Empirical Bernstein Inequality for Martingales : Application to Online Learning

Thomas Peel, Sandrine Anthoine, Liva Ralaivola

Empirical Bernstein Inequality for Martingales : Application to Online Learning

Thomas Peel[1,2] **Sandrine Anthoine**[1] **Liva Ralaivola**[2]
[1]Aix-Marseille Université - CNRS
LATP, UMR 7353
Marseille, France
[2]Aix-Marseille Université - CNRS
LIF, UMR 7279
Marseille, France

Abstract

In this article we present a new empirical Bernstein inequality for bounded martingale difference sequences. This inequality refines the one by Freedman [1975] and is then used in order to bound the average risk of the hypotheses during an online learning process. We show theoretical and empirical evidences of the tightness of our result compared with the state of the art bound provided by Cesa-Bianchi and Gentile [2008].

1 INTRODUCTION

The motivation behind this work comes from the wish to analyze the risk of the models (or hypotheses) produced by an online learning algorithm. Such an algorithm works incrementally on a sequence of independent and identically distributed (i.i.d.) random variables. At each step, it receives an example that is used in order to update the current model parameters. Once this update is done, the performance of the new hypothesis is measured by evaluating its loss on the next example of the sequence and so on. By averaging these losses, one can define a statistic $\hat{R}_n$ called *empirical instantaneous risk*. The risk of a model is simply the expectation of its loss on a new unseen example given the sequence of data used in its construction. In their recent works, Cesa-Bianchi et al. [2004] and Cesa-Bianchi and Gentile [2008] show how the statistic $\hat{R}_n$ can be used for selecting a hypothesis with a low risk. The key tool in their analyses is the use of concentration inequalities for martingales (Azuma-Hoeffding, Bernstein). Indeed, the dependencies existing between the hypotheses that are inherent to online learning processes prevent the use of standard concentration inequalities that require independence.

Bernstein (second-order) inequalities are known to be tighter than their first-order counterparts. However, the variance is in general unknown and need to be upper bounded. Recent works in the *batch* setting have proposed an empirical (data-dependent) version of the Bernstein inequality [Maurer and Pontil, 2009, Peel et al., 2010] where an estimator of the variance is used as upper bound. However, these inequalities are not applicable to the online learning setting. In this paper, we propose a new Bernstein inequality for bounded martingale difference sequences (Theorem 2) that takes advantage of the statistic $\hat{V}_n$, an instantaneous estimator of the variance. This inequality is then used in order to refine the tail bound by Cesa-Bianchi and Gentile [2008]. Briefly, we show that under the same assumptions they make, the average risk of the hypotheses produced by an online learning algorithm is bounded with high probability by

$$\hat{R}_n + \frac{1}{n}\sqrt{\beta_n \ln\left(\frac{2}{\delta}\right)} + \frac{2}{3n}\ln\left(\frac{2}{\delta}\right),$$

where β_n is a function of $\hat{V}_n$ we will detail later. This bound can be applied to any online algorithm and as an example we show how to use it to characterize the average risk of the hypotheses produced by Pegasos [Shalev-Shwartz et al., 2011], a stochastic method for solving the SVM optimization problem.

We want to emphasize that the scope of our new empirical Bernstein inequality for martingales goes far beyond any application to online learning processes.

The paper is organized as follows. In Section 2 we recall a few fundamental notions about martingales and the classical concentration inequalities associated with this kind of random processes. Section 3 presents the main result of this paper, a concentration inequality that takes advantage of a second order empirical information in the martingale setting. This one is then applied in Section 4 to get a bound on the mean generalization error made by the hypotheses learned during an online learning process. This bound substantially improves the results mentionned above. We end this

paper with Section 5, a direct consequence of the previous inequalities that let us bound the mean risk of the weight vectors generated during a run of the Pegasos algorithm.

2 PRELIMINARIES

This section briefly reminds basic notions about the martingale theory and the classical concentration inequalities associated with this kind of stochastic processes.

2.1 Martingale and Martingale Difference Sequence

Definition 1 (Martingale)**.** A sequence $\{M_n : 0 \leq n < \infty\}$ of random variables is said to be a *martingale* with respect to the sequence of random variables $\{X_n : 1 \leq n < \infty\}$ if the sequence $\{M_0, \ldots, M_n\}$ has two basic properties. The first one is that for each $n \geq 1$ there is a function $f_n : \mathbb{R}^n \to \mathbb{R}$ such that $M_n = f_n(X_1, X_2, \ldots, X_n)$. The second property is that the sequence $\{M_n\}$ satisfies for all $n \geq 1$:

$$\mathbb{E}\left[|M_n|\right] < \infty \tag{1}$$
$$\mathbb{E}\left[M_n \,|\, X_1, \ldots, X_{n-1}\right] = M_{n-1}. \tag{2}$$

Given this definition of a martingale, we now define a martingale difference sequence.

Definition 2 (Martingale difference sequence)**.** We say that a sequence of random variables $\{Y_n : 0 \leq n < \infty\}$ is a *martingale difference sequence* (MDS) if the sequence $\{Y_n\}$ satisfies the following properties for all $n \geq 1$:

$$\mathbb{E}\left[|Y_n|\right] < \infty \tag{3}$$
$$\mathbb{E}\left[Y_n \,|\, Y_1, \ldots, Y_{n-1}\right] = 0. \tag{4}$$

By construction, this implies that if the sequence $\{M_n\}$ is a martingale then the sequence $\{Y_n = M_n - M_{n-1}\}$ is a martingale difference sequence. We now introduce two well-known concentration inequalities about the sum of the increments of a MDS that we will use in the next sections.

2.2 Azuma-Hoeffding Inequality

The Azuma-Hoeffding inequality [Hoeffding, 1963, Azuma, 1967] gives a result about the concentration of the values of a martingale with bounded increments around his initial value M_0.

