Optimization with Markovian Noise

First and Zero Order Lower Bounds

Boris Prokhorov

Adviser - Aleksandr Beznosikov

May 17, 2024

Table of Contents

Table of Contents

Stochastic optimization problem

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

Is well-studied [4].

We consider generalization:

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_{Z \sim \pi}[F(x, Z)] \tag{2}$$

Where Z is a Markov Chain and π is it's invariant distribution.

Introduction

Quick intro to Markov Chains

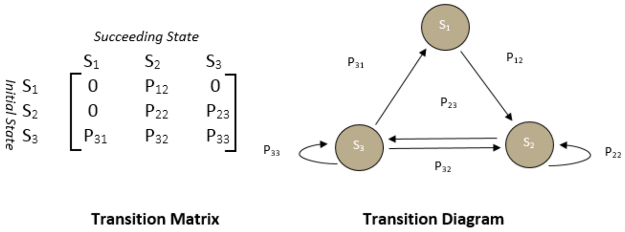

Table of Contents

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 can be solved (in terms of $\mathbb{E}\left[\|x^{(k)} - x^*\|^2\right] \le \varepsilon$) in

$$\tilde{\mathcal{O}}\left(\tau\left[(1+\delta^2)\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon} + \frac{\sigma^2}{\mu^2\varepsilon}\right]\right) \quad \text{oracle calls}\,. \tag{3}$$

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

Our main focus is on the optimality of the bound (3).

The main result of our work is the
Theorem. For any $\delta \geq 1$ there is a problem instance f and markov chain Z satisfying A1-A4 such that any first order method needs at least

$$\tilde{\mathcal{O}}\left(\tau\left[(1+\delta^2)\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right]\right) \quad \text{oracle calls} \tag{4}$$

in order to achieve $\mathbb{E}\left[\|x^{(k)}-x^*\|^2\right] \leq \varepsilon$.

Thus we obtain optimal convergence rates for $\delta \geq 1$.

Table of Contents

Small noise

Previous Work

The case $\delta < 1$ is special. (A4):

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

Theorem (Polyak [4]) Under A1-A2, A4 with $\sigma = 0$ the problem can be solved (in terms of $\|x^{(k)} - x^*\|^2 \leq \varepsilon$) in

$$\tilde{\mathcal{O}}\left(\frac{1}{1-\delta}\frac{L}{\mu}\log\frac{1}{\varepsilon}\right) \quad \text{oracle calls} \tag{5}$$

Which means that GD is robust to arbitrary bad noise ($\tau = \infty$).

When $\delta \lesssim \sqrt{\frac{\mu}{L}}$, the Accelerated GD [3] is also robust.

Theorem [2] Under A1-A2, A4 with $\delta \lesssim \sqrt{\frac{\mu}{L}}$, $\sigma = 0$ the problem can be solved (in terms of $\|x^{(k)} - x^*\|^2 \leq \varepsilon$) in

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right) \quad \text{oracle calls} \tag{6}$$

But for $\sqrt{\frac{\mu}{L}} \lesssim \delta < 1$ the optimal rate is still unknown.

Small noise

New Result

In adversarial noise regime we established the following

Theorem. For any $\delta < 1$ there is a problem instance f such that any first order method needs at least

$$\tilde{\mathcal{O}}\left(\frac{1}{1-\delta}\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right) \quad \text{oracle calls} \tag{7}$$

in order to achieve $\|x^{(k)} - x^*\|^2 \leq \varepsilon$

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:

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

Table of Contents

Contribution

- For $\delta \geq 1$ the rate (3)

$$\tilde{\mathcal{O}}\left(\tau\left[(1+\delta^2)\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right]\right)$$

  is shown to be optimal in markovian settings.

- For $\delta < 1$ the rate (5)

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$$

  is shown to be optimal in terms of δ for adversarial noise.

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.