Competitive Gradient Optimization

Abhijeet Vyas [1] **Brian Bullins** [1] **Kamyar Azizzadenesheli** [2]

Abstract

We study the problem of convergence to a stationary point in zero-sum games. We propose competitive gradient optimization (*CGO*), a gradient-based method that incorporates the interactions between two players in zero-sum games for its iterative updates. We provide a continuous-time analysis of *CGO* and its convergence properties while showing that in the continuous limit, previous methods degenerate to their gradient descent ascent (*GDA*) variants. We further provide a rate of convergence to stationary points in the discrete-time setting. We propose a generalized class of α-coherent functions and show that for strictly α-coherent functions, *CGO* ensures convergence to a saddle point. Moreover, we propose optimistic *CGO* (o*CGO*), an optimistic variant, for which we show a convergence rate of $O(\frac{1}{n})$ to saddle points for α-coherent functions.

1. Introduction

We study the zero-sum simultaneous two-player optimization problem of the following form,

$$\min_{x\in\mathcal{X}} f(x,y), \quad \max_{y\in\mathcal{Y}} f(x,y) \tag{1}$$

where x and y are players' moves with $\mathcal{X}\subseteq\mathbb{R}^m, \mathcal{Y}\subseteq\mathbb{R}^n$ and f is a scalar value map from $\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$. Such an optimization problem, also known as minimax optimization problems, has numerous applications in machine learning and decision theory, some examples including competitive Markov decision processes (Filar and Vrieze, 1996), e.g., game of StarCraft, Go, soccer, and car racing (Vinyals et al., 2019; Silver et al., 2016; Prajapat et al., 2021), adversarial and robust learning (Sinha et al., 2017; Namkoong and Duchi, 2016; Madry et al., 2017), generative adversarial networks (GAN) (Goodfellow et al., 2014; Radford et al., 2015; Arjovsky et al., 2017), and risk assessment (Artzner et al., 1999).

Gradient descent ascent (*GDA*) is the standard first-order approach to the minimax optimization problem in Eq. (1) and is known to converge for *strictly*-coherent functions (Mertikopoulos et al., 2019) which subsumes the *strictly* convex-concave function class (Facchinei and Pang, 2003). Yet, *GDA* cycles or **diverges on simple functions** with interactive terms between the players, e.g., a function like $f(x,y) = y^\top x$ (Mertikopoulos et al., 2019) which are not *strictly*-coherent. To tackle this issue, Schäfer and Anandkumar (2019) propose competitive gradient descent (*CGD*) which includes the bilinear approximation of the function as opposed to only the linear approximation used in *GDA* to formulate the local update. In this approach, despite being bilinear, the game approximation per player is linear. With this update, *CGD* is able to utilize the interaction terms to guarantee convergence in some non-convex non-concave problems rather than be impeded by them. Schäfer and Anandkumar (2019) further asks the question whether a local optimality result analogous to (Lee et al., 2016) for the minimization problem can be obtained for minimax optimization. Letcher (2020) answers this to the negative by constructing an example with a local-Nash equilibrium to which no 'reasonable' algorithm can converge. Yet, the constructed function is one with **discontinuous** $\nabla_{xy} f$ the continuity of which is a key assumption in *CGD*. Diakonikolas et al. (2021) proposes a generalized version of the extragradient method that solves the problem in Eq. (1) for the $weak$-MVI condition which extends the MVI condition in (Mertikopoulos et al., 2019). This is achieved by decoupling the learning rates in the two steps of the extragradient method.

Daskalakis et al. (2018) extend the online learning algorithm optimistic mirror descent ascent (*OMDA*) (Rakhlin and Sridharan, 2013) to two player games and shows convergence of the method for all bilinear games of the form $f(x,y) = y^\top Ax$ (thereby for $y^\top x$). Mertikopoulos et al. (2019) use the extragradient version of *OMDA* to show convergence for all coherent saddle points which includes the saddle points in bilinear-games of the form $f(x,y) = y^\top Ax$. However, we show, *CGD* and *OMDA* (as defined in (Mertikopoulos et al., 2019)) reduce to *GDA* and mirror descent ascent (*MDA*) respectively in the continuous-time limit (gradient-flow). The continuous-time regime has

[1]Department of Computer Science, Purdue University, West Lafayette, IN 47907 [2]Nvidia, Santa Clara, CA 95050. Correspondence to: Abhijeet Vyas <vyas26@purdue.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

given insights into the behavior of single-player optimization algorithms (Wilson et al.; Lee et al., 2016) and has been used to study games in (Mazumdar et al., 2020).

In light of this, we propose competitive gradient optimization (*CGO*), an optimization method that incorporates players' interaction in order to come up with gradient updates. *CGO* considers a local linear approximation of the game and introduces the interaction terms in the linear model. The algorithm solves the problem in Eq. (1) for a class of α-coherent functions. This class is strictly larger than the coherent class in Mertikopoulos et al. (2019) and contains examples that are not in the *weak*-MVI class in Diakonikolas et al. (2021). At an iteration point (x, y), the *CGO* update is as follows,

$$\begin{aligned} &\arg\min_{\delta x \in \mathcal{X}} \delta x^\top \nabla_x f + \frac{\alpha}{\eta} \delta x^\top \nabla^2_{xy} f \delta y + \delta y^\top \nabla_y f \\ &\qquad\qquad + \frac{1}{2\eta} \delta x^\top \delta x \\ &\arg\max_{\delta y \in \mathcal{Y}} \delta y^\top \nabla_y f + \frac{\alpha}{\eta} \delta y^\top \nabla^2_{yx} f \delta x + \delta x^\top \nabla_x f \\ &\qquad\qquad - \frac{1}{2\eta} \delta y^\top \delta y. \end{aligned} \tag{2}$$

where the first term in the update is a local linear approximation of the game. The second term is the interaction term between players which is scaled with α to represent the importance of incorporating the interaction in the update. This scaling is analogous to scaling in Newton methods with varying learning rates. This approximation results in a local bilinear approximation of the game. Finally, η is the learning rate appearing in the penalty term. It is important noting that fixing the other player, the optimization for each player is a linear approximation of the game. Since the game approximation for each player is linear in its action, we consider this update yet a linear update. The solution to the *CGO* update is the following,

$$\begin{bmatrix} \delta x \\ \delta y \end{bmatrix} = -\eta g_\alpha := -\eta \begin{bmatrix} I & \alpha \nabla_{xy} f \\ -\alpha \nabla_{yx} f & I \end{bmatrix}^{-1} \begin{bmatrix} \nabla_x f \\ -\nabla_y f \end{bmatrix} \tag{3}$$

where g_α is the gradient-based update at point (x, y). *CGO* is a generalization of its predecessors, in the sense that, setting $\alpha = 0$ recovers *GDA*, and setting $\alpha = \eta$ recovers *CGD*. *CGO* gives greater flexibility for the updates in the hyper-parameters and gives rise to a distinct algorithm in continuous-time. In large-scale practical and deep learning settings, this update can be efficiently and directly computed using an optimized implementation of conjugate gradient and Hessian vector products.

Further, we introduce generalized versions of the Stampacchia and Minty variational inequality (Facchinei and Pang, 2003) and extend the definition of coherent saddle points (Mertikopoulos et al., 2019) to α-coherent saddle points and show the convergence of *CGO* under α-coherence (which contains the bilinear function class and thus explains the success of CGD). Finally, we propose optimistic *CGO* which converges to the saddle points for α-coherent saddle point problems which are not *strictly* α-coherent.

Our main contributions are as follows:

- We propose *CGO* that utilizes bilinear approximation of the game in Eq. (1) and accordingly weights the interaction terms between agents in the updates.
- In order to study whether *CGO* provides a fundamentally new component, we study *CGO*'s and its predecessors' behaviors in continuous-time. We observe that in the limit of the learning rate approaching zero, i.e., continuous-time regime, the *CGD* and *OMDA reduce* to their *GDA* and *MDA* counterparts and *CGO* gives rise to a **distinct update in the continuous-time.**
- Using the Lyapunov analysis in the continuous time regime, we show that while *CGD* and *GDA* converge for *strictly convex-concave functions*, *CGO* allows for a deviation below zero in eigenvalues of the pure Hessian block of the minimizer and above zero for that of the maximizer. A deviation from the convex-concave condition **proportional to the lowest eigenvalue** of the cross-terms of the Hessian is allowed.
- We extend the definition of coherence function class (Mertikopoulos et al., 2019) to α-coherent functions for which we show the optimistic variant of *CGO*, optimistic *CGO* (*oCGO*) converges to saddle points **with a rate of** $O\left(\frac{1}{n}\right)$ while *CGO* converges to the saddle points which satisfy the *strict α-coherence* condition. We provide functions that are not coherent and explain the success of *CGD* in bilinear functions by setting $\alpha = \eta$.

2. Related works

Largely the algorithms proposed to solve the minimax optimization problem can be divided into two groups: those containing simultaneous update which solve a simultaneous game locally at each iteration and those containing sequential updates. While our work focuses on the simultaneous updates, sequential updates are relevant due to their close proximity and the fact that they often give rise to relevant solutions. We discuss the work done in the aforementioned categories below.

Sequential updates. A sequential version *GDA* in an alternating form is alternating gradient descent ascent (*AGDA*) is often time shown to be more stable than its simultaneous counter-part (Gidel et al., 2019; Bailey et al., 2020). Yang et al. (2020) introduces the 2-sided Polyak-Lojasiewicz (PL)-inequality. The PL-inequality was first introduced by Polyak (1963) as a sufficient condition for gradient descent to achieve a linear convergence rate, Yang et al. (2020) shows that the same can be extended to achieve convergence of *AGDA* to saddle points, which are the only stationary points for the said functions. Yet, *AGDA* also cycles in several problems including bi-linear functions showing the persisting difficulty of cycling behavior for any *GDA* algorithm. To solve this problem, 2-time scale gradient descent ascent has been proposed (Heusel et al., 2017; Goodfellow et al., 2014; Metz et al., 2016; Prasad et al., 2015) which uses different learning rates for the descent and ascent. Heusel et al. (2017) proves its convergence to local Nash-equilibrium (saddle points). Jin et al. (2020) discusses the limit points of 2-time scale *GDA* by defining local minimax points, analogues of the local Nash-equilibrium in the sequential game setting and shows that for vanishing learning rate for the descent, 2-time scale *GDA* provably converges to local mini-max points. Another line of work concerns itself with finding stationary points of the function $F(x) = \max_y f(x, y)$, (Lin et al., 2020; Rafique et al., 2018; Nouiehed et al., 2019; Jin et al., 2019). The 2-time scale approaches mainly rely on the convergence of one player per update step of the other player, which makes these updates generally slow to converge.

Simultaneous updates. Simultaneous update methods preserve the simultaneous nature of the game at each step, such methods include *OMDA* (Daskalakis et al., 2018), its extra-gradient version (Mertikopoulos et al., 2019), ConOpt (Mescheder et al., 2017), *CGD* (Schäfer and Anandkumar, 2019), LOLA (Foerster et al., 2017), predictive update (Yadav et al., 2017) and symplectic gradient adjustment (Balduzzi et al., 2018). Of the above, (Daskalakis et al., 2018; Mertikopoulos et al., 2019; Foerster et al., 2017) are inspired from no-regret strategies formulated in (Rakhlin and Sridharan, 2013; Jadbabaie et al., 2015) based on follow the leader (Shalev-Shwartz and Singer, 2006; Grnarova et al., 2017) for online learning. (Schäfer and Anandkumar, 2019) uses the cross-term of the Hessian, while (Mescheder et al., 2017) uses the pure terms to come up with a second order update. (Balduzzi et al., 2018) proposes an update based on the asymmetric part of the game Hessian obtained from its Helmholtz decomposition. Some of these algorithms converge to stationary points that need not correspond to saddle points, Daskalakis and Panageas (2018) shows that *GDA*, as well as optimistic *GDA*, may converge to stationary points which are not saddle points. ConOpt is shown to converge to stationary points which are not local Nash equilibrium in the experiments (Schäfer and Anandkumar, 2019).

3. Preliminaries

In this section, we describe the simultaneous minimax optimization problem and notations to express the properties of functions we use in the analysis. We discuss the class of α-coherent functions which extends the definition of coherence in (Mertikopoulos et al., 2019) and for different versions of which *CGO* and *oCGO* converge to the saddle point.

Throughout the paper we often denote the concatenation of the arguments x and y to be $z := (x, y)$.

Definition 3.1 (First order stationary point)**.** *A point $z^* = (x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is a stationary point of the optimization Eq. (1) if it satisfies the following,*

$$\nabla_x f(x^*, y^*) = \boldsymbol{0}, \nabla_y f(x^*, y^*) = \boldsymbol{0} \tag{4}$$

We say a function f is L Lipschitz continuous if for any two points $z_1 := (x_1, y_1) \in \mathcal{X} \times \mathcal{Y}$ and $z_2 =: (x_2, y_2) \in \mathcal{X} \times \mathcal{Y}$, it satisfies

$$|f(z_1) - f(z_2)| \leq L\|z_1 - z_2\|_2$$

where $|\cdot|$ denote the absolute value and $\|\cdot\|_2$ denote the corresponding 2-norm in the product space $\mathcal{X} \times \mathcal{Y}$. Similarly, for a given function f, we say it has L'-Lipschitz continuous gradient if for any two points $z_1 := (x_1, y_1) \in \mathcal{X} \times \mathcal{Y}$ and $z_2 =: (x_2, y_2) \in \mathcal{X} \times \mathcal{Y}$, it satisfies

$$\|\nabla f(z_1) - \nabla f(z_2)\|_2 \leq L'\|z_1 - z_2\|_2$$

And finally, we say a function has (L_{xx}, L_{yy}, L_{xy})-Lipschitz continuous Hessian, if similarly, for any two points $z_1 := (x_1, y_1) \in \mathcal{X} \times \mathcal{Y}$ and $z_2 := (x_2, y_2) \in \mathcal{X} \times \mathcal{Y}$ the followings hold,

$$\|\nabla^2_{xx} f(z_1) - \nabla^2_{xx} f(z_2)\|_2 \leq L_{xx}\|z_1 - z_2\|_2$$
$$\|\nabla^2_{yy} f(z_1) - \nabla^2_{yy} f(z_2)\|_2 \leq L_{yy}\|z_1 - z_2\|_2$$
$$\|\nabla^2_{xy} f(z_1) - \nabla^2_{xy} f(z_2)\|_2 \leq L_{xy}\|z_1 - z_2\|_2$$

where all the norms are 2-norms with respect to their corresponding suitable definition of native spaces. We present the notation for the minimum and maximum value of matrices derived from the Hessian of f. The extrema are evaluated over the complete domain of f.

Table 1. Eigenvalue notations for matrices derived from 2[nd] derivates to simplify notation

Matrix	Min Eigenvalue	Max Eigenvalue
$\nabla^2_{xx} f$	$\underline{\lambda_{xx}}$	$\overline{\lambda_{xx}}$
$\nabla^2_{yy} f$	$\underline{\lambda_{yy}}$	$\overline{\lambda_{yy}}$
$\nabla^2_{xy} f \nabla^2_{yx} f$	$\underline{\lambda_{xy}}$	$\overline{\lambda_{xy}}$
$\nabla^2_{yx} f \nabla^2_{xy} f$	$\underline{\lambda_{yx}}$	$\overline{\lambda_{yx}}$

Further we define $\overline{\lambda_1} = \max(\overline{\lambda_{xx}}, -\underline{\lambda_{yy}}), \overline{\lambda_2} = \max(\overline{\lambda_{xx}}, \overline{\lambda_{yy}})$. We also have $\underline{\lambda_{xy}}, \underline{\lambda_{yx}} \geq 0$ since $\nabla^2_{xy} f \nabla^2_{yx} f, \nabla^2_{yx} f \nabla^2_{xy} f$ are positive semi-definite.

Definition 3.2 (Bregman Divergence). *The Bregman divergence with a strongly convex and differentiable potential function h is defined as*

$$\mathcal{B}_h(x, y) = h(x) - h(y) - \langle x - y, \nabla h(y) \rangle$$

Saddle point (SP). We define the solutions of the following problems to be min-max and $\max - \min$ saddle points respectively,

- min-max *saddle point:* $$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) \quad (5)$$
- max-min *saddle point:* $$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f(x, y) \quad (6)$$

We now introduce modified forms of the Stampacchia and Minty variational inequalities and present the definition of α-coherent saddle point problems.

Definition 3.3 (α-Variational inequalities). *α-coherence generalizes the definition of coherent saddle points in (Mertikopoulos et al., 2019) which sets $\alpha = 0$. The definition of α-coherence hinges on the following two variational inequalities (g_α as in Eq. (2)),*

- *α-MVI : $g_\alpha(x, y)^\top (z - z^*) \geq 0$ for all $z : (x, y) \in \mathcal{X} \times \mathcal{Y}$*
- *α-SVI : $g_\alpha(x^*, y^*)^\top (z - z^*) \geq 0$ for all $z : (x, y) \in \mathcal{X} \times \mathcal{Y}$*

Definition 3.4 (α-coherence). *We say that* min-max *SP problem is α-coherent if,*

- *Every solution of α-SVI is also a* min-max *SP*
- *There exists a* min-max *SP, p that satisfies α-MVI*
- *Every* min-max *SP, (x^*, y^*) satisfies α-MVI locally, i.e., for all (x, y) sufficiently close to (x^*, y^*)*

The α-coherent max-min *SP problem is defined similarly.*

In the above, if α-MVI holds as a strict inequality whenever x is not a solution thereof, SP problem will be called *strictly* α-coherent; by contrast, if α-MVI holds as an equality for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we will say that the SP problem is null α-coherent. Note that in the unconstrained setting α-SVI is satisfied iff $g_\alpha = \mathbf{0}$ which occurs iff $g_0 = \mathbf{0}$

4. Motivation

In this section, we present the main motivations of our approach. The first is the popularity of the damped Newton method (Algorithm 9.5, (Boyd and Vandenberghe, 2004)) which scales the second-order term in the Taylor-series expansion of the function to come up with the local update. The second is the observation that both *CGD* and *OMDA* reduce to *GDA* and *MDA* in the continuous-time limit which calls for a new algorithm that is distinct from *GDA* and *MDA* in continuous-time. The third is the observation that several functions give rise to (SP) problems which are *strictly* α-coherent, $\quad \forall \alpha > 0$ but not *strictly*-coherent as defined in (Mertikopoulos et al., 2019) which coincides with α-coherence when we set $\alpha = 0$

4.1. Adjustable learning rate Newton method

The celebrated Newton method in one player optimization gives rise to an update which is the solution to the following local optimization problem. For a function $f : \mathcal{X} \to \mathbb{R}, x \in \mathcal{X}, \mathcal{X} \subseteq \mathbb{R}^m$, we have,

$$\min_{\delta x \in \mathcal{X}} \nabla f^\top \delta x + \frac{1}{2} \delta x^\top \nabla^2_{xx} f \delta x \quad (7)$$

which does not have the notion of learning rate. However, prior work provides strong learning and regret guarantees, even in adversarial cases, for the adjusted Newton method where the Newton term is replaced with its weighted version $\frac{\alpha}{2} \delta x^\top \nabla^2_{xx} f \delta x$ (Hazan et al., 2007). This scaling allows for different learning updates that adjust how much the update weighs the second term. This is a similar approach taken in *CGO* update.

4.2. Continuous-time version of *CGD* and *OMDA*

In this section, we analyze the continuous-time versions of *CGD*, *OMDA*, and *CGO*. We show that while, in continuous-time, *CGD* and *OMDA* reduce to their *GDA* and *MDA* counterparts, *CGO* gives rise to a distinct update.

CGD. Following the discussion in the introduction, *CGD* can be obtained by setting $\alpha = \eta$ in the *CGO* update rule. Doing so in Eq. (2), we obtain

$$\delta x = -\eta \left(I + \eta^2 \nabla^2_{xy} f \nabla^2_{yx} f \right)^{-1} \left(\nabla_x f + \eta \nabla^2_{xy} f \nabla_y f \right) \quad (8)$$

$$\delta y = -\eta \left(I + \eta^2 \nabla^2_{yx} f \nabla^2_{xy} f \right)^{-1} \left(-\nabla_y f + \eta \nabla^2_{yx} f \nabla_x f \right). \quad (9)$$

For the continuous-time analysis, the learning rate η corresponds to the time discretization Δt with scaling factor β, i.e., $\eta = \beta \Delta t$. The ratios of the changes in $x := \delta x$ and $y := \delta y$ to η then become the time derivative of x and y in the limit $\eta \to 0$. Ergo, for *CGD* update we obtain,

$$\dot{x} = -\beta \nabla_x f \quad (10)$$

$$\dot{y} = \beta \nabla_y f \quad (11)$$

Where $\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t}, \dot{y} = \frac{\mathrm{d}y}{\mathrm{d}t}$ are the time derivatives of x and y. This is the *same* as the update rule for *GDA* and the interaction information is lost in continuous-time.

