Characterization of Local Convergence for Min-max Saddle Point Problems

Name: Jie Wang

ID: 116010214

Fourth Year Undergraduate Student (Pure Math Major)

Abstract

In this project paper, we review the convergence properties of iterations designed for the min-max saddle point problem, by revisiting and extending the classic results raised by Prof. Yin Zhang 20 years ago [9]. We give necessary and sufficient conditions for a stationary point to be a point of strong attraction in the iteration process. This concept not simply gives interpretations of the convergence behaviors certain types of algorithms in the literature, but also motivates the new design of heuristics that may outperform the current state-of-the-art algorithms.

Index Terms

Attraction, repulsion, minimax problem.

I. Introduction

In this paper, we consider the *min-max saddle point problem* of the form

$$\min_{x\in\mathbb{R}^m}\max_{y\in\mathbb{R}^n} f(x,y) \tag{1}$$

where the objective function $f:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$ is smooth. This problem has applications in a variety of topics, such as unsupervised learning [3], statistics [2], game theory [6], and otherwise. In practice, the iteration for solving the min-max problem can be represented as

$$\begin{pmatrix} x^{k+1}\\ y^{k+1}\end{pmatrix} = \begin{pmatrix} x^{k}\\ y^{k}\end{pmatrix} - \alpha^k \begin{pmatrix} I_m & 0\\ 0 & -I_n\end{pmatrix}\cdot B^k\cdot \nabla f(x^k,y^k) \tag{2}$$

where $\alpha^k>0$ denotes the step size, and $B^k\in\mathbb{R}^{(m+n)\times(m+n)}$ is some properly chosen matrix at each iteration k. For instance, for gradient descent-ascent (GDA) algorithm [1],

$$B^k\equiv I,\quad \alpha^k\equiv\alpha;$$

for consensus-type regularization (CR) algorithm [7],

$$B^k\equiv I+\gamma\cdot\begin{pmatrix} 0 & I_m\\ I_n & 0\end{pmatrix},\quad \alpha^k\equiv\alpha.$$

However, neither GDA nor CR algorithm could guarantee convergence for most cases, and it is not clear how to pick better α^k, B^k to obtain more satisfying results.

In this paper, we try to shed some light on the convergence analysis for certain types of iteration formulas with the form (2), by revisiting and extending some simple, potentially important, but rarely noticed results published 20 years ago by Prof. Yin Zhang [9], in which he characterized the local convergence and non-convergence behavior for the standard iterative framework in unconstrained minimization problems, and then found the selective minimization property, i.e., the iterates for certain algorithms will only seek global or good local minima, but escaping the remaining minima. Prof. Zhang's paper naturally motivates the following question:

Given certain conditions on α^k and B^k, what type of minimax saddle points of $f(x, y)$ could guarantee the local convergence or non-convergence of iteration (2)?

A. Contributions

In this paper, we give partial answers to the question above, and present several Heuristic algorithms that may outperform the current state-of-the-art one. The main contributions are summarized below:

1) Review the definition of points of attraction and points of repulsion, which means whether the given algorithm could achieve convergence to the desired point locally or not.
2) Provide the necessary and sufficient conditions for the desired point to be a point of strong attraction in Theorem 2 and Corollary 1. Then we give interpretations of the convergence behaviors of GDA and CR in Proposition 1.
3) Design new type of algorithms based on the characterization of points of attraction, i.e., GDA with Newton's Heuristic and CR with Newton's Heuristic.

B. Fixed Point Iteration of (2)

The iteration formula (2) can be represented as a fixed point iteration:

$$(x^{k+1}, y^{k+1}) = T(x^k, y^k), \tag{3a}$$

where T is some function from $\mathbb{R}^{m+n}$ to $\mathbb{R}^{m+n}$ with the form

$$T(x, y) = \begin{pmatrix} x \\ y \end{pmatrix} - \alpha(x, y) \begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} B(x, y) \nabla f(x, y) \tag{3b}$$

By assuming $B(x, y)$ is invertible, it's clear that the point (x^*, y^*) is a first order stationary point (FOSP) of $f(x, y)$ if and only if (x^*, y^*) is a fixed point of $T(x, y)$. However, we want to study the convergence to min-max saddle points, which further requires classifying fixed points of $T(x, y)$. We will discuss the classification in Section II.

