Journal of Machine Learning Research 22 (2021) 1-39 Submitted 9/20; Revised 4/21; Published 6/21

A Unified Analysis of First-Order Methods for Smooth Games via Integral Quadratic Constraints

Guodong Zhang[1,3] GDZHANG@CS.TORONTO.EDU
Xuchan Bao[1,3] JENNYBAO@CS.TORONTO.EDU
Laurent Lessard[2] L.LESSARD@NORTHEASTERN.EDU
Roger Grosse[1,3] RGROSSE@CS.TORONTO.EDU
[1] *Department of Computer Science, University of Toronto*
[2] *Mechanical and Industrial Engineering, Northeastern University*
[3] *Vector Institute*

Editor: Sebastian Nowozin

Abstract

The theory of integral quadratic constraints (IQCs) allows the certification of exponential convergence of interconnected systems containing nonlinear or uncertain elements. In this work, we adapt the IQC theory to study first-order methods for smooth and strongly-monotone games and show how to design tailored quadratic constraints to get tight upper bounds of convergence rates. Using this framework, we recover the existing bound for the gradient method (GD), derive sharper bounds for the proximal point method (PPM) and optimistic gradient method (OG), and provide *for the first time* a global convergence rate for the negative momentum method (NM) with an iteration complexity $\mathcal{O}(\kappa^{1.5})$, which matches its known lower bound. In addition, for time-varying systems, we prove that the gradient method with optimal step size achieves the fastest provable worst-case convergence rate with quadratic Lyapunov functions. Finally, we further extend our analysis to stochastic games and study the impact of multiplicative noise on different algorithms. We show that it is impossible for an algorithm with one step of memory to achieve acceleration if it only queries the gradient once per batch (in contrast with the stochastic strongly-convex optimization setting, where such acceleration has been demonstrated). However, we exhibit an algorithm which achieves acceleration with two gradient queries per batch. Our code is made public at `https://github.com/gd-zhang/IQC-Game`.

Keywords: Smooth Game Optimization, Monotone Variational Inequality, First-Order Methods, Integral Quadratic Constraints, Dynamical Systems

1. Introduction

Gradient-based optimization algorithms have played a prominent role in machine learning and underpinned a significant fraction of the recent successes in deep learning (Krizhevsky et al., 2012; Silver et al., 2017). Typically, the training of many models can be formulated as a single-objective optimization problem, which can be efficiently solved by gradient-based optimization methods. However, there are a growing number of models that involve multiple interacting objectives. For example, generative adversarial networks (Goodfellow et al., 2014; Radford et al., 2015; Arjovsky et al., 2017), adversarial training (Madry et al., 2018) and primal-dual reinforcement learning (Du et al., 2017; Dai et al., 2018) all require

the joint minimization of several objectives. Hence, there is a surge of interest in coupling machine learning and game theory by modeling problems as smooth games.

Smooth games, and the closely related framework of variational inequalities, are generalizations of the standard single-objective optimization framework, allowing us to model multiple players and objectives. However, new issues and challenges arise in solving smooth games or variational inequalities. Due to the conflict of optimizing different objectives, standard gradient-based algorithms may exhibit rotational behaviors (Mescheder et al., 2017; Letcher et al., 2019) and hence converge slowly. To combat this problem, several algorithms have been introduced specifically for smooth games, including negative momentum (NM) (Gidel et al., 2019), optimistic gradient method (OG) (Popov, 1980; Rakhlin and Sridharan, 2013; Daskalakis et al., 2018; Mertikopoulos et al., 2018) and extra-gradient (EG) (Korpelevich, 1976; Nemirovski, 2004). While these algorithms were motivated by provable convergence bounds, many such analyses were limited to quadratic problems with linear dynamics, or to proving local convergence (so that the dynamics could be linearized) (Gidel et al., 2019; Azizian et al., 2020b; Zhang and Wang, 2021). Other analyses proved global convergence rates, but relied on deep insight[1] to design Lyapunov functions on a case-by-case basis (Gidel et al., 2018; Azizian et al., 2020a; Mokhtari et al., 2020a).

In this paper, we aim to provide a systematic framework for analyzing first-order methods in solving smooth and strongly-monotone games using techniques from control theory. In particular, we view common optimization algorithms as feedback interconnections and adopt the theory of integral quadratic constraints (IQCs) (Megretski and Rantzer, 1997) to model the nonlinearities and uncertainties in the system. While enforcing common assumptions in optimization would seem to require infinitely many IQCs, Lessard et al. (2016) showed that it was possible to certify tight convergence bounds for first-order optimization algorithms using a small number of IQCs. The result of their analysis was a largely mechanical procedure for converting questions about convergence into small semidefinite programs which could be solved efficiently. We perform an analogous analysis in the more complex setting of smooth games, arriving at a very different, but similarly compact, set of IQCs. Particularly, we show that only a few pointwise IQCs are sufficient to certify tight convergence bounds for a variety of algorithms — an even more parsimonious description than in the optimization setting. The end result of our analysis is a unified and automated method for analyzing convergence of first-order methods for smooth games.

Using this framework, we are able to recover or even improve known convergence bounds for a variety of algorithms, which we summarize as follows:

- We recover the known convergence rate of the gradient method for smooth and strongly-monotone games by solving a 2×2 semidefinite program (SDP) analytically.
- Similarly, we derive an analytical convergence bound for the proximal point method that is sharper than the best available result (Mokhtari et al., 2020a, Theorem 2).
- We derive a slightly improved convergence rate for the optimistic gradient method (even though the existing analysis (Gidel et al., 2018) is fairly involved).

1. Designing a Lyapunov function is largely regarded as a black art. The proofs of OG/EG are based on the insight that they both approximate the proximal point method (Mokhtari et al., 2020a). Typically these insights do not generalize to other algorithms.

We emphasize that all of the above results are obtainable from our unified framework through a mechanical procedure of deriving and solving an SDP. Beyond these results, we can gain new insights and derive new results that were previously unknown and are difficult to obtain using existing approaches:

- We prove that, for time-varying systems, the gradient method with optimal step size achieves the fastest provable convergence rate with quadratic Lyapunov functions among any algorithm representable as a linear time-invariant system with finite state.
- We provide the first global convergence rate guarantee for the negative momentum method for smooth and strongly-monotone games, matching the known lower bound (Zhang and Wang, 2021).
- We also show that the optimistic gradient method achieves the optimal convergence rate provable in our framework among algorithms with one step of memory (6).

Further, we adapt the IQC framework to analyze stochastic games. We model stochasticity using the strong growth condition (Schmidt and Roux, 2013; Vaswani et al., 2019), which has been used to model multiplicative noise in the optimization setting, but has not been investigated in the game setting. The key is to model optimization algorithms as stochastic jump systems as in Hu et al. (2017). We demonstrate that GD is robust to noise in the sense that it can attain the same $\mathcal{O}(\kappa^2)$ convergence rate as the deterministic case (where the constant depends on noise level). By contrast, OG and NM are degraded to an $\mathcal{O}(\kappa^2)$ convergence rate, in contrast with their $\mathcal{O}(\kappa)$ and $\mathcal{O}(\kappa^{1.5})$ rates in the deterministic setting. We show this is an instance of a more general phenomenon: with large enough noise, no first-order algorithm with at most one step of memory can be proved under our analysis to improve upon GD's convergence rate. (This is in contrast to the setting of smooth and strongly convex optimization, where such acceleration has been proved (Jain et al., 2018; Vaswani et al., 2019).) Nonetheless, we exhibit an algorithm which achieves acceleration in the stochastic setting by querying the vector field twice for each batch of data.

We believe our IQC framework is a powerful tool for exploratory algorithmic research, since it allows us to quickly ask and answer a variety of questions about the convergence of algorithms for smooth games.

1.1 Other Related Works

For the general monotone setting (without strong monotonicity), it is known that the optimal rate of convergence for first-order methods is $\mathcal{O}(1/T)$, and this rate is achieved by both the EG and OG algorithms (Nemirovski, 2004; Tseng, 2008; Hsieh et al., 2019; Mokhtari et al., 2020b) for the averaged (ergodic) iterates. Later, (Golowich et al., 2020b,a) derived a $\mathcal{O}(1/\sqrt{T})$ bound for the last iterate of EG and OG.

Beyond the monotone setting, non-monotone games (e.g., nonconvex-nonconcave minimax problems) have recently gained more attention due to their generality. However, there might be no Nash (or even local Nash) equilibria in that setting due to the loss of strong duality. To overcome that, different notions of equilibrium were introduced by taking into account the sequential structure of games (Jin et al., 2020; Fiez et al., 2019; Farnia and Ozdaglar, 2020; Mangoubi and Vishnoi, 2020). In that setting, the main challenge is to

find the right equilibrium and some algorithms (Wang et al., 2019; Adolphs et al., 2019; Mazumdar et al., 2019) have been proposed to achieve that.

For smooth game optimization, there are also many algorithms using high-order information. For example, consensus optimization (Mescheder et al., 2017), Hamiltonian gradient descent (Letcher et al., 2019; Abernethy et al., 2019), competitive gradient descent (Schäfer and Anandkumar, 2019), follow-the-ridge (Wang et al., 2019) and LEAD (Hemmat et al., 2020) all used second-order information to accelerate the convergence. Currently, our IQC framework is primarily designed for first-order algorithms. Exploring new types of IQCs that can be used to analyze algorithms using high-order information would be an interesting future direction.

2. Preliminaries

2.1 Variational Inequality Formulation of Smooth Games

We begin by presenting the basic variational inequality framework that we consider in the sequel. Let Ω be a nonempty convex subset of $\mathbb{R}^d$, and let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a continuous mapping on $\mathbb{R}^d$. In its most general form, the variational inequality (VI) problem (Harker and Pang, 1990) associated to F and Ω can be stated as:

$$\text{find } z^* \in \Omega \text{ such that } F(z^*)^\top (z - z^*) \geq 0 \text{ for all } z \in \Omega. \tag{1}$$

In the case of $\Omega = \mathbb{R}^d$, it reduces to finding z^* such that $F(z^*) = 0$. To provide some intuition about variational inequalities, we discuss two important examples below:

Example 1 (Minimization) *Suppose that $F = \nabla_z f$ for a smooth function f on $\mathbb{R}^d$, then the variational inequality problem amounts to finding the critical points of f. In the case where f is convex, any solution of (1) is a global minimizer.*

Example 2 (Minimax Games) *Consider a convex-concave minimax optimization problem (saddle-point problem). Our objective is to solve the problem $\min_x \max_y f(x, y)$, where f is a smooth function. It is easy to show that minimax optimization is a special case of (1) with $F(z) = [\nabla_x f(x, y)^\top, -\nabla_y f(x, y)^\top]^\top$, where $z = [x^\top, y^\top]^\top$.*

To be noted, the vector field F in Example 2 is not necessarily conservative, i.e., it might not be the gradient of any function. In addition, if f in minimax problem is convex-concave, any solution $z^* = [x^{*\top}, y^{*\top}]^\top$ of (1) is a global *Nash Equilibrium* (Von Neumann and Morgenstern, 1944):

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*) \quad \text{for all } x \text{ and } y \in \mathbb{R}^d. \tag{2}$$

In this work, we are particularly interested in the case of f being a strongly-convex-strongly-concave and smooth function, which basically implies that the vector field F is strongly-monotone and Lipschitz (see Fallah et al. (2020, Lemma 2.6) for more details). Here we state our assumptions formally.

Assumption 1 (Strongly Monotone) *The vector field F is m-strongly-monotone:*

$$(F(z_1) - F(z_2))^\top (z_1 - z_2) \geq m \|z_1 - z_2\|_2^2 \quad \textit{for all } z_1, z_2 \in \mathbb{R}^d. \tag{3}$$

Assumption 2 (Lipschitz) *The vector field F is L-Lipschitz:*

$$\|F(z_1) - F(z_2)\|_2 \leq L\|z_1 - z_2\|_2 \quad \textit{for all } z_1, z_2 \in \mathbb{R}^d. \tag{4}$$

In the context of variational inequalites, Lipschitzness and (strong) monotonicity are fairly standard and have been used in many classical works (Tseng, 1995; Chen and Rockafellar, 1997; Nesterov, 2007; Nemirovski, 2004). With these two assumptions in hand, we define the condition number $\kappa \triangleq L/m$, which measures the hardness of the problem. In the following, we turn to suitable optimization techniques for the variational inequality.

2.2 Optimization Algorithms as Dynamical Systems

Borrowing the notations from Lessard et al. (2016), we frame various first-order algorithms as a unified linear dynamical system[2] in feedback with a nonlinearity $\phi : \mathbb{R}^d \to \mathbb{R}^d$,

$$\begin{aligned} \xi_{k+1} &= A\xi_k + Bu_k \\ y_k &= C\xi_k + Du_k \\ u_k &= \phi(y_k). \end{aligned} \tag{5}$$

At each iteration $k = 0, 1, ...$, $u_k \in \mathbb{R}^d$ is the control input, $y_k \in \mathbb{R}^d$ is the output, and $\xi_k \in \mathbb{R}^{nd}$ is the state for algorithms with n step of memory. The state matrices A, B, C, D differ for various algorithms. For most algorithms we consider in the paper, they have the general form:

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{cc|c} (1+\beta)\mathbf{I}_d & -\beta\mathbf{I}_d & -\eta\mathbf{I}_d \\ \mathbf{I}_d & \mathbf{0}_d & \mathbf{0}_d \\ \hline (1+\alpha)\mathbf{I}_d & -\alpha\mathbf{I}_d & \mathbf{0}_d \end{array}\right],$$

where $\mathbf{I}_d$ and $\mathbf{0}_d$ are the identity and zero matrix of size $d \times d$, respectively. One can then reduce linear dynamical system (5) to a second-order difference equation by setting $\xi_k \coloneqq \left[z_k^\top, z_{k-1}^\top\right]^\top$ and $\phi \coloneqq F$, which we term *algorithms with one step of memory*:

$$z_{k+1} = (1+\beta)z_k - \beta z_{k-1} - \eta F((1+\alpha)z_k - \alpha z_{k-1}), \tag{6}$$

where η is a constant step size. By choosing different α, β, we can recover different methods[3] (see Table 1). For instance, optimistic gradient method (OG) (Daskalakis et al., 2018) is typically written in the following form ($\alpha = 1$ and $\beta = 0$).

$$z_{k+1} = z_k - 2\eta F(z_k) + \eta F(z_{k-1}). \tag{7}$$

For smooth and strongly-monotone games, Azizian et al. (2020b, Corollary 1) showed a lower bound on convergence rate for *any* algorithm of the form (5):

$$\|\xi_k - \xi^*\|_2 \geq \rho_{\text{opt}}^k \|\xi_0 - \xi^*\|_2 \quad \text{with} \quad \rho_{\text{opt}} = 1 - \frac{2m}{m+L}. \tag{8}$$

Also, one can show that the lower bound for GD is $\sqrt{1 - 1/\kappa^2}$ and the lower bound for NM (Zhang and Wang, 2021) is $1 - c\kappa^{-1.5}$ where c is a constant independent of κ.

2. This linear dynamical system can represent any first-order methods.
3. One can also model EG using the same dynamical system (5), but it does not fit into (6).