OMDA. To present the updates for *OMDA* we first define the proximal map,

$$P_z(p) = \arg\min_{z' \in \mathcal{X} \times \mathcal{Y}} \{\langle p, z - z' \rangle + \mathcal{B}_h(z', z)\} \tag{12}$$

Where $\mathcal{B}_h$ is the Bregman Divergence with the potential function h. The *OMDA* update rule is then given by,

$$\begin{aligned} z_{n+\frac{1}{2}} &= P_{z_n}(-\eta g_n) = \nabla h^{-1}(\nabla h(z_n) - \eta g_n) \\ &= z_n - \eta \nabla(\nabla h^{-1})^\top g_n + o(\eta g_n) \qquad (13) \\ z_{n+1} &= P_{z_n}(-\eta g_{n+\frac{1}{2}}) = \nabla h^{-1}(\nabla h(z_n) - \eta g_{n+\frac{1}{2}}) \\ &= z_n - \eta \nabla(\nabla h^{-1})^\top g_{n+\frac{1}{2}} + o(\eta g_{n+\frac{1}{2}}) \qquad (14) \end{aligned}$$

where $g_n, g_{n+\frac{1}{2}}$ are the vector $(\nabla_x f(x,y), -\nabla_y f(x,y))$ evaluated at $z_n, z_{n+\frac{1}{2}}$ respectively. We now analyze the updates of *OMDA* in continuous-time. In the limit $\eta \to 0$ we have $\frac{\partial z}{\partial t} = -\beta(\nabla(\nabla h)^{-1}(z))^\top \nabla f(z)$ which is the same as the update rule of *MDA* and the effect of half time stepping vanishes in continuous-time. It is interesting to note that Diakonikolas et al. (2021) use different learning rates in the two steps of (13) and avoid this issue, which allows them to show convergence of the augmented *OMDA* for a larger class of *weak*-MVI functions.

CGO. Taking the continuous-time limit $\eta \to 0$ of the *CGO* updates Eq. (2) we obtain,

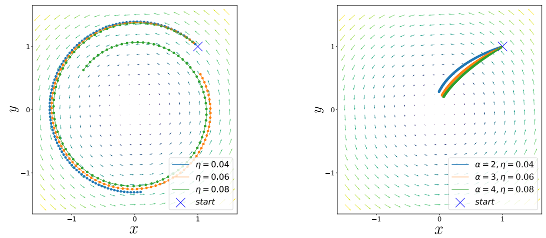
$$\begin{aligned} \dot{x} &= -\beta \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \left(\nabla_x f + \alpha \nabla^2_{xy} f \nabla_y f\right) \\ \dot{y} &= -\beta \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \left(-\nabla_y f + \alpha \nabla^2_{yx} f \nabla_x f\right), \end{aligned} \tag{15}$$

which is a distinct update from *GDA* and the interaction information is preserved in continuous-time. We simulate the continuous-time setting by using a very small learning rate and observe that while *CGD* cycles around the origin (Figure (7a)), *CGO* is able to take a somewhat direct path to the saddle point solution (Figure (1b)). This is an encouraging experiment, validating our hypothesis on the importance of *CGO* update.

4.3. Families of functions which give rise to α-coherent SP

The following examples establish a few families of α-coherent functions. First we present the important result that all bi-linear games $f = x^\top Ay$, are *strictly* α-coherent.

Example 4.1. *All functions of the form* $f(x,y) = x^\top Ay, A \in \mathbb{R}^{m \times n}$*, give rise to strictly α-coherent* min-max *SP problems* $\forall \alpha > 0$ *and are null coherent for* $\alpha = 0$.



(a) *CGD* (b) *CGO*

Figure 1. Modeling of the continuous-time regime for $f(x,y) = xy$: *CGD* cycles while *CGO* takes a direct path. The exact analysis of the resulting ODE's is provided in Appendix E.1

Proof Sketch. The origin is the only saddle point of the above function, we evaluate SVI and α-SVI at the origin,

i) We have $\langle g_0, z \rangle$, $g_0 = (Ay, -A^\top x)$. Hence, $\langle g_0, z \rangle = x^\top Ay - y^\top A^\top x = 0, \quad \forall\, (x,y) \in \mathcal{X} \times \mathcal{Y}$

ii) Also we have

$$\begin{aligned} \langle g_\alpha, z \rangle \geq\ & \alpha \lambda_{\min}((I + \alpha^2 AA^\top)^{-1} AA^\top)\|x\|^2 \\ & + \alpha \lambda_{\min}((I + \alpha^2 A^\top A)^{-1} A^\top Ay)\|y\|^2 > 0 \end{aligned} \tag{16}$$

Where the final inequality follows from the fact that $\min(\lambda_{\min}(A^\top A), \lambda_{\min}(AA^\top)) > 0, \quad \forall A \in \mathbb{R}^{m \times n}$. See Appendix A for a detailed proof. □

We present another family of functions parameterized by a scalar k. For $k \geq 0$ the functions exhibit a min-max saddle point at the origin (a max-min saddle point is at $(\infty, -\infty)$), while for $k < 0$ the function has a max-min saddle point at the origin (a min-max saddle point is at $(-\infty, \infty)$). For both cases, origin satisfies the α-variational inequalities for $\alpha \geq -k$, strictly for $\alpha > k$.

Example 4.2. *The family of functions* $f_k(x,y) = \frac{k}{2}(x^2 - y^2) + xy$ *with* $k \geq 0$ *gives rise to an α-coherent* min-max *SP problem for* $\alpha = -k$ *and a strictly α-coherent* min-max *SP problem* $\forall \alpha > -k$. *For* $k < 0$*, it gives rise to an α-coherent* max-min *SP problem for* $\alpha = -k$ *and a strictly α-coherent* max-min *SP problem* $\forall \alpha > -k$

Proof Sketch. We evaluate the variational inequalities at the origin. For g_0 we have

$$\langle g_0, z \rangle = x(kx + y) - y(-ky + x) = kx^2 + ky^2 > 0 \tag{17}$$

For g_α we have:

$$\langle g_\alpha, z \rangle = \frac{k + \alpha}{1 + \alpha^2}(x^2 + y^2) > 0, \quad \forall \alpha > -k \tag{18}$$

□

Finally we present α-coherent functions that do not satisfy the weak-MVI condition and establish that the class is strictly larger than the coherent class.

Example 4.3. *Consider the family of functions $f_k(x, y) = x^2y + kxy$. For $k = 0$ it gives rise to a* min-max *SP problem such that the region where the α-MVI is not satisfied, shrinks as α increases. For $k = 1$, it gives rise to a* min-max *SP problem which is α-coherent for large α. Furthermore, both the above mentioned problems do not satisfy the weak-MVI condition in Diakonikolas et al. (2021).*

Proof Sketch. Since the Nash Equilibrium is at the origin for both the problems we evaluate the α-VIs at the origin and obtain the conditions for α-coherence as,

- $x^2y + 2\alpha(x^4 + 2x^2y^2) \geq 0$ or $y + 2\alpha(x^2 + 4y^2) \geq 0$ since x can be non-zero.
- $x^2y + \alpha(x^2(x + 1)(2x + 1) + y^2(2x + 1)^2) \geq 0$

The first condition is satisfied for all but sufficiently small (x, y), furthermore it is clear from the expression that the region where it is not satisfied reduces as α increases, this is also shown in Figure (6a). The second condition is satisfied for large $\alpha(\sim 10)$ for the restricted domain $x \geq -\frac{1}{3}$. We numerically verify this through plots in Appendix C

We numerically verify that the *weak*-MVI condition is not satisfied for both the above examples and demonstrate it through heat maps in the Appendix C. ☐

5. Convergence results of *CGO* and the *oCGO* algorithm

In this section we present the convergence results of our *CGO* algorithm. We first consider the convergence to stationary points and present the conditions and rate for the continuous-time and discrete-time regimes. Then, we state the convergence results of the *CGO* algorithm to *strictly* α-coherent saddle points. Then, we introduce the *oCGO* updates and present its rate of convergence to α-coherent saddle points. Finally we showcase the working of *CGO* and *oCGO* by simulating them on a few benchmark functions from the families presented in Subsection 4.3.

5.1. Convergence analysis in continuous-time

We present our first result for convergence of *CGO* in continuous-time. We present the proof sketch and refer the readers to Appendix D for the complete proof. To highlight the difference in convergence rate and condition of *CGO* from *GDA* we also derive the conditions for convergence of *GDA* using a Lyapunov-style analysis. By carefully choosing the parameter α we show that we can accommodate arbitrary deviation from the strictly convex-concave condition which is required for the convergence of continuous-time *GDA*.

Theorem 5.1. *Continuous-time CGO runs on a twice differentiable function f with parameters α, β on functions satisfying $\lambda > 0$ where*

$$\lambda := \beta \min(2\underline{\lambda_{xx}} - 2\alpha\overline{\lambda_{xx}}^2 + c\frac{\lambda_{xy}}{1 + \alpha^2\underline{\lambda_{xy}}}, \quad -2\overline{\lambda_{yy}} - 2\alpha\overline{\lambda_{yy}}^2 + c\frac{\lambda_{yx}}{1 + \alpha^2\underline{\lambda_{yx}}}) \tag{19}$$

converges exponentially to a stationary point with rate λ. Where $c = \beta(\alpha - 2\alpha^2\overline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)$.

Proof Sketch. We choose $\|g_0\|^2$ to be our Lyapnuov function where,

$$g_0 := (\nabla_x(f(x, y), -\nabla_y f(x, y))$$

Evaluating the time derivative of $\|g\|^2$, we obtain

$$\begin{aligned}\frac{\mathrm{d}\|g_0\|^2}{\mathrm{d}t} &= 2g_0^\top \dot{g}_0 \\ &= 2\begin{bmatrix}\nabla_x f^\top & -\nabla_y f^\top\end{bmatrix}\begin{bmatrix}\nabla_{xx} f & \nabla_{xy} f \\ -\nabla_{xy} f^\top & -\nabla_{yy} f\end{bmatrix}\begin{bmatrix}\dot{x} \\ \dot{y}\end{bmatrix} \\ &= 2\dot{x}^\top \nabla_{xx} f \nabla_x f + 2\nabla_x f^\top \nabla_{xy} f \dot{y} \\ &\quad + 2\dot{y}^\top \nabla_{yy} f \nabla_y f + 2\nabla_y f^\top \nabla_{xy} f^\top \dot{x}\end{aligned} \tag{20}$$

By plugging in *CGO* updates and manipulating we show that

$$\frac{\mathrm{d}\|g_0\|^2}{\mathrm{d}t} \leq -\lambda\|g_0\|^2$$

where λ is as stated in the theorem. The detailed proof is in Appendix E.

To compare, we also derive the conditions for *GDA* Eq. (10) in continuous-time in Appendix D. We obtain

$$\begin{aligned}\frac{\mathrm{d}\|g_0\|^2}{\mathrm{d}t} &\leq -\|g_0\|^2 \min(\lambda_{\min}(2\beta\nabla_{xx}), \lambda_{\min}(-2\beta\nabla_{yy})) \\ &= -2\beta\|g_0\|^2 \min(\underline{\lambda_{xx}}, -\overline{\lambda_{yy}})\end{aligned} \tag{21}$$

For convergence, we require $\min(\underline{\lambda_{xx}}, -\overline{\lambda_{yy}}) \geq 0$ which is the convex-concave condition. ☐

This theorem implies that in the presence of interaction, particularly, when $\frac{\lambda_{xy}}{1+\alpha^2\underline{\lambda_{xy}}}$ and $\frac{\lambda_{yx}}{1+\alpha^2\underline{\lambda_{yx}}}$ are positive, it allows to break free from the convex-concave condition by appropriately setting α.

We set α such that $\overline{\lambda_{xx}} \leq \frac{1}{5\alpha}; \underline{\lambda_{xx}} \geq -\frac{1}{5\alpha}; \underline{\lambda_{yx}}, \underline{\lambda_{xy}} \sim \frac{K}{\alpha^2}; \underline{\lambda_{yy}} \geq -\frac{1}{5\alpha}; \overline{\lambda_{yy}} \leq \frac{1}{5\alpha}; K \gg 1$ which implies $\overline{\lambda_1}, \overline{\lambda_2} < \frac{1}{5\alpha}$ and we obtain $\lambda_{min} \geq \frac{1}{50\alpha}$. This shows that continuous-time *CGO* allows **arbitrary deviation** of $\underline{\lambda_{xx}}, \overline{\lambda_{yy}}$ (from the convex-concave condition i.e. $\underline{\lambda_{xx}} \geq 0, \overline{\lambda_{yy}} \leq 0$), if $\underline{\lambda_{yx}}, \underline{\lambda_{xy}}$ are proportional to the square of the deviation of the pure terms.

5.2. Convergence analysis in discrete-time

Convergence to stationary points. We derive the conditions required for *CGO* to converge to a stationary point and show that large singular values of the interaction terms help in convergence. By tuning the hyperparameters we are able to control the influence of this interactive term and obtain faster convergence.

Theorem 5.2. *CGO with parameters α and η when initialized in the neighborhood of a first-order stationary point z^* on a Lipschitz-continuous and thrice differentiable function f that has Lipschitz-continuous gradients and Hessian and $1 \geq \lambda > 0$ where*

$$\lambda := \min(\eta(2\underline{\lambda_{xx}} - 2\frac{10\eta + 8\alpha}{\eta}\overline{\lambda_{xx}}^2) + c\frac{\underline{\lambda_{xy}}}{1+\alpha^2\underline{\lambda_{xy}}}, \tag{22}$$

$$-\eta(2\underline{\lambda_{yy}} + 2\frac{10\eta + 8\alpha}{\eta}\overline{\lambda_{yy}}^2) + c\frac{\underline{\lambda_{yx}}}{1+\alpha^2\underline{\lambda_{yx}}}))$$

converges exponentially to z^* with rate $r(\lambda) = 1-\lambda$. Where c is a polynomial function of $\eta, \alpha, \overline{\lambda_1}, \overline{\lambda_2}$.

Similar to the continuous-time setting, the terms $\frac{\underline{\lambda_{xy}}}{1+\alpha^2\underline{\lambda_{xy}}}$, $\frac{\underline{\lambda_{yx}}}{1+\alpha^2\underline{\lambda_{yx}}}$ are non-negative and appropriately choosing α and η allows us to tune c and obtain convergence for functions not satisfying the convex-concave condition. *CGD* restricts the flexibility of c by choosing $\alpha = \eta$ and *CGO* utilizes this extra degree of freedom granted by α to allow convergence for a larger class of functions. The proof of the above theorem is provided in Appendix G, for completeness we also provide the analysis of discrete time *GDA* in the Appendix F.

Convergence to *strictly* α-coherent saddle points. Now we discuss the convergence properties of *CGO* for the class of *strictly* α-coherent functions. The detailed proof can be found in Appendix H.

Theorem 5.3. *Suppose that a Lipschitz-continuous function f has Lipschitz-continuous gradients and Hessian and gives rise to a strictly α-coherent SP. If CGO is run with perfect gradient and competitive Hessian oracles and parameter α and parameter sequence $\{\eta_n\}$ such that $\sum_1^\infty \eta_n^2 < \infty$ and $\sum_1^\infty \eta_n = \infty$, then the sequence of CGD iterates $\{z_n\}$, converges to a solution of SP.*

Convergence to α-coherent saddle points. For convergence to the saddle points for α-coherent functions which are not *strictly* α-coherent, we propose the optimistic *CGO* algorithm.

Optimistic *CGO* The update rule is given by:

$$z_{n+\frac{1}{2}} = P_{z_n}(-\eta g_{\alpha,n}) \overset{(a)}{=} z_n - \eta g_{\alpha,n} \tag{23}$$

$$z_{n+1} = P_{z_n}(-\eta g_{\alpha,n+\frac{1}{2}}) \overset{(b)}{=} z_n - \eta g_{\alpha,n+\frac{1}{2}} \tag{24}$$

where $g_{\alpha,n}$, is as in (2) and η is the learning rate. Where a and b hold for the unconstrained setting .

Theorem 5.4. *Suppose that a L-Lipschitz-continuous function f that has L'-Lipschitz-continuous gradients and L_{xy} Lipschitz-continuous Hessian gives rise to an α-coherent SP. If oCGO is run with parameter α and parameter sequence $\{\eta_n\}$ such that,*

- $0 < \alpha^2 < \frac{\sqrt{L'^4 + 4L_{xy}^2 L^2} - L'^2}{2L_{xy}^2 L^2}$
- $0 < \eta_n < \frac{\sqrt{\alpha^2 L^2 L_{xy}^2 + L'^2 - 2\alpha^4 L^2 L'^2 L_{xy}^2 - \alpha^2 L'^4} - \alpha^3 L_0^2 L_{xy}^2}{\alpha^2 L^2 L_{xy}^2 + L'^2}, \forall n$

then the sequence of iterates z_n converges to z^* where $z^* := (x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is a saddle point. Moreover, the oCGO converges with the rate of $\frac{1}{n}$, i.e., for the average of the gradients, we have,

$$\frac{1}{n}\sum_{k=1}^{n} \|g_{\alpha,k}\|^2 = O\left(\frac{1}{n}\right)$$

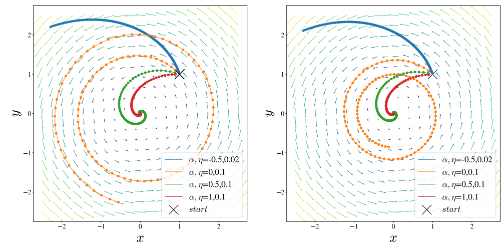
The detailed proof of the above theorem is provided in Appendix H.2.

5.3. Simulation of *CGO* and *oCGO* on families from Section 4

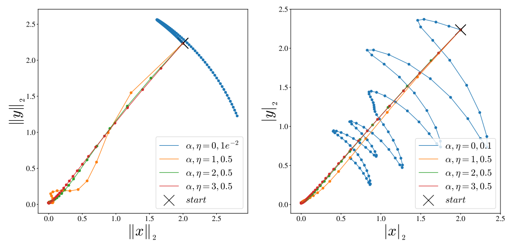
We now evaluate the performance[1] of *CGO* and *oCGO* on families discussed in examples (4.1) and (4.2). We first consider a function $f(x, y) = x^\top Ay, A \in \mathbb{R}^{4\times5}, x \in \mathbb{R}^4, y \in \mathbb{R}^5$. We sample all the entries of A independently from a standard Gaussian, $A = (a_{ij}), a_{ij} \sim \mathcal{N}(0, 1)$. We consider the plot of the L^2 norm of x vs. that of y, since the only saddle point is the origin, the desired solution is $\|x\|_2, \|y\|_2 \to 0$. We plot the iterates of *CGO* and *oCGO* for different α, and observe that *oCGO* converges to the saddle point for $\alpha \geq 0$ (at a very slow rate for $\alpha = 0$) while *CGO* does so for $\alpha > 0$. The results at $\alpha = 0$ are that of *GDA* and optimistic *GDA*. We see similar results for the case where A is the scalar 1, i.e. $f(x, y) = xy$. This is in accordance with the analysis in example (4.1).

We then proceed to perform experiments on the family from example (4.2) for $k = 2, -2$. For both values of k we see that *oCGO* converges to the origin for $\alpha \geq -k$ and *CGO*

[1]The code for the experiments is available at Link to code



(a) *CGO* (b) optimistic *CGO*



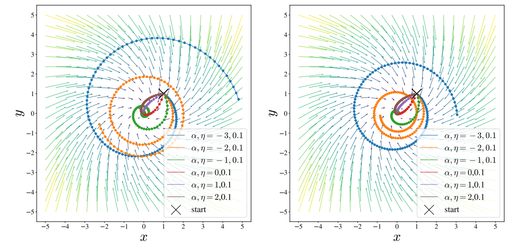
(c) *CGO* (d) optimistic *CGO*

Figure 2. *CGO* and optimistic *CGO* on bilinear functions $f(x, y) = xy, (x, y) \in \mathbb{R}^2$: (a,b) and $f(x, y) = x^\top Ay, x \in \mathbb{R}^4, y \in \mathbb{R}^5$: (c,d) for 100 iterations.

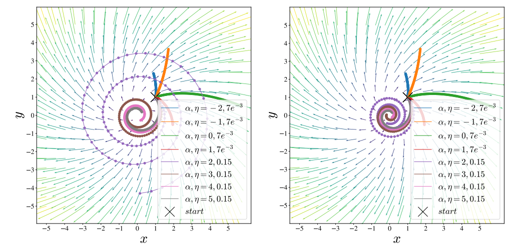
converges for $\alpha > -k$, following the analysis in example (4.2). For $k = 2$ the origin is a min-max saddle point, while for $k = -2$ it is a max-min saddle point. Finally we perform experiments on α-coherent functions from example (4.3) that do not satisfy the *weak*-MVI assumption.