Before that, we need to derive the Jacobian of $T(x, y)$ at FOSP (x^*, y^*), say $T'(x^*, y^*)$. From elementary real analysis knowledge [8], we know that the existence of $T'(x^*, y^*)$ does not require the differentiability of (3b:1), but only the continuity of this term at x^*:

Theorem 1. *Let (x^*, y^*) be a FOSP of $f(x, y)$. Suppose that $\alpha(x, y), B(x, y)$ are continuous at (x^*, y^*). Then $T(x, y)$ is differentiable at (x^*, y^*), with*

$$T'(x^*, y^*) = I - \alpha(x^*, y^*) \begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} B(x^*, y^*)\nabla^2 f(x^*, y^*).$$

Proof. Since $B(x, y)$ is non-singular, define

$$H(x, y) \equiv \alpha(x, y) \begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} B(x, y).$$

It suffices to show that the derivative of $H(x, y)\nabla f(x, y)$ at (x^*, y^*) is $H(x^*, y^*)\nabla^2 f(x^*, y^*)$:

$$\begin{aligned}
&\frac{(H \cdot \nabla f)(x^* + \Delta x, y^* + \Delta y) - (H \cdot \nabla f)(x^*, y^*) - (H \cdot \nabla^2 f)(x^*, y^*)(\Delta x, \Delta y)}{\|(\Delta x, \Delta y)\|} \\
=&H(x^* + \Delta x, y^* + \Delta y)\frac{\nabla f(x^* + \Delta x, y^* + \Delta y) - \nabla f(x^*, y^*) - \nabla^2 f(x^*, y^*)(\Delta x, \Delta y)}{\|(\Delta x, \Delta y)\|} \\
&+ [H(x^* + \Delta x, y^* + \Delta y) - H(x^*, y^*)]\frac{\nabla^2 f(x^*, y^*)(\Delta x, \Delta y)}{\|(\Delta x, \Delta y)\|}
\end{aligned}$$

Taking $\|(\Delta x, \Delta y)\| \to 0$, the first term on the RHS vanishes due to the differentiability of $f(x, y)$ at (x^*, y^*); the second term vanishes due to the continuity of $H(x, y)$. □

II. Points of Attraction and Repulsion

In this section, we talk about how to classify fixed points of $T(x, y)$, by introducing the notion of points of attraction and repulsion. Some basic results related to these definitions are also covered.

Definition 1. *A fixed point (x^*, y^*) of $f(x, y)$ is said to be a point of attraction of the iteration (3a) if there exists an open ball N of (x^*, y^*), such that for any initial guess $(x^0, y^0) \in N$, the iterates $\{(x^k, y^k)\}$ generated by (3a) all lie in N, and converge to (x^*, y^*). Otherwise, it is said to be a point of repulsion of (3a).*

Definition 2. *A fixed point (x^*, y^*) of $f(x, y)$ is said to be a point of strong attraction of the iteration (3a) if $T(x, y)$ is differentiable at (x^*, y^*), and $\rho(T'(x^*, y^*)) < 1$.*

Definition 3. *A fixed point (x^*, y^*) of $f(x, y)$ is said to be a point of strong repulsion of the iteration (3a) if $T(x, y)$ is differentiable at (x^*, y^*), and $\rho(T'(x^*, y^*)) > 1$.*

Remark 1.
1) *If (x^*, y^*) is a point of strong attraction, then it is a point of attraction. The converse is not necessarily true.*
2) *If (x^*, y^*) is a point of strong repulsion, then it is a point of repulsion. The converse is not necessarily true.*
3) *The convergence to a point of repulsion is highly unlikely in practice. However, we need rigorous and quantitative research on this claim.*

III. Necessary Conditions for Point of Strong Attraction

We now discuss necessary and sufficient conditions for a stationary point of $f(x,y)$ to be a point of strong attraction of the iteration (3a). It suffices to give some characterziation of the condition $\rho(T'(x^*,y^*)) < 1$, which leads to some new observations that haven't been fully studied in literature.