Method	Parameter Chioce	Complexity	Reference
GD	$\alpha = 0$, $\beta = 0$	$\mathcal{O}(\kappa^2)$	Ryu and Boyd (2016); Azizian et al. (2020a)
OG	$\alpha = 1$, $\beta = 0$	$\mathcal{O}(\kappa)$	Gidel et al. (2018); Mokhtari et al. (2020a)
NM	$\alpha = 0$, $\beta < 0$	$\mathcal{O}(\kappa^{1.5})$	Section 3.4 of **this paper**

Table 1: Global convergence rates of algorithms for smooth and strongly-monotone games.

2.3 IQCs for Exponential Convergence Rates

We now present the theory of IQCs and connect it with exponential convergence. IQCs provide a convenient framework for analyzing interconnected dynamical systems that contain components that are nonlinear, uncertain, or otherwise difficult to model. The idea is to replace these troublesome components by quadratic constraints on its inputs and outputs that are known to be satisfied by all possible instances of the component.

In our case, the vector field F is the troublesome function we wish to analyze (currently, the IQC framework is limited to first-order algorithms). Although we do not know F exactly, we assume to have some knowledge of the constraints it imposes on the input-output pair (y, u). For example, we already assume F to be L-Lipschitz, which implies $\|u_k - u^*\|_2 \leq L\|y_k - y^*\|_2$ for all k with $u^* = F(y^*)$ as a fixed point. In matrix form, this is

$$\begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix}^\top \begin{bmatrix} L^2\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix} \begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix} \geq 0. \tag{9}$$

Notably, the above constraint is very special in that it only manifests itself as separate quadratic constraints on each (y_k, u_k). It is possible to specify quadratic constraints that couple different k values. To achieve that, we follow Lessard et al. (2016) and adopt auxiliary sequences ζ, s together with a map Ψ characterized by matrices $(A_\Psi, B^y_\Psi, B^u_\Psi, C_\Psi, D^y_\Psi, D^u_\Psi)$:

$$\begin{aligned} \zeta_{k+1} &= A_\Psi \zeta_k + B^y_\Psi y_k + B^u_\Psi u_k, \\ s_k &= C_\Psi \zeta_k + D^y_\Psi y_k + D^u_\Psi u_k. \end{aligned} \tag{10}$$

The equations (10) define an affine map $s = \Psi(y, u)$, where s_k could be a function of all past y_i and u_i with $i \leq k$. We consider the quadratic form $(s_k - s^*)^\top M(s_k - s^*)$ for a given matrix M with s^* and ξ^* fixed points of (10). We note that the quadratic form is a function of $(y_0, \ldots, y_k, u_0, \ldots, u_k)$ that is determined by our choice of (Ψ, M). In particular, we can recover constraint (9) with

$$\Psi = \left[\begin{array}{c|cc} A_\Psi & B^y_\Psi & B^u_\Psi \\ \hline C_\Psi & D^y_\Psi & D^u_\Psi \end{array}\right] = \left[\begin{array}{c|cc} \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \hline \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d \end{array}\right], \qquad M = \begin{bmatrix} L^2\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix}. \tag{11}$$

In general, this sort of quadratic constraints are called IQCs. There are different types of IQCs (see Lessard et al. (2016, Definition 3)), but we will only need *pointwise* IQCs as quadratic Lyapunov functions turn out to be expressive enough.

Definition 1 *A **Pointwise IQC** defined by (Ψ, M) satisfies*

$$(s_k - s^*)^\top M(s_k - s^*) \geq 0 \quad \textit{for all } k \geq 0.$$

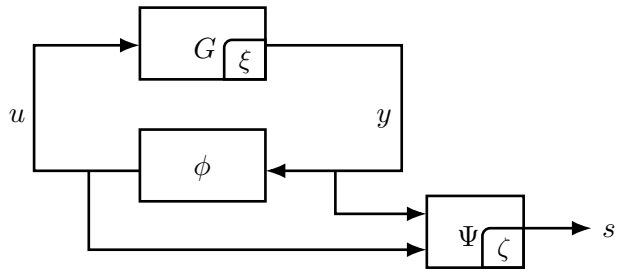


Figure 1: Feedback interconnection between a system G (optimization algorithm) with state matrices (A, B, C, D) and a nonlinearity ϕ. An IQC is a constraint on (y, u) satisfied by ϕ and we are mostly interested in the case where $\phi = F$.

Combining the dynamics (5) with the map Ψ (by eliminating y_k), we obtain

$$\begin{aligned} \begin{bmatrix} \xi_{k+1} \\ \zeta_{k+1} \end{bmatrix} &= \begin{bmatrix} A & 0 \\ B_\Psi^y C & A_\Psi \end{bmatrix} \begin{bmatrix} \xi_k \\ \zeta_k \end{bmatrix} + \begin{bmatrix} B \\ B_\Psi^u + B_\Psi^y D \end{bmatrix} u_k, \\ s_k &= \begin{bmatrix} D_\Psi^y C & C_\Psi \end{bmatrix} \begin{bmatrix} \xi_k \\ \zeta_k \end{bmatrix} + \begin{bmatrix} D_\Psi^u + D_\Psi^y D \end{bmatrix} u_k. \end{aligned} \tag{12}$$

More succinctly, (12) can be written as

$$\begin{aligned} x_{k+1} &= \hat{A} x_k + \hat{B} u_k \\ s_k &= \hat{C} x_k + \hat{D} u_k \end{aligned}, \qquad \text{where } x_k \triangleq \begin{bmatrix} \xi_k \\ \zeta_k \end{bmatrix}. \tag{13}$$

With these definitions in hand, we now state the main result of verifying exponential convergence. Basically, we build a Linear Matrix Inequality (LMI) to guide the search for the parameters of quadratic Lyapunov function in order to establish a rate bound.

Theorem 1 *Consider the dynamical system* (5)*. Suppose the vector field F satisfies the pointwise IQC (Ψ, M) and define $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ according to* (11)–(13)*. Consider the following linear matrix inequality (LMI):*

$$\begin{bmatrix} \hat{A}^\top P \hat{A} - \rho^2 P & \hat{A}^\top P \hat{B} \\ \hat{B}^\top P \hat{A} & \hat{B}^\top P \hat{B} \end{bmatrix} + \lambda \begin{bmatrix} \hat{C} & \hat{D} \end{bmatrix}^\top M \begin{bmatrix} \hat{C} & \hat{D} \end{bmatrix} \preceq 0. \tag{14}$$

If this LMI is feasible for some $P \succ 0$, $\lambda \geq 0$ and $\rho > 0$[4]*, we have*

$$(x_{k+1} - x^*)^\top (P \otimes \mathbf{I}_d)(x_{k+1} - x^*) \leq \rho^2 (x_k - x^*)^\top (P \otimes \mathbf{I}_d)(x_k - x^*). \tag{15}$$

Consequently, for any ξ_0 and $\zeta_0 = \zeta^*$, we obtain

$$\|\xi_k - \xi^*\|_2^2 \leq \mathrm{cond}(P) \rho^{2k} \|\xi_0 - \xi^*\|_2^2. \tag{16}$$

Remark 1 *The LMI* (14) *can be extended to the case of multiple constraints with (Ψ_i, M_i) (see Lessard et al. (2016, Page 12) for details).*

4. Note that ρ is not necessarily smaller than 1.

Remark 2 *The positive definite quadratic function* $V(x) \triangleq (x-x^*)^\top (P \otimes \mathbf{I}_d)(x-x^*)$ *is a Lyapunov function that certifies exponential convergence. This is the main difference from Lessard et al. (2016, Theorem 4) in that the function* $V(x)$ *in their case cannot serve as a Lyapunov function because it does not strictly decrease over all trajectories.*

To apply Theorem 1, we seek to solve the semidefinite program (SDP) of finding the minimal ρ such that the LMI (14) is feasible. For simple algorithms, one can typically solve the SDP analytically. Nevertheless, one may only get a numerical proof when the algorithm of interest is complicated and the resulting SDP is hard to solve. The solution yields the best convergence rate that can be certified by quadratic Lyapunov functions. We remark that it *automatically* searches for a quadratic Lyapunov function for proving exponetial convergence by solving the SDP. This is extremely convenient compared to designing ad-hoc Lyapunov functions on an algorithm-by-algorithm basis. Moreover, by inspecting the corresponding λ_i of constraint (Ψ_i, M_i), we could tell if the constraint or assumption is redundant or not. Moreover, this framework makes it easy to analyze the performance of optimization algorithms for time-varying systems, as we will show in the next section.

3. IQCs for Variational Inequalities

To apply Theorem 1 to smooth and strongly-monotone variational inequalities, we will derive two sets of IQCs describing the vector field F: sector IQCs and *off-by-one pointwise* IQCs. According to Assumptions 1 and 2, the constraints (3) and (4) hold over the whole domain. Hence, it has essentially infinite number of constraints and is therefore hard to use. The key idea of the following two sets of IQCs is to find the necessary conditions (but not sufficient) of smoothness and strongly-monotonicity by discretizing the constraints to finite number of (z_1, z_2) pairs. This is equivalent to a relaxation to the original problem since functions which are not L-Lipschitz and m-strongly-monotone can potentially satisfy the discretized conditions. Hence in principle, we need to make the discretized conditions to be as close to the original necessary and sufficient conditions as possible.

We first introduce two sector IQCs, which takes the discretization of (y_k, y^*) with y_k the output of iteration k and y^* the output of the stationary state.

Lemma 1 (Sector IQCs) *Suppose vector field* F_k *is* m*-strongly monotone and* L*-Lipschitz for all* k*, if* $u_k = F_k(y_k)$*, then* $\phi := (F_0, F_1, ...)$ *satisfies the* ***pointwise IQCs*** *defined by*

$$\Psi_1 = \Psi_2 = \left[\begin{array}{c|cc} \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \hline \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d \end{array}\right], \quad M_1 = \begin{bmatrix} L^2\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix}, \quad M_2 = \begin{bmatrix} -2m\mathbf{I}_d & \mathbf{I}_d \\ \mathbf{I}_d & \mathbf{0}_d \end{bmatrix}. \tag{17}$$

We have corresponding quadratic inequalities:

$$\begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix}^\top \begin{bmatrix} L^2\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix} \begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix} \geq 0 \quad \textit{and} \quad \begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix}^\top \begin{bmatrix} -2m\mathbf{I}_d & \mathbf{I}_d \\ \mathbf{I}_d & \mathbf{0}_d \end{bmatrix} \begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix} \geq 0.$$

As we will show in the next few sections, the introduced set of sector IQCs is far from sufficient for some algorithms because it allows the vector field F_k to be time-varying[5],

5. In convex optimization (Hazan, 2016), it is equivalent to allowing the losses to be adversarially chosen.

therefore leading to very conservative estimate of convergence rates. As a remedy, we can add (y_{k-1}, y_k) pairs to enforce the consistency of the vector field over time, which leads to the following off-by-one pointwise IQCs. We stress that the proposed off-by-one pointwise IQC is different from the one in Lessard et al. (2016, Lemma 8) in that their off-by-one IQC is a more complicated ρ-hard IQC (see Definition 3 in Lessard et al. (2016) for details) rather than a pointwise IQC. This is due to the fact that the first-order oracle in convex minimization involves the function value f and they have to use the ρ-hard IQC to get tight bounds.

Lemma 2 (Off-by-one pointwise IQCs) *Suppose F is m-strongly monotone and L-Lipschitz. If $u_k = F(y_k)$, then $\phi \coloneqq (F, F, ...)$ satisfies the **pointwise IQCs** defined by*

$$\Psi_1 = \Psi_2 = \left[\begin{array}{cc|c|c} \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d \\ \hline -\mathbf{I}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d & \mathbf{0}_d & \mathbf{I}_d \end{array}\right], \quad M_1 = \begin{bmatrix} L^2\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix}, \quad M_2 = \begin{bmatrix} -2m\mathbf{I}_d & \mathbf{I}_d \\ \mathbf{I}_d & \mathbf{0}_d \end{bmatrix}. \tag{18}$$

We have corresponding quadratic inequalities:

$$\begin{aligned} L^2\|y_{k+1} - y_k\|_2^2 - \|u_{k+1} - u_k\|_2^2 \geq 0, \\ (y_{k+1} - y_k)^\top (u_{k+1} - u_k - m(y_{k+1} - y_k)) \geq 0. \end{aligned} \tag{19}$$

In principle, a convex combination of the sector and off-by-one pointwise IQCs is still not sufficient, though we can further add off-by-n pointwise IQCs (i.e., (y_{k-n}, y_k)) to make it less conservative. To exactly characterize the nonlinearity of vector field F, it requires us to introduce the following interpolation condition (insipred by Taylor et al. (2017)):

Definition 2 ($\{m, L\}$-interpolation) *Let I be an index set, and consider the set of tuples $S = \{(y_i, u_i)\}_{i \in I}$. Then set S is $\{m, L\}$-interpolable if and only if there exists a m-strongly monotone and L-Lipschitz vector field F such that $u_i = F(y_i)$ for all $i \in I$.*

At first glance, it might seem that all pairs (y_i, y_j) of indices $i \in I$ and $j \in I$ satisfying (3) and (4) would be necessary and sufficient for $\{m, L\}$-interpolation. However, it was shown in Ryu et al. (2020, Proposition 3) that it is not the case. In other words, including all off-by-n pointwise IQCs ($n = 1, 2, ...$) is still not sufficient to fully describe the nonlinearity. So in general, the rate bounds certified by our IQC framework could be loose even with all off-by-n IQCs and one might like to use as many IQCs as possible to make the obtained bounds less conservative. Nevertheless, we find in practice that it is possible to certify tight convergence bounds using only a small number of IQCs for algorithms we consider. In particular, the sector IQCs in Lemma 1 are sufficient[6] to provide a tight bound for GD, and adding more IQCs will not improve the rate bound obtained by our IQC framework[7]. Formally, we offer the following conjecture based on our numerical simulations:

Conjecture 1 *For first-order algorithms with T steps of memory, we only need off-by-n pointwise IQCs up to T to get the tightest convergence rate in our framework. In other words, adding off-by-n pointwise IQCs with $n > T$ in SDP* (14) *will not improve the bound.*

6. We mean the obtained rate bound matches the known lower bound exactly.
7. In particular, the corresponding λ_i of newly added constraint (Ψ_i, M_i) after solving the SDP would be zero up to numerical precision, manifesting the constraint is redundant.

We numerically verified our conjecture for GD and algorithms with one step of memory (see Figure 11)[8]. For instance, we notice that a combination of the sector and off-by-one pointwise IQCs is enough for algorithms with one step of memory (6). Interestingly, such combination is far from enough to get tight bounds for minimization problem.

3.1 Warm-up: Analysis of Gradient Method

We first warm up with the simplest algorithm – GD. The recursion is given by

$$z_{k+1} = z_k - \eta F(z_k). \tag{20}$$

We will analyze this algorithm by applying Theorem 1. The first thing is to find proper IQCs for GD. By the assumptions, we know the vector field F is m-strongly monotone and L-Lipschitz. We may start with the sector IQCs defined in Lemma 1[9].

