TIGHT CONVERGENCE ANALYSIS OF ALTERNATING GRADIENT DESCENT-ASCENT WITH NEGATIVE MOMENTUM VIA PARAMETRIC SDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient descent-ascent is a widely used method in minimax optimization, but it cannot guarantee convergence in general convex-concave problems. Alternating gradient descent-ascent with negative momentum has been proposed to resolve this issue, but the existing convergence guarantees cover only a particular choice of the stepsize and the momentum parameter. In this work, we improve the convergence guarantees so that a broader range of parameters is covered, including arbitrarily small negative momentum parameters and stepsizes. For convex-concave objectives, we show both convergence upper and lower bounds that match up to constant factors, establishing a convergence rate tight in the stepsize, momentum, and iteration count. For strongly-convex-strongly-concave objectives, we prove that the method enjoys a worst-case optimal rate. Our convergence proofs use a computer-assisted search for Lyapunov potentials inspired by the performance estimation problem framework. Covering the aforementioned parameter range leads to semidefinite programs with multiple parameters, and to make solving them tractable, we develop techniques for reducing the number of parameters involved.

1 INTRODUCTION

Minimax problems, which are optimization problems of the form

$$\min_x \max_y \; f(x, y), \tag{1}$$

appear in a diversity of interesting machine learning problems, such as generative adversarial networks (Arjovsky et al., 2017; Goodfellow et al., 2014), adversarial training (Mądry et al., 2018), reinforcement learning (Littman, 1994; Pinto et al., 2017) and neural optimal transport (Fan et al., 2022; Rout et al., 2022), but are also infamous for being notoriously hard to solve. Indeed, the naïve *gradient descent-ascent* (GDA) method, which applies gradient descent in x and gradient ascent in y, can fail to converge even when f is convex-concave, *i.e.*, convex in x and concave in y (He et al., 2014; Mescheder et al., 2018).

Classical remedies for this nonconvergence issue include the *proximal point* method (Martinet, 1970), the *extragradient* method (Korpelevich, 1976), and the *optimistic gradient* method (Popov, 1980). In short, the proximal point method uses an implicit update, while extragradient uses a predictor–corrector type update and optimistic gradient combines current and past gradient information. Nonetheless, we ask whether a few familiar algorithmic ingredients can stabilize GDA.

As an answer to this question, we study the *alternating gradient descent-ascent method with negative momentum* (AltGDA-NeMo)

$$x_{k+1} = x_k - \eta \nabla_x f(x_k, y_k) + \beta(x_k - x_{k-1}) \tag{2a}$$

$$y_{k+1} = y_k + \eta \nabla_y f(x_{k+1}, y_k) + \beta(y_k - y_{k-1}) \tag{2b}$$

where $\eta > 0$ is the *stepsize* and $\beta < 0$ is the *momentum parameter*. With the initial point (x_0, y_0), to ensure that (2) is well-defined when $k = 0$, we formally set $(x_{-1}, y_{-1}) = (x_0, y_0)$. Notice that an *alternating update* is used, where the minimization step (2a) and the maximization step (2b) are executed in turn, rather than simultaneously. More importantly, the momentum parameter β is set to be a *negative* number, contrary to the common practice employed in solving minimization problems.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

As we discuss in Section 2, we would like to remark that the method design itself is not new. In particular, Shugart et al. (2026) gave a proof of convergence of AltGDA-NeMo with $(\eta, \beta) = (\frac{1}{5L}, -\frac{1}{2})$, where L is the smoothness constant of f. More precisely, they showed that the method converges to a stationary point when f is convex-concave, and in addition at a worst-case optimal rate when f is strongly-convex-strongly-concave. While this is, to our knowledge, the first convergence result for AltGDA-NeMo for general convex-concave objectives, the proof is specialized to a specific (η, β) pair, leaving open the problem of obtaining a convergence proof applicable to a more general setting.

In this work, we significantly generalize the convergence proof of AltGDA-NeMo, establishing the guarantee over the triangular region $\{(\eta, \beta) : 0 < \eta L \leq -\beta \leq \frac{1}{2}\}$. Inspired by Shugart et al. (2026), our approach is built upon the recent developments in the *performance estimation problem* (PEP) framework and semidefinite programming (SDP) approaches to construct Lyapunov potentials for convergence proofs. Along the way, we also develop a systematic pipeline of reducing the number of parameters involved when using the SDP approach, which we expect to be applicable to proving convergence of other methods that are similarly formulatable. Our contributions are as follows.

- We prove the convergence of AltGDA-NeMo for convex-concave and strongly-convex-strongly-concave problems for all parameters that satisfy $0 < \eta L \leq -\beta \leq \frac{1}{2}$. Interestingly, AltGDA-NeMo converges for an arbitrarily small negative momentum parameter. With an admissible choice $\eta = \Theta(1/L)$, our convergence rate for strongly-convex-strongly-concave problems is worst-case optimal up to constant factors, matching the lower bound of Zhang et al. (2022b).
- Complementing our convergence upper bound for convex-concave problems, we prove a lower bound on the best-iterate squared gradient norm that matches the upper bound up to a constant factor. In particular, our results show that the found rate is tight, not only in terms of the number of iterations but also in terms of the parameters η and β, within the certified triangular region of (η, β).
- Our convergence upper bounds build upon the SDP based Lyapunov analysis framework. To construct Lyapunov certificates throughout the parameter region, we reduce the SDP search over the stepsize, momentum, and strong-convexity parameters to two successive one-parameter SDPs, as detailed in Section 6. This reduction combines a stepsize rescaling with a development of a sufficient condition for extending convex-concave certificates to the strongly-convex-strongly-concave setting.

2 Related Works

Alternating updates in minimax optimization At least since Martinet (1970), the *saddle gradient* operator $\boldsymbol{F} = (\nabla_x f, -\nabla_y f)$ has provided a classical viewpoint on convex-concave minimax optimization. This viewpoint makes simultaneous updates of x and y, based on evaluating both components of $\boldsymbol{F}$ at the same point, a natural choice. Yet, alternating updates have also been studied, most notably in *primal-dual hybrid gradient* methods (Chambolle & Pock, 2011; Zhu & Chan, 2008). More recent works have shown that alternating updates can be beneficial in various settings (Gidel et al., 2019a;b; Kim & Kim, 2024; Lee et al., 2024; Zhang et al., 2022a). Nonetheless, one should note that alternating updates alone do not guarantee convergence of the iterates in general convex-concave problems (Gidel et al., 2019a; He et al., 2014).

Negative momentum Popularized by Polyak (1964), momentum is a common algorithmic ingredient in designing optimization methods, typically with a positive momentum parameter. The benefits of negative momentum in minimax optimization were demonstrated by Gidel et al. (2019b), and have been further investigated in subsequent work (Feng et al., 2025; Lorraine et al., 2022; Zhang et al., 2021). However, only recently did Shugart et al. (2026) show that AltGDA-NeMo converges for general smooth convex-concave problems and achieves a worst-case optimal rate for strongly-convex-strongly-concave problems. Nonetheless, as briefly mentioned in Section 1, their analysis leaves open analytical convergence bounds over a broader range of stepsizes and momentum parameters, as well as the tightness of the convex-concave rates. We establish such bounds and matching lower bounds for the convex-concave best-iterate squared gradient norm, showing tight dependence on the stepsize, momentum, and iteration count within our parameter region.

Computer-assisted convergence proofs via semidefinite programming The *performance estimation problem (PEP)* (Drori & Teboulle, 2014; Taylor et al., 2017) framework provides a systematic way to formulate worst-case performance analyses of many first-order methods as SDPs. The PEP framework can also be adapted to search for Lyapunov potentials via SDP formulations (Suh et al., 2026; Taylor et al., 2018). Shugart et al. (2026) used this latter approach to obtain a symbolic convergence proof for AltGDA-NeMo at a single fixed (η, β) pair, while highlighting the joint search for algorithm parameters and Lyapunov potentials as a notorious challenge in PEP. That said, we extend these guarantees to all (η, β) satisfying $0 < \eta L \leq -\beta \leq \frac{1}{2}$, establishing parameter-dependent convergence rates in both convex-concave and strongly-convex-strongly-concave settings.

3 Preliminaries

3.1 Notations

Let $\mathcal{H}$ and $\mathcal{K}$ be real Hilbert spaces. For a (Fréchet) differentiable objective $f : \mathcal{H} \times \mathcal{K} \to \mathbb{R}$, we consider the minimax optimization problem (1) where we minimize with respect to $x \in \mathcal{H}$ and maximize with respect to $y \in \mathcal{K}$. Let ∇f denote the gradient, or more precisely the Riesz representation vector of the (Fréchet) derivative, of f.

For two collections of n vectors $\boldsymbol{w} = (w_1, \ldots, w_n)$ and $\boldsymbol{z} = (z_1, \ldots, z_n)$, all from the same Hilbert space, and an $n \times n$ real matrix $\boldsymbol{A} = (a_{ij})_{1 \leq i,j \leq n}$, we define $\langle \boldsymbol{w} \,|\, \boldsymbol{A} \,|\, \boldsymbol{z} \rangle \coloneqq \sum_{1 \leq i,j \leq n} a_{ij} \langle w_i, z_j \rangle$.

For $x \in \mathcal{H}$ and $y \in \mathcal{K}$, by (x, y) we mean an ordered pair in the Cartesian product $\mathcal{H} \times \mathcal{K}$, equipped with the canonical inner product $\langle (x, y), (x', y') \rangle \coloneqq \langle x, x' \rangle + \langle y, y' \rangle$ and the norm induced by it.

3.2 Problem setting

On the objective function f, we further impose the following assumptions.

Assumption 1 (Convex-concavity)**.** We say that the objective function f is *convex-concave (C-C)* if $f(\cdot, y)$ is convex for any $y \in \mathcal{K}$ and $f(x, \cdot)$ is concave for any $x \in \mathcal{H}$.

Moreover, for some fixed constant $\mu > 0$, if $f(\cdot, y)$ is μ-strongly convex for any $y \in \mathcal{K}$ and $f(x, \cdot)$ is μ-strongly concave for any $x \in \mathcal{H}$, we say that f is *μ-strongly-convex-strongly-concave (μ-SC-SC)*.

Assumption 2 (Smoothness)**.** There exists some fixed constant $L > 0$ such that f is *L-smooth*, *i.e.*, $\|\nabla f(x, y) - \nabla f(x', y')\| \leq L \|(x, y) - (x', y')\|$ holds for any $(x, y), (x', y') \in \mathcal{H} \times \mathcal{K}$.

Assumption 3 (Attainability)**.** There exists a point (x_*, y_*) which is an optimal solution of the problem (1), satisfying $f(x_*, y) \leq f(x_*, y_*) \leq f(x, y_*)$ for any $x \in \mathcal{H}$ and $y \in \mathcal{K}$.

By a slight abuse of notation, we may, and will, say that C-C functions are 0-SC-SC. That said, we will say that f is strongly-convex-strongly-concave (SC-SC) only if f is μ-SC-SC for some $\mu > 0$.

4 Main Convergence Results

Let us first state the convergence rates of AltGDA-NeMo we are to establish in this work.

Theorem 4.1. *Under Assumption 3, suppose that f is convex-concave and L-smooth. For any η and β such that $0 < \eta L \leq -\beta \leq \frac{1}{2}$, suppose we run AltGDA-NeMo with stepsize η and momentum β. Then, for any $T \geq 1$, it holds that*

$$\frac{1}{T} \sum_{k=0}^{T-1} \|\nabla f(x_k, y_k)\|^2 \leq \frac{375}{32|\beta|\eta^2 T} \left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2 \right)$$

Theorem 4.2. *Under Assumption 3, suppose that f is L-smooth and μ-strongly-convex-strongly-concave for some $\mu \in (0, L)$. For any η and β such that $0 < \eta L \leq -\beta \leq \frac{1}{2}$, say we run AltGDA-NeMo with stepsize η and momentum β. Then, for any $T \geq 1$, it holds that*

$$\|x_T - x_*\|^2 + \|y_T - y_*\|^2 \leq \frac{75}{17} \left(1 - \frac{3\eta\mu}{10} \right)^T \left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2 \right).$$

Notice that our results apply to all (η, β) such that $0 < \eta L \leq -\beta \leq \frac{1}{2}$. As our results are derived with this as the primary goal, they are not identical to the ones found by Shugart et al. (2026), even when $(\eta, \beta) = (\frac{1}{5L}, -\frac{1}{2})$, which is the specific choice fixed in their work.

Moreover, the upper bounds above are tight up to constant factors in the following senses. In the SC-SC case, Theorem 4.2 with an admissible stepsize choice of $\eta = \Theta(1/L)$ matches the known lower bound by Zhang et al. (2022b) for first-order methods in SC-SC problems; see Section F. For the C-C case, we provide a matching lower bound for AltGDA-NeMo as follows.

Theorem 4.3. *Let η, β, and L be given such that $0 < \eta L \leq -\beta \leq \frac{1}{2}$. Then, for any integer $T \geq 1$, there exists a convex-concave and L-smooth function f satisfying Assumption 3, where running AltGDA-NeMo with stepsize η and momentum β produces the iterates $\{(x_k, y_k)\}$ such that*

$$\min_{k=0,1,\dots,T-1} \|\nabla f(x_k, y_k)\|^2 \geq \frac{\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2}{83} \min\left\{L^2, \frac{1}{|\beta|\eta^2 T}\right\}.$$

5 Semidefinite Programming Approach for Convergence Proofs

In this section, we construct an SDP whose feasible solutions provide a Lyapunov potential that yields the convergence rates of AltGDA-NeMo stated in Section 4. The derivation of these rates can be found in Section E. Our construction uses quadratic Lyapunov functions and interpolation inequalities, as in Shugart et al. (2026), but targets a different decrease inequality with η and β left as parameters. We give the SDP formulation explicitly to support the subsequent parameter reductions and exact proof reconstruction.

Denoting the momenta by $u_k = x_k - x_{k-1}$ and $v_k = y_k - y_{k-1}$, we can rewrite (2) as

$$\begin{aligned} u_{k+1} &= \beta u_k - \eta \nabla_x f(x_k, y_k), & v_{k+1} &= \beta v_k + \eta \nabla_y f(x_{k+1}, y_k), \\ x_{k+1} &= x_k + u_{k+1}, & y_{k+1} &= y_k + v_{k+1}. \end{aligned} \tag{3}$$

Notice that the quantities at step $k+1$ are completely determined by those at step k, suggesting the state variables

$$\boldsymbol{\xi}_k \coloneqq \big(x_k - x_*,\ \eta \nabla_x f(x_k, y_k),\ u_k\big) \in \mathcal{H}^3, \quad \boldsymbol{\theta}_k \coloneqq \big(y_k - y_*,\ \eta \nabla_y f(x_k, y_k),\ v_k\big) \in \mathcal{K}^3.$$

Using these notations, we seek a suitable Lyapunov potential that could yield the desired convergence proofs. To make the search tractable, we restrict the search space to quadratics. That is, for positive semidefinite matrices $\boldsymbol{Q}_x$ and $\boldsymbol{Q}_y$, both independent of k, we wish the Lyapunov potentials to take the form of

$$\mathcal{E}_k \coloneqq \langle \boldsymbol{\xi}_k \,|\, \boldsymbol{Q}_x \,|\, \boldsymbol{\xi}_k \rangle + \langle \boldsymbol{\theta}_k \,|\, \boldsymbol{Q}_y \,|\, \boldsymbol{\theta}_k \rangle \tag{4}$$

where the notation $\langle \cdot \,|\, \cdot \,|\, \cdot \rangle$ is as introduced in Section 3.1.

As noted by Shugart et al. (2026, Remark 4.2), we may assume without loss of generality that $L = 1$, as we may consider the rescaled function $\frac{1}{L} f$, which is 1-smooth, in place of f.

By the so-called *interpolation condition* for convex functions (Taylor & Drori, 2023, Theorem 2; Upadhyaya et al., 2026, Proposition 1.3), f being μ-SC-SC for some $\mu \in [0, 1)$ implies that

$$\begin{aligned} M^{\mu}_{\mathrm{cvx},i,j,p} &\coloneqq f(x_i, y_p) - f(x_j, y_p) - \langle \nabla_x f(x_j, y_p), x_i - x_j \rangle - \frac{1}{2}\|\nabla_x f(x_i, y_p) - \nabla_x f(x_j, y_p)\|^2 \\ &\qquad - \frac{\mu}{2(1-\mu)}\|x_i - x_j - (\nabla_x f(x_i, y_p) - \nabla_x f(x_j, y_p))\|^2, \\ M^{\mu}_{\mathrm{ccv},i,j,p} &\coloneqq f(x_p, y_j) - f(x_p, y_i) + \langle \nabla_y f(x_p, y_j), y_i - y_j \rangle - \frac{1}{2}\|\nabla_y f(x_p, y_i) - \nabla_y f(x_p, y_j)\|^2 \\ &\qquad - \frac{\mu}{2(1-\mu)}\|y_i - y_j + (\nabla_y f(x_p, y_i) - \nabla_y f(x_p, y_j))\|^2, \end{aligned} \tag{5}$$

for any $i, j, p \in \{*, 0, 1, \dots\}$, are all nonnegative. Also, the 1-smoothness of f implies that

$$M^{\mu}_{\mathrm{smth},i,j,p,q} \coloneqq \|x_i - x_p\|^2 + \|y_j - y_q\|^2 - \|\nabla f(x_i, y_j) - \nabla f(x_p, y_q)\|^2 \tag{6}$$

is also nonnegative for any $i, j, p, q \in \{*, 0, 1, \dots\}$. Although μ is not involved here, we include the superscript μ for consistency with (5).

That said, without any assumptions on f imposed yet, let us consider the ansatz that there exist multipliers $\boldsymbol{\lambda} \succeq \mathbf{0}$, independent of k, such that the inequality

$$S_k \coloneqq \mathcal{E}_k - \mathcal{E}_{k+1} + c\beta\eta^2 \|\nabla f(x_k, y_k)\|^2 - \sum_\alpha \lambda_\alpha M^\mu_{k,\alpha} \geq 0 \tag{7}$$

holds for all $k = 0, 1, \ldots$, where $c > 0$ is some fixed constant and $M_{k,\alpha}$ are the quantities defined in (5) and (6), reindexed by α so that all possible combinations of $i, j, p, q \in \{k, k+1, *\}$ would occur in the sum. Then, under the assumption that f is μ-SC-SC, we have $M^\mu_{k,\alpha} \geq 0$ for all α, so it would follow from (7) that

$$c|\beta|\eta^2 \|\nabla f(x_k, y_k)\|^2 \leq \mathcal{E}_k - \mathcal{E}_{k+1}, \tag{8}$$

from which a telescopic sum with appropriate bounds on $\mathcal{E}_k$ would yield the convergence rates as we detail in Section E. The choice of c is simply a matter of rescaling, as dividing each of $\boldsymbol{Q}_x$, $\boldsymbol{Q}_y$, and $\boldsymbol{\lambda}$ by c gives the normalized case $c = 1$. On the other hand, the factor $-\beta\eta^2$ is part of the ansatz, based on observations for bilinear f that eventually led to Theorem F.1.

