A vector-contraction inequality for Rademacher complexities

Andreas Maurer

Abstract. The contraction inequality for Rademacher averages is extended to Lipschitz functions with vector-valued domains, and it is also shown that in the bounding expression the Rademacher variables can be replaced by arbitrary iid symmetric and sub-gaussian variables. Example applications are given for multi-category learning, K-means clustering and learning-to-learn.

1 Introduction

The method of Rademacher complexities has become a popular tool to prove generalization in learning theory. One has the following result [1], which gives a bound on the estimation error, uniform over a loss class $\mathcal{F}$.

Theorem 1. *Let $\mathcal{X}$ be any set, $\mathcal{F}$ a class of functions $f : \mathcal{X} \to [0,1]$ and let $X, X_1, ..., X_n$ be iid random variables with values in $\mathcal{X}$. Then for $\delta > 0$ with probability at least $1-\delta$ in , $\mathbf{X} = (X_1, ..., X_n)$ we have for every $f \in \mathcal{F}$ that*

$$\mathbb{E}f(X) \leq \frac{1}{n}\sum f(X_i) + \frac{2}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i f(X_i)\,|\mathbf{X}\right] + \sqrt{\frac{9\ln 2/\delta}{2n}}.$$

Here the $\epsilon_1, ..., \epsilon_n$ are (and will be throughout this paper) independent Rademacher variables, uniformly distributed on $\{-1, 1\}$. For any class $\mathcal{F}$ of real, not necessarily $[0,1]$-valued, functions defined on $\mathcal{X}$, and any vector $\mathbf{x} = (x_1, ..., x_n) \in \mathcal{X}^n$, the quantity

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i f(x_i)$$

is called the Rademacher complexity of the class $\mathcal{F}$ on the sample $x = (x_1, ..., x_n) \in \mathcal{X}^n$. Here we omit the customary factor $2/n$, as this will simplify most of our statements below.

Most applications of the method at some point or another use the so-called contraction inequality. For functions $h_i : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant L, the scalar contraction inequality states that

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i h_i(f(x_i)) \leq L\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i f(x_i).$$

There are situations when it is desirable to extend this result to the case, when the class $\mathcal{F}$ consists of vector-valued functions and the loss functions are Lipschitz functions defined on a more than one-dimensional space. Such occurs for example in the analysis of multi-class learning, K-means clustering or learning-to-learn. At present one has dealt with these problems by passing to Gaussian averages and using Slepian's inequality (see e.g. Theorem 14 in [1]). This is sufficient for many applications, but there are two drawbacks: 1. the proof relies on a highly nontrivial result (Slepian's inequality) and 2. while Rademacher complexities are tightly bounded in terms of Gaussian complexities, it is well known ([12], [4]) that bounding the latter in terms of the former incurs a factor logarithmic in the number of variables, potentially resulting in an unnecessary weakening of the results (see e.g. [13]).

In this paper we will prove the vector contraction inequality

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i h_i\left(f\left(x_i\right)\right)\leq\sqrt{2}L\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sum_{k=1}^{K}\epsilon_{ik}f_k\left(x_i\right), \tag{1}$$

where the members of $\mathcal{F}$ take values in $\mathbb{R}^K$ with component functions $f_k\left(.\right)$, the h_i are L-Lipschitz functions from $\mathbb{R}^K$ (with Euclidean norm) to $\mathbb{R}$, and the ϵ_{ik} are an $n\times K$ matrix of independent Rademacher variables. It is also shown that the ϵ_{ik} on the right hand side of (1) can be replaced by arbitrary iid random variables as long as they are symmetric and sub-gaussian, and $\sqrt{2}$ is replaced by a suitably chosen constant. Furthermore the result extends to infinite dimensions in the sense that $\mathbb{R}^K$ can be replaced by the Hilbert space ℓ_2. The proof given is self-contained and independent of Slepian's inequality.

We illustrate applications of this inequality by showing that it applies to loss function in a variety of relevant cases. In Section 3 we discuss multi-class learning, K-means clustering and learning-to-learn. We also give some indications of how the vector-valued complexity on the right hand side of (1) may be bounded. An example pertaining to the truly infinite dimensional case is given, generalizing some bounds for least-squares regression with operator valued kernels ([19], [5]) to more general loss-functions.

The inequality (1) is perhaps not the most natural form of a vector-contraction inequality, and, since the right hand side is sometimes difficult to bound, one is led to look for alternatives. An attractive conjecture might be the following.

Conjecture 1. Let $\mathcal{X}$ be any set, $n\in\mathbb{N}$, $(x_1,...,x_n)\in X^n$, let $\mathcal{F}$ be a class of functions $f:\mathcal{X}\to\ell_2$ and let $h:\ell_2\to\mathbb{R}$ have Lipschitz norm L. Then

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i}\epsilon_i h\left(f\left(x_i\right)\right)\leq KL\mathbb{E}\sup_{f\in\mathcal{F}}\left\|\sum_{i}\epsilon_i f\left(x_i\right)\right\|,$$

where K is some universal constant.

This conjecture is false and will be disproved in the sequel.

