat{J}(L) \\ J_{\min} &= \hat{J}(m) \,+\, \hat{J}(L) \,+\, (n \,-\, 2)\hat{J}(1/\alpha) \end{aligned}$$

where $\hat{J}(m)$, $\hat{J}(L)$, $\hat{J}(1/\alpha)$ are given by Lemma 1, and $J_{\max}$, $J_{\min}$ are the extreme values of J when the algorithm is applied to $f \in \mathcal{Q}_m^L$ with condition number $\kappa = L/m$.

Proof: The result follows from combining Lemma 1 and the expression $J = \sum_{i=1}^n \hat{J}(\lambda_i)$ established in Theorem 5. In particular, J is maximized when Q has $n-1$ eigenvalues at m and one at L, and it is minimized when, apart from the extreme eigenvalues m and L, the rest are at $\lambda = 1/\alpha$. □

We next establish order-wise tight upper and lower bounds on $\hat{J}_{\max}/(1-\rho)$ and $\hat{J}_{\min}/(1-\rho)$ in terms of κ.

Lemma 2 *For the parameterized family of Nesterov-like methods* (3.38)*, the largest and smallest modal contributions to variance amplification established in Lemma 1 satisfy*

$$\sigma_w^2\omega_1 r\,\kappa(r\,\kappa\,+\,1)\,\leq\,\hat{J}_{\max}\times T_s\,\leq\,\sigma_w^2\omega_2 r\,\kappa(r\,\kappa\,+\,1)$$
$$\sigma_w^2\sqrt{3\kappa+1}/2\,\leq\,\hat{J}_{\min}\times T_s\,\leq\,\sigma_w^2(\kappa\,+\,1)/2$$

where the scalar $\omega_1 := (1+\rho)^{-3}(1-c)^2(1-c+c\rho)^{-2}/2$*,* $\omega_2 := (1+\rho)\omega_1$*, and we have* $(1\,+\,\rho)^{-5} \leq \omega_1 \leq (1\,+\,\rho)^{-3}$*.*

Proof: To obtain the upper and lower bounds on $\hat{J}_{\max}\times T_s = \hat{J}(m)\times T_s$, we combine (B.5b) and (B.5c) to write

$$\hat{J}(m)\;=\;\frac{r\kappa(r\kappa\,+\,1)}{2h(m)(l(m))^2/d(m)}$$

where we set $\sigma_w = 1$. This equation in conjunction with the trivial inequalities

$$1\;\leq\;(1\,-\,\rho)h(m)/d(m)\;\leq\;1\,+\,\rho$$

allows us to write

$$\frac{r\kappa(r\kappa\,+\,1)}{2(1\,+\,\rho)l^2(m)}\;\leq\;\frac{\hat{J}(m)}{1\,-\,\rho}\;\leq\;\frac{r\kappa(r\kappa\,+\,1)}{2l^2(m)}. \tag{B.6}$$

Combining (B.5a) and (B.6) yields the desired bounds on $\hat{J}_{\max}/(1-\rho)$. Finally, the bounds on $\hat{J}_{\min}\times T_s = \sigma_w^2/(1-\rho)$ can be obtained by noting that $T_s = 1/(1-\rho) \in [\sqrt{3\kappa+1}/2, (\kappa+1)/2]$ as shown in Lemma 1. □

Similar to the heavy-ball-like methods, $\hat{J}_{\max}\times T_s = \Theta(\kappa^2)$. However, the upper and lower bounds on $\hat{J}_{\min}\times T_s$ scale linearly with κ and $\sqrt{\kappa}$, respectively. We next use this result to bound $J\times T_s$ and complete the proof of Proposition 4. In particular, we have

$$(n-1)\omega_1 r\kappa(r\kappa\,+\,1)\,+\,\frac{\sqrt{3\kappa+1}}{2}\leq\frac{J_{\max}\times T_s}{\sigma_w^2}\leq n\,\omega_2 r\kappa(r\kappa\,+\,1)$$
$$\omega_1 r\kappa(r\kappa\,+\,1)+(n-1)\frac{\sqrt{3\kappa+1}}{2}\leq\frac{J_{\min}\times T_s}{\sigma_w^2}\,\leq\omega_2 r\kappa(r\kappa\,+\,1)+(n-1)\frac{\kappa+1}{2}$$

where the scalar $r \in [1,3]$, and ω_1 and ω_2 are given by Lemmas 1 and 2, respectively. To see this, note that as shown in the proof of Corollary 1, $J/(1-\rho)$ is maximized when Q has $n-1$ eigenvalues at m and one at L, and is minimized when, apart from the extreme eigenvalues m and L, the rest are placed at $\lambda = 1/\alpha$. Employing the bounds on $\hat{J}_{\max} = \hat{J}(m)$ and $\hat{J}_{\min} = \hat{J}(1/\alpha)$ provided by Lemma 2 and noting that $\hat{J}(L) \in [\hat{J}_{\min}, \hat{J}_{\max}]$ completes the proof.

B.5 Proofs of Section 3.6

B.5.1 Proof of Lemma 5

Stability can be verified using the Routh-Hurwitz criterion applied to the characteristic polynomial $F(s) = \det(sI - M) = s^2 + bs + a$. Similarly, conditions for ρ-exponential stability can be obtained by applying the Routh-Hurwitz criterion to the characteristic polynomial $F_\rho(s)$ associated with the matrix $M + \rho I$, i.e.,

$$F_\rho(s) \;=\; s^2 \;+\; (b \;-\; 2\rho)s \;+\; \rho^2 \;-\; \rho b \;+\; a$$

and noting that strict inequalities become non-strict as we require $\Re(\mathrm{eig}(M + \rho I))$ to be non-positive.

B.5.2 Proof of Proposition 5

The ρ-exponential stability of (3.41b) with $\alpha = 1/L$ is equivalent to the inclusion of the line segment $(b(\lambda), a(\lambda)),\ \lambda \in [m, L]$, in the triangle Δ_ρ in (3.44), where $a(\lambda)$ and $b(\lambda)$ are given by (3.42c). In addition, using the convexity of Δ_ρ, this condition further reduces to the end points $(b(L), a(L))$ and $(b(m), a(m))$ belonging to Δ_ρ. Now since $a(L) = 1$, $a(m) = 1/\kappa$, the above condition implies

$$a_{\max}/a_{\min} \;\geq\; \kappa \tag{B.7}$$

where $a_{\max}$ and $a_{\min}$ are the largest and smallest values that a can take among all $(b, a) \in \Delta_\rho$. It is now easy to verify that $a_{\max} = 1$ and $a_{\min} = \rho^2$ correspond to the edge $Y_\rho Z_\rho$ and the vertex X_ρ of Δ_ρ, respectively; see Figure 3.5. Thus, inequality (B.7) yields the upper bound $\rho \leq 1/\sqrt{\kappa}$ and we can achieve this rate with,

$$(b(m), a(m)) \;=\; X_\rho,\ (b(L), a(L)) \;=\; E_v \tag{B.8}$$

where $E_v := (b_v, 1) = (2\rho + v(\rho - 1/\rho), 1)$, $v \in [0, 1]$, parameterizes the edge $Y_\rho Z_\rho$. Solving the equations in (B.8) for γ and β yields the optimal values of parameters. Finally, letting $\gamma = 0$ and $\gamma = \beta$ yields the conditions on v for the heavy-ball and Nesterov's method, respectively. The condition $\kappa \geq 4$ for Nesterov's method stems from the fact that, for $\alpha = 1/L$, setting $\gamma = \beta$ yields $b(L) = 1$. Thus, we have the necessary condition $2\rho \leq 1$ to ensure $(b(L), a(L)) \in \Delta_\rho$; see Figure 3.6. This completes the proof.

B.5.3 Proof of Theorem 7

Let $G := (b_G, a_G)$ be the point on the edge $X_\rho Z_\rho$ of the triangle Δ_ρ in (3.44) such that

$$a_G \;=\; a(m),\ b_G \;=\; a(m)/\rho \;+\; \rho.$$

Using $(b(m), a(m)) \in \Delta_\rho$, it is easy to verify that $b_G \geq b(m)$. This allows us to write

$$\frac{\hat{J}(m)}{\rho} \;=\; \frac{1}{2a(m)b(m)\,\rho} \;\geq\; \frac{1}{2a(m)b_G\,\rho} \;=\; \frac{1}{2a(m)(a(m)+\rho^2)}.$$

Combining the above inequality with $a(m) = 1/\kappa$ and the upper bound $\rho \leq 1/\sqrt{\kappa}$ from Lemma 5 yields

$$\hat{J}(m)/\rho \;\geq\; \kappa^2/4. \tag{B.9}$$

Noting that among the points in Δ_ρ, the modal contribution $\hat{J} = 1/(2ab)$ takes its minimum value

$$\hat{J}_{\min} \;=\; 1/(2\rho + 2/\rho) \tag{B.10}$$

at the vertex $Z_\rho = (1,\, \rho + 1/\rho)$, we can write

$$\frac{J}{\rho} \;=\; \frac{\hat{J}(m)}{\rho} + \sum_{i=1}^{n-1} \frac{\hat{J}(\lambda_i)}{\rho} \;\geq\; \frac{\kappa^2}{4} + \frac{n\,-\,1}{2(1\,+\,\rho^2)}$$

where we use (B.9) to lower bound the first term $\hat{J}(m)/\rho$. This completes the proof of (3.48b).

To prove the lower bound in (3.48a), we consider a quadratic objective function for which the Hessian has $n-1$ eigenvalues at $\lambda = m$ and one eigenvalue at $\lambda = L$. For such a function, we can write

$$J_{\max} \;\geq\; J \;=\; \hat{J}(m)(n\,-\,1) \,+\, \hat{J}(L).$$

Finally, we lower bound the right hand-side using (B.9) and (B.10) to complete the proof.

B.6 Lyapunov equations and the steady-state variance

For the discrete-time LTI system in (3.4a), the covariance matrix $P^t := \mathbb{E}\left(\psi^t(\psi^t)^T\right)$ of the state vector ψ^t satisfies the linear recursion

$$P^{t+1} \;=\; A\,P^t A^T \,+\, BB^T \tag{B.11a}$$

and its steady-state limit

$$P \;:=\; \lim_{t\,\to\,\infty} \mathbb{E}\left[\psi^t(\psi^t)^T\right] \tag{B.11b}$$

is the unique solution to the algebraic Lyapunov equation [41],

$$P \;=\; A\,PA^T \,+\, BB^T. \tag{B.11c}$$

For stable LTI systems, performance measure (3.8) can be computed using

$$J \;=\; \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t} \operatorname{trace}\left(Z^k\right) \;=\; \operatorname{trace}\left(Z\right) \tag{B.11d}$$

where $Z = CPC^T$ is the steady-state limit of the output covariance matrix

$$Z^t \;:=\; \mathbb{E}\left[z^t (z^t)^T\right] \;=\; CP^t C^T.$$

We can prove Theorem 5 by finding the solution P to (B.11c) for the two-step momentum algorithm. The above results carry over to the continuous-time case with the only difference that the Lyapunov equation for the steady-state covariance matrix of $\psi(t)$ is given by

$$AP \;+\; PA^T \;=\; -BB^T.$$

Appendix C

Supporting proofs for Chapter 4

C.1 Proofs of Section 4.2

We first present a technical lemma that we use in our proofs.

Lemma 1 *For any $\rho \in [1/\mathrm{e}, 1)$, $a(t) := t\rho^t$ satisfies*

$$\operatorname*{argmax}_{t \geq 1} a(t) \;=\; -1/\log(\rho), \;\; \max_{t \geq 1} a(t) \;=\; -1/(\mathrm{e}\log(\rho)).$$

Proof: Follows from the fact that $\mathrm{d}a/\mathrm{d}t = \rho^t(1 + t\log(\rho))$ vanishes at $t = -1/\log(\rho)$. $\square$

C.1.1 Proof of Lemma 1

For $\mu_1 \neq \mu_2$, the eigenvalue decomposition of M is determined by

$$M \;=\; \frac{1}{\mu_2 - \mu_1} \begin{bmatrix} 1 & 1 \\ \mu_1 & \mu_2 \end{bmatrix} \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix} \begin{bmatrix} \mu_2 & -1 \\ -\mu_1 & 1 \end{bmatrix}.$$

Computing the tth power of the diagonal matrix and multiplying throughout completes the proof for $\mu_1 \neq \mu_2$. For $\mu_1 = \mu_2 =: \mu$, M admits the Jordan canonical form

$$M \;=\; \begin{bmatrix} 1 & 0 \\ \mu & 1 \end{bmatrix} \begin{bmatrix} \mu & 1 \\ 0 & \mu \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\mu & 1 \end{bmatrix}$$

and the proof follows from

$$\begin{bmatrix} \mu & 1 \\ 0 & \mu \end{bmatrix}^t = \begin{bmatrix} \mu^t & t\,\mu^{t-1} \\ 0 & \mu^t \end{bmatrix}.$$

C.1.2 Proof of Lemma 2

From Lemma 1, it follows

$$\begin{bmatrix} 1 & 0 \end{bmatrix} M^t = \begin{bmatrix} -\sum_{i=0}^{t-2} \mu_1^{i+1}\mu_2^{t-1-i} & \sum_{i=0}^{t-1} \mu_1^{i}\mu_2^{t-1-i} \end{bmatrix}$$

where μ_1 and μ_2 are the eigenvalues of M. Moreover,

$$\begin{aligned} |\sum_{i=0}^{t-2} \mu_1^{i+1}\mu_2^{t-1-i}| &\leq \sum_{i=0}^{t-2} |\mu_1^{i+1}\mu_2^{t-1-i}| \leq \sum_{i=0}^{t-2} \rho^t \leq (t-1)\rho^t \\ |\sum_{i=0}^{t-1} \mu_1^{i}\mu_2^{t-1-i}| &\leq \sum_{i=0}^{t-1} |\mu_1^{i}\mu_2^{t-1-i}| \leq \sum_{i=0}^{t-1} \rho^{t-1} \leq t\rho^{t-1} \end{aligned}$$

by triangle inequality. Finally, for $\mu_1 = \mu_2 \in \mathbb{R}$, we have $\rho = |\mu_1| = |\mu_2|$ and the above inequalities become equalities.