Observing the right hand side of (7), one can notice that the expression defining S_{k+1} is identical to that of S_k, up to shifting the index from k to $k+1$. Thus, without loss of generality, it suffices to show that there exist $\boldsymbol{Q}_x$, $\boldsymbol{Q}_y$, and $\boldsymbol{\lambda}$, all independent of k, that satisfy (7) when $k = 0$. To this end, let us drop the subscript k from $M_{k,\alpha}$ for simplicity of notation, set

$$S \coloneqq \mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2 \|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M^\mu_\alpha, \tag{9}$$

and aim for a guarantee that $S \geq 0$. We remark that, although the initialization sets $u_0 = 0$ and $v_0 = 0$, we aim for the inequality $S \geq 0$ without assuming anything about u_0 and v_0, so there is indeed no loss of generality.

Fix any $\eta > 0$, $\beta < 0$, and $\mu \in [0, 1)$. By expanding the expression for S, one can observe that S is a quadratic of the $\mathcal{H}$-vectors $x_i - x_*$ and $\nabla_x f(x_i, y_j)$, $\mathcal{K}$-vectors $y_j - y_*$ and $\nabla_y f(x_i, y_j)$, and function values $f_{ij} \coloneqq f(x_i, y_j)$, where $i, j \in \{0, 1, *\}$. That is, with the details deferred to Section A for simplicity of the demonstration, denoting the collection of $\mathcal{H}$-vectors by $\boldsymbol{\chi}$, that of $\mathcal{K}$-vectors by $\boldsymbol{\psi}$, and that of function values by $\boldsymbol{\phi}$, one can conclude that there exist symmetric matrices $\boldsymbol{P}_x, \boldsymbol{P}_y$ and a vector $\boldsymbol{q}$, all affine in $\boldsymbol{Q}_x$, $\boldsymbol{Q}_y$, and $\boldsymbol{\lambda}$, such that

$$S = \langle \boldsymbol{\chi} \,|\, \boldsymbol{P}_x \,|\, \boldsymbol{\chi} \rangle + \langle \boldsymbol{\psi} \,|\, \boldsymbol{P}_y \,|\, \boldsymbol{\psi} \rangle + \langle \boldsymbol{q}, \boldsymbol{\phi} \rangle .$$

To have a guarantee that $S \geq 0$, it suffices to have $\boldsymbol{P}_x \succeq \mathbf{0}$, $\boldsymbol{P}_y \succeq \mathbf{0}$, and $\boldsymbol{q} = \mathbf{0}$. Thus, our goal can be reduced to finding a solution to a feasibility problem

$$\begin{aligned} \min_{\boldsymbol{Q}_x, \boldsymbol{Q}_y, \boldsymbol{\lambda}} \quad & 0 \\ \text{s.t.} \quad & \boldsymbol{Q}_x \succeq \mathbf{0}, \boldsymbol{Q}_y \succeq \mathbf{0}, \boldsymbol{\lambda} \succeq \mathbf{0}, \\ & \boldsymbol{P}_x \succeq \mathbf{0}, \boldsymbol{P}_y \succeq \mathbf{0}, \boldsymbol{q} = \mathbf{0}. \end{aligned} \tag{10}$$

6 Parameter Reduction Techniques

In principle, solving (10) will yield a convergence proof of AltGDA-NeMo. However, the issue here is that (10) is a *parametric* SDP which involves multiple parameters, namely η, β, and μ; solving it poses substantial challenges. Also, even if we fix certain values of (η, β, μ), the only tractable approach to solving (10) is via numerical methods. A numerical solution provides strong evidence that a convergence proof is possible, but it does not itself yield a fully rigorous proof. We address the first issue here and the second in Section 7.

6.1 Convex-concave to strongly-convex-strongly-concave conversion

The discussion in Section 5 applies both when f is C-C and when it is SC-SC. Indeed, the descent inequality (8) yields the C-C bound in Theorem 4.1, as detailed in Section E.1. However, it does not exploit the additional SC-SC structure, which is expected to allow much faster convergence. For SC-SC problems, we refine this inequality, so that it eventually leads to obtaining linear convergence.

Let us introduce some notations. For each α, let K_α be the perfect square term that satisfies $M_\alpha^\mu = M_\alpha^0 - \frac{\mu}{2(1-\mu)} K_\alpha$. Notice that each K_α only involves the iterates and the (partial) gradients of f, and moreover $K_\alpha = 0$ when α is the index associated with the smoothness quantity (6). Then, the following proposition decouples μ from (10), leaving two SDPs involving only η and β.

Proposition 6.1. *Suppose that we have the solution $\boldsymbol{Q}_x$, $\boldsymbol{Q}_y$, and $\boldsymbol{\lambda}$ of the SDP* (10) *when $\mu = 0$, so that the inequality*

$$\mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2 \|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M_\alpha^0 \geq 0$$

holds unconditionally. For that solution of the SDP, if there exists a scalar $\tau \geq 0$ which satisfies

$$R \coloneqq (1-\tau\eta)\mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2 \|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M_\alpha^0 + \frac{1}{2}\sum_\alpha \lambda_\alpha K_\alpha \geq 0, \tag{11}$$

then for any $\mu \in [0, 1)$ it also holds that

$$(1-\tau\eta\mu)\mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2 \|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M_\alpha^\mu \geq 0. \tag{12}$$

Notice that (12) is a *stronger* version of (9). In particular, (12) with $\mu > 0$ and $\tau > 0$ gives geometric decay of $\mathcal{E}_k$. Technically, any $\tau > 0$ feasible for (11) would suffice. Yet, the larger τ is, the faster the decay, so we seek to maximize τ from the perspective of searching for better convergence rates.

As shown in Lemma B.2, the condition (11) can be reduced to linear matrix inequalities affine in τ, so maximizing τ subject to these constraints is an SDP. Moreover, this SDP does not involve μ. Thus, it indeed remains to solve two SDPs that do not involve μ; we first solve the feasibility SDP (10) with $\mu = 0$, and then maximize τ for the resulting certificate. See Section D for more details.

In Section E.2, we establish the bounds relating the potential $\mathcal{E}_k$ to the squared distance from an optimal solution, which in turn results in the convergence rates claimed in Theorem 4.2.

6.2 Eliminating the stepsize

Here, we would like to observe that a trick similar to the one used to eliminate the smoothness constant L can also eliminate η from the SDP (10). More precisely, for any $\epsilon \in (0, 1]$, it turns out that considering ϵf in place of f makes the following hold.

Proposition 6.2. *Fix any $\epsilon \in (0, 1]$. Then, with the same initialization, AltGDA-NeMo applied to f with parameters $(\epsilon\eta, \beta)$ generates the same iterates as AltGDA-NeMo applied to ϵf with parameters (η, β).*

Notice that scaling by a positive constant preserves convex-concavity and saddle points. Moreover, if f is 1-smooth and μ-SC-SC, then ϵf is 1-smooth and $\epsilon\mu$-SC-SC. Thus, Proposition 6.2 implies that, for a function $H(\beta)$ depending on β, if we show that AltGDA-NeMo converges with stepsize $H(\beta)$ and momentum β, we automatically get that AltGDA-NeMo with momentum β also converges under any choice of the stepsize $0 < \eta \leq H(\beta)$. As we have mentioned in Section 1, our focus is on the simplest case $H(\beta) = -\beta$.

Hence, without loss of generality, it suffices to consider $\eta = -\beta$, eliminating η as an independent parameter. We nevertheless retain both η and β in the notation for clarity.

7 Reconstructing the Convergence Proof Algebraically

7.1 Solving one-parameter SDPs

Following the reductions in Section 6, it suffices to set $L = 1$ and $\eta = -\beta$. It remains to solve the SDP (10) with $\mu = 0$, and then the SDP induced by (11). Both SDPs have a single remaining parameter, which, using the fact that $\eta = -\beta$, we take to be η. Let us elaborate on how we approach (10), as the same steps also apply to the SDP induced from (11).

To emphasize the dependency of $\boldsymbol{P}_x$ on the parameter η, let us write $\boldsymbol{P}_x(\eta)$, and also similarly write $\boldsymbol{P}_y(\eta)$ and $\boldsymbol{q}(\eta)$. It is natural to expect a feasible solution which is also dependent on η, so let us

accordingly write its components as $\boldsymbol{Q}_x(\eta)$, $\boldsymbol{Q}_y(\eta)$, and $\boldsymbol{\lambda}(\eta)$. Inspired by Dong et al. (2025), we parameterize them as polynomials in η. The required degree is unknown a priori, but increasing it enlarges the search space. For our convergence bounds, it turns out that the degree 1 ansätze

$$\boldsymbol{Q}_x(\eta) = \boldsymbol{Q}_{x,0} + \eta \boldsymbol{Q}_{x,1}, \qquad \boldsymbol{Q}_y(\eta) = \boldsymbol{Q}_{y,0} + \eta \boldsymbol{Q}_{y,1}, \qquad \boldsymbol{\lambda}(\eta) = \eta \boldsymbol{\lambda}_1 \tag{13}$$

suffice, where all the coefficient matrices and the vector $\boldsymbol{\lambda}_1$ are independent of η.

Under (13), as we are to solve a one-parameter SDP, the discretization technique (Hettich & Kortanek, 1993; Still, 2001) provides a tractable approach. More precisely, we choose points $\eta_1, \ldots, \eta_n$ from the interval $[0, \frac{1}{2}]$, and replace the η-dependent SDP constraints by the constraints imposed at each of $\eta_1, \ldots, \eta_n$ with the same unknown $\boldsymbol{Q}_{x,0}$, $\boldsymbol{Q}_{x,1}$, $\boldsymbol{Q}_{y,0}$, $\boldsymbol{Q}_{y,1}$, and $\boldsymbol{\lambda}_1$. This results in solving

$$\begin{aligned} \min_{\boldsymbol{Q}_{x,0}, \boldsymbol{Q}_{x,1}, \boldsymbol{Q}_{y,0}, \boldsymbol{Q}_{y,1}, \boldsymbol{\lambda}_1} \quad & 0 \\ \text{s.t.} \quad & \boldsymbol{Q}_{x,0} \succeq \boldsymbol{0}, \;\; \boldsymbol{Q}_{y,0} \succeq \boldsymbol{0}, \\ & \boldsymbol{Q}_{x,0} + \frac{1}{2}\boldsymbol{Q}_{x,1} \succeq \boldsymbol{0}, \;\; \boldsymbol{Q}_{y,0} + \frac{1}{2}\boldsymbol{Q}_{y,1} \succeq \boldsymbol{0}, \;\; \boldsymbol{\lambda}_1 \succeq \boldsymbol{0}, \\ & \boldsymbol{P}_x(\eta_i) \succeq \boldsymbol{0}, \;\; \boldsymbol{P}_y(\eta_i) \succeq \boldsymbol{0}, \;\; \boldsymbol{q}(\eta_i) = \boldsymbol{0} \qquad \text{for all } i = 1, \ldots, n \end{aligned} \tag{14}$$

as a proxy of the parametric SDP we wish to ultimately solve. Since $\boldsymbol{Q}_x(\eta)$ and $\boldsymbol{Q}_y(\eta)$ are affine in η and the positive semidefinite cone is convex, we note that the endpoint constraints ensure their positive semidefiniteness throughout the interval $[0, \frac{1}{2}]$. Also, because $\boldsymbol{\lambda}(\eta) = \eta\boldsymbol{\lambda}_1$ is nonnegative throughout $[0, \frac{1}{2}]$ if and only if $\boldsymbol{\lambda}_1 \succeq \boldsymbol{0}$, it suffices to impose only the latter.

7.2 Simplifying the solution

There are 72 quantities M_α in (9), so $\boldsymbol{\lambda} \in \mathbb{R}^{72}$. However, familiar convergence proofs for other well-known methods suggest that fewer inequalities may suffice. Thus, it is natural to aim for a sparse solution; see Thomsen et al. (2026) for a recent discourse on this. To this end, in solving (14), we add an ℓ_1 regularization term on $\boldsymbol{\lambda}_1$, following the common heuristic of promoting solution sparsity. Also, aiming to stabilize the numerical search, we add small ℓ_2 regularization terms for $\boldsymbol{Q}_{x,0}$, $\boldsymbol{Q}_{x,1}$, $\boldsymbol{Q}_{y,0}$, and $\boldsymbol{Q}_{y,1}$. These terms modify only the objective, leaving the feasibility constraints unchanged.

The numerical solution suggests that just 10 of the 72 inequalities may suffice; the retained multipliers are listed in Section C.3. We then re-solve the SDP with all other multipliers fixed to zero and obtain a numerical solution for the reduced problem.

As we soon discuss in Section 7.3, constructing an exact algebraic convergence proof involves rationalizing the solution of the SDP. However, rationalizing the numerical solution is essentially a perturbation, so if the solution found is close to the boundary of the feasible set, rationalized solutions may be infeasible. Hence, to leave room for such perturbations, we introduce additional techniques that can increase the feasibility margins. However, as those techniques are largely specific to our SDP, we defer the details to Section C.1.

7.3 Rationalization for an algebraic proof

To obtain an algebraically verifiable convergence proof, we rationalize the numerically computed entries of the coefficient matrices and vector in the degree-one ansätze (13), replacing floating-point coefficients with nearby rational values, as in Peyrl & Parrilo (2008). Once a rationalization is obtained, exact arithmetic and symbolic computation can be used to check whether the rationalized candidate satisfies (10) throughout the parameter interval.

However, rounding the coefficients all independently can violate feasibility. To remedy this issue, we use an integer programming approach. With the details deferred to Section C.2, we sketch the main ideas. We first identify the identities that must be preserved. Straightforwardly, $\boldsymbol{q} = \boldsymbol{0}$ gives such identities. Some identities also arise from positive semidefiniteness conditions. For instance, if a diagonal entry of a positive semidefinite matrix is identically zero, every entry in its row and column must also vanish. Equating coefficients in these identities yields linear relations among the coefficients to be rationalized.

We group coefficients coupled by these relations. Then, for each group, we rationalize the coefficients simultaneously. More precisely, for each candidate common denominator d, we write the coefficients within a group as p_j/d with integer numerators p_j. Then, the relations become linear constraints on the numerators. We solve a mixed integer linear program subject to these constraints, minimizing the sum of absolute deviations of the rational coefficients from the numerical solution. For each group, we retain the closest approximation over the candidate denominators.

This construction preserves the extracted identities. We then verify the remaining nonnegativity and positive semidefiniteness conditions exactly throughout $\eta \in (0, \frac{1}{2}]$. We also verify in exact arithmetic that $\tau = \frac{3}{10}$, obtained by solving the SDP induced by (11), satisfies (11) throughout $\eta \in (0, \frac{1}{2}]$, as detailed in Section D. By Proposition 6.1, this yields the contraction $\mathcal{E}_{k+1} \leq (1 - \frac{3}{10}\eta\mu)\mathcal{E}_k$ for SC-SC objectives. Section C gives the full reconstruction procedure and exact certificates, and Section E gives the resulting convergence proofs.

8 Negative Momentum with Practical Optimizers

The Adam optimizer (Kingma & Ba, 2015) is widely used in modern machine learning. The main feature of Adam is its deftly designed momentum tracking mechanism. Our convergence results for AltGDA-NeMo raise the question of whether negative momentum can be incorporated into Adam.

We show that it is possible, to a certain extent. Consider Adam with alternating updates for the minimax problem (1). To isolate the role of β_1, we temporarily omit other components, including second-moment tracking and bias correction. The minimization update is then

$$\mathtt{u}_{k+1} = \beta_1 \mathtt{u}_k + (1-\beta_1)\mathtt{g}_{x,k} \tag{15a}$$

$$\mathtt{x}_{k+1} = \mathtt{x}_k - \alpha \mathtt{u}_{k+1} \tag{15b}$$

where $\alpha > 0$ is the learning rate and $\mathtt{g}_{x,k}$ is an estimator of $\nabla_x f(\mathtt{x}_k, \mathtt{y}_k)$. Noting that the previous update step of (15b) implies that $\alpha \mathtt{u}_k = \mathtt{x}_{k-1} - \mathtt{x}_k$, one can substitute (15a) into (15b) to obtain

$$\begin{aligned} \mathtt{x}_{k+1} &= \mathtt{x}_k - \alpha\big(\beta_1 \mathtt{u}_k + (1-\beta_1)\mathtt{g}_{x,k}\big) \\ &= \mathtt{x}_k - \alpha(1-\beta_1)\mathtt{g}_{x,k} + \beta_1(\mathtt{x}_k - \mathtt{x}_{k-1}). \end{aligned} \tag{16}$$

Through an analogous derivation, one finds that the update performed by the Adam optimizer for maximization is equivalent to

$$\mathtt{y}_{k+1} = \mathtt{y}_k + \alpha(1-\beta_1)\mathtt{g}_{y,k} + \beta_1(\mathtt{y}_k - \mathtt{y}_{k-1}) \tag{17}$$

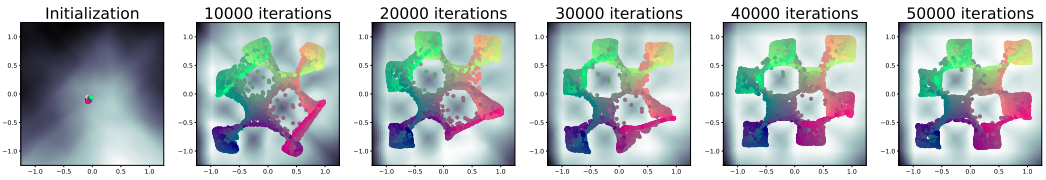
where $\mathtt{g}_{y,k}$ is now an estimator of $\nabla_y f(\mathtt{x}_{k+1}, \mathtt{y}_k)$.

Comparing (16) and (17) to the AltGDA-NeMo updates (2), we assert that one can view AltGDA-NeMo as a simplified version of "alternating Adam" with $\beta_1 < 0$, albeit with the "effective" learning rate $\alpha(1-\beta_1)$. This motivates us to consider *alternating Adam with negative momentum (AltAdam-NeMo)*, where two Adam optimizers, one for minimization and the other for maximization, perform alternating updates with momentum parameters (β_1, β_2) satisfying $\beta_1 < 0$.

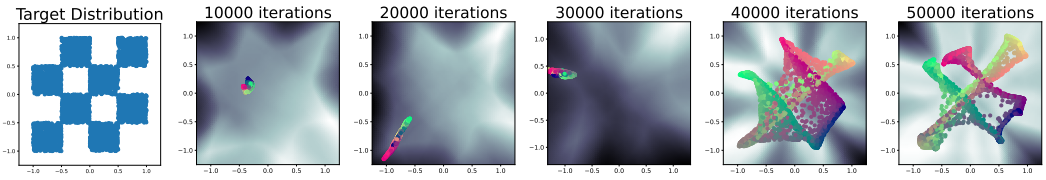
We remark that similar approaches, such as *ExtraAdam* (Gidel et al., 2019a) and *Optimistic Adam* (Daskalakis et al., 2018), exist, where minimax optimization methods convergent for C-C problems, in this case extragradient and optimistic gradient, respectively, are modified to compute the update direction in the style of Adam. Nonetheless, AltAdam-NeMo has an advantage over those methods in ease of implementation. Unlike ExtraAdam and Optimistic Adam which require substantial modifications to the base Adam optimizer, AltAdam-NeMo can be implemented by simply constructing two Adam optimizers using existing implementations as is, one for minimization variables and the other for maximization variables, but with $\beta_1 < 0$.[1]

In addition, we believe that the momentum mechanism à la (15a) is a common algorithmic ingredient which can be found across numerous optimizers, in particular in modern machine learning. Hence, while we focused on Adam in this section to avoid distractions, a similar manipulation can be applied, which will then reveal that the negative momentum trick discussed in this section is quite universal rather than specific to Adam.

