Optimization with Markovian Noise

First and Zero Order Methods

Boris Prokhorov

Adviser - Aleksandr Beznosikov

April 9, 2024

Table of Contents

Table of Contents

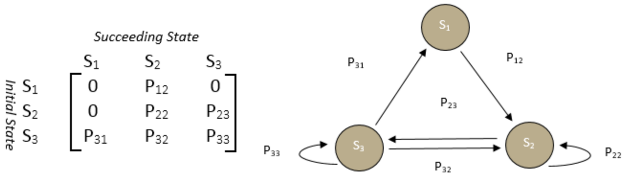
We study the minimization problem

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_{Z \sim \pi}[F(x, Z)] \tag{1}$$

Where Z is a Markov Chain.

Introduction

Quick intro to Markov Chains



Transition Matrix **Transition Diagram**

Table of Contents

Our research aims to expand the results of [1].
Following the article, we consider additional assumptions:

- (A1,A2) f is $(\mu,\, L)$-smooth.
- (A3) Z is uniformly geometrically ergodic with mixing time τ.
- (A4) F is $(\sigma,\, \delta)$-bounded:

$$\|\nabla F(x, Z) - \nabla f(x)\|^2 < \sigma^2 + \delta^2 \|\nabla f(x)\|^2$$

The main result of [1] is the following

Theorem. Under A1-A4 the problem (1) can be solved (in terms of $\mathbb{E}\left[\|x^{(k)} - x^*\|^2\right] \leq \varepsilon$) in

$$\tilde{\mathcal{O}}\left(\tau\left[(1+\delta^2)\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon} + \frac{\sigma^2}{\mu^2\varepsilon}\right]\right) \quad \text{oracle calls}\,. \tag{2}$$

And the bound is tight for $\delta = 1$.

Our main focus is on the optimality of the bound (2).

Summary of the previous results:

1. For $\delta = 1$ the bound is tight. (**Theorem**, [1])
2. For $\delta > 1$ the bound is tight. (**Seminar 1**)
   The proof generalizes the previous result.
3. For $\delta < 1$ the bound is **not** tight. (Polyak, [4])
   Gradient Descent converges even if the noise is adversarial.
4. For $\delta \lesssim \sqrt{\frac{L}{\mu}}$ the bound is **not** tight. ([2])
   Accelerated GD [3] converges even if the noise is adversarial.

$\sqrt{\frac{L}{\mu}} \lesssim \delta < 1$ is still an open problem. Current progress:

1. For $\delta \lesssim \sqrt{\frac{L}{\mu}}$ the optimal bound is $\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right)$ (see [2])
2. For $\delta < 1$ the upper bound is $\tilde{\mathcal{O}}\left(\frac{1}{1-\delta} \frac{L}{\mu} \log \frac{1}{\varepsilon}\right)$ (see [4])
3. For $\delta < 1$ the lower bound is $\tilde{\mathcal{O}}\left(\frac{1}{1-\delta} \sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right)$ (**New**)

We also found a much shorter proof for the 1st bound.

Table of Contents

Future work

Zero Order

We keep main assumptions A1-A4.

Now only values of $F(x, Z)$ are available, instead of $\nabla F(x, Z)$.

Using finite-difference schemes we can approximate the gradients:

$$\langle \nabla F(x, Z), v \rangle = \frac{\partial F}{\partial v} \approx \frac{F(x + hv, Z) - F(x, Z)}{h}$$

Our main focus is on lower bounds.

Table of Contents

Bibliography I

[1] Aleksandr Beznosikov et al. *First Order Methods with Markovian Noise: from Acceleration to Variational Inequalities*. 2023. arXiv: 2305.15938 [math.OC].

[2] Nikita Kornilov et al. *Intermediate Gradient Methods with Relative Inexactness*. 2023. arXiv: 2310.00506 [math.OC].

[3] Yurii Nesterov. "A method for solving the convex programming problem with convergence rate $O(1/k^2)$". In: *Proceedings of the USSR Academy of Sciences* 269 (1983), pp. 543–547.

[4] Boris Polyak. *Introduction to Optimization*. July 2020.