An $O(s^r)$-Resolution ODE Framework for Discrete-Time Optimization Algorithms and Applications to Minimax Problems

Haihao (Sean) Lu

University of Chicago

INS, SJTU, Nov. 2020

Two papers under review:

Haihao Lu. "An $O(s^r)$-Resolution ODE Framework for Discrete-Time Optimization Algorithms and Applications to Linear Convergence of Minimax Problems."

Benjamin Grimmer, Haihao Lu, Pratik Worah, Vahab Mirrokni. "Limiting Behaviors of Nonconvex-Nonconcave Minimax Optimization via Continuous-Time Systems."

Discrete-Time Algorithms and Ordinary Differential Equations

- Discrete-Time Algorithms (DTA):

$$z^+ = g(z, s)$$

- Ordinary Differential Equations (ODE):

$$\dot{Z} = G(Z)$$

- Comparisons between DTA and ODE
  - DTA is easy to be computed numerically
  - ODE is easy to be analyzed theoretically

Numerical ODE and ODE for DTA

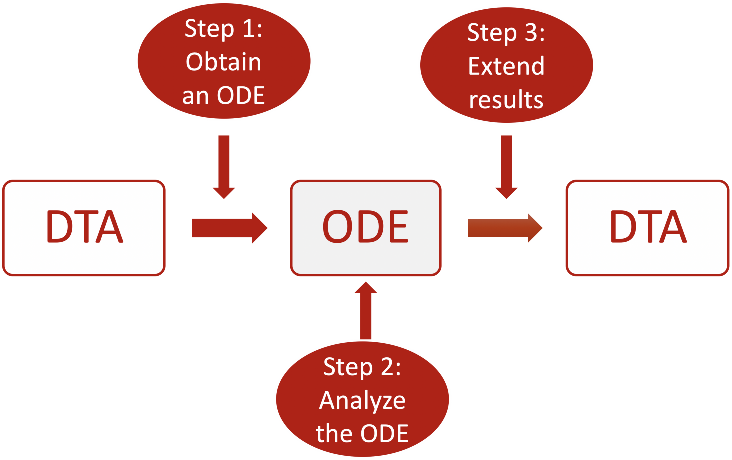
Numerical ODE:

This work:

Using ODEs to Understand Optimization Methods

- History
  - There is a history of using ODE to understand optimization method [Schropp and Singer, 2000]
  - Renewed spark recently [Su, Boyd, Candes, 2014]
  - Hundreds of papers on this topic in the past six years
- Two fundamental open question:
  - How to obtain a suitable ODE from a DTA?
  - What is the connection between the convergence of the ODE and the convergence of the DTA?

Three Major Steps



Motivating Example

- Unconstrained minimax problem

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

- Goal: Find a first-order Nash Equilibrium (x^*, y^*)

$$\nabla_x L(x^*, y^*) = 0 \text{ and } \nabla_y L(x^*, y^*) = 0$$

- New notations

$$z = (x, y) \in \mathbb{R}^{n+m} \text{ and } F(z) = [\nabla_x L(x, y), -\nabla_y L(x, y)] \in \mathbb{R}^{n+m}$$

- Applications: game theory, generative adversarial networks (GANs), robust optimization/robust machine learning

Classic DTAs for Minimax Problems

- Gradient Method (GM):

$$z_+ = z - sF(z)$$

- Proximal Point Method (PPM):

$$z_+ = z - sF(z_+)$$

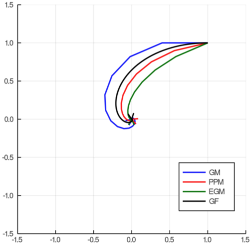
- Extra-Gradient Method (EGM) (it is a also special case of Mirror Prox Algorithm):

$$\tilde{z} = z - sF(z), z_+ = z - sF(\tilde{z})$$

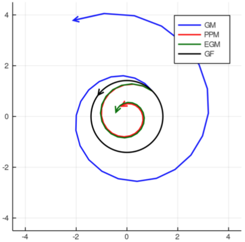
- When $s \to 0$, all above three algorithms converge to gradient flow:

$$\dot{Z} = -F(Z)$$

Behaviors of Different Algorithms



(a) The trajectories of GM, PPM, EGM and GF for solving $\min_x \max_y \frac{1}{2}x^2 + xy - \frac{1}{2}y^2$ with step-size $s = 0.3$ and initial solution $(1, 1)$.



