Journal of Machine Learning Research 6 (2005) 307–340 Submitted 6/04; Revised 12/04; Published 3/05

Generalization Bounds and Complexities Based on Sparsity and Clustering for Convex Combinations of Functions from Random Classes

Savina Andonova Jaeger JAEGER@HCP.MED.HARVARD.EDU
Harvard Medical School
Harvard University
Boston, MA 02115, USA

Editor: John Shawe-Taylor

Abstract

A unified approach is taken for deriving new generalization data dependent bounds for several classes of algorithms explored in the existing literature by different approaches. This unified approach is based on an extension of Vapnik's inequality for VC classes of sets to random classes of sets - that is, classes depending on the random data, invariant under permutation of the data and possessing the increasing property. Generalization bounds are derived for convex combinations of functions from random classes with certain properties. Algorithms, such as SVMs (support vector machines), boosting with decision stumps, radial basis function networks, some hierarchies of kernel machines or convex combinations of indicator functions over sets with finite VC dimension, generate classifier functions that fall into the above category. We also explore the individual complexities of the classifiers, such as sparsity of weights and weighted variance over clusters from the convex combination introduced by Koltchinskii and Panchenko (2004), and show sparsity-type and cluster-variance-type generalization bounds for random classes.

Keywords: complexities of classifiers, generalization bounds, SVM, voting classifiers, random classes

1. Introduction

Statistical learning theory explores ways of estimating functional dependency from a given collection of data. It, also referred to as the theory of finite samples, does not rely on a priori knowledge about a problem to be solved. Note that "to control the generalization in the framework of this paradigm, one has to take into account two factors, namely, the quality of approximation of given data by the chosen function and the capacity of the subset of functions from which the approximating function was chosen" (Vapnik, 1998). Typical measures of the capacity of sets of functions are entropy measures, VC-dimensions and V-γ dimensions. Generalization inequalities such as Vapnik's inequalities for VC-classes, which assert the generalization performance of learners from *fixed* class of functions and take into account the quality of approximation of given data by the chosen function and the capacity of the class of functions, were proven to be useful in building successful learning algorithms such as SVMs (Vapnik, 1998).

An extension of Vapnik's inequality, for VC classes of sets (Vapnik, 1998; Anthony and Shawe-Taylor, 1993) and VC-major classes of functions to classes of functions satisfying Dudley's uniform entropy conditions, was shown by Panchenko (2002). A class of functions $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$

satisfies Dudley's uniform entropy condition if

$$\int_0^\infty \log^{1/2} D(\mathcal{F},u)du < \infty,$$

where $D(\mathcal{F},u)$ denotes Koltchinksii packing numbers defined for example by Dudley (1999) or Panchenko (2002). Applications of the inequality were shown in several papers (Koltchinskii and Panchenko, 2002; Koltchinskii et al., 2003a; Koltchinskii and Panchenko, 2004) which explored the generalization ability of ensemble classification methods, that is, learning algorithms that combine several classifiers into new voting classifiers with better performance. "The study of the convex hull, conv$(\mathcal{H})$, of a given base function class $\mathcal{H}$ has become an important object of study in machine learning literature" (Koltchinskii and Panchenko, 2004). New measures of individual complexities of voting classifiers derived in related work (Koltchinskii et al., 2003a; Koltchinskii and Panchenko, 2004; Koltchinskii et al., 2003b) were shown theoretically and experimentally to play an important role in the generalization performance of the classifiers from conv$(\mathcal{H})$ of a given base function class $\mathcal{H}$. In order to do so, the base class $\mathcal{H}$ is assumed to have Koltchinskii packing numbers satisfying the following condition

$$D(\mathcal{H},u) \leq K(V)u^{-V},$$

for some $V > 0$, and where K depends only on V. "New margin type bounds that are based to a greater extent on complexity measures of individual classifier functions from the convex hull, are more adaptive and more flexible than previously shown bounds" (Koltchinskii and Panchenko, 2004).

Here, we are interested in studying the generalization performance of functions from a convex hull of *random* class of functions (random convex hull), that is, the class of learners is no longer fixed and depends on the data. This is done by deriving a new version of Vapnik's inequality applied to random classes, that is, a bound for relative deviations of frequencies from probabilities for random classes of events. The proof of the inequality mirrors the proofs of Vapnik's inequality for non-random classes of sets (see Vapnik et al., 1974; Vapnik, 1998; Anthony and Shawe-Taylor, 1993) but with the observation that the symmetrization step of the proof can be carried out for random classes of sets. The new version of Vapnik's inequality is then applied to derive flexible and adaptive bounds on the generalization errors of learners from random convex hulls. We exploit techniques previously used in deriving generalization bounds for convex combinations of functions from non-random classes in (Koltchinskii and Panchenko, 2004), and several measures of individual classifier complexities, such as effective dimension, pointwise variance and weighted variance over clusters, similar to the measures introduced by Koltchinskii and Panchenko (2004).

Surprisingly, the idea of studying random convex hulls allows one simultaneously to prove generalization results, and incorporate measures of classifier complexities in the bounds, for several existing algorithms such as SVMs, boosting with decision stumps, radial basis function networks and combinations of indicator functions over sets with finite VC dimension. It is also noteworthy that an extension on the VC theory of statistical learning to data dependent spaces of classifiers was recently found by Cannon et al., 2002, who defined a measure of complexity for data dependent hypothesis classes and provide data dependent versions of bounds on error deviance and estimation error.

2. Definition of Random Classes

First, an inequality that concerns the uniform relative deviation over a random class of events of relative frequencies from probabilities is exhibited. This inequality is an extension of the following Vapnik's inequality for a fixed VC-class $\mathcal{C}$ (with finite VC-dimension V) of sets (see Vapnik et al. (1974); Vapnik (1998); Anthony and Shawe-Taylor (1993)):

$$\mathbb{P}^n\Big(\sup_{C\in\mathcal{C}}\Big[\frac{\mathbb{P}(C)-\frac{1}{n}\sum_{i=1}^n I(x_i\in C)}{\sqrt{(\mathbb{P}(C))}}\Big]\geq t\Big)\leq 4\Big(\frac{2en}{V}\Big)^V e^{-\frac{nt^2}{4}}. \tag{2.1}$$

Inequality (2.1) allows one to prove stronger generalization results for several problems discussed in (Vapnik, 1998). In order to extend the above inequality to random classes of sets, we introduce the following definitions. Let $(\mathcal{Z},\mathcal{S},\mathbb{P})$ be a probability space. For a sample $\{z_1,\ldots,z_n\}$, $z_i\in\mathcal{Z}$, $i=1,\ldots,n$, define $z^n=(z_1,\ldots,z_n)$ and let $I(z^n)=\{z_i:1\leq i\leq n\}$. Let $\mathcal{C}(z^n)\in\mathcal{S}$ be a class of sets, possibly dependent on the sample $z^n=(z_1,\ldots,z_n)\in\mathcal{Z}^n$.
The integer $\Delta_{\mathcal{C}(z^n)}(z^n)$ is defined to be the number of distinct sets of the form $A\bigcap I(z^n)$, where A runs through $\mathcal{C}(z^n)$, that is, $\Delta_{\mathcal{C}(z^n)}(z^n)=\mathrm{card}\{A\bigcap\{z_1,\ldots,z_n\},A\in\mathcal{C}(z^n)\}$. The random collection of level sets $\mathcal{C}(z^n)=\Big\{A=\{z\in\mathcal{Z}:h(z)\leq 0\},h\in\mathcal{H}(z_1,\ldots,z_n)\Big\}$, where $\mathcal{H}(z^n)$ is a random class of functions possibly depending on z^n serves as a useful example. We call $A\bigcap I(z^n)$ a representation of the sample z^n by the set A. $\Delta_{\mathcal{C}(z^n)}(z^n)$ is the number of different representation of $\{z_1,\ldots,z_n\}$ by functions from $\mathcal{H}(z^n)$.

Now consider the random collection $\mathcal{C}(z^n)$ of $\mathcal{S}$-measurable subsets of $\mathcal{Z}$,

$$\mathcal{C}(z^n)=\{A:A\in\mathcal{S}\},$$

having the following properties:

$$1.)\ \mathcal{C}(z^n)\subseteq\mathcal{C}\Big(z^n\bigcup y\Big),\ z^n\in\mathcal{Z}^n,y\in\mathcal{Z} \tag{2.2}$$

(*the incremental property*)

$$2.)\ \mathcal{C}(z_{\pi(1)},\ldots,z_{\pi(n)})\equiv\mathcal{C}(z_1,\ldots,z_n), \tag{2.3}$$

for any permutation π of $\{z_1,\ldots,z_n\}$ (*the permutation property*).

The relative frequency of $A\in\mathcal{C}(z^n)$ on $z^n=(z_1,\ldots,z_n)\in\mathcal{Z}^n$ is defined to be

$$\mathbb{P}_{z^n}(A)=\frac{1}{n}|\{i:z_i\in A\}|=\frac{1}{n}\,|I(z^n)\cap A|\,,$$

where $|A|$ denotes the cardinality of a set A.

Let $\mathbb{P}^n$ be the product measure on n copies of $(\mathcal{Z},\mathcal{S},\mathbb{P})$, and $\mathbb{E}_n$ the expectation with respect to $\mathbb{P}^n$. Define

$$G(n)=\mathbb{E}_n\Delta_{\mathcal{C}(z^n)}(z^n).$$

3. Main Results

Given the above definitions, the following theorem holds.

Theorem 1 *For any $t > 0$,*

$$\mathbb{P}^n\Big\{z^n \in \mathcal{Z}^n : \sup_{A\in\mathcal{C}(z^n)} \frac{\mathbb{P}(A) - \mathbb{P}_{z^n}(A)}{\sqrt{\mathbb{P}(A)}} \geq t\Big\} \leq 4G(2n)e^{-\frac{nt^2}{4}}. \tag{3.4}$$

The proof of this theorem is given in the following Section 4. Observe that if the random collection $\mathcal{C}$ of sets is a VC-class (Vapnik, 1998), then the inequality (3.4) is the same as Vapnik's inequality (2.1) for VC-classes. Based on this theorem and the above definitions, several results on the generalization performance and the complexity of classifiers from random classes are exhibited below.

The following notation and definitions will be used from here on. Let $(\mathcal{X},\mathcal{A})$ be a measurable space (space of instances) and take $\mathcal{Y} = \{-1,1\}$ to be the set of labels. Let $\mathbb{P}$ be the probability measure on $\left(\mathcal{X}\times\mathcal{Y}, \mathcal{A}\times 2^{\{-1,1\}}\right)$ and let $(X_i,Y_i), i = 1,\ldots,n$ be i.i.d random pairs in $\mathcal{X}\times\mathcal{Y}$, randomly sampled with respect to the distribution $\mathbb{P}$ of a random variable (X,Y). The probability measure on the main sample space on which all of the random variables are defined will be denoted by P. Let $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, $Z_i = (X_i,Y_i), i = 1,\ldots,n$ and $Z^n = (Z_1,\ldots,Z_n)$. We will also define several random classes of functions and show how several learning algorithms generate functions from the convex hulls of random classes.

Consider the following four problems for which bounds on the generalization errors will be shown using inequality (3.4).

Problem 1. *Support vector machine (SVM) classifiers with uniformly bounded kernels.*

Consider any solution of an SVM algorithm $f(x) = \sum_{i=1}^n \lambda_i Y_i K(X_i,x)$, where $K(.,.) : \mathcal{X}\times\mathcal{X} \to [0,1]$ is the kernel and $\lambda_i \geq 0$. $\text{sign}(f(x))$ is used to classify $x \in \mathcal{X}$ in class $+1$ or -1. Take the random function class

$$\mathcal{H}(Z^n) = \{Y_i K(X_i,x) : i = 1,\ldots,n\},$$

which depends on the random sample $Z^n \in \mathcal{Z}^n$. The classifier function

$$f'(x) = \sum_{i=1}^n \lambda_i' Y_i K(X_i,x),\ \lambda_i' = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j},\ i = 1,\ldots,n$$

belongs to $\text{conv}(\mathcal{H}(Z^n))$ and the probability of error $\mathbb{P}(Yf(X) \leq 0) = \mathbb{P}(Yf'(X) \leq 0)$.

Problem 2. *Classifiers, built by some two-level component based hierarchies of SVMs (Heisele et al. (2001);Andonova (2004)) or kernel-based classifiers (like the one produced by radial basis function (RBF) networks).*

We explore component based hierarchies, such that the first level of the hierarchy is formed by SVM classifiers (with kernel K) built on each component (formed for example by projecting of the input space $\mathcal{X} \subseteq \mathbb{R}^m$ of instances onto subspace of $\mathbb{R}^l$, $l < m$) and the second level of the hierarchy is a linear combination of the real-valued outputs on each component of the classifier functions from the first level (for example, applying SVM with linear kernel or boosting methods on the output from the first level). In our formulation, the components of the hierarchy can depend on the training data (for example, found through dimensionality reduction algorithms, such as self-organizing maps (SOM, Kohonen (1990))). The type of the hierarchical classifier functions are of

this form sign$(f(x))$, where

$$f(x,\alpha,Q,w_2)=\sum_{j=1}^{d} w_2^j \sum_{i=1}^{n} \alpha_i^j Y_i K(Q^j X_i, Q^j x), Y_i=\pm 1,$$

where Q^j are the projections of the instances (determining the "components"), $w_2^j \in \mathbb{R}, \alpha_i^j \geq 0$. One can consider Q^j being nonlinear transformation of the instance space, for example applying filter functions. Let $|K(x,t)| \leq 1, \forall x,t \in \mathcal{X}$. Consider the random function class

$$\mathcal{H}(X_1,\ldots,X_n)=\{\pm K(Q^j X_i, Q^j x) : i \leq n, j=1,\ldots,d\},$$

where n is the number of training points (X_i,Y_i) and d is the number of the components.

In the case of RBF networks with one hidden layer and a linear threshold, the classifier function is of the form

$$f(x)=\sum_{j=1}^{d}\sum_{i=1}^{\hat{n}} \alpha_j^i K_{\sigma_j}(c_i,x),$$

where $c_i, i=1,\ldots,\hat{n}$ are centers of clusters, formed by clustering the training points $\{X_1,\ldots,X_n\}$ and σ_j (they can depend on the training data $(X_i,Y_i), i=1,\ldots,n$) are different widths for the Gaussian kernel, $K_{\sigma_j}(c_i,x)=e^{-\frac{||c_i-x||^2}{\sigma_j^2}}$. Consider the following random function class

$$\mathcal{H}(Z^n)=\{\pm K_{\sigma_j}(c_i,x) : i \leq \hat{n}, j \leq d\},$$

where $\hat{n}$ is the number of clusters, which is bounded by the number n of training points, and the cluster centers $\{c_i\}_{i=1}^{\hat{n}}$ depend on the training instances $\{X_i\}_{i=1}^{n}$.

Without loss of generality, we can consider $f \in \text{conv}(\mathcal{H}(Z^n))$ in both of the above described algorithms, after normalizing the classifier function with the sum of the absolute values of the coefficients in front of the random functions.

Problem 3. *Boosting over decision stumps.*

Given a finite set of d functions $\{h_i : \mathcal{X}\times\mathcal{X} \to [-1,1]\}$ for $i \leq d$, define the random class of as $\mathcal{H}(X_1,\ldots,X_n)=\{h_i(X_j,x) : j \leq n, i \leq d\}$, where n is the number of training points (X_i,Y_i). This type of random class is used for example in aggregating combined classifier by boosting over decision stumps. Indeed, decision stumps are simple classifiers, h, of the types $2I(x^i \leq a)-1$ or $2I(x^i \geq a)-1$, where $i \in \{1,\ldots,m\}$ is the direction of splitting ($\mathcal{X} \subset \mathbb{R}^m$) and $a \in \mathbb{R}$ is the threshold. It is clear that the threshold a can be chosen among $X_1^i,\ldots,X_n^i$ (the performance of the stump will remain the same on the training data). In this case, take $h_i(X_j,x)=2I(x^i \leq X_j^i)-1$ or $\widetilde{h}_i(X_j,x)=2I(x^i \geq X_j^i)-1$ and $\mathcal{H}(X_1,\ldots,X_n)=\{h_i(X_j,x),\widetilde{h}_i(X_j,x) : j \leq n, i \leq m\}$ and take $d=2m$.

