Byzantine Collaborative Filtering: Securely Listening to the Silent Majority

Lê-Nguyên Hoang[1]

[1]Tournesol Association

Abstract

This paper aims to leverage the covariances of different contributors' preferences to perform Byzantine collaborative filtering. Interestingly, we show that this also allows to "listen to the silent majority", thereby correcting for *activity bias* in a secure way. We believe that our results represent a major milestone in the quest for fair, ethical and secure algorithms.

1 Introduction

Today's most influential learning algorithms rely on massive amounts of user-generated data. However, because some users are much more active on social medias than others, this creates a very concerning algorithmic *activity bias*. Namely, algorithms are shaped by users that engage the most in online debates. Such users are unlikely to be representative of all users; as a results, today's algorithms fail to account for the actual diversity of online users. New solutions are needed to "listen to the silent majority".

The combination of collaborative filtering and personalized federated learning seems to yield a promising framework to precisely this. Intuitively, this will correspond to amplifying the voice of those who seem to speak in the name of a large silent majority, which is underrepresented in the set of active users.

However, such a system may be especially vulnerable to Byzantine attacks. Namely, disinformation campaigns may create fake accounts, which claim to belong to underrepresented groups. In this paper, we propose the first voting mechanism that corrects for *activity bias*, while providing security guarantees against Byzantine attacks.

Licchavi is a framework introduced by [FGH21].

2 The quadratically regularized geometric median primitives

In this section, we introduce two key primitives, which have very desirable properties to guarantee fairness and security. Their construction builds upon the "one person, one unit force" principle [EFGH21, FGH21], while also accounting for uncertainty and strong Byzantine resilience guarantees [AGHV22].

2.1 Quadratically regularized geometric median for vectors

Consider a set $[N] \triangleq \{1, \dots, N\}$ of users, each with a given voting right $w_n \geq 0$ and an unknown preferred vector $x_n \in \mathbb{R}^d$. Suppose that our knowledge of this preferred vector is best described by a (Bayesian) prior $\tilde{x}_n$ over $\mathbb{R}^d$. Then, for any voting resilience hyperparameter $W \geq 0$, we define the quadratically regularized geometric median of the users' preferred vectors by

$$\textsc{QrGm}_{W,x_0}(w_{1:N}, \tilde{x}_{1:N}) \triangleq \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \left\{ \frac{1}{2} W \left\| x - x_0 \right\|_2^2 + \sum_{n \in [N]} w_n \mathbb{E}_{x_n \leftarrow \tilde{x}_n} \left\| x - x_n \right\|_2 \right\}. \tag{1}$$

Clearly, for $W = 0$ and deterministic priors $\tilde{x}_n$, we retrieve the classical geometric median. However, intuitively, $\textsc{QrGm}_W$ has a more Bayesian flavor. On one hand, it allows to consider non-deterministic priors $\tilde{x}_n$. On the other hand, the term $\frac{1}{2} W \left\| x - x_0 \right\|_2^2$ can be regarded as a prior on the geometric median, which we assume to be more likely to be close to x_0 a priori. Interestingly, this term also guarantees the existence and uniqueness of the quadratically regularized geometric median.

Proposition 1. *For any voting resilience $W > 0$, voting rights $w_{1:N}$ and priors $\tilde{x}_{1:N}$, the quadratically regularized geometric median $\textsc{QrGm}_{W,x_0}(w_{1:N}, \tilde{x}_{1:N})$ is well-defined.*

Proof. The regularization $\frac{1}{2} W \left\| x - x_0 \right\|_2^2$ guarantees W-strong convexity, which implies the existence and uniqueness of a minimum. □

More importantly, QrGm guarantees an interesting form of Byzantine resilience, which was first studied by [AGHV22].

Conjecture 1 (99% credence in truth and our capability to prove it)**.** *For any $W \geq 0$ and $x_0 \in \mathbb{R}^d$, $\textsc{QrGm}_{W,x_0}$ is W-Byzantine resilient. More precisely, for any subset $F \subset [N]$, then ignoring the inputs from such users only moves $\textsc{QrGm}_{W,x_0}$ by at most their cumulative voting rights divided by the voting resilience W, i.e.*

$$\left\| \textsc{QrGm}_{W,x_0}(w_{1:N}, \tilde{x}_{1:N}) - \textsc{QrGm}_{W,x_0}(w_{[N]-F}, \tilde{x}_{[N]-F}) \right\|_2 \leq \frac{1}{W} \sum_{f \in F} w_f. \tag{2}$$

Another desirable property of QrGm is that it belongs to the convex hull of its inputs. More precisely, define $\textsc{Supp}(\tilde{x})$ the support of the distribution $\tilde{x}$, and $\textsc{Hull}(x_1, \dots, x_N)$ the convex hull of its input vectors. We also generalize straightforwadly Hull to inputs that are subsets of vectors. Then the following holds.

