

---

# Empirical Bernstein Bounds and Sample Variance Penalization

---

**Andreas Maurer**  
Adalbertstrasse 55  
D-80799 München, Germany  
andreasmaurer@compuserve.com

**Massimiliano Pontil\***  
Dept. of Computer Science  
University College London  
Malet Pl., WC1E, London, UK  
m.pontil@cs.ucl.ac.uk

## Abstract

We give improved constants for data dependent and variance sensitive confidence bounds, called empirical Bernstein bounds, and extend these inequalities to hold uniformly over classes of functions whose growth function is polynomial in the sample size  $n$ . The bounds lead us to consider *sample variance penalization*, a novel learning method which takes into account the empirical variance of the loss function. We give conditions under which sample variance penalization is effective. In particular, we present a bound on the excess risk incurred by the method. Using this, we argue that there are situations in which the excess risk of our method is of order  $1/n$ , while the excess risk of empirical risk minimization is of order  $1/\sqrt{n}$ . We show some experimental results, which confirm the theory. Finally, we discuss the potential application of our results to sample compression schemes.

## 1 Introduction

The method of empirical risk minimization (ERM) is so intuitive, that some of the less plausible alternatives have received little attention by the machine learning community. In this work we present sample variance penalization (SVP), a method which is motivated by some variance-sensitive, data-dependent confidence bounds, which we develop in the paper. We describe circumstances under which SVP works better than ERM and provide some preliminary experimental results which confirm the theory.

In order to explain the underlying ideas and highlight the differences between SVP and ERM, we begin with a discussion of the confidence bounds most frequently used in learning theory.

**Theorem 1 (Hoeffding's inequality)** *Let  $Z, Z_1, \dots, Z_n$  be i.i.d. random variables with values in  $[0, 1]$  and let  $\delta > 0$ . Then with probability at least  $1 - \delta$  in  $(Z_1, \dots, Z_n)$  we have*

$$\mathbb{E}Z - \frac{1}{n} \sum_{i=1}^n Z_i \leq \sqrt{\frac{\ln 1/\delta}{2n}}.$$

---

\*This work was partially supported by EPSRC Grants GR/T18707/01 and EP/D071542/1.

It is customary to call this result Hoeffding's inequality. It appears in a stronger, more general form in Hoeffding's 1963 milestone paper [4]. Proofs can be found in [4] or [8]. We cited Hoeffding's inequality in form of a confidence-dependent bound on the deviation, which is more convenient for our discussion than a deviation-dependent bound on the confidence. Replacing  $Z$  by  $1 - Z$  shows that the confidence interval is symmetric about  $\mathbb{E}Z$ .

Suppose some underlying observation is modeled by a random variable  $X$ , distributed in some space  $\mathcal{X}$  according to some law  $\mu$ . In learning theory Hoeffding's inequality is often applied when  $Z$  measures the loss incurred by some hypothesis  $h$  when  $X$  is observed, that is,

$$Z = \ell_h(X).$$

The expectation  $\mathbb{E}_{X \sim \mu} \ell_h(X)$  is called the risk associated with hypothesis  $h$  and distribution  $\mu$ . Since the risk depends only on the function  $\ell_h$  and on  $\mu$  we can write the risk as

$$P(\ell_h, \mu),$$

where  $P$  is the expectation functional. If an i.i.d. vector  $\mathbf{X} = (X_1, \dots, X_n)$  has been observed, then Hoeffding's inequality allows us to estimate the risk, for fixed hypothesis, by the empirical risk

$$P_n(\ell_h, \mathbf{X}) = \frac{1}{n} \sum_i \ell_h(X_i)$$

within a confidence interval of length  $2\sqrt{(\ln 1/\delta)/(2n)}$ .

Let us call the set  $\mathcal{F}$  of functions  $\ell_h$  for all different hypotheses  $h$  the *hypothesis space* and its members *hypotheses*, ignoring the distinction between a hypothesis  $h$  and the induced loss function  $\ell_h$ . The bound in Hoeffding's inequality can easily be adjusted to hold uniformly over any finite hypothesis space  $\mathcal{F}$  to give the following well known result [1].

**Corollary 2** *Let  $X$  be a random variable with values in a set  $\mathcal{X}$  with distribution  $\mu$ , and let  $\mathcal{F}$  be a finite class of hypotheses  $f: \mathcal{X} \rightarrow [0, 1]$  and  $\delta > 0$ . Then with probability at least  $1 - \delta$  in  $\mathbf{X} = (X_1, \dots, X_n) \sim \mu^n$*

$$P(f, \mu) - P_n(f, \mathbf{X}) \leq \sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{2n}}, \quad \forall f \in \mathcal{F},$$

where  $|\mathcal{F}|$  is the cardinality of  $\mathcal{F}$ .

This result can be further extended to hold uniformly over hypothesis spaces whose complexity can be controlled with different covering numbers which then appear in place of the cardinality  $|\mathcal{F}|$  above. A large body of literature exists on the subject of such uniform bounds to justify hypothesis selection by empirical risk minimization, see [1] and references therein. Given a sample  $\mathbf{X}$  and a hypothesis space  $\mathcal{F}$ , empirical risk minimization selects the hypothesis

$$ERM(\mathbf{X}) = \arg \min_{f \in \mathcal{F}} P_n(f, \mathbf{X}).$$

A drawback of Hoeffding's inequality is that the confidence interval is independent of the hypothesis in question, and always of order  $\sqrt{1/n}$ , leaving us with a uniformly blurred view of the hypothesis class. But for hypotheses of small variance better estimates are possible, such as the following, which can be derived from what is usually called Bennett's inequality (see e.g. Hoeffding's paper [4]).

**Theorem 3 (Bennett's inequality)** *Under the conditions of Theorem 1 we have with probability at least  $1 - \delta$  that*

$$\mathbb{E}Z - \frac{1}{n} \sum_{i=1}^n Z_i \leq \sqrt{\frac{2\mathbb{V}Z \ln 1/\delta}{n}} + \frac{\ln 1/\delta}{3n},$$

where  $\mathbb{V}Z$  is the variance  $\mathbb{V}Z = \mathbb{E}(Z - \mathbb{E}Z)^2$ .

The bound is symmetric about  $\mathbb{E}Z$  and for large  $n$  the confidence interval is now close to  $2\sqrt{\mathbb{V}Z}$  times the confidence interval in Hoeffding's inequality. A version of this bound which is uniform over finite hypothesis spaces, analogous to Corollary 2, is easily obtained, involving now for each hypothesis  $h$  the variance  $\mathbb{V}h(X)$ . If  $h_1$  and  $h_2$  are two hypotheses then  $2\sqrt{\mathbb{V}h_1(X)}$  and  $2\sqrt{\mathbb{V}h_2(X)}$  are always less than or equal to 1 but they can also be much smaller, or one of them can be substantially smaller than the other one. For hypotheses of zero variance the diameter of the confidence interval decays as  $O(1/n)$ .

Bennett's inequality therefore provides us with estimates of lower accuracy for hypotheses of large variance, and higher accuracy for hypotheses of small variance. Given many hypotheses of equal and nearly minimal empirical risk it seems intuitively safer to select the one whose true risk can be most accurately estimated (a point to which we shall return). But unfortunately the right hand side of Bennett's inequality depends on the unobservable variance, so our view of the hypothesis class remains uniformly blurred.

## 1.1 Main results and SVP algorithm

We are now ready to describe the main results of the paper, which provide the motivation for the SVP algorithm.

Our first result provides a purely data-dependent bound with similar properties as Bennett's inequality.