By exploiting the structure of the problem (see Lessard et al. (2016, section 4.2)), we are able to reduce the problem to the following SDP by simply setting $P = 1$ without loss of generality:

$$\begin{bmatrix} 1-\rho^2 & -\eta \\ -\eta & \eta^2 \end{bmatrix} + \lambda_1 \begin{bmatrix} L^2 & 0 \\ 0 & -1 \end{bmatrix} + \lambda_2 \begin{bmatrix} -2m & 1 \\ 1 & 0 \end{bmatrix} \preceq 0. \tag{21}$$

We remark that the SDP (21) is independent of the dimension d. Using Schur complements (Haynsworth, 1968), it is equivalent to

$$\lambda_1 \geq \eta^2 \quad \lambda_2 \geq 0 \quad \rho^2 \geq 1 + \lambda_1 L^2 - 2\lambda_2 m + \frac{(\lambda_2 - \eta)^2}{\lambda_1 - \eta^2}. \tag{22}$$

By analyzing the lower bound on ρ^2 in (22), one can easily show that $\rho^2 \geq 1 - 2m\eta + L^2\eta^2$. Optimizing over η, we get $\rho^2 \geq 1 - \frac{1}{\kappa^2}$, matching the lower bound of convergence rate of GD (Azizian et al., 2020b). Notably, we only impose sector-bounded constraints in this section, which means the vector field can change over time (time-varying system).

After giving the warm-up example, we now present a deep result of GD for its optimality on time-varying systems. Particularly, we show that GD with stepsize $\eta = m/L^2$ achieves the fastest possible worst-case convergence rate not only among all tunings of GD, but among any algorithm where z_{k+1} depends linearly on $\{z_k, z_{k-1}, ..., z_{k-l}\}$ for some fixed l.

Theorem 2 *With only the sector IQCs (i.e., the system can be time-varying), the best worst-case convergence rate in solving SDP* (14) *is achieved by GD with stepsize $\eta = m/L^2$ among all algorithms representable as a linear time-invariant system with finite state.*

To put it differently, when the system is time-varying, we cannot improve our upper bound by using more complex algorithms.

3.2 Analysis of Proximal Point Method

While GD is discretizing vector field flow with forward Euler method and suffers from overshotting problem, proximal point method (PPM) (Rockafellar, 1976; Parikh and Boyd,

8. We also tested on some algorithms with two step of memory with randomly sampled parameters.
9. As we argue in Conjecture 1, adding more IQCs probably will not improve the bound of GD.

2014) adopts backward Euler method and is more stable.

$$z_{k+1} = z_k - \eta F(z_{k+1}). \tag{23}$$

Although PPM is in general not efficiently implementable, it is largely regarded as a "conceptual" guiding principle for accelerating optimization algorithms (Drusvyatskiy, 2017; Ahn, 2020). Indeed, Mokhtari et al. (2020a) showed that both OG and EG are approximating PPM in the context of smooth games. Nevertheless, the convergence analysis of PPM is in general more involved than gradient descent method. Here we follow the same IQC pipeline to analyze its convergence rate. Notably, PPM can still be expressed as a discrete linear system as in (5) but with the matrix $D = -\eta \mathbf{I}_d$. Similar to GD, we impose the sector IQCs and reduce the problem to the following SDP:

$$\begin{bmatrix} 1-\rho^2 & -\eta \\ -\eta & \eta^2 \end{bmatrix} + \lambda_1 \begin{bmatrix} L^2 & -\eta L^2 \\ -\eta L^2 & \eta^2 L^2 - 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} -2m & 2\eta m + 1 \\ 2\eta m + 1 & -2\eta^2 m - 2\eta \end{bmatrix} \preceq 0. \tag{24}$$

Our goal is to find the minimal ρ such that this LMI is feasible. To achieve that, we can use Schur complements and optimize λ_1, λ_2 to lower bound ρ. Towards this end, we are able to prove the exponential convergence of PPM by solving the LMI (24).

Theorem 3 *Under Assumption 1 and 2, PPM converges linearly with any positive η.*

$$\|z_k - z^*\|_2^2 \leq \left(\frac{1}{1+2\eta m}\right)^k \|z_0 - z^*\|_2^2, \quad \textit{for all } k \geq 0. \tag{25}$$

As opposed to the rate bound of GD, the convergence rate $\rho^2 = \frac{1}{1+2\eta m}$ does not depend on the Lipschitz constant L and is strictly smaller than 1 for all positive stepsize η. Moreover, our bound is better than the one in Mokhtari et al. (2020a, Theorem 2) with rate $\rho^2 = \frac{1}{1+\eta m}$. It is important to know that PPM can converge arbitrarily faster with large η, but the computation of $F(z_{k+1})$ would become expensive.

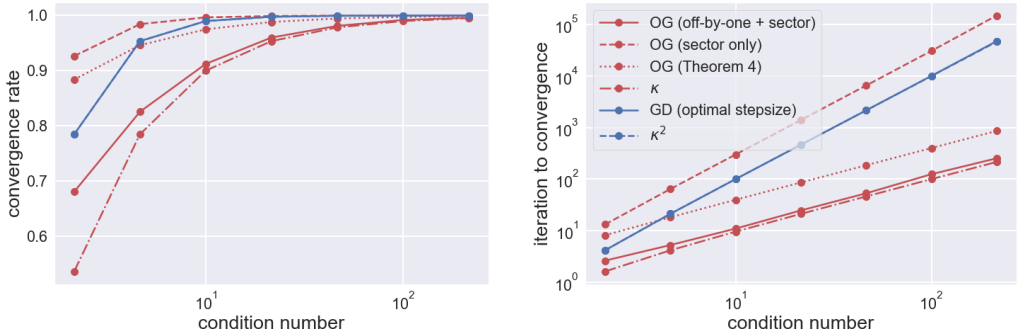
3.3 Accelerating Smooth Games With Optimism

Optimistic gradient method (OG) was shown to be an approximation to PPM (Mokhtari et al., 2020a) and it approximates $F(z_{k+1})$ with a lookahead step. From this standpoint, one may expect OG to inherit the merits of PPM and potentially improve upon plain gradient method. In this section, we analyze OG with our IQC framework and show that indeed it converges faster than GD, as also shown in Gidel et al. (2018). In particular, we study the recursion of (7). In this case, OG is also an approximation to Extra-gradient (EG) (Korpelevich, 1976) method by using past gradient (Gidel et al., 2018; Hsieh et al., 2019):

Proposition 1 *OG is an approximation to EG but using the past gradient:*

$$\begin{aligned} z_{k+1/2} &= z_k - \eta F(z_{k-1/2}), \\ z_{k+1} &= z_k - \eta F(z_{k+1/2}). \end{aligned} \tag{26}$$

Rewriting OG as a variant of EG, one can derive the following convergence result.



(a) Convergence Rate **(b)** Iteration Complexity

Figure 2: Upper bounds of convergence rate and iteration complexity for GD and OG. We test both the sector IQCs and the combination of the sector and off-by-one pointwise IQCs for OG. We tuned the step sizes of OG and GD using grid search. Compared to the rate in Theorem 4, we are able to improve the bound by roughly a factor of 4. Two blue lines are virtually identical.

Theorem 4 (Gidel et al. (2018)) *Under Assumption 1 and 2, if we take* $\eta = 1/(4L)$*, then*

$$\|z_k - z^*\|_2^2 \leq \left(1 - \frac{1}{4\kappa}\right)^k \|z_0 - z^*\|_2^2, \quad \textit{for all } k \geq 0. \tag{27}$$

Theorem 4 suggests that OG has an iteration complexity of $\mathcal{O}(\kappa)$, which indeed accelerates GD substantially. Notably, this also implies that OG is *near optimal* in the sense that it matches the lower bound (8) up to a constant (Azizian et al., 2020b, Corollary 1). However, the proof of Theorem 4 is quite involved and relies on a cleverly designed Lyapunov function. Here, we improve the rate bound using IQC machinery.

We compute the rate bounds using Theorem 1 with either the sector IQCs in Lemma 1 or a combination of the sector and off-by-one pointwise IQCs in Lemma 2. In contrast to GD, the SDP problem induced by OG is not analytically solvable anymore, thus we use bisection search to find the optimal rate ρ. For fixed ρ and κ, the SDP (14) become an LMI and can be efficiently solved using interior-point methods (Boyd et al., 2004). For all simulations in the paper, we use CVXPY (Diamond and Boyd, 2016) package with Mosek solver.

With the sector IQCs alone, we observe that OG may diverge if we take $\eta = 1/4L$ in the sense that the best rate achieved is $\rho \geq 1$ even for very small condition numbers. To understand why, recall from Lemma 1 that the sector IQCs allow for F_k to be different at each iteration. Unlike GD, OG is not robust to having a changing F_k. We further conjecture that the divergence of OG is caused by the aggressive step size choice (compared to m/L^2 in GD), we therefore tune the step size for OG. Figure 2 shows the certified convergence rates of OG. We find that the optimal step size for OG in that setting is much smaller than $1/4L$. Moreover, its iteration complexity scales quadratically with condition number κ and is perhaps worse than GD by a constant. This matches the prediction of Theorem 2 that GD is provably optimal for time-varying systems.

On the other hand, if we add off-by-one pointwise IQCs to the LMI, the bound for OG does improve upon that of GD, especially when the condition number is large (see red solid lines in Figure 2). This suggests that enforcing the consistency of two consecutive vector

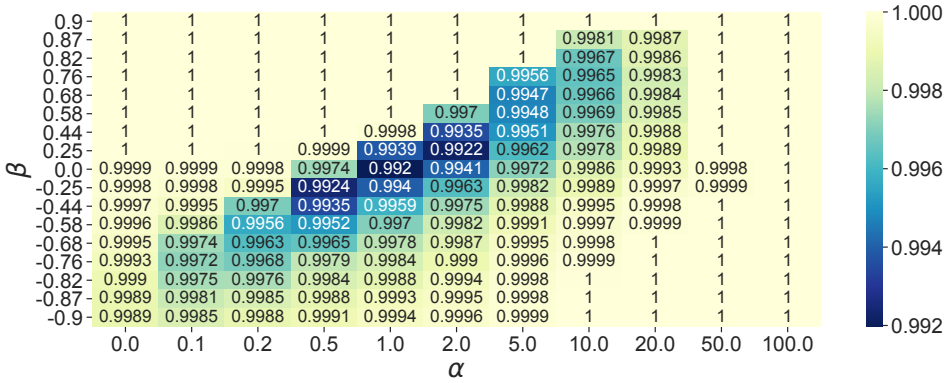


Figure 3: Grid search over algorithms with one step memory (6), where α and β are two parameters for this algorithm family. Here, we compute the convergence rates using the condition number $\kappa = 100$. It turns out that optimistic gradient method (OG) with $\alpha = 1$ and $\beta = 0$ has the best convergence rate over all combinations of α, β.

field quries is important for the acceleration of OG. In this case, the complexity of OG scales linearly with condition number, matching existing bounds. Moreover, the convergence rate improves upon that of Gidel et al. (2018) (see Theorem 4) by roughly a constant factor of 4, highlighting the usefulness of IQCs for certifying sharp bounds.

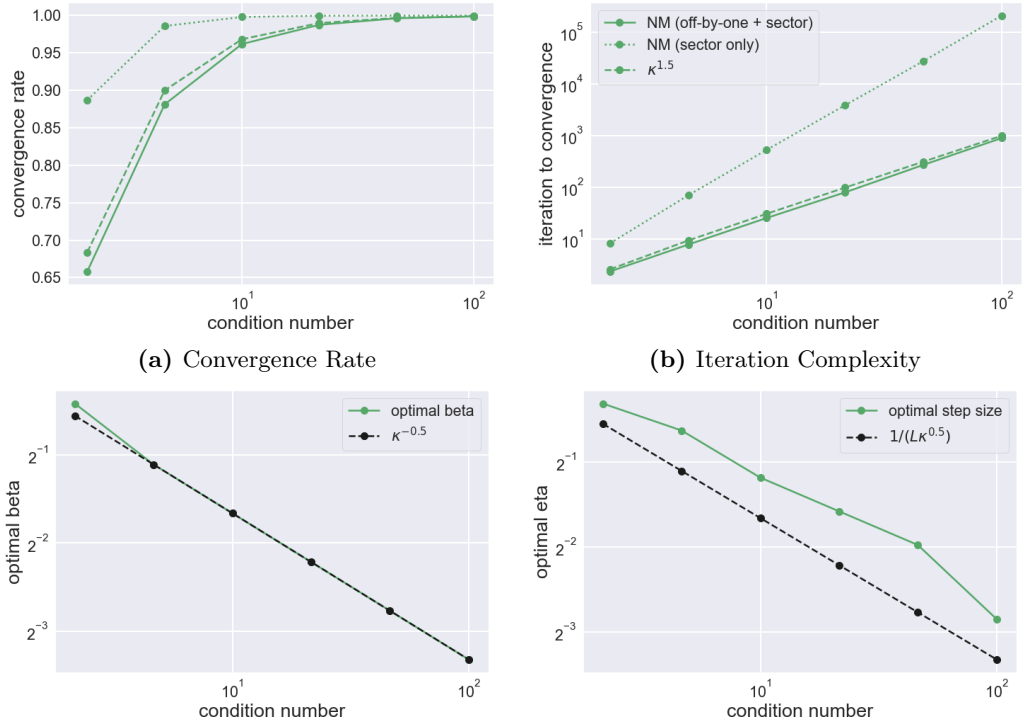
Lastly, we may ask the question how OG performs over the family of algorithms with one step of memory (6). This is easy to carry out numerically since one can search for the minimal ρ^2 for different combinations of α, β. To be precise, we conduct grid search over α, β, η and interestingly it turns out that the optimal parameters are $\alpha = 1$ and $\beta = 0$, corresponding exactly to OG (see Figure 3). In other words, OG appears likely to be optimal within the family of algorithms with one step of memory.

3.4 Global Convergence of Negative Momentum

In the preceding sections, we recovered or improved previously known convergence rates of GD, PPM and OG either analytically or numerically. One may further ask whether we can provide the convergence rates of some algorithms which were unknown before based on our IQC analysis. We answer this question in the affirmative for deriving a novel convergence bound of negative momentum, which essentially refers to Polyak momentum with a negative damping parameter.

Negative momentum was first studied in Gidel et al. (2019) on simple bilinear games. Later, it was shown by Zhang and Wang (2021) that negative momentum converges *locally* with an iteration complexity of $\mathcal{O}(\kappa^{1.5})$ for smooth and strongly-monotone variational inequality problems. More importantly, Zhang and Wang (2021) showed that the bound is tight asymptotically by proving a lower bound of $\Omega(\kappa^{1.5})$. Yet, it is unclear whether negative momentum can converge globally with the same rate. In general, it is highly non-trivial to prove an explicit global convergence rate for Polyak momentum. For example, Ghadimi et al. (2015) can only show that Polyak momentum converges globally with properly chosen parameters but no explicit rate was provided.

