Beyond Mean: a Principled Approach for Robust Aggregation

Aymeric Dieuleveut, École Polytechnique, Paris
Hadrien Hendrikx, INRIA Grenoble

We are looking for exceptional candidates for an internship and PhD thesis, jointly supervised by Aymeric Dieuleveut and Hadrien Hendrikx, (École Polytechnique and Inria Grenoble), on the topic of robust aggregation. Competitive funding for the PhD is already available (with travel support, etc.).

The internship and PhD will focus on decentralized optimization, with a theoretically focused approach. The objective is to understand and improve stochastic learning algorithms in the decentralized context, with different constraints.

This direction of research is very dynamic, offers great possibilities and results in numerous applications. We briefly describe the topic hereafter.

Feel free to contact us for questions of applications (with a CV and transript): Aymeric Dieuleveut aymeric.dieuleveut@gmail.com, & Hadrien Hendrikx hadrien.hendrikx@epfl.ch.

1 Introduction

In the vast majority of distributed algorithms, nodes collaborate by averaging their parameters. In decentralized optimization, nodes rely on so-called 'gossip' communications, *i.e.*, frequent approximate averaging through multiplication by a gossip matrix W. In Federated Averaging, nodes perform infrequent full averaging of their parameters. However, averaging models might not always be the best way to communicate, for instance:

- In byzantine-robust optimization, when nodes actively try to perturb optimization. The mean is a brittle estimate in this case, as it can be arbitrarily changed by only one participant.
- In non-convex optimization, where different nodes might be in different modes. In this case, averaging two good models can lead to a bad one.
- In privacy-preserving optimization, where less sensitive aggregation procedures can lead to better privacy guarantees.
- To reduce the communication cost. Most current approaches are based on compression + averaging, but different procedures might be more efficient.

Ad-hoc solutions have been developed for each problem, involving for instance computing medians instead of means, or clipping the individual gradients before averaging them. However, the convergence guarantees of these solutions are not always very clear, especially when it comes down to precisely evaluating convergence speed, or speeding up convergence using standard tricks on top.

The goal of this project is to obtain new communication procedures through a principled approach. This will then pave the way for more robust distributed algorithms, that leverage the power of standard optimization analyses out of the box.

2 Dual approach

(Primal-)Dual methods are a very efficient way to build decentralized algorithms, but their standard formulation leads to 'gossip communications' . Thus, a natural idea is to modify this framework to obtain similar algorithms with different guarantees. In standard problems, we write:

$$\min_{x\in\mathbb{R}^d}\sum_{i=1}^n f_i(x) \Leftrightarrow \min_{X\in\mathbb{R}^{n\times d}, X_i=X_j, i\sim j}\sum_{i=1}^n f_i(X_i)$$

Then, we dualize the right part and we apply standard algorithms on the resulting problem. In the case of robust optimization, we probably don't want a strict equivalence, because solving the initial problem exactly can lead to arbitrary results in the presence of an omniscient adversary (that could for instance send gradients of $f_0 = g - \sum_{i>0} f_i$. Thus, the first step is to choose the right reformulation. Some natural ones that come to mind are the 'f-robust problem':

$$\min_{X\in\mathbb{R}^{n\times d}, \sum_{j\sim i}\mathbb{1}\{X_i\neq X_j\}\leq f}\sum_{i=1}^n f_i(x), \tag{1}$$

in which each node is allowed to disagree with at most f neighbors. A more subtle formulation is the median one:

$$\min_{X\in\mathbb{R}^{n\times d}, X_i=\mathrm{med}(\{X_j, i\sim j\})}\sum_{i=1}^n f_i(x). \tag{2}$$

In this case, the constraints are again satisfied by the solution to the global problem, but there might be some other solutions that have lower error. Note that in the previous formulations, it might be better to state that some nodes do not need to agree with anyone (so that some arbitrarily different node does not change the objective too much). Other constraint sets that might be suited to different cases are $\{d(X_i, X_j)\leq\varepsilon\}$ for some norm (edge-version), or the $\{\sum_{j\sim i} d(X_i, X_j)\leq\varepsilon\}$ (node-version). Note that the latter generalizes the f-robust problem, that is obtained by taking $d(x, y) = \|x - y\|_0$ and $\varepsilon = f$.

When we change the problem in this way, the first thing we need to think about is **what is the new solution? Does it have desirable properties?**.

3 Deriving new algorithms

From these new problems, we would like to derive new algorithms. We detail in this section the standard dual approach [SBB+17, HBM21]. In particular, we can rewrite those as:

$$\min_{X\in\mathbb{R}^{n\times d}} F(X) + \delta(X). \tag{3}$$

Then, Lagrangian multipliers can be introduced for the linear constraints. Minimizing over X, we obtain the convex conjugate of the initial objective function. More specifically, we rewrite the problem as:

$$\min_{X\in\mathbb{R}^{n\times d}}\max_{Y\in\mathbb{R}^{n\times d}} \delta(X) + X^\top Y - F^*(Y) \Leftrightarrow \max_{Y\in\mathbb{R}^{n\times d}} -\delta^*(-Y) - F^*(Y) \tag{4}$$

The standard dual formulation is obtained by using the fact that if $\delta(X) = \mathbb{1}\{A^\top X = 0\}$, then $\delta^*(-Y)$ finite implies that $Y = A\lambda$ for some λ. In particular, the dual problem becomes:

$$\min_{\lambda\in\mathbb{R}^{n\times d}} F^*(A\lambda), \tag{5}$$

which is very convenient because taking gradients of this objective has a direct decentralized interpretation. Then we can leverage standard tools such as acceleration, splitting etc...

The main question for us in this case is to **evaluate δ^*, and see if it has some nice properties (we can 'easily' take gradients/prox for instance)**. We can also consider $\delta + \alpha\|\cdot\|^2$ by transfering some of the strong convexity from F if it helps (does this also make sense in the standard formulation?).

4 The project

The main challenge of the project now comes down to finding the right formulation and in particular the right constraint indicator δ to balance the two aspects presented above:

- The solutions of the problem have 'good properties' (robustness for instance)
- Efficient algorithms can be derived to solve it (because we know how to compute prox or gradients of δ or δ^* for instance).

In the standard case, the 'good property' is to not change the minimum, and the efficient algorithm comes from the fact that the dual objective has a very simple form amenable to decentralized optimization. **Now, let's be creative and find out others!**

We will first focus on two main problems:

- (Byzantine) Robustness, in which we want to be robust against a set of nodes that would behave arbitrarily [BEMGS17].
- Non-convex optimization, in which we would like aggregators that make sense even though agents end up at different modes. A first step could be to study diagonal linear networks [VKR19, PPVF21].

5 Material Conditions

The internship is expected to lead to a Ph.D. thesis, starting September 2023, for which funding is available, as well as support funding for travel and missions.

The thesis will be co-supervised by Aymeric Dieuleveut and Hadrien Hendrikx, and will be hosted at École Polytechnique, CMAP, with the possibility to spend time at Grenoble on a regular basis.

École Polytechnique offers competitive Ph.D. salaries and an extremely dynamic environment for research. During the course of the thesis, an internship abroad may be organized, e.g., at EPFL or in the US.

References

[BEMGS17] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

[HBM21] Hadrien Hendrikx, Francis Bach, and Laurent Massoulie. An optimal algorithm for decentralized finite-sum optimization. *SIAM Journal on Optimization*, 31(4):2753–2783, 2021.

[PPVF21] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

[SBB+17] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *international conference on machine learning*, pages 3027–3036. PMLR, 2017.

[VKR19] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.