First-order Methods Almost Always Avoid Saddle Points:
The Case of Vanishing Stepsizes

Ioannis Panageas[1] Georgios Piliouras[1] Xiao Wang[1]

[1] Singapore University of Technology and Design

Question: Do first-order methods avoid saddle points with vanishing stepsizes?

Motivation

- In many applications the stepsize of optimization algorithm is adaptive or vanishing.
- The choice of stepsize is really crucial. Changing the stepsize can change the convergence properties or the rate of convergence.
- In the paper Lee et al. it is proved that first-order methods avoid saddle points almost always with constant stepsize. The case of vanishing stepsize left as an open question.

Gradient Descent

- **Intuitive Example** Let $f(x) = \frac{1}{2}x^\top Ax$, $A = diag(\lambda_1, ..., \lambda_n)$, gradient descent has the form of

$$x_{k+1} = diag\left(\prod_{t=0}^{k}(1-\alpha_t\lambda_1), ..., \prod_{t=0}^{k}(1-\alpha_t\lambda_n)\right)x_0$$

For α_k being $\Omega\left(\frac{1}{k}\right)$, $\lim_{k\to\infty} x_k = 0$, the stable manifold is spanned by eigenvectors with positive eigenvalues so has measure 0.

- **General Case** If f is general C^2 function, the Taylor expansion of gradient descent at saddle x^* is

$$x_{k+1} = (I - \alpha_k\nabla^2 f(x^*))(x_k - x^*) + \eta(k, x_k)$$

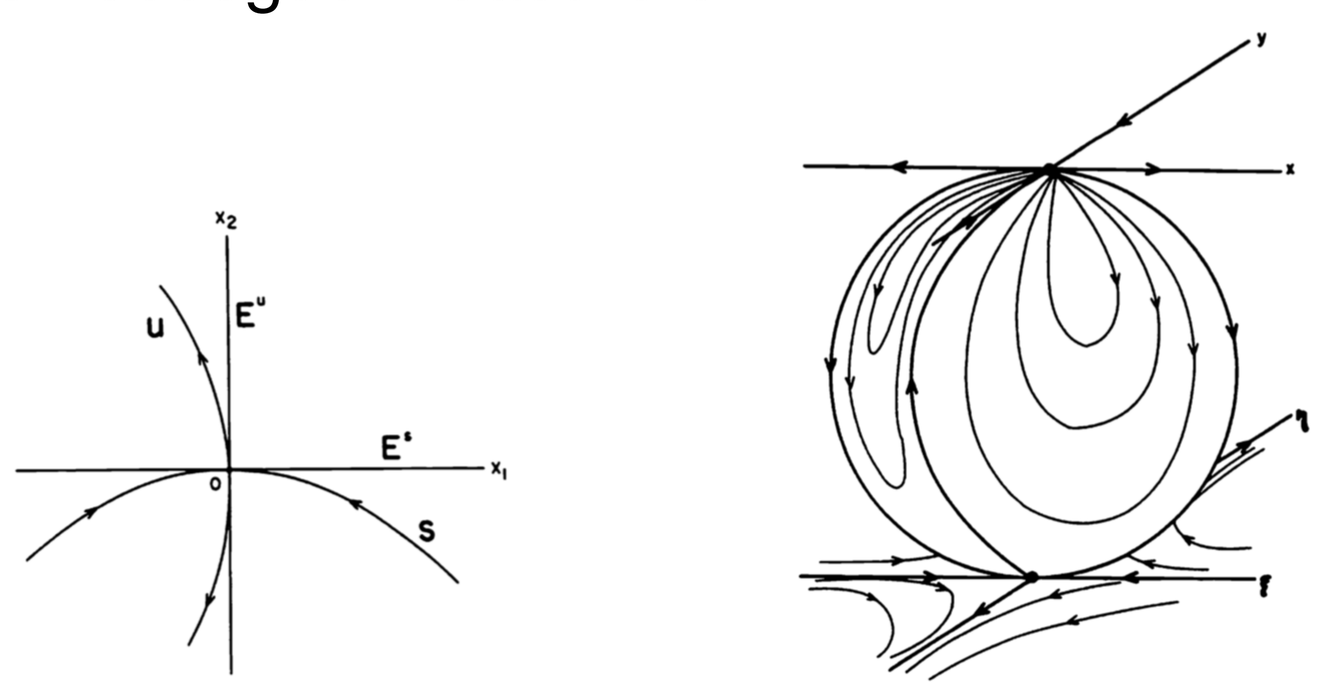
where $\eta(k, x^*) = x^*$ and $\eta(k, x)$ is of order $o(||x - x^*||)$ around x^*.
The stable manifold is the graph of certain differentiable function $\varphi : E^s \to E^u$, where E^s and E^u are the stable-unstable subspaces w.r.t the eigenvalues of $\nabla^2 f(x^*)$.

