Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 50-1 (2017) 14106–14111

High-dimensional Online Adaptive Filtering ⋆

Sholeh Yasini * **Kristiaan Pelckmans** *

* *Department of Information Technology, Uppsala University, 752 37 Uppsala (e-mails: sholeh.yasini@ it.uu.se, kp@it.uu.se)*

Abstract: While recent advances in online learning arising from the *universal prediction* perspective have led to algorithms with prediction performance guarantee, these techniques are not able to cope with high-dimensional data. This paper analyses a random projection gradient descent (RP-GD) algorithm which addresses this challenge and yields low theoretical regret bound and computationally efficient algorithm. It is shown that the performance of the algorithm converges to the performance of the best offline, computationally complex, high-dimensional algorithm in hindsight.

Keywords: High-dimensional data, online gradient descent, random projection, regret bound.

1. INTRODUCTION

Classical stochastic online gradient descent (GD) methods such as Least Mean Square (LMS) and Recursive Least Square (RLS) algorithms have been extensively used for signal estimation and engineering applications (see e.g. the survey of Sayed (2008)). Their respective prediction and tracking performance are further substantiated by their relation to the Kalman filter (Kailath et al. (2000)). However, the standard ways of analysing these methods are limited in their applicability since (a) they typically assume stochastic assumptions on the random signals and (b) they rely on strong assumptions concerning a generative model of observations.

A class of prediction methods, receiving broad attention within the machine learning community, analyses the performance of online learning algorithms in a *minimax* worst-case setting, where no stochastic assumptions are made about the data, see Cesa-Bianchi and Lugosi (2006) for a comprehensive survey. This approach is based on the *universal prediction* perspective which attempts to perform well on any sequence of observations without assuming a generative model of the data. Instead of making assumption on the data generating process, the aim is to predict the sequence of data almost as well as the best offline reference predictor (comparator) in the comparison class of predictors that has access to all the data simultaneously.

The online learning algorithm is formulated as an iterative game between the *player* and the *environment*. At each round of this game, the player receives an input instance from the environment and is challenged to predict an unknown output value also generated by the environment. Then, the true output is revealed and the player suffers a (convex) loss function. The goal of the player is to minimize the total loss he suffers over all rounds. As no assumption is made on how the environment generates the sequence of outputs, the player would accumulate an arbitrary high loss. To set a reasonable goal, a class of *reference predictors*, denoted by $\mathcal{C}$, is fixed. The difference between the player's accumulated loss and that of a reference predictor in $\mathcal{C}$ is called *regret*. The aim of the player is to construct prediction strategies that guaratee a low regret with respect to any reference predictor in $\mathcal{C}$. In particular, the goal of the player is to achieve an average loss of prediction that is not too large compared to the average loss of the best reference predictor *in hindsight* (In this case, in hindsight means that the reference predictor corresponds to the output of an offline algorithm with access to all the data simultaneously). The possibility of achieving a low regret depends on the structure of the class of reference predictors and on the loss function.

Recently, there has been an increasing interest in the machine learning community for analyzing the Least Mean Square (LMS) algorithm, see Cesa-Bianchi et al. (1996), Cesa-Bianchi (1999), Hazan et al. (2007), Bartlett et al. (2008), and the Recursive Least Square (RLS) algorithm (Froster (1999), Vaits et al. (2015)) in a worst-case setting, see also Anava et al. (2013). Although efficient, however, these methods are not applicable to the setting where the dimensionality of the parameter to be estimated is very high since the derived regret bounds deteriorate fast in terms of the dimension.

The availability of *big data* today allows for identification of more complex physical models with larger number of parameters. Working with high dimensions data poses significant challenges, both in computational complexity and in accuracy. Modern developments in the research of machine learning, theoretical computer science and applied statistics has brought forward sufficient tools to handle large dimensional problems. Specifically, the method of compressed sensing as in Candes and Romberg (2005) has emerged as a complete framework (Foucart and Rauhut

⋆ This work is supported by Swedish Research Council under contract 621- 2007-6364. S. Yasini (corresponding author) and K. Pelckmans are with the Division of Systems and Control, Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden. (Emails: sholeh.yasini@it.uu.se , kp@it.uu.se).

10.1016/j.ifacol.2017.08.1851

(2013) and Yonina and Kutyniok (2012)) to tackle this problem under the assumption that the problem has a proper sparsity structure. More recent work has studied the role of sparsity in online learning methods such as RLS in Angelosante et al. (2010) and Dynamic Mirror Descent (DMD) algorithms in Hall and Willett (2015).

Random projection have recently emerged as a powerful technique for dimensionality reduction. This technique is an example of the power and elegance of the *probabilistic method* that has played an important role in a large variety of applications ranging from numerical analysis Sarlos (2006) to image and text processing Bingham and Mannila (2001). Theoretical results establish that the method preserves the pairwise Euclidean distances quite well (Vempala (2004)). It hence came as a surprise that this technique is not prominently studied in a context of automatic control and signal processing.

In this paper we analyse the random projected-based approximation of the online GD algorithm, in which the original high-dimensional GD algorithm is approximated by the solution of a lower-dimensional problem without making sparsity assumptions. Such dimensionality reduction is essential in case where the dimension of the covariates is larger than the number of rounds and where the regret bound grows fast with the dimension leading to poor performance of the algorithm. We prove a probabilistic regret bound for the proposed algorithm, *Random Projected GD* (RP-GD), which characterizes how well the algorithm performs relative to the best offline high-dimensional comutationally intractable GD algorithm.

