Convergence Behaviour of Some Gradient-Based Methods on Bilinear Zero-Sum Games

Guojun Zhang and Yaoliang Yu
University of Waterloo, Waterloo AI Institute, Vector Institute
{guojun.zhang,yaoliang.yu}@uwaterloo.ca

Abstract

Min-max formulations have attracted much attention in the ML community due to the rise of deep generative models and adversarial methods, and understanding the dynamics of (stochastic) gradient algorithms for solving such formulations has been a grand challenge. As a first step, we restrict to bilinear zero-sum games and give a systematic analysis of popular gradient updates, for both simultaneous and alternating versions. We provide exact conditions for their convergence and find the optimal parameter setup and convergence rates. In particular, our results offer formal evidence that alternating updates converge "better" than simultaneous ones.

1 Introduction

Min-max optimization has received significant attention due to the popularity of generative adversarial networks (GANs) [1] and adversarial training [2], just to name some examples. Formally, given a (bivariate) objective function $f(x, y)$, we aim to find a *saddle point* (x^*, y^*) such that

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*),\ \forall x \in \mathbb{R}^n,\ \forall y \in \mathbb{R}^n.$$

Since the beginning of game theory, various algorithms have been proposed for finding saddle points [e.g. 3–11]. Due to its recent resurgence in ML, new algorithms designed for training GANs were proposed [e.g. 12–15]. However, due to non-convexity in deep learning formulations, our understanding of the convergence behaviour of new and classic gradient algorithms is still limited, and existing analysis mostly focused on bilinear games [e.g. 12, 14] or strongly-convex-strongly-concave games [e.g. 16–18]. Non-zero-sum bilinear games, on the other hand, are PPAD-complete [19] (for the definition see [20]; for finding approximate Nash equilibria, see e.g. [21]).

In this work, we focus on bilinear zero-sum games as a first step towards understanding general min-max optimization, although our results apply to some simple GAN settings [22]. It is well-known that certain gradient algorithms converge at a linear rate on bilinear zero-sum games [6, 7, 16, 17]. These iterative algorithms usually come with two versions: *Jacobi* style or *Gauss–Seidel* (GS) style. In Jacobi style, we update the two sets of parameters (i.e., x and y) *simultaneously* whereas in GS style we update them *alternatively* (i.e., one after the other). Thus, Jacobi style updates are naturally amenable to parallelization while GS style updates have to be sequential, although the latter are usually found to converge faster (and more stable). In numerical linear algebra, the celebrated Stein–Rosenberg theorem [23] formally proves that in solving certain linear systems, GS updates converge *strictly* faster than their Jacobi counterparts, and often with a larger set of convergent instances. However, this result does not readily apply to bilinear zero-sum games (see §3).

Our main goal here is to answer the following questions about solving bilinear zero-sum games:

- When exactly does a gradient-type algorithm converge?
- What is the optimal convergence rate by tuning the step size or other parameters?
- Can we prove something similar to the Stein–Rosenberg theorem for Jacobi and GS updates?

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NeurIPS 2019), Vancouver, Canada.

Table 1: Comparisons between Jacobi and Gauss–Seidel updates. The second and third columns show when exactly an algorithm converges, with Jacobi or GS updates. The last column shows whether the convergence region of the Jacobi update is contained in the GS convergence region.

Algorithm	Jacobi	Gauss–Seidel	Contained?
EG	Theorem 3.1	Theorem 3.1	if $2\beta + \alpha^2 < 2/\sigma_1^2$
OGD	Theorem 3.2	Theorem 3.2	yes
momentum	does not converge	Theorem 3.3	yes

Table 2: Optimal convergence rates. In the second column, β_* denotes a specific parameter that depends on σ_1 and σ_n. In the third column, the linear rates are asymptotic given large κ. The optimal parameters for both Jacobi and Gauss–Seidel EG algorithms are the same.

Algorithm	α	β_1	β_2	rate exponent	Comment
EG	~ 0	$2/(\sigma_1^2 + \sigma_n^2)$	β_1	$\sim 1 - 2/\kappa^2$	Jacobi and Gauss–Seidel
Jacobi OGD	$2\beta_1$	β_*	β_1	$\sim 1 - 1/(6\kappa^2)$	$\beta_1 = \beta_2 = \alpha/2$
GS OGD	$\sqrt{2}/\sigma_1$	$\sqrt{2}\sigma_1/(\sigma_1^2 + \sigma_n^2)$	0	$\sim 1 - 1/\kappa^2$	β_1 and β_2 can switch

Contributions We summarize our main results from §3 and §4 in Table 1 and 2 resp., with supporting experiments given in §5. We use σ_1 and σ_n to denote the largest and the smallest singular values of matrix E (cf. eq. (2.1)), and $\kappa := \sigma_1/\sigma_n$ denotes the condition number. The algorithms will be introduced in §2. Note that we generalize gradient-type algorithms but retain the same names.

