SOLVE MINIMAX OPTIMIZATION BY ANDERSON ACCELERATION

Huan He, Shifan Zhao, Yuanzhe Xi, Joyce C Ho ∗ Yousef Saad †
Department of Computer Science Department of Computer Science and Engineering
Emory University University of Minnesota
Atlanta, GA 30329, USA Minneapolis, MN 55455, USA

ABSTRACT

Many modern machine learning algorithms such as generative adversarial networks (GANs) and adversarial training can be formulated as minimax optimization. Gradient descent ascent (GDA) is the most commonly used algorithm due to its simplicity. However, GDA can converge to non-optimal minimax points. We propose a new minimax optimization framework, GDA-AM, that views the GDA dynamics as a fixed-point iteration and solves it using Anderson Mixing to converge to the local minimax. It addresses the diverging issue of simultaneous GDA and accelerates the convergence of alternating GDA. We show theoretically that the algorithm can achieve global convergence for bilinear problems under mild conditions. We also empirically show that GDA-AM solves a variety of minimax problems and improves adversarial training on several datasets. Codes are available on Github.

1 INTRODUCTION

Minimax optimization has received a surge of interest due to its wide range of applications in modern machine learning, such as generative adversarial networks (GAN), adversarial training and multi-agent reinforcement learning [Goodfellow et al., 2014; Madry et al., 2018; Li et al., 2019]. Formally, given a bivariate function \( f(x, y) \), the objective is to find a stable solution where the players cannot improve their objective, i.e., to find the Nash equilibrium of the underlying game [von Neumann & Morgenstern, 1944]:

\[
\arg\min_{x \in X} \arg\max_{y \in Y} f(x, y).
\] (1)

It is commonplace to use simple algorithms such as gradient descent ascent (GDA) to solve such problems, where both players take a gradient update simultaneously or alternatively. Despite its simplicity, GDA is known to suffer from a generic issue for minimax optimization: it may cycle around a stable point, exhibit divergent behavior, or converge very slowly since it requires very small learning rates [Gidel et al., 2019a; Mertikopoulos et al., 2019]. Given the widespread usage of gradient-based methods for solving machine learning problems, first-order optimization algorithms to solve minimax problems have gained considerable popularity in the last few years. Algorithms such as optimistic Gradient Descent Ascent (OG) [Daskalakis et al., 2018; Mertikopoulos et al., 2019] and extra-gradient (EG) [Gidel et al., 2019a] can alleviate the issue of GDA for some problems. Yet, it has been shown that these methods can still diverge or cycle around a stable point [Adolphs et al., 2019; Mazumdar et al., 2019; Parker-Holder et al., 2020]. For example, these algorithms even fail to find a local minimax (the set of local minimax is a superset of local Nash [Jin et al., 2020; Wang et al., 2020]) as shown in Figure 1. This leads to the following question: Can we design better algorithms for minimax problems? We answer this in the affirmative, by introducing GDA-AM. We cast the GDA dynamics as a fixed-point iteration problem and compute the iterates effectively using an advanced nonlinear extrapolation method. We show that indeed our algorithm has theoretical and empirical guarantees across a broad range of minimax problems, including GANs.

∗ hhe37, szhao89, yxi26, jho31@emory.edu
† saad@umn.edu

https://github.com/hehuannb/GDA-AM
We also provide empirical results for GAN training across two different datasets, CIFAR10 and CelebA. Given the limited computational overhead of our method, the results suggest that an extrapolation add-on to GDA can lead to significant performance gains. Moreover, the convergence behavior across a variety of problems and the ease-of-use demonstrate the potential of GDA-AM with table size $p = 3$ always exhibits desirable behaviors.

Our contributions: In this paper, we propose a different approach to solve minimax optimization. Our starting point is to cast the GDA dynamics as a fixed-point iteration. We then highlight that the fixed-point iteration can be solved effectively by using advanced non-linear extrapolation methods such as Anderson Mixing (Anderson, 1965), which we name as GDA-AM. Although first mentioned in Azizian et al. (2020), to our best knowledge, this is still the first work to investigate and improve the GDA dynamics by tapping into advanced fixed-point algorithms. We demonstrate that GDA dynamics can benefit from Anderson Mixing. In particular, we study bilinear games and give a systematic analysis of GDA-AM across a variety of problems and the ease-of-use demonstrate the potential of GDA-AM can achieve global convergence guarantees under mild conditions.

We complement our theoretical results with numerical simulations across a variety of minimax problems. We show that for some convex-concave and non-convex-concave functions, GDA-AM can converge to the optimal point with little hyper-parameter tuning whereas existing first-order methods are prone to divergence and cycling behaviors.

We also provide empirical results for GAN training across two different datasets, CIFAR10 and CelebA. Given the limited computational overhead of our method, the results suggest that an extrapolation add-on to GDA can lead to significant performance gains. Moreover, the convergence behavior across a variety of problems and the ease-of-use demonstrate the potential of GDA-AM to become the minimax optimization workhorse.

2 PRELIMINARIES AND BACKGROUND

2.1 MINIMAX OPTIMIZATION

Definition 1. Point $(x^*, y^*)$ is a local Nash equilibrium of $f$ if there exists $\delta > 0$ such that for any $(x, y)$ satisfying $\|x - x^*\| \leq \delta$ and $\|y - y^*\| \leq \delta$ we have: $f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*)$.

To find the Nash equilibria, common algorithms including GDA, EG and OG, can be formulated as follows. For the two variants of GDA, simultaneous GDA (SimGDA) and alternating GDA (AltGDA), the updates have the following forms:

Simultaneous: $x_{t+1} = x_t - \eta_1 \nabla_x f(x_t, y_t)$, $y_{t+1} = y_t + \eta_2 \nabla_y f(x_t, y_t)$

Alternating: $x_{t+1} = x_t - \eta_1 \nabla_x f(x_t, y_t)$, $y_{t+1} = y_t + \eta_2 \nabla_y f(x_{t+1}, y_t)$.

(2)

The EG update has the following form:

$x_{t+\frac{1}{2}} = x_t - \eta \nabla_x f(x_t, y_t)$, $y_{t+\frac{1}{2}} = y_t + \eta \nabla_y f(x_t, y_t)$

$x_{t+1} = x_t - \eta \nabla_x f(x_{t+\frac{1}{2}}, y_{t+\frac{1}{2}})$, $y_{t+1} = y_t + \eta \nabla_y f(x_{t+\frac{1}{2}}, y_{t+\frac{1}{2}})$.

(3)
The OG update has the following form:

\[ x_{t+1} = x_t - \eta \nabla_x f(x_t, y_t) + \frac{\eta}{2} \nabla_x f(x_{t-1}, y_{t-1}), \]

\[ y_{t+1} = y_t + \eta \nabla_y f(x_t, y_t) - \frac{\eta}{2} \nabla_y f(x_{t-1}, y_{t-1}). \]

(4)

2.2 Fixed-Point Iteration and Anderson Mixing (AM)

**Definition 2.** \( w^* \) is a fixed point of the mapping \( g \) if \( w^* = g(w^*) \).

Consider the simple fixed-point iteration \( w_{t+1} = g(w_t) \) which produces a sequence of iterates \( \{w_0, w_1, \ldots, w_N\} \). In most cases, this converges to the fixed-point, \( w^* = g(w^*) \). Take gradient descent as an example, it can be viewed as iteratively applying the operation: \( w_{t+1} = g(w_t) \). We know that the limit is the fixed-point \( w^* = g(w^*) \) (i.e. \( \nabla f(w_t) = 0 \)). SimGDA updates can be defined as the repeated application of a nonlinear operator:

\[ w_{t+1} = G_{w_t} \triangleq w_t - \eta V(w_t) \text{ with } w = [x, y], V(w) = \left[ \nabla_x f(x, y), -\nabla_y f(x, y) \right] \]

Similarly, we can write AltGDA updates as \( w_{t+1} = G_{w_t}^{(alt)} \). An issue with fixed-point iteration is that it does not always converge, and even in the cases where it does converge, it might do so very slowly. GDA is one example that it could result in the possibility of the operator converging to a limit cycle instead of a single point for the GDA dynamic. A way of dealing with these problems is to use acceleration methods, which can potentially speed up the convergence process and in some cases even decrease the likelihood for divergence.

There are many different acceleration methods, but we will put our focus on an algorithm which we refer to as Anderson Mixing (or Anderson Acceleration). In short, Anderson Mixing (AM) shares the same idea as Nesterov’s acceleration. Given a fixed-point iteration \( w_t = g(w_{t-1}) \), Anderson Mixing argues that a good approximation to the final solution \( w^* \) can be obtained as a linear combination of the previous \( p \) iterates \( w_{t+1} = \sum_{i=t-p}^{t} \beta_i g(w_{t-p+i}) \). Since obtaining the proper coefficients \( \beta_i \) is a nonlinear procedure, Anderson Mixing is also known as a nonlinear extrapolation method. The general form of Anderson Mixing is shown in Algorithm 1. For efficiency, we prefer a ‘restarted’ version with a small table size \( p \) that cleans up the table \( F \) every \( p \) iterations because it avoids solving a linear system of increasing size.

**Algorithm 1:** Anderson Mixing Prototype (truncated version)

**Input:** Initial point \( w_0 \), Anderson restart dimension \( p \), fixed-point mapping \( g : \mathbb{R}^n \rightarrow \mathbb{R}^n \).

**Output:** \( w_{t+1} \)

for \( t = 0, 1, \ldots \) do

\[ p_t = \min \{t, p\} \]

Set \( F_t = [f_{t-p}, \ldots, f_1] \), where \( f_i = g(w_i) - w_i \) for each \( i \in \{t - p_t \ldots t\} \).

Determine weights \( \beta = (\beta_0, \ldots, \beta_{p_t})^T \) that solves \( \min_{\beta} \| F_t \beta \|_2 \), s. t. \( \sum_{i=0}^{p_t} \beta_i = 1 \).

Set \( w_{t+1} = \sum_{i=0}^{p_t} \beta_i g(w_{t-p_t+i}) \).

end

2.3 AM and Generalized Minimal Residual (GMRES)

Developed by Saad & Schultz [1986], Generalized Minimal Residual method (GMRES) is a Krylov subspace method for solving linear system equations. The method approximates the solution by the vector in a Krylov subspace with minimal residual, which is described below.

**Definition 3.** Assume we have the linear system of equations \( Ax = b \) with \( A \in \mathbb{R}^{n \times n} \), \( b \in \mathbb{R}^n \) and an initial guess \( x_0 \). Then we denote the initial residual by \( r_0 = b - Ax_0 \) and define the \( t \)th Krylov subspace as \( \mathcal{K}_t = \text{span}\{r_0, Ar_0, \ldots, A^{t-1}r_0\} \).

The \( t \)th iterate \( x_t \) of GMRES minimizes the norm of the residual \( r_t = b - Ax_t \) in \( \mathcal{K}_t \), that is, \( x_t \) solves

\[ \min_{x_t \in \mathcal{K}_t} \| b - Ax_t \|_2 \].
We propose to exploit the dynamic information present in the GDA iterates to “smartly” combine the past iterates. We propose a novel minimax optimizer, called \( \text{GDA-AM} \), that is inspired by recent advances in parameter (or weight) averaging (Wu et al., 2020; Yazici et al., 2019). We argue that a nonlinear adaptive average (combination) is a more appropriate choice for minimax optimization.

