Advanced Topics in Machine Learning 2024

Nirupam Gupta

Assignment 6: Introduction to Federated Learning

Deadline: 23:59, Thursday, 31 October 2024

The assignments must be submitted individually – each student must write and submit a personal solution but we do not prevent you from discussing high-level ideas in small groups. If you use any LLM tool such as ChatGPT, please specify the purpose and manner in which you have utilized it.

We are interested in how *you solved the problems, and not just in the final answers. Please explain your solutions and ideas as clearly as you can.*

Late Penalty and multiple Submissions *Late submissions will incur a penalty of 10% of the total marks for every hour of delay (rounded up) with a maximum allowed delay of 5 hours after which the submission server will close. If you submit multiple submissions, only the last submission will be considered relevant both for grading answers as well as late penalty.*

Submission format: *Please upload your answers in a single* `.pdf` *file. If you have created code please include at least the key parts in the PDF as well as a link to the full code (on Github, Colab, or similar).*

Learning points: *The goal of this assignment is to provide a better understanding of the concepts covered in the class. It will help you grasp additional details that were skipped during the lecture.*

1. **Peer-to-peer DGD.** In the class, we looked at the *distributed gradient descent* (DGD) method in a peer-to-peer setting. We studied the growth of the model drifts. We showed that the local models maintained by each client approach each other when the underlying communication topology $\mathcal{G} = (\mathcal{V} = [n], \mathcal{E})$ is connected, provided the gossip rate and local learning rate, i.e., α and γ, are small enough. This analysis proves useful in obtaining a convergence guarantee of the overall algorithm. We assumed that for each client i, local loss function $\mathcal{L}_i(\theta)$ is λ-Lipschitz smooth over $\Theta = \Re^d$. We also assumed that there exists $\zeta \in \Re$ such that for all $\theta \in \Re^d$, the *gradient dissimilarity* is uniformly bounded as follows:

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla \mathcal{L}_i(\theta) - \nabla \mathcal{L}(\theta)\|^2 \leq \zeta^2,$$

where $\mathcal{L}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta)$.

Question 1 (25 points)**.** Complete the convergence proof, showing that there exits $\gamma \in \Re$ such that if $\gamma_t = \gamma$ for all $t \in [T] = \{1, \ldots, T\}$ then

$$\frac{1}{T}\sum_{t\in[T]} \|\nabla \mathcal{L}(\theta_t)\|^2 \in \mathcal{O}\left(\frac{\zeta^2}{T^{2/3}} + \frac{\rho^2\lambda}{T}\right). \tag{1}$$

Recall that $\rho = 1 - \alpha\mu_{\mathcal{G}}$, where α is gossip rate and $\mu_{\mathcal{G}}$ denotes the *algebraic connectivity* of $\mathcal{G}$.

Question 2 (25 points)**.** Empirically check the dependence of the error in (1) on (a) *algebraic connectivity* and (b) *gradient dissimilarity*. While you are free to choose the learning problem, it suffices to consider a squared-loss minimization problem wherein, for each client i, let $\mathcal{L}_i(\theta) = \frac{1}{2}\|\theta - \theta_i\|^2$ where $\theta \in \Re^2$ and θ_i is fixed point in $\Re^2$. (You can also consider the scalar case). Consider a system of at least $n = 4$ clients.
Since the global loss function in this case is *strongly convex*, you can plot actual values of the global loss function (or the error with respect to the minimum value) over the different iterations $t = 1, 2, \ldots$ to see the impact of the two parameters (a) and (b).

Question 3 (25 points)**.** Consider the following generalization of the gossip rule discussed in the class. Specifically, for all $i \in [n] = \{1, \ldots, n\}$ and t,

$$\theta^{(i)}_{t+\frac{1}{2}} = \sum_{j \in [n]} w_{ij}\ \theta^{(j)}_t,$$

where

$$w_{ij} = \begin{cases} > 0 & , \quad (i,j) \in \mathcal{E} \text{ (with } (i,i) \in \mathcal{E}) \\ 0 & , \quad \text{otherwise} \end{cases}.$$

To preserve symmetry in the weighted communication topology, let $w_{ij} = w_{ji}$. Recall that $\theta_t := \frac{1}{n}\sum_{i=1}^n \theta^{(i)}_t$, and $\Gamma\left(\theta^{(1)}_t, \ldots, \theta^{(n)}_t\right) = \frac{1}{n}\sum_{i=1}^n \left\|\theta^{(i)}_t - \theta_t\right\|^2$, for all t.

- What is a sufficient condition on the weights w_{ij}'s such that, for all t,

$$\Gamma\left(\theta^{(1)}_{t+\frac{1}{2}}, \ldots, \theta^{(n)}_{t+\frac{1}{2}}\right) \leq c\,\Gamma\left(\theta^{(1)}_t, \ldots, \theta^{(n)}_t\right),$$

  where $c \in [0, 1)$?
- If we additionally desire that $\theta_{t+\frac{1}{2}} = \theta_t$, how would the above condition change?
- How does the convergence result, shown in (1), change according to the above gossip rule?
- **(10 Bonus points)** Design a convex programming problem solving which yields the smallest value for c, subject to the above constraints.

2. **Distributed stochastic gradient descent (DSGD).** Consider the DSGD method in the server-based architecture, presented in the class, under synchronicity. For each client i, let $\sigma_i = \sigma$.

Question 4 (15 points)**.** Show that there exists $\gamma \in \Re$ such that if $\gamma_t = \gamma$ for all $t \in [T]$ then

$$\mathbb{E}\left[\frac{1}{T}\sum_{t \in [T]} \|\nabla \mathcal{L}(\theta_t)\|^2\right] \in \mathcal{O}\left(\sqrt{\frac{\sigma^2}{nT}}\right), \tag{2}$$

where $\mathbb{E}[\cdot]$ denotes the expectation over the randomness in the algorithm.

Question 5 (10 points)**.** Compare (2) with the convergence rate of the *local SGD* method, which was analyzed in the class, in the special case when the number of local steps $\tau = 1$. Which of the two convergence rates is tighter? How can we fix the gap?