Theorem 1 (Azuma-Hoeffding inequality)**.** *Let $\{M_n\}$ be a martingale and define $\{Y_n = M_n - M_{n-1}\}$ the associated martingale difference sequence such that $|Y_i| \leq c_i$ for all $1 \leq i \leq n$. Then, for all $\epsilon > 0$*

$$\mathbb{P}\left[\sum_{i=1}^{n} Y_i = M_n - M_0 \geq \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^{n} {c_i}^2}\right). \tag{5}$$

This result makes it possible to extend the Hoeffding inequality [Hoeffding, 1963] to the case where the random variables of interest are not necessarily independent.

Corollary 1. *Let $X_1, \ldots, X_n$ be a sequence of random variables such that for all $1 \leq i \leq n$ we have $|\mathbb{E}\left[X_i \,|\, X_1, \ldots, X_{i-1}\right] - X_i| \leq c_i$. Set $S_n = \sum_{i=1}^{n} X_i$, then for all $\epsilon > 0$*

$$\mathbb{P}\left[\sum_{i=1}^{n} \mathbb{E}\left[X_i \,|\, X_1, \ldots, X_{i-1}\right] - S_n \geq \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^{n} {c_i}^2}\right). \tag{6}$$

Proof. A direct application of Theorem 1 to the martingale difference sequence $\{Y_n\}$ such that $Y_i = \mathbb{E}\left[X_i \,|\, X_1, \ldots, X_{i-1}\right] - X_i$ gives the result. □

2.3 Bernstein Inequality for Martingales

The inequality we recall in the following lemma is a consequence of the Bernstein inequality for martingales given in Freedman [1975]. This lemma extends the classical Bernstein inequality [Bennett, 1962] which requires independence between the random variables X_i under consideration. This limitation is overcome by looking at the martingale difference sequence $\{Y_n = \mathbb{E}\left[X_n | X_1, \ldots, X_{n-1}\right] - X_n\}$.

Lemma 1 (Bernstein inequality for martingales)**.** *Suppose $X_1, \ldots, X_n$ is a sequence of random variables such that $0 \leq X_i \leq 1$. Define the martingale difference sequence $\{Y_n = \mathbb{E}\left[X_n | X_1, \ldots, X_{n-1}\right] - X_n\}$ and note K_n the sum of the conditional variances*

$$K_n = \sum_{t=1}^{n} \mathbb{V}\left[X_n | X_1, \ldots, X_{n-1}\right]. \tag{7}$$

Let $S_n = \sum_{i=1}^{n} X_i$, then for all $\epsilon, v \geq 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n} \mathbb{E}\left[X_n | X_1, \ldots, X_{n-1}\right] - S_n \geq \epsilon, K_n \leq k\right] \leq \exp\left(-\frac{\epsilon^2}{2k + 2\epsilon/3}\right). \tag{8}$$

As we shall see, this lemma is central in our analysis as it was in the work by Cesa-Bianchi and Gentile [2008].

3 EMPIRICAL BERNSTEIN INEQUALITY FOR MARTINGALES

Second order Bernstein inequalities are known to be tighter than their first order counterparts thanks to the variance term. However, in practice, this term often can not be evaluated and it is common to upper bound it by the expectation (making the assumption that the random variables under interest are bounded by 1) in order to compute the whole inequality. We propose another approach based on the use of an instantaneous estimator of the variance instead of the usual approach. By doing so, we hope to get a tighter inequality without any a priori assumption on the underlying distribution of the random variables. This section presents the main result of the paper, a refined version of Bernstein inequality for martingales recalled above where the sum of conditional variances is upper bounded using an instantaneous estimator. We first introduce the inequality reversal lemma, which allows us to transform tail inequalities into upper bounds (or confidence intervals). This lemma has been used by Peel et al. [2010] to prove their empirical Bernstein inequality for U-Statistics.

Lemma 2 (Inequality reversal lemma)**.** *Let X be a random variable and $a, b>0, c, d\geq 0$ such that*

$$\forall \varepsilon > 0,\ \mathbb{P}_X\left[|X| \geq \varepsilon\right] \leq a \exp\left\{-\frac{b\varepsilon^2}{c+d\varepsilon}\right\}, \quad (9)$$

then, with probability at least $1-\delta$

$$|X| \leq \sqrt{\frac{c}{b}\ln\frac{a}{\delta}} + \frac{d}{b}\ln\frac{a}{\delta}. \quad (10)$$

Proof. Solving for ε such that the right hand side of (9) is equal to δ gives:

$$\varepsilon = \frac{1}{2b}\left(d\ln\frac{a}{\delta} + \sqrt{d^2\ln^2\frac{a}{\delta} + 4bc\ln\frac{a}{\delta}}\right).$$

Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ gives an upper bound on ε and provides the result. □

We use the notation $f_{\{Z_t\}}$ in order to indicate a function determined by the sequence of random variables $\{Z_t\} = \{Z_1, \ldots, Z_t\}$ i.e. the expression of $f_{\{Z_t\}}$ is fixed by the sequence $\{Z_t\}$. The next theorem is the main result of this paper.

