Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions

Ioannis Panageas
MIT-SUTD
Previously Georgia Tech

joint work with **Georgios Piliouras** (SUTD)

Outline

Basics

Results and Proof sketch

Examples

Previous and Future work

Definitions

Problem

Let $f : \mathbb{R}^N \to \mathbb{R}$ and f is C^2:

$$\min_{\mathbf{x} \in \mathbb{R}^N} f(\mathbf{x}).$$

Definitions

Problem

Let $f : \mathbb{R}^N \to \mathbb{R}$ and f is C^2:

$$\min_{\mathbf{x} \in \mathbb{R}^N} f(\mathbf{x}).$$

Typical way; Gradient Descent (GD)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k), \tag{1}$$

with constant $\alpha > 0$. A discrete dynamical system $\mathbf{x}_{k+1} = g(\mathbf{x}_k)$.

Definitions

Problem

Let $f : \mathbb{R}^N \to \mathbb{R}$ and f is C^2:

$$\min_{\mathbf{x} \in \mathbb{R}^N} f(\mathbf{x}).$$

Typical way; Gradient Descent (GD)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k), \tag{1}$$

with constant $\alpha > 0$. A discrete dynamical system $\mathbf{x}_{k+1} = g(\mathbf{x}_k)$.

Question: Great but any guarantees?

Definitions

Problem

Let $f : \mathbb{R}^N \to \mathbb{R}$ and f is C^2:

$$\min_{\mathbf{x} \in \mathbb{R}^N} f(\mathbf{x}).$$

Typical way; Gradient Descent (GD)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k), \tag{1}$$

with constant $\alpha > 0$. A discrete dynamical system $\mathbf{x}_{k+1} = g(\mathbf{x}_k)$.

Question: Great but any guarantees?

- Answer: If ∇f is L-Lipschitz and $\alpha \leq \frac{1}{L}$ then GD converges to fixed points.

Definitions

Problem

Let $f : \mathbb{R}^N \to \mathbb{R}$ and f is C^2:

$$\min_{\mathbf{x} \in \mathbb{R}^N} f(\mathbf{x}).$$

Typical way; Gradient Descent (GD)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k), \tag{1}$$

with constant $\alpha > 0$. A discrete dynamical system $\mathbf{x}_{k+1} = g(\mathbf{x}_k)$.

Question: Great but any guarantees?

- Answer: If ∇f is L-Lipschitz and $\alpha \leq \frac{1}{L}$ then GD converges to fixed points.
- Folklore: $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|_2^2$.
- $\Rightarrow$ f is decreasing $\Rightarrow$ set-wise convergence (not point-wise!).

Definitions (cont.)

Question: What if f is non-convex?

Definitions (cont.)

Question: What if f is non-convex?

- Answer: The best we can hope for is convergence to local minimum!

Definitions (cont.)

Question: What if f is non-convex?

- Answer: The best we can hope for is convergence to local minimum!
- And we will have it... (under "mild" assumptions)

Important definitions

- $\mathbf{x}^*$ is a critical point of f if $\nabla f(\mathbf{x}^*) = \mathbf{0}$ (uncountably many!).
- $\mathbf{x}^*$ is isolated if there is a U around $\mathbf{x}^*$ and $\mathbf{x}^*$ is the only critical point in U.
- $\mathbf{x}^*$ is a saddle point if for all U around $\mathbf{x}^*$ there are $\mathbf{y}, \mathbf{z} \in U$ such that $f(\mathbf{z}) \leq f(\mathbf{x}^*) \leq f(\mathbf{y})$.
- $\mathbf{x}^*$ of f is a strict saddle if $\lambda_{\min}(\nabla^2 f(\mathbf{x}^*)) < 0$.
- Set $\mathcal{S}$ is called *forward or positively invariant* w.r.t $h : \mathcal{E} \to \mathbb{R}^N$ with $\mathcal{S} \subseteq \mathcal{E} \subseteq \mathbb{R}^N$ if $h(\mathcal{S}) \subseteq \mathcal{S}$.

Previous work and our results

Theorem (Lee, Simchowitz, Jordan, Recht 16')

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a C^2 function, ∇f is globally L-Lipschitz and $\mathbf{x}^*$ be a strict saddle. Assume that $0 < \alpha < \frac{1}{L}$, then

$$\Pr(\lim_k \mathbf{x}_k = \mathbf{x}^*) = 0.$$

If the strict saddle points are isolated, then GD converges to saddle points with probability zero.