**Theorem 4** *Under the conditions of Theorem 1 we have with probability at least  $1 - \delta$  in the i.i.d. vector  $\mathbf{Z} = (Z_1, \dots, Z_n)$  that*

$$\mathbb{E}Z - \frac{1}{n} \sum_{i=1}^n Z_i \leq \sqrt{\frac{2V_n(\mathbf{Z}) \ln 2/\delta}{n}} + \frac{7 \ln 2/\delta}{3(n-1)},$$

where  $V_n(\mathbf{Z})$  is the sample variance

$$V_n(\mathbf{Z}) = \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} (Z_i - Z_j)^2.$$

We next extend Theorem 4 over a finite function class.

**Corollary 5** *Let  $X$  be a random variable with values in a set  $\mathcal{X}$  with distribution  $\mu$ , and let  $\mathcal{F}$  be a finite class of hypotheses  $f : \mathcal{X} \rightarrow [0, 1]$ . For  $\delta > 0$ ,  $n \geq 2$  we have with probability at least  $1 - \delta$  in  $\mathbf{X} = (X_1, \dots, X_n) \sim \mu^n$  that*

$$P(f, \mu) - P_n(f, \mathbf{X}) \leq \sqrt{\frac{2V_n(f, \mathbf{X}) \ln(2|\mathcal{F}|/\delta)}{n}} + \frac{7 \ln(2|\mathcal{F}|/\delta)}{3(n-1)}, \quad \forall f \in \mathcal{F},$$

where  $V_n(f, \mathbf{X}) = V_n(f(X_1), \dots, f(X_n))$ .

Theorem 4 makes the diameter of the confidence interval observable. The corollary is obtained from a union bound over  $\mathcal{F}$ , analogous to Corollary 2, and provides us with a view of the loss class which is blurred for hypotheses of large sample variance, and more in focus for hypotheses of small sample variance.

We note that an analogous result to Theorem 4 is given by Audibert et al. [2]. Our technique of proof is new and the bound we derive has a slightly better constant. Theorem 4 itself resembles Bernstein's or Bennett's inequality, in confidence bound form, but in terms of observable quantities. For this reason it has been called an *empirical Bernstein bound* in [9]. In [2] Audibert et al. apply their result to the analysis of algorithms for the multi-armed bandit problem and in [9] it is used to derive stopping rules for sampling procedures. We will prove Theorem 4 in Section 2, together with some useful confidence bounds on the standard deviation, which may be valuable in their own right.

Our next result extends the uniform estimate in Corollary 5 to infinite loss classes whose complexity can be suitably controlled. Beyond the simple extension involving covering numbers for  $\mathcal{F}$  in the uniform norm  $\|\cdot\|_\infty$ , we can use the following complexity measure, which is also fairly commonplace in the machine learning literature [1], [3].

For  $\epsilon > 0$ , a function class  $\mathcal{F}$  and an integer  $n$ , the "growth function"  $\mathcal{N}_\infty(\epsilon, \mathcal{F}, n)$  is defined as

$$\mathcal{N}_\infty(\epsilon, \mathcal{F}, n) = \sup_{\mathbf{x} \in \mathcal{X}^n} \mathcal{N}(\epsilon, \mathcal{F}(\mathbf{x}), \|\cdot\|_\infty),$$

where  $\mathcal{F}(\mathbf{x}) = \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n$  and for  $A \subseteq \mathbb{R}^n$  the number  $\mathcal{N}(\epsilon, A, \|\cdot\|_\infty)$  is the smallest cardinality  $|A_0|$  of a set  $A_0 \subseteq A$  such that  $A$  is contained in the union of  $\epsilon$ -balls centered at points in  $A_0$ , in the metric induced by  $\|\cdot\|_\infty$ .

**Theorem 6** *Let  $X$  be a random variable with values in a set  $\mathcal{X}$  with distribution  $\mu$  and let  $\mathcal{F}$  be a class of hypotheses  $f : \mathcal{X} \rightarrow [0, 1]$ . Fix  $\delta \in (0, 1)$ ,  $n \geq 16$  and set*

$$\mathcal{M}(n) = 10\mathcal{N}_\infty(1/n, \mathcal{F}, 2n).$$

Then with probability at least  $1 - \delta$  in the random vector  $\mathbf{X} = (X_1, \dots, X_n) \sim \mu^n$  we have

$$P(f, \mu) - P_n(f, \mathbf{X}) \leq \sqrt{\frac{18V_n(f, \mathbf{X}) \ln(\mathcal{M}(n)/\delta)}{n}} + \frac{15 \ln(\mathcal{M}(n)/\delta)}{n-1}, \quad \forall f \in \mathcal{F}.$$

The structure of this bound is very similar to Corollary 5, with  $2|\mathcal{F}|$  replaced by  $\mathcal{M}(n)$ . In a number of practical cases polynomial growth of  $\mathcal{N}_\infty(1/n, \mathcal{F}, n)$  in  $n$  has been established. For instance, we quote [3, equation (28)] which states that for the bounded linear functionals in the reproducing kernel Hilbert space associated with Gaussian kernels one has  $\ln \mathcal{N}_\infty(1/n, \mathcal{F}, 2n) = O(\ln^{3/2} n)$ . Composition with fixed Lipschitz functions preserves this property, so we can see that Theorem 6 is applicable to a large family of function classes which occur in machine learning. We will prove Theorem 6 in Section 3.

Since the minimization of uniform upper bounds is frequent practice in machine learning, one could consider minimizing the bounds in Corollary 5 or Theorem 6. This leads to *sample variance penalization*, a technique which selects the hypothesis

$$SV P_\lambda(\mathbf{X}) = \arg \min_{f \in \mathcal{F}} P_n(f, \mathbf{X}) + \lambda \sqrt{\frac{V_n(f, \mathbf{X})}{n}},$$

where  $\lambda \geq 0$  is some regularization parameter. For  $\lambda = 0$  we recover empirical risk minimization. The last term on the right hand side can be regarded as a data-dependent regularizer.

Why, and under which circumstances, should sample variance penalization work better than empirical risk minimization? If two hypotheses have the same empirical risk, why should we discard the one with higher sample variance? After all, the empirical risk of the high variance hypothesis may be just as much overestimating the true risk as underestimating it. In Section 4 we will argue that the decay of the excess risk of sample variance penalization can be bounded in terms of the variance of an optimal hypothesis (see Theorem 15) and if there is an optimal hypothesis with zero variance, then the excess risk decreases as  $1/n$ . We also give an example of such a case where the excess risk of empirical risk minimization cannot decrease faster than  $O(1/\sqrt{n})$ . We then report on the comparison of the two algorithms in a toy experiment.

Finally, in Section 5 we present some preliminary observations concerning the application of empirical Bernstein bounds to sample-compression schemes.

## 1.2 Notation

We summarize the notation used throughout the paper. We define the following functions on the cube  $[0, 1]^n$ , which will be used throughout. For every  $\mathbf{x} = (x_1, \dots, x_n) \in [0, 1]^n$  we let

$$P_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n x_i$$

and

$$V_n(\mathbf{x}) = \frac{1}{n(n-1)} \sum_{i,j=1}^n \frac{(x_i - x_j)^2}{2}.$$

If  $\mathcal{X}$  is some set,  $f: \mathcal{X} \rightarrow [0, 1]$  and  $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$  we write  $f(\mathbf{x}) = (f(x_1), \dots, f(x_n))$ ,  $P_n(f, \mathbf{x}) = P_n(f(\mathbf{x}))$  and  $V_n(f, \mathbf{x}) = V_n(f(\mathbf{x}))$ .

