Some Impossibility Results for Budgeted Learning

Nicolò Cesa-Bianchi CESA-BIANCHI@DSI.UNIMI.IT
DSI, Università degli Studi di Milano, Italy

Shai Shalev-Shwartz SHAIS@CS.HUJI.AC.IL
The Hebrew University, Jerusalem, Israel

Ohad Shamir OHADSH@CS.HUJI.AC.IL
The Hebrew University, Jerusalem, Israel

Abstract

We prove two impossibility results for budgeted learning with linear predictors. The first result shows that no budgeted learning algorithm can in general learn an ϵ-accurate d-dimensional linear predictor while observing less than d/ϵ attributes at training time. Our second result deals with the setting studied by Greiner et al. (2002), where the learner has all the information at training time and at test time he has to form a prediction after observing a fixed amount of attributes per each instance. We formally prove that while it is possible to learn a consistent predictor accessing at most two attributes of each example at training time, it is not possible (even with an infinite amount of training examples) to build an active classifier that uses at most two attributes of each example at test time, and whose error will be smaller than a constant.

1. Preliminaries

We consider the problem of learning linear predictors on a budget. In linear regression each example is an instance-target pair, $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$. We refer to $\mathbf{x}$ as a vector of attributes and the goal of the learner is to find a linear predictor $\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$, where we refer to $\mathbf{w} \in \mathbb{R}^d$ as the predictor. To do so, the learning algorithm receives a training set of m examples, $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, which are assumed to be sampled i.i.d. from an unknown distribution $\mathcal{D}$ over pairs $(\mathbf{x}, y)$. We distinguish between four scenarios:

- **Full information:** The learner receives the entire training set. This is the traditional linear regression setting.
- **Local Budget Constraint:** For each individual example, $(\mathbf{x}_i, y_i)$, the learner receives the target y_i but is only allowed to see k attributes of $\mathbf{x}_i$, where k is a parameter of the problem. The learner has the freedom to actively choose *which* of the attributes will be revealed, as long as at most k of them will be given. This setting was first proposed in (Ben-David and Dichterman, 1998), where it is called "learning with restricted focus of attention", and in the context of regression it was recently studied by (Cesa-Bianchi et al., 2010).
- **Global Budget Constraint:** The total number of training attributes the learner is allowed to see is bounded by k. As in the local budget constraint setting, the learner has the freedom to actively choose which of the attributes will be revealed. In contrast to the local budget constraint setting, the learner can choose to reveal more than k/m attributes from specific examples as long as the global number of attributes is bounded by k. This setting was recently studied by several authors — see for example (Deng et al., 2007; Kapoor and Greiner, 2005) and the references therein.
- **Prediction on a budget:** The learner receives the entire training set, however, at test time, the predictor can see at most k attributes of each instance and then must form a prediction. The predictor is allowed to actively choose which of the attributes will be revealed. This setting was proposed and studied by Greiner et al. (2002).

In all cases the goal of the of the learner is to find a predictor with low risk, defined as the expected loss $L_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\langle \mathbf{w}, \mathbf{x} \rangle, y)]$. For simplicity we focus on the squared loss function, $\ell(a, b) = (a-b)^2$. We

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

denote the training loss by $L_S(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}(\langle \mathbf{w}, \mathbf{x}_i\rangle - y_i)^2$.

2. Main Results

2.1. Learning on a budget

In this section we show that any budget learning algorithm (local or global) needs in general a budget of d/ϵ attributes for learning a d-dimensional, ϵ-accurate, linear predictor.

Theorem 1 *For any $\epsilon \in (0, 1/16)$, there exists a distribution over examples and a weight vector $\mathbf{w}^\star \in \mathbb{R}^d$, with $\|\mathbf{w}^\star\|_0 = 1$ and $\|\mathbf{w}^\star\|_2 = \|\mathbf{w}^\star\|_1 = 2\sqrt{\epsilon}$, such that any learning algorithm must see $\Omega\left(\frac{d}{\epsilon}\right)$ attributes in expectation in order to learn a linear predictor $\mathbf{w}$ with $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathcal{D}}(\mathbf{w}^\star) < \epsilon$.*

The proof is given in the next section. Note that Cesa-Bianchi et al. (2010) proved that under the assumptions given in Theorem 1, it is possible to learn an ϵ-accurate predictor using a local budget of two attributes per examples and using total $O(d^2/\epsilon)$ training examples. Therefore, $O(d^2/\epsilon)$ attributes are sufficient for learning an ϵ-accurate predictor in this case. That is, we have a gap of factor d between the lower bound and the upper bound, and it remains open to bridge this gap.

2.2. Predicting on a budget

We now consider the "Prediction on a budget" setting. Greiner et al. (2002) studied this setting and showed positive results regarding (agnostic)-PAC learning of *k-active* predictors, namely predictors that are restricted to use at most k attributes per test example. In particular, they show that it is possible to learn a k-active predictor from training examples whose performance is slightly worse than that of the best k-active predictor.

But, how good are the predictions of the best k-active predictor? In this section we show that even in simple cases in which there exists a linear predictor $\mathbf{w}^\star$ with $L_{\mathcal{D}}(\mathbf{w}^\star) = 0$, the risk of the best k-active predictor can be high.