(b) The trajectories of GM, PPM, EGM and GF for solving $\min_x \max_y xy$ with step-size $s = 0.3$ and initial solution $(1, 1)$.

Under What Conditions does PPM/EGM Have Linear Convergence?

Problem of interest:

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

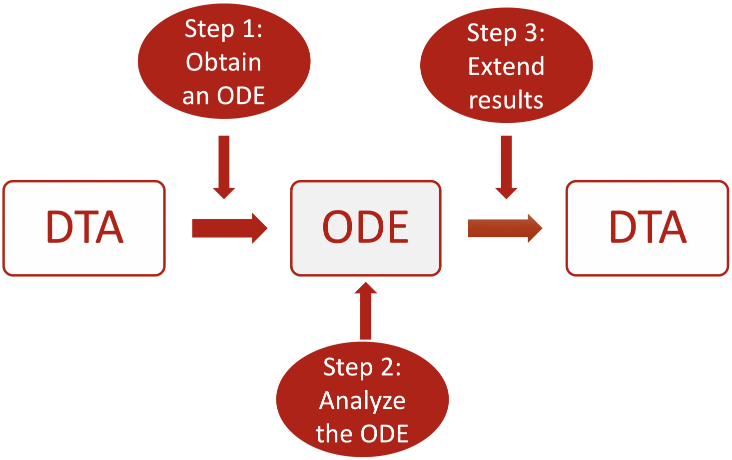
Previous works show that PPM/EGM appear linear convergence when

- $L(x, y)$ is strongly convex-strongly concave, or
- $L(x, y) = x^T By$ is a bilinear function

Question:

- Is there a unified or more fundamental condition and how to obtain it?
- How about nonconvex-nonconcave minimax problems?

Three Major Steps



Step 1: Obtain an ODE from a DTA

- Question: How to obtain a suitable ODE from a DTA?
- Previous works:
  - Mostly let step-size s go to 0
  - Exception [Shi et al, 2018]: high-order resolution ODE to distinguish heavy ball method and accelerated method
- However:
  - Step-size s is never 0 in practice
  - The solution path of a DTA and 0-step-size ODE can be topologically different
  - Different DTAs may collapse to one ODE
- This work:
  - An $O(s^r)$-resolution ODE framework:
    A framework to obtain the unique ODE with certain order of accuracy in normal form

Step 2: Analyze the Convergent Properties of the ODEs

- Previous works:
  - Given the class of problems and an ODE, identify a decaying energy function
- However:
  - It may not always be easy to identify a perfect energy function for this class of problems
- This work:
  - Given the ODE and a reasonable energy fuction, identify the class of problems that the energy function decays

Step 3: Extend the Results from ODEs to DTAs

Question: What is the connection between the convergence of the ODE and the convergence of the DTA?

- Previous works:
  - Prove independently the energy function still decays for the DTA
- However:
  - Some modification of the energy function may be needed
  - Such proof can be highly non-trivial and independent from the proofs for ODEs
- This work:
  - Propose the properness of an energy function
  - Show that the DTAs have linear convergence whenever the $O(s^r)$-resolution ODEs have linear convergence w.r.t. a proper energy function

The $O(s^r)$-Resolution ODE of a DTA

Step 1: Obtain a “good” ODE from a DTA:

The $O(s^r)$-Resolution ODE of a DTA

Generic DTAs

We consider a generic DTA with iterate update:

$$z^{+} = g(z, s) \ ,$$

where

- z is the iterate input
- z^{+} is the iterate output
- s is the step-size of the algorithm
- $g(z, s)$ is sufficiently differentiable in z, s
- $g(z, 0) = z$

Definition of the $O(s^r)$-Resolution ODE

Definition: The $O(s^r)$-Resolution ODE of a DTA

We say an ODE system with the following normal form

$$\dot{Z} = f^{(r)}(Z, s) := f_0(Z) + sf_1(Z) + \cdots + s^r f_r(Z)$$

the $O(s^r)$-resolution ODE of the discrete-time algorithm with iterate update $z^+ = g(z, s)$ if it satisfies that for any z that

$$\|Z(s) - z^+\| = o(s^{r+1}) \qquad (\text{ or } O(s^{r+2}))\ , \qquad (*)$$

where $Z(s)$ is the solution obtained at $t = s$ following the above ODE with initial solution $Z(0) = z$.