Although GMRES is applied to a system of linear equations not a fixed-point problem, there is a strong connection between Anderson Mixing and GMRES. In AM we are looking for a fixed-point \( \tilde{x} \) such that \( G\tilde{x} = 0 \) and by rearranging this equation we get

\[
\tilde{x} = (I - G)^{-1}b.
\]

Theorem 2.2 shows that if GMRES is applied to the system \((I - G)\tilde{x} = b\) and AM is applied to \(g(x) = Gx + b\) with the same initial guess and \(I - G\) is non-singular, then these are equivalent in the sense that the iterates of each algorithm can be obtained directly from the iterates of the other algorithm.

Theorem 2.2 can also be generalized to the restart version of AM and GMRES as well.

**3 GDA-AM: GDA WITH ANDERSON MIXING**

We propose a novel minimax optimizer, called \( \text{GDA-AM} \), that is inspired by recent advances in parameter (or weight) averaging (Wu et al., 2020; Yazici et al., 2019). We argue that a nonlinear adaptive average (combination) is a more appropriate choice for minimax optimization.

**3.1 GDA with Naïve Anderson Mixing**

We propose to exploit the dynamic information present in the GDA iterates to “smartly” combine the past iterates. This is in contrast to the classical averaging methods (moving averaging and exponential moving averaging) (Yang et al., 2019) that “blindly” combine past iterates. A naïve adoption of Anderson Mixing using the past \( p \) GDA iterates for both SimGDA and AltGDA has the following form:

\[
\text{Anderson mixing: } x_{t+1} = \sum_{i=0}^{p} \beta_i x_{t-p+i}, y_{t+1} = \sum_{i=0}^{p} \beta_i y_{t-p+i}.
\]

Since Zhang et al. (2021); Gidel et al. (2019b) show the AltGDA is superior to SimGDA in many aspects, we briefly summarized both Simultaneous and Alternating GDA-AM in Algorithms 2 and 3 with the truncated Anderson Mixing Algorithm using a table size \( p \).
We also investigate bilinear-quadratic games from a spectrum analysis perspective. In addition, we show that analysis based on the numerical range (Bollapragada et al., 2018) can be also extended to zero-sum bilinear games: Simultaneous GDA can be written in the following matrix form: Suppose $x_0$ and $y_0$ can be found in Appendix C.1 and C.4.1. such games, although it can not help derive a convergent bound for equation 7. Detailed discussion shows that it is important to note that the Anderson Mixing form shown in Algorithm 1 is for illustrative purpose and not computationally efficient. For example, only one column of $F_t$ needs to be updated at each iteration. In addition, the solution of the least-square problem in Algorithm 1 can also be solved by a quick QR update scheme which costs $(2n + 1)p^2$ (Walker & Ni, 2011a). Thus, from Algorithms 2 and 3 we can see that the major cost of GDA-AM arises from solving the additional linear least squares problem compared to regular GDA at each iteration. Additional implementation details are provided in the Appendix.

4 Convergence Results for GDA-AM

In this section, we show that both simultaneous and alternating version GDA-AM converge to the equilibrium for bilinear problems. First, we do not require the learning rate to be sufficiently small. Second, we explicitly provide a linear convergence rate that is faster than EG and OG. More importantly, we derive nonasymptotic rates from the spectrum analysis perspective because existing theoretical results can not help us derive a convergent rate (see C.1).

4.1 Bilinear Games

Bilinear games are often regarded as an important simple example for theoretically analyzing and understanding new algorithms and techniques for solving general minimax problems (Gidel et al., 2019a; Merrikh-Bayat et al., 2019; Schaefer & Anandkumar, 2019). In this section, we analyze the convergence property of simultaneous GDA-AM and alternating GDA-AM schemes on the following zero-sum bilinear games:

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^n} f(x, y) = x^T A y + b^T x + c^T y, \quad A \text{ is full rank.}$$

(7)

The Nash equilibrium to the above problem is given by $(x^*, y^*) = (-A^{-T} c, -A^{-1} b)$.

We also investigate bilinear-quadratic games from a spectrum analysis perspective. In addition, we show that analysis based on the numerical range (Bollapragada et al., 2018) can be also extended to such games, although it can not help derive a convergent bound for equation 7. Detailed discussion can be found in Appendix C.1 and C.4.1.

4.2 Simultaneous GDA-AM

Suppose $x_0$ and $y_0$ are the initial guesses for $x^*$ and $y^*$, respectively. Then each iteration of simultaneous GDA can be written in the following matrix form:

$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} I & -\eta A^T \\ \eta A & I \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} - \eta \begin{bmatrix} b \\ c \end{bmatrix} + w_{t+1},$$

(8)
It has been shown that the iteration in equation 8 often cycles and fails to converge for the bilinear problem due to the poor spectrum/numerical range of the fixed point operator $G^{(Sim)}$ (Gidel et al., 2019a; Azizian et al., 2020; Mokhtari et al., 2020a). Next we show that the convergence can be improved with Algorithm 2.

**Theorem 4.1.** [Global convergence for simultaneous GDA-AM on bilinear problem] Denote the distance between the stationary point $w^*$ and current iterate $w_{(k+1)p}$ of Algorithm 2 with table size $p$ as $N_{(k+1)p} = \|w^* - w_{(k+1)p}\|$. Then we have the following bound for $N_t$:

$$N_{t+1}^2 \leq \rho(A) N_t^2,$$

where $\rho(A) = \left( \frac{1}{T_p(1+\kappa(A)^2-1)} \right)^2$. Here, $T_p$ is the Chebyshev polynomial of first kind of degree $p$ and $\frac{1}{T_p(1+\kappa(A)^2-1)} < 1$ since $1 + \frac{\kappa^2(A)^2}{\kappa(A)^2-1} > 1$.

It is worthy emphasizing that the convergence rate of Algorithm 2 is independent of learning rate $\eta$ while the convergence results of other methods like EG and OG depend on the learning rate.

**Remark 4.1.1.** Both EG and OG have the following form of convergence rate (Mokhtari et al., 2020a) for bilinear problem

$$N_{t+1}^2 \leq (1 - \frac{c}{\kappa(A)^2}) N_t^2,$$

where $c$ is a positive constant independent of the problem parameters.

![Figure 2a](image)

The blue line is the spectrum of matrix $G^{(Sim)}$ while the red line is spectrum of matrix $I - G^{(Sim)}$. Our method transforms the divergent problem to a convergent problem due to the transformed spectrum. **Figure 2b** Convergence rate comparison between SimGDA-AM and EG for different condition numbers of $A$ and fixed table size $p = 10, 20, 50$. **Figure 2c** Convergence rate comparison between SimGDA-AM and EG for increasing table size on a matrix $A$ with condition number $100$.

### 4.3 Alternating GDA-AM

The underlying fixed point iteration in Algorithm 3 can be written in the following matrix form:

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \begin{pmatrix} I & -\eta A \\ \eta A^T & I - \eta^2 A^T A \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} - \eta \begin{pmatrix} b \\ c \end{pmatrix}.$$

According to the equivalence between truncated Anderson acceleration and GMRES with restart, we can analyze the convergence of Algorithm 3 through the convergence analysis of applying GMRES to solve linear systems associated with $G = I - G^{(Alt)}$:

$$G = \begin{bmatrix} 0 & \eta A \\ -\eta A^T & \eta^2 A^T A \end{bmatrix}.$$
Theorem 4.2. [Global convergence for alternating GDA–AM on bilinear problem] Denote the distance between the stationary point $w^*$ and current iterate $w_{(k+1)p}$ of Algorithm 3 with table size $p$ as $N_{(k+1)p} = \|w^* - w_{(k+1)p}\|$. Assume $A$ is normalized such that its largest singular value is equal to 1. Then when the learning rate $\eta$ is less than $2$, we have the following bound for $N_t$:

$$N_{(k+1)p}^2 \leq \sqrt{1 + \frac{2\eta}{2 - \eta} \left(\frac{c}{r}\right)^p N_{kp}^2}$$

where $c$ and $r$ are the center and radius of a disk $D(c, r)$ which includes all the eigenvalues of $G$. Especially, $\frac{c}{r} < 1$.

Theorem 4.2 shows that when $p > \frac{\log \sqrt{2} - \frac{2\eta}{2 - \eta}}{\log \frac{c}{r}}$, alternating GDA–AM will converge globally.

4.4 Discussion of obtained rates

We would like to first explain on why taking Chebyshev polynomial of degree $p$ at the point $1 + \frac{2}{\kappa - 1}$. We evaluate the Chebyshev polynomial at this specific point because the reciprocal of this value gives the minimal value of infinite norm of the all polynomials of degree $p$ defined on the interval $I = [\eta^2 \sigma_{min}^2(A), \eta^2 \sigma_{max}^2(A)]$ based on Theorem 6.25 (page 209) (Saad, 2003). In other words, taking the function value at this point leads to the tight bound.

When comparing between existing bounds, we would like to point our our derived bounds are hard to compare directly. The numerical experiments in figure 2b numerically verify that our bound is smaller than EG. We wanted to numerically compare our rate with EG with positive momentum. However the bound of EG with positive momentum is asymptotic. Moreover, it does not specify the constants so we can not numerically compare them. We do provide empirical comparison between GDA-AM and EG with positive momentum for bilinear problems in Appendix D.1. It shows GDA-AM outperforms EG with positive momentum. Regarding alternating GDA–AM, we would like to note that the bound in Theorem 4.2 depends on the eigenvalue distribution of the matrix $G$. Condition number is not directly related to the distribution of eigenvalues of a nonsymmetric matrix $G$. Thus, the condition number is not a precise metric to characterize the convergence. If these eigenvalues are clustered, then our bound can be small. On the other hand, if these eigenvalues are evenly distributed in the complex plane, then the bound can very close to 1.

More importantly, we would like to stress several technical contributions.

1 : Our obtained Theorem 4.1 and 4.2 provide nonasymptotic guarantees, while most other work are asymptotic. For example, EG with positive momentum can achieve a asymptotic rate of $1 - O(1/\sqrt{\kappa})$ under strong assumptions (Azizian et al., 2020).

2 : Our contribution is not just about fix the convergence issue of GDA by applying Anderson Mixing; another contribution is that we arrive at a convergent and tight bound on the original work and not just adopting existing analyses. We developed Theorem 4.1 and 4.2 from a new perspective because applying existing theoretical results fail to give us neither convergent nor tight bounds.

3 : Theorem 4.1 and 4.2 only requires mild conditions and reflects how the table size $p$ controls the convergence rate. Theorem 4.1 is independent of the learning rate $\eta$. However, the convergence results of other methods like EG and OG depend on the learning rate, which may yield less than desirable results for ill-specified learning rates.

5 Experiments

In this section, we conduct experiments to see whether GDA–AM improves GDA for minimax optimization from simple to practical problems. We first investigate performance of GDA–AM on bilinear games. In addition, we evaluate the efficacy of our approach on GANs.
5.1 Bilinear Problems

In this section, we answer following questions: **Q1:** How is GDA-AM perform in terms of iteration number and running time? **Q2:** How is the scalability of GDA-AM? **Q3:** How is the performance of GDA-AM using different table size \( p \)? **Q4:** Does GDA-AM converge for large step size \( \eta \)?

We compare the performance with SimGDA, AltGDA, EG, and OG and EG with Negative Momentum (Azizian et al. (2020)) on bilinear minimax games shown in equation 7 without any constraint. \( A, b, c \), and initial points are generated using normally distributed random number. We set the maximum iteration number as \( 1 \times 10^6 \), stopping criteria \( 1 \times 10^{-5} \) and depict convergence by use of the norm of distance to optima, which is defined as \( ||w^* - w_t|| \). Similarly to Azizian et al. (2020); Wei et al. (2021a), the step size is set as 1 after rescaling \( A \) to have 2-norm 1. We present results of different settings in Figures 4, 5, and 6.