2 Preliminaries

Mathematically, zero-sum *bilinear* games can be formulated as the following min-max problem:

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^n} \quad x^\top E y + b^\top x + c^\top y. \tag{2.1}$$

(Throughout for simplicity we assume E is invertible.) For bilinear games, it is well-known that simultaneous gradient descent does not converge [10] and other gradient-based algorithms tailored for min-max optimization have been proposed [6, 12, 15, 22]. These iterative algorithms all belong to the class of general linear dynamical systems (LDSs), and they can be described as:

$$z^{(t)} = \sum_{i=1}^{k} A_i z^{(t-i)} + d, \quad z^{(t)} := (x^{(t)}, y^{(t)}). \tag{2.2}$$

The following well-known result decides when such a k-step LDS converges for any initialization:

Theorem 2.1 (e.g. [24])**.** *The LDS $z^{(t)} = \sum_{i=1}^{k} A_i z^{(t-i)} + d$ converges for any initialization $(z^{(0)}, \ldots, z^{(k-1)})$ iff the spectral radius $r := \max\{|\lambda| : \det(\lambda^k I - \sum_{i=1}^{k} A_i \lambda^{k-i}) = 0\} < 1$, in which case $\{z^{(t)}\}$ converges linearly with (asymptotic) exponent r.*

Therefore, understanding the bilinear game dynamics reduces to spectral analysis. The (sufficient and necessary) convergence condition reduces to that all roots of the characteristic polynomial lie in the unit circle, which can be conveniently analyzed through the celebrated Schur's theorem [25].

Let us formally define Jacobi and GS updates: Jacobi updates take the form

$$x^{(t)} = T_1(x^{(t-1)}, y^{(t-1)}, \ldots, x^{(t-k)}, y^{(t-k)}), \; y^{(t)} = T_2(x^{(t-1)}, y^{(t-1)}, \ldots, x^{(t-k)}, y^{(t-k)}),$$

while Gauss–Seidel updates replace $x^{(t-i)}$ with the more recent $x^{(t-i+1)}$ in operator T_2, where $T_1, T_2 : \mathbb{R}^{nk} \times \mathbb{R}^{nk} \to \mathbb{R}^n$ can be any update functions. For LDS updates in (2.2) we find a nice relation between the characteristic polynomials of Jacobi and GS updates:

Theorem 2.2 (Jacobi vs. Gauss–Seidel). *Let $p(\lambda, \gamma) = \det(\sum_{i=1}^{k} (\gamma L_i + U_i) \lambda^{k-i} - \lambda^k I)$, where $A_i = L_i + U_i$ and L_i is strictly lower block triangular. Then, the characteristic polynomial of the Jacobi update is $p(\lambda, 1)$ while that of the Gauss–Seidel update is $p(\lambda, \lambda)$.*

Next, we define some popular gradient algorithms for finding saddle points in the min-max problem $\min_x \max_y f(x, y)$. Unlike their usual presentations, we introduced more "step sizes" for refined analysis, as the enlarged parameter space often contain choices for faster linear convergence (see §4). We only define the Jacobi updates, while the GS counterparts can be easily inferred.

Extra-gradient (EG) We study a generalized version of EG, defined as follows:

$$x^{(t+1/2)} = x^{(t)} - \gamma_2 \nabla_x f(x^{(t)}, y^{(t)}), \qquad y^{(t+1/2)} = y^{(t)} + \gamma_1 \nabla_y f(x^{(t)}, y^{(t)}); \tag{2.3}$$

$$x^{(t+1)} = x^{(t)} - \alpha_1 \nabla_x f(x^{(t+1/2)}, y^{(t+1/2)}),\ y^{(t+1)} = y^{(t)} + \alpha_2 \nabla_y f(x^{(t+1/2)}, y^{(t+1/2)}). \tag{2.4}$$

EG was first proposed in [6] with the restriction $\alpha_1 = \alpha_2 = \gamma_1 = \gamma_2$, under which linear convergence was proved for bilinear games. A slightly more generalized version was analyzed in [16] where $\alpha_1 = \alpha_2$, $\gamma_1 = \gamma_2$, again with linear convergence proved. For later convenience we define $\beta_i = \alpha_i \gamma_i$.