Stepsizes

- The stepsize cannot converge too fast, i.e. $\alpha_k \in \Omega\left(\frac{1}{k}\right)$. If $\alpha_k < \frac{1}{k}$, see Figure.
- In the contrast to the stochastic approximation, the condition $\sum_k \alpha_k^2 < \infty$ is **not necessary** in deterministic methods.

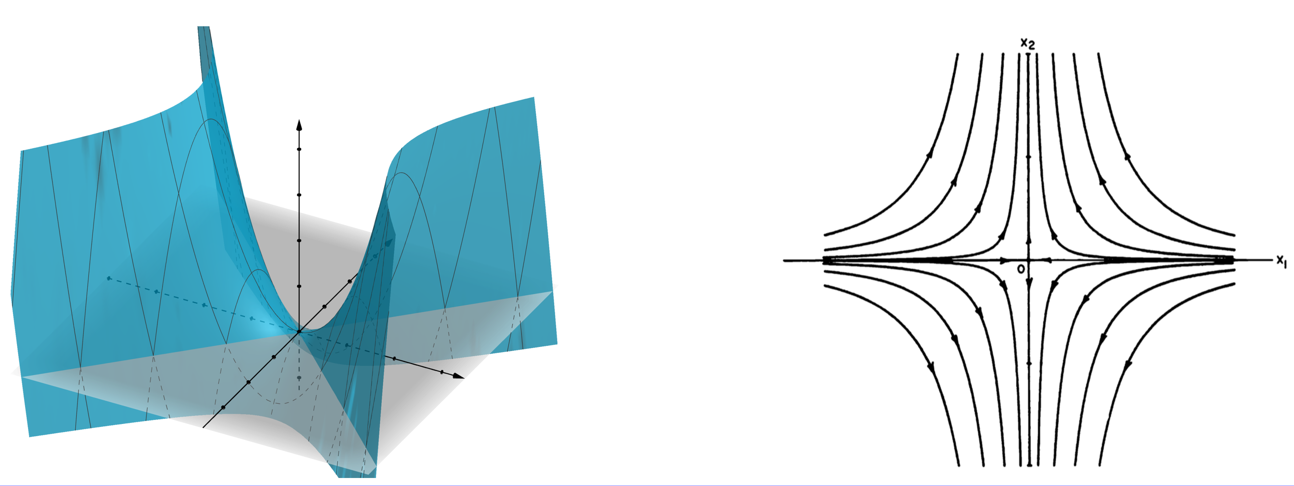
Stable Manifolds: GD and Manifold GD

If $f \in C^2$ is non-convex, the stable manifold of a saddle point has co-dimension at least 1, so has Lebesgue measure 0.