We first generate different problem size \( (n = 100, 1000, 5000) \) and present results of convergence in terms of iteration number in Figure 4. It can be observed that GDA-AM converges in much fewer iterations for different problem sizes. Note that EG, EG-NM, and OG converge in the end but requires many iterations, thus we plot only a portion for illustrative purposes. Figure 5 depicts the convergence for all methods in terms of time. It can be observed that the running time of GDA-AM is faster than EG. Although slower than OG, we can observe GDA-AM converges in much less time for all problems. Figure 4 and Figure 5 answer Q1 and Q2; although there is additional computation for GDA-AM, it does not hinder the benefits of adopting Anderson Mixing. Even for a large problem size, GDA-AM still converges in much less time than the baselines.

Next, we run GDA-AM using different table size \( p \) and show the results in Figure 6a and Figure 6b. Figure 6a indicates an increasing of table size results in faster convergence in terms of iteration number, which also verifies our claim in Theorem 4.1. However, we also observe an increased running time when using a larger table size in Figure 6b. Further, we can see that \( p = 50 \) converges in a comparable time and iterations to \( p = 100 \). Similar results are found in repeated experiments as well. As a result, our answer to Q3 is that although a larger \( p \) means less iterations, a medium \( p \) is sufficient and a small \( p \) still outperforms the baselines. The optimal choice of \( p \) is related to the condition number and step size, which is another interesting topic in the Anderson Mixing community.

Next, we answer Q4 on convergence under different step sizes. Although GDA-AM usually converges with suitable step size, our theorem suggests it requires a larger table size when combined with a extremely aggressive step size. Figure 6c shows the convergence under such circumstance. We can observe that although a very large step size goes the wrong way in the beginning, Anderson Mixing can still make it back on track except when \( \eta > 1 \). It answers the question and confirms our claim that GDA-AM can achieve global convergence for bilinear problems for a large step size \( \eta > 0 \).

![Figure 4: Comparison in terms of iteration:](image)

Figure 4: Comparison in terms of iteration: \( \min_x \max_y f(x, y) = x^T A y + b^T x + c^T y \). We use different problem size and fix \( p = 10, \eta = 1 \) for all experiments.

5.2 GAN Experiments: Image Generation

We apply our method to the CIFAR10 dataset (Krizhevsky, 2009) and use the ResNet architecture with WGAN-GP (Gulrajani et al., 2017) and SNGAN (Miyato et al., 2018) objective. We also compared
Published as a conference paper at ICLR 2022

Figure 5: Comparison between methods in terms of time.

(a) Effects of $p$ in terms of iteration (b) Time comparison for Figure 4a (c) Effect of step size $\eta$, $p = 10$

Figure 6: Effects of table size $p$ and step size $\eta$, $n = 500$

Table 1: Best inception scores and FID for Cifar10 and FID for CelebA (IS is a less informative metric for celebA).

<table>
<thead>
<tr>
<th>Method</th>
<th>IS  $\uparrow$</th>
<th>FID $\downarrow$</th>
<th>IS  $\uparrow$</th>
<th>FID $\downarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam</td>
<td>7.76 ±.11</td>
<td>22.45 ±.65</td>
<td>8.43 ±.05</td>
<td>8.21 ±.05</td>
</tr>
<tr>
<td>EG</td>
<td>7.83 ±.08</td>
<td>20.73 ±.22</td>
<td>8.15 ±.06</td>
<td>8.15 ±.07</td>
</tr>
<tr>
<td>Ours (GDA-RAM)</td>
<td><strong>8.05 ±.06</strong></td>
<td><strong>19.32 ±.16</strong></td>
<td><strong>7.82 ±.06</strong></td>
<td><strong>8.38 ±.04</strong></td>
</tr>
</tbody>
</table>

6 Conclusion

We prove the convergence property of GDA-AM and obtain a faster convergence rate than EG and OG on the bilinear problem. Empirically, we verify our claim for such a problem and show the efficacy of GDA-AM in a deep learning setting as well. We believe our work is different from previous approaches and takes an important step towards understanding and improving minimax optimization by exploiting the GDA dynamic and reforming it with numerical techniques.
REFERENCES


Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.


A RELATED WORK

There is a rich literature on different strategies to alleviate the issue of minimax optimization. A useful add-on technique, Momentum, has been shown to be effective for bilinear games and strongly-convex-strongly-concave settings \cite{ZhangWang2021,Gidel2019,Azizian2020}. Several second-order methods \cite{Adolphe2017,Mazumdar2019,Parker-Holder2020} show that their stable fixed points are exactly either Nash equilibria or local minima by incorporating second-order information. However, such methods are computationally expensive and thus unsuitable for large applications such as image generation. Focusing on variants of GDA, EG and OG are two widely studied algorithms on improving the GDA dynamics. EG proposed to apply extra-gradient to overcome the cycling behaviour of GDA. OG, originally proposed in Popov (1980) and rediscovered in Daskalakis et al. (2018); Mertikopoulos et al. (2019), is more efficient by storing and re-using the extrapolated gradient for the extrapolation step. Without projection, OG is equivalent to extrapolation from past. \cite{Mokhtari2020b} shows that both of these algorithms can be interpreted as approximations of the classical proximal point method and did a unified analysis for bilinear games. These approaches mentioned the GDA dynamics can be viewed as a fixed-point iteration, but none of them further provides a solution to improve it. In this work, we fill this gap by proposing the application of the extrapolation method directly on the entire GDA dynamics. Unlike OG, EG and their variants \cite{Hsieh2019,Lei2021,Thekumparampl2019,Yang2019}, which regard minimax problems as variational inequality problems \cite{Bruck1977,Nemirovski2004}, our work is from a new perspective and thus orthogonal to these previous approaches.

In addition, several recent works consider nonconvex-concave minimax problems. \cite{ZhangWang2021} introduced a “smoothing” scheme combined with GDA to stabilize the dynamic of GDA. Luo et al. (2020) proposed a method called Stochastic Recursive gradEnt Descent Ascent (SREDA) for nonconvex-strongly-concave minimax problems, by estimating gradients recursively and reducing its variance. \cite{Lin2020} showed that the two-timescale GDA can find a stationary point of nonconvex-concave minimax problems effectively. \cite{Ostrovski2021} proposed a variant of Nesterov’s accelerated algorithm to find \(\epsilon\)-first-order Nash equilibrium that is a stronger criterion than the commonly used proximal gradient norm. \cite{Nouiehed2019} proposed a iterative method that finds \(\epsilon\)-first-order Nash equilibrium in \(O(\epsilon^{-2})\) iterations under Polyak-Lojasiewicz (PL) condition. Focusing on nonconvex minimax problems, they studied an interesting and difficult problem. Since our work cast insight on the effectiveness of solving minimax optimization via Anderson Mixing, we expect the extension of this algorithm to general nonconvex problems can be further investigated in the future.

B ANDERSON MIXING IMPLEMENTATION DETAILS

In this section, we discuss the efficient implementation of Anderson Mixing. We start with generic Anderson Mixing prototype (Algorithm 4) and then present the idea of Quick QR-update Anderson Mixing implementation as described in \cite{Walker2011}, which is commonly used in practice. For each iteration \(t \geq 0\), AM prototype solves a least squares problem with a normalization constraint. The intuition is to minimize the norm of the weighted residuals of the previous \(m\) iterates.

\begin{algorithm}
\caption{Anderson Mixing Prototype (truncated version)}
\textbf{Input:} Initial point \(w_0\), Anderson restart dimension \(p\), fixed-point mapping \(g : \mathbb{R}^n \rightarrow \mathbb{R}^n\).
\textbf{Output:} \(w_{t+1}\)

\textbf{for} \(t = 0, 1, \ldots\) \textbf{do}
\begin{itemize}
  \item Set \(p_t = \min\{t, p\}\).
  \item Set \(F_t = \{f_{t-p_t}, \ldots, f_t\}\), where \(f_i = g(w_i) - w_i\) for each \(i \in [t - p_t, \ldots, t]\).
  \item Determine weights \(\beta = (\beta_0, \ldots, \beta_{p_t})^T\) that solves \(\min_\beta \|F_t \beta\|_2\), s. t. \(\sum_{i=0}^{p_t} \beta_i = 1\).
  \item Set \(w_{t+1} = \sum_{i=0}^{p_t} \beta_i g(w_{t-p_t+i})\).
\end{itemize}
\end{algorithm}

The constrained linear least-squares problem in Algorithm AA can be solved in a number of ways. Our preference is to recast it in an unconstrained form suggested in \cite{FangSaad2009,Walker2011} that is straightforward to solve and convenient for implementing efficient updating of QR.
Define $f_i = g(w_i) - w_i$, $\Delta f_i = f_{i+1} - f_i$ for each $i$ and set $F_i = [f_{t-p_1}, \ldots, f_t]$, $\mathcal{F}_t = [\Delta f_{t-p_1}, \ldots, \Delta f_t]$. Then solving the least-squares problem $\min \| f_t \beta \|_2$, s.t. $\sum_{i=0}^{p_1} \beta_i = 1$ is equivalent to

$$\min_{\gamma = (\gamma_0, \ldots, \gamma_{p_1-1})} \| f_t - \mathcal{F}_t \gamma \|_2$$

(10)

where $\alpha$ and $\gamma$ are related by $\alpha_0 = \gamma_0, \alpha_i = \gamma_i - \gamma_{i-1}$ for $1 \leq i \leq p_t - 1$, and $\alpha_{p_t} = 1 - \gamma_{p_t-1}$.

Now the inner minimization subproblem can be efficiently solved as an unconstrained least squares problem by a simple variable elimination. This unconstrained least-squares problem leads to a modified form of Anderson Mixing

$$w_{t+1} = g(w_t) - \sum_{i=0}^{p_t-1} \gamma_i(t) [g(w_{t-p_t+i+1}) - g(w_{t-p_t+i})] = g(w_t) - G_t \gamma(t)$$

where $G_t = [\Delta g_{t-p_t}, \ldots, \Delta g_{t-1}]$ with $\Delta g_i = g(w_{i+1}) - g(w_i)$ for each $i$.

To obtain $\gamma(t) = (\gamma_0(t), \ldots, \gamma_{p_t-1}(t))^T$ by solving equation 10 efficiently, we show how the successive least-squares problems can be solved efficiently by updating the factors in the QR decomposition $F_t = Q_t R_t$ as the algorithm proceeds. We assume a think QR decomposition, for which the solution of the least-squares problem is obtained by solving the $p_t \times p_t$ linear system $R \gamma = Q^T f_t$. Each $F_t$ is $n \times p_t$ and is obtained from $F_{t-1}$ by adding a column on the right and, if the resulting number of columns is greater than $p_t$, also cleaning up (re-initialize) the table. That is, we never need to delete the left column because cleaning up the table stands for a restarted version of AM. As a result, we only need to handle two cases: 1) the table is empty(cleaned). 2) the table is not full. When the table is empty, we initialize $F_t = Q_t R_t$ with $Q_t = \Delta f_0 / \| \Delta f_0 \|_2$ and $R = \| \Delta f_0 \|_2$. If the table size is smaller than $p_t$, we add a column on the right of $F_{t-1}$. Have $F_{t-1} = QR$, we update $Q$ and $R$ so that $F_t = [F_{t-1}, \Delta f_{t-1}] = QR$. It is a single modified Gram–Schmidt sweep that is described as follows:

**Algorithm 5:** QR-updating procedures

for $i = 1, \ldots, p_{t-1}$ do
  Set $R(i, p_t) = Q(:, i)' \cdot \Delta f_{t-1}$.
  Update $\Delta f_{t-1} \leftarrow \Delta f_{t-1} - R(i, p_t) \cdot Q(:, i)$
end

Set $Q(:, p_t) = \Delta f_{t-1} / \| \Delta f_{t-1} \|_2$ and $R(p_t, p_t) = \| \Delta f_{t-1} \|_2$

Note that we do not explicitly conduct QR decomposition in each iteration, instead we update the factors ($O(p_t^2 n)$) and then solve a linear system using back substitution which has a complexity of $O(p_t^3)$. Based on this complexity analysis, we can find Anderson Mixing with QR-updating scheme has limited computational overhead than GDA (or OG). This explains why GDA–AM is faster than EG but slower than OG in terms of running time of each iteration.