Without loss of generality, we can consider $f \in \text{conv}(\mathcal{H}(X_1,\ldots,X_n))$, after normalizing with the sum of the absolute values of the coefficients.

Problem 4. *VC-classes of sets.*

Let the random class of functions $\mathcal{H}(Z^n)$ has the property that for all $h \in \mathcal{H}(Z^n)$, $h \in \{-1,1\}$ the VC dimension V of the class of sets $\{\{x \in \mathcal{X} : h(x)=1\}, h \in \mathcal{H}(Z^n)\}$ is finite.

A classifier is formed by taking convex combinations of functions from the class $\mathcal{H}(Z^n)$. Problem 4, in the case when the class $\mathcal{H}$ is not depending on the random sample Z^n, was approached

before with the previously existing VC-inequalities for indicator functions (Vapnik, 1998; Vapnik et al., 1968). The results shown here for Problem 4 in the case when $\mathcal{H}$ is a random class, are comparable to those derived before for indicators over class of sets with finite VC dimension.

In *general*, all of the above four types of problems, consider the convex combinations of functions from the random convex hull

$$\mathcal{F}(Z^n) = \text{conv}(\mathcal{H}(Z^n)) = \bigcup_{T\in\mathbb{N}} \mathcal{F}_T(Z^n),$$
$$\mathcal{F}_T(Z^n) = \left\{\Sigma_{i=1}^T \lambda_i h_i, \lambda_i \geq 0, \Sigma_{i=1}^T \lambda_i = 1, h_i \in \mathcal{H}(Z^n)\right\}, \tag{3.5}$$

where $\mathcal{H}(Z^n)$ is for example one of the random classes defined above, such that either $|\mathcal{H}(Z^n)| = \text{card}(\mathcal{H}(Z^n))$ is finite, or $\mathcal{H}(Z^n)$ is a collection of indicators over random class of sets with finite VC-dimension.

General problem:

We are interested is the following general problem:

Let $\mathcal{H}$ be a general-base class of uniformly bounded functions with values in $[-1,1]$. Let $Z_1,\ldots,Z_n, Z_i = (X_i,Y_i) \in \mathcal{X}\times\mathcal{Y}$ be i.i.d. (training data) sampled with respect to the distribution $\mathbb{P}$. Assume that based on the data $Z_1,\ldots,Z_n$ one selects a class of functions $\mathcal{H}(Z^n) \subseteq \mathcal{H}$ that is either

i) with *finite cardinality* depending on the data, such that
$\frac{\ln(\sup_{Z^n}|\mathcal{H}(Z^n)|)\ln n}{n} \to 0$ for $n\to\infty$, or

ii) $\mathcal{H}(Z^n) \subseteq \mathcal{H}$ is a collection of indicator functions $\{2I_C - 1 : C \in \mathcal{C}_{Z^n}\}$ over a class of sets $\mathcal{C}_{Z^n}$ with *finite VC-dimension* V.

We will call $\mathcal{H}(Z^n)$ a *random-base class* of functions. We are interested in studying the generalization errors for classifier functions $f \in \text{conv}(\mathcal{H}(Z^n))$ that are produced by broad classes of algorithms. Let us take

$$G^*(n,\mathcal{H}) = \sup_{Z^n\in(\mathcal{X}\times\mathcal{Y})^n} |\mathcal{H}(Z^n)|,$$

when $\mathcal{H}$ is the general-base class and the random-base classes $\mathcal{H}(Z^n)$ are with finite cardinality $\text{H}(Z^n)$, and take

$$G^*(n,\mathcal{H}) = \left(\frac{ne}{V}\right)^V,$$

when $\mathcal{H}$ is the general-base and the random-base classes $\mathcal{H}(Z^n)$ are collections of indicators over class of sets with finite VC-dimension V (Problem 4).

From the definitions and Problems 1, 2 and 3, it is clear that $G^*(n,\mathcal{H}) \leq 2n$ for Problem 1 and $G^*(n,\mathcal{H}) \leq 2nd$ for Problems 2 and 3. For completeness of the results in case of $\mathcal{H}(Z^n)$ being a collection of indicators over class of sets with finite VC-dimension V, we will assume that $n \geq \frac{V}{2e}$.

Following the notation by Koltchinskii and Panchenko (2004), let $\mathcal{P}(\mathcal{H}(Z^n))$ be the collection of all discrete distributions over the random-base class $\mathcal{H}(Z^n)$. Let $\lambda \in \mathcal{P}(\mathcal{H}(Z^n))$ and $f(x) = \int h(x)\lambda(dh)$, which is equivalent to $f \in \text{conv}(\mathcal{H}(Z^n))$. The generalization error of any classifier function is defined as

$$\mathbb{P}(\text{sign}(f(x)) \neq y) = \mathbb{P}(yf(x) \leq 0) = \mathbb{E}(I(yf(x) \leq 0)).$$

Given an i.i.d sample $(X_i,Y_i), i = 1,\dots n$ from the distribution $\mathbb{P}$, let $\mathbb{P}_n$ denote the empirical distribution and for any measurable function g on $\mathcal{X}\times\mathcal{Y}$, let

$$\mathbb{P}g = \int g(x,y)d\mathbb{P}(x,y),\ \ \mathbb{P}_n g = \frac{1}{n}\sum_{i=1}^{n} g(X_i,Y_i).$$

The first bound we show for the generalization errors of functions from random classes of functions is the following:

Theorem 2 *Let $\mathcal{H}$ be a general-base class of functions. For any $t>0$, with probability at least $1-e^{-t}$, for any n i.i.d. random pairs $(X_1,Y_1),\dots,(X_n,Y_n)$ randomly drawn with respect to the distribution $\mathbb{P}$, for all $\lambda\in\mathcal{P}(\mathcal{H}(Z^n))$ and $f(x)=\int h(x)\lambda(dh)$,*

$$\mathbb{P}(yf(x)\le 0) \le \inf_{0<\delta\le 1}\left(U^{\frac{1}{2}} + (\mathbb{P}_n(yf(x)\le 2\delta) + \frac{1}{n} + U)^{\frac{1}{2}}\right)^2 + \frac{1}{n}, \tag{3.6}$$

where

$$U = \frac{1}{n}\left(t + \ln\frac{4}{\delta} + \frac{8\ln n}{\delta^2}\ln G^*(2n,\mathcal{H}) + \ln(8n+4)\right).$$

The proof of this theorem is given in Section 4. It is based on random approximation of a function and Hoeffding-Černoff inequality as in (Koltchinskii and Panchenko, 2004), exploring the properties of random class of the level sets of the margins of the approximating functions, defined in the proof and Inequality (3.4).

The first result for the generalization error of classifiers from $\mathrm{conv}(\mathcal{H})$, where $\mathcal{H}$ is a fixed VC-class, was achieved by Schapire et al. (1998). They explained the generalization ability of classifiers from $\mathrm{conv}(\mathcal{H})$ in terms of the empirical distribution $\mathbb{P}_n(yf(x)\le\delta)$, $f\in\mathrm{conv}(\mathcal{H})$ of the quantity $yf(x)$, known in the literature as *margin* ("confidence" of prediction of the example x) and proposed several boosting algorithms that are built on the idea of maximizing the margin. Important properties, development, improvements and optimality of the generalization results of this type for broader fixed classes of functions $\mathcal{H}$ were shown by Koltchinskii and Panchenko (2004). The bound on the generalization error shown here is valid for random classes of functions and is not optimized for convergence with respect to n. Here, we have a different goal: to prove generalization results for random classes of functions that relate to broader classes of algorithms. Exploring the optimality of this result remains an open question.

In the rest of the paper, we will explore the individual complexity of classifier $f\in\mathrm{conv}(\mathcal{H})$, following the line of investigation begun by Koltchinskii and Panchenko (2004). We will explore the structure of the *random convex hull* and show bounds similar to the ones by Koltchinskii and Panchenko (2004) that reflect some measures of complexity of convex combinations.

First, we explore how the sparsity of the weights of a function from a random convex hull influences the generalization performance of the convex combination. Here, recall from Koltchinskii and Panchenko (2004), by sparsity of the weights in convex combination, we mean rapidity of decrease. For $\delta>0$ and $f(x)=\sum_{i=1}^{T}\lambda_i h_i(x), \sum_i\lambda_i=1, \lambda_i\ge 0$, let us define the *dimension* function by

$$e_n(f,\delta) = \left(T - \sum_{k=1}^{T}(1-\lambda_k)^{\frac{8\ln n}{\delta^2}}\right). \tag{3.7}$$

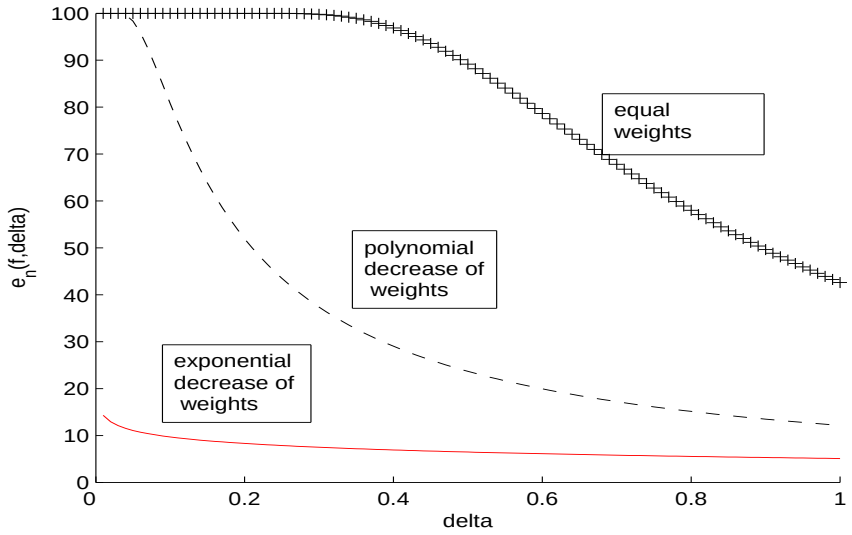

Figure 1: Dimension Function; From top to bottom: equal,polynomial, exponential decay; the x-axis is δ, the y-axis is dimension function value.

The name of this function is motivated by the fact that it can be interpreted as a dimension of a subset of the random convex hull $\mathrm{conv}(\mathcal{H})$ containing a function approximating f "well enough" (Koltchinskii and Panchenko, 2004). In a way, the dimension function measures the sparsity of the weights in the convex combination f. We plot the dimension function (see Fig. 1) in the cases when $T = 100, n = 1000$ and the weights $\{\lambda_i\}_{i=1}^T$ are equal, polynomially decreasing ($\lambda_i = i^{-2}$) and exponentially decreasing ($\lambda_i = e^{(-i+1)}$). One can see in Fig. 1 that when the weights decrease faster, the dimension function values are uniformly smaller with respect to $\delta \in (0,1]$. (For different sparsity measures of voting classifiers from convex hulls of VC-major classes see (Koltchinskii et al. (2003a); Koltchinskii and Panchenko (2004); Andonova (2004).)

Theorem 3 *(Sparsity bound) For any $t > 0$, with probability at least $1 - e^{-t}$, for any n i.i.d. random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ randomly drawn with respect to the distribution $\mathbb{P}$, for all $\lambda \in \mathcal{P}(\mathcal{H}(Z^n))$ and $f(x) = \sum_{i=1}^T \lambda_i h_i(x)$,*

$$\mathbb{P}(yf(x) \le 0) \le \inf_{0<\delta\le 1} \left(U^{1/2} + (\mathbb{P}_n(yf(x) \le 2\delta) + U + \frac{1}{n})^{1/2} \right)^2 + \frac{1}{n}, \tag{3.8}$$

where

$$U = \frac{1}{n}\left(t + \ln\frac{4}{\delta} + e_n(f,\delta)\ln G^*(2n,\mathcal{H}) + e_n(f,\delta)\ln(\frac{8}{\delta^2}\ln n) + \ln(8n+4) \right).$$

The proof of this theorem is also shown in Section 4. It is based on random approximation of functions similarly to the proof of Theorem 2, Hoeffding-Černoff inequality, properties of conditional expectation, exploring the capacity of random class of the level sets of the margins of the approximating functions and Inequality (3.4). The constants are explicit. For many experimental results,

showing the behavior of the above bound in the case of convex combinations of decision stumps, such as those produced by various boosting algorithms (see Andonova, 2004). There, it is shown experimentally that the sparsity bound is indicative of the generalization performance of a classifier from the convex hull of stumps. For further results and new algorithms, utilizing various notions of sparsity of the weight of convex combinations (see Koltchinskii et al., 2003a). In the case of hard margin SVM classifiers $f(x) = \sum_{i=1}^{T} \lambda_i K(X_i, x)$ with uniformly bounded kernels, the bound with margin δ becomes of order

$$\mathbb{P}(yf(x) \leq 0) \preceq \frac{\min(T, 8\delta^{-2} \ln n)}{n} \ln \frac{n}{\delta},$$

because

$$e_n(f, \delta) \leq \min(T, 8\delta^{-2} \ln n).$$

The inequality for $e_n(f,\delta)$ follows from the inequality $(1-\lambda)^p \geq 1 - p\lambda$ for $\lambda \in [0,1]$ and $p \geq 1$, and the fact that $\sum_{k=1}^{T} \lambda_k \leq 1$.

This zero-error bound is comparable to the compression bound (Littlestone and Warmuth, 1986) of order $\frac{T}{n-T} \ln \frac{n}{T}$, and the bounds of Bartlett and Shawe-Taylor (1999), where $U \sim \frac{R^2}{n\delta^2} \ln^2 n$ and $R \leq 1$ in case of $K(x,y) \leq 1$. When $T \ll n$ the bound in (3.8) is an improvement of the last bound. For example, $T \ll n$ when SVMs produce very "sparse" solutions (small number of support vectors), that is, the vector of weights $(\lambda_1, \ldots, \lambda_T)$ is sparse. The sparseness (in the sense of there being a small number of support vectors) of the solutions of SVMs was recently explored by Steinwart (2003), where lower bounds (of order $O(n)$) on the number T of support vectors for specific types of kernels were shown; in those cases, the bound in (3.8), relaxed to the upper bound of $e_n(f,\delta) \leq \min(T, 8\delta^{-2} \ln n)$, is probably not a significant improvement of the result of Bartlett and Shawe-Taylor (1999). The sparsity of weights of the solutions of SVMs, understood as rapidity of decrease of weights, is in need of further exploration, as it would provide better insight into the bound (3.8) of the generalizations error.

We now notice also that, because $e_n(f,\delta) \leq \min(T, 8\delta^{-2} \ln n)$ and $G^*(n, \mathcal{H}) \leq 2n$ for Problem 1 and $G^*(n, \mathcal{H}) \leq 2nd$ for Problems 2, 3 and $G^*(n, \mathcal{H}) = \left(\frac{ne}{V}\right)^V$, the bound (3.8) is extension of the results of Breiman (1999) for zero-error case, and is similar in nature to the result of Koltchinskii and Panchenko (2004) and Koltchinskii et al. (2003b), but now holding for *random* classes of functions.

Motivations for considering different bounds on the generalization error of classifiers that take into account measures of closeness of random functions in convex combinations and their clustering properties were given by Koltchinskii and Panchenko (2004). We now review those complexities and show bounds on the generalization error, that are similar to the ones proven by Koltchinskii and Panchenko (2004), but applied for different classes of functions. The proofs of the results are similar to those exhibited by Koltchinskii and Panchenko (2004).

Recall that a pointwise variance of h with respect to the distribution $\lambda \in \mathcal{P}(\mathcal{H}(Z^n))$ is defined by

$$\sigma_\lambda^2(x) = \int \Big(h(x) - \int h(x)\lambda(dh)\Big)^2 \lambda(dh), \tag{3.9}$$

where, $\sigma_\lambda^2(x) = 0$ if and only if $h_1(x) = h_2(x)$ for all $h_1, h_2 \in \mathcal{H}(Z^n)$ (Koltchinskii and Panchenko, 2004). The following theorem holds:

Theorem 4 *For any $t > 0$ with probability at least $1 - e^{-t}$, for any n i.i.d. random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ randomly drawn with respect to the distribution $\mathbb{P}$, for all $\lambda \in \mathcal{P}(\mathcal{H}(Z^n))$ and $f(x) = \int h(x)\lambda(dh)$,*

$$\mathbb{P}(yf(x) \leq 0) \leq \inf_{0<\delta\leq\gamma\leq 1}\Big(2\mathbb{P}_n(yf(x)\leq 2\delta)+4\mathbb{P}_n(\sigma^2_\lambda(x)\geq\frac{\gamma}{3})+ + \frac{8}{n}\left(\frac{56\gamma}{\delta^2}(\ln n)\ln G^*(2n,\mathcal{H})+\ln(8n+4)+t+\ln\frac{2\gamma}{\delta}\right)+\frac{6}{n}\Big). \quad (3.10)$$

The proof is given in Section 4. This time, the proof incorporates random approximations of the classifier function and its variance, Bernstein's inequality as in (Koltchinskii and Panchenko, 2004), exploring the capacity of random class of the level sets of the margins of the approximating functions and Inequality (3.4).