Previous work and our results

Theorem (Lee, Simchowitz, Jordan, Recht 16')

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a C^2 function, ∇f is globally L-Lipschitz and $\mathbf{x}^*$ be a strict saddle. Assume that $0 < \alpha < \frac{1}{L}$, then

$$\Pr(\lim_k \mathbf{x}_k = \mathbf{x}^*) = 0.$$

If the strict saddle points are isolated, then GD converges to saddle points with probability zero.

Theorem (Main)

Let $f : \mathcal{S} \to \mathbb{R}$ be C^2 in an open convex set $\mathcal{S} \subseteq \mathbb{R}^N$ and $\sup_{\mathbf{x} \in \mathcal{S}} \left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L < \infty$. If $g(\mathcal{S}) \subseteq \mathcal{S}$ then the set of initial conditions $\mathbf{x} \in \mathcal{S}$ so that gradient descent with $0 < \alpha < 1/L$ converges to a strict saddle point is of (Lebesgue) measure zero, without the assumption that critical points are isolated.

Remarks and proof steps

Corollary

Assume furthermore that $\lim_k \mathbf{x}_k$ *exists and let* ν *be a prior measure (support* $\mathcal{S}$*) which is absolutely continuous w.r.t Lebesgue measure. Then with probability 1, GD converges to local minima.*

Remarks and proof steps

Corollary

Assume furthermore that $\lim_k \mathbf{x}_k$ *exists and let* ν *be a prior measure (support* $\mathcal{S}$*) which is absolutely continuous w.r.t Lebesgue measure. Then with probability 1, GD converges to local minima.*

Remarks

Lee et al. result is generalized in two ways:

- No global Lipschitz condition.
- Critical points do not have to be isolated.

Remarks and proof steps

Corollary

Assume furthermore that $\lim_k \mathbf{x}_k$ *exists and let* ν *be a prior measure (support* $\mathcal{S}$*) which is absolutely continuous w.r.t Lebesgue measure. Then with probability 1, GD converges to local minima.*

Remarks

Lee et al. result is generalized in two ways:

- No global Lipschitz condition.
- Critical points do not have to be isolated.

Proof steps

- 1. Convergence: Show that GD converges (already).

Remarks and proof steps

Corollary

Assume furthermore that $\lim_k \mathbf{x}_k$ *exists and let* ν *be a prior measure (support* $\mathcal{S}$*) which is absolutely continuous w.r.t Lebesgue measure. Then with probability 1, GD converges to local minima.*

Remarks

Lee et al. result is generalized in two ways:

- No global Lipschitz condition.
- Critical points do not have to be isolated.

Proof steps

- 1. Convergence: Show that GD converges (already).
- 2. Diffeomorphism: Prove that g is a diffeomorphism in $\mathcal{S}$ (eigenvalue analysis, show Jacobian is invertible).

Remarks and proof steps

Corollary

Assume furthermore that $\lim_k \mathbf{x}_k$ *exists and let* ν *be a prior measure (support* $\mathcal{S}$*) which is absolutely continuous w.r.t Lebesgue measure. Then with probability 1, GD converges to local minima.*

Remarks

Lee et al. result is generalized in two ways:

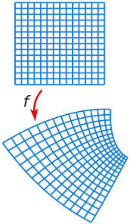
- No global Lipschitz condition.
- Critical points do not have to be isolated.

Proof steps

- 1. Convergence: Show that GD converges (already).
- 2. Diffeomorphism: Prove that g is a diffeomorphism in $\mathcal{S}$ (eigenvalue analysis, show Jacobian is invertible).
- 3. Measure zero: Use center-stable manifold along with Lindelof lemma.

Why are these technicalities important?

- Manifold: Topological space that "looks like" Euclidean space near each point.
- **Diffeomorphism**
  A diffeomorphism is a map between manifolds which is continuously differentiable and has a continuously differentiable inverse.



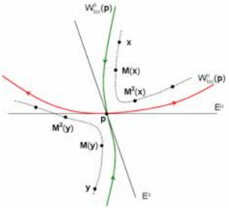
This is a useful technical smoothness condition that allows us to apply standard theorems about dynamical systems. (e.g., Center-Stable Manifold theorem).