A version of the scalar contraction inequality occurs in [12], Theorem 4.12. There the absolute value of the Rademacher sum is used, and a necessary factor of two appears on right hand side. With the work of Koltchinski and Panchenko [11] and Bartlett and Mendelson [1] Rademacher averages became attractive to the machine learning community, there was an increased interest in contraction inqualities and it was realized that the absolute value was unnecessary for most of the new applications. Meir and Zhang [18] gave a nice and simple proof of the scalar contraction inequality as stated above. Our proof of (1) is an extension of their method.

2 The vector-contraction inequality

All random variables are assumed to be defined on some probability space (Ω, Σ). The space $L_p(\Omega, \Sigma)$ is abbreviated L_p. We use ℓ_2 to denote the Hilbert space of square summable sequences of real numbers. The norm on ℓ_2 and the Euclidean norm on $\mathbb{R}^K$ are denoted with $\|.\|$.

A real random variable X is called *symmetric* if $-X$ and X are identically distributed. It is called *sub-gaussian* if there exists a constant $b = b(X)$ such that for every $\lambda \in \mathbb{R}$

$$\mathbb{E}e^{\lambda X} \leq e^{\frac{\lambda^2 b^2}{2}}.$$

We call b the *sub-gaussian parameter* of X. Rademacher and standard normal variables are symmetric and sub-gaussian.

The following is the main result of this paper.

Theorem 2. *Let X be nontrivial, symmetric and subgaussian. Then there exists a constant $C < \infty$, depending only on the distribution of X, such that for any countable set $\mathcal{S}$ and functions $\psi_i : \mathcal{S} \to \mathbb{R}$, $\phi_i : \mathcal{S} \to \ell_2$, $1 \leq i \leq n$ satisfying*

$$\forall s, s' \in \mathcal{S},\ \psi_i(s) - \psi(s_i') \leq \|\phi_i(s) - \phi_i(s')\|$$

we have.

$$\mathbb{E}\sup_{s\in\mathcal{S}} \sum_i \epsilon_i \psi_i(s) \leq C\ \mathbb{E}\sup_{s\in\mathcal{S}} \sum_{i,k} X_{ik}\phi_i(s)_k,$$

where the X_{ik} are independent copies of X for $1 \leq i \leq n$ and $1 \leq k \leq \infty$, and $\phi_i(s)_k$ is the k-th coordinate of $\phi_i(s)$.

If X is a Rademacher variable we may choose $C = \sqrt{2}$, if X is standard normal $C = \sqrt{\pi/2}$.

For applications in learning theory we can at once substitute a Rademacher variable for X and $\sqrt{2}$ for C. For $\mathcal{S}$ we take a class $\mathcal{F}$ of vector valued functions $f : \mathcal{X} \to \ell_2$, for the ϕ_i the evaluation functionals on a sample $(x_1, ..., x_n)$, so that $\phi_i(f) = f(x_i)$ and for ψ_i we take the evaluation functionals composed with a Lipschitz loss function $h : \ell_2 \to \mathbb{R}$ of Lipschitz norm L. We obtain the

Corollary 1. *Let $\mathcal{X}$ be any set, $(x_1, ..., x_n) \in \mathcal{X}^n$, let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \ell_2$ and let $h_i : \ell_2 \to \mathbb{R}$ have Lipschitz norm L. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_i \epsilon_i h_i \left(f \left(x_i \right) \right) \leq \sqrt{2} L \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i,k} \epsilon_{ik} f_k \left(x_i \right),$$

where ϵ_{ik} is an independent doubly indexed Rademacher sequence and $f_k \left(x_i \right)$ is the k-th component of $f \left(x_i \right)$.

Clearly finite dimensional versions are obtained by restricting to the subspace spanned by the first K coordinate functions in ℓ_2.

3 Examples of loss functions

We give some examples of seemingly complicated loss functions to which Theorem 2 and Corollary 1 can be applied. These examples are not exhaustive, in fact it seems that many applications of Slepian's inequality in the machine learning literature can be circumvented by Theorem 2 (see also [6], [7], [17] for applications to information retrieval and generalization of autoregressive models).

3.1 Multi-class classification

Consider the problem of assigning to inputs taken from a space $\mathcal{X}$ a label corresponding to one of K classes. We are given a labelled iid sample $\mathbf{z} = \left(\left(x_1, y_1 \right), ..., \left(x_n, y_n \right) \right)$ drawn from some unknown distribution on $\mathcal{X} \times \{1, ..., K\}$, where the points x_i are inputs $x_i \in \mathcal{X}$ and the y_i are corresponding labels, $y_i \in \{1, ..., K\}$. Many approaches assume that there is class $\mathcal{F}$ of vector valued functions $f : \mathcal{X} \to \mathbb{R}^{K'}$, where $K' = o(K)$ (typically $K' = K$, for 1-versus-all classification, or $K' = K - 1$ for simplex coding [20]), a classification rule $c : \mathbb{R}^{K'} \to \{1, ..., K\}$, and for each label $k \in \{1, ..., c\}$ a loss function $\ell_k : \mathbb{R}^K \to \mathbb{R}_+$. The loss function ℓ_k is designed so as to upper bound, or approximate the indicator function of the set $\left\{ z \in \mathbb{R}^{K'} : c(z) \neq k \right\}$ (see [9], [13]). In most cases the loss functions are Lipschitz on $\mathbb{R}^K$ relative to the euclidean norm, with some Lipschitz constant L. The empirical error incurred by a function $f \in \mathcal{F}$ is

$$\frac{1}{n} \sum_i \ell_{y_i} \left(f \left(x_i \right) \right).$$

The Rademacher complexity, which would lead to the uniform bound on the estimation error, is

