arXiv:1507.03040v3 [stat.ML] 2 Jul 2016

Tight Risk Bounds for Multi-Class Margin Classifiers

Yury Maximov
Predictive Modeling and Optimization Department
Institute of Information Transmission Problems
Moscow, Bolshoy Karenty 19/1, 127051
yurymaximov@iitp.ru

Daria Reshetova
Predictive Modeling and Optimization Laboratory
Moscow Institute of Physics and Technology
Moscow, Kerchenskaya 1a/1, 117303

Predictive Modeling and Optimization Department
Institute of Information Transmission Problems
Moscow, Bolshoy Karenty 19/1, 127051
reshetova@phystech.edu

Abstract

We consider a problem of risk estimation for large-margin multi-class classifiers. We propose a novel risk bound for the multi-class classification problem. The bound involves the marginal distribution of the classifier and the Rademacher complexity of the hypothesis class. We prove that our bound is tight in the number of classes. Finally, we compare our bound with the related ones and provide a simplified version of the bound for the multi-class classification with kernel based hypotheses.

Keywords: statistical learning, multi-class classification, excess risk bound

1 Introduction

The principal goal of the statistical learning theory is to provide a framework for studying the problems of a statistical nature and characterize the performance of learning algorithms in order to facilitate the design of better learning algorithm.

The statistical learning theory of supervised binary classification is by now pretty well developed, while its multi-class extension contains numerous statistical challenges. Multi-class classification problems widely arise in everyday practice in various domains, ranging from ranking to computer vision.

For binary classification problems a quite good distribution-free characterization of risk bounds is given via VC dimension. Tighter data-dependent bounds are known in terms of Rademacher complexity or covering numbers. These bounds correctly describe a finite sample performance of learning algorithms.

Bounding classification risk for multi-class problems is much less straightforward. Recently, finite sample performance of multi-class learning algorithms was given by means of Natarajan dimension (Daniely and Shalev-Shwartz 2014, Daniely, Sabato, Ben-David, and Shalev-Shwartz 2011). An interesting VC-dimension based bound for the risk of large margin mutti-class classifiers is provided in (Guermeur 2007).

These estimates give a quite tight data-independent bound on the risk of multi-class classification methods. On the other hand data-dependent characterization of algorithm quality usually give much better estimates for practical problems.

Rademacher complexity bounds seem to be one of the tightest way to estimate data-dependent finite-sample performance of learning algorithms (Koltchinskii and Panchenko 2002, Bartlett and Mendelson 2003). There is a lot of progress in risk estimation for binary classification problems (Bartlett, Bousquet, and Mendelson 2005, Boucheron, Lugosi, and Massart 2013).

For multi-class learning problems the situation is more delicate. A seminal paper of Koltchinskii & Panchenko (Koltchinskii and Panchenko 2002) provides Rademacher complexity based margin risk bound. The main drawback of this bound is a quadratic dependence on the number of classes, which makes the bound hardly applicable to real-life huge-scale problems of computer vision or text classification. In spite of numerous research there was only a slight improvement of this bound (Mohri, Rostamizadeh, and Talwalkar 2012, Cortes, Mohri, and Rostamizadeh 2013).

Contribution. The main contributions of this paper are

a) a new Rademacher complexity based bound for large-margin multi-class classifiers. The bound is linear in the number of classes which improves quadratic dependence of formerly the best Rademacher complexity bounds (Koltchinskii and Panchenko 2002, Cortes, Mohri, and Rostamizadeh 2013);

b) a new lower bound on the Rademacher complexity of multi-class margin classification methods. This means that sub-linear in the number of classes Rademacher complexity based bound is hardly possible for multi-class margin classifiers in a standard (unconstrained) model. But it is still possible to provide better bounds in terms of their dependence on the number of classes under other models or extra assumptions (Allwein, Schapire, and Singer 2001, Dietterich and Bakiri 1995, Zhang 2004).

Paper structure. The paper consists of four parts. In the second part of the paper, we present the theoretical contribution, namely new Rademacher complexity bounds. It is followed by a discussion of related works and comparison the proposed bound with other multi-class complexity bounds.

2 Multi-class learning guarantees

We consider a standard multi-class classification framework. Let $\mathcal{X}$ be a set of observations and $\mathcal{Y}$, $|\mathcal{Y}| < \infty$ be a set of labels respectively. Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}, P)$ be a probability space and let $\mathcal{F}$ be a class of measurable functions from $(\mathcal{X}, \mathcal{A})$ into $\mathbb{R}$. Let $\{(x_i, y_i)\}$ be a sequence of i.i.d. random variables taking values in $(\mathcal{X} \times \mathcal{Y}, \mathcal{A})$ with common distribution P. We assume that this sequence is defined on a probability space $(\Omega, \Sigma, \mathbb{P})$. Let P_n be the empirical measure associated with the sample $S = \{(x_i, y_i)\}_{i=1}^n$.

