First Order Methods with Markovian Noise: from Acceleration to Variational Inequalities

Aleksandr Beznosikov [1,2,3,4] Sergey Samsonov [5] Marina Sheshukova [5] Alexander Gasnikov [3,2,6] Alexey Naumov [5] Eric Moulines [7]

[1]Innopolis University [2]Skoltech [3]Moscow Institute of Physics and Technology [4]Yandex [5] HSE University [6] Ecole polytechnique

Problem

• We study the minimization problem

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

where the access to the function f and its gradient is available only through the noisy oracle $F(x,Z)$ and $\nabla F(x,Z)$, respectively.

Setting

• **The function f is L-smooth on $\mathbb{R}^d$ with** $L>0$, i.e., it is differentiable and there is a constant $L>0$ such that the following inequality holds for all $x,y\in\mathbb{R}^d$:

$$\|\nabla f(x)-\nabla f(y)\| \le L\|x-y\|.$$

• **The function f is μ-strongly convex on** $\mathbb{R}^d$, i.e., it is continuously differentiable and there is a constant $\mu>0$ such that the following inequality holds for all $x,y\in\mathbb{R}^d$:

$$\frac{\mu}{2}\|x-y\|^2 \le f(x)-f(y)-\langle\nabla f(y),x-y\rangle.$$

We consider here the general setting of $\{Z_i\}_{i=0}^{\infty}$ being a time-homogeneous Markov chain.

• $\{Z_i\}_{i=0}^{\infty}$ **is a stationary Markov chain on** $(\mathsf{Z},\mathcal{Z})$ with Markov kernel Q and unique invariant distribution π. Moreover, Q is **uniformly geometrically ergodic with mixing time** $\tau\in\mathbb{N}$, i.e., for every $k\in\mathbb{N}$,

$$\triangle(\mathrm{Q}^k) = \sup_{z,z'\in\mathsf{Z}} (1/2)\|\mathrm{Q}^k(z,\cdot)-\mathrm{Q}^k(z',\cdot)\|_{\mathsf{TV}} \le (1/4)^{\lfloor k/\tau\rfloor}.$$

Next we specify our assumptions on stochastic gradient:

• For all $x\in\mathbb{R}^d$ it holds that $\mathbb{E}_\pi[\nabla F(x,Z)]=\nabla f(x)$. Moreover, for all $z\in\mathsf{Z}$ and $x\in\mathbb{R}^d$ it holds that

$$\|\nabla F(x,z)-\nabla f(x)\|^2 \le \sigma^2+\delta^2\|\nabla f(x)\|^2.$$

Main Contributions

◇ **Accelerated SGD.** We provide the first analysis of SGD, including the Nesterov accelerated SGD method, with Markov noise without the assumption of bounded domain and uniformly bounded stochastic gradient estimates. Our results are summarised in Table and cover both strongly convex and non-convex scenarios.

◇ **Lower bounds.** We give the lower bounds showing that the presence of mixing time in the upper complexity bounds is not an artefact of the proof.

◇ **Extensions.** We provide, as far as we know, the first analysis for variational inequalities with general stochastic Markov oracle, arbitrary optimization set, and arbitrary composite term. Our finite-time performance analysis provides complexity bounds in terms of oracle calls that scale linearly with the mixing time of the underlying chain.

Algorithm

Algorithm 1 `Randomized Accelerated GD`

Parameters: stepsize $\gamma>0$, momentums θ,η,β,p, number of iterations N, batchsize limit M
Initialization: choose $x^0=x_f^0$
for $k=0,1,2,\dots,N-1$ **do**
  $x_g^k=\theta x_f^k+(1-\theta)x^k$
  Sample $J_k\sim$ Geom $(1/2)$
  $g^k=g_0^k+\begin{cases}2^{J_k}\left(g_{J_k}^k-g_{J_k-1}^k\right), & \text{if } 2^{J_k}\le M\\ 0, & \text{otherwise}\end{cases}$
  with $g_j^k=2^{-j}B^{-1}\sum_{i=1}^{2^jB}\nabla f(x_g^k,Z_{T^k+i})$
  $x_f^{k+1}=x_g^k-p\gamma g^k$
  $x^{k+1}=\eta x_f^{k+1}+(p-\eta)x_f^k+(1-p)(1-\beta)x^k+(1-p)\beta x_g^k$
  $T^{k+1}=T^k+2^{J_k}B$
end for

Key lemma

Let assumptions are valid. Then for the gradient estimates g^k from Algorithm 1 it holds that

$$\mathbb{E}_k[g^k] = \mathbb{E}_k[g^k_{\lfloor\log_2 M\rfloor}],$$
$$\mathbb{E}_k[\|\nabla f(x^k)-g^k\|^2] \lesssim \left(\tau B^{-1}\log_2 M+\tau^2B^{-2}\right)(\sigma^2+\delta^2\|\nabla f(x^k)\|^2),$$
$$\|\nabla f(x^k)-\mathbb{E}_k[g^k]\|^2 \lesssim \tau^2M^{-2}B^{-2}(\sigma^2+\delta^2\|\nabla f(x^k)\|^2).$$

Summary

◇ $\|\nabla f(x^k)-\mathbb{E}_k[g^k]\|^2\sim M^{-2}$.
◇ M can be super big, but $\mathbb{E}[2^{J_k}]=\mathcal{O}(1)$.
◇ It gives that $\|\nabla f(x^k)-\mathbb{E}_k[g^k]\|^2$ can be killed for free.