[1] Yet, in the current version (`2.14`) of `PyTorch`, simply passing a negative value to β_1 will raise a `ValueError`. Still, this can be resolved by modifying just one line in the implementation of the Adam optimizer so that the asserted condition becomes $-1 < \beta_1 < 1$.



(a) AltAdam-NeMo with $(\beta_1, \beta_2) = (-0.5, 0.9)$



(b) Ordinary alternating Adam with $(\beta_1, \beta_2) = (0.0, 0.9)$

Figure 1: Training of GANs on the checkerboard dataset.

9 Experiments

Let us compare plain alternating Adam, which is the standard optimizer used in training neural networks when the task is formulated as a minimax optimization problem, with AltAdam-NeMo. The empirical investigations in this work aim to demonstrate that changing only the first-moment parameter β_1 to a negative value can substantially improve training. We consider generative adversarial networks (GANs) and neural optimal transport on both toy and real-world datasets. Accordingly, we hold the remaining training configuration fixed for both methods, comparing the commonly used standard choice $(\beta_1, \beta_2) = (0.0, 0.9)$ with $(-0.5, 0.9)$. With details of the training/plotting procedure and further discussions deferred to Section G, let us summarize.

For GANs, we revisit the most basic form (Goodfellow et al., 2014), which is known to be especially challenging to train and to suffer from *mode collapse* (Goodfellow, 2016). Figure 1 shows the results of training on the checkerboard toy dataset on a 4×4 grid. As demonstrated visually, changing only β_1 from 0 to -0.5 yields a striking improvement. AltAdam-NeMo finds all eight modes early in training and gradually recovers their shapes, nearly reproducing the squares at 50k iterations. Ordinary alternating Adam still exhibits mode collapse at 30k iterations and only roughly covers all eight modes at 50k iterations.

We also train the optimal transport plan (OTP) model of Choi et al. (2025b) on toy datasets and image-to-image translation tasks. We observe that AltAdam-NeMo trains more robustly, avoiding mode collapse and achieving better quantitative results. The experimental settings and detailed results can be found in Section G.2.

10 Conclusion

We established convergence guarantees for AltGDA-NeMo over a broader range of stepsizes and negative momentum parameters, including momentum arbitrarily close to zero with sufficiently small stepsizes. For convex-concave problems, matching upper and lower bounds establish tight dependence of convergence to stationarity on the stepsize, momentum, and iteration count. For strongly-convex-strongly-concave problems, we proved linear convergence, with worst-case optimal rates under an appropriate parameter choice. To construct Lyapunov certificates throughout the parameter region, we developed reductions that simplify the search over stepsize, momentum, and strong-convexity parameters to two successive one-parameter SDPs. Our toy experiment also illustrates the benefits of negative momentum.

AI USE STATEMENT

In this work, we used generative AI tools for getting help in formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and proposing or refining hypotheses.

We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, and interpreting results.

Assisting with translation, cleaning and reformatting dataset, and supporting qualitative and thematic data analysis are not applicable to this work.

Additionally, we used generative AI tools for drafting parts of a research paper, summarizing or analyzing existing literature, editing a research paper to improve readability, and identifying relevant literature.

We have reviewed all AI-assisted work. Let us elaborate on how AI has assisted in developing our work and how we have reviewed its contributions.

First, we discuss formulating mathematical claims, providing critical ingredients for proving mathematical claims, and proposing or refining hypotheses. In constructing the ansatz (7), the main difficulty was identifying that $-\beta\eta^2$, which is $\Theta(\eta^3)$ under our reduction $\eta = -\beta$, should be in the coefficient of the gradient norm squared term. Both the authors and AI initially (incorrectly) conjectured that the coefficient would be $\Theta(\eta^2)$, but of course this led to the construction of an infeasible SDP. By trial and error, the authors have found out that a $\Theta(\eta^3)$ coefficient gives an SDP with a feasible solution, and made a corrected conjecture that the coefficient should be a multiple of $\beta\eta^2$. To verify this, knowing that if f is bilinear then the update rule of AltGDA-NeMo reduces to a linear recurrence, the authors asked AI to work out the recurrence and possibly a closed-form expression for the iterates. The AI successfully completed this request, confirming the $\beta\eta^2$ scaling in the bilinear case.

Later, in proving the lower bound result (Theorem 4.3), based on prior work (Golowich et al., 2020a;b) which demonstrated that the lower bound instances of the extragradient and the optimistic gradient methods come from bilinear objectives, the authors conjectured that the previous bilinear computation would also yield a lower bound instance of AltGDA-NeMo in the convex-concave regime. The authors thus asked the AI to identify the similarities between the previous bilinear computations and the existing prior work by Golowich et al. (2020a;b), providing the context that the authors were attempting to construct a lower bound instance. The AI, provided with this context, autonomously came up with the Lyapunov potential (47), leaving only simple routine computations. The authors were indeed able to carry out the remaining computations that lead to Theorem 4.3.

All the proofs are written by the authors. We nonetheless hereby declare that AI assisted in writing the proofs, because AI did help us in finding typos and grammatical mistakes, and accordingly revising the wording. Still, actually typing in those revisions was manually done by the authors.

AI was also used in identifying, summarizing, and analyzing existing literature. The authors reviewed the sources found by the AI and decided which ones to cite after reading them. Also, AI was used to draft some parts of the paper, such as the conclusion section.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The `Julia` code we used to solve the SDP and construct an algebraic proof from it, as detailed in Sections 5–7 and C, is provided as supplementary material along with the submission. The theoretical results are provided with full proofs in the appendices, with appropriate assumptions included in each of the statements. For the experiments in Sections 9 and G, we provide detailed descriptions of the experimental settings and procedures in Section G to support reproducibility. If further details are needed, we plan to provide the code used for these experiments upon request.

REFERENCES

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

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Generative modeling through the semi-dual formulation of unbalanced optimal transport. *Advances in Neural Information Processing Systems*, 36, 2023.

Jaemoo Choi, Yongxin Chen, and Jaewoong Choi. Improving neural optimal transport via displacement interpolation. In *International Conference on Learning Representations*, 2025a.

Jaemoo Choi, Jaewoong Choi, and Dohyun Kwon. Overcoming spurious solutions in semi-dual neural optimal transport: A smoothing approach for learning the optimal transport plan. In *International Conference on Machine Learning*, 2025b.

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

Qiran Dong, Paul Grigas, and Vishal Gupta. Beyond discretization: Learning the optimal solution path. In *International Conference on Artificial Intelligence and Statistics*, pp. 4528–4536. PMLR, 2025.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014.

Jiaojiao Fan, Shu Liu, Shaojun Ma, Yongxin Chen, and Hao-Min Zhou. Scalable computation of monge maps with general costs. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

Yi Feng, Kaito Fujii, Stratis Skoulakis, Xiao Wang, and Volkan Cevher. Continuous-time analysis of heavy ball momentum in min-max games. In *International Conference on Machine Learning*, 2025.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019a.

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

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

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

Ian Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Paul J. Goulart and Yuwen Chen. Clarabel: An interior-point solver for conic programs with quadratic objectives. *arXiv preprint arXiv:2405.12762*, 2024.

Shashi Gowda, Yingbo Ma, Alessandro Cheli, Maja Gwóźdź, Viral B. Shah, Alan Edelman, and Christopher Rackauckas. High-performance symbolic-numerics via multiple dispatch. *ACM Communications in Computer Algebra*, 55(3):92–96, 2021.

Bingsheng He, Yanfei You, and Xiaoming Yuan. On the convergence of primal-dual hybrid gradient algorithm. *SIAM Journal on Imaging Sciences*, 7(4):2526–2537, 2014.

Rainer Hettich and Kenneth O. Kortanek. Semi-infinite programming: theory, methods, and applications. *SIAM review*, 35(3):380–429, 1993.

Qi Huangfu and J. A. Julian Hall. Parallelizing the dual revised simplex method. *Mathematical programming computation*, 10(1):119–142, 2018.

Kyuwon Kim and Donghwan Kim. Double-step alternating extragradient with increasing timescale separation for finding local minimax points: Provable improvements. In *International Conference on Machine Learning*, 2024.

Diederik P. Kingma and Jimmy Lei Ba. A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

Jaewook Lee, Hanseul Cho, and Chulhee Yun. Fundamental benefit of alternating updates in minimax optimization. In *International Conference on Machine Learning*, 2024.

Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 34, 2021.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *International Conference on Machine Learning*, 1994.

Jonathan P Lorraine, David Acuna, Paul Vicol, and David Duvenaud. Complex momentum for optimization in games. In *International Conference on Artificial Intelligence and Statistics*, pp. 7742–7765. PMLR, 2022.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Bernard Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Revue Française d'informatique et de Recherche opérationelle*, 4:154–158, 1970.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *International Conference on Machine Learning*, 2018.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Helfried Peyrl and Pablo A. Parrilo. Computing sum of squares decompositions with rational coefficients. *Theoretical Computer Science*, 409(2):269–281, 2008.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International conference on machine learning*, 2017.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 4(5):791–803, 1964.

Leonid D. Popov. A modification of the Arrow-Hurwitz method of search for saddle points. *Matematicheskie Zametki*, 28(5):777–784, 1980.

Franz Rellich. *Perturbation theory of eigenvalue problems*. Gordon and Breach Science Publishers, 1969.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022.

Henry Shugart, Shuyi Wang, and Jason M. Altschuler. Negative momentum for convex-concave optimization. *arXiv preprint arXiv:2604.17145*, 2026.

Motahareh Sohrabi, Jianxin You, Simon Lacoste-Julien, Eduard Gorbunov, and Gauthier Gidel. Accelerated and stable convergence with anchored generalized optimistic method. In *International Conference on Machine Learning*, 2026.

Georg Still. Discretization in semi-infinite programming: the rate of convergence. *Mathematical programming*, 91(1):53–69, 2001.

Jaewook J. Suh, TaeHo Yoon, Edward Duc Hien Nguyen, Bicheng Ying, and Shiqian Ma. Peppy: An AI-assisted workflow for tight convergence analysis of optimization algorithms. In *ICML 2026 Workshop: AI as a Tool for Mathematics, Computer Science, and Machine Learning*, 2026.

Adrien Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, 199(1):557–594, 2023.

Adrien Taylor, Bryan Van Scoy, and Laurent Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *International Conference on Machine Learning*, pp. 4897–4906. PMLR, 2018.

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

Daniel Berg Thomsen, Manu Upadhyaya, Baptiste Goujaud, Aymeric Dieuleveut, and Adrien Taylor. Finding simple proofs for first-order optimization. In *ICML 2026 Workshop: AI as a Tool for Mathematics, Computer Science, and Machine Learning*, 2026.

Manu Upadhyaya, Daniel Berg Thomsen, Aymeric Dieuleveut, and Adrien B. Taylor. An optimal first-order method for smooth and strongly convex composite optimization and its stationary limit. *arXiv preprint arXiv:2605.22929*, 2026.

TaeHo Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning*, 2021.

Guodong Zhang, Xuchan Bao, Laurent Lessard, and Roger Grosse. A unified analysis of first-order methods for smooth games via integral quadratic constraints. *Journal of Machine Learning Research*, 22(103):1–39, 2021.

Guodong Zhang, Yuanhao Wang, Laurent Lessard, and Roger Grosse. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 7659–7679. PMLR, 2022a.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1):901–935, 2022b.

Mingqiang Zhu and Tony Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report*, 34(2), 2008.

A Further Details on Deriving the Feasibility SDP

Let us begin by making explicit the following property of the $\langle \cdot \,|\, \cdot \,|\, \cdot \rangle$ notation, which has been implicitly used in the previous sections.

Proposition A.1. *Suppose that $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a positive semidefinite matrix. Then, for any collection of n vectors, $\boldsymbol{z} = (z_1, \dots, z_n)$, all from the same Hilbert space, it holds that $\langle \boldsymbol{z} \,|\, \boldsymbol{A} \,|\, \boldsymbol{z} \rangle \geq 0$.*

Proof. As $\boldsymbol{A}$ is positive semidefinite, it admits a Cholesky decomposition $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$. Let us denote the (i, j) entry of $\boldsymbol{L}$ by l_{ij}. Then, the claimed inequality follows from the fact that

$$\begin{aligned}
\langle \boldsymbol{z} \,|\, \boldsymbol{A} \,|\, \boldsymbol{z} \rangle &= \left\langle \boldsymbol{z} \,\middle|\, \boldsymbol{L}\boldsymbol{L}^\top \,\middle|\, \boldsymbol{z} \right\rangle \\
&= \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} l_{ik} l_{jk} \right) \langle z_i, z_j \rangle \\
&= \sum_{k=1}^{n} \left\langle \sum_{i=1}^{n} l_{ik} z_i, \sum_{j=1}^{n} l_{jk} z_j \right\rangle \\
&= \sum_{k=1}^{n} \left\| \sum_{i=1}^{n} l_{ik} z_i \right\|^2 \\
&\geq 0
\end{aligned}$$

where the third line follows from swapping the order of summation and then using the bilinearity of inner products. □

For any $i, j \in \{0, 1, *\}$, let us write

$$r_i \coloneqq x_i - x_*, \quad s_j \coloneqq y_j - y_*, \quad f_{ij} \coloneqq f(x_i, y_j), \quad g_{ij} \coloneqq \nabla_x f(x_i, y_j), \quad h_{ij} \coloneqq \nabla_y f(x_i, y_j).$$

Notice that the optimality of (x_*, y_*) asserts that $g_{**} = h_{**} = 0$. Letting $\boldsymbol{\omega} = (\boldsymbol{\chi}, \boldsymbol{\psi}, \boldsymbol{\phi})$ where

$$\boldsymbol{\chi} \coloneqq (r_0, u_0, g_{00}, g_{01}, g_{0*}, g_{10}, g_{11}, g_{1*}, g_{*0}, g_{*1}), \tag{18a}$$

$$\boldsymbol{\psi} \coloneqq (s_0, v_0, h_{00}, h_{01}, h_{0*}, h_{10}, h_{11}, h_{1*}, h_{*0}, h_{*1}), \tag{18b}$$

$$\boldsymbol{\phi} \coloneqq (f_{00}, f_{01}, f_{0*}, f_{10}, f_{11}, f_{1*}, f_{*0}, f_{*1}, f_{**}), \tag{18c}$$

it is clear that $\boldsymbol{\xi}_0$ and $\boldsymbol{\theta}_0$ are linear on $\boldsymbol{\omega}$, and (3) asserts that $\boldsymbol{\xi}_1$ and $\boldsymbol{\theta}_1$ are linear in $\boldsymbol{\omega}$. Hence, under the quadratic Lyapunov assumption (4), one can observe that $\mathcal{E}_0 - \mathcal{E}_1$ is a quadratic of $\boldsymbol{\omega}$, in the sense that it is a linear combination of the twenty-nine quantities in (18) that comprise $\boldsymbol{\omega}$ and their inner products. Moreover, from the fact that $\|\nabla f(x_0, y_0)\|^2 = \langle g_{00}, g_{00} \rangle + \langle h_{00}, h_{00} \rangle$ and the definitions (5) and (6) of each M_α, we conclude that those are also quadratics of $\boldsymbol{\omega}$ in the same sense. Hence, S is also quadratic on $\boldsymbol{\omega}$, and hence, there exist symmetric matrices $\boldsymbol{P}_x, \boldsymbol{P}_y \in \mathbb{R}^{10\times 10}$ and a vector $\boldsymbol{q} \in \mathbb{R}^9$ such that

$$S = \langle \boldsymbol{\chi} \,|\, \boldsymbol{P}_x \,|\, \boldsymbol{\chi} \rangle + \langle \boldsymbol{\psi} \,|\, \boldsymbol{P}_y \,|\, \boldsymbol{\psi} \rangle + \langle \boldsymbol{q}, \boldsymbol{\phi} \rangle \,. \tag{19}$$

Comparing (9) and (19), and following the steps described above, it is not hard to verify that $\boldsymbol{P}_x$, $\boldsymbol{P}_y$, and $\boldsymbol{q}$ are affine in $\boldsymbol{Q}_x$, $\boldsymbol{Q}_y$, and $\boldsymbol{\lambda}$. Therefore, the constraints in (10) consist of affine equalities and linear matrix inequalities, from which it follows that (10) is indeed an SDP.

By Proposition A.1, if $\boldsymbol{P}_x, \boldsymbol{P}_y \succeq \boldsymbol{0}$ and $\boldsymbol{q} = \boldsymbol{0}$, then indeed $S \geq 0$ for any $\boldsymbol{\chi}$, $\boldsymbol{\psi}$, and $\boldsymbol{\phi}$.

B Missing Details and Proofs for Section 6

Recall that K_α are the perfect square terms that satisfy $M_\alpha^\mu = M_\alpha^0 - \frac{\mu}{2(1-\mu)} K_\alpha$; for example, if α is the index that corresponds to $M^\mu_{\text{cvx},i,j,p}$, then we simply have

$$K_\alpha = \left\| x_i - x_j - \left(\nabla_x f(x_i, y_p) - \nabla_x f(x_j, y_p) \right) \right\|^2,$$

and if α is the index that corresponds to $M^\mu_{\text{smth},i,j,p,q}$ then $K_\alpha = 0$.