- There can be multiple ODEs satisfying $(*)$, but the one of the normal form is unique.

Definition of $O(s^r)$-Resolution ODE, continued

Definition: $O(s^r)$-Resolution ODE

We say an ODE system with the following normal form

$$\dot{Z} = f^{(r)}(Z, s) := f_0(Z) + sf_1(Z) + \cdots + s^r f_r(Z)$$

the $O(s^r)$-resolution ODE of the discrete-time algorithm with iterate update $z^+ = g(z, s)$ if it satisfies that for any z that

$$\|Z(s) - z^+\| = o(s^{r+1}) \quad (\text{ or } O(s^{r+2})) , \quad (*)$$

where $Z(s)$ is the solution obtained at $t = s$ following the above ODE with initial solution $Z(0) = z$.

- There can be multiple ODEs satisfying $(*)$, but the one of the normal form is unique.

Definition of $O(s^r)$-Resolution ODE, continued

Definition: $O(s^r)$-Resolution ODE

We say an ODE system with the following normal form

$$\dot{Z} = f^{(r)}(Z, s) := f_0(Z) + sf_1(Z) + \cdots + s^r f_r(Z)$$

the $O(s^r)$-resolution ODE of the discrete-time algorithm with iterate update $z^+ = g(z, s)$ if it satisfies that for any z that

$$\|Z(s) - z^+\| = o(s^{r+1}) \quad (\text{ or } O(s^{r+2})) \,, \quad (*)$$

where $Z(s)$ is the solution obtained at $t = s$ following the above ODE with initial solution $Z(0) = z$.

- There can be multiple ODEs satisfying $(*)$, but the one of the normal form is unique.

How to Obtain the $O(s^r)$-Resolution ODE?

Theorem: Obtaining the $O(s^r)$-resolution ODE from $g(z,s)$

Consider a discrete-time algorithm with iterate update $z_+ = g(z,s)$, where $g(z,0) = z$ and $g(z,s)$ is $(r+1)$-th order differentiable over s for any z. Then the i-th coefficient function in the $O(s^r)$-resolution ODE can be obtained recursively by

$$f_i(Z) = g_{i+1}(Z) - \sum_{l=2}^{i+1} \frac{1}{l!} h_{l,i+1-l}(Z)\ , \ \text{for } i = 0, 1, \ldots, r,$$

where $g_i(z)$ is the i-th Taylor's expansion of $g(z,s)$:

$$g(z,s) = \sum_{j=0}^{r+1} g_j(z) s^j + o(s^{r+1})$$

$h_{l,i+1-l}(Z)$ is a function of $f_0(Z), \ldots, f_{i-1}(Z)$ defined as the coefficient function of s^i in the expansion of $\frac{d^j}{dt^j} Z$:

$$\frac{d^j}{dt^j} Z = \sum_{i=0}^{r+1} h_{j,i}(Z) s^i + o(s^{r+1}).$$

How to Obtain the $O(s^r)$-Resolution ODE?

Theorem: Obtaining the $O(s^r)$-resolution ODE from $g(z, s)$

Consider a discrete-time algorithm with iterate update $z_+ = g(z, s)$, where $g(z, 0) = z$ and $g(z, s)$ is $(r + 1)$-th order differentiable over s for any z. Then the i-th coefficient function in the $O(s^r)$-resolution ODE can be obtained recursively by

$$f_i(Z) = g_{i+1}(Z) - \sum_{l=2}^{i+1} \frac{1}{l!} h_{l,i+1-l}(Z) \ , \ \text{for } i = 0, 1, \dots, r,$$

where $g_i(z)$ is the i-th Taylor's expansion of $g(z, s)$:

$$g(z, s) = \sum_{i=0}^{r+1} g_i(z) s^i + o(s^{r+1})$$

$h_{l,i+1-l}(Z)$ is a function of $f_0(Z), \dots, f_{i-1}(Z)$ defined as the coefficient function of s^i in the expansion of $\frac{d^j}{dt^j} Z$:

$$\frac{d^j}{dt^j} Z = \sum_{i=0}^{r+1} h_{j,i}(Z) s^i + o(s^{r+1}).$$

How to Obtain the $O(s^r)$-Resolution ODE?