6. Conclusion

We propose the *CGO* algorithm which allows us to control the effect of the cross derivative term in *CGD*. This increases the size of the class of functions for which the algorithm converges. In the realm of continuous-time we observe that *CGD* reduces to *GDA*, *CGO* on the other hand gives rise to a distinct update which allows for a margin of deviation from the strictly convex-concave convergence condition of *GDA*. Furthermore, we generalize the definition of coherent saddle point problems defined in (Mertikopoulos et al., 2019) to α-coherent saddle points for which we prove convergence of Optimistic *CGO* and of *CGO* in the strict version of α-coherence, we show order $O(\frac{1}{n})$ rate of the average gradients for *CGO*. Finally we present a short experiment study on some α-coherent functions. Future work would involve using *CGO* in various machine learning tasks such as GANs, competitive reinforcement learning (RL) and adversarial machine learning.

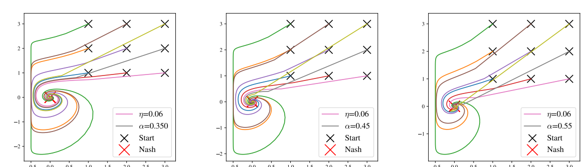


(a) *CGO* (b) *oCGO*

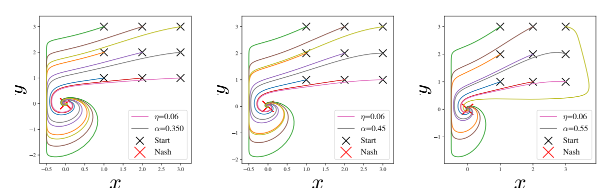


(c) *CGO* (d) *oCGO*

Figure 3. *CGO* and optimistic *CGO* on functions from the family $f(x, y) = \frac{k}{2}(x^2 - y^2) - xy$. $k = 2$: (a,b) and $k = -2$: (c,d) for 100 iterations.

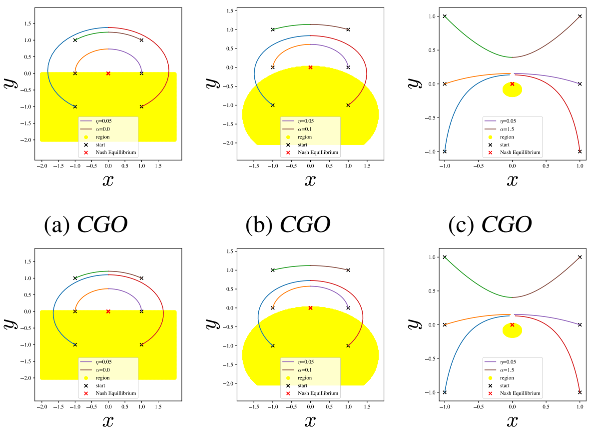


(a) *CGO* (b) *CGO* (c) *CGO*



(d) *oCGO* (e) *oCGO* (f) *oCGO*

Figure 4. *CGO* and optimistic *CGO* on the function $f(x, y) = x^2y + xy$ from multiple initializations for 500 iterations with increasing α.



(a) *CGO* (b) *CGO* (c) *CGO*

(d) *oCGO* (e) *oCGO* (f) *oCGO*

Figure 5. *CGO* and optimistic *CGO* on the function $f(x, y) = x^2y$ from multiple initializations for 500 iterations with increasing α. The shrinking yellow region is where α-MVI is not satisfied.

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

J. P. Bailey, G. Gidel, and G. Piliouras. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In *Conference on Learning Theory*, pages 391–407. PMLR, 2020.

D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 354–363. PMLR, 2018.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJJySbbAZ.

J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. 01 1996. ISBN 978-1-4612-8481-9. doi: 10.1007/978-1-4612-4054-9.

J. N. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326*, 2017.

G. Gidel, R. A. Hemmat, M. Pezeshki, R. Le Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811. PMLR, 2019.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

P. Grnarova, K. Y. Levy, A. Lucchi, T. Hofmann, and A. Krause. An online learning approach to generative adversarial networks. *arXiv preprint arXiv:1706.03269*, 2017.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

A. Jadbabaie, A. Rakhlin, S. Shahrampour, and K. Sridharan. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pages 398–406. PMLR, 2015.

C. Jin, P. Netrapalli, and M. I. Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

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

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR, 2016.

A. Letcher. On the impossibility of global convergence in multi-loss optimization. *arXiv preprint arXiv:2005.12649*, 2020.

T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. 06 2017.

E. Mazumdar, L. J. Ratliff, and S. S. Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg8jjC9KQ.

L. Mescheder, S. Nowozin, and A. Geiger. The numerics of gans. *Advances in neural information processing systems*, 30, 2017.

L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

M. Prajapat, K. Azizzadenesheli, A. Liniger, Y. Yue, and A. Anandkumar. Competitive policy optimization. In *Uncertainty in Artificial Intelligence*, pages 64–74. PMLR, 2021.

H. Prasad, P. LA, and S. Bhatnagar. Two-timescale algorithms for learning nash equilibria in general-sum stochastic games. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1371–1379, 2015.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

H. Rafique, M. Liu, Q. Lin, and T. Yang. Non-convex min–max optimization: provable algorithms and applications in machine learning (2018). *arXiv preprint arXiv:1810.02060*, 2018.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

F. Schäfer and A. Anandkumar. Competitive gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.

S. Shalev-Shwartz and Y. Singer. Convex repeated games and fenchel duality. *Advances in neural information processing systems*, 19, 2006.

D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016. doi: 10.1038/nature16961.

A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. 10 2017.

O. Vinyals, I. Babuschkin, W. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. Agapiou, M. Jaderberg, and D. Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575, 11 2019. doi: 10.1038/s41586-019-1724-z.

A. Wilson, B. Recht, and M. Jordan. A lyapunov analysis of momentum methods in optimization (2016). *arXiv preprint arXiv:1611.02635*.

A. Yadav, S. Shah, Z. Xu, D. Jacobs, and T. Goldstein. Stabilizing adversarial nets with prediction methods. *arXiv preprint arXiv:1705.07364*, 2017.

J. Yang, N. Kiyavash, and N. He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33:1153–1165, 2020.

In this section we present proofs for statements pertaining to example (4.1) and (4.2).

A. Proof of example (4.1)

For clarity we restate the statement of the example (4.1). All functions of the form $x^\top Ay$ are strictly α-coherent $\forall\alpha > 0$ and are null coherent for $\alpha = 0$.

Proof of example (4.1). In order to show the above mentioned statement, we first note that the origin is the only saddle point of this function. We now evaluate $\langle g_0, z\rangle$, where $g_0 = (Ay, -A^\top x)$. Hence, $\forall\ (x, y) \in \mathcal{X} \times \mathcal{Y}$. we have,

$$\langle g_0, z\rangle = x^\top Ay - y^\top A^\top x = 0.$$

ergo, the function $x^\top Ay$ is null-coherent.

Similarly, we evaluate the α-SVI. We observe for the function $x^\top Ay$,

$$g_\alpha = ((I + \alpha^2 AA^\top)^{-1}(Ay + \alpha AA^\top x), (I + \alpha^2 A^\top A)^{-1}(-A^\top x + \alpha A^\top Ay))$$

Hence for $\langle g_\alpha, z\rangle$ we have,

$$\langle g_\alpha, z\rangle = x^\top (I + \alpha^2 AA^\top)^{-1}(Ay + \alpha AA^\top x) + y^\top (I + \alpha^2 A^\top A)^{-1}(-A^\top x + \alpha A^\top Ay) \tag{25}$$

We further observe that, following the statement of Lemma (E.1), we have

$$(I + \alpha^2 AA^\top)^{-1} A = A(I + \alpha^2 A^\top A)^{-1},$$

and therefore, incorporating it in to the Eq. (25), we have,

$$x^\top (I + \alpha^2 AA^\top)^{-1} Ay = x^\top A(I + \alpha^2 A^\top A)^{-1} y = y^\top (I + \alpha^2 A^\top A)^{-1} A^\top x.$$

Thus, for $\langle g_\alpha, z\rangle$ we have,

$$\begin{aligned}\langle g_\alpha, z\rangle =& x^\top (I + \alpha^2 AA^\top)^{-1}\alpha AA^\top x + y^\top (I + \alpha^2 A^\top A)^{-1}\alpha A^\top Ay \\ \geq& \alpha\lambda_{min}((I + \alpha^2 AA^\top)^{-1} AA^\top)\|\Delta x\|^2 + \alpha\lambda_{min}((I + \alpha^2 A^\top A)^{-1} A^\top Ay)\|\Delta y\|^2\end{aligned}$$

Finally, observing that $\min(\lambda_{min}(A^\top A), \lambda_{min}(AA^\top)) \geq 0$ for any A, and following the statement in the Lemma (E.3) we also have,

$$\alpha\lambda_{min}((I + \alpha^2 AA^\top)^{-1} AA^\top)\|\Delta x\|^2 + \alpha\lambda_{min}((I + \alpha^2 A^\top A)^{-1} A^\top Ay)\|\Delta y\|^2 > 0, \quad \forall\alpha > 0,$$

and hence $\langle g_\alpha, z\rangle > 0, \quad \forall\alpha > 0$. Ergo, the function $x^\top Ay$ is *strictly* α coherent. □

B. Proof of example (4.2)

Now, we restate the statement of the example (4.2). The family ofunctions $f_k(x, y) = \frac{k}{2}(x^2 - y^2) + xy$ for $k \geq 0$ gives rise to

- min-max α-coherent SP problem when $\alpha = -k$,
- min-max *strictly* α-coherent SP problem when $\alpha > -k$.

and for $k < 0$ the family gives rise to,

- $\max - \min$ α-coherent SP problem when $\alpha = -k$,
- $\max - \min$ *strictly* α-coherent SP problem when $\alpha > -k$.

Proof of example (4.2). We first note that the origin is the only saddle point of the above family. Further, the origin is a min-max saddle point when $k \geq 0$ and a $\max - \min$ saddle point when $k < 0$.

For this family we evaluate $\langle g_\alpha, z\rangle$,

$$\begin{aligned}\langle g_\alpha, z\rangle =& x((1 + \alpha^2)^{-1}(kx - y - \alpha(-x - ky))) + y((1 + \alpha^2)^{-1}(x + ky - \alpha(kx - y))) \\ =& (1 + \alpha^2)^{-1}(kx^2 - xy + \alpha x^2 + \alpha kxy + xy + ky^2 - \alpha kxy + \alpha y^2)\end{aligned}$$

□

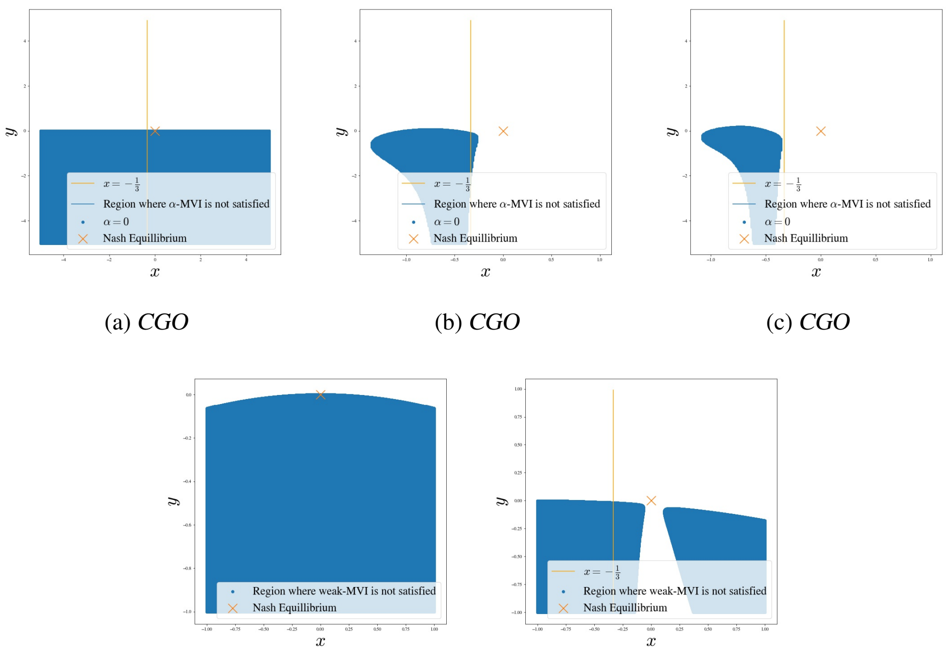
Simplifying this expression for $\alpha > -k$ we obtain,

$$\langle g_\alpha, z \rangle = \frac{k+\alpha}{1+\alpha^2}(x^2+y^2) > 0, \ \forall \alpha > -k$$

Ergo, the above mentioned function class is *strictly* α-coherent when $\alpha > -k$. Furthermore, when $\alpha = -k$ we have $\langle g_\alpha, z \rangle = 0$, ergo the class is null α-coherent for $\alpha = -k$.

C. Proof of example (4.3)

Beyond the explanation in the main text we provide numerically generated heat-maps for the *weak*-MVI condition for the counter-examples provided in example (4.3) and the α-coherence region for $f(x, y) = x^2y + xy, x \geq -\frac{1}{3}$



(a) *CGO* (b) *CGO* (c) *CGO*

(d) *CGO* (e) *CGO*

Figure 6. (a),(b) and (c) : where α-MVI condition is not satisfied for the function $f(x, y) = x^2y + xy$ for increasing α. (d) and (e) : *weak*-MVI condition for $f(x, y) = x^2y$ and $f(x, y) = x^2y + xy; x \geq -\frac{1}{3}$

.

A detailed simulation of the α-MVI condition for $f(x, y) = x^2y + xy; x \geq -\frac{1}{3}$ is available here.

D. Continuous time *GDA*

In this section, we state the update rule for *GDA* and derive sufficient convergence conditions using Lyapunov analysis. The update rule of *GDA* is computed through the following optimization problem,

$$\begin{aligned} &\min_{\delta x \in \mathbb{R}^m} \delta x^\top \nabla_x f + \delta y^\top \nabla_y f + \frac{1}{2\eta} \delta x^\top \delta x \\ &\max_{\delta y \in \mathbb{R}^n} \delta y^\top \nabla_y f + \delta x^\top \nabla_x f - \frac{1}{2\eta} \delta y^\top \delta y. \end{aligned} \tag{26}$$

Which gives the following closed form update,

$$\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} = -\eta \begin{bmatrix}\nabla_x f\\ -\nabla_y f\end{bmatrix} \tag{27}$$

where η is the learning rate. Taking the limit $\eta \to 0$ and scaling the flow of time with β we get the continuous time dynamics as follows,

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = -\beta \begin{bmatrix}\nabla_x f\\ -\nabla_y f\end{bmatrix} = -\beta g_0 \tag{28}$$

where $g_0 = \begin{bmatrix}\nabla_x f\\ -\nabla_y f\end{bmatrix}$ is the concatenation of the gradients. Furthermore, for the second order curvature of this dynamics, i.e., the gradient of g_0, we have,

$$\dot{g}_0 = \begin{bmatrix}\nabla^2_{xx} f & \nabla^2_{xy} f\\ -\nabla^2_{yx} f & -\nabla^2_{yy} f\end{bmatrix} \begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} \tag{29}$$

For the Lyapunov analysis, we now choose $\|g_0\|^2$ as our Lyapunov function and evaluate its time-derivative, i.e.,

$$\begin{aligned}\dot{\|g_0\|^2} = \frac{\mathrm{d}\|g_0\|^2}{\mathrm{d}t} = 2 g_0^\top \dot{g}_0 = 2\left[\nabla_x f^\top \quad -\nabla_y f^\top\right] \begin{bmatrix}\nabla^2_{xx} f & \nabla^2_{xy}\\ -\nabla^2_{yx} f & -\nabla^2_{yy} f\end{bmatrix} \begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix}\\ = 2\dot{x}^\top \nabla^2_{xx} f \nabla_x f + 2\nabla_x f^\top \nabla^2_{xy} f \dot{y} + 2\dot{y}^\top \nabla^2_{yy} f \nabla_y f + 2\nabla_y f^\top \nabla^2_{yx} f \dot{x}\end{aligned}$$

Using the update rule of *GDA*, i.e., Eq. (28), we substitute $\dot{x}$ and $\dot{y}$ in the above equation and have,

$$\begin{aligned}\dot{\|g_0\|^2} &= -2\beta \nabla_x f^\top \nabla^2_{xx} f \nabla_x f + 2\beta \nabla_y f^\top \nabla^2_{yy} f \nabla_y f\\ &\qquad - 2\beta \nabla_x f^\top \nabla^2_{xy} f \nabla_y f + 2\beta \nabla_y f^\top \nabla^2_{yx} f \nabla_x f\\ &= -2\beta \nabla_x f^\top \nabla^2_{xx} f \nabla_x f - (-2\beta \nabla_y f^\top \nabla^2_{yy} f \nabla_y f)\end{aligned} \tag{30}$$

For the right hand side, we know,

$$2\beta \nabla_x f^\top \nabla^2_{xx} f \nabla_x f + (-2\beta \nabla_y f^\top \nabla^2_{yy} f \nabla_y f) \geq \lambda_{min}(2\beta \nabla^2_{xx} f)\|\nabla_x f\|^2 + \lambda_{min}(-2\beta \nabla^2_{yy} f)\|\nabla_y f\|^2$$

Therefore, following the Eq. (30), we have,

$$-\dot{\|g_0\|^2} \geq \lambda_{min}(2\beta \nabla^2_{xx} f)\|\nabla_x f\|^2 + \lambda_{min}(-2\beta \nabla^2_{yy} f)\|\nabla_y f\|^2$$

Resulting in the following Lyapunov key inequality,

$$\dot{\|g_0\|^2} \leq -\|g_0\|^2 \min\{\lambda_{\min}(2\beta \nabla^2_{xx} f), \lambda_{\min}(-2\beta \nabla^2_{yy} f)\}$$

Since, for convex-concave functions, $\min\{\lambda_{\min}(2\beta \nabla^2_{xx} f), \lambda_{\min}(-2\beta \nabla^2_{yy} f)\}$ is always non-negative, which guarantees convergence of this dynamical system.

E. Continuous time *CGO*

In this section, we first derive the continuous-time update rule of *CGO* and then show convergence by choosing the norm squared of the gradient of f as the Lyapunov function. Taking the *CGO* update rule,

$$\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} = -\eta \begin{bmatrix}I & \alpha \nabla^2_{xy} f\\ -\alpha \nabla_{yx} f & I\end{bmatrix}^{-1} \begin{bmatrix}\nabla_x f\\ -\nabla_y f\end{bmatrix}$$

and taking the limit $\eta \to 0$, treating η as time, and scaling time with β, we get,

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = -\beta \begin{bmatrix} I & \alpha \nabla^2_{xy} f \\ -\alpha \nabla^2_{yx} f & I \end{bmatrix}^{-1} \begin{bmatrix} \nabla_x f \\ -\nabla_y f \end{bmatrix} \tag{31}$$

We further simplify Eq. (31) by re-arranging the matrix inverse,

$$\begin{bmatrix} \dot{x} + \alpha \nabla^2_{xy} f \dot{y} \\ -\alpha \nabla^2_{yx} f \dot{x} + \dot{y} \end{bmatrix} = \begin{bmatrix} -\beta \nabla_x f \\ \beta \nabla_y f \end{bmatrix} \tag{32}$$

The above form will be useful in showing convergence. By solving for variable $\dot{x}, \dot{y}$, we get the explicit form,

$$\begin{aligned} \dot{x} &= -\beta \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f \right)^{-1} \left(\nabla_x + \alpha \nabla^2_{xy} f \nabla_y \right) \\ \dot{y} &= -\beta \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f \right)^{-1} \left(\alpha \nabla^2_{yx} f \nabla_x - \nabla_y \right) \end{aligned} \tag{33}$$

We use this construction to prove Theorem (5.1).