Convergence and complexity

Theorem

Let assumptions are valid and let a problem be solved by Algorithm 1. Then for any $b\in\mathbb{N}^*$, $\gamma\in(0;\frac{3}{4L}]$, and β,θ,η,p,M,B satisfying

$$p\simeq(1+(1+\gamma L)[\delta^2\tau b^{-1}+\delta^2\tau^2b^{-2}])^{-1},\quad \beta\simeq\sqrt{p^2\mu\gamma},$$
$$\eta\simeq\sqrt{\tfrac{1}{\mu\gamma}},\quad \theta\simeq\tfrac{p\eta^{-1}-1}{\beta p\eta^{-1}-1},$$
$$M\simeq\max\{2;\sqrt{p^{-1}(1+p/\beta)}\},\quad B=\lceil b\log_2 M\rceil,$$

it holds that

$$\mathbb{E}\left[\|x^N-x^*\|^2+\frac{6}{\mu}(f(x_f^N)-f(x^*))\right]$$
$$\lesssim \exp\left(-N\sqrt{\frac{p^2\mu\gamma}{3}}\right)\left[\|x^0-x^*\|^2+\frac{6}{\mu}(f(x^0)-f(x^*))\right]$$
$$+\frac{p\sqrt{\gamma}}{\mu^{3/2}}\left(\sigma^2\tau b^{-1}+\sigma^2\tau^2b^{-2}\right).$$

Corollary

Under the conditions of Theorem, choosing $b=\tau$ and γ as

$$\gamma\simeq\min\left\{\frac{1}{L};\frac{1}{p^2\mu N^2}\right\}$$

in order to achieve ε-approximate solution (in terms of $\mathbb{E}[\|x-x^*\|^2]\lesssim\varepsilon$) it takes

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

Summary

◇ Repeats the estimate for independent noise, but with an additional τ multiplier due to Markov oracle.
◇ It is unclear whether this estimate is unprovable. It may be possible to eliminate τ for some summands, i.e., $\tilde{\mathcal{O}}\left((1+\delta^2)\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}+\tau\frac{\sigma^2}{\mu^2\varepsilon}\right)$.

Lower bounds

Theorem

There exists an instance of the optimization problem satisfying assumptions with $\delta=1$ and arbitrary $\sigma\ge0, L,\mu>0,\tau\in\mathbb{N}^*$, such that for any first-order gradient method it takes at least

$$N=\Omega\left(\tau\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}+\frac{\tau\sigma^2}{\mu^2\varepsilon}\right)$$

oracle calls in order to achieve $\mathbb{E}[\|x^N-x^*\|^2]\le\varepsilon$.

Theorem

There exists an instance of the optimization problem satisfying assumptions with arbitrary $L,\mu>0,\tau\in\mathbb{N}^*$, $\delta=\frac{L}{\mu}$, and $\sigma=0$, such that for any first-order gradient method it takes at least

$$N=\Omega\left(\tau\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$$

gradient calls in order to achieve $\mathbb{E}[\|x^N-x^*\|^2]\le\varepsilon$.

Theorem

There exists an instance of the optimisation problem satisfying assumptions with with arbitrary $L,\mu>0,\tau\in\mathbb{N}^*$, $\sigma=1,\delta=0$, such that for any first-order gradient method it takes at least

$$N=\Omega\left(\left(\tau+\sqrt{\frac{L}{\mu}}\right)\log\frac{1}{\varepsilon}\right)$$

oracle calls in order to achieve $\mathbb{E}[\|x^N-x^*\|^2]\le\varepsilon$.

Summary

◇ Only for particular cases.
◇ In particular cases lower bounds show the optimality of Algorithm 1.
◇ BUT Lower bounds in the general case remain an open question, and thus the overall optimality of the proposed algorithm is not proved.

Table 1: This table summarizes our results on first-order method with Markovian noise. The columns of the table indicate whether the authors consider optimization over bounded domain, potentially unbounded gradients, and whether or not they assume additional restrictions on the Markovian noise (finite state space or reversibility). For ease of comparison we provide the respective results on SGD and ASGD (accelerated SGD) in the i.i.d. setting.