We assume that the labels take values in a finite set $\mathcal{Y}$ with $|\mathcal{Y}| = k$. Let $\tilde{\mathcal{F}}$ be a class of functions from S into $\mathbb{R}$. A function $f \in \tilde{\mathcal{F}}$ predicts a label $y \in \mathcal{Y}$ for an example $x \in S$ iff

$$f(x, y) > \max_{y \neq y'} f(x, y') \tag{1}$$

The margin of a labeled example (x, y) is defined as

$$m_f(x, y) := f(x, y) - \max_{y \neq y'} f(x, y'), \tag{2}$$

so f misclassifies the labeled example (x, y) iff $m_f(x, y) \leq 0$.

Let

$$\mathcal{F} := \{f(\cdot, y) : y \in \mathcal{Y}, f(\cdot, y) \in \tilde{\mathcal{F}}_y\}.$$

In a more common situation all scoring function belongs to same class $\tilde{\mathcal{F}}$.

We refer to the empirical Rademacher complexity of the class $\mathcal{F}$ as

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i),$$

where $\varepsilon_1, \ldots, \varepsilon_n$ is independent $\{\pm 1\}$-valued random variables. Then the Rademacher complexity of $\mathcal{F}$ is $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}\widehat{\mathfrak{R}}_n(\mathcal{F})$.

The following theorem states an upper bound for the classification error of k-class classifier. This result improves theorem 11 of (Koltchinskii and Panchenko 2002), theorem 1 of (Cortes, Mohri, and Rostamizadeh 2013) and theorem 8.1 of (Mohri, Rostamizadeh, and Talwalkar 2012) by a factor of k.

Theorem 1. *For all $t > 0$,*

$$\mathbb{P}\Bigg\{\exists f \in \tilde{\mathcal{F}} : P\{m_f \le 0\} > \inf_{\delta\in(0,1]}\left[P_n\{m_f \le \delta\} + \frac{4k}{\delta}\mathfrak{R}_n(\mathcal{F}) + \left(\frac{\log\log_2(2/\delta)}{n}\right)^{1/2} + \frac{t}{\sqrt{n}}\right]\Bigg\} \le 2\exp(-2t^2)$$

Later we show that theorem 1 give a tight bound on the multi-class complexity.

Let $\mathcal{M}_k(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ be a class of functions such that

$$\mathcal{M}_k(\mathcal{F}_1, \ldots, \mathcal{F}_k) = \{\forall m \in \mathcal{M}_k : m(x,y) = f(x,y) - \max_{y\neq y'} f(x,y'), f(x,y) \in \mathcal{F}_y\}. \tag{3}$$

Prior to the proof of the theorem one needs to proof the following lemma.

Lemma 1. *Let $\mathcal{M}_k(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ be a class of margin functions over $\mathcal{F}_1, \ldots, \mathcal{F}_k$ defined in* **??**. *Then for any i.i.d. sample $S_n = \{(x_i, y_i)\}_{i=1}^n$ of size n holds*

$$\widehat{\mathfrak{R}}_n(\mathcal{M}_k(\mathcal{F}_1, \ldots, \mathcal{F}_k)) \le \sum_{j=1}^{k} \widehat{\mathfrak{R}}_n(\mathcal{F}_j).$$

Proof. We provide a proof of the lemma in the case $\mathcal{F} \doteq \mathcal{F}_1 = \cdots = \mathcal{F}_k$. It can be easily extended into a more general case. For a single class $\mathcal{F}$ the class of margin functions $\mathcal{M}_k(\mathcal{F})$ has a form

$$\mathcal{M}_k(\mathcal{F}) \doteq \{\forall m \in \mathcal{M}_k : m(x,y) = f(x,y) - \max_{y\neq y'} f(x,y')\}.$$

Let $m_f(x,y)^{(\mathcal{Y}'|\mathcal{Y})}$ be a partial margin of the object (x,y) taken with respect to the subset $\mathcal{Y}'$ of the set of classes, $\mathcal{Y}' \subseteq \mathcal{Y}$:

$$m_f(x,y)^{(\mathcal{Y}'|\mathcal{Y})} \doteq \begin{cases} f(x,y) - \max\limits_{\substack{y'\in\mathcal{Y}' \\ y'\neq y}} f(x,y'), & \text{if } y \in \mathcal{Y}' \\ -\max\limits_{y'\in\mathcal{Y}'} f(x,y'), & \text{if } y \notin \mathcal{Y} \end{cases}$$

Let $\mathcal{M}_k^{\mathcal{Y}'}(\mathcal{F}) \doteq \{\forall m \in \mathcal{M}_k^{\mathcal{Y}'}(\mathcal{F}) : m = m_f^{(k|\mathcal{Y}')}(x_i, y_i), f \in \mathcal{F}\}$.

The proof is by induction on the size of $\mathcal{Y}'$. Note that $\mathcal{M}_k^{\mathcal{Y}}(\mathcal{F}) = \mathcal{M}_k(\mathcal{F})$ and

$$\mathcal{M}_k^{\{1\}}(x,y) = \begin{cases} f(x,y), & \text{if } y = 1 \\ -f(x,y), & \text{if } y \neq 1 \end{cases}.$$

Denote by $\delta(y, y')$ the indicator of $y = y'$

$$\delta(y, y') = \begin{cases} 1, & \text{if } y = y' \\ 0, & \text{if } y \neq y' \end{cases}$$

Then for $\mathcal{Y}' = \{y\}$ holds

$$\widehat{\mathfrak{R}}_n(\mathcal{M}_k^{\mathcal{Y}'}(\mathcal{F})) = \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i (2\delta(y_i, y) - 1) f(x_i) = \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) = \widehat{\mathfrak{R}}_n(\mathcal{F}),$$

because a binary sequence $\delta(y_i, y)$ is independent of the class of functions $\mathcal{F}$ and the Rademacher variables $\{\varepsilon_i\}_{i=1}^n$. Therefore, the induction base is proved.