Using a combination of the sector and off-by-one pointwise IQCs, we evaluate the rates of negative momentum numerically by doing bisection search on ρ and grid search on the pa-



(a) Convergence Rate **(b)** Iteration Complexity

(c) Optimal momentum parameter **(d)** Optimal step size

Figure 4: Top: Curves of convergence rate and iteration complexity for negative momentum with tuned β and η. The dashed line is the known lower bound of NM $\rho^2 = 1 - \kappa^{-1.5}$ up to a unspecified constant. The dotted line is obtained by only using the sector IQCs with $\beta = \kappa^{-0.5} - 1$. **Bottom**: Optimal momentum value and step size for negative momentum as functions of condition number κ. For momentum parameter, we plot $\beta + 1$ for clarity.

rameter β and step size η. We report the results in Figure 4. Unexpectedly, the complexity curve of negative momentum has a slope of 1.5, suggesting it attains the same complexity of $\mathcal{O}(\kappa^{1.5})$ globally. Also, the rate matches the known lower bound tightly (see the dashed line in Figure 4). This is surprising, in that Polyak momentum fails to achieve the same accelerated convergence rate (i.e., its local convergence rate) globally in the convex optimization setting (Lessard et al., 2016). Furthermore, we find that the optimal step size and momentum parameter follow simple functions of condition number κ. According to our simulations, the optimal step size is roughly $\frac{1}{L\sqrt{\kappa}}$ while the optimal momentum value is $\kappa^{-0.5} - 1$, as shown in Figure 4. To the best of our knowledge, we provide the *first* global convergence rate guarantee for negative momentum using our IQC framework, something which is otherwise difficult to prove.

4. IQCs for Stochastic Games

We have been discussing handling the nonlinear element F of the variational inequality with IQCs. Here, we further extend it to model the uncertainty in computing the vector field F. Of particular interest to us is the situation where $F(z) = \mathbb{E}_\epsilon F(z; \epsilon)$ is the expectation with respect to random variable ϵ of the random operator $F(z; \epsilon)$. Inspired by recent works on

the interpolation regime, we consider a *strong growth condition* (Vaswani et al., 2019):

$$\mathbb{E}_\epsilon \|F(z;\epsilon)\|_2^2 \le \delta \|F(z)\|_2^2. \tag{28}$$

Equivalently, in the finite-sum setting:

$$\mathbb{E}_i \|F_i(z)\|_2^2 \le \delta \|F(z)\|_2^2. \tag{29}$$

For this inequality to hold, if $F(z) = 0$, then $F_i(z) = 0$ for all i. This noise model is an instance of multiplicative noise in the sense that the perturbation noise is a function of the state z. Note that this noise model has been shown to hold for overparameterized models (Ma et al., 2018; Liu and Belkin, 2018) and underlies exponential convergence of stochastic gradient based algorithms (see Strohmer and Vershynin (2009); Moulines and Bach (2011); Ma et al. (2018)). It is also possible to include additive noise by using the *bias-variance* decomposition (Bach and Moulines, 2013; Fallah et al., 2020). For numerical tractability, we here focus on the finite-sum setting with n examples. Later, we will show in Theorem 6 that the convergence rate is independent of n for all $n \ge 2$. In that case, we can model optimization algorithms as stochastic jump systems (Costa et al., 2006):

$$\begin{aligned} \xi_{k+1} &= A\xi_k + B_{i_k} u_k \\ y_k &= C\xi_k + D u_k \\ u_k &= [F_1(y_k)^\top, ..., F_n(y_k)^\top]^\top. \end{aligned} \tag{30}$$

For the gradient method, the matrix B_{i_k} is simply $(-\eta \mathbf{e}_{i_k}^\top) \otimes \mathbf{I}_d$ where $\mathbf{e}_{i_k}$ is a one-hot vector with i_k-entry being 1. Similar to the deterministic system, we can impose quadratic constraints by designing $s = \Psi(y, u)$ and matrix M. For example, in the case of $n = 2$, we can enforce L-Lipschitzness of F together with the strong growth condition as follows (where we ignore the dimension since we can factorize all the matrices as Kronecker products):

$$\Psi = \left[\begin{array}{c|c|cc} 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right], \quad M_1 = \begin{bmatrix} L^2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \delta/2 - 1 & \delta/2 \\ 0 & 0 & \delta/2 & \delta/2 - 1 \end{bmatrix}.$$

Again, combining the dynamics (30) with Ψ, we have the following compact form:

$$\begin{aligned} x_{k+1} &= \hat{A} x_k + \hat{B}_{i_k} u_k \\ s_k &= \hat{C} x_k + \hat{D} u_k \end{aligned}, \quad \text{where } x_k = \begin{bmatrix} \xi_k \\ \zeta_k \end{bmatrix}. \tag{31}$$

Assuming i_k is drawn uniformly in an i.i.d manner, we have

Theorem 5 (Hu et al. (2017, Theorem 1)) *Consider the stochastic jump system* (30)*. Suppose F satisfies the pointwise IQC specified by (Ψ, M), and consider the following LMI:*

$$\begin{bmatrix} \hat{A}^\top P \hat{A} - \rho^2 P & \frac{1}{n}\sum_{i=1}^n \hat{A}^\top P \hat{B}_i \\ \frac{1}{n}\sum_{i=1}^n \hat{B}_i^\top P \hat{A} & \frac{1}{n}\sum_{i=1}^n \hat{B}_i^\top P \hat{B}_i \end{bmatrix} + \lambda \begin{bmatrix} \hat{C} & \hat{D} \end{bmatrix}^\top M \begin{bmatrix} \hat{C} & \hat{D} \end{bmatrix} \preceq 0. \tag{32}$$

If this LMI is feasible with $P \succ 0$ and $\lambda \geq 0$, then the following inequality holds,

$$\mathbb{E}[(x_{k+1} - x^*)^\top (P \otimes \mathbf{I}_d)(x_{k+1} - x^*)] \leq \rho^2 (x_k - x^*)^\top (P \otimes \mathbf{I}_d)(x_k - x^*). \tag{33}$$

Consequently, we have $\mathbb{E}\|\xi_k - \xi^*\|_2^2 \leq \mathrm{cond}(P)\rho^{2k}\|\xi_0 - \xi^*\|_2^2$ for any ξ_0 and $k \geq 1$.

Similar to Theorem 1, when ρ^2 is given, the condition (32) is linear with respect to P and λ. Therefore, it is an LMI whose feasible set is convex and can be effectively solved using the state-of-the-art convex optimization techniques, such as interior-point method (Boyd et al., 2004). Besides, it is important to know that the size of the LMI condition (32) scales proportionally with n. Nevertheless, one can show that the optimal ρ^2 is independent of n under the strong growth condition for algorithms we consider.

Theorem 6 *For algorithms with one step of memory, the feasible set of the LMI* (32) *is independent of n when $n \geq 2$, hence the optimal solution ρ^2 of the SDP in Theorem 5 under the strong growth condition is independent of n when $n \geq 2$.*

Therefore, we could safely choose $n = 2$ in all our numerical simulations.

4.1 The Robustness of Gradient Method

We now analyze the dynamics of GD to determine if it is robust to noisy gradients. It is easy to show (without using Theorem 5) that with properly scaled step size, GD maintains the iteration complexity of $\mathcal{O}(\kappa^2)$ which we derived for the deterministic case.

Theorem 7 (GD with the strong growth condition) *Under Assumptions 1 and 2, if we further assume the vector field F satisfies the strong growth condition* (28) *with parameter δ and take $\eta = 1/(L\kappa\delta)$, then we have*

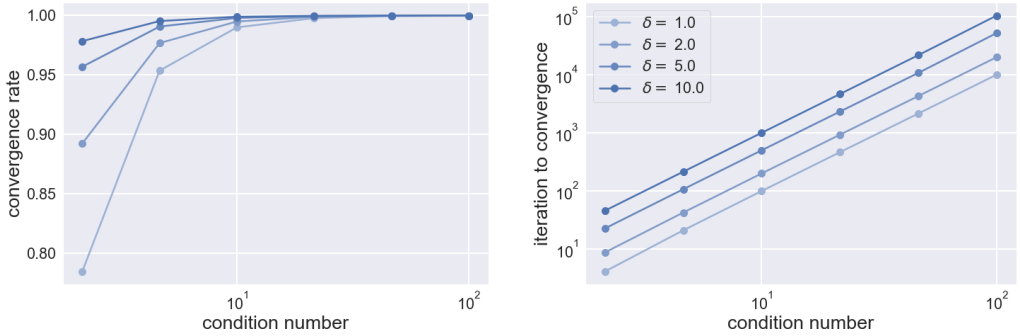
$$\mathbb{E}\|z_k - z^*\|_2^2 \leq \left(1 - \frac{1}{\kappa^2\delta}\right)^k \|z_0 - z^*\|_2^2. \tag{34}$$

Compared to the rate of deterministic setting, the rate in Theorem 7 is worse by the constant factor δ. And as expected, the noisier the vector field computation (i.e., larger δ), the slower the convergence. Nevertheless, the scaling with the condition number κ matches the deterministic setting, manifesting the robustness of GD.

To sanity check our IQC framework, we also compute the rate bounds using Theorem 5 by setting $n = 2$ for convenience. We note that the result is independent of the value of n, choosing $n = 2$ makes the SDP problem easy to solve. We observe that the numerical rates obtained by our IQC framework match the prediction of Theorem 7 exactly, as shown in Figure 5. This also implies that the upper bound in Theorem 7 is probably sharp.

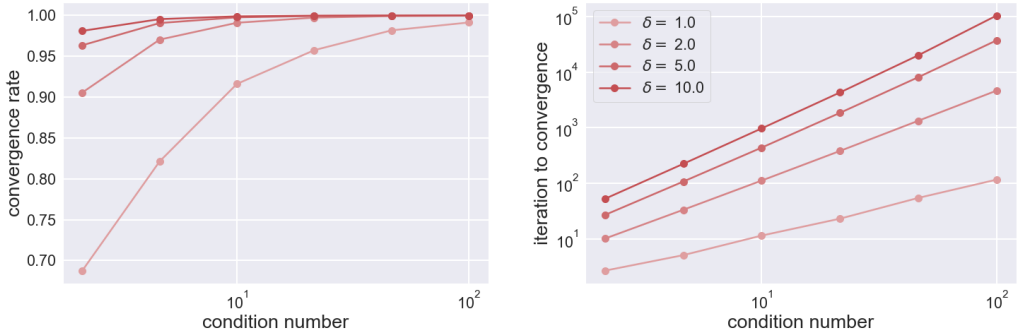
4.2 The Brittleness of Optimistic Gradient Method and Negative Momentum

As discussed in the last section, GD is robust to multiplicative noise when it satisfies the strong growth condition (28). It is natural to ask whether the same is true of OG and NM. Namely, are they able to match their respective deterministic convergence rates and hence accelerate GD in the stochastic setting?



(a) Convergence Rate **(b)** Iteration Complexity

Figure 5: Convergence rate (or iteration complexity) of GD with tuned stepsize as a function of condition number under different noise levels. $\delta = 1$ is basically the deterministic setting we studied. For a given δ, it takes $\mathcal{O}(\kappa^2)$ iterations to converge no matter how large δ is.



(a) Convergence Rate **(b)** Iteration Complexity

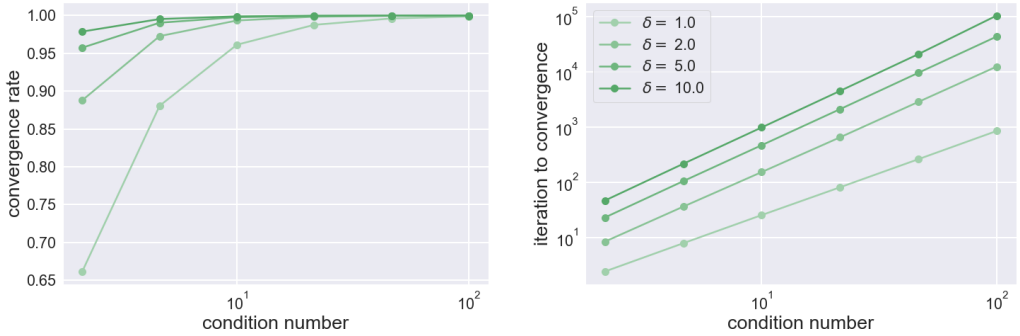
Figure 6: Convergence rate (or iteration complexity) of OG with tuned stepsize as a function of condition number under different noise levels. In the case of $\delta = 1$ (i.e., the deterministic setting), it takes $\mathcal{O}(\kappa)$ iterations for OG to converge. Increasing the noise level with a larger δ degrades the rate. When $\delta = 10$, the iteration complexity is close to $\mathcal{O}(\kappa^2)$, which is no better than GD.

We compute the convergence rates of OG using Theorem 5 together with the sector and off-by-one IQCs. We search for the optimal step size η using grid-search. As shown in Figure 6, the convergence rate of optimally tuned OG deteriorates as we use $\delta > 1$ and the complexity is roughly $\mathcal{O}(\kappa^2)$ when $\delta \gg 1$. In other words, the convergence rate of OG is no better than that of GD in the stochastic setting.

We also analyze negative momentum (NM) using Theorem 5 with the momentum parameter $\beta = \kappa^{-0.5} - 1$ and tuned step size. Figure 7 shows the plots of convergence rate and iteration complexity for different noise levels. Similar to OG, NM suffers as we gradually increase the noise level δ from 1 to 10. In particular, its complexity scales quadratically as a function of condition number when $\delta \gg 1$. This is to be expected by analogy with the minimization case that momentum method is fragile to injected noise.

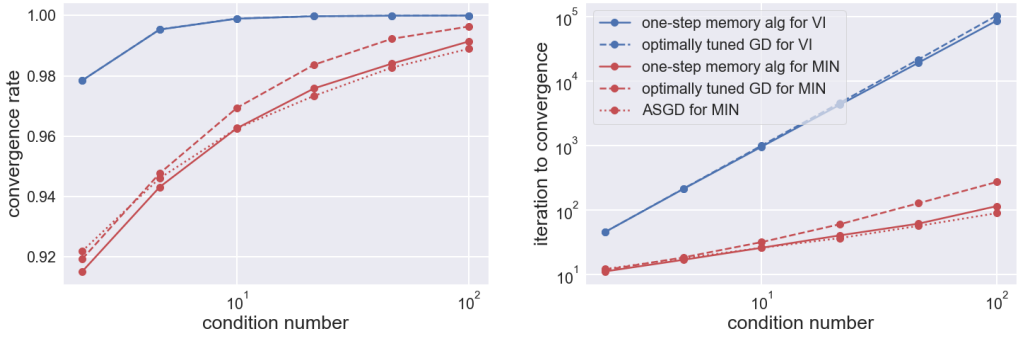
4.3 Is It Possible to Accelerate GD in the Stochastic Setting?

In the last section, we showed that both OG and NM fail to accelerate GD in the presence of noise. One may ask: *does there exist any algorithm with only one step of memory that can*



(a) Convergence Rate **(b)** Iteration Complexity

Figure 7: Convergence rate (or iteration complexity) of NM with tuned stepsize as a function of condition number under different noise levels. In the case of $\delta = 1$ (i.e., the deterministic setting), it takes $\mathcal{O}(\kappa^{1.5})$ iterations for NM to converge. Increasing the noise level with a larger δ degrades the rate. When $\delta = 10$, the iteration complexity is close to $\mathcal{O}(\kappa^2)$, which is no better than GD.

