Lemma 17 in SCAFFOLD, and its implication for Nonconvex Rates

Durmus Alp Emre Acar
Boston University

Convergence of non-convex loss functions in SCAFFOLD [1] appears to hinge on Lemma 17, and as such, we are unable to verify the validity of the claim. We present a numerical example to point out that we cannot verify whether the left-hand-side is indeed equal to the right-hand-side in Lemma 17. In general it is not clear whether the expressions can be related.

We follow the notation in [1]. We consider two devices ($N = 2$). The devices functions are,

$$f_1(x) = (x-4)^2, \quad f_2(x) = 2(x-6)^2,$$

and the initial server model as $x^0 = 0$. In each round, one device is chosen at random, namely, $S = 1$. We do one gradient descent update in devices ($K = 1$, $\sigma = 0$), and let learning rates be $\eta_g = 0$ and $\eta_l = 0.1$. Let us adopt Option 1 for the updates of states in each round.

We recall definitions in [1] for ease of reference.

- *Active States.* $\mathcal{S}^r$ with cardinality S.
- *Server Model.* $\boldsymbol{x}^r = \frac{1}{S}\sum_{i\in\mathcal{S}^r}\boldsymbol{y}^r_{i,1}$.
- *Active Device Model Update.* $\boldsymbol{y}^r_{i,1} = \boldsymbol{x}^{r-1} - \eta_l\left(\nabla f_i(\boldsymbol{x}^{r-1}) - \boldsymbol{c}_i^{r-1} + \boldsymbol{c}^{r-1}\right)$ and $\boldsymbol{y}^r_{i,0} = \boldsymbol{x}^{r-1}$.
- *Device/Server States.* $\boldsymbol{c}_i^r = \nabla f_i(\boldsymbol{x}^{r-1})$, $\boldsymbol{c}^r = \frac{1}{2}\sum_i \boldsymbol{c}_i^r$
- *Device Parameters.* $\boldsymbol{\alpha}^r_{i,0}$

$$\boldsymbol{\alpha}^r_{i,0} = \begin{cases}\boldsymbol{y}^r_{i,0} & \text{if } i\in\mathcal{S}^r\,, \\ \boldsymbol{\alpha}^{r-1}_{i,0} & \text{otherwise}\,.\end{cases}$$

$\boxed{\text{Lemma 17}}$ claims

$$\Xi_r = \frac{1}{KN}\sum_{i,k}\mathbb{E}\|\boldsymbol{\alpha}^r_{i,k-1} - \boldsymbol{x}^r\|^2 = \left(1 - \frac{S}{N}\right)\cdot \underbrace{\frac{1}{KN}\sum_{k,i}\mathbb{E}\|\boldsymbol{\alpha}^{r-1}_{i,k-1} - \boldsymbol{x}^r\|^2}_{\mathcal{T}_5} + \frac{S}{N}\cdot\underbrace{\frac{1}{KN}\sum_{k,i}\mathbb{E}\|\boldsymbol{y}^r_{i,k-1} - \boldsymbol{x}^r\|^2}_{\mathcal{T}_6}\,.$$

where, the expectation is with respect to the randomness in device selection.

Summary of our Computation. We compute the LHS and RHS for the first two rounds. For the the first two rounds, largely due to initialization, the LHS is equal to RHS. On the third round we show that LHS is actually greater than RHS. It is not hard to find other examples where the opposite is true. Putting these together, we find that LHS and RHS cannot be related. As such, without this Lemma, we are unable to validity of the main theorem.

We rewrite the statement in our simple case,

$$\Xi_r = \frac{1}{2}\sum_i \mathbb{E}\|\boldsymbol{\alpha}^r_{i,0} - \boldsymbol{x}^r\|^2 = \frac{1}{2}\cdot\underbrace{\frac{1}{2}\sum_i\mathbb{E}\|\boldsymbol{\alpha}^{r-1}_{i,0} - \boldsymbol{x}^r\|^2}_{\mathcal{T}_5} + \frac{1}{2}\cdot\underbrace{\frac{1}{2}\sum_i\mathbb{E}\|\boldsymbol{y}^r_{i,0} - \boldsymbol{x}^r\|^2}_{\mathcal{T}_6}\,. \tag{1}$$

Let us note the initial parameters of SCAFFOLD algorithm for reference,

Table 1: Realizations of variables for round three.