Theorem 2 (Empirical Bernstein inequality for martingales)**.** *Let $Z_1, \ldots, Z_n$ be a sequence of random variables following the same probability distribution $\mathcal{D}$ such that Z_{t+1}, Z_{t+2} are conditionally independent given $\{Z_t\}$, for all $1 \leq t \leq n$. Suppose $\{f_{\{Z_t\}}\}_{t=1}^n$ is a family of functions which take their values in $[0,1]$, then for all $0 < \delta \leq 1$ we have with probability at least $1-\delta$*

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[f_{\{Z_t\}}(Z_{t+1}) \,|\, Z_1, \ldots, Z_t\right] \leq \frac{1}{n}\sum_{i=1}^{n} f_{\{Z_t\}}(Z_{t+1}) + \frac{1}{n}\sqrt{\beta_n \ln\left(\frac{2}{\delta}\right)} + \frac{2}{3n}\ln\left(\frac{2}{\delta}\right), \quad (11)$$

where

$$\beta_n = n\hat{V}_n + \sqrt{\frac{n}{2}\ln\left(\frac{2}{\delta}\right)}, \quad (12)$$

and

$$\hat{V}_n = \frac{1}{2n}\sum_{t=1}^{n}\left(f_{\{Z_t\}}(Z_{t+1}) - f_{\{Z_t\}}(Z_{t+2})\right)^2. \quad (13)$$

In a nutshell, the message carried by this theorem is that it is possible to use an instantaneous variance estimator to quantify the deviation of the sum $\frac{1}{n}\sum_{i=1}^{n} f_{\{Z_t\}}(Z_{t+1})$ from its expected value. In order to prove the previous concentration inequality, we need an intermediate result about the conditional variance estimator introduced in Equation (13). In essence, the following lemma allows us to quantify the deviation of this estimator from the sum V_n of the conditional variances:

$$V_n = \sum_{t=1}^{n}\mathbb{V}\left[f_{\{Z_t\}}(Z) \,|\, Z_1, \ldots, Z_t\right]. \quad (14)$$

Lemma 3. *Let $Z_1, \ldots, Z_n$ be a sequence of random variables following the same probability distribution $\mathcal{D}$ such that Z_{t+1}, Z_{t+2} are conditionally independent given $\{Z_t\}$, for all $1 \leq t \leq n$. Suppose $\{f_{\{Z_t\}}\}_{t=1}^n$ is a family of functions which take their values in $[0,1]$, then for all $0 < \delta \leq 1$,*

$$\mathbb{P}\left[V_n \geq n\hat{V}_n + \sqrt{\frac{n}{2}\ln\left(\frac{1}{\delta}\right)}\right] \leq \delta\,. \quad (15)$$

Proof. We begin this proof by defining the sequence of random variables $\{M_n\}$ such that for all $1 \leq t \leq n$,

$$M_t = \frac{1}{2}\left(f_{\{Z_t\}}(Z_{t+1}) - f_{\{Z_t\}}(Z_{t+2})\right)^2,$$

and the associated martingale difference sequence

$$\{A_n = \mathbb{E}\left[M_n \,|\, Z_1, \ldots, Z_n\right] - M_n\}\,.$$

Using the fact that the Z_t follow the same distribution and that Z_{t+1}, Z_{t+2} are conditionally independent we get that

$$\mathbb{E}\left[M_t \,|\, Z_1, \ldots, Z_t\right] = \mathbb{V}\left[f_{\{Z_t\}}(Z) \,|\, Z_1, \ldots, Z_t\right].$$

It follows that

$$\frac{1}{n}\sum_{t=1}^{n} A_t = \frac{1}{n}\sum_{t=1}^{n} \mathbb{V}\left[f_{\{Z_t\}}(Z) \,|\, Z_1, \ldots, Z_t\right] - \hat{V}_n$$
$$= \frac{1}{n} V_n - \hat{V}_n.$$

Noting that $M_t \in [0, \frac{1}{2}]$ because f takes its values in $[0,1]$ entails $\mathbb{E}\left[M_t \,|\, Z_1, \ldots, Z_t\right] \in [0, \frac{1}{2}]$ and furthermore each term of the sequence $\{A_n\}$ is bounded:

$$-\frac{1}{2} \leq A_t \leq \frac{1}{2}.$$

Consequently $\{A_n\}$ is a bounded martingale difference sequence on which we can apply the Azuma-Hoeffding inequality (Theorem 1) to obtain

$$\mathbb{P}\left[\frac{1}{n} V_n - \hat{V}_n \geq \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{n}\right).$$

We conclude the proof by using Lemma 2. □

Thanks to this first result, we can now prove Theorem 2.

Proof. (Theorem 2) Define the sequence of random variables $\{M_n\}$ such that

$$M_i = f_{\{Z_i\}}(Z),$$

and the associated martingale difference sequence

$$\{A_n = \mathbb{E}\left[M_n \,|\, Z_1, \ldots, Z_n\right] - M_n\}.$$

Remark that for β_n as in Equation (12) and s fixed

$$\mathbb{P}\left[\sum_{t=1}^{n} A_t \geq s\right] = \mathbb{P}\left[\sum_{t=1}^{n} A_t \geq s, V_n \geq \beta_n\right] + \mathbb{P}\left[\sum_{t=1}^{n} A_t \geq s, V_n < \beta_n\right].$$

We need to upper bound the two parts of the right hand side of the previous equation in order to get the desired bound on the left hand side. Remark that $\mathbb{P}\left[\sum_{t=1}^{n} A_t \geq s, V_n \geq \beta_n\right] \leq \mathbb{P}\left[V_n \geq \beta_n\right]$. We use Lemma 3 to bound $\mathbb{P}\left[V_n \geq \beta_n\right]$ and obtain

$$\mathbb{P}\left[\sum_{t=1}^{n} A_t \geq s, V_n \geq \beta_n\right] \leq \frac{\delta}{2}. \tag{16}$$

Then, by using the Bernstein inequality for martingales (Lemma 1) on the martingale difference sequence $\{A_n\}$ we have

$$\mathbb{P}\left[\sum_{t=1}^{n} A_t \geq s, V_n < b\right] \leq \exp\left(-\frac{s^2}{2b + 2s/3}\right), \tag{17}$$

which we can write alternatively

$$\mathbb{P}\left[\sum_{t=1}^{n} A_t \geq \sqrt{b \ln\left(\frac{2}{\delta}\right)} + \frac{2}{3}\ln\left(\frac{2}{\delta}\right), V_n < b\right] \leq \frac{\delta}{2}, \tag{18}$$

thanks to Lemma 2. We conclude the proof by setting $b = \beta_n$ in (18) and

$$s = \sqrt{\beta_n \ln\left(\frac{2}{\delta}\right)} + \frac{2}{3}\ln\left(\frac{2}{\delta}\right),$$

in Equation (16). □

In the upcoming section, we use Theorem 2 in an online learning setting. More precisely, we employ our result with the intention of characterizing the mean of the risks

$$\frac{1}{n}\sum_{t=0}^{n-1} R(h_t)$$

associated with the hypotheses learned during such a process.

