DISTRIBUTED SEQUENCE PREDICTION: A CONSENSUS+INNOVATIONS APPROACH

Anit Kumar Sahu and Soummya Kar

Department of Electrical and Computer Engineering,
Carnegie Mellon University, Pittsburgh PA 15213
{anits,soummyak}@andrew.cmu.edu

ABSTRACT

This paper focuses on the problem of distributed sequence prediction in a network of sparsely interconnected agents, where agents collaborate to achieve provably reasonable predictive performance. An expert assisted online learning algorithm in a distributed setup of the *consensus+innovations* form is proposed, in which the agents update their weights for the experts' predictions by simultaneously processing the latest network losses (*innovations*) and the cumulative losses obtained from neighboring agents (*consensus*). This paper characterizes the regret of the agents' prediction in lieu of the proposed distributed online learning algorithm and establishes the sub-linear regret of the agents' predictions with respect to the best forecasting expert.

***Index Terms*—** Distributed Inference, Online Learning, Expert-assisted Learning, Sequence Prediction, Multi-agent Networks.

1. INTRODUCTION

The ubiquitous nature of online learning has made it an area of interest across different fields, especially in the era of big data. Due to heterogeneity and the enormity of data nowadays, it is very difficult to figure out the statistical properties of distributions so as to apply appropriate learning and inference algorithms based on the distributions. In such situations, where the statistical characterizations of distributions from which the data is sampled from is unknown, many offline learning algorithms are rendered ineffective as they are based on assumptions which cannot be verified with the given set of data. The difficulties encountered in such situations motivates the use of online learning algorithms which do not make any assumptions on the distributions from which the data is sampled from. In the context of online algorithms, expert assisted online algorithms have gained a lot of prominence. Broadly speaking, in expert assisted learning scenarios, an online learner uses the learning capabilities of *reference learners* or experts. In the context of expert assisted online learning, we specifically look at expert assisted sequence prediction in this paper, where the forecaster aggregates the predictions from the *reference predictors* or the experts and then weighs the predictions of the experts in a systematic way by considering the past predictive performance of the experts (see, for example [1, 2]). In real world scenarios, such as predicting stock prices, it is practically not possible for a forecaster to have access to all the reference stock price forecasters or experts. Also, there might be privacy concerns for an expert to share the predictions with every forecaster. Motivated by the above discussion, we propose a distributed sequence prediction algorithm of the *consensus + innovations* type ([3, 4]) in a network of forecasting agents, where each agent (forecaster) has access to its expert predictor but computes the weights for other experts by simultaneously processing neighborhood information concerning the tracked losses (*consensus*) and latest observed losses (*innovations*). Moreover, the inter-agent communication conforms to a preassigned possibly sparse communication graph. Depending on the scenario under consideration, the inter-agent communication might reflect the dynamics of communications between forecasters with respect to privacy and competitiveness. Each agent in the proposed setup has access to a genie which takes the weights of the experts as generated by an agent and informs the agent of its loss. In spite of such a constrained setup, where the agents' do not have access to all the experts, we establish the sub-linear regret of each forecasting agent under some mild assumptions, where the regret is respect to the best performing *reference forecaster*.

Prediction with expert advice in literature have been extensively studied (see, for example [1, 2, 5]). The use of potential functions in sequential prediction was introduced in [6]. However, the unique nature of forecasting based on exponential potential which makes the weighting scheme a function of the previous time instant loss makes it particularly attractive in practice. Exponentially weighted average forecasting was first proposed and analyzed in [1]. The sub-linear regret of such a scheme for a fixed horizon was first established in [7]. Sub-linear regret bounds that hold uniformly over time for exponentially weighted average forecasting, which involve a doubling trick was established in [8]. To the best of our knowledge, this is the first time the problem of online sequence prediction is being addressed in a distributed setup. For the proposed distributed algorithm of the *consen-*

This work was supported in part by NSF under grant CCF-1513936.

sus+innovations form, we not only establish sub-linear regret for a fixed horizon but also sub-linear regret bounds that hold uniformly over time.