	$\mathcal{S}^1=\{1\}$	$\mathcal{S}^1=\{1\}$	$\mathcal{S}^1=\{1\}$	$\mathcal{S}^1=\{1\}$	$\mathcal{S}^1=\{2\}$	$\mathcal{S}^1=\{2\}$	$\mathcal{S}^1=\{2\}$	$\mathcal{S}^1=\{2\}$
	$\mathcal{S}^2=\{1\}$	$\mathcal{S}^2=\{1\}$	$\mathcal{S}^2=\{2\}$	$\mathcal{S}^2=\{2\}$	$\mathcal{S}^2=\{1\}$	$\mathcal{S}^2=\{1\}$	$\mathcal{S}^2=\{2\}$	$\mathcal{S}^2=\{2\}$
	$\mathcal{S}^3=\{1\}$	$\mathcal{S}^3=\{2\}$	$\mathcal{S}^3=\{1\}$	$\mathcal{S}^3=\{2\}$	$\mathcal{S}^3=\{1\}$	$\mathcal{S}^3=\{2\}$	$\mathcal{S}^3=\{1\}$	$\mathcal{S}^3=\{2\}$
c_1^1	−8.000	−8.000	−8.000	−8.000	0.000	0.000	0.000	0.000
$y_{1,0}^2$	0.800	0.800	0.800	0.800	2.400	2.400	2.400	2.400
$\alpha_{1,0}^2$	0.800	0.800	0.000	0.000	2.400	2.400	0.000	0.000
$y_{1,1}^2$	1.040	1.040	1.040	1.040	3.920	3.920	3.920	3.920
c_1^2	−6.400	−6.400	−8.000	−8.000	−3.200	−3.200	0.000	0.000
$y_{1,0}^3$	1.040	1.040	3.280	3.280	3.920	3.920	2.640	2.640
$\alpha_{1,0}^3$	1.040	0.800	3.280	0.000	3.920	2.400	2.640	0.000
$y_{1,1}^3$	1.312	1.312	4.064	4.064	4.976	4.976	3.632	3.632
c_2^1	0.000	0.000	0.000	0.000	−24.000	−24.000	−24.000	−24.000
$y_{2,0}^2$	0.800	0.800	0.800	0.800	2.400	2.400	2.400	2.400
$\alpha_{2,0}^2$	0.000	0.000	0.800	0.800	0.000	0.000	2.400	2.400
$y_{2,1}^2$	3.280	3.280	3.280	3.280	2.640	2.640	2.640	2.640
c_2^2	0.000	0.000	−20.800	−20.800	−24.000	−24.000	−14.400	−14.400
$y_{2,0}^3$	1.040	1.040	3.280	3.280	3.920	3.920	2.640	2.640
$\alpha_{2,0}^3$	0.000	1.040	0.800	3.280	0.000	3.920	2.400	2.640
$y_{2,1}^3$	3.344	3.344	3.728	3.728	3.712	3.712	3.264	3.264
x^1	0.800	0.800	0.800	0.800	2.400	2.400	2.400	2.400
c^1	−4.000	−4.000	−4.000	−4.000	−12.000	−12.000	−12.000	−12.000
x^2	1.040	1.040	3.280	3.280	3.920	3.920	2.640	2.640
c^2	−3.200	−3.200	−14.400	−14.400	−13.600	−13.600	−7.200	−7.200
x^3	1.312	3.344	4.064	3.728	4.976	3.712	3.632	3.264

- $y_{1,0}^1 = y_{2,0}^1 = \alpha_{1,0}^0 = \alpha_{2,0}^0 = c_1^0 = c_2^0 = c^0 = x^0 = 0$,
- $y_{1,1}^1 = x^0 - \eta_l(\nabla f_1(x^0) - c_1^0 + c^0) = 0.8$,
- $y_{2,1}^1 = x^0 - \eta_l(\nabla f_2(x^0) - c_2^0 + c^0) = 2.4$.