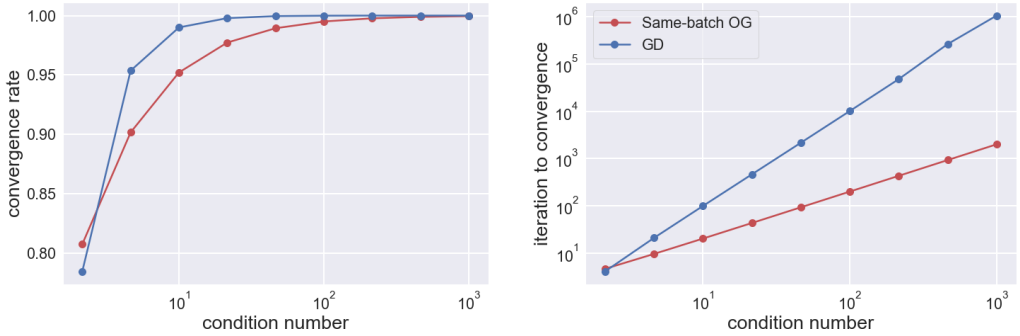


(a) Convergence Rate **(b)** Iteration Complexity

Figure 8: Comparison between optimally tuned GD and optimally tuned one step memory algorithm in the case $\delta = 10$. For one step memory algorithm, we search $\beta + 1$ or $1 - \beta$ in log space uniformly from $[\frac{1}{\kappa\delta}, 1.0]$. For α, we search over $[0.0, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0, 50.0, 100.0]$. We use VI for the abbreviation of variational inequality and MIN for minimization. Accelerated Stochastic Gradient Descent (ASGD) (Jain et al., 2018) is a variant of the Nesterov Accelerated Gradient which is able to accelerate SGD for minimizing strongly-convex functions under the strong growth condition. We provide a detailed proof for the acceleration effect of ASGD in Appendix B.2.

achieve acceleration in the stochastic setting? In this section, we first show that acceleration is impossible if the algorithm queries each batch of data only once before moving on to the next one. We then answer our question in the affirmative by showing there exists an algorithm achieving acceleration by querying each batch of data twice.

We first search over algorithms with one step of memory (6) by doing a grid search over values of α and β for every particular condition number κ. In particular, we set δ to be 10 since the slope of resulting curve stays unchanged with larger δ. This experiment is easy to carry out in our framework, because choosing new values of α and β simply amounts to changing parameters in the LMI. We find that no algorithm is provable (under our IQC model) to obtain a faster convergence rate than the $\mathcal{O}(\kappa^2)$ rate obtained for GD (see Figure 8). This is in stark contrast to minimizing a strongly-convex function, where there is an algorithm with one step of memory accelerating GD under the strong growth



(a) Convergence Rate **(b)** Iteration Complexity

Figure 9: Convergence rates and iteration complexities of GD and same-batch OG under the assumptions (39). By using the same stochastic operator (i.e., query the same batch twice) in (38), same-batch OG accelerates GD with an iteration complexity of $\mathcal{O}(\kappa)$.

condition (Jain et al., 2018; Vaswani et al., 2019). More importantly, we do match the rate of this algorithm by conducting the same grid search over α and β (see red solid line).

The previous result inspires us to re-examine the reason why OG can accelerate GD in the first place when F is noiseless. By inspecting λ_i in the final solution of the LMI (14)[10], we notice the convergence analysis of deterministic OG heavily relies on the following property:

$$\|z_{k+1} - z_{k+1/2}\| = \eta\|F(z_{k+1/2}) - F(z_{k-1/2})\|_2 \leq \eta L\|z_{k+1/2} - z_{k-1/2}\|_2. \tag{35}$$

where we used the L-Lipschitz assumption of F. However, when the stochastic update is used, this property no longer holds. Recall the OG update in the stochastic setting:

$$z_{k+1} = z_k - \eta F(2z_k - z_{k-1}; \epsilon_k). \tag{36}$$

According to Proposition 1, OG can be rewritten as the following form:

$$\begin{aligned} z_{k+1/2} &= z_k - \eta F(z_{k-1/2}; \epsilon_{k-1}), \\ z_{k+1} &= z_k - \eta F(z_{k+1/2}; \epsilon_k). \end{aligned} \tag{37}$$

Observe that two different stochastic operators $F(\cdot; \epsilon_{k-1})$ and $F(\cdot; \epsilon_k)$ are used, so (35) need not hold for any value of L. One can fix this problem by sharing the same stochastic operator (e.g. using the same batch of data) to compute the updates, which we term *same-batch OG*. This fix was first proposed in Mishchenko et al. (2020) for extra-gradient.

$$\begin{aligned} z_{k+1/2} &= z_k - \eta F(z_{k-1/2}; \epsilon_k), \\ z_{k+1} &= z_k - \eta F(z_{k+1/2}; \epsilon_k). \end{aligned} \tag{38}$$

To prove convergence, we have to replace Assumption (1) and (2) with stronger assumptions (z_1, z_2 could depend on ϵ):

$$\begin{aligned} \mathbb{E}[\|F(z_1; \epsilon) - F(z_2; \epsilon)\|_2^2] &\leq \mathbb{E}[L(\epsilon)^2\|z_1 - z_2\|_2^2] \leq L^2\mathbb{E}[\|z_1 - z_2\|_2^2], \\ \mathbb{E}[(F(z_1; \epsilon) - F(z_2; \epsilon))^\top (z_1 - z_2)] &\geq \mathbb{E}[m(\epsilon)\|z_1 - z_2\|_2^2] \geq m\mathbb{E}[\|z_1 - z_2\|_2^2]. \end{aligned} \tag{39}$$

10. In all four quadratic constraints, only the strongly-monotone sector IQC and the Lipschitz off-by-one pointwise IQC are used with non-zero λ.

Basically, we allow different $F(\cdot;\epsilon)$ to have distinct Lipschitz and monotone constants. With minor modifications to our IQC analysis (see Appendix B.3 for details), we can show same-batch OG (38) accelerates GD with an iteration complexity of $\mathcal{O}(\kappa)$, as seen in Figure 9. We remark that assumptions (39) are less restrictive than the ones used in (Mishchenko et al., 2020) where they require $F(\cdot;\epsilon)$ to be almost surely strongly-monotone and Lipschitz.

5. Discussion

Smooth game optimization has recently emerged as a new paradigm for many models in machine learning due to its flexibility to model multiple players and their interactions. Nevertheless, the dynamics of games are more complicated than their single-objective counterparts, and raise new algorithmic challenges. We believe a unified and systematic analysis framework is crucial, since it could save us from the pain of analyzing algorithms in a case-by-case manner. To this end, we argue that the introduced IQC framework is a very powerful tool to study game dynamics, especially when the system contains nonlinear and uncertain ingredients.

We note that our current framework is limited to strongly-monotone and smooth games, but other techniques from control theory (e.g., dissipativity theory (Hu and Lessard, 2017b)) may allow us to certify sublinear rates in general monotone games. Similarly to Lessard et al. (2016), our IQC framework could also be extended to the non-smooth setting. However, the numerical results might be less interpretable because most algorithms fail to attain linear convergence in the non-smooth setting. Another limitation is that our IQC framework is not generally applicable for algorithms accessing higher-order information (e.g., competitive gradient descent (Schäfer and Anandkumar, 2019)). Nevertheless, for algorithms that can be written as a first-order method on modified utility functions (e.g., consensus optimization (Mescheder et al., 2017)[11]), it is possible to apply our IQC framework for tight convergence analysis. Exploring new types of IQCs that can be used to analyze algorithms using high-order information would be an interesting future direction.

So far for all the algorithms we analyzed, we have shown that our IQC framework provides tight bound certification as long as the algorithm can fit into the variational inequality framework. To be noted, our framework can also be used to analyze the extra-gradient method which we did not discuss in the paper. Nonetheless, for problems with additional structure (e.g., the bilinear saddle point problem $\min_x \max_y f(x) - g(y) + x^\top By$), additional modifications are required to take into account the structural information for tight bounds.

Finally, one of the biggest limitations of our IQC framework is that it provides only a numerical proof, except in simpler cases where the SDP can be solved analytically. However, as noted in Lessard et al. (2016), it might be possible to find analytical proofs for complex SDPs using tools from algebraic geometry (Grayson and Stillman, 2002; Rostalski and Sturmfels, 2010). Furthermore, there might exist examples that require large numbers of IQCs to get tight bounds, making the corresponding SDPs hard to solve. In our own investigations, a handful of IQCs have sufficed to obtain tight bounds, and we expect that other limited memory algorithms can be analyzed with similarly compact IQCs.

11. Consensus optimization can be viewed as gradient descent algorithm on modified objectives that include additional gradient norm penalties.

Acknowledgments

We thank Bryan Van Scoy and Yuanhao Wang for many helpful discussions. We thank Shengyang Sun, Xuechen Li and Guojun Zhang for detailed comments on early drafts. We also thank Adrien Taylor for pointing out a mistake of the necessary and sufficient condition for $\{m, L\}$-interpolation in the first version of our paper. Besides, we thank the anonymous JMLR reviewers for their useful feedback on earlier versions of this manuscript.

GZ would like to thank for the supports from Borealis AI fellowship and Ontario Graduate Scholarship. RG acknowledges support from the CIFAR Canadian AI Chairs program.

Appendix A. Proofs for Theoretical Results

A.1 Proofs for Section 2

Theorem 1 *Consider the dynamical system* (5)*. Suppose the vector field F satisfies the pointwise IQC (Ψ, M) and define $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ according to* (11)–(13)*. Consider the following linear matrix inequality (LMI):*

$$\begin{bmatrix} \hat{A}^\top P \hat{A} - \rho^2 P & \hat{A}^\top P \hat{B} \\ \hat{B}^\top P \hat{A} & \hat{B}^\top P \hat{B} \end{bmatrix} + \lambda \begin{bmatrix} \hat{C} & \hat{D} \end{bmatrix}^\top M \begin{bmatrix} \hat{C} & \hat{D} \end{bmatrix} \preceq 0. \tag{14}$$

If this LMI is feasible for some $P \succ 0$, $\lambda \geq 0$ and $\rho > 0$[12]*, we have*

$$(x_{k+1} - x^*)^\top (P \otimes \mathbf{I}_d)(x_{k+1} - x^*) \leq \rho^2 (x_k - x^*)^\top (P \otimes \mathbf{I}_d)(x_k - x^*). \tag{15}$$

Consequently, for any ξ_0 and $\zeta_0 = \zeta^*$, we obtain

$$\|\xi_k - \xi^*\|_2^2 \leq \mathrm{cond}(P) \rho^{2k} \|\xi_0 - \xi^*\|_2^2. \tag{16}$$

Proof Let x, u, s be a set of sequences that satisfies (13). Suppose (P, λ) is a solution of SDP (14). Multiply (14) on the left and right by $[(x_k - x^*)^\top, (u_k - u^*)^\top]$ and its transpose, respectively. Making use of (13) and (10), we obtain

$$(x_{k+1} - x^*)^\top P (x_{k+1} - x^*) - \rho^2 (x_k - x^*)^\top P (x_k - x^*) + \lambda (s_k - s^*)^\top M (s_k - s^*) \leq 0 \tag{40}$$

Because F satisfies the pointwise IQC definied by (Ψ, M), therefore we obtain

$$(x_{k+1} - x^*)^\top P (x_{k+1} - x^*) \leq \rho^2 (x_k - x^*)^\top P (x_k - x^*)$$

for all k and consequently $\|x_k - x^*\|_2 \leq \sqrt{\mathrm{cond}(P)} \rho^k \|x_0 - x^*\|_2$. Recall from (13) that $x_k = (\xi_k, \zeta_k)$ and $\zeta_0 = \zeta^*$, we therefore have

$$\begin{aligned} \|\xi_k - \xi^*\|_2^2 &\leq \|x_k - x^*\|_2^2 \\ &\leq \mathrm{cond}(P) \rho^{2k} \|x_0 - x^*\|_2^2 \\ &= \mathrm{cond}(P) \rho^{2k} (\|\xi_0 - \xi^*\|_2^2 + \|\zeta_0 - \zeta^*\|_2^2) \\ &= \mathrm{cond}(P) \rho^{2k} \|\xi_0 - \xi^*\|_2^2 \end{aligned}$$

and this completes the proof. ■

A.2 Proofs for Section 3

Lemma 1 (Sector IQCs) *Suppose vector field F_k is m-strongly monotone and L-Lipschitz for all k, if $u_k = F_k(y_k)$, then $\phi := (F_0, F_1, ...)$ satisfies the **pointwise IQCs** defined by*

$$\Psi_1 = \Psi_2 = \left[\begin{array}{c|cc} \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \hline \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d \end{array}\right], \quad M_1 = \begin{bmatrix} L^2 \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix}, \quad M_2 = \begin{bmatrix} -2m \mathbf{I}_d & \mathbf{I}_d \\ \mathbf{I}_d & \mathbf{0}_d \end{bmatrix}. \tag{17}$$

12. Note that ρ is not necessarily smaller than 1.

We have corresponding quadratic inequalities:

$$\begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix}^\top \begin{bmatrix} L^2\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix} \begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix} \geq 0 \quad \textit{and} \quad \begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix}^\top \begin{bmatrix} -2m\mathbf{I}_d & \mathbf{I}_d \\ \mathbf{I}_d & \mathbf{0}_d \end{bmatrix} \begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix} \geq 0.$$

Proof Two quadratic inequalities follows immediately from (3) and (4). ■

Lemma 2 (Off-by-one pointwise IQCs) *Suppose F is m-strongly monotone and L-Lipschitz. If $u_k = F(y_k)$, then $\phi \coloneqq (F, F, ...)$ satisfies the **pointwise IQCs** defined by*

$$\Psi_1 = \Psi_2 = \left[\begin{array}{cc|c|c} \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d \\ \hline -\mathbf{I}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d & \mathbf{0}_d & \mathbf{I}_d \end{array}\right], \quad M_1 = \begin{bmatrix} L^2\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix}, \quad M_2 = \begin{bmatrix} -2m\mathbf{I}_d & \mathbf{I}_d \\ \mathbf{I}_d & \mathbf{0}_d \end{bmatrix}. \quad (18)$$

We have corresponding quadratic inequalities:

$$\begin{aligned} L^2\|y_{k+1} - y_k\|_2^2 - \|u_{k+1} - u_k\|_2^2 \geq 0, \\ (y_{k+1} - y_k)^\top (u_{k+1} - u_k - m(y_{k+1} - y_k)) \geq 0. \end{aligned} \quad (19)$$

Proof We note two quadratic inequalities follows immediately from (3) and (4) by using $(z_1, z_2) \to (y_{k+1}, y_k)$. To verify the IQC factorization, we note the state equations for Ψ given in Lemma 2 are

$$\zeta_{k+1} = \begin{bmatrix} y_k \\ u_k \end{bmatrix} \quad \text{and} \quad s_k = \begin{bmatrix} y_k - y_{k-1} \\ u_k - u_{k-1} \end{bmatrix}$$

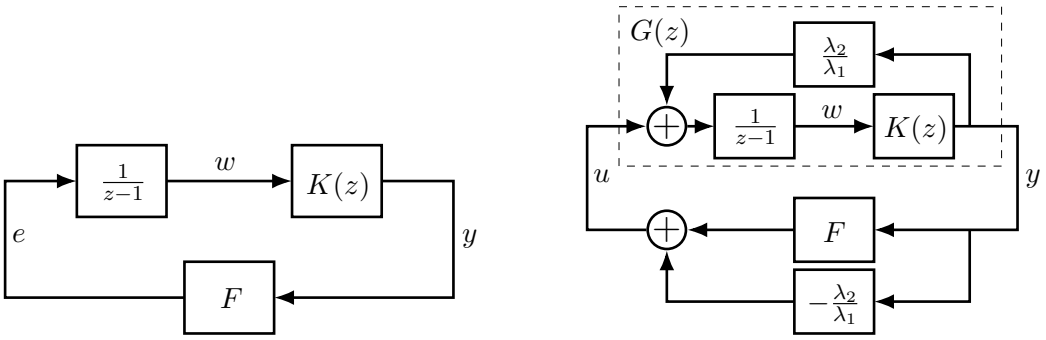
and it follows that $(s_k - s^*)^\top M_1 (s_k - s^*)$ and $(s_k - s^*)^\top M_2 (s_k - s^*)$ are equivalent to quadratic constraints (19), as required. ■

Theorem 2 *With only the sector IQCs (i.e., the system can be time-varying), the best worst-case convergence rate in solving SDP* (14) *is achieved by GD with stepsize $\eta = m/L^2$ among all algorithms representable as a linear time-invariant system with finite state.*

Proof To prove the Theorem, we need to first define a family of algorithms which are expressive enough. Following on Hu and Lessard (2017a); Lessard and Seiler (2020), we will consider algorithms set up as in Figure 10a.