Theorem: Obtaining the $O(s^r)$-resolution ODE from $g(z, s)$

Consider a discrete-time algorithm with iterate update $z_+ = g(z, s)$, where $g(z, 0) = z$ and $g(z, s)$ is $(r + 1)$-th order differentiable over s for any z. Then the i-th coefficient function in the $O(s^r)$-resolution ODE can be obtained recursively by

$$f_i(Z) = g_{i+1}(Z) - \sum_{l=2}^{i+1} \frac{1}{l!} h_{l,i+1-l}(Z) \text{ , for } i = 0, 1, \dots, r,$$

where $g_i(z)$ is the i-th Taylor's expansion of $g(z, s)$:

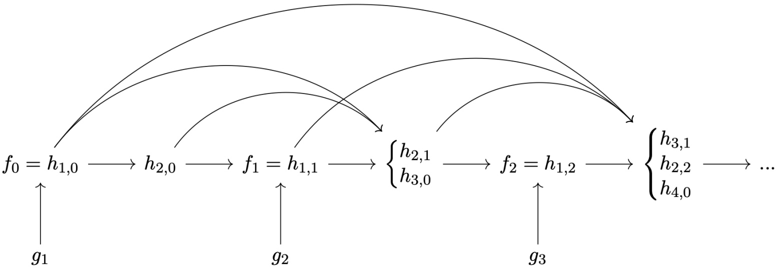
$$g(z, s) = \sum_{j=0}^{r+1} g_j(z) s^j + o(s^{r+1})$$

$h_{l,i+1-l}(Z)$ is a function of $f_0(Z), \dots, f_{i-1}(Z)$ defined as the coefficient function of s^i in the expansion of $\frac{d^j}{dt^j} Z$:

$$\frac{d^j}{dt^j} Z = \sum_{i=0}^{r+1} h_{j,i}(Z) s^i + o(s^{r+1}).$$

The Logic Flow of Computing the $O(s^r)$-Resolution ODEs

Given $f_0, g_1, g_2, g_3, \ldots$



- $O(s^r)$-resolution ODE gives the first r terms of the $O(s^{r+1})$-resolution ODE
- How to determine r? Try it out!

Going Back to Minimax Problems

Corollary: $O(1)$-resolution and $O(s)$-resolution ODE of GM, PPM and EGM

(i) The $O(1)$-resolution ODEs of GM, PPM and EGM are the same, that is, GF:

$$\dot{Z} = -F(Z) \ .$$

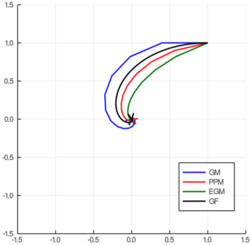
(ii) The $O(s)$-resolution ODE of GM is

$$\dot{Z} = -F(Z) - \frac{s}{2}\nabla F(Z)F(Z) \ .$$

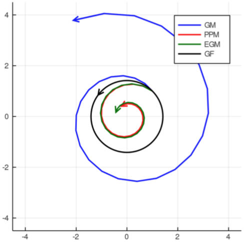
(iii) The $O(s)$-resolution ODEs of PPM and of EGM are the same:

$$\dot{Z} = -F(Z) + \frac{s}{2}\nabla F(Z)F(Z) \ .$$

Behaviors of Different Algorithms



(a) The trajectories of GM, PPM, EGM and GF for solving $\min_x \max_y \frac{1}{2}x^2 + xy - \frac{1}{2}y^2$ with step-size $s = 0.3$ and initial solution $(1, 1)$.



(b) The trajectories of GM, PPM, EGM and GF for solving $\min_x \max_y xy$ with step-size $s = 0.3$ and initial solution $(1, 1)$.