Conjecture 2 (99% credence in truth and our capability to prove it)**.** *For any $W \geq 0$ and $x_0 \in \mathbb{R}^d$, $\textsc{QrGm}_{W,x_0}$ belongs to the convex hull of x_0 and of the supports of its input vector distributions, i.e.,*

$$\textsc{QrGm}_{W,x_0}(w_{1:N}, \tilde{x}_{1:N}) \in \textsc{Hull}\left(x_0, \textsc{Supp}(\tilde{x}_1), \dots, \textsc{Supp}(\tilde{x}_N)\right). \tag{3}$$

2.2 Quadratically regularized geometric median for matrices

Note that $\textsc{QrGm}_W$ can be generalized to matrix distributions $\tilde{S}_{1:N}$ instead of vector distributions $\tilde{\theta}_{1:N}$, by using the Frobenius norm instead of the Euclidean norm. Recall that the Frobenius norm

is given by

$$\|S\|_{\text{FROB}}^2 \triangleq \text{TR}\ SS^T = \sum_{i=1}^{d}\sum_{j=1}^{d} S_{ij}^2. \tag{4}$$

The quadratically regularized geometric median for matrices is then given by

$$\text{QRGM}_{W,S_0}(w_{1:N}, \tilde{S}_{1:N}) \triangleq \operatorname*{arg\,min}_{S\in\mathbb{R}^{d\times d}} \left\{ \frac{1}{2} W \|S - S_0\|_{\text{FROB}}^2 + \sum_{n\in[N]} w_n \mathbb{E}_{S_n\leftarrow\tilde{S}_n} \|S - S_n\|_{\text{FROB}} \right\}. \tag{5}$$

Evidently, the lemmas of the previous section still apply to this operator. Interestingly, this has the following consequence.

Conjecture 3. *Suppose S_0 and all matrices S_n's are positive semi-definite with probability 1. Then $\text{QRGM}_{W,S_0}(w_{1:N}, \tilde{S}_{1:N})$ is also positive semi-definite.*

3 The covariant Licchavi algorithm

As earlier, we assume that each user $n \in [N]$ has an unknown preferred model $\theta_n \in \mathbb{R}^d$. However, we now consider that each user provides a signed dataset $\mathcal{D}_n$ for $n \in [N]$ which partially reveals this preferred model. This dataset may provide personal socio-democraphic data about the user, or a set of answers that the user provided to different queries. The goal of collaborative filtering is to determine a posterior distribution $\tilde{\theta}_n$ on θ_n, given not only the user's data $\mathcal{D}_n$, but also given other users' data $\mathcal{D}_m$, for $m \neq n$. However, we want to achieve this in a Byzantine-resilient manner, meaning that $\tilde{\theta}_n^{\text{CL}}$ should remain reasonable, despite the presence of a subset $F \subset [N]$ of malicious users. In particular, $\tilde{\theta}_n^{\text{CL}}$ should not be too distant from what it would have been, if the subset F of malicious users had been identified and discarded.

In this section, we introduce a new algorithm to achieve this, called CovariantLicchavi (in short, CL). CovariantLicchavi is composed of three stages. First, a local posterior $\tilde{\theta}_n^{\text{LOCAL}}$ is determined for each user $n \in [N]$, based solely on the user's local dataset $\mathcal{D}_n$. Second, the primitives introduced in Section 2 are leveraged to determine a Byzantine-resilient covariant matrix Σ^{CL}, which determines correlations between different coordinates of users' preferred models. Third, and finally, the Licchavi algorithm [FGH21] is adapted, to leverage this covariant matrix and guarantee collaborative filtering, thereby computing CovariantLicchavi posteriors $\tilde{\theta}_n^{\text{CL}}$ for all users $n \in [N]$. Note that a fourth step may be added, to perform Byzantine-resilient Bayesian voting, by simply also outputting the global model $\tilde{\rho}^{\text{CL}}$ computed by our adaptation of Licchavi.

3.1 Purely local models

In principle, a user n's purely local distribution $\tilde{\theta}_n^{\text{LOCAL}}$ can be obtained by consider a prior distribution $\tilde{\theta}_n^{\text{PRIOR}}$, and by conditioning it to the dataset $\mathcal{D}_n$. In practice, however, this operation may be computationally intractable. There may then be different ways to recover an approximation of the posterior given dataset $\mathcal{D}_n$ (or at least, to sample from this posterior distribution).