The rest of the paper is organized as follows. Spectral graph theory, preliminaries and notation are discussed next. The expert-assisted online learning setup is described in Section 2, where we also review some preliminaries concerning online sequence prediction. Section 3 presents the proposed distributed sequence prediction algorithm, while Section 4 concerns with the main results of the paper. Finally, Section 5 concludes the paper.

Spectral Graph Theory. The inter-agent communication network is a simple[1] undirected graph $G = (V, E)$, where V denotes the set of agents or vertices with cardinality $|V| = N$, and E the set of edges with $|E| = M$. If there exists an edge between agents i and j, then $(i, j) \in E$. A path between agents i and j of length m is a sequence $(i = p_0, p_1, \cdots, p_m = j)$ of vertices, such that $(p_t, p_{t+1}) \in E$, $0 \leq t \leq m-1$. A graph is connected if there exists a path between all possible agent pairs. The neighborhood of an agent n is given by $\Omega_n = \{j \in V | (n, j) \in E\}$. The degree of agent n is given by $d_n = |\Omega_n|$. The structure of the graph is represented by the symmetric $N \times N$ adjacency matrix $\mathbf{A} = [A_{ij}]$, where $A_{ij} = 1$ if $(i, j) \in E$, and 0 otherwise. The degree matrix is given by the diagonal matrix $\mathbf{D} = diag(d_1 \cdots d_N)$. The graph Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$. The Laplacian is a positive semidefinite matrix, hence its eigenvalues can be ordered and represented as $0 = \lambda_1(\mathbf{L}) \leq \lambda_2(\mathbf{L}) \leq \cdots \lambda_N(\mathbf{L})$. Furthermore, a graph is connected if and only if $\lambda_2(\mathbf{L}) > 0$ (see [9] for instance).

2. ONLINE LEARNING: SEQUENCE PREDICTION

In this section we discuss preliminaries about online sequence prediction (see, [10] for example). Sequence prediction in an online learning framework is applicable to fairly generic prediction scenarios as it is assumption free as far as the sequence is concerned and is an online algorithm. Expert assisted sequence prediction involves sequential decision making where a forecaster's goal is to predict an unknown sequence $\{y_i\}_{i=1}$ whose elements come from an action space $\mathcal{Y}$ by aggregating the predictions of *reference forecasters* or experts. The forecaster's predictions which are denoted as $\{\hat{p}_i\}$ belong to a decision space $\mathcal{D}$, which is taken to be a convex subset of a vector space. The forecaster computes his prediction in an online sequential manner and the predictive performance is compared to that of *reference forecasters* or experts. Technically speaking, the forecaster at time t has access to predictions $\{f_{i,t}\}_{i=1}^{N}$ from N *reference forecasters* or experts. Based on these predictions, the forecaster comes up with its own prediction $\hat{p}_t$ which is when the true outcome y_t is revealed. The predictions of the forecaster and the experts are evaluated using a non-negative loss function $l : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$. We formalize some assumptions on the loss function $l(\cdot, \cdot)$ and the decision space $\mathcal{D}$ before proceeding further.

Assumption A1. *The loss function $l : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is convex in its first argument and it takes values in $[0, 1]$.*

Assumption A2. *The decision space $\mathcal{D}$ is a convex subset of a vector space.*

The goal of the forecaster is to keep the cumulative regret with respect to each *reference forecaster* as low as possible. Formally, the regret with respect to expert n is defined as

$$R_{n,t} = \sum_{s=1}^{t} \left(l\left(\hat{p}_s, y_s\right) - l(f_{n,s}, y_s) \right) = \hat{L}_t - L_{n,t}, \tag{1}$$

where $\hat{L}_t = \sum_{s=1}^{t} l\left(\hat{p}_s, y_s\right)$ and $L_{n,t} = \sum_{s=1}^{t} l\left(f_{n,s}, y_s\right)$ which denote the cumulative loss of the forecaster and the expert n at time t respectively. Technically speaking, the forecaster's goal is to attain a regret as possible across all sequences of outcomes possible. Formally, the goal can be represented as