Toy Example $L(x, y) = xy$ with $z^* = (0, 0)$

- GF circles:

$$\langle \dot{Z}, Z \rangle = Z^T \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} Z = 0$$

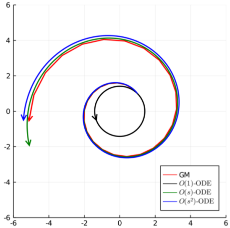
- GM diverges:

$$\dot{Z} = -F(Z) + \frac{s}{2} Z$$

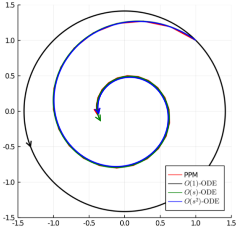
- PPM and EGM converges:

$$\dot{Z} = -F(Z) - \frac{s}{2} Z$$

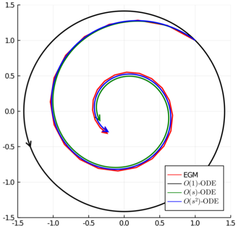
Toy Example $L(x, y) = xy$ with $z^* = (0, 0)$, continued



(a) The trajectories of GM and its corresponding ODEs.



(b) The trajectories of PPM and its corresponding ODEs.



(c) The trajectories of EGM and its corresponding ODEs.

- The higher the order of resolution, the closer the trajectoris between the DTA and the ODE
- PPM and EGM are different in their $O(s^2)$ terms

Extensions to Bilinear Minimax Problem $L(x, y) = x^T By$

- The $O(s)$-resolution ODE of PPM and EGM is a linear ODE

$$\dot{Z} = \begin{bmatrix} -\frac{s}{2}BB^T & -B \\ B^T & -\frac{s}{2}B^T B \end{bmatrix} Z$$

- After changing basis, it leads to independent evolving 2-d ODE with close form solution:

$$\hat{x}_i(t) = c_i e^{-\frac{s}{2}\lambda_i^2 t} \cos(\lambda_i t + \delta_i)$$
$$\hat{y}_i(t) = c_i e^{-\frac{s}{2}\lambda_i^2 t} \sin(\lambda_i t + \delta_i)$$

- Explains the Linear convergence rate of PPM and EGM

- Similarly, the $O(s)$-resolution ODE of GM diverges linearly
- PPM/EGM is superior to GM for solving minimax problems

Step 2: $O(s^r)$ Linear Convergence Condition

Step 2: Analyze the $O(s^r)$-Resolution ODE:

The $O(s^r)$ Linear Convergence Condition

The Standard Steps to Show Linear Convergence

The standard steps to show linear convergence of a dynamic:

- Identify an energy function E such that $\mathrm{E}(z^*) = 0$ and $E(z) \geq 0$
- Continuous-time dynamic:

$$\frac{d}{dt}\mathrm{E}(Z) \leq -\rho(s)\mathrm{E}(Z)$$

- Discrete-time algorithm:

$$\mathrm{E}(z^{k+1}) \leq (1 - s\rho(s))\mathrm{E}(z^k)$$

The $O(s^r)$ Linear Convergence Condition of a DTA

Definition: $O(s^r)$ Linear Convergence Condition of a Discrete-Time Algorithm w.r.t. an Energy Function E

We say a condition the $O(s^r)$ linear convergence condition of a discrete-time algorithm w.r.t. an energy function E following the dynamic of its $O(s^r)$-resolution ODE decays linearly:

$$\frac{d}{dt}\mathrm{E}(Z) \leq -\rho(s)\mathrm{E}(Z) \ .$$

- $\rho(s)$ is usually lower-bounded by a r-th order polynomial of s

Linear Convergence Condition of PPM, EGM and GM

We choose the energy function $E(z) = \frac{1}{2}\|F(z)\|^2$

- $E(z) = 0$ iff z is an optimal minimax solution

Let us introduce new notations:

$$A = \nabla_{xx}L(x, y), B = \nabla_{xy}L(x, y), C = -\nabla_{yy}L(x, y)$$

$O(1)$ Linear Convergence Condition of PPM, EGM and GM

Proposition: $O(1)$ linear convergence condition

The $O(1)$ linear convergence condition of PPM, EGM and GM w.r.t. $E(z)$ is

$$F(Z)^T \begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix} F(Z) \geq \frac{1}{2}\rho \|F(Z)\|^2 ,$$

and a sufficient condition is strongly convex-strongly concave:

$$A \succ 0, C \succ 0.$$

Proof. Recall that the $O(1)$-resolution ODE of PPM, EGM and GM is $\dot{Z} = -F(Z)$. Thus

$$\frac{d}{dt}\frac{1}{2}\|F(Z)\|^2 = F(Z)^T \nabla F(Z)\dot{Z} = -F(Z)^T \nabla F(Z) F(Z) = -F(Z)^T \begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix} F(Z) .$$