The following theorem shows that if the only constraint on $\mathbf{w}^\star$ is bounded ℓ_2 norm, then the risk can be as high as $1 - k/d$. We use the notation $L_{\mathcal{D}}(A)$ to denote the expected loss of the k-active predictor on a test example.

Theorem 2 *There exists a weight vector $\mathbf{w}^\star \in \mathbb{R}^d$ and a distribution $\mathcal{D}$ such that $\|\mathbf{w}^\star\|_2 = 1$ and $L_{\mathcal{D}}(\mathbf{w}^\star) = 0$ while any algorithm A that gives predictions $A(\mathbf{x})$ while viewing only $k < d$ attributes of each $\mathbf{x}$ must have $L_{\mathcal{D}}(A) \geq 1 - k/d$.*

The proof is given in the next Section. Note that the risk of the constant prediction of zero is 1. Therefore, the theorem tells us that no active predictor can get an improvement over the naive predictor of more than k/d.

It is well known that a low ℓ_1 norm of $\mathbf{w}^\star$ encourages sparsity of the learned predictor, which naturally helps in designing active predictors. The following theorem shows that even if we restrict $\mathbf{w}^\star$ to have $\|\mathbf{w}^\star\|_1 = 1$, $L_{\mathcal{D}}(\mathbf{w}^\star) = 0$, and $\|\mathbf{w}^\star\|_0 > k$, we still have that the risk of the best k-active predictor can be non-vanishing.

Theorem 3 *There exists a weight vector $\mathbf{w}^\star \in \mathbb{R}^d$ and a distribution $\mathcal{D}$ such that $\|\mathbf{w}^\star\|_1 = 1$, $L_{\mathcal{D}}(\mathbf{w}^\star) = 0$, and $\|\mathbf{w}^\star\|_0 = ck$ (for $c > 1$) such that any algorithm A that gives predictions $A(\mathbf{x})$ while viewing only $k < ck \leq d$ attributes of each $\mathbf{x}$ must have $L_{\mathcal{D}}(A) \geq \left(1 - \frac{1}{c}\right)\frac{1}{ck}$.*

The proof is given in the next Section. Two examples are given below:

- Choose $c = 2$, then $\|\mathbf{w}^\star\|_0 = 2k$ and $L_{\mathcal{D}}(A) \geq 1/(4k)$
- Choose $c = (k+1)/k$, then $\|\mathbf{w}^\star\|_0 = k+1$ and $L_{\mathcal{D}}(A) \geq \frac{1}{(k+1)^2}$

Note that if $\|\mathbf{w}^\star\|_0 \leq k$ there is a trivial way to predict on a budget of k attributes by always querying the attributes corresponding to the non-zero elements of $\mathbf{w}^\star$.

These negative results highlight an interesting phenomenon: In (Cesa-Bianchi et al., 2010) it is shown that one can learn an arbitrarily accurate predictor $\mathbf{w}$ with a local budget of $k = 2$. However, here we show that even if we know the optimal $\mathbf{w}^\star$, we might not be able to accurately predict a new partially observed example unless k is very large. Therefore, learning on a budget is much easier than predicting on a budget.

3. Proofs

3.1. Proof of Theorem 1

The proof is an extension to global budget constraints of the proof of Theorem 3 in (Cesa-Bianchi et al., 2010) for local budget. Here we only sketch the main differences. We define the data distribution as follows: First, $j \in \{1, \ldots, d\}$ is drawn uniformly at random. Then, we generate $y_1, y_2, \ldots \in \{\pm 1\}$ i.i.d. according

to $\mathbb{P}\big[y_t = 1\big] = \mathbb{P}\big[y_t = -1\big] = \frac{1}{2}$. Given j and y_t, $\mathbf{x}_t \in \{\pm 1\}$ is generated according to $\mathbb{P}\big[x_{t,i} = y_t\big] = \frac{1}{2} + \mathbb{1}_{[i=j]}\sqrt{\epsilon}$. Just like in the proof of Theorem 3 of Cesa-Bianchi et al. (2010), one can show that for this distribution the value of j can be identified from any ϵ-good predictor; that is, any $\mathbf{w} \in \mathbb{R}^d$ whose risk $L_{\mathcal{D}}(\mathbf{w}^\star)$ is strictly smaller than $L_{\mathcal{D}}(\mathbf{w}^\star) + \epsilon$, where $\mathbf{w}^\star$ is the linear predictor with smallest risk.

We now define an instance of the multi-armed bandit problem based on this data distribution. Each coordinate $i \in \{1, \ldots, d\}$ is an arm and the reward of pulling i at time t is $\frac{1}{2}\big|x_{N_{i,t},i} + y_{N_{i,t}}\big| \in \{0, 1\}$, where $N_{i,t}$ denotes the number of times arm i has been pulled in the first t plays. Hence the expected reward of pulling i is $\frac{1}{2} + \mathbb{1}_{[i=j]}\sqrt{\epsilon}$. At the end of each round t the player observes $x_{N_{i,t},i}$ and $y_{N_{i,t}}$. Note that if $N_{i',s} = N_{i,t}$ for some $i' \neq i$ and $s < t$, then $y_{N_{i,t}}$ was already observed at play s, but this does not provide additional information to the player as $\mathbb{P}\big[x_{i,s} = y_s\big] = \mathbb{P}\big[x_{i,s} = y_s \mid y_s\big]$ for all s.