	Method	Unbounded: Domain	Unbounded: Gradient noise	General MC	Acceleration	Oracle complexity (Smooth and non-convex)	Oracle complexity (Smooth and strongly convex)
i.i.d.	SGD [1,2,3]	✓	✗	N/A	✗	$\tilde{\mathcal{O}}\left(L(f(x^0)-f(x^*))\left[\frac{1}{\varepsilon^2}+\frac{\sigma^2}{\varepsilon^4}\right]\right)$	$\tilde{\mathcal{O}}\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right)$
i.i.d.	ASGD [4,5]	✓	✓	N/A	✓	$\tilde{\mathcal{O}}\left(L(f(x^0)-f(x^*))\left[\frac{1+\delta^2}{\varepsilon^2}+\frac{\sigma^2}{\varepsilon^4}\right]\right)$	$\tilde{\mathcal{O}}\left((1+\delta^2)\sqrt{\frac{L}{\mu}}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right)$
Markovian	EMD [6]	✗	✗	✓	✗	$\tilde{\mathcal{O}}\left(\frac{\tau G^2D^2}{\varepsilon^4}\right)$	✗
Markovian	MC SGD [7]	✓	✗	✗	✗	$\tilde{\mathcal{O}}\left(h(G,L)\left(\frac{\tau}{\varepsilon^2}\right)^{1/(1-q)}\right)$	✗
Markovian	MC SGD [8]	✓	✓	✓	✗	$\tilde{\mathcal{O}}\left(\frac{\tau L^2(1+\|x^*\|^2+\|x^0-x^*\|^2)}{\varepsilon^4}\right)$	$\tilde{\mathcal{O}}\left(e^{\tau(L/\mu)^2}\left[h(\frac{L}{\mu})\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\tau^2L^2(1+\|x^*\|^2)}{\mu^2\varepsilon}\right]\right)$
Markovian	ASGD [9]	✗	✗	✗	✓	$\tilde{\mathcal{O}}\left(\frac{1}{\varepsilon^4}\left[B^2+G^6(L^2\tau^2+1)\right]\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\frac{\|x^0-x^*\|^2}{\varepsilon^{1/2}}+\frac{\tau^2(G^2+\mu GD+\mu LD^2)}{\mu^2\varepsilon}\right)$
Markovian	MAG [10]	✓	✗	✓	✗	$\tilde{\mathcal{O}}\left(\frac{\tau(G+L+B)^2G^2}{\varepsilon^4}\right)$	✗
Markovian	MC SGD [11]	✓	✗	✗	✗	$\mathcal{O}\left(\frac{\tau(L(f(x^0)-f(x^*))+\sigma^2)}{\varepsilon^2}+\frac{\tau(L(f(x^0)-f(x^*))+\sigma^2)\sigma^2}{\varepsilon^4}\right)$	$\mathcal{O}\left(\frac{\tau L}{\mu}\log\frac{(f(x^0)-f(x^*))/\mu+\sigma^2/(\mu L)}{\varepsilon}+\frac{\tau\sigma^2}{\mu^2\varepsilon}\right)$
Markovian	MC SGD [11]	✓	✗	✗	✗	✗	$O\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{L\tau\sigma_*^2}{\mu^3\varepsilon}\right)$
Markovian	RASGD (ours)	✓	✓	✓	✓	$\tilde{\mathcal{O}}\left(\tau L(f(x^0)-f(x^*))\left[\frac{1+\delta^2}{\varepsilon^2}+\frac{\sigma^2}{\varepsilon^4}\right]\right)$	$\tilde{\mathcal{O}}\left(\tau\left[(1+\delta^2)\sqrt{\frac{L}{\mu}}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right]\right)$

notation: $G=\sup_{x,z}\|\nabla F(x,z)\|$. Note that $G\ge L$. We also set $B=\sup_x|f(x)|$; x^0 - starting point, x^* - solution, $\mathcal{D}$ - optimization domain; $D=\sup_{x\in\mathcal{D}}\|x-x^*\|$, σ_* - stochasticity parameter in x^*, ε - accuracy of the solution, measured as $\mathbb{E}[\|\nabla f(x)\|^2]\lesssim\varepsilon^2$ for non-convex problems and $\mathbb{E}[\|x-x^*\|^2]\lesssim\varepsilon$ for the strongly convex ones. Functions $h(L/\mu)$ and $h(G,L)$ stands for an implicit dependence of the respective parameters.

References

[1] H. Robbins & S. Monro. A Stochastic Approximation Method. AMS, 1951
[2] E. Moulines & F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. NeurIPS, 2011
[3] S. Ghadimi et al. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Mathematical Programming, 2016
[4] S. Vaswani et al. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. AISTATS, 2019.
[5] Y. Chen et al. Convergence analysis of accelerated stochastic gradient descent under the growth condition. 2020.
[6] J. Duchi et al. Ergodic mirror descent. SIAM 2012.
[7] T. Sun et al. On Markov chain gradient descent. NeurIPS, 2018.
[8] T. Doan. Finite-time analysis of markov gradient descent. IEEE, 2023.
[9] T. Doan et al. Convergence rates of accelerated markov gradient descent with applications in reinforcement learning. arXiv, 2020.
[10] R. Dorfman & K. Levy. Adapting to mixing time in stochastic optimization with markovian data. ICML, 2022.
[11] M. Even. Stochastic gradient descent under Markovian sampling schemes. arXiv, 2023.