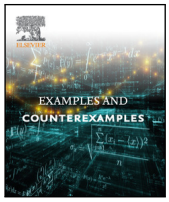
4 (2023) 100123

Contents lists available at ScienceDirect

Examples and Counterexamples

journal homepage: www.elsevier.com/locate/exco



Counterexamples for Noise Models of Stochastic Gradients

Vivak Patel

Department of Statistics, University of Wisconsin – Madison, 1300 University Ave, Madison, 53703, WI, USA

ARTICLE INFO

Keywords:
Stochastic Gradient Descent
Noise Models

ABSTRACT

Stochastic Gradient Descent (SGD) is a widely used, foundational algorithm in data science and machine learning. As a result, analyses of SGD abound making use of a variety of assumptions, especially on the noise behavior of the stochastic gradients. While recent works have achieved a high-degree of generality on assumptions about the noise behavior of the stochastic gradients, it is unclear that such generality is necessary. In this work, we construct a simple example that shows that less general assumptions will be violated, while the most general assumptions will hold.

1. Introduction

Stochastic Gradient Descent (SGD) is a foundational algorithm for stochastic optimization that is essential to machine learning and data science. As a result, SGD has been widely analyzed with a number of remarkable recent results about its global convergence behavior [2,4–7], greedy global complexity behavior [3,8], local convergence behavior [6,9,10], and unstable saddle-point behavior [6,11,12].

These analyses of SGD make a number of different assumptions about the stochastic optimization problem, especially on the noise behavior of the stochastic gradients. These assumptions range from highly restrictive—the stochastic gradients having a uniformly bounded variance, see A1—to highly general—there exists an $\epsilon \in (0, 1]$ such that $1 + \epsilon$ moment of the stochastic gradients is bounded by an arbitrary upper semi-continuous function, see A6. While the generality is very appealing, SGD users and experts often argue that SGD can be practically limited to a bounded region which would render the most general assumption equivalent to the most restrictive. This argument raises two questions. First, is there a stochastic optimization problem that is limited to a bounded region for which more general assumptions are necessary? Moreover, even if the most general assumption is necessary, is there a case where $\epsilon \neq 1$?

In this work, we will construct a simple example that answers both of these questions affirmatively. As a result, we argue that the more general assumptions (e.g., A5 and A6) are not a special case of the more restrictive assumption (e.g., A1), and merit analyzing SGD under these more general assumptions.

2. Problem formulation & assumptions

The stochastic optimization problem is as follows. We are given a function $f : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ and we want to solve

$$\min_{\theta \in \mathbb{R}^p} \{F(\theta) := \mathbb{E}[f(\theta, X)]\}, \tag{1}$$

where X is a random variable taking value in a measurable space $\mathcal{X}$; and $\mathbb{E}$ is the corresponding expectation operator. Given that SGD is a gradient-based algorithm, we will require that f is differentiable with respect to its first argument with probability one, denoted $\dot{f}(\theta, X)$, which we refer to as stochastic gradients. Moreover, we will keep things simple by requiring that $\mathbb{E}[\dot{f}(\theta, X)] = \dot{F}(\theta)$, where $\dot{F}(\theta)$ is the gradient of F evaluated at θ.[1]

As mentioned, a number of different assumptions about the stochastic gradients, $\dot{f}(\theta, X)$, are made.[2] Letting $\dot{F}(\theta)$ denote the gradient of F evaluated at θ, these assumptions about the stochastic gradients include (in roughly increasing order of generality):

- A1 (Bounded Variance) There exists $C_1 > 0$ such that $\forall \theta \in \mathbb{R}^p$, $\mathbb{E}[\|\dot{f}(\theta, X)\|_2^2] \leq C_1 + \|\dot{F}(\theta)\|_2^2$.
- A2 There exists $C_1 > 0$ and $C_2 > 1$ such that $\forall \theta \in \mathbb{R}^p$, $\mathbb{E}[\|\dot{f}(\theta, X)\|_2^2] \leq C_1 + C_2\|\dot{F}(\theta)\|_2^2$ [1].
- A3 There exists a $C_1 \in \mathbb{R}$, $C_3 > 0$ such that $\forall \theta \in \mathbb{R}^p$, $\mathbb{E}[\|\dot{f}(\theta, X)\|_2^2] \leq C_1 + \|\dot{F}(\theta)\|_2^2 + C_3 F(\theta)$ [3].
- A4 (Expected Smoothness) There exists $C_1 \in \mathbb{R}$, $C_2 > 1$, $C_3 > 0$ such that $\forall \theta \in \mathbb{R}^p$, $\mathbb{E}[\|\dot{f}(\theta, X)\|_2^2] \leq C_1 + C_2\|\dot{F}(\theta)\|_2^2 + C_3 F(\theta)$ [3].

E-mail address: vivak.patel@wisc.edu.