Proposition B.1 (Proposition 6.1)**.** *Suppose that we have the solution $\boldsymbol{Q}_x$, $\boldsymbol{Q}_y$, and $\boldsymbol{\lambda}$ of the SDP* (10) *when $\mu = 0$, so that the inequality*

$$\mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2\|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M_\alpha^0 \geq 0 \tag{20}$$

holds unconditionally. For that solution of the SDP, if there exists a scalar $\tau \geq 0$ which satisfies

$$R \coloneqq (1-\tau\eta)\mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2\|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M_\alpha^0 + \frac{1}{2}\sum_\alpha \lambda_\alpha K_\alpha \geq 0, \tag{21}$$

then for any $\mu \in [0, 1)$ it also holds that

$$(1-\tau\eta\mu)\mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2\|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M_\alpha^\mu \geq 0. \tag{22}$$

Proof. Suppose we have a convergence proof for the convex-concave case, meaning that for some given $c > 0$ we have already found $\boldsymbol{Q}_x$, $\boldsymbol{Q}_y$, and $\boldsymbol{\lambda}_\alpha \succeq 0$ that satisfy (20). Multiplying the left hand side of (22) by $(1-\mu)$ and then rearranging the terms, we get

$$\begin{aligned}
&(1-\mu)\left((1-\tau\eta\mu)\mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2\|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M_\alpha^\mu\right) \\
&\quad= (1-\mu)\left(\mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2\|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M_\alpha^0\right) - \tau\eta\mu(1-\mu)\mathcal{E}_0 + \frac{\mu}{2}\sum_\alpha \lambda_\alpha K_\alpha \\
&\quad= (1-\mu)^2\left(\mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2\|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M_\alpha^0\right) \\
&\qquad+ \mu(1-\mu)\left(\mathcal{E}_0 - \mathcal{E}_1 + c\beta\eta^2\|\nabla f(x_0, y_0)\|^2 - \sum_\alpha \lambda_\alpha M_\alpha^0 - \tau\eta\mathcal{E}_0 + \frac{1}{2}\sum_\alpha \lambda_\alpha K_\alpha\right) \\
&\qquad+ \frac{\mu^2}{2}\sum_\alpha \lambda_\alpha K_\alpha.
\end{aligned}$$

In the last line, the first term is nonnegative by (20), and the third term is nonnegative as each K_α is a perfect square. Hence, for (22) to hold, it suffices to have $\tau \geq 0$ which satisfies (21). □

Lemma B.2. *Given any η and β, the condition for τ to satisfy* (11) *can be formulated as a set of linear matrix inequalities.*

Proof. Repeating the steps we took in Sections 5 and A, one can notice that R, the left hand side of (11), is also a quadratic of $\boldsymbol{\omega}$. Moreover, comparing the definitions (9) of S and (11) of R, and noticing that the affine terms on $\boldsymbol{\omega}$ can only come from M_α^0, one can observe that the affine terms on $\boldsymbol{\omega}$ in S and R must be identical, both equal to $\langle \boldsymbol{q}, \boldsymbol{\phi}\rangle$. Hence, because $\boldsymbol{Q}_x$, $\boldsymbol{Q}_y$, and $\boldsymbol{\lambda}$ being a solution to the SDP (10) implies that $\boldsymbol{q} = \boldsymbol{0}$, we get that R is a quadratic on $\boldsymbol{\omega}$ without any affine terms. This implies that R takes the form of

$$R = \langle \boldsymbol{\chi} \,|\, \boldsymbol{R}_x \,|\, \boldsymbol{\chi}\rangle + \langle \boldsymbol{\psi} \,|\, \boldsymbol{R}_y \,|\, \boldsymbol{\psi}\rangle$$

for some matrices $\boldsymbol{R}_x$ and $\boldsymbol{R}_y$. Since R is affine in τ, we can reformulate the above as the statement that there exist matrices $\boldsymbol{R}_{x,0}, \boldsymbol{R}_{x,1}, \boldsymbol{R}_{y,0}, \boldsymbol{R}_{y,1} \in \mathbb{R}^{10\times 10}$ such that

$$R = \langle \boldsymbol{\chi} \,|\, \boldsymbol{R}_{x,0} + \tau\boldsymbol{R}_{x,1} \,|\, \boldsymbol{\chi}\rangle + \langle \boldsymbol{\psi} \,|\, \boldsymbol{R}_{y,0} + \tau\boldsymbol{R}_{y,1} \,|\, \boldsymbol{\psi}\rangle .$$

Therefore, for $R \geq 0$ to hold it suffices for τ to satisfy

$$\boldsymbol{R}_{x,0} + \tau\boldsymbol{R}_{x,1} \succeq \boldsymbol{0} \quad \text{and} \quad \boldsymbol{R}_{y,0} + \tau\boldsymbol{R}_{y,1} \succeq \boldsymbol{0} \tag{23}$$

which indeed are linear matrix inequalities. □

Proposition B.3 (Proposition 6.2)**.** *Fix any $\epsilon \in (0, 1]$. Then, with the same initialization, AltGDA-NeMo applied to f with parameters $(\epsilon\eta, \beta)$ generates the same iterates as AltGDA-NeMo applied to ϵf with parameters (η, β).*

Proof. For AltGDA-NeMo with stepsize $\epsilon\eta$ and momentum parameter β, notice that its update rule can be written as

$$x_{k+1} = x_k - \eta\big(\epsilon\nabla_x f(x_k, y_k)\big) + \beta(x_k - x_{k-1})$$
$$y_{k+1} = y_k + \eta\big(\epsilon\nabla_y f(x_{k+1}, y_k)\big) + \beta(y_k - y_{k-1}).$$

That is, the trajectory of the iterates is exactly the same as if AltGDA-NeMo is run with stepsize η and momentum parameter β, on the objective function ϵf. As f is 1-smooth and $\epsilon \leq 1$, we have

$$\|\epsilon\nabla f(x, y) - \epsilon\nabla f(x', y')\| = \epsilon\|\nabla f(x, y) - \nabla f(x', y')\| \leq \|(x, y) - (x', y')\|,$$

which implies that ϵf is 1-smooth. As the optimality of a point (x_*, y_*) is invariant under scaling the objective function by a positive scalar, the conclusion follows. □

C Solution of the Feasibility SDP

Here we describe the actual solution of the feasibility SDP (10), obtained by following the procedure elaborated in Sections 6 and 7. We chose $c = 8$ when obtaining these solutions. `Julia` is used as the main programming language in implementing and executing the tasks related to this section. In particular, the numerical solutions, before rationalization, are computed using the open source convex optimization solver `Clarabel` (Goulart & Chen, 2024), and the integer programming of rationalization is solved using another open source solver `HiGHS` (Huangfu & Hall, 2018).

C.1 Securing feasibility margins

The computations to be done in this section involve 10×10 matrices with complicated polynomials as their entries. To avoid the inconvenience of having to perform such computations by hand, we utilize the computer algebra system package `Symbolics.jl` (Gowda et al., 2021) in `Julia`.

As we briefly discussed in Section 7.2, after fixing the sparsity pattern, we refine the numerical solution to facilitate rationalization, making it robust against the perturbation incurred by it. Eventually, under the notations of (19), we wish $\boldsymbol{P}_x$ and $\boldsymbol{P}_y$ to be positive semidefinite, and for them to be robust under the rationalization step, we wish them to be away from the boundary of the positive semidefinite cone with margins as much as possible.

As the same logic would also apply to $\boldsymbol{P}_y$, let us temporarily focus on $\boldsymbol{P}_x$. As $\boldsymbol{P}_x$ is a 10×10 matrix, studying its properties would be easier if we can break it down into smaller matrices. In particular, it would be desirable if there existed a permutation that, when applied simultaneously to the rows and columns of $\boldsymbol{P}_x$, would transform it into a block matrix with smaller blocks. This is because applying the same permutation on the rows and columns simultaneously is equivalent to a conjugation by a permutation matrix, so $\boldsymbol{P}_x$ would be positive semidefinite if and only if the block diagonal entries obtained after the permutation are all positive semidefinite.

To this end, one can view $\boldsymbol{P}_x \in \mathbb{R}^{10\times 10}$ as an adjacency matrix of some graph with 10 nodes, where there is an edge between nodes i and j if and only if its (i, j) entry, $[\boldsymbol{P}_x]_{i,j}$, is nonzero. Each of the connected components will then form a block of $\boldsymbol{P}_x$, revealing the hidden block diagonal structure of $\boldsymbol{P}_x$. For $\boldsymbol{P}_x$, it turns out that there are four such connected components; $\{1, 2, 3, 6, 7, 9, 10\}$, $\{4\}$, $\{5\}$, and $\{8\}$. Hence, by the discussions from the previous paragraph, $\boldsymbol{P}_x$ is positive semidefinite if and only if $[\boldsymbol{P}_x]_{4,4} \geq 0$, $[\boldsymbol{P}_x]_{5,5} \geq 0$, $[\boldsymbol{P}_x]_{8,8} \geq 0$, and the principal submatrix $\check{\boldsymbol{P}}_x \in \mathbb{R}^{7\times 7}$ of $\boldsymbol{P}_x$ obtained by extracting only rows and columns 1, 2, 3, 6, 7, 9, and 10 is positive semidefinite.

As $[\boldsymbol{P}_x]_{4,4} = [\boldsymbol{P}_x]_{5,5} = 0$ and $[\boldsymbol{P}_x]_{8,8} = \frac{1}{2}\eta\lambda_{1,\mathrm{cvx},1,*,*}$, these quantities will automatically be nonnegative as long as the rationalization is executed respecting the constraint $\boldsymbol{\lambda} \succeq \mathbf{0}$. In addition, the reduction by inspecting the sparsity pattern can be pushed one step further, as $[\check{\boldsymbol{P}}_x]_{1,1}$, the $(1, 1)$ entry of the principal submatrix $\check{\boldsymbol{P}}_x$, is identically zero. This enforces all the entries in the first row and the first column of $\check{\boldsymbol{P}}_x$ to be also identically zero. For instance, as the $(1, 2)$ entry is

$$[\check{\boldsymbol{P}}_x]_{1,2} = [\boldsymbol{Q}_{x,0}]_{1,3} + \eta\big([\boldsymbol{Q}_{x,0}]_{1,1} + [\boldsymbol{Q}_{x,0}]_{1,3} + [\boldsymbol{Q}_{x,1}]_{1,3}\big) + \eta^2\big([\boldsymbol{Q}_{x,1}]_{1,1} + [\boldsymbol{Q}_{x,1}]_{1,3}\big),$$

it is necessary that

$$[\boldsymbol{Q}_{x,0}]_{1,3} = 0, \quad [\boldsymbol{Q}_{x,0}]_{1,1} + [\boldsymbol{Q}_{x,0}]_{1,3} + [\boldsymbol{Q}_{x,1}]_{1,3} = 0, \quad [\boldsymbol{Q}_{x,1}]_{1,1} + [\boldsymbol{Q}_{x,1}]_{1,3} = 0. \tag{24}$$

As we mentioned in Section 7.3, this is how the positive semidefiniteness condition imposes the equality constraints that have to be preserved in the rationalization step. Given that all the entries in the first row and the first column of $\check{\boldsymbol{P}}_x$ are zero, it suffices to have the positive semidefiniteness of its principal submatrix $\boldsymbol{X} \in \mathbb{R}^{6\times 6}$, obtained by removing the first row and column from $\check{\boldsymbol{P}}_x$, in order to have $\check{\boldsymbol{P}}_x \succeq \boldsymbol{0}$.

Our desideratum now amounts to having $\boldsymbol{X} \succeq \boldsymbol{0}$, and in order to ensure robustness of the positive semidefiniteness, the farther away from zero the eigenvalues are the better. Let us recall that $\boldsymbol{X}$ depends on η, and to emphasize this, let us write $\boldsymbol{X}(\eta)$.

Intuitively, the endpoints $\eta = 0$ and $\eta = \frac{1}{2}$ would be the most vulnerable in terms of the positive semidefiniteness of $\boldsymbol{X}(\eta)$ being preserved under rationalization. For the endpoint $\eta = \frac{1}{2}$, an easily implementable heuristic is, in the discretization step à la obtaining (14), to sample the points from $[0, \frac{1}{2} + \varepsilon]$ for a small $\varepsilon > 0$, so that the numerical solutions become a proof certificate of a slightly larger interval $[0, \frac{1}{2} + \varepsilon]$ from the first place, giving a surplus that can be freely lost in the rationalization step. However, this heuristic works only because we are exploiting the a priori known result by Shugart et al. (2026) which numerically verified that AltGDA-NeMo also converges for $\eta = -\beta$ slightly larger than $\frac{1}{2}$. In particular, the same trick cannot be applied to the other endpoint $\eta = 0$, as we already know well that alternating GDA with positive momentum does not converge in general C-C problems (Gidel et al., 2019b), let alone alternating GDA itself without momentum.

Hence, for $\boldsymbol{X}(\eta)$ around $\eta = 0$, we cannot help but carry out a precise characterization of the spectrum of it, in order to ensure that all of its eigenvalues are nonnegative. To this end, let us note that, as observable by a direct computation, there exist a scalar-valued polynomial $a_x(\eta)$, an $\mathbb{R}^5$-valued polynomial $\boldsymbol{b}_x(\eta)$, and an $\mathbb{R}^{5\times 5}$-valued polynomial $\boldsymbol{C}_x(\eta)$ such that

$$\boldsymbol{X}(\eta) = \begin{bmatrix} a_x(\eta) & \eta \boldsymbol{b}_x(\eta)^\top \\ \eta \boldsymbol{b}_x(\eta) & \eta \boldsymbol{C}_x(\eta) \end{bmatrix}. \tag{25}$$

It holds that $a_x(0) = [\boldsymbol{Q}_{x,0}]_{3,3}$, and the numerical solution suggests that $[\boldsymbol{Q}_{x,0}]_{3,3}$ is unambiguously strictly greater than zero. Hence, ignoring the possibly at most finitely many points where $a_x(\eta) = 0$, we have $\boldsymbol{X}(\eta) \succeq \boldsymbol{0}$ if and only if its Schur complement $\eta \boldsymbol{S}_x(\eta)$, where

$$\boldsymbol{S}_x(\eta) \coloneqq \boldsymbol{C}_x(\eta) - \frac{\eta}{a_x(\eta)} \boldsymbol{b}_x(\eta) \boldsymbol{b}_x(\eta)^\top,$$

is positive semidefinite. We wish to have $\eta \boldsymbol{S}_x(\eta) \succeq \boldsymbol{0}$ for all $\eta \in [0, \frac{1}{2}]$, so we need $\boldsymbol{S}_x \succeq \boldsymbol{0}$. Since $\boldsymbol{S}_x(\eta)$ is a rational function of η that is analytic on a neighborhood of $\eta = 0$, we have

$$\boldsymbol{S}_x(\eta) = \boldsymbol{C}_x(0) + \eta \left(\boldsymbol{C}_x'(0) - \frac{\boldsymbol{b}_x(0) \boldsymbol{b}_x(0)^\top}{a_x(0)} \right) + O(\eta^2). \tag{26}$$

Let us now recall a well-known result from perturbation theory.

Theorem C.1 (Rellich, 1969)**.** *Let $\boldsymbol{M}(\eta) \in \mathbb{C}^{d\times d}$ be a matrix-valued function whose entries are analytic on a neighborhood of $\eta = 0$. Moreover, suppose that $\boldsymbol{M}(\eta)$ is Hermitian for $\eta \in \mathbb{R}$. Then, for any eigenvalue γ_0 of $\boldsymbol{M}(0)$, for ν denoting the multiplicity of γ_0, there exist power series $\gamma_1(\eta), \dots, \gamma_\nu(\eta) \in \mathbb{C}$ and $\boldsymbol{\zeta}_1(\eta), \dots, \boldsymbol{\zeta}_\nu(\eta) \in \mathbb{C}^d$, all convergent in a neighborhood of $\eta = 0$, with $\{\boldsymbol{\zeta}_1(\eta), \dots, \boldsymbol{\zeta}_\nu(\eta)\}$ linearly independent for each η in that neighborhood, such that each $\boldsymbol{\zeta}_j(\eta)$ is an eigenvector of $\boldsymbol{M}(\eta)$ associated with eigenvalue $\gamma_j(\eta)$ satisfying $\gamma_j(0) = \gamma_0$.*

By Theorem C.1, the eigenvalues of $\boldsymbol{S}_x(\eta)$ can be written as analytic functions on a neighborhood of $\eta = 0$, and they converge to the eigenvalues of $\boldsymbol{C}_x(0)$ as $\eta \to 0$. By computing the characteristic polynomial of $\boldsymbol{C}_x(0)$, one gets that 0 is a simple eigenvalue of $\boldsymbol{C}_x(0)$. That is, as $\eta \to 0$, among the five eigenvalues of $\boldsymbol{S}_x(\eta)$, four of them are $\Theta(1)$, while one is $O(\eta)$. Clearly, the $O(\eta)$ eigenvalue is the most vulnerable one under rationalization, and thus it would be helpful to have additional measures which promote its nonnegativity.

To this end, let us further identify the $O(\eta)$ eigenvalue of $\boldsymbol{S}_x(\eta)$. First, let us write that eigenvalue as $\eta \gamma_x(\eta)$, where $\gamma_x(\eta)$ is some analytic function. In addition, one can observe that $\boldsymbol{1}_5 = (1, 1, 1, 1, 1)$ is in fact an eigenvector of $\boldsymbol{C}_x(0)$ associated to the eigenvalue 0. Hence, by Theorem C.1, there exists

an eigenvector of $\boldsymbol{S}_x(\eta)$ associated to $\eta\gamma_x(\eta)$, say $\boldsymbol{\zeta}_x(\eta)$, which admits a power series expansion $\boldsymbol{\zeta}_x(\eta) = \mathbf{1}_5 + O(\eta)$. Then, as $\boldsymbol{C}_x(0)\mathbf{1}_5 = \mathbf{0}$, it follows from (26) that

$$\boldsymbol{\zeta}_x(\eta)^\top \boldsymbol{S}_x(\eta)\boldsymbol{\zeta}_x(\eta) = \eta \mathbf{1}_5^\top \left(\boldsymbol{C}_x'(0) - \frac{\boldsymbol{b}_x(0)\boldsymbol{b}_x(0)^\top}{a_x(0)} \right) \mathbf{1}_5 + O(\eta^2). \tag{27}$$

But at the same time, since $\boldsymbol{\zeta}_x(\eta)$ is an eigenvector, it also holds that

$$\begin{aligned} \boldsymbol{\zeta}_x(\eta)^\top \boldsymbol{S}_x(\eta)\boldsymbol{\zeta}_x(\eta) &= \boldsymbol{\zeta}_x(\eta)^\top \big(\eta\gamma_x(\eta)\boldsymbol{\zeta}_x(\eta)\big) \\ &= \eta\gamma_x(\eta)\big(\boldsymbol{\zeta}_x(\eta)^\top \boldsymbol{\zeta}_x(\eta)\big) \\ &= \eta\gamma_x(\eta)\big(\mathbf{1}_5^\top \mathbf{1}_5 + O(\eta)\big) \\ &= 5\eta\gamma_x(0) + O(\eta^2) \end{aligned} \tag{28}$$

where the last line follows from the fact that $\mathbf{1}_5^\top \mathbf{1}_5 = 5$. Therefore, comparing (27) and (28), we get

$$5\gamma_x(0) = \mathbf{1}_5^\top \left(\boldsymbol{C}_x'(0) - \frac{\boldsymbol{b}_x(0)\boldsymbol{b}_x(0)^\top}{a_x(0)} \right) \mathbf{1}_5.$$

Using the equality constraints we wish to preserve in the rationalization step and the explicit formulae for $a_x(\eta)$, $\boldsymbol{b}_x(\eta)$, and $\boldsymbol{C}_x(\eta)$, we furthermore get

$$5\gamma_x(0) = \frac{\lambda_{1,\mathrm{cvx},0,1,0} - \lambda_{1,\mathrm{ccv},0,1,1}}{2} - \frac{([\boldsymbol{Q}_{x,0}]_{2,3} - [\boldsymbol{Q}_{x,0}]_{3,3})^2}{[\boldsymbol{Q}_{x,0}]_{3,3}}. \tag{29}$$

Let us recall that, to have $\boldsymbol{S}_x(\eta) \succeq \mathbf{0}$ for all sufficiently small η, it is necessary that $\gamma_x(0) \geq 0$.