Proof of Theorem (5.1). We choose $\|g_0\|^2$ as our Lyapunov function and evaluate its time derivative to observe,

$$\begin{aligned} \dot{\|g_0\|^2} &= \frac{\mathrm{d}\|g_0\|^2}{\mathrm{d}t} \\ &= 2 g_0^\top \dot{g}_0 \\ &= 2 \begin{bmatrix} \nabla_x f^\top & -\nabla_y f^\top \end{bmatrix} \begin{bmatrix} \nabla^2_{xx} f & \nabla^2_{xy} f \\ -\nabla^2_{yx} f & -\nabla^2_{yy} f \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} \\ &= 2 \dot{x}^\top \nabla^2_{xx} f \nabla_x f + 2 \nabla_x f^\top \nabla^2_{xy} f \dot{y} + 2 \dot{y}^\top \nabla^2_{yy} f \nabla_y f + 2 \nabla_y f^\top \nabla^2_{yx} f \dot{x} \end{aligned} \tag{34}$$

Ignoring the factor 2, we expand the terms containing $\nabla^2_{xy} f$ in Eq. (34) by replacing $\dot{x}$ and $\dot{y}$ using Eq. (33) as follows,

$$\begin{aligned} \dot{x}^\top \nabla^2_{xy} f \nabla_y f + \nabla_x f^\top \nabla^2_{xy} f \dot{y} &= -\beta \left(\nabla_x + \alpha \nabla^2_{xy} f \nabla_y \right)^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f \right)^{-1} \nabla^2_{xy} f \nabla_y f \\ &\quad - \nabla_x f^\top \nabla^2_{xy} f \beta \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f \right)^{-1} \left(\alpha \nabla^2_{yx} f \nabla_x - \nabla_y \right) \\ &= -\beta \nabla_x f^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f \right)^{-1} \nabla^2_{xy} f \nabla_y f \\ &\quad - \alpha\beta \nabla_y f^\top \nabla^2_{yx} f \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f \right)^{-1} \nabla^2_{xy} f \nabla_y f \\ &\quad + \beta \nabla_x f^\top \nabla^2_{xy} f \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f \right)^{-1} \nabla_y f \\ &\quad - \alpha\beta \nabla_x f^\top \nabla^2_{xy} f \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f \right)^{-1} \nabla^2_{yx} f \nabla_y f \end{aligned} \tag{35}$$

Using the equality proven in Lemma (E.1) we have,

$$\begin{aligned} \dot{x}^\top \nabla^2_{xy} f \nabla_y f + \nabla_x f^\top \nabla^2_{xy} f \dot{y} &= -\alpha\beta \nabla_x f^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f \right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f \\ &\quad - \alpha\beta \nabla_y f^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f \right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f \end{aligned}$$

Using the expanded terms in RHS of Eq. (35) back into Eq. (34), we obtain a unified expression,

$$\begin{aligned} \dot{\|g_0\|^2} &= 2 \dot{x}^\top \nabla^2_{xx} f \nabla_x f - 2\alpha\beta \nabla_x f^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f \right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f \\ &\quad + 2 \dot{y}^\top \nabla^2_{yy} f \nabla_y f - 2\alpha\beta \nabla_y f^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f \right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f \end{aligned}$$

We now observe that $\alpha \nabla^2_{xy} f \dot{y} + \beta \nabla_x f = -\dot{x}$ and $\alpha \nabla^2_{yx} f \dot{x} + \beta \nabla_y f = \dot{y}$, yielding in,

$$\begin{aligned} \dot{\|g_0\|^2} &= -2\beta \nabla_x f^\top \nabla^2_{xx} f \nabla_x f - 2\alpha\beta \nabla_x f^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f \right)^{-1} \nabla^2_{xy} f \nabla^2_{xy} f \nabla_x f \\ &\quad - 2\alpha \dot{y}^\top \nabla^2_{yx} f \nabla^2_{xx} f \nabla_x \end{aligned} \tag{36}$$

$$\begin{aligned} &+ 2\beta \nabla_y f^\top \nabla^2_{yy} f \nabla_y f - 2\alpha\beta \nabla_y f^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f \right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f \\ &\quad + 2\alpha \dot{x}^\top \nabla^2_{xy} f \nabla^2_{yy} f \nabla_y f \end{aligned} \tag{37}$$

Substituting $\nabla_x f$ and $\nabla_y f$ in lines (36) and (37) with their equivalences in Eq. (32), we get,

$$\begin{aligned}\|\dot{g_0}\|^2 = &-2\beta\nabla_x f^\top\nabla^2_{xx}f\nabla_x f - 2\alpha\beta\nabla_x f^\top\left(I+\alpha^2\nabla^2_{xy}f\nabla^2_{yx}f\right)^{-1}\nabla^2_{xy}f\nabla^2_{xy}f\nabla_x f\\ &-2\alpha\dot{y}^\top\nabla^2_{yx}f\nabla^2_{xx}f(-\frac{\dot{x}+\alpha\nabla^2_{xy}f\dot{y}}{\beta}) \quad (38)\\ &+2\beta\nabla_y f^\top\nabla^2_{yy}f\nabla_y f - 2\alpha\beta\nabla_y f^\top\left(I+\alpha^2\nabla^2_{yx}f\nabla^2_{xy}f\right)^{-1}\nabla^2_{yx}f\nabla^2_{xy}f\nabla_y f\\ &+2\alpha\dot{x}^\top\nabla^2_{xy}f\nabla^2_{yy}f\frac{\dot{y}-\alpha\nabla^2_{yx}f\dot{x}}{\beta} \quad (39)\end{aligned}$$

Taking transpose of the final terms in lines (38) and (39), we obtain,

$$\begin{aligned}\|\dot{g_0}\|^2 = &-2\beta\nabla_x f^\top\nabla^2_{xx}f\nabla_x f - 2\alpha\beta\nabla_x f^\top\left(I+\alpha^2\nabla^2_{xy}f\nabla^2_{yx}f\right)^{-1}\nabla^2_{xy}f\nabla^2_{xy}f\nabla_x f^\top\\ &+\frac{2}{\beta}\left(\alpha\dot{x}+\alpha^2\nabla^2_{xy}f\dot{y}\right)^\top\nabla^2_{xx}f\nabla^2_{xy}f\dot{y}\\ &+2\beta\nabla_y f^\top\nabla^2_{yy}f\nabla_y f - 2\alpha\beta\nabla_y f^\top\left(I+\alpha^2\nabla^2_{yx}f\nabla^2_{xy}f\right)^{-1}\nabla^2_{yx}f\nabla^2_{xy}f\nabla_y f^\top\\ &+\frac{2}{\beta}\left(\alpha\dot{y}-\alpha^2\nabla^2_{yx}f\dot{x}\right)^\top\nabla^2_{yy}f\nabla^2_{yx}f\dot{x} \quad (40)\end{aligned}$$

We utilize the Peter-Paul inequality to further expand $\nabla^2_{xx}f$ and $\nabla^2_{yy}f$ terms in Eq. (40). In particular, we derive the following inequalities,

$$2\dot{x}^\top\nabla^2_{xx}f\nabla^2_{xy}f\dot{y} \le \|\dot{x}^\top\nabla^2_{xx}f\|^2 + \|\nabla^2_{xy}f\dot{y}\|^2$$

and

$$2\dot{x}^\top\nabla^2_{xx}f\nabla^2_{xy}f\dot{y} \le \|\dot{x}^\top\nabla^2_{xx}f\|^2 + \|\nabla^2_{xy}f\dot{y}\|^2.$$

Using these inequalities in Eq. (40), we have,

$$\begin{aligned}\|\dot{g}_0\|^2 \le &-2\beta\nabla_x f^\top\nabla^2_{xx}f\nabla_x f - 2\alpha\beta\nabla_x f^\top\left(I+\alpha^2\nabla^2_{xy}f\nabla^2_{yx}f\right)^{-1}\nabla^2_{xy}f\nabla^2_{xy}f\nabla_x f\\ &+\frac{1}{\beta}\dot{y}^\top\nabla^2_{yx}f\left(\alpha I+2\alpha^2\nabla^2_{xx}f\right)\nabla^2_{xy}f\dot{y}+2\beta\nabla_y f^\top\nabla^2_{yy}f\nabla_y f\\ &-2\alpha\beta\nabla_y f^\top\left(I+\alpha^2\nabla^2_{yx}f\nabla^2_{xy}f\right)^{-1}\nabla^2_{yx}f\nabla^2_{xy}f\nabla_y f+\frac{1}{\beta}\dot{x}^\top\nabla^2_{xy}f\left(\alpha I-2\alpha^2\nabla^2_{yy}f\right)\nabla^2_{yx}f\dot{x}\\ &+\frac{\alpha}{\beta}\dot{y}^\top\nabla^2_{yy}f\nabla^2_{yy}f\dot{y}+\frac{\alpha}{\beta}\dot{x}^\top\nabla^2_{xx}f\nabla^2_{xx}f\dot{x}\end{aligned}$$

Considering that $\nabla^2_{xx}f$ and $\nabla^2_{yy}f$ are symmetric matrices, we have,

$$\begin{aligned}\|\dot{g_0}\|^2 \le &-2\beta\nabla_x f^\top\nabla^2_{xx}f\nabla_x f - 2\alpha\beta\nabla_x f^\top\left(I+\alpha^2\nabla^2_{xy}f\nabla^2_{yx}f\right)^{-1}\nabla^2_{xy}f\nabla^2_{xy}f\nabla_x f\\ &+\frac{\alpha}{\beta}\dot{x}^\top\nabla^2_{xx}f\nabla^2_{xx}f\dot{x}+\frac{1}{\beta}\left(\alpha+2\alpha^2\overline{\lambda_{xx}}\right)\|\nabla^2_{xy}f\dot{y}\|^2\\ &+2\beta\nabla_y f^\top\nabla^2_{yy}f\nabla_y f - 2\alpha\beta\nabla_y f^\top\left(I+\alpha^2\nabla^2_{yx}f\nabla^2_{xy}f\right)^{-1}\nabla^2_{yx}f\nabla^2_{xy}f\nabla_y f\\ &+\frac{\alpha}{\beta}\dot{y}^\top\nabla^2_{yy}f\nabla^2_{yy}f\dot{y}+\frac{1}{\beta}\left(\alpha-2\alpha^2\underline{\lambda_{yy}}\right)\|\nabla^2_{yx}f\dot{x}\|^2 \quad (41)\end{aligned}$$

Setting $\overline{\lambda_1} = \max(\overline{\lambda_{xx}}, -\underline{\lambda_{yy}})$ we obtain,

$$\begin{aligned}\|\dot{g_0}\|^2 \le &-2\beta\nabla_x f^\top\nabla^2_{xx}f\nabla_x f - 2\alpha\beta\nabla_x f^\top\left(I+\alpha^2\nabla^2_{xy}f\nabla^2_{yx}f\right)^{-1}\nabla^2_{xy}f\nabla^2_{xy}f\nabla_x f\\ &+\frac{\alpha}{\beta}\dot{x}^\top\nabla^2_{xx}f\nabla^2_{xx}f\dot{x}+\frac{1}{\beta}\left(\alpha+2\alpha^2\overline{\lambda_1}\right)\|\nabla^2_{xy}f\dot{y}\|^2\\ &+2\beta\nabla_y f^\top\nabla^2_{yy}f\nabla_y f - 2\alpha\beta\nabla_y f^\top\left(I+\alpha^2\nabla^2_{yx}f\nabla^2_{xy}f\right)^{-1}\nabla^2_{yx}f\nabla^2_{xy}f\nabla_y f\\ &+\frac{\alpha}{\beta}\dot{y}^\top\nabla^2_{yy}f\nabla^2_{yy}f\dot{y}+\frac{1}{\beta}\left(\alpha+2\alpha^2\overline{\lambda_1}\right)\|\nabla^2_{yx}f\dot{x}\|^2 \quad (42)\end{aligned}$$

Using the update rule in Eq. (33), we compute,

$$\left\|\nabla_{yx}^2 f\dot{x}\right\|^2 = \beta^2\left(\nabla_x f + \alpha\nabla_{xy}^2 f\nabla_y f\right)^\top\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-2}\nabla_{xy}^2 f\nabla_{yx}^2 f\left(\nabla_x f + \alpha\nabla_{xy}^2 f\nabla_y f\right)$$
$$\left\|\nabla_{xy}^2 f\dot{y}\right\|^2 = \beta^2\left(-\nabla_y f + \alpha\nabla_{yx}^2 f\nabla_x f\right)^\top\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-2}\nabla_{yx}^2 f\nabla_{xy}^2 f\left(-\nabla_y f + \alpha\nabla_{yx}^2 f\nabla_x f\right).$$

by adding up the two equalities above, we obtain,

$$\begin{aligned}\left\|\nabla_{yx}^2 f\dot{x}\right\|^2 + \left\|\nabla_{xy}^2 f\dot{y}\right\|^2 = {} & \beta^2\nabla_x f^\top\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-2}\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_x f \\ & + \beta^2\nabla_y f^\top\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-2}\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_y f \\ & + \alpha\beta^2\nabla_x f^\top\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-2}\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_y f \\ & + \alpha\beta^2\nabla_y f^\top\nabla_{yx}^2 f\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-2}\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_x f \\ & - \underbrace{\alpha\beta^2\nabla_x f^\top\nabla_{xy}^2 f\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-2}\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_y f}_{(i)} \\ & - \underbrace{\alpha\beta^2\nabla_y f^\top\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-2}\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_x f}_{(ii)} \\ & + \underbrace{\alpha^2\beta^2\nabla_x f^\top\nabla_{xy}^2 f\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-2}\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_x f}_{(iii)} \\ & + \underbrace{\alpha^2\beta^2\nabla_y f^\top\nabla_{yx}^2 f\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-2}\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_y f}_{(iv)}\end{aligned} \tag{43}$$

We further analyze the last four terms of the Eq. (43). In particular, we utilize the statement of Lemma (E.1) and for the term (i) in the above equality, we have,

$$\begin{aligned}&\alpha\beta^2\nabla_x f^\top\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-2}\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_y f \\ &\quad = \alpha\beta^2\nabla_x f^\top\nabla_{xy}^2 f\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-2}\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_y f\end{aligned}$$

correspondingly, for the term (ii), we have,

$$\begin{aligned}&\alpha\beta^2\nabla_y f^\top\nabla_{yx}^2 f\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-2}\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_x f \\ &\quad = \alpha\beta^2\nabla_y f^\top\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-2}\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_x f\end{aligned}$$

for the term (iii), we have,

$$\begin{aligned}&\alpha^2\beta^2\nabla_x f^\top\nabla_{xy}^2 f\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-2}\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_x f \\ &\quad = \alpha^2\beta^2\nabla_x f^\top\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-2}\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_x f\end{aligned}$$

correspondingly, for the term (iv), we have,

$$\begin{aligned}&\alpha^2\beta^2\nabla_y f^\top\nabla_{yx}^2 f\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-2}\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_y f \\ &\quad = \alpha^2\beta^2\nabla_y f^\top\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-2}\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_y f\end{aligned}$$

Putting these equalities together in Eq. (43), we have,

$$\begin{aligned}&\left\|\nabla_{yx}^2 f\dot{x}\right\|^2 + \left\|\nabla_{xy}^2 f\dot{y}\right\|^2 \\ &\quad = \beta^2\nabla_x f^\top\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-2}\left(\nabla_{xy}^2 f\nabla_{yx}^2 f + \alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_{yx}^2 f\right)\nabla_x f \\ &\qquad + \beta^2\nabla_y f^\top\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-2}\left(\nabla_{yx}^2 f\nabla_{xy}^2 f + \alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_{xy}^2 f\right)\nabla_y f \\ &\quad = \beta^2\nabla_x f^\top\left(I+\alpha^2\nabla_{xy}^2 f\nabla_{yx}^2 f\right)^{-1}\nabla_{xy}^2 f\nabla_{yx}^2 f\nabla_x f \\ &\qquad + \beta^2\nabla_y f^\top\left(I+\alpha^2\nabla_{yx}^2 f\nabla_{xy}^2 f\right)^{-1}\nabla_{yx}^2 f\nabla_{xy}^2 f\nabla_y f\end{aligned} \tag{44}$$

Plugging this into Eq. (42) we obtain,

$$\begin{aligned}
\dot{\|g_0\|^2} \leq & -2\beta\nabla_x f^\top \nabla^2_{xx} f \nabla_x f + \beta(2\alpha^2\overline{\lambda_1} - \alpha)\nabla_x f^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{xy} f \nabla_x f \\
& + \frac{\alpha}{\beta}\dot{x}^\top \nabla^2_{xx} f \nabla^2_{xx} f \dot{x} + \frac{\alpha}{\beta}\dot{y}^\top \nabla^2_{yy} f \nabla^2_{yy} f \dot{y} \\
& + 2\beta\nabla_y f^\top \nabla^2_{yy} f \nabla_y f \\
& + \beta(2\alpha^2\overline{\lambda_1} - \alpha)\nabla_y f^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f
\end{aligned} \tag{45}$$

We now do the following set of computations,

$$\begin{aligned}
\dot{x}^\top \nabla^2_{xx} f \nabla^2_{xx} f \dot{x} + \dot{y}^\top \nabla^2_{yy} f \nabla^2_{yy} f \dot{y} & \stackrel{(a)}{=} (-\alpha\nabla^2_{xy} f \dot{y} - \beta\nabla_x f)^\top \nabla^2_{xx} f \nabla^2_{xx} f (-\alpha\nabla^2_{xy} f \dot{y} - \beta\nabla_x f) \\
& \quad + (\alpha\nabla^2_{yx} f \dot{x} + \beta\nabla_y f)^\top \nabla^2_{yy} f \nabla^2_{yy} f (\alpha\nabla^2_{yx} f \dot{x} + \beta\nabla_y f) \\
& \stackrel{(b)}{=} \|(\alpha\nabla^2_{xy} f \dot{y} + \beta\nabla_x f)^\top \nabla^2_{xx} f\|^2 \\
& \quad + \|(\alpha\nabla^2_{yx} f \dot{x} + \beta\nabla_y f)^\top \nabla^2_{yy} f\|^2 \\
& \stackrel{(c)}{\leq} 2\alpha^2 \|\nabla^2_{xy} f \dot{y}\|^2 \|\nabla^2_{xx} f\|^2 + 2\alpha^2 \|\nabla^2_{yx} f \dot{x}\|^2 \|\nabla^2_{yy} f\|^2 \\
& \quad + 2\beta^2 \|\nabla_x f \nabla^2_{xx} f\|^2 + 2\beta^2 \|\nabla_y f \nabla^2_{yy} f\|^2 \\
& \stackrel{(d)}{\leq} 2\beta^2\alpha^2 \overline{\lambda_{xx}}^2 \|\nabla^2_{xy} f \dot{y}\|^2 + 2\beta^2\alpha^2 \overline{\lambda_{yy}}^2 \|\nabla^2_{yx} f \dot{x}\|^2 \\
& \quad + 2\beta^2 \overline{\lambda_{xx}}^2 \nabla_x f^\top \nabla_x f + 2\beta^2 \overline{\lambda_{yy}}^2 \nabla_y f^\top \nabla_y f \\
& \stackrel{(e)}{\leq} 2\beta^2\alpha^2 \overline{\lambda_2}^2 (\|\nabla^2_{xy} f \dot{y}\|^2 + \|\nabla^2_{yx} f \dot{x}\|^2) \\
& \quad + 2\beta^2 \overline{\lambda_{xx}}^2 \nabla_x f^\top \nabla_x f + 2\beta^2 \overline{\lambda_{yy}}^2 \nabla_y f^\top \nabla_y f \\
& \stackrel{(f)}{\leq} 2\beta^2\alpha^2 \overline{\lambda_2}^2 \nabla_x f^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f \\
& \quad + 2\beta^2\alpha^2 \overline{\lambda_2}^2 \nabla_y f^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f \\
& \quad + 2\beta^2 \overline{\lambda_{xx}}^2 \nabla_x f^\top \nabla_x f + 2\beta^2 \overline{\lambda_{yy}}^2 \nabla_y f^\top \nabla_y f
\end{aligned}$$

Where for (a) we use Eq. (32) to substitute Δx and Δy, in (b) we re-write the terms as norms, in (c) we use the inequality $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, in (d) we bound the terms using the maximum eigenvalues, in (e) we set $\overline{\lambda_2} = \max(\overline{\lambda_{xx}}, \overline{\lambda_{yy}})$ and finally for (f) we use Eq. (44).