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_i \epsilon_i \ell_{y_i} \left(f \left(x_i \right) \right).$$

Using Corollary 1 with $h_i = \ell_{y_i}$ we can immediately eliminate the loss functions ℓ_{y_i}

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_i \epsilon_i \ell_{y_i} \left(f \left(x_i \right) \right) \leq \sqrt{2} L \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i,k} \epsilon_{ik} f_k \left(x_i \right).$$

How we proceed to further bound this now depends on the nature of the vector-valued class $\mathcal{F}$. Some techniques to bound the Rademacher complexity of vector valued classes are sketched in Section 4 below.

3.2 K-means clustering

Let H be a Hilbert space and $\mathbf{x} = (x_1, ..., x_n)$ a sample of points in the unit ball B_1 of H. The algorithm seeks centers $c = (c_1, ..., c_K) \in \mathcal{S} = B_1^K$ to represent the sample.

$$c^* = \arg \min_{(c_1,...,c_K)\in\mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} \min_{k=1}^{K} \|x_i - c_k\|^2 .$$

The corresponding Rademacher average to bound the estimation error is

$$R(\mathcal{S}, \mathbf{x}) = \mathbb{E} \sup_{c\in\mathcal{S}} \sum_i \epsilon_i \min_{k=1}^{K} \|x_i - c_k\|^2 = \mathbb{E} \sup_{c\in\mathcal{S}} \sum_i \epsilon_i \psi_i(c),$$

where we define $\psi_i(c) = \min_k \|x_i - c_k\|^2$ in preparation of an application of Theorem 2. The next step is to search for an appropriate Lipschitz property of the ψ_i. We have, for $c, c' \in \mathcal{S}$,

$$\begin{aligned}
\psi_i(c) - \psi_i(c') &= \min_k \|x_i - c_k\|^2 - \min_k \|x_i - c'_k\|^2 \\
&\leq \max_k \|x_i - c_k\|^2 - \|x_i - c'_k\|^2 \\
&\leq \left(\sum_k \left(\|x_i - c_k\|^2 - \|x_i - c'_k\|^2 \right)^2 \right)^{1/2} \\
&= \|\phi_i(c) - \phi_i(c')\| .
\end{aligned}$$

Where we defined $\phi_i : \mathcal{S} \to \mathbb{R}^K$ by $\phi_i(c) = \left(\|x_i - c_1\|^2, ..., \|x_i - c_K\|^2 \right)$. We can now apply Theorem 2 with $L = 1$ and obtain

$$\begin{aligned}
2^{-1/2} R(\mathcal{S}, \mathbf{x}) &\leq \mathbb{E} \sup_{c\in\mathcal{S}} \sum_{ik} \epsilon_{ik} \|x_i - c_k\|^2 \\
&\leq 2\mathbb{E} \sup_{c\in\mathcal{S}} \sum_{ik} \epsilon_{ik} \langle x_i, c_k \rangle + \mathbb{E} \sup_{c\in\mathcal{S}} \sum_{ik} \epsilon_{ik} \|c_k\|^2 \\
&\leq K \left(2\mathbb{E} \left\| \sum_i \epsilon_i x_i \right\| + \mathbb{E} \left| \sum_i \epsilon_i \right| \right) \\
&\leq 3K\sqrt{n}.
\end{aligned}$$

Dividing by n we obtain generalization bounds as in [3] or [15]. In this simple case it was very easy to explicitely bound the complexity of the vector-valued class.

3.3 Learning to learn or meta-learning

With input space $\mathcal{X}$ suppose we have a class $\mathcal{H}$ of feature maps $h : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}^K$ and a loss class $\mathcal{F}$ of functions $f : \mathcal{Y} \to [0,1]$. The loss class could be used for classification or function estimation or also in some unsupervised setting. We assume that every function $f \in \mathcal{F}$ is Lipschitz with Lipschitz constant L and that $\mathcal{F}$ is small enough for good generalization in the sense that for some $B < \infty$

$$\mathbb{E}_{Y_1,...,Y_n}\left[\sup_{f\in\mathcal{F}} \mathbb{E}_Y f(Y) - \frac{1}{n}\sum_{i=1}^{n} f(Y_i)\right] \leq \frac{B}{\sqrt{n}} \tag{2}$$

for any $\mathcal{Y}$-valued random variable Y and iid copies $Y_1, ..., Y_n$. Such conditions might be established using standard techniques, for example also Rademacher complexities.

We now want to learn a feature map $h \in \mathcal{H}$, such that empirical risk minimization (ERM) with the h-dependent function class $\mathcal{F} \circ h = \{x \mapsto f(h(x)) : f \in \mathcal{F}\}$ gives good results on future, yet unseen, tasks. Of course this depends on the tasks in question, and a good feature map h can only be chosen on the basis of some kind of experience made with these tasks.