## C THEORETICAL RESULTS

### C.1 DIFFICULTY OF ANALYSIS ON GDA WITH ANDERSON MIXING

In the analysis, we study the inherent structures of the dynamics of the fixed point iteration and provide the convergence analysis for both simultaneous and alternating schemes. We want to emphasize that the direct application of existing convergence results of GMRES can not lead to convergent results. A recent paper Bollapragada et al. [2018] study the convergence acceleration schemes for multi-step optimization algorithms using Regularized Nonlinear Acceleration. We also want to point out that a naïve application of Crouzeix’s bound to the minimax optimization problem can not be used to derive the convergent result.

**Theorem C.1** (Fischer & Freund [1991]). Let $n \geq 5$ be an integer, $r > 1$, and $c \in \mathbb{R}$. Consider the following constrained polynomial minimax problem

$$\min_{p \in P_n \setminus \{c\}} \max_{z \in \mathcal{E}_r} |p(z)|$$

(11)

where

$$\mathcal{E}_r := \left\{ z \in \mathbb{C} \mid |z - 1| + |z + 1| \leq r + \frac{1}{r} \right\}$$

(12)
and $c \in \mathbb{C} \setminus \mathcal{E}_r$. Then this problem can be solved uniquely by

$$t_n(z; c) := \frac{T_n(z)}{T_n(c)},$$

where

$$T_n(z) = \frac{1}{2} \left( v^n + \frac{1}{v^n} \right), \quad z = \frac{1}{2} \left( v + \frac{1}{v} \right)$$

if (a) $|c| \geq \frac{1}{2} \left( r + \sqrt{r^2 - 1} \right)$ or
(b) $|c| \geq \frac{1}{2a_r} \left( 2a_r^2 - 1 + \sqrt{2a_r^2 - a_r^2 + 1} \right)$, where $a_r := \frac{1}{2} \left( r + \frac{1}{r} \right)$.

This is because the point 0 where all the residual polynomials take the fixed value of 1 is included in the numerical range of the iteration matrix, which violates the assumption of Theorem C.1. As a result, it cannot be used to prove that the residual norm is decreasing based on this approach. Instead, we show that although the coefficient matrix is non-normal, it is diagonalizable. We then give a result, it can not be used to prove that the residual norm is decreasing based on this approach. Instead, we show that although the coefficient matrix is non-normal, it is diagonalizable. We then give a result, it can not be used to prove that the residual norm is decreasing based on this approach.

Writing this linear system in the block form:

$$(I - G^{(Alt)})w = b^{(Alt)}, \quad \text{with } w_0 = w_0^{(Alt)}.$$  \hfill (15)

Writing this linear system in the block form:

$$
\begin{bmatrix}
0 & \eta A \\
-\eta A^T & \eta^2 A^T A
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_2
\end{bmatrix}
= b^{(Alt)}.
$$

The residual norm bound for GMRES reads:

$$\|r_i\|_2 = \min_{p \in \mathbb{F}^2} \|p(I - G^{(Alt)})r_0\|_2.$$  \hfill (17)

Notice that the matrix $(I - G^{(Alt)})$ is non-normal. If we apply Crouzeix’s bound in Crouzeix & Palencia (2017) to our problem as Bollapragada et al. (2018) did, then we have the following bound

$$\|r_i\|_2 \leq \min_{p \in \mathbb{F}^2} \|p(I - G^{(Alt)})\| \leq (1 + \sqrt{2}) \min_{p \in \mathbb{F}^2} \sup_{z \in W(I - G^{(Alt)})} \|p(z)\|.$$  \hfill (18)

where $W(I - G^{(Alt)}) = \{z^* (I - G^{(Alt)})z, \forall z \in \mathbb{C}^2 \setminus \{0\}, \|z\| = 1\}$ is the numerical range for $I - G^{(Alt)}$. In order to simplify the upper bound in the previous theorem, we study the numerical range of $I - G^{(Alt)}$ similar to Bollapragada et al. (2018). Writing $z = [z_1, z_2]$ and computing the numerical range of $I - G^{(Alt)}$ explicitly yields:

$$[z_1^*, z_2^*] \begin{bmatrix} 0 & \eta A \\ -\eta A^T & \eta^2 A^T A \end{bmatrix} \begin{bmatrix} z_1 \\
z_2 \end{bmatrix} = \eta^2 z_2^* A^T A z_2 + \eta z_1^* A z_2 - \eta z_2^* A^T z_1.$$  \hfill (19)

For a general matrix $A$, there is no special structure about the numerical range of $I - G^{(Alt)}$. However, when $A$ is symmetric, we can decompose $A$ as $A = \sum_{i=1}^n \lambda_i v_i v_i^T$ where $\{\lambda_i\}_{i=1}^n$ are eigenvalues of $A$ in decreasing order and $\{v_i\}_{i=1}^n$ are associated eigenvectors, and write $A^T A = \sum_{i=1}^n \lambda_i^2 v_i v_i^T$.

Then we can compute the numerical range of $G^{(Alt)}$ as follows:

$$\sum_{i=1}^n [z_1^i, z_2^i] \begin{bmatrix} 0 & \eta \lambda_i v_i v_i^T \\ -\eta \lambda_i v_i v_i^T & \eta^2 \lambda_i^2 v_i v_i^T \end{bmatrix} \begin{bmatrix} z_1 \\
z_2 \end{bmatrix} = \sum_{i=1}^n [z_1^i v_i, z_2^i v_i] \begin{bmatrix} 0 & \eta \lambda_i \\ -\eta \lambda_i & \eta^2 \lambda_i^2 \end{bmatrix} \begin{bmatrix} v_i^T z_1 \\
v_i^T z_2 \end{bmatrix}.$$  \hfill (20)

Following the techniques proposed in Bollapragada et al. (2018) to analyze the numerical range of general $2 \times 2$ matrices, we can show that the numerical range of $I - G^{(Alt)}$ is equal to the convex hull of the union of the numerical range of

$$G_i = \begin{bmatrix} 0 & \eta \lambda_i \\ -\eta \lambda_i & \eta^2 \lambda_i^2 \end{bmatrix}, \quad i = 1, \ldots, n.$$  \hfill (21)
And the boundary of numerical range of $G_i$ is an ellipse whose axes are the line segments joining the points $x$ to $y$ and $w$ to $z$, respectively, with

$$x = 0, \quad y = \eta^2 \lambda_i^2, \quad w = \frac{\eta^2 \lambda_i^2}{2} - \sqrt{\eta} |\lambda_i|, \quad z = \frac{\eta^2 \lambda_i^2}{2} + \sqrt{\eta} |\lambda_i|. \quad (22)$$

Thus, the numerical range of $I - G(Alt)$ can be spanned by convex hull of the union of the numerical range of a set of 2-by-2 matrices and the numerical range of each such a 2-by-2 matrix is an ellipse. We can compute the center $o$ and focal distance $d$ of the ellipse generated by numerical range of $I - G(Alt)$ explicitly. Then a linear transformation enables us to use Theorem C.1 to show that the near-optimal polynomial for the minimax problem on the numerical range of $I - G(Alt)$ is given by $t_n(z;c) := \frac{T_p(z)}{T_p(z_c)}$ if 0 is excluded from the numerical range of $I - G(Alt)$. However, according to equation (22), the numerical range includes the point 0 where the residual polynomial takes value 1, thus the analysis based on numerical range cannot help derive the convergent result as the upper bound is not guaranteed to be less than 1.

C.2 Proofs of Theorem

We first provide proof of Theorem 4.1.

**Theorem C.2** (Global convergence for simultaneous GDA–AM on bilinear problem). Denote the distance between the stationary point $w^*$ and current iterate $w_{(k+1)p}$ of Algorithm 2 with Anderson restart dimension $p$ as $N_{(k+1)p} = \text{dist}(w^*, w_{(k+1)p})$. Then we have the following bound for $N_i$ Algorithm 2 is unconditionally convergent

$$N_{(k+1)p} \leq \frac{1}{T_p(1 + \frac{2}{\kappa(A^2A) - 1})} N_{kp} \quad (23)$$

where $T_p$ is the Chebyshev polynomial of first kind of degree $p$ and $\frac{1}{T_p(1 + \frac{2}{\kappa(A^2A) - 1})} < 1$ since $1 + \frac{2}{\kappa(A^2A) - 1} > 1$.

**Proof of Theorem C.1** Note that $I - G(\text{Sim})$ is a normal matrix which will be denoted as $G$ for notational simplicity. Thus it admits the following eigendecomposition:

$$G = U \Lambda U^T, \quad UU^T = I, \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{2n}). \quad (24)$$

Based on the equivalence between GMRES and Anderson Mixing, we know that the convergence rate of simultaneous GDA–AM can be estimated by the spectrum of $G$. Especially, it holds that

$$r_{(k+1)p} = U f_p(A)U^T r_{kp}, \quad f_p \in P_p \quad (25)$$

where $P_p$ is the family of residual polynomials with degree $p$ such that $f_p(0) = 1, \forall f_p \in P_p$. According to Lemma 2.1, we have the following estimation

$$||r_{(k+1)p}||_2 = \min_{f_p \in P_p} \|f_p(G) r_{kp}\|_2 \leq \min_{f_p \in P_p} \max_{f_p(\lambda_i)} \|f_p(\lambda_i)\| r_{kp} \|_2. \quad (26)$$

Due to the block structure of $G$, the eigenvalues of $G$ can be computed explicitly as

$$\pm \eta \sigma_i \sqrt{-1}, \quad i = 1, \ldots, n, \quad (27)$$

where $\sigma_i$ is the $i$th largest singular value of matrix $A$. This shows that the eigenvalues of $G$ are $n$ pairs of purely imaginary numbers excluding 0 since $A$ has full rank.