4 APPLICATION TO ONLINE LEARNING

Before stating the main theorem of this section, we recall the online learning setting and define a new instantaneous estimator of the conditional variance well suited for an online learning procedure.

4.1 Online Learning and Instantaneous Conditional Variance Estimator

There is no formal definition of an online learning process, even in reference works as Littlestone et al. [1995] or Shalev-Shwartz [2007]. One generally defines it as follows. Consider a dataset $\underline{Z}_n = \{z_i\}_{i=1}^n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ of independent and identically distributed random variables with respect to an unknown probability distribution $\mathcal{D}$ on the product space $\mathcal{X} \times \mathcal{Y}$. An online learning algorithm working with the set $\underline{Z}_n$ produces a set $\{h_0, \ldots, h_n\}$ of hypotheses where each $h_t : \mathcal{X} \to \tilde{\mathcal{Y}}$ aims at predicting the class of a new example $\boldsymbol{x}$ drawn from $\mathcal{D}$. From an initial hypothesis h_0 and the first datum $(\boldsymbol{x}_1, y_1)$ the algorithm produces a new hypothesis h_1. This new hypothesis is a function of the random variable $z_1 = (\boldsymbol{x}_1, y_1)$ (and the hypothesis h_0). It then uses the next example $(\boldsymbol{x}_2, y_2)$ and the hypothesis h_1 to generate a second hypothesis h_2 and so on. At the end of the learning process, the algorithm outputs the set $\{h_0, \ldots, h_n\}$ where each hypothesis h_t is constructed using the previous hypothesis h_{t-1} and the example $(\boldsymbol{x}_t, y_t)$. Thus each hypothesis h_t depends on the sequence of random

variables $\{z_1, \dots, z_t\}$. We use a bounded loss function $\ell : \tilde{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}^+$ in order to evaluate the performance of an hypothesis. The risk of the hypothesis h_t, denoted by $R(h_t) = \mathbb{E}\left[\ell(h_t(X), Y) \,|\, z_1, \dots, z_t\right]$, is simply the expectation of the loss function ℓ conditionally to the random variables $\{z_1, \dots, z_t\}$. Obviously, this quantity is unknown since $\mathcal{D}$ is unknown. In this article, we assume that the loss function is such that $\ell \in [0,1]^{\tilde{\mathcal{Y}} \times \mathcal{Y}}$. It is important to notice that this assumption does not limit the scope of the results presented hereafter.

A common wish in online learning is to characterize the mean risk

$$\frac{1}{n}\sum_{t=0}^{n-1} R(h_t) = \frac{1}{n}\sum_{t=0}^{n-1} \mathbb{E}\left[\ell(h_t(X), Y) \,|\, z_1, \dots, z_t\right], \quad (19)$$

associated with the hypotheses produced by an algorithm using an online estimator $\hat{R}_n$ such that

$$\hat{R}_n = \hat{R}_n(\underline{Z}_n) = \frac{1}{n}\sum_{t=0}^{n-1} \ell(h_t(\boldsymbol{x}_{t+1}), y_{t+1}). \quad (20)$$

The hypothesis h_n is discarded for purely technical reasons. The quantity $\hat{R}_n$ is often referred to as the *average instantaneous risk*. It is central in many online learning analysis (see by example Cesa-Bianchi et al. [2004]). Each term of the previous sum is an estimator of the risk $R(h_t)$ associated with the hypothesis h_t (conditionally to the examples $z_1, \dots, z_t$) :

$$\mathbb{E}\left[\ell(h_t(\boldsymbol{x}_{t+1}), y_{t+1}) \,|\, z_1, \dots, z_t\right] = R(h_t).$$

The term *instantaneous* comes from the fact that $\hat{R}_n$ only relies on the example $(\boldsymbol{x}_{t+1}, y_{t+1})$ appearing at iteration $t+1$ in order to evaluate the risk of h_t. A state of the art result due to Cesa-Bianchi and Gentile [2008] links $\hat{R}_n$ to $\frac{1}{n}\sum_{t=0}^{n-1} R(h_t)$:

Proposition 1. *Let $h_0, \dots, h_{n-1}$ be the set of hypotheses generated by an online learning algorithm using the bounded loss function $\ell \in [0,1]^{\tilde{\mathcal{Y}} \times \mathcal{Y}}$. Then, for all $0 < \delta \leq 1$, we have with probability at least $1-\delta$*

$$\frac{1}{n}\sum_{t=0}^{n-1} R(h_t) \leq \hat{R}_n + 2\sqrt{\frac{\hat{R}_n}{n}\ln\left(\frac{n\hat{R}_n + 3}{\delta}\right)} + \frac{36}{n}\ln\left(\frac{n\hat{R}_n + 3}{\delta}\right). \quad (21)$$

Remark 1. The Gibbs classifier [McAllester, 1999] is a stochastic classifier obtained by selecting randomly a hypothesis among a set of hypotheses, given a probability distribution on these hypotheses. $\frac{1}{n}\sum_{t=0}^{n-1} R(h_t)$ can thus be seen as the risk of the Gibbs classifier for an uniform distribution on the set $\{h_0, \dots, h_{n-1}\}$.