As mentioned around the beginning of this section, the same logic applies to $\boldsymbol{P}_y$, and in fact with surprising similarities. Viewing $\boldsymbol{P}_y \in \mathbb{R}^{10\times 10}$ as an adjacency matrix on 10 nodes, there are five connected components, namely $\{1,2,3,6,7,8\}$, $\{4\}$, $\{5\}$, $\{9\}$, and $\{10\}$. One can check that the diagonal entries of $\boldsymbol{P}_y$ are also quantities that are nonnegative so long as the rationalization is executed respecting the required conditions, so it suffices to have the principal submatrix $\check{\boldsymbol{P}}_y$ of $\boldsymbol{P}_y$, obtained by extracting only rows and columns 1, 2, 3, 6, 7, and 8, be positive semidefinite. But $[\check{\boldsymbol{P}}_y]_{1,1}$ turns out to be also identically zero, so it is required that all the entries in the first row and column of $\check{\boldsymbol{P}}_y$ are identically zero and its principal submatrix $\boldsymbol{Y}$ obtained by removing the first row and column from $\check{\boldsymbol{P}}_y$ is positive semidefinite.

As in (25), there exist a scalar-valued polynomial $a_y(\eta)$, an $\mathbb{R}^4$-valued polynomial $\boldsymbol{b}_y(\eta)$, and an $\mathbb{R}^{4\times 4}$-valued polynomial $\boldsymbol{C}_y(\eta)$ such that

$$\boldsymbol{Y}(\eta) = \begin{bmatrix} a_y(\eta) & \eta\boldsymbol{b}_y(\eta)^\top \\ \eta\boldsymbol{b}_y(\eta) & \eta\boldsymbol{C}_y(\eta) \end{bmatrix}.$$

It holds that $a_y(0) = [\boldsymbol{Q}_{y,0}]_{3,3}$, and the numerical solution suggests that $[\boldsymbol{Q}_{y,0}]_{3,3}$ is unambiguously strictly greater than zero. Hence, we wish the Schur complement $\eta\boldsymbol{S}_y(\eta)$, which admits the series expansion

$$\boldsymbol{S}_y(\eta) = \boldsymbol{C}_y(0) + \eta \left(\boldsymbol{C}_y'(0) - \frac{\boldsymbol{b}_y(0)\boldsymbol{b}_y(0)^\top}{a_y(0)} \right) + O(\eta^2),$$

to be positive semidefinite. Curiously, as one can verify, $\boldsymbol{C}_y$ also has 0 as a simple eigenvalue, with an associated eigenvector $\mathbf{1}_4 = (1,1,1,1)$. Hence, by Theorem C.1, letting $\eta\gamma_y(\eta)$ be the $O(\eta)$ eigenvalue of $\boldsymbol{S}_y(\eta)$, a computation analogous to (27) and (28) yields

$$4\gamma_y(0) = \mathbf{1}_4^\top \left(\boldsymbol{C}_y'(0) - \frac{\boldsymbol{b}_y(0)\boldsymbol{b}_y(0)^\top}{a_y(0)} \right) \mathbf{1}_4.$$

But here, using the equality constraints we wish to preserve in the rationalization step and the explicit formulae for $a_y(\eta)$, $\boldsymbol{b}_y(\eta)$, and $\boldsymbol{C}_y(\eta)$, we end up with

$$4\gamma_y(0) = -\frac{\lambda_{1,\mathrm{cvx},0,1,0} - \lambda_{1,\mathrm{ccv},0,1,1}}{2} - \frac{([\boldsymbol{Q}_{y,0}]_{2,3} + [\boldsymbol{Q}_{y,0}]_{3,3})^2}{[\boldsymbol{Q}_{y,0}]_{3,3}}. \tag{30}$$

While we need both $\gamma_x(0) \geq 0$ and $\gamma_y(0) \geq 0$ to hold, (29) and (30) together assert that that can happen only when $\gamma_x(0) = \gamma_y(0) = 0$. As a side remark, let us notice that this asserts the constraints

$$\lambda_{1,\mathrm{cvx},0,1,0} = \lambda_{1,\mathrm{ccv},0,1,1}, \qquad [\boldsymbol{Q}_{x,0}]_{2,3} - [\boldsymbol{Q}_{x,0}]_{3,3} = 0, \qquad [\boldsymbol{Q}_{y,0}]_{2,3} + [\boldsymbol{Q}_{y,0}]_{3,3} = 0. \tag{31}$$

Anyhow, we then have that $\eta\gamma_x(\eta)$ and $\eta\gamma_y(\eta)$, the eigenvalues of $\boldsymbol{S}_x(\eta)$ and $\boldsymbol{S}_y(\eta)$, respectively, are of $O(\eta^2)$. In terms of the Schur complements $\eta\boldsymbol{S}_x(\eta)$ and $\eta\boldsymbol{S}_y(\eta)$ of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, this means that the largest positive semidefiniteness margin we can hope for is to have

$$\eta\boldsymbol{S}_x(\eta) \succeq \rho\eta^3\boldsymbol{I} \quad \text{and} \quad \eta\boldsymbol{S}_y(\eta) \succeq \rho\eta^3\boldsymbol{I} \tag{32}$$

for some $\rho > 0$.

In the reverse direction of obtaining the Schur complements $\eta\boldsymbol{S}_x(\eta)$ and $\eta\boldsymbol{S}_y(\eta)$ from their original matrices, one gets that the Loewner inequalities in (32) are equivalent to

$$\begin{bmatrix} a_x(\eta) & \eta\boldsymbol{b}_x(\eta)^\top \\ \eta\boldsymbol{b}_x(\eta) & \eta\boldsymbol{C}_x(\eta) - \rho\eta^3\boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0} \quad \text{and} \quad \begin{bmatrix} a_y(\eta) & \eta\boldsymbol{b}_y(\eta)^\top \\ \eta\boldsymbol{b}_y(\eta) & \eta\boldsymbol{C}_y(\eta) - \rho\eta^3\boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0}.$$

These linear matrix inequalities are added when solving the sparsified SDP.

Speaking of the sparsified SDP, notice that we no longer need the ℓ_1 regularization term for $\boldsymbol{\lambda}$, but we do wish to maximize ρ so that the positive semidefiniteness margin is secured as much as possible in view of (32). Therefore, the objective of the sparsified SDP is to minimize $-\rho$, whereas the small ℓ_2 regularization terms are still applied.

C.2 Further details on the rationalization step

As $\boldsymbol{Q}_{x,0}$, $\boldsymbol{Q}_{x,1}$, $\boldsymbol{Q}_{y,0}$, and $\boldsymbol{Q}_{y,1}$ are 3×3 symmetric matrices, there are six independent entries each. In addition, sparsifying $\boldsymbol{\lambda}_1$ makes all of its entries zero except for ten of them, so essentially there are ten independent entries in $\boldsymbol{\lambda}_1$ remaining. We pack those 34 independent entries into a single vector $\boldsymbol{\iota} \in \mathbb{R}^{34}$.

As we have seen in Section C.1, positive semidefiniteness constraints give birth to homogeneous linear equality constraints on $\boldsymbol{\iota}$, such as the ones we saw in (24) and (31). The constraint $\boldsymbol{q} = \boldsymbol{0}$ from (10) also yields such linear constraints. Hence, as sketched in Section 7.3, we can collect those linear constraints to form a system $\boldsymbol{A}\boldsymbol{\iota} = \boldsymbol{0}$ where we here use $\boldsymbol{A}$ to denote the matrix of coefficients.

Through Gaussian elimination, we can split this system into independent subsystems where the variables of those subsystems form a partition of the entries in $\boldsymbol{\iota}$. For each of the subsystems, say it has the form of $\acute{\boldsymbol{A}}\boldsymbol{h} = \boldsymbol{0}$, where $\boldsymbol{h} \in \mathbb{R}^m$ is a subvector of $\boldsymbol{\iota}$ and $\acute{\boldsymbol{A}}$ is the reduced matrix of coefficients. Let us denote the numerical value found for $\boldsymbol{h}$, to be rationalized, as $\hat{\boldsymbol{h}}$.

We fix a denominator d, which could vary from subsystem to subsystem, and aim for the best rational approximation $\hat{\boldsymbol{h}} \approx \frac{1}{d}\boldsymbol{p}$, where $\boldsymbol{p} \in \mathbb{Z}^m$, subject to the constraint $\acute{\boldsymbol{A}}\boldsymbol{h} = \boldsymbol{0}$ which must be preserved. For a fixed d, we can cast this into a mixed integer linear programming problem

$$\begin{aligned} \min_{\boldsymbol{p}\in\mathbb{Z}^m, \boldsymbol{e}\in\mathbb{R}^m} \quad & \sum_{j=1}^{m} e_j \\ \text{s.t.} \quad & \acute{\boldsymbol{A}}\boldsymbol{p} = \boldsymbol{0}, \\ & -\boldsymbol{e} \leqslant \frac{1}{d}\boldsymbol{p} - \hat{\boldsymbol{h}} \leqslant \boldsymbol{e} \end{aligned}$$

where $\leqslant$ denotes the elementwise inequality.

We solve this problem for each $d = 1, \ldots, 15$ and retain the candidate with the smallest total deviation $\sum_{j=1}^{m} e_j$. The remaining nonnegativity and positive semidefiniteness conditions are verified after reconstruction, as described next.

C.3 The exact rational solution

As discussed in Section C.1, we strategically select the discretization points $\eta_1, \ldots, \eta_n$ from $[0, \frac{1}{2} + \varepsilon]$ for some $\varepsilon > 0$. We set $\eta_1, \ldots, \eta_n$ to be equidistant points over the interval $[0, 0.505]$,

with $n = 256$. The ℓ_2 regularization terms are always added after being scaled down by a factor of 10^{-4}. On the other hand, the ℓ_1 regularization term is added without any scaling, because it is added only when solving the feasibility SDP (14) whose objective function is identically zero, so the scale with respect to the objective function is irrelevant.

The sparsity pattern of $\boldsymbol{\lambda}$ reveals that it suffices to only include ten among the 72 quantities M_α^μ, namely

$$\begin{aligned}&M^\mu_{\text{cvx},0,1,0}, M^\mu_{\text{cvx},1,*,*}, M^\mu_{\text{cvx},*,0,0}, M^\mu_{\text{cvx},*,1,1},\\ &M^\mu_{\text{ccv},0,1,1}, M^\mu_{\text{ccv},0,*,*}, M^\mu_{\text{ccv},1,*,*}, M^\mu_{\text{ccv},*,0,1},\\ &M^\mu_{\text{smth},1,0,1,1}, M^\mu_{\text{smth},0,0,1,0}.\end{aligned}$$

Solving the simplified version of (14) with only these 10 quantities included and then rationalizing, we obtain the solution

$$\begin{aligned}\boldsymbol{Q}_{x,0} &= \begin{bmatrix} \frac{189}{5} & -\frac{189}{10} & 0 \\ -\frac{189}{10} & \frac{201}{10} & \frac{149}{14} \\ 0 & \frac{149}{14} & \frac{149}{14} \end{bmatrix}, \qquad \boldsymbol{Q}_{x,1} = \begin{bmatrix} \frac{189}{5} & 0 & -\frac{189}{5} \\ 0 & -16 & \frac{137}{14} \\ -\frac{189}{5} & \frac{137}{14} & \frac{227}{10} \end{bmatrix},\\ \boldsymbol{Q}_{y,0} &= \begin{bmatrix} \frac{189}{5} & 0 & 0 \\ 0 & 15 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \boldsymbol{Q}_{y,1} = \begin{bmatrix} \frac{189}{5} & 0 & -\frac{189}{5} \\ 0 & \frac{121}{15} & -\frac{611}{15} \\ -\frac{189}{5} & -\frac{611}{15} & \frac{769}{13} \end{bmatrix}, \end{aligned} \tag{33}$$

$$\lambda_{1,\text{cvx},1,*,*} = \lambda_{1,\text{ccv},*,0,1} = \frac{378}{5}, \quad \lambda_{1,\text{smth},1,0,1,1} = \frac{77}{3}, \quad \lambda_{1,\text{smth},0,0,1,0} = \frac{159}{7},$$

$$\lambda_{1,\text{cvx},0,1,0} = \lambda_{1,\text{cvx},*,0,0} = \lambda_{1,\text{cvx},*,1,1} = \lambda_{1,\text{ccv},0,1,1} = \lambda_{1,\text{ccv},0,*,*} = \lambda_{1,\text{ccv},1,*,*} = \frac{189}{5}.$$

One can easily check the following indeed holds, as required by the feasibility condition.

Lemma C.2. *For the matrices in* (33)*, it holds that* $\boldsymbol{Q}_{x,0} + \eta\boldsymbol{Q}_{x,1} \succeq \mathbf{0}$ *and* $\boldsymbol{Q}_{y,0} + \eta\boldsymbol{Q}_{y,1} \succeq \mathbf{0}$ *for any* $\eta \in [0, \frac{1}{2}]$.

Proof. As the positive semidefinite cone is a convex set, to see that $\boldsymbol{Q}_{x,0} + \eta\boldsymbol{Q}_{x,1} \succeq \mathbf{0}$ for all $\eta \in [0, \frac{1}{2}]$ it suffices to check that $\boldsymbol{Q}_{x,0} \succeq \mathbf{0}$ and $\boldsymbol{Q}_{x,0} + \frac{1}{2}\boldsymbol{Q}_{x,1} \succeq \mathbf{0}$. This amounts to verifying the positive semidefiniteness of two fixed numeric matrices, which can be easily done, for example via Sylvester's criterion. □

For the rational coefficients in (33), direct substitution gives $\boldsymbol{q}(\eta) \equiv \mathbf{0}$ and $\boldsymbol{\lambda}(\eta) \succeq \mathbf{0}$ throughout $\eta \in [0, \frac{1}{2}]$. It remains to verify the positive semidefiniteness of $\boldsymbol{P}_x(\eta)$ and $\boldsymbol{P}_y(\eta)$.

Positive semidefiniteness is equivalent to nonnegativity of all principal minors. We apply this criterion blockwise, reducing the verification to sign checks for univariate polynomials with rational coefficients. Identically zero minors require no further check. For each nonzero minor, Sturm's theorem confirms that it has no root in $(0, \frac{1}{2})$. Its value at $\eta = \frac{1}{4}$ is positive, and its values at both endpoints are nonnegative. All these checks are performed in exact rational arithmetic.

Consequently, $\boldsymbol{P}_x(\eta) \succeq \mathbf{0}$ and $\boldsymbol{P}_y(\eta) \succeq \mathbf{0}$ throughout $[0, \frac{1}{2}]$. Together with Lemma C.2, these verifications establish that (33), interpreted through (13), satisfies (10) for $L = 1$, $\mu = 0$, $\beta = -\eta$, and $c = 8$ throughout this interval.

D On the SDP for the Strongly-convex-strongly-concave Case

With the solution (33), we then seek τ that satisfies (11), so that we can invoke Proposition 6.1, which would lead to our convergence proof in the SC-SC case. The smaller the contraction factor is, the more desirable, as then the convergence will be faster; one can see that this actually turns out

to be the case in the following Section E. According to the observations made in Lemma B.2, using the notations therein, such a search can be formulated into another SDP

$$\begin{aligned} &\max_{\tau} \ \tau \\ &\text{s.t. } \ \boldsymbol{R}_{x,0} + \tau \boldsymbol{R}_{x,1} \succeq \mathbf{0}, \ \boldsymbol{R}_{y,0} + \tau \boldsymbol{R}_{y,1} \succeq \mathbf{0}. \end{aligned} \tag{34}$$

This is also a one-parameter SDP with parameter η (or, equivalently, β), so we apply the same techniques we used in solving the other parametric SDP, namely (10), as described in Section 7.1. That being said, in this case we seek a constant scalar τ independent of η; although this is an assumption made for simplification purposes, one can understand that the dependency of η is already decoupled from τ through the implicit ansatz made in (11) that the contraction factor is $1-\tau\eta$, where one could have alternatively written it as $1-\tau$.

Using the same discretization technique of choosing $\eta_1, \ldots, \eta_n$ to be equidistant points over the interval $[0, 0.505]$ with $n = 256$, the numerical SDP solver gave $\tau^* = 0.305582\ldots$ as the optimal solution for the discretized proxy of (34). Based on this result, we conservatively choose $\tau = \frac{3}{10}$ as the value to be used in the algebraic proof recovery. Applying the same exact verification procedure as described in Section C.3, using Sylvester's criterion and Sturm's theorem, to the matrices in (23) confirms that both linear matrix inequality constraints in (34) are indeed satisfied over the whole interval $\eta \in [0, \frac{1}{2}]$. Therefore, the required certificate for the SC-SC case follows from Proposition 6.1. See Section E for the resulting convergence bounds.

E Proofs of the Convergence Upper Bounds

As the proofs of the convergence rates in Theorems 4.1 and 4.2 require the exact solutions of the SDP constructed in Section 5, they had to be unavoidably deferred until here. As we now have the exact solutions (33), we are ready to establish the convergence results.

E.1 Convex-concave Case

Let us begin by observing the following, which is an immediate corollary of Proposition A.1.

Corollary E.1. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a positive symmetric matrix. Then, for any collection of n vectors $\boldsymbol{z} = (z_1, \ldots, z_n)$, all from the same Hilbert space, it holds that*

$$\langle \boldsymbol{z} \,|\, \boldsymbol{A} \,|\, \boldsymbol{z} \rangle \leq \|\boldsymbol{A}\| \sum_{i=1}^{n} \|z_i\|^2.$$

Proof. Let $\|\boldsymbol{A}\| = \Lambda$. Then, $\Lambda \boldsymbol{I} - \boldsymbol{A}$ is positive semidefinite, so by Proposition A.1 we get

$$\Lambda \sum_{i=1}^{n} \|z_i\|^2 - \langle \boldsymbol{z} \,|\, \boldsymbol{A} \,|\, \boldsymbol{z} \rangle = \langle \boldsymbol{z} \,|\, \Lambda \boldsymbol{I} - \boldsymbol{A} \,|\, \boldsymbol{z} \rangle \geq 0$$

which completes the proof. □

Let us then observe some properties of the obtained solution.

Proposition E.2. *For any two symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same size, suppose that there exist scalars t_0, t_1, and $\Lambda \geq 0$ such that*

$$\|\boldsymbol{A} + t_0 \boldsymbol{B}\| \leq \Lambda \qquad \text{and} \qquad \|\boldsymbol{A} + t_1 \boldsymbol{B}\| \leq \Lambda.$$

Then, for any $t \in [t_0, t_1]$, it also holds that $\|\boldsymbol{A} + t\boldsymbol{B}\| \leq \Lambda$.

Proof. We use the triangle inequality, which holds for any norm. For any fixed $t \in [t_0, t_1]$, letting $\omega = \frac{t-t_0}{t_1-t_0} \in [0, 1]$, we can observe that

$$\begin{aligned} \|\boldsymbol{A} + t\boldsymbol{B}\| &= \|(1-\omega)(\boldsymbol{A} + t_0\boldsymbol{B}) + \omega(\boldsymbol{A} + t_1\boldsymbol{B})\| \\ &\leq (1-\omega)\|\boldsymbol{A} + t_0\boldsymbol{B}\| + \omega\|\boldsymbol{A} + t_1\boldsymbol{B}\| \leq \Lambda \end{aligned}$$

as claimed. □

Corollary E.3. *For the matrices in* (33)*, both* $\|\boldsymbol{Q}_{x,0} + \eta\boldsymbol{Q}_{x,1}\| \leq 75$ *and* $\|\boldsymbol{Q}_{y,0} + \eta\boldsymbol{Q}_{y,1}\| \leq 75$ *hold for any* $\eta \in [0, \frac{1}{2}]$.