Since the eigenvalues of $G$ are distributed in two intervals excluding the origin

$$I = [-\eta \sigma_{\text{max}}(A) \sqrt{-1}, -\eta \sigma_{\text{min}}(A) \sqrt{-1}] \cup [\eta \sigma_{\text{min}}(A) \sqrt{-1}, \eta \sigma_{\text{max}}(A) \sqrt{-1}],$$

it can be shown that the following $p$-th degree polynomial with value 1 at the origin that has the minimal maximum deviation from 0 on $I$ is given by:

$$f_p(z) = \frac{T_p(q(\sqrt{-1}z))}{T_p(q(0))}, \quad q(\sqrt{-1}z) = 1 - \frac{2(\sqrt{-1}z - \eta \sigma_{\text{min}}(A) \sqrt{-1})}{(\eta \sigma_{\text{max}}(A))^2 - (\eta \sigma_{\text{min}}(A))^2}, \quad (28)$$
where \( l = \left\lfloor \frac{p}{2} \right\rfloor \) and \( T_l \) is the Chebyshev polynomial of first kind of degree \( l \). The function \( q(\sqrt{-1}z) \) maps \([-1, 1]\) to \([-1, 1]\). Thus the numerator of the polynomial \( f_p \) is bounded by 1 on \([-1, 1]\). The size of denominator can be determined by the method discussed in Chapter 3 of [Greenbaum, 1997]. Assume \( q(0) = \frac{1}{2}(y + y^{-1}) \), then \( T_l(q(0)) = \frac{1}{2}(y^l + y^{-l}) \). Then \( y \) can be determined by solving

\[
q(0) = \frac{(\eta \sigma_{\text{max}}(A))^2 + (\sigma_{\text{min}}(A))^2}{(\eta \sigma_{\text{max}}(A))^2 - (\sigma_{\text{min}}(A))^2}. \tag{29}
\]

The solutions to this equation are

\[
y_1 = \frac{\eta \sigma_{\text{max}}(A) + \sigma_{\text{min}}(A)}{\eta \sigma_{\text{max}}(A) - \sigma_{\text{min}}(A)} \quad \text{or} \quad y_2 = \frac{\eta \sigma_{\text{max}}(A) - \sigma_{\text{min}}(A)}{\eta \sigma_{\text{max}}(A) + \sigma_{\text{min}}(A)} \tag{30}
\]

Then plugging the value of \( q(0) \) into the polynomial \( f_p \) yields

\[
\frac{\|r_{(k+1)p}\|}{\|r_{kp}\|} \leq 2 \left( \frac{\sqrt{\eta^2 \sigma_{\text{max}}^2(A)} - \sqrt{\eta^2 \sigma_{\text{min}}^2(A)}}{\sqrt{\eta^2 \sigma_{\text{max}}^2(A)} + \sqrt{\eta^2 \sigma_{\text{min}}^2(A)}} \right)^l.
\]

\[
= 2 \left( \frac{\sigma_{\text{max}}(A) - \sigma_{\text{min}}(A)}{\sigma_{\text{max}}(A) + \sigma_{\text{min}}(A)} \right)^l = 2 \left( \frac{\eta \lambda - 1}{\eta \lambda + 1} \right)^l. \tag{31}
\]

Note that \( N_t \) and \( r_t \) is related through \( G(w_t - w^*) = r_t \). Therefore,

\[
N_{(k+1)p} = \|w_{(k+1)p} - w^*\|_2 = \|G^{-1}r_{(k+1)p}\|_2 = \min_{f_p \in \mathbb{P}_p} \|G^{-1}f_p(G)(w_{kp} - w^*)\|_2
\]

\[
\leq \min_{f_p \in \mathbb{P}_p} \max_i |f_p(\lambda_i)| \|w_{kp} - w^*\|_2 \leq 2 \left( 1 - \frac{2}{\eta \lambda + 1} \right) \leq \frac{T_p(1 + 2 \sigma_{\text{min}}^2)}{\eta \lambda + 1} N_{kp}. \tag{32}
\]

Actually a tighter bound can be proved after noting that the problem is essentially equivalent to polynomial min-max problem on the interval:

\[
\tilde{I} = [\eta^2 \sigma_{\text{min}}^2, \eta^2 \sigma_{\text{max}}^2].
\]

Then it is well known that,

\[
N_{(k+1)p} \leq \frac{1}{T_p(1 + 2 \sigma_{\text{min}}^2)} N_{kp}
\]

\[
\leq \frac{1}{T_p(1 + \frac{1}{\eta \lambda + 1})} N_{kp} \tag{33}
\]

where \( T_p \) Chebyshev polynomial of degree \( p \) of the first kind and \( \frac{1}{T_p(1 + \frac{1}{\eta \lambda + 1})} < 1 \). Explicitly,

\[
T_p(1 + \frac{2}{\eta \lambda + 1}) = \frac{1}{2} \left[ (1 + \frac{2}{\eta \lambda + 1}) + \sqrt{(1 + \frac{2}{\eta \lambda + 1})^2 - 1} \right]^p
\]

\[
+ \left( 1 + \frac{2}{\eta \lambda + 1} \right) + \sqrt{(1 + \frac{2}{\eta \lambda + 1})^2 - 1} \right]^{-p}.
\]

Next, we give the proof of Theorem 4.2

**Theorem C.3** (Global convergence for alternating GDA-AM on bilinear problem). Denote the distance between the stationary point \( w^* \) and current iterate \( w_{(k+1)p} \) of Algorithm 4.1 with Anderson restart dimension \( p \) as \( N_{(k+1)p} = \text{dist}(w^*, w_{(k+1)p}) \). Assume \( A \) is normalized such that its largest singular value is equal to 1. Then when the learning rate \( \eta \) is less than 2, we have the following bound for \( N_t \)

\[
N_{(k+1)p}^2 \leq \sqrt{1 + \frac{2\eta}{2 - \eta} \frac{C}{c} \frac{r}{c} N_{kp}^2}
\]

where \( c \) and \( r \) are the center and radius of a disk \( D(c, r) \) which includes all the eigenvalues of \( G \) in equation 4.3. Especially, \( \frac{r}{c} < 1 \).
Proof. Since the residual $\mathbf{r}_p$ of AA at p-th iteration has the form of

$$\mathbf{r}_p = (\mathbf{I} - \sum_{i=1}^{p} \mathbf{G}^i)\mathbf{r}_0,$$

and AA minimizes the residual, we have

$$\|\mathbf{r}_{(k+1)p}\|_2^2 \leq \min_{\beta} \|\mathbf{r}_{kp} - \beta \mathbf{G}^i \mathbf{r}_{kp}\|_2^2 \leq \min_{f_p \in \mathcal{P}_p} \|f_p(\mathbf{G})\mathbf{r}_{kp}\|_2^2,$$

where $\mathcal{P}_p$ is the family of polynomials with degree p such that $f_p(0) = 1$, $\forall f_p \in \mathcal{P}_p$. It’s easy to see that $\mathbf{G}$ is unitarily similar to a block diagonal matrix $\Lambda$ with $2 \times 2$ blocks as follows:

$$\begin{bmatrix} 0 & \eta \sigma_i \\ -\eta \sigma_i & (\eta \sigma_i)^2 \end{bmatrix} \quad \forall \ i \in [n].$$

Thus the eigenvalues of $\mathbf{G}$ can be easily identified as

$$\lambda_{\pm i} = \frac{(\eta \sigma_i (\eta \sigma_i \pm \sqrt{[(\eta \sigma_i)^2 - 4]))}{2}, \quad i \in [n].$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ are the singular values of $\mathbf{A}$. Furthermore, the eigenvector and eigenvalue associated with each $2 \times 2$ diagonal block are

$$\begin{bmatrix} 0 & \eta \sigma_i \\ -\eta \sigma_i & (\eta \sigma_i)^2 \end{bmatrix} \begin{bmatrix} \lambda_{\pm i} \\ \eta \sigma_i \end{bmatrix} = \lambda_{\pm i} \begin{bmatrix} 1 \\ \eta \sigma_i \end{bmatrix}$$

Thus $\mathbf{G}$ is diagonalizable and denote the matrix with the columns of eigenvectors of $\mathbf{G}$ by $\mathbf{X}$. The real part of the eigenvalues of $\mathbf{G}$ are at least

$$\mathcal{R}(\lambda_{\pm i}) \geq \frac{(\eta \sigma_i)^2}{2}, \quad i \in [n].$$

And since $|\eta \sigma_i| \geq \sqrt{[(\eta \sigma_i)^2 - 4]}$, all the eigenvalues will be included in a disk $D(c, r)$ which is included in the right half plane. Moreover, both $c$ and $r$ being greater than zero indicates that $\frac{c}{r} < 1$. Start from the following inequality:

$$N_{(k+1)p} = \|w_{(k+1)p} - w^*\|_2^2 = \|\mathbf{G}^{-1}\mathbf{r}_{(k+1)p}\|_2^2 \leq \min_{f_{p} \in \mathcal{P}_p} \|\mathbf{G}^{-1}f_{p}(\mathbf{G})\mathbf{r}_{kp}\|_2^2 = \min_{f_{p} \in \mathcal{P}_p} \|\mathbf{G}^{-1}_p(\mathbf{G})(w_{kp} - w^*)\|_2^2 = \min_{f_{p} \in \mathcal{P}_p} \|f_{p}(\mathbf{G})(w_{kp} - w^*)\|_2$$

We will use the eigendecomposition of $\mathbf{G}$ and the special polynomial $(\frac{c}{r})^p$ to derive the inequality in Theorem 3. Now we know $\frac{c}{r} < 1$. If we choose $g_{p}(t) = (\frac{c}{r})^p$, we can obtain

$$\min_{f_{p} \in \mathcal{P}_p} \|f_{p}(\mathbf{G})(w_{kp} - w^*)\|_2 \leq \|g_{p}(\mathbf{G})(w_{kp} - w^*)\|_2$$

which implies

$$\min_{f_{p} \in \mathcal{P}_p} \|f_{p}(\mathbf{G})(w_{kp} - w^*)\|_2 \leq \|g_{p}(\mathbf{X}\mathbf{X}^{-1})|||(w_{kp} - w^*)|||_2$$

Since $\mathbf{G}$ is diagonalizable (which has been shown above), we assume the eigendecomposition of $\mathbf{G}$ is

$$\mathbf{G} = \mathbf{X}\Lambda\mathbf{X}^{-1}.$$

Then

$$\min_{f_{p} \in \mathcal{P}_p} \|g_{p}(\mathbf{G})(w_{kp} - w^*)\|_2 \leq \|\mathbf{X}\|\|\mathbf{X}^{-1}\|\max_{\lambda_{i} \in \{\lambda_{n}\}} \|g_{p}(\lambda)||||(w_{kp} - w^*)|||_2 \leq \kappa_{\mathbf{G}}(\frac{r}{c})^p \|w_{kp} - w^*\|_2$$

where $\kappa_{\mathbf{G}}$ is the condition number of $\mathbf{X}$. The last inequality comes from Lemma 6.26 and Proposition 6.32 in [Saad, 2003]. Since $\mathbf{G}$ and $\Lambda$ are unitarily similar, $\kappa_{\mathbf{G}}$ is equal to the condition number of the eigenvector matrix of $\Lambda$. The eigenvector matrix of $\Lambda$ is a block diagonal matrix with the $i$th block as

$$\begin{bmatrix} \lambda_{\pm i} \\ \frac{1}{\eta \sigma_i} \lambda_{\pm i} \end{bmatrix}.$$ 

Thus the singular values of the eigenvector matrix of $\Lambda$ are equal to the union of the singular values of these 2-by-2 blocks. Under the assumption that the largest singular value of $\mathbf{A}$ is equal to 1 and the learning rate is less than 2, it is easy to find the singular values of the eigenvector matrix of $\Lambda$ are $\sqrt{2} \pm \eta \sigma_i$. Thus,

$$\kappa_{\mathbf{G}} = \frac{2\sqrt{2} \eta \sigma_{\text{max}}}{\sqrt{2} - \eta \sigma_{\text{max}}} = \frac{2\sqrt{2} \eta}{\sqrt{2} - \eta} = \sqrt{1 + \frac{4 \eta}{2 - \eta}}.$$
C.3 Discussion of Obtained Rates

We would like to first explain why taking Chebyshev polynomial of degree $p$ at the point $1 + \frac{2}{\kappa - 1}$. We evaluate the Chebyshev polynomial at this specific point because the reciprocal of this value gives the minimal value of infinite norm of all polynomials of degree $p$ defined on the interval $\tilde{I} = [\eta^2 \sigma^2_{\min}(A), \eta^2 \sigma^2_{\max}(A)]$ based on Theorem 6.25 (page 209) [Saad, 2003]. In other words, taking the function value at this point leads to the tight bound.