Optimistic gradient descent (OGD) We study a generalized version of OGD, defined as follows:

$$x^{(t+1)} = x^{(t)} - \alpha_1 \nabla_x f(x^{(t)}, y^{(t)}) + \beta_1 \nabla_x f(x^{(t-1)}, y^{(t-1)}), \tag{2.5}$$

$$y^{(t+1)} = y^{(t)} + \alpha_2 \nabla_y f(x^{(t)}, y^{(t)}) - \beta_2 \nabla_y f(x^{(t-1)}, y^{(t-1)}). \tag{2.6}$$

The original version of OGD was given in [12] with $\alpha_1 = \alpha_2 = 2\beta_1 = 2\beta_2$, and its linear convergence for bilinear games was proved in [16]. A slightly generalized version with $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ was analyzed in [17], again with linear convergence proved.

Momentum method Generalized heavy ball method was proposed and analyzed in [14]:

$$x^{(t+1)} = x^{(t)} - \alpha_1 \nabla_x f(x^{(t)}, y^{(t)}) + \beta_1 (x^{(t)} - x^{(t-1)}), \tag{2.7}$$

$$y^{(t+1)} = y^{(t)} + \alpha_2 \nabla_y f(x^{(t)}, y^{(t)}) + \beta_2 (y^{(t)} - y^{(t-1)}), \tag{2.8}$$

as a modification of Polyak's heavy ball (HB) [26], which also motivated Nesterov's accelerated gradient algorithm (NAG) [27]. For bilinear games, HB and NAG are the same and hence we call both the momentum method. For this algorithm our result below improves those obtained in [14].

3 Exact conditions

With tools from §2, we give necessary and sufficient conditions under which a gradient-based algorithm converges for bilinear games. For simplicity, we mostly take the parameters for the two sets of variables to be the same, i.e., $\alpha_1 = \alpha_2 = \alpha$, $\beta_1 = \beta_2 = \beta$ and $\gamma_1 = \gamma_2 = \gamma$ (if available). The same conditions for more general algorithms can be found in our complete paper.

Theorem 3.1 (**EG**). *For generalized EG with $\alpha_1 = \alpha_2 = \alpha$ and $\gamma = \beta/\alpha$, linear convergence is achieved iff for any singular value σ of E, we have $\alpha^2\sigma^2 + (\beta\sigma^2 - 1)^2 < 1$ for the Jacobi update, and $0 < \beta\sigma^2 < 2$ and $|\alpha\sigma| < 2 - \beta\sigma^2$ for the GS update. If $2\beta + \alpha^2 < 2/\sigma_1^2$, the convergence region of GS updates **strictly** include that of Jacobi updates.*

Theorem 3.2 (**OGD**). *For generalized OGD with $\alpha_1 = \alpha_2 = \alpha$, linear convergence is achieved iff for any singular value σ of E, we have: $0 < \beta\sigma < 1$, $\beta < \alpha < \beta\frac{3-\beta^2\sigma^2}{1+\beta^2\sigma^2}$ for the Jacobi update, and $|\alpha + \beta|\sigma < 2$, $|1 + \alpha\beta\sigma^2| > 1 + \beta^2\sigma^2$ for the GS update. The convergence region of GS updates **strictly** include that of Jacobi updates.*

Theorem 3.3 (**momentum**). *For generalized momentum with $\alpha_1 = \alpha_2 = \alpha$, the Jacobi update never converges, while the GS update with $\beta_1 = \beta_2 = \beta$ converges iff for any singular value σ of E, we have $-1 < \beta < 0$, $|\alpha\sigma| < 2(1 + \beta)$. If $\beta_2 = 0$, the exact condition is $-1 < \beta_1 < 0$ and $0 < \alpha\sigma_1 < 2\sqrt{1 + \beta_1}$.*

Prior to our work, only sufficient conditions for linear convergence are given for the usual EG and OGD; cf. §2 above. For the momentum method, our result improves upon [14] where the authors only considered specific cases of parameters. For example, they only considered $\beta \geq -1/16$ for Jacobi momentum, and $\beta_1 = -1/2$, $\beta_2 = 0$ for GS momentum. Our Theorem 3.3 gives a more complete picture. (For an even more general result please refer to our full paper.)

In the theorems above, we use the term "convergence region" to denote a set of the parameters (α, β or γ) where the algorithm converges. Our result shares similarity with the Stein–Rosenberg theorem [23], which only applies to solving linear systems with non-negative matrices. If one applies it to our case, the matrix E in (2.1) should be the trivial zero matrix (see the full paper). In this sense, our results extend the Stein–Rosenberg theorem to cover nontrivial bilinear games.