Let us consider round three. We have the events as $\left(\mathcal{S}^1 = \{i\}, \mathcal{S}^2 = \{j\}, \mathcal{S}^3 = \{k\}\right)$ with equal probability. We give all possible realizations in Table 1.

Now we can explicitly write the expectations of LHS and RHS in Eq. 1 at $r = 3$.

LHS becomes,

$$\begin{aligned}
\Xi_3 =& \frac{1}{2}\sum_i \mathbb{E}\|\boldsymbol{\alpha}_{i,0}^3 - \boldsymbol{x}^3\|^2 = \frac{1}{8}\sum_{j,k,l}\left(\frac{1}{2}\sum_i \mathbb{E}\left[\|\boldsymbol{\alpha}_{i,0}^3 - \boldsymbol{x}^3\|^2 | \mathcal{S}^1 = \{j\}, \mathcal{S}^2 = \{k\}, \mathcal{S}^3 = \{l\}\right]\right) \\
=& \frac{1}{8}\left(\frac{1}{2}(1.040 - 1.312)^2 + \frac{1}{2}(0.000 - 1.312)^2\right) + \frac{1}{8}\left(\frac{1}{2}(0.800 - 3.344)^2 + \frac{1}{2}(1.040 - 3.344)^2\right) \\
&+ \frac{1}{8}\left(\frac{1}{2}(3.280 - 4.064)^2 + \frac{1}{2}(0.800 - 4.064)^2\right) + \frac{1}{8}\left(\frac{1}{2}(0.000 - 3.728)^2 + \frac{1}{2}(3.280 - 3.728)^2\right) \\
&+ \frac{1}{8}\left(\frac{1}{2}(3.920 - 4.976)^2 + \frac{1}{2}(0.000 - 4.976)^2\right) + \frac{1}{8}\left(\frac{1}{2}(2.400 - 3.712)^2 + \frac{1}{2}(3.920 - 3.712)^2\right) \\
&+ \frac{1}{8}\left(\frac{1}{2}(2.640 - 3.632)^2 + \frac{1}{2}(2.400 - 3.632)^2\right) + \frac{1}{8}\left(\frac{1}{2}(0.000 - 3.264)^2 + \frac{1}{2}(2.640 - 3.264)^2\right) \\
=& 5.008
\end{aligned}$$

Similarly, let's find RHS by calculating $\mathcal{T}_5$ and $\mathcal{T}_6$,

$$
\begin{aligned}
\mathcal{T}_5 =& \frac{1}{2}\sum_i \mathbb{E}\|\boldsymbol{\alpha}_{i,0}^2 - \boldsymbol{x}^3\|^2 = \frac{1}{8}\sum_{j,k,l}\left(\frac{1}{2}\sum_i \mathbb{E}\left[\|\boldsymbol{\alpha}_{i,0}^2 - \boldsymbol{x}^3\|^2 | \mathcal{S}^1 = \{j\}, \mathcal{S}^2 = \{k\}, \mathcal{S}^3 = \{l\}\right]\right) \\
=& \frac{1}{8}\left(\frac{1}{2}(0.800 - 1.312)^2 + \frac{1}{2}(0.000 - 1.312)^2\right) + \frac{1}{8}\left(\frac{1}{2}(0.800 - 3.344)^2 + \frac{1}{2}(0.000 - 3.344)^2\right) \\
&+ \frac{1}{8}\left(\frac{1}{2}(0.000 - 4.064)^2 + \frac{1}{2}(0.800 - 4.064)^2\right) + \frac{1}{8}\left(\frac{1}{2}(0.000 - 3.728)^2 + \frac{1}{2}(0.800 - 3.728)^2\right) \\
&+ \frac{1}{8}\left(\frac{1}{2}(2.400 - 4.976)^2 + \frac{1}{2}(0.000 - 4.976)^2\right) + \frac{1}{8}\left(\frac{1}{2}(2.400 - 3.712)^2 + \frac{1}{2}(0.000 - 3.712)^2\right) \\
&+ \frac{1}{8}\left(\frac{1}{2}(0.000 - 3.632)^2 + \frac{1}{2}(2.400 - 3.632)^2\right) + \frac{1}{8}\left(\frac{1}{2}(0.000 - 3.264)^2 + \frac{1}{2}(2.400 - 3.264)^2\right) \\
=& 8.8928
\end{aligned}
$$