When comparing between existing bounds, we would like to point our our derived bounds are hard to compare directly. Alternatively, we can derive another bound for comparison with existing bounds for simultaneous GDA–AM. If we use the inequality that $T_p(t) \geq \frac{1}{2}(t + \sqrt{t^2 - 1})^p$, we can obtain the bound $\rho(A) = 4(\sqrt{\kappa(A^T A)^{-1}})^2 = 4(1 - O(\frac{1}{\sqrt{\kappa(A^T A)}}))$, which is in a form that is comparable with EG and can compete with EG + positive momentum. The numerical experiments in figure 2b numerically verify that our bound is smaller than EG. We wanted to numerically compare our rate with EG with positive momentum. However the bound of EG with positive momentum is asymptotic. Moreover, it does not specify the constants so we can not numerically compare them. We do provide empirical comparison between GDA-AM and EG with positive momentum for bilinear problems in Appendix D.1. It shows GDA-AM outperforms EG with positive momentum. Regarding alternating GDA-AM, we would like to note that the bound in Theorem 4.2 depends on the eigenvalue distribution of the matrix $G$. Condition number is not directly related to the distribution of eigenvalues of a nonsymmetric matrix $G$. Thus, the condition number is not a precise metric to characterize the convergence. If these eigenvalues are clustered, then our bound can be small. On the other hand, if these eigenvalues are evenly distributed in the complex plane, then the bound can very close to 1.

More importantly, we would like to stress several technical contributions.

1. Our obtained Theorem 4.1 and 4.2 provide nonasymptotic guarantees, while most other work are asymptotic. For example, EG with positive momentum can achieve an asymptotic rate of $1 - O(1/\sqrt{\kappa})$ under strong assumptions [Azizian et al., 2020].
2. Our contribution is not just about fix the convergence issue of GDA by applying Anderson Mixing; another contribution is that we arrive at a convergent and tight bound on the original work and not just adopting existing analyses. We developed Theorem 4.1 and 4.2 from a new perspective because applying existing theoretical results fail to give us neither convergent nor tight bounds.
3. Theorem 4.1 and 4.2 only requires mild conditions and reflects how the table size $p$ controls the convergence rate. Theorem 4.1 is independent of the learning rate $\eta$. However, the convergence results of other methods like EG and OG depend on the learning rate, which may yield less than desirable results for ill-specified learning rates.

C.4 Convex-Concave and General Case

Given the widespread usage of minimax problems in applications of machine learning, it is natural to ask about its properties when being applied to general nonconvex-nonconcave settings. If $f$ is a nonconvex-nonconcave function, the problem of finding global Nash equilibrium is NP-hard in general. Recently, Jin et al. (2020) show that local or global Nash equilibrium may not exist in nonconvex-nonconcave settings and propose a new notation local minimax as defined below:

**Definition 4.** A point $(x^*, y^*)$ is said to be a local minimax point of $f$, if there exists $\delta_0 > 0$ and a function $h$ satisfying $h(\delta) \to 0$ as $\delta \to 0$, such that for any $\delta \in (0, \delta_0]$, and any $(x, y)$ satisfying $\|x - x^*\| \leq \delta$ and $\|y - y^*\| \leq \delta$, we have

$$f(x^*, y) \leq f(x^*, y^*) \leq \max_{y': \|y' - y^*\| \leq h(\delta)} f(x, y').$$

Jin et al. (2020) also establishes the following first- and second-order conditions to characterize local minimax:

**Proposition 1** (First-order Condition). Any local minimax point $(x^*, y^*)$ satisfies $\nabla f(x^*, y^*) = 0$.

**Proposition 2** (Second-order Necessary Condition). Any local minimax point $(x^*, y^*)$ satisfies $\nabla_{yy} f(x^*, y^*) \preceq 0$ and $\nabla_{xx} f(x^*, y^*) - \nabla_{xy} f(x^*, y^*) (\nabla_{yy} f(x^*, y^*))^{-1} \nabla_{yx} f(x^*, y^*) \succeq 0$. 

20
Proposition 3 (Second-order Sufficient Condition). Any stationary point \((x^*, y^*)\) satisfies 
\[
\nabla_{yy} f(x^*, y^*) < 0 \quad \text{and} \quad \nabla_{xx} f(x^*, y^*) - \nabla_{xy} f(x^*, y^*) (\nabla_{yy} f(x^*, y^*))^{-1} \nabla_{yx} f(x^*, y^*) \succ 0
\]

is a local minimax point.

Given the second-order conditions of local minimax, it turns out that above question is extremely challenging—GDA-AM is a first-order method. But we can prove the following result for GDA-AM:

**Theorem C.4** (Local minimax as subset of limiting points of GDA-AM). Consider a general objective function \(f(x, y)\). The set of limiting points of GDA-AM for minimax problem

\[
\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^n} f(x, y)
\]

includes the local minimax points of this function.

The definition of local minimax is stronger than that of first order \(\epsilon\) point. The convergence analysis for complexity of finding \(\epsilon\) stationary point is included in the next section. The proof of Theorem C.4 needs the result from the following theorem.

**Theorem C.5** (Calvetti et al. (2002)). Let \(\delta\) satisfy \(0 < \delta \leq \delta_0\) for some constant \(\delta_0 > 0\) (refer to Calvetti et al. (2002) for details), and let \(b^\ell \in X\) satisfy \(\|b - b^\ell\| \leq \delta\). Let \(k \leq \ell\) and let \(x_k^\ell\) denote the \(k\)th iterate determined by the GMRES method applied to equation \(Ax = b^\ell\), with initial guess \(x_0 = 0\). Similarly, let \(x_k\) denote the \(k\)th iterate determined by the GMRES method applied to equation \(Ax = b\) with initial guess \(x_0 = 0\). Then, there are constants \(\sigma_k\) independent of \(\delta\), such that

\[
\|x_k - x_k^\ell\| \leq \sigma_k \delta, \quad 1 \leq k \leq \ell
\]

Then, we give the proof of Theorem C.4.

**Proof of Theorem C.4**. For notational simplicity, we will denote \(\nabla_{xx} f(x^*, y^*)\), \(\nabla_{xy} f(x^*, y^*)\) and \(\nabla_{yy} f(x^*, y^*)\) by \(H_{xx}, H_{xy}, H_{yy}\), respectively. Simultaneous GDA can be written as

\[
w_{t+1} = \begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} x_t - \eta \nabla_x f(x_t, y_t) \\ y_t + \eta \nabla_y f(x_t, y_t) \end{bmatrix}.
\]

Since the function is differentiable, Taylor expansion holds for \(\nabla_x f(x_t, y_t)\) and \(\nabla_y f(x_t, y_t)\) at a local minimax point \(w^* = (x^*, y^*)\),

\[
\begin{align*}
\nabla_x f(x_t, y_t) &= \nabla_x f(x^*, y^*) + H_{xx}(x_t - x^*) + H_{xy}(y_t - y^*) + o(\|w_t - w^*\|_2) \\
\nabla_y f(x_t, y_t) &= \nabla_y f(x^*, y^*) + H_{yy}(y_t - y^*) + H_{yx}(x_t - x^*) + o(\|w_t - w^*\|_2).
\end{align*}
\]

Use the fact that \(\nabla f(x^*, y^*) = 0\) to simplify the above equations and obtain

\[
\begin{align*}
\nabla_x f(x_t, y_t) &= H_{xx}(x_t - x^*) + H_{xy}(y_t - y^*) + o(\|w_t - w^*\|_2) \\
\nabla_y f(x_t, y_t) &= H_{yy}(y_t - y^*) + H_{yx}(x_t - x^*) + o(\|w_t - w^*\|_2).
\end{align*}
\]

Inserting the above formulas into the iteration scheme, it yields

\[
w_{t+1} = \begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} I - \eta H_{xx} & -\eta H_{xy} \\ \eta H_{yx} & I + \eta H_{yy} \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} \eta H_{xx} x^* + \eta H_{xy} y^* + \epsilon \\ -\eta H_{yx} x^* - \eta H_{yy} y^* - \eta H_{xy} x^* + \epsilon \end{bmatrix}
\]

where \(\epsilon\) denotes the higher order error \(o(\|w_t - w^*\|_2)\). According to Theorem 2.2, we know that simultaneous GDA-AM is equivalent to applying GMRES to solve the following linear system

\[
(I - \begin{bmatrix} (1 - \alpha)I - \eta H_{xx} & -\eta H_{xy} \\ \eta H_{yx} & (1 - \alpha)I + \eta H_{yy} \end{bmatrix})w = \begin{bmatrix} \alpha I + \eta H_{xx} & \eta H_{xy} \\ -\eta H_{yx} & \alpha I - \eta H_{yy} \end{bmatrix} w = b + \epsilon
\]

where \(b = \begin{bmatrix} \eta H_{xx} x^* + \eta H_{xy} y^* \\ -\eta H_{yx} y^* - \eta H_{xx} x^* \end{bmatrix}\). We now know that GDA-AM is equivalent to GMRES being applied to solve the following linear system

\[
\begin{bmatrix} \alpha I + \eta H_{xx} & \eta H_{xy} \\ -\eta H_{yx} & \alpha I - \eta H_{yy} \end{bmatrix} w = b
\]
The symmetric part of the coefficient matrix of the above linear system is
\[
\begin{bmatrix}
\alpha I + \eta H_{xx^*} & \eta H_{xy^*} \\
0 & \alpha I - \eta H_{yy^*}
\end{bmatrix}.
\]

According to Proposition 2, \(\alpha I - \eta H_{yy^*}\) is positive definite since \(H_{yy^*} \succ 0\). If \(H_{xx^*}\) is positive semidefinite, then \(\alpha I + \eta H_{xx^*}\) is positive definite and we’re done. Otherwise, assume \(\lambda_{\min}(H_{xx^*}) < 0\). Then for fixed \(\alpha\), when \(\eta < -\frac{\lambda_{\min}(H_{xx^*})}{2}\), \(\alpha I + \eta H_{xx^*}\) will be positive definite. Then according to Theorem 2.2, we know GDA-AM indeed converges. Let’s create a new companion linear system as follows
\[
\begin{bmatrix}
\alpha I + \eta H_{xx^*} & \eta H_{xy^*} \\
0 & \alpha I - \eta H_{yy^*}
\end{bmatrix} \tilde{w} = b + \alpha w^*.
\]

Note that \(\tilde{w} = w^*\) and GMRES on this companion linear system is convergent under suitable choice of learning rate \(\eta\). Let the iterates of GMRES for \(\tilde{w}, w\) be denoted by \(\tilde{w}_1, \tilde{w}_2, \tilde{w}_t\). Then \(\|\tilde{w}_1 - \tilde{w}_2\| \leq \|\tilde{w}_2 - \tilde{w}_3\| + \|\tilde{w}_3 - \tilde{w}_4\|\). According to Theorem C.3, we also have \(\|\tilde{w}_1 - \tilde{w}_k\| \leq \sigma_k \epsilon, 1 \leq k \leq t\). Further more, again according to Theorem C.3, we know \(\|\tilde{w}_1 - \tilde{w}_t\| \leq \sigma_t \epsilon \|w^* + \epsilon\|\).

Starting from an initial point very close to \(w^*\) and let \(t \to \infty\) and \(\epsilon \to 0\), \(\tilde{w}_t\) will converge to \(w^* = (x^*, y^*)\), which means the local minimax point \(w^* = (x^*, y^*)\) is a limiting point of GDA-AM.

Theorem C.6. For strongly-convex-strongly-concave function \(f(x, y)\), GDA-AM will converge to the Nash equilibrium of this function.