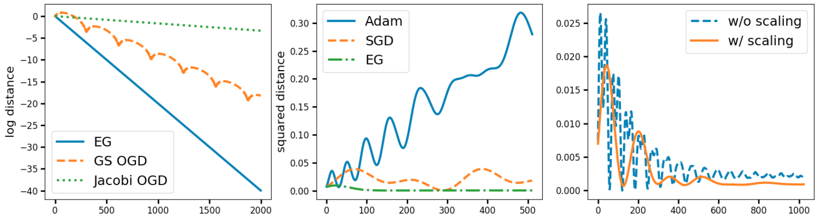

Figure 1: **Left:** linear convergence of optimal EG, Jacobi OGD, Gauss–Seidel OGD in a bilinear game. **Middle:** comparison among Adam, SGD and EG in learning the mean of a Gaussian with WGAN. **Right:** Comparison between EG with ($\alpha = 0.02$, $\gamma = 2.0$) and without scaling ($\alpha = \gamma = 0.2$).

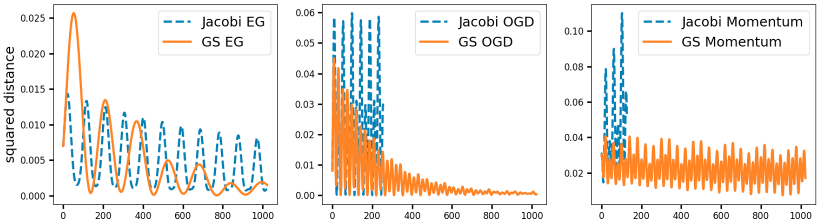

Figure 2: Jacobi vs. GS updates. **Left:** EG with $\gamma = 0.2, \alpha = 0.02$; **Middle:** OGD with $\alpha = 0.2$, $\beta_1 = 0.1$, $\beta_2 = 0$; **Right:** Momentum with $\alpha = 0.08$, $\beta = -0.1$. We plot only a few epochs for Jacobi updates if they do not converge.

4 Optimal rates

In this section we study the optimal convergence rates of EG and OGD. We define the exponent of linear convergence as $r = \lim_{t\to\infty} ||z^{(t)}||/||z^{(t-1)}||$. For ease of presentation we fix $\alpha_1 = \alpha_2 = \alpha > 0$ and we use r_* to denote the optimal rate (w.r.t. the parameters α, β, γ). In Theorem 4.2, the exact formula β_* in Jacobi OGD, as well as more relevant results, can be found in our full paper.

Theorem 4.1 (**EG optimal**). *Both Jacobi and GS EG achieve the optimal exponent of linear convergence $r_* = (\kappa^2 - 1)/(\kappa^2 + 1)$ at $\alpha \to 0$ and $\beta_1 = \beta_2 = 2/(\sigma_1^2 + \sigma_n^2)$. As $\kappa \to \infty$, $r_* \to 1 - 2/\kappa^2$.*

Theorem 4.2 (**OGD optimal**). *For Jacobi OGD with $\beta_1 = \beta_2 = \beta$, to achieve the optimal linear convergence, we must have $\alpha \le 2\beta$. At $\beta = \alpha/2 = \beta_*$, $r_* \sim 1 - 1/(6\kappa^2)$ at large κ. For GS OGD with $\beta_2 = 0$, $r_* = \sqrt{(\kappa^2 - 1)/(\kappa^2 + 1)} \sim 1 - 1/\kappa^2$, at $\alpha = \sqrt{2}/\sigma_1$ and $\beta_1 = \sqrt{2}\sigma_1/(\sigma_1^2 + \sigma_n^2)$.*

5 Experiments

Bilinear game We experiment on a bilinear game and choose the optimal parameters as suggested in Theorem 4.1 and 4.2. The results, shown in the left panel of Figure 1, agree with our theory.

WGAN example As in [12], we consider a WGAN [28] that learns the mean of a Gaussian:

$$\min_\phi \max_\theta f(\phi, \theta) := \mathbb{E}_{x\sim\mathcal{N}(v,\sigma^2 I)}[s(\theta^T x)] - \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I)}[s(\theta^T(z + \phi))], \tag{5.1}$$

with $s(x)$ the sigmoid function. Near the saddle point $(\theta^*, \phi^*) = (0, v)$ the min-max optimization can be treated as a bilinear game (see the full paper). Since we are doing stochastic versions of the algorithms, we should not expect they will converge exactly to a saddle point. Instead, convergence to a neighborhood is good enough.