This result is an improvement of the above margin-bound in the case that the total pointwise variance is small, that is, the classifier functions h_i in the convex combination f are close to each other. The constants in the bound are explicit. From the Remark of Theorem 3 in (Koltchinskii and Panchenko, 2004) and the above inequality (3.10), one can see that the quantity $\mathbb{P}_n\sigma^2_\lambda$ might provide a complexity penalty in the *general class* of problems defined above.

A result that improves inequality (3.10) by exploring the clustering properties of the convex combination from a *random* convex hull will now be shown.

Given $\lambda\in\mathcal{P}(\mathcal{H}(Z^n))$ and $f(x)=\int h(x)\lambda(dh)$, represent f as

$$f=\sum_{j=1}^{p}\alpha_j\sum_{k=1}^{T}\lambda_k^{(j)}h_k^{(j)}$$

with $\sum_{j\leq p}\alpha_j=1, T\leq\infty, h_k^{(j)}\in\mathcal{H}(Z^n)$.

Consider an element $c\in\mathcal{C}^p(\lambda)$, that is, $c=(\alpha_1,\ldots,\alpha_p,\lambda^1,\ldots,\lambda^p)$, such that $\alpha_i\in\Delta(m)=\left\{t_km^{-k}, k\in\mathbb{N}, t_k\in\{1,2,3,\ldots,m^k\}\right\}$, $m\in\mathbb{N}$, $\lambda=\sum_{i=1}^p\alpha_i\lambda^i$, and $\lambda^i\in\mathcal{P}(\mathcal{H}(Z^n)), i=1,\ldots,p$. Denote by $\alpha_c^*=\min_{i\in\{1,\ldots,p\}}\alpha_i$, where $\{\alpha_i\}_{i=1}^p$ are called the cluster weights. c is interpreted as a decomposition of λ into p clusters as in (Koltchinskii and Panchenko, 2004). For an element $c\in\mathcal{C}^p(\lambda)$, a weighted variance over clusters is defined by

$$\sigma^2(c;x)=\sum_{k=1}^{p}\alpha_k^2\sigma^2_{\lambda^k}(x), \quad (3.11)$$

where $\sigma^2_{\lambda^k}(x)$ are defined in (3.9). One can see in Fig. 2 that when the number of the clusters increases, the weighted variance over clusters uniformly decreases (shifts to the left). If there are p small clusters among functions $h_1,\ldots,h_T$, then one should be able to choose element $c\in\mathcal{C}^p(\lambda)$ so that $\sigma^2(c;x)$ will be small on the majority of data points $X_1,\ldots,X_n$ (Koltchinskii and Panchenko, 2004). The following theorem holds:

Theorem 5 *For any $m\in\mathbb{N}$, for any $t>0$ with probability at least $1-e^{-t}$, the following is true for any n i.i.d. random pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$ randomly drawn with respect to the distribution $\mathbb{P}$, for any $p\geq 1$, $c\in\mathcal{C}^p(\lambda)$, $\lambda=\sum_{i=1}^p\alpha_i\lambda^i\in\mathcal{P}(\mathcal{H}(Z^n))$, such that $\alpha_1,\ldots,\alpha_p\in\Delta(m)$ with $\sum_i\alpha_i\leq 1$ and $f(x)=\int h(x)\lambda(dh)$*

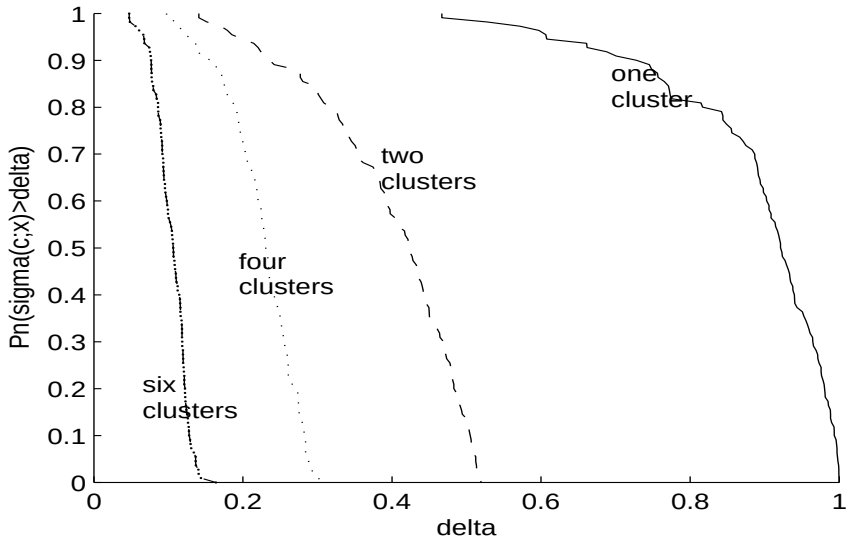

Figure 2: Empirical distribution of weighted variance over clusters; From right to left: one, tow, four, six clusters; the x-axis is δ.

$$\begin{aligned}\mathbb{P}(yf(x)\le 0) \;\le\; & \inf_{0<\delta\le\gamma\le 1}\Big(2\mathbb{P}_n(yf(x)\le 2\delta)+4\mathbb{P}_n(\sigma^2(c;x)\ge\gamma/3)+\\ & + \frac{8}{n}\Big(56p\frac{\gamma}{\delta^2}(\ln n)\ln G^*(2n,\mathcal{H})+\ln(8n+4)+2\sum_{j=1}^{p}\ln\Big(\log_m\frac{\alpha_j}{\alpha_c^*}+1\Big)+\\ & + 2\ln\Big(\log_m\frac{1}{\alpha_c^*}+1\Big)+2p\ln 2+t+\ln\frac{p^2\pi^4\gamma}{18\delta}\Big)+\frac{6}{n}\Big). \qquad (3.12)\end{aligned}$$

The proof is given in the following Section 4. Here, the proof incorporates more sophisticated random approximations of the classifier function and its weighted variance over clusters, Bernstein's inequality as in (Koltchinskii and Panchenko, 2004), exploring the capacity of random class of the level sets of the margins of the approximating functions and Inequality (3.4). The above bound can be simplified in the following way:

$$\begin{aligned}\mathbb{P}(yf(x)\le 0) \;\le\; & \inf_{0<\delta\le\gamma\le 1}\Big(2\mathbb{P}_n(yf(x)\le 2\delta)+4\mathbb{P}_n(\sigma^2(c;x)\ge\gamma/3)+\\ & + \frac{8}{n}\Big(56p\frac{\gamma}{\delta^2}(\ln n)\ln G^*(2n,\mathcal{H})+\ln(8n+4)+\\ & + 2(p+1)\ln\Big(\log_m\frac{1}{\alpha_c^*}+1\Big)+2p\ln 2+t+\ln\frac{p^2\pi^4\gamma}{18\delta}\Big)\Big).\end{aligned}$$

Define the number $\widehat{p}_\lambda(m,n,\gamma,\delta)$ of (γ,δ)-clusters of λ as the smallest p, for which there exists $c\in\mathcal{C}_\lambda^p$ such that (Koltchinskii and Panchenko, 2004)

$$\mathbb{P}_n(\sigma^2(c;x)\ge\gamma)\le 56p\frac{\gamma}{n\delta^2}(\ln n)\ln G^*(2n,\mathcal{H}).$$

Recall that $G^*(n,\mathcal{H}) \leq 2n$ for Problem 1 and $G^*(n,\mathcal{H}) \leq 2nd$ for Problems 2, 3 and $G^*(n,\mathcal{H}) = \left(\frac{ne}{V}\right)^V$. Then the above simplified bound implies that for all $\lambda = \sum_{i=1}^{p} \alpha_i \lambda^i \in \mathbb{P}(\mathcal{H}(Z^n))$, such that $\alpha_1,\ldots,\alpha_p \in \Delta(m)$ with $\sum_i \alpha_i \leq 1$,

$$\mathbb{P}(yf(x) \leq 0) \leq K \inf_{0<\delta\leq\gamma\leq 1}\Big(\mathbb{P}_n(yf(x)\leq\delta) \quad + \quad \widehat{p}_\lambda(m,n,\gamma,\delta)\frac{\gamma}{n\delta^2}(\ln n)\ln G^*(2n,\mathcal{H}) \\ + \quad \widehat{p}_\lambda(m,n,\gamma,\delta)\frac{\ln\left(\log_m \frac{1}{\alpha_c^*}+1\right)}{n}\Big).$$

Observe that if $\gamma = \delta$, then

$$\mathbb{P}(yf(x) \leq 0) \leq K\Big(\mathbb{P}_n(yf(x)\leq\delta) \quad + \quad \widehat{p}_\lambda(m,n,\delta,\delta)\frac{(\ln n)\ln G^*(2n,\mathcal{H})}{n\delta} \\ + \quad \widehat{p}_\lambda(m,n,\delta,\delta)\frac{\ln\left(\log_m \frac{1}{\alpha_c^*}+1\right)}{n}\Big).$$

The above bound is an improvement of the previous bounds in the case when there is a small number $\widehat{p}_\lambda$ of clusters so that the resulting weighted variance over clusters is small, and provided that the minimum of the cluster weights α_c^* is not too small. The bounds shown above are similar in nature to the bounds by Koltchinskii and Panchenko (2004) for base-classes $\mathcal{H}$ satisfying a general entropy condition. The advantages of the current results are that they are applicable for random classes of functions. The bounds derived here are with explicit constants. For more information regarding the empirical performance of the bounds and the complexities in the case of boosting with stumps and decision trees (see Koltchinskii et al., 2003b; Andonova, 2004). There, it is shown experimentally that generalization bounds based on weighted variance over clusters and margin capture the generalization performance of classifiers produced by several boosting algorithms over decision stumps. Our goal here is to show theoretically the impact of the complexity terms on the generalization performance of functions from random convex hulls, which happen to capture well known algorithms such as SVMs. More experimental evidences are needed to explore the above complexities in the setting of the *general problem* defined here.

4. Proofs

First we will prove the following lemma that will be used in the proof of Theorem 1.

Lemma 6 *For n large enough, if X is a random variable with values in* $\{0,1\}$, $P(X=1)=p,\ p \in \left[\frac{2}{n},1\right]$ *and* $X_1,\ldots,X_n$ *are independent random realizations of X (Bernoulli trials), then*

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p\right) \geq \frac{1}{4}.$$

Sketch of the Proof of Lemma 6.
We want to prove that

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p\right) = \sum_{k\geq np}\binom{n}{k} p^k (1-p)^{n-k} \geq \frac{1}{4}.$$

Observe that if $\frac{n-1}{n} < p \leq 1$, then $n \geq np > n-1$ and the inequality becomes $p^n > \left(\frac{n-1}{n}\right)^n \geq \frac{1}{4}$, which is true for $n \geq 2$.

Assume that $p \leq \frac{n-1}{n}$. The proof of the inequality in this case relies on Poisson and Gaussian approximation to binomial distribution. Let $S_n = \sum_{i=1}^n X_i$ and $Z_n = \frac{\sum_{i=1}^n (X_i - p)}{\sqrt{np(1-p)}}$. Notice that

$$P\left(\frac{1}{n}\sum_{i=1}^n X_i \geq p\right) = P(S_n \geq np) = P(Z_n \geq 0).$$

We want to show that there is n_0, such that for any $n \geq n_0$ the following is true for any $p \in \left[\frac{2}{n}, 1-\frac{1}{n}\right]$

$$P\left(\frac{1}{n}\sum_{i=1}^n X_i \geq p\right) \geq \frac{1}{4}$$

From the Poisson-Verteilung approximation theorem, (see Borowkow, 1976, Theorem 7, chapter 5, page 85) it follows that

$$P(S_n \geq \mu) \geq \sum_{k \geq np} \frac{\mu^k}{k!} e^{-\mu} - \frac{\mu^2}{n},$$

where $\mu = np \geq 2$. From the properties of the Poisson cumulative distribution function $F(x|\mu) = e^{-\mu}\sum_{i=0}^{\lfloor x \rfloor} \frac{\mu^i}{i!}$, one can see that $1 - F(x|\mu) > 1 - F(2|2) > 0.32$ for $x < \mu$ and $\mu \geq 2$. Therefore,

$$P(S_n \geq \mu) \geq 1 - F(x|\mu) - \frac{\mu^2}{n} > 0.32 - \frac{\mu^2}{n} = 0.32 - np^2.$$

Now, from the Berry-Esséen Theorem (see Feller, 1966, chapter XVI, page 515) one can derive that

$$|P(Z_n \geq 0) - 0.5| < \frac{33}{4} \cdot \frac{E(X-EX)^3}{\sqrt{n(E(X-EX)^2)^3}} = \frac{33}{4} \cdot \frac{p^2 + (1-p)^2}{\sqrt{np(1-p)}}.$$

Therefore, $P(Z_n \geq 0) > 0.5 - \frac{33}{4} \cdot \frac{p^2+(1-p)^2}{\sqrt{np(1-p)}}$. The goal is to find n_0 such that for any $n \geq n_0$ and $p \in \left[\frac{2}{n}, 1-\frac{1}{n}\right]$ the following is true:

$$\max\left\{0.32 - np^2, 0.5 - \frac{33}{4} \cdot \frac{p^2 + (1-p)^2}{\sqrt{np(1-p)}}\right\} \geq \frac{1}{4}.$$

Let $x = np^2$. One can see that the first term $0.32 - np^2 = 0.32 - x$ is decreasing with respect to x and the second term $0.5 - \frac{33}{4} \cdot \frac{p^2+(1-p)^2}{\sqrt{np(1-p)}} = 0.5 - \frac{33}{4} \cdot \frac{p^2+(1-p)^2}{\sqrt{(1-p)}(nx)^{\frac{1}{4}}}$ is increasing with respect to x. The solution $x(n)$ of the equation

$$0.32 - x = 0.5 - \frac{33}{4} \cdot \frac{x/n + (1-x/n)^2}{\sqrt{\left(1-\sqrt{x/n}\right)}(nx)^{\frac{1}{4}}}$$

is decreasing with respect to n and therefore one can find n_0, such that for $n > n_0$ the inequality $0.32 - x(n) \geq 0.25$ is true.

Remark: A shorter proof could be achieved if one directly shows that for $p \in \left[\frac{2}{n}, 1\right]$,

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p\right) = \sum_{k \geq np} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \geq \frac{1}{4}.$$

A stronger version of the above inequality for any p and n was used in (Vapnik (1998), page 133); however, a reference to a proof of this inequality appears currently to be unavailable.

□

Proof of Theorem 1.

The proof of Inequality (3.4) for random collection of sets of Theorem 1 follows the three main steps - Symmetrization, Randomization and Tail Inequality (see Vapnik (1998); Anthony and Shawe-Taylor (1993)). The difference with other approaches is that the symmetrization step of the proof is carried out for random classes invariant under permutation, after one combines the training set with a ghost sample and uses the incremental property of the random class. Note that symmetrization for a random subset under similar incremental and permutation properties was proved for the "standard" Vapnik's inequality by Gat (1999) (bounding the absolute deviation).

Let $t > 0$ be fixed. Assume that $n \geq 2/t^2$, otherwise if $n < 2/t^2$, then $4\exp^{-nt^2/4} > 1$; nothing more need be proved.

Denote the set

$$A = \left\{ x = (x_1, \ldots, x_n) \in \mathcal{Z}^n : \sup_{C \in \mathcal{C}(x)} \frac{\mathbb{P}(C) - \frac{1}{n}\sum I(x_i \in C)}{\sqrt{\mathbb{P}(C)}} \geq t \right\}.$$

Assume there exist a set C_x, such that

$$\frac{\mathbb{P}(C_x) - \frac{1}{n}\sum I(x_i \in C_x)}{\sqrt{\mathbb{P}(C_x)}} \geq t. \tag{4.13}$$

Then $\mathbb{P}(C_x) \geq t^2$. We have assumed that $t^2 \geq \frac{2}{n}$, therefore $\mathbb{P}(C_x) \geq \frac{2}{n}$.