The key of the result exposed in the previous proposition lies in the use of a second order concentration inequality for martingales (proposed by Freedman [1975]) which introduces the sum V_n of the conditional variances of the loss of each hypothesis:

$$V_n = \sum_{t=0}^{n-1} \mathbb{V}\left[\ell(h_t(\boldsymbol{x}_{t+1}), y_{t+1}) \,|\, z_1, \dots, z_t\right].$$

As R_n, this quantity can not be computed since the distribution $\mathcal{D}$ is unknown. Cesa-Bianchi and Gentile [2008] proposed to upper bound this sum using a stratification process in order to get their inequality. In this section we improve the previous bound by employing Theorem 2 together with an online estimator $\hat{V}_n$ of the sum V_n, which allows for a better control of the former. The average *empirical instantaneous variance* $\hat{V}_n$ is simply defined as

$$\hat{V}_n = \frac{1}{2(n-1)}\sum_{t=0}^{n-2}\Big(\ell(h_t(\boldsymbol{x}_{t+1}), y_{t+1}) - \ell(h_t(\boldsymbol{x}_{t+2}), y_{t+2})\Big)^2. \quad (22)$$

Again, we discard the hypotheses h_{n-1} et h_n from this quantity for technical reasons. Each term of this sum is an estimator of the conditional variance of $\ell(h_t(\boldsymbol{x}), y)$:

$$\mathbb{E}\left[\Big(\ell(h_t(\boldsymbol{x}_{t+1}), y_{t+1}) - \ell(h_t(\boldsymbol{x}_{t+2}), y_{t+2})\Big)^2 \,|\, z_1, \dots, z_t\right] = 2\mathbb{V}\left[\ell(h_t(\boldsymbol{x}), y) \,|\, z_1, \dots, z_t\right]. \quad (23)$$

$\hat{V}_n$ may be easily computed during an online learning process and plays a central role in the theorem we present here.

4.2 Empirical Bernstein Inequalities for Online Learning

In the following theorem, we use Theorem 2 and the instantaneous estimators $\hat{R}_n$ et $\hat{V}_n$ in order to bound $\frac{1}{n}\sum_{t=0}^{n-1} R(h_t)$, the mean of the risks of the hypotheses learned by an online algorithm.

Theorem 3 (Empirical Bernstein inequality for online learning)**.** *Let $h_0, \dots, h_{n-1}$ be the set of hypotheses generated from the sample $\underline{Z}_n = \{z_i\}_{i=1}^n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ of i.i.d. random variables by an online learning algorithm using the bounded loss function $\ell \in [0,1]^{\tilde{\mathcal{Y}} \times \mathcal{Y}}$. Then, for all $0 < \delta \leq 1$ we have with probability at least $1-\delta$:*

$$\frac{1}{n}\sum_{t=0}^{n-1} R(h_t) \leq \hat{R}_n + \frac{1}{n}\sqrt{\beta_n \ln\left(\frac{2}{\delta}\right)} + \frac{2}{3n}\ln\left(\frac{2}{\delta}\right), \quad (24)$$

where

$$\beta_n = (n-1)\hat{V}_n + \sqrt{\frac{n-1}{2}\ln\left(\frac{2}{\delta}\right)}. \quad (25)$$

Proof. (Theorem 3) The proof is direct. Consider the set $\underline{Z}_n = \{z_i\}_{i=1}^n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ of i.i.d. random variables and the family of functions $\{\ell(h_i(\cdot), \cdot)\}_{i=0}^n$ where each function $\ell(h_t(\cdot), \cdot)$ only depends on the variables $z_1, \ldots, z_t$ by definition of h_t. Noting that z_{t+1}, z_{t+2} are independent with respect to $z_1, \ldots, z_t$ (by definition of $\underline{Z}_n$), we simply apply Theorem 2 and adjust the indexes to obtain the result. □

We now want to emphasize the comparison with the bound by Cesa-Bianchi and Gentile [2008] (Equation (21)). Our result firstly improves the constants involved in the bound which is very appreciable when the bound is computed with a small number of hypotheses (when n is small, the last term in the bound can not be neglected). In order to analyze the behavior of our result when we have a sufficient number of hypotheses to omit the last term, we have to pay attention to

$$\frac{1}{n}\sqrt{\beta_n \ln\left(\frac{2}{\delta}\right)} \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)\hat{V}_n}{n}} + \left(\frac{\ln\left(\frac{2}{\delta}\right)}{n}\right)^{3/4}, \quad (26)$$

where we used the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ to get the upper bound. Thus, omitting the constant terms, our bound tends to $\hat{R}_n$ at least in

$$\mathcal{O}\left(\sqrt{\frac{\hat{V}_n}{n}} + \frac{1}{n^{3/4}} + \frac{1}{n}\right),$$

when the one by Cesa-Bianchi and Gentile [2008] tends to $\hat{R}_n$ in

$$\mathcal{O}\left(\sqrt{\hat{R}_n \frac{\ln(n\hat{R}_n)}{n}} + \frac{\ln(n\hat{R}_n)}{n}\right).$$

In order to study the difference between the two rates of convergence, we need to compare the two terms $\hat{V}_n$ and $\hat{R}_n$.

$$\begin{aligned}
\hat{V}_n &= \frac{1}{2(n-1)} \sum_{t=0}^{n-2} \big(\ell(h_t(\boldsymbol{x}_{t+1}), y_{t+1}) \\
&\quad - \ell(h_t(\boldsymbol{x}_{t+2}), y_{t+2})\big)^2 \\
&\leq \frac{1}{2(n-1)} \left(\sum_{t=0}^{n-2} \ell(h_t(\boldsymbol{x}_{t+1}), y_{t+1})^2 \right. \\
&\quad \left. + \sum_{t=0}^{n-2} \ell(h_t(\boldsymbol{x}_{t+2}), y_{t+2})^2 \right) \\
&\leq \frac{1}{2(n-1)} \left(\sum_{t=0}^{n-2} \ell(h_t(\boldsymbol{x}_{t+1}), y_{t+1}) \right. \\
&\quad \left. + \sum_{t=0}^{n-2} \ell(h_t(\boldsymbol{x}_{t+2}), y_{t+2}) \right).
\end{aligned}$$