Using the above inequality in Eq. (45), we have,

$$\begin{aligned}
\dot{\|g_0\|^2} \leq & -\nabla_x f^\top (2\beta\nabla^2_{xx} f - 2\beta\alpha\overline{\lambda_{xx}}^2 I)\nabla_x f + \nabla_y f^\top (2\beta\nabla^2_{yy} f + 2\beta\alpha\overline{\lambda_{yy}}^2 I)\nabla_y f \\
& - \beta(\alpha - 2\alpha^2\overline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)\nabla_y f^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f \\
& - \beta(\alpha - 2\alpha^2\overline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)\nabla_x f^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{xy} f \nabla_x f
\end{aligned}$$

By rearranging the above inequality, we get,

$$\begin{aligned}
\dot{\|g_0\|^2} \leq & -\nabla_x f^\top \left((2\beta\nabla^2_{xx} f - 2\beta\alpha\overline{\lambda_{xx}}^2 I) + \beta(\alpha - 2\alpha^2\overline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)\left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f\right) \nabla_x f \\
& - \nabla_y f^\top \left(-(2\beta\nabla^2_{yy} f + 2\beta\alpha\overline{\lambda_{yy}}^2 I) + \beta(\alpha - 2\alpha^2\underline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)\left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f\right) \nabla_y f \\
\leq & -\|g_0\|^2 \min\{\lambda_{\min}((2\beta\nabla^2_{xx} f - 2\beta\alpha\overline{\lambda_{xx}}^2 I) + \beta(\alpha - 2\alpha^2\overline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)\left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f), \\
& \lambda_{\min}(-(2\beta\nabla^2_{yy} f + 2\beta\alpha\overline{\lambda_{yy}}^2 I) + \beta(\alpha - 2\alpha^2\underline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)\left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f)\}
\end{aligned}$$

which is the key Lyapunov inequality. Thus, under the conditions expressed in the statement of the main Theorem, i.e., λ, as defined in the following is positive,

$$\lambda := \min\{\lambda_{\min}((2\beta\nabla^2_{xx}f - 2\beta\alpha\overline{\lambda_{xx}}^2 I) + \beta(\alpha - 2\alpha^2\overline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)\left(I + \alpha^2\nabla^2_{xy}f\nabla^2_{yx}f\right)^{-1}\nabla^2_{xy}f\nabla^2_{yx}f), \tag{46}$$

$$\lambda_{\min}(-(2\beta\nabla^2_{yy}f + 2\beta\alpha\overline{\lambda_{yy}}^2 I) + \beta(\alpha - 2\alpha^2\underline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)\left(I + \alpha^2\nabla^2_{yx}f\nabla^2_{xy}f\right)^{-1}\nabla^2_{yx}f\nabla^2_{xy}f)\} \tag{47}$$

the quantity $\|g_0\|^2$ converges to zero exponentially fast with the rate at least λ. ☐

Now, we simplify the above expression of the rate using Lemmas (E.2) and (E.3) to address the 1^{st} and 2^{nd} terms respectively in lines (46) and (47),

$$\lambda_{min} \geq \beta\min\{2\underline{\lambda_{xx}} - 2\alpha\overline{\lambda_{xx}}^2 + \beta(\alpha - 2\alpha^2\overline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)\frac{\lambda_{xy}}{1 + \alpha^2\underline{\lambda_{xy}}},$$
$$-2\overline{\lambda_{yy}} - 2\alpha\overline{\lambda_{yy}}^2 + \beta(\alpha - 2\alpha^2\overline{\lambda_1} - 2\alpha^3\overline{\lambda_2}^2)\frac{\lambda_{yx}}{1 + \alpha^2\underline{\lambda_{yx}}}\}$$

To better understand the above results, we set some relations between the quantities in the above expression. If we set α such that $\overline{\lambda_{xx}} \leq \frac{1}{5\alpha}; \underline{\lambda_{xx}} \geq -\frac{1}{5\alpha}; \underline{\lambda_{yx}}, \underline{\lambda_{xy}} \sim \frac{K}{\alpha^2}; \underline{\lambda_{yy}} \geq -\frac{1}{5\alpha}; \overline{\lambda_{yy}} \leq \frac{1}{5\alpha}; K \gg 1$. We have $\overline{\lambda_1}, \overline{\lambda_1} \leq \frac{1}{5\alpha}$ and we obtain $\lambda_{min} \geq \frac{1}{50\alpha}$.
This shows that as long as the interaction terms $\underline{\lambda_{yx}}, \underline{\lambda_{xy}}$ are of the order of the square of the deviation of the pure terms $\underline{\lambda_{xx}}, \overline{\lambda_{yy}}$ (from the convex-concave condition i.e. $\underline{\lambda_{xx}} \geq 0, \overline{\lambda_{yy}} \leq 0$), we can guarantee convergence for *CGO*

Statements and proofs of the Lemmas used in the above derivation are provided below,

Lemma E.1. *The following equality holds,*

$$\nabla^2_{yx}f(I + \alpha^2\nabla^2_{xy}f\nabla^2_{yx}f)^{-1} = (I + \alpha^2\nabla^2_{yx}f\nabla^2_{xy}f)^{-1}\nabla^2_{yx}f.$$

Proof. To prove this equality statement, we write,

$$\nabla^2_{yx}f + \alpha^2\nabla^2_{yx}f\nabla^2_{xy}f\nabla^2_{yx}f = (I + \alpha^2\nabla^2_{yx}f\nabla^2_{xy}f)\nabla^2_{yx}f$$

and at the same time,

$$\nabla^2_{yx}f + \alpha^2\nabla^2_{yx}f\nabla^2_{xy}f\nabla^2_{yx}f = \nabla^2_{yx}f(I + \alpha^2\nabla^2_{xy}f\nabla^2_{yx}f)$$

therefore, we have,

$$(I + \alpha^2\nabla^2_{yx}f\nabla^2_{xy}f)\nabla^2_{yx}f = \nabla^2_{yx}f(I + \alpha^2\nabla^2_{xy}f\nabla^2_{yx}f)$$

Multiplying both sides with the inverse of $(I + \alpha^2\nabla^2_{yx}f\nabla^2_{xy}f)$ from the left, and the inverse of $(I + \alpha^2\nabla^2_{xy}f\nabla^2_{yx}f)$ from the right results in,

$$\nabla^2_{yx}f(I + \alpha^2\nabla^2_{xy}f\nabla^2_{yx}f)^{-1} = (I + \alpha^2\nabla^2_{yx}f\nabla^2_{xy}f)^{-1}\nabla^2_{yx}f$$

which is the statement of the Lemma. ☐

Lemma E.2. *The following inequality holds,* $\lambda_{\min}(A + B) \geq \lambda_{\min}(A) + \lambda_{\min}(B), \quad \forall A, B \in \mathcal{S}^n_+$.

Proof. We know,

$$\|\Delta x\|^2\lambda_{min}(A + B) \geq x^\top(A + B)x = x^\top Ax + x^\top Bx, \quad \forall x$$

the following also holds,

$$x^\top Ax + x^\top Bx \geq \|\Delta x\|^2\lambda_{min}(A) + \|\Delta x\|^2\lambda_{min}(B), \quad \forall x$$

Choosing x not equal to zero we complete the proof. ☐

Lemma E.3. *Let $B \in \mathcal{S}_+^n$, if (I+B) is invertible, $\lambda_{min}((I+B)^{-1}B) \geq \frac{\underline{\lambda_b}}{1+\underline{\lambda_b}}$, where $\underline{\lambda_b} = \lambda_{min}(B)$*

Proof. We can write the following,

$$(I+B)^{-1}B = (I+B)^{-1}B = I - (I+B)^{-1}$$

From the statement of Lemma (E.2) we can write,

$$\lambda_{min}((I+B)^{-1}B) \geq \lambda_{min}(I) + \lambda_{min}(-(I+B)^{-1})$$

Hence we have,

$$\lambda_{min}((I+B)^{-1}B) \geq 1 + (-\frac{1}{1+\underline{\lambda_b}}) = \frac{\underline{\lambda_b}}{1+\underline{\lambda_b}}$$

which is the statement of the Lemma. □

E.1. Sample continuous time analysis

For the function $f(x, y) = xy$. The continuous time equations for **GDA/CGD** are,

$$\dot{x} = -\beta y \tag{48}$$
$$\dot{y} = \beta x \tag{49}$$

The solution to the above ODE is,

$$x = c_1 cos(\beta t) - c_2 sin(\beta t) \tag{50}$$
$$y = c_1 sin(\beta t) + c_2 cos(\beta t) \tag{51}$$

For the aforementioned x and y we have,

$$x^2 + y^2 = c_1^2 + c_2^2$$

which is the equation of a circle indicating that the iterates circle around the nash equillibrium.

for **CGO** we have,

$$\dot{x} = -\frac{\beta}{1+\alpha^2}(y + \alpha x) \tag{52}$$
$$\dot{y} = -\frac{\beta}{1+\alpha^2}(-x + \alpha y) \tag{53}$$

The solution is,

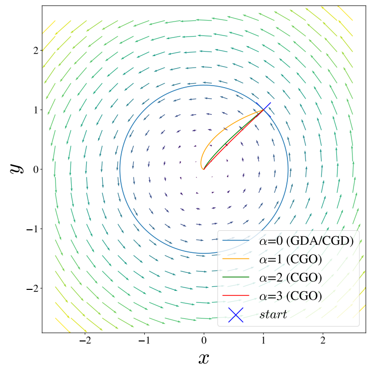
$$x(t) = c_1 \frac{e^{-\alpha\beta t}}{\alpha^2+1} cos(\frac{\beta t}{\alpha^2+1}) - c_2 \frac{e^{-\alpha\beta t}}{\alpha^2+1} sin(\frac{\beta t}{\alpha^2+1}) \tag{54}$$
$$y(t) = c_1 \frac{e^{-\alpha\beta t}}{\alpha^2+1} sin(\frac{\beta t}{\alpha^2+1}) + c_2 \frac{e^{-\alpha\beta t}}{\alpha^2+1} cos(\frac{\beta t}{\alpha^2+1}) \tag{55}$$

Thus x and y satisfy,

$$x^2 + y^2 = (c_1^2 + c_2^2)\frac{e^{-2\alpha\beta t}}{(\alpha^2+1)^2}$$

Indicating that the distance of the iterates from center falls exponentially.

The figure below illustrates the trajectories of the 2 algorithms.



(a) Exact Trajectories

Figure 7. The exact trajectories of *GDA* and *CGO* in continuous time with time scale $\beta = 1, t \in [0, 2\pi]$ and starting point $x_0, y_0 = 1, 1$

F. Discrete time *GDA*

In this section, we present the analysis of the discrete time *GDA* algorithm for completeness. We first present the optimization problem and then derive *GDA* convergence conditions and convergence rate.

To come up with the update rule, we solve the below optimization problem,

$$\begin{aligned} &\min_{\delta x \in \mathbb{R}^m} \delta x^\top \nabla_x f + \delta y^\top \nabla_y f + \frac{1}{2\eta}\delta x^\top \delta x \\ &\max_{\delta y \in \mathbb{R}^n} \delta y^\top \nabla_y f + \delta x^\top \nabla_x f - \frac{1}{2\eta}\delta y^\top \delta y. \end{aligned} \tag{56}$$

Which gives,

$$\begin{bmatrix}\Delta x \\ \Delta y\end{bmatrix} = -\eta \begin{bmatrix}\nabla_x f \\ -\nabla_y f\end{bmatrix} \tag{57}$$

We now write the Taylor expansion of $\nabla_x f, \nabla_y f$ around the (x, y),

$$\begin{aligned} \nabla_x f(\Delta x + x, \Delta y + y) &= \nabla_x f(x,y) + \nabla^2_{xx} f \Delta x + \nabla^2_{xy} f \Delta y + \mathcal{R}_x(\Delta x, \Delta y) \\ \nabla_y f(\Delta x + x, \Delta y + y) &= \nabla_y f(x,y) + \nabla^2_{yy} f \Delta y + \nabla^2_{yx} f \Delta x + \mathcal{R}_y(\Delta x, \Delta y) \end{aligned}$$

where the remainder terms $\mathcal{R}_x$ and $\mathcal{R}_y$ are defined as,

$$\mathcal{R}_x(\Delta x, \Delta y) := \int_0^1 \left(\left(\nabla^2_{xx} f(t\Delta x + x, t\Delta y + y) - \nabla^2_{xx} f\right)\Delta x + \left(\nabla^2_{xy} f(t\Delta x + x, t\Delta y + y) - \nabla^2_{xy} f\right)\Delta y \right) \mathrm{d}t \tag{58}$$

$$\mathcal{R}_y(\Delta x, \Delta y) := \int_0^1 \left(\left(\nabla^2_{yy} f(t\Delta x + x, t\Delta y + y) - \nabla^2_{yy} f\right)\Delta y + \left(\nabla^2_{yx} f(t\Delta x + x, t\Delta y + y) - \nabla^2_{yx} f\right)\Delta x \right) \mathrm{d}t$$

Using this equality, we obtain,

$$\begin{aligned}
&\|\nabla_x f(\Delta x + x, \Delta y + y)\|^2 + \|\nabla_y f(\Delta x + x, \Delta y + y)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&\quad = 2\Delta x^\top \nabla^2_{xx} f \nabla_x f(x,y) + 2\nabla_x f(x,y)^\top \nabla^2_{xy} f \Delta y + \Delta x^\top \nabla^2_{xx} f \nabla^2_{xx} f \Delta x \\
&\qquad + 2\Delta y^\top \nabla^2_{yy} f \nabla_y f(x,y) + 2\nabla_y f(x,y)^\top \nabla^2_{yx} f \Delta x + \Delta y^\top \nabla^2_{yy} f \nabla^2_{yy} f \Delta y + \\
&\qquad + \Delta y^\top \nabla^2_{yx} f \nabla^2_{xy} f \Delta y + \Delta x^\top \nabla^2_{xy} f \nabla^2_{yx} f \Delta x + 2\Delta x^\top \nabla^2_{xx} f \nabla^2_{xy} f \Delta y + 2\Delta y^\top \nabla^2_{yy} f \nabla^2_{yx} f \Delta x \\
&\qquad + 2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y) + 2\Delta x^\top \nabla^2_{xx} f \mathcal{R}_x(\Delta x, \Delta y) + 2\Delta y^\top \nabla^2_{yx} f \mathcal{R}_x(\Delta x, \Delta y) + \|\mathcal{R}_x(\Delta x, \Delta y)\|^2 \\
&\qquad + 2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y) + 2\Delta y^\top \nabla^2_{yy} f \mathcal{R}_y(\Delta x, \Delta y) + 2\Delta x^\top \nabla^2_{xy} f \mathcal{R}_y(\Delta x, \Delta y) + \|\mathcal{R}_y(\Delta x, \Delta y)\|^2
\end{aligned}$$

Substituting $\Delta x = -\eta \nabla_x f(x,y)$ and $\Delta y = \eta \nabla_y f(x,y)$ we obtain,

$$\begin{aligned}
&\|\nabla_x f(\Delta x + x, \Delta y + y)\|^2 + \|\nabla_y f(\Delta x + x, \Delta y + y)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&\quad = -2\eta \nabla_x f(x,y)^\top \nabla^2_{xx} f \nabla_x f(x,y) + 2\eta \nabla_y f(x,y)^\top \nabla^2_{yy} f \nabla_y f(x,y) \\
&\qquad + 2\eta^2 \nabla_y f(x,y)^\top \nabla^2_{yy} f \nabla^2_{yx} f \nabla_x f(x,y) - 2\eta^2 \nabla_x f(x,y)^\top \nabla^2_{xx} f \nabla^2_{xy} f \nabla_y f(x,y) \\
&\qquad + \underbrace{2\eta \nabla_x f(x,y)^\top \nabla^2_{xy} f \nabla_y f(x,y)}_{(i)} - \underbrace{2\eta \nabla_y f(x,y)^\top \nabla^2_{yx} f \nabla_x f(x,y)}_{(ii)} \\
&\qquad + \eta^2 \nabla_y f(x,y)^\top \nabla^2_{yy} f \nabla^2_{yy} f \nabla_y f(x,y) + \eta^2 \nabla_x f(x,y)^\top \nabla^2_{xx} f \nabla^2_{xx} f \nabla_x f(x,y) \\
&\qquad + \eta^2 \nabla_y f(x,y)^\top \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f(x,y) + \eta^2 \nabla_x f(x,y)^\top \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f(x,y) \\
&\qquad + 2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y) + 2\Delta x^\top \nabla^2_{xx} f \mathcal{R}_x(\Delta x, \Delta y) + 2\Delta y^\top \nabla^2_{yx} f \mathcal{R}_x(\Delta x, \Delta y) + \|\mathcal{R}_x(\Delta x, \Delta y)\|^2 \\
&\qquad + 2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y) + 2\Delta y^\top \nabla^2_{yy} f \mathcal{R}_y(\Delta x, \Delta y) + 2\Delta x^\top \nabla^2_{xy} f \mathcal{R}_y(\Delta x, \Delta y) + \|\mathcal{R}_y(\Delta x, \Delta y)\|^2
\end{aligned}$$

The terms (i) and (ii) in the RHS cancel out. Using the Cauchy-Schwarz inequality we obtain,

$$\begin{aligned}
2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y) &\leq 2\|\nabla_x f(x,y)\| \|\mathcal{R}_x(\Delta x, \Delta y)\| \\
2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y) &\leq 2\|\nabla_y f(x,y)\| \|\mathcal{R}_y(\Delta x, \Delta y)\|
\end{aligned} \tag{59}$$

Using the upper bounds on $2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y)$ and $2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y)$ derived in Eq. (59) we obtain,

$$\begin{aligned}
&\|\nabla_x f(\Delta x + x, \Delta y + y)\|^2 + \|\nabla_y f(\Delta x + x, \Delta y + y)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&\quad = -2\eta \nabla_x f(x,y)^\top \nabla^2_{xx} f \nabla_x f(x,y) + 2\eta \nabla_y f(x,y)^\top \nabla^2_{yy} f \nabla_y f(x,y) \\
&\qquad + 2\eta^2 \nabla_y f(x,y)^\top \nabla^2_{yy} f \nabla^2_{yx} f \nabla_x f(x,y) - 2\eta^2 \nabla_x f(x,y)^\top \nabla^2_{xx} f \nabla^2_{xy} f \nabla_y f(x,y) \\
&\qquad + 2\eta^2 \nabla_y f(x,y)^\top \nabla^2_{yy} f \nabla^2_{yy} f \nabla_y f(x,y) + 2\eta^2 \nabla_x f(x,y)^\top \nabla^2_{xx} f \nabla^2_{xx} f \nabla_x f(x,y) \\
&\qquad + \eta^2 \nabla_y f(x,y)^\top \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f(x,y) + \eta^2 \nabla_x f(x,y)^\top \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f(x,y) \\
&\qquad + 2\|\nabla_x f(x,y)\| \|\mathcal{R}_x(\Delta x, \Delta y)\| + 2\|\nabla_y f(x,y)\| \|\mathcal{R}_y(\Delta x, \Delta y)\| \\
&\qquad + 4\|\mathcal{R}_x(\Delta x, \Delta y)\|^2 + 4\|\mathcal{R}_y(\Delta x, \Delta y)\|^2
\end{aligned}$$

Rearranging we obtain,

$$\begin{aligned}
&\|\nabla_x f(\Delta x + x, \Delta y + y)\|^2 + \|\nabla_y f(\Delta x + x, \Delta y + y)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&\quad = \nabla_x f(x,y)^\top \left(\eta^2 \nabla^2_{xx} f^2 - 2\eta \nabla^2_{xx} f + \eta^2 \nabla^2_{xy} f \nabla^2_{yx} f\right) \nabla_x f(x,y) \\
&\qquad + \nabla_y f(x,y)^\top \left(\eta^2 \nabla^2_{yy} f^2 + 2\eta \nabla^2_{yy} f + \eta^2 \nabla^2_{yx} f \nabla^2_{xy} f\right) \nabla_y f(x,y) \\
&\qquad + 2\eta^2 \nabla_x f(x,y)^\top \left(\nabla^2_{xy} f \nabla^2_{yy} f - \nabla^2_{xx} f \nabla^2_{xy} f\right) \nabla_y f(x,y) \\
&\qquad + 4\|\mathcal{R}_x(\Delta x, \Delta y)\|^2 + 4\|\mathcal{R}_y(\Delta x, \Delta y)\|^2 + 2\|\nabla_x f(x,y)\| \|\mathcal{R}_x(\Delta x, \Delta y)\| \\
&\qquad + 2\|\nabla_y f(x,y)\| \|\mathcal{R}_y(\Delta x, \Delta y)\|
\end{aligned}$$

To conclude, we need to bound the $\mathcal{R}$ terms. Using the Lipschitz-continuity of the Hessian and Eq. (58), we can bound the remainder terms as,

$$\|\mathcal{R}_x(\Delta x, \Delta y)\|, \|\mathcal{R}_y(\Delta x, \Delta y)\| \leq L_{xy}(\|\Delta x\| + \|\Delta y\|)^2 \tag{60}$$

Using Eq. (57), we get,

$$\|\Delta x\|^2 + \|\Delta y\|^2 = \eta^2(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)$$

Hence we have,

$$L_{xy}(\|\Delta x\| + \|\Delta y\|)^2 \leq 2L_{xy}(\|\Delta x\|^2 + \|\Delta y\|^2) \leq 2\eta^2 L_{xy}(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)$$