and

$$
\begin{aligned}
\mathcal{T}_6 =& \frac{1}{2}\sum_i \mathbb{E}\|\boldsymbol{y}_{i,0}^3 - \boldsymbol{x}^3\|^2 = \frac{1}{8}\sum_{j,k,l}\left(\frac{1}{2}\sum_i \mathbb{E}\left[\|\boldsymbol{y}_{i,0}^3 - \boldsymbol{x}^3\|^2 | \mathcal{S}^1 = \{j\}, \mathcal{S}^2 = \{k\}, \mathcal{S}^3 = \{l\}\right]\right) \\
=& \frac{1}{8}\left(\frac{1}{2}(1.040 - 1.312)^2 + \frac{1}{2}(1.040 - 1.312)^2\right) + \frac{1}{8}\left(\frac{1}{2}(1.040 - 3.344)^2 + \frac{1}{2}(1.040 - 3.344)^2\right) \\
&+ \frac{1}{8}\left(\frac{1}{2}(3.280 - 4.064)^2 + \frac{1}{2}(3.280 - 4.064)^2\right) + \frac{1}{8}\left(\frac{1}{2}(3.280 - 3.728)^2 + \frac{1}{2}(3.280 - 3.728)^2\right) \\
&+ \frac{1}{8}\left(\frac{1}{2}(3.920 - 4.976)^2 + \frac{1}{2}(3.920 - 4.976)^2\right) + \frac{1}{8}\left(\frac{1}{2}(3.920 - 3.712)^2 + \frac{1}{2}(3.920 - 3.712)^2\right) \\
&+ \frac{1}{8}\left(\frac{1}{2}(2.640 - 3.632)^2 + \frac{1}{2}(2.640 - 3.632)^2\right) + \frac{1}{8}\left(\frac{1}{2}(2.640 - 3.264)^2 + \frac{1}{2}(2.640 - 3.264)^2\right) \\
=& 1.0912
\end{aligned}
$$

Plugging the values we get $\text{LHS} = 5.008, \text{RHS} = \frac{1}{2}\mathcal{T}_5 + \frac{1}{2}\mathcal{T}_6 = 4.992$ in Eq. 1. Indeed, LHS is greater than RHS in this example.

The reason for this inconsistency appears to arise from the fact that the server model is a function of the active device set. Hence, we should explicitly write the conditional expectations in $\mathcal{T}_5$ and $\mathcal{T}_6$ splits. In general, we should write,

$$
\begin{aligned}
\Xi_r &= \frac{1}{KN}\sum_{i,k}\mathbb{E}\|\boldsymbol{\alpha}_{i,k-1}^r - \boldsymbol{x}^r\|^2 \\
&= \left(1 - \frac{S}{N}\right)\cdot \underbrace{\frac{1}{KN}\sum_{k,i}\mathbb{E}\left[\mathbb{E}_{r-1}\|\boldsymbol{\alpha}_{i,k-1}^{r-1} - \boldsymbol{x}^r\|^2 | i \notin \mathcal{S}^r\right]}_{\mathcal{T}_5} + \frac{S}{N}\cdot \underbrace{\frac{1}{KN}\sum_{k,i}\mathbb{E}\left[\mathbb{E}_{r-1}\|\boldsymbol{y}_{i,k-1}^r - \boldsymbol{x}^r\|^2 | i \in \mathcal{S}^r\right]}_{\mathcal{T}_6}.
\end{aligned}
$$

With this observation in place, it is not clear whether the expressions can be further simplified, namely, the conditioning could possibly be removed.

Bounding with Device Independent Constants. Another possibility that we examined was whether one could bound RHS with some sufficiently large constant. However, apart from scaling with number of devices, we were unable to find tighter bounds. Also, scaling the functions generally results in scaling the difference between LHS and RHS. For instance, let us scale functions by 100. We have $f_1(x) = 10(x-4)^2$, $f_2(x) = 20(x-6)^2$. If we use the same learning rate as 0.1, we get LHS-RHS=376.

References

[1] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for on-device federated learning. *CoRR*, abs/1910.06378, 2019.