$O(s)$ Linear Convergence Condition of PPM and EGM

Proposition: $O(s)$ linear convergence condition

The $O(s)$ linear convergence condition of PPM and EGM w.r.t. $E(z)$ is

$$F(Z)^T \begin{bmatrix} A - \frac{s}{2}A^2 + \frac{s}{2}BB^T & 0 \\ 0 & C - \frac{s}{2}C^2 + \frac{s}{2}B^TB \end{bmatrix} F(Z) \geq \rho(s)\|F(Z)\|^2 \ ,$$

and a sufficient condition with $s \leq \frac{1}{\gamma}$ is

$$A + sBB^T \succ 0, C + sB^TB \succ 0.$$

Proof. Recall that the $O(s)$-resolution ODE of PPM and EGM is $\dot{Z} = -F(Z) + \frac{s}{2}\nabla F(Z)F(Z)$. Thus

$$\begin{aligned} \frac{d}{dt}\frac{1}{2}\|F(Z)\|^2 &= -F(Z)^T\nabla F(Z)F(Z) + \frac{s}{2}F(Z)^T(\nabla F(Z))^2F(Z) \\ &= -F(Z)^T \begin{bmatrix} A - \frac{s}{2}A^2 + \frac{s}{2}BB^T & 0 \\ 0 & C - \frac{s}{2}C^2 + \frac{s}{2}B^TB \end{bmatrix} F(Z) \ . \end{aligned}$$

$O(s)$ linear convergence condition of PPM and EGM, continued

Proposition: $O(s)$ linear convergence condition

The $O(s)$ linear convergence condition of PPM and EGM w.r.t. $E(z)$ is

$$F(Z)^T \begin{bmatrix} A - \frac{s}{2}A^2 + \frac{s}{2}BB^T & 0 \\ 0 & C - \frac{s}{2}C^2 + \frac{s}{2}B^T B \end{bmatrix} F(Z) \geq \rho(s)\|F(Z)\|^2 ,$$

and a sufficient condition with $s \leq \frac{1}{\gamma}$ is

$$A + sBB^T \succ 0, C + sB^T B \succ 0.$$

- This unifies the two conditions PPM/EGM has linear convergence
- More cases when PPM/EGM has linear convergence:
  - $L(x, y) = f(x) + x^T By - g(y)$ with strongly convex f and full column rank B
  - $L(x, y) = f(C_1 x) + x^T By - g(C_2 y)$ with strongly convex f and g
  - $L(x, y)$ is nonconvex-nonconcave with large enough interaction terms

Step 3: Extend the Convergent Results of ODEs to DTAs

Step 3: Extend the Convergent Results of ODEs back to DTAs

Fundamental Questions to Answer

Questions:

- What are the connections between the convergence of a DTA and the convergence of its $O(s^r)$-resolution ODEs?
- How to choose the energy function?

Our answer (informal):
With a "proper" energy function, if the $O(s^r)$-resolution ODE converges linearly to an optimal solution, then the DTA converges linearly to an optimal solution.

Proper Energy Function

Recall by definition of the $O(s^r)$-resolution ODE that:

- $\|Z(s) - z^+\| \le O(s^{r+2})$.

Definition: Proper Energy Function

We say an energy function $E(z) = \frac{1}{2}e(z)^2$ with $e(z) \ge 0$ is proper for studying the $O(s^r)$-resolution ODE of a DTA $z^+ = g(z,s)$ if there exists a and c such that it holds for any $z \in \{e(z) \le \delta\}$ that

$$\|Z(s) - z^+\| \le cs^{r+2}e(z).$$

How to Check Whether an Energy Function is Proper?

Recall that

$$g(z,s)=\sum_{j=0}^{r+1} g_j(z)s^j + o(s^{r+1})$$

Theorem: Sufficient Conditions for Proper Energy Functions

Suppose $g_j(z)$ is $(2r+3-j)$-th order differentiable over z, and it holds for any $z \in \{e(z)\le\delta\}$ that

$$\|g_j(z)\| \le O(e(z)) \text{ and } \|\nabla^k g_j(z)\| \le O(1)$$

for $j=1,\ldots,r+2$ and $k=1,\ldots,2r+3-j$. Then the energy function $E(z)=\frac{1}{2}e(z)^2$ is proper.

