Verifying Assumptions Used in Federated Learning Analysis

Michal Yemini

August 2023

1 Introduction

Many papers that analyze FL algorithms assume a form of a combination of the following three assumptions: 1) unbiasedness of stochastic gradients, 2) boundedness of stochastic gradients, 3) μ-strong convexity of the deterministic function that the stochastic gradients approximate. It seems that this combination is impossible, in what follows we investigate it further.

Note, that it is also assumed in many works that the function to be optimized also has L-Lipschitz continuous gradients, though it is not part of the discussion in this write-up.

1.1 Proving that these assumptions cannot coexist

Let $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$ be ***unbounded*** and let $f(\theta) : \boldsymbol{\Theta} \to \mathbb{R}$ be a differentiable deterministic real function with gradient $\nabla f(\theta)$.

Definition 1 (μ-strongly convex functions)**.** *A differentiable function f is called strongly convex with parameter $\mu > 0$ if the following inequality holds for all points x, y in its domain:*

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \mu \|x - y\|_2^2$$

or, more generally,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$$

where $\langle \cdot, \cdot \rangle$ is any inner product, and $\| \cdot \|$ is the corresponding norm.

Hereafter, $\| \cdot \|$ denotes the L_2 norm. Additionally, we denote by $g(\theta)$ the stochastic gradient of the deterministic function $f(\theta)$.

Furthermore, we note that we use the specific form of Definition 1 for μ strongly functions since it leads more easily to the conclusion that the gradients of f on $\boldsymbol{\Theta}$ generally cannot be bounded. Nonetheless, we can also show using the following version of the definition

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2.$$

Lemma 1. *Suppose that $g(\theta)$ is a real stochastic function of θ, then the following assumptions cannot coexist:*

1. $\mathbb{E}(g(\theta)) = \nabla f(\theta)$ *(this condition is fulfilled, for example, when the samples are chosen randomly and uniformly from the batch), for every* $\theta \in \mathbf{\Theta}$
2. $\mathbb{E}\left(\|g(\theta)\|^2\right) \leq M^2$, *for every* $\theta \in \mathbf{\Theta}$, *(an even stronger assumption is that* $\|g(\theta)\| \leq M$ *since it readily implies that* $E\|g(\theta)\|^2 \leq M^2$*)*
3. f *is* μ*-strongly convex.*

Proof. We prove this lemma by contradiction. First, we use the identity

$$g(\theta) = (g(\theta) - \nabla f(\theta)) + \nabla f(\theta)$$

to show that the first two assumptions imply that the deterministic function f has bounded gradients. However, this contradicts the third assumption of the μ-strong convexity of f.

We start from the first assumption, and then use the unbiasedness of the stochastic gradients g:

$$\begin{aligned}
M^2 &\geq \mathbb{E}\left(\|g(\theta)\|^2\right) \\
&= \mathbb{E}\left(\|(g(\theta) - \nabla f(\theta)) + \nabla f(\theta)\|^2\right) \\
&= \mathbb{E}\left(\|g(\theta) - \nabla f(\theta)\|^2 + 2\nabla f(\theta)^T(g(\theta) - \nabla f(\theta)) + \|\nabla f(\theta)\|^2\right) \\
&= \mathbb{E}\left(\|g(\theta) - \nabla f(\theta)\|^2\right) + 2\nabla f(\theta)^T(\underbrace{\mathbb{E}\left(g(\theta)\right)}_{=\nabla f(\theta)} - \nabla f(\theta)) + \|\nabla f(\theta)\|^2 \\
&= \underbrace{\mathbb{E}\left(\|g(\theta) - \nabla f(\theta)\|^2\right)}_{\geq 0} + \|\nabla f(\theta)\|^2 \\
&\geq \|\nabla f(\theta)\|^2. \qquad (1)
\end{aligned}$$

Thus, $\|\nabla f(\theta)\|^2 \leq M^2$ for every θ. However, such f cannot be a μ-strongly convex function since a μ-strongly convex function cannot have a bounded gradient.[1] Thus, we reached a contradiction.

□

1.2 Extending this contradiction to weaker conditions

We can further show that even if we relax the unbiasedness condition to a bounded one, there is still a persisting issue.

Lemma 2. *Suppose that $g(\theta)$ is a real stochastic function of θ, then the following assumptions cannot coexist*

[1] This can be readily proved by using the Cauchy Schwarz Inequality on the LHS of Definition 1 and fixing y.

