Multitask Cooperative Online Learning

Juliette Achddou, Nicolò Cesa-Bianchi, Pierre Laforgue

LAILA lab, Università degli Studi di Milano

Funded by the European Union

Motivation and objective

Motivation: many real-world applications (recommendation, health monitoring) require to provide a personalized service to multiple different clients.

Central idea: communicating information (models, privatized gradients) between agents in a decentralized fashion along the training process.

Informal objective: designing an algorithm with small multitask regret when neighbors in the communication network have similar tasks.

Relevant related works: cooperative online learning [1], multitask online learning [2], distributed online optimization [3].

Model

N **agents**, organized in communication network described by an undirected **graph** G. A hidden sequence of **convex loss functions** $\ell_1, \ell_2, \dots$ chosen adversarially. For $t = 1, 2, \dots$

1. agent $i_t \in [N]$ is activated
2. i_t may *fetch* information from its neighbors
3. i_t predicts $x_t \in \mathcal{X}$
4. i_t pays $\ell_t(x_t)$ and observes $g_t \in \partial \ell_t(x_t)$
5. i_t may *send* information to its neighbors

We aim at minimizing for any comparator $U \in \mathcal{U}$ and horizon T the **multitask regret**

$$R_T(U) = \sum_{i=1}^{N} \sum_{t\,:\,i_t = i} \Big(\ell_t(x_t) - \ell_t([U]_{i:}) \Big)\,.$$

Algorithm

Algorithm 2 `COOL-CN`

Requires: Base algorithm `AlgoClique`, right stochastic matrix W

for $t = 1, 2, \dots$ **do**

  Active agent i_t:

    fetches $\big[Y_t^{(j)}\big]_{i_t:}$ from each $j \in \mathcal{N}_{i_t}$

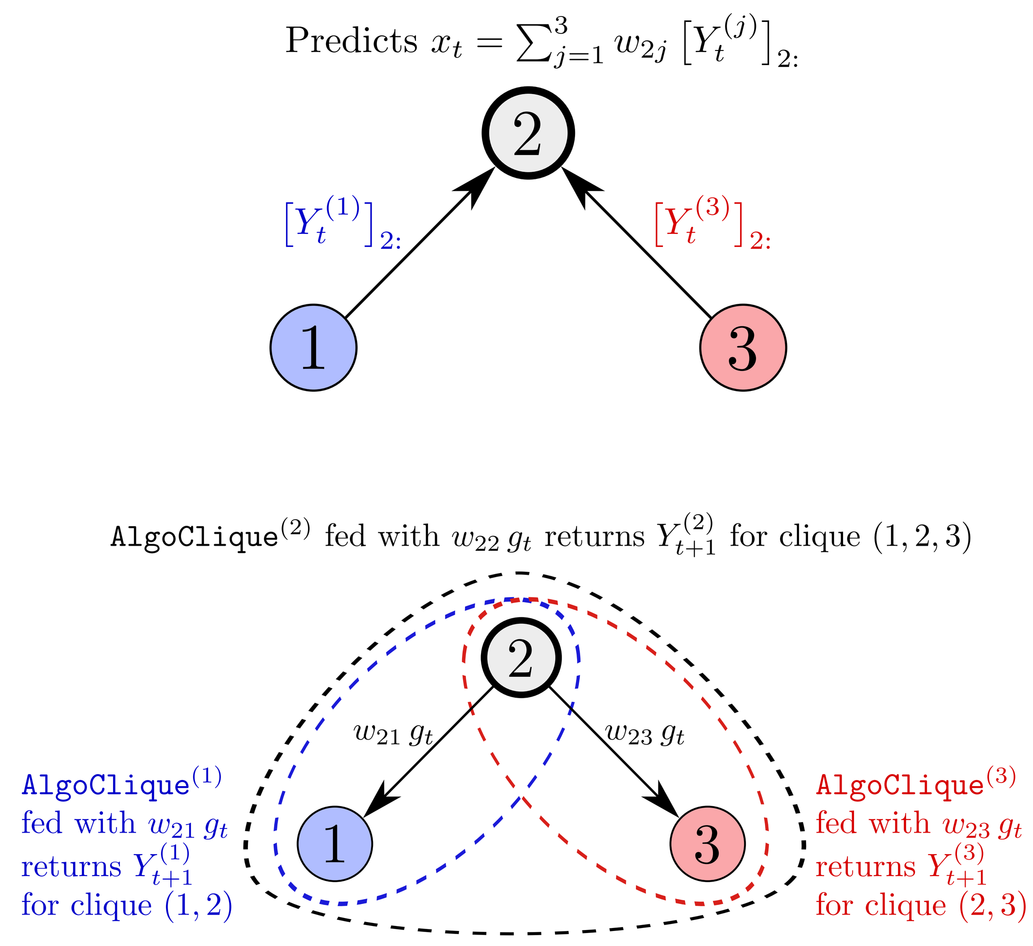
    predicts $x_t = \sum_{j \in \mathcal{N}_{i_t}} w_{i_t j} \big[Y_t^{(j)}\big]_{i_t:}$

    pays $\ell_t(x_t)$ and observes $g_t \in \partial \ell_t(x_t)$

    sends $\big(i_t, w_{i_t j}\, g_t\big)$ to each neighbor $j \in \mathcal{N}_{i_t}$

  for $j \in \mathcal{N}_{i_t}$ **do**

    Agent j feeds the linear loss $\langle w_{i_t j}\, g_t, \cdot \rangle$ to their local instance of `AlgoClique`, and obtains $Y_{t+1}^{(j)}$



Theoretical guarantees

Why this (meta-)algorithm?