The induction hypothesis is that for any $\mathcal{Y}' \subset \mathcal{Y}$, $|\mathcal{Y}'| \leq t$ the Rademacher complexity of $\mathcal{M}_k^{\mathcal{Y}'}$ satisfies

$$\widehat{\mathfrak{R}}_n(\mathcal{M}_k^{\mathcal{Y}'}(\mathcal{F})) \leq |\mathcal{Y}'| \widehat{\mathfrak{R}}_n(\mathcal{F}). \tag{4}$$

If $\mathcal{Y}' = \mathcal{Y}$ the statement is proved, otherwise the set $\mathcal{Y} \setminus \mathcal{Y}'$ is not empty. Then for any $\tilde{y} \in \mathcal{Y} \setminus \mathcal{Y}'$ and i.i.d. sample $S = \{(x_i, y_i)\}_{i=1}^n$ holds

$$\widehat{\mathfrak{R}}_n(\mathcal{M}_k^{\mathcal{Y}' \cup \tilde{y}}(\mathcal{F})) = \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \Bigg\{ \sum_{\substack{(x_i, y_i) \in S \\ y_i = \tilde{y}}} \varepsilon_i \{ f(x_i, y_i) - \max_{\substack{y \in \mathcal{Y}' \\ y \neq y_i}} f(x_i, y) \} - \sum_{\substack{(x_i, y_i) \in S \\ y_i \neq \tilde{y}}} \varepsilon_i \max\{ f(x_i, \tilde{y}), \max_{y \in \mathcal{Y}'} f(x_i, y) \} \Bigg\}$$

Note that $\max\{f_1, f_2\} = \frac{f_1 + f_2}{2} + \frac{|f_1 - f_2|}{2}$.

Then

$$\begin{aligned}
\widehat{\mathfrak{R}}_n(\mathcal{M}_k^{\mathcal{Y}' \cup \tilde{y}}(\mathcal{F})) = \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \Bigg\{ & \sum_{\substack{(x_i, y_i) \in S \\ y_i = \tilde{y}}} \varepsilon_i (f(x_i, y_i) - \max_{\substack{y \in \mathcal{Y}' \\ y \neq y_i}} f(x_i, y)) - \\
& \frac{1}{2} \sum_{\substack{(x_i, y_i) \in S \\ y_i \neq \tilde{y}}} \varepsilon_i \left\{ f(x_i, \tilde{y}) + \max_{y \in \mathcal{Y}'} f(x_i, y)) - \left| f(x_i, \tilde{y}) - \max_{y \in \mathcal{Y}'} f(x_i, y) \right| \right\} \Bigg\} \leq \\
& \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^n \varepsilon_i (2\delta(y_i, \tilde{y}) - 1) f(x_i, \tilde{y}) + \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^n \varepsilon_i (1 - 2\delta(y_i, \tilde{y})) \max_{y \in \mathcal{Y}'} f(x_i, y) + \\
& \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^n \varepsilon_i \left\{ \delta(y_i, \tilde{y}) (f(x_i, \tilde{y}) - \max_{y \in \mathcal{Y}'} f(x_i, y)) + (1 - \delta(y_i, \tilde{y})) \left| f(x_i, \tilde{y}) - \max_{y \in \mathcal{Y}'} f(x_i, y) \right| \right\} = \\
& \frac{\widehat{\mathfrak{R}}_n(\mathcal{F})}{2} + \frac{\widehat{\mathfrak{R}}_n(\mathcal{M}_k^{\mathcal{Y}'}(\mathcal{F}))}{2} + \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^n \varepsilon_i \Bigg\{ \delta(y_i, \tilde{y}) (f(x_i, \tilde{y}) - \max_{y \in \mathcal{Y}'} f(x_i, y)) \\
& \qquad + (1 - \delta(y_i, \tilde{y})) \left| f(x_i, \tilde{y}) - \max_{y \in \mathcal{Y}'} f(x_i, y) \right| \Bigg\}
\end{aligned}$$

Note, that $x + y \to x + |y|$ is a 1-Lipschitz. Thus by Talagrand's contraction inequality (see theorem 4.12, p. 112–114 of (Ledoux and Talagrand 1991) and more appropriate lemma 4.2, p. 78–79 of