C.1.3 Proof of Theorem 1

Let μ_{1i} and μ_{2i} be the eigenvalues and let $\rho_i = \max\{|\mu_{1i}|, |\mu_{2i}|\}$ be the spectral radius of A_i. We can use Lemma 2 with $M := A_i$ to obtain

$$\begin{aligned} \max_{i \leq r} \|C_i A_i^t\|_2^2 &\leq \max_{i \leq r} \left((t-1)^2\rho_i^{2t} + t^2\rho_i^{2t-2}\right) \\ &\leq (t-1)^2\rho^{2t} + t^2\rho^{2t-2} \end{aligned} \tag{C.1}$$

where $\rho := \max_{i \leq r} \rho_i$. For the parameters provided in Table 4.1, the matrices A_1 and A_r, that correspond to the largest and smallest non-zero eigenvalues of Q, i.e., $\lambda_1 = L$ and $\lambda_r = m$, respectively, have the largest spectral radius [93, Eq. (64)],

$$\rho = \rho_1 = \rho_r \geq \rho_i, \; i = 2, \ldots, r-1 \tag{C.2}$$

and A_r has repeated eigenvalues. Thus, we can write

$$\begin{aligned} &\max_{i \leq r} \|C_i A_i^t\|_2^2 \geq \| \begin{bmatrix} 1 & 0 \end{bmatrix} A_r^t\|_2^2 = \\ &(t-1)^2\rho_r^{2t} + t^2\rho_r^{2t-2} = (t-1)^2\rho^{2t} + t^2\rho^{2t-2} \end{aligned} \tag{C.3}$$

where the first equality follows from Lemma 2 applied to $M := A_r$ and the second equality follows from (C.2). Finally, combining (C.1) and (C.3) with $\beta < \rho$ and Proposition 1 completes the proof.

C.1.4 Proof of Theorem 2

Let $a(t) := t\rho^t$. Theorem 1 implies $J^2(t) = \rho^2 a^2(t-1) + \rho^{-2}a^2(t)$ and, for $t \geq 1$, $J(t)$ has only one critical point, which is a maximizer. Moreover, since $\mathrm{d}J^2(t)/\mathrm{d}t$ is positive at $t = -1/\log(\rho)$ and negative at $t = 1 - 1/\log(\rho)$, we conclude that the maximizer lies

between $-1/\log(\rho)$ and $1-1/\log(\rho)$. Regarding $\max_t J(t)$, we note that $\sqrt{2}\rho a(t-1) \leq J(t) \leq \sqrt{2}a(t)/\rho$ and the proof follows from $\max_{t\geq 1} a(t) = -1/(\mathrm{e}\log(\rho))$ (cf. Lemma 1).

C.1.5 Proof of Proposition 2

Since for all $a \leq 1$, we have [167]

$$a \;\leq\; -\log(1-a) \;\leq\; a/(1-a)$$

$\rho_{\mathrm{hb}} = 1-2/(\sqrt{\kappa}+1)$ and $\rho_{\mathrm{na}} = 1-2/(\sqrt{3\kappa+1})$ satisfy

$$\begin{aligned} 2/(\sqrt{\kappa}+1) \;&\leq\; -\log(\rho_{\mathrm{hb}}) \;\leq\; 2/(\sqrt{\kappa}-1) \\ 2/\sqrt{3\kappa+1} \;&\leq\; -\log(\rho_{\mathrm{na}}) \;\leq\; 2/(\sqrt{3\kappa+1}-2). \end{aligned}$$

The conditions on κ ensure that ρ_{hb} and ρ_{na} are not smaller than $1/\mathrm{e}$ and we combine the above bounds with Theorem 2 to complete the proof.

C.2 Proof of Theorem 3

The condition $x_0 = x_1$ is equivalent to $\hat{x}_i^0 = \hat{x}_i^1$ in (4.5). Thus, for $\lambda_i = 0$, equation (4.12) yields $\hat{x}_i^t = \hat{x}_i^0 = \hat{x}_i^\star$. For $\lambda_i \neq 0$, we have $\hat{\psi}_i^0 - \hat{\psi}_i^\star = \begin{bmatrix} \hat{x}_i^0 & \hat{x}_i^0 \end{bmatrix}^T$ and, hence,

$$\frac{\|x^t - x^\star\|_2}{\|x^0 - x^\star\|_2} \;\leq\; \max_{i\leq r} \frac{|\hat{x}_i^t - \hat{x}_i^\star|}{|\hat{x}_0^t - \hat{x}_i^\star|} \;=\; \max_{i\leq r} \left| C_i A_i^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right| \tag{C.4a}$$

where the equality follows from (4.10). To bound the right-hand side, we use Lemma 1 with $M = A_i$ to obtain

$$C_i A_i^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} \;=\; \begin{bmatrix} 1 & 0 \end{bmatrix} A_i^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} \;=\; \omega_t(\mu_{1i}, \mu_{2i}) \tag{C.4b}$$

where μ_{1i} and μ_{2i} are the eigenvalues of A_i and

$$\omega_t(z_1, z_2) \;:=\; \sum_{i=0}^{t-1} z_1^i z_2^{t-1-i} \;-\; \sum_{i=1}^{t-1} z_1^i z_2^{t-i} \tag{C.5}$$

for any $t \in \mathbb{N}$ and $z_1, z_2 \in \mathbb{C}$.

For Nesterov's accelerated method, the characteristic polynomial $\det(zI - A_i) = z^2 - (1+\beta)h_i z + \beta h_i$ yields $\mu_{1i}, \mu_{2i} = ((1+\beta)h_i \pm \sqrt{(1+\beta)^2 h_i^2 - 4\beta h_i})/2$, where λ_i is the ith the eigenvalue of Q and $h_i := 1 - \alpha\lambda_i$. For the parameters provided in Table 4.1, it is easy to show that:

- For $\lambda_i \in [m, 1/\alpha]$, we have $h_i \in [0, 4\beta/(1+\beta)^2]$ and μ_{1i} and μ_{2i} are complex conjugates of each other and lie on a circle of radius $\beta/(1+\beta)$ centered at $z = \beta/(1+\beta)$.
- For $\lambda_i \in (1/\alpha, L]$, μ_{1i} and μ_{2i} are real with opposite signs and can be sorted to satisfy $|\mu_{2i}| < |\mu_{1i}|$ with $-1 \leq \mu_{1i} \leq 0 \leq \mu_{2i} \leq 1/3$.

The next lemma provides a unit bound on $|\omega_t(\mu_{1i}, \mu_{2i})|$ for both of the above cases.

Lemma 2 *For any $z = l\cos(\theta)\mathrm{e}^{i\theta} \in \mathbb{C}$ with $|\theta| \leq \pi/2$ and $0 \leq l \leq 1$, and for any real scalars (z_1, z_2) such that $-1 \leq z_1 \leq 0 \leq z_2 \leq 1/3$, and $z_2 < -z_1$, the function ω_t in (C.5) satisfies $|\omega_t(z, \bar{z}\,)| \leq 1$ and $|\omega_t(z_1, z_2)| \leq 1$ for all $t \in \mathbb{N}$, where $\bar{z}$ is the complex conjugate of z.*

Proof: Since $\omega_1(z_1, z_2) = 1$, we assume $t \geq 2$. We first address $\theta = 0$, i.e., $z = l \in \mathbb{R}$ and $\omega_t(z, \bar{z}\,) = tl^{t-1} - (t-1)l^t$. We note that $\mathrm{d}\omega_t/\mathrm{d}l = t(t-1)(l^{t-2} - l^{t-1}) = 0$ only if $l \in \{0, 1\}$. This in combination with $l \in [0, 1]$ yield $|\omega_t(l, l)| \leq \max\{|\omega_t(1, 1)|\,, |\omega_t(0, 0)|\} \leq 1$.

To address $\theta \neq 0$, we note that $b(t) := \sin(t\theta)/t$ satisfies

$$|b(t)| \;\leq\; |\sin(\theta)| \tag{C.6}$$

which follows from

$$|\sin(t\theta)| \;=\; |\sin((t-1)\theta)\cos(\theta) \,+\, \cos((t-1)\theta)\sin(\theta)| \;\leq\; |\sin((t-1)\theta)| \,+\, |\sin(\theta)|\,.$$

For $z = l\cos(\theta)\mathrm{e}^{i\theta}$, we have

$$\begin{aligned}\omega_t(z, \bar{z}\,) \;&=\; (z^t - \bar{z}^{\,t} - z\bar{z}\,(z^{t-1} - \bar{z}^{\,t-1}))/(z - \bar{z}\,) \\ &=\; (l\cos(\theta))^{t-1}(\sin(t\theta) - l\cos(\theta)\sin((t-1)\theta))/\sin(\theta).\end{aligned}$$

Thus, $\mathrm{d}\omega_t/\mathrm{d}l = 0$ only if $l = 0$, 1, or $l^\star := b(t)/(b(t{-}1)\cos(\theta))$. Moreover, it is straightforward to show that

$$\omega_t(z, \bar{z}\,) \;=\; \begin{cases} 0, & l \,=\, 0 \\ (\cos(\theta))^{t-1}\cos((t-1)\theta), & l \,=\, 1 \\ (l^\star\cos(\theta))^{t-1}b(t)/\sin(\theta), & l \,=\, l^\star. \end{cases}$$

Combining this with (C.6) completes the proof for complex z.

To address the case of z_1, $z_2 \in \mathbb{R}$, we note that

$$\omega_t(z_1, z_2) \;=\; \big(z_1^t(1 \,-\, z_2) \,-\, z_2^t(1 \,-\, z_1)\big)/(z_1 \,-\, z_2).$$

Thus, differentiating with respect to z_1 yields

$$\frac{\mathrm{d}\omega_t}{\mathrm{d}z_1} \;=\; (1 - z_2)\frac{(t-1)z_1^{t-1} - z_2\sum_{i=0}^{t-2} z_1^{t-2-i}z_2^i}{z_1 - z_2}.$$

Moreover, from $|z_2| < |z_1|$, it follows that

$$(t-1)\left|z_1^{t-1}\right| \;>\; |z_2|\sum_{i=0}^{t-2}\left|z_1^{t-2-i}z_2^i\right| \;>\; \left|z_2\sum_{i=0}^{t-2} z_1^{t-2-i}z_2^i\right|.$$

Therefore, $\mathrm{d}\omega_t/\mathrm{d}z_1 \neq 0$ over our range of interest for z_1, z_2. Thus, $\omega_t(z_1, z_2)$ may take its extremum only at the boundary $z_1 \in \{0, -1\}$, i.e. $|\omega_t(z_1, z_2)| \leq \max\{|\omega_t(0, z_2)|, |\omega_t(1, z_2)|\}$. Finally, it is easy to show that $|\omega_t(0, z_2)| = \left|z_2^{t-1}\right| < 1$, and

$$|\omega_t(-1, z_2)| \;=\; \left|(-1)^t(z_2 \,-\, 1) \,+\, 2z_2^t\right|/(1 \,+\, z_2) \;\leq\; 1.$$

□

We complete the proof of Theorem 3 by noting that the eigenvalues of the matrices A_i for Nesterov's algorithm with parameters provided in Table 4.1 satisfy the conditions in Lemma 2.

C.3 Proofs of Section 4.3

C.3.1 Proof of Lemma 3

For any $f \in \mathcal{F}_m^L$, the L-Lipschitz continuity of the gradient ∇f,

$$f(x^{t+2}) \,-\, f(y^t) \;\leq\; (\nabla f(y^t))^T(x^{t+2} - y^t) \,+\, \frac{L}{2}\,\|x^{t+2} - y^t\|_2^2 \tag{C.7a}$$

and the m-strong convexity of f,

$$f(y^t) \,-\, f(x^{t+1}) \;\leq\; (\nabla f(y^t))^T(y^t - x^{t+1}) \,-\, \frac{m}{2}\,\|y^t - x^{t+1}\|_2^2 \tag{C.7b}$$

can be used to show that (4.20) for the solution of Nesterov's accelerated algorithm (4.18). In particular, for (4.18) we have $u^t := \nabla f(y^t)$ and

$$\begin{aligned} x^{t+2} \,-\, y^t \;&=\; -\alpha u^t \\ y^t \,-\, x^{t+1} \;&=\; \beta(x^{t+1} \,-\, x^t) \;=\; \left[\begin{array}{cc} -\beta I & \beta I \end{array}\right]\psi^t. \end{aligned} \tag{C.8}$$

Substituting (C.8) into (C.7a) and (C.7b) and adding the resulting inequalities completes the proof.

C.3.2 Proof of Lemma 4

Pre- and post-multiplication of LMI (4.21) by $(\eta^t)^T$ and $\eta^t := [\,(\psi^t)^T\ (u^t)^T\,]^T$ yields

$$\begin{aligned} 0 \,\geq\; (\eta^t)^T \left[\begin{array}{cc} A^T X\, A - X & A^T X\, B \\ B^T\, X\, A & B^T\, X\, B \end{array}\right]\eta^t \,+\, \theta_1(\eta^t)^T M_1 \eta^t \,+\, \theta_2(\eta^t)^T M_2 \eta^t \\ \geq\; (\eta^t)^T \left[\begin{array}{cc} A^T X\, A - X & A^T X\, B \\ B^T\, X\, A & B^T\, X\, B \end{array}\right]\eta^t \,+\, \theta_2(\eta^t)^T M_2 \eta^t \end{aligned}$$

where the second inequality follows from (4.19c). This yields

$$0 \;\leq\; \hat{V}(\psi^t) \,-\, \hat{V}(\psi^{t+1}) \,-\, \theta_2(\eta^t)^T M_2 \eta^t \tag{C.9}$$

where $\hat{V}(\psi) := \psi^T X \psi$. Also, since Lemma 3 implies

$$-(\eta^t)^T M_2 \eta^t \;\leq\; 2\left(f(x^{t+1}) \;-\; f(x^{t+2})\right) \tag{C.10}$$

combining (C.9) and (C.10) yields

$$\hat{V}(\psi^{t+1}) \;+\; 2\theta_2 f(x^{t+2}) \;\leq\; \hat{V}(\psi^t) \;+\; 2\theta_2 f(x^{t+1}).$$

Thus, using induction, we obtain the uniform upper bound

$$\hat{V}(\psi^t) \;+\; 2\theta_2 f(x^{t+1}) \;\leq\; \hat{V}(\psi^0) \;+\; 2\theta_2 f(x^1). \tag{C.11}$$

This allows us to bound $\hat{V}$ by writing

$$\lambda_{\min}(X)\|\psi\|_2^2 \;\leq\; \hat{V}(\psi) \;\leq\; \lambda_{\max}(X)\|\psi\|_2^2. \tag{C.12a}$$

We can also upper and lower bound $f \in \mathcal{F}_m^L$ as

$$m\|x\|_2^2 \;\leq\; 2f(x) \;\leq\; L\|x\|_2^2. \tag{C.12b}$$

Finally, combining (C.11) and (C.12) yields

$$\lambda_{\min}(X)\|\psi^t\|_2^2 \;+\; m\,\theta_2\|x^{t+1}\|_2^2 \;\leq\; \lambda_{\max}(X)\|\psi^0\|_2^2 \;+\; L\,\theta_2\|x^1\|_2^2.$$

We complete the proof by noting that $\|x^{t+1}\|_2 \leq \|\psi^t\|_2$.

