Adaptive Communication Bounds for Distributed Online Learning

Michael Kamp[1], Mario Boley[1], Michael Mock[2], Daniel Keren[3], Assaf Schuster[4], and Izchak Sharfman[4]

[1]Fraunhofer IAIS & University Bonn, {surname.name}@iais.fhg.de
[2]Fraunhofer IAIS, michael.mock@iais.fhg.de
[3]Haifa University, dkeren@cs.haifa.ac.il
[4]Technion, Israel Institute of Technology, {assaf,tsachis}@technion.ac.il

Abstract

We consider distributed online learning protocols that control the exchange of information between local learners in a round-based learning scenario. The learning performance of such a protocol is intuitively optimal if approximately the same loss is incurred as in a hypothetical serial setting. If a protocol accomplishes this, it is inherently impossible to achieve a strong communication bound at the same time. In the worst case, every input is essential for the learning performance, even for the serial setting, and thus needs to be exchanged between the local learners. However, it is reasonable to demand a bound that scales well with the hardness of the serialized prediction problem, as measured by the loss received by a serial online learning algorithm. We provide formal criteria based on this intuition and show that they hold for a simplified version of a previously published protocol.

1 Introduction

We consider round-based learning scenarios on multiple connected dynamic data streams where a real-time service is provided by a **distributed online learning system** of $k \in \mathbb{N}$ local learners. A **distributed online learning protocol** controls the exchange of information between the learners with the goal of providing a service quality—measured by a loss function—that is close to the one of a hypothetical serial learner. Such an optimal predictive behavior can be trivially achieved by centralizing all data but at the expense of a total communication cost in $\Omega(kT)$ for a total time horizon $T \in \mathbb{N}$. This amount of communication in practice often exceeds network capacities or delays the service beyond its time limits. No communication, on the other hand, leads to a significantly higher loss that increases with the number of learners. Earlier research focused on static communication strategies that retain the service quality of a hypothetical centralized learner by communicating periodically, thereby reducing communication by a fixed factor, but not changing the asymptotic behavior. In this paper, we introduce the notion of an adaptive protocol that invests communication only if it thereby significantly reduces its loss. As a consequence, the communication bound of an adaptive protocol depends on the loss of the serial setting instead of the total time. We show for a simplified version of dynamic averaging, a previously published protocol, that it satisfies this notion of adaptivity while it retains the optimal predictive behavior of a serial learner.

2 Performance Bounds for Distributed Online Learning Protocols

In the following, we assume an online learning algorithm $A = (W, \varphi, f)$ run on each local learner $l \in [k]$ in the distributed system to maintain a local model $w_{t,l} \in W$. At each time point $t \in \mathbb{N}$,

each learner observes an input $z_{t,l}$ drawn independently from an input space Z from a time-variant distribution $\mathcal{D}_t : Z \to [0, 1]$. Based on this input and the local model, the local learner provides a service whose quality is measured by a loss function $f : Z \times W \to \mathbb{R}_+$. After providing the service, the local learner updates its local model using an update rule $\varphi : Z \times W \to W$. The performance of A within a time horizon $T \in \mathbb{N}$ is measured by its **cumulative loss**

$$L_A(T) = \sum_{t=1}^{T} f(z_t, w_t) \ .$$

Performance guarantees for online learning algorithms are typically given by a **loss bound** $\mathbf{L}_A(T)$, i.e., for all input sequences $z_1, ..., z_T \in Z$ it holds that $L_A(T) \leq \mathbf{L}_A(T)$. Loss bounds can be defined with respect to a sequence of reference models, in which case they are referred to as regret bounds. Such regret bounds are typically sub-linear in T, e.g., for scenarios with a static distribution an optimal regret bound is in $\mathcal{O}(\sqrt{T})$ (see, e.g., Cesa-Bianchi and Lugosi [2006]), which is achieved by many online learning algorithms, including stochastic gradient descent [Zinkevich et al., 2010] and passive aggressive updates [Crammer et al., 2006]. A distributed online learning protocol $P = (A, \sigma)$ runs algorithm A on a distributed online learning system and interchanges information between the local learners by synchronizing their local models $w_{t,1}, ..., w_{t,k}$ using a **synchronization operator** $\sigma : W^k \to W^k$. The cumulative loss of a distributed online learning protocol P is defined as