$$\begin{aligned} &\max_{n=1,\cdots,N} R_{n,t} = o(t), \text{or, equivalently} \\ &\lim_{t\to\infty} \frac{1}{t} \left(\hat{L}_t - L_{n,t} \right) = 0, \end{aligned} \tag{2}$$

where the convergence is uniform across all sequence of outcomes. For the rest of the paper, we focus on weighted average forecasters. In case of weighted average forecasting, the prediction at any time t are computed as follows:

$$\hat{p}_t = \frac{\sum_{n=1}^{N} w_{n,t-1} f_{n,t}}{\sum_{n=1}^{N} w_{n,t-1}}, \tag{3}$$

where $w_{n,t}$ is the weight associated with the n-th expert's prediction at time t. Note, that the prediction $\hat{p}_t \in \mathcal{D}$ as it is a convex combination of the experts' predictions. As sequence prediction is an online sequential task, it is natural to see that the weights assigned to the experts at time t depends on the regret or the individual losses of the experts till time $t-1$. Hence, the weight assigned to expert n can be considered to be an increasing function of the regret of the forecaster with respect to expert n. One of the most widely used weighting schemes is the exponential weighting scheme, where the weights for the prediction at time t is computed as

$$w_{n,t-1} = \frac{e^{\eta R_{n,t-1}}}{\sum_{n=1}^{N} e^{\eta R_{n,t-1}}}, \tag{4}$$

where η is positive and referred to as the *learning parameter*. Then, the prediction of the forecaster at time t is given by,

$$\hat{p}_t = \frac{\sum_{n=1}^{N} e^{-\eta L_{n,t-1}} f_{n,t}}{\sum_{n=1}^{N} e^{-\eta L_{n,t-1}}}. \tag{5}$$

It is interesting that with exponentially weighted average prediction depends just on thepast performance of the experts

[1] A graph is said to be simple if it is devoid of self loops and multiple edges.

and not on the past predictions $\{\hat{p}_s\}_{s\leq t}$. Under rather weak assumptions on the loss functions and decision space,i.e., assumptions A1-A2 sub-linear regret has been established for exponentially weighted average prediction.
The following result characterizes the regret bound for a specific time instant t.

Theorem 2.1 (Theorem 2.2 in [10]). *Let Assumptions A1-A2 hold. Consider the exponentially weighted average predictor as discussed in* (4)-(5)*. For any fixed time t and $\eta > 0$, and for all sequence of outcomes $\{y_s\}_{s=1}^{t} \in \mathcal{Y}$ the regret of the exponentially weighted average predictor satisfies*

$$\hat{L}_t - \min_{n=1,\cdots,N} L_{n,t} \leq \frac{\ln N}{\eta} + \frac{t\eta}{8}. \tag{6}$$

In particular, if η is chosen to be $\sqrt{\frac{8\ln N}{t}}$, the right hand side (RHS) in (6) *becomes $\sqrt{\frac{t\ln N}{2}}$.*

The above result which characterizes the regret upto a fixed time t can be extended to hold for all times by switching to a time-varying version of the parameter η. Formally, let $\eta_t = \sqrt{\frac{8\ln N}{t}}$. Then, we have:

Theorem 2.2 (Theorem 2.3 in [10]). *Let Assumptions A1-A2 hold. Consider the exponentially weighted average predictor as discussed in* (4)-(5)*. For all $t \geq 1$, and for all sequence of outcomes $\{y_s\}_{s=1}^{t} \in \mathcal{Y}$ the regret of the exponentially weighted average predictor satisfies*

$$\hat{L}_t - \min_{n=1,\cdots,N} L_{n,t} \leq \sqrt{\frac{\ln N}{8}} + \sqrt{2t\ln N}. \tag{7}$$