To formalize this Baxter [2] has introduced the notion of an *environment* η, which is a distribution on the set of tasks, where each task t is characterized by some distribution μ_t (e.g. on inputs and outputs). For each task $t \sim \eta$ we can then also draw an iid training sample $\mathbf{x}^t = (x_1^t, ..., x_n^t) \sim \mu_t^n$. In this way the environment also induces a distribution on the set of training samples. Now we can make our problem more precise:

Suppose we have T tasks and corresponding training samples $\bar{\mathbf{x}} = \left(\mathbf{x}^1, ..., \mathbf{x}^T\right)$ drawn iid from the environment η. For $h \in \mathcal{H}$ let

$$\psi_t(h) = \min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f\left(h\left(x_i^t\right)\right)$$

be the training error obtained by the use of the feature map h. We propose to use the feature map

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \frac{1}{T}\sum_{t=1}^{T} \psi_t(h).$$

To give a performance guarantee for ERM using $\mathcal{F} \circ \hat{h}$, we now seek to bound the expected *training error* $\mathbb{E}_{t\sim\eta}[\psi_t(h)]$ for a new task drawn from the environment (with corresponding training sample), in terms of the average of the observed $\psi_t(h)$, uniformly over the set of feature maps $h \in \mathcal{H}$. Observe that, given the bound on (2) such a bound will also give a bound on the expected true error when using $\hat{h}$ on new tasks in the environment η, a meta-generalization bound, so to speak (for more details on this type of argument see [14] or [16]).

The Rademacher average in question is

$$R(\mathcal{H}, \bar{\mathbf{x}}) = \mathbb{E}\sup_{h\in\mathcal{H}} \sum_{t=1}^{T} \epsilon_t \psi_t(h).$$

To apply Theorem 2 we look for a Lipschitz property of the ψ_t. For $h \in \mathcal{H}$ define $\phi_t(h) \in \mathbb{R}^{Kn}$ by $[\phi_t(h)]_{k,i} = h_k(x_i^t)$. Then for $h, h' \in \mathcal{H}$

$$\begin{aligned}\psi_t(h) - \psi_t(h') &\leq \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(h(x_i)) - f(h(x_i)) \\ &\leq \frac{L}{n} \sum_{i=1}^{n} \|h(x_i) - h'(x_i)\| \leq \frac{L}{\sqrt{n}} \|\phi_t(h) - \phi_t(h')\|,\end{aligned}$$

where the first inequality comes from the Lipschitz property of the functions in $\mathcal{F}$ and the second from Jensen's inequality. From Theorem 2 we conclude that

$$R(\mathcal{H}, \bar{\mathbf{x}}) \leq \frac{L}{\sqrt{n}} \mathbb{E} \sup_{h \in \mathcal{H}} \sum_{tki} \epsilon_{tki} h_k\left(x_i^t\right).$$

How to proceed depends on the nature of the feature maps in $\mathcal{H}$. Examples are given in [14] or [16], but see also the next section.

4 Bounding the Rademacher complexity of vector-valued classes

At first glance the expression

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i,k} \epsilon_{ik} f_k(x_i)$$

appears difficult to bound. Nevertheless there are some general techniques which can be used, such as the reduction to scalar classes, or the use of duality for linear classes. We also give an example in a truly infinite dimensional setting.

4.1 Reduction to component classes

Suppose $\mathcal{F}_1, ..., \mathcal{F}_K$ are classes of scalar valued functions and define a vector-valued class $\prod_k \mathcal{F}_k$ with values in $\mathbb{R}^K$ by

$$\prod_k \mathcal{F}_k = \{x \mapsto (f_1(x), ..., f_K(x)) : f_k \in \mathcal{F}_k\}.$$

Then, since the constraints are independent, the Rademacher average of the product class

$$\mathbb{E} \sup_{f \in \prod_k \mathcal{F}_k} \sum_{i,k} \epsilon_{ik} f_k(x_i) = \sum_k \mathbb{E} \sup_{f \in \mathcal{F}_k} \sum_i \epsilon_i f(x_i) \tag{3}$$

is just the sum of the Rademacher averages of the scalar valued component classes. Now let $\mathcal{F}$ be any function class with values in $\mathbb{R}^K$ and for $k \in \{1, ..., K\}$ define a scalar-valued class $\mathcal{F}_k$ by

$$\mathcal{F}_k = \{x \mapsto f_k(x) : f = (f_1, ..., f_k, ..., f_K) \in \mathcal{F}\}.$$

Then $\mathcal{F} \subseteq \prod_k \mathcal{F}_k$, so by the identity (3)

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i,k} \epsilon_{ik} f_k (x_i) \leq \sum_k \mathbb{E} \sup_{f \in \mathcal{F}_k} \sum_i \epsilon_i f (x_i) .$$

This is loose in many interesting cases, but for product classes it is unimprovable.