Let $x' = (x'_1, \ldots, x'_n)$ be independent copy of $x = (x_1, \ldots, x_n)$. It can be observed (see Lemma 6 and Anthony and Shawe-Taylor (1993), Theorem 2.1) that since $\mathbb{P}(C_x) = \mathbb{E}(I(y \in C_x)) \geq \frac{2}{n}$, then with probability at least $1/4$

$$\mathbb{P}(C_x) \leq \frac{1}{n}\sum I(x'_i \in C_x). \tag{4.14}$$

From the assumption (4.13) and (4.14), then since $\frac{x-a}{\sqrt{x+a}}$ is a monotone and increasing function in $x > 0$ ($a > 0$), we have that

$$\begin{aligned} 0 < t \qquad &\leq \frac{\mathbb{P}(C_x) - \frac{1}{n}\sum I(x_i \in C_x)}{\sqrt{\mathbb{P}(C_x)}} \\ &\leq \frac{\mathbb{P}(C_x) - \frac{1}{n}\sum I(x_i \in C_x)}{\sqrt{\frac{1}{2}(\mathbb{P}(C_x) + \frac{1}{n}\sum I(x_i \in C_x))}} \\ &\leq \frac{\frac{1}{n}\sum I(x'_i \in C_x) - \frac{1}{n}\sum I(x_i \in C_x)}{\sqrt{\frac{1}{2}(\frac{1}{n}\sum I(x'_i \in C_x) + \frac{1}{n}\sum I(x_i \in C_x))}}. \end{aligned}$$

From (4.14) and the above inequality,

$$\begin{aligned}\frac{1}{4}I(x\in A) \quad &\le \mathbb{P}_{x'}\left(\mathbb{P}(C_x)\le \frac{1}{n}\sum I(x_i'\in C_x)\right) I(x\in A)\\ &\le \mathbb{P}_{x'}\left(\frac{\frac{1}{n}\sum I(x_i'\in C_x)-\frac{1}{n}\sum I(x_i\in C_x)}{\sqrt{\frac{1}{2}(\frac{1}{n}\sum I(x_i'\in C_x)+\frac{1}{n}\sum I(x_i\in C_x))}}\ge t\right)\\ &\le \mathbb{P}_{x'}\left(\sup_{C\in\mathcal{C}(x)}\frac{\frac{1}{n}\sum I(x_i'\in C)-\frac{1}{n}\sum I(x_i\in C)}{\sqrt{\frac{1}{2}(\frac{1}{n}\sum I(x_i'\in C)+\frac{1}{n}\sum I(x_i\in C))}}\ge t\right).\end{aligned}$$

Taking the expectation $\mathbb{E}_x$ of both sides,

$$\begin{aligned}&\mathbb{P}_x\left(\sup_{C\in\mathcal{C}(x)}\frac{\mathbb{P}(C)-\frac{1}{n}\sum_i I(x_i\in C)}{\sqrt{\mathbb{P}(C)}}\ge t\right)\le\\ &\le 4\mathbb{P}_{x,x'}\left(\sup_{C\in\mathcal{C}(x)}\frac{\frac{1}{n}\sum_i I(x_i'\in C)-\frac{1}{n}\sum_i I(x_i\in C)}{\sqrt{\frac{1}{2}(\frac{1}{n}\sum_i I(x_i'\in C)+\frac{1}{n}\sum_i I(x_i\in C))}}\ge t\right)\end{aligned}$$

(using increasing property)

$$\le 4\mathbb{P}_{x,x'}\left(\sup_{C\in\mathcal{C}(x,x')}\frac{\frac{1}{n}\sum_i I(x_i'\in C)-\frac{1}{n}\sum_i I(x_i\in C)}{\sqrt{\frac{1}{2}(\frac{1}{n}\sum_i I(x_i'\in C)+\frac{1}{n}\sum_i I(x_i\in C))}}\ge t\right)$$

(using permutation property)

$$= 4\mathbb{P}_{x,x',\varepsilon}\left(\sup_{C\in\mathcal{C}(x,x')}\frac{\frac{1}{n}\sum_i \varepsilon_i(I(x_i'\in C)-\sum_i I(x_i\in C))}{\sqrt{\frac{1}{2}(\frac{1}{n}\sum_i I(x_i'\in C)+\frac{1}{n}\sum_i I(x_i\in C))}}\ge t\right)$$

(using Hoeffding-Azuma's inequality)

$$\begin{aligned}&\le 4\mathbb{E}\left(\Delta_{\mathcal{C}(x,x')}(x_1,\ldots,x_n,x_1',\ldots,x_n')\dot{\exp}\left(-\frac{nt^2}{4\frac{\sum_i(I_i-I_i')^2}{\sum_i(I_i+I_i')}}\right)\right)\\ &\le 4\mathbb{E}_{x,x'}\left(\Delta_{\mathcal{C}(x,x')}(x_1,\ldots,x_n,x_1',\ldots,x_n')\dot{\exp}\left(-\frac{nt^2}{4}\right)\right)=\\ &= 4G(2n)\exp\left(-\frac{nt^2}{4}\right).\end{aligned}$$

Here the increasing (2.2) and permutation (2.3) properties of the random collection of sets have been used .

□

The following lemma will be useful in the proofs of Theorems 2, 3, 4 and 5.

Lemma 7 *Let $Z_1,\dots,Z_n$ be n i.i.d. random variables randomly drawn with respect to the distribution $\mathbb{P}$, $Z_i=(X_i,Y_i)\in\mathcal{X}\times\mathcal{Y}$. Let*

$$\mathcal{C}_{N,k}(Z^n)=\{C:C=\{(x,y)\in\mathcal{X}\times\mathcal{Y}:yg(x)\le\delta\},\ g\in\mathcal{G}_{N,k}(Z^n),\delta\in[0,1]\},$$

where

$$\mathcal{G}_{N,k}(Z^n)=\left\{g:g(z)=\frac{1}{N}\sum_{i=1}^{N}k_ih_i(z),h_i\in\mathcal{H}(Z^n),1\le k_i\le N-k+1,k_i\in\mathbb{N}\right\},\ N,k\in\mathbb{N}$$

and $\mathcal{H}(Z^n)$ is a random-base class from the general problem. Then

$$G(n)=\mathbb{E}_n\Delta_{\mathcal{C}_{N,k}(Z^n)}(Z^n)\le\min\left((n+1)(N-k+1)^k(G^*(n,\mathcal{H}))^k,2^n\right).$$

If $k=N$, then $k_i=1$ and $\mathcal{G}_{N,N}(Z^n)=\left\{g:g(z)=\frac{1}{N}\sum_{i=1}^{N}h_i(z),h_i\in\mathcal{H}(Z^n)\right\}$, where $N\in\mathbb{N}$. In this case, it is clear that $G(n)=\mathbb{E}_n\Delta_{\mathcal{C}_{N,N}(Z^n)}(Z^n)\le\min\left((n+1)(G^*(n,\mathcal{H}))^N,2^n\right)$.

Proof.

Following the notation we have to prove that if $\mathcal{H}(Z^n)$ is with finite cardinality $H(Z^n)$, then

$$G(n)=\mathbb{E}_n\Delta_{\mathcal{C}_{N,k}(Z^n)}(Z^n)\le\min\left((n+1)(N-k+1)^k\mathbb{E}_n\left(H(Z^n)^k\right),2^n\right)$$

and if $\mathcal{H}(Z^n)$ is a collection of indicators from the general problem, then

$$G(n)=\mathbb{E}_n\Delta_{\mathcal{C}_{N,k}(Z^n)}(Z^n)\le\min\left((n+1)(N-k+1)^k\left(\frac{ne}{V}\right)^{Vk},2^n\right).$$

First, let $\mathcal{H}(Z^n)$ be with finite cardinality $H(Z^n)$. Then

$$\operatorname{card}\mathcal{G}_{N,k}(Z^n)\le(N-k+1)^kH(Z^n)^k,$$

because for each $g\in\mathcal{G}_{N,k}(Z^n)$ there are k different functions $h_i\in\mathcal{H}(Z^n)$ participating in the convex combination and the integer coefficients $k_i\in\{1,\dots,N-k+1\}$. Also, for fixed $g\in\mathcal{G}_{N,k}(Z^n)$, it follows that

$$\operatorname{card}\left\{\{yg(x)\le\delta\}\bigcap\{z_1,\dots,z_n\},\delta\in[-1,1]\right\}\le(n+1).$$

(This is clear after re-ordering $Y_1g(X_1),\dots,Y_ng(X_n)\to Y_{i_1}g(X_{i_1})\le\ldots\le Y_{i_n}g(X_{i_n})$ and taking for values of $\delta\in\{Y_{i_1}g(X_{i_1}),\dots,Y_{i_n}g(X_{i_n}),1\}$.) Therefore,

$$G(n)=\mathbb{E}_n\Delta_{\mathcal{C}_{N,k}(Z^n)}(Z^n)\le\min\left((n+1)(N-k+1)^k\mathbb{E}_nH(Z^n)^k,2^n\right)\le$$

$$\le\min\left((n+1)(N-k+1)^k(G^*(n,\mathcal{H}))^k,2^n\right).$$

Next, let $\mathcal{H}(Z^n)$ be a collection of indicators over class of sets with finite VC-dimension V. Then, for fixed $\delta\in[0,1]$, the number of possible representations of $(Z_1,\dots,Z_n)$ by the class $\mathcal{C}_{N,k}(Z^n,\delta)=$

$\{C : C = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : yg(x) \leq \delta\}, g \in \mathcal{G}_{N,k}(Z^n)\}$ is bounded by $(N-k+1)^k \left(\frac{ne}{V}\right)^{Vk}$. Similarly to the previous case, for fixed $g \in \mathcal{G}_{N,k}(Z^n)$,

$$\text{card}\left\{\{yg(x) \leq \delta\} \bigcap \{z_1, \ldots, z_n\}, \delta \in [0,1]\right\} \leq (n+1),$$

and therefore

$$G(n) = \mathbb{E}_n \Delta_{\mathcal{C}_{N,k}(Z^n)}(Z^n) \leq \min\left((n+1)(N-k+1)^k \left(\frac{ne}{V}\right)^{Vk}, 2^n\right) =$$
$$= \min\left((n+1)(N-k+1)^k (G^*(n,\mathcal{H}))^k, 2^n\right).$$

□

Next, the proofs of Theorem 2,3, 4 and 5 are shown. They follow closely the proofs given by Koltchinskii and Panchenko (2004) and Koltchinskii et al. (2003b) for non random classes of functions. We adjust the proofs to hold for random classes of functions by using Inequality 3.4 from Theorem 1.

Define the function

$$\phi(a,b) = \frac{(a-b)^2}{a} I(a \geq b),$$

that is convex for $a > 0$ and increasing with respect to a, decreasing with respect to b.

Proof of Theorem 2.

Let $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ be i.i.d samples randomly drawn with respect to the distribution $\mathbb{P}$. Let us first fix $\delta \in (0,1]$ and let $f = \sum_{k=1}^T \lambda_k h_k \in \text{conv}(\mathcal{H}(Z^n))$ be any function from the convex hull of $\mathcal{H}(Z^n)$, where $\mathcal{H}(Z^n)$ is the random-base class defined in the general problem.
Given $N \geq 1$, generate i.i.d sequence of functions $\xi_1, \ldots, \xi_N$ according to the distribution $\lambda = (\lambda_1, \ldots, \lambda_T)$, $\mathbb{P}_\xi(\xi_i = h_k) = \lambda_k$ for $k = 1, \ldots, T$ and ξ_i are independent of $\{(X_k, Y_k)\}_{k=1}^n$. Then $\mathbb{E}_\xi \xi_i(x) = \sum_{k=1}^T \lambda_k h_k(x)$.

Consider a function

$$g(x) = \frac{1}{N} \sum_{k=1}^N \xi_k(x),$$

which plays the role of a random approximation of f in the following sense:

$$\begin{aligned}
\mathbb{P}(yf(x) \leq 0) &= \mathbb{P}\Big(yf(x) \leq 0, yg(x) \leq \delta\Big) + \mathbb{P}\Big(yf(x) \leq 0, yg(x) > \delta\Big) \\
&\leq \mathbb{P}\Big(yg(x) \leq \delta\Big) + \mathbb{E}_{x,y}\mathbb{P}_\xi\Big(\mathbb{E}_\xi yg(x) \leq 0, yg(x) \geq \delta\Big) \\
&\leq \mathbb{P}\Big(yg(x) \leq \delta\Big) + \mathbb{E}_{x,y}\mathbb{P}_\xi\Big(yg(x) - \mathbb{E}_\xi yg(x) \geq \delta\Big) \\
&= \mathbb{P}\Big(yg(x) \leq \delta\Big) + \mathbb{E}_{x,y}\mathbb{P}_\xi\left(\sum_{k=1}^N (y\xi_i(x) - y\mathbb{E}_\xi \xi_i(x)) \geq N\delta\right) \\
&\leq \mathbb{P}\Big(yg(x) \leq \delta\Big) + \exp\Big(\frac{-N\delta^2}{2}\Big), \qquad (4.15)
\end{aligned}$$

where in the last step is applied Hoeffding-Černoff inequality. Then,

$$\mathbb{P}\Big(yf(x) \leq 0\Big) \leq \mathbb{P}\Big(yg(x) \leq \delta\Big) + \exp(-N\delta^2/2). \qquad (4.16)$$

Similarly to the above inequality, one can derive that,

$$\mathbb{E}_\xi \mathbb{P}_n\Big(yg(x) \le \delta\Big) \le \mathbb{P}_n\Big(yf(x) \le 2\delta\Big) + \exp(-N\delta^2/2). \tag{4.17}$$

For any random realization of the sequence $\xi_1,\dots,\xi_N$, the random function g belongs to the class $\mathcal{G}_N(Z^n) = \Big\{\frac{1}{N}\sum_{i=1}^N h_i(x) : h_i \in \mathcal{H}(Z^n)\Big\}$.
Consider the random collection of level sets for fixed $N \in \mathbb{N}$,

$$\mathcal{C}(Z^n) = \Big\{C = \{(x,y) \in \mathcal{X}\times\mathcal{Y} : yg(x) \le \delta\}, g \in \mathcal{G}_N(Z^n), \delta \in (0,1]\Big\}.$$

Clearly $\mathcal{C}(Z^n)$ satisfies conditions (2.2) and (2.3). In order to apply the inequality for the random collection of sets (3.4), one has to estimate $G(n) = \mathbb{E}^n \Delta_{\mathcal{C}(Z^n)}(Z^n)$. By Lemma 7 it follows that $G(n) \le (G^*(n,\mathcal{H}))^N (n+1)$.
From this and Theorem 1, we have

$$\begin{aligned}\mathbb{P}^n\left(\sup_{C\in\mathcal{C}(Z^n)} \frac{\mathbb{P}(C) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C)}{\sqrt{\mathbb{P}(C)}} \ge t\right) &\le 4G(2n)e^{-\frac{nt^2}{4}} \le \\ &\le 4\left(G^*(2n,\mathcal{H})\right)^N (2n+1)e^{-\frac{nt^2}{4}} = e^{-u},\end{aligned}$$

where a change of variables $t = \sqrt{\frac{4}{n}(u + N\ln(G^*(2n,\mathcal{H})) + \ln(8n+4))}$ is made. So, for a fixed $\delta \in (0,1]$, for any $u > 0$ with probability at least $1 - e^{-u}$, it follows that

$$\frac{\mathbb{P}(yg(x) \le \delta) - \frac{1}{n}\sum_{i=1}^n I(Y_i g(X_i) \le \delta)}{\sqrt{\mathbb{P}(yg(x) \le \delta)}} \le \sqrt{\frac{4}{n}(u + N\ln(G^*(2n,\mathcal{H})) + \ln(8n+4))}. \tag{4.18}$$

The function $\phi(a,b), a > 0$ is convex. Therefore,

$$\mathbb{E}_\xi \phi\Big(\mathbb{P}(yg(x) \le \delta), \mathbb{P}_n(yg(x) \le \delta)\Big) \ge \phi\Big(\mathbb{E}_\xi \mathbb{P}(yg(x) \le \delta), \mathbb{E}_\xi \mathbb{P}_n(yg(x) \le \delta)\Big).$$