This paper is organized as follow. In the next section, the problem formulation and the random projection technique for adaptive filtering is introduced. Then, the main results are given in Section 3 through analysing the performance of the RP-GD algorithm. Section 4 illustrates the theoretical results using simulation followed by the summary of this research.

2. RANDOM PROJECTION IN ADAPTIVE FILTERING

2.1 Preliminaries

We study the game-theoretic-based formulation of the online GD algorithm. To analyse the performance of the algorithm, we combine the parameter update mechanism from the classical GD (Sayed, 2008) with the online learning setting (regret setting) from universal prediction (Cesa-Bianchi and Lugosi (2006)). Consider the data $(\mathbf{x}_t, y_t), t = 1, \dots, T$ where $\mathbf{x}_t \in \mathbb{R}^d$ is the input vector assumed to be known, and $y_t \in \mathbb{R}$ is the output. We are interested in the setting where the dimensionality of the covariates d is very large, thus extending the previous results by Cesa-Bianchi et al. (1996), Cesa-Bianchi (1999) to the high-dimensional setting.

The game (as implemented here by the GD algorithm) proceeds in trials $t = 1, \dots, T$, where T is assumed to be known. The player maintains a parameter vector (hypothesis), denoted by $\mathbf{w}_{t-1} \in \mathbb{R}^d$. At each trial t the player receives the input vector $\mathbf{x}_t$ and makes a prediction $\hat{y}_t = \mathbf{w}_{t-1}^\top \mathbf{x}_t$. Then the actual output is revealed and the player incurs the corresponding loss $\ell(y_t, \hat{y}_t)$. Here, it is assumed that the loss function $\ell(y_t, \hat{y}_t)$ is convex, bounded and differentiable in its second argument with $\ell'(y_t, x)$ the derivative evaluated at $x = \hat{y}_t$.

Algorithm 1 Game-theoretic-based GD

Require: $\gamma_t > 0$.
Initialization: set $\mathbf{w}_0 = \mathbf{0}$.
for t = 1, 2, ..., T
(1) Player receives input instance $\mathbf{x}_t$
(2) Player predicts $\hat{y}_t = \mathbf{w}_{t-1}^T \mathbf{x}_t$.
(3) Environment reveals y_t
(4) Player scores loss $\ell(y_t, \hat{y}_t)$ and update estimate

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma_t \ell'(y_t, \hat{y}_t)\mathbf{x}_t. \quad (1)$$

end for

To any arbitrary reference predictor $\bar{\mathbf{w}}$, we associate a linear predictor $\bar{p}_t(\bar{\mathbf{w}}) \equiv \bar{y}_t = \bar{\mathbf{w}}^\top \mathbf{x}_t$, defined as $\bar{\mathbf{w}} \in \mathbb{R}^d$. Then any set $\mathfrak{C} \subseteq \mathbb{R}^d$ of vectors defines a comparison class $\mathcal{C}$ of linear predictors by $\mathcal{C} = \{\bar{p}_t(\bar{\mathbf{w}}) | \bar{\mathbf{w}} \in \mathfrak{C}, t = 1, \dots, T\}$. The loss of the reference predictor $\bar{p}_t(\bar{\mathbf{w}})$ is defined as $\ell(y_t, \bar{y}_t)$. It is assumed that $\ell(y_t, \bar{y}_t)$ belongs to a bounded family of loss functions for all t. The goal of the player is to minimize the *regret*, defined as the difference between the total loss of the player and the total loss of the reference predictor $\bar{p}(\bar{\mathbf{w}}) \in \mathcal{C}$

$$\mathcal{R}_T(\bar{\mathbf{w}}) = \sum_{t=1}^{T} \ell(y_t, \mathbf{w}_{t-1}^\top \mathbf{x}_t) - \sum_{t=1}^{T} \ell(y_t, \bar{\mathbf{w}}^\top \mathbf{x}_t). \quad (2)$$

In particular, the aim of the player is to incur total loss that is not much larger than the smallest total loss of any reference predictor in the comparison class $\mathcal{C}$

$$\begin{aligned} \mathcal{R}_T &\equiv \sup_{\bar{\mathbf{w}} \in \mathcal{C}} \mathcal{R}_T(\bar{\mathbf{w}}) \\ &= \sum_{t=1}^{T} \ell(y_t, \mathbf{w}_{t-1}^\top \mathbf{x}_t) - \inf_{\bar{\mathbf{w}} \in \mathcal{C}} \sum_{t=1}^{T} \ell(y_t, \bar{\mathbf{w}}^\top \mathbf{x}_t). \end{aligned} \quad (3)$$

We restate the players aim as having a *low regret*, by which we mean that $\mathcal{R}_T$ grows sublinearly with the number of rounds T, i.e., $\mathcal{R}_T = o(T)$. Intuitively, the regret measures how sorry the player is, after seeing all the data (i.e., in hindsight), for not having followed the predictions of the best reference predictor in the comparison class $\mathcal{C}$. Previous work Cesa-Bianchi (1999) and Froster (1999) obtained sub-linear regret bounds for the online GD and online second order algorithms. However, these bounds deteriorate fast in terms of the dimension of covariates d. We show how the regret grows in the online setting where the dimension d is large.

2.2 Random Projected Gradient Descent (RP-GD)

The celebrated Johnson-Lindenstrauss lemma states that, given an arbitrary set of T points in a (high-dimensional) Euclidean space, there exists a random linear embedding of these points in a k-dimensional Euclidean space such that all pairwise distances are preserved within a factor of $1 \pm \epsilon$ if k is proportional to $(\log T/\epsilon^2)$ (Vempala (2004) and Boucheron et al. (2012)).