C.3.3 Proof of Theorem 4

To prove (4.23a), we need to find a feasible solution for θ_1, θ_2 and X in terms of the condition number κ. Let us define

$$\begin{aligned} X \;&:=\; \begin{bmatrix} x_1 I & x_0 I \\ x_0 I & x_2 I \end{bmatrix} \;=\; x_2 \begin{bmatrix} \beta^2 I & -\beta I \\ -\beta I & I \end{bmatrix} \\ \theta_2 \;&:=\; \theta_1(L+m)\beta/(1-\beta) \\ x_2 \;&:=\; ((L+m)\theta_1 \;+\; \theta_2)/\alpha \;=\; \theta_2/(\alpha\beta). \end{aligned} \tag{C.13}$$

If (C.13) holds, it is easy to verify that $X \succeq 0$ with $\lambda_{\min}(X) = 0$, $\lambda_{\max}(X) = (1+\beta^2)x_2 = \theta_2(1+\beta^2)/(\alpha\beta)$, and $A^T X A - X = 0$. Moreover, the matrix W on the left-hand-side of (4.21) is block-diagonal, $W := \operatorname{diag}(W_1, W_2)$, and negative semi-definite for all $\alpha \leq 1/L$, where

$$\begin{aligned} W_1 \;&=\; -m(2\theta_1 L\, C_y^T C_y + \theta_2\, C_2^T C_2) \;\preceq\; 0 \\ W_2 \;&=\; -\left((2-\alpha(L+m))\,\theta_1 \;+\; \alpha(1-\alpha L)\,\theta_2\right) I \;\preceq\; 0. \end{aligned}$$

Thus, the choice of (θ_1, θ_2, X) in (C.13) satisfies the conditions of Lemma 4. Using the expressions for the largest and smallest eigenvalues of the matrix X in equation (4.22) in Lemma 4, leads to the upper bound for $\|x^t\|_2^2$ in (4.23a). Furthermore, from (4.23a) we have

$$\|x^t\|_2^2 \;\leq\; \kappa \left(1 + (1+\beta^2)/(\alpha\beta L)\right) \|\psi^0\|_2^2$$

and the upper bound in (4.23c) follows from the fact that, for α and β in (4.23b), $1 + (1 + \beta^2)/(\alpha\beta L) = 3 + 4/(\kappa - 1)$.

To obtain the lower bound in (4.23c), we employ our framework for quadratic objective functions in Section 4.2. In particular, for the parameters α and β in (4.23b), the largest spectral radius $\rho(A_i)$ corresponds to A_n, which is associated with the smallest eigenvalue $\lambda_n = m$ of Q. Since A_n has repeated real eigenvalues $\rho = 1 - 1/\sqrt{\kappa}$, using similar arguments as in Theorem 1 for quadratic problems we obtain,

$$\begin{aligned} J(t_{\max}) \;&=\; \sqrt{(t_{\max}-1)^2\rho^{2t_{\max}} \;+\; t_{\max}^2\rho^{2(t_{\max}-1)}} \\ &\geq\; \sqrt{2}\,(t_{\max} \;-\; 1)\,\rho^{t_{\max}} \;\geq\; \sqrt{2}(\sqrt{\kappa}-1)^2/(\mathrm{e}\sqrt{\kappa}) \end{aligned}$$

which completes the proof.

Appendix D

Supporting proofs for Chapter 6

D.1 Lack of convexity of function f

The function f is nonconvex in general because its effective domain, namely, the set of stabilizing feedback gains $\mathcal{S}_K$ can be nonconvex. In particular, for $A = 0$ and $B = -I$, the closed-loop A-matrix is given by $A - BK = K$. Now, let

$$K_1 = \begin{bmatrix} -1 & 2-2\epsilon \\ 0 & -1 \end{bmatrix}, \; K_2 = \begin{bmatrix} -1 & 0 \\ 2-2\epsilon & -1 \end{bmatrix}, \; K_3 = \frac{K_1+K_2}{2} = \begin{bmatrix} -1 & 1-\epsilon \\ 1-\epsilon & -1 \end{bmatrix} \tag{D.1}$$

where $0 \leq \epsilon \ll 1$. It is straightforward to show that for $\epsilon > 0$, the entire line-segment $\overline{K_1K_2}$ lies in $\mathcal{S}_K$. However, if we let $\epsilon \to 0$, while the endpoints K_1 and K_2 converge to stabilizing gains, the middle point K_3 converges to the boundary of $\mathcal{S}_K$. Thus, $f(K_1)$ and $f(K_2)$ are bounded whereas $f(K_3) \to \infty$. This implies the existence of a point on the line-segment $\overline{K_1K_2}$ for some $\epsilon \ll 1$ for which the function f has negative curvature. For $\epsilon = 0.1$, Fig. D.1 illustrates the value of the LQR objective function $f(K(\gamma))$ associated with the above example and the problem parameters $Q = R = \Omega = I$, where $K(\gamma) := \gamma K_1 + (1-\gamma)K_2$ is the line-segment $\overline{K_1K_2}$. We observe the negative curvature of f around the middle point K_3. Alternatively, we can verify the negative curvature using the second-order term $\langle J, \nabla^2 f(K); J\rangle$ in the Taylor series expansion of $f(K+J)$ around K given in Appendix D.7. For the above example, letting $J = (K_1 - K_2)/\|K_1 - K_2\|$ yields the negative value $\langle J, \nabla^2 f(K_3); J\rangle = -135.27$.

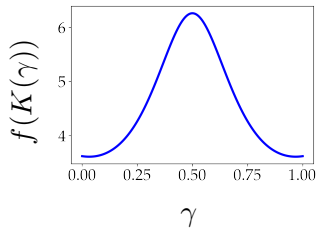


Figure D.1: The LQR objective function $f(K(\gamma))$, where $K(\gamma) := \gamma K_1 + (1-\gamma)K_2$ is the line-segment between K_1 and K_2 in (D.1) with $\epsilon = 0.1$.

D.2 Invertibility of the linear map $\mathcal{A}$

The invertibility of the map $\mathcal{A}$ is equivalent to the matrices A and $-A^T$ not having any common eigenvalues. If $\mathcal{A}$ is non-invertible, we can use $K^0 \in \mathcal{S}_K$ to introduce the change of variables $\hat{K} := K - K^0$ and $\hat{Y} := \hat{K}X$ and obtain $f(K) = \hat{h}(X, \hat{Y}) := \operatorname{trace}(Q^0 X + X^{-1}\hat{Y}^T R\,\hat{Y} + 2\hat{Y}^T R K^0)$ for all $K \in \mathcal{S}_K$, where $Q^0 := Q + (K^0)^T R\,K^0$. Moreover, X and $\hat{Y}$ satisfy the affine relation $\mathcal{A}_0(X) - \mathcal{B}(\hat{Y}) + \Omega = 0$, where $\mathcal{A}_0(X) := (A - BK^0)X + X(A - BK^0)^T$. Since the matrix $A - BK^0$ is Hurwitz, the map $\mathcal{A}_0$ is invertible. This allows us to write X as an affine function of $\hat{Y}$, $X(\hat{Y}) = \mathcal{A}_0^{-1}(\mathcal{B}(\hat{Y}) - \Omega)$. Since the function $\hat{h}(\hat{Y}) := \hat{h}(X(\hat{Y}), \hat{Y})$ has a similar form to $h(Y)$ except for the linear term $2\operatorname{trace}(\hat{Y}^T R\,K^0)$, the smoothness and strong convexity of $h(Y)$ established in Proposition 1 carry over to the function $\hat{h}(Y)$.

D.3 Proof of Proposition 1

The second-order term in the Taylor series expansion of $h(Y+\tilde{Y})$ around Y is given by [121, Lemma 2]

$$\left\langle \tilde{Y}, \nabla^2 h(Y;\tilde{Y}) \right\rangle \;=\; 2\,\|R^{\frac{1}{2}}(\tilde{Y} \;-\; K\tilde{X})X^{-\frac{1}{2}}\|_F^2 \tag{D.2}$$

where $\tilde{X}$ is the unique solution to $\mathcal{A}(\tilde{X}) = \mathcal{B}(\tilde{Y})$. We show that this term is upper and lower bounded by $L\|\tilde{Y}\|_F^2$ and $\mu\|\tilde{Y}\|_F^2$, where L and μ are given by (6.10a) and (6.10b), respectively. The proof for the upper bound is borrowed from [121, Lemma 1]; we include it for completeness. We repeatedly use the bounds on the variables presented in Lemma 15; see Appendix D.11.

Smoothness

For any $Y \in \mathcal{S}_Y(a)$ and $\tilde{Y}$ with $\|\tilde{Y}\|_F = 1$,

$$\begin{aligned}
\left\langle \tilde{Y}, \nabla^2 h(Y;\tilde{Y}) \right\rangle \;&=\; 2\|R^{\frac{1}{2}}(\tilde{Y} \;-\; K\tilde{X})X^{-\frac{1}{2}}\|_F^2 \;\le\; 2\|R\|_2\|X^{-1}\|_2\|\tilde{Y} \;-\; K\mathcal{A}^{-1}\mathcal{B}(\tilde{Y})\|_F^2 \\
&\le\; \frac{2\,\|R\|_2}{\lambda_{\min}(X)}\left(\|\tilde{Y}\|_F \;+\; \|K\|_2\|\mathcal{A}^{-1}\mathcal{B}\|_2\|\tilde{Y}\|_F\right)^2 \\
&\le\; \frac{2a\|R\|_2}{\nu}\left(1 \;+\; \frac{a\|\mathcal{A}^{-1}\mathcal{B}\|_2}{\sqrt{\nu\lambda_{\min}(R)}}\right)^2 =:\; L.
\end{aligned}$$

Here, the first and second inequalities are obtained from the definition of the 2-norm in conjunction with the triangle inequality, and the third inequality follows from (D.36b) and (D.36c). This completes the proof of smoothness.

Strong convexity

Using the positive definiteness of matrices R and X, the second-order term (D.2) can be lower bounded by

$$\left\langle \tilde{Y}, \nabla^2 h(Y;\tilde{Y}) \right\rangle \;\geq\; \frac{2\lambda_{\min}(R)\|H\|_F^2}{\|X\|_2} \tag{D.3}$$

where $H := \tilde{Y} - K\tilde{X}$. Next, we show that

$$\frac{\|H\|_F}{\|\tilde{X}\|_F} \;\geq\; \frac{\lambda_{\min}(\Omega)\lambda_{\min}(\Omega)}{a\,\|\mathcal{B}\|_2}. \tag{D.4}$$

We substitute $H + K\tilde{X}$ for $\tilde{Y}$ in $\mathcal{A}(\tilde{X}) = \mathcal{B}(\tilde{Y})$ to obtain

$$\Gamma \;=\; \mathcal{B}(H) \tag{D.5}$$

where $\Gamma := \mathcal{A}_K(\tilde{X})$. The closed-loop stability implies $\tilde{X} = \mathcal{A}_K^{-1}(\Gamma)$ and from Eq. (D.5) we have

$$\|H\|_F \;\geq\; \frac{\|\Gamma\|_F}{\|\mathcal{B}\|_2}. \tag{D.6}$$

This allows us to use Lemma 17, presented in Appendix D.12, to write

$$a\|\Gamma\|_F \;\geq\; \lambda_{\min}(\Omega)\lambda_{\min}(Q)\|\tilde{X}\|_F.$$

This inequality in conjunction with (D.6) yield (D.4). Next, we derive an upper bound on $\|\tilde{Y}\|_F$,

$$\|\tilde{Y}\|_F \;=\; \|H \,+\, K\tilde{X}\|_F \;\leq\; \|H\|_F \,+\, \|K\|_F\|\tilde{X}\|_F \;\leq\; \|H\|_F\left(1 \,+\, a^2\eta\right) \tag{D.7}$$

where η is given by (6.10c) and the second inequality follows from (D.36d) and (D.4). Finally, inequalities (D.3) and (D.7) yield

$$\frac{\left\langle \tilde{Y}, \nabla^2 f(Y;\tilde{Y}) \right\rangle}{\|\tilde{Y}\|_F^2} \;\geq\; \frac{2\lambda_{\min}(R)\|H\|_F^2}{\|X\|_2\|\tilde{Y}\|_F^2} \;\geq\; \frac{2\lambda_{\min}(R)}{\|X\|_2(1 \,+\, a^2\eta)^2} \;\geq\; \frac{2\lambda_{\min}(R)\lambda_{\min}(Q)}{a(1 \,+\, a^2\eta)^2} \;=:\; \mu \tag{D.8}$$

where the last inequality follows from (D.36a).

D.4 Proofs for Section 6.5

Proof of Lemma 1

The gradients are given by $\nabla f(K) = EX$ and $\nabla h(Y) = E + 2B^T(P - W)$, where $E := 2(RK - B^TP)$, P is determined by (6.6a), and W is the solution to (6.11b). Subtracting

the equation in (6.11b) from (6.6b) yields $A^T(P-W)+(P-W)A = -\frac{1}{2}\left(K^TE+E^TK\right)$, which in turn leads to

$$\|P-W\|_F \;\le\; \|\mathcal{A}^{-1}\|_2\|K\|_F\|E\|_F \;\le\; \frac{a\|\mathcal{A}^{-1}\|_2\|E\|_F}{\sqrt{\nu\lambda_{\min}(R)}}$$

where the second inequality follows from (D.36d) in Appendix D.11. Thus, by applying the triangle inequality to $\nabla h(Y)$, we obtain

$$\frac{\|\nabla h(Y)\|_F}{\|E\|_F} \;\le\; 1 \;+\; \frac{2a\|\mathcal{A}^{-1}\|_2\|B\|_2}{\sqrt{\nu\lambda_{\min}(R)}}.$$

Moreover, using the lower bound (D.36c) on $\lambda_{\min}(X)$, we have

$$\|\nabla f(K)\|_F \;=\; \|EX\|_F \ge (\nu/a)\|E\|_F.$$

Combining the last two inequalities completes the proof.

Proof of Lemma 2

For any pair of stabilizing feedback gains K and $\hat{K} := K+\tilde{K}$, we have [152, Eq. (2.10)], $f(\hat{K})-f(K) = \mathrm{trace}\big(\tilde{K}^T\big(R(K+\hat{K})-2B^T\hat{P}\big)X\big)$, where $X = X(K)$ and $\hat{P} = P(\hat{K})$ are given by (6.4a) and (6.6a), respectively. Letting $\hat{K} = K^\star$ in this equation and using the optimality condition $B^T\hat{P} = R\hat{K}$ completes the proof.