4.2 Linear classes defined by norms

Let H be a separable real Hilbert-space and let $\mathcal{B}\left(H, \mathbb{R}^K\right)$ be the set of bounded linear transformations from H to $\mathbb{R}^K$. Then every member of $\mathcal{B}\left(H, \mathbb{R}^K\right)$ is characterized by a sequence of weight vectors $(w_1, ..., w_K)$ with $w_k \in H$. Let $|||.|||$ be a norm on $\mathcal{B}\left(H, \mathbb{R}^K\right)$ with dual norm $|||.|||_*$. Fix some real number B, and define a class $\mathcal{F}$ of functions from H to $\mathbb{R}^K$ by

$$\mathcal{F} = \left\{ x \mapsto Wx : W \in \mathcal{B}\left(H, \mathbb{R}^K\right), |||W||| \leq B \right\} .$$

Then

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i,k} \epsilon_{ik} f_k (x_i) &= \mathbb{E} \sup_{|||(w_1,...,w_K)||| \leq B} \sum_k \left\langle w_k, \sum_i \epsilon_{ik} x_i \right\rangle \\ &= \mathbb{E} \sup_{|||W||| \leq B} tr (D^* W) \leq B \, \mathbb{E} \, |||D^*|||_* , \end{aligned}$$

where $D \in \mathcal{B}\left(H, \mathbb{R}^K\right)$ is the random transformation

$$v \mapsto \left(\left\langle v, \sum_i \epsilon_{i1} x_i \right\rangle, ..., \left\langle v, \sum_i \epsilon_{iK} x_i \right\rangle \right) .$$

The details of bounding $\mathbb{E} \, |||D^*|||_*$ then depend on the nature of the norm $|||.|||$. The simplest case is the Hilbert-Schmidt or Frobenius norm, where

$$\mathbb{E} \, |||D^*|||_* = \mathbb{E} \sqrt{\sum_k \left\| \sum_i \epsilon_{ik} x_i \right\|^2} \leq \sqrt{K \sum_i \|x_i\|^2} .$$

More interesting are mixed norms or the trace norm. A valuable reference for this approach is [10].

4.3 Operator valued kernels

We give an example in a truly infinite dimensional setting and refer to the mechanism of learning vector valued functions as exposed in [19]. There is a generic separable Hilbert space H and a kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(H)$ satisfying certain positivity and regularity properties as decribed in [19], where $\mathcal{X}$ is some arbitrary input space. Then there exists an induced feature-map $\Phi : \mathcal{X} \to \mathcal{L}(\ell_2, H)$ such that the kernel is given by

$$\kappa (x, y) = \Phi (x) \Phi^* (y)$$

and the class of H-valued functions to be learned is

$$\{x \mapsto f_w(x) = \Phi(x)\, w : \|w\| \leq B\},$$

where also $\|f_w(x)\|_H = \|w\| \sqrt{\kappa(x,x)}$. Then for any sample $\mathbf{x} = (x_1, ..., x_n) \in \mathcal{X}^n$ and L-Lipschitz loss functions $h_i : H \to \mathbb{R}$ we have

$$\begin{aligned}
\mathbb{E} \sup_{\|w\| \leq 1} \sum_i \epsilon_i h_i(\Phi(x_i)\, w) &\leq \sqrt{2} L \mathbb{E} \sup_{\|w\| \leq 1} \sum_{i,k} \epsilon_{ik} \langle \Phi(x_i)\, w, e_k \rangle \\
&= \sqrt{2} L \mathbb{E} \sup_{\|w\| \leq 1} \left\langle w, \sum_{i,k} \epsilon_{ik} \Phi(x_i)^* e_k \right\rangle \\
&\leq \sqrt{2} L B \mathbb{E} \left\| \sum_{i,k} \epsilon_{ik} \Phi(x_i)^* e_k \right\| \\
&\leq \sqrt{2} L B \left(\sum_{i,k} \left\| \Phi(x_i)^* e_k \right\|^2 \right)^{1/2} \\
&= \sqrt{2} L B \left(\sum_i tr\ \kappa(x_i, x_i) \right)^{1/2}.
\end{aligned}$$

Here we used Corollary 1 in the first and Cauchy-Schwarz in the second inequality. Then we use Jensen's inequality combined with orthonormality of the Rademacher sequence. For the result to make sense we need the $\kappa(x_i, x_i)$ to be trace class. In the case $H = \mathbb{R}$ we obtain the standard result for the scalar case, as in [1]. The bound above can be used to prove a non-asymptotic upper bound for the algorithm described in [5], where vector-valued regression with square loss and Tychonov regularization in $\|f_w\| = \|w\|$ is considered.

5 Proof of the contraction inequality

We start with some simple observations on subgaussian random variables.

Lemma 1. *If X is subgaussian with subgaussian-constant b and v is a unit vector in $\mathbb{R}^K$ then*

$$\Pr\left\{ \left| \sum_{k=1}^K v_k X_k \right| > t \right\} \leq 2e^{-t^2/(2b^2)},$$

where $X_1, ..., X_K$ are independent copies of X.