It is readily seen that the regret bound in (7) satisfies the condition that $\lim_{t\to\infty} \frac{1}{t}\left(\hat{L}_t - L_{n,t}\right) = 0$.
Theorems 2.1-2.2 require the forecaster to have access to all of the experts' predictions at all times, or equivalently the cumulative losses at all times. However, in many practical application scenarios, the access to predictions of all experts might have privacy concerns and it might not be possible for a forecaster to track all the losses of the experts. A best motivating example would be predicting the stock prices in a stock exchange, where multiple agents are trying to predict stock prices where they have access to a few experts or a few agents in their proximal neighborhood with whom they exchange summarized information, i.e., losses and not the predictions. No agent has access to all the experts in such a setting. Based on such practical application scenarios, where the information exchange in a multi-agent network setting is constrained to an agent's neighborhood, we propose a distributed approach for sequence prediction scheme. To obtain a reasonable predictive performance, we propose a distributed online learning algorithm in the *consensus + innovations* framework, where every forecasting agent incorporates the information obtained from the neighbors and the latest sensed information simultaneously.

3. DISTRIBUTED SEQUENCE PREDICTION

In this section, we introduce and develop the distributed sequence prediction algorithm. In section 4 we state the main results concerning the regret bounds for the proposed distributed sequence prediction algorithm. We skip the proofs due to space limitations.
There are N forecasting agents in the network. Each agent n has access to its local expert n. At each time instant t, an agent observes the loss of its own local expert and shares with its neighborhood a cumulative loss type quantity (to be specified soon) and the latest prediction loss and, in turn, receives the same from its neighbors. Formally, every agent n tracks the network losses across all the experts, albeit in a constrained manner by updating $\mathbf{S}_{d,n}(t) \in \mathbb{R}^N$. Formally the update can be represented as,

$$\begin{aligned}\mathbf{S}_{d,n}(t+1) = &\underbrace{w_{nn}\mathbf{S}_{d,n}(t) + \sum_{l\in\Omega_n} w_{nl}\mathbf{S}_{d,l}(t)}_{\text{neighborhood consensus}} \\ &\underbrace{+ w_{nn}\mathbf{L}_{d,n}(t) + \sum_{l\in\Omega_n} w_{nl}\mathbf{L}_{d,l}(t)}_{\text{innovation}},\end{aligned} \tag{8}$$

where $\mathbf{L}_{d,n}(t) \in \mathbb{R}^N$, Ω_n and w_{ln}'s denote the vector of losses at forecasting agent n, the neighborhood of agent n and the weight of the link from the agent n to agent l in the inter-agent communication graph respectively. As forecasting agent n has access only to the loss of its own expert, $\mathbf{L}_{d,n}(t)$ has all its entries to be zero except the n-th entry. The update in (8) can be written in a compact manner as follows:

$$\mathbf{S}_d(t+1) = (\mathbf{I}_N \otimes \mathbf{W})\left(\mathbf{S}_d(t) + \mathbf{L}_d(t)\right), \tag{9}$$

where $\mathbf{S}_d(t+1) = \left[\mathbf{S}_{d,1}^\top(t+1), \cdots, \mathbf{S}_{d,N}^\top(t+1)\right]^\top$ and $\mathbf{L}_d(t+1) = \left[\mathbf{L}_{d,1}^\top(t+1), \cdots, \mathbf{L}_{d,N}^\top(t+1)\right]^\top$. The information exchange in the update (8) is limited to a pre-specified possibly sparse inter-agent communication graph, where the weights are designed according to $\mathbf{W} = \mathbf{I} - \delta\mathbf{L}$, where $\mathbf{L}$ is the graph Laplacian. It is to be noted that the entries of the weight matrix $\mathbf{W} = \mathbf{I} - \delta\mathbf{L}$ are designed in such a way that $\mathbf{W}$ is non-negative, symmetric, irreducible and stochastic, i.e., each row of $\mathbf{W}$ sums to one. Furthermore, the second largest eigenvalue in magnitude of $\mathbf{W}$, denoted by r, is strictly less than one (see [11]). Moreover, by the stochasticity of $\mathbf{W}$, the quantity r satisfies $r = ||\mathbf{W} - \mathbf{J}||$, where $\mathbf{J} = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$. The quantity r corresponds to the information flow in the network. For example, $r = 0$ corresponds to the completely connected setting, while $r = 1$ corresponds to the setting where each agent is by itself. The choice of δ is taken to be $\frac{2}{\lambda_2(\mathbf{L})+\lambda_N(\mathbf{L})}$ (see [12]). We state an assumption on the inter-agent communication graph before proceeding further,