Proof of Lemma 3

We show that the second-order term $\left\langle \tilde{K}, \nabla^2 f(K;\tilde{K})\right\rangle$ in the Taylor series expansion of $f(K+\tilde{K})$ around K is upper bounded by $L_f\|\tilde{K}\|_F^2$ for all $K\in\mathcal{S}_K(a)$. From [168, Eq. (2.3)], it follows

$$\left\langle \tilde{K}, \nabla^2 f(K;\tilde{K})\right\rangle \;=\; 2\,\mathrm{trace}\,(\tilde{K}^TR\,\tilde{K}X - 2\tilde{K}^TB^T\tilde{P}X)$$

where $\tilde{P} = (\mathcal{A}_K^*)^{-1}(C)$ and $C := \tilde{K}^T(B^TP-RK)+(B^TP-RK)^T\tilde{K}$. Here, $X = X(K)$ and $P = P(K)$ are given by (6.4a) and (6.6a) respectively. Thus, using basic properties of the matrix trace and the triangle inequality, we have

$$\frac{\left\langle \tilde{K}, \nabla^2 f(K;\tilde{K})\right\rangle}{\|\tilde{K}\|_F^2} \;\le\; 2\|X\|_2\left(\|R\|_2 + \frac{2\|B\|_2\|\tilde{P}\|_F}{\|\tilde{K}\|_F}\right). \tag{D.9}$$

Now, we use Lemma 17 to upper bound the norm of $\tilde{P}$, $\|\tilde{P}\|_F \le a\|C\|_F/(\lambda_{\min}(\Omega)\lambda_{\min}(Q))$. Moreover, from the definition of C, the triangle inequality, and the submultiplicative property

of the 2-norm, we have $\|C\|_F \leq 2\|\tilde{K}\|_F(\|B\|_2\|P\|_2 + \|R\|_2\|K\|_2)$. Combining the last two inequalities gives

$$\frac{\|\tilde{P}\|_F}{\|\tilde{K}\|_F} \;\leq\; \frac{2a}{\lambda_{\min}(\Omega)\lambda_{\min}(Q)}\left(\|B\|_2\|P\|_2 + \|R\|_2\|K\|_2\right)$$

which in conjunction with (D.9) lead to

$$\frac{\left\langle \tilde{K}, \nabla^2 f(K;\tilde{K})\right\rangle}{\|\tilde{K}\|_F^2} \;\leq\; 2\|X\|_2\left(\|R\|_2 \;+\; \frac{4a}{\lambda_{\min}(\Omega)\lambda_{\min}(Q)}\left(\|B\|_2^2\|P\|_2 + \|B\|_2\|R\|_2\|K\|_2\right)\right).$$

Finally, we use the bounds provided in Appendix D.11 to obtain

$$\frac{\left\langle \tilde{K}, \nabla^2 f(K;\tilde{K})\right\rangle}{\|\tilde{K}\|_F^2} \;\leq\; \frac{2a\|R\|_2}{\lambda_{\min}(Q)} \;+\; \frac{8a^3}{\lambda_{\min}^2(Q)\lambda_{\min}(\Omega)}\left(\frac{\|B\|_2^2}{\lambda_{\min}(\Omega)} \;+\; \frac{\|B\|_2\|R\|_2}{\sqrt{\nu\lambda_{\min}(R)}}\right)$$

which completes the proof.

D.5 Proofs for Section 6.6.1.1

We first present two technical lemmas.

Lemma 1 *Let $Z \succ 0$ and let the Hurwitz matrix F satisfy*

$$\begin{bmatrix} \delta^2 I + F^T Z + ZF & Z \\ Z & -I \end{bmatrix} \;\prec\; 0. \tag{D.10}$$

Then $F + \delta\Delta$ is Hurwitz for all Δ with $\|\Delta\|_2 \leq 1$.

Proof: The matrix $F + \delta\Delta$ is Hurwitz if and only if the linear map from w to x with the state-space realization $\{\dot{x} = Fx + w + u,\, z = \delta x\}$ in feedback with $u = \Delta z$ is input-output stable. From the small-gain theorem [86, Theorem 8.2], this system is stable for all Δ in the unit ball if and only if the induced gain of the map $u \mapsto z$ with the state-space realization $\{\dot{x} = Fx + u,\, z = \delta x\}$ is smaller than one. The KYP Lemma [86, Lemma 7.4] implies that this norm condition is equivalent to (D.10). □

Lemma 2 *Let the matrices F, $X \succ 0$, and $\Omega \succ 0$ satisfy*

$$FX \;+\; XF^T \;+\; \Omega \;=\; 0. \tag{D.11}$$

Then the matrix $F + \Delta$ is Hurwitz for all Δ that satisfy $\|\Delta\|_2 < \lambda_{\min}(\Omega)/(2\|X\|_2)$.

Proof: From (D.11), we obtain that F is Hurwitz and $F\hat{X} + \hat{X}F^T + I \preceq 0$ where $\hat{X} := X/\lambda_{\min}(\Omega)$. Multiplication of this inequality from both sides by $\hat{X}^{-1}$ and division by 2 yields $ZF + F^T Z + 2\,Z^2 \preceq 0$ where $Z := (2\hat{X})^{-1}$. For any positive scalar $\delta < \lambda_{\min}(Z) = \lambda_{\min}(\Omega)/(2\|X\|_2)$ the last matricial inequality implies $\delta^2 I + ZF + F^T Z + Z^2 \prec 0$. The result follows from Lemma 1 by observing that the last inequality is equivalent to (D.10) via the use of Schur complement. □

Proof of Proposition 3

For any feedback gain $\hat{K}$ such that $\|\hat{K} - K\|_2 < \zeta$, the closed-loop matrix $A - B\hat{K}$ satisfies $\|A - B\hat{K} - (A - BK)\|_2 \leq \|K - \hat{K}\|_2 \|B\|_2 < \zeta\|B\|_2$. This bound on the distance between the closed-loop matrices $A - BK$ and $A - B\hat{K}$ allows us to apply Lemma 2 with $F := A - BK$ and $X := X(K)$ to complete the proof.

We next present a technical lemma.

Lemma 3 *For any $K \in \mathcal{S}_K$ and $\hat{K} \in \mathbb{R}^{m\times n}$ such that $\|\hat{K} - K\|_2 < \delta$, with*

$$\delta \;:=\; \frac{1}{4\,\|B\|_F}\,\min\left\{\frac{\lambda_{\min}(\Omega)}{\mathrm{trace}\,(X(K))},\,\frac{\lambda_{\min}(Q)}{\mathrm{trace}\,(P(K))}\right\}$$

the feedback gain matrix $\hat{K} \in \mathcal{S}_K$, and

$$\|X(\hat{K}) - X(K)\|_F \;\leq\; \epsilon_1\|\hat{K} - K\|_2 \tag{D.12a}$$
$$\|P(\hat{K}) - P(K)\|_F \;\leq\; \epsilon_2\|\hat{K} - K\|_2 \tag{D.12b}$$
$$\|\nabla f(\hat{K}) - \nabla f(K)\|_F \;\leq\; \epsilon_3\|\hat{K} - K\|_2 \tag{D.12c}$$
$$|f(\hat{K}) - f(K)| \;\leq\; \epsilon_4\|\hat{K} - K\|_2 \tag{D.12d}$$

where $X(K)$ and $P(K)$ are given by (6.4a) *and* (6.6a)*, respectively. Furthermore, the parameters ϵ_i which only depend on K and problem data are given by*

$$\begin{aligned}
\epsilon_1 &:= \|X(K)\|_2/\delta\\
\epsilon_2 &:= 2\,\mathrm{trace}(P)(2\,\|P\|_2\|B\|_F + (\delta + 2\|K\|_2)\|R\|_F)/\lambda_{\min}(Q)\\
\epsilon_4 &:= \epsilon_2\|\Omega\|_F\\
\epsilon_3 &:= 2(\epsilon_1\|K\|_2 + 2\|X(K)\|_2)\|R\|_F + 2\epsilon_1(\|P(K)\|_2 + 2\epsilon_2\|X(K)\|_2)\|B\|_F.
\end{aligned}$$

Proof: Note that $\delta \leq \zeta$, where ζ is given in Proposition 3. Thus, we can use Proposition 3 to show that $\hat{K} \in \mathcal{S}_K$. We next prove (D.12a). For K and $\hat{K} \in \mathcal{S}_K$, we can represent $X = X(K)$ and $\hat{X} = X(\hat{K})$ as the positive definite solutions to

$$(A - BK)X + X(A - BK)^T + \Omega = 0 \tag{D.13a}$$
$$(A - B\hat{K})\hat{X} + \hat{X}(A - B\hat{K})^T + \Omega = 0. \tag{D.13b}$$

Subtracting (D.13a) from (D.13b) and rearranging terms yield

$$(A - BK)\tilde{X} + \tilde{X}(A - BK)^T = B\tilde{K}\hat{X} + \hat{X}(B\tilde{K})^T$$

where $\tilde{X} := \hat{X} - X$ and $\tilde{K} := \hat{K} - K$. Now, we use Lemma 16, presented in Appendix D.12, with $F := A - BK$ to upper bound the norm of $\tilde{X} = \mathcal{F}(-B\tilde{K}\hat{X} - \hat{X}(B\tilde{K})^T)$, where the linear map $\mathcal{F}$ is defined in (D.40), as follows

$$\begin{aligned}
\|\tilde{X}\|_F &\leq \|\mathcal{F}\|_2 \|B\tilde{K}\hat{X} + \hat{X}(B\tilde{K})^T\|_F \leq \frac{\operatorname{trace}(X)}{\lambda_{\min}(\Omega)} \|B\tilde{K}\hat{X} + \hat{X}(B\tilde{K})^T\|_F \\
&\leq \frac{2\operatorname{trace}(X)\|B\|_F\|\tilde{K}\|_2}{\lambda_{\min}(\Omega)} \Big(\|X\|_2 + \|\tilde{X}\|_2\Big) \\
&\leq \frac{2\operatorname{trace}(X)\|B\|_F\|\tilde{K}\|_2\|X\|_2}{\lambda_{\min}(\Omega)} + \frac{1}{2}\|\tilde{X}\|_F. \qquad \text{(D.14)}
\end{aligned}$$

Here, the second inequality follows from Lemma 16, the third inequality follows from a combination of the sub-multiplicative property of the Frobenius norm and the triangle inequality, and the last inequality follows from $\|\tilde{K}\| \leq \delta$ and $\|\tilde{X}\|_2 \leq \|\tilde{X}\|_F$. Rearranging the terms in (D.14) completes the proof of (D.12a).

We next prove (D.12b). Similar to the proof of (D.12a), subtracting the Lyapunov equation (6.6b) from that of $\hat{P} = P(\hat{K})$ yields $(A - BK)^T\tilde{P} + \tilde{P}(A - BK) = W$ where $\tilde{P} := \hat{P} - P$ and $W := (B\tilde{K})^T\hat{P} + \hat{P}B\tilde{K} - \tilde{K}^T R\,\tilde{K} - \tilde{K}^T R\,K - K^T R\,\tilde{K}$. This allows us to use Lemma 16, presented in Appendix D.12, with $F := (A - BK)^T$ to upper bound the norm of $\tilde{P} = \mathcal{F}(-W)$, where the linear map $\mathcal{F}$ is defined in (D.40), as follows

$$\begin{aligned}
\|\tilde{P}\|_F &\leq \|\mathcal{F}\|_2\|W\|_F \leq \frac{\operatorname{trace}(\mathcal{F}(Q + K^T R K))}{\lambda_{\min}(Q + K^T R K)} \|W\|_F \\
&= \frac{\operatorname{trace}(P)}{\lambda_{\min}(Q + K^T R K)} \|W\|_F \leq \frac{\operatorname{trace}(P)}{\lambda_{\min}(Q)} \|W\|_F.
\end{aligned}$$

Here, the second inequality follows from Lemma 16. This inequality in conjunction with applying the triangle inequality to the definition of W yield

$$\begin{aligned}
\|\tilde{P}\|_F &\leq \frac{\operatorname{trace}(P)}{\lambda_{\min}(Q)} \times \\
&\Big(\|(B\tilde{K})^T\tilde{P} + \tilde{P}\,B\tilde{K}\|F + \|(B\tilde{K})^T P + P\,B\tilde{K} - \tilde{K}^T R\,\tilde{K} - \tilde{K}^T R\,K - K^T R\,\tilde{K}\|_F\Big) \\
&\leq \frac{\|\tilde{P}\|_F}{2} + \frac{\operatorname{trace}(P)}{\lambda_{\min}(Q)}\Big(2\|P\|_2\|B\|_F + (\delta + 2\|K\|_2)\|R\|_F\Big)\|\tilde{K}\|_2.
\end{aligned}$$

The second inequality is obtained by bounding the two terms on the left-hand side using basic properties of norm, where, for the first term, $\|\tilde{K}\|_2 \leq \delta \leq \lambda_{\min}(Q)/(4\|B\|_F \operatorname{trace}(P(K)))$ and, for the second term, $\|\tilde{K}\|_2 \leq \delta$. Rearranging the terms in above completes the proof of (D.12b).

We next prove (D.12c). It is straightforward to show that the gradient (6.5) satisfies

$$\tilde{\nabla} \;:=\; \nabla f(\hat{K}) - \nabla f(K) \;=\; 2R(\tilde{K}X + K\tilde{X} + \tilde{K}\tilde{X}) \;-\; 2B^T(\tilde{P}X + P\tilde{X} + \tilde{P}\tilde{X})$$

where $P := P(K)$ and $\tilde{P} := \hat{P} - P$. The triangle inequality in conjunction with $\|\tilde{X}\|_F \leq \epsilon_1\|\tilde{K}\|_2$, $\|\tilde{P}\|_F \leq \epsilon_2\|\tilde{K}\|_2$, and $\|\tilde{K}\|_2 < \delta$, yield $\|\tilde{\nabla}\|_F/\|\tilde{K}\|_2 \leq 2\|R\|_F\,(\|X\|_2 + \epsilon_1(\|K\|_2 + \delta)) + 2\|B\|_F\,(\epsilon_2\|X\|_2 + \epsilon_1(\|P\|_2 + \epsilon_2\delta))$. Rearranging terms completes the proof of (D.12c).

Finally, we prove (D.12d). Using the definitions of $f(K)$ in (6.3b) and $P(K)$ in (6.6a), it is easy to verify that $f(K) = \text{trace}\,(P(K)\Omega)$. Application of the Cauchy-Schwartz inequality yields $|f(\hat{K}) - f(K)| = |\text{trace}\,(\tilde{P}\Omega)| \leq \|\tilde{P}\|_F\|\Omega\|_F$, which completes the proof. □

Proof of Lemma 4

For any $K \in \mathcal{S}_K(a)$, we can use the bounds provided in Appendix D.11 to show that $c_1/a \leq \delta$ and $\epsilon_4 \leq c_2a^2$, where δ and ϵ_4 are given in Lemma 3 and each c_i is a positive constant that depends on the problem data. Now, Lemma 3 implies $f(K+r(a)U)-f(K) \leq \epsilon_4 r(a)\|U\|_2 \leq a$ where $r(a) := \min\{c_1, 1/c_2\}/(a\sqrt{mn})$. This inequality together with $f(K) \leq a$ complete the proof.

D.6 Proof of Proposition 4

We first present two technical lemmas.