Proof. For any $\lambda \in \mathbb{R}$

$$\begin{aligned}\mathbb{E}\exp\left(\lambda\sum_k v_k X_k\right) &= \prod_k \mathbb{E}\exp\left(\lambda v_k X_k\right)\\ &\leq \prod_k \exp\left(\lambda^2\frac{b^2}{2}v_k^2\right)\\ &= \exp\left(\frac{\lambda^2 b^2}{2}\right).\end{aligned}$$

The first line follows from independence of the X_i, the next because X is subgaussian, and the last because v is a unit vector. It then follows from Markov's inequality that

$$\begin{aligned}\Pr\left\{\sum_k v_k X_k > t\right\} &\leq \mathbb{E}\exp\left(\lambda\left(\sum_k v_k X_k - t\right)\right)\\ &\leq \exp\left(\frac{\lambda^2 b^2}{2} - \lambda t\right)\\ &= e^{-t^2/\left(2b^2\right)},\end{aligned}$$

where the last identity is obtained by optimizing in λ. The conclusion follows from a union bound.

For the purpose of vector-contraction inequalities the crucial property of subgaussian random variables is the following.

Proposition 1. *Let X be nontrivial and subgaussian with subgaussian parameter b and let $\mathbf{X} = (X_1, ..., X_K, ...)$ be an infinite sequence of independent copies of X. Then*

(i) For every $v \in \ell_2$ the sequence of random variables $Y_K = \sum_{i=1}^K X_k v_k$ converges in L_p for $1 \leq p < \infty$ to a random variable denoted by $\sum_{k=1}^\infty X_k v_k$. The map $v \mapsto \sum_{k=1}^\infty X_k v_k$ is a bounded linear transformation from ℓ_2 to L_p.

(ii) There exists a constant $C < \infty$ such that for every $v \in \ell_2$

$$\|v\| \leq C\mathbb{E}\left|\sum_{k=1}^\infty X_k v_k\right|.$$

The proof, given below, is easy and modeled after the proof of the Khintchine inequalities in [12].

For Rademacher variables the best constant is $C = \sqrt{2}$ ([22], see also inequality (4.3) in [12] or Theorem 5.20 in [**?**]). In the standard normal case the inequality in (ii) becomes equality with $C = \sqrt{\pi/2}$. This is an easy consequence of the rotation invariance of isonormal processes.

Proof (Proof of Proposition 1). Let X have subgaussian-constant b.

(i) Assume first that $\|v\| = 1$. For $1 \leq p < \infty$ it follows fom integration by parts that for any $v \in \ell_2$

$$\begin{aligned}\mathbb{E}\left|\sum_{k=1}^{K} v_k X_k\right|^p &= p\int_0^\infty t^{p-1}\Pr\left\{\left|\sum_{k=1}^{K} v_k X_k\right| > t\right\} dt \\ &\leq 2p\int_0^\infty t^{p-1} e^{-t^2/\left(2b^2\right)} dt,\end{aligned}$$

where the last inequality follows from Lemma 1. The last integral is finite and depends only on p and b. By homogeneity it follows that for some constant B and any $v \in \ell_2$

$$\left(\mathbb{E}\left|\sum_{k=1}^{K} v_k X_k\right|^p\right)^{1/p} \leq B\left(\sum_{k=1}^{K} v_k^2\right)^{1/2}$$

which implies convergence in L_p. This proves existence and boundedness of the map $v \mapsto \sum_{k=1}^{\infty} v_k X_k$. Linearity is established with standard arguments.

(ii) Let C be the finite constant

$$C := \frac{\left(8\int_0^\infty t^3 e^{-t^2/\left(2b^2\right)} dt\right)^{1/2}}{\mathbb{E}\left[X^2\right]^{3/2}}.$$

It suffices to prove the conclusion for unit vectors $v \in \ell_2$. From the first part we obtain

$$\mathbb{E}\left|\sum_k v_k X_k\right|^4 \leq 8\int_0^\infty t^3 e^{-t^2/\left(2b^2\right)} dt$$

Combined with Hölder's inequality this implies

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \mathbb{E}\left(\left|\sum_k v_k X_k\right|^2\right) = \mathbb{E}\left(\left|\sum_k v_k X_k\right|^{4/3}\left|\sum_k v_k X_k\right|^{2/3}\right) \\ &\leq \left(\mathbb{E}\left|\sum_k v_k X_k\right|^4\right)^{1/3}\left(\mathbb{E}\left|\sum_k v_k X_k\right|\right)^{2/3} \\ &\leq \left(8\int_0^\infty t^3 e^{-t^2/\left(2b^2\right)} dt\right)^{1/3}\left(\mathbb{E}\left|\sum_k v_k X_k\right|\right)^{2/3}.\end{aligned}$$

Dividing by $\mathbb{E}\left[X^2\right]$ and taking the power of $3/2$ gives

$$1 \leq C\mathbb{E}\left|\sum_k v_k X_k\right|.$$

To prove the main vector contraction result we first consider only a single Rademacher variable ϵ and then complete the proof by induction.