1. *There exists* $b \geq 0$ *such that* $\|\mathbb{E}[g(\theta) - \nabla f(\theta)]\| \leq b$ *for every* $\theta \in \boldsymbol{\Theta}$,

2. $\mathbb{E}\left(\|g(\theta)\|^2\right) \leq M^2$ *for every* $\theta \in \boldsymbol{\Theta}$,

3. f *is* μ*-strongly convex.*

Proof. Using the identity

$$g(\theta) = (g(\theta) - \nabla f(\theta)) + \nabla f(\theta),$$

as in the proof of Lemma 1, we have that

$$\begin{aligned}
M^2 &\geq \mathbb{E}\left(\|g(\theta)\|^2\right) \\
&= \mathbb{E}\left(\|(g(\theta) - \nabla f(\theta)) + \nabla f(\theta)\|^2\right) \\
&= \mathbb{E}\left(\|g(\theta) - \nabla f(\theta)\|^2 + 2\nabla f(\theta)^T(g(\theta) - \nabla f(\theta)) + \|\nabla f(\theta)\|^2\right) \\
&= \mathbb{E}\left(\|g(\theta) - \nabla f(\theta)\|^2\right) + 2\nabla f(\theta)^T\mathbb{E}((g(\theta)) - \nabla f(\theta)) + \|\nabla f(\theta)\|^2 \\
&\overset{(a)}{\geq} \|\mathbb{E}\left[g(\theta) - \nabla f(\theta)\right]\|^2 + 2\nabla f(\theta)^T\mathbb{E}((g(\theta)) - \nabla f(\theta)) + \|\nabla f(\theta)\|^2 \\
&\overset{(b)}{\geq} \|\mathbb{E}\left[g(\theta) - \nabla f(\theta)\right]\|^2 - 2\|\nabla f(\theta)\| \cdot \|\mathbb{E}[g(\theta) - \nabla f(\theta)]\| + \|\nabla f(\theta)\|^2 \\
&= \left(\|\nabla f(\theta)\| - \underbrace{\|\mathbb{E}\left[g(\theta) - \nabla f(\theta)\right]\|}_{\triangleq b(\theta)}\right)^2 \\
&\geq \min_{b(\theta)\in[0,b]} \left\{\left(\|\nabla f(\theta)\| - b(\theta)\right)^2\right\} \\
&\overset{(c)}{=} \left(\max\{0, \|\nabla f(\theta)\| - b\}\right)^2, \qquad (2)
\end{aligned}$$

where (a) follows from the convexity of $\|\cdot\|^2$ and Jensen's Inequality, (b) follows from the Cauchy Schwarz Inequality, and (c) follows from the non-negativity of the norm.

We can now conclude the proof by observing that the above inequality yields that for every $\theta \in \boldsymbol{\Theta}$

$$b \geq \|\nabla f(\theta)\| \qquad \text{or} \qquad M^2 \geq \left(\|\nabla f(\theta)\| - b\right)^2.$$

Due to the non-negativity of the norm, this condition is equivalent to:

$$b \geq \|\nabla f(\theta)\| \qquad \text{or} \qquad M + b \geq \|\nabla f(\theta)\|, \ \forall\, \theta \in \boldsymbol{\Theta}$$

which contradicts the μ-strongly convex assumption. □

2 Where the Conditions Come From

Some works cites these two papers in their analysis [1, 2]. However, it seems that the paper [1] relies on an additional projection step and dedicates a discussion to this point and the conditions on the parameter set. Similarly, the work

[2] assumes that the parameter set is *compact* and requires only local strong convexity. Then in Assumption 5 , the work [2] uses the additional assumption for *strong* convexity on the *compact* parameter set, however, since the optimization in Step 1 in their SAVGM algorithm is done over the parameter set their algorithm, the results stand correct.

3 If There is an Issue, How Can it be Fixed?

Best way: Remove the stochastic boundedness assumption and use a different method to prove convergence without this assumption.

An alternative solution: Remove the strong convexity assumption and keep the assumption regarding the bounded gradients (though the family of functions with bounded gradients is not very large).

Quickest way: Assume that the parameter set $\mathbf{\Theta}$ is compact and covnvex, and add a projection step onto the parameter set after each local gradient step. Then, one can use the non-expansiveness of the projection operator. There is a caveat where previous analysis should be carefully checked to make sure that it did not use the fact that the gradient in the optimal point is zero since the global optimal point is not necessarily in the compact and convex set $\mathbf{\Theta}$.

References

[1] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, p. 1571–1578, 2012.

[2] Y. Zhang, J. C. Duchi, and M. J. Wainwright, "Communication-efficient algorithms for statistical optimization," *J. Mach. Learn. Res.*, vol. 14, p. 3321–3363, jan 2013.