Lemma 4 *Let the matrices F, $X \succ 0$, and $\Omega \succ 0$ satisfy $FX + XF^T + \Omega = 0$. Then, for any $t \geq 0$,*

$$\|\mathrm{e}^{Ft}\|_2^2 \;\leq\; (\|X\|_2/\lambda_{\min}(X))\,\mathrm{e}^{-(\lambda_{\min}(\Omega)/\|X\|_2)\,t}.$$

Proof: The function $V(x) := x^TXx$ is a Lyapunov function for $\dot{x} = F^Tx$ because $\dot{V}(x) = -x^T\Omega x \leq -cV(x)$, where $c := \lambda_{\min}(\Omega)/\|X\|_2$. For any initial condition x_0, this inequality together with the comparison lemma [157, Lemma 3.4] yield $V(x(t)) \leq V(x_0)\,\mathrm{e}^{-ct}$. Noting that $x^T(t) = x_0^T\mathrm{e}^{Ft}$, we let x_0 be the normalized left singular vector associated with the maximum singular value of e^{Ft} to obtain

$$\|\mathrm{e}^{Ft}\|_2^2 \;=\; \|x(t)\|^2 \;\leq\; \frac{V(x(t))}{\lambda_{\min}(X)} \;\leq\; \frac{V(x_0)}{\lambda_{\min}(X)}\,\mathrm{e}^{-ct}$$

which along with $V(x_0) \leq \|X\|_2$ complete the proof. □

Lemma 5 establishes an exponentially decaying upper bound on the difference between $f_{x_0}(K)$ and $f_{x_0,\tau}(K)$ over any sublevel set $\mathcal{S}_K(a)$ of the LQR objective function $f(K)$.

Lemma 5 *For any $K \in \mathcal{S}_K(a)$ and $v \in \mathbb{R}^n$, $|f_v(K) - f_{v,\tau}(K)| \leq \|v\|^2\kappa_1(a)\mathrm{e}^{-\kappa_2(a)\tau}$, where the positive functions $\kappa_1(a)$ and $\kappa_2(a)$, given by* (D.17)*, depend on problem data.*

Proof: Since $x(t) = \mathrm{e}^{(A-BK)t}v$ is the solution to (6.1b) with $u = -Kx$ and the initial condition $x(0) = v$, it is easy to verify that $f_{v,\tau}(K) = \operatorname{trace}\big((Q + K^T R\, K)\, X_{v,\tau}(K)\big)$ and $f_v(K) = \operatorname{trace}\big((Q\, +\, K^T R\, K)\, X_v(K)\big)$, where

$$X_{v,\tau}(K) \;:=\; \int_0^\tau \mathrm{e}^{(A-BK)t}\, vv^T \mathrm{e}^{(A-BK)^T t}\, \mathrm{d}t$$

and $X_v := X_{v,\infty}$. Using the triangle inequality, we have

$$\|X_v(K) - X_{v,\tau}(K)\|_F \;\le\; \|v\|^2 \int_\tau^\infty \|\mathrm{e}^{(A-BK)t}\|_2^2\, \mathrm{d}t. \tag{D.15}$$

Equation (6.4b) allows us to use Lemma 4 with $F := A - BK$, $X := X(K)$ to upper bound $\|\mathrm{e}^{(A-BK)t}\|_2,\ \lambda_{\min}(X)\|\mathrm{e}^{(A-BK)t}\|_2^2 \;\le\; \|X\|_2\, \mathrm{e}^{-(\lambda_{\min}(\Omega)/\|X\|_2)\, t}$. Integrating this inequality over $[\tau, \infty]$ in conjunction with (D.15) yield

$$\|X_v(K) - X_{v,\tau(K)}\|_F \;\le\; \|v\|^2 \kappa_1'\, \mathrm{e}^{-\kappa_2'\tau} \tag{D.16}$$

where $\kappa_1' := \|X(K)\|_2^2/(\lambda_{\min}(\Omega)\lambda_{\min}(X(K)))$ and $\kappa_2' := \lambda_{\min}(\Omega)/\|X(K)\|_2$. Furthermore,

$$\begin{aligned} |f_v(K) - f_{v,\tau}(K)| \;&=\; \big|\operatorname{trace}\big((Q + K^T R\, K)\,(X_v - X_{v,\tau})\big)\big| \;\le \\ &(\|Q\|_F + \|R\|_2\|K\|_F^2)\|X_v - X_{v,\tau}\|_F \;\le\; \|v\|^2(\|Q\|_F + \|R\|_2\|K\|_F^2)\, \kappa_1' \mathrm{e}^{-\kappa_2'\tau} \end{aligned}$$

where we use the Cauchy-Schwartz and triangle inequalities for the first inequality and (D.16) for the second inequality. Combining this result with the bounds on the variables provided in Lemma 15 completes the proof with

$$\kappa_1(a) \;:=\; \left(\|Q\|_F + \frac{a^2\|R\|_2}{\nu\lambda_{\min}(R)}\right) \frac{a^3}{\nu\lambda_{\min}(\Omega)\lambda_{\min}^2(Q)} \tag{D.17a}$$

$$\kappa_2(a) \;:=\; \lambda_{\min}(\Omega)\lambda_{\min}(Q)/a \tag{D.17b}$$

where the constant ν is given by (6.10d). □

Proof of Proposition 4

Since $K \in \mathcal{S}_K(a)$ and $r \le r(a)$, Lemma 4 implies that $K \pm rU_i \in \mathcal{S}_K(2a)$. Thus, $f_{x_i}(K \pm rU_i)$ is well defined for $i = 1, \ldots, N$, and

$$\begin{aligned} \widetilde{\nabla} f(K) - \overline{\nabla} f(K) \;&=\; \frac{1}{2rN} \times \\ &\Big(\sum_i \big(f_{x_i}(K + rU_i) - f_{x_i,\tau}(K + rU_i)\big)U_i - \sum_i \big(f_{x_i}(K - rU_i) - f_{x_i,\tau}(K - rU_i)\big)U_i\Big). \end{aligned}$$

Furthermore, since $K \pm rU_i \in \mathcal{S}_K(2a)$, we can use triangle inequality and apply Lemma 5, $2N$ times, to bound each term individually and obtain

$$\|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F \;\leq\; (\sqrt{mn}/r) \max_i \|x_i\|^2 \kappa_1(2a) \mathrm{e}^{-\kappa_2(2a)\tau}$$

where we used $\|U_i\|_F = \sqrt{mn}$. This completes the proof.

D.7 Proof of Proposition 5

We first establish bounds on the smoothness parameter of $\nabla f(K)$. For $J \in \mathbb{R}^{m\times n}$, $v \in \mathbb{R}^n$, and $f_v(K)$ given by (6.24a), let $j_v(K) := \langle J, \nabla^2 f_v(K; J)\rangle$, denote the second-order term in the Taylor series expansion of $f_v(K+J)$ around K. Following similar arguments as in [168, Eq. (2.3)] leads to $j_v(K) = 2\,\mathrm{trace}\,(J^T(RJ - 2B^T D)X_v)$, where X_v and D are the solutions to

$$\begin{aligned} \mathcal{A}_K(X_v) \;&=\; -vv^T && \text{(D.18a)} \\ \mathcal{A}_K^*(D) \;&=\; J^T(B^T P - R\,K) + (B^T P - R\,K)^T J && \text{(D.18b)} \end{aligned}$$

and P is given by (6.6a). The following lemma provides an analytical expression for the gradient $\nabla j_v(K)$.

Lemma 6 *For any $v \in \mathbb{R}^n$ and $K \in \mathcal{S}_K$, $\nabla j_v(K) = 4\big(B^T W_1 X_v + (RJ - B^T D)W_2 + (RK - B^T P)W_3\big)$, where W_i are the solutions to the linear equations*

$$\begin{aligned} \mathcal{A}_K^*(W_1) \;&=\; J^T R\,J - J^T B^T D - DBJ && \text{(D.19a)} \\ \mathcal{A}_K(W_2) \;&=\; BJX_v + X_v J^T B^T && \text{(D.19b)} \\ \mathcal{A}_K(W_3) \;&=\; BJ\,W_2 + W_2 J^T B^T. && \text{(D.19c)} \end{aligned}$$

Proof: We expand $j_v(K + \epsilon\tilde{K})$ around K and to obtain

$$j_v(K+\epsilon\tilde{K}) - j_v(K) \;=\; 2\,\epsilon\,\mathrm{trace}\,(J^T(R\,J - 2B^T D)\tilde{X}_v) - 4\,\epsilon\,\mathrm{trace}\,(J^T B^T \tilde{D} X_v) + o(\epsilon).$$

Here, $o(\epsilon)$ denotes higher-order terms in ϵ, whereas $\tilde{X}_v$, $\tilde{D}$, and $\tilde{P}$ are obtained by perturbing Eqs. (D.18a), (D.18b), and (6.6b), respectively,

$$\begin{aligned} \mathcal{A}_K(\tilde{X}_v) \;&=\; B\tilde{K}X_v + X_v\tilde{K}^T B^T && \text{(D.20a)} \\ \mathcal{A}_K^*(\tilde{D}) \;&=\; \tilde{K}^T B^T D + DB\tilde{K} \;+\; \mathcal{A}_K^*(\tilde{D}) \;=\; J^T(B^T\tilde{P} - R\tilde{K}) + (B^T\tilde{P} - R\tilde{K})^T J && \text{(D.20b)} \\ \mathcal{A}_K^*(\tilde{P}) \;&=\; \tilde{K}^T B^T P + PB\tilde{K} - K^T R\tilde{K} - \tilde{K}^T RK. && \text{(D.20c)} \end{aligned}$$

Applying the adjoint identity on Eqs. (D.20a) and (D.20b) yields

$$\begin{aligned} j_v(K+\epsilon\tilde{K}) - j_v(K) \;&\approx\; 2\epsilon\,\mathrm{trace}\,((B\tilde{K}X_v + X_v\tilde{K}^TB^T)W_1) \\ &\quad - 2\epsilon\,\mathrm{trace}\,((\tilde{K}^TB^TD + DB\tilde{K} + J^T(B^T\tilde{P} - R\tilde{K}) + (B^T\tilde{P} - R\tilde{K})^TJ)W_2) \\ &=\; 4\epsilon\,\mathrm{trace}\,(\tilde{K}^TB^TW_1X_v) - 4\epsilon\,\mathrm{trace}\,(\tilde{K}^T(B^TD - RJ)W_2) - 4\epsilon\,\mathrm{trace}\,(W_2J^TB^T\tilde{P}) \end{aligned}$$

where we have neglected $o(\epsilon)$ terms, and W_1 and W_2 are given by (D.19a) and (D.19b), respectively. Moreover, the adjoint identity applied to (D.20c) allows us to simplify the last term as,

$$2\,\mathrm{trace}\,(W_2J^TB^T\tilde{P}) \;=\; \mathrm{trace}\,((\tilde{K}^TB^TP + PB\tilde{K} - K^TR\tilde{K} - \tilde{K}^TRK)W_3)$$

where W_3 is given by (D.19c). Finally, this yields

$$j(K+\epsilon\tilde{K}) - j(K) \;\approx\; 4\epsilon\,\mathrm{trace}\,(\tilde{K}^T((RK - B^TP)W_3 + B^TW_1X_v + (RJ - B^TD)W_2)).$$

□

We next establish a bound on $\|\nabla j_v(K)\|_F$.

Lemma 7 *Let $K, K' \in \mathbb{R}^{m\times n}$ be such that the line segment $K + t(K' - K)$ with $t \in [0,1]$ belongs to $\mathcal{S}_K(a)$ and let $J \in \mathbb{R}^{m\times n}$ and $v \in \mathbb{R}^n$ be fixed. Then, the function $j_v(K)$ satisfies $|j_v(K_1) - j_v(K_2)| \le \ell(a)\|J\|_F^2\|v\|^2\|K_1 - K_2\|_F$, where $l(a)$ is a positive function given by*

$$\ell(a) \;:=\; ca^2 \;+\; c'a^4 \tag{D.21}$$

and c, c' are positive scalars that depend only on problem data.

Proof: We show that the gradient $\nabla j_v(K)$ given by Lemma 6 is upper bounded by $\|\nabla j_v(K)\|_F \;\le\; \ell(a)\|J\|_F^2\|v\|^2$. Applying Lemma 17 on (D.18), the bounds in Lemma 15, and the triangle inequality, we have $\|X_v\|_F \le c_1a\|v\|^2$ and $\|D\|_F \le c_2a^2\|J\|_F$, where c_1 and c_2 are positive constants that depend on problem data. We can use the same technique to bound the norms of W_i in Eq. (D.19), $\|W_1\|_F \le (c_3a + c_4a^3)\|J\|_F^2$, $\|W_2\|_F \le c_5a^2\|v\|^2\|J\|_F$, $\|W_3\|_F \;\le\; c_6a^3\|v\|^2\|J\|_F^2$, where $c_3, \ldots, c_6$ are positive constants that depend on problem data. Combining these bounds with the Cauchy-Schwartz and triangle inequalities applied to $\nabla f_v(K)$ completes the proof. □

D.7.1 Proof of Proposition 5

Since $r \le r(a)$, Lemma 4 implies that $K \pm sU \in \mathcal{S}_K(2a)$ for all $s \le r$. Also, the mean-value theorem implies that, for any $U \in \mathbb{R}^{m\times n}$ and $v \in \mathbb{R}^n$,

$$f_v(K \pm rU) \;=\; f_v(K) \;\pm\; r\,\langle \nabla f_v(K), U\rangle \;+\; \frac{r^2}{2}\,\big\langle U, \nabla^2 f_v(K \pm\, s_\pm\, U; U)\big\rangle$$

where $s_\pm \in [0, r]$ are constants that depend on K and U. Now, if $\|U\|_F = \sqrt{mn}$, the above identity yields

$$\begin{aligned}\frac{1}{2r}(f_v(K + rU) - f_v(K - rU)) - \langle \nabla f_v(K), U\rangle \\ = \frac{r}{4}\left(\left\langle U, \nabla^2 f_v(K + s_+U; U)\right\rangle - \left\langle U, \nabla^2 f_v(K - s_-U; U)\right\rangle\right) \\ \leq \frac{r}{4}(s_+ + s_-)\|U\|_F^3\,\ell(2a)\,\|v\|^2 \leq \frac{r^2}{2}mn\sqrt{mn}\,\ell(2a)\,\|v\|^2\end{aligned}$$

where the first inequality follows from Lemma 7. Combining this inequality with the triangle inequality applied to the definition of $\widehat{\nabla} f(K) - \widetilde{\nabla} f(K)$ completes the proof.