[1] There exist problems for which this assumption needs to be relaxed, and the strategies for doing so are discussed by Bottou et al. [1].

[2] See Patel [2] and Khaled and Richtárik [3] for general discussions of assumptions and their relationships.

[3] This assumption was designed in the context of solving (1) when F is convex.

https://doi.org/10.1016/j.exco.2023.100123
Received 22 April 2022; Received in revised form 13 April 2023; Accepted 16 August 2023
Available online 27 August 2023

A5 There is a non-decreasing function $H : \mathbb{R}_{\geq 0} \to [0, \infty)$ such that $\forall \theta \in \mathbb{R}^p$, $\mathbb{E}[\|\dot{f}(\theta, X)\|_2^2] \leq H(\text{dist}(\theta, \Theta))$, where Θ is the set of solutions to (1) [7].[3]

A6 There is an $\epsilon \in (0, 1]$ and there exists an upper semi-continuous function $G : \mathbb{R}^p \to [0, \infty)$ such that $\forall \theta \in \mathbb{R}^p$, $\mathbb{E}[\|\dot{f}(\theta, X)\|_2^{1+\epsilon}] \leq G(\theta)$ [5].

We will now construct a simple $f(\theta, X)$ such that θ belongs to a finite interval and for which A1 to A4 will fail to hold, while A5 and A6 hold. Then, with a small modification we will show that even A5 and A6 fail to hold if $\epsilon = 1$.

3. Example 1

To construct our example, let $f : [1, e^{1/2}) \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ such that $f(\theta, x) = \theta^x$. Moreover, let X be an exponential random variable with probability distribution function $h(x) = e^{-x}$. We now compute $F(\theta), \dot{F}(\theta), \mathbb{E}[\dot{f}(\theta, X)]$, and $\mathbb{E}[[\dot{f}(\theta, X)]^2]$.

First, $F(\theta) = \mathbb{E}[f(\theta, X)] = \mathbb{E}[e^{X \log(\theta)}]$. Notice, this is just the moment generating function of the exponential random variable with parameter $\log(\theta)$, which exists since $\log(\theta) < 1$ for $\theta \in [1, \exp(1/2))$. Hence, $F(\theta) = (1 - \log(\theta))^{-1}$.

Second, by a direct calculation, $\dot{F}(\theta) = [\theta(1-\log(\theta))^2]^{-1}$. Third, since $\dot{f}(\theta, X) = X\theta^{X-1}$,

$$\mathbb{E}\left[\dot{f}(\theta, X)\right] = \frac{1}{\theta}\int_0^\infty x e^{-(1-\log(\theta))x} dx \tag{2}$$

$$= \frac{1}{\theta(1-\log(\theta))}\int_0^\infty x \frac{e^{-(1-\log(\theta))x}}{(1-\log(\theta))^{-1}} dx. \tag{3}$$

Notice, the last term in the integral is the expected value of an exponential random variable with parameter $1 - \log(\theta)$ which is positive given the interval on which θ exists. It follows that $\mathbb{E}[\dot{f}(\theta, X)] = [\theta(1 - \log(\theta))^2]^{-1}$.

Fourth, $\mathbb{E}[[\dot{f}(\theta, X)]^2] = \mathbb{E}[X^2\theta^{2X-2}]$. Therefore,

$$\mathbb{E}[[\dot{f}(\theta, X)]^2] = \frac{1}{\theta^2}\int_0^\infty x^2 e^{-(1-2\log(\theta))x} dx \tag{4}$$

$$= \frac{1}{\theta^2(1-2\log(\theta))}\int_0^\infty x^2 \frac{e^{-(1-2\log(\theta))x}}{(1-2\log(\theta))^{-1}} dx. \tag{5}$$

The last expression is just the second moment of an exponential random variable with parameter $1 - 2\log(\theta)$, which is positive on the given interval for θ. Hence,

$$\mathbb{E}\left[[\dot{f}(\theta, X)]^2\right] = \frac{2}{\theta^2(1-2\log(\theta))^3}. \tag{6}$$

With these calculations, we collect some facts in the following lemma.

Lemma 1. *The function $F(\theta)$ is minimized at $\theta = 1$ on the interval $[1, e^{1/2})$. The stochastic gradients are unbiased (i.e., $\dot{F}(\theta) = \mathbb{E}[\dot{f}(\theta, X)]$). On the interval $[1, e^{1/2})$, $F(\theta)$ is bounded by 2 and $\dot{F}(\theta)$ is bounded by 4. Finally, $\lim_{\theta \uparrow e^{1/2}} \mathbb{E}[[\dot{f}(\theta, X)]^2] = \infty$.*

From this lemma, we see that $F(\theta)$ and $\dot{F}(\theta)$ are bounded on the interval while the second moment diverges. As a result, we conclude that A1, A2, A3 and A4 fail to hold for this example. Using (6), we can set

$$H(z) = \frac{2}{(z+1)^2(1-2\log(z+1))^3} \tag{7}$$

to see that A5 holds. We can set $G(\theta)$ to (6) to conclude that A6 holds.