Theorem 2. *Let (x^*,y^*) be a stationary point of $f(x,y)$, and $T(x,y)$ be defined in (3a). Assume that $\alpha(x,y) > 0$ and $B(x,y), \alpha(x,y)$ are continuous at (x^*,y^*). Then $\rho(T'(x^*,y^*)) \leq 1$ if and only if*

$$\alpha(x^*,y^*) \leq 2 \min_i \cos\left(\lambda_i\left[\begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} B(x^*,y^*)\nabla^2 f(x^*,y^*)\right]\right) \tag{4}$$

Moreover, $\rho(T'(x^*,y^*)) < 1$ (i.e., x^* is a point of strong attraction) if and only if strict inequalities hold in (4).

Proof. Suppose that the i-th eigenvalue

$$\lambda_i\left[\begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} B(x^*,y^*)\nabla^2 f(x^*,y^*)\right] = a_i + ib_i,$$

and $\alpha^* := \alpha(x^*,y^*)$. Therefore, $\rho(T'(h^*)) \leq 1$ is equivalent to

$$|1-\alpha^*(a_i+ib_i)| \leq 1, \forall i \Longleftrightarrow \alpha^* \leq 2\min_i \cos\left(\lambda_i\left[\begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} B(x^*,y^*)\nabla^2 f(x^*,y^*)\right]\right).$$

□

Corollary 1. *Continue the setting in Theorem 2. For sufficiently small step size $\alpha(x^*,y^*)$, x^* is a point of attraction if*

$$\textit{Re}\left(\lambda_i\left[\begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} B(x^*,y^*)\nabla^2 f(x^*,y^*)\right]\right) > 0, \qquad \forall i. \tag{5}$$

By considering some popular choices of $B(h^*)$, we obtain the following facts:

Proposition 1. *Assume that $\alpha(x)$ is continuous at x^*, which is the point of our interest.*

1) *(GDA Algorithm) For $B(x^*,y^*) = I$, any minimax saddle point (x^*,y^*) of $f(x,y)$ is a point of strong attraction, if the point (x^*,y^*) is a **Nash-Equalibrium** minimax saddle point such that $\nabla^2_{xx} f(x^*,y^*) \succ 0$ and $\nabla^2_{yy} f(x^*,y^*) \prec 0$.*
2) *(CR Algorithm) Assume that $m = n$. For $B(x^*,y^*)$ with the form*

$$B(x^*,y^*) = I + \gamma \cdot \begin{pmatrix} 0 & I_m \\ I_n & 0 \end{pmatrix}$$

any minimax saddle point (x^*,y^*) of $f(x,y)$ is a point of strong attraction for sufficiently small $\alpha(x^*,y^*)$, if

$$\textit{Real } \lambda_i\left\{\left[I - \gamma\begin{pmatrix} 0 & I_m \\ I_n & 0 \end{pmatrix}\right]\begin{pmatrix} \nabla^2_{xx} f(x^*,y^*) & \nabla^2_{xy} f(x^*,y^*) \\ -\nabla^2_{yx} f(x^*,y^*) & -\nabla^2_{yy} f(x^*,y^*) \end{pmatrix}\right\} > 0, \quad \forall i$$

3) (Pure Newton's Method) For

$$B(x^*,y^*)=\begin{pmatrix} I_m & 0\\ 0 & -I_n\end{pmatrix}[\nabla^2 f(x^*,y^*)]^{-1},$$

any non-degenerate stationary point of $f(x,y)$ is a point of strong attraction, provided that $\alpha(x^*,y^*)\in(0,2)$.