Proof: Since strongly-convex-strongly-concave function \(f(x, y)\) has unique Nash equilibrium which is also the unique minimax point, this minimax point must be the limiting point of GDA-AM according to Theorem C.3.

C.4.1 BILINEAR-QUADRATIC GAMES

Moreover, we can further show that the GDA-AM converges on bilinear-quadratic games. Consider a quadratic problem as follows,
\[
\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^n} f(x, y) = x^T A y + x^T B x - y^T C y + b^T x + c^T y,
\]

where \(A\) is full rank, \(B\) and \(C\) are both positive definite.

Theorem C.7. [Global convergence for simultaneous GDA-AM on bilinear-quadratic problem] Let \(r_t^{(Sim)}\) be the residual of Algorithm 2 being applied to problem equation 36. For some constant \(\rho < 1\),
\[
\|r_t^{(Sim)}\|_2 \leq \left(1 - \frac{(\lambda_{\min}(J^T + J))^2}{4 \lambda_{\max}(J^T J)}\right)^{t/2} \|r_0\|_2,
\]

where \(J = \begin{bmatrix} \eta B & \eta A \\ -\eta A^T & \eta C \end{bmatrix}\) and \(\lambda_{\min}\) and \(\lambda_{\max}\) denote the smallest and largest eigenvalue, respectively.

The convergence property of GMRES has been studied in the next theorem. We use this theorem to show the convergence rate of GDA-AM for bilinear-quadratic games.

Theorem C.8 [Elman (1982)]. Consider solving a linear system \(Ex = b\) using GMRES. Let \(r_t = b - Ex_t\) be the residual at \(t\)th iteration. If the Hermitian part of \(E\) is positive definite, then for some positive constant \(\rho < 1\), it holds that
\[
\|r_t\|_2 \leq \left(1 - \frac{(\lambda_{\min}(E^H + E))^2}{4 \lambda_{\max}(E^H E)}\right)^{t/2} \|r_0\|_2.
\]
We know that we need to study the convergence properties of GMRES for solving the following linear system
\[
\begin{bmatrix}
\eta B & \eta A \\
-\eta A^T & \eta C
\end{bmatrix}
\begin{bmatrix}
x^t \\
y^t
\end{bmatrix}
\begin{bmatrix}
w_t \quad \omega_t
\end{bmatrix}
\begin{bmatrix}
x_{t+1} \\
y_{t+1}
\end{bmatrix}
= \begin{bmatrix}
-\eta b \\
-\eta c
\end{bmatrix}.
\] (39)

We know that we need to study the convergence properties of GMRES for solving the following linear system
\[
\begin{bmatrix}
\eta B & \eta A \\
-\eta A^T & \eta C
\end{bmatrix}
w = b.
\] (40)

For notational simplicity, the superscripts has been dropped. Denote the coefficient matrix \(\begin{bmatrix}
\eta B & \eta A \\
-\eta A^T & \eta C
\end{bmatrix}\) by \(J\). The symmetric part of \(J\) is \(J + J^T\), which is positive definite. Then immediately by Theorem C.8 the following convergence rate holds
\[
\|r_t\|_2 = \min_{p \in R^t} \|p(J)r_0\|_2 \leq \left(1 - \frac{\lambda_{\min}(J + J^T)^2}{4\lambda_{\max}(J^TJ)}\right)^{t/2} \|r_0\|_2
\] (41)
\[=ho^{t/2} \|r_0\|_2.
\]

Note that the convergence of GDA–AM for bilinear-quadratic games can also be analyzed by numerical range as shown in (Bollapragada et al., 2018). Although we previously show that analysis based on the numerical range can not help us derive a convergent bound for bilinear games, we show analysis in Bollapragada et al. (2018) can be extended to bilinear-quadratic games. When \(B\) and \(C\) are positive definite, 1 is outside of the numerical range of matrix \(G^{(\text{Quad-sim})}\) as shown in 7a. When \(B\) or \(C\) is not positive definite, 1 can be included in the numerical range of matrix \(G^{(\text{Quad-sim})}\) as shown in 7b. That is saying analysis based on the numerical range (Crouzeix & Palencia (2017) Bollapragada et al. (2018)) to the bilinear-quadratic problem can lead to a convergent result when \(B\) and \(C\) are positive definite. And analysis based on the numerical range can not help us derive convergent results when \(B\) or \(C\) is not positive definite.

### C.5 Stochastic convex-nonconvex case

In this section, we study the convergence of GDA–AM for convex-nonconvex problem in the stochastic setting with the same assumptions in Wei et al. (2021b); Xu et al. (2021). The recent work Wei et al. (2021b) proves the convergence of the stochastic gradient descent with Anderson Mixing for min optimization. The convergence of GDA–AM for minimax optimization builds on top of it with several modifications. The minimax problem is equivalent to minimizing a function \(\Phi(\cdot) = \max_{y \in Y} f(\cdot, y)\) (Lin et al., 2020). And we are interested in complexity of a pair of \(\epsilon\)-stationary point \((x, y)\) instead of analysis of a point \(x\).

**Definition.** (Lin et al., 2020) A pair of points \((x, y)\) is an \(\epsilon\)-stationary point \((\epsilon \geq 0)\) of a differentiable function \(\Phi\) if
\[
\|\nabla_x f(x, y)\| \leq \epsilon
\]
\[
\|P_{Y}(y + (1/\ell)\nabla_y f(x, y)) - y\| \leq \epsilon/\ell
\]

**Assumption 1.** \(f : \mathbb{R}^d \to \mathbb{R}\) is continuously differentiable. \(f(x) \geq f^{\text{low}} > -\infty\) for any \(x \in \mathbb{R}^d\). \(\nabla f\) is globally \(L\)-Lipschitz continuous; namely \(\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2\) for any \(x, y \in \mathbb{R}^d\).
To ensure the hessian matrix for $G$ is positive definite, we want to argue that similar arguments in Wei et al. (2021b) can be applied to the problem here. Consider Assumption 2.

For any iteration $k$, the stochastic gradient $\nabla f_{\xi_k}(x_k)$ satisfies $E_{\xi_k}[\nabla f_{\xi_k}(x_k)] = \nabla f(x_k)$, $E_{\xi_k}[\|\nabla f_{\xi_k}(x_k) - \nabla f(x_k)\|^2_2] \leq \sigma^2$, where $\sigma > 0$, and $\xi_k, k = 0, 1, \ldots$, are independent samples that are independent of $\{x_i\}_k$.

Theorem C.9. For a general convex-nonconcave function $f$, suppose that Assumptions 1 and 2 hold. Batch size $n_t = n$ for $t = 0, \ldots, N - 1$. $C > 0$ is a constant. $\beta_t = \frac{n}{\ell(1+t\ell - 1)}$, $\delta_t \geq C\beta_t^{-2}$, $0 \leq \alpha_t \leq \min \{1, \beta_t^{\frac{1}{2}}\}$ and $\alpha_t$ is chosen to make sure the positive definiteness of $H_t$. Let $R$ be a random variable following $P_R(t) = \{R = t\} = 1/N$, and $N$ be the total number of stochastic GDA-AM calls needed to calculate stochastic gradients $\nabla f_{x_t}(w_t)$ in our algorithm. To ensure $E[\|\nabla f(w_R)\|_2^2] < \epsilon$, total number of stochastic GDA-AM calls needed to calculate stochastic gradients $\nabla f_{x_t}(w_t)$ is $O(\epsilon^{-2})$.

Recall that we can recast GDA scheme as the following fixed point iteration.

$$w_{t+1} = G^{(\text{sim})}_\eta(w_t) = w_t + \eta V(w_t)$$

with $w = \begin{bmatrix} x \\ y \end{bmatrix}$, $V(w) = \begin{bmatrix} -\nabla_x f(x, y) \\ \nabla_y f(x, y) \end{bmatrix}$.

Ignoring the stepsize $\eta$ and let $W_t$ and $R_t$ record the first and second order difference of recent m iterates:

$$W_t = [\Delta w_{t-m}, \Delta w_{t-m+1}, \ldots, \Delta w_{t-1}], R_t = [\Delta V_{t-m}, \Delta V_{t-m+1}, \ldots, \Delta V_{t-1}]$$

Similarly as Wei et al. (2021b), the Anderson mixing can be decoupled into

$$\bar{w}_{t+1} = w_t - W_t \Gamma_t, \quad (\text{Projection step})$$

$$\bar{w}_{t+1} = w_t + \beta_t \bar{V}_t, \quad (\text{Mixing step})$$

where $\beta_t$ is the mixing parameter, and $\bar{V}_t = V_t - W_t \bar{\Gamma}_t$ and $\bar{\Gamma}_t$ is solved by

$$\Gamma_t = \arg \min_{\Gamma \in \mathbb{R}^m} \|V_t - R_t \Gamma\|_2 + \delta_t \|\Gamma\|_2$$

We want to argue that similar arguments in Wei et al. (2021b) can be applied to the problem here. To see why Anderson mixing works for minimax optimization, we assume function $f$ is smooth. Then the hessian matrix for $G^{(\text{sim})}_\eta$ is

$$H = \begin{bmatrix} -\nabla^2_{xx} f & -\nabla^2_{xy} f \\ \nabla^2_{yx} f & -\nabla^2_{yy} f \end{bmatrix}$$
Notice that in a small neighborhood of $w_{t+1}$, we have

$$R_t = -HW_t = \begin{pmatrix} \nabla^2_{xx}f & \nabla^2_{xy}f \\ -\nabla^2_{yx}f & -\nabla^2_{yy}f \end{pmatrix} W_t$$

Thus $\|V_t - R_t \Gamma\|_2 \approx \|V_t + HW_t \Gamma\|_2$, which is equivalent to solving for a vector $p_t$ such that $H_p = V_t$. This is exactly the second order method for the fixed point iteration problem. Also at each step the AM is minimizing the residual, the reason that AM is equivalent to GMRES for linear problem is that this quadratic approximation is exact. Finally, we rewrite AM as the quasi-newton framework as [Wei et al. (2021b)] did. $w_{t+1} = w_t + H_t V_t$ where

$$\min_{H_t} \|H_t - \beta I\|_F \text{ subject to } H_t R_t = -X_t$$

Finally, with damping parameter, Anderson mixing has the following form

$$W_{t+1} = W_t + \beta_t V_t - \alpha_t (W_t + \beta_t R_t) \Gamma_t$$

we can also apply the very similar arguments to prove key results in lemma 1, lemma 2 in [Wei et al. (2021b)]. There is also a key difference with [Wei et al. (2021b)]. Here we are considering minimax optimization problem. Thus our gradient is actually $V(w) = \begin{bmatrix} -\nabla_x f(x,y) \\ \nabla_y f(x,y) \end{bmatrix}$ rather than $\nabla f(w) = \begin{bmatrix} \nabla_x f(x,y) \\ \nabla_y f(x,y) \end{bmatrix}$. This will introduce some difficulty to the dynamics of the fixed point iteration. However, noticing that $\|V\| = \|\nabla f(w)\|$ and

$$f(w_{t+1}) \leq f(w_t) + \nabla f(w_t)^T(w_{t+1} - w_t) + \frac{L}{2}\|w_{t+1} - w_t\|_2^2$$

$$\leq f(w_t) + \nabla f(w_t)^T(w_{t+1} - w_t) + \frac{L}{2}\|H_t V_t\|_2^2$$

$$= f(w_t) + \nabla f(w_t)^T H_t V_t + \frac{L}{2}\|H_t V_t\|_2^2$$

where

$$\nabla f(w_t) = \begin{bmatrix} -\nabla_x f(x,y) \\ \nabla_y f(x,y) \end{bmatrix}$$

we call this the ascent-descent gradient (ADG) which is the gradient for minimax optimization problem

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} f(x,y).$$

To see why $\nabla f(w_t)^T(w_{t+1} - w_t) \leq \nabla f(w_t)^T(w_{t+1} - w_t)$, we consider their difference

$$(\nabla - \nabla)f(w_t)^T(w_{t+1} - w_t) = -2\nabla_x f(x_t, y_t)^T(x_{t+1} - x_t)$$

For fixed $y_t$, $f(x_t, y_{t+1})$ has the Taylor expansion:

$$f(x_{t+1}, y_t) = f(x_t, y_t) + \nabla_x f(x_t, y_t)^T(x_{t+1} - x_t) + \nabla_y f(x_t, y_t)^T(y_{t+1} - y_t)$$

Assuming $f$ is convex w.r.t $x$ and apply safeguard to ensure $f(x_{t+1}, y_t) \leq f(x_t, y_t)$ can guarantee $$(\nabla - \nabla)f(w_t)^T(w_{t+1} - w_t) \geq 0.$$ Now applying lemmas in [Wei et al. (2021b)], we can derive the convergence of our method for general convex-nonconcave function similarly.