Based on the monotonic properties of $\phi(a,b)$ and inequalities (4.16) and (4.17), it is obtained that for any $\delta \in (0,1]$, for any $u > 0$ with probability at least $1 - e^{-u}$,

$$\begin{aligned}&\phi\Big(\mathbb{P}(yf(x) \le 0) - \exp(-N\delta^2/2), \mathbb{P}_n(yf(x) \le 2\delta + \exp(-N\delta^2/2))\Big) \le \\ &\le \frac{4}{n}(u + N\ln(G^*(2n,\mathcal{H})) + \ln(8n+4)).\end{aligned} \tag{4.19}$$

Choose $N = \frac{2\ln n}{\delta^2}$, such that $\exp(-N\delta^2/2) = \frac{1}{n}$. Take

$$U = \frac{1}{n}\left(u + \frac{2\ln n}{\delta^2}\ln(G^*(2n,\mathcal{H})) + \ln(8n+4)\right).$$

Solving the above inequality with respect to $\mathbb{P}(yf(x) \le 0)$, it follows that

$$\mathbb{P}(yf(x) \le 0) \le \left(\sqrt{U} + \sqrt{\mathbb{P}_n(yf(x) \le 2\delta) + \frac{1}{n} + U}\right)^2 + \frac{1}{n}.$$

In order to make the bound uniform with respect to $\delta \in (0,1]$, we apply standard union bound techniques (Koltchinskii and Panchenko, 2004). First, we prove the uniformity for $\delta \in \Delta = \{2^{-k}, k = 0,1,\ldots..\}$. Apply the above inequality for fixed $\delta \in \Delta$ by replacing u by $u + \ln\frac{2}{\delta}$ and hence e^{-u} replaced by $\frac{\delta}{2}e^{-u}$. Denote

$$U' = \frac{1}{n}\left(u + \ln\frac{2}{\delta} + \frac{2\ln n}{\delta^2}\ln(G^*(2n,\mathcal{H})) + \ln(8n+4)\right).$$

Then

$$\mathrm{P}\left[\bigcap_{\delta\in\Delta}\left\{\mathbb{P}(yf(x)\le 0) \le \left(\sqrt{U'} + \sqrt{\mathbb{P}_n(yf(x)\le 2\delta) + \frac{1}{n} + U'}\right)^2 + \frac{1}{n}\right\}\right] \ge$$
$$\ge 1 - e^{-u}\sum_{k=1}^{\infty} 2^{-k} \ge 1 - e^{-u}.$$

Now, in order to make the bound for any $\delta \in (0,1]$, observe that if $\delta_0 \in (0,1]$ then there is $k \in \mathbb{Z}_+$, $2^{-k-1} \le \delta_0 < 2^{-k}$.

Therefore, if the above bound holds for fixed $\delta_0 \in (0,1]$, then

$$\mathbb{P}_n(yf(x) \le \delta_0) \le \mathbb{P}_n\left(yf(x) \le 2^{-k}\right)$$

and

$$1/\delta_0^2 \le 2^{2k+2},\ \ln\frac{2}{\delta_0} \le \ln 2^{k+2}.$$

So, changing the constants in the bound, denote

$$U = \frac{1}{n}\left(t + \ln\frac{4}{\delta} + \frac{8\ln n}{\delta^2}\ln(G^*(2n,\mathcal{H})) + \ln(8n+4)\right).$$

It follows that, for any $t > 0$ with probability at least $1 - e^{-t}$ for any $\delta \in (0,1]$, the following holds:

$$\mathbb{P}(yf(x) \le 0) \le \left(\sqrt{U} + \sqrt{\mathbb{P}_n(yf(x)\le 2\delta) + \frac{1}{n} + U}\right)^2 + \frac{1}{n}$$

Thus, the Theorem 2 and inequality (3.6) hold.

□

Now, the **proof of Sparsity bound of Theorem 3** will be shown.

Denote $\Delta = \{2^{-k} : k \ge 1\}$ and $z = (x,y)$, $Z^n = \left((X_1,Y_1),\ldots,(X_n,Y_n)\right)$.

Let us fix $f(x) = \sum_{k=1}^{T}\lambda_k h_k(x) \in \ \mathrm{conv}(\mathcal{H}(Z^n))$. Given $N \ge 1$, generate an i.i.d. sequence of functions $\xi_1,\ldots,\xi_N$ according to the distribution $\mathbb{P}_\xi(\xi_i(x) = h_k(x)) = \lambda_k$ for $k = 1,\ldots,T$ and independent of $\{(X_i,Y_i)\}_{i=1}^n$. Clearly, $\mathbb{E}_\xi\xi_i(x) = \sum_{k=1}^{T}\lambda_k h_k(x)$. Consider the function

$$g(x) = \frac{1}{N}\sum_{k=1}^{N}\xi_k(x),$$

which plays the role of a random approximation of f and $\mathbb{E}_\xi g(x) = f(x)$. One can write,

$$\begin{aligned}\mathbb{P}(yf(x) \le 0) &= \mathbb{E}_\xi \mathbb{P}\Big(yf(x) \le 0, yg(x) < \delta\Big) + \mathbb{E}_\xi \mathbb{P}\Big(yf(x) \le 0, yg(x) \ge \delta\Big) \le \\ &\le \mathbb{E}_\xi \mathbb{P}\Big(yg(x) \le \delta\Big) + \mathbb{E}\mathbb{P}_\xi\Big(yg(x) \ge \delta, \mathbb{E}_\xi yg(x) \le 0\Big).\end{aligned}$$

In the last term for a fixed $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned}\mathbb{P}_\xi\Big(yg(x) \ge \delta, \mathbb{E}_\xi yg(x) \le 0\Big) &\le \mathbb{P}_\xi\Big(yg(x) - \mathbb{E}_\xi yg(x) \ge \delta\Big) = \\ &= \mathbb{P}_\xi\left(\sum_{i=1}^N (y\xi_i(x) - y\mathbb{E}_\xi \xi_i(x)) \ge N\delta\right) \le \exp\big(-N\delta^2/2\big).\end{aligned}$$

where in the last step Hoeffding-Černoff inequality has been applied. Hence,

$$\mathbb{P}(yf(x) \le 0) - e^{-N\delta^2/2} \le \mathbb{E}_\xi \mathbb{P}(yg(x) \le \delta). \tag{4.20}$$

Similarly,

$$\mathbb{E}_\xi \mathbb{P}_n(yg(x) \le \delta) \le \mathbb{P}_n(yf(x) \le 2\delta) + e^{-N\delta^2/2}. \tag{4.21}$$

Clearly, for any random realization of the sequence $\xi_1, \ldots, \xi_N$, the function $g(x)$ belongs to the class

$$F_{N,k}(Z^n) = \left\{ \frac{1}{N} \sum_{i=1}^k k_i h_i(x) : \sum_{i=1}^k k_i = N, 1 \le k_i \le N, h_i \in \mathcal{H}(Z^n) \right\},$$

for some $k \in \mathbb{N}$, which is the number of different indices i and $k_i \in \mathbb{N}$ is the number of repeating function h_i in the representation of g. Recall, $\mathcal{H}(Z^n)$ is the random-base class from the general problem. Then, $1 \le k \le \min(T,N)$. Let $p_{k,N} = \mathbb{P}_\xi(g \in F_{N,k}(Z^n))$.

Then the expectation $\mathbb{E}_\xi$ can be represented as

$$\mathbb{E}_\xi(L(g)) = \sum_{k \ge 1} p_{k,N} \mathbb{E}_\xi \left(L(g) | g \in F_{N,k}(Z^n) \right),$$

where L is a real valued measurable function and g is the random function

$$g(x) = \frac{1}{N} \sum_{k=1}^N \xi_k(x).$$

Now consider the random collection of sets

$$\mathcal{C}_{N,k}(Z^n) = \Big\{ C : C = \{(x,y) : yg(x) \le \delta\}, g \in F_{N,k}(Z^n), \delta \in (0,1] \Big\},$$

where $N,k \in \mathbb{N}$. Clearly $\mathcal{C}_{N,k}(Z^n)$ satisfies conditions (2.2) and (2.3). In order to apply the inequality for random collection of sets (3.4), one has to estimate $G'(n) = \mathbb{E}_n \Delta_{\mathcal{C}_{N,k}(Z^n)}(Z^n)$.

By Lemma 7, it follows that

$$G'(n) \le (G^*(n,\mathcal{H}))^k (N-k+1)^k (n+1) \le (G^*(n,\mathcal{H}))^k N^k (n+1).$$

Now apply Inequality (3.4) for the random collection of sets $\mathcal{C}_{N,k}(Z^n)$. Then, with probability at least $1-e^{-t}$

$$\frac{(\mathbb{P}_{x,y}(yg(x)\le\delta)-\mathbb{P}_n(yg(x)\le\delta))^2}{\mathbb{P}_{x,y}(yg(x)\le\delta)}\le\frac{4}{n}(t+k\ln G^*(2n,\mathcal{H})+k\ln N+\ln(8n+4)).$$

The function $\phi(a,b), a>0$ is convex, so $\phi(\mathbb{E}_\xi a,\mathbb{E}_\xi b)\le\mathbb{E}_\xi\phi(a,b)$ for $a>0$.

Therefore,

$$\frac{(\mathbb{E}_\xi\mathbb{P}_{x,y}(yg(x)\le\delta)-\mathbb{E}_\xi\mathbb{P}_n(yg(x)\le\delta))^2}{\mathbb{E}_\xi\mathbb{P}_{x,y}(yg(x)\le\delta)}\le\mathbb{E}_\xi\frac{(\mathbb{P}_{x,y}(yg(x)\le\delta)-\mathbb{P}_n(yg(x)\le\delta))^2}{\mathbb{P}_{x,y}(yg(x)\le\delta)}=$$

$$=\sum_{k\ge1}p_{k,N}\mathbb{E}_\xi\Big(\frac{(\mathbb{P}_{x,y}(yg(x)\le\delta)-\mathbb{P}_n(yg(x)\le\delta))^2}{\mathbb{P}_{x,y}(yg(x)\le\delta)}|g\in F_{N,k}(Z^n)\Big)\le$$

$$\le\sum_{k\ge1}p_{k,N}\frac{4}{n}(t+k\ln G^*(2n,\mathcal{H})+k\ln N+\ln(8n+4)).$$

Observe that

$$\sum_{k\ge1}kp_{k,N}=\mathbb{E}\text{card}\{k:k'th\text{ index is picked at least once }\}=$$

$$\sum_{k=1}^{T}\mathbb{E}I(k\text{ is picked at least once })=\sum_{k=1}^{T}(1-(1-\lambda_k)^N).$$

Denote $e_n(f,\delta)=\sum_{k=1}^{T}\Big(1-(1-\lambda_k)^N\Big)$. Let $N=\frac{2}{\delta^2}\ln n$, so that $e^{-N\delta^2/2}=\frac{1}{n}$.
The function $\phi(a,b)$ is increasing in a and decreasing in b. Combine the last result with (4.20) and (4.21):

$$\phi\Big(\mathbb{P}(yf(x)\le0)-n^{-1},\mathbb{P}_n(yf(x)\le2\delta)+n^{-1}\Big)\le$$
$$\le\frac{4}{n}(t+e_n(f,\delta)\ln G^*(2n,\mathcal{H})+e_n(f,\delta)\ln(\frac{2}{\delta^2}\ln n)+\ln(8n+4)).$$

Denote

$$W=\frac{1}{n}(t+e_n(f,\delta)\ln G^*(2n,\mathcal{H})+e_n(f,\delta)\ln(\frac{2}{\delta^2}\ln n)+\ln(8n+4)).$$

After solving the above inequality for $\mathbb{P}(yf(x)\le0)$, one can get that, for a fixed $\delta\in\{2^{-k}:k\ge1\}$, for every $t>0$ with probability at least $1-e^{-t}$ the following holds

$$\mathbb{P}(yf(x)\le0)\le\Big(\sqrt{W}+\sqrt{\mathbb{P}_n(yf(x)\le2\delta)+\frac{1}{n}+W}\Big)^2+\frac{1}{n}.\tag{4.22}$$

It remains to make the bound uniform over $\delta\in(0,1]$, which is done again by using standard union bound techniques shown in the proof of Theorem 2 and the observation that if $\delta_0\in(0,1]$, then there is $k\in\mathbb{Z}_+$, $2^{-k-1}<\delta_0\le2^{-k}$ and $e_n(f,\delta_0)\le\sum_{k=1}^{T}(1-(1-\lambda_i)^{8(\ln n)2^{2k}})$.

Redefine $e_n(f,\delta) = \sum_{k=1}^{T}(1-(1-\lambda_i)^{\frac{8\ln n}{\delta^2}})$.

So, by changing the constants in the bound, it follows that for any $t>0$, with probability at least $1-e^{-t}$ for any $\delta \in (0,1]$ the following holds:

$$\mathbb{P}(yf(x)\le 0) \le \left(\sqrt{U}+\sqrt{\mathbb{P}_n(yf(x)\le 2\delta)+\frac{1}{n}+U}\right)^2+\frac{1}{n},$$

where

$$U=\frac{1}{n}\left(t+\ln\frac{4}{\delta}+e_n(f,\delta)\ln G^*(2n,\mathcal{H})+e_n(f,\delta)\ln\left(\frac{8}{\delta^2}\ln n\right)+\ln(8n+4)\right).$$

Thus, the Theorem 3 and inequality (3.8) hold.

□

We now show the **proof** for the bound with the total variance in **Theorem 4**, using Theorem 1.

Given $f(x)=\sum_{k=1}^{T}\lambda_k h_k(x)$, and given $N\ge 1$, first generate an i.i.d. sequence of functions $\xi_1,\dots,\xi_N$ independently of $\{(X_i,Y_i)\}$ and according to the distribution $\mathbb{P}_\xi(\xi_i=h_k)=\lambda_k$, for $k=1,\dots,T$, and consider a function

$$g(x)=\frac{1}{N}\sum_{i=1}^{N}\xi_i(x),$$

which plays the role of random approximation of f.