4. Example 2

Of course, if we extend the right side of the interval from $[1, e^{1/2})$ of the previous example to something slightly greater than $e^{1/2}$, then the second moment of the stochastic gradient fails to exist, and A1 through A6 will all fail to hold. For a more interesting case, consider choosing an interval $[1, e^{1/(1+\epsilon)})$ for some $\epsilon \in (0, 1)$.

Then, $\mathbb{E}[[\dot{f}(\theta, X)]^{1+\epsilon}] = \theta^{-1-\epsilon}\mathbb{E}[X^{1+\epsilon}\theta^{(1+\epsilon)X}]$. By Young's inequality, $x^{1+\epsilon} \leq x^2(1+\epsilon)/2 + (1-\epsilon)/2$. Applying this inequality,

$$\mathbb{E}\left[[\dot{f}(\theta, X)]^{1+\epsilon}\right] \leq \frac{1+\epsilon}{2\theta^{1+\epsilon}}\mathbb{E}\left[X^2\theta^{(1+\epsilon)X}\right] + \frac{1-\epsilon}{2\theta^{1+\epsilon}}\mathbb{E}\left[\theta^{(1+\epsilon)X}\right] \tag{8}$$

$$= \frac{1+\epsilon}{2\theta^{1+\epsilon}[1-(1+\epsilon)\log(\theta)]^3} + \frac{1-\epsilon}{2\theta^{1+\epsilon}[1-(1+\epsilon)\log(\theta)]}, \tag{9}$$

where we have made use of the same tricks as before to compute the integrals. We see that if we relax the second moment condition, then A6 will still hold by setting $G(\theta)$ equal to (9). Hence, there is value in consider $\epsilon \neq 1$.

5. Conclusion

We considered the noise model assumptions that are commonly used in the analysis of stochastic gradient descent (SGD) (see A1 to A6). We pointed out an argument that raised the two questions:

1. Is there a stochastic optimization problem that is limited to a bounded region for which more general assumptions (A5 and A6) are necessary?
2. If A6 is needed, is there a case where $\epsilon \neq 1$?

In Section 3, we constructed a simple example that answered the first question affirmatively. In Section 4, we extended the aforementioned example to answer the second question affirmatively. Owing to these examples, we showed that A5 and A6 are not simply interesting to analyze for the sake of generality, but that they have utility in realistic problems.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The author is supported by funds from the Wisconsin Alumni Research Foundation, United States. Part of this research was performed while the author was visiting the Institute for Mathematical and Statistical Innovation (IMSI), which is supported by the National Science Foundation, USA (Grant No. DMS-1929348).

References

[1] L. Bottou, F.E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, Siam Rev. 60 (2) (2018) 223–311.
[2] V. Patel, Stopping criteria for, and strong convergence of, stochastic gradient descent on Bottou-Curtis-Nocedal functions, Math. Program. (2021) 1–42.
[3] A. Khaled, P. Richtárik, Better theory for SGD in the nonconvex world, 2020, arXiv preprint arXiv:2002.03329.
[4] V. Patel, S. Zhang, Stochastic gradient descent on nonconvex functions with general noise models, 2021, arXiv preprint arXiv:2104.00423.

[5] V. Patel, B. Tian, S. Zhang, Global convergence and stability of stochastic gradient descent, 2021, arXiv preprint arXiv:2110.01663.
[6] P. Mertikopoulos, N. Hallak, A. Kavis, V. Cevher, On the almost sure convergence of stochastic gradient descent in non-convex problems, 2020, arXiv preprint arXiv:2006.11144.
[7] H. Asi, J.C. Duchi, Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity, SIAM J. Optim. 29 (3) (2019) 2257–2290.
[8] R.M. Gower, O. Sebbouh, N. Loizou, SGD for structured nonconvex functions: Learning rates, minibatching and interpolation, 2020, arXiv preprint arXiv:2006.10311.
[9] T. Ko, X. Li, A local convergence theory for the stochastic gradient descent method in non-convex optimization with non-isolated local minima, 2022, arXiv preprint arXiv:2203.10973.
[10] J. Liu, Y. Yuan, On almost sure convergence rates of stochastic gradient methods, 2022, arXiv preprint arXiv:2202.04295.
[11] C. Fang, Z. Lin, T. Zhang, Sharp analysis for nonconvex sgd escaping from saddle points, in: Conference on Learning Theory, PMLR, 2019, pp. 1192–1234.
[12] C. Jin, P. Netrapalli, R. Ge, S.M. Kakade, M.I. Jordan, On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points, J. ACM 68 (2) (2021) 1–29.