Assumption A3. *The inter-agent communication network modeling the information exchange among the forecasting agents is connected, i.e., $\lambda_2(\mathbf{L}) > 0$, where $\mathbf{L}$ denotes the associated graph Laplacian matrix.*

The weight for the forecasting expert l at forecasting agent n at time t, $w_{l,t}^n$ is computed as follows:

$$w_{l,t}^n = \frac{e^{\mathbf{e}_l^\top \eta S_{d,n}(t)}}{\sum_{m=1}^N e^{\mathbf{e}_m^\top \eta S_{d,n}(t)}}, \tag{10}$$

where η is positive and referred to as the *learning parameter*. Note that, $S_{d,n}(t)$ has losses incorporated into it till time $t-1$. It is to be noted that forecasting agent doesn't have access to the predictions of other forecasting experts except its own. Hence, we assume that, there is a *genie* which takes the weights assigned to the forecasting experts at each time t and then computes the prediction of the forecasting agent n, $\hat{p}_{n,t}$ in the following manner

$$\hat{p}_{n,t} = \frac{\sum_{l=1}^N w_{l,t}^n f_{l,t}}{\sum_{l=1}^N w_{l,t}^n}. \tag{11}$$

Moreover, due to the nature of the update for the proposed distributed algorithm, the prediction incorporates all the past information unlike the centralized case in (5), which makes the analysis for the distributed case highly non-trivial. We state another assumption pertaining to the best performing forecasting expert before proceeding further.

Assumption A4. *The cumulative loss of the best performing forecasting expert satisfies*

$$\min_{n=1,\cdots,N} L_{n,t} = o(t) \tag{12}$$

Assumption A4 ensures that the experts' predictions are reasonable. In case, when assumption A4 does not hold, the performance of not only the distributed algorithm, but also that of the centralized algorithm would be bad, i.e., the cumulative losses would be linear in time.

4. MAIN RESULTS

In this section, we specifically characterize the regret bounds for the proposed distributed sequence prediction algorithm. The first result concerns with the regret bounds for a fixed time t.

Theorem 4.1. *Consider the distributed sequence prediction algorithm in* (8)-(11) *under Assumptions A1-A4. For any fixed time t and $\eta > 0$, and for all sequence of outcomes $\{y_s\}_{s=1}^t \in \mathcal{Y}$ the regret of the distributed exponentially weighted average predictor satisfies*

$$\hat{L}_{n,t} \leq e^{\frac{2\eta N r}{1-r}} \left(\min_{n=1,\cdots,N} L_{n,t} + \frac{\ln N}{\eta} + \frac{t\eta}{8} + \frac{2\eta N r}{1-r} \right), \tag{13}$$

where $\hat{L}_{n,t}$ and r denote the loss cumulative loss of forecasting agent n and $\|\mathbf{W} - \mathbf{J}\|$ respectively.

As such we do not need Assumption A4 to establish the regret bound in (13). However, Assumption A4 ensures that the regret bound is sub-linear with respect to the best performing expert. It is to be noted that if η is chosen as $\sqrt{\frac{8(1-r)\ln N}{t+2Nr}}$, then the regret bound becomes

$$\hat{L}_{n,t} \leq e^{\frac{2\sqrt{8\ln N}}{\sqrt{(t+2Nr)(1-r)}}} \left(\min_{n=1,\cdots,N} L_{n,t} + \frac{\sqrt{\ln N(t+2Nr)}}{\sqrt{8(1-r)}} + \frac{t\sqrt{(1-r)\ln N}}{\sqrt{8\,(t+2Nr)}} + \frac{2\sqrt{8\ln N}}{\sqrt{(t+2Nr)(1-r)}} \right). \tag{14}$$

It is readily seen that when $r = 0$, the regret bound reduces to that of the one in Theorem 2.1. Moreover, it can also be seen that $\lim_{t\to\infty} \frac{1}{t}\left(\hat{L}_{n,t} - L_{n,t}\right) = 0$ and that the regret bound is a function of network connectivity in terms of r. The next result concerns with the regret bounds that hold uniformly over time.

