AdaGrad Avoids Saddle Points

Kimon Antonakopoulos (EPFL)

with Panayotis Mertikopoulos (CNRS), Georgios Piliouras (SUTD)
and Xiao Wang (SUFE)

Setup

Problem:

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

Assumptions:

- $\inf_{x \in \mathbb{R}^d} f(x) > -\infty$, this holds for loss functions.
- There exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^d.$$

- The undesirable critical points is the set of *strict saddle points*,

$$\|\nabla f(x^*)\| = 0, \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x^*)) < 0.$$

Algorithm: AdaGrad family

The AdaGrad algorithms we consider are of the following forms:

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

where

- AdaNorm:

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

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

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

- FullAdaGrad: $\Gamma_t = G_t^{-\frac{1}{2}}$,

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

Our techniques

Stabilization of step matrix

The adaptive step-size matrices Γ_t converge to symmetric positive definite matrices.

Proposition

Let Γ_t be one of adaptive step-size policies of AdaNorm, AdaDiag and FullAdaGrad. Then the following hold:

- For each initial point x_0, the eigenvalue of Γ_t converges to strictly positive numbers.
- For each sequence $\{x_t\}_{t\in\mathbb{N}}$ generated by AdaGrad, the sum of square of gradient norm is finite.
- $\{\Gamma_t\}_{t\in\mathbb{N}}$ exists and in particular, the limit is positive definite.

Local structure of AdaGrad

By Taylor expansion of AdaGrad dynamics at saddle point (assumed to be $\mathbf{0}$ W.L.O.G), combining with stabilization result, we can write the AdaGrad in the following way:

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

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

Stable manifold theorem

The dynamical system

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

converges to an unstable fixed point only if the initial point is taken from a certain lower dimension manifold.

Our Result

Main Theorem

AdaGrad (Ada-Norm, Ada-Diag, and FullAdaGrad) algorithms do not converge to undesirable critical points.

Thank you for your attention!