The last inequality is obtained by using $\ell \in [0,1]^{\tilde{\mathcal{Y}} \times \mathcal{Y}}$. Suppose that the error made by each hypothesis h_t on the example z_{t+2} is not too different from the error made by the same hypothesis on z_{t+1}:

$$\ell(h_t(\boldsymbol{x}_{t+2}), y_{t+2}) \approx \ell(h_t(\boldsymbol{x}_{t+1}), y_{t+1}).$$

In this case, the previous right hand side is almost

$$\frac{1}{(n-1)} \sum_{t=0}^{n-2} \ell(h_t(\boldsymbol{x}_{t+1}), y_{t+1}) \approx \hat{R}_n$$

thus it follows that $\hat{V}_n \leq \hat{R}_n$. A setting studied by Cesa-Bianchi and Gentile [2008] is when the empirical cumulative risk $n\hat{R}_n$ is in $\mathcal{O}(1)$ i.e. $n\hat{R}_n$ is bounded. Their result thus reaches an asymptotic behavior in $\mathcal{O}(\frac{1}{n})$ (the terms involving $\ln(n\hat{R}_n)$ vanishes as a constant). With the assumption that $\hat{V}_n$ is in $\mathcal{O}(1)$ as well, our bound shows a rate of convergence slightly worse in $\mathcal{O}(\frac{1}{n^{3/4}})$. However, as soon as the cumulative risk $n\hat{R}_n$ increases with n, the bound by Cesa-Bianchi and Gentile [2008] converges at the rate $\mathcal{O}(\sqrt{\ln n/n})$ whereas ours reaches a $\mathcal{O}(\sqrt{1/n})$ rate.

Case of a Convex Loss Function When an online algorithm uses a convex loss function ℓ, we can use Theorem 3 in order to characterize the risk associated with the *mean hypothesis* $\bar{h}$:

$$\bar{h} = \frac{1}{n} \sum_{t=0}^{n-1} h_t. \quad (27)$$

When the decision space $\tilde{\mathcal{Y}}$ associated with the classifiers $h_t : \mathcal{X} \to \tilde{\mathcal{Y}}$ is convex then the hypothesis $\bar{h}$ belongs to the same function class as each of the h_t, $\bar{h} : \mathcal{X} \to \tilde{\mathcal{Y}}$. The mean hypothesis is thus a deterministic classifier, by opposition to the Gibbs classifier defined earlier, which shares the same bound on its risk.

Corollary 2. *Let $h_0, \ldots, h_{n-1}$ be the set of all the hypotheses generated by an online learning algorithm using the convex loss function ℓ such that $\ell \in [0,1]^{\tilde{\mathcal{Y}} \times \mathcal{Y}}$. Then, for all $0 < \delta \leq 1$ with probability at least $1 - \delta$*

$$R(\bar{h}) \leq \hat{R}_n + \frac{1}{n}\sqrt{\beta_n \ln\left(\frac{2}{\delta}\right)} + \frac{2}{3n} \ln\left(\frac{2}{\delta}\right), \quad (28)$$

where

$$\beta_n = (n-1)\hat{V}_n + \sqrt{\frac{n-1}{2} \ln\left(\frac{2}{\delta}\right)}.$$

Proof. Using Jensen's inequality and linearity of the

expectation, it is easy to show that

$$\begin{aligned} R(\bar{h}) &= \mathbb{E}\left[\ell(\frac{1}{n}\sum_{t=0}^{n-1} h_t(X), Y)\right] \\ &\leq \frac{1}{n}\sum_{t=0}^{n-1}\mathbb{E}\left[\ell(h_t(X), Y)\right] \\ &= \frac{1}{n}\sum_{t=0}^{n-1} R(h_t). \end{aligned}$$

To conclude the proof, we just need to combine this result with Theorem 3. □

5 BOUNDING THE AVERAGE RISK OF PEGASOS

In this section, we use the previous corollary in order to derive a bound on the mean risk of the hypotheses generated by the Pegasos [Shalev-Shwartz et al., 2011] algorithm.

5.1 Pegasos

Pegasos is an algorithm designed to solve the primal SVM problem. Recall that given a sample $\underline{Z}_n = \{z_i\}_{i=1}^n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ the SVM objective function is given by:

$$F(\mathbf{w}) := \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\ell_{hinge}(\mathbf{w}, \boldsymbol{x}_i, y_i), \qquad (29)$$

where $\ell_{hinge}(\mathbf{w}, \boldsymbol{x}, y) = \max\{0, 1 - y\langle \mathbf{w}, \boldsymbol{x}\rangle\}$. Pegasos works in an online fashion by doing a stochastic sub-gradient descent on the SVM objective function. At time t, Pegasos randomly selects an example $Z_{i_t} = (\boldsymbol{x}_{i_t}, y_{i_t})$ and aims at minimizing the approximation

$$f(\mathbf{w}^t, Z_{i_t}) = \frac{\lambda}{2}\left\|\mathbf{w}^t\right\|_2^2 + \ell_{hinge}(\mathbf{w}^t, \boldsymbol{x}_{i_t}, y_{i_t}),$$

of the SVM objective function. It considers the following sub-gradient, taken at point $\mathbf{w}^t$, of the previous function which is given by

$$\nabla_t = \nabla_{\mathbf{w}^t} f(\mathbf{w}^t, Z_{i_t}) = \lambda\mathbf{w}^t - \mathbb{1}_{\left[y_{i_t}\langle \mathbf{w}^t, \boldsymbol{x}_{i_t}\rangle < 1\right]} y_{i_t}\boldsymbol{x}_{i_t},$$

and it updates the current weight vector $\mathbf{w}^t$ to $\mathbf{w}^{t+1}$ by $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t\nabla_t$ using a step $\eta_t = 1/(\lambda t)$. So, we get at each iteration the vector

$$\mathbf{w}^{t+1} \leftarrow \left(1 - \frac{1}{t}\right)\mathbf{w}^t + \eta_t \mathbb{1}_{\left[y_{i_t}\langle \mathbf{w}^t, \boldsymbol{x}_{i_t}\rangle < 1\right]} y_{i_t}\boldsymbol{x}_{i_t}.$$