The basic solution is to consider a local loss function $\mathcal{L}^{\text{LOCAL}}(\theta_n|\mathcal{D}_n)$, which may be defined as the negative log-posterior on θ_n, given $\mathcal{D}_n$. A minimum θ_n^{LOCAL} of this loss would then correspond to a maximum-a-posteriori. Assuming that the local loss $\mathcal{L}^{\text{LOCAL}}$ is twice differentiable, then the Hessian matrix $\nabla^2\mathcal{L}^{\text{LOCAL}}(\theta_n^{\text{LOCAL}}|\mathcal{D}_n)$ can be regarded as an estimation of the posterior covariance

on θ_n^{LOCAL}, given the local dataset $\mathcal{D}_n$. More precisely, defining $\Sigma_n^{\text{LOCAL}} \triangleq \frac{1}{2}\nabla^2 \mathcal{L}^{\text{LOCAL}}(\theta_n^{\text{LOCAL}}|\mathcal{D}_n)^{-1}$, we may approximate the posterior on θ_n by the normal distribution $\mathcal{N}\left(\theta_n^{\text{LOCAL}}, \Sigma_n^{\text{LOCAL}}\right)$.

An alternative solution, which allows sampling without the costly computation of Σ_n^{LOCAL}, consists of adding a random noise ξ_n^{LOCAL} to θ_n^{LOCAL}, and to assign a weight to this noise proportional to $\exp\left(-\mathcal{L}^{\text{LOCAL}}(\theta_n^{\text{LOCAL}} + \xi_n^{\text{LOCAL}}|\mathcal{D}_n)\right)$.

3.2 Covariant matrix estimator

We now consider a first Byzantine mean estimation ρ^{LOCAL} of the users' preferred models, by simply applying a zero-centered quadratically regularized geometric median primitive, i.e.,

$$\rho^{\text{LOCAL}} \triangleq \text{QRGM}_{W,0}(w_{1:N}, \tilde{\theta}_{1:N}^{\text{LOCAL}}). \tag{6}$$

For each user $n \in [N]$, we then define the user's CovariantLicchavi matrix distribution $\tilde{\Sigma}_n^{\text{CL}}$, obtained by drawing θ_n^{LOCAL} from $\tilde{\theta}_n^{\text{LOCAL}}$ and returning the positive semi-definite matrix $(\rho^{\text{LOCAL}} - \theta_n^{\text{LOCAL}})(\rho^{\text{LOCAL}} - \theta_n^{\text{LOCAL}})^T$. Intuitively, this is a distribution over matrices that describe how the user's purely local model θ_n^{LOCAL} probably diverges from the Byzantine mean estimation ρ^{LOCAL} of all users' preferred models. The aggregate CovariantLicchavi matrix Σ^{CL} is then obtained by aggregating all users' matrix distributions $\tilde{\Sigma}_n^{\text{CL}}$, using the identity-centered quadratically regularized geometric median primitive, i.e.

$$\Sigma^{\text{CL}} \triangleq \text{QRGM}_{W,I}(w_{1:N}, \tilde{\Sigma}_{1:N}^{\text{CL}}). \tag{7}$$

3.3 Skewed Licchavi

We now propose to adapt Licchavi by leveraging the CovariantLicchavi covariant matrix Σ^{CL}. We do so by penalizing the discrepancy between a local model and the global model, especially when they are distant according to the covariant matrix Σ^{CL}. More precisely, for any semi-definite positive matrix $S \succeq 0$, define the Mahalanobis norm $\|x\|_S^2 \triangleq x^T S^{-1} x$. Note that this norm can take infinite values, when x does not belong to the image of S. We then define the CovariantLicchavi loss by

$$\text{CL}(\rho, \theta_{1:N}|W, \mathcal{D}_{1:N}) = \frac{1}{2} W \|\rho\|_2^2 + \sum_{n\in[N]} \mathcal{L}^{\text{LOCAL}}(\theta_n|\mathcal{D}_n) + \sum_{n\in[N]} w_n \|\rho - \theta_n\|_{\Sigma^{\text{CL}}}. \tag{8}$$

4 Algorithm

Describe an algorithm. Implement. Test performances.

5 Theorems

Conjecture 4 (To be defined)**.** *Covariant-*Licchavi *is Byzantine-resilient.*

Conjecture 5. *There's a connection with SVD (to be found).*

Conjecture 6. *The covariance matrix is PAC-learned by covariant-*Licchavi*.*

6 Conclusion

This is breakthrough paper.

References

[AGHV22] Youssef Allouah, Rachid Guerraoui, Lê Nguyên Hoang, and Oscar Villemaud. Robust sparse voting. *CoRR*, abs/2202.08656, 2022.

[EFGH21] El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, and Lê Nguyên Hoang. On the strategyproofness of the geometric median. *CoRR*, abs/2106.02394, 2021.

[FGH21] Sadegh Farhadkhani, Rachid Guerraoui, and Lê Nguyên Hoang. Strategyproof learning: Building trustworthy user-generated datasets. *CoRR*, abs/2106.02398, 2021.