Overview of Saddle Point Escaping Problem

Jingqi Zhu

The University of Manchester

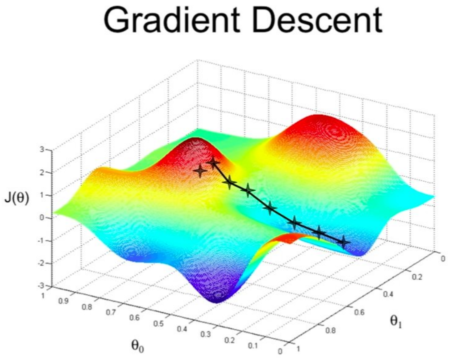
September 10, 2021

Overview

1. What's the problem about saddle points?
2. Why would people care about saddle points?
3. How to solve the problem?

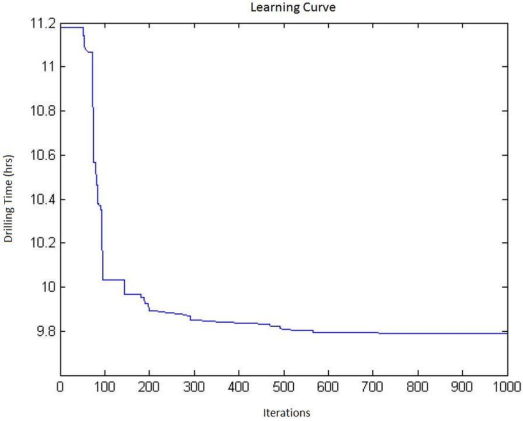
What's the problem about saddle points?

- Ultimate mission in optimization:
  $\underset{x}{minimize}\, f(x), s.t. x \in \mathcal{C}, where\ f : \mathbb{R}^d \to \mathbb{R}$
- GD: $x_{s+1} = x_s - \eta \nabla f(x_s)$ is guaranteed to converge to a stationary point. Local minima, local maxima, saddle point?



Why would people care about saddle points?

- No need to worry in convex models. i.e. linear or softmax regression.
- Unfortunately, most ML models are non-convex: MLP, CNN, RNN...
- More importantly, GD is more likely to converge to saddle points compared to local minima!
- As a result, saddle points flatten the learning curve!



How to solve the problem?

- Hessian matrix? Too expensive. Gradient-based!
- 1.Adding perturbations: By adding noise at each time, GD can escape saddle points in polynomial time, provided that the objective function satisfies strict saddle property. (*Escaping From Saddle Points – Online Stochastic Gradient for Tensor Decomposition*, Ge et al. 2015.)
- $O(d^4)$ still not good enough

How to solve the problem?

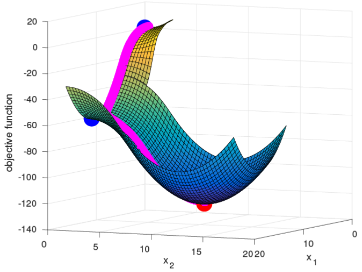
- 2.Random initialization: GD almost always escape saddle points asymptotically with random initialization. (*Gradient Descent Converges to Minimizers*, Lee et al. 2016 COLT)
- the Stable Manifold Theorem based on dynamical systems
- No bound of the iteration number.
- Even with random initialization, GD can take exponential time to escape saddle points. (*Gradient Descent Can Take Exponential Time to Escape Saddle Points*, Du et al. 2017 NIPS.)

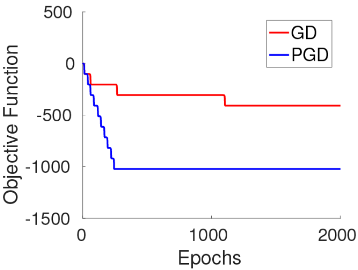
How to solve the problem?

- Before this paper, the runtime was at least polynomial function of dimension d. This is the first nearly dimension-free result: $O(log^4(d))$ (*How to Escape Saddle Points Efficiently*, Jin et al. 2017. ICML.)
- e.g.$d = 100,\ d^3 = 1 million,\ log^4(d) = 16$

Algorithm	**Iterations**	**Oracle**
Ge et al. (2015)	$O(\text{poly}(d/\epsilon))$	Gradient
Levy (2016)	$O(d^3\text{poly}(1/\epsilon))$	Gradient
This Work	$O(\log^4(d)/\epsilon^2)$	Gradient
Agarwal et al. (2016)	$O(\log(d)/\epsilon^{7/4})$	Hessian-vector
Carmon et al. (2016)	$O(\log(d)/\epsilon^{7/4})$	Hessian-vector
Carmon & Duchi (2016)	$O(\log(d)/\epsilon^2)$	Hessian-vector
Nesterov & Polyak (2006)	$O(1/\epsilon^{1.5})$	Hessian
Curtis et al. (2014)	$O(1/\epsilon^{1.5})$	Hessian

How to solve the problem?

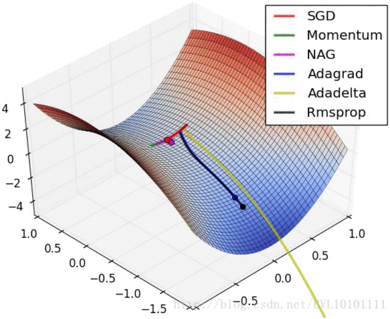




- Perturbed gradient descent can escape saddle points efficiently!
- Explained why gradient-based algorithms can work surprisingly well in actually non-convex optimization.
- Open problem: Will adding momentum improve the convergence rate to a second-order stationary point?
- More detailed estimate of the remainder of the Taylor expansion near saddle point. (*On Nonconvex Optimization for Machine Learning: Gradients, Stochasticity, and Saddle Points*, Jin et al. 2019 ICML.)

How to solve the problem

- Provide the first second-order convergence result for any adaptive method. Moreover, Adaptive methods escape saddle points faster than SGD. (*Escaping Saddle Points with Adaptive Gradient Methods*, Staib, 2020)
- Adaptive gradient methods can be viewed as preconditioned SGD where noise is isotropic near stationary points, which helps escape saddle points.



References

Rong Ge, Furong Huang, Chi Jin, Yang Yuan (2015)
Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition
https://arxiv.org/abs/1503.02101

Jason D. Lee, Max Simchowitz, Michael I. Jordan, Benjamin Recht (2016)
Gradient Descent Converges to Minimizers
https://arxiv.org/abs/1602.04915

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, Michael I. Jordan(2017)
How to Escape Saddle Points Efficiently
https://arxiv.org/abs/1703.00887

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, Michael I. Jordan (2019)
On Nonconvex Optimization for Machine Learning: Gradients, Stochasticity, and Saddle Points
https://arxiv.org/abs/1902.04811

Matthew Staib, Sashank J. Reddi, Satyen Kale, Sanjiv Kumar, Suvrit Sra (2020)
Escaping Saddle Points with Adaptive Gradient Methods
https://arxiv.org/abs/1901.09149