A projection step (optional) that we detail in the sequel ends up the iteration t. Pegasos stops when $t = T$,

Algorithm 1 Pegasos

Require: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n, \lambda \geq 0$ and $T \geq 0$
Ensure: $\mathbf{w}^{T+1}$
  $\mathbf{w}^0 \leftarrow \mathbf{0}$
  for $t \leftarrow 0$ to T **do**
    Pick randomly $i_t \in \{1, \ldots, n\}$
    Define $\eta_t = \frac{1}{\lambda t}$
    if $y_{i_t}\langle \mathbf{w}^t, \boldsymbol{x}_{i_t}\rangle < 1$ **then**
      $\mathbf{w}^{t+1} \leftarrow (1 - \eta_t\lambda)\mathbf{w}^t + \eta_t y_{i_t}\boldsymbol{x}_{i_t}$
    else
      $\mathbf{w}^{t+1} \leftarrow (1 - \eta_t\lambda)\mathbf{w}^t$
    end if
    $\mathbf{w}^{t+1} = \min\left[1, \frac{1/\sqrt{\lambda}}{\|\mathbf{w}^{t+1}\|_2}\right]\mathbf{w}^{t+1}$
  end for

where T is a number of iteration given as a parameter. Thus Pegasos can be seen as an online algorithm working with the sequence of examples $Z_{i_1}, \ldots, Z_{i_T}$ constructed by picking randomly at each iteration an example from $\underline{Z}_n$. Algorithm 1 sums up the different steps of Pegasos.

5.2 Bounding the Mean Risk of the Hypotheses Generated by Pegasos

In order to apply Theorem 3 we need the loss function to be bounded. It can be shown that $\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w})$ satisfies $\|\mathbf{w}^*\|_2 \leq 1/\sqrt{\lambda}$. Thus, we can limit the search space to the ball of radius $1/\sqrt{\lambda}$ by incorporating a projection step as mentioned above

$$\mathbf{w}^{t+1} = \min\left[1, \frac{1/\sqrt{\lambda}}{\|\mathbf{w}^{t+1}\|_2}\right]\mathbf{w}^{t+1}.$$

With the assumption that $\|\boldsymbol{x}\|_2 \leq M$, we can bound the hinge loss function:

$$\ell_{hinge}(\mathbf{w}, \boldsymbol{x}, y) \leq 1 + \|\boldsymbol{x}\|_2\|\mathbf{w}\|_2 \leq 1 + \frac{M}{\sqrt{\lambda}} = C.$$

Thereby, the loss function used by Pegasos can be adjusted to satisfy the assumption of Theorem 3 and we can use it to prove the following corollary.

Corollary 3. *Let $\mathbf{w}^0, \ldots, \mathbf{w}^T$ be the sequence of weight vectors generated by the Pegasos algorithm from a sample $\underline{Z}_n$ where $\|\boldsymbol{x}_i\|_2 \leq M$, $1 \leq i \leq n$. Then for all $0 < \delta \leq 1$, we have with probability at least $1 - \delta$,*

$$\frac{1}{n}\sum_{t=0}^{n-1} R(\mathbf{w}^t) \leq \hat{R}_n + \frac{C}{n}\sqrt{\tilde{\beta}_n \ln\left(\frac{2}{\delta}\right)} + \frac{2C}{3n}\ln\left(\frac{2}{\delta}\right),$$

where

$$\tilde{\beta}_n = \frac{(n-1)\hat{V}_n}{C^2} + \sqrt{\frac{n-1}{2}\ln\left(\frac{2}{\delta}\right)}. \qquad (30)$$