Questions of measurability will be ignored throughout, if necessary this is enforced through finiteness assumptions. If  $X$  is a real valued random variable we use  $\mathbb{E}X$  and  $\mathbb{V}X$  to denote its expectation and variance, respectively. If  $X$  is a random variable distributed in some set  $\mathcal{X}$  according to a distribution  $\mu$ , we write  $X \sim \mu$ . Product measures are denoted by the symbols  $\times$  or  $\prod$ ,  $\mu^n$  is the  $n$ -fold product of  $\mu$  and the random variable  $\mathbf{X} = (X_1, \dots, X_n) \sim \mu^n$  is an i.i.d. sample generated from  $\mu$ . If  $X \sim \mu$  and  $f: \mathcal{X} \rightarrow \mathbb{R}$  then we write  $P(f, \mu) = \mathbb{E}_{X \sim \mu} f(X) = \mathbb{E}f(X)$  and  $V(f, \mu) = \mathbb{V}_{X \sim \mu} f(X) = \mathbb{V}f(X)$ .

## 2 Empirical Bernstein bounds and variance estimation

In this section, we prove Theorem 4 and some related useful results, in particular concentration inequalities for the variance of a bounded random variable, (5) and (6) below, which may be of independent interest. For future use we derive our results for the more general case where the  $X_i$  in the sample are independent, but not necessarily identically distributed.

We need two auxiliary results. One is a concentration inequality for self-bounding random variables (Theorem 13 in [7]):

**Theorem 7** *Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a vector of independent random variables with values in some set  $\mathcal{X}$ . For  $1 \leq k \leq n$  and  $y \in \mathcal{X}$ , we use  $\mathbf{X}_{y,k}$  to denote the vector obtained from  $\mathbf{X}$  by replacing  $X_k$  by  $y$ . Suppose that  $a \geq 1$  and that  $Z = Z(\mathbf{X})$  satisfies the inequalities*

$$Z(\mathbf{X}) - \inf_{y \in \mathcal{X}} Z(\mathbf{X}_{y,k}) \leq 1, \quad \forall k \quad (1)$$

$$\sum_{k=1}^n \left( Z(\mathbf{X}) - \inf_{y \in \mathcal{X}} Z(\mathbf{X}_{y,k}) \right)^2 \leq aZ(\mathbf{X}) \quad (2)$$

almost surely. Then, for  $t > 0$ ,

$$\Pr \{ \mathbb{E}Z - Z > t \} \leq \exp \left( \frac{-t^2}{2a\mathbb{E}Z} \right).$$

If  $Z$  satisfies only the self-boundedness condition (2) we still have

$$\Pr \{ Z - \mathbb{E}Z > t \} \leq \exp \left( \frac{-t^2}{2a\mathbb{E}Z + at} \right).$$

The other result we need is a technical lemma on conditional expectations.

**Lemma 8** *Let  $X, Y$  be i.i.d. random variables with values in an interval  $[a, a+1]$ . Then*

$$\mathbb{E}_X \left[ \mathbb{E}_Y (X - Y)^2 \right]^2 \leq (1/2) \mathbb{E}(X - Y)^2.$$

**Proof:** The right side of the above inequality is of course the variance  $\mathbb{E}[X^2 - XY]$ . One computes

$$\mathbb{E}_X \left[ \mathbb{E}_Y (X - Y)^2 \right]^2 = \mathbb{E}[X^4 + 3X^2Y^2 - 4X^3Y].$$

We therefore have to show that  $\mathbb{E}[g(X, Y)] \geq 0$  where

$$g(X, Y) = X^2 - XY - X^4 - 3X^2Y^2 + 4X^3Y$$

A rather tedious computation gives

$$\begin{aligned} g(X, Y) + g(Y, X) &= \\ &= X^2 - XY - X^4 - 3X^2Y^2 + 4X^3Y + \\ &\quad + Y^2 - XY - Y^4 - 3X^2Y^2 + 4Y^3X \\ &= (X - Y + 1)(Y - X + 1)(Y - X)^2. \end{aligned}$$

The latter expression is clearly nonnegative, so

$$2[\mathbb{E}g(X, Y)] = \mathbb{E}[g(X, Y) + g(Y, X)] \geq 0,$$

which completes the proof.  $\blacksquare$

When the random variables  $X$  and  $Y$  are uniformly distributed on a finite set,  $\{x_1, \dots, x_n\}$ , Lemma 8 gives the following useful corollary.

**Corollary 9** Suppose  $\{x_1, \dots, x_n\} \subset [0, 1]$ . Then

$$\frac{1}{n} \sum_k \left( \frac{1}{n} \sum_j (x_k - x_j)^2 \right)^2 \leq \frac{1}{2n^2} \sum_{k,j} (x_k - x_j)^2.$$

We first establish confidence bounds for the standard deviation.

**Theorem 10** Let  $n \geq 2$  and  $\mathbf{X} = (X_1, \dots, X_n)$  be a vector of independent random variables with values in  $[0, 1]$ . Then for  $\delta > 0$  we have, writing  $\mathbb{E}V_n$  for  $\mathbb{E}_{\mathbf{X}} V_n(\mathbf{X})$ ,

$$\Pr \left\{ \sqrt{\mathbb{E}V_n} > \sqrt{V_n(\mathbf{X})} + \sqrt{\frac{2 \ln 1/\delta}{n-1}} \right\} \leq \delta \quad (3)$$

$$\Pr \left\{ \sqrt{V_n(\mathbf{X})} > \sqrt{\mathbb{E}V_n} + \sqrt{\frac{2 \ln 1/\delta}{n-1}} \right\} \leq \delta. \quad (4)$$

**Proof:** Write  $Z(\mathbf{X}) = nV_n(\mathbf{X})$ . Now fix some  $k$  and choose any  $y \in [0, 1]$ . Then

$$\begin{aligned} Z(\mathbf{X}) - Z(\mathbf{X}_{y,k}) &= \\ &= \frac{1}{n-1} \sum_j \left( (X_k - X_j)^2 - (y - X_j)^2 \right) \\ &\leq \frac{1}{n-1} \sum_j (X_k - X_j)^2. \end{aligned}$$

It follows that  $Z(\mathbf{X}) - \inf_{y \in \Omega} Z(\mathbf{X}_{y,k}) \leq 1$ . We also get

$$\begin{aligned} \sum_k \left( Z(\mathbf{X}) - \inf_{y \in [0,1]} Z(\mathbf{X}_{y,k}) \right)^2 &\leq \\ &\leq \sum_k \left( \frac{1}{n-1} \sum_j (X_k - X_j)^2 \right)^2 \\ &\leq \frac{n^3}{(n-1)^2} \frac{1}{2n^2} \sum_{kj} (X_k - X_j)^2 \\ &= \frac{n}{n-1} Z(\mathbf{X}), \end{aligned}$$

where we applied Corollary 9 to get the second inequality. It follows that  $Z$  satisfies (1) and (2) with  $a = n/(n-1)$ . From Theorem 7 and

$$\Pr \{ \pm \mathbb{E}V_n \mp V_n(\mathbf{X}) > s \} = \Pr \{ \pm \mathbb{E}Z \mp Z(\mathbf{X}) > ns \}$$

we can therefore conclude the following concentration result for the sample variance: For  $s > 0$

$$\Pr \{ \mathbb{E}V_n - V_n(\mathbf{X}) > s \} \leq \exp \left( \frac{-(n-1)s^2}{2\mathbb{E}V_n} \right) \quad (5)$$

$$\Pr \{ V_n(\mathbf{X}) - \mathbb{E}V_n > s \} \leq \exp \left( \frac{-(n-1)s^2}{2\mathbb{E}V_n + s} \right). \quad (6)$$

