The Limit Points of (Optimistic) Gradient Descent in Min-Max Optimization

Constantinos Daskalakis[1] Ioannis Panageas[2]

[1] MIT, [2] Singapore University of Technology and Design

Q: How does vanilla Gradient Descent/Ascent behave in min-max problems?

Min-max problem and GANs

Motivated by GANs:

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}). \quad (1)$$

- $\mathbf{x}$ parameters of the generator deep neural net,
- $\mathbf{y}$ parameters of the discriminator neural net,
- $f(\mathbf{x}, \mathbf{y})$ some measure of how close the distribution generated by the generator appears to the true distribution from the perspective of the discriminator.

Gradient Descent/Ascent (GDA)

Most *natural* approach to solve (1) is by doing gradient descent on $\mathbf{x}$ and gradient ascent on $\mathbf{y}$, i.e.,

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - \alpha \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \\ \mathbf{y}_{t+1} &= \mathbf{y}_t + \alpha \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \end{aligned} \quad (2)$$

where $\alpha > 0$ is a small constant (stepsize).

Optimistic Gradient Descent/Ascent (OGDA)

The update rule of OGDA (“negative momentum”) is:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - 2\alpha \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) + \alpha \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), \\ \mathbf{y}_{t+1} &= \mathbf{y}_t + 2\alpha \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) - \alpha \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}). \end{aligned} \quad (3)$$

where $\alpha > 0$ is a small constant (stepsize).

Local min-max (Local saddles)

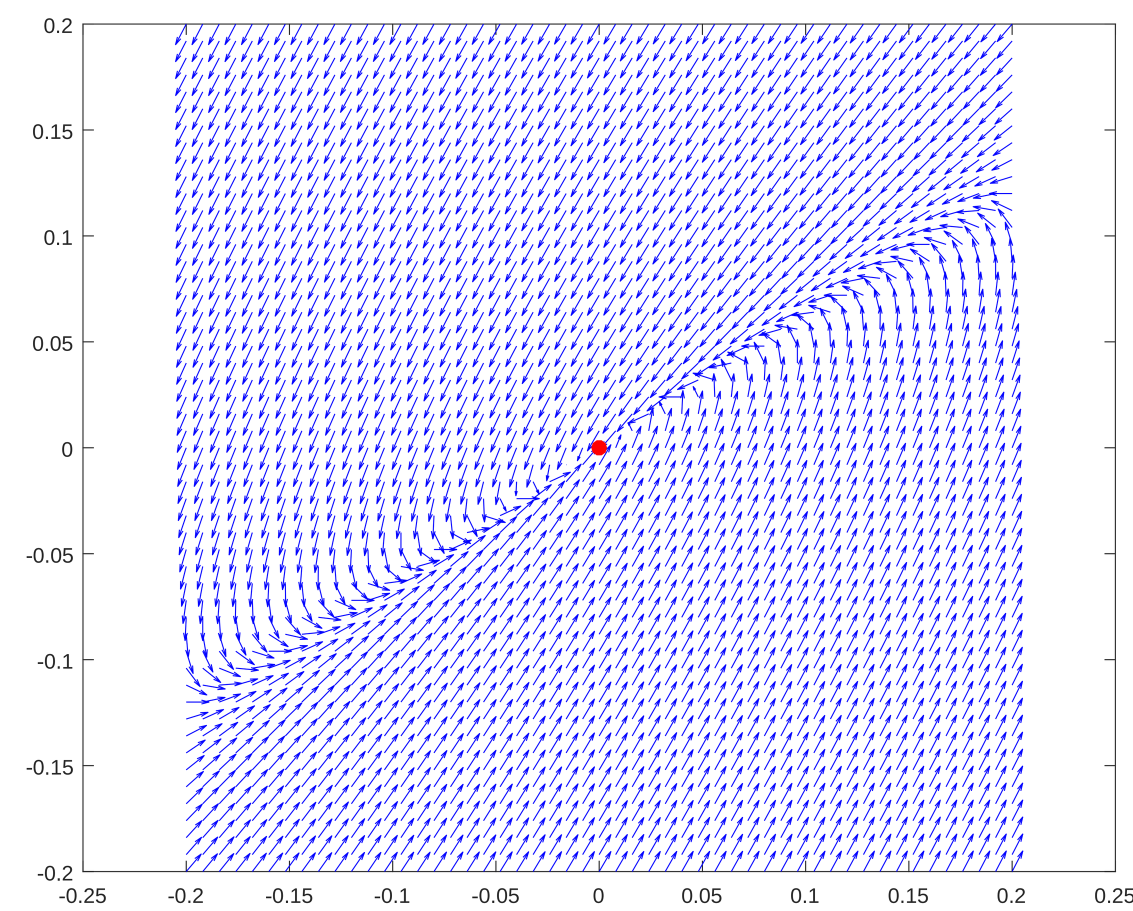
- Critical point if $\nabla f(\mathbf{x}^*, \mathbf{y}^*) = \mathbf{0}$.
- **Local min-max point or local saddle** if there exists a neighborhood U around critical point $(\mathbf{x}^*, \mathbf{y}^*)$ so that for all $(\mathbf{x}, \mathbf{y}) \in U$ we have that $f(\mathbf{x}^*, \mathbf{y}) \le f(\mathbf{x}^*, \mathbf{y}^*) \le f(\mathbf{x}, \mathbf{y}^*)$.
- Strongly local min-max point if $\lambda_{\min}(\nabla^2_{\mathbf{xx}} f(\mathbf{x}^*, \mathbf{y}^*)) > 0$ and $\lambda_{\max}(\nabla^2_{\mathbf{yy}} f(\mathbf{x}^*, \mathbf{y}^*)) < 0$.
- Linearly stable or just stable if, for the Jacobian J of the update rule of the dynamics (GDA or OGDA) computed at $(\mathbf{x}^*, \mathbf{y}^*)$ holds that its spectral radius $\rho(J) < 1$.
- **Remark:** There are other notions of stability, this is the most convenient for us.