The main difference from the proof of the above theorems is that in equation (4.15) the condition on the variance $\sigma_\lambda^2(x)$ is also introduced. Namely, one can write

$$\begin{aligned}\mathbb{P}(yf(x)\le 0)\le \mathbb{E}_\xi\mathbb{P}(yg(x)\le\delta) \quad &+ \quad \mathbb{P}\left(\sigma_\lambda^2(x)\ge\gamma\right)+\\ &+ \quad \mathbb{E}\mathbb{P}_\xi\left(yg(x)\ge\delta, yf(x)\le 0,\sigma_\lambda^2(x)\le\gamma\right).\end{aligned}$$

The variance of ξ_i's, for a fixed $x\in\mathcal{X}$, is

$$\mathrm{Var}_\xi(\xi_i(x))=\sigma_\lambda^2(x).$$

$-1\le\xi_i(x)\le 1$, as well. Bernstein's inequality,

$$\begin{aligned}&\mathbb{P}_\xi\left(yg(x)\ge\delta, yf(x)\le 0,\sigma_\lambda^2(x)\le\gamma\right)\le\\ &\le\mathbb{P}_\xi\left(\sum_{i=1}^{N}(y\xi_i(x)-y\mathbb{E}_\xi\xi_i(x))\ge N\delta|\mathrm{Var}_\xi(\xi_1(x))\le\gamma\right)\le\\ &\le 2\exp\Big(-\frac{1}{4}\min\Big(\frac{N\delta^2}{\gamma},N\delta\Big)\Big)=2\exp\Big(-\frac{1}{4}\frac{N\delta^2}{\gamma}\Big),\end{aligned}$$

is used, since it is assumed that $\gamma\ge\delta$. Making this term negligible by taking $N=4(\frac{\gamma}{\delta^2})\ln n$,

$$\mathbb{P}(yf(x)\le 0)\le\mathbb{E}_\xi\mathbb{P}(yg(x)\le\delta)+\mathbb{P}\left(\sigma_\lambda^2(x)\ge\gamma\right)+n^{-1}. \tag{4.23}$$

Similarly,

$$\begin{aligned}\mathbb{E}_\xi \mathbb{P}_n(yg(x) \le \delta) \le \mathbb{P}_n(yf(x) \le 2\delta) &+ \mathbb{P}_n\left(\sigma_\lambda^2(x) \ge \gamma\right) + \\ &+ \mathbb{P}_n \mathbb{P}_\xi\left(yg(x) \le \delta, yf(x) \ge 2\delta, \sigma_\lambda^2(x) \le \gamma\right).\end{aligned}$$

Applying Bernstein's inequality to the last term with the same choice of $N = 4(\frac{\gamma}{\delta^2})\ln n$, one has

$$\mathbb{E}_\xi \mathbb{P}_n(yg(x) \le \delta) \le \mathbb{P}_n(yf(x) \le 2\delta) + \mathbb{P}_n\left(\sigma_\lambda^2(x) \ge \gamma\right) + \frac{1}{n}. \tag{4.24}$$

Now, similarly to the proof of Theorem 2, we derive inequality (4.18). For any $\gamma \ge \delta \in (0,1], N = 4(\frac{\gamma}{\delta^2})\ln n$, for any $t > 0$ with probability at least $1 - e^{-t}$, the following holds:

$$\begin{aligned}&\phi\Big(\mathbb{E}_\xi \mathbb{P}(yg(x) \le \delta), \mathbb{E}_\xi \mathbb{P}_n(yg(x) \le \delta)\Big) \le \mathbb{E}_\xi \phi\Big(\mathbb{P}(yg(x) \le \delta), \mathbb{P}_n(yg(x) \le \delta)\Big) \le \\ &\le \frac{4}{n}\Big(4\frac{\gamma}{\delta^2}(\ln n)\ln G^*(2n, \mathcal{H}) + \ln(8n+4) + t\Big),\end{aligned} \tag{4.25}$$

where the fact that the function $\phi(a,b) = \frac{(a-b)^2}{a} I(a \ge b), a > 0$ is convex has been used; so, $\phi(\mathbb{E}_\xi a, \mathbb{E}_\xi b) \le \mathbb{E}_\xi \phi(a,b)$. The function $\phi(a,b)$ is increasing in a and decreasing in b; combining the last result with (4.23) and (4.24), one has

$$\begin{aligned}&\phi\Big(\mathbb{P}(yf(x) \le 0) - \mathbb{P}(\sigma_\lambda^2(x) \ge \gamma) - n^{-1}, \mathbb{P}_n(yf(x) \le 2\delta) + \mathbb{P}_n(\sigma_\lambda^2(x) \ge \gamma) + n^{-1}\Big) \\ &\le \frac{4}{n}(t + 4\frac{\gamma}{\delta^2}(\ln n)\ln G^*(2n, \mathcal{H}) + \ln(8n+4)).\end{aligned}$$

After solving this inequality for $\mathbb{P}(yf(x) \le 0)$, one has that, for any $\delta \in (0,1]$, any $1 \ge \gamma \ge \delta$, for any $t > 0$ with probability at least $1 - e^{-t}$, the following inequality holds

$$\begin{aligned}\mathbb{P}(yf(x) \le 0) &\le \mathbb{P}(\sigma^2(x) \ge \gamma) + \frac{1}{n} + \\ &+ \left(\left(\mathbb{P}_n(yf(x) \le 2\delta) + \mathbb{P}_n(\sigma^2(x) \ge \gamma) + \frac{1}{n} + U\right)^{\frac{1}{2}} + U^{\frac{1}{2}}\right)^2,\end{aligned} \tag{4.26}$$

where

$$U = \frac{1}{n}\left(t + 4\frac{\gamma}{\delta^2}(\ln n)\ln G^*(2n, \mathcal{H}) + \ln(8n+4)\right).$$

Next, in (4.23),(4.24) and (4.26), the term $\mathbb{P}(\sigma_\lambda^2(x) \ge \gamma)$ is related to the term $\mathbb{P}_n(\sigma_\lambda^2(x) \ge \gamma)$ that appears. In order to be able to do this, generate two independent sequences ξ_k^1 and ξ_k^2 as above and consider

$$\sigma_N^2(x) = \frac{1}{2N}\Sigma_{k=1}^N(\xi_k^2(x) - \xi_k^1(x))^2 = \frac{1}{N}\Sigma_{k=1}^N \xi_k(x),$$

where

$$\xi_k(x) = \frac{1}{2}\Big(\xi_k^1(x) - \xi_k^2(x)\Big)^2.$$

Notice that $\xi_k(x)$ are i.i.d. random variables and $\mathbb{E}_\xi \xi_k(x) = \sigma_\lambda^2(x)$. Since $\xi_k^1, \xi_k^2 \in \mathcal{H}(Z^n)$, then $|\xi_k^1(x) - \xi_k^2(x)| \leq 2$. The variance

$$\mathrm{Var}_\xi(\xi_1(x)) \leq \mathbb{E}_\xi \xi_1^2(x) \leq 2\mathbb{E}_\xi \xi_1(x) = 2\sigma_\lambda^2(x).$$

Bernstein's inequality implies that for any $c > 0$,

$$\mathbb{P}_\xi \left(\sigma_N^2(x) - \sigma^2(x) \leq 2\sqrt{\frac{\sigma_\lambda^2(x)\gamma}{c}} + 8\frac{\gamma}{3c} \right) \geq 1 - e^{(-\frac{N\gamma}{c})}$$

and

$$\mathbb{P}_\xi \left(\sigma_\lambda^2(x) - \sigma_N^2(x) \leq 2\sqrt{\frac{\sigma_\lambda^2(x)\gamma}{c}} + 8\frac{\gamma}{3c} \right) \geq 1 - e^{(-\frac{N\gamma}{c})}.$$

Let choose $c = 18$. If $\sigma_\lambda^2(x) \leq \gamma$, then with probability at least $1 - e^{-N\gamma/18}$, it follows from the first inequality that $\sigma_N^2(x) \leq 2\gamma$. On the other hand, if $\sigma_N^2(x) \leq 2\gamma$, then with probability at least $1 - e^{-N\gamma/18}$, it follows from the second inequality that $\sigma_\lambda^2(x) \leq 3\gamma$. Based on this,

$$\mathbb{P}_\xi \left(\sigma_N^2(x) \geq 2\gamma, \sigma_\lambda^2(x) \leq \gamma \right) \leq e^{(-\frac{N\gamma}{18})},$$

and

$$\mathbb{P}_\xi \left(\sigma_N^2(x) \leq 2\gamma, \sigma_\lambda^2(x) \geq 3\gamma \right) \leq e^{(-\frac{N\gamma}{18})}.$$

One can write

$$\begin{aligned} \mathbb{P}(\sigma_\lambda^2(x) \geq 3\gamma) &= \mathbb{E}_\xi \mathbb{P}\left(\sigma_\lambda^2(x) \geq 3\gamma, \sigma_N^2(x) \geq 2\gamma\right) + \mathbb{E}_\xi \mathbb{P}\left(\sigma_\lambda^2(x) \geq 3\gamma, \sigma_N^2(x) \leq 2\gamma\right) \\ &\leq \mathbb{E}_\xi \mathbb{P}\left(\sigma_N^2(x) \geq 2\gamma\right) + \mathbb{E}\mathbb{P}_\xi\left(\sigma_N^2(x) \leq 2\gamma, \sigma_\lambda^2(x) \geq 3\gamma\right) \end{aligned}$$

and

$$\mathbb{E}_\xi \mathbb{P}_n \left(\sigma_N^2(x) \geq 2\gamma\right) \leq \mathbb{P}_n \left(\sigma_\lambda^2(x) \geq \gamma\right) + \mathbb{E}_\xi \mathbb{P}_n \left(\sigma_N^2(x) \geq 2\gamma, \sigma_\lambda^2(x) \leq \gamma\right).$$

Setting $N = c\gamma^{-1} \ln n$, then

$$\mathbb{P}\left(\sigma_\lambda^2(x) \geq 3\gamma\right) \leq \mathbb{E}_\xi \mathbb{P}\left(\sigma_N^2(x) \geq 2\gamma\right) + \frac{1}{n}, \tag{4.27}$$

and

$$\mathbb{E}_\xi \mathbb{P}_n \left(\sigma_N^2(x) \geq 2\gamma\right) \leq \mathbb{P}_n \left(\sigma_\lambda^2(x) \geq \gamma\right) + \frac{1}{n}. \tag{4.28}$$

For any realization of $\xi_k^{j,1}, \xi_k^{j,2}$, the functions σ_N^2 belong to the class

$$\mathcal{F}_N(Z^n) = \left\{ \frac{1}{2N} \sum_{k=1}^{N} (h_k^{j,1} - h_k^{j,2})^2 : h_k^{j,1}, h_k^{j,2} \in \mathcal{H}(Z^n) \right\},$$

where $\mathcal{H}(Z^n)$ is defined as the random-base function class in the general problem.

Now, consider the random collection of sets

$$\mathcal{C}(Z^n) = \left\{C : C = \left\{x \in \mathcal{X} : \{\sigma_N^2(x) \geq \gamma\}\right\}, \sigma_N^2 \in \mathcal{F}_N(Z^n), \gamma \in (0,1]\right\}.$$

In order to bound $G'(n) = \mathbb{E}^n \Delta_{\mathcal{C}(Z^n)}(Z^n)$ take into account that if $\mathcal{H}(Z^n)$ is a random-base class of finite cardinality, then $\mathrm{card}\mathcal{F}_N(Z^n) \leq G^*(n,\mathcal{H})^{2N}$. In the case of the base-random class $\mathcal{H}(Z^n)$ being a collection of indicators, similarly to the proof of Lemma 7, one can count the maximum number of different representations of $\{X_1,\ldots,X_n\}$ by

$$\mathcal{C}(Z^n,\gamma) = \left\{C : C = \left\{x \in \mathcal{X} : \{\sigma_N^2(x) \geq \gamma\}\right\}, \sigma_N^2 \in \mathcal{F}_N(Z^n)\right\}$$

for a fixed $\gamma \in (0,1]$. It is bounded by $\left(\frac{ne}{V}\right)^{2N}$. Then varying γ over the ordered discrete set $\{1, \sigma_N^2(X_{i_1}), \sigma_N^2(X_{i_2}), \ldots, \sigma_N^2(X_{i_n})\}$ for a fixed $\sigma_N^2 \in \mathcal{F}_N(Z^n)$, one can see that $G'(n) \leq (n+1)G^*(n,\mathcal{H})^{2N}$. Now, we apply Theorem 1 for the random collection of sets $\mathcal{C}$, for $N = 18\gamma^{-1}\ln n$. Then for any $t > 0$ with probability at least $1 - e^{-t}$ for any sample Z^n, the following holds

$$\begin{aligned}\phi\Big(\mathbb{E}_\xi \mathbb{P}(\sigma_N^2(x) \geq \gamma), \mathbb{E}_\xi \mathbb{P}_n(\sigma_N^2(x) \geq \gamma)\Big) &\leq \mathbb{E}_\xi \phi\Big(\mathbb{P}(\sigma_N^2(x) \geq \gamma), \mathbb{P}_n(\sigma_N^2(x) \geq \gamma)\Big) \leq \\ &\leq \frac{4}{n}\Big(2N \ln G^*(2n,\mathcal{H}) + \ln(8n+4) + t\Big).\end{aligned}$$

Here, the monotonic property of $\phi(a,b)$ is used together with (4.27) and (4.28), in order to obtain the following bound under the above conditions:

$$\begin{aligned}&\phi\Big(\mathbb{P}(\sigma_\lambda^2(x) \geq 3\gamma) - \frac{1}{n}, \mathbb{P}_n(\sigma_\lambda^2(x) \geq \gamma) + \frac{1}{n}\Big) \leq \\ &\leq \frac{4}{n}\Big(36\gamma^{-1}(\ln n)\ln G^*(2n,\mathcal{H}) + \ln(8n+4) + t\Big),\end{aligned}$$

Solving the above inequality for $\mathbb{P}(\sigma_\lambda^2(x) \geq \gamma)$, we obtain

$$\mathbb{P}(\sigma_\lambda^2(x) \geq \gamma) \leq \frac{1}{n} + \left(W^{\frac{1}{2}} + \left(W + \mathbb{P}_n(\sigma_\lambda^2(x) \geq \gamma/3) + \frac{1}{n}\right)^{\frac{1}{2}}\right)^2,$$

where

$$W = \frac{1}{n}\left(t + \frac{108}{\gamma}(\ln n)\ln G^*(2n,\mathcal{H}) + \ln(8n+4)\right).$$

Combining the above inequality with the inequality (4.26) and using the inequalities $(a+b)^2 \leq 2a^2 + 2b^2$ and $\frac{1}{\gamma} \leq \frac{\gamma}{\delta^2}$ for $\gamma \geq \delta$, one has that, for any $\delta \in (0,1]$ and any $\gamma \in (0,1], \gamma \geq \delta$, for all $t > 0$ with probability at least $1 - e^{-t}$, for any random sample Z^n, for any $\lambda \in \mathcal{P}(\mathcal{H}(Z^n))$ and $f(x) = \int h(x)\lambda(dh)$,

$$\mathbb{P}(yf(x) \leq 0) \leq 2\mathbb{P}_n(yf(x) \leq 2\delta) + 2\mathbb{P}_n(\sigma_\lambda^2(x) \geq \gamma) + 2\mathbb{P}_n(\sigma_\lambda^2(x) \geq \gamma/3) +$$

$$+\frac{8t}{n} + \frac{8\ln(8n+4)}{n} + \frac{6}{n} + \frac{448\gamma(\ln n)\ln G^*(2n,\mathcal{H})}{n\delta^2}.$$

Observe that $\mathbb{P}_n(\sigma^2_\lambda(x) \geq \gamma) \leq \mathbb{P}_n(\sigma^2_\lambda(x) \geq \gamma/3)$. Rewrite

$$\begin{aligned}\mathbb{P}(yf(x) \leq 0) &\leq 2\mathbb{P}_n(yf(x) \leq 2\delta) + 4\mathbb{P}_n\left(\sigma^2_\lambda(x) \geq \frac{\gamma}{3}\right) + \\ &+ \frac{448\gamma(\ln n)\ln G^*(2n,\mathcal{H})}{n\delta^2} + \frac{8t}{n} + \frac{8\ln(8n+4)}{n} + \frac{6}{n}.\end{aligned}$$

Next, the bound is made uniform with respect to $\gamma \in (0,1]$ and $\delta \in (0,1]$. First, one makes the bound uniform when $\gamma \in \Delta = \{2^{-k}, k \in \mathbb{Z}_+\}$, and $\delta \in \Delta$. Apply the above inequality for fixed $\delta \leq \gamma \in \Delta$ by replacing t by $t' + \ln\frac{2\gamma}{\delta}$ and, hence, e^{-t} replaced by $e^{-t'} = e^{-t}\frac{\delta}{2\gamma}$, where δ and $\gamma \in \Delta = \{2^{-k} : k \geq 0\}$.

$$\begin{aligned}\mathrm{P}\Big[\bigcap_{\delta,\gamma}\Big\{ &\mathbb{P}(yf(x) \leq 0) \leq \left(2\mathbb{P}_n(yf(x) \leq 2\delta) + 4\mathbb{P}_n\left(\sigma^2_\lambda(x) \geq \frac{\gamma}{3}\right) + \frac{6}{n} + \right. \\ &+ \left.\frac{8}{n}\left(t + \ln\frac{2\gamma}{\delta} + \ln(8n+4) + \frac{56\gamma}{\delta^2}(\ln n)\ln G^*(2n,\mathcal{H})\right)\right)\Big\}\Big] \geq \\ &\geq 1 - \sum_{l\in\mathbb{Z}_+}\frac{2^{-l}}{2}.e^{-t} \geq 1 - e^{-t},\end{aligned}$$

where is used $\sum_{l\in\mathbb{Z}_+} 2^{-l} < 2$. Then the union bound should be applied in the whole range of $\delta, \gamma \in (0,1]$.