Lemma 2. *Let X be nontrivial, symmetric and subgaussian. Then there exists a constant $C < \infty$ such that for any countable set $\mathcal{S}$ and functions $\psi : \mathcal{S} \to \mathbb{R}$, $\phi : \mathcal{S} \to \ell_2$ and $f : \mathcal{S} \to \mathbb{R}$ satisfying*

$$\forall s, s' \in \mathcal{S},\ \psi(s) - \psi(s') \leq \|\phi(s) - \phi(s')\|$$

we have.

$$\mathbb{E} \sup_{s \in \mathcal{S}} \epsilon\psi(s) + f(s) \leq C\mathbb{E} \sup_{s \in \mathcal{S}} \sum_k X_k \phi(s)_k + f(s),$$

where the X_k are independent copies of X for $1 \leq k \leq \infty$, and $\phi(s)_k$ is the k-th coordinate of $\phi(s)$.

Proof. For C we take the constant of Proposition 1 and we let $Y = CX$ and $Y_k = CX_k$ so that for every $v \in \ell_2$

$$\|v\| \leq \mathbb{E} \left| \sum_k v_k Y_k \right|. \tag{4}$$

Let $\delta > 0$ be arbitrary. Then, by definition of the Rademacher variable,

$$\begin{aligned}
&2\mathbb{E} \sup_{s \in \mathcal{S}} (\epsilon\psi(s) + f(s)) - \delta \\
&= \sup_{s_1, s_2 \in \mathcal{S}} \psi(s_1) + f(s_1) - \psi(s_2) + f(s_2) - \delta \\
&\leq \psi(s_1^*) - \psi(s_2^*) + f(s_1^*) + f(s_2^*) && (5) \\
&\leq \|\phi(s_1^*) - \phi(s_2^*)\| + f(s_1^*) + f(s_2^*) && (6) \\
&\leq \mathbb{E} \left| \sum_k Y_k (\phi(s_1^*)_k - \phi(s_2^*)_k) \right| + f(s_1^*) + f(s_2^*) && (7) \\
&\leq \mathbb{E} \sup_{s_1, s_2 \in \mathcal{S}} \left| \sum_k Y_k \phi(s_1)_k - \sum_k Y_k \phi(s_2)_k \right| + f(s_1) + f(s_2) && (8) \\
&= \mathbb{E} \sup_{s_1 \in \mathcal{S}} \sum_k Y_k \phi(s_1)_k + f(s_1) + \mathbb{E} \sup_{s_2 \in \mathcal{S}} - \sum_k Y_k \phi(s_2)_k + f(s_2) && (9) \\
&= 2 \left(\mathbb{E} \sup_{s \in \mathcal{S}} \sum_k Y_k \phi(s)_k + f(s) \right). && (10)
\end{aligned}$$

In (5) we pass to approximate maximizers $s_1^*, s_2^* \in \mathcal{S}$, in (6) we use the assumed Lipschitz property relating ψ and ϕ, and in (7) we apply inequality (4). In (8) we use linearity and bound by a supremum in s_1 and s_2. In this expression we can simply drop the absolute value, because for any fixed configuration of the Y_k the maximum will be attained when the difference is positive, since the remaining

expression $f(s_1)+f(s_2)$ is invariant under the exchange of s_1 and s_2. This gives (9). The identity (10) then follows from the symmetry of the variables Y_k. Since $\delta > 0$ was arbitrary, the result follows.

Proof (Proof of Theorem 2). The constant C and the Y_k are chosen as in the previous Lemma. We prove by induction that $\forall\ m \in \{0, ..., n\}$

$$\mathbb{E} \sup_{s\in\mathcal{S}} \sum_i \epsilon_i \psi_i(s) \leq \mathbb{E}\left[\sup_{s\in\mathcal{S}} \sum_{i:1\leq i\leq m} \sum_k Y_{ik}\phi_i(s)_k + \sum_{i:m<i\leq n}^{n} \epsilon_i\psi_i(s)\right].$$

The result then follows for $m = n$. The case $m = 0$ is an obvious identity. Assume the claim to hold for fixed $m-1$, with $m \leq n$. We denote $\mathbb{E}_m = \mathbb{E}[.|\{\epsilon_i, Y_{ik} : i \neq m\}]$ and define $f : \mathcal{S} \to \mathbb{R}$ by

$$f(s) = \sum_{i:1\leq i<m} \sum_k Y_{ik}\phi_i(s)_k + \sum_{i:m<i\leq n}^{n} \epsilon_i\psi_i(s).$$

Then

$$\begin{aligned}
\mathbb{E} \sup_{s\in\mathcal{S}} \sum_i \sigma_i\psi_i(\mathbf{c}) &\leq \mathbb{E}\left[\sup_{s\in\mathcal{S}} \sum_{i:1\leq i<m} \sum_k Y_{ik}\phi_i(s)_k + \sum_{i:m\leq i\leq n}^{n} \epsilon_i\psi_i(s)\right] \\
&= \mathbb{E}\ \mathbb{E}_m \sup_{s\in\mathcal{S}} (\epsilon_m\psi_m(s) + f(s)) \\
&\leq \mathbb{E}\ \mathbb{E}_m \sup_{s\in\mathcal{S}} \sum_k Y_{mk}\phi_m(s)_k + f(s) \\
&= \mathbb{E}\ \sup_{s\in\mathcal{S}} \sum_{i:1\leq i\leq m} \sum_k Y_{ik}\phi_i(s)_k + \sum_{i:m<i\leq n} \epsilon_i\psi_i(s).
\end{aligned}$$

The first inequality is the induction hypothesis, the second is Lemma 2.