(Mohri, Rostamizadeh, and Talwalkar 2012)) holds

$$\widehat{\mathfrak{R}}_n(\mathcal{M}^{\mathcal{Y}'\cup\tilde{y}|\mathcal{Y}}) \le \frac{\widehat{\mathfrak{R}}_n(\mathcal{F})}{2} + \frac{\widehat{\mathfrak{R}}_n(\mathcal{M}^{\mathcal{Y}'})}{2} + \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \frac{1}{2n}\sum_{i=1}^n \varepsilon_i(2\delta(y_i,\tilde{y})-1)(f(x_i,m) - \max_{y\in\mathcal{Y}'} f(x_i,y)) \le$$

$$\frac{\widehat{\mathfrak{R}}_n(\mathcal{F})}{2} + \frac{\widehat{\mathfrak{R}}_n(\mathcal{M}^{\mathcal{Y}'})}{2} + \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \frac{1}{2n}\sum_{i=1}^n \varepsilon_i(2\delta(y_i,\tilde{y})-1) f(x_i,m) + \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \frac{1}{2n}\sum_{i=1}^n \varepsilon_i(1-2\delta(y_i,\tilde{y})) \max_{y\in\mathcal{Y}'} f(x_i,y) =$$

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) + \widehat{\mathfrak{R}}_n(\mathcal{M}_k^{\mathcal{Y}'}(\mathcal{F})) \le (|\mathcal{Y}'|+1)\widehat{\mathfrak{R}}_n(\mathcal{F}),$$

where the last but one inequality holds by the inductive hypothesis, ineq. **??**). This completes the inductive proof. □

Proof of the theorem 1. Following to (Koltchinskii and Panchenko 2002) consider 2 sequences $\{\delta_j\}_{j\ge1}$ and $\{\varepsilon_j\}_{j\ge1}$, $\varepsilon_j \in (0,1)$.

The standard Rademacher complexity margin bound (theorem 4.4, p. 81–82 of (Mohri, Rostamizadeh, and Talwalkar 2012)) gives for any fixed δ_t and ε_t:

$$\mathbb{P}\left\{P(m_f(x,y)<0) - P_n(m_f(x,y)<\delta_t) \ge \frac{2}{\delta_j}\mathfrak{R}(\mathcal{M}_t(\mathcal{F})) + \varepsilon_j\right\} \le \exp(-2n\varepsilon_j^2).$$

Then by choosing $\varepsilon_j = \frac{t}{\sqrt{n}} + \sqrt{\frac{\log j}{n}}$ and applying the union bound

$$\mathbb{P}\left\{\exists\, j : P(m_f(x,y)<0) - P_n(m_f(x,y)<\delta_j) \ge \frac{2}{\delta_j}\mathfrak{R}(\mathcal{M}_k(\mathcal{F})) + \varepsilon_j\right\}$$

$$\le \sum_{j\ge1}\exp(-2n\varepsilon_j^2) \le \exp(-2t^2)\sum_{j\ge1}\exp(-2\log j) = \frac{\pi^2}{6}\exp(-2nt^2) < 2\exp(-2nt^2).$$

We choose $\delta_k = 1/2^k$, then $2/\delta_j \le 4/\delta$. By lemma 1 we have $\mathfrak{R}(\mathcal{M}_k(\mathcal{F})) \le k\mathfrak{R}(\mathcal{F})$ which proofs the theorem. □

Below we present a Rademacher complexity bounds for multi-class kernel learning in a simplified form. Let $\mathfrak{K} : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a positive definite symmetric kernel and $\Phi : \mathcal{X}\to\mathbb{H}$ be a feature mapping associated to $\mathfrak{K}$. In the multi-class setting a family of kernel-based hypotheses $\mathcal{H}_{k,p}$ is defined for any $p \ge 1$ as

$$\mathcal{H}_{\mathfrak{K},p} = \{(x,y)\in\mathcal{X}\times\mathcal{Y}\to w_y\cdot\Phi(x) : W = (w_1,\ldots,w_k)^{\mathrm{T}}, \|W\|_{\mathbb{H},p} \le \Lambda\},$$

where $\|W\|_{\mathbb{H}}^p = (\sum_{i=1}^k \|w_i\|_{\mathbb{H}}^p)^{1/p}$. The labels are assigned according to $\arg\max_{y\in\mathcal{Y}}\langle w_y, \Phi(x)\rangle$.

The following bound is a corollary of the theorem 1.

Theorem 2. *Let $\mathfrak{K} : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a positive definite symmetric kernel and let $\Phi : \mathcal{X}\to\mathbb{H}$ be the associated feature mapping function. Assume that there exists $R > 0$ such that $\mathfrak{K}(x,x) \le R^2$ for all $x\in\mathcal{X}$. Then, for any $t > 0$ the following multi-class classification generalization bounds hold for all hypotheses $h\in\mathbb{H}_{K,p}$*

$$\mathbb{P}\left\{\exists f\in\tilde{\mathcal{F}} : P\{m_f\le0\} > P_n\{m_f\le\delta\} + \frac{2k}{\delta}\sqrt{\frac{R^2\Lambda^2}{n}} + \frac{t}{\sqrt{n}}\right\} \le \exp(-2t^2)$$

Below we proof that the bound on the Rademacher complexity of the class $\mathfrak{R}_n(\mathcal{M}_k(\mathcal{F}_1,\dots,\mathcal{F}_k)$ is tight. Let $\mathcal{F}_t^j = \{f : \mathbb{R} \to [-1;+1]\}$ be a class of functions such that

$$\mathcal{F}_t^j \ni f(x) = \begin{cases} -1, & \text{if } x \notin [j;j+1] \\ +1 \text{ or } -1, & \text{if } x \in [j;j+1] \end{cases}$$

and moreover each $f \in \mathcal{F}_t^j$ has in $(j,j+1)$ no more than t discontinuity points. We refer to $\mathcal{F}_0$ as the class of functions takes -1 over real line.