From the lower tail bound (5) we obtain with probability at least  $1 - \delta$  that

$$\mathbb{E}V_n - 2\sqrt{\mathbb{E}V_n} \sqrt{\frac{\ln 1/\delta}{2(n-1)}} \leq V_n(\mathbf{X}).$$

Completing the square on the left hand side, taking the square-root, adding  $\sqrt{\ln(1/\delta)/(2(n-1))}$  and using  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$  gives (3). Solving the right side of (6) for  $s$  and using the same square-root inequality we find that with probability at least  $1 - \delta$  we have

$$\begin{aligned} V_n(\mathbf{X}) &\leq \mathbb{E}V_n + 2\sqrt{\frac{\mathbb{E}V_n \ln 1/\delta}{2(n-1)}} + \frac{\ln 1/\delta}{(n-1)} \\ &= \left( \sqrt{\mathbb{E}V_n} + \sqrt{\frac{\ln 1/\delta}{2(n-1)}} \right)^2 + \frac{\ln 1/\delta}{2(n-1)}. \end{aligned}$$

Taking the square-root and using the root-inequality again gives (4).  $\blacksquare$

We can now prove the empirical Bernstein bound, which reduces to Theorem 4 for identically distributed variables.

**Theorem 11** Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a vector of independent random variables with values in  $[0, 1]$ . Let  $\delta > 0$ . Then with probability at least  $1 - \delta$  in  $\mathbf{X}$  we have

$$\mathbb{E}[P_n(\mathbf{X})] \leq P_n(\mathbf{X}) + \sqrt{\frac{2V_n(\mathbf{X}) \ln 2/\delta}{n}} + \frac{7 \ln 2/\delta}{3(n-1)}.$$

**Proof:** Write  $W = (1/n) \sum_i \mathbb{V} X_i$  and observe that

$$W \leq \frac{1}{n} \sum_i \mathbb{E}(X_i - \mathbb{E}X_i)^2 \quad (7)$$

$$+ \frac{1}{2n(n-1)} \sum_{i \neq j} (\mathbb{E}X_i - \mathbb{E}X_j)^2 \quad (8)$$

$$\begin{aligned} &= \frac{1}{2n(n-1)} \sum_{i,j} \mathbb{E}(X_i - X_j)^2 \\ &= \mathbb{E}V_n. \end{aligned} \quad (9)$$

Recall that Bennett's inequality, which holds also if the  $X_i$  are not identically distributed (see [8]), implies with probability at least  $1 - \delta$

$$\begin{aligned} \mathbb{E}P_n(\mathbf{X}) &\leq P_n(\mathbf{X}) + \sqrt{\frac{2W \ln 1/\delta}{n}} + \frac{\ln 1/\delta}{3n} \\ &\leq P_n(\mathbf{X}) + \sqrt{\frac{2\mathbb{E}V_n \ln 1/\delta}{n}} + \frac{\ln 1/\delta}{3n}, \end{aligned}$$

so that the conclusion follows from combining this inequality with (3) in a union bound and some simple estimates. ■

### 3 Empirical Bernstein bounds for function classes of polynomial growth

We now prove Theorem 6. We will use the classical double-sample method ([10], [1]), but we have to pervert it somewhat to adapt it to the nonlinearity of the empirical standard-deviation functional. Define functions  $\Phi, \Psi : [0, 1]^n \times \mathbb{R}_+ \rightarrow \mathbb{R}$  by

$$\Phi(\mathbf{x}, t) = P_n(\mathbf{x}) + \sqrt{\frac{2V_n(\mathbf{x})t}{n}} + \frac{7t}{3(n-1)},$$

$$\Psi(\mathbf{x}, t) = P_n(\mathbf{x}) + \sqrt{\frac{18V_n(\mathbf{x})t}{n}} + \frac{11t}{n-1}.$$

We first record some simple Lipschitz properties of these functions.

**Lemma 12** For  $t > 0$ ,  $\mathbf{x}, \mathbf{x}' \in [0, 1]^n$  we have

$$(i) \quad \Phi(\mathbf{x}, t) - \Phi(\mathbf{x}', t) \leq (1 + 2\sqrt{t/n}) \|\mathbf{x} - \mathbf{x}'\|_\infty,$$

$$(ii) \quad \Psi(\mathbf{x}, t) - \Psi(\mathbf{x}', t) \leq (1 + 6\sqrt{t/n}) \|\mathbf{x} - \mathbf{x}'\|_\infty.$$

**Proof:** One verifies that

$$\sqrt{V_n(\mathbf{x})} - \sqrt{V_n(\mathbf{x}')} \leq \sqrt{2} \|\mathbf{x} - \mathbf{x}'\|_\infty,$$

which implies (i) and (ii). ■

Given two vectors  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$  and  $\sigma \in \{-1, 1\}^n$  define  $(\sigma, \mathbf{x}, \mathbf{x}') \in \mathcal{X}^n$  by  $(\sigma, \mathbf{x}, \mathbf{x}')_i = x_i$  if  $\sigma_i = 1$  and  $(\sigma, \mathbf{x}, \mathbf{x}')_i = x'_i$  if  $\sigma_i = -1$ . In the following the  $\sigma_i$  will be independent random variables, uniformly distributed on  $\{-1, 1\}$ .

**Lemma 13** Let  $\mathbf{X} = (X_1, \dots, X_n)$  and  $\mathbf{X}' = (X'_1, \dots, X'_n)$  be random vectors with values in  $\mathcal{X}$  such that all the  $X_i$  and  $X'_i$  are independent and identically distributed. Suppose that  $F : \mathcal{X}^{2n} \rightarrow [0, 1]$ . Then

$$\mathbb{E}F(\mathbf{X}, \mathbf{X}') \leq \sup_{(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^{2n}} \mathbb{E}_\sigma F((\sigma, \mathbf{x}, \mathbf{x}'), (-\sigma, \mathbf{x}, \mathbf{x}')).$$

**Proof:** For any configuration  $\sigma$  and  $(\mathbf{X}, \mathbf{X}')$ , the configuration  $((\sigma, \mathbf{X}, \mathbf{X}'), (-\sigma, \mathbf{X}, \mathbf{X}'))$  is obtained from  $(\mathbf{X}, \mathbf{X}')$  by exchanging  $X_i$  and  $X'_i$  whenever  $\sigma_i = -1$ . Since  $X_i$  and  $X'_i$  are identically distributed this does not affect the expectation. Thus

$$\begin{aligned} \mathbb{E}F(\mathbf{X}, \mathbf{X}') &= \mathbb{E}_\sigma \mathbb{E}F((\sigma, \mathbf{X}, \mathbf{X}'), (-\sigma, \mathbf{X}, \mathbf{X}')) \\ &\leq \sup_{(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^{2n}} \mathbb{E}_\sigma F((\sigma, \mathbf{x}, \mathbf{x}'), (-\sigma, \mathbf{x}, \mathbf{x}')). \end{aligned}$$

The next lemma is where we use the concentration results in Section 2. ■

**Lemma 14** Let  $f : \mathcal{X} \rightarrow [0, 1]$  and  $(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^{2n}$  be fixed. Then

$$\Pr_\sigma \{ \Phi(f(\sigma, \mathbf{x}, \mathbf{x}'), t) > \Psi(f(-\sigma, \mathbf{x}, \mathbf{x}'), t) \} \leq 5e^{-t}.$$