6 A negative result

Conjecture 1 can be refuted by a simple counterexample. Let $\mathcal{X} = \ell_2$ with canonical basis (e_i) and set $x_i = e_i$ for $1 \leq i \leq n$. Let $\mathcal{F}$ be the unit ball in the set of bounded operators $\mathcal{B}(\ell_2)$, and for h we take the function $h : x \in \ell_2 \mapsto \|x\|$, which has Lipschitz norm equal to one.

If the conjecture was true then there is a universal constant K such that

$$\mathbb{E} \sup_{T\in\mathcal{B}(H):\|T\|_\infty\leq 1} \sum_i \epsilon_i \|Tx_i\| \leq K\mathbb{E} \sup_{T\in\mathcal{B}(H):\|T\|_\infty\leq 1} \left\|\sum_i \epsilon_i Tx_i\right\|. \tag{11}$$

For any Rademacher sequence $\epsilon = (\epsilon_i)$ we let T_ϵ be the operator defined by $T_\epsilon e_i = e_i$ if $i \leq n$ and $\epsilon_i = 1$, and $T_\epsilon = 0$ in all other cases. Clearly T_ϵ has norm

$\|T_\epsilon\|_\infty \leq 1$ (it is the orthogonal projection to the subspace spanned by the basis vectors e_i such that $\epsilon_i = 1$). Then

$$\frac{n}{2} = \mathbb{E}\left|\{i : \epsilon_i = 1\}\right| = \mathbb{E}\sum_i \epsilon_i \left\|T_\epsilon x_i\right\| \leq \mathbb{E}\sup_{T\in\mathcal{B}(H):\|T\|_\infty\leq 1}\sum_i \epsilon_i \left\|Tx_i\right\|.$$

But on the other hand, the orthonormality of the Rademacher sequence implies that

$$\mathbb{E}\sup_{T\in\mathcal{B}(H):\|T\|_\infty\leq 1}\left\|\sum_i \epsilon_i T x_i\right\| \leq \mathbb{E}\left\|\sum_i \epsilon_i e_i\right\| \leq \sqrt{n}.$$

With (11) we obtain $n/2 \leq K\sqrt{n}$ for some universal constant K, which is absurd.

References

1. P.L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
2. J. Baxter. A Model of Inductive Bias Learning, *Journal of Artificial Intelligence Research* 12:149–198, 2000.
3. Biau, G., Devroye, L., & Lugosi, G. (2008). On the performance of clustering in Hilbert spaces. Information Theory, IEEE Transactions on, 54(2), 781-790.
4. S. Boucheron, G. Lugosi, P. Massart. *Concentration Inequalities*, Oxford University Press, 2013
5. A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.
6. Chapelle, O., & Wu, M. (2010). Gradient descent optimization of smoothed information retrieval metrics. Information retrieval, 13(3), 216-235.
7. Chaudhuri, S., & Tewari, A. Generalization Bounds for Convex Surrogates in Learning to Rank.
8. Ciliberto, C., Poggio, T., & Rosasco, L. (2015). Convex learning of multiple tasks and their structure. arXiv preprint arXiv:1504.03101.
9. Crammer, K., & Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. The Journal of Machine Learning Research, 2, 265-292.
10. S. M. Kakade, S. Shalev-Shwartz, A. Tewari. Regularization Techniques for Learning with Matrices. *Journal of Machine Learning Research* 13:1865–1890, 2012.
11. V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50, 2002.
12. M. Ledoux, M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer, Berlin, 1991.
13. Lei, Y., Dogan, U., Binder, A., & Kloft, M. (2015). Multi-class SVMs: From Tighter Data-Dependent Generalization Bounds to Novel Algorithms. In Advances in Neural Information Processing Systems (pp. 2026-2034).
14. Maurer, A. (2009). Transfer bounds for linear feature learning. Machine learning, 75(3), 327-350.
15. Maurer, A., & Pontil, M. (2010). K-Dimensional Coding Schemes in Hilbert Spaces. Information Theory, IEEE Transactions on, 56(11), 5839-5846.

16. Maurer, A., Pontil, M., & Romera-Paredes, B. (2015). The Benefit of Multitask Representation Learning. arXiv preprint arXiv:1505.06279.
17. McDonald, D. J., Shalizi, C. R., & Schervish, M. (2011). Generalization error bounds for stationary autoregressive models. arXiv preprint arXiv:1103.0942.
18. R. Meir and T. Zhang, "Generalization error bounds for Bayesian mixture algorithms," *JMLR*, vol. 4, pp. 839–860, 2003.
19. C.A. Michelli and M. Pontil, On learning vector–valued functions, *JMLR* 6 (2005), 615–637.
20. Mroueh, Y., Poggio, T., Rosasco, L., & Slotine, J. J. (2012). Multiclass learning with simplex coding. In Advances in Neural Information Processing Systems (pp. 2789-2797).
21. D. Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Tech. J.*, 41: 463–501, 1962.
22. S. Szarek. On the best constants in the Khintchine inequality. Studia Math. 58, 197–208, 1976.