$$L_P(T, k) \sum_{t=1}^{T} \sum_{l=1}^{k} f(z_{t,l}, w_{t,l}) \ .$$

We measure the amount of communication of σ by its **communication cost** $c_\sigma : W^k \times \mathbb{N} \to \mathbb{N}$, i.e., $c_\sigma(\mathbf{w}_t, t)$ is the number of bytes transmitted at time point t when σ is applied to the current model configuration $\mathbf{w}_t = w_{t,1}, ..., w_{t,k}$. The communication cost depends on the system architecture, e.g., in a system with a designated coordinator node all models can be sent to the coordinator, processed and the new models can be send back to the learners using $2k$ messages each containing one model of fixed size, thus $c_\sigma(\mathbf{w}_t, t) \in \Theta(k)$. We denote the cumulative amount of communication required for synchronization by

$$C_P(T, k) = \sum_{t=1}^{T} c_\sigma(\mathbf{w}_t, t) \ .$$

There is a natural trade-off between communication and loss of a distributed online learning system. A loss similar to a serial setting can be trivially achieved by permanent centralization. On the other hand, communication can be entirely omitted. If the cumulative loss of an online learning algorithm A is bounded by $\mathbf{L}_A(T)$, the loss of a distributed system with k local learners running A without any synchronization is bounded by $\mathbf{L}_{\text{nosync}}(T, k) = k\mathbf{L}_A(T)$, whereas the communication is $C_P(T) = 0$. Such a distributed system processes kT inputs. The loss of a permanently centralizing system on the other hand is bounded by $\mathbf{L}_{\text{central}}(T, k) = \mathbf{L}_A(kT)$, i.e., the loss bound of a serial online learning algorithm processing kT inputs. For example in a static scenario with an optimal loss bound of $\mathbf{L}_A(kT) \in \mathcal{O}(\sqrt{kT})$, the loss bound of centralization is superior by a factor of $\sqrt{k}$. However, the communication cost $C_P(T)$ is in $\Theta(kT)$. Current communication strategies that retain the service quality of a hypothetical centralized learner are based on communicating periodically [Mcdonald et al., 2009, Dekel et al., 2012] after a fixed number $b \in \mathbb{N}$ of data points have been processed, thereby reducing communication by a fixed factor of $1/b$, but the communication bound remains in $\Theta(kT)$. The communication bound of an adaptive protocol should only depend on $\mathbf{L}_A(T)$ and not on T, while at the same time retaining the loss bound of the serial setting. In the following definition we formalize this.

Definition 1. *A distributed online learning protocol $P = (A, \sigma)$ processing kT inputs is **consistent** if it retains the loss bound of the serial online learning algorithm A processing kT inputs, i.e.,*

$$\boldsymbol{L}_P(T, k) \in \mathcal{O}\left(\boldsymbol{L}_A(kT)\right) \ .$$

*The protocol is **adaptive** if its communication bound is linear in the number of local learners k and the loss bound $\boldsymbol{L}_A(kT, k)$ of the serial online learning algorithm, i.e.,*

$$\boldsymbol{C}_P(T, k) \in \mathcal{O}\left(k\boldsymbol{L}_A(kT)\right) \ .$$

An efficient protocol is adaptive and consistent at the same time. In the following we will present such a protocol.

3 An Adaptive and Consistent Distributed Online Learning Protocol

In Kamp et al. [2014] we presented a dynamic protocol, denoted **dynamic averaging**, that adapts its communication to the loss incurred. For specific scenarios this protocol is consistent. In particular, the protocol is consistent for an online learning scenario where each learner maintains a linear model, i.e., $W = \mathbb{R}^n$, using a specific type of update rules denoted f-proportional convex update rules for a loss function f. That is, there exists a constant $\gamma > 0$, a closed convex set $\Gamma_z \subseteq \mathbb{R}^n$, and $\tau_z \in (0, 1]$ such that for all $w \in \mathbb{R}^n$ and $z \in Z$ it holds that

(i) $\|w - \varphi(z, w)\| \geq \gamma f(z, w)$, i.e., the update magnitude is a true fraction of the loss incurred, and

(ii) $\varphi(z, w) = w + \tau_z (\pi_\Gamma (w) - w)$ where $\pi_\Gamma (w)$ denotes the projection of w onto Γ_z, i.e., the update direction is identical to the direction of a convex projection that only depends on the training example.