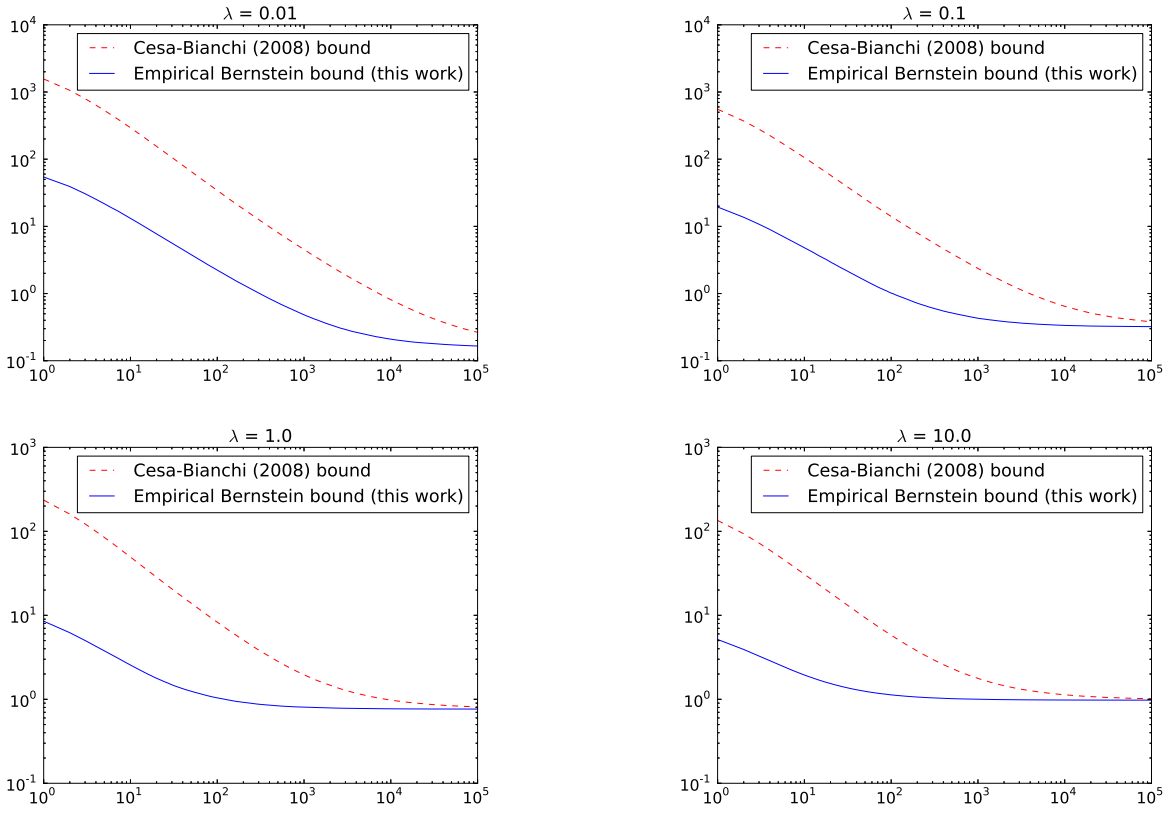

Figure 1: Comparison of the Bounds From Proposition 1 and Corollary 3 Computed for the Pegasos Algorithm on a Toy Linearly Separable Dataset.

5.3 Proof of Concept

In this section we want to highlight experimentally the performance of our empirical Bernstein inequality applied to online learning. In order to do that, we compare the bound provided by Corollary 3 for the Pegasos algorithm to the one exposed in Proposition 1. We use a linearly separable toy dataset and compare the convergence of the empirical risk to the mean risk of the hypotheses $\mathbf{w}^0, \dots, \mathbf{w}^T$. We generate random vectors $x_i \in [-1,1]^2$ to which we assign the class $y_i = sign(\langle \mathbf{w}^*, \boldsymbol{x}_i \rangle) \in \{+1, -1\}$ for a vector $\mathbf{w}^* \in [-1,1]^2$ also randomly generated. We work with a learning sample containing 200000 points and report in Figure 1 the values of the right hand sides appearing in Proposition 1 [Cesa-Bianchi and Gentile, 2008] and in Corollary 3 computed with a confidence of 95% ($\delta = 0.05$). We ran the experiment 20 times for many values of the parameter λ and averaged the results. We can see that our inequality is far tighter than the one by Cesa-Bianchi and Gentile [2008] during the first iterations, as it was sounded in the theoretical comparison done in Section 3. The gap between the two inequalities tightens when the number of hypotheses considered increases but remains in our favor.

6 CONCLUSION AND OUTLOOKS

In this article, we present a new empirical Bernstein concentration inequality for martingales. We applied this result to the online learning setting in order to bound the mean risk of the hypotheses learned during such learning processes. Because we introduce of a new instantaneous variance estimator, our inequality is well suited for the online learning setting and improves the state of the art. This improvement is mainly noticeable when the number of hypotheses considered is small as shown in the empirical section of this work.

There are many outlooks opened by this work. First of all, we can think about a new online learning algorithm that aims at minimizing our empirical Bernstein bound as it is done in the *batch* setting [*Variance Penalizing AdaBoost*, Shivaswamy and Jebara, 2011, by example]. Then, it will be of interest to derive new kind of bounds for online algorithms taking advantage of our result (by example on the *excess risk* as it is done in the work by Kakade and Tewari [2009]). The last perspective that we want to mention is the comparison of our bound with the very recent PAC-Bayes-Empirical-Bernstein Inequality by Tolstikhin and Seldin [2013].

References

Kazuoki Azuma. Weighted Sums of Certain Dependent Random Variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.

George Bennett. Probability Inequalities for the Sum of Independent Random Variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Nicolò Cesa-Bianchi and Claudio Gentile. Improved Risk Tail Bounds for On-Line Algorithms. *IEEE Transactions on Information Theory*, 54(1):386–390, 2008.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the Generalization Ability of On-Line Learning Algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

David A. Freedman. On Tail Probabilities for Martingales. *The Annals of Probability*, 3(1):100 – 118, 1975.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Sham M. Kakade and Ambuj Tewari. On the Generalization Ability of Online Strongly Convex Programming Algorithms. In *Advances in Neural Information Processing Systems 21 - NIPS '08*, pages 801–808, 2009.

Nicholas Littlestone, Philip Long, and Manfred Warmuth. On-line Learning of Linear Functions. *Computational Complexity*, 5(1):1–23, 1995.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample Variance Penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory - COLT '09*, 2009.

David A. McAllester. PAC-Bayesian Model Averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory - COLT '99*, pages 164–170, 1999.

Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical Bernstein Inequalities for U-Statistics. In *Advances in Neural Information Processing Systems 23 - NIPS '10*, pages 1903–1911, 2010.

Shai Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. PhD thesis, Hebrew University, 2007.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal Estimated Sub-Gradient Solver for SVM. *Mathematical Programming*, 127(1):3–30, 2011.

Pannagadatta K. Shivaswamy and Tony Jebara. Variance Penalizing AdaBoost. In *Advances in Neural Information Processing Systems 24 - NIPS '11*, pages 1908–1916, 2011.

Ilya Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein Inequality. In *Advances in Neural Information Processing Systems - NIPS '13*, 2013.