Denote

$$\mathcal{F}_t^* = \left\{\max\{f_1,f_2,\dots,f_k\}, f_i \in \mathcal{F}_t^j\right\} \text{ and } \mathcal{F}_t = \bigcup_{j=1}^{m} \mathcal{F}_t^j.$$

Note, that all the classes $\{\mathcal{F}_t^j\}_{j=1}^k$, $\mathcal{F}_t$ and $\{\mathcal{F}_t^*\}$ for a fixed t satisfy the conditions of the central limit theorem.

Let $\mathfrak{R}_n^*(\mathcal{F}_t^j)$ be a Rademacher complexity of $\mathcal{F}_t^j$ defined with respect to the interval $(j,j+1)$ only

$$\mathfrak{R}_n^*(\mathcal{F}_t^j) = \sup_{f\in\mathcal{F}_t^j} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i) \mathbb{1}_{x_i\in(j,j+1)}.$$

Lemma 2. *Let $P^{\mathcal{X}}$ be a uniform distribution over the domain $\mathcal{X} = [1;k+1]$. Then for any $C > 0$ there exists $t = t(C,k)$ such that for any sample $S_n = \{x_i\}_{i=1}^n$ of size n drawn i.i.d. from $P^{\mathcal{X}}$ and any j, $1 \le j \le k$ holds*

$$\mathfrak{R}_n^*(\mathcal{F}_t^j) \ge C\,\mathfrak{R}_n(\mathcal{F}_0)$$

since $n \ge n_0$, $n_0 = n_0(t)$.

Proof. By theorem 5.3.3. of (Talagrand 2014) for any sequences $t_1,\dots,t_m$ in ℓ^2 such that

$$\ell \ne \ell' \Rightarrow \|t_\ell - t_{\ell'}\| \ge a$$

and

$$\forall \ell \le m \Rightarrow \|t_\ell\|_\infty \le b$$

the following lower bound for Rademacher process holds

$$\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(x_i)\varepsilon_i \ge \frac{1}{L}\min\left\{a\sqrt{\log m}, \frac{a^2}{b}\right\}, \tag{5}$$

for some absolute constant L.

By the standard chaining argument the Rademacher complexity of the class $\mathcal{F}_0$ satisfies

$$\mathfrak{R}_n(\mathcal{F}_0) \le \frac{C_0}{\sqrt{n}},$$

for some absolute constant $C_0 = C_0(\mathcal{F}) > 0$ independent of n.

Let objects $x^1,\dots,x^{n_j}$ belong to $(j,j+1)$ are ordered in such a way that $(x^i - x^j)(i-j) \ge 0$ for all i,j. Note that for any such sequence there exist functions $\{f_1,\dots,f_{2^{\lfloor n_j/t\rfloor}}\} \in \mathcal{F}_{t+1}$ such that the function f_j assigns $+1$ to objects $\{x^{st+1},\dots,x^{st+t}\}$, $s : 1 \le s \le \lfloor n_j/t\rfloor$ iff a binary representation of j contains 1 in s-th digit from the right. Otherwise it assigns to -1 to $\{x^{st+1},\dots,x^{st+t}\}$.

Then by the equation **??** the following lower bound on Rademacher complexity of the class $\widehat{\mathcal{F}}_j$, $\widehat{\mathcal{F}}_j = \{f_1,\dots,f_{2^{\lfloor n_j/t\rfloor}}\}$ takes place

$$\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i)\mathbb{1}_{x\in(j,j+1)} \ge \frac{1}{L}\min\left\{\frac{n_j}{n}\sqrt{2-\frac{2t}{n_j}}, \frac{2t\sqrt{n_j}}{n}\right\}, \quad f \in \widehat{\mathcal{F}}_j$$

for some absolute constant L stated by the inequality **??**.

Remind that the median for Binomial distribution with parameter $1/k$ is one of the integers $\{\lfloor n/k \rfloor - 1, \lfloor n/k \rfloor, \lfloor n/k \rfloor + 1\}$. Then the number of objects in $(j, j+1)$ is $n/k - 2$ or more with probability at least $1/2$.