The iterative algorithm must contain a pure integrator, i.e., its transfer function must take the form $K(z)\frac{1}{z-1}$. where $K(z)$ is an LTI system that represents the algorithm. Assume $K(z)$ has a state space representation (A_K, B_K, C_K, D_K). Let $w \in \mathbb{R}^{13}$ and $q \in \mathbb{R}^{n_K}$ be the state of integrator and $K(z)$, respectively. The order of $K(z)$, denoted n_K, is unspecified

13. Without loss of generality, we assume the whole system is single-input and single-output.



(a) Block diagram representing an iterative optimization algorithm.

(b) Loop transformation.

Figure 10: Block Diagram for control systems.

at this point, i.e., the algorithm may have a finite but arbitrary amount of memory. A realization of the whole algorithm then is given by

$$\begin{aligned} w_{k+1} &= w_k + e_k \\ q_{k+1} &= A_K q_k + B_K w_k \\ y_k &= C_K q_k + D_K w_k \end{aligned} \tag{41}$$

We remark that this family of algorithms is very general and can easily represent most algorithms. For example, we can recover momentum method by taking $K(z) = \frac{-\eta z}{z-\beta}$.

Under current assumptions about F, we know that we have the following quadratic constraint on the input-output pair if we only consider the sector IQCs:

$$\begin{bmatrix} y_k - y^* \\ e_k - e^* \end{bmatrix}^\top \begin{bmatrix} \lambda_1 L^2 - 2\lambda_2 m & \lambda_2 \\ \lambda_2 & -\lambda_1 \end{bmatrix} \begin{bmatrix} y_k - y^* \\ e_k - e^* \end{bmatrix} \geq 0 \tag{42}$$

where λ_1 and λ_2 are non-negative scalars. An crucial step is now to diagonalize the quadratic constraint. In particular, we perform a loop transformation as shown in Figure 10b. After the transformation, the constraint becomes

$$\begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix}^\top \begin{bmatrix} \lambda_1 L^2 - 2\lambda_2 m + \frac{\lambda_2^2}{\lambda_1} & 0 \\ 0 & -\lambda_1 \end{bmatrix} \begin{bmatrix} y_k - y^* \\ u_k - u^* \end{bmatrix} \geq 0 \tag{43}$$

Notice that the input to $K(z)\frac{1}{z-1}$ is transformed in the form: $e_k = u_k + \frac{\lambda_2}{\lambda_1} y_k$. Therefore, we obtain the following state space realization of $G(z)$ in terms of (q_k, w_k):

$$\left[\begin{array}{c|c} A_G & B_G \\ \hline C_G & D_G \end{array}\right] = \left[\begin{array}{cc|c} 0 & 0 & 0 \\ 0 & 1 & 1 \\ \hline 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{cc} \mathbf{I} & 0 \\ 0 & \frac{\lambda_2}{\lambda_1} \\ \hline 0 & 1 \end{array}\right] \underbrace{\begin{bmatrix} A_K & B_K \\ C_K & D_K \end{bmatrix}}_{K} \left[\begin{array}{cc|c} \mathbf{I} & 0 & 0 \\ 0 & 1 & 0 \end{array}\right] \tag{44}$$

Combining it with the map Ψ defined in Lemma 1, we have

$$\left[\begin{array}{c|c} \hat{A} & \hat{B} \\ \hline \hat{C} & \hat{D} \end{array}\right] = \left[\begin{array}{c|c} A_G & B_G \\ \hline \left[\begin{smallmatrix} C_G \\ 0 \end{smallmatrix}\right] & \left[\begin{smallmatrix} D_G \\ 1 \end{smallmatrix}\right] \end{array}\right]$$

By Theorem 1, iterates converge with rate $\rho \in (0, 1]$ if there exists $P \succ 0$ such that

$$\begin{bmatrix} A_G & B_G \\ C_G & D_G \end{bmatrix}^\top \begin{bmatrix} P & 0 \\ 0 & \lambda_1 L^2 - 2\lambda_2 m + \frac{\lambda_2^2}{\lambda_1} \end{bmatrix} \begin{bmatrix} A_G & B_G \\ C_G & D_G \end{bmatrix} - \begin{bmatrix} \rho^2 P & 0 \\ 0 & \lambda_1 \end{bmatrix} \preceq 0. \tag{45}$$

According to Schur complements, we can write the equivalent condition:

$$\begin{bmatrix} \rho^2 P & 0 & A_G^\top & C_G^\top \\ 0 & \lambda_1 & B_G^\top & D_G^\top \\ A_G & B_G & P^{-1} & 0 \\ C_G & D_G & 0 & H^{-1} \end{bmatrix} \succeq 0, \quad \text{and} \quad P \succ 0 \tag{46}$$

where $H \triangleq (\lambda_1 L^2 - 2\lambda_2 m + \frac{\lambda_2^2}{\lambda_1}) \geq 0$. Substituting (44) into (46), we obtain

$$\underbrace{\begin{bmatrix} \rho^2 P & \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right] & \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right] & \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right] \\ \left[\begin{smallmatrix} 0 & 0 \end{smallmatrix}\right] & \lambda_1 & \left[\begin{smallmatrix} 0 & 1 \end{smallmatrix}\right] & 0 \\ \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right] & \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right] & P^{-1} & \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right] \\ \left[\begin{smallmatrix} 0 & 0 \end{smallmatrix}\right] & 0 & \left[\begin{smallmatrix} 0 & 0 \end{smallmatrix}\right] & H^{-1} \end{bmatrix}}_{\Theta} + \text{sym} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \left[\begin{smallmatrix} \mathbf{I} \\ 0 \end{smallmatrix}\right] & \left[\begin{smallmatrix} 0 \\ \frac{\lambda_2}{\lambda_1} \end{smallmatrix}\right] \\ 0 & 1 \end{bmatrix} K \begin{bmatrix} \left[\begin{smallmatrix} \mathbf{I} & 0 \end{smallmatrix}\right] & 0 & 0 & 0 \\ \left[\begin{smallmatrix} 0 & 1 \end{smallmatrix}\right] & 0 & 0 & 0 \end{bmatrix} \succeq 0, \tag{47}$$

where $\text{sym}X \triangleq X + X^\top$. We need to introduce a Lemma to further simplify the problem:

Lemma 3 (Gahinet and Apkarian (1994)) *Given a symmetric matrix $\Theta \in \mathbb{R}^{n \times n}$ and two matrices P, Q of column dimension n, consider the problem of finding some matrix Ξ of compatible dimensions such that*

$$\Theta + P^\top \Xi^\top Q + Q^\top \Xi P \preceq 0 \tag{48}$$

Denote by W_P, W_Q any matrices whose columns form bases for the null spaces of P and Q respectively. Then there exists Ξ satisfying (48) if and only if

$$W_P^\top \Theta W_P \preceq 0 \quad \textit{and} \quad W_Q^\top \Theta W_Q \preceq 0 \tag{49}$$

By this Lemma, we know that (47) is feasible if and only if a pair of conditions hold. In that case, the conditions are:

$$\begin{bmatrix} \rho^2 P & \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right] & \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right] \\ \left[\begin{smallmatrix} 0 & 0 \end{smallmatrix}\right] & \lambda_1 & 1 \\ \left[\begin{smallmatrix} 0 & 1 \end{smallmatrix}\right] & 1 & \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]^\top P^{-1} \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right] + (\frac{\lambda_2}{\lambda_1})^2 H^{-1} \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \lambda_1 & \left[\begin{smallmatrix} 0 & 1 \end{smallmatrix}\right] & 0 \\ \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right] & P^{-1} & \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right] \\ 0 & \left[\begin{smallmatrix} 0 & 0 \end{smallmatrix}\right] & H^{-1} \end{bmatrix} \succeq 0 \tag{50}$$

Using Schur complements again, we have ($r \triangleq \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]^\top P^{-1} \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$)

$$r + (\tfrac{\lambda_2}{\lambda_1})^2 H^{-1} - (\rho^{-2} r + \tfrac{1}{\lambda_1}) \geq 0 \quad \text{and} \quad r \geq \tfrac{1}{\lambda_1} \tag{51}$$

After some manipulations, we have

$$\rho^2 \geq 1 - 2\tfrac{\lambda_1}{\lambda_2} m + (\tfrac{\lambda_1}{\lambda_2})^2 L^2 \tag{52}$$

Optimizing over $\frac{\lambda_1}{\lambda_2}$ yields $\rho^2 \geq 1 - \frac{1}{\kappa^2}$, which is exactly the convergence rate of GD in this setting. ■

Theorem 3 *Under Assumption 1 and 2, PPM converges linearly with any positive η.*

$$\|z_k - z^*\|_2^2 \leq \left(\frac{1}{1+2\eta m}\right)^k \|z_0 - z^*\|_2^2, \quad \textit{for all } k \geq 0. \tag{25}$$

Proof By Theorem 1, we get the following SDP by setting $P = 1$:

$$\begin{bmatrix} 1-\rho^2 & -\eta \\ -\eta & \eta^2 \end{bmatrix} + \lambda_1 \begin{bmatrix} L^2 & -\eta L^2 \\ -\eta L^2 & \eta^2 L^2 - 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} -2m & 2\eta m + 1 \\ 2\eta m + 1 & -2\eta^2 m - 2\eta \end{bmatrix} \preceq 0$$

Using Schur complements, the SDP is equivalent to

$$\begin{aligned} &\eta^2(1 + \lambda_1 L^2 - 2\lambda_2 m) - \lambda_1 - 2\lambda_2 \eta \leq 0, \lambda_1 \geq 0, \lambda_2 \geq 0 \\ &\rho^2 \geq 1 + \lambda_1 L^2 - 2\lambda_2 m + \frac{[(1 + \lambda_1 L^2 - 2\lambda_2 m)\eta - \lambda_2]^2}{\lambda_1 + 2\lambda_2 \eta - \eta^2(1 + \lambda_1 L^2 - 2\lambda_2 m)} \end{aligned}$$

For notational convenience, we let $\Delta \triangleq 1 + \lambda_1 L^2 - 2\lambda_2 m$ and then we have

$$\begin{aligned} &\eta^2 \Delta - \lambda_1 - 2\lambda_2 \eta \leq 0, \lambda_1 \geq 0, \lambda_2 \geq 0 \\ &\rho^2 \geq \Delta + \frac{(\eta\Delta - \lambda_2)^2}{\lambda_1 + 2\lambda_2 \eta - \eta^2 \Delta} = \frac{\lambda_1 \Delta + \lambda_2^2}{\lambda_1 + 2\lambda_2 \eta - \eta^2 \Delta} \end{aligned} \tag{53}$$

We notice that ρ^2 yields the smallest value when $\eta = \lambda_2/\Delta$. Therefore, we have

$$\eta(1 + \lambda_1 L^2 - 2\lambda_2 m) = \lambda_2 \Rightarrow \lambda_2 = \frac{\eta(1 + \lambda_1 L^2)}{1 + 2\eta m} \tag{54}$$

Plugging (54) back into (53), we obtain

$$\rho^2 \geq \frac{\lambda_2}{\eta} = \frac{1 + \lambda_1 L^2}{1 + 2\eta m} \geq \frac{1}{1 + 2\eta m}$$

where the last inequality follows from the fact that $\lambda_1 \geq 0$. We finish the proof. ■

Proposition 1 *OG is an approximation to EG but using the past gradient:*

$$\begin{aligned} z_{k+1/2} &= z_k - \eta F(z_{k-1/2}), \\ z_{k+1} &= z_k - \eta F(z_{k+1/2}). \end{aligned} \tag{26}$$

Proof With some manipulations, we have

$$z_{k+1} = z_k - \eta F(z_k - \eta F(z_{k-1/2}))$$

Notice that $\eta F(z_{k-1/2}) = z_{k-1} - z_k$, we then get

$$z_{k+1} = z_k - \eta F(2z_k - z_{k-1})$$

We therefore conclude that OG is an approximation to EG using the past gradient. ■

Theorem 4 (Gidel et al. (2018)) *Under Assumption 1 and 2, if we take $\eta = 1/(4L)$, then*

$$\|z_k - z^*\|_2^2 \leq \left(1 - \frac{1}{4\kappa}\right)^k \|z_0 - z^*\|_2^2, \quad \textit{for all } k \geq 0. \tag{27}$$

Proof Recall Proposition 1, we have

$$\begin{aligned}\|z_{k+1} - z^*\|_2^2 &= \|z_k - \eta F(z_{k+1/2}) - z^*\|_2^2 \\ &= \|z_k - z^*\|_2^2 - 2\eta F(z_{k+1/2})^\top (z_{k+1} - z^*) - \|z_{k+1} - z_k\|_2^2 \end{aligned} \tag{55}$$

Also, we notice that

$$\begin{aligned}\|z_{k+1} - z_k\|_2^2 &= \|z_{k+1} - z_{k+1/2} - \eta F(z_{k-1/2})\|_2^2 \\ &= \|z_{k+1} - z_{k+1/2}\|_2^2 + \|z_{k+1/2} - z_k\|_2^2 - 2\eta F(z_{k-1/2})^\top (z_{k+1} - z_{k+1/2}) \end{aligned} \tag{56}$$

Plugging (56) back into (55), we have

$$\begin{aligned}\|z_{k+1} - z^*\|_2^2 &= \|z_k - z^*\|_2^2 + \|z_{k+1} - z_{k+1/2}\|_2^2 - \|z_{k+1/2} - z_k\|_2^2 \\ &\quad - 2\eta F(z_{k+1/2})^\top (z_{k+1/2} - z^*) \\ &\leq \|z_k - z^*\|_2^2 - \|z_{k+1/2} - z_k\|_2^2 + \eta^2 L^2 \|z_{k+1/2} - z_{k-1/2}\|_2^2 \\ &\quad - 2\eta F(z_{k+1/2})^\top (z_{k+1/2} - z^*) \end{aligned} \tag{57}$$

where we used the Lipschtiz assumption of vector field F. Also by strongly monotonicity, we have

$$\begin{aligned}-2F(z_{k+1/2})^\top (z_{k+1/2} - z^*) &\leq -2m\|z_{k+1/2} - z^*\|_2^2 \\ &\leq -m\|z_k - z^*\|_2^2 + 2m\|z_{k+1/2} - z_k\|_2^2 \end{aligned} \tag{58}$$