**Proof:** Define the random vector  $\mathbf{Y} = (Y_1, \dots, Y_n)$ , where the  $Y_i$  are independent random variables, each  $Y_i$  being uniformly distributed on  $\{f(x_i), f(x'_i)\}$ . The  $Y_i$  are of course not identically distributed. Within this proof we use the shorthand notation  $\mathbb{E}P_n = \mathbb{E}_\mathbf{Y} P_n(\mathbf{Y})$  and  $\mathbb{E}V_n = \mathbb{E}_\mathbf{Y} V_n(\mathbf{Y})$ , and let

$$A = \mathbb{E}P_n + \sqrt{\frac{8\mathbb{E}V_n t}{n}} + \frac{14t}{3(n-1)}.$$

Evidently

$$\begin{aligned} \Pr_\sigma \{ \Phi(f(\sigma, \mathbf{x}, \mathbf{x}'), t) > \Psi(f(-\sigma, \mathbf{x}, \mathbf{x}'), t) \} &\leq \\ &\leq \Pr_\sigma \{ \Phi(f(\sigma, \mathbf{x}, \mathbf{x}'), t) > A \} + \\ &\quad + \Pr_\sigma \{ A > \Psi(f(-\sigma, \mathbf{x}, \mathbf{x}'), t) \} = \\ &= \Pr_\mathbf{Y} \{ \Phi(\mathbf{Y}, t) > A \} + \Pr_\mathbf{Y} \{ A > \Psi(\mathbf{Y}, t) \}. \end{aligned}$$

To prove our result we will bound these two probabilities in turn.

Now

$$\begin{aligned} \Pr_\mathbf{Y} \{ \Phi(\mathbf{Y}, t) > A \} &\leq \\ &\leq \Pr \left\{ P_n(\mathbf{Y}) > \mathbb{E}P_n + \sqrt{\frac{2\mathbb{E}V_n t}{n}} + \frac{t}{3(n-1)} \right\} + \\ &\quad + \Pr \left\{ \sqrt{\frac{2V_n(\mathbf{Y})t}{n}} > \sqrt{\frac{2\mathbb{E}V_n t}{n}} + \frac{2t}{n-1} \right\}. \end{aligned}$$

Since  $\sum_i \mathbb{V}(f(Y_i)) \leq n\mathbb{E}V_n$  by equation (7), the first of these probabilities is at most  $e^{-t}$  by Bennett's inequality, which also holds for variables which are not identically distributed.

That the second of these probabilities is bounded by  $e^{-t}$  follows directly from Theorem 10(4). We conclude that  $\Pr_\mathbf{Y} \{ \Phi(\mathbf{Y}, t) > A \} \leq 2e^{-t}$ .

Since  $\sqrt{2} + \sqrt{8} = \sqrt{18}$  we have

$$\begin{aligned} \Pr_\mathbf{Y} \{ A > \Psi(\mathbf{Y}, t) \} &\leq \\ &\leq \Pr \left\{ \mathbb{E}P_n > P_n(\mathbf{Y}) + \sqrt{\frac{2V_n(\mathbf{Y})t}{n}} + \frac{7t}{3(n-1)} \right\} + \\ &\quad + \Pr \left\{ \sqrt{\frac{8\mathbb{E}V_n t}{n}} > \sqrt{\frac{8V_n(\mathbf{Y})t}{n}} + \frac{4t}{n-1} \right\}. \end{aligned}$$

The first probability in the sum is at most  $2e^{-t}$  by Theorem 11, and the second is at most  $e^{-t}$  by Theorem 10(3). Hence  $\Pr_\mathbf{Y} \{ A > \Psi(f(\mathbf{Y}), t) \} \leq 3e^{-t}$ , so it follows that

$$\Pr_\sigma \{ \Phi(f(\sigma, \mathbf{x}, \mathbf{x}'), t) > \Psi(f(-\sigma, \mathbf{x}, \mathbf{x}'), t) \} \leq 5e^{-t}. \quad \blacksquare$$

**Proof of Theorem 6.** It follows from Theorem 11 that for  $t > \ln 4$  we have for any  $f \in \mathcal{F}$  that

$$\Pr \{ \Phi(f(\mathbf{X}), t) > P(f, \mu) \} \geq 1/2.$$

In other words, the functional

$$f \mapsto \Lambda(f) = \mathbb{E}_{\mathbf{X}'} \mathbb{1} \{ \Phi(f(\mathbf{X}'), t) > P(f, \mu) \}$$

satisfies  $1 \leq 2\Lambda(f)$  for all  $f$ . Consequently, for any  $s > 0$  we have, using  $\mathbb{1}A$  to denote the indicator function of  $A$ , that

$$\begin{aligned} & \Pr_{\mathbf{X}} \{ \exists f \in \mathcal{F} : P(f, \mu) > \Psi(f(\mathbf{X}), t) + s \} \\ &= \mathbb{E}_{\mathbf{X}} \sup_{f \in \mathcal{F}} \mathbb{1} \{ P(f, \mu) > \Psi(f(\mathbf{X}), t) + s \} \\ &\leq \mathbb{E}_{\mathbf{X}} \sup_{f \in \mathcal{F}} \mathbb{1} \{ P(f, \mu) > \Psi(f(\mathbf{X}), t) + s \} 2\Lambda(f) \\ &= 2\mathbb{E}_{\mathbf{X}} \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{X}'} \mathbb{1} \left\{ \begin{array}{l} P(f, \mu) > \Psi(f(\mathbf{X}), t) + s \\ \text{and } \Phi(f(\mathbf{X}'), t) > P(f, \mu) \end{array} \right\} \\ &\leq 2\mathbb{E}_{\mathbf{X}\mathbf{X}'} \sup_{f \in \mathcal{F}} \mathbb{1} \left\{ \begin{array}{l} P(f, \mu) > \Psi(f(\mathbf{X}), t) + s \\ \text{and } \Phi(f(\mathbf{X}'), t) > P(f, \mu) \end{array} \right\} \\ &\leq 2\mathbb{E}_{\mathbf{X}\mathbf{X}'} \sup_{f \in \mathcal{F}} \mathbb{1} \{ \Phi(f(\mathbf{X}'), t) > \Psi(f(\mathbf{X}), t) + s \} \\ &\leq 2 \sup_{(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^{2n}} \Pr_{\sigma} \{ \exists f \in \mathcal{F} : \Phi(f(\sigma, \mathbf{x}, \mathbf{x}'), t) \\ &\quad > \Psi(f(-\sigma, \mathbf{x}, \mathbf{x}'), t) + s \}, \end{aligned}$$

where we used Lemma 13 in the last step.