Therefore, if $n \geq 16kt^2$, $t \geq 1$

$$\mathfrak{R}_n(\widehat{\mathcal{F}}_j) = \mathbb{E}\,\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(x_i)\varepsilon_i \geq \frac{1}{L} \min\left\{ \frac{1}{2k}\sqrt{2 - \frac{4tk}{n}}, \frac{2t}{\sqrt{nk}} \right\} \geq$$

$$\min\left\{ \frac{1}{2kL}, \frac{2t}{L\sqrt{nk}} \right\} = \frac{2t}{L\sqrt{nk}}.$$

Then it is sufficient to choose $t \geq C_0 CL\sqrt{k}/2$ and $n \geq 16kt^2$ as above to satisfy the conditions of the lemma. □

Theorem 3. *Let $P^{\mathcal{X}}$ be a uniform distribution over the domain $\mathcal{X} = [1; k+1]$ and $P^{\mathcal{Y}}$ concentrated on a single class $k+1$. Then for any sample $S_n = \{(x_i, y_i)\}_{i=1}^n$ of size n drawn i.i.d. from $P^{\mathcal{X}} \times P^{\mathcal{Y}}$ and any $\varepsilon > 0$ for the Rademacher complexity of the margin class $\mathcal{M}_{k+1} = (\mathcal{F}_t^1, \ldots, \mathcal{F}_t^{k+1}, \mathcal{F}_0)$ holds*

$$\mathfrak{R}_n(\mathcal{M}_{k+1}) \geq (1 - \varepsilon) \sum_{j=1}^{k} \mathfrak{R}_n(\mathcal{F}_t^j)$$

for some large enough $t = t(\varepsilon, k)$ independent of n and all $n \geq n_0$, $n_0 = n_0(t)$.

Proof. By the symmetry under negation of classes $\mathcal{F}_t^j$ in $(j, j+1)$ and definition of the class $\mathcal{F}_t^*$ we have

$$\mathfrak{R}_n(\mathcal{F}_t^*) = \sum_{j=1}^{n} \mathfrak{R}^*(\mathcal{F}_t^*) \geq \left(1 - \frac{1}{C}\right) \sum_{j=1}^{k} \mathfrak{R}_n(\mathcal{F}_t^j) = k\left(1 - \frac{1}{C}\right) \mathfrak{R}_n(\mathcal{F}_t^j), \quad j' : 1 \leq j \leq k$$

where $C = 1/\varepsilon$ is defined in accordance with the lemma 2.

Note that the Rademacher complexity of $\mathcal{M}_{k+1}(\mathcal{F}_t^1, \ldots, \mathcal{F}_t^k, \mathcal{F}_0)$ is at least the same as the Rademacher complexity of $\mathcal{F}_t^*$ by the construction of $\mathcal{M}_{k+1}$ and $\mathcal{F}_t^1, \ldots, \mathcal{F}_t^k$. This proofs the lemma. □

A similar bound holds for the Rademacher complexity of the classes $\mathcal{F}_t$ and $\mathcal{M}_{k+1}(\mathcal{F}_t)$ respectively. Note, that this bound is effectively the lower bound to the estimate of the theorem 1 in the sense that the bound there can not improved based on the Rademacher complexity estimates only if one put no assumptions on the behavior of the function class (e.g. small covering number bound or small VC dimension).

3 Related works and discussion

A number of works are devoted to bounding the risk of multi-class classification methods. One popular approach to solving a problem with multiple classes is to reduce it to a sequence of binary classification problems. In terms of risk dependence on the number of classes a great breakthrough was done with the design of error–correcting output codes (ECOC) for multi-class classification (Dietterich and Bakiri 1995, Allwein, Schapire, and Singer 2001, Beygelzimer, Langford, and Ravikumar 2009).

In spite of some very promising results concerning ECOC Rifkin & Klautau argued in (Rifkin and Klautau 2004) that the classical approaches, such as one-vs-all classification, is at least as preferable as error-correcting codes from the practical point of view.

Another approach is to define a score function on the point-label pairs and choose a label with the highest score (one-vs-all classification method can be considered from this point of view as well). It is natural to characterize the risk bounds of these methods in terms of classification margin δ equals to the gap between the highest score and the second highest score (see def. **??** for details).

Multi-class SVM extension. Among the methods that share scoring-based paradigm, one should mention the Weston & Watkins multi-class extension of SVM (Weston and Watkins 1998). An improved version multi-class SVM as well as the improved margin risk bound of the order $\tilde{O}(k^2/n\delta^2)$ were presented by Crammer & Singer in (Crammer and Singer 2002b, Crammer and Singer 2002a).

Rademacher complexity bounds. Currently Rademacher complexity as well as combinatorial dimension estimates seem to be among of the most powerful tools to get strong enough risk bounds for multi-class classification. The important property of Rademacher complexity based bounds is that the bounds are applicable in arbitrary Banach spaces and do not depend on the dimension of the feature space directly.

Koltchinksii & Panchenko introduced a margin-based bound for multi-class classification in terms of Rademacher complexities (Koltchinskii and Panchenko 2002, Koltchinskii, Panchenko, and Lozano 2001). The bound was slightly improved (by a constant factor prior to the Rademacher complexity term) in a series of subsequent works (Mohri, Rostamizadeh, and Talwalkar 2012, Cortes, Mohri, and Rostamizadeh 2013).

The main drawback of these state-of-the-art bounds for multi-class classification is a quadratic dependence on the number of classes which makes the bounds unreliable for practical problems with a considerable number of classes.

The principal contribution of this paper is a new Rademacher complexity based upper bound with a linear complexity w.r.t. the number of classes. Moreover we provide the lower bound on Rademacher complexity of margin-based multi-class algorithms. Up to a constant factor it matches to the upper bound. Than means that the bound can not be improved without further assumptions.