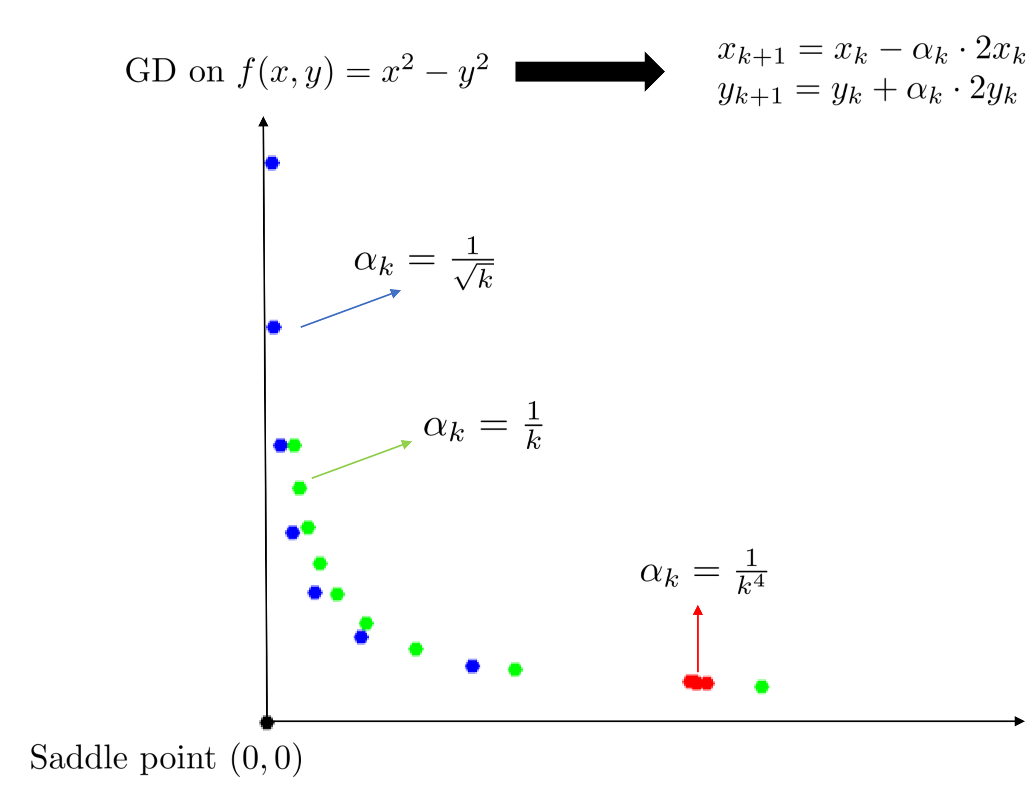
Example

Let $f(x_1, x_2) = \frac{1}{2}(x_1^2 - x_2^2)$, 0 is the saddle and x_1-axis is the stable manifold of Gradient Descent.



Stepsizes

- $\alpha_k = \frac{1}{k}$ and $\frac{1}{\sqrt{k}}$, GD converges to critical point and avoids saddle,
- $\alpha_k = \frac{1}{k^4}$, GD converges to a non-critical point.



Main Results

Theorem

Gradient Descent, Mirror Descent, Proximal Point and Manifold Gradient Descent with vanishing stepsize α_k of order $\Omega\left(\frac{1}{k}\right)$ avoid the set of strict saddle points (isolated and non-isolated) almost surely under random initialization.

Technical Overview

- **Lyapunov-Perron Method** The dynamical systems from variant first-order methods can be reduced to

$$x_{k+1} = A(k,0)x_0 + \sum_{i=0}^{k} A(k, i+1)\eta(i, x_i) \quad (1)$$

and the integral operator T written as $(Tx)_{k+1} =$

$$\begin{pmatrix} B(k,0)x_0^+ + \sum_{i=0}^{k} B(k,i+1)\eta^+(i,x_i) \\ -\sum_{i=0}^{\infty} C(k+1+i,k+1)^{-1}\eta^-(k+1+i,x_{k+1+i}) \end{pmatrix}. \quad (2)$$

has unique fixed point (a sequence) as the solution of (1) with initial condition x_0, where $B(m,n)$ and $C(m,n)$ are stable and unstable integral operators.

- **Banach Fixed Point Theorem** Let (X, d) be a complete metric space, then each contraction map $T : X \to X$ has unique fixed point.

The metric space X of sequences converging to 0 is complete. The operator T is a contraction map on X. Then from Banach Fixed Point Theorem, there exists unique function $\varphi : E^s \to E^u$ whose graph contains all initial conditions that converge to saddle point.

References

1. Lee et al. First-order methods almost always avoid saddle points, *Math. Programming* 2019.
2. Perko, Differential Equations and Dynamical Systems, 2001, Springer.

ARXIV

- `https://arxiv.org/abs/1906.07772`