Now take an arbitrary learning algorithm that finds an ϵ-good predictor under a global budget constraint of k. The expected reward of the bandit player that runs the learner for the first k rounds, and for the remaining $T-k$ rounds always chooses the coordinate j identified from the learned predictor, is then at least $\frac{k}{2} + (T - k)\big(\frac{1}{2} + \sqrt{\epsilon}\big) = \frac{T}{2} + (T - k)\sqrt{\epsilon}$. Moreover, using the bandit lower bound of Auer et al. (2003), this expected reward is at most $\frac{T}{2} + T\sqrt{\epsilon}\left(\frac{1}{d} + \sqrt{\frac{6}{d}T\epsilon}\right)$. Combining upper and lower bound, choosing T of order $\frac{d}{\epsilon}$, and solving for k gives $k = \Omega\big(\frac{d}{\epsilon}\big)$, as desired.

3.2. Proof of Theorem 2

For any $d > k$ let $\mathbf{w}^\star = \big(1/\sqrt{d}, \ldots, 1/\sqrt{d}\big)$. Let $\mathbf{x} \in \{\pm 1\}^d$ be distributed uniformly at random and y is determined deterministically to be $\langle \mathbf{w}^\star, \mathbf{x}\rangle$. Then, $L_{\mathcal{D}}(\mathbf{w}^\star) = 0$ and $\|\mathbf{w}^\star\|_2 = 1$. Without loss of generality, suppose the prediction algorithm asks for the first k attributes of a test example and form its prediction to be $\hat{y}$. Since the generation of attributes is independent, we have that the value of $x_{k+1}, \ldots, x_d$ does not depend on $x_1, \ldots, x_k$, and on $\hat{y}$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[(\hat{y} - \langle \mathbf{w}^\star, \mathbf{x}\rangle)^2\right] \\
&= \mathbb{E}\left[\left(\hat{y} - \sum_{i=1}^{k} w_i^\star x_i - \sum_{i>k} w_i^\star x_i\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{y} - \sum_{i=1}^{k} w_i^\star x_i\right)^2\right] + \sum_{i>k} (w_i^\star)^2\, \mathbb{E}[x_i^2] \\
&\geq 0 + \frac{d-k}{d} = 1 - \frac{k}{d}
\end{aligned}$$

which concludes our proof.

3.3. Proof of Theorem 3

Let

$$\mathbf{w}^\star = \left(\underbrace{\tfrac{1}{ck}, \ldots, \tfrac{1}{ck}}_{ck \text{ elements}}, 0, \ldots, 0\right)$$

and let $\mathbf{x} \in \{\pm 1\}^d$ be distributed uniformly at random and y is determined deterministically to be $\langle \mathbf{w}^\star, \mathbf{x}\rangle$. Then, $L_{\mathcal{D}}(\mathbf{w}^\star) = 0$, $\|\mathbf{w}^\star\|_1 = 1$, and $\|\mathbf{w}^\star\|_0 = ck$. Without loss of generality, suppose the prediction algorithm asks for the first k attributes of a test example and form its prediction to be $\hat{y}$. Since the generation of attributes is independent, we have that the value of $x_{k+1}, \ldots, x_d$ does not depend on $x_1, \ldots, x_k$, and on $\hat{y}$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[(\hat{y} - \langle \mathbf{w}^\star, \mathbf{x}\rangle)^2\right] \\
&= \mathbb{E}\left[\left(\hat{y} - \sum_{i=1}^{k} w_i^\star x_i - \sum_{i>k} w_i^\star x_i\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{y} - \sum_{i=1}^{k} w_i^\star x_i\right)^2\right] + \sum_{i>k} (w_i^\star)^2\, \mathbb{E}[x_i^2] \\
&\geq 0 + \frac{ck-k}{(ck)^2} \\
&= \frac{c-1}{c^2 k} = \big(1 - \tfrac{1}{c}\big)\, \frac{1}{ck}
\end{aligned}$$

which concludes our proof.

References

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SICOMP: SIAM Journal on Computing*, 32, 2003.

S. Ben-David and E. Dichterman. Learning with restricted focus of attention. *JCSS: Journal of Computer and System Sciences*, 56, 1998.

Nicolò Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Efficient learning with partially observed attributes. In *ICML*, 2010.

Kun Deng, Chris Bourke, Stephen D. Scott, Julie Sunderman, and Yaling Zheng. Bandit-based algorithms for budgeted learning. In *ICDM*, pages 463–468. IEEE Computer Society, 2007.

Russell Greiner, Adam J. Grove, and Dan Roth. Learning cost-sensitive active classifiers. *Artificial Intelligence*, 139(2):137–174, 2002.

Aloak Kapoor and Russell Greiner. Learning and classifying under hard budgets. In *ECML*, pages 170–181, 2005.