Proof. 1) It suffices to charaterzie the eigenvalues of the matrix

$$J_1\triangleq\begin{pmatrix} I_m & 0\\ 0 & -I_n\end{pmatrix}\nabla^2 f(x^*,y^*)=\begin{pmatrix} \nabla^2_{xx}f(x^*,y^*) & \nabla^2_{xy}f(x^*,y^*)\\ -\nabla^2_{yx}f(x^*,y^*) & -\nabla^2_{yy}f(x^*,y^*)\end{pmatrix}$$

We can decompose this matrix as:

$$J_1=P_1+Q_1,\quad \text{where } P_1=\begin{pmatrix} \nabla^2_{xx}f(x^*,y^*) & 0\\ 0 & -\nabla^2_{yy}f(x^*,y^*)\end{pmatrix},\quad Q_1=\begin{pmatrix} 0 & \nabla^2_{xy}f(x^*,y^*)\\ -(\nabla^2_{xy}f(x^*,y^*))^{\mathrm{T}} & 0\end{pmatrix}$$

Let (λ,v) be any eigen-pair of J_1, where $v=v_1+iv_2$. It follows that

$$\begin{aligned}\text{Real}(\lambda)&=\frac{1}{2}(\bar{v}^{\mathrm{T}}J_1v+v^{\mathrm{T}}J_1\bar{v})\\ &=v_1^{\mathrm{T}}J_1v_1+v_2^{\mathrm{T}}J_1v_2=(v_1^{\mathrm{T}}P_1v_1+v_2^{\mathrm{T}}P_1v_2)+(v_1^{\mathrm{T}}Q_1v_1+v_2^{\mathrm{T}}Q_1v_2)\end{aligned}$$

Since (x^*,y^*) is a Nash-Equalibrium, we have $P_1\succ 0$, i.e.,

$$v_1^{\mathrm{T}}P_1v_1+v_2^{\mathrm{T}}P_1v_2>0.$$

Moreove, since Q_1 is skew-symmetric, $v_1^{\mathrm{T}}Q_1v_1=v_2^{\mathrm{T}}Q_1v_2=0$. This proves $\text{Re}(\lambda)>0$.

2) It suffices to charaterzie the eigenvalues of the matrix

$$\begin{aligned}J_2&\triangleq\begin{pmatrix} I_m & 0\\ 0 & -I_n\end{pmatrix}B(x^*,y^*)\nabla^2 f(x^*,y^*)\\ &=J_1+\gamma\begin{pmatrix} (\nabla^2_{xy}f(x^*,y^*))^{\mathrm{T}} & \nabla^2_{yy}f(x^*,y^*)\\ -\nabla^2_{xx}f(x^*,y^*) & -\nabla^2_{xy}f(x^*,y^*)\end{pmatrix}\\ &=\left[I-\gamma\begin{pmatrix} 0 & I_m\\ I_n & 0\end{pmatrix}\right]J_1\end{aligned}$$

3) Substituting $B(x^*,y^*)$ into (4) gives the desired result.

□

IV. Heurstic Design and Simulations

We observe that Newton's method search all FOSPs, but the GDA algorithm only converges to the Nash-Equilibrium minimax saddle point. The conjecture is that whether it is possible to design an algorithm converges to the minimax saddle point (not necessarily Nash-Equilibrium) instead of the remaining FOSPs.

Choose $\alpha(x,y)\equiv\alpha$, we wish to pick various kinds of $B(x,y)$ in (3a) to examine the corresponding convergence behaviors, i.e., check whether these algorithms would make the optimum the point of strong attraction or not:

1) Newton's method:

$$B(x,y) = \begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} [\nabla^2 f(x,y)]^{-1},$$

2) GDA Algorithm:

$$B(x,y) = I$$

3) CR Algorithm:

$$B(x,y) = I + \gamma \cdot \begin{pmatrix} 0 & I_m \\ I_n & 0 \end{pmatrix}$$

4) GDA with Newton's Heuristic:

$$B(x,y) = I + \gamma \cdot \begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} [\nabla^2 f(x,y)]^{-1},$$