Proof. By Proposition E.2, it suffices to verify that the claimed inequalities hold when $\eta = 0$ and $\eta = \frac{1}{2}$. These verifications are straightforward, so we omit the details here; for example, verifying that $\left\|\boldsymbol{Q}_{x,0} + \frac{1}{2}\boldsymbol{Q}_{x,1}\right\| \leq 75$ holds can be done by showing that the matrix $75\boldsymbol{I} - \left(\boldsymbol{Q}_{x,0} + \frac{1}{2}\boldsymbol{Q}_{x,1}\right)$ is positive semidefinite, for example using Sylvester's criterion. □

Now we are ready to prove Theorem 4.1.

Theorem E.4 (Theorem 4.1)**.** *Under Assumption 3, suppose that* f *is convex-concave and* L*-smooth. For any* η *and* β *such that* $0 < \eta L \leq -\beta \leq \frac{1}{2}$*, say we run* AltGDA-NeMo *with stepsize* η *and momentum* β*. Then, for any* $T \geq 1$*, it holds that*

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla f(x_k, y_k)\|^2 \leq \frac{375}{32|\beta|\eta^2 T}\left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2\right)$$

Proof. Let us first establish the result in the special case where $\eta = -\beta$ and $L = 1$, which will be used to prove the convergence in the general case.

From (33) being a solution to the feasibility SDP (10), we know that (7) holds. As f being convex-concave implies that $M^0_{k,\alpha} \geq 0$ for all α, from (7) we obtain that

$$\begin{aligned}\mathcal{E}_k - \mathcal{E}_{k+1} &\geq -8\beta\eta^2\|\nabla f(x_k, y_k)\|^2 + \sum_\alpha \lambda_\alpha M_{k,\alpha} \\ &\geq -8\beta\eta^2\|\nabla f(x_k, y_k)\|^2 \\ &= 8|\beta|\eta^2\|\nabla f(x_k, y_k)\|^2.\end{aligned}$$

Taking the telescopic sum of the above yields

$$\sum_{k=0}^{T-1} 8|\beta|\eta^2\|\nabla f(x_k, y_k)\|^2 \leq \mathcal{E}_0 - \mathcal{E}_T. \tag{35}$$

By Lemma C.2, we know that $\mathcal{E}_T \geq 0$. Meanwhile, notice that Corollaries E.1 and E.3 together imply that

$$\begin{aligned}\mathcal{E}_0 &= \langle \boldsymbol{\xi}_0 \,|\, \boldsymbol{Q}_x \,|\, \boldsymbol{\xi}_0\rangle + \langle \boldsymbol{\theta}_0 \,|\, \boldsymbol{Q}_y \,|\, \boldsymbol{\theta}_0\rangle \\ &\leq 75\Big(\|x_0 - x_*\|^2 + \|\eta\nabla_x f(x_0, y_0)\|^2 + \|u_0\|^2 + \|y_0 - y_*\|^2 + \|\eta\nabla_y f(x_0, y_0)\|^2 + \|v_0\|^2\Big).\end{aligned}$$

Since the 1-smoothness of f and the bound $\eta \leq \frac{1}{2}$ give

$$\begin{aligned}&\|\eta\nabla_x f(x_0, y_0)\|^2 + \|\eta\nabla_y f(x_0, y_0)\|^2 \\ &\quad= \eta^2\|\nabla_x f(x_0, y_0) - \nabla_x f(x_*, y_*)\|^2 + \eta^2\|\nabla_y f(x_0, y_0) - \nabla_y f(x_*, y_*)\|^2 \\ &\quad\leq \frac{1}{4}\left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2\right)\end{aligned}$$

while the initialization asserts that $u_0 = 0$ and $v_0 = 0$, it follows that

$$\mathcal{E}_0 \leq 75\left(1 + \frac{1}{4}\right)\left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2\right). \tag{36}$$

Therefore, plugging this bound into (35), we get

$$8|\beta|\eta^2\sum_{k=0}^{T-1}\|\nabla f(x_k, y_k)\|^2 \leq \frac{375}{4}\left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2\right).$$

Dividing by $8|\beta|\eta^2 T$ leads to

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla f(x_k, y_k)\|^2 \leq \frac{375}{32|\beta|\eta^2 T}\left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2\right) \tag{37}$$

which gives us the convergence rate in the 1-smooth case with $\eta = -\beta$.

Now, to prove the result in the general case, fix any $\beta \in [-\frac{1}{2}, 0)$. Let f be L-smooth. Put $\epsilon = \frac{\eta L}{-\beta}$, and suppose that we are applying AltGDA-NeMo on $\frac{\epsilon}{L} f$ with stepsize $-\beta$ and momentum parameter β. Then, the update rule of the minimization step (2a) reads

$$\begin{aligned} x_{k+1} &= x_k - \frac{(-\beta)\epsilon}{L} \nabla_x f(x_k, y_k) + \beta(x_k - x_{k-1}) \\ &= x_k - \eta \nabla_x f(x_k, y_k) + \beta(x_k - x_{k-1}), \end{aligned} \tag{38}$$

which is identical to the update rule one would get as a minimization step of AltGDA-NeMo with stepsize η and momentum parameter β applied to f. Similarly, the update rule of the maximization step (2b) reads

$$y_{k+1} = y_k + \eta \nabla_y f(x_{k+1}, y_k) + \beta(y_k - y_{k-1}),$$

from which we conclude that the iterates $\{(x_k, y_k)\}$ one gets by applying AltGDA-NeMo with stepsize η and momentum β on f are identical to the iterates one would get by applying AltGDA-NeMo with stepsize $-\beta$ and momentum β on $\frac{\epsilon}{L} f$. As $0 < \epsilon \leq 1$ by assumption, we know that $\frac{\epsilon}{L} f$ is 1-smooth, and thus (37) asserts that

$$\frac{\epsilon^2}{L^2 T} \sum_{k=0}^{T-1} \|\nabla f(x_k, y_k)\|^2 \leq \frac{375}{32(-\beta)^3 T} \left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2 \right).$$

Dividing both sides by $\frac{\epsilon^2}{L^2}$ and then using the definition $\epsilon = \frac{\eta L}{-\beta}$ yields the claimed bound. □

E.2 Strongly-convex-strongly-concave Case

For the SC-SC result, we in addition use the following lower bound result that mirrors Corollary E.3.

Lemma E.5. *Let $\boldsymbol{E}_{ij}$ be the matrix whose (i, j) entry is 1 and all the other entries are 0. Then, for the matrices in* (33)*, both $\boldsymbol{Q}_{x,0} + \eta \boldsymbol{Q}_{x,1} \succeq 28\boldsymbol{E}_{11} - 27\boldsymbol{E}_{22}$ and $\boldsymbol{Q}_{y,0} + \eta \boldsymbol{Q}_{y,1} \succeq 28\boldsymbol{E}_{11} - 27\boldsymbol{E}_{22}$ hold for any $\eta \in [0, \frac{1}{2}]$.*

Proof. Let us focus on the former positive semidefiniteness claim, as the latter can be shown in the exact same manner. Once we rewrite the Loewner inequality as

$$\left(\boldsymbol{Q}_{x,0} - 28\boldsymbol{E}_{11} + 27\boldsymbol{E}_{22}\right) + \eta \boldsymbol{Q}_{x,1} \succeq \boldsymbol{0},$$

thanks to the convexity of the positive semidefinite cone, in order to show that the above holds for all $\eta \in [0, \frac{1}{2}]$ it suffices to show that it holds at the endpoints $\eta = 0$ and $\eta = \frac{1}{2}$. The rest is straightforward, as we now only need to verify the positive semidefiniteness of two fixed numeric matrices. □

Corollary E.6. *For any $k \geq 0$, consider $\mathcal{E}_k$ defined with the quantities specified in* (33)*. Then, for any $\eta \in [0, \frac{1}{2}]$, it holds that*

$$\mathcal{E}_k \geq 28 \left(\|x_k - x_*\|^2 + \|y_k - y_*\|^2 \right) - 27\eta^2 \|\nabla f(x_k, y_k)\|^2.$$

Proof. By Proposition A.1 and Lemma E.5, we have

$$\begin{aligned} 0 &\leq \langle \boldsymbol{\xi}_k \,|\, \boldsymbol{Q}_{x,0} + \eta \boldsymbol{Q}_{x,1} - 28\boldsymbol{E}_{11} + 27\boldsymbol{E}_{22} \,|\, \boldsymbol{\xi}_k \rangle \\ &= \langle \boldsymbol{\xi}_k \,|\, \boldsymbol{Q}_x \,|\, \boldsymbol{\xi}_k \rangle - 28\|x_k - x_*\|^2 + 27\|\eta \nabla_x f(x_k, y_k)\|^2 \end{aligned}$$

and similarly

$$0 \leq \langle \boldsymbol{\theta}_k \,|\, \boldsymbol{Q}_y \,|\, \boldsymbol{\theta}_k \rangle - 28\|y_k - y_*\|^2 + 27\|\eta \nabla_y f(x_k, y_k)\|^2.$$

As $\mathcal{E}_k = \langle \boldsymbol{\xi}_k \,|\, \boldsymbol{Q}_x \,|\, \boldsymbol{\xi}_k \rangle + \langle \boldsymbol{\theta}_k \,|\, \boldsymbol{Q}_y \,|\, \boldsymbol{\theta}_k \rangle$, adding the two inequalities above side by side and then rearranging the terms gives us the claimed inequality. □

Theorem E.7 (Theorem 4.2)**.** *Under Assumption 3, suppose that f is L-smooth and μ-strongly-convex-strongly-concave for some $\mu \in (0, L)$. For any η and β such that $0 < \eta L \leq -\beta \leq \frac{1}{2}$, say we run AltGDA-NeMo with stepsize η and momentum β. Then, for any $T \geq 1$, it holds that*

$$\|x_T - x_*\|^2 + \|y_T - y_*\|^2 \leq \frac{75}{17} \left(1 - \frac{3\eta\mu}{10}\right)^T \left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2 \right).$$

Proof. As in the proof of Theorem 4.1, namely the C-C case, let us first establish the result in the special case where $\eta = -\beta$ and $0 < \mu < L = 1$, and then extend that result into the general case.

As discussed in Section D, the quantities (33) and $\tau = \frac{3}{10}$ satisfy the feasibility SDP (10) with $c = 8$ and the linear matrix inequalities in (11). Therefore, by Proposition 6.1, it holds that

$$(1 - \tau\eta\mu)\mathcal{E}_k - \mathcal{E}_{k+1} - 8|\beta|\eta^2\|\nabla f(x_k, y_k)\|^2 - \sum_\alpha \lambda_\alpha M^\mu_{k,\alpha} \geq 0.$$

That f is μ-SC-SC implies that $M^\mu_\alpha \geq 0$ for all α, so we get from the above that

$$0 \leq 8|\beta|\eta^2\|\nabla f(x_k, y_k)\|^2 + \sum_\alpha \lambda_\alpha M^\mu_{k,\alpha} \leq \left(1 - \frac{3\eta\mu}{10}\right)\mathcal{E}_k - \mathcal{E}_{k+1}.$$

Unrolling the recurrence yields

$$\mathcal{E}_T \leq \left(1 - \frac{3\eta\mu}{10}\right)^T \mathcal{E}_0. \tag{39}$$

Here, by the 1-smoothness of f, Corollary E.6, and the bound $\eta \leq \frac{1}{2}$, we have

$$\begin{aligned}
\mathcal{E}_T &\geq 28\left(\|x_T - x_*\|^2 + \|y_T - y_*\|^2\right) - 27\eta^2\|\nabla f(x_T, y_T)\|^2 \\
&\geq \left(28 - 27\eta^2\right)\left(\|x_T - x_*\|^2 + \|y_T - y_*\|^2\right) \\
&\geq \left(28 - \frac{27}{4}\right)\left(\|x_T - x_*\|^2 + \|y_T - y_*\|^2\right)
\end{aligned}$$

Plugging this lower bound and the upper bound (36) of $\mathcal{E}_0$ into (39), we obtain

$$\frac{85}{4}\left(\|x_T - x_*\|^2 + \|y_T - y_*\|^2\right) \leq \frac{375}{4}\left(1 - \frac{3\eta\mu}{10}\right)^T \left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2\right).$$

Dividing both sides by $\frac{85}{4}$ gives us

$$\|x_T - x_*\|^2 + \|y_T - y_*\|^2 \leq \frac{75}{17}\left(1 - \frac{3\eta\mu}{10}\right)^T \left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2\right). \tag{40}$$

Now we extend this result into the general case. Fix any $\beta \in [-\frac{1}{2}, 0)$, and let f be L-smooth and μ-SC-SC. As we did in the proof of Theorem 4.1, let us set $\epsilon = \frac{\eta L}{-\beta}$ and suppose that we are applying AltGDA-NeMo on $\frac{\epsilon}{L}f$. Then, by steps analogous to those in (38), we can again verify that the iterates $\{(x_k, y_k)\}$ one gets by applying AltGDA-NeMo with stepsize η and momentum β on f are identical to the iterates one would get by applying AltGDA-NeMo with stepsize $-\beta$ and momentum β on $\frac{\epsilon}{L}f$. As $0 < \epsilon \leq 1$ by assumption, $\frac{\epsilon}{L}f$ is 1-smooth and $\frac{\epsilon\mu}{L}$-SC-SC, so (40) applies to $\frac{\epsilon}{L}f$, only to yield

$$\begin{aligned}
\|x_T - x_*\|^2 + \|y_T - y_*\|^2 &\leq \frac{75}{17}\left(1 - \frac{3(-\beta)\epsilon\mu}{10L}\right)^T \left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2\right) \\
&= \frac{75}{17}\left(1 - \frac{3\eta\mu}{10}\right)^T \left(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2\right)
\end{aligned}$$

where in the last line the identity $-\beta\epsilon = \eta L$ which follows from the definition of ϵ is used. □

F On the Convergence Lower Bounds

F.1 Strongly-convex-strongly-concave case

Consider the class of objective functions f that are L-smooth and μ-SC-SC for some $\mu > 0$. Suppose that we are given the task of, given $\epsilon > 0$, finding a point (x, y) such that

$$\|x - x_*\|^2 + \|y - y_*\|^2 \leq \epsilon. \tag{41}$$

Zhang et al. (2022b) have established that for any given first-order method, there exists an instance that requires $\Omega\Big(\frac{L}{\mu}\log\big(\frac{1}{\epsilon}\big)\Big)$ gradient evaluations in order to find such a point. It is clear that the upper bound provided by Theorem 4.2 matches this lower bound. Indeed, for any given $\epsilon > 0$, if we let $C = \frac{75}{17}\big(\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2\big)$, then Theorem 4.2 and an elementary inequality $1 - x \le e^{-x}$ which holds for all $x \in \mathbb{R}$ imply that

$$\|x - x_*\|^2 + \|y - y_*\|^2 \le Ce^{-3\eta\mu T/10}.$$

Hence, to achieve (41), it suffices to choose $\eta = \frac{|\beta|}{L}$ and have T satisfy

$$\begin{aligned} Ce^{-3\eta\mu T/10} \le \epsilon \quad &\iff \quad -\frac{3\eta\mu T}{10} \le \log\left(\frac{\epsilon}{C}\right) \\ &\iff \quad T \ge \frac{10L}{3|\beta|\mu}\log\left(\frac{C}{\epsilon}\right). \end{aligned}$$

Notice that, for any fixed β, this asserts that $\Theta\Big(\frac{L}{\mu}\log\big(\frac{1}{\epsilon}\big)\Big)$ iterations are indeed sufficient.

F.2 Convex-concave case

We remark that the convergence lower bounds for the extragradient method (Golowich et al., 2020b) and the optimistic gradient method (Golowich et al., 2020a) are both known, being $\Omega(1/T)$. Interestingly, the lower bound instances for those two methods turn out to be both bilinear problems. This suggests that bilinear problems may also provide a lower bound instance for AltGDA-NeMo. The following theorem shows that this is indeed the case.

Theorem F.1 (Theorem 4.3). *Let η, β, and L be given such that $0 < \eta L \le -\beta \le \frac{1}{2}$. Then, for any integer $T \ge 1$, there exists a convex-concave and L-smooth function f satisfying Assumption 3, where running AltGDA-NeMo with stepsize η and momentum β produces the iterates $\{(x_k, y_k)\}$ such that*

$$\min_{k=0,1,\ldots,T-1}\|\nabla f(x_k, y_k)\|^2 \ge \frac{\|x_0 - x_*\|^2 + \|y_0 - y_*\|^2}{83}\min\left\{L^2, \frac{1}{|\beta|\eta^2 T}\right\}.$$

Proof. It suffices to establish the existence of an objective such that $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, as for the general case one can fix 1-dimensional subspaces each from $\mathcal{H}$ and $\mathcal{K}$ and let the objective function vary only along those subspaces.

We claim that one of the lower bound instances that provides the claimed bound is $f(x, y) = \sigma xy$ where $\sigma = \min\left\{\frac{L}{6}, \frac{1}{6\eta\sqrt{|\beta|T}}\right\}$. To this end, throughout this proof, let us fix f as such. Being a bilinear function, it is clearly convex-concave, and the optimal solution is its stationary point $(x_*, y_*) = (0, 0)$. Moreover, as $\nabla f(x, y) = (\sigma y, \sigma x)$, it holds that

$$\|\nabla f(x, y) - \nabla f(x', y')\| = \sigma\|(y - y', x - x')\|,$$

and as $\sigma \le \frac{L}{6} \le L$ we can indeed observe that f is L-smooth.