We therefore have

$$\begin{aligned}\|z_{k+1} - z^*\|_2^2 \leq (1 - \eta m)\|z_k - z^*\|_2^2 - (1 - 2\eta m)\|z_{k+1/2} - z_k\|_2^2 \\ + \eta^2 L^2 \|z_{k+1/2} - z_{k-1/2}\|_2^2 \end{aligned} \tag{59}$$

By further noticing that

$$\begin{aligned}2\|z_{k+1/2} - z_{k-1/2}\|_2^2 &\leq 4\|z_{k+1/2} - z_k\|_2^2 + 4\|z_k - z_{k-1/2}\|_2^2 \\ &\leq 4\|z_{k+1/2} - z_k\|_2^2 + 4\eta^2 L^2 \|z_{k-1/2} - z_{k-3/2}\|_2^2 \end{aligned} \tag{60}$$

where we used the Lipschitz assumption again. Finally, combining (59) and (60), we get

$$\|z_{k+1} - z^*\|_2^2 \leq (1-\eta m)\|z_k - z^*\|_2^2 - (1 - 2\eta m - 4\eta^2 L^2)\|z_{k+1/2} - z_k\|_2^2 \\ + 4\eta^4 L^4 \|z_{k-1/2} - z_{k-3/2}\|_2^2 - \eta^2 L^2 \|z_{k+1/2} - z_{k-1/2}\|_2^2 \quad (61)$$

Taking $\eta = 1/(4L)$, we have

$$\|z_{k+1} - z^*\|_2^2 + \frac{1}{16}\|z_{k+1/2} - z_{k-1/2}\|_2^2 \leq \left(1 - \frac{1}{4L}\right)\left(\|z_k - z^*\|_2^2 + \frac{1}{16}\|z_{k-1/2} - z_{k-3/2}\|_2^2\right)$$

This completes the proof. ■

A.3 Proofs for Section 4

Theorem 5 (Hu et al. (2017, Theorem 1)) *Consider the stochastic jump system* (30)*. Suppose F satisfies the pointwise IQC specified by (Ψ, M), and consider the following LMI:*

$$\begin{bmatrix} \hat{A}^\top P \hat{A} - \rho^2 P & \frac{1}{n}\sum_{i=1}^n \hat{A}^\top P \hat{B}_i \\ \frac{1}{n}\sum_{i=1}^n \hat{B}_i^\top P \hat{A} & \frac{1}{n}\sum_{i=1}^n \hat{B}_i^\top P \hat{B}_i \end{bmatrix} + \lambda \begin{bmatrix} \hat{C} & \hat{D} \end{bmatrix}^\top M \begin{bmatrix} \hat{C} & \hat{D} \end{bmatrix} \preceq 0. \quad (32)$$

If this LMI is feasible with $P \succ 0$ and $\lambda \geq 0$, then the following inequality holds,

$$\mathbb{E}[(x_{k+1} - x^*)^\top (P \otimes \mathbf{I}_d)(x_{k+1} - x^*)] \leq \rho^2 (x_k - x^*)^\top (P \otimes \mathbf{I}_d)(x_k - x^*). \quad (33)$$

Consequently, we have $\mathbb{E}\|\xi_k - \xi^*\|_2^2 \leq \mathrm{cond}(P)\rho^{2k}\|\xi_0 - \xi^*\|_2^2$ for any ξ_0 and $k \geq 1$.

Proof Let x, u, s be a set of sequences that satisfies (31). Take the Lynapunov function with the form $V(x_k) = (x_k - x^*)^\top P (x_k - x^*)$, we then have the following relation:

$$\mathbb{E}[V(x_{k+1})] = \sum_{i=1}^n [\hat{A}(x_k - x^*) + \hat{B}_i(u_k - u^*)]^\top P [\hat{A}(x_k - x^*) + \hat{B}_i(u_k - u^*)]$$

Multiply (32) on the left and right by $[(x_k - x^*)^\top, (u_k - u^*)^\top]$ and its transpose, respectively. We then have for all k

$$\mathbb{E}[V(x_{k+1})] \leq \rho^2 V(x_k)$$

Consequently, we have $\mathbb{E}\|x_k - x^*\|_2^2 \leq \mathrm{cond}(P)\rho^{2k}\|x_0 - x^*\|_2^2$. Given that $\|x_k - x^*\|_2^2 = \|\xi_k - \xi^*\|_2^2 + \|\zeta_k - \zeta^*\|_2^2$, we finish the proof. ■

Theorem 6 *For algorithms with one step of memory, the feasible set of the LMI* (32) *is independent of n when $n \geq 2$, hence the optimal solution ρ^2 of the SDP in Theorem 5 under the strong growth condition is independent of n when $n \geq 2$.*

Proof For algorithms with one step of memory, we impose the sector and off-by-one pointwise IQCs. Along with the strong growth condition (28), it yields the following state-space matrices in (31):

$$\hat{A} = \begin{bmatrix} 1+\beta & -\beta & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1+\alpha & -\alpha & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \ \hat{B}_i = \begin{bmatrix} -\eta \mathbf{e}_i^\top \\ \mathbf{0}_{1\times n} \\ \mathbf{0}_{1\times n} \\ \frac{1}{n}\mathbf{1}_{1\times n} \end{bmatrix}, \ \hat{C} = \begin{bmatrix} 1+\alpha & -\alpha & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1+\alpha & -\alpha & -1 & 0 \\ 0 & 0 & 0 & -1 \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times 1} \end{bmatrix}$$
$$\hat{D} = \begin{bmatrix} \mathbf{0}_{1\times n} \\ \frac{1}{n}\mathbf{1}_{1\times n} \\ \mathbf{0}_{1\times n} \\ \frac{1}{n}\mathbf{1}_{1\times n} \\ \mathbf{I}_n \end{bmatrix}, M = \begin{bmatrix} \lambda_1 L^2 - 2\lambda_2 m & \lambda_2 & 0 & 0 & \mathbf{0}_{n\times 1} \\ \lambda_2 & -\lambda_1 & 0 & 0 & \mathbf{0}_{n\times 1} \\ 0 & 0 & \lambda_3 L^2 - 2\lambda_4 m & \lambda_4 & \mathbf{0}_{n\times 1} \\ 0 & 0 & \lambda_4 & -\lambda_3 & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times 1} & \frac{\delta\lambda_5}{n}\mathbf{1}_{n\times n} - \lambda_5 \mathbf{I}_n \end{bmatrix} \tag{62}$$

By Theorem 5, we have the following condition to hold:

$$\begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix} \triangleq \begin{bmatrix} \hat{A}^\top P \hat{A} - \rho^2 P + \hat{C}^\top M \hat{C} & \frac{1}{n}\sum_{i=1}^n \hat{A}^\top P \hat{B}_i + \hat{C}^\top M \hat{D} \\ \frac{1}{n}\sum_{i=1}^n \hat{B}_i^\top P \hat{A} + \hat{D}^\top M \hat{C} & \frac{1}{n}\sum_{i=1}^n \hat{B}_i^\top P \hat{B}_i + \hat{D}^\top M \hat{D} \end{bmatrix} \preceq 0 \tag{63}$$

By Schur complements, we have the following two equivalent conditions:

$$\Theta_{11} \preceq 0, \ \Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12} \preceq 0 \tag{64}$$

Note that the first condition is independent of n, so we only need to check the second one. After some basic manipulations, we have the second condition as follows:

$$\underbrace{(\tfrac{P_{11}\eta^2}{n} - \lambda_5)\mathbf{I}_n + (\tfrac{P_{44}}{n^2} - \tfrac{P_{14}\eta}{n^2} - \tfrac{P_{41}\eta}{n^2} - \tfrac{\lambda_1}{n^2} - \tfrac{\lambda_3}{n^2} + \tfrac{\lambda_5\delta}{n})\mathbf{1}_{n\times n}}_{\Theta_{22}} + \underbrace{\tfrac{K}{n^2}\mathbf{1}_{n\times n}}_{\Theta_{21}\Theta_{11}^{-1}\Theta_{12}} \preceq 0, \tag{65}$$

where K is a scalar that does not depend on n. If $n \geq 2$, then we know that one necessary condition for (65) to hold is $\frac{P_{11}\eta^2}{n} - \lambda_5 \leq 0$. Let $\lambda_5' \triangleq \frac{\lambda_5}{n}$, we have

$$(P_{11}\eta^2 - \lambda_5')\mathbf{I}_n + (P_{44} - P_{14}\eta - P_{41}\eta - \lambda_1 - \lambda_3 + \lambda_5'\delta + K)\frac{\mathbf{1}_{n\times n}}{n} \preceq 0. \tag{66}$$

To further simplify (66), we need to introduce the following Lemma.

Lemma 4 *For $a\mathbf{I}_n + b\frac{\mathbf{1}_{n\times n}}{n} \preceq 0$ to hold ($n \geq 2$), we have $a \leq 0$ and $a + b \leq 0$.*

This Lemma can be proved immediately by showing that the matrix $a\mathbf{I}_n + b\frac{\mathbf{1}_{n\times n}}{n}$ has eigenvalues $\lambda_1 = ... = \lambda_{n-1} = a$ and $\lambda_n = a + b$. One caveat is that this Lemma only hold for $n \geq 2$. Hence, one can show the necessary and sufficient condition of (66) is as follows.

$$P_{11}\eta^2 - \lambda_5' + P_{44} - P_{14}\eta - P_{41}\eta - \lambda_1 - \lambda_3 + \lambda_5'\delta + K \leq 0, \ P_{11}\eta^2 - \lambda_5' \leq 0. \tag{67}$$

It is important to note that two conditions in (67) are all independent of the choice of n. Therefore, we conclude that for any choice of $P, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5'$, if the LMI (32) is feasible for a particular $n \geq 2$, then it is feasible for all $n \geq 2$. In other words, the feasible set of the LMI (32) is invariant to the choice of n, which further implies the optimal ρ of the corresponding SDP is independent of n. This completes the proof. ■

Theorem 7 (GD with the strong growth condition) *Under Assumptions 1 and 2, if we further assume the vector field F satisfies the strong growth condition* (28) *with parameter δ and take $\eta = 1/(L\kappa\delta)$, then we have*

$$\mathbb{E}\|z_k - z^*\|_2^2 \leq \left(1 - \frac{1}{\kappa^2\delta}\right)^k \|z_0 - z^*\|_2^2. \tag{34}$$

Proof For any $k \geq 0$, we have

$$\begin{aligned}\|z_{k+1} - z^*\|_2^2 &= \|z_k - \eta F(z_k;\epsilon_k) - z^*\|_2^2 \\ &= \|z_k - z^*\|_2^2 - 2\eta F(z_k;\epsilon_k)^\top (z_k - z^*) + \eta^2 \|F(z_k;\epsilon_k)\|_2^2\end{aligned}$$

We then take the expectation over ϵ_k and obtain

$$\begin{aligned}\mathbb{E}[\|z_{k+1} - z^*\|_2^2] &= \|z_k - z^*\|_2^2 - 2\eta \mathbb{E}[F(z_k;\epsilon_k)]^\top (z_k - z^*) + \eta^2 \mathbb{E}[\|F(z_k;\epsilon_k)\|_2^2] \\ &= \|z_k - z^*\|_2^2 - 2\eta F(z_k)^\top (z_k - z^*) + \eta^2 \mathbb{E}[\|F(z_k;\epsilon_k)\|_2^2] \\ &\leq \|z_k - z^*\|_2^2 - 2\eta m \|z_k - z^*\|_2^2 + \eta^2 \mathbb{E}[\|F(z_k;\epsilon_k)\|_2^2] \\ &\leq (1 - 2\eta m + \eta^2 \delta L^2)\|z_k - z^*\|_2^2 \\ &= \left(1 - \frac{1}{\kappa^2\delta}\right)\|z_k - z^*\|_2^2\end{aligned}$$

where we used the strongly-monotone assumption in the first inequality, Lipschitz assumption and the strong growth condition in the second inequality. By repeatedly taking expectation over $\epsilon_{k-1}, \epsilon_{k-2}, ...$, we conclude

$$\mathbb{E}\|z_k - z^*\|_2^2 \leq \left(1 - \frac{1}{\kappa^2\delta}\right)^k \|z_0 - z^*\|_2^2$$

Hence, we prove this Theorem. ■

Appendix B. Additional Results

B.1 Evidences for Conjecture 1

B.2 Proof of ASGD under the Strong Growth Condition

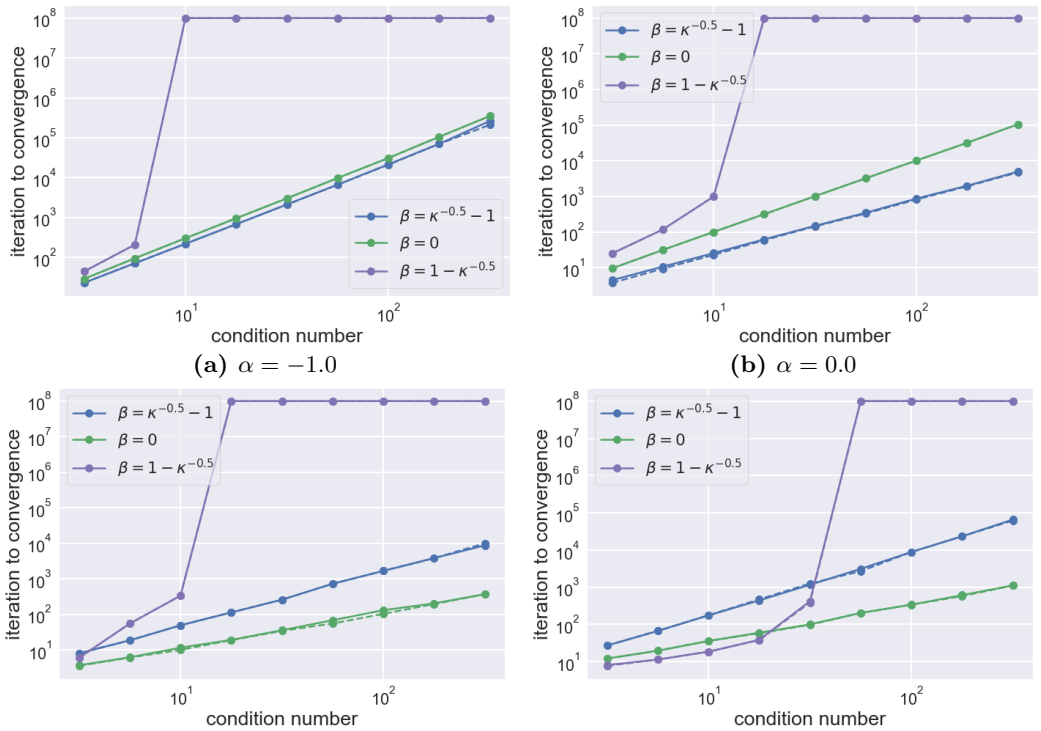
We note that the original ASGD (Jain et al., 2018) was proposed for least squares regression, here we generalize the result to general smooth and strongly-convex functions under the strong growth condition.

