Review: AdaGrad Avoids Saddle Points

田皓原, 孙伟嘵, 郑嘉业

Shanghaitech University, Shanghai, China

2023 年 5 月

Outline

Background

- $AdaGrad$ is a family of adaptive first-order methods that automatically adapt an optimization algorithm's step size.
- The convergence guarantee leaves open the issue of converging to **undesirable saddle points**.
- We need rigorous proof to persuade the community.

Intuition

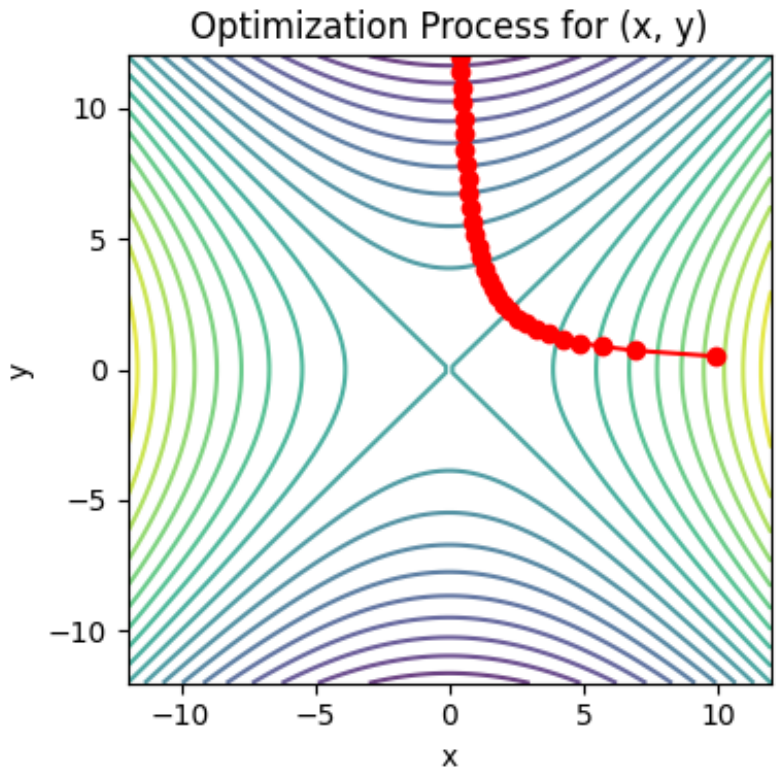

Figure: An empirical evidence that $AdaGrad$ avoids saddle point

Outline

AdaGrad Family

The *AdaGrad* algorithm is considered in the form:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Gamma_t \nabla_t f(\mathbf{x}_t)$$

where

- *AdaNorm*:

$$\Gamma_t = \frac{1}{\sqrt{\delta_0^2 + \sum_t^{s=0} \|\nabla f(x_s)\|_2^2}}$$

- *AdaGrad*-Diag: $\Gamma_t = G_t^{-\frac{1}{2}}$

$$G_t = \delta_0^2 I + diag(\sum_{s=0}^{t} \nabla f(x_s) \nabla f(x_s)^\top)$$

- *AdaGrad*-Full: $\Gamma_t = G_t^{-\frac{1}{2}}$

$$G_t = \delta_0^2 I + \sum_{s=0}^{t} \nabla f(x_s) \nabla f(x_s)^\top$$

Problem Statement

Problem:

$$\min_{x \in \mathbb{R}^d} f(x)$$

Assumptions:

- $\inf_{x \in \mathbb{R}^d} f(x) > -\infty$, this holds for loss function.
- Lipschitz continuity holds for $f(x)$
- The undesirable critical points is a set of strict saddle points

$$\|\nabla f(\mathbf{x}^*)\| = 0 \wedge \lambda_{\min}(\nabla^2 f(\mathbf{x}^*)) < 0$$

Analysis

The proof is classified into a few steps:

- Note that iterative algorithm $x_{t+1} \leftarrow x_t - \Gamma_t \nabla f(x)$ is a **dynamical system**.
- Adaptive step-size matrices Γ_t converge to **symmetric positive** definite matrices.
- Local Structure of $AdaGrad$ dynamical system(trivially near $\mathbf{0}$) can be expanded via Taylor's Theorem into linear and nonlinear parts,

$$\mathbf{x}_{t+1} = (I - \Gamma \nabla^2 f(\mathbf{0}))\mathbf{x}_t - \Gamma\theta(\mathbf{x}_t) - (\gamma_t - \Gamma)\nabla f(x_t)$$

where $\theta(\cdot)$ is the remainder of Taylor expansion.

Analysis (cont.)

- Stable manifold theorem
  The dynamical system

$$x_{t+1} \leftarrow x_t - \Gamma_t \nabla f(x)$$

  converges to an unstable fixed point only if the initial point $\mathbf{x}_0$ is taken from a certain lower-dimension manifold.

- Extend the local proposition to $\mathbb{R}^d$ via proving the map $\varphi(\mathbf{x}) = \mathbf{x} - \Gamma_t \nabla f(\mathbf{x})$ is diffeomorphism.

Outline

Result

Main Theorem
$AdaGrad$ algorithm family does not converge to undesirable critical points.

Q&A

Thank you!

感谢您的聆听和反馈