Covering number based bounds. Zhang in (Zhang 2004, Zhang 2002) studied covering number bounds for the risk of the multi-class margin classification. Based on the ℓ^∞ covering number bound estimate for the Rademacher complexity of kernel learning problem he obtained asymptotically better rates in the number of classes k (see tab. 1) than those proposed in our paper.

Note, that Zhang's analysis is based on some extra assumptions (not really too restrictive) about underlying hypothesis class and the loss function used. We suppose that the results of (Zhang 2004) are appreciated from the theoretical point of view but still quite limited for practice. This is due to high overestimate (from a practical perspective) of the Rademacher complexity of the hypothesis class by a ℓ^∞ covering number based bound. It should also be noted that Zhang's bound are valid only for learning kernel-based hypothesis and have some extra poly-logarithmic dependence on the number of labeled examples.

Related results for metric spaces with low doubling dimension were obtained by Kontorovich (Kontorovich and Weiss 2014), who used nearest neighbors method to improve the dependence on the number of classes in favor of (doubling) dimension dependence. We should note as well that his approach allows to speed-up multi-class learning algorithms.

We gather margin based bounds applicable for learning functions in Hilbert space the tab. 1.

Combinatorial dimension bounds. Natarajan dimension was introduced in (Natarajan 1989) in order to characterize multi-class PAC learnability. It exactly matches the notion of Vapnik-Chervonenkis dimension in the case of two classes. A number of results concerning risk bounds in terms of Natarajan

Upper bound, $\tilde{O}(\cdot)$	Paper
$\frac{k^2}{\delta\sqrt{n}}$	Koltchinskii & Panchenko, (Koltchinskii and Panchenko 2002)
$\frac{k}{\delta^2\sqrt{n}}$	Cortes et al., (Cortes, Mohri, and Rostamizadeh 2013), Mohri et al. (Mohri, Rostamizadeh, and Talwalkar 2012) Guermeur, (Guermeur 2010)
$\frac{1}{\delta}\sqrt{\frac{k}{n}}$	Zhang, (Zhang 2004)
$\frac{k^2}{\delta^2 n}$	Crammer & Singer, (Crammer and Singer 2002b)
$\frac{k}{\delta\sqrt{n}}$	this paper

Table 1: Dimension-free margin-based bounds for multi-class classification.

dimension were proved in (Daniely, Sabato, Ben-David, and Shalev-Shwartz 2011, Daniely and Shalev-Shwartz 2014, Ben-David, Cesabianchi, Haussler, and Long 1995, Daniely, Sabato, and Shalev-Shwartz 2012). A closely related but more powerful notion of graph dimension was introduced in (Daniely, Sabato, Ben-David, and Shalev-Shwartz 2011, Daniely and Shalev-Shwartz 2014). VC-dimension based bounds for multi-class learning problems were obtained in (Allwein, Schapire, and Singer 2001).

Natarajan and graph dimensions are very useful tools for obtaining multi-class classification risk bounds. The main drawback of these bounds is that they are data-independent. In this sense, we believe that the bounds proposed in this paper are much stronger than the Natarajan/graph dimension bounds same as that of Rademacher complexity bounds are stronger than the VC dimension bounds for binary classification.

We also note that VC dimension bounds as well as Natarajan dimension bounds are usually dimension dependent (Daniely and Shalev-Shwartz 2014), which makes them hardly applicable for practical huge scale problems (such as typical computer vision problems).

Guermeur in (Guermeur 2007, Guermeur 2010) gave a bound for scale-sensitive analog of Natarajan dimension $\tilde{d}_{Nat}$. In Hilbert space for a class of linear functions it can be bounded in terms of the margin as $\tilde{O}(k^2/\delta^2)$ which leads to the risk decay rate of the order $\tilde{O}(k/\delta^2\sqrt{n})$ (see tab. 1).

We gather the bounds above in the tab. 2. Note, that the bound of the order $\tilde{O}(d_{Nat}/n)$ is valid in a separable case only. A clear comparison between various multi-class classification methods is provided

Upper bound, $\tilde{O}(\cdot)$	Paper
$\frac{\log k}{\delta}\sqrt{\frac{d_{VC}}{n}}$	Allwein et al., (Allwein, Schapire, and Singer 2001)
$\frac{\log k}{\delta}\sqrt{\frac{\tilde{d}_{Nat}}{n}}$	Guermeur, (Guermeur 2010)
$\frac{d_{Nat}}{n}$	Daniely et al., (Daniely and Shalev-Shwartz 2014)

Table 2: Combinatorial dimension based upper bounds for multi-class classification.

in (Daniely, Sabato, and Shalev-Shwartz 2012). Lower bounds on Natarajan dimension and sample complexity of multi-class classification methods provided in (Daniely, Sabato, Ben-David, and Shalev-Shwartz 2011, Daniely and Shalev-Shwartz 2014). It was shown in (Daniely, Sabato, Ben-David, and Shalev-Shwartz 2011, Daniely and Shalev-Shwartz 2014) that for multi-class linear classifiers the bounds on Natarajan dimension can be as poor as $\Omega(dk)$, where d is a feature space dimension and k are a number of classes. In this work we provide a linear (in the number of classes) lower bound on the Rademacher complexity of the multi-class margin class of functions (see th. 3 for details).