D.8 Proof of Proposition 6

From inequality (6.28a), it follows that G is a descent direction of the function $f(K)$. Thus, we can use the descent lemma [158, Eq. (9.17)] to show that $K^+ := K - \alpha G$ satisfies

$$f(K^+) - f(K) \leq (L_f\alpha^2/2)\,\|G\|_F^2 - \alpha\,\langle \nabla f(K), G\rangle \tag{D.22}$$

for any α for which the line segment between K^+ and K lies in $\mathcal{S}_K(a)$. Using (6.28), for any $\alpha \in [0, 2\mu_1/(\mu_2 L_f)]$, we have

$$(L_f\alpha^2/2)\,\|G\|_F^2 - \alpha\,\langle \nabla f(K), G\rangle \leq (\alpha\,(L_f\mu_2\alpha - 2\mu_1)/2)\,\|\nabla f(K)\|_F^2 \leq 0 \tag{D.23}$$

and the right-hand side of inequality (D.22) is nonpositive for $\alpha \in [0, 2\mu_1/(\mu_2 L_f)]$. Thus, we can use the continuity of the function $f(K)$ along with inequalities (D.22) and (D.23) to conclude that $K^+ \in \mathcal{S}_K(a)$ for all $\alpha \in [0, 2\mu_1/(\mu_2 L_f)]$, and

$$f(K^+) - f(K) \leq (\alpha\,(L_f\mu_2\alpha - 2\mu_1)/2)\,\|\nabla f(K)\|_F^2.$$

Combining this inequality with the PL condition (6.18), it follows that

$$f(K^+) - f(K) \leq -(\mu_1\alpha/2)\,\|\nabla f(K)\|_F^2 \leq -\mu_f\mu_1\alpha\,(f(K) - f(K^\star))$$

for all $\alpha \in [0, c_1/(c_2 L_f)]$. Subtracting $f(K^\star)$ and rearranging terms complete the proof.

D.9 Proofs of Section 6.6.2.1

We first present two technical results. Lemma 8 extends [164, Theorem 3.2] on the norm of Gaussian matrices presented in Appendix D.10 to random matrices with uniform distribution on the sphere $\sqrt{mn}\,S^{mn-1}$.

Lemma 8 *Let $E \in \mathbb{R}^{m\times n}$ be a fixed matrix and let $U \in \mathbb{R}^{m\times n}$ be a random matrix with* $\mathrm{vec}(U)$ *uniformly distributed on the sphere $\sqrt{mn}\,S^{mn-1}$. Then, for any $s \geq 1$ and $t \geq 1$,*

we have $\mathbb{P}(\mathsf{B}) \leq 2\mathrm{e}^{-s^2q-t^2n} + \mathrm{e}^{-mn/8}$*, where* $\mathsf{B} := \left\{\|E^TU\|_2 > c'\left(s\|E\|_F + t\sqrt{n}\|E\|_2\right)\right\}$*, and* $q := \|E\|_F^2/\|E\|_2^2$ *is the stable rank of* E*.*

Proof: For a matrix G with i.i.d. standard normal entries, we have

$$\|E^TU\|_2 \;\sim\; \sqrt{mn}\|E^TG\|_2/\|G\|_F.$$

Let the constant κ be the ψ_2-norm of the standard normal random variable and let us define two auxiliary events,

$$\begin{aligned}\mathsf{C}_1 \;&:=\; \{\sqrt{mn} > 2\|G\|_F\}\\ \mathsf{C}_0 \;&:=\; \{\sqrt{mn}\,\|E^TG\|_2 > 2c\kappa^2\|G\|_F\left(s\|E\|_F + t\sqrt{n}\|E\|_2\right)\}.\end{aligned}$$

For $c' := 2c\kappa^2$, we have $\mathbb{P}(\mathsf{B}) = \mathbb{P}(\mathsf{C}_0) \leq \mathbb{P}(\mathsf{C}_1 \cup \mathsf{A}) \leq \mathbb{P}(\mathsf{C}_1) + \mathbb{P}(\mathsf{A})$, where the event A is given by Lemma 13. Here, the first inequality follows from $\mathsf{C}_0 \subset \mathsf{C}_1 \cup \mathsf{A}$ and the second follows from the union bound. Now, since $\|\cdot\|_F$ is Lipschitz continuous with parameter 1, from the concentration of Lipschitz functions of standard normal Gaussian vectors [160, Theorem 5.2.2], it follows that $\mathbb{P}(C_1) \leq \mathrm{e}^{-mn/8}$. This in conjunction with Lemma 13 complete the proof. □

Lemma 9 *In the setting of Lemma 8, we have* $\mathbb{P}\left\{\|E^TU\|_F > 2\sqrt{n}\,\|E\|_F\right\} \leq \mathrm{e}^{-n/2}$.

Proof: We begin by observing that $\|E^TU\|_F = \|\mathrm{vec}(E^TU)\|_F = \|\left(I \otimes E^T\right)\mathrm{vec}(U)\|_F$, where $\otimes$ denotes the Kronecker product. Thus, it is easy to verify that $\|E^TU\|_F$ is a Lipschitz continuous function of U with parameter $\|I \otimes E^T\|_2 = \|E\|_2$. Now, from the concentration of Lipschitz functions of uniform random variables on the sphere $\sqrt{mn}\,S^{mn-1}$ [160, Theorem 5.1.4], for all $t > 0$, we have $\mathbb{P}\left\{\|E^TU\|_F > \sqrt{\mathbb{E}\left[\|E^TU\|_F^2\right]} + t\right\} \leq \mathrm{e}^{-t^2/(2\|E\|_2^2)}$. Now, since

$$\begin{aligned}\mathbb{E}\left[\|E^TU\|_F^2\right] \;&=\; \mathbb{E}\left[\|(I \otimes E^T)\,\mathrm{vec}(U)\|_F^2\right]\\ &=\; \mathbb{E}\left[\mathrm{trace}\left((I \otimes E^T)\mathrm{vec}(U)\mathrm{vec}(U)^T(I \otimes E)\right)\right]\\ &=\; \mathrm{trace}\left((I \otimes E^T)(I \otimes E)\right) \;=\; n\|E\|_F^2\end{aligned}$$

we can rewrite the last inequality for $t = \sqrt{n}\|E\|_F$ to obtain

$$\mathbb{P}\left\{\|E^TU\|_F > 2\sqrt{n}\,\|E\|_F\right\} \;\leq\; \mathrm{e}^{-n\|E\|_F^2/(2\|E\|_2^2)} \;\leq\; \mathrm{e}^{-n/2}$$

where the last inequality follows from $\|E\|_F \geq \|E\|_2$. □

Proof of Lemma 5

We define the auxiliary events

$$\mathsf{D}_i \;:=\; \{\|\mathcal{M}^*(E^T U_i)\|_2 \le c\sqrt{n}\,\|M^*\|_S\|E\|_F\} \cap \{\|\mathcal{M}^*(E^T U_i)\|_F \;\le\; 2\sqrt{n}\|M^*\|_2\|E\|_F\}$$

for $i = 1, \ldots, N$. Since

$$\|\mathcal{M}^*(E^T U_i)\|_2 \le \|\mathcal{M}^*\|_S\|E^T U_i\|_2$$

and

$$\|\mathcal{M}^*(E^T U_i)\|_F \le \|\mathcal{M}^*\|_2\|E^T U_i\|_F$$

we have

$$\mathbb{P}(\mathsf{D}_i) \;\ge\; \mathbb{P}\left(\{\|E^T U_i\|_2 \;\le\; c\sqrt{n}\|E\|_F\} \;\cap\; \{\|E^T U_i\|_F \le 2\sqrt{n}\|E\|_F\}\right).$$

Applying Lemmas 8 and 9 to the right-hand side of the above events together with the union bound yield $\mathbb{P}(\mathsf{D}_i^c) \le 2\mathrm{e}^{-n} + \mathrm{e}^{-mn/8} + \mathrm{e}^{-n/2} \le 4\mathrm{e}^{-n/8}$, where D_i^c is the complement of D_i. This in turn implies

$$\mathbb{P}(\mathsf{D}^c) \;=\; \mathbb{P}(\bigcup_{i=1}^{N} \mathsf{D}_i^c) \;\le\; \sum_{i=1}^{N} \mathbb{P}(\mathsf{D}_i^c) \;\le\; 4N\mathrm{e}^{-\frac{n}{8}} \tag{D.24}$$

where $\mathsf{D} := \cap_i \mathsf{D}_i$. We can now use the conditioning identity to bound the failure probability,

$$\begin{aligned}
\mathbb{P}\{|a| > b\} \;&=\; \mathbb{P}\left\{|a| > b \,\middle|\, \mathsf{D}\right\}\mathbb{P}(\mathsf{D}) + \mathbb{P}\left\{|a| > b \,\middle|\, \mathsf{D}^c\right\}\mathbb{P}(\mathsf{D}^c) \\
&\le\; \mathbb{P}\left\{|a| > b \,\middle|\, \mathsf{D}\right\}\mathbb{P}(\mathsf{D}) + \mathbb{P}(\mathsf{D}^c) \\
&=\; \mathbb{P}\left\{|a\,\mathbb{1}_{\mathsf{D}}| \;> b\right\} \;+\; \mathbb{P}(\mathsf{D}^c) \\
&\le\; \mathbb{P}\left\{|a\,\mathbb{1}_{\mathsf{D}}| > b\right\} \;+\; 4N\mathrm{e}^{-n/8}
\end{aligned} \tag{D.25}$$

where

$$\begin{aligned}
a \;&:=\; (1/N)\sum_i \langle E\left(X_i - X\right), U_i\rangle \langle EX, U_i\rangle \\
b \;&:=\; \delta\|EX\|_F\|E\|_F
\end{aligned}$$

and $\mathbb{1}_{\mathsf{D}}$ is the indicator function of D. It is now easy to verify that

$$\mathbb{P}\{|a\,\mathbb{1}_{\mathsf{D}}| > b\} \;\le\; \mathbb{P}\{|Y| > b\}$$

where

$$\begin{aligned} Y &:= (1/N)\sum_i Y_i \\ Y_i &:= \langle E(X_i - X), U_i\rangle \langle EX, U_i\rangle \mathbb{1}_{\mathsf{D}_i}. \end{aligned}$$

The rest of the proof uses the $\psi_{1/2}$-norm of Y to establish an upper bound on $\mathbb{P}\{|Y| > b\}$.

Since Y_i are linear in the zero-mean random variables $X_i - X$, we have $\mathbb{E}[Y_i|U_i] = 0$. Thus, the law of total expectation yields $\mathbb{E}[Y_i] = \mathbb{E}[\mathbb{E}[Y_i|U_i]] = 0$. Therefore, Lemma 14 implies

$$\|Y\|_{\psi_{1/2}} \leq (c'/\sqrt{N})(\log N) \max_i \|Y_i\|_{\psi_{1/2}}. \tag{D.26}$$

Now, using the standard properties of the ψ_α-norm, we have

$$\begin{aligned} \|Y_i\|_{\psi_{1/2}} &\leq c''\| \langle E(X_i - X), U_i\rangle \mathbb{1}_{\mathsf{D}_i}\|_{\psi_1} \| \langle EX, U_i\rangle \|_{\psi_1} \\ &\leq c'''\| \langle E(X_i - X), U_i\rangle \mathbb{1}_{\mathsf{D}_i}\|_{\psi_1} \|EX\|_F \end{aligned} \tag{D.27}$$

where the second inequality follows from [160, Theorem 3.4.6],

$$\| \langle EX, U_i\rangle \|_{\psi_1} \leq \| \langle EX, U_i\rangle \|_{\psi_2} \leq c_0 \|EX\|_F. \tag{D.28}$$

We can now use

$$\begin{aligned} \langle E(X_i - X), U_i\rangle &= \langle X_i - X, E^T U_i\rangle \\ &= \langle \mathcal{M}(x_i x_i^T), E^T U_i\rangle - \langle \mathcal{M}(I), E^T U_i\rangle \\ &= x_i^T \mathcal{M}^*(E^T U_i) x_i - \operatorname{trace}(\mathcal{M}^*(E^T U_i)) \end{aligned}$$

to bound the right-hand side of (D.27). This identity allows us to use the Hanson-Write inequality (Lemma 12) to upper bound the conditional probability

$$\mathbb{P}\left\{|\langle E(X_i - X), U_i\rangle| > t \,\middle|\, U_i\right\} \leq 2\mathrm{e}^{-\hat{c}\min\{\frac{t^2}{\kappa^4\|\mathcal{M}^*(E^T U_i)\|_F^2}, \frac{t}{\kappa^2\|\mathcal{M}^*(E^T U_i)\|_2}\}}.$$

Thus, we have

$$\begin{aligned} \mathbb{P}\{|\langle E(X_i - X), U_i\rangle \mathbb{1}_{\mathsf{D}_i}| > t\} &= \mathbb{E}_{U_i}\left[\mathbb{1}_{\mathsf{D}_i} \mathbb{E}_{x_i}\left[\mathbb{1}_{\{|\langle E(X_i - X), U_i\rangle| > t\}}\right]\right] \\ &= \mathbb{E}_{U_i}\left[\mathbb{1}_{\mathsf{D}_i} \mathbb{P}\left\{|\langle E(X_i - X), U_i\rangle| > t \,\middle|\, U_i\right\}\right] \\ &\leq \mathbb{E}_{U_i}\left[\mathbb{1}_{\mathsf{D}_i} 2\mathrm{e}^{-\hat{c}\min\{\frac{t^2}{\kappa^4\|\mathcal{M}^*(E^T U_i)\|_F^2}, \frac{t}{\kappa^2\|\mathcal{M}^*(E^T U_i)\|_2}\}}\right] \\ &\leq 2\mathrm{e}^{-\hat{c}\min\{\frac{t^2}{4n\kappa^4\|\mathcal{M}^*\|_2^2\|E\|_F^2}, \frac{t}{c\sqrt{n}\kappa^2\|\mathcal{M}^*\|_S\|E\|_F}\}} \end{aligned}$$

where the definition of D_i was used to obtain the last inequality. The above tail bound implies [169, Lemma 11]

$$\| \langle E\left(X_i - X\right), U_i \rangle \mathbb{1}_{\mathsf{D}_i} \|_{\psi_1} \;\leq\; \tilde{c}\kappa^2\sqrt{n}(\|\mathcal{M}^*\|_2 + \|\mathcal{M}^*\|_S)\|E\|_F. \tag{D.29}$$

Using (6.30), it is easy to obtain the lower bound on the number of samples,

$$N \;\geq\; C'\left(\beta^2\kappa^2/\delta\right)^2(\|\mathcal{M}^*\|_2 \,+\, \|\mathcal{M}^*\|_S)^2\, n\, \log^6 N.$$

We can now combine (D.26), (D.29) and (D.27) to obtain

$$\|Y\|_{\psi_{1/2}} \;\leq\; C'\kappa^2\frac{\sqrt{n}\log N}{\sqrt{N}}(\|\mathcal{M}^*\|_2 + \|\mathcal{M}^*\|_S)\|E\|_F\|EX\|_F \;\leq\; \frac{\delta}{\beta^2\log^2 N}\,\|E\|_F\|EX\|_F$$

where the last inequality follows from the above lower bound on N. Combining this inequality and (D.35) with $t := \delta\|E\|_F\|EX\|_F/\|Y\|_{\psi_{1/2}}$ yields $\mathbb{P}\{|Y| > \delta\|E\|_F\|EX\|_F\} \leq 1/N^\beta$, which completes the proof.