Lemma 1. *The regret of* `COOL-CN` *satisfies*

$$R_T(U) \le \sum_{j=1}^{N} R_T^{\text{clique-}j}\big(U^{(j)}\big)\,.$$

where $R_T^{\text{clique-}j}$ *is the regret suffered by* `AlgoClique` *on the linear losses* $\langle w_{i_t j}\, g_t, \cdot\rangle$ *over the rounds* $t \le T$ *such that* $i_t \in \mathcal{N}_j$, *and* $U^{(j)}$ *contains the rows* $U_{i:}$ *for* $i \in \mathcal{N}_j$.

What choice for `Algoclique`?

A valid choice of `AlgoClique` is `MT-FTRL` [2], in its variance adaptive version. It is an algorithm designed for the case without communication constraints (i.e., G is a clique), which satisfies

$$R_T^{\text{clique-}j}\big(U^{(j)}\big) = \tilde{\mathcal{O}}\left(\sqrt{1 + \sigma_j^2 (N_j - 1)}\sqrt{T}\right),$$

where $\sigma_j^2 = \frac{1}{2N_j(N_j-1)} \sum_{i,i' \in \mathcal{N}_j} \left\| U_{i:} - U_{i':} \right\|_2^2$ is a measure of the local variance.

Upper bounds for adversarial activations.

Theorem 2. *For general weights* w_{ij} *we have*

$$R_T(U) \overset{\tilde{\mathcal{O}}}{=} \sum_{j=1}^{N} \max_{i \in \mathcal{N}_j} w_{ij} \sqrt{1 + \sigma_j^2 (N_j - 1)} \sqrt{\sum_{i \in \mathcal{N}_j} T_i}\,.$$

Setting $w_{ij} = \mathbb{I}\{j \in \mathcal{N}_i\}/N_i$, *it becomes*

$$R_T(U) \overset{\tilde{\mathcal{O}}}{=} \sqrt{1 + \sigma_{\max}^2 (N_{\max} - 1)} \sqrt{\frac{N N_{\max} T}{N_{\min}^2}}\,,$$

that particularizes well, e.g., for ρ*-regular graphs*

$$R_T(U) \overset{\tilde{\mathcal{O}}}{=} \sqrt{1 + \rho\, \sigma_{\max}^2} \sqrt{\frac{NT}{\rho + 1}}\,.$$

Upper bounds for stochastic activations.

Theorem 3. *Using appropriate* w_{ij}, *with* $\alpha(G)$ *the independence number of* G, *we have*

$$\mathbb{E}[R_T(U)] \overset{\tilde{\mathcal{O}}}{=} \sqrt{1 + \sigma_{\max}^2 (N_{\max} - 1)} \sqrt{\alpha(G)\, T}\,.$$

Lower bounds for adversarial activations.

Theorem 4. *There exists a sequence of activations and gradients such that for any algorithm*

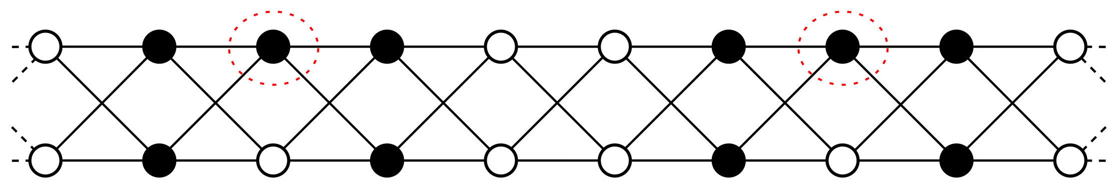
$$R_T \ge \frac{1}{3} \max\left(\sqrt{1 + \sigma^2 (N-1)},\, \sqrt{\alpha_2(G)} \right) \sqrt{T}\,,$$

where $\alpha_2(G)$ *is the double independence number.*

Theorem 5. *For any even number* ρ, *there exists a* ρ*-regular graph and a sequence of activations and gradients such that for any algorithm*

$$R_T \ge \frac{1}{5} \sqrt{1 + \rho\, \sigma_{\min}^2} \sqrt{\frac{NT}{\rho}}\,.$$

Sketch of proof. Activations are restricted to a doubly independent set.



Other results and directions

Privacy.

With `AlgoClique` set as `MT-FTRL`, it is possible to make the algorithm ϵ**-private**, while preserving regret guarantees. We exhibit a **cut-off privacy value** ϵ, below which communicating is useless.

Research Directions.

1. Consider communication delays
2. Consider sending different information
3. Consider communication constraints (e.g., size of the messages sent)
4. Bridging our setting with distributed online learning with multiple tasks

References

[1] N. Cesa-Bianchi, T. Cesari, and C. Monteleoni. Cooperative online learning: Keeping your neighbors updated. In *Algorithmic learning theory*. PMLR, 2020.

[2] N. Cesa-Bianchi, P. Laforgue, A. Paudice, and M. Pontil. Multitask online mirror descent. *Transactions on Machine Learning Research*, 2021.

[3] S. Hosseini, A. Chapman, and M. Mesbahi. Online distributed optimization via dual averaging. In *IEEE Conference on Decision and Control*. IEEE, 2013.

Check out our video