Now we fix  $(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^{2n}$  and let  $\epsilon > 0$  be arbitrary. We can choose a finite subset  $\mathcal{F}_0$  of  $\mathcal{F}$  such that  $|\mathcal{F}_0| \leq \mathcal{N}(\epsilon, \mathcal{F}, 2n)$  and that  $\forall f \in \mathcal{F}$  there exists  $\hat{f} \in \mathcal{F}_0$  such that  $|f(x_i) - \hat{f}(x_i)| < \epsilon$  and  $|f(x'_i) - \hat{f}(x'_i)| < \epsilon$ , for all  $i \in \{1, \dots, n\}$ . Suppose there exists  $f \in \mathcal{F}$  such that

$$\Phi(f(\sigma, \mathbf{x}, \mathbf{x}'), t) > \Psi(f(-\sigma, \mathbf{x}, \mathbf{x}'), t) + \left(2 + 8\sqrt{\frac{t}{n}}\right) \epsilon.$$

It follows from the Lemma 12 (i) and (ii) that there must exist  $\hat{f} \in \mathcal{F}_0$  such that

$$\Phi(\hat{f}(\sigma, \mathbf{x}, \mathbf{x}'), t) > \Psi(\hat{f}(-\sigma, \mathbf{x}, \mathbf{x}'), t).$$

We conclude from the above that

$$\begin{aligned} & \Pr_{\sigma} \left\{ \begin{array}{l} \exists f \in \mathcal{F} : \Phi(f(\sigma, \mathbf{x}, \mathbf{x}'), t) > \\ > \Psi(f(-\sigma, \mathbf{x}, \mathbf{x}'), t) + \left(2 + 8\sqrt{\frac{t}{n}}\right) \epsilon \end{array} \right\} \\ &\leq \Pr_{\sigma} \{ \exists f \in \mathcal{F}_0 : \Phi(f(\sigma, \mathbf{x}, \mathbf{x}'), t) > \Psi(f(-\sigma, \mathbf{x}, \mathbf{x}'), t) \} \\ &\leq \sum_{f \in \mathcal{F}_0} \Pr_{\sigma} \{ \Phi(f(\sigma, \mathbf{x}, \mathbf{x}'), t) > \Psi(f(-\sigma, \mathbf{x}, \mathbf{x}'), t) \} \\ &\leq 5\mathcal{N}(\epsilon, \mathcal{F}, 2n) e^{-t}, \end{aligned}$$

where we used Lemma 14 in the last step. We arrive at the statement that

$$\begin{aligned} & \Pr_{\mathbf{X}} \left\{ \exists f \in \mathcal{F} : P(f, \mu) \geq \Psi(f(\mathbf{X}), t) + \left(2 + 8\sqrt{\frac{t}{n}}\right) \epsilon \right\} \\ &\leq 10\mathcal{N}(\epsilon, \mathcal{F}, 2n) e^{-t}. \end{aligned}$$

Equating this probability to  $\delta$ , solving for  $t$ , substituting  $\epsilon = 1/n$  and using  $8\sqrt{t/n} \leq 2t$ , for  $n \geq 16$  and  $t \geq 1$ , give the result.  $\blacksquare$

We remark that a simplified version of the above argument gives uniform bounds for the standard deviation  $\sqrt{V(f, \mu)}$ , using Theorem 10 (4) and (3).

## 4 Sample variance penalization versus empirical risk minimization

Since empirical Bernstein bounds are observable, have estimation errors which can be as small as  $O(1/n)$  for small sample variances, and can be adjusted to hold uniformly over realistic function classes, they suggest a method which minimizes the bounds of Corollary 5 or Theorem 6. Specifically we consider the algorithm

$$SVP_{\lambda}(\mathbf{X}) = \arg \min_{f \in \mathcal{F}} P_n(f, \mathbf{X}) + \lambda \sqrt{\frac{V_n(f, \mathbf{X})}{n}}, \quad (10)$$

where  $\lambda$  is a non-negative parameter. We call this method sample variance penalization (SVP). Choosing the regularization parameter  $\lambda = 0$  reduces the algorithm to empirical risk minimization (ERM).

It is intuitively clear that SVP will be inferior to ERM if losses corresponding to better hypotheses have larger variances than the worse ones. But this seems to be a somewhat unnatural situation. If, on the other hand, there are some optimal hypotheses of small variance, then SVP should work well. To make this rigorous we provide a result, which can be used to bound the excess risk of  $SVP_{\lambda}$ . Below we use Theorem 6, but it is clear how the argument is to be modified to obtain better constants for finite hypothesis spaces.

**Theorem 15** *Let  $X$  be a random variable with values in a set  $\mathcal{X}$  with distribution  $\mu$ , and let  $\mathcal{F}$  be a class of hypotheses  $f : \mathcal{X} \rightarrow [0, 1]$ . Fix  $\delta \in (0, 1)$ ,  $n \geq 2$  and set  $\mathcal{M}(n) = 10\mathcal{N}_{\infty}(1/n, \mathcal{F}, 2n)$  and  $\lambda = \sqrt{18 \ln(3\mathcal{M}(n)/\delta)}$ .*

*Fix  $f^* \in \mathcal{F}$ . Then with probability at least  $1 - \delta$  in the draw of  $\mathbf{X} \sim \mu^n$ ,*

$$\begin{aligned} & P(SVP_{\lambda}(\mathbf{X}), \mu) - P(f^*, \mu) \\ &\leq \sqrt{\frac{32V(f^*, \mu) \ln(3\mathcal{M}(n)/\delta)}{n}} \\ &\quad + \frac{22 \ln(3\mathcal{M}(n)/\delta)}{n-1}. \end{aligned}$$

**Proof:** Denote the hypothesis  $SVP_{\lambda}(\mathbf{X})$  by  $\hat{f}$ . By Theorem 6 we have with probability at least  $1 - \delta/3$  that

$$\begin{aligned} P(\hat{f}, \mu) &\leq P_n(\hat{f}, \mathbf{X}) + \lambda \sqrt{\frac{V_n(\hat{f}, \mathbf{X})}{n}} + \frac{15\lambda^2}{18(n-1)} \\ &\leq P_n(f^*, \mathbf{X}) + \lambda \sqrt{\frac{V_n(f^*, \mathbf{X})}{n}} + \frac{15\lambda^2}{18(n-1)}. \end{aligned}$$

The second inequality follows from the definition of  $SV P_\lambda$ . By Bennett's inequality (Theorem 3) we have with probability at least  $1 - \delta/3$  that

$$P_n(f^*, \mathbf{X}) \leq P(f^*, \mu) + \sqrt{\frac{2V(f^*, \mu) \ln 3/\delta}{n}} + \frac{\ln 3/\delta}{3n}$$

and by Theorem 10 (4) we have with probability at least  $1 - \delta/3$  that

$$\sqrt{V_n(f^*, \mathbf{X})} \leq \sqrt{V(f^*, \mu)} + \sqrt{\frac{2 \ln 3/\delta}{n-1}}.$$

Combining these three inequalities in a union bound and using  $\ln(3\mathcal{M}(n)/\delta) \geq 1$  and some other crude but obvious estimates, we obtain with probability at least  $1 - \delta$

$$P(\hat{f}, \mu) \leq P(f^*, \mu) + \sqrt{\frac{32V(f^*, \mu) \ln(3\mathcal{M}(n)/\delta)}{n}} + \frac{22 \ln(3\mathcal{M}(n)/\delta)}{n-1}.$$

■

If we let  $f^*$  be an optimal hypothesis we obtain a bound on the excess risk. The square-root term in the bound scales with the standard deviation of this hypothesis, which can be quite small. In particular, if there is an optimal (minimal risk) hypothesis of zero variance, then the excess risk of the hypothesis chosen by SVP decays as  $(\ln \mathcal{M}(n))/n$ . In the case of finite hypothesis spaces  $\mathcal{M}(n) = |\mathcal{F}|$  is independent of  $n$  and the excess risk then decays as  $1/n$ . Observe that apart from the complexity bound on  $\mathcal{F}$  no assumption such as convexity of the function class or special properties of the loss functions were needed to derive this result.

To demonstrate a potential competitive edge of SVP over ERM we will now give a very simple example of this type, where the excess risk of the hypothesis chosen by ERM is of order  $O(1/\sqrt{n})$ .