For any $t > 0$ with probability at least $1 - e^{-t}$, for any $\lambda \in \mathcal{P}(\mathcal{H})$ and $f(x) = \int h(x)\lambda(dh)$,

$$\mathbb{P}(yf(x) \leq 0) \leq \inf_{0<\delta\leq\gamma\leq 1}\Big(2\mathbb{P}_n(yf(x) \leq 2\delta) + 4\mathbb{P}_n(\sigma^2_\lambda(x) \geq \frac{\gamma}{3}) +$$

$$+\frac{8}{n}\left(t + \ln\frac{2\gamma}{\delta} + \ln(8n+4) + \frac{56\gamma}{\delta^2}(\ln n)\ln G^*(2n,\mathcal{H})\right) + \frac{6}{n}\Big).$$

□

Now the **proof of Theorem 5** regarding cluster-variance bound is given. Let us fix

$$\alpha_1, \ldots, \alpha_p, \sum_{i=1}^p \alpha_i \leq 1, \alpha_i > 0$$

used for the weights of the clusters in

$$c = (\alpha_1, \ldots, \alpha_p, \lambda^1, \ldots, \lambda^p),\ \lambda = \sum_{i=1}^p \alpha_i\lambda^i,\ \lambda^i \in \mathcal{P}(\mathcal{H}(Z^n)).$$

Generate functions from each cluster independently from each other and independently of the data and take their sum to approximate $f(x) = \int h(x)\lambda(dh) = \sum_{i=1}^T \lambda_i h_i(x)$. Given $N \geq 1$, generate independent $\xi^j_k(x), k \leq N, j \leq p$, where for each j, $\xi^j_k(x)$'s are i.i.d. and have the distribution $\mathbb{P}_\xi(\xi^j_k(x) = h_i(x)) = \lambda^j_i, i \leq T$. Consider a function that plays role of a random approximation of f

$$g(x) = \frac{1}{N}\sum_{j=1}^p \alpha_j \sum_{k=1}^N \xi^j_k(x) = \frac{1}{N}\sum_{k=1}^N g_k(x),$$

where $g_k(x) = \sum_{j=1}^{p} \alpha_j \xi_k^j(x)$.

For a fixed $x \in \mathcal{X}$ and $k \leq N$, the expectation of g_k with respect to the distribution $\mathbb{P}_\xi = \mathbb{P}_{\xi^1} \times, \ldots, \times \mathbb{P}_{\xi^p}$ is

$$\mathbb{E}_\xi(g_k(x)) = \Sigma_{j=1}^{p} \alpha_j \mathbb{E}_\xi(\xi_k^j(x)) = f(x);$$

its variance is

$$\mathrm{Var}_\xi(g_k(x)) = \sum_{j=1}^{p} \mathrm{Var}_\xi(\xi_k^j(x)) = \sum_{j=1}^{p} \alpha_j^2 \sigma_{\lambda^j}^2(x) = \sigma^2(c;x).$$

Then

$$\begin{aligned} \mathbb{P}(yf(x) \leq 0) &\leq \mathbb{E}_\xi \mathbb{P}(yg(x) \leq \delta) + \mathbb{P}\left(\sigma^2(c;x) \geq \gamma\right) + \\ &+ \mathbb{E}\mathbb{P}_\xi\left(yg(x) \geq \delta, yf(x) \leq 0, \sigma^2(c;x) \leq \gamma\right). \end{aligned}$$

Using Bernstein's inequality, $\gamma \geq \delta > 0$, $|g_k(x)| \leq 1$ and taking $N = \lceil 2 + 4/3 \rceil (\frac{\gamma}{\delta^2}) \ln n = 4\frac{\gamma}{\delta^2} \ln n$ will make the last term negligible. Thus,

$$\mathbb{P}(yf(x) \leq 0) \leq \mathbb{E}_\xi \mathbb{P}(yg(x) \leq \delta) + \mathbb{P}\left(\sigma^2(c;x) \geq \gamma\right) + \frac{1}{n}. \tag{4.29}$$

Also,

$$\begin{aligned} \mathbb{E}_\xi \mathbb{P}_n(yg(x) \leq \delta) &\leq \mathbb{P}_n(yf(x) \leq 2\delta) + \mathbb{P}_n\left(\sigma^2(c;x) \geq \gamma\right) + \\ &+ \mathbb{P}_n \mathbb{P}_\xi\left(yg(x) \geq \delta, yf(x) \leq 2\delta, \sigma^2(c;x) \leq \gamma\right). \end{aligned}$$

Applying Bernstein's inequality to the last term with the same choice of $N = 4\frac{\gamma}{\delta^2} \ln n$, it follows that

$$\mathbb{E}_\xi \mathbb{P}_n(yg(x) \leq \delta) \leq \mathbb{P}_n(yf(x) \leq 2\delta) + \mathbb{P}_n\left(\sigma^2(c;x) \geq \gamma\right) + \frac{1}{n}. \tag{4.30}$$

Now, consider the random collection of level sets

$$\mathcal{C}(Z^n) = \{C : C = \{(x,y) : yg(x) \leq \delta, (x,y) \in \mathcal{X} \times \mathcal{Y}\}, g \in F_N(Z^n), \delta \in [-1,1]\},$$

where

$$F_N(Z^n) = \left\{ \frac{1}{N} \Sigma_{i=1}^{N} g_i, g_i \in G(\alpha_1, \ldots, \alpha_p)_{[Z^n]} \right\}$$

and

$$G(\alpha_1, \ldots, \alpha_p)_{[Z^n]} = \left\{ g_k(x) = \sum_{j=1}^{p} \alpha_j \xi_k^j(x), \xi_k^j \in \mathcal{H}(Z^n) \right\},$$

where $\mathcal{H}(Z^n)$ is the random-base class of functions, defined in the general problem.

Similarly to the proof of Theorem 2, for fixed $g \in F_N(Z^n)$, we have

$$\mathrm{card}\left\{C \bigcap Z_1, \ldots, Z_n\right\} \leq (n+1)$$

and

$$\mathrm{card} F_N(Z^n) \leq G^*(n, \mathcal{H})^{Np}.$$

Therefore, $G'(n) = \mathbb{E}^n \Delta_{\mathcal{C}}(Z^n) \leq (n+1) G^*(n, \mathcal{H})^{Np}$. Apply Inequality (3.4) from Theorem 1 for random the collection of sets $\mathcal{C}$. Then, with probability at least $1 - e^{-t}$

$$\frac{(\mathbb{P}_{x,y}(yg(x) \leq \delta) - \mathbb{P}_n(yg(x) \leq \delta))^2}{\mathbb{P}_{x,y}(yg(x) \leq \delta)} \leq \frac{4}{n}(t + Np \ln G^*(2n, \mathcal{H}) + \ln(8n+4)).$$

The function $\phi(a,b) = \frac{(a-b)^2}{a} I(a \geq b), a > 0$ is convex, so $\phi(\mathbb{E}_\xi a, \mathbb{E}_\xi b) \leq \mathbb{E}_\xi \phi(a,b)$

$$\frac{(\mathbb{E}_\xi \mathbb{P}_{x,y}(yg(x) \leq \delta) - \mathbb{E}_\xi \mathbb{P}_n(yg(x) \leq \delta))^2}{\mathbb{E}_\xi \mathbb{P}_{x,y}(yg(x) \leq \delta)} \leq \frac{4}{n}(t + Np \ln G^*(2n, \mathcal{H}) + \ln(8n+4)).$$

The function $\phi(a,b)$ is increasing in a and decreasing in b and combined with the last result with (4.29) and (4.30) (recall that $N = 4(\frac{\gamma}{\delta^2}) \ln n$)

$$\phi\Big(\mathbb{P}(yf(x) \leq 0) - \mathbb{P}(\sigma^2(c;x) \geq \gamma) - \frac{1}{n}, \mathbb{P}_n(yf(x) \leq 2\delta) + \mathbb{P}_n(\sigma^2(c;x) \geq \gamma) + \frac{1}{n}\Big) \leq$$
$$\leq \frac{4}{n}\left(t + 4p\frac{\gamma}{\delta^2}(\ln n) \ln G^*(2n, \mathcal{H}) + \ln(8n+4)\right).$$

After solving this inequality for $\mathbb{P}(yf(x) \leq 0)$, one can get that, for any $\gamma, \delta \in (0,1]$, $\gamma \geq \delta$, and $\alpha^1, \ldots, \alpha^p, \sum \alpha_i \leq 1, \alpha_i > 0$ for any $t > 0$ with probability at least $1 - e^{-t}$,

$$\mathbb{P}(yf(x) \leq 0) \leq \mathbb{P}\left(\sigma^2(c;x) \geq \gamma\right) +$$
$$+ \left(\left(\mathbb{P}_n(yf(x) \leq 2\delta) + \mathbb{P}_n(\sigma^2(c;x) \geq \gamma) + \frac{1}{n} + U\right)^{\frac{1}{2}} + U^{\frac{1}{2}} + \frac{1}{n}\right)^2, \tag{4.31}$$

where

$$U = \frac{1}{n}\left(t + 4\frac{p\gamma}{\delta^2}(\ln n) \ln G^*(2n, \mathcal{H}) + \ln(8n+4)\right),$$

$c \in \mathcal{C}^p(\lambda), \lambda = \sum_{j=1}^p \alpha_j \lambda^j$, $\lambda_j \in \mathcal{P}(\mathcal{H})$.

Now, $\mathbb{P}(\sigma^2(c;x) \geq \gamma)$ has to be estimated. Generate two independent random sequences of functions $\xi_k^{j,1}(x)$ and $\xi_k^{j,2}(x)$, $j = 1, \ldots p$, $k = 1, \ldots, N$ as before ($\mathbb{P}_\xi(\xi_k^{j,1}(x) = h_i(x)) = \lambda_i^j$, $\mathbb{P}_\xi(\xi_k^{j,2}(x) = h_i(x)) = \lambda_i^j$) and consider

$$\sigma_N^2(c;x) = \frac{1}{2N}\Sigma_{k=1}^N \left(\Sigma_{j=1}^p \alpha^j (\xi_k^{j,2}(x) - \xi_k^{j,1}(x))\right)^2 = \frac{1}{N}\Sigma_{k=1}^N \xi_k(x),$$

where

$$\xi_k(x) = \frac{1}{2}\left(\Sigma_{j=1}^p \alpha^j (\xi_k^{j,1}(x) - \xi_k^{j,2}(x))\right)^2. \tag{4.32}$$

Then $\xi_k(x)$ are i.i.d. random variables and $\mathbb{E}_\xi \xi_k(x) = \sigma^2(c;x)$. Since $\xi_k^{j,1}, \xi_k^{j,2} \in \mathcal{H}$, then $|\xi_k^{j,1}(x) - \xi_k^{j,2}(x)| \leq 2$. The variance satisfies the following inequality

$$\mathrm{Var}_\xi(\xi_1(x)) \leq \mathbb{E}_\xi \xi_1^2(x) \leq 2\mathbb{E}_\xi \xi_1(x) = 2\sigma^2(c;x).$$

Bernstein's inequality implies that

$$\mathbb{P}_\xi\Big(\sigma_N^2(c;x)-\sigma^2(c;x)\le 2\sqrt{\frac{\sigma^2(c;x)}{K}}+8\frac{\gamma}{3K}\Big)\ge 1-e^{(-\frac{N\gamma}{K})}$$

and

$$\mathbb{P}_\xi\Big(\sigma^2(c;x)-\sigma_N^2(c;x)\le 2\sqrt{\frac{\sigma^2(c;x)}{K}}+8\frac{\gamma}{3K}\Big)\ge 1-e^{(-\frac{N\gamma}{K})}.$$

Based on this, for large enough $K>0$ ($K=18$ is sufficient),

$$\mathbb{P}_\xi\left(\sigma_N^2(c;x)\ge 2\gamma,\sigma^2(c;x)\le\gamma\right)\le e^{(-\frac{N\gamma}{K})},$$

and

$$\mathbb{P}_\xi\left(\sigma_N^2(c;x)\le 2\gamma,\sigma^2(c;x)\ge 3\gamma\right)\le e^{(-\frac{N\gamma}{K})}.$$

One can write

$$\mathbb{P}\left(\sigma^2(c;x)\ge 3\gamma\right)\le\mathbb{E}_\xi\mathbb{P}\left(\sigma_N^2(c;x)\ge 2\gamma\right)+\mathbb{E}\mathbb{P}_\xi\left(\sigma_N^2(c;x)\le 2\gamma,\sigma^2(c;x)\ge 3\gamma\right),$$

and

$$\mathbb{E}_\xi\mathbb{P}_n\left(\sigma_N^2(c;x)\ge 2\gamma\right)\le\mathbb{P}_n\left(\sigma_N^2(c;x)\ge\gamma\right)+\mathbb{P}_n\mathbb{P}_\xi\left(\sigma_N^2(c;x)\ge 2\gamma,\sigma^2(c;x)\le\gamma\right).$$

Choose $N=K\gamma^{-1}\ln n$; then

$$\mathbb{P}\left(\sigma^2(c;x)\ge 3\gamma\right)\le\mathbb{E}_\xi\mathbb{P}\left(\sigma_N^2(c;x)\ge 2\gamma\right)+\frac{1}{n},\tag{4.33}$$

and

$$\mathbb{E}_\xi\mathbb{P}_n\left(\sigma_N^2(c;x)\ge 2\gamma\right)\le\mathbb{P}_n\left(\sigma_N^2(c;x)\ge\gamma\right)+\frac{1}{n}.\tag{4.34}$$

Now, consider the random collection of sets

$$\mathcal{C}_{Z^n}=\left\{C:C=\left\{x:\sigma_N^2(c;x)\ge 2\gamma\right\},\sigma_N^2(c;x)\in\mathcal{F}_N(Z^n),\gamma\in(0,1]\right\},$$

where

$$\mathcal{F}_N(Z^n)=\left\{\frac{1}{2N}\Sigma_{k=1}^N\Big(\Sigma_{j=1}^p\alpha^j(h_k^{j,1}-h_k^{j,2})\Big)^2,h_k^{j,1},h_k^{j,2}\in\mathcal{H}(Z^n)\right\}.$$

For any $\{x_1,\ldots,x_n\}$ and a fixed $\sigma_N^2(c;.)\in\mathcal{F}_N(Z^n)$, it follows that

$\operatorname{card}\{C\bigcap\{X_1,\ldots,X_n\}\}\le(n+1)$ and $\operatorname{card}\mathcal{F}_N(Z^n)\le G^*(n,\mathcal{H})^{2Np}$

if the base-random class $\mathcal{H}(Z^n)$ is of finite cardinality. Therefore, $G'_{\mathcal{C}}(n)=\mathbb{E}^n\Delta_{\mathcal{C}_{Z^n}}(Z^n)\le(n+1)G^*(n,\mathcal{H})^{2Np}$. The case of $\mathcal{H}(Z^n)$ being a collection of indicators as in the general problem is similar and dealt with in the previous proofs of the theorems.