Q: Are the stable limit points of GDA and OGDA locally min-max solutions?

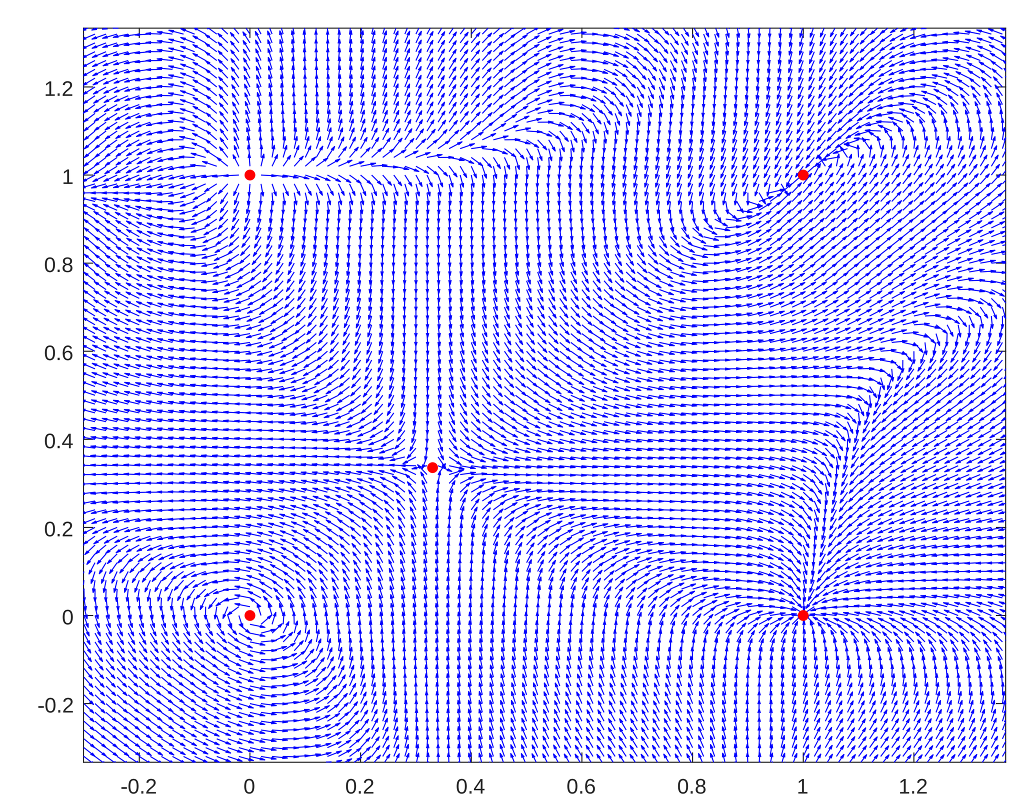
Stable under GDA, Not local min-max

We plot the "vector field" of GDA for the quadratic function:

$$f(x, y) = -\frac{1}{8}x^2 - \frac{1}{2}y^2 + \frac{6}{10}xy.$$



An example with all possible scenarios



Critical point	GDA-stable	OGDA-stable	Local min-max
$(0,0)$	NO	YES	NO
$(0,1)$	NO	NO	NO
$(1,0)$	YES	YES	YES
$(1,1)$	YES	YES	NO
$(0.33, 0.336)$	NO	NO	NO

Table: Summary of critical points.

Q: How do the stable limit points of GDA and OGDA relate to each other?

Inclusion results

Local min-max $\subset$ GDA-stable $\subset$ OGDA-stable

Theorem

Assume f is twice differentiable and ∇f is Lipschitz with constant L.

1. *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be a local min max critical point that satisfies some genericity assumption[a] [b]. For $\alpha > 0$ sufficiently small it holds that $(\mathbf{x}^*, \mathbf{y}^*)$ is GDA-stable. Moreover, there is a function with a critical point $(\mathbf{x}^*, \mathbf{y}^*)$ which is not local min-max but it is GDA-stable (for sufficiently small $\alpha > 0$).*
2. *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be a GDA-stable. For $0 < \alpha < \frac{1}{2L}$ it holds that $(\mathbf{x}^*, \mathbf{y}^*)$ is OGDA-stable. Similarly the inclusion is strict.*

[a] The hessian with opposite the x-rows has no eigenvalue with real part 0.
[b] No assumption needed if strong local min-max.

Diffeomorphism and avoiding unstable points

Theorem

Let f be twice differentiable and ∇f is Lipschitz with constant L. Assume that $0 < \alpha < \frac{1}{L}$.

- *The update rule of the GDA dynamics is a local diffeomorphism[a].*
- *If $\nabla^2 f$ is invertible, then the update rule of OGDA is a local diffeomorphism.*

Corollary

Using the center-stable manifold theorem it can be shown that the set of initial conditions so that GDA (respectively OGDA) converges to unstable points is of measure zero.

[a] A local diffeomorphism is a function that locally is invertible, smooth and its (local) inverse is also smooth.

Take away messages and Questions

- Optimism makes critical points "more stable".
- GDA, OGDA do not converge ONLY to local saddles (robust solutions). Another method?
- What about vanishing stepsizes?
- What is a good solution concept?
- `https://arxiv.org/abs/1807.03907`