Examples of such update rules are stochastic gradient descend, as well as passive aggressive updates and their regularized variants. For these update rules the update magnitude is bounded by a multiple of the loss, i.e., there exists a constant $C \in \mathbb{R}_+$ such that for all inputs $z \in Z$ it holds that $\|w - \varphi(z, w)\| \leq C f(z, w)$. This is true, e.g., for stochastic gradient descent with $C = D_Z D_W$, where D denotes the diameter. In this setting dynamic averaging retains the regret bounds of static averaging, i.e., synchronizing every $b \in \mathbb{N}$ rounds. In case of a fixed target distribution $\mathcal{D}$ it has been shown by Dekel et al. [2012] that static averaging retains the optimality of stochastic gradient descent, i.e., it retains the regret bound of the serial online learning algorithm. Thus, both static and dynamic averaging are consistent. However, static averaging has communication costs of $C_{\text{static}}(T, k) = \mathbf{c}_k \lceil T/b \rceil$ and is thus not adaptive. In the following we will define dynamic averaging and provide a communication bound in $\mathbf{L}_A(kT)$.

The dynamic averaging protocol $P = (A, \sigma_\Delta)$ synchronizes the local learners using a **dynamic averaging operator** σ_Δ. This operator only communicates when the **model divergence** $\delta(\mathbf{w}_t) = 1/k \sum_{l=1}^{k} \|w_{t,l} - \overline{\mathbf{w}}_t\|^2$ exceeds a divergence threshold Δ, where $\overline{\mathbf{w}}_t = 1/k \sum_{l=1}^{k} w_{t,l}$ denotes the average model. The dynamic averaging operator is defined as

$$\sigma_\Delta(\mathbf{w}_t) = \begin{cases} (\overline{\mathbf{w}}_t, \ldots, \overline{\mathbf{w}}_t), & \text{if } \delta(\mathbf{w}_t) > \Delta \\ \mathbf{w}_t, & \text{otherwise} \end{cases} .$$

In order to decide when to communicate, each local learner $l \in [k]$ monitors the **local condition** $\|w_{t,l} - r_t\|^2 \leq \Delta$ for a **reference vector** $r_t \in W$ that is common among all learners (see Keren et al. [2012], Sharfman et al. [2007], Gabel et al. [2014], Giatrakos et al. [2012] for a more general description of this method). The local conditions guarantee that if none of them is violated the divergence does not exceed the threshold Δ. If a local condition is violated, a synchronization is triggered. We refer to the averaging of all models as **full synchronization**. Using an architecture with k local learners and an additional coordinator node, a full synchronization can be performed with communication linear in k, i.e., $c_{\sigma_\Delta}(\mathbf{w}_t, t) = \mathbf{c}_k \in \mathcal{O}(k)$. In order to further reduce communication, instead of a full synchronization, the distributed system can try to resolve the violation by averaging a small subset of local models so that the local conditions hold after this averaging. We refer to this operation as **partial synchronization**. The subset is augmented by local models until either all local conditions hold or a stop criterion of an appropriate hedging strategy is met. The hedging strategy ensures that the communication in each time step cannot exceed $\mathbf{c}_k$. In the following we show that dynamic averaging is adaptive.

Theorem 2. *The communication $C_P(T, k)$ of a distributed online learning protocol using the dynamic averaging operator and an f-proportional convex update rule with $\|w - \varphi(z, w)\| \leq C f(z, w)$ is bounded by*

$$\boldsymbol{C}_P(T, k) = \boldsymbol{c}_k \frac{C}{\sqrt{\Delta}} \boldsymbol{L}_P(T, k)$$

where $\boldsymbol{c}_k$ is an upper bound on the amount of communication per time point t.

Proof. The dynamic averaging protocol communicates only if a violation of a local condition $\|w_{t,l} - r_t\|^2 \leq \Delta$ occurs. At each time point with at least one violation dynamic averaging has

communication costs of at most $\mathbf{c}_k$, i.e., the cost of a full synchronization. Thus, we can bound the amount of communication by bounding the number of violations. That is, we derive a bound for $V_l(T)$, the number of time points $t \in [T]$ where the local condition of learner l is violated. For that, assume that at $t = 1$ all models are initialized with $w_{1,1} = \cdots = w_{1,k}$ and $r_1 = \overline{W}_1$, i.e., for all local conditions it holds that $\|w_{1,l} - r_1\| = 0$. A violation, i.e., $\|w_{t,l} - r\| > \sqrt{\Delta}$, occurs if one local learner drifts away from r_t by more than $\sqrt{\Delta}$. After a violation a full synchronization is performed and $r_t = \overline{W}_t$, hence $\|w_{t,l} - r_t\| = 0$ and the situation is again similar to the initial setup for $t = 1$. In the worst case, a local learner drifts continuously in one direction until a violation occurs. Hence, we can bound the number of violations $V_l(T)$ by the sum of its drifts divided by $\sqrt{\Delta}$:

$$V_l(T) \leq \frac{1}{\sqrt{\Delta}} \sum_{t=1}^{T} \|w_{t,l} - w_{t+1,l}\| = \frac{1}{\sqrt{\Delta}} \sum_{t=1}^{T} \|w_{t,l} - \varphi(z_{t,l}, w_{t,l})\| \leq \frac{1}{\sqrt{\Delta}} \sum_{t=1}^{T} C f(z_{t,l}, w_{t,l}) \ .$$

To bound the communication we need to bound the number of time points $t \in [T]$ where at least one learner l has a violation, denoted $V(T)$. In the worst case, all violations at all local learners occur at different time points,so that we can upper bound $V(T)$ by the sum of local violations $V_l(T)$ which is again upper bounded by the cumulative sum of drifts of all local models:

$$V(T) \leq \sum_{l=1}^{k} V_l(T) \leq \frac{1}{\sqrt{\Delta}} \sum_{t=1}^{T} \sum_{l=1}^{k} C f(z_{t,l}, w_{t,l}) = \frac{C}{\sqrt{\Delta}} L_P(T, k) \ .$$

Since dynamic averaging has communication costs of at most $\mathbf{c}_k$ per time point, the total amount of communication is

$$C_P(T, k) = \mathbf{c}_k V(T) \leq \mathbf{c}_k \frac{C}{\sqrt{\Delta}} L_P(T, k) \leq \underbrace{\mathbf{c}_k}_{\in \mathcal{O}(k)} \frac{C}{\sqrt{\Delta}} \mathbf{L}_P(T, k) \in \mathcal{O}\left(k \mathbf{L}_P(T, k)\right)$$

□

Since dynamic averaging is consistent in this setting, $\mathbf{L}_P(T, k) \in \mathcal{O}(\mathbf{L}_A(kT))$ and we can follow that $\mathbf{C}_P(T, k) \in \mathcal{O}\left(k \mathbf{L}_A(kT)\right)$, i.e., dynamic averaging is adaptive.

4 Conclusion

The analysis of the dynamic averaging protocol constitutes a first example of a relatively specific setting in which adaptivity and consistency can be achieved at the same time. Central for the future is to adapt dynamic averaging so that this holds also for more general settings. Moreover, tighter communication bounds are desirable which could be achieved by taking into account the communication reduction potential of partial synchronization and hedging strategies.

Acknowledgments

This research has been partially supported by the EU FP7-ICT-2013-11 under grant 619491 (FERARI).

References

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

Moshe Gabel, Daniel Keren, and Assaf Schuster. Communication-efficient distributed variance monitoring and outlier detection for multivariate time series. In *Proceedings of the 28th International Parallel and Distributed Processing Symposium (IPDPS)*, pages 37–47. IEEE, 2014.

Nikos Giatrakos, Antonios Deligiannakis, Minos Garofalakis, Izchak Sharfman, and Assaf Schuster. Prediction-based geometric monitoring over distributed data streams. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 265–276. ACM, 2012.

Michael Kamp, Mario Boley, Daniel Keren, Assaf Schuster, and Izchak Sharfman. Communication-efficient distributed online prediction by dynamic model synchronization. In *Machine Learning and Knowledge Discovery in Databases*, pages 623–639. Springer, 2014.

Daniel Keren, Izchak Sharfman, Assaf Schuster, and Avishay Livne. Shape sensitive geometric monitoring. *IEEE Transactions on Knowledge and Data Engineering*, 24(8):1520–1535, 2012.

Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon S Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 22, pages 1231–1239, 2009.

Izchak Sharfman, Assaf Schuster, and Daniel Keren. A geometric approach to monitoring threshold functions over distributed data streams. *Transactions on Database Systems (TODS)*, 32(4), 2007.

Martin Zinkevich, Markus Weimer, Alexander J. Smola, and Lihong Li. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2595–2603, 2010.