The rest of the arguments are similar to the proof of the Theorem 4. Apply the inequality (3.4) from Theorem 1 for random collection of sets $\mathcal{C}_{Z^n}$, and based on convexity of $\phi(a,b)$, one has that for $\gamma\in(0,1]$, $\alpha_1,\ldots,\alpha_p,\sum_{j=1}^p\alpha_j\le 1,\alpha_j>0$ and for any $t>0$ with probability at least $1-e^{-t}$

$$\frac{(\mathbb{E}_\xi \mathbb{P}\left(\sigma_N^2(c;x) \geq \gamma\right) - \mathbb{E}_\xi \mathbb{P}_n\left(\sigma_N^2(c;x) \geq \gamma\right))^2}{\mathbb{E}_\xi \mathbb{P}\left(\sigma_N^2(c;x) \geq \gamma\right)} \leq \frac{2}{n}(t + 2Np \ln G^*(2n, \mathcal{H}) + \ln(8n+4)),$$

for any $\lambda^j \in \mathcal{P}(\mathcal{H}(Z^n)), j = 1, \ldots, p$. Combining the last result ($\phi(a,b)$ is increasing in a and decreasing in b) with (4.33) and (4.34) (recall that $N = 18\gamma^{-1}\ln n$),

$$\phi\Big(\mathbb{P}\left(\sigma^2(c;x) \geq 3\gamma\right) - n^{-1}, \mathbb{P}_n\left(\sigma^2(c;x) \geq \gamma\right) + n^{-1}\Big) \leq$$
$$\leq \frac{4}{n}\left(t + 36p\frac{\gamma}{\delta^2}(\ln n)\ln G^*(2n,\mathcal{H}) + \ln(8n+4)\right).$$

Solving the above inequality for $\mathbb{P}(\sigma_\lambda^2(x) \geq \gamma)$, then with probability at least $1 - e^{-t}$

$$\mathbb{P}\left(\sigma_\lambda^2(x) \geq \gamma\right) \leq \frac{1}{n} + \left(W^{\frac{1}{2}} + \left(W + \mathbb{P}_n\left(\sigma_\lambda^2(x) \geq \gamma/3\right) + \frac{1}{n}\right)^{\frac{1}{2}}\right)^2,$$

where

$$W = \frac{1}{n}\left(t + \frac{108p}{\gamma}(\ln n)\ln G^*(2n,\mathcal{H}) + \ln(8n+4)\right).$$

Finally, combining this with (4.31) and using the inequalities $(a+b)^2 \leq 2a^2 + 2b^2$ and $\frac{1}{\gamma} \leq \frac{\gamma}{\delta^2}$ for $\gamma \geq \delta$, one obtains: for any $\delta \in (0,1]$ and any $\gamma \in (0,1], \gamma \geq \delta$, for all $t > 0$ with probability at least $1 - e^{-t}$, for any random sample Z^n, for any $\lambda \in \mathcal{P}(\mathcal{H}(Z^n))$ and $f(x) = \int h(x)\lambda(dh)$,

$$\mathbb{P}(yf(x) \leq 0) \leq 2\mathbb{P}_n(yf(x) \leq 2\delta) + 2\mathbb{P}_n\left(\sigma_\lambda^2(x) \geq \gamma\right) + 2\mathbb{P}_n\left(\sigma_\lambda^2(x) \geq \gamma/3\right) +$$

$$+\frac{8t}{n} + \frac{8\ln(8n+4)}{n} + \frac{6}{n} + \frac{448p\gamma(\ln n)\ln G^*(2n,\mathcal{H})}{n\delta^2}.$$

Observe that $\mathbb{P}_n\left(\sigma_\lambda^2(x) \geq \gamma\right) \leq \mathbb{P}_n\left(\sigma_\lambda^2(x) \geq \gamma/3\right)$. Then, rewrite

$$\begin{aligned}\mathbb{P}(yf(x) \leq 0) \quad &\leq \quad 2\mathbb{P}_n(yf(x) \leq 2\delta) + 4\mathbb{P}_n\left(\sigma_\lambda^2(x) \geq \frac{\gamma}{3}\right) + \\ &+ \quad \frac{448p\gamma(\ln n)\ln G^*(2n,\mathcal{H})}{n\delta^2} + \frac{8t}{n} + \frac{8\ln(8n+4)}{n} + \frac{6}{n}.\end{aligned}$$

The next step is to make this bound uniform with respect to $\alpha_j > 0, j = 1, \ldots, p, \sum_{j=1}^p \alpha_j \leq 1$.

First, consider simply $\alpha_i \in \Delta = \{2^{-j}, j = 1, 2, \ldots\}$. The case of $\alpha_i = 1$ is proven in the previous Theorem 4 for total variance. Let $\alpha_j = 2^{-l_j}$. Redefine cluster $c(l_1, \ldots, l_p) := c(\alpha_1, \ldots, \alpha_p, \lambda^1, \ldots, \lambda^p)$. Then consider the event

$$A_{c(l_1,\ldots,l_p)} = \Big\{\forall f \in \mathcal{F}_d : \mathbb{P}(yf(x) \leq 0) \leq 2\mathbb{P}_n(yf(x) \leq 2\delta) + 4\mathbb{P}_n\left(\sigma_\lambda^2(x) \geq \frac{\gamma}{3}\right) +$$
$$+\frac{448p\gamma(\ln n)\ln G^*(2n,\mathcal{H})}{n\delta^2} + \frac{8t}{n} + \frac{8\ln(8n+4)}{n} + \frac{6}{n}\Big\},$$

that holds with probability $1-e^{-t}$. Make change of variables t by $t+2\sum_{j=1}^{p}\ln l_j + p\ln 4$ in the last bound. With this choice, the event $A_{c(l_1,\ldots,l_p)}$ can be rewritten as

$$A_{c(l_1,\ldots,l_p)} = \Big\{\forall f\in\mathcal{F}_d : \mathbb{P}(yf(x)\le 0) \le 2\mathbb{P}_n(yf(x)\le 2\delta) + 4\mathbb{P}_n\Big(\sigma^2_\lambda(x)\ge\frac{\gamma}{3}\Big) + \\ + \frac{448p\gamma(\ln n)\ln G^*(2n,\mathcal{H})}{n\delta^2} + \frac{8t}{n} + \frac{16\sum_{j=1}^{p}\ln l_j}{n} + \frac{8p\ln 4}{n} + \frac{8\ln(8n+4)}{n} + \frac{6}{n}\Big\},$$

which holds with probability at least

$$\mathrm{P}(A_{c(l_1,\ldots,l_p)}) \ge 1 - \prod\frac{1}{l_j^2}e^{-t}4^{-p}.$$

This implies the probability of the intersection

$$\mathrm{P}\left(\bigcap_{l_1,\ldots,l_p} A_{c(l_1,\ldots,l_p)}\right) \ge 1 - \sum_{l_1,\ldots,l_p\in\mathbb{N}}\prod\frac{1}{l_j^2}e^{-t}4^{-p} =$$

$$= 1-4^{-p}e^{-t}\Big(\sum_{l_i\in\mathbb{N}}\frac{1}{l_i^2}\Big)^p = 1-4^{-p}e^{-t}(1+\pi^2/6)^p \ge 1-e^{-t} \ge 1-e^{-t}$$

and $\sum\ln l_j = \sum\ln(|\log_2\alpha_j|)$. For fixed $p\ge 1$ and $1\ge\gamma\ge\delta>0$ and $\forall t>0$ with probability at least $1-e^{-t}$, the following is true for any $\alpha_1,\ldots,\alpha_p\in\Delta$, $\sum_{i=1}^{p}\alpha_i\le 1$, $\Delta=\{2^{-j}, j=0,1,\ldots\}$:

$$\mathbb{P}(yf(x)\le 0) \le \Big(2\mathbb{P}_n(yf(x)\le 2\delta) + 4\mathbb{P}_n\left(\sigma^2(c;x)\ge\gamma/3\right) + \\ + \frac{448p\gamma(\ln n)\ln G^*(2n,\mathcal{H})}{n\delta^2} + \frac{8t}{n} + \frac{16\sum_{j=1}^{p}\ln(|\log_2\alpha_j|)}{n} + \frac{8p\ln 4}{n} + \frac{8\ln(8n+4)}{n} + \frac{6}{n}\Big).$$

Next, for a fixed $m\in\mathbb{N}$, consider the following discretization of $\alpha_j = t_j m^{-s}$, for a fixed a priori $s\in\mathbb{Z}_+$, and $t_j\in\{1,2,3,\ldots,m^s\}$. Therefore $s+\log_m\alpha_j\ge 0$.

For any $\alpha_j = t_j m^{-s}$ there is $l_j\in\mathbb{Z}_+$, such that

$$m^{-l_j-1} < \alpha_j = t_j.m^{-s} \le m^{-l_j}.$$

That is $s-l_j-1<\log_m t_j\le s-l_j$, $l_j\le s$.

This time we make the change of variables $t' = t+\sum_{j=1}^{p}2\ln(s+\log_m\alpha_j+1)+2p\ln 2$ and apply the bound for that t'.

Then $e^{-t'} = e^{-t}\sum_{j=1}^{p}\frac{1}{(\log_m t_j+1)^2}4^{-p} \le e^{-t}\sum_{j=1}^{p}\frac{1}{(s-l_j+1)^2}4^{-p}$. Applying union bound trick as before, shows that for any $t>0$, with probability at least

$$1-e^{-t}4^{-p}\Big(\sum_{j=1}^{s}\frac{1}{(s-l_j+1)^2}\Big)^p > 1-e^{-t},$$

for any $\alpha_j = t_j m^{-s}$, $t_j\in\{1,2,3,\ldots,m^s\}$, $j=1,\ldots,p$, the following bound holds:

$$\begin{aligned}\mathbb{P}(yf(x)\le 0) \quad &\le \quad \Big(2\mathbb{P}_n(yf(x)\le 2\delta) + 4\mathbb{P}_n\left(\sigma^2(c;x)\ge\gamma/3\right) + \\ &+ \quad \frac{448p\gamma(\ln n)\ln G^*(2n,\mathcal{H})}{n\delta^2} + \frac{8t}{n} + \frac{16\sum_{j=1}^{p}\ln(s+\log_m\alpha_j+1)}{n} + \\ &+ \quad \frac{16p\ln 2}{n} + \frac{8\ln(8n+4)}{n} + \frac{6}{n}\Big).\end{aligned}$$

In order to make the bound uniform for all $s, p \geq 1$ and $1 \geq \gamma \geq \delta > 0$, apply the above inequality for fixed $p \in \mathbb{N}, \delta \leq \gamma \in \Delta = \{2^{-k} : k \geq 1\}$ by replacing t by $t' = t + \ln \frac{s^2 p^2 \pi^4 \gamma}{\delta 18}$ and hence replacing e^{-t} by $e^{-t'} = e^{-t} \frac{\delta 18}{s^2 p^2 \pi^4 \gamma}$, where δ and $\gamma \in \Delta = \{2^{-k} : k \geq 1\}$.

$$
\begin{aligned}
P\Big[\bigcap_{\delta,\gamma,p}\Big\{ \quad & \mathbb{P}(yf(x) \leq 0) \leq \Big(2\mathbb{P}_n(yf(x) \leq 2\delta) + 4\mathbb{P}_n\left(\sigma^2(c;x) \geq \gamma/3\right) + \\
& + \frac{448p\gamma(\ln n)\ln G^*(2n,\mathcal{H})}{n\delta^2} + \frac{8t}{n} + \frac{8\ln\frac{s^2p^2\pi^4\gamma}{18\delta}}{n} + \frac{16\sum_{j=1}^{p}\ln(s+\log_m \alpha_j + 1)}{n} + \\
& + \frac{16p\ln 2}{n} + \frac{8\ln(8n+4)}{n} + \frac{6}{n}\Big)\Big\}\Big] \\
& \geq 1 - \sum_{l\in\mathbb{Z}_+} \frac{2^{-l}36}{2\pi^4} \cdot (\sum_k \frac{1}{k^2})^2 e^{-t} \geq 1 - e^{-t},
\end{aligned}
$$

where we have applied $\sum_{k\in\mathbb{Z}_+} \frac{1}{k^2} \leq \frac{\pi^2}{6}$ and $\sum_{l\in\mathbb{Z}_+} 2^{-l} \leq 2$.

Finally, $\forall t > 0$ with probability at least $1 - e^{-t}$, the following is true for all $s \in \mathbb{N}, \alpha_1, \ldots, \alpha_p \in \Delta = \left\{t_j m^{-s}, 0 < t_j \leq m^s, t_j \in \mathbb{N}\right\}$, $p \in \mathbb{N}$ and $1 \geq \gamma \geq \delta > 0$

$$
\begin{aligned}
\mathbb{P}(yf(x) \leq 0) \quad \leq \quad & \Big(2\mathbb{P}_n(yf(x) \leq 2\delta) + 4\mathbb{P}_n\left(\sigma^2(c;x) \geq \gamma/3\right) + \\
& + \frac{1}{n}\Big(\frac{448p\gamma(\ln n)\ln G^*(2n,\mathcal{H})}{\delta^2} + 8t + 8\ln\frac{s^2p^2\pi^4\gamma}{18\delta} + \\
& + 16\sum_{j=1}^{p}\ln(s+\log_m \alpha_j + 1) + 16p\ln 2 + 8\ln(8n+4) + 6\Big)\Big).
\end{aligned}
$$

From here, by replacing s with $\lceil \log_m(\frac{1}{\alpha_c^*}) \rceil$ in the above inequality, the result (3.12) follows.

□

5. Conclusions

Here, we showed unified data-dependent generalization bounds for classifiers from *random* convex hulls in the setting of the *general problem* defined above. Such classifiers are generated, for example, by broad classes of algorithms such as SVMs, RBF networks and boosting. The bounds involve the individual complexities of the classifiers introduced by Koltchinskii and Panchenko (2004), such as sparsity of weights and weighted variance over clusters. This was achieved by proving a version of Vapnik's inequality applied to random classes, that is, a bound for relative deviations of frequencies from probabilities for random classes of events (Theorem 1). The results show how various algorithms fit in a single, *general class*. Also, it was indicated that algorithms controlling the individual complexities of the classifiers can produce classifiers with good generalization ability (see Koltchinskii et al. (2003a); Koltchinskii et al. (2003b); Andonova (2004) for some experimental results in the setting of various boosting algorithms). Experimental investigations of the above complexities in the setting of the *general problem* are desirable.

Acknowledgments

The author expresses special thanks to Dmitry Panchenko for introducing me to the area of statistical learning and for his valuable advice and input during the exploration of the above problems. The author also expresses thanks to Vladimir Koltchinskii for valuable communication and input. Valuable remarks and suggestions about the paper and references were given by Sayan Mukherjee, the reviewers and the editor.

References

Andonova, S. *Theoretical and experimental analysis of the generalization ability of some statistical learning algorithms.* PhD thesis, Department of Mathematics and Statistics, Boston University, Boston, MA, 2004.

Anthony, M., Shawe-Taylor, J. A result of Vapnik with applications. *Discrete Applied Mathematics,* 47(3): 207–217, 1993.

Bartlett, P., Shawe-Taylor, J. Generalization performance of support vector machines and other pattern classifiers. *Advances in Kernel Methods. Support Vector Learning. Schölkopf, Burges and Smola (Eds.),* 1, The MIT Press, Cambridge, 1999.

Borowkow, A. A. *Wahrscheinlichkeits–theorie.* Birkhäuser Verlag, 1976.

Cannon, A., Ettinger, M., Hush, D., Scovel, C. Machine Learning with Data Dependent Hypothesis Classes. *Journal of Machine Learning Research,* 2: 335–358, 2002.

Breiman, L. Prediction games and arcing algorithms. *Neural Computation,* 11(7): 1493–1517, 1999.

Dudley, R. M. *Uniform Central Limit Theorems.* Cambridge University Press, 1999.

Feller, W. *An Introduction to probability theory and its applications, volume II.* John Wiley, 1966.

Gat, Y. A bound concerning the generalization ability of a certain class of learning algorithms. Technical Report 548, University of California, Berkeley, CA, 1999.

Heisele, B., Serre, T., Mukherjee, S., Poggio, T. Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2: 18–24, 2001.

Kohonen, T. The self-organizing map. In *Proceedings of IEEE,* 78: 1464–1479, 1990.

Koltchinskii, V., Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics,* 30(1), 2002.

Koltchinskii, V., Panchenko, D. Complexities of convex combinations and bounding the generalization error in classification. *submitted,* 2004.

Koltchinskii, V., Panchenko, D., Lozano, F. Bounding the generalization error of convex combinations of classifiers: balancing the dimensionality and the margins. *The Annals of Applied Probability,* 13(1), 2003a.

Koltchinskii, V., Panchenko, D., Andonova, S. Generalization bounds for voting classifiers based on sparsity and clustering. In *Proceedings of the Annual Conference on Computational Learning Theory, Lecture Notes in Artificial Intelligence*, M. Warmuth and B. Schoelkopf (eds.). Springer, New York, 2003b.

Littlestone, N., Warmuth, M. Relating data compression and learnability. Technical Report, University of California, Santa Cruz, CA, 1986.

Panchenko, D. Some extensions of an inequality of Vapnik and Červonenkis. *Electronic Communications in Probability,* 7: 55–65, 2002.

Schapire, R., Freund, Y., Bartlett, P., Lee, W. S. Boosting the margin: A new explanation of effectiveness of voting methods. *The Annals of Statistics,* 26: 1651–1687, 1998.

Schapire, R., Singer, Y. Improved Boosting Algorithms using Confidence-Rated Predictions. *Machine Learning,* 37: 297–336, 1999.

Steinwart, I. Sparseness of Support Vector Machines. *Journal of Machine Learning Research,* 2: 1071–1105, 2003.

Vapnik, V. N., Červonenkis, A. Ya. On the uniform convergence of relative frequencies of event to their probabilities. *Soviet Math. Dokl.,* 9: 915 – 918, 1968.

Vapnik, V. N., Červonenkis A. Ya. *Theory of Pattern Recognition.* Nauka, Moscow (in Russian), 1974.

Vapnik, V. *Statistical Learning Theory.* John Wiley & Sons, New York, 1998.

Vapnik, V. *Estimation of Dependencies Based on Empirical Data.* SpringerVerlag, New York, 1982.