Suppose that  $\mathcal{F}$  consists of only two hypotheses  $\mathcal{F} = \{c_{1/2}, b_{1/2+\epsilon}\}$ . The underlying distribution  $\mu$  is such that  $c_{1/2}(X) = 1/2$  almost surely and  $b_{1/2+\epsilon}(X)$  is a Bernoulli variable with expectation  $1/2 + \epsilon$ , where  $\epsilon \leq 1/\sqrt{8}$ . The hypothesis  $c_{1/2}$  is optimal and has zero variance, the hypothesis  $b_{1/2+\epsilon}$  has excess risk  $\epsilon$  and variance  $1/4 - \epsilon^2$ . We are given an i.i.d. sample  $\mathbf{X} = (X_1, \dots, X_n) \sim \mu^n$  on which we are to base the selection of either hypothesis.

It follows from the previous theorem (with  $f^* = c_{1/2}$ ), that the excess risk of  $SV P_\lambda$  decays as  $1/n$ , for suitably chosen  $\lambda$ . To make our point we need to give a lower bound for the excess risk of empirical risk minimization. We use the following inequality due to Slud which we cite in the form given in [1, p. 363].

**Theorem 16** *Let  $B$  be a binomial  $(n, p)$  random variable with  $p \leq 1/2$  and suppose that  $np \leq t \leq n(1-p)$ . Then*

$$\Pr\{B > t\} \geq \Pr\left\{Z > \frac{t - np}{\sqrt{np(1-p)}}\right\},$$

where  $Z$  is a standard normal  $N(0, 1)$ -distributed random variable.

Now ERM selects the inferior hypothesis  $b_{1/2+\epsilon}$  if

$$P_n(b_{1/2+\epsilon}, \mathbf{X}) < P_n(c_{1/2}, \mathbf{X}) = 1/2.$$

We therefore obtain from Theorem 16, with

$$B = n(1 - P_n(b_{1/2+\epsilon}(\mathbf{X}))),$$

$p = 1/2 - \epsilon$  and  $t = n/2$  that

$$\begin{aligned} \Pr\{ERM(\mathbf{X}) = b_{1/2+\epsilon}\} &= \Pr\{P_n(b_{1/2+\epsilon}(\mathbf{X})) < 1/2\} \\ &\geq \Pr\{B > t\} \\ &\geq \Pr\left\{Z > \frac{\sqrt{n}\epsilon}{\sqrt{1/4 - \epsilon^2}}\right\} \end{aligned}$$

A well known bound for standard normal random variables gives for  $\eta > 0$

$$\begin{aligned} \Pr\{Z > \eta\} &\geq \frac{1}{\sqrt{2\pi}} \frac{\eta}{1 + \eta^2} \exp\left(-\frac{\eta^2}{2}\right) \\ &\geq \exp(-\eta^2), \text{ if } \eta \geq 2. \end{aligned}$$

If we assume  $n \geq \epsilon^{-2}$  we have  $\sqrt{n}\epsilon/\sqrt{1/4 - \epsilon^2} \geq 2$ , so

$$\Pr\{ERM(\mathbf{X}) = b_{1/2+\epsilon}\} \geq \exp\left(-\frac{n\epsilon^2}{1/4 - \epsilon^2}\right) \geq e^{-8n\epsilon^2},$$

where we used  $\epsilon \leq 1/\sqrt{8}$  in the last inequality. Since this is just the probability that the excess risk is  $\epsilon$  we arrive at the following statement: For every  $n \geq \epsilon^{-2}$  there exists  $\delta (= e^{-8n\epsilon^2})$  such that the excess risk of the hypothesis generated by ERM is at least

$$\epsilon = \sqrt{\frac{\ln 1/\delta}{8n}},$$

with probability at least  $\delta$ . Therefore the excess risk for ERM cannot have a faster rate than  $O(1/\sqrt{n})$ .

This example is of course a very artificial construction, chosen as a simple illustration. It is clear that the conclusions do not change if we add any number of deterministic hypotheses with risk larger than  $1/2$  (they simply have no effect), or if we add any number of Bernoulli hypotheses with risk at least  $1/2 + \epsilon$  (they just make things worse for ERM).

To obtain a more practical insight into the potential advantages of SVP we have conducted a simple experiment, where  $\mathcal{X} = [0, 1]^K$  and the random variable  $X \in \mathcal{X}$  is distributed according to  $\prod_{k=1}^K \mu_{a_k, b_k}$  where

$$\mu_{a,b} = (1/2)(\delta_{a-b} + \delta_{a+b}).$$

Each coordinate  $\pi_k(X)$  of  $X$  is thus a binary random variable, assuming the values  $a_k - b_k$  and  $a_k + b_k$  with equal probability, having expectation  $a_k$  and variance  $b_k^2$ .

The distribution of  $X$  is itself generated at random by selecting the pairs  $(a_k, b_k)$  independently:  $a_k$  is chosen from the uniform distribution on  $[B, 1 - B]$  and the standard deviation  $b_k$  is chosen from the uniform distribution on the interval  $[0, B]$ . Thus  $B$  is the only parameter governing the generation of the distribution.

As hypotheses we just take the  $K$  coordinate functions  $\pi_k$  in  $[0, 1]^K$ . Selecting the  $k$ -th hypothesis then just means that

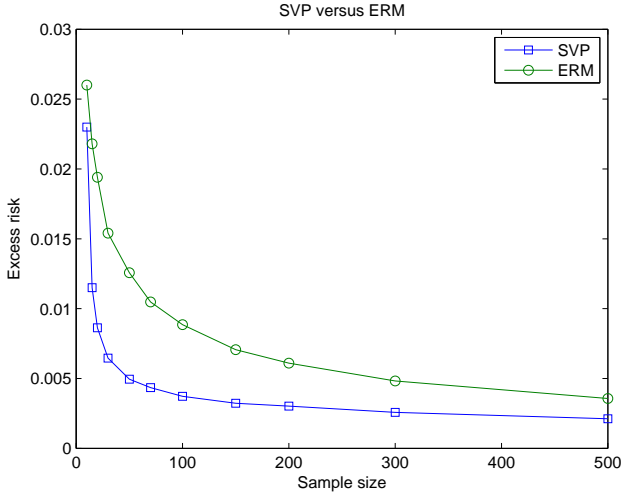


Figure 1: Comparison of the excess risks of the hypotheses returned by ERM (circled line) and SVP with  $\lambda = 2.5$  (squared line) for different sample sizes.

we select the corresponding distribution  $\mu_{a_k, b_k}$ . Of course we want to find a hypothesis of small risk  $a_k$ , but we can only observe  $a_k$  through the corresponding sample, the observation being obscured by the variance  $b_k^2$ .

We chose  $B = 1/4$  and  $K = 500$ . We tested the algorithm (10) with  $\lambda = 0$ , corresponding to ERM, and  $\lambda = 2.5$ . The sample sizes ranged from 10 to 500. We recorded the true risks of the respective hypotheses generated, and averaged these risks over 10000 randomly generated distributions. The results are reported in Figure 1 and show clearly the advantage of SVP in this particular case. It must however be pointed out that this advantage, while being consistent, is small compared to the risk of the optimal hypotheses (around  $1/4$ ).

If we try to extract a practical conclusion from Theorem 15, our example and the experiment, then it appears that SVP might be a good alternative to ERM, whenever the optimal members of the hypothesis space still have substantial risk (for otherwise ERM would do just as good), but there are optimal hypotheses of very small variance. These two conditions seem to be generic for many noisy situations: when the noise arises from many independent sources, but does not depend too much on any single source, then the loss of an optimal hypothesis should be sharply concentrated around its expectation (e.g. by the bounded difference inequality - see [8]), resulting in a small variance.