5) CR with Newton's Heuristic:

$$B(x,y) = B(x,y) = I + \gamma_1 \cdot \begin{pmatrix} 0 & I_m \\ I_n & 0 \end{pmatrix} + \gamma_2 \cdot \begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} [\nabla^2 f(x,y)]^{-1}$$

We consider the special example

$$f(x,y) = -x^2 + 2xy - \frac{1}{2}y^2.$$

It's easy to see that $(0,0)$ is a minimax saddle point since it satisfies the second-order sufficient minimax optimality condition [5]. We will concentrate our attention at the square $-5 \leq x, y \leq 5$. See the Figure for a contour plot of $f(x,y)$ in the square of our interst:

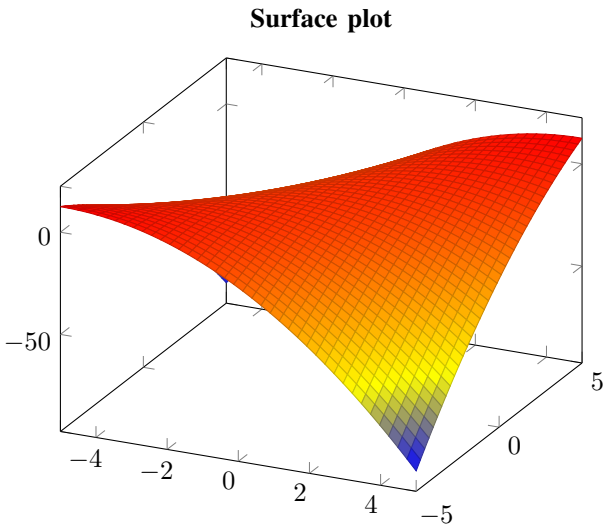

Fig. 1: Funtion plot of $f(x,y) = -x^2 + 2xy - \frac{1}{2}y^2$

By checking the condition (5), we can check whether the saddle point is the point of strong attraction of iteration (3a) or not:

Type of Algorithms	If $(0,0)$ a point of strong attraction
Newton's Method	Y
GDA Algorithm	N
CR Algorithm with $\gamma = 2$	N
GDA with Newton's Heuristic ($\gamma = 2$)	Y
CR with Newton's Heuristic ($\gamma_1 = 0.6, \gamma_2 = 2$)	Y

We run three methods in which $(0,0)$ a point of strong attraction, from the following grid of initial points in the region $-5 \leq x, y \leq 5$:

$$(x_i, y_i) = (i, j)/4, \quad -20 \leq i, j \leq 20.$$

For each method, we record whether or not the iterates converge to teh minimax saddle point, or do not converge within a prescribed maximum number of iterations, which is set to 100 in our experiments. The numerical results for the three methods are summarized below:

1) Newton: About $1/3$ of the starting points led to the minimax saddle point, and the remaining to other stationary points outside the square, or were such that the method idd not terminate after 100 iterations.
2) GDA with Newton's Heuristic: The starting points which converge to the minimax saddle point are along the stright line, i.e., only careful choice of starting points will lead to the convergence to the minimax saddle point
3) CR with Newton's Heuristic: only some points close to the optimum led to the minimax saddle point.

In Figures 2-4, the regions of attraction of the minimax saddle point for these three methods are plotted, where the red dots represent points from which a method converge to the saddle point, and the blue ones refer to the points fail to do so.

V. Final Remarks

Concerning the general iteration

$$\begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ y^k \end{pmatrix} - \alpha^k \begin{pmatrix} I_m & 0 \\ 0 & -I_n \end{pmatrix} \cdot B^k \cdot \nabla f(x^k, y^k),$$

Theorem 2 and Corollary 1 provides interpretations to various choices of α^k and B^k to guarantee local convergence. This has also been formulated using the dynamical linear system in [1], but we think our characterization is more clear and straightforward.

However, it is not clear why the CR algorithm can outperform the GDA algorithm from this perspective now. We hope to answer this question by simplifying the conditions in Theorem 2 in the future. It is also possible to construct new algorithms that skip minimizers but only converge to good minimax saddle points, by combining with some random sampling techniques such as simulated annealing [4]. We call this phenomenon the *selective min-maximization*, which is the generalized concept of selective minimization shown in Prof. Zhang's paper [9].