Proof of Lemma 6

The marginals of a uniform random variable have bounded sub-Gaussian norm (see the inequality in (D.28)). Thus, [160, Lemma 2.7.6] implies

$$\| \langle W, U_i\rangle^2 \|_{\psi_1} \;=\; \| \langle W, U_i\rangle \|_{\psi_2}^2 \;\leq\; \hat{c}\|W\|_F^2$$

which together with the triangle inequality yield

$$\| \langle W, U_i\rangle^2 - \|W\|_F^2\|_{\psi_1} \;\leq\; c'\|W\|_F^2.$$

Now since $\langle W, U_i\rangle^2 - \|W\|_F^2$ are zero-mean and independent, we can apply the Bernstein inequality (Lemma 11) to obtain

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^N \langle W, U_i\rangle^2 - \|W\|_F^2\right| > t\|W\|_F^2\right\} \;\leq\; 2\mathrm{e}^{-cN\min\{t^2,t\}} \tag{D.30}$$

which together with the triangle inequality complete the proof.

D.10 Proofs for Section 6.6.2.2 and probabilistic toolbox

We first present a technical lemma.

Lemma 10 *Let $v_1, \ldots, v_N \in \mathbb{R}^d$ be i.i.d. random vectors uniformly distributed on the sphere $\sqrt{d}\,S^{d-1}$ and let $a \in \mathbb{R}^d$ be a fixed vector. Then, for any $t \geq 0$, we have*

$$\mathbb{P}\left\{\frac{1}{N}\|\sum_{j=1}^N \langle a, v_j\rangle\, v_j\| \;>\; (c + c\frac{\sqrt{d}+t}{\sqrt{N}})\|a\|\right\} \;\leq\; 2\mathrm{e}^{-t^2} + N\mathrm{e}^{-d/8} + 2\mathrm{e}^{-\hat{c}N}.$$

Proof: It is easy to verify that $\sum_j \langle a, v_j \rangle v_j = Vv$, where $V := [v_1 \cdots v_n] \in \mathbb{R}^{d \times N}$ is the random matrix with the jth column given by v_j and $v := V^T a \in \mathbb{R}^N$. Thus,

$$\| \sum_j \langle a, v_j \rangle v_j \| \;=\; \|Vv\| \;\leq\; \|V\|_2 \|v\|.$$

Now, let $G \in \mathbb{R}^{d \times N}$ be a random matrix with i.i.d. standard normal Gaussian entries and let $\hat{G} \in \mathbb{R}^{d \times N}$ be a matrix obtained by normalizing the columns of G as $\hat{G}_j := \sqrt{d}\, G_j / \|G_j\|$, where G_j and $\hat{G}_j$ are the jth columns of G and $\hat{G}$, respectively. From the concentration of norm of Gaussian vectors [160, Theorem 5.2.2], we have $\|G_j\| \geq \sqrt{d}/2$ with probability not smaller than $1 - \mathrm{e}^{-d/8}$. This in conjunction with a union bound yield $\|\hat{G}\|_2 \leq 2\|G\|_2$ with probability not smaller than $1 - N\mathrm{e}^{-d/8}$. Furthermore, from the concentration of Gaussian matrices [160, Theorem 4.4.5], we have $\|G\|_2 \leq C(\sqrt{N} + \sqrt{d} + t)$ with probability not smaller than $1 - 2\mathrm{e}^{-t^2}$. By combining this inequality with the above upper bound on $\|\hat{G}\|_2$, and using $V \sim \hat{G}$ in conjunction with a union bound, we obtain

$$\|V\|_2 \;\leq\; 2C(\sqrt{N} \,+\, \sqrt{d} \,+\, t) \tag{D.31}$$

with probability not smaller than $1 - 2\mathrm{e}^{-t^2} - N\mathrm{e}^{-d/8}$. Moreover, using (D.30) in the proof of Lemma 6, gives $\|v\| \leq C'\sqrt{N}\|a\|$ with probability not smaller than $1 - 2\mathrm{e}^{-\hat{c}N}$. Combining this inequality with (D.31) and employing a union bound complete the proof. □

Proof of Lemma 7

We begin by noting that

$$\| \sum_{i=1}^{N} \langle E\left(X_i - X\right), U_i \rangle U_i \|_F \;=\; \|Uu\| \;\leq\; \|U\|_2 \|u\| \tag{D.32}$$

where $U \in \mathbb{R}^{mn \times N}$ is a matrix with the ith column $\mathrm{vec}(U_i)$ and $u \in \mathbb{R}^N$ is a vector with the ith entry $\langle E\left(X_i - X\right), U_i \rangle$. Using (D.31) in the proof of Lemma 10, for $s \geq 0$, we have

$$\|U\|_2 \;\leq\; c(\sqrt{N} \,+\, \sqrt{mn} \,+\, s) \tag{D.33}$$

with probability not smaller than $1 - 2\mathrm{e}^{-s^2} - N\mathrm{e}^{-mn/8}$. To bound the norm of u, we use similar arguments as in the proof of Lemma 5. In particular, let D_i be defined as above and let $\mathsf{D} := \cap_i \mathsf{D}_i$. Then for any $b \geq 0$,

$$\mathbb{P}\{\|u\| > b\} \;\leq\; \mathbb{P}\left\{\|u \mathbb{1}_{\mathsf{D}}\| > b\right\} \,+\, 4N\mathrm{e}^{-n/8} \tag{D.34}$$

where $\mathbb{1}_{\mathsf{D}}$ is the indicator function of D; cf. (D.25). Moreover, it is straightforward to verify that $\|u\mathbb{1}_{\mathsf{D}}\| \leq \|z\|$, where the entries of $z \in \mathbb{R}^N$ are given $z_i = u_i \mathbb{1}_{\mathsf{D}_i}$. Since $\|\|z\|^2\|_{\psi_{1/2}} = \|\sum_i z_i^2\|_{\psi_{1/2}}$, we have

$$\begin{aligned}
\|\sum_{i=1}^N z_i^2\|_{\psi_{1/2}} &\overset{(a)}{\leq} \|\sum_{i=1}^N z_i^2 - \mathbb{E}[z_i^2]\|_{\psi_{1/2}} + N\|\mathbb{E}[z_1^2]\|_{\psi_{1/2}} \\
&\overset{(b)}{\leq} \bar{c}_1 \|z_1^2\|_{\psi_{1/2}} \sqrt{N} \log N + \bar{c}_2 N \|z_1\|_{\psi_1}^2 \\
&\overset{(c)}{\leq} \bar{c}_3 N \|z_1\|_{\psi_1}^2 \\
&\overset{(d)}{\leq} \bar{c}_4 N \kappa^4 n (\|\mathcal{M}^*\|_2 + \|\mathcal{M}^*\|_S)^2 \|E\|_F^2.
\end{aligned}$$

Here, (a) follows from the triangle inequality, (b) follows from combination of Lemma 14, applied to the first term, and $\mathbb{E}[z_1^2] \leq \tilde{c}_0 \|z_1\|_{\psi_1}^2$ (e.g., see [160, Proposition 2.7.1]) applied to the second term, (c) follows from $\|z_1^2\|_{\psi_{1/2}} \leq \tilde{c}_1 \|z_1\|_{\psi_1}^2$, and (d) follows from (D.29). This allows us to use (D.35) with $\xi = \|z\|^2$ and $t = r^2$ to obtain

$$\mathbb{P}\{\|z\| > r\sqrt{nN}\kappa^2(\|\mathcal{M}^*\|_2 + \|\mathcal{M}^*\|_S)\|E\|_F\} \leq \bar{c}_5 \mathrm{e}^{-r}$$

for all $r > 0$. Combining this inequality with (D.34) yield

$$\mathbb{P}\left\{\|u\| > r\sqrt{nN}\kappa^2(\|\mathcal{M}^*\|_2 + \|\mathcal{M}^*\|_S)\|E\|_F\right\} \leq \bar{c}_5 \, \mathrm{e}^{-r} + 4N\mathrm{e}^{-n/8}.$$

Finally, substituting $r = \beta \log n$ in the last inequality and letting $s = \sqrt{mn}$ in (D.33) yield

$$\begin{aligned}
\mathbb{P}\left\{\frac{1}{N}\|\sum_i \langle E\left(X_i - X\right), U_i\rangle U_i\|_F > c_1 \beta \sqrt{mn} \log n \kappa^2 (\|\mathcal{M}^*\|_2 + \|\mathcal{M}^*\|_S)\|E\|_F\right\} \leq \\
c_0 n^{-\beta} + 2\mathrm{e}^{-mn} + N\mathrm{e}^{-mn/8} + 4N\mathrm{e}^{-n/8} \leq c_2(n^{-\beta} + N\mathrm{e}^{-n/8})
\end{aligned}$$

where we used inequality (D.32), $N \geq c_0 n$, and applied the union bound. This completes the proof.

Proof of Lemma 8

This result is obtained by applying Lemma 10 to the vectors $\mathrm{vec}(U_i)$ and setting $t = \sqrt{mn}$.

Probabilistic toolbox

In this subsection, we summarize known technical results which are useful in establishing bounds on the correlation between the gradient estimate and the true gradient. Herein, we use c, c', and c_i to denote positive absolute constants. For any positive scalar α, the ψ_α-norm of a random variable ξ is given by [170, Section 4.1], $\|\xi\|_{\psi_\alpha} := \inf_t \{t > 0 \,|\, \mathbb{E}\left[\psi_\alpha(|\xi|/t)\right] \leq 1\}$,

where $\psi_\alpha(x) := \mathrm{e}^{x^\alpha} - 1$ (linear near the origin when $0 < \alpha < 1$ in order for ψ_α to be convex) is an Orlicz function. Finiteness of the ψ_α-norm implies the tail bound

$$\mathbb{P}\left\{|\xi| > t\|\xi\|_{\psi_\alpha}\right\} \leq c_\alpha \mathrm{e}^{-t^\alpha} \text{ for all } t \geq 0 \tag{D.35}$$

where c_α is an absolute constant that depends on α; e.g., see [171, Section 2.3] for a proof. The random variable ξ is called sub-Gaussian if its distribution is dominated by that of a normal random variable. This condition is equivalent to $\|\xi\|_{\psi_2} < \infty$. The random variable ξ is sub-exponential if $\|\xi\|_{\psi_1} < \infty$. It is also well-known that for any random variables ξ and ξ' and any positive scalar α, $\|\xi\,\xi'\|_{\psi_\alpha} \leq \hat{c}_\alpha \|\xi\|_{\psi_{2\alpha}} \|\xi'\|_{\psi_{2\alpha}}$ and the above inequality becomes equality with $c_\alpha = 1$ if $\alpha \geq 1$.

Lemma 11 (Bernstein inequality [160, Corollary 2.8.3]) *Let the vectors $\xi_1, \ldots, \xi_N$ be independent, zero-mean, sub-exponential random variables with $\kappa \geq \|\xi_i\|_{\psi_1}$. Then, for any scalar $t \geq 0$, $\mathbb{P}\{|(1/N)\sum_i \xi_i| > t\} \leq 2\mathrm{e}^{-cN\min\{t^2/\kappa^2, t/\kappa\}}$.*

Lemma 12 (Hanson-Wright inequality [164, Theorem 1.1]) *Let A be a fixed matrix in $\mathbb{R}^{N\times N}$ and let $x \in \mathbb{R}^N$ be a random vector with independent entries that satisfy $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i^2] = 1$, and $\|x_i\|_{\psi_2} \leq \kappa$. Then, for any nonnegative scalar t, we have*

$$\mathbb{P}\{\left|x^T A\, x - \mathbb{E}[x^T A\, x]\right| > t\} \leq 2\mathrm{e}^{-c\min\{t^2/(\kappa^4\|A\|_F^2), t/(\kappa^2\|A\|_2)\}}.$$

Lemma 13 (Norms of random matrices [164, Theorem 3.2]) *Let E be a fixed matrix in $\mathbb{R}^{m\times n}$ and let $G \in \mathbb{R}^{m\times n}$ be a random matrix with independent entries that satisfy $\mathbb{E}[G_{ij}] = 0$, $\mathbb{E}[G_{ij}^2] = 1$, and $\|G_{ij}\|_{\psi_2} \leq \kappa$. Then, for any scalars $s, t \geq 1$,*

$$\mathbb{P}(\mathsf{A}) \leq 2\mathrm{e}^{-s^2 q - t^2 n}$$

where $q := \|E\|_F^2/\|E\|_2^2$ is the stable rank of E and

$$\mathsf{A} := \{\|E^T G\|_2 > c\kappa^2\left(s\|E\|_F + t\sqrt{n}\|E\|_2\right)\}.$$

The next lemma provides us with an upper bound on the ψ_α-norm of sum of random variables that is by Talagrand. This result is a straightforward consequence of combining the results in [170, Theorem 6.21] and [172, Lemma 2.2.2]; see e.g. [173, Theorem 8.4] for a formal argument.

Lemma 14 *For any scalar $\alpha \in (0, 1]$, there exists a constant C_α such that for any sequence of independent random variables $\xi_1, \ldots, \xi_N$ we have*

$$\|\sum_i \xi_i - \mathbb{E}\left[\sum_i \xi_i\right]\|_{\psi_\alpha} \leq C_\alpha(\max_i \|\xi_i\|_{\psi_\alpha})\sqrt{N}\log N.$$

D.11 Bounds on optimization variables

Building on [152], in Lemma 15 we provide useful bounds on the matrices K, $X = X(K)$, $P = P(K)$, and $Y = KX(K)$.

Lemma 15 *Over the sublevel set $\mathcal{S}_K(a)$ of the LQR objective function $f(K)$, we have*

$$\begin{aligned} \operatorname{trace}(X) &\leq a/\lambda_{\min}(Q) && \text{(D.36a)} \\ \|Y\|_F &\leq a/\sqrt{\lambda_{\min}(R)\lambda_{\min}(Q)} && \text{(D.36b)} \\ \nu/a &\leq \lambda_{\min}(X) && \text{(D.36c)} \\ \|K\|_F &\leq a/\sqrt{\nu\lambda_{\min}(R)} && \text{(D.36d)} \\ \operatorname{trace}(P) &\leq a/\lambda_{\min}(\Omega) && \text{(D.36e)} \end{aligned}$$

where the constant ν is given by (6.10d)*.*

Proof: For $K \in \mathcal{S}_K(a)$, we have

$$\operatorname{trace}(QX + Y^T R Y X^{-1}) \leq a \tag{D.37}$$

which along with $\operatorname{trace}(QX) \geq \lambda_{\min}(Q)\|X^{1/2}\|_F^2$ yield (D.36a). To establish (D.36b), we combine (D.37) with

$$\operatorname{trace}(RYX^{-1}Y^T) \geq \lambda_{\min}(R)\|YX^{-1/2}\|_F^2$$

to obtain $\|YX^{-1/2}\|_F^2 \leq a/\lambda_{\min}(R)$. Thus, $\|Y\|_F^2 \leq a\|X\|_2/\lambda_{\min}(R)$. This inequality along with (D.36a) give (D.36b). To show the inequality in (D.36c), let v be the normalized eigenvector corresponding to the smallest eigenvalue of X. Multiplication of Eq. (6.8a) from the left and the right by v^T and v, respectively, gives

$$v^T(DX^{1/2} + X^{1/2}D^T)\,v = \sqrt{\lambda_{\min}(X)}\,v^T(D + D^T)\,v = -v^T\Omega\,v$$

where $D := AX^{1/2} - BYX^{-1/2}$. Thus,

$$\lambda_{\min}(X) = \frac{(v^T\Omega\,v)^2}{(v^T(D + D^T)\,v)^2} \geq \frac{\lambda_{\min}^2(\Omega)}{4\,\|D\|_2^2} \tag{D.38}$$

where we applied the Cauchy-Schwarz inequality on the denominator. Using the triangle inequality and submultiplicative property of the 2-norm, we can upper bound $\|D\|_2$,

$$\|D\|_2 \leq \|A\|_2\|X^{1/2}\|_2 + \|B\|_2\|YX^{-1/2}\|_2 \leq \sqrt{a}\,(\|A\|_2/\sqrt{\lambda_{\min}(Q)} + \|B\|_2/\sqrt{\lambda_{\min}(R)}) \tag{D.39}$$

where the last inequality follows from (D.36a) and the upper bound on $\|YX^{-1/2}\|_F^2$.