The minimization step (2a) can then be written as

$$x_{k+1} = x_k - \eta\sigma y_k + \beta(x_k - x_{k-1}), \tag{42}$$

and plugging this into the maximization step (2b) yields

$$\begin{aligned} y_{k+1} &= y_k + \eta\sigma x_{k+1} + \beta(y_k - y_{k-1}) \\ &= y_k + \eta\sigma x_k - \eta^2\sigma^2 y_k + \eta\sigma\beta(x_k - x_{k-1}) + \beta(y_k - y_{k-1}). \end{aligned} \tag{43}$$

Let $\boldsymbol{E}_{ij}$ be the matrix whose (i, j) entry is 1 and all the other entries are 0. For the antisymmetric matrix $\boldsymbol{J} \coloneqq \boldsymbol{E}_{21} - \boldsymbol{E}_{12}$, recalling the notations $u_k = x_k - x_{k-1}$ and $v_k = y_k - y_{k-1}$ for the momenta, we can collect (42) and (43) compactly into

$$\begin{bmatrix} u_{k+1} \\ v_{k+1} \end{bmatrix} = \beta(\boldsymbol{I} + \eta\sigma\boldsymbol{E}_{21})\begin{bmatrix} u_k \\ v_k \end{bmatrix} + \eta\sigma(\boldsymbol{J} - \eta\sigma\boldsymbol{E}_{22})\begin{bmatrix} x_k \\ y_k \end{bmatrix}. \tag{44}$$

Let us use the notations $\boldsymbol{z}_k = (x_k, y_k)$ and $\boldsymbol{w}_k = (u_k, v_k)$, and set $M = \max_{0 \le k < T} \|\boldsymbol{z}_k\|$. As $\|\boldsymbol{E}_{21}\| = \|\boldsymbol{E}_{22}\| = \|\boldsymbol{I}\| = \|\boldsymbol{J}\| = 1$, we can apply the triangle inequality to (44) to obtain that

$$\begin{aligned}\|\boldsymbol{w}_{k+1}\| &\le |\beta|\big(\|\boldsymbol{I}\| + \eta\sigma\|\boldsymbol{E}_{21}\|\big)\|\boldsymbol{w}_k\| + \eta\sigma\big(\|\boldsymbol{J}\| + \eta\sigma\|\boldsymbol{E}_{22}\|\big)\|\boldsymbol{z}_k\| \\ &\le |\beta|(1+\eta\sigma)\|\boldsymbol{w}_k\| + \eta\sigma(1+\eta\sigma)M.\end{aligned}$$

Because this holds for all $k = 0, 1, \ldots, T-1$, we can unroll it to get

$$\begin{aligned}\|\boldsymbol{w}_k\| &\le |\beta|(1+\eta\sigma)\|\boldsymbol{w}_{k-1}\| + \eta\sigma(1+\eta\sigma)M \\ &\le |\beta|^2(1+\eta\sigma)^2\|\boldsymbol{w}_{k-2}\| + |\beta|\eta\sigma(1+\eta\sigma)^2 M + \eta\sigma(1+\eta\sigma)M \\ &\le \ldots \\ &\le |\beta|^k(1+\eta\sigma)^k\|\boldsymbol{w}_0\| + \eta\sigma(1+\eta\sigma)M\sum_{j=0}^{k-1}|\beta|^j(1+\eta\sigma)^j.\end{aligned}$$

Here, as $\boldsymbol{w}_0 = (0,0)$ and $\eta\sigma \le \frac{\eta L}{6} \le \frac{1}{12}$, for any $k = 0, 1, \ldots, T$ we have

$$\begin{aligned}\|\boldsymbol{w}_k\| &\le \eta\sigma(1+\eta\sigma)M\sum_{j=0}^{k-1}|\beta|^j(1+\eta\sigma)^j \\ &\le \frac{13\eta\sigma M}{12}\sum_{j=0}^{\infty}\left(\frac{13}{24}\right)^j \\ &= \frac{26\eta\sigma M}{11}.\end{aligned} \tag{45}$$

Meanwhile, let us now observe that

$$\begin{aligned}\|\boldsymbol{z}_{k+1}\|^2 - \|\boldsymbol{z}_k\|^2 &= x_{k+1}^2 - x_k^2 + y_{k+1}^2 - y_k^2 \\ &= (x_{k+1} - x_k)(x_{k+1} + x_k) + (y_{k+1} - y_k)(y_{k+1} + y_k) \\ &= (-\eta\sigma y_k + \beta(x_k - x_{k-1}))(x_{k+1} + x_k) \\ &\quad + (\eta\sigma x_{k+1} + \beta(y_k - y_{k-1}))(y_{k+1} + y_k) \\ &= \eta\sigma(x_{k+1}y_{k+1} - x_k y_k) + \beta\left\langle \boldsymbol{z}_k - \boldsymbol{z}_{k-1}, \boldsymbol{z}_{k+1} + \boldsymbol{z}_k\right\rangle \\ &= \eta\sigma(x_{k+1}y_{k+1} - x_k y_k) \\ &\quad + \beta\Big(\langle \boldsymbol{z}_k - \boldsymbol{z}_{k-1}, \boldsymbol{z}_{k+1} - \boldsymbol{z}_{k-1}\rangle + \|\boldsymbol{z}_k\|^2 - \|\boldsymbol{z}_{k-1}\|^2\Big) \\ &= \eta\sigma(x_{k+1}y_{k+1} - x_k y_k) + \beta\Big(\langle \boldsymbol{w}_k, \boldsymbol{w}_{k+1} + \boldsymbol{w}_k\rangle + \|\boldsymbol{z}_k\|^2 - \|\boldsymbol{z}_{k-1}\|^2\Big) \\ &= \eta\sigma(x_{k+1}y_{k+1} - x_k y_k) \\ &\quad + \beta\left(\frac{1}{2}\|\boldsymbol{w}_{k+1} + \boldsymbol{w}_k\|^2 + \frac{1}{2}\|\boldsymbol{w}_k\|^2 - \frac{1}{2}\|\boldsymbol{w}_{k+1}\|^2 + \|\boldsymbol{z}_k\|^2 - \|\boldsymbol{z}_{k-1}\|^2\right).\end{aligned} \tag{46}$$

Hence, if we define the Lyapunov potentials

$$\mathcal{V}_k \coloneqq \|\boldsymbol{z}_k\|^2 - \eta\sigma x_k y_k + \frac{\beta}{2}\|\boldsymbol{w}_k\|^2 - \beta\|\boldsymbol{z}_{k-1}\|^2, \tag{47}$$

then by rearranging the terms in (46) and noting that $\beta < 0$, we get that

$$\mathcal{V}_{k+1} - \mathcal{V}_k = \frac{\beta}{2}\|\boldsymbol{w}_{k+1} + \boldsymbol{w}_k\|^2 \le 0. \tag{48}$$

In particular, $\{\mathcal{V}_k\}$ is indeed monotone decreasing. That said, let us seek an expression for $\mathcal{V}_k$ which involves only the quantities indexed with k. Since $\boldsymbol{w}_k = \boldsymbol{z}_k - \boldsymbol{z}_{k-1}$, we have

$$\|\boldsymbol{z}_{k-1}\|^2 = \|\boldsymbol{z}_k - \boldsymbol{w}_k\|^2 = \|\boldsymbol{z}_k\|^2 + \|\boldsymbol{w}_k\|^2 - 2\left\langle \boldsymbol{z}_k, \boldsymbol{w}_k\right\rangle.$$

Substituting this into (47) yields

$$\begin{aligned}\mathcal{V}_k &= \|\boldsymbol{z}_k\|^2 - \eta\sigma x_k y_k + \frac{\beta}{2}\|\boldsymbol{w}_k\|^2 - \beta\left(\|\boldsymbol{z}_k\|^2 + \|\boldsymbol{w}_k\|^2 - 2\left\langle \boldsymbol{z}_k, \boldsymbol{w}_k\right\rangle\right) \\ &= (1-\beta)\|\boldsymbol{z}_k\|^2 - \frac{\beta}{2}\|\boldsymbol{w}_k\|^2 - \eta\sigma x_k y_k + 2\beta\left\langle \boldsymbol{z}_k, \boldsymbol{w}_k\right\rangle\end{aligned} \tag{49}$$

as desired. Here, by the AM–GM inequality it holds that

$$\begin{aligned}|-\eta\sigma x_k y_k + 2\beta\,\langle \boldsymbol{z}_k, \boldsymbol{w}_k\rangle| &\le \eta\sigma\,|x_k y_k| + 2|\beta|\,|\langle \boldsymbol{z}_k, \boldsymbol{w}_k\rangle| \\ &\le \frac{\eta\sigma}{2}\big(x_k^2 + y_k^2\big) + |\beta|\big(\|\boldsymbol{z}_k\|^2 + \|\boldsymbol{w}_k\|^2\big) \\ &= \left(\frac{\eta\sigma}{2} - \beta\right)\|\boldsymbol{z}_k\|^2 - \beta\|\boldsymbol{w}_k\|^2 \\ &\le \left(\frac{1}{24} - \beta\right)\|\boldsymbol{z}_k\|^2 - \beta\|\boldsymbol{w}_k\|^2,\end{aligned}$$

so with (49) we get that

$$\frac{23}{24}\|\boldsymbol{z}_k\|^2 + \frac{\beta}{2}\|\boldsymbol{w}_k\|^2 \le \mathcal{V}_k \le \left(\frac{25}{24} - 2\beta\right)\|\boldsymbol{z}_k\|^2 - \frac{3\beta}{2}\|\boldsymbol{w}_k\|^2 \tag{50}$$

Let $k^\star \in \{0, 1, \dots, T-1\}$ be the index such that $M = \|\boldsymbol{z}_{k^\star}\|$, whose existence is guaranteed because M is defined as a maximum over a finite set. Then we have

$$\begin{aligned}\mathcal{V}_0 \ge \mathcal{V}_{k^\star} &\ge \frac{23}{24}\|\boldsymbol{z}_{k^\star}\|^2 + \frac{\beta}{2}\|\boldsymbol{w}_{k^\star}\|^2 \\ &\ge \frac{23M^2}{24} - \frac{1}{4}\left(\frac{13M}{66}\right)^2 \\ &= \frac{16529M^2}{17424}\end{aligned} \tag{51}$$

where the second line follows from a bound implied by (45) and $\eta\sigma \le \frac{1}{12}$, namely

$$\|\boldsymbol{w}_{k^\star}\| \le \frac{26\eta\sigma M}{11} \le \frac{13M}{66}.$$

Meanwhile, as $\boldsymbol{w}_0 = \boldsymbol{z}_0 - \boldsymbol{z}_{-1} = \boldsymbol{0}$, putting $k = 0$ in (49) leads to

$$\mathcal{V}_0 = (1-\beta)\|\boldsymbol{z}_0\|^2 - \eta\sigma x_0 y_0, \tag{52}$$

where the AM–GM inequality implies that

$$|-\eta\sigma x_0 y_0| \le \frac{\eta\sigma}{2}\big(x_0^2 + y_0^2\big) \le \frac{1}{24}\|\boldsymbol{z}_0\|^2. \tag{53}$$

Equations (52) and (53) together yield

$$\begin{aligned}\left(\frac{23}{24} - \beta\right)\|\boldsymbol{z}_0\|^2 \le \mathcal{V}_0 &\le \left(\frac{25}{24} - \beta\right)\|\boldsymbol{z}_0\|^2 \\ &\le \frac{37}{24}\|\boldsymbol{z}_0\|^2.\end{aligned} \tag{54}$$

Combining this with (51), we obtain the upper bound

$$M^2 \le \frac{26862}{16529}\|\boldsymbol{z}_0\|^2. \tag{55}$$

Now, resuming from (48), notice that a telescopic sum results in

$$\begin{aligned}\frac{2}{\beta}(\mathcal{V}_k - \mathcal{V}_0) &= \sum_{j=0}^{k-1}\|\boldsymbol{w}_{j+1} + \boldsymbol{w}_j\|^2 \\ &\le \sum_{j=0}^{k-1} 2\big(\|\boldsymbol{w}_{j+1}\|^2 + \|\boldsymbol{w}_j\|^2\big) \\ &\le \frac{2704k\eta^2\sigma^2M^2}{121} \\ &\le \frac{676kM^2}{1089|\beta|T}\end{aligned}$$

where in the second line we used Young's inequality, in the third line the bound (45), and in the last line that $\sigma \leq \frac{1}{6\eta\sqrt{|\beta|T}}$. As $\beta \leq 0$ and $k \leq T$, it follows that

$$\mathcal{V}_0 - \mathcal{V}_k \leq \frac{338M^2}{1089}.$$

Recalling that $\boldsymbol{w}_0 = \mathbf{0}$, we then can get from (50) and (54) that

$$\begin{aligned}\left(\frac{25}{24} - 2\beta\right)\|\boldsymbol{z}_k\|^2 - \frac{3\beta}{2}\|\boldsymbol{w}_k\|^2 \geq \mathcal{V}_k &\geq \mathcal{V}_0 - \frac{338M^2}{1089} \\ &\geq \left(\frac{23}{24} - \beta\right)\|\boldsymbol{z}_0\|^2 - \frac{338M^2}{1089}.\end{aligned}$$

Again, using (45), with $\eta\sigma \leq \frac{1}{12}$, to bound $\|\boldsymbol{w}_k\|$, one can moreover obtain that

$$\begin{aligned}\left(\frac{25}{24} - 2\beta\right)\|\boldsymbol{z}_k\|^2 &\geq \left(\frac{23}{24} - \beta\right)\|\boldsymbol{z}_0\|^2 - \frac{338M^2}{1089} + \frac{3\beta}{2}\|\boldsymbol{w}_k\|^2 \\ &\geq \left(\frac{23}{24} - \beta\right)\|\boldsymbol{z}_0\|^2 - \frac{338M^2}{1089} + \frac{3\beta}{2}\left(\frac{13M}{66}\right)^2 \\ &= \left(\frac{23}{24} - \beta\right)\|\boldsymbol{z}_0\|^2 - \left(\frac{338}{1089} - \frac{169\beta}{2904}\right)M^2 \\ &\geq \left(\frac{23}{24} - \beta\right)\|\boldsymbol{z}_0\|^2 - \frac{26862}{16529}\left(\frac{338}{1089} - \frac{169\beta}{2904}\right)\|\boldsymbol{z}_0\|^2 \\ &= \left(\frac{180071}{396696} - \frac{59863\beta}{66116}\right)\|\boldsymbol{z}_0\|^2 \\ &\geq \frac{180071}{413225}\left(\frac{25}{24} - 2\beta\right)\|\boldsymbol{z}_0\|^2\end{aligned}$$

where the fourth line follows from (55), and the last line follows from the fact that

$$\frac{180071}{396696} - \frac{59863\beta}{66116} = \frac{180071}{413225}\left(\frac{25}{24} - 2\beta\right) - \frac{56007\beta}{1652900}.$$

Therefore, for any $k = 0, 1, \ldots, T-1$, it holds that

$$\|\boldsymbol{z}_k\|^2 \geq \frac{180071}{413225}\|\boldsymbol{z}_0\|^2.$$

Finally, as $\|\nabla f(x_k, y_k)\|^2 = \|(\sigma y_k, \sigma x_k)\|^2 = \sigma^2(x_k^2 + y_k^2) = \sigma^2\|\boldsymbol{z}_k\|^2$, we get that

$$\begin{aligned}\min_{k=0,1,\ldots,T-1}\|\nabla f(x_k, y_k)\|^2 &= \min_{k=0,1,\ldots,T-1}\sigma^2\|\boldsymbol{z}_k\|^2 \\ &\geq \frac{180071}{413225}\sigma^2\|\boldsymbol{z}_0\|^2 \\ &= \frac{180071}{413225}\min\left\{\frac{L^2}{36}, \frac{1}{36|\beta|\eta^2 T}\right\}(x_0^2 + y_0^2) \\ &= \frac{180071}{14876100}\min\left\{L^2, \frac{1}{|\beta|\eta^2 T}\right\}(x_0^2 + y_0^2).\end{aligned}$$

The claimed bound then follows from simplifying the coefficient using $\frac{180071}{14876100} > \frac{1}{83}$, and noticing that $x_0^2 + y_0^2 = (x_0 - x_*)^2 + (y_0 - y_*)^2$ because $(x_*, y_*) = (0, 0)$. □

Nonetheless, Yoon & Ryu (2021) have shown that the universal lower bound of first-order methods for the class of convex-concave objective functions is $\Omega(1/T^2)$. Several methods that attain a matching upper bound have been proposed (Lee & Kim, 2021; Sohrabi et al., 2026; Yoon & Ryu, 2021). Those are all variants of the extragradient and the optimistic gradient methods, amended to include a mechanism known as *anchoring*, which amounts to repeatedly taking a convex combination of the current iterate with the initial point. To the best of our knowledge, a clever incorporation of anchoring to existing methods seems to be the crux of achieving the accelerated rate of $O(1/T^2)$, rather than having to invent a completely new method. In accordance to this status quo, we second the question raised by Shugart et al. (2026) on whether AltGDA-NeMo can be also accelerated.

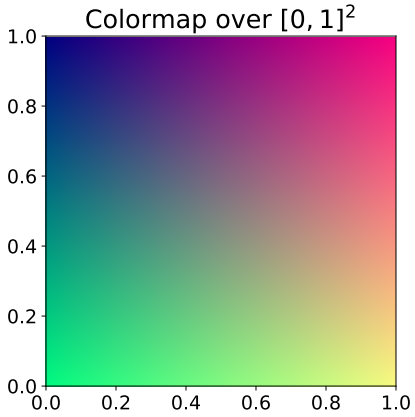


Figure 2: The *Schumann* colormap used to encode positions of the latent vectors in $[0, 1] \times [0, 1]$

G Further Experiments

Let us further investigate the practical performance of AltAdam-NeMo, in particular in comparison to the ordinary alternating Adam with the widely used standard choice $(\beta_1, \beta_2) = (0.0, 0.9)$, to see how much a possibly simplest change of just β_1 into a negative value can benefit the training.

G.1 Generative adversarial networks on toy datasets

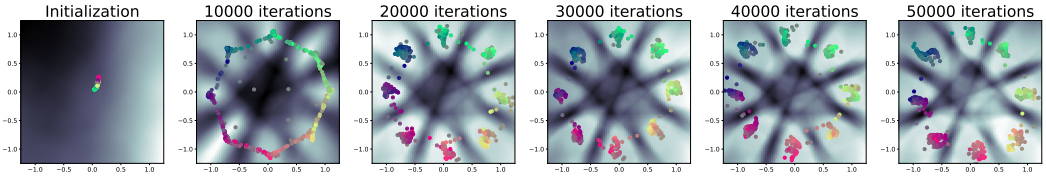
Let us first examine the performance of AltAdam-NeMo in training GANs on simple low-dimensional distributions, the so-called toy datasets. It is well known that the training of GANs can exhibit instability and the *mode collapse* phenomenon, even when learning such toy datasets, and addressing these failures has been a substantive problem in minimax optimization in its own right (Goodfellow, 2016; Metz et al., 2017). Moreover, the known structure of these toy datasets permits direct inspection of mode coverage and discrepancies between the generated and target samples, thus offering observations that can be elusive in the high-dimensional real-world datasets. Nonetheless, we did also evaluate AltAdam-NeMo on real-world datasets through *neural optimal transport* experiments, as will be reported in Section G.2.

G.1.1 Checkerboard dataset

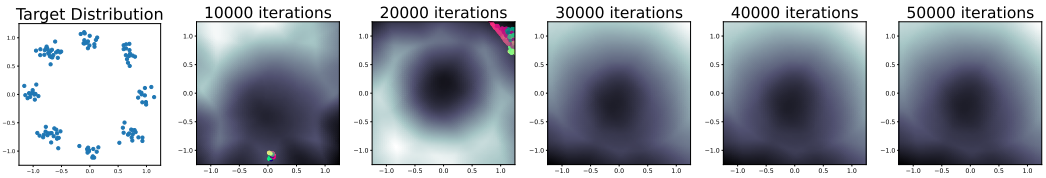
Let us first describe the settings of the experiments discussed in Section 9. We set the latent distribution to be a uniform distribution over the square $[0, 1] \times [0, 1] \subset \mathbb{R}^2$. The target distribution is the uniform distribution over the 4×4 checkerboard pattern, as shown in the lower leftmost subfigure of Figure 1.

For the generator network, we use a 3-layer multilayer perceptron, with hidden layers of width 1024 and LeakyReLU of negative slope 0.01 as activations. The discriminator network is constructed exactly the same way except that the width of the hidden layers is 256. Typically, discriminators in GANs have a final sigmoid activation layer to ensure that the outputs are confined within the range $[0, 1]$, but in our experiments, we integrated the computation of the sigmoid function to the loss function instead of docking it as a part of the network. The batch size is set to be 128, and the learning rate of the Adam optimizers is set to be 0.0001.