## 5 Application to sample compression

Sample compression schemes [6] provide an elegant method to reduce a potentially very complex function class to a finite, data-dependent subclass. With  $\mathcal{F}$  being as usual, assume that some algorithm  $A$  is already specified by a fixed function

$$A : \mathbf{X} \in \bigcup_{n=1}^{\infty} \mathcal{X}^n \mapsto A_{\mathbf{X}} \in \mathcal{F}.$$

The function  $A_S$  can be interpreted as the hypothesis chosen by the algorithm on the basis of the training set  $S$ , composed with the fixed loss function. For  $x \in \mathcal{X}$  the quantity  $A_S(x)$  is thus the loss incurred by training the algorithm from  $S$  and applying the resulting hypothesis to  $x$ .

The idea of sample compression schemes [6] is to train the algorithm on subsamples of the training data and to use the remaining data points for testing. A comparison of the different results then leads to the choice of a subsample and a corresponding hypothesis. If this hypothesis has small risk, we can say that the problem-relevant information of the sample is present in the subsample in a compressed form, hence the name.

Since the method is crucially dependent on the quality of the individual performance estimates, and empirical Bernstein bounds give tight, variance sensitive estimates, a combination of sample compression and SVP is promising. For simplicity we only consider compression sets of a fixed size  $d$ . We introduce the following notation for a subset  $I \subset \{1, \dots, n\}$  of cardinality  $|I| = d$ .

- $A_{\mathbf{X}[I]}$  is the hypothesis trained with  $A$  from the subsample  $\mathbf{X}[I]$  consisting of those examples whose indices lie in  $I$ .

- For  $f \in \mathcal{F}$ , we let

$$P_{I^c}(f) = P_{n-d}(f(\mathbf{X}[I^c])) = \frac{1}{n-d} \sum_{i \notin I} f(X_i),$$

the empirical risk of  $f$  computed on the subsample  $\mathbf{X}[I^c]$  consisting of those examples whose indices do not lie in  $I$ .

- For  $f \in \mathcal{F}$ , we let

$$\begin{aligned} V_{I^c}(f) &= V_{n-d}(f(\mathbf{X}[I^c])) \\ &= \frac{1}{2(n-d)(n-d-1)} \sum_{i, j \notin I} (f(X_i) - f(X_j))^2, \end{aligned}$$

the sample variance of  $f$  computed on  $\mathbf{X}[I^c]$ .

- $\mathcal{C}$  is the collection of subsets  $I \subset \{1, \dots, n\}$  of cardinality  $|I| = d$ .

With this notation we define our sample compression scheme as

$$\begin{aligned} \text{SVP}_{\lambda}(\mathbf{X}) &= A_{\mathbf{X}[\hat{I}]} \\ \hat{I} &= \arg \min_{I \in \mathcal{C}} P_{I^c}(A_{\mathbf{X}[I]}) + \lambda \sqrt{V_{I^c}(A_{\mathbf{X}[I]})}. \end{aligned}$$

As usual,  $\lambda = 0$  gives the classical sample compression schemes. The performance of this algorithm can be guaranteed by the following result.

**Theorem 17** *With the notation introduced above fix  $\delta \in (0, 1)$ ,  $n \geq 2$  and set  $\lambda = \sqrt{2 \ln(6|\mathcal{C}|/\delta)}$ . Then with probability at least  $1 - \delta$  in the draw of  $\mathbf{X} \sim \mu^n$ , we have for every  $I^* \in \mathcal{C}$*

$$\begin{aligned} &P(\text{SVP}_{\lambda}(\mathbf{X}), \mu) - P(A_{\mathbf{X}[I^*]}, \mu) \\ &\leq \sqrt{\frac{8V(A_{\mathbf{X}[I^*]}, \mu) \ln(6|\mathcal{C}|/\delta)}{n-d}} + \frac{14 \ln(6|\mathcal{C}|/\delta)}{3(n-d-1)} \end{aligned}$$

**Proof:** Use a union bound and Theorem 4 to obtain an empirical Bernstein bound uniformly valid over all  $A_{\mathbf{X}[I]}$  with  $I \in \mathcal{C}$  and therefore also valid for  $SV P_\lambda(\mathbf{X})$ . Then follow the proof of Theorem 15. Since now  $I^* \in \mathcal{C}$  is chosen *after* seeing the sample, uniform versions of Bennett’s inequality and Theorem 10 (4) have to be used, and are again readily obtained with union bounds over  $\mathcal{C}$ . ■

The interpretation of this result as an excess risk bound is more subtle than for Theorem 15, because the optimal hypothesis is now sample-dependent. If we define

$$I^* = \arg \min_{I \in \mathcal{C}} P(A_{\mathbf{X}[I]}, \mu),$$

then the theorem tells us how close we are to the choice of the optimal subsample. This will be considerably better than what we get from Hoeffding’s inequality if the variance  $V(A_{\mathbf{X}[I^*]}, \mu)$  is small and sparse solutions are sought in the sense that  $d/n$  is small (observe that  $\ln |\mathcal{C}| \leq d \ln(ne/d)$ ).

This type of relative excess risk bound is of course more useful if the minimum  $P(A_{\mathbf{X}[I^*]}, \mu)$  is close to some true optimum arising from some underlying generative model. In this case we can expect the loss  $A_{\mathbf{X}[I^*]}$  to behave like a noise variable centered at the risk  $P(A_{\mathbf{X}[I^*]}, \mu)$ . If the noise arises from many independent sources, each of which makes only a small contribution, then  $A_{\mathbf{X}[I^*]}$  will be sharply concentrated and have a small variance  $V(A_{\mathbf{X}[I^*]}, \mu)$ , resulting in tight control of the excess risk.

## 6 Conclusion

We presented sample variance penalization as a potential alternative to empirical risk minimization and analyzed some of its statistical properties in terms of empirical Bernstein bounds and concentration properties of the empirical standard deviation. The promise of our method is that, in simple but perhaps practical scenarios the excess risk of our method is guaranteed to be substantially better than that of empirical risk minimization.

The present work raises some questions. Perhaps the most pressing issue is to find an efficient implementation of the method, to deal with the fact that sample variance penalization is non-convex in many situations when empirical risk minimization is convex, and to compare the two methods on some real-life data sets. Another important issue is to further investigate the application of empirical Bernstein bounds to sample compression schemes.

## References

- [1] M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK, 1999.
- [2] J. Y. Audibert, R. Munos, C. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. To appear in *Theoretical Computer Science*.
- [3] Y. G. Guo, P. L. Bartlett, J. Shawe-Taylor, R. C. Williamson. Covering numbers for support vector machines. *Proceedings of COLT*, 1999.
- [4] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13-30, 1963.
- [5] W. S. Lee, P. L. Bartlett, R. C. Williamson. The Importance of Convexity in Learning with Squared Loss. *IEEE Trans. Info. Theory* 44(5):1974-1980, 1998.
- [6] N. Littlestone and M. K. Warmuth. Relating data compression and learnability. Technical report, University of California Santa Cruz, Santa Cruz, CA, 1986.
- [7] A. Maurer. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms*, 29:121–138, 2006.
- [8] C. McDiarmid. *Concentration*. In *Probabilistic Methods of Algorithmic Discrete Mathematics*, pages 195–248. Springer, 1998.
- [9] V. Mnih, C. Szepesvári, J. Y. Audibert. Empirical Bernstein Stopping. In Proc. ICML 2008.
- [10] V. Vapnik. *The Nature of Statistical Learning Theory*, Springer 1995.