Theorem 8 (AGSD) *Under L-smoothness and m-strong-convexity, if f satisfies the strong growth condition with constant δ, then ASGD in the form of* (6) *with the following choice of parameters:*

$$\alpha = \frac{\sqrt{\kappa}\delta - 1}{\sqrt{\kappa} + 1},\ \beta = \frac{\sqrt{\kappa}\delta - 1}{\sqrt{\kappa}\delta + 1},\ \eta = \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa}\delta + 1}\frac{1}{L\delta} \tag{68}$$

results in the following convergence rate:

$$\mathbb{E}[f(z_k)] - f(z^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}\delta}\right)^k \left(f(z_0) - f(z^*) + \frac{1}{2}m\|z^*\|_2^2\right) \tag{69}$$



(a) $\alpha = -1.0$ **(b)** $\alpha = 0.0$

(c) $\alpha = 1.0$ **(d)** $\alpha = 5.0$

Figure 11: Curves of iteration complexity for algorithms with one step of memory (6). We pick some representative algorithms with different α, β. We then do grid search over step size for every algorithm. The solid curves are the ones with only either the sector IQCs (for GD) or a combination of the sector and off-by-one pointwise IQCs while the dashed curves are obtained by adding off-by-two pointwise IQCs. Basically, we observe that off-by-two IQCs are totally redundant. In the special case of $\alpha = \beta = 0$, off-by-one IQCs are also redundant. With some particular choices of parameters, the algorithm fails to converge and we set the upper bound to be 10^8 for visual clarity.

Proof To begin with, we rewrite ASGD in the following form:

$$\begin{aligned} x_k &= z_{k-1} - \eta_1 \nabla_z f(z_{k-1}; \epsilon_{k-1}) \\ y_k &= z_{k-1} - \eta_2 \nabla_z f(z_{k-1}; \epsilon_{k-1}) \\ v_k &= (1-\beta_1) v_{k-1} + \beta_1 x_k \\ z_k &= (1-\beta_2) v_k + \beta_2 y_k \end{aligned} \tag{70}$$

We note that the equations (70) can be compactly written as a second-order difference equation in the form of (6). We choose $x_0 = y_0 = v_0 = 0$ for convenience. In addition, we also stress that all stationary states are the same in the sense of $x^* = y^* = v^* = z^*$. Without loss of generality, we assume the minimal loss $f(z^*) = 0$. Now we choose the Lyapunov function of $V(k) \triangleq f(y_k) + \frac{1}{2} m \|v_k - v^*\|_2^2$, then it suffice to prove

$$\mathbb{E}[V(k+1)] \leq \left(1 - \frac{1}{\sqrt{\kappa}\delta}\right) V(k). \tag{71}$$

First, we notice that

$$\begin{aligned}\mathbb{E}[f(y_{k+1})] &= \mathbb{E}[f(z_k - \eta_2 \nabla_z f(z_k; \epsilon_k))] \\ &\leq f(z_k) - \eta_2 \|\nabla_z f(z_k)\|_2^2 + \frac{L}{2}\eta_2^2 \mathbb{E}[\|\nabla_z f(z_k;\epsilon_k)\|_2^2] \\ &\leq f(z_k) - \eta_2 \|\nabla_z f(z_k)\|_2^2 + \frac{L}{2}\eta_2^2 \delta \|\nabla_z f(z_k)\|_2^2\end{aligned} \tag{72}$$

where the first inequality we used the L-smoothness of f and the second inequality we used the strong growth condition. Further, we consider the other term $\frac{1}{2}m\|v_{k+1} - v^*\|_2^2$.

$$\begin{aligned}\frac{1}{2}m\mathbb{E}[\|v_{k+1} - v^*\|_2^2] &= \frac{1}{2}m\mathbb{E}[\|(1-\beta_1)(v_k - v^*) + \beta_1(z_k - z^*) - \beta_1\eta_1 \nabla_z f(z_k;\epsilon_k)\|_2^2] \\ &\leq \frac{1}{2}m(1-\beta_1)\|v_k - v^*\|_2^2 + \frac{1}{2}m\beta_1\|z_k - z^*\|_2^2 + \frac{1}{2}m\beta_1^2\eta_1^2\mathbb{E}[\|\nabla_z f(z_k;\epsilon_k)\|_2^2] \\ &\qquad - m\beta_1\eta_1((1-\beta_1)(v_k - v^*) + \beta_1(z_k - z^*))^\top \nabla_z f(z_k) \\ &\leq \frac{1}{2}m(1-\beta_1)\|v_k - v^*\|_2^2 + \frac{1}{2}m\beta_1\|z_k - z^*\|_2^2 + \frac{1}{2}m\beta_1^2\eta_1^2\delta\|\nabla_z f(z_k)\|_2^2 \\ &\qquad - m\beta_1\eta_1((1-\beta_1)(v_k - v^*) + \beta_1(z_k - z^*))^\top \nabla_z f(z_k)\end{aligned} \tag{73}$$

where we used Jensen inequality in the first inequality and then the strong growth condition in the second inequality. Next, we observe that

$$\begin{aligned}(1-\beta_1)(v_k - v^*) + \beta_1(z_k - z^*) &= (1-\beta_1)\frac{z_k - \beta_2 y_k}{1-\beta_2} + \beta_1 z_k - z^* \\ &= (1-\beta_1)\frac{\beta_2}{1-\beta_2}(z_k - y_k) + z_k - z^*.\end{aligned} \tag{74}$$

Therefore, we have

$$\begin{aligned}&((1-\beta_1)(v_k - v^*) + \beta_1(z_k - z^*))^\top \nabla_z f(z_k) \\ &\qquad \geq f(z_k) + \frac{1}{2}m\|z_k - z^*\|_2^2 + (1-\beta_1)\frac{\beta_2}{1-\beta_2}(f(z_k) - f(y_k))\end{aligned} \tag{75}$$

where we used the m-strong-convexity of f. Plugging (75) back into (73), we have

$$\begin{aligned}\frac{1}{2}m\mathbb{E}[\|v_{k+1} - v^*\|_2^2] &\leq \frac{1}{2}m(1-\beta_1)\|v_k - v^*\|_2^2 + \frac{1}{2}m(\beta_1 - \beta_1\eta_1 m)\|z_k - z^*\|_2^2 \\ &\quad + \frac{1}{2}m\beta_1^2\eta_1^2\delta\|\nabla_z f(z_k)\|_2^2 - m\beta_1\eta_1 f(z_k) - m\beta_1\eta_1(1-\beta_1)\frac{\beta_2}{1-\beta_2}(f(z_k) - f(y_k))\end{aligned} \tag{76}$$

Recall that our goal is to prove $V(k) \triangleq f(y_k) + \frac{1}{2}m\|v_k - v^*\|_2^2$ is a valid Lyapunov function and also it decreases at every iteration. To achieve that, we have to choose the value of $\eta_1, \eta_2, \beta_1, \beta_2$ carefully. By inspecting (72) and (76), we find the following parameters might work:

$$\eta_1 = 1/m, \quad \beta_1\frac{\beta_2}{1-\beta_2} = 1, \quad \frac{1}{2}m\beta_1^2\eta_1^2\delta - \eta_2 + \frac{L}{2}\eta_2^2\delta \leq 0 \tag{77}$$

Combining (72) and (76), we have

$$\mathbb{E}[V(k+1)] \leq (1-\beta_1)V(k). \tag{78}$$

Therefore, we could choose $\eta_1 = \frac{1}{m}$, $\eta_2 = \frac{1}{L\delta}$, $\beta_1 = \frac{1}{\sqrt{\kappa}\delta}$, $\beta_2 = \frac{\sqrt{\kappa}\delta}{\sqrt{\kappa}\delta+1}$. Comparing (70) to (6), we find that this choice of parameters $\eta_1, \eta_2, \beta_1, \beta_2$ exactly corresponds to (68). Hence, we finish the proof. ■

Remark 3 *By setting δ to be 1 (i.e., deterministic setting), the algorithm of ASGD is exactly Nesterov's accelerated method (NAG) (Nesterov, 1983) and we also recover the convergence rate of NAG.*

B.3 IQC Analysis for Sample-batch Optimistic Gradient Method

Recall the same-batch OG update in the finite-sum setting

$$\begin{aligned} z_{k+1/2} &= z_k - \eta F_{i_k}(z_{k-1/2}) \\ z_{k+1} &= z_k - \eta F_{i_k}(z_{k+1/2}) \end{aligned} \tag{79}$$

To model this algorithm as a discrete dynamical system, we need to take $\xi_k = [z_k^\top, z_{k-1/2}^\top]^\top$. We then have the state matrices as follows (in the case of $n = 2$)

$$\left[\begin{array}{c|c} A & B_{i_k} \\ \hline C & D \end{array}\right] = \left[\begin{array}{cc|cc} 1 & 0 & \begin{bmatrix}0&0\end{bmatrix} & -\eta \mathbf{e}_{i_k}^\top \\ 1 & 0 & -\eta \mathbf{e}_{i_k}^\top & \begin{bmatrix}0&0\end{bmatrix} \\ \hline 0 & 1 & \begin{bmatrix}0&0\end{bmatrix} & \begin{bmatrix}0&0\end{bmatrix} \\ 0 & 1 & \begin{bmatrix}0&0\end{bmatrix} & \begin{bmatrix}0&0\end{bmatrix} \\ 1 & 0 & \begin{bmatrix}-\eta,0\end{bmatrix} & \begin{bmatrix}0&0\end{bmatrix} \\ 1 & 0 & \begin{bmatrix}0,-\eta\end{bmatrix} & \begin{bmatrix}0&0\end{bmatrix} \end{array}\right] \otimes \mathbf{I}_d$$

where $\mathbf{e}_{i_k}$ is a one-hot vector with i_k-entry being 1. We also have the map Ψ in the following form for sector IQCs:

$$\Psi_1 = \Psi_2 = \left[\begin{array}{c|cccc|cccc} \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \hline \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d \end{array}\right] \tag{80}$$

Similarly, for off-by-one pointwise IQCs, we have

$$\Psi_3 = \Psi_4 = \left[\begin{array}{c|cccc|cccc} \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \hline \mathbf{0}_d & -\mathbf{I}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & -\mathbf{I}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & -\mathbf{I}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & -\mathbf{I}_d & \mathbf{0}_d & \mathbf{I}_d \end{array}\right] \tag{81}$$

In addition, we have the following matrices describing the assumptions:

$$M_1 = M_3 = \begin{bmatrix} L^2\mathbf{I}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \mathbf{0}_d & L^2\mathbf{I}_d & \mathbf{0}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & -\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix} \text{ and } M_2 = M_4 = \begin{bmatrix} -2m\mathbf{I}_d & \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -2m\mathbf{I}_d & \mathbf{0}_d & \mathbf{I}_d \\ \mathbf{I}_d & \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d & \mathbf{0}_d \end{bmatrix}$$

Finally, by (31) and Theorem 5, we can reduce the problem to a small SDP.

References

Jacob Abernethy, Kevin A Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization. *arXiv preprint arXiv:1906.02027*, 2019.

Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 486–495, 2019.

Kwangjun Ahn. From proximal point method to nesterov's acceleration. *arXiv preprint arXiv:2005.08304*, 2020.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873, 2020a.

Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *The 23rd International Conference on Artificial Intelligence and Statistics*, pages 1705–1715, 2020b.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781, 2013.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

George HG Chen and R Tyrrell Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

Oswaldo Luiz Valle Costa, Marcelo Dutra Fragoso, and Ricardo Paulino Marques. *Discrete-time Markov jump linear systems*. Springer Science & Business Media, 2006.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134, 2018.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations*, 2018.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Dmitriy Drusvyatskiy. The proximal point method revisited. *arXiv preprint arXiv:1712.06038*, 2017.

Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058, 2017.

Alireza Fallah, Asuman Ozdaglar, and Sarath Pattathil. An optimal multistage stochastic gradient method for minimax problems. *arXiv preprint arXiv:2002.05683*, 2020.

Farzan Farnia and Asuman Ozdaglar. Do GANs always have Nash equilibria? In *Proceedings of the 37th International Conference on Machine Learning*, pages 3029–3039, 2020.

Tanner Fiez, Benjamin Chasnov, and Lillian J Ratliff. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

Pascal Gahinet and Pierre Apkarian. A linear matrix inequality approach to H∞ control. *International journal of robust and nonlinear control*, 4(4):421–448, 1994.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811, 2019.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in Neural Information Processing Systems*, 33, 2020a.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 1758–1784, 2020b.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Daniel R Grayson and Michael E Stillman. Macaulay2, a software system for research in algebraic geometry, 2002.

Patrick T Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1-3):161–220, 1990.

Emilie V Haynsworth. On the schur complement. Technical report, BASEL UNIV (SWITZERLAND) MATHEMATICS INST, 1968.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

Reyhane Askari Hemmat, Amartya Mitra, Guillaume Lajoie, and Ioannis Mitliagkas. Lead: Least-action dynamics for min-max optimization. *arXiv preprint arXiv:2010.13846*, 2020.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, pages 6938–6948, 2019.

Bin Hu and Laurent Lessard. Control interpretations for first-order optimization methods. In *2017 American Control Conference (ACC)*, pages 3114–3119. IEEE, 2017a.

Bin Hu and Laurent Lessard. Dissipativity theory for nesterov's accelerated method. In *International Conference on Machine Learning*, pages 1549–1557. PMLR, 2017b.

Bin Hu, Peter Seiler, and Anders Rantzer. A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints. *Proceedings of Machine Learning Research vol*, 65:1–33, 2017.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604, 2018.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR, 2020.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Laurent Lessard and Peter Seiler. Direct synthesis of iterative algorithms with bounds on achievable worst-case convergence rate. In *2020 American Control Conference (ACC)*, pages 119–125. IEEE, 2020.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1): 57–95, 2016.

Alistair Letcher, David Balduzzi, Sébastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. Differentiable game mechanics. *Journal of Machine Learning Research*, 20:1–40, 2019.

Chaoyue Liu and Mikhail Belkin. Mass: an accelerated stochastic method for over-parametrized learning. *arXiv preprint arXiv:1810.13395*, 2018.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Oren Mangoubi and Nisheeth K Vishnoi. A second-order equilibrium in nonconvex-nonconcave min-max optimization: Existence and algorithm. *arXiv preprint arXiv:2006.12363*, 2020.

Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

Alexandre Megretski and Anders Rantzer. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830, 1997.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*, 2018.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *Advances in Neural Information Processing Systems*, pages 1825–1835, 2017.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582, 2020.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507, 2020a.

Aryan Mokhtari, Asuman E Ozdaglar, and Sarath Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate o(k^2). In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Philipp Rostalski and Bernd Sturmfels. Dualities in convex algebraic geometry. *arXiv preprint arXiv:1006.4894*, 2010.

Ernest K Ryu and Stephen Boyd. Primer on monotone operator methods. *Appl. Comput. Math*, 15(1):3–43, 2016.

Ernest K Ryu, Adrien B Taylor, Carolina Bergeling, and Pontus Giselsson. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *SIAM Journal on Optimization*, 30(3):2251–2271, 2020.

Florian Schäfer and Anima Anandkumar. Competitive gradient descent. In *33rd Conference on Neural Information Processing Systems*, pages Art–No. Neural Information Processing Systems Foundation, Inc., 2019.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262, 2009.

Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204, 2019.

J Von Neumann and O Morgenstern. Theory of games and economic behavior. 1944.

Yuanhao Wang, Guodong Zhang, and Jimmy Ba. On solving minimax optimization locally: A follow-the-ridge approach. In *International Conference on Learning Representations*, 2019.

Guodong Zhang and Yuanhao Wang. On the suboptimality of negative momentum for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2098–2106. PMLR, 2021.