Some typical examples of $e(z)$:

- $e(z)=\|F(z)\|$, $e(z)=\|z-z^*\|$
- $e(z)=\sqrt{f(z)-f^*}$ for convex optimization

$\frac{1}{2}\|F(z)\|^2$ is a proper energy function for GM, PPM and EGM

Connections between DTAs and ODEs

Theorem: Connections between DTAs and ODEs

Consider a DTA and its $O(s^r)$-resolution ODE with a proper energy function $E(z)$. Suppose the $O(s^r)$-linear-convergence condition is satisfied, i.e.,

$$\frac{d}{dt}\mathrm{E}(Z) \leq -\rho(s)\mathrm{E}(Z) \ ,$$

and it holds for any $z \in \{e(z) \leq \delta\}$ that $\|\nabla e(z)\| \leq \gamma$. If the step-size s satisfies $\gamma c s^{r+2} \leq \min\left(1, \frac{s\rho(s)}{16}\right)$, it holds for any $k \geq 0$ that

$$E(z^k) \leq \left(1 - \frac{s\rho(s)}{4}\right)^k E(z^0) \ .$$

- $\rho(s) \geq O(s^r)$, thus there exists s^* such that the step-size condition holds when $s \leq s^*$

Connections between DTAs and ODEs

Theorem: Connections between DTAs and ODEs

Consider a DTA and its $O(s^r)$-resolution ODE with a proper energy function $E(z)$. Suppose the $O(s^r)$-linear-convergence condition is satisfied, i.e.,

$$\frac{d}{dt}\mathrm{E}(Z) \leq -\rho(s)\mathrm{E}(Z) \ ,$$

and it holds for any $z \in \{e(z) \leq \delta\}$ that $\|\nabla e(z)\| \leq \gamma$. If the step-size s satisfies $\gamma c s^{r+2} \leq \min\left(1, \frac{s\rho(s)}{16}\right)$, it holds for any $k \geq 0$ that

$$E(z^k) \leq \left(1 - \frac{s\rho(s)}{4}\right)^k E(z^0) \ .$$

- $\rho(s) \geq O(s^r)$, thus there exists s^* such that the step-size condition holds when $s \leq s^*$

Connections between DTAs and ODEs

Theorem: Connections between DTAs and ODEs

Consider a DTA and its $O(s^r)$-resolution ODE with a proper energy function $E(z)$. Suppose the $O(s^r)$-linear-convergence condition is satisfied, i.e.,

$$\frac{d}{dt}\mathrm{E}(Z) \leq -\rho(s)\mathrm{E}(Z) ,$$

and it holds for any $z \in \{e(z) \leq \delta\}$ that $\|\nabla e(z)\| \leq \gamma$. If the step-size s satisfies $\gamma c s^{r+2} \leq \min\left(1, \frac{s\rho(s)}{16}\right)$, it holds for any $k \geq 0$ that

$$E(z^k) \leq \left(1 - \frac{s\rho(s)}{4}\right)^k E(z^0) .$$

- $\rho(s) \geq O(s^r)$, thus there exists s^* such that the step-size condition holds when $s \leq s^*$

Applications: Nonconvex-Nonconcave Minimax Problems

Applications:

Nonconvex-Nonconcave Minimax Problems

Nonconvex-Nonconcave Minimax Problems

The problem of interest is

$$\min_x \max_y L(x, y) \ ,$$

where $L(x, y)$ may not be convex in x nor concave in y.

Many applications:

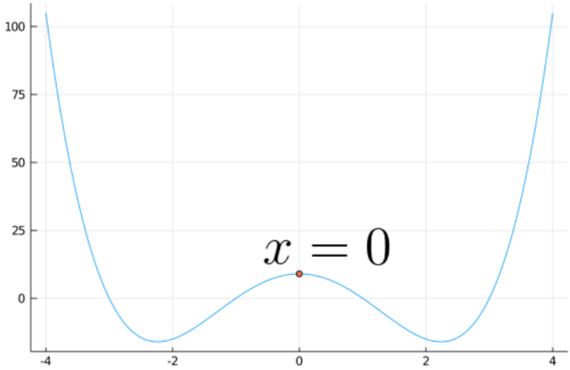
- Generative Adversarial Nets (GANs)
- Robust Neural Networks

A Simple 2-d Problem

Consider simple 2-d nonconvex-nonconcave problem with bilinear interaction term:

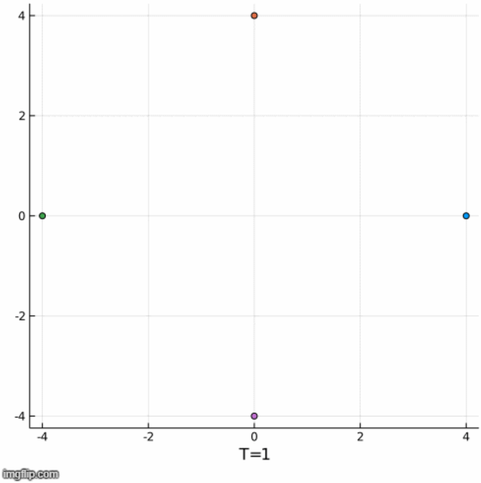
$$\min_x \max_y L(x, y) = f(x) + xAy - g(y),$$

where $f(x) = g(x) = (x-3)(x-1)(x+1)(x+3)$.



Why are Nonconvex-Nonconcave Problems Hard?

Cycling is a fundamental part of nonconvex-nonconcave problems (trajectory of PPM):

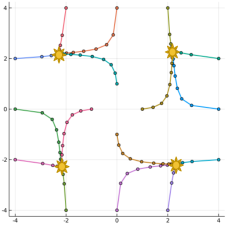


The Landscape of Nonconvex-Nonconcave Problems

Consider simple 2-d nonconvex-nonconcave problem with bilinear interaction term:

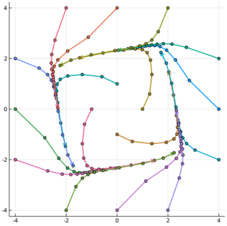
$$\min_x \max_y L(x, y) = f(x) + x^T Ay - g(y),$$

where $f(x) = g(x) = (x-3)(x-1)(x+1)(x+3)$.



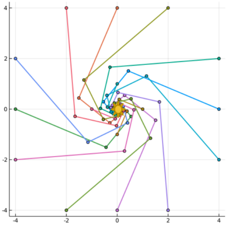
A=1

Linear Convergence to a Local Solution if **Interaction Weak**



A=10

Cycling is possible if **Interaction Moderate**



A=50

Linear Convergence to a Global Solution if **Interaction Dominate**

The Landscape of Nonconvex-Nonconcave Problems

The above structure extends to every nonconvex-nonconcave bilinear problem:

$$L(x, y) = f(x) + x^T Ay - g(y)$$

- A Large Enough: PPM has **global linear convergence** to a stationary point
- A Middle Size: PPM may **cycle** indefinitely
- A Small Enough: PPM has **local linear convergence** to a stationary point with a good initialization

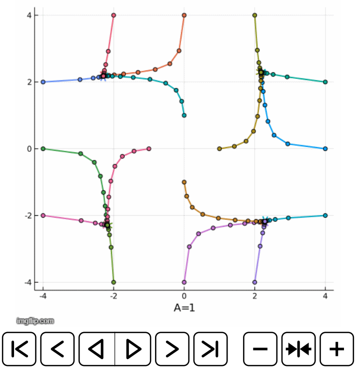
Recall the $O(s)$-linear-convergence condition for bilinear nonconvex-nonconcave problem:

$$\nabla^2 f(x) + sAA^T \succeq \rho(s)I, \nabla^2 g(x) + sA^T A \succeq \rho(s)I$$

- The first case globally satisfies the above condition; The third case locally satisfies this condition.

The Landscape of Nonconvex-Nonconcave Problems

A more smoothed phase shift:



- The phase transition can be characterized by Hopf Bifurcation of the $O(s)$-resolution ODE

Summary

$O(s^r)$-**Resolution ODE framework**

- First Step — Obtain ODEs from a DTA:
  - The $O(s^r)$-resolution ODEs, and how to obtain them
  - Examples for PPM, EGM and GM
- Second Step — Analyze the $O(s^r)$-resolution ODEs:
  - $O(s^r)$-linear convergence condition
- Third Step — Going back to DTAs:
  - Proper energy function
  - The connection between the ODEs and the DTAs
  - How to check whether an energy function is proper
- Application — Nonconvex-Nonconcave Minimax Problems

Thank you!