## D Additional Experiments

### D.1 Comparison with EG with Positive Momentum

In this section, we include additional comparison between GDA-AM and EG with positive momentum. GDA-AM has two big theoretical advantages over EG with positive momentum. First, convergence of GDA-AM does not require strong assumptions on choices of hyperparameters. Second, [4.1] and [4.2] provide nonasymptotic guarantees while convergence of EG with positive momentum is asymptotic. Experimental results are shown in [8]. It indicates GDA-AM outperforms EG with positive momentum. Finding a good choice of the inner and outer step size of EG and momentum term is hard. For EG with positive momentum, we set the step size of extrapolation step as 1, the step size of update as 0.5, and the positive momentum term as 0.3 after grid search as shown in [Rb] and [Rc]. On the other hand, GDA-AM converges fast for different step size without hyperparameter tuning.
D.2 1D Minimax functions

We begin with investigating the empirical performance of GDA-AM for 6 non-trivial 1d bivariate functions. We set initial points as (3, 3) and m as 20 or 5 for all functions. We use optimal learning rates for all methods on each problem. Results are shown in Figure 9, 10, 11, 12, 13 and 14. We observe GDA-AM consistently outperforms all baselines and improves convergence. It is worthwhile to mention that the difference between GDA-AM and traditional averaging is twofold. First, traditional averaging does not involve an adaptive averaging scheme and thus blindly converge to (0, 0) for all 1d bivariate functions. In contrast, GDA-AM obtains optimal weights by solving a small linear system on past iterates. Using different weights for each iteration, GDA-AM is able to minimize the residual of past iterates and thus find the solution of a fixed-point iteration. More importantly, averaging does not change the GDA dynamic because averaging generates a new sequence of parameters based on GDA iterates. This means averaging is independent with base training algorithm (GDA here). However, GDA-AM changes the dynamic directly by overwriting the latest iterate. It means Anderson Mixing interacts with GDA, which is another major difference from averaging.

Figure 9: $f(x, y) = (x - \frac{1}{2})(y - \frac{1}{2}) + \frac{1}{2} e^{-(x-0.25)^2-(y-0.75)^2}$. The optima for this function is not (0, 0). Because averaging blindly converges to (0, 0), it can never find the correct solution.

D.3 Density estimation

To test our proposed method, we evaluate our method on two low-dimension density estimation problems, mixture of 25 Gaussians and Swiss roll. For both generator and discriminator, we use fully connected neural networks with 3 hidden layers and 128 hidden units in each layer. Except for the output layer of discriminator that uses a sigmoid activation, we use tanh-activation for all other layers. We run Adam and GDA-AM for 50000 steps. The learning rate is set as $2 \times 10^{-4}$ and $\beta_1 = 0$, $\beta_2 = 0.9$ after an extensive grid search, which is close to the maximal possible stepsize under which the methods rarely diverge. Figure 15 and 16 show the output after $\{1K, 10K, 30K, 50K\}$ iterations. It can be seen that our method converges faster to the target distribution offers a improvement over Adam. In addition, we can observe that the generated samples using our method gather around the circle and are less connected with other circles.
Figure 10: $f(x, y) = (4x^2 - (y - 3x + 0.05x^3)^2 - 0.1y^4)e^{-0.01(x^2+y^2)}$. All baselines except averaging are cycling around the optima. Averaging is converging slowly.

Figure 11: $f(x, y) = -3x^2 - y^2 + 4xy$. Baselines tend to diverge. Averaging is converging slowly again because averaging can only blindly converge to $(0, 0)$ and the optima for this function is $(0, 0)$.

D.4 Robust Neural Network Training

In this section, we test the effectiveness of $GDA-AM$ by training a robust neural network on MNIST data set against adversarial attacks (Madry et al., 2019; Goodfellow et al., 2015; Kurakin et al., 2017). The optimization formulation is

$$\min_w \sum_{i=1}^{N} \max_{\delta_i, s.t. |\delta_i|_\infty \leq \varepsilon} \ell(f(x_i + \delta_i; w), y_i)$$

(45)

where $w$ is the parameter of the neural network, the pair $(x_i, y_i)$ denotes the $i$-th data point, and $\delta_i$ is the perturbation added to data point $i$. The accuracy of our formulation against popular attacks, FGSM (Goodfellow et al., 2015) and PGD (Kurakin et al., 2017), are summarized in Table 2. Since solving such problem is computationally challenging, Nouiehed et al. (2019) proposed an approximation of the above optimization problem with a new objective function as the following nonconvex-concave problem:

$$\min_w \sum_{i=1}^{N} \max_{t^* \in T} \sum_{j=0}^{9} t_j \ell(f(x^K_{ij}, w), y_i), \ T = \left\{(t_1, \cdots, t_m) | \sum_{i=1}^{m} t_i = 1, t_i \geq 0 \right\}$$

(46)

where $K$ is a parameter in the approximation, and $x^K_{ij}$ is an approximated attack on sample $x_i$ by changing the output of the network to label $j$. We use the public available implementation (Nouiehed et al., 2019)\(^2\). We apply our algorithm on top of (Nouiehed et al., 2019) and compare our results ($p = 50$) with (Madry et al., 2019; Zhang et al., 2019; 2020; Nouiehed et al., 2019). Results are summarized in table 2. We can observe that $GDA-AM$ leads to a comparable or slightly better performance to the other methods. In addition, $GDA-AM$ does not exhibit a significant drop in accuracy when $\varepsilon$ is larger and this suggests the learned model is more robust.

\(^2\)https://github.com/optimization-for-data-driven-science/Robust-NN-Training
In this section, we provide additional experimental results that are not given in Section 5. Figure 18a and 18b show the Inception Score for CIFAR10 using WGAN-GP and SNGAN. It can be observed that our method consistently performs better than Adam and EG during training. Further, on CIFAR-10 using WGAN-GP and SNGAN, GDA-AM is slightly slower than Adam (about 110-115% computational time), but significantly faster than EG (about 65-75% computational time).

D.6 DETAILS ON THE EXPERIMENTS

For our experiments, we used the PyTorch deep learning framework. Experiments were run on one NVIDIA V100 GPU. The residual network architecture for generator and discriminator are summarized in Table 3 and 4. We use a WGAN-GP loss, with gradient penalty $\lambda = 10$. When using the gradient penalty (WGAN-GP), we remove the batch normalization layers in the discriminator. When using SNGAN, we replace the batch normalization layers with spectral normalization. Hyperparameters of Adam are selected after grid search. We use a learning rate of $2 \times 10^{-4}$ and batch size of 64. For table size of GDA-AM, we set it as 120 for CIFAR10 and 150 for CelebA. We set $\beta_1 = 0.0$ and $\beta_2 = 0.9$ as we find it gives us better models than default settings.

Table 2: Test accuracies under FGSM and PGD attack. Trade refers to Zhang et al. (2019).

<table>
<thead>
<tr>
<th></th>
<th>FGSM $L_\infty$</th>
<th>PGD$^{40}L_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon = 0.2$</td>
<td>$\epsilon = 0.3$</td>
</tr>
<tr>
<td>Madry et al. (2019)</td>
<td>98.58%</td>
<td>96.09%</td>
</tr>
<tr>
<td>Trade: $\epsilon = 0.3$</td>
<td>97.37%</td>
<td>95.47%</td>
</tr>
<tr>
<td>Trade: $\epsilon = 0.4$</td>
<td>97.21%</td>
<td>96.19%</td>
</tr>
<tr>
<td>Nouiehed et al. (2019)</td>
<td>98.20%</td>
<td>97.04%</td>
</tr>
<tr>
<td>Zhang et al. (2019)</td>
<td>98.89%</td>
<td>97.87%</td>
</tr>
<tr>
<td>GDA-AM</td>
<td>98.61%</td>
<td>97.75%</td>
</tr>
</tbody>
</table>

1https://pytorch.org/
Figure 14: \( f(x, y) = 2x^2 + y^2 + 4xy + \frac{4}{3}y^3 - \frac{1}{4}y^4 \).

(a) Adam  
(b) GDA-AM

Figure 15: 25 Gaussians: Evolution plot of Adam and GDA-AM. Green dots are observed points and red dots are generated points.

(a) Adam  
(b) GDA-AM

Figure 16: Swiss roll: Evolution plot of Adam and GDA-AM. Green dots are observed points and red dots are generated points.
Figure 17: FID (lower or ↓ is better) for CIFAR 10

Figure 18: **Left:** IS for CIFAR10 using WGAN-GP. **Middle:** IS for CIFAR10 using SNGAN. **Right:** FID for CelebA using WGAN-GP.

Figure 19: Generated Images for CIFAR10 and CelebA using WGAN-GP(ResNet)
### Table 3: ResNet architecture used for our CIFAR-10 experiments.

<table>
<thead>
<tr>
<th>Generator</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: $z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear 128 $\rightarrow$ 256 $\times$ 4 $\times$ 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 128 $\rightarrow$ 128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 256 $\rightarrow$ 256</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 256 $\rightarrow$ 256</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batch Normalization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReLu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>transposed conv. (256, kernel: $3 \times 3$, stride: 1, pad: 1)</td>
<td>$\tanh(\cdot)$</td>
<td>$\tanh(\cdot)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discriminator</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: $x \in \mathbb{R}^{3\times32\times32}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear 128 $\rightarrow$ 128 $\times$ 4 $\times$ 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 128 $\rightarrow$ 128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 128 $\rightarrow$ 128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 128 $\rightarrow$ 128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear 128 $\rightarrow$ 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: ResNet architecture used for our CelebA (64 $\times$ 64) experiments.

<table>
<thead>
<tr>
<th>Generator</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: $z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear 128 $\rightarrow$ 512 $\times$ 8 $\times$ 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 512 $\rightarrow$ 256</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 256 $\rightarrow$ 128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 128 $\rightarrow$ 64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batch Normalization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReLu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>transposed conv. (64, kernel: $3 \times 3$, stride: 1, pad: 1)</td>
<td>$\tanh(\cdot)$</td>
<td>$\tanh(\cdot)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discriminator</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: $x \in \mathbb{R}^{3\times64\times64}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear 128 $\rightarrow$ 128 $\times$ 4 $\times$ 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 128 $\rightarrow$ 128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 128 $\rightarrow$ 256</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResBlock 256 $\rightarrow$ 512</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear 512 $\rightarrow$ 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>