A preliminary version of the upper bounds (theorem 1) with slightly poor dependence on k was presented by the first author in context of semi-supervised multi-class classification on the workshop "Frontiers of High Dimensional Statistics, Optimization, and Econometrics" in February 2015. The risk

bounds stated in this paper were presented in the final form on March 25-th at the main seminar of Institute for Information Transmission Problems (IITP RAS). In July 2015 the authors were notified be their colleagues that similar results were proposed independently by Kuznetsov et al. and presented on ICML Workshop on Extreme Classification.[1] and in (Kuznetsov, Mohri and Syed 2014). Still we suppose that the bounds presented in this paper are much stronger than the ones presented by Kuznetsov et al. in the sense that we prove explicit lower bounds as well. This shows that the bound which we proved in theorem 1 is tight, i.e. linear dependence on the number of classes is inevitable if no further assumptions are made.

4 Conclusion.

In this paper we propose new state-of-the-art Rademacher complexity based upper bounds for the risk of multi-class margin classifiers. The bound depends linearly in in the number of classes. We prove as well that the bound can not be further improved based on the Rademacher complexities only. Still it is possible to provide a better estimates for the excess risk of multi-class classification using other techniques or supplementary assumptions.

5 Acnowledgement.

We are grateful to Massih-Reza Amini and Zaid Harchaoui for the problem setting and useful suggestions. We would also like to thank Anatoli Juditsky, Grigorii Kabatianski, Vladimir Koltchinskii, Axel Munk, Arkadi Nemirovski and Vladimir Spokoiny for helpful discussions.

References

Allwein, E. L., R. E. Schapire, and Y. Singer (2001). Reducing multiclass to binary: A unifying approach for margin classifiers. *The Journal of Machine Learning Research 1*, 113–141.

Bartlett, P. L., O. Bousquet, and S. Mendelson (2005). Local rademacher complexities. *Annals of Statistics*, 1497–1537.

Bartlett, P. L. and S. Mendelson (2003). Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research 3*, 463–482.

Ben-David, S., N. Cesabianchi, D. Haussler, and P. M. Long (1995). Characterizations of learnability for classes of {0,..., n}-valued functions. *Journal of Computer and System Sciences 50*(1), 74–86.

Beygelzimer, A., J. Langford, and P. Ravikumar (2009). Error-correcting tournaments. In *Algorithmic Learning Theory*, pp. 247–262. Springer.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Cortes, C., M. Mohri, and A. Rostamizadeh (2013). Multi-class classification with maximum margin multiple kernel. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 46–54.

[1] Vitaly Kuznetsov, Mehryar Mohri and Umar Syed. Rademacher complexity margin bounds for learning with a large number of classes. In *In ICML 2015 Workshop on Extreme Classification. Lille, France, July 2015*

Crammer, K. and Y. Singer (2002a). On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research 2*, 265–292.

Crammer, K. and Y. Singer (2002b). On the learnability and design of output codes for multiclass problems. *Machine Learning 47*(2-3), 201–233.

Daniely, A., S. Sabato, S. Ben-David, and S. Shalev-Shwartz (2011). Multiclass learnability and the erm principle. *JMLR - Proceedings Track 19*, 207–232.

Daniely, A., S. Sabato, and S. Shalev-Shwartz (2012). Multiclass learning approaches: A theoretical comparison with implications. In *Advances in Neural Information Processing Systems*, pp. 485–493.

Daniely, A. and S. Shalev-Shwartz (2014). Optimal learners for multiclass problems. In *Proceedings of The 27th Conference on Learning Theory*, pp. 287–316.

Dietterich, T. and G. Bakiri (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 263–286.

Guermeur, Y. (2007). Vc theory of large margin multi-category classifiers. *The Journal of Machine Learning Research 8*, 2551–2594.

Guermeur, Y. (2010). Ensemble methods of appropriate capacity for multi-class support vector machines. *SMTDA 10*, 311–318.

Koltchinskii, V. and D. Panchenko (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 1–50.

Koltchinskii, V., D. Panchenko, and F. Lozano (2001). Some new bounds on the generalization error of combined classifiers. In *Advances in Neural Information Processing Systems*, pp. 245–251.

Kontorovich, A. and R. Weiss (2014). Maximum margin multiclass nearest neighbors. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 892–900.

Kuznetsov, V., M. Mohri, and U. Syed (2014). Multi-class deep boosting. In *Advances in Neural Information Processing Systems*, pp. 2501–2509.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*, Volume 23. Springer Science & Business Media.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012). *Foundations of machine learning*. MIT press.

Natarajan, B. K. (1989). On learning sets and functions. *Machine Learning 4*(1), 67–97.

Rifkin, R. and A. Klautau (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research 5*, 101–141.

Talagrand, M. (2014). *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*, Volume 60. Springer Science & Business Media.

Weston, J. and C. Watkins (1998). Multi-class support vector machines.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research 2*, 527–550.

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *The Journal of Machine Learning Research 5*, 1225–1251.