Thus,

$$\begin{aligned}
&\|\nabla_x f(\Delta x + x, \Delta y + y)\|^2 + \|\nabla_y f(\Delta x + x, \Delta y + y)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&\quad \leq \nabla_x f(x,y)^\top \Big(\eta^2 \nabla^2_{xx} f^2 - 2\eta \nabla^2_{xx} f + 2\eta^2 \nabla^2_{xy} f \nabla^2_{yx} f \\
&\qquad\qquad + 4\eta^2 L_{xy}(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|)\Big) \nabla_x f(x,y) \\
&\qquad + \nabla_y f(x,y)^\top \Big(\eta^2 \nabla^2_{yy} f^2 + 2\eta \nabla^2_{yy} f + 2\eta^2 \nabla^2_{yx} f \nabla^2_{xy} f \\
&\qquad\qquad + 4\eta^2 L_{xy}(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|)\Big) \nabla_y f(x,y) \\
&\qquad + \underbrace{2\eta^2 \nabla_x f(x,y)^\top \left(\nabla^2_{xy} f \nabla^2_{yy} f - \nabla^2_{xx} f \nabla^2_{xy} f\right) \nabla_y f(x,y)}_{(i)} \\
&\qquad + 8\eta^2 L_{xy}(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)
\end{aligned}$$

We further use the following inequality,

$$a^\top A b = \frac{1}{2} a^\top A b + \frac{1}{2} b^\top A^\top a \overset{(a)}{\leq} \frac{1}{4}(a^\top (AA^\top + I)a + b^\top (A^\top A + I)b)$$

(where in (a) we use the Peter-Paul inequality on both the terms) to bound the term (i) in the above inequality. We obtain,

$$\begin{aligned}
&\|\nabla_x f\left(\Delta x + x, \Delta y + y\right)\|^2 + \|\nabla_y f\left(\Delta x + x, \Delta y + y\right)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&\quad = \nabla_x f(x,y)^\top \Big((\eta^2 \nabla^2_{xx} f^2 - 2\eta \nabla^2_{xx} f + 2\eta^2 \nabla^2_{xy} f \nabla^2_{yx} f) \\
&\qquad\qquad + I(8\eta^2 L_{xy} + 4\eta^2 L_{xy}(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|))\Big) \nabla_x f(x,y) \\
&\qquad + \nabla_y f(x,y)^\top \Big((\eta^2 \nabla^2_{yy} f^2 + 2\eta \nabla^2_{yy} f + 2\eta^2 \nabla^2_{yx} f \nabla^2_{xy} f) \\
&\qquad\qquad + I(8\eta^2 L_{xy} + 4\eta^2 L_{xy}(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|))\Big) \nabla_y f(x,y) \\
&\qquad + \eta^2 \nabla_x f(x,y)^\top \left((\nabla^2_{xy} f \nabla^2_{yy} f - \nabla^2_{xx} f \nabla^2_{xy} f)(\nabla^2_{xy} f \nabla^2_{yy} f - \nabla^2_{xx} f \nabla^2_{xy} f)^\top \right) \nabla_x f(x,y) \\
&\qquad + \eta^2 \nabla_y f(x,y)^\top \left((\nabla^2_{xy} f \nabla^2_{yy} f - \nabla^2_{xx} f \nabla^2_{xy} f)^\top (\nabla^2_{xy} f \nabla^2_{yy} f - \nabla^2_{xx} f \nabla^2_{xy} f) \right) \nabla_y f(x,y) \\
&\qquad + \eta^2/2(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)
\end{aligned}$$

This gives,

$$\|\nabla_x f\left(\Delta x + x, \Delta y + y\right)\|^2 + \|\nabla_y f\left(\Delta x + x, \Delta y + y\right)\|^2 \leq (1 - \lambda_{min})(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)$$

Where,

$$\begin{aligned}
\lambda_{min} = \eta \ \min \Big\{ &\lambda_{min}(2\nabla^2_{xx} f - \eta(\nabla^2_{xx} f^2 - 2\nabla^2_{xy} f \nabla^2_{yx} f - I(8L_{xy} + 4L_{xy}(\|\nabla_x f\| + \|\nabla_y f\|) + \frac{1}{2}) \\
&\quad - (\nabla^2_{xy} f \nabla^2_{yy} f - \nabla^2_{xx} f \nabla^2_{xy} f)(\nabla^2_{xy} f \nabla^2_{yy} f - \nabla^2_{xx} f \nabla^2_{xy} f)^\top)), \\
&\lambda_{min}(-2\nabla^2_{yy} f - \eta(\nabla^2_{yy} f^2 - 2\nabla^2_{yx} f \nabla^2_{xy} f - I(8L_{xy} + 4L_{xy}(\|\nabla_x f\| + \|\nabla_y f\|) + \frac{1}{2}) \\
&\quad - (\nabla^2_{xy} f \nabla^2_{yy} f - \nabla^2_{xx} f \nabla^2_{xy} f)^\top (\nabla^2_{xy} f \nabla^2_{yy} f - \nabla^2_{xx} f \nabla^2_{xy} f))) \Big\}
\end{aligned}$$

Hence for $1 \geq \lambda_{min} > 0$ we have exponentially fast convergence. For sufficiently small η, we have convergence for all strongly convex-concave functions with rate $1 - \lambda_{min}$ where,

$$\lambda_{min} = \eta(\min\{\lambda_{min}(2\nabla^2_{xx} f), \lambda_{min}(-2\nabla^2_{yy} f)\})$$

G. Discrete time *CGO*

In this section, we restate the update rule for the *CGO* algorithm and then derive its convergence rate and a condition for convergence. Recall the update rule for *CGO*,

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = -\eta \begin{bmatrix} I & \alpha \nabla^2_{xy} f \\ -\alpha \nabla^2_{yx} f & I \end{bmatrix}^{-1} \begin{bmatrix} \nabla_x f \\ -\nabla_y f \end{bmatrix}$$

The following form of the above equation will be useful in the proof,

$$\begin{aligned}
\Delta x &= -\eta \nabla_x f - \alpha \nabla^2_{xy} f \Delta y \\
\Delta y &= \eta \nabla_y f + \alpha \nabla^2_{yx} f \Delta x
\end{aligned} \tag{61}$$

Finally, writing the updates explicitly,

$$\begin{aligned}
\Delta x &= -\eta \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f \right)^{-1} \left(\nabla_x f + \alpha \nabla^2_{xy} f \nabla_y f \right) \\
\Delta y &= \eta \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f \right)^{-1} \left(\nabla_y f - \alpha \nabla^2_{yx} f \nabla_x f \right),
\end{aligned} \tag{62}$$

Proof of Theorem (5.2). Using the Taylor expansion of $(\nabla_x f, \nabla_y f)$, around the point (x, y) we obtain,

$$\nabla_x f(\Delta x + x, \Delta y + y) = \nabla_x f(x, y) + \nabla^2_{xx} f \Delta x + \nabla^2_{xy} f \Delta y + \mathcal{R}_x(\Delta x, \Delta y)$$
$$\nabla_y f(\Delta x + x, \Delta y + y) = \nabla_y f(x, y) + \nabla^2_{yy} f \Delta y + \nabla^2_{yx} f \Delta x + \mathcal{R}_y(\Delta x, \Delta y)$$

where the remainder terms $\mathcal{R}_x$ and $\mathcal{R}_y$ are defined as,

$$\mathcal{R}_x(\Delta x, \Delta y) := \int_0^1 \left(\left(\nabla^2_{xx} f(t\Delta x + x, t\Delta y + y) - \nabla^2_{xx} f\right) \Delta x + \left(\nabla^2_{xy} f(t\Delta x + x, t\Delta y + y) - \nabla^2_{xy} f\right) \Delta y \right) \mathrm{d}t \quad (63)$$

$$\mathcal{R}_y(\Delta x, \Delta y) := \int_0^1 \left(\left(\nabla^2_{yy} f(t\Delta x + x, t\Delta y + y) - \nabla^2_{yy} f\right) \Delta y + \left(\nabla^2_{yx} f(t\Delta x + x, t\Delta y + y) - \nabla^2_{yx} f\right) \Delta x \right) \mathrm{d}t \quad (64)$$

Using these equalities, we obtain the value of the difference between norm of the vector $(\nabla_x f, \nabla_y f)$ at points (x, y) and updated ones, $(\Delta x + x_k, \Delta y + y_k)$.

$$\begin{aligned}
&\|\nabla_x f\left(\Delta x + x, \Delta y + y\right)\|^2 + \|\nabla_y f\left(\Delta x + x, \Delta y + y\right)\|^2 - \|\nabla_x f(x, y)\|^2 - \|\nabla_y f(x, y)\|^2 \\
&\quad = 2\Delta x^\top \nabla^2_{xx} f \nabla_x f(x, y) + 2\nabla_x f(x, y)^\top \nabla^2_{xy} f \Delta y + \Delta x^\top \nabla^2_{xx} f \nabla^2_{xx} f \Delta x + 2\Delta x^\top \nabla^2_{xx} f \nabla^2_{xy} f \Delta y \\
&\qquad + 2\Delta y^\top \nabla^2_{yy} f \nabla_y f(x, y) + 2\nabla_y f(x, y)^\top \nabla^2_{yx} f \Delta x + \Delta y^\top \nabla^2_{yy} f \nabla^2_{yy} f \Delta y + 2\Delta y^\top \nabla^2_{yy} f \nabla^2_{yx} f \Delta x \\
&\qquad + \Delta y^\top \nabla^2_{yx} f \nabla^2_{xy} f \Delta y + \Delta x^\top \nabla^2_{xy} f \nabla^2_{yx} f \Delta x \\
&\qquad + 2\nabla_x f(x, y)^\top \mathcal{R}_x(\Delta x, \Delta y) + 2\Delta x^\top \nabla^2_{xx} f \mathcal{R}_x(\Delta x, \Delta y) + 2\Delta y^\top \nabla^2_{yx} f \mathcal{R}_x(\Delta x, \Delta y) + \|\mathcal{R}_x(\Delta x, \Delta y)\|^2 \\
&\qquad + 2\nabla_y f(x, y)^\top \mathcal{R}_y(\Delta x, \Delta y) + 2\Delta y^\top \nabla^2_{yy} f \mathcal{R}_y(\Delta x, \Delta y) + 2\Delta x^\top \nabla^2_{xy} f \mathcal{R}_y(\Delta x, \Delta y) + \|\mathcal{R}_y(\Delta x, \Delta y)\|^2
\end{aligned} \quad (65)$$

We now observe using Eq. (61) that,

$$\Delta y^\top \nabla^2_{yx} f \nabla^2_{xy} f \Delta y = -\Delta y^\top \nabla^2_{yx} f \frac{\Delta x + \eta \nabla_x f(x, y)}{\alpha} \quad (66)$$

$$\Delta x^\top \nabla^2_{xy} f \nabla^2_{yx} f \Delta x = \Delta x^\top \nabla^2_{xy} f \frac{\Delta y - \eta \nabla_y f(x, y)}{\alpha} \quad (67)$$

Adding up Eq. (66) and Eq. (67) we obtain,

$$\Delta x^\top \nabla^2_{xy} f \nabla^2_{yx} f \Delta x + \Delta y^\top \nabla^2_{yx} f \nabla^2_{xy} f \Delta y = -\frac{\eta}{\alpha}(\Delta y^\top \nabla^2_{yx} f \nabla_x f(x, y) + \Delta x^\top \nabla^2_{xy} f \nabla_y f(x, y))$$

Substituting this into Eq. (65) yields,

$$\begin{aligned}
&\|\nabla_x f\left(\Delta x + x, \Delta y + y\right)\|^2 + \|\nabla_y f\left(\Delta x + x, \Delta y + y\right)\|^2 - \|\nabla_x f(x, y)\|^2 - \|\nabla_y f(x, y)\|^2 \\
&\quad = 2\Delta x^\top \nabla^2_{xx} f \nabla_x f(x, y) + (2 - \frac{\eta}{\alpha})\nabla_x f(x, y)^\top \nabla^2_{xy} f \Delta y + \Delta x^\top \nabla^2_{xx} f \nabla^2_{xx} f \Delta x + 2\Delta x^\top \nabla^2_{xx} f \nabla^2_{xy} f \Delta y \\
&\qquad + 2\Delta y^\top \nabla^2_{yy} f \nabla_y f(x, y) + (2 - \frac{\eta}{\alpha})\nabla_y f(x, y)^\top \nabla^2_{yx} f \Delta x + \Delta y^\top \nabla^2_{yy} f \nabla^2_{yy} f \Delta y + 2\Delta y^\top \nabla^2_{yy} f \nabla^2_{yx} f \Delta x \\
&\qquad + 2\nabla_x f(x, y)^\top \mathcal{R}_x(\Delta x, \Delta y) + 2\Delta x^\top \nabla^2_{xx} f \mathcal{R}_x(\Delta x, \Delta y) + 2\Delta y^\top \nabla^2_{yx} f \mathcal{R}_x(\Delta x, \Delta y) + \|\mathcal{R}_x(\Delta x, \Delta y)\|^2 \\
&\qquad + 2\nabla_y f(x, y)^\top \mathcal{R}_y(\Delta x, \Delta y) + 2\Delta y^\top \nabla^2_{yy} f \mathcal{R}_y(\Delta x, \Delta y) + 2\Delta x^\top \nabla^2_{xy} f \mathcal{R}_y(\Delta x, \Delta y) + \|\mathcal{R}_y(\Delta x, \Delta y)\|^2
\end{aligned}$$

We use the update rule of *CGO*, Eq. (62) to substitute Δx and Δy and observe that $\nabla^2_{yx} f(I + \nabla^2_{xy} f \nabla^2_{yx} f)^{-1} = (I + \nabla^2_{yx} f \nabla^2_{xy} f)^{-1} \nabla^2_{yx} f$ as stated in Lemma (E.1) to obtain the following equality,

$$\begin{aligned}
&\Delta x^\top \nabla^2_{xy} f \nabla_y f(x,y) + \nabla_x f(x,y)^\top \nabla^2_{xy} f \Delta y \\
&\quad = -\eta\alpha \nabla_x f(x,y)^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f(x,y) \\
&\quad\quad - \eta\alpha \nabla_y f(x,y)^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f(x,y).
\end{aligned}$$

Yielding,

$$\begin{aligned}
&\|\nabla_x f\left(\Delta x + x, \Delta y + y\right)\|^2 + \|\nabla_y f\left(\Delta x + x, \Delta y + y\right)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&= 2\Delta x^\top \nabla^2_{xx} f \nabla_x f(x,y) - \eta\alpha(2 - \frac{\eta}{\alpha}) \nabla_x f(x,y)^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f(x,y) \\
&\quad + 2\Delta x^\top \nabla^2_{xx} f \nabla^2_{xy} f \Delta y + 2\Delta y^\top \nabla^2_{yy} f \nabla^2_{yx} f \Delta x + \Delta x^\top \nabla^2_{xx} f \nabla^2_{xx} f \Delta x + \Delta y^\top \nabla^2_{yy} f \nabla^2_{yy} f \Delta y \\
&\quad + 2\Delta y^\top \nabla^2_{yy} f \nabla_y f(x,y) - \eta\alpha(2 - \frac{\eta}{\alpha}) \nabla_y f(x,y)^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f(x,y) \\
&\quad + 2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y) + 2\Delta x^\top \nabla^2_{xx} f \mathcal{R}_x(\Delta x, \Delta y) + 2\Delta y^\top \nabla^2_{yx} f \mathcal{R}_x(\Delta x, \Delta y) + \|\mathcal{R}_x(\Delta x, \Delta y)\|^2 \\
&\quad + 2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y) + 2\Delta y^\top \nabla^2_{yy} f \mathcal{R}_y(\Delta x, \Delta y) + 2\Delta x^\top \nabla^2_{xy} f \mathcal{R}_y(\Delta x, \Delta y) + \|\mathcal{R}_y(\Delta x, \Delta y)\|^2.
\end{aligned}$$

We now substitute Δx and Δy using Eq. (61) yielding,

$$\begin{aligned}
&\|\nabla_x f\left(\Delta x + x, \Delta y + y\right)\|^2 + \|\nabla_y f\left(\Delta x + x, \Delta y + y\right)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&= -2\eta \nabla_x f(x,y)^\top \nabla^2_{xx} f \nabla_x f(x,y) + 2\eta \nabla_y f(x,y)^\top \nabla^2_{yy} f \nabla_y f(x,y) \\
&\quad - \eta\alpha(2 - \frac{\eta}{\alpha}) \nabla_x f(x,y)^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f(x,y) \\
&\quad + \Delta x^\top \nabla^2_{xx} f \nabla^2_{xx} f \Delta x + \frac{2}{\eta} \left(\underbrace{(\alpha + \eta)\Delta x}_{(i)} + \alpha^2 \nabla^2_{xy} f \Delta y \right)^\top \nabla^2_{xx} f \nabla^2_{xy} f \Delta y \\
&\quad + \Delta y^\top \nabla^2_{yy} f \nabla^2_{yy} f \Delta y + \frac{2}{\eta} \left(\underbrace{(\alpha + \eta)\Delta y}_{(ii)} - \alpha^2 \nabla^2_{yx} f \Delta x \right)^\top \nabla^2_{yy} f \nabla^2_{yx} f \Delta x \\
&\quad - \eta\alpha(2 - \frac{\eta}{\alpha}) \nabla_y f(x,y)^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f(x,y) \\
&\quad + 2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y) + \underbrace{2\Delta x^\top \nabla^2_{xx} f \mathcal{R}_x(\Delta x, \Delta y)}_{(iii)} + 2\Delta y^\top \nabla^2_{yx} f \mathcal{R}_x(\Delta x, \Delta y) + \|\mathcal{R}_x(\Delta x, \Delta y)\|^2 \\
&\quad + 2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y) + \underbrace{2\Delta y^\top \nabla^2_{yy} f \mathcal{R}_y(\Delta x, \Delta y)}_{(iv)} + 2\Delta x^\top \nabla^2_{xy} f \mathcal{R}_y(\Delta x, \Delta y) + \|\mathcal{R}_y(\Delta x, \Delta y)\|^2
\end{aligned} \tag{68}$$

Now we use Peter-Paul inequality, and bound the terms (i) and (ii) respectively as follows,

$$\begin{aligned}
\frac{2(\alpha + \eta)}{\eta} \Delta x^\top \nabla^2_{xx} f \nabla^2_{xy} f \Delta y &\leq 8\frac{\alpha + \eta}{\eta} \|\Delta x^\top \nabla^2_{xx} f\|^2 + \frac{\alpha + \eta}{8\eta} \|\nabla^2_{xy} f \Delta y\|^2 \\
\frac{2(\alpha + \eta)}{\eta} \Delta y^\top \nabla^2_{yy} f \nabla^2_{yx} f \Delta x &\leq 8\frac{\alpha + \eta}{\eta} \|\Delta y^\top \nabla^2_{yy} f\|^2 + \frac{\alpha + \eta}{8\eta} \|\nabla^2_{yx} f \Delta x\|^2
\end{aligned} \tag{69}$$

and terms (iii) and (iv) as,

$$\begin{aligned}
2\Delta x^\top \nabla^2_{xy} f \mathcal{R}_y(\Delta x, \Delta y) &\leq \|\mathcal{R}_y(\Delta x, \Delta y)\|^2 + \|\nabla^2_{xy} f \Delta y\|^2 \\
2\Delta y^\top \nabla^2_{yx} f \mathcal{R}_x(\Delta x, \Delta y) &\leq \|\mathcal{R}_x(\Delta x, \Delta y)\|^2 + \|\nabla^2_{yx} f \Delta x\|^2
\end{aligned} \tag{70}$$

Using the bounds obtained in Eqs. (69) and (70) we get,

$$
\begin{aligned}
&\|\nabla_x f(\Delta x + x, \Delta y + y)\|^2 + \|\nabla_y f(\Delta x + x, \Delta y + y)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&\quad \leq -2\eta \nabla_x f(x,y)^\top \nabla^2_{xx} f \nabla_x f(x,y) \\
&\qquad - \eta\alpha(2 - \frac{\eta}{\alpha}) \nabla_x f(x,y)^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f(x,y) \\
&\qquad + \frac{10\eta + 8\alpha}{\eta} \Delta x^\top \nabla^2_{xx} f \nabla^2_{xx} f \Delta x + \Delta y^\top \nabla^2_{yx} f \left(\frac{\alpha + \eta}{8\eta} I + \frac{2\alpha^2 \nabla^2_{xx} f}{\eta} \right) \nabla^2_{xy} f \Delta y \\
&\qquad + 2\eta \nabla_y f(x,y)^\top \nabla^2_{yy} f \nabla_y f(x,y) \\
&\qquad - \eta\alpha(2 - \frac{\eta}{\alpha}) \nabla_y f(x,y)^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f(x,y) \\
&\qquad + \frac{10\eta + 8\alpha}{\eta} \Delta y^\top \nabla^2_{yy} f \nabla^2_{yy} f \Delta y + \Delta x^\top \nabla^2_{xy} f \left(\frac{\alpha + \eta}{8\eta} I - \frac{2\alpha^2 \nabla^2_{yy} f}{\eta} \right) \nabla^2_{yx} f \Delta x \\
&\qquad + 2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y) + \underbrace{2\Delta y^\top \nabla^2_{yx} f \mathcal{R}_x(\Delta x, \Delta y)}_{(i)} + 2\|\mathcal{R}_x(\Delta x, \Delta y)\|^2 \\
&\qquad + 2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y) + \underbrace{2\Delta x^\top \nabla^2_{xy} f \mathcal{R}_y(\Delta x, \Delta y)}_{(ii)} + 2\|\mathcal{R}_y(\Delta x, \Delta y)\|^2.
\end{aligned}
$$

