Accelerated Stochastic Three Point Method

Kutenkov Roman, Kuznetsov Ivan

Scientific advisor: Solodkin Vladimir

April 17, 2025

Table of contents

Introduction

Problem

We consider unconstrained minimization problem in **DFO** setting

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

with smooth target function, that is bounded from below.
In this setting we only have access to a function evaluation oracle, it may be either because function gradient is impractical to evaluate, noisy or it is inaccessable at all.

Related work

Stochastic Three Points (STP) method

Our work is based on STP method, proposed in [1]. It requires distribution law $\mathcal{D}$ and stepsizes $\{\alpha_k\}_{k=0}^{\infty}$. The algorithm is

Sample random vector $s^k \sim \mathcal{D}$
Update $x^{k+1} = \arg\min\{f(x^k), f(x^k + \alpha_k s^k), f(x^k - \alpha_k s^k)\}$

Stochastic Momentum Three Points (SMTP) method

SMTP is a modification of STP, that uses momentum technique as described in [2]. This algorithm requires additional parameter β that is momentum. The main change is

Sample random vector $s^k \sim \mathcal{D}$ and set $v_{\pm}^{k+1} = \beta v^k \pm s^k$
Use points $x^k - \eta_k v_+^k$ and $x^k - \eta_k v_-^k$ to update x^{k+1} and v^{k+1}

where η_k is the combination of α_k and β

Motivation

Overview of FO optimization

The idea of using momentum was first applied in GD method. It works well in practice, however there are no proved boost of theoretical global convergence.

Achieving acceleration with linear coupling

Linear coupling [3] is the algorithm that essentially combines MD and GD. It has proven accelerated rates of convergence for strongly convex problems and it works in general $\|\cdot\|$-norm setup.

Motivation

The idea is to adopt this concept to ZO setup and achieve acceleration in theory and in practice.

Goals

- Develop algorithm based on three points method
- Proove its convergence and get accelerated convergence rate
- Consider practical choices of required parameters
- Compare its performance with STP and SMTP, as well as with other ZO methods

Initial idea of the algorithm

$$y^{k+1} = \arg\min\{f(x^k), f(x^k + \gamma_k s^k), f(x^k - \gamma_k s^k)\}$$

$$\mathrm{z}^{k+1} = \begin{cases} \mathrm{z}^k & \text{if } y^{k+1} = x^k, \\ \mathrm{z}^k + \alpha_k s^k & \text{if } y^{k+1} = x^k + \gamma_k s^k, \\ \mathrm{z}^k - \alpha_k s^k & \text{if } y^{k+1} = x^k - \gamma_k s^k \end{cases}$$

$$\mathrm{x}^{k+1} = \begin{cases} x^k & \text{if } y^{k+1} = x^k, \\ \tau_k \mathrm{z}^{k+1} + (1 - \tau_k) y^{k+1} & \text{otherwise} \end{cases}$$

My part in the project

Implement proposed algorithm and run experiments

- Run preliminary experiments on $f(x) = \frac{1}{2}x^TAx - b^Tx$
- Adapted code from [4] and set up the environment
- Run more experiments on fine-tuning LLM with some modifications of the algorithm
- Incorporated Optuna framework

Most of the runs are available at my WandB account

Assist with proof and try to come up with new ideas

- Considered resetting z^k every fixed number of steps
- Considered updating z^k only if condition of sort $f(\mathrm{z}^{k+1}) \leq f(x^k) + \delta_k$ is satisfied for some $\delta_k > 0$, otherwise reset $\mathrm{z}^{k+1} = y^{k+1}$
- Considered added second $\arg\min$ to ensure that z^k converges

Preliminary experiments

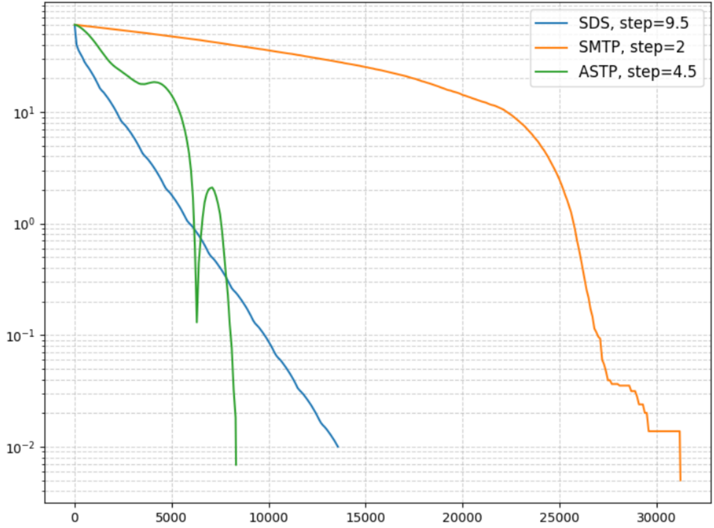

Figure: Comparison of the best case scenario for ASTP, SMTP and DDS. Parameter search was done on a unit grid

Experiments visualisation

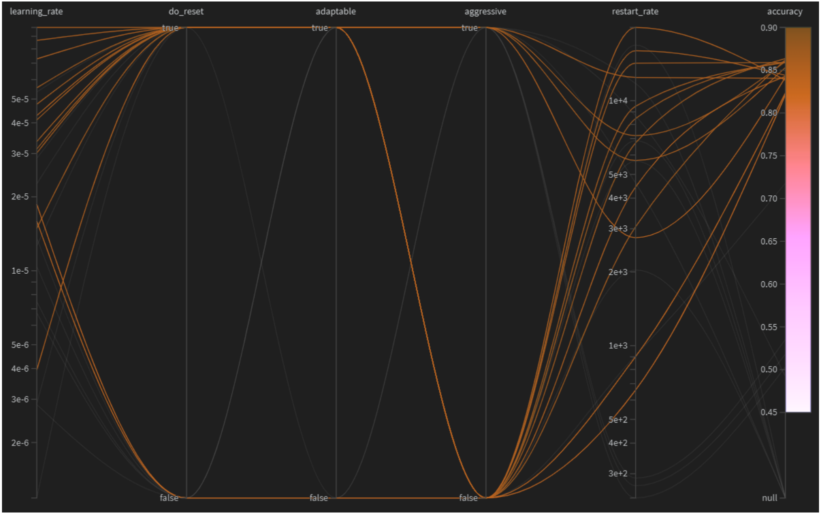

Figure: Visualisation of experiments on SST2 task in WandB

Conclusion

Proof of theoretical convergence

We were not able to proove accelerated convergence. For modification when z^k is apdated only if $f(z^{k+1}) \leq f(x^k) + \delta_k$ we obtained theoretical bounds that are similar to STP and we suspect that this method behaves almost the same as STP. The main difficulty is the choice of update rule for z^k, we have limited options and yet need it to be impactful.

Results of experiments

I ran experiments that involved fine-tuning LLM on two task: SST2 and RTE. First task is about classifying statement sentiment, I used facebook/opt-125m model and achieved similar performance to SMTP and zo-SGD method described in [4]. The second task is about recognizing whether one statement implies the other, it is more complicated than SST2, however I had troubles running it and could not even reproduce results from [4].

Conclusion

Task	ASTP	SMTP	zo-SGD
SST2	85.6%	86.1%	89.4%
RTE	58.1%	58.8%	68.7%

Table: Best accuracy achieved by each method*

RTE experiments

I used eval loss as a metric to see what happened during fine-tuning – the problem was it oscilated and did not steadily decrease. I implemented support for various stepsizes selection strategies and tried linear and cosine schedulers. Linear scheduler with warmup steps made eval loss descrease steadily, however it was slow.

References

- [1] El Bergou, Eduard Gorbunov, and Peter Richtarik – *stochastic three points method for unconstrained smooth minimization*
- [2] El Bergou, Eduard Gorbunov, Peter Richtarik, Adel Bibi, Ozan Sener – *a stochastic derivative free optimization method with momentum*
- [3] Zeyuan Allen-Zhu and Lorenzo Orecchia – *An Ultimate Unification of Gradient and Mirror Descent*
- [4] Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li and others – *revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: a benchmark*

- github repository from above with my code added