Inequality (D.36c), with ν given by (6.10d), follows from combining (D.38) and (D.39). To show (D.36d), we use the upper bound on $\|YX^{-1/2}\|_F^2$, which is equivalent to

$$\|KX^{1/2}\|_F^2 \leq a/\lambda_{\min}(R)$$

to obtain

$$\|K\|_F^2 \leq a/\lambda_{\min}(R)\lambda_{\min}(X) \leq a^2/(\nu\lambda_{\min}(R)).$$

Here, the second inequality follows from (D.36c). Finally, to prove (D.36e), note that the definitions of $f(K)$ in (6.3b) and P in (6.6a) imply $f(K) = \operatorname{trace}(P\,\Omega)$. Thus, from $f(K) \leq a$, we have $\operatorname{trace}(P) \leq a/\lambda_{\min}(\Omega)$, which completes the proof.

□

D.12 The norm of the inverse Lyapunov operator

Lemma 16 provides an upper bound on the norm of the inverse Lyapunov operator for stable LTI systems.

Lemma 16 *For any Hurwitz matrix $F \in \mathbb{R}^{n\times n}$, the linear map $\mathcal{F}$: $\mathbb{S}^n \to \mathbb{S}^n$*

$$\mathcal{F}(W) \;:=\; \int_0^\infty \mathrm{e}^{Ft}\, W\, \mathrm{e}^{F^T t}\, \mathrm{d}t \tag{D.40}$$

is well defined and, for any $\Omega \succ 0$,

$$\|\mathcal{F}\|_2 \;\leq\; \operatorname{trace}(\mathcal{F}(I)) \;\leq\; \operatorname{trace}(\mathcal{F}(\Omega))/\lambda_{\min}(\Omega). \tag{D.41}$$

Proof: Using the triangle inequality and the sub-multiplicative property of the Frobenius norm, we can write

$$\|\mathcal{F}(W)\|_F \;\leq\; \int_0^\infty \|\mathrm{e}^{F\,t}\, W\, \mathrm{e}^{F^T t}\|_F\, \mathrm{d}t \;\leq\; \|W\|_F \int_0^\infty \|\mathrm{e}^{F\,t}\|_F^2\, \mathrm{d}t \;=\; \|W\|_F \operatorname{trace}(\mathcal{F}(I))\,. \tag{D.42}$$

Thus, $\|\mathcal{F}\|_2 = \max_{\|W\|_F=1} \|\mathcal{F}(W)\|_F \leq \operatorname{trace}(\mathcal{F}(I))$, which proves the first inequality in (D.41). To show the second inequality, we use the monotonicity of the linear map $\mathcal{F}$, i.e., for any symmetric matrices W_1 and W_2 with $W_1 \preceq W_2$, we have $\mathcal{F}(W_1) \preceq \mathcal{F}(W_2)$. In particular, $\lambda_{\min}(\Omega) I \preceq \Omega$ implies $\lambda_{\min}(\Omega)\mathcal{F}(I) \preceq \mathcal{F}(\Omega)$ which yields $\lambda_{\min}(\Omega)\operatorname{trace}(\mathcal{F}(I)) \leq \operatorname{trace}(\mathcal{F}(\Omega))$ and completes the proof. □

We next use Lemma 16 to establish a bound on the norm of the inverse of the closed-loop Lyapunov operator $\mathcal{A}_K$ over the sublevel sets of the LQR objective function $f(K)$.

Lemma 17 *For any $K \in \mathcal{S}_K(a)$, the closed-loop Lyapunov operators $\mathcal{A}_K$ given by* (6.7) *satisfies $\|\mathcal{A}_K^{-1}\|_2 = \|(\mathcal{A}_K^*)^{-1}\|_2 \leq a/\lambda_{\min}(\Omega)\lambda_{\min}(Q)$.*

Proof: Applying Lemma 16 with $F = A - BK$ yields

$$\|\mathcal{A}_K^{-1}\|_2 \;=\; \|(\mathcal{A}_K^*)^{-1}\|_2 \;\leq\; \operatorname{trace}(X)/\lambda_{\min}(\Omega).$$

Combining this inequality with (D.36a) completes the proof. □

Parameter $\theta(a)$ in Theorem 4

As discussed in the proof, over any sublevel set $\mathcal{S}_K(a)$ of the function $f(K)$, we require the function θ in Theorem 4 to satisfy

$$(\|(\mathcal{A}_K^*)^{-1}\|_2 + \|(\mathcal{A}_K^*)^{-1}\|_S)/\lambda_{\min}(X) \;\leq\; \theta(a)$$

for all $K \in \mathcal{S}_K(a)$. Clearly, Lemma 17 in conjunction with Lemma 15 can be used to obtain $\|(\mathcal{A}_K^*)^{-1}\|_2 \leq a/(\lambda_{\min}(Q)\lambda_{\min}\Omega)$ and $\lambda_{\min}^{-1}(X) \leq a/\nu$, where ν is given by (6.10d). The existence of $\theta(a)$, follows from the fact that there is a scalar $M(n) > 0$ such that $\|\mathcal{A}\|_S \leq M\|\mathcal{A}\|_2$ for all linear operators $\mathcal{A}$: $\mathbb{S}^n \to \mathbb{S}^n$.

Appendix E

Supporting proofs for Chapter 7

E.1 Proof of Proposition 1

Since G has a positive inner product with the gradient of the function $f(K)$, we can use the descent lemma [158, Eq. (9.17)] to show that $K^+ := K - \alpha G$ satisfies

$$f(K^+) \;-\; f(K) \;\leq\; (L_f(a)\alpha^2/2)\,\|G\|_F^2 \;-\; \alpha\,\langle \nabla f(K), G\rangle \tag{E.1}$$

for any α for which the line segment between K^+ and K lies in $\mathcal{S}_K(a)$. Using the inequalities in (7.11), for any $\alpha \in [0, 2\mu_1/(\mu_2 L_f(a))]$, we have

$$(L_f(a)\alpha^2/2)\,\|G\|_F^2 \;-\; \alpha\,\langle \nabla f(K), G\rangle \;\leq\; (\alpha\,(L_f(a)\mu_2\alpha - 2\mu_1)/2)\,\|\nabla f(K)\|_F^2 \;\leq\; 0 \tag{E.2}$$

and the right-hand side of inequality (E.1) is nonpositive for $\alpha \in [0, 2\mu_1/(\mu_2 L_f(a))]$. Thus, we can use the continuity of the function $f(K)$ along with inequalities (E.1) and (E.2) to conclude that $K^+ \in \mathcal{S}_K(a)$ for all $\alpha \in [0, 2\mu_1/(\mu_2 L_f(a))]$, and

$$f(K^+) - f(K) \;\leq\; (\alpha\,(L_f(a)\mu_2\alpha - 2\mu_1)/2)\,\|\nabla f(K)\|_F^2.$$

Combining this inequality with the PL condition, it follows that

$$f(K^+) - f(K) \;\leq\; -(\mu_1\alpha/2)\,\|\nabla f(K)\|_F^2 \;\leq\; -\mu_f(a)\mu_1\alpha\,(f(K) - f(K^\star))$$

for all $\alpha \in [0, \mu_1/(\mu_2 L_f(a))]$. Subtracting $f(K^\star)$ and rearranging terms complete the proof.

E.2 Proof of Proposition 2

We first present two technical lemmas.

Lemma 1 *Let the matrices F, $X \succ 0$, and $\Omega \succ 0$ satisfy*

$$FXF^T \;-\; X \;+\; \Omega \;=\; 0.$$

Then, we have $\|F^t\|_2^2 \leq c\rho^t$ for all $t \in \mathbb{N}$, where

$$c \;:=\; \|X\|_2/\lambda_{\min}(X), \qquad \rho \;:=\; 1 - \lambda_{\min}(\Omega)/\|X\|_2.$$

Proof: Using the trivial inequalities $\Omega \succeq \lambda_{\min}(\Omega) I$ and $X \preceq \|X\|_2 I$, we can write

$$F X F^T \;=\; X - \Omega \;\preceq\; \rho X$$

where $\rho := 1 - \lambda_{\min}(\Omega)/\|X\|_2$. This matrix inequality implies that $V(x) := x^T X x$ is a Lyapunov function for $x^{t+1} = F^T x^t$ because $V(x^{k+1}) \leq \rho V(x^k)$. Thus, for any initial condition x^0, we have $V(x^t) \leq \rho^t V(x^0)$. Noting that $x^t = (F^T)^t x^0$, we let x^0 be the normalized left singular vector associated with the maximum singular value of F^t to obtain

$$\|F^t\|_2^2 \;=\; \|x^t\|^2 \;\leq\; \frac{V(x^t)}{\lambda_{\min}(X)} \;\leq\; \rho^t \frac{V(x^0)}{\lambda_{\min}(X)}$$

which along with $V(x^0) \leq \|X\|_2$ complete the proof. □

Lemma 2 establishes an exponentially decaying upper bound on the difference between $f_\zeta(K) = f_{\zeta,\infty}(K)$ and $f_{\zeta,\tau}(K)$ over any sublevel set $\mathcal{S}_K(a)$ of the LQR objective function $f(K)$.

Lemma 2 *For any $K \in \mathcal{S}_K(a)$ and $\zeta \in \mathbb{R}^n$,*

$$|f_\zeta(K) - f_{\zeta,\tau}(K)| \;\leq\; \|\zeta\|^2 \kappa_1(a)(1 - \kappa_2(a))^\tau$$

where κ_1 and $\kappa_2 < 1$ are positive rational functions that depend on the problem data.

Proof: Since $x^t = (A - BK)^t \zeta$ is the solution to (7.1a) with $u = -Kx$ and the initial condition $x^0 = \zeta$, it is easy to verify that $f_{\zeta,\tau}(K) = \left\langle Q + K^T R\, K, X_{\zeta,\tau}(K) \right\rangle$, where

$$X_{\zeta,\tau}(K) \;:=\; \sum_{t=0}^{\tau} (A - BK)^t\, \zeta \zeta^T ((A - BK)^T)^t.$$

Using the triangle inequality, we have

$$\|X_\zeta(K) - X_{\zeta,\tau}(K)\|_F \;\leq\; \|\zeta\|^2 \sum_{t\,=\,\tau}^{\infty} \|(A - BK)^t\|_2^2 \tag{E.3}$$

where $X_\zeta(K) = X_{\zeta,\infty}(K)$ is given by (7.3). The Lyapunov equation in (7.5) allows us to use Lemma 1 with $F := A - BK$, $X := X(K)$ to upper bound $\|(A - BK)^t\|_2$,

$$\lambda_{\min}(X) \|(A - BK)^t\|_2^2 \;\leq\; \|X\|_2\, (1 - \lambda_{\min}(\Omega)/\|X\|_2)^t.$$

Summing this inequality from $t = \tau$ onward in conjunction with (E.3) yield

$$\|X_\zeta(K) - X_{\zeta,\tau(K)}\|_F \;\leq\; \|\zeta\|^2 \kappa_1'\, (1 - \kappa_2')^\tau \tag{E.4}$$

where $\kappa_1' := \|X(K)\|_2^2/(\lambda_{\min}(\Omega)\lambda_{\min}(X(K)))$ and $\kappa_2' := \lambda_{\min}(\Omega)/\|X(K)\|_2$. Furthermore,

$$\begin{aligned} |f_\zeta(K) - f_{\zeta,\tau}(K)| \;&=\; \left|\mathrm{trace}\left((Q + K^T R\, K)\,(X_\zeta - X_{\zeta,\tau})\right)\right| \\ &\leq\; (\|Q\|_F + \|R\|_2\|K\|_F^2)\|X_\zeta - X_{\zeta,\tau}\|_F \\ &\leq\; \|\zeta\|^2(\|Q\|_F + \|R\|_2\|K\|_F^2)\,\kappa_1'(1-\kappa_2')^\tau \end{aligned}$$

where we use the Cauchy-Schwartz and triangle inequalities for the first inequality and (E.4) for the second inequality. Combining this result with trivial upper bounds on the norms $\|K\|_F$, $\|X(K)\|_2$, and $X(X) \succeq \Omega \succ 0$ completes the proof. See [14, Lemma 23] for derivation of these bounds. □

We are now ready to prove the first inequality in Proposition 2. Since $K \in \mathcal{S}_K(a)$ and $r \leq r(a)$, Lemma 1 implies that $K \pm rU_i \in \mathcal{S}_K(2a)$. Thus, $f_{\zeta^i}(K \pm rU_i)$ is well defined for $i = 1, \ldots, N$, and

$$\begin{aligned} \widetilde{\nabla} f(K) - \overline{\nabla} f(K) \;&=\; \frac{1}{2rN} \times \\ &\sum_i \left(\left(f_{\zeta^i}(K + rU_i) - f_{\zeta^i,\tau}(K + rU_i)\right)U_i \;-\; \left(f_{\zeta^i}(K - rU_i) - f_{\zeta^i,\tau}(K - rU_i)\right)U_i\right). \end{aligned}$$

Furthermore, since $K \pm rU_i \in \mathcal{S}_K(2a)$, we can use triangle inequality and apply Lemma 2, $2N$ times, to bound each term individually and obtain

$$\|\widetilde{\nabla} f(K) \;-\; \overline{\nabla} f(K)\|_F \;\leq\; (\sqrt{mn}/r)\max_i \|\zeta^i\|^2 \kappa_1(2a)(1 \;-\; \kappa_2(2a))^\tau$$

where we used $\|U_i\|_F = \sqrt{mn}$. This completes the proof of the first inequality. The proof of the second inequality follows similar arguments as in [14, Propopsition 5], which exploits the third derivatives of the functions f_ζ.