Center-Stable Manifold theorem

Center-Stable Manifold theorem (informally)

If the rule of the dynamics is a diffeomorphism then

- For every fixed point $\mathbf{p}$, there exists an open ball $B_{\mathbf{p}}$ so that if trajectory $q(n)$ is inside $B_{\mathbf{p}}$ for all $n \geq 0$ then $p(0)$ belongs to a (local) center stable manifold $W_{sc}(\mathbf{p})$ which has dimension equal to the dimension of the space spanned by eigenvectors of the Jacobian (at $\mathbf{p}$) with eigenvalues of absolute value ≤ 1.



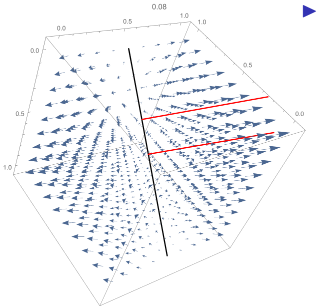
Step 3. Measure zero

Proof Sketch of Step 3

- Every strict saddle critical point $\mathbf{p}$ has a (local) center stable manifold $W_{sc}(\mathbf{p})$ of dimension lower than $N-1$), hence measure zero in $\mathbb{R}^{N-1}$
- Consider the union of all $B_{\mathbf{p}}$ and pick a countable subcover (Lindelof's lemma: every open cover in $\mathbb{R}^k$ has a countable subcover.)
- g^{-1} is C^1, maps null sets to null sets, the set of points that converge to some $B_{\mathbf{p}}$ is measure zero.
- Countable union of measure zero sets is measure zero.

Examples - Non-isolated critical points

$$f(x, y, z) = 2xy + 2xz - 2x - y - z,$$
$$\Rightarrow \nabla f(x, y, z) = (2y + 2z - 2, 2x - 1, 2x - 1).$$



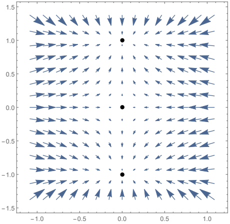
- Strict saddle points correspond to the line $(1/2, w, 1 - w)$ for $w \in \mathbb{R}$ (min eigenvalue is $-2\sqrt{2}$).

Therefore...

Set of initial conditions in $\mathcal{R}^3$ so that GD converges to black line has measure zero.

Examples (cont.) - Forward invariant set

$$f(x,y) = \frac{x^2}{2} + \frac{y^4}{4} - \frac{y^2}{2}, \text{Hessian } J = \begin{pmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{pmatrix}.$$



- f is not globally Lipschitz (Lee et al result does not apply!)
- Critical points are $(0,0),(0,1),(0,-1)$.
- For $\mathcal{S} = (-1,1) \times (-2,2)$, $\sup_{(x,y)\in\mathcal{S}} \left\|\nabla^2 f(x,y)\right\|_2 \leq 11$ (for $y = 2$ maximum).
- Choose $\alpha = \frac{1}{12} < \frac{1}{11}$, hence $g(x,y) = (\frac{11x}{12}, \frac{13y}{12} - \frac{y^3}{12}) \Rightarrow g(\mathcal{S}) \subseteq \mathcal{S}$

Therefore...

Set of initial conditions in $\mathcal{S}$ so that GD converges to $(0,0)$ has measure zero. Start at random, then GD converges to $(0,1),(0,-1)$ with probability 1.

Previous and Future work

- Vector flows perturbed by noise cannot converge to unstable fixed points [Pemantle 90'].
- Other dynamics? Results for replicator dynamics (evolution, game theory) [Mehta, P, Piliouras 15'].
- Mirror Descent (mirror map strongly convex). Ongoing work [Lee, P, Simchowitz, Jordan, Piliouras, Recht 16'].
- Non-negative matrix factorization (NMF)? Ongoing work [P, Piliouras, Tetali] analyzing Lee and Seung.
- Quantitative versions (stronger assumptions) [Ge, Huang, Jin, Yuan 15'].
- Many more...

Thank you!

Postdoc positions open!

Where: Singapore

On What: Game Theory, Algorithms, Dynamical Systems

georgios@sutd.edu.sg