The intermediate and the final training results are shown in Figure 1. The outputs from the generator network are depicted by scatter plots, where the color of each marker represents where the position of the corresponding latent vector is following the colormap in Figure 2. The shades in the background of each plot are depictions of the outputs from the discriminator network.



(a) AltAdam-NeMo with $(\beta_1, \beta_2) = (-0.5, 0.9)$



(b) Ordinary alternating Adam with $(\beta_1, \beta_2) = (0.0, 0.9)$

Figure 3: Training of GANs on the 8 Gaussians dataset

G.1.2 8 GAUSSIANS DATASET

To show that the advantage of AltAdam-NeMo is not limited to a certain dataset, we run the same experiment with all the specifics preserved as in Section G.1.1, but with the dataset changed to the 8 Gaussians dataset, another popular toy dataset for benchmarking GANs.

First, we tested with the distribution where the eight Gaussian modes each have a variance of 0.1^2 and are placed so that their centers are along the unit circle; *i.e.*, placed at a distance 1 apart from the origin. The results are plotted in Figure 3.

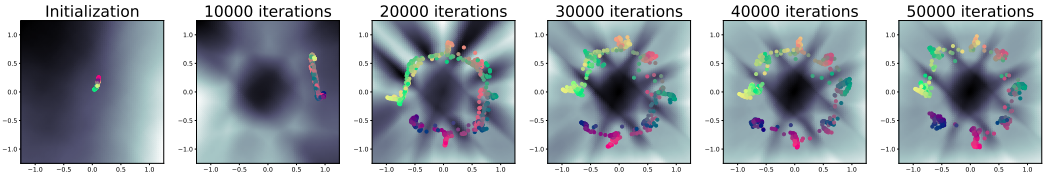
Notably, the scatter plots for ordinary alternating Adam from 30k to 50k iterations seem to be missing. This is because the training *diverged*, in the sense that the distribution generated by the generator network drifted far away from the origin and did not come back. On the other hand, networks trained by AltAdam-NeMo are stably well trained, in particular successfully recovering all eight modes.

While this stark contrast already effectively demonstrates the potential benefits of AltAdam-NeMo over ordinary alternating Adam, it would also be informative to compare the two methods in a setting where ordinary alternating Adam learns the target distribution reasonably well. To this end, to forcefully enhance the stability of the training, we change the 8 Gaussians dataset by scaling down by a factor of 0.8, and add a $\tanh$ layer to the generator network as its final layer to confine the outputs within the square $[-1, 1] \times [-1, 1]$. This resulted in the plots shown in Figure 4. As we can see, training by AltAdam-NeMo has also ended up reasonably well. At 50k iterations, ordinary alternating Adam and AltAdam-NeMo produce comparable results overall, although visual inspection suggests that AltAdam-NeMo yields more clearly separated modes and better coverage within each mode.

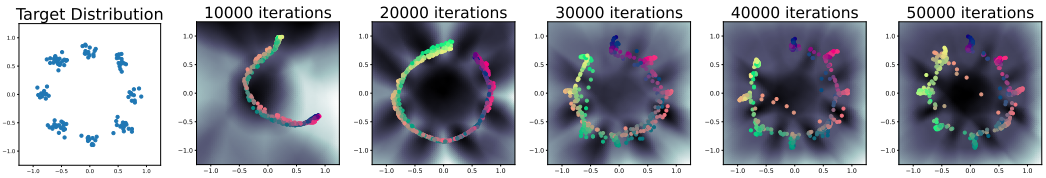
G.2 NEURAL OPTIMAL TRANSPORT

Neural optimal transport is a machine learning task where we aim to train a neural network so that it approximates an optimal transport map between two given probability distributions. Among several approaches, the one based on the semidual formulation of the optimal transport problem (Fan et al., 2022; Rout et al., 2022), which casts the task into a minimax optimization, has shown highly promising progress (Choi et al., 2023; 2025a;b).

Neural optimal transport is a rapidly growing field, and there are several different approaches even within the realm of minimax formulation. Here, to test AltAdam-NeMo, we specifically focus on the *optimal transport plan* (OTP) model, a framework proposed by Choi et al. (2025b). More precisely, given a source distribution ξ and a target distribution ν, and the cost function c used in defining the optimal transport problem, we aim to train two neural networks, namely the *optimal transport map*



(a) AltAdam-NeMo with $(\beta_1, \beta_2) = (-0.5, 0.9)$



(b) Ordinary alternating Adam with $(\beta_1, \beta_2) = (0.0, 0.9)$

Figure 4: Training of GANs on the rescaled 8 Gaussians dataset

network T and the *potential* network V, via a maximin problem

$$\max_V \min_T \ \mathbb{E}_{x\sim\xi}\left[c(x,T(x))^2 - V(T(x))\right] + \mathbb{E}_{y\sim\nu}\left[V(y)\right].$$

However, oftentimes the optimal transport map from ξ to ν may not exist, depending on the choice of ξ, ν, and c, while only an optimal transport plan (*i.e.*, a *stochastic* optimal transport map) does. For this, Choi et al. (2025b) propose a smoothing approach, where instead of using ξ per se we use the resulting distribution from taking the convolution of ξ with a Gaussian distribution, and show that this smoothing approach indeed enables the learning of an optimal transport plan. Therein, two ways to smooth ξ are suggested; a direct Gaussian convolution $\xi_\varepsilon = \xi * \mathcal{N}(0, \varepsilon \boldsymbol{I})$ and a *variance-preserving* convolution $\xi_\varepsilon = (\sqrt{1-\varepsilon}\boldsymbol{I})_\sharp \xi * \mathcal{N}(0, \varepsilon \boldsymbol{I})$. As a larger ε makes the learning of the optimal transport plan easier while having ξ_ε close to ξ requires small ε, we schedule ε so that early in the training we use large ε, whereas it gradually decreases. For more details we refer to the original paper (Choi et al., 2025b).

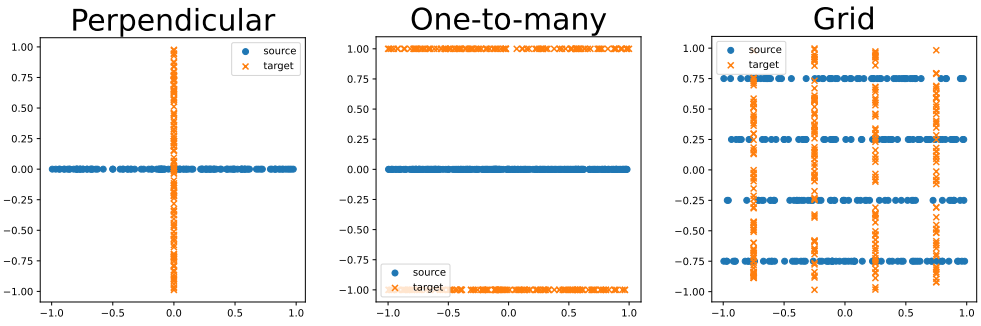
G.2.1 Toy datasets

First, we train the OTP model on the tasks involving toy datasets in $\mathbb{R}^2$, whose source and target distributions are configured as follows, and compare how the optimizers train the models.

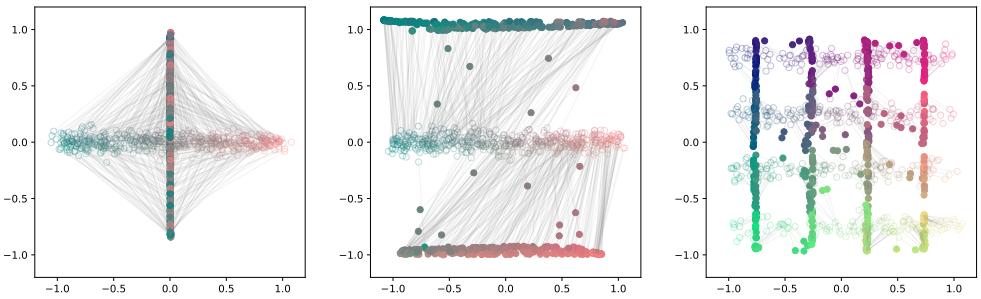
- Perpendicular
  - source: uniform distribution over the horizontal line $[-1, 1] \times \{0\}$
  - target: uniform distribution over the vertical line $\{0\} \times [-1, 1]$
- One-to-many
  - source: uniform distribution over a horizontal line $[-1, 1] \times \{0\}$
  - target: uniform distribution over two horizontal lines $[-1, 1] \times \{-1, 1\}$
- Grid
  - source: uniform distribution over four horizontal lines $[-1, 1] \times \{\pm\frac{1}{4}, \pm\frac{3}{4}\}$
  - target: uniform distribution over four vertical lines $\{\pm\frac{1}{4}, \pm\frac{3}{4}\} \times [-1, 1]$

The architectures we use for the optimal transport map network T and the potential network V are exactly the same as those of the generator network and the discriminator network, respectively, in the previous GAN experiments in Section G.1.

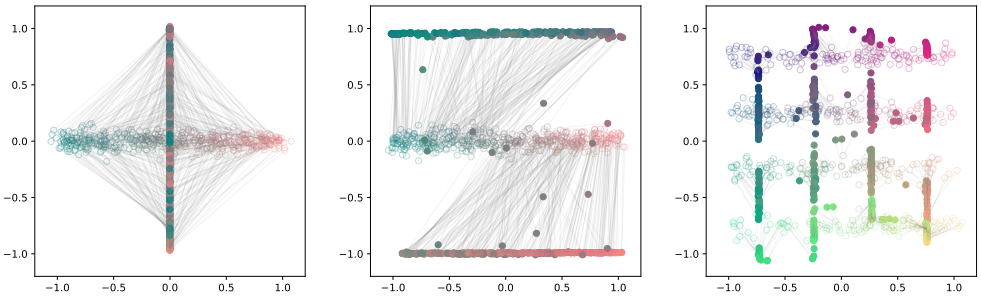
For each of the runs, the batch size is 256, and the learning rates for the Adam optimizers are set to be 0.0001. The total number of iterations is set to be $K = 20000$, and the scheduling of ε is done



(a) Sampled points from the source and target distributions for the experiments in Section G.2.1



(b) AltAdam-NeMo with $(\beta_1, \beta_2) = (-0.5, 0.9)$



(c) Ordinary alternating Adam with $(\beta_1, \beta_2) = (0.0, 0.9)$

Figure 5: Optimal transport plan models on 2D toy datasets

with a cosine schedule. More precisely, if we let ε_k be the value of ε at the kth iteration, then for the initial noise level $\varepsilon_{\text{init}} = 0.2$ and the terminal noise level $\varepsilon_{\text{term}} = 0.05$, we set

$$\varepsilon_k = \varepsilon_{\text{term}} + (\varepsilon_{\text{init}} - \varepsilon_{\text{term}}) \left(\frac{1}{2} + \frac{1}{2} \cos \left(\frac{k\pi}{K} \right) \right).$$

This is slightly different from the noise schedule used by Choi et al. (2025b) on the same tasks, but it makes ε decrease gradually, while also being easier to implement.

The results are shown in Figure 5. The three subplots in the first row depict some sample points from the source distribution, plotted with blue circles, and from the target distribution, plotted with orange crosses. The next three subplots in the second row depict the results when training is done using AltAdam-NeMo. The hollow circles are the sample points from the source distribution, colored by their positions in $\mathbb{R}^2$ following the 2D colormap[2] in Figure 2. The filled circles are the generated

[2]To be more precise, because the domain of the colormap is $[0, 1] \times [0, 1]$, we rescale the samples from the source distribution by an affine map $[-1, 1] \times [-1, 1] \to [0, 1] \times [0, 1]$ when assigning colors to them.

Table 1: FID scores obtained in image translation tasks

Task	AltAdam-NeMo	Ordinary alternating Adam
MNIST → CMNIST	4.52 ± 1.75 (256 channels) 8.84 ± 1.58 (64 channels)	75.81 ± 18.42 (256 channels) 15.62 ± 5.02 (64 channels)
Wild → Cat	9.05	9.36

points, obtained by passing the samples from the source distribution through the optimal transport map network T. The half-transparent lines depict that mapping by T. The last three subplots in the third row are drawn in the same way as the subplots in the second row, but they depict the results of training with ordinary alternating Adam.

Our interpretation of Figure 5 is twofold. On one hand, in the Grid task, AltAdam-NeMo shows a visually clear advantage over the ordinary alternating Adam, as in the previous GAN examples. On the other hand, in the One-to-many task, it also visually seems that ordinary alternating Adam resulted in a slightly better performing model than AltAdam-NeMo. That said, the model trained by AltAdam-NeMo still arguably performs comparably. Hence, we cannot say that AltAdam-NeMo is an unconditionally superior optimizer compared to ordinary alternating Adam; indeed, in training neural networks, it is almost impossible for an optimizer to be unconditionally superior to another. Nonetheless, at least the empirical results we have seen so far, and also the ones in the following section, assert that AltAdam-NeMo in general gives a better result, and if not at least a result that is on par, compared to ordinary alternating Adam.

G.2.2 Image-to-image translation

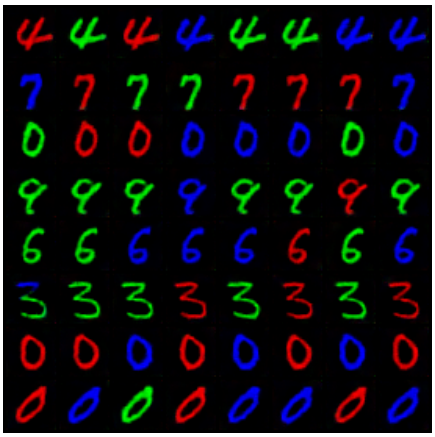
We also train the OTP model on image-to-image translation tasks, which is one of the most representative tasks in optimal transport. We consider two tasks, MNIST-to-CMNIST and Wild-to-Cat, which have also been studied by Choi et al. (2025b).

We exactly follow the configuration described by Choi et al. (2025b), except that in the Wild-to-Cat task, ε is reduced every 600 training iterations instead of every 100 iterations as specified by $P = 100$ in equation (39) of their work. Still, both optimizers use the same experimental configuration within each comparison. The results we report should be interpreted as comparisons within these experimental settings, and the specifics, such as FID scores, would differ from those reported in Choi et al. (2025b).

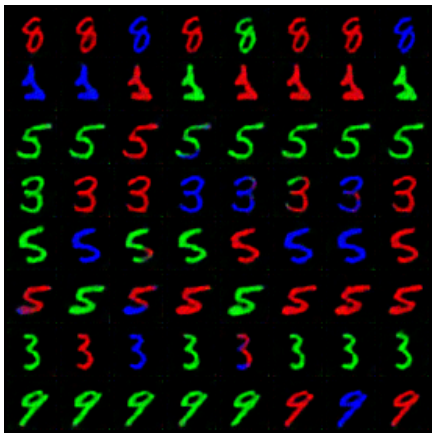
For the MNIST-to-CMNIST task, we ran the experiment five times and computed the final FIDs. We, however, should state that, when we tried to replicate the experiments in Choi et al. (2025b), ordinary alternating Adam suffered from the mode collapse phenomenon, resulting in the digits being colored all by the same color, as depicted in Figure 6c. This yielded an FID of 75.81 ± 18.42. To remedy this issue, we tried changing the model architecture, and in the end, changing the number of channels from 256 to 64 fixed the mode collapse. This mode collapse issue did not happen with AltAdam-NeMo, but to have a fair comparison, we ran AltAdam-NeMo on the models with 256 channels and 64 channels also.

The mean and the standard deviation across the five runs are reported in the first row of Table 1, and the qualitative results are depicted in Figure 6. AltAdam-NeMo shows more robust training, avoiding mode collapse, and in addition consistently shows better quantitative results even in settings where mode collapse did not occur. One run using AltAdam-NeMo even resulted in the model with FID 2.84, as shown in Figure 6a, which improves upon the FID of 3.18 reported by Choi et al. (2025b).

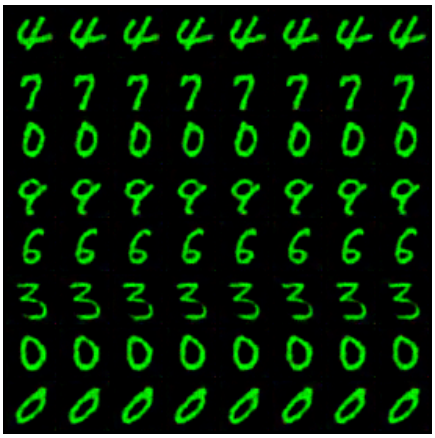
For the Wild-to-Cat task, the much larger networks limited us to a single training run per optimization method within our computational budget. We thus evaluated the saved checkpoints from each run, and report the lowest FID obtained in the second row of Table 1. The generated images from each of the lowest FID checkpoints are shown in Figure 7.



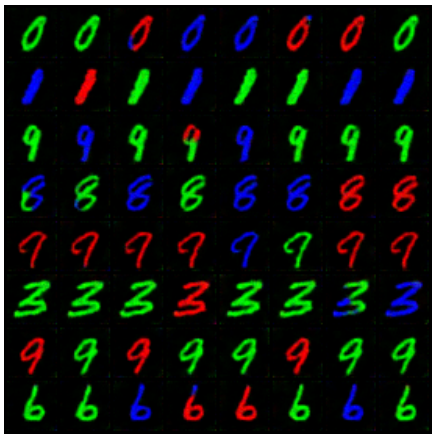
(a) AltAdam-NeMo with $(\beta_1, \beta_2) = (-0.5, 0.9)$
256-channel model, FID 2.84



(b) AltAdam-NeMo with $(\beta_1, \beta_2) = (-0.5, 0.9)$
64-channel model, FID 7.44



(c) Alternating Adam with $(\beta_1, \beta_2) = (0.0, 0.9)$
256-channel model, FID 50.06



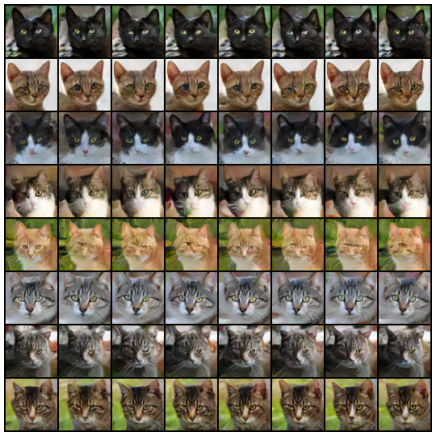
(d) Alternating Adam with $(\beta_1, \beta_2) = (0.0, 0.9)$
64-channel model, FID 8.01

Figure 6: Results of the MNIST to CMNIST translation task with 32×32 images

The reported FIDs and the qualitative results show that AltAdam-NeMo is a competitive choice in the training of neural networks via minimax optimization, not only on toy datasets but also in practical tasks involving real-world datasets.



(a) AltAdam-NeMo with $(\beta_1, \beta_2) = (-0.5, 0.9)$



(b) Alternating Adam with $(\beta_1, \beta_2) = (0.0, 0.9)$

Figure 7: Results of the Wild to Cat translation task with 64×64 images