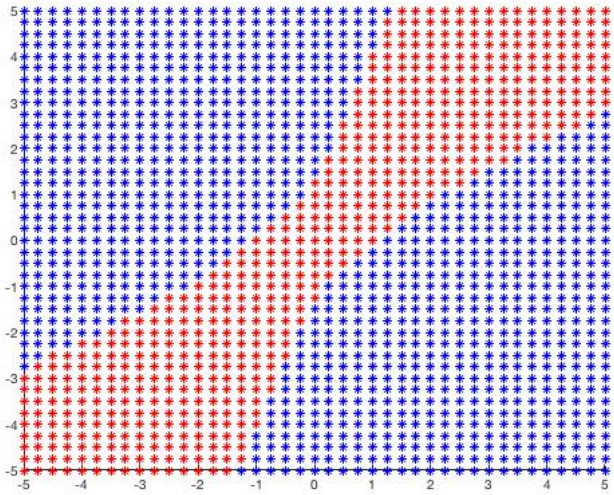

Fig. 2: Newton's method

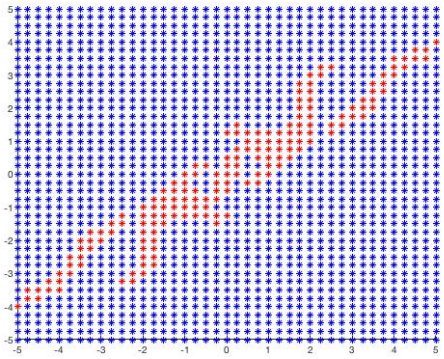

Fig. 3: GDA with Newton's Heuristic

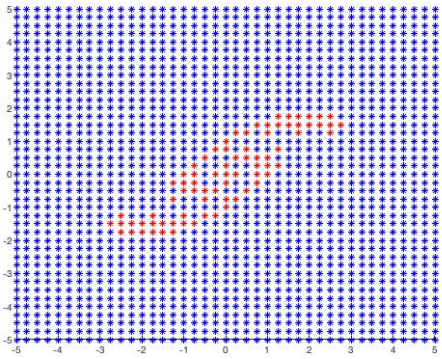

Fig. 4: CR with Newton's Heuristic

REFERENCES

[1] C. DASKALAKIS AND I. PANAGEAS, *The limit points of (optimistic) gradient descent in min-max optimization*, in Proceedings of the 32Nd International Conference on Neural Information Processing Systems, NIPS'18, USA, 2018, Curran Associates Inc., pp. 9256–9266.

[2] F. FARNIA AND D. TSE, *A minimax approach to supervised learning*, in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., Curran Associates, Inc., 2016, pp. 4240–4248.

[3] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, AND Y. BENGIO, *Generative adversarial nets*, in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., Curran Associates, Inc.,

2014, pp. 2672–2680.

[4] V. GRANVILLE, M. KRIVANEK, AND J. . RASSON, *Simulated annealing: a proof of convergence*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 16 (1994), pp. 652–656.

[5] C. JIN, P. NETRAPALLI, AND M. I. JORDAN, *What is local optimality in nonconvex-nonconcave minimax optimization?*, 2019.

[6] E. KALAI, *Game theory: Analysis of conflict: By roger b. myerson, harvard univ. press, cambridge, ma, 1991. 568 pp., 45.00*, Games and Economic Behavior, 3 (1991), pp. 387–391.

[7] L. MESCHEDER, S. NOWOZIN, AND A. GEIGER, *The numerics of gans*, in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Curran Associates, Inc., 2017, pp. 1825–1835.

[8] J. WANG, *Mat2006: Elementary real analysis*, 2019. Available at the link `walterbabyrudin.github.io`.

[9] Y. ZHANG, R. TAPIA, AND L. VELAZQUEZ, *On convergence of minimization methods: Attraction, repulsion, and selection*, Journal of Optimization Theory and Applications, 107 (2000), pp. 529–546.