With GS updates, we find that Adam diverges, SGD goes around a limit cycle, and EG converges, as shown in the middle panel of Figure 1. Our next experiment shows that generalized algorithms may have an advantage over traditional ones. Inspired by Theorem 4.1, we compare the convergence of two EGs with the same parameter $\beta = \alpha\gamma$, and find that with scaling EG converges faster to a neighborhood of the saddle point with less oscillation, as shown in the right panel of Figure 1. Note that we always use the squared distance as a measure of convergence.

Finally, we compare Jacobi updates with GS updates. In Figure 2, GS updates converge even when the corresponding Jacobi updates do not. In the left panel, Jacobi EG and GS EG do not differ so much because they are very similar given small α (see the full paper). In this case, we can parallelize Jacobi EG to obtain faster convergence.

Acknowledgement

We thank the reviewers for their comments, and Argyrios Deligkas and Sarath Pattathil for pointing out additional references. This work is supported by NSERC.

References

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

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks". In: *ICLR*. 2018.

[3] K.J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in linear and non-linear programming*. 1958.

[4] V. F. Dem'yanov and A. B. Pevnyi. "Numerical methods for finding saddle points". In: *USSR Computational Mathematics and Mathematical Physics* 12.5 (1972), pp. 11–52.

[5] E. G. Gol'shtein. "A generalized gradient method for finding saddlepoints". In: *Ekonomika i matematicheskie metody* 8.4 (1972), pp. 569–579.

[6] GM Korpelevich. "The extragradient method for finding saddle points and other problems". In: *Matecon* 12 (1976), pp. 747–756.

[7] R Tyrrell Rockafellar. "Monotone operators and the proximal point algorithm". In: *SIAM journal on control and optimization* 14.5 (1976), pp. 877–898.

[8] Ronald E. Bruck. "On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space". In: *Journal of Mathematical Analysis and Applications* 61.1 (1977), pp. 159–164.

[9] P. L. Lions. "Une methode iterative de resolution d'une inequation variationnelle". In: *Israel Journal of Mathematics* 31.2 (1978), pp. 204–208.

[10] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* 1983.

[11] Yoav Freund and Robert E Schapire. "Adaptive game playing using multiplicative weights". In: *Games and Economic Behavior* 29.1-2 (1999), pp. 79–103.

[12] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. "Training gans with optimism". In: *ICLR*. 2018.

[13] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *ICLR*. 2015.

[14] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Gabriel Huang, Remi Lepriol, Simon Lacoste-Julien, and Ioannis Mitliagkas. "Negative momentum for improved game dynamics". In: *AISTATS*. 2019.

[15] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. "The numerics of GANs". In: *Advances in Neural Information Processing Systems*. 2017, pp. 1825–1835.

[16] Tengyuan Liang and James Stokes. "Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks". In: *AISTATS*. 2019.

[17] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. "A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach". In: *arXiv preprint arXiv:1901.08511* (2019).

[18] Paul Tseng. "On linear convergence of iterative methods for the variational inequality problem". In: *Journal of Computational and Applied Mathematics* 60.1-2 (1995), pp. 237–252.

[19] Xi Chen, Xiaotie Deng, and Shang-Hua Teng. "Settling the complexity of computing two-player Nash equilibria". In: *Journal of the ACM* 56.3 (2009), p. 14.

[20] Christos H Papadimitriou. "On the complexity of the parity argument and other inefficient proofs of existence". In: *Journal of Computer and system Sciences* 48.3 (1994), pp. 498–532.

[21] Argyrios Deligkas, John Fearnley, Rahul Savani, and Paul Spirakis. "Computing approximate Nash equilibria in polymatrix games". In: *Algorithmica* 77.2 (2017), pp. 487–514.

[22] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. "A variational inequality perspective on generative adversarial networks". In: *ICLR*. 2019.

[23] P Stein and RL Rosenberg. "On the solution of linear simultaneous equations by iteration". In: *Journal of the London Mathematical Society* 1.2 (1948), pp. 111–118.

[24] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials*. Academic Press, 1982.

[25] I. Schur. "Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind." In: *Journal für die reine und angewandte Mathematik* 147 (1917), pp. 205–232. English translation: "On power series which are bounded in the interior of the unit circle: I & II" in *Operator theory: Advances and applications*, vol. 18, 1986, edited by I. Gohberg.

[26] B. T. Polyak. "Some methods of speeding up the convergence of iteration methods". In: *USSR Computational Mathematics and Mathematical Physics* 4.5 (1964), pp. 1–17.

[27] Yurii Nesterov. "A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$". In: *Doklady AN USSR* 269 (1983), pp. 543–547.

[28] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein Generative Adversarial Networks". In: *ICML*. 2017.