Theorem 4.2. *Let the hypotheses of Theorem 4.1 hold. For all $t \geq 1$, and for all sequence of outcomes $\{y_s\}_{s=1}^t \in \mathcal{Y}$ the regret of the distributed exponentially weighted average predictor with time-varying learning parameter $\eta_t = \sqrt{8\ln N/t}$ satisfies*

$$\hat{L}_{n,t} \leq e^{\frac{2Nr\sqrt{8\ln N}}{1-r}} \left(\min_{n=1,\cdots,N} L_{n,t} + \sqrt{\frac{\ln N}{8}} + \sqrt{2t\ln N} + \frac{2Nr\sqrt{8\ln N}}{1-r} \right). \tag{15}$$

As in the case of Theorem 4.1, we do not need Assumption A4 in order to establish the regret bound in (15). However, Assumption A4 ensures that the regret bound is sub-linear with respect to the best performing expert. As a consistency check, it can be verified that with $r = 0$, the regret bound reduces to the bound derived in Theorem 2.2. The regret bound derived in Theorem 4.2 is a function of the network connectivity and the bound is smaller when the network connectivity is better. Finally, it can also be seen that $\lim_{t\to\infty} \frac{1}{t}\left(\hat{L}_{n,t} - L_{n,t}\right) = 0$ which establishes that the regret is sub-linear with respect to the best performing expert.

5. CONCLUSION

In this paper, we have proposed a *consensus + innovations* type algorithm for distributed expert assisted sequence prediction, in which every agent computes the weights of the experts' predictions by simultaneous processing of neighborhood information and local newly sensed information and where the inter-agent collaboration is restricted to a possibly sparse communication graph. Under rather generic assumptions, we have established the sub-linear regret bounds for each agent with respect to the best performing expert in the fixed time setting and also in the case where the bounds hold uniformly over time. A natural direction for future research consists of establishing minimax regret for settings involving noisy information exchange in a distributed information processing setup.

6. REFERENCES

[1] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," in *Foundations of Computer Science, 1989., 30th Annual Symposium on*. IEEE, 1989, pp. 256–261.

[2] V. G. Vovk, "Aggregating strategies," in *Proc. Third Workshop on Computational Learning Theory*. Morgan Kaufmann, 1990, pp. 371–383.

[3] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 674–690, 2011.

[4] ——, "Consensus+ innovations distributed inference over networks: cooperation and sensing in networked systems," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 99–109, 2013.

[5] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, "How to use expert advice," *Journal of the ACM (JACM)*, vol. 44, no. 3, pp. 427–485, 1997.

[6] N. Cesa-Bianchi and G. Lugosi, "Potential-based algorithms in on-line prediction and game theory," *Machine Learning*, vol. 51, no. 3, pp. 239–261, 2003.

[7] N. Cesa-Bianchi, "Analysis of two gradient-based algorithms for on-line regression," in *Proceedings of the tenth annual conference on Computational learning theory*. ACM, 1997, pp. 163–170.

[8] P. Auer, N. Cesa-Bianchi, and C. Gentile, "Adaptive and self-confident on-line learning algorithms," *Journal of Computer and System Sciences*, vol. 64, no. 1, pp. 48–75, 2002.

[9] F. R. Chung, *Spectral graph theory*. American Mathematical Soc., 1997, vol. 92.

[10] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[11] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[12] S. Kar, S. Aldosari, and J. M. F. Moura, "Topology for distributed inference on graphs," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2609–2613, June 2008.