We use the Peter-Paul inequality to bound the term (i) as,

$$
2\Delta y^\top \nabla^2_{yx} f \mathcal{R}_x(\Delta x, \Delta y) \leq 4\|\mathcal{R}_x(\Delta x, \Delta y)\|^2 + \frac{1}{4}\|\nabla^2_{yx} f \Delta x\|^2
$$

and the term (ii) as,

$$
2\Delta x^\top \nabla^2_{xy} f \mathcal{R}_y(\Delta x, \Delta y) \leq 4\|\mathcal{R}_y(\Delta x, \Delta y)\|^2 + \frac{1}{4}\|\nabla^2_{xy} f \Delta y\|^2
$$

Substituting the above obtained bounds and noting that $\nabla_{xx} f$ and $\nabla_{yy} f$ are symmetric matrices we obtain,

$$
\begin{aligned}
&\|\nabla_x f(\Delta x + x, \Delta y + y)\|^2 + \|\nabla_y f(\Delta x + x, \Delta y + y)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&\quad \leq -2\eta \nabla_x f(x,y)^\top \nabla^2_{xx} f \nabla_x f(x,y) \\
&\qquad - \eta\alpha(2 - \frac{\eta}{\alpha}) \nabla_x f(x,y)^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f(x,y) \\
&\qquad + \frac{10\eta + 8\alpha}{\eta} \Delta x^\top \nabla^2_{xx} f \nabla^2_{xx} f \Delta x + \left(\frac{\alpha + 3\eta}{8\eta} + \frac{2\alpha^2 \underline{\underline{\lambda_{xx}}}}{\eta} \right) \|\nabla^2_{xy} f \Delta y\|^2 \\
&\qquad + 2\eta \nabla_y f(x,y)^\top \nabla^2_{yy} f \nabla_y f(x,y) \\
&\qquad - \eta\alpha(2 - \frac{\eta}{\alpha}) \nabla_y f(x,y)^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f(x,y) \\
&\qquad + \frac{10\eta + 8\alpha}{\eta} \Delta y^\top \nabla^2_{yy} f \nabla^2_{yy} f \Delta y + \left(\frac{\alpha + 3\eta}{8\eta} - \frac{2\alpha^2 \underline{\underline{\lambda_{yy}}}}{\eta} \right) \|\nabla^2_{yx} f \Delta x\|^2 \\
&\qquad + 2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y) + 6\|\mathcal{R}_x(\Delta x, \Delta y)\|^2 \\
&\qquad + 2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y) + 6\|\mathcal{R}_y(\Delta x, \Delta y)\|^2
\end{aligned}
$$

Using Eq. (61) to substitute Δx and Δy we compute,

$$
\begin{aligned}
\left\|\nabla^2_{yx} f \Delta x\right\|^2 = \eta^2 &\left(\nabla_x f(x,y) + \nabla^2_{xy} f \nabla_y f(x,y)\right)^\top \\
&\left(I + \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-2} \nabla^2_{xy} f \nabla^2_{yx} f \left(\nabla_x f(x,y) + \nabla^2_{xy} f \nabla_y f(x,y)\right)
\end{aligned}
$$

And,

$$\left\|\nabla_{xy}^2 f \Delta y\right\|^2 = \eta^2 \left(-\nabla_y f(x,y) + \nabla_{yx}^2 f \nabla_x f(x,y)\right)^\top \\ \left(I + \nabla_{yx}^2 f \nabla_{xy}^2 f\right)^{-2} \nabla_{yx}^2 f \nabla_{xy}^2 f \left(-\nabla_y f(x,y) + \nabla_{yx}^2 f \nabla_x f(x,y)\right)$$

By adding up the two, we obtain,

$$\begin{aligned}\left\|\nabla_{yx}^2 f \Delta x\right\|^2 + \left\|\nabla_{xy}^2 f \Delta y\right\|^2 &= \eta^2 \nabla_x f(x,y)^\top \left(I + \nabla_{xy}^2 f \nabla_{yx}^2 f\right)^{-2} \left(\nabla_{xy}^2 f \nabla_{yx}^2 f + \nabla_{xy}^2 f \nabla_{yx}^2 f\right) \nabla_x f(x,y) \\ &\quad + \eta^2 \nabla_y f(x,y)^\top \left(I + \nabla_{yx}^2 f \nabla_{xy}^2 f\right)^{-2} \left(\nabla_{yx}^2 f \nabla_{xy}^2 f + \nabla_{yx}^2 f \nabla_{xy}^2 f\right) \nabla_y f(x,y) \\ &= \eta^2 \nabla_x f(x,y)^\top \left(I + \alpha^2 \nabla_{xy}^2 f \nabla_{yx}^2 f\right)^{-1} \nabla_{xy}^2 f \nabla_{yx}^2 f \nabla_x f(x,y) \\ &\quad + \eta^2 \nabla_y f(x,y)^\top \left(I + \alpha^2 \nabla_{yx}^2 f \nabla_{xy}^2 f\right)^{-1} \nabla_{yx}^2 f \nabla_{xy}^2 f \nabla_y f(x,y)\end{aligned} \tag{71}$$

Setting $\overline{\lambda_1} = \max(\overline{\lambda_{xx}}, -\underline{\lambda_{yy}})$ and using Eq. (71) to substitute $\|\nabla_{xy}^2 f \Delta y\|^2 + \|\nabla_{yx}^2 f \Delta x\|^2$, we have,

$$\begin{aligned}&\|\nabla_x f\left(\Delta x + x, \Delta y + y\right)\|^2 + \|\nabla_y f\left(\Delta x + x, \Delta y + y\right)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\ &\quad \leq -2\eta \nabla_x f(x,y)^\top \nabla_{xx}^2 f \nabla_x f(x,y) + \frac{10\eta + 8\alpha}{\eta} \underbrace{\Delta x^\top \nabla_{xx}^2 f \nabla_{xx}^2 f \Delta x}_{(i)} \\ &\qquad + \eta \left(\frac{11\eta + 16\alpha^2 \overline{\lambda_1}}{8} - \frac{15}{8}\alpha\right) \nabla_x f(x,y)^\top \left(I + \alpha^2 \nabla_{xy}^2 f \nabla_{yx}^2 f\right)^{-1} \nabla_{xy}^2 f \nabla_{yx}^2 f \nabla_x f(x,y) \\ &\qquad + 2\eta \nabla_y f(x,y)^\top \nabla_{yy}^2 f \nabla_y f(x,y) + \frac{10\eta + 8\alpha}{\eta} \underbrace{\Delta y^\top \nabla_{yy}^2 f \nabla_{yy}^2 f \Delta y}_{(ii)} \\ &\qquad + \eta \left(\frac{11\eta + 16\alpha^2 \overline{\lambda_1}}{8} - \frac{15}{8}\alpha\right) \nabla_y f(x,y)^\top \left(I + \alpha^2 \nabla_{yx}^2 f \nabla_{xy}^2 f\right)^{-1} \nabla_{yx}^2 f \nabla_{xy}^2 f \nabla_y f(x,y) \\ &\qquad + 2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y) + 6\|\mathcal{R}_x(\Delta x, \Delta y)\|^2 \\ &\qquad + 2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y) + 6\|\mathcal{R}_y(\Delta x, \Delta y)\|^2\end{aligned} \tag{72}$$

Substituting Δx and Δy from Eq. (61) we bound the sum of terms (i) and (ii) as follows,

$$\begin{aligned}&\Delta x^\top \nabla_{xx}^2 f \nabla_{xx}^2 f \Delta x + \Delta y^\top \nabla_{yy}^2 f \nabla_{yy}^2 f \Delta y \\ &\quad = (-\alpha \nabla_{xy}^2 f \Delta y - \eta \nabla_x f(x,y))^\top \nabla_{xx}^2 f \nabla_{xx}^2 f (-\alpha \nabla_{xy}^2 f \Delta y - \eta \nabla_x f(x,y)) \\ &\qquad + (\alpha \nabla_{yx}^2 f \Delta x + \eta \nabla_y f(x,y))^\top \nabla_{yy}^2 f \nabla_{yy}^2 f (\alpha \nabla_{yx}^2 f \Delta x + \eta \nabla_y f(x,y)) \\ &\quad = \|\nabla_{xx}^2 f (\nabla_{xy}^2 f \Delta y + \eta \nabla_x f(x,y))\|^2 \\ &\qquad + \|\nabla_{yy}^2 f (\alpha \nabla_{yx}^2 f \Delta x + \eta \nabla_y f(x,y))\|^2 \\ &\quad \leq 2\alpha^2 \|\nabla_{xy}^2 f \Delta y\|^2 \|\nabla_{xx}^2 f\|^2 + 2\alpha^2 \|\nabla_{yx}^2 f \Delta x\|^2 \|\nabla_{yy}^2 f\|^2 \\ &\qquad + 2\eta^2 \|\nabla_{xx}^2 f \nabla_x f(x,y)\|^2 + 2\eta^2 \|\nabla_{yy}^2 f \nabla_y f(x,y)\|^2 \\ &\quad = 2\alpha^2 \|\nabla_{xy}^2 f \Delta y\|^2 \overline{\lambda_{xx}}^2 + 2\alpha^2 \|\nabla_{yx}^2 f \Delta x\|^2 \overline{\lambda_{yy}}^2 \\ &\qquad + 2\eta^2 \|\nabla_{xx}^2 f \nabla_x f(x,y)\|^2 + 2\eta^2 \|\nabla_{yy}^2 f \nabla_y f(x,y)\|^2\end{aligned}$$

Setting $\overline{\lambda_2} = \max(\overline{\lambda_{xx}}, \overline{\lambda_{yy}})$ and using Eq. (71) to substitute $\|\nabla^2_{xy} f \Delta y\|^2 + \|\nabla^2_{yx} f \Delta x\|^2$,

$$\begin{aligned}
\Delta x^\top \nabla^2_{xx} f \nabla^2_{xx} f \Delta x + \Delta y^\top \nabla^2_{yy} f \nabla^2_{yy} f \Delta y \\
\leq\ & 2\alpha^2 \overline{\lambda_2}^2 (\|\nabla^2_{xy} f \Delta y\|^2 + \|\nabla^2_{yx} f \Delta x\|^2) \\
& + 2\eta^2 \|\nabla^2_{xx} f \nabla_x f(x,y)\|^2 + 2\eta^2 \|\nabla^2_{yy} f \nabla_y f(x,y)\|^2 \\
\leq\ & 2\alpha^2 \eta^2 \overline{\lambda_2}^2 \nabla_x f(x,y)^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f(x,y) \\
& + 2\alpha^2 \eta^2 \overline{\lambda_2}^2 \nabla_y f(x,y)^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f(x,y) \\
& + 2\overline{\lambda_{xx}}^2 \nabla_x f(x,y)^\top \nabla_x f(x,y) \\
& + 2\overline{\lambda_{yy}}^2 \nabla_y f(x,y)^\top \nabla_y f(x,y)
\end{aligned}$$

Substituting the above bound in Eq. (72) we obtain,

$$\begin{aligned}
& \|\nabla_x f(\Delta x + x, \Delta y + y)\|^2 + \|\nabla_y f(\Delta x + x, \Delta y + y)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
& \leq -\eta \nabla_x f(x,y)^\top \left(2\nabla^2_{xx} f - 2\frac{10\eta + 8\alpha}{\eta} \overline{\lambda_{xx}}^2\right) \nabla_x f(x,y) \\
& \quad + \eta \nabla_y f(x,y)^\top \left(2\nabla^2_{yy} f + 2\frac{10\eta + 8\alpha}{\eta} \overline{\lambda_{yy}}^2\right) \nabla_y f(x,y) \\
& \quad + \left(\eta\left(\frac{11\eta + 16\alpha^2 \overline{\lambda_1}}{8} - \frac{15}{8}\alpha\right) + 2(10\eta + 8\alpha)\alpha^2 \eta \overline{\lambda_2}^2\right) \\
& \qquad (\nabla_y f(x,y)^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f(x,y) \\
& \quad + \nabla_x f(x,y)^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f(x,y)) \\
& \quad + 2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y) + 6\|\mathcal{R}_x(\Delta x, \Delta y)\|^2 \\
& \quad + 2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y) + 6\|\mathcal{R}_y(\Delta x, \Delta y)\|^2.
\end{aligned} \tag{73}$$

To conclude, we need to bound the $\mathcal{R}$-terms. Using the Lipschitz-continuity of the Hessian, and equations Eq. (63) and Eq. (64) we can bound,

$$\|\mathcal{R}_x(\Delta x, \Delta y)\|, \|\mathcal{R}_y(\Delta x, \Delta y)\| \leq L_{xy}(\|\Delta x\| + \|\Delta y\|)^2 \tag{74}$$

Using Eq. (61) we get,

$$\begin{aligned}
(\|\Delta x\|^2 + \|\Delta y\|^2) =\ & \eta^2 (\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2) \\
& + 2\eta\alpha(\nabla_x f(x,y)^\top \nabla^2_{yx} f \Delta x + \nabla_y f(x,y)^\top \nabla^2_{xy} f \Delta y) \\
& + \alpha^2 (\|\nabla^2_{yx} f \Delta x\|^2 + \|\nabla^2_{xy} f \Delta y\|^2)
\end{aligned}$$

From Eq. (71) we have,

$$\begin{aligned}
\alpha^2 (\left\|\nabla^2_{yx} f \Delta x\right\|^2 + \left\|\nabla^2_{xy} f \Delta y\right\|^2) =\ & \eta^2 \nabla_x f(x,y)^\top \left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f \nabla_x f(x,y) \\
& + \eta^2 \nabla_y f(x,y)^\top \left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f \nabla_y f(x,y) \\
\leq\ & \eta^2 (\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)
\end{aligned} \tag{75}$$

And observe,

$$
\begin{aligned}
\nabla_x f(x,y)^\top \nabla^2_{yx} f \Delta x + \nabla_y f(x,y)^\top \nabla^2_{xy} f \Delta y &= (\nabla_x f(x,y), \nabla_y f(x,y))^\top (\nabla^2_{yx} f \Delta x, \nabla^2_{xy} f \Delta y) \\
&\overset{(c)}{\leq} \|(\nabla_x f(x,y), \nabla_y f(x,y))\| \|(\nabla^2_{yx} f \Delta x, \nabla^2_{xy} f \Delta y)\| \\
&\overset{(d)}{\leq} \frac{\eta}{\alpha} (\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)
\end{aligned} \tag{76}
$$

Where in (c) we use the Cauchy-Schwarz inequality and in (d) we use the bound derived in Eq. (75). We then substitute Δx and Δy using Eq. (62) to obtain,

$$
\begin{aligned}
L_{xy}(\|\Delta x\| + \|\Delta y\|)^2 &\leq 2L_{xy}(\|\Delta x\|^2 + \|\Delta y\|^2) \\
&\leq 2L_{xy}(\eta^2(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2) \\
&\quad + 2\eta\alpha \underbrace{(\nabla_x f(x,y)^\top \nabla^2_{yx} f \Delta x + \nabla_y f(x,y)^\top \nabla^2_{xy} f \Delta y)}_{(i)} \\
&\quad + \alpha^2 \underbrace{(\|\nabla^2_{yx} \Delta x\|^2 + \|\nabla^2_{xy} f \Delta y\|^2))}_{(ii)} \\
&\overset{(e)}{\leq} 8\eta^2 L_{xy}(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)
\end{aligned} \tag{77}
$$

Where in (e) we have used Eq. (76) to bound term (i) and Eq. (75) to bound (ii). Combining Eq. (74) and Eq. (77) we obtain,

$$
\|\mathcal{R}_x(\Delta x, \Delta y)\|, \|\mathcal{R}_y(\Delta x, \Delta y)\| \leq 8\eta^2 L_{xy}(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2) \tag{78}
$$

Also we have,

$$
\begin{aligned}
&2\nabla_x f(x,y)^\top \mathcal{R}_x(\Delta x, \Delta y) + 2\nabla_y f(x,y)^\top \mathcal{R}_y(\Delta x, \Delta y) \\
&\overset{(a)}{\leq} 2(\|\nabla_x f(x,y)\| \|\mathcal{R}_x(\Delta x, \Delta y)\| + \|\nabla_y f(x,y)\| \|\mathcal{R}_y(\Delta x, \Delta y)\|) \\
&\overset{(b)}{\leq} 16L_{xy}\eta^2(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|)(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)
\end{aligned} \tag{79}
$$

Where we use Cauchy-Schwarz inequality in (a) and Eq. (74) in (b). Finally we use the bounds in Eq. (78) and Eq. (79) to bound the terms containing $\mathcal{R}_x(\Delta x, \Delta y)$ and $\mathcal{R}_y(\Delta x, \Delta y)$ in Eq. (73) and further set $k = \eta(\frac{11\eta + 16\alpha^2\overline{\lambda_1}}{8} - \frac{15}{8}\alpha) + 2(10\eta + 8\alpha)\alpha^2\eta\overline{\lambda_2}^2$ to obtain,

$$
\begin{aligned}
&\|\nabla_x f(\Delta x + x, \Delta y + y)\|^2 + \|\nabla_y f(\Delta x + x, \Delta y + y)\|^2 - \|\nabla_x f(x,y)\|^2 - \|\nabla_y f(x,y)\|^2 \\
&\leq -\nabla_x f(x,y)^\top (\eta \left(2\nabla^2_{xx} f - 2\frac{10\eta + 8\alpha}{\eta}\overline{\lambda_{xx}}^2\right) + k\left(I + \alpha^2 \nabla^2_{xy} f \nabla^2_{yx} f\right)^{-1} \nabla^2_{xy} f \nabla^2_{yx} f \\
&\quad - 16\eta^2 L_{xy}\left(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|\right) \\
&\quad - 384\eta^4 L^2_{xy}(\|\nabla_x f(x,y)\|^2 + \|\nabla_x f(x,y)\|^2))\nabla_x f(x,y) \\
&\quad - \nabla_y f(x,y)^\top (-\eta \left(2\nabla^2_{yy} f + 2\frac{10\eta + 8\alpha}{\eta}\overline{\lambda_{yy}}^2\right) + k\left(I + \alpha^2 \nabla^2_{yx} f \nabla^2_{xy} f\right)^{-1} \nabla^2_{yx} f \nabla^2_{xy} f \\
&\quad - 16\eta^2 L_{xy}\left(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|\right) \\
&\quad - 384\eta^4 L^2_{xy}(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2))\nabla_y f(x,y)
\end{aligned}
$$

Rearranging we obtain,

$$\|(\nabla_x f(x+x, y+y), \nabla_y f(x+x, y+y)\| \leq (1-\lambda_{\min})\|(\nabla_x f(x,y), \nabla_y f(x,y)\|$$

Thus for $1 \geq \lambda_{min} > 0$ where,

$$\begin{aligned}\lambda_{min} = \min\Big\{&\lambda_{min}(\eta\left(2\nabla^2_{xx} f - 2\frac{10\eta+8\alpha}{\eta}\overline{\lambda_{xx}}^2\right) + k\left(I + \alpha^2\nabla^2_{xy} f\nabla^2_{yx} f\right)^{-1}\nabla^2_{xy} f\nabla^2_{yx} f\\ &- 16\eta^2 L\left(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|\right)\\ &- 384\eta^4 L^2(\|\nabla_x f(x,y)\|^2 + \|\nabla_x f(x,y)\|^2)),\\ &\lambda_{min}(-\eta\left(2\nabla^2_{yy} f + 2\frac{10\eta+8\alpha}{\eta}\overline{\lambda_{yy}}^2\right) + k\left(I + \alpha^2\nabla^2_{yx} f\nabla^2_{xy} f\right)^{-1}\nabla^2_{yx} f\nabla^2_{xy} f\\ &- 16\eta^2 L\left(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|\right)\\ &- 384\eta^4 L^2(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)\Big\}\end{aligned}$$

□

we have exponential convergence with rate $(1-\lambda_{min})$.

Now, we simplify the above expression using Lemmas (E.2) and (E.3) to obtain,

$$\begin{aligned}\lambda_{min} \geq \min\Big\{&\eta(2\underline{\lambda_{xx}} - 2\frac{10\eta+8\alpha}{\eta}\overline{\lambda_{xx}}^2) + k\frac{\underline{\lambda_{xy}}}{1+\alpha^2\underline{\lambda_{xy}}} - 16\eta^2 L\left(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|\right)\\ &- 384\eta^4 L^2(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2),\\ &-\eta(2\underline{\lambda_{yy}} + 2\frac{10\eta+8\alpha}{\eta}\overline{\lambda_{yy}}^2) + k\frac{\underline{\lambda_{yx}}}{1+\alpha^2\underline{\lambda_{yx}}} - 16\eta^2 L\left(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|\right)\\ &- 384\eta^4 L^2(\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2)\Big\}\end{aligned}$$

When initializing close to the stationary point, the Lipschitz-continuity of the gradient guarantees that the terms $(\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\|)$ and $\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2$ are small and we have,

$$\begin{aligned}\lambda_{min} \geq \min\Big\{&\eta(2\underline{\lambda_{xx}} - 2\frac{10\eta+8\alpha}{\eta}\overline{\lambda_{xx}}^2) + k\frac{\underline{\lambda_{xy}}}{1+\alpha^2\underline{\lambda_{xy}}},\\ &-\eta(2\underline{\lambda_{yy}} + 2\frac{10\eta+8\alpha}{\eta}\overline{\lambda_{yy}}^2) + k\frac{\underline{\lambda_{yx}}}{1+\alpha^2\underline{\lambda_{yx}}}\Big\}\end{aligned}$$

which is the statement of our Theorem.

H. Convergence for α-coherent functions

H.1. *CGO* converges to a saddle point under strictly α-coherent functions

Proof of Theorem (5.3). We prove the convergence through contradiction. Let us assume that the algorithm does not converge to a saddle point. Let $z_n := (x_n, y_n)$ denote the parameters at the n'th iterate of the algorithm. $g_{\alpha,n} := g_\alpha(z_n)$ denote the vector g_α evaluated at z_n. Let the set of saddle points be $\mathcal{Z}^*$, and let all the iterates of the algorithm lie in a compact set $\mathcal{C}$. Then from the assumption we have $\mathcal{Z}^* \cap \mathcal{C} = \phi$. Now from the definition of strict coherence we have $\langle g_{\alpha,n}, z - z^*\rangle \geq a$ for some $a > 0$ and $z^* \in \mathcal{Z}^*$ and $, \forall z \in \mathcal{C}$. Such a z^* is guaranteed by definition (3.4)(2^{nd} point).

Recall the proximal map defined in Eq. (12),

$$z^+ = P_z(y) = \arg\min_{z'\in\mathcal{Z}}\{\langle y, z - z'\rangle + D(z', z)\} = \arg\max_{z'\in\mathcal{Z}}\{\langle y + \nabla h(z), z'\rangle - h(z')\}. \quad (80)$$

Then for Bregmann Divergence $D_h(x, y)$ with K-strongly convex potential function h and 2-norm $\|.\|$ we have, (Mertikopoulos et al., 2019)(Proposition B.3),

$$D(p, z^+) \leq D(p, z) + \langle y, z - p\rangle + \frac{K}{2}\|\Delta y\|^2. \tag{81}$$

To obtain the *CGO* update we substitute $y = -\eta_n g_{\alpha,n}, z = z_n, z^+ = z_{n+1}, p = z^*, h = \frac{\|.\|_2^2}{2}$ in Eq. (81) we get,

$$D(z^*, z_{n+1}) = D(z^*, P_{z_n}(-\eta_n g_{\alpha,n})) \leq D(z^*, z_n) - \eta_n\langle g_{\alpha,n}, z_n - z^*\rangle + \frac{\eta_n^2\|g_{\alpha,n}\|^2}{2}$$

Note that the above substitution in Eq.(81) is equivalent to *CGO* only in the interior of the domain. At the boundary we need an additional projection step since in Eq. (81) we only look within the domain for the minimum.

Since the saddle point is α-coherent we have $\langle g_{\alpha,n}, z - z^*\rangle \geq a$ for some $a > 0$.

$$D(z^*, z_{n+1}) \leq D(z^*, z_n) - \eta_n a + \frac{\eta_n^2\|g_{\alpha,n}\|^2}{2} \leq D(z^*, z_0) - (a - \frac{\sum_{k=1}^n \|\eta_k\|^2}{2\sum_{k=1}^n \eta_k})\sum_{k=1}^n \eta_k$$

Since we have $\sum_{k=1}^n \eta_k = \infty$ and $\sum_{k=1}^n \|\eta_k\|^2 < \infty$, we obtain $\lim_{n\to\infty} D_n = -\infty$, which is a contradiction since the divergence is positive. Hence *CGO* converges to a saddle point. □

H.2. *oCGO* converges to a saddle point under α-coherent functions

Proof of Theorem (5.4). Let $P_z(y)$ be as in Eq. (12) and $z_1^+ = P_z(y_1), z_2^+ = P_z(y_2)$. We then have for Bregmann Divergence $D_h(x, y)$ with K-strongly convex potential function h, 2-norm $\|.\|$ and a fixed point p (Mertikopoulos et al., 2019)(Proposition B.4),

$$D(p, x_2^+) \leq D(p, x) + \langle y_2, x_1^+ - p\rangle + \frac{1}{2K}\|\Delta y_2 - y_1\|^2 - \frac{K}{2}\|\Delta x_1^+ - x\|^2. \tag{82}$$

Let p^* be a solution of the SP problem such that α-MVI holds $\forall z \in \mathcal{X} \times \mathcal{Y}$, the existence of such a p is guaranteed via the definition of α-coherence Def. (3.4)(2^{nd} point).

In order to obtain the *oCGO* update we substitute $y_1 = -\eta_n g_{\alpha,n}, y_2 = -\eta_n g_{\alpha,n+\frac{1}{2}}, x = z_n, x_1^+ = z_{n+\frac{1}{2}}, x_2^+ = z_{n+1}, p = p^*$ and set $h = \frac{\|.\|^2}{2}$ (for this h we have $K = 1$),

$$D(x^*, z_{n+1}) \leq D(x^*, z_n) - \eta_n\langle g_{\alpha,n+\frac{1}{2}}, z_{n+\frac{1}{2}} - x^*\rangle + \frac{\eta_n^2}{2}\|g_{\alpha,n+\frac{1}{2}} - g_{\alpha,n}\|^2 - \frac{1}{2}\|z_{n+\frac{1}{2}} - z_n\|^2$$

From coherence condition we have,

$$D(p^*, z_{n+1}) \leq D(p^*, z_n) + \frac{\eta_n^2}{2}\|g_{\alpha,n+\frac{1}{2}} - g_{\alpha,n}\|^2 - \frac{1}{2}\|z_{n+\frac{1}{2}} - z_n\|^2 \tag{83}$$

Using Eq. (61) we get,

$$\|g_{\alpha,n+\frac{1}{2}} - g_{\alpha,n}\|^2 \leq \|g_{0,n+\frac{1}{2}} - g_{0,n}\|^2 + \frac{\alpha^2}{\eta_n^2}(\|\nabla_{xy,n+\frac{1}{2}} f\Delta y_{n+\frac{1}{2}} - \nabla_{xy,n} f\Delta y_n\|^2 + \|\nabla_{xy,n+\frac{1}{2}} f^\top \Delta x_{n+\frac{1}{2}} - \nabla_{xy,n} f^\top \Delta x_n\|^2)$$

Where $(\Delta x_n, \Delta y_n) = -\eta_n g_{\alpha,n}$, $(\Delta x_{n+\frac{1}{2}}, \Delta y_{n+\frac{1}{2}}) = -\eta_n g_{\alpha,n+\frac{1}{2}}$ and $\nabla_{xy,n} f, \nabla_{xy,n+\frac{1}{2}} f$ are the second order cross terms evaluated at $z_n, z_{n+\frac{1}{2}}$. We can re-write the above as,

$$\begin{aligned}
\frac{1}{\eta_n^2}\|(\Delta x_{n+\frac{1}{2}} - \Delta x_n, \Delta y_{n+\frac{1}{2}} - \Delta y_n)\|^2 \\
&\leq \|g_{0,n+\frac{1}{2}} - g_{0,n}\|^2 \\
&\quad + \frac{\alpha^2}{\eta_n^2}\|\nabla_{xy,n+\frac{1}{2}} f \Delta y_{n+\frac{1}{2}} - \nabla_{xy,n+\frac{1}{2}} f \Delta y_n + \nabla_{xy,n+\frac{1}{2}} f \Delta y_n - \nabla_{xy,n} f \Delta y_n\|^2 \\
&\quad + \frac{\alpha^2}{\eta_n^2}\|\nabla_{xy,n+\frac{1}{2}} f^\top \Delta x_{n+\frac{1}{2}} - \nabla_{xy,n+\frac{1}{2}} f^\top \Delta x_n + \nabla_{xy,n+\frac{1}{2}} f^\top \Delta x_n - \nabla_{xy,n} f^\top \Delta x_n\|^2 \\
&\leq \|g_{0,n+\frac{1}{2}} - g_{0,n}\|^2 \\
&\quad + \frac{\alpha^2}{\eta_n^2}(\|\nabla_{xy,n+\frac{1}{2}} f\|^2 \|\Delta y_{n+\frac{1}{2}} - \Delta y_n\|^2 + \|\Delta y_n\|^2 \|\nabla_{xy,n+\frac{1}{2}} f - \nabla_{xy,n} f\|^2) \\
&\quad + \frac{\alpha^2}{\eta_n^2}(\|\nabla_{xy,n+\frac{1}{2}} f^\top\|^2 \|\Delta x_{n+\frac{1}{2}} - \Delta x_n\|^2 + \|\Delta x_n\|^2 \|\nabla_{xy,n+\frac{1}{2}} f^\top - \nabla_{xy,n} f^\top\|^2)
\end{aligned}$$

Using the Lipschitz continuity of the Hessian terms, setting $\alpha^2\|\nabla_{xy,n+\frac{1}{2}} f\|^2 = \alpha^2\|\nabla_{xy,n+\frac{1}{2}} f^\top\|^2 = \alpha^2 L_{xy}^2 \leq 1$, and rearranging we get,

$$\begin{aligned}
\frac{1}{\eta_n^2}\|\Delta x_{n+\frac{1}{2}} - \Delta x_n, \Delta y_{n+\frac{1}{2}} - \Delta y_n\|^2 &\leq \frac{1}{1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2}\|g_{0,n+\frac{1}{2}} - g_{0,n}\|^2 \\
&\quad + \frac{\alpha^2}{\eta_n^2(1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2)}(\|\Delta y_n\|^2 \|\nabla_{xy,n+\frac{1}{2}} f - \nabla_{xy,n} f\|^2) \\
&\quad + \frac{\alpha^2}{\eta_n^2(1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2)}(\|\Delta x_n\|^2 \|\nabla_{xy,n+\frac{1}{2}} f^\top - \nabla_{xy,n} f^\top\|^2) \\
&\leq \frac{L^2}{\eta_n^2(1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2)}\|z_{n+\frac{1}{2}} - z_n\|^2 \\
&\quad + \frac{L_{xy}^2 \alpha^2}{\eta_n^2(1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2)}(\|\Delta x_n\|^2 + \|\Delta y_n\|^2)\|z_{n+\frac{1}{2}} - z_n\|^2
\end{aligned}$$

Finally we have,

$$\begin{aligned}
\|g_{\alpha,n+\frac{1}{2}} - g_{\alpha,n}\|^2 \leq & \frac{L^2 + L_{xy}^2 \alpha^2(\|\Delta x_n\|^2 + \|\Delta y_n\|^2)}{\eta_n^2(1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2)}\|z_{n+\frac{1}{2}} - z_n\|^2 \\
\leq & \frac{L^2 + L_{xy}^2 \alpha^2(\alpha + \eta)^2 \|g_{0,n}\|^2}{\eta_n^2(1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2)}\|z_{n+\frac{1}{2}} - z_n\|^2
\end{aligned} \tag{84}$$

Substituting in Eq. (83) we get,

$$\begin{aligned}
D(p^*, z_{n+1}) \leq & D(p^*, z_n) + \|z_{n+\frac{1}{2}} - z_n\|^2 \left(\frac{\eta_n^2 L'^2 + L_{xy}^2 \alpha^2 (\alpha + \eta_n)^2 \|g_{0,n}\|^2}{2(1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2)} - \frac{1}{2}\right) \\
\leq & D(p^*, z_n) + \|z_{n+\frac{1}{2}} - z_n\|^2 \left(\frac{\eta_n^2 L'^2 + L_{xy}^2 \alpha^2 (\alpha + \eta_n)^2 L^2}{2(1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2)} - \frac{1}{2}\right)
\end{aligned} \tag{85}$$

Hence if α satisfies the following,

$$\alpha^4 L_{xy}^2 L^2 + \alpha^2 L'^2 - 1 < 0$$

or equivalently we have,

$$-\sqrt{\frac{\sqrt{L'^4 + 4L_{xy}^2 L^2} - L'^2}{2L_{xy}^2 L^2}} < \alpha < \sqrt{\frac{\sqrt{L'^4 + 4L_{xy}^2 L^2} - L'^2}{2L_{xy}^2 L^2}} \tag{86}$$

and also η_n satisfying the following,

$$0 < \eta_n < \frac{\sqrt{\alpha^2 L^2 L_{xy}^2 + L'^2 - 2\alpha^4 L^2 L'^2 L_{xy}^2 - \alpha^2 L'^4} - \alpha^3 L_0^2 L_{xy}^2}{\alpha^2 L^2 L_{xy}^2 + L'^2} \tag{87}$$

We have,

$$\frac{\eta_n^2 L'^2 + L_{xy}^2 \alpha^2 (\alpha + \eta_n)^2 L^2}{2(1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2)} - \frac{1}{2} < 0$$

and the divergence decreases at each step. By telescoping Eq. (85) we obtain,

$$\sum_{k=1}^{n} \|z_{k+\frac{1}{2}} - z_k\|^2 (1 - \frac{\eta_k^2 L'^2 + L_{xy}^2 \alpha^2 (\alpha + \eta_k)^2 L^2}{(1 - \|\nabla_{xy,k+\frac{1}{2}}\|^2 \alpha^2)}) \leq 2D(x^*, z_1). \tag{88}$$

We also know $z_{k+\frac{1}{2}} - z_k = -\eta_k g_{\alpha,k}$, thus for α and η_n satisfying Eq. (86) and Eq. (87), we have,

$$\frac{1}{n} \sum_{k=1}^{n} \|z_{k+\frac{1}{2}} - z_k\|^2 = \frac{1}{n} \sum_{k=1}^{n} \eta_k^2 \|g_{\alpha,k}\|^2 \leq \frac{2}{nc} D(x^*, z_1). \tag{89}$$

Where $1 - \frac{\eta_k^2 L'^2 + L_{xy}^2 \alpha^2 (\alpha + \eta_k)^2 L^2}{(1 - \|\nabla_{xy,k+\frac{1}{2}}\|^2 \alpha^2)} > c, \forall \eta_k$. If we assume without loss of generality that η_k converges to η, then we have from Eq. (89) that the average of $\|g_{\alpha,n}\|$ and $\|z_{n+\frac{1}{2}} - z_n\|$ falls with order $O(\frac{1}{n})$ where n is the iteration count.
Taking limit of $z_{k+\frac{1}{2}}$ we have,

$$z^* = \lim_{k \to \infty} z_{k+\frac{1}{2}} = P_{z^*}(-\eta g_\alpha(z^*)),$$

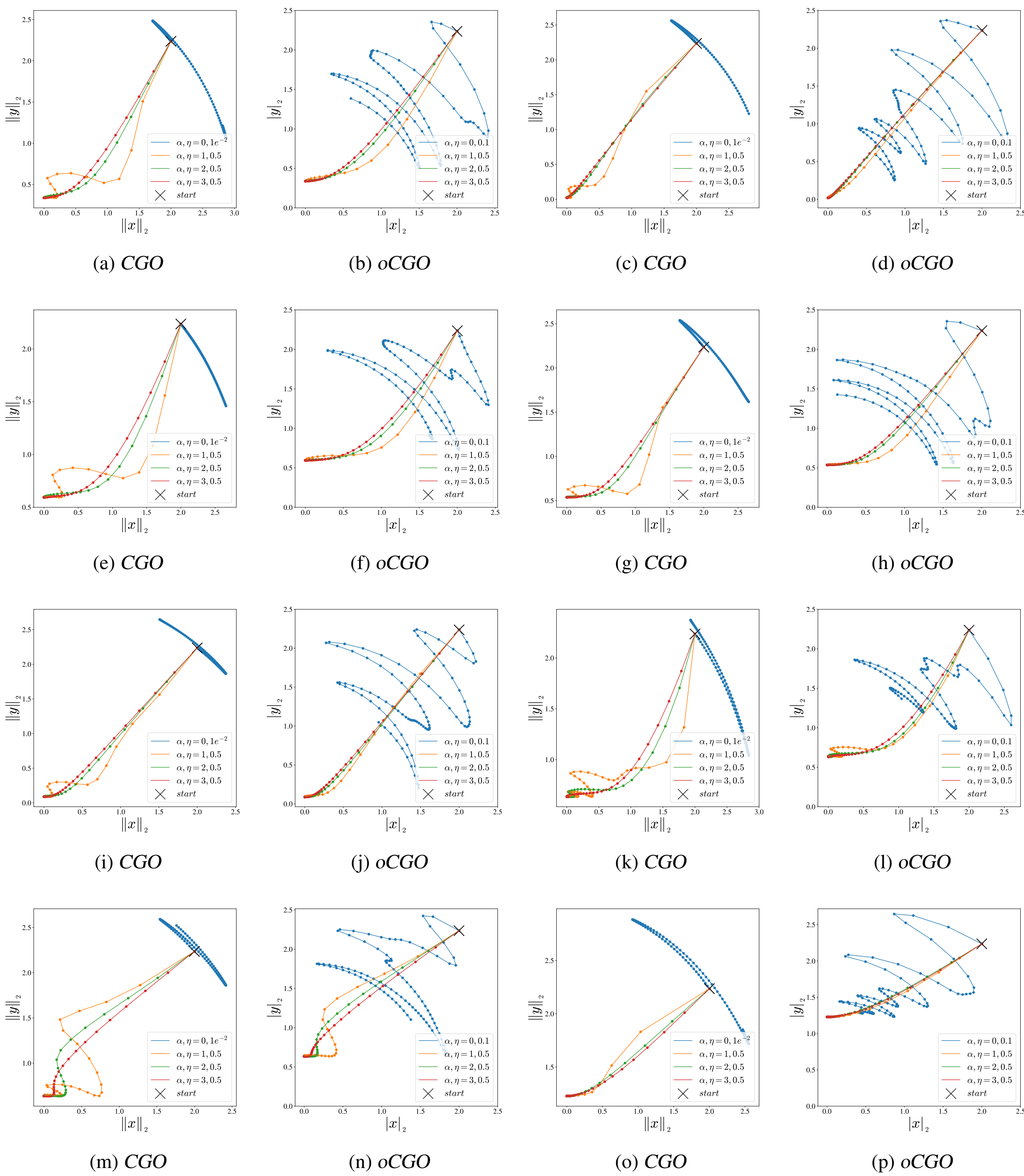
this implies z^* satisfies α-SVI and is hence a solution of the SP problem via definition of α-coherence Def. (3.4) (1^{st} point). Coherence condition Def. (3.4)(3^{rd} point) implies that α-MVI holds locally around z^*. Thus for, α and η_n satisfying Eq. (86) and Eq. (87) respectively, and sufficiently large n, we have,

$$D(z^*, z_{n+1}) \leq D(z^*, z_n) + \|z^* - z_n\|^2 (\frac{\eta_n^2 L'^2 + L_{xy}^2 \alpha^2 (\alpha + \eta_n)^2 L^2}{2(1 - \|\nabla_{xy,n+\frac{1}{2}} f\|^2 \alpha^2)} - \frac{1}{2}) \overset{(a)}{\leq} D(z^*, z_n)$$

Where the equality in (a) holds if and only if $z^* = z_n$. Thus $D(z^*, z_n)$ is non-increasing and $z_n \to z^*$ which is a saddle point. □

I. Additional examples and simulations

We now present some more simulations of *CGO* and *oCGO* on the function $x^\top Ay$ with multiple samples of the matrix $A = (a_{ij}), a_{ij} \sim \mathcal{N}(0, 1)$.



(a) *CGO* (b) *oCGO* (c) *CGO* (d) *oCGO*

(e) *CGO* (f) *oCGO* (g) *CGO* (h) *oCGO*

(i) *CGO* (j) *oCGO* (k) *CGO* (l) *oCGO*

(m) *CGO* (n) *oCGO* (o) *CGO* (p) *oCGO*

Figure 8. CGO and *oCGO* on bilinear function family $f(x, y) = x^\top Ay, x \in \mathbb{R}^4, y \in \mathbb{R}^5$ for 100 iterations. In each row, the 1^{st} and 2^{nd} as well as the 3^{rd} and 4^{th} figures correspond to the same sample of A