entropy

MDPI

Article

Distribution-Dependent Weighted Union Bound †

Luca Oneto [1,*] and Sandro Ridella [2]

[1] Department of Computer Science, Bioengineering, Robotics and Systems Engineering, University of Genoa, Via Opera Pia 11a, 16145 Genova, Italy
[2] Department of Biophysical and Electronic Engineering, University of Genoa, Via Opera Pia 11a, 16145 Genova, Italy; sandro.ridella@unige.it
* Correspondence: luca.oneto@unige.it
† This paper is an extended version of our paper published in European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Brugge, Belgium, 2–4 October 2020.

Abstract: In this paper, we deal with the classical Statistical Learning Theory's problem of bounding, with high probability, the true risk $R(h)$ of a hypothesis h chosen from a set $\mathcal{H}$ of m hypotheses. The Union Bound (UB) allows one to state that $\mathbb{P}\{\mathtt{L}(\hat{R}(h),\delta q_h) \leq R(h) \leq \mathtt{U}(\hat{R}(h),\delta p_h)\} \geq 1-\delta$ where $\hat{R}(h)$ is the empirical errors, if it is possible to prove that $\mathbb{P}\{R(h)\geq \mathtt{L}(\hat{R}(h),\delta)\}\geq 1-\delta$ and $\mathbb{P}\{R(h)\leq \mathtt{U}(\hat{R}(h),\delta)\}\geq 1-\delta$, when h, q_h, and p_h are chosen before seeing the data such that $q_h, p_h \in [0,1]$ and $\sum_{h\in\mathcal{H}}(q_h+p_h)=1$. If no *a priori* information is available q_h and p_h are set to $1/2m$, namely equally distributed. This approach gives poor results since, as a matter of fact, a learning procedure targets just particular hypotheses, namely hypotheses with small empirical error, disregarding the others. In this work we set the q_h and p_h in a distribution-dependent way increasing the probability of being chosen to function with small true risk. We will call this proposal Distribution-Dependent Weighted UB (DDWUB) and we will retrieve the sufficient conditions on the choice of q_h and p_h that state that DDWUB outperforms or, in the worst case, degenerates into UB. Furthermore, theoretical and numerical results will show the applicability, the validity, and the potentiality of DDWUB.

Keywords: union bound; weighted union bound; distribution-dependent weights; statistical learning theory; finite number of hypothesis

Citation: Oneto, L.; Ridella, S. Distribution-Dependent Weighted Union Bound. *Entropy* **2021**, *23*, 101. https://doi.org/10.3390/e23010101

Received: 11 November 2020
Accepted: 10 January 2021
Published: 12 January 2021

1. Introduction

Statistical learning theory [1–4] deals with the problem of understanding and estimating the performance of a statistical learning procedure. The goal is to better understand the factors that influence its behavior and to suggest ways to improve it. Although asymptotic analysis is a crucial first step in this direction, finite sample error bounds are of more value as they allow the design of model selection procedures [5–7]. These error bounds typically have the following form: with high probability, the generalization error of the selected hypothesis, chosen in a space of possible ones, is bounded by an empirical estimate of the generalization error plus a penalty term which depends on the size of the hypothesis space and the number of samples available. The latter term basically considers that the learning procedure selects a hypothesis in a set of possible ones based on the available data. Every data-dependent choice implies a risk, and the penalty term is exactly the measure of this risk. When the hypothesis space is composed of an arbitrary finite number of hypothesis, and no additional information is provided, the evaluation of the total risk is usually made with the Union Bound (UB) [2,7,8]. The UB is an ubiquitous building block in statistical learning theory and is exploited in many context and in many different ways to derive the final result: in the Vapnik–Chervonenkis theory [2], in the Rademacher Complexity theory [9,10], in the Algorithmic Stability theory [11], in the Compression Bound [12], in the PAC-Bayes theory [13], and more recently in the Differential Privacy theory [14].

Let us consider the classical binary classification framework (The extension to the general supervised learning characterized by bounded loss functions will be discussed

later during the presentation.). Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{-1, +1\}$ be the set of binary output labels. Let $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$, $\forall i \in \{1, \cdots, n\}$, be a sequence of $n \in \mathbb{N}^*$ samples drawn independently from an unknown probability distribution μ over $\mathcal{X} \times \mathcal{Y}$. Let us consider a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ chosen from a finite set $\mathcal{H}$ of possible hypotheses of cardinality $m \in \mathbb{N}^*$ such that $\mathcal{H} = \{h_i : i \in \mathcal{I}\}$ where $\mathcal{I} = \{1, \cdots, m\}$. The error of h in approximating $\mathbb{P}\{Y|X\}$ is measured by a prescribed loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Since we are dealing with binary classification problems the most natural choice is the loss function which counts the number of errors $\ell(h(X), Y) = \mathbb{1}\{Y \neq h(X)\} \in \{0, 1\}$. The generalization error of h is defined as

$$R(h) = \mathbb{E}\{\ell(h(X), Y)\} \in [0, 1].$$

Since the probability measure μ is usually unknown, the generalization error cannot be computed; however, we can compute the empirical error

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i) \in [0, 1].$$

If the choice of $h \in \mathcal{H}$ does not depend on $\mathcal{D}_n$, namely if we want to bound the generalization error of a single hypothesis in the hypothesis space chosen before seeing the data, it is possible to prove that (Please note that for simplicity, we will refer to $R(h)$ and $\hat{R}(h)$ with R and $\hat{R}$ respectively, when it is clear from the context.)

$$\mathbb{P}\{R \geq \mathtt{L}(\hat{R}, \delta)\} \geq 1 - \delta, \quad \mathbb{P}\{R \leq \mathtt{U}(\hat{R}, \delta)\} \geq 1 - \delta,$$

where $\delta \in (0, 1)$ while $\mathtt{L}$ and $\mathtt{U}$ are respectively lower and upper bounds of the generalization error (see, for example, [15–17]).

Since the generalization error cannot be smaller than zero or larger than one consequently we have that $\mathtt{L}(\hat{r}, \delta) \in [0, 1)$ and $\mathtt{U}(\hat{r}, \delta) \in (0, 1]$ $\forall \hat{r} \in [0, 1]$ and $\forall \delta \in (0, 1)$ [15]. When, instead of [15,16], or similar results, are exploited it is necessary to truncate them.

In general, the choice of $h \in \mathcal{H}$ does depend on $\mathcal{D}_n$: in this case we must estimate the risk due to this data-dependent choice.

As an example, common practice for choosing $h \in \mathcal{H}$ based on $\mathcal{D}_n$ is to choose the hypothesis with minimum empirical error

$$\arg\min_{h \in \mathcal{H}} \hat{R}(h),$$

and this approach is called Empirical Risk Minimization [2,18], but others possibilities exist such as the Structural Risk Minimization [2,19,20], or the penalized (regularized) Empirical Risk Minimization [21–23].

To guarantee a prescribed confidence level, or risk, of the chosen hypothesis, the UB can be applied. The UB can be expressed in two forms (Please note that for simplicity, we will refer to $R(h_i)$ and $\hat{R}(h_i)$ with R_i and $\hat{R}_i$ respectively, when it is clear from the context.) [8]: a simplified version (Theorem 1) and a generalized version (Theorem 2).

Theorem 1 (Simple UB)**.** *The following bounds hold*

$$\mathbb{P}\left\{\mathtt{L}\left(\hat{R}_i, \frac{\delta}{2m}\right) \leq R_i \leq \mathtt{U}\left(\hat{R}_i, \frac{\delta}{2m}\right) \forall i \in \mathcal{I}\right\} \geq 1 - \delta.$$

Theorem 2 (Generalized UB)**.** *Let $q(h_i) \in (0, 1)$ and $p(h_i) \in (0, 1)$ be some weight associated with h_i with $i \in \mathcal{I}$ before seeing the data (Please note that for simplicity, we will refer to $q(h_i)$ and $p(h_i)$ with q_i and p_i respectively, when it is clear from the context.) and such that $\sum_{i \in \mathcal{I}}(q_i + p_i) = 1$, then the following bounds hold*

$$\mathbb{P}\{\mathtt{L}(\hat{R}_i, \delta q_i) \leq R_i \leq \mathtt{U}(\hat{R}_i, \delta p_i) \; \forall i \in \mathcal{I}\} \geq 1 - \delta.$$

Theorem 1 is a special case of Theorem 2 when $q_i = p_i = 1/2m \; \forall i \in \{1, \cdots, m\}$.

Theorem 2 introduces a weight for each risk associated with each choice. Weighting more the risk associated with useful choices leads to tighter bounds on the generalization error of hypotheses that will be selected by the algorithm (hypotheses characterized by small empirical error) and looser estimates over the others (hypotheses characterized by high empirical error). Unfortunately, this approach is mainly theoretical since the weights must be chosen before seeing the data and consequently we cannot set them without an *a priori* knowledge about the problem. Finally, Theorem 2 does not propose any solution for the choice of these weights.

For this reason, in this work, we propose a Distribution-Dependent Weighted UB (DDWUB) where the weights depend on some parameters of the distribution which generated them, extending our preliminary work [24]. In particular, we define a set of functions $f_i^p : \mathbb{R}^m \to \mathbb{R}$ and $f_i^q : \mathbb{R}^m \to \mathbb{R}$ with $i \in \mathcal{I}$ such that

$$\begin{aligned} &q_i = f_i^q(R_1, \cdots, R_m), p_i = f_i^p(R_1, \cdots, R_m) \in (0,1), \quad \forall i \in \mathcal{I}, \\ &\sum_{i \in \mathcal{I}} \Big(f_i^q(R_1, \cdots, R_m) + f_i^p(R_1, \cdots, R_m) \Big) = 1. \end{aligned}$$

Please note that f_i^q, f_i^p with $i \in \mathcal{I}$ are quite general and are data independent (Please note that in the framework of the paper (binary classification with a loss function which counts the number of misclassified samples), the generalization error is the only parameter of the distribution which is a Binomial). It is surely possible to consider even more general data independent functions for defining the weights, but we think that our definition is general enough to contemplate a wide variety of cases.

At this point the proposed DDWUB for bounding the generalization error of a hypothesis chosen from a finite set of possible ones can be stated.

Theorem 3. *If* $\forall r_1, \cdots, r_m \in [0,1]$

$$\begin{aligned} &f_i^q(r_1, \cdots, r_m), f_i^p(r_1, \cdots, r_m) \in (0,1), \quad \forall i \in \mathcal{I}, \\ &\sum_{i \in \mathcal{I}} \Big(f_i^q(r_1, \cdots, r_m) + f_i^p(r_1, \cdots, r_m) \Big) = 1, \end{aligned}$$

then the following bound holds

$$\mathbb{P}\Big\{ \mathtt{L}\Big(\hat{R}_i, \delta f_i^q(R_1, \cdots, R_m) \Big) \leq R_i \leq \mathtt{U}\Big(\hat{R}_i, \delta f_i^p(R_1, \cdots, R_m) \Big) \; \forall i \in \mathcal{I} \Big\} \geq 1 - \delta.$$

The proof is a direct consequence of Theorem 2.

DDWUB allow the binding of the generalization error of each hypothesis in the space of hypotheses but, to prove that the DDWUB outperforms the UB, we will require some sufficient conditions that $\mathtt{L}$, $\mathtt{U}$, f_i^q, and f_i^p with $i \in \mathcal{I}$ must satisfy. These sufficient conditions define a class of functions and an open problem would be to find special and simpler classes of functions which satisfy them.

Nevertheless, we will show that it is possible to find simple classes of functions which satisfy these conditions. For example, if one is interested in having tighter upper bound of the generalization error of the empirical minimizer DDWUB suggest combining classical $\mathtt{L}$ and $\mathtt{U}$, such as [15] or [16], and set

$$f_i^q(R_1, \cdots, R_m) = \frac{1}{2m}, \quad f_i^p(R_1, \cdots, R_m) = \frac{1}{2} \frac{e^{-\gamma \max[\theta, R_i]}}{\sum_{j \in \mathcal{I}} e^{-\gamma \max[\theta, R_j]}}, \quad \forall i \in \mathcal{I},$$

with particular values of $\gamma \in [0, \infty)$ and $\theta \in [0,1]$.

As a last remark we would like to note that all our results easily extend to multiclass classification problems and regression problems if the loss function is bounded.

DDWUB is a distribution-dependent form of the UB analogously to the Computable Shell Decomposition Bounds (CSDB) [20]. The CSDB splits the hypothesis space in shells based on the generalization error of each hypothesis and, instead of taking into account the risk of each hypothesis in the space, show that it is possible to just take into account the risk of choosing one shell and the risk associated with each hypothesis in the shell. