STOCHASTIC ZEROTH-ORDER PROXIMAL GRADIENT DESCENT WITH ADAPTIVE MOMENTUM (SZOPGD-AM)

Varun Gambhir
MT24161
Indraprastha Institute of Information Technology, Delhi

ABSTRACT

This project explores the design and analysis of a novel optimization algorithm, **Stochastic Zeroth-Order Proximal Gradient Descent with Adaptive Momentum (SZOPGD-AM)**, for minimizing composite objectives of the form $F(x) = f(x) + g(x)$, where $f(x)$ is smooth but gradients are unavailable, and $g(x)$ is non-smooth and convex. The algorithm combines zeroth-order gradient estimation, proximal operators, and adaptive momentum to address challenges in stochastic and non-smooth optimization. A detailed mathematical convergence analysis under standard assumptions is stated, showing that the algorithm achieves a sublinear convergence rate of $O(\log T/\sqrt{T})$.

1 Introduction

Optimization algorithms are critical in machine learning and data science, especially for solving large-scale problems where gradients may be unavailable or expensive to compute. In this project, a novel algorithm is proposed; **Stochastic Zeroth-Order Proximal Gradient Descent with Adaptive Momentum (SZOPGD-AM)**, designed for composite objectives:

$$F(x) = f(x) + g(x),$$

where:

- $f(x)$: Smooth but gradients are unavailable (zeroth-order setting).

- $g(x)$: Non-smooth and convex.

This report is organized as follows. Section 2 describes the proposed algorithm. Section 3 provides a detailed convergence analysis. Section 4 includes a dry-run example. Section 5 concludes with a discussion of potential extensions and future work.

2 Algorithm

The proposed algorithm, **SZOPGD-AM**, is outlined in Algorithm 1. It integrates zeroth-order gradient estimation, momentum updates, and proximal operators to handle stochasticity and non-smoothness effectively.

Algorithm 1 Stochastic Zeroth-Order Proximal Gradient Descent with Adaptive Momentum (SZOPGD-AM)

Input: Initial point x_0, initial step size η_0, momentum parameter β, smoothing parameter μ, number of iterations T.
Output: Final iterate x_T
Initialize: $v_0 \leftarrow 0, t \leftarrow 0$
for $t = 0, 1, \ldots, T-1$ **do**
  Sample random perturbation $u_t \sim \mathcal{N}(0, I_d)$
  Compute zeroth-order gradient estimate:

$$\nabla f_{\text{hat}}(x_t) = \frac{f(x_t + \mu u_t) - f(x_t)}{\mu} u_t$$

  Update momentum term:

$$v_{t+1} = \beta v_t + (1-\beta)\nabla f_{\text{hat}}(x_t)$$

  Compute adaptive step size:

$$\eta_t = \frac{\eta_0}{\sqrt{t+1}}$$

  Perform proximal update:

$$x_{t+1} = \text{prox}_{\eta_t g}(x_t - \eta_t v_{t+1})$$

$\triangleright$ $\text{prox}_{\eta_t g}(y) = \arg\min_x \left\{ g(x) + \frac{1}{2\eta_t}\|x-y\|^2 \right\}$
end for
Return: x_T

3 Convergence Analysis

We analyze the convergence of **SZOPGD-AM** under the following assumptions:

[Smoothness of $f(x)$] The function $f(x)$ is L-smooth, i.e., its gradient is Lipschitz continuous:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|, \quad \forall x, y.$$

[Convexity of $F(x)$] The composite objective function $F(x) = f(x) + g(x)$ is convex:

$$F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y), \quad \forall x, y, \lambda \in [0,1].$$

[Bounded Variance of Zeroth-Order Gradient Estimator] The zeroth-order gradient estimator satisfies:

$$\mathbb{E}[\nabla f_{\text{hat}}(x)] = \nabla f(x), \quad \mathbb{E}[\|\nabla f_{\text{hat}}(x) - \nabla f(x)\|^2] \leq \sigma^2.$$

[Step Size Schedule] The step size decreases as:

$$\eta_t = \frac{\eta_0}{\sqrt{t+1}},$$

where $\eta_0 > 0$ is the initial step size.

[Momentum Parameter] The momentum parameter satisfies $\beta \in [0, 1)$, controlling the weight of past gradients.

3.1 Lyapunov Function

Define the Lyapunov function:

$$V_t = \|x_t - x^*\|^2 + \frac{\eta_t}{1-\beta}\|v_t\|^2,$$

where x^* is an optimal solution to $F(x)$. This function tracks both the distance to the optimal solution and the contribution of the momentum term.

3.2 Proof of Convergence

3.2.1 Step 1: Expand $\|x_{t+1} - x^*\|^2$

Using the proximal operator property:

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - \eta_t v_{t+1} - x^*\|^2 - \|x_t - \eta_t v_{t+1} - x_{t+1}\|^2.$$

Expand $\|x_t - \eta_t v_{t+1} - x^*\|^2$:

$$\|x_t - \eta_t v_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 - 2\eta_t \langle v_{t+1}, x_t - x^* \rangle + \eta_t^2 \|v_{t+1}\|^2.$$

Thus:

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - 2\eta_t \langle v_{t+1}, x_t - x^* \rangle + \eta_t^2 \|v_{t+1}\|^2 - \|x_t - \eta_t v_{t+1} - x_{t+1}\|^2.$$

3.2.2 Step 2: Bound $\|v_{t+1}\|^2$

From the momentum update:

$$v_{t+1} = \beta v_t + (1 - \beta) \nabla f_{\text{hat}}(x_t).$$

Taking norms and applying the triangle inequality:

$$\|v_{t+1}\|^2 \leq \beta \|v_t\|^2 + (1 - \beta) \|\nabla f_{\text{hat}}(x_t)\|^2.$$

Using the bounded variance assumption:

$$\mathbb{E}[\|\nabla f_{\text{hat}}(x_t)\|^2] \leq 2\|\nabla f(x_t)\|^2 + 2\sigma^2.$$

Thus:

$$\mathbb{E}[\|v_{t+1}\|^2] \leq \beta \mathbb{E}[\|v_t\|^2] + (1 - \beta)(2\mathbb{E}[\|\nabla f(x_t)\|^2] + 2\sigma^2).$$

3.2.3 Step 3: Combine Terms into the Lyapunov Function

Substitute the bounds for $\|x_{t+1} - x^*\|^2$ and $\|v_{t+1}\|^2$ into the Lyapunov function V_t:

$$\mathbb{E}[V_{t+1}] \leq \mathbb{E}[V_t] - 2\eta_t \mathbb{E}[\langle v_{t+1}, x_t - x^* \rangle] + \eta_t^2 \mathbb{E}[\|v_{t+1}\|^2] - \mathbb{E}[\|x_t - \eta_t v_{t+1} - x_{t+1}\|^2] + \frac{\eta_{t+1}}{1 - \beta} \mathbb{E}[\|v_{t+1}\|^2].$$

Using the convexity of $F(x)$ and the unbiasedness of $\nabla f_{\text{hat}}(x_t)$, we have:

$$\mathbb{E}[\langle v_{t+1}, x_t - x^* \rangle] \geq \mathbb{E}[F(x_t) - F(x^*)].$$

Thus:

$$\mathbb{E}[V_{t+1}] \leq \mathbb{E}[V_t] - c_1 \eta_t \mathbb{E}[\|\nabla f(x_t)\|^2] + c_2 \eta_t^2 (\sigma^2 + \mathbb{E}[\|v_t\|^2]),$$

where $c_1, c_2 > 0$ are constants depending on L, β, and other parameters.

3.2.4 Step 4: Telescoping Sum

Summing over $t = 0, 1, \ldots, T - 1$, and using the step size schedule $\eta_t = \frac{\eta_0}{\sqrt{t+1}}$:

$$\sum_{t=0}^{T-1} \eta_t \mathbb{E}[\|\nabla f(x_t)\|^2] \leq C_1 + C_2 \sum_{t=0}^{T-1} \eta_t^2,$$

where $C_1, C_2 > 0$ are constants.
Since $\eta_t = \frac{\eta_0}{\sqrt{t+1}}$, we have:

$$\sum_{t=0}^{T-1} \eta_t^2 = O(\log T).$$

Thus:

$$\mathbb{E}[F(x_T) - F(x^*)] = O\left(\frac{\log T}{\sqrt{T}}\right).$$

3.2.5 Step 5: Final Convergence Rate

Under the assumptions, the algorithm achieves a sublinear convergence rate:

$$\boxed{\mathbb{E}[F(x_T) - F(x^*)] = O\left(\frac{\log T}{\sqrt{T}}\right)}.$$

4 Dry-Run Example

To validate the algorithm, let us perform a dry-run for one iteration ($t = 0$):

- **Initial Values:**

$$x_0 = [1, 1]^\top, \quad v_0 = [0, 0]^\top, \quad \eta_0 = 0.1, \quad \beta = 0.9, \quad \mu = 0.01.$$

- **Random Perturbation:** Sample $u_0 \sim \mathcal{N}(0, I_2)$. Suppose $u_0 = [0.5, -0.3]^\top$.
- **Zeroth-Order Gradient Estimate:**

$$\nabla f_{\text{hat}}(x_0) = \frac{f(x_0 + \mu u_0) - f(x_0)}{\mu} u_0.$$

Suppose $f(x_0 + \mu u_0) = 1.05$, $f(x_0) = 1.0$:

$$\nabla f_{\text{hat}}(x_0) = \frac{1.05 - 1.0}{0.01}[0.5, -0.3]^\top = [5, -3]^\top.$$

- **Momentum Update:**

$$v_1 = \beta v_0 + (1 - \beta)\nabla f_{\text{hat}}(x_0).$$

With $\beta = 0.9$:

$$v_1 = 0.9[0, 0]^\top + 0.1[5, -3]^\top = [0.5, -0.3]^\top.$$

- **Proximal Update:** Compute:

$$x_1 = \text{prox}_{\eta_0 g}(x_0 - \eta_0 v_1).$$

Suppose $g(x) = \|x\|_1$. Solve:

$$x_1 = \arg\min_x \left\{ \|x\|_1 + \frac{1}{2 \cdot 0.1} \|x - ([1, 1]^\top - 0.1[0.5, -0.3]^\top)\|^2 \right\}.$$

Simplify:

$$x_1 = \arg\min_x \left\{ \|x\|_1 + 5\|x - [0.95, 1.03]^\top\|^2 \right\}.$$

Use soft-thresholding to compute x_1.

5 Conclusion

In this project, **SZOPGD-AM**, a novel optimization algorithm combining zeroth-order gradient estimation, proximal operators, and adaptive momentum is proposed and analyzed. Our convergence analysis shows that the algorithm achieves a sublinear rate of $O(\log T/\sqrt{T})$, making it suitable for stochastic and non-smooth optimization problems.