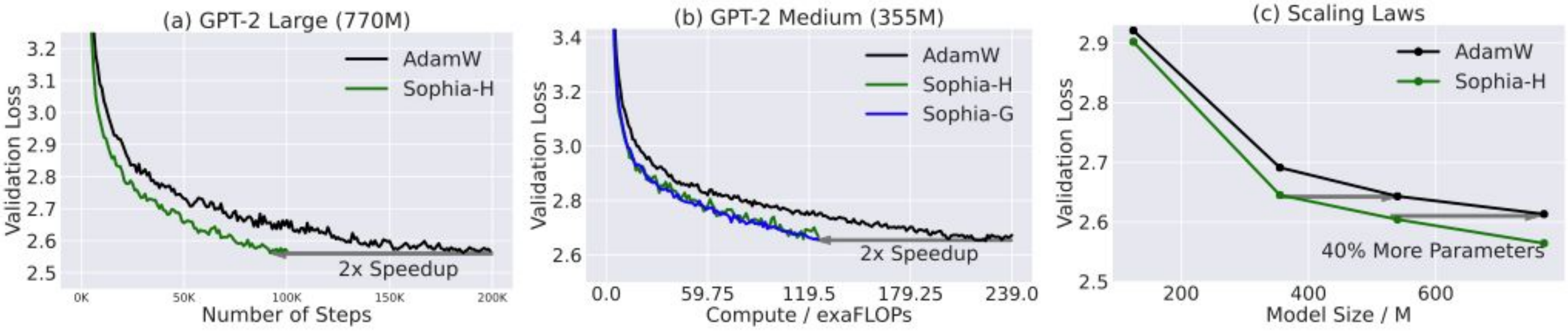
Sophia

Second-**o**rder Cli**p**ped Stoc**h**astic Opt**i**miz**a**tion

Abstract



Sophia achieves a **2x speed-up** compared with Adam in the **number of steps**, **total compute**, and **wall-clock time**.

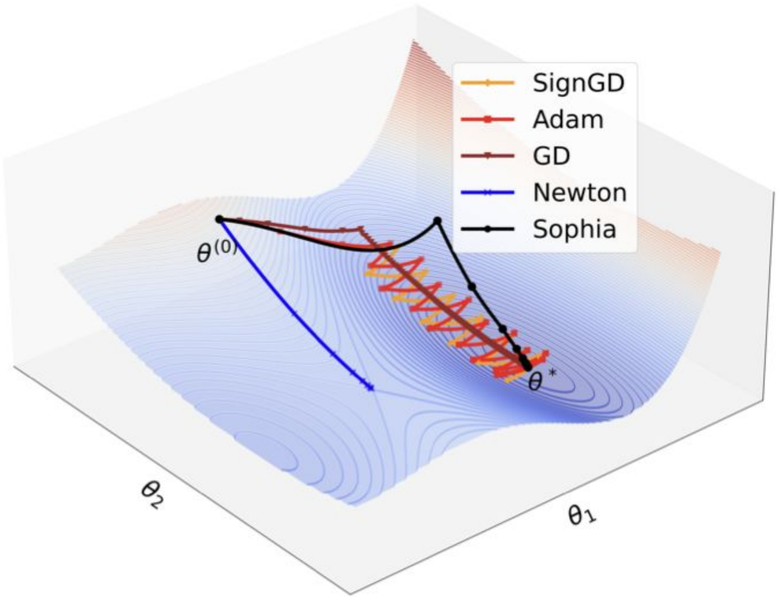
Motivation

Adam

Adam is a method that is commonly used for optimization, but it has limitations when dealing with different types of curves. It doesn't adapt well to curves that have varying shapes or curvatures.

Newton

Newton's method is effective for optimizing convex functions (functions that have a U-shaped curve), but it has weaknesses when dealing with negative curves or curves that change rapidly.



※ Source: https://arxiv.org/abs/2305.14342

Motivation

Adam

… limitations when dealing with different types of curves …

Newton

… weaknesses when dealing with negative curves or curves that change rapidly…

① EMA of diagonal Hessian estimates

② Pre-coordinate clipping

Introduces a new optimizer called **SOPHIA**

Method

① EMA of diagonal Hessian estimates

Exponential Moving Average (EMA)

The EMA of a sequence of values is a weighted average where **more recent values are given higher weights**. It provides a way to **smoothen values, reduce noise**, and adaptively adjust parameters based on recent information.

※ *EMA of gradient gt*

$$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t.$$

Method

① EMA of diagonal Hessian estimates

The Hessian matrix is a square matrix that contains the **second-order partial derivatives of the loss function with respect to pairs of parameters**. If we have a loss function with multiple parameters (θ1, θ2, ..., θn), the Hessian matrix will have dimensions n x n.

There are 2 options to calculate the diagonal of Hessian matrix: **Hutchinson's unbiased estimator** and **Gauss-Newton-Bartlett (GNB) estimator** *(Appendix)*

Method

EMA of diagonal Hessian estimates

$$h_t = \beta_2 h_{t-k} + (1 - \beta_2)\hat{h}_t \ \text{ if } t \bmod k = 1; \ \text{else } h_t = h_{t-1} \, .$$

Sophia uses a diagonal Hessian-based pre-conditioner, which directly **adjusts the update size of different parameter dimensions according to their curvatures**. To mitigate the overhead, we only estimate the Hessian every k steps (k = 10 in our implementation). At time step t with t mod k = 1, the estimator returns an estimate hˆt of the diagonal of the Hessian of the mini-batch loss.

Method

② Pre-coordinate clipping

The idea is to consider only the **positive entries of the diagonal Hessian**, discarding the negative entries. The update rule for parameter θ is then modified as follows:

$$\begin{cases} \theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \mathrm{clip}(m_t / \max\{h_t, \epsilon\}, \rho), \\ \mathrm{clip}(z, \rho) = \max\{\min\{z, \rho\}, -\rho\} \end{cases}$$

Method

② Pre-coordinate clipping

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \operatorname{clip}(m_t / \max\{h_t, \epsilon\}, \rho),$$

When any entry of h_t is negative, e.g., $h_t[i] < 0$, the corresponding entry in the pre-conditioned gradient $m_t[i]/\max\{h_t[i], \epsilon\} = m_t[i]/\epsilon$ is extremely large and has the same sign as $m_t[i]$, and thus $\eta \cdot \operatorname{clip}(m_t[i]/\max\{h_t[i], \epsilon\}, \rho) = \eta\rho \cdot \operatorname{sign}(m_t[i])$, which is the same as stochastic momentum SignSGD.

In the worst case, the update = ηp is still larger than the worst update size η in stochastic momentum SignSGD → Avoid vanishing gradient problem.

Appendix Diagonal Hessian Estimators

Option 1: Hutchinson's unbiased estimator

A method used to estimate the diagonal elements of the Hessian matrix, which describes the curvature of a loss function with respect to the parameters.

$$\mathbb{E}[\hat{h}] = \mathrm{diag}(\nabla^2 \ell(\theta))\,. \qquad (1)$$

Option 2: Gauss-Newton-Bartlett (GNB) estimator

A biased stochastic estimator used to approximate the diagonal elements of the Hessian matrix, which measures the curvature of a loss function.

$$\mathbb{E}_{\hat{y}_b's}\left[B \cdot \nabla_\theta \widehat{L}(\theta) \odot \nabla_\theta \widehat{L}(\theta)\right] = \mathbb{E}_{\hat{y}_b's}\left[\frac{1}{B}\sum_{b=1}^{B} \nabla \ell_{ce}(f(\theta, x_b), \hat{y}_b) \odot \nabla \ell_{ce}(f(\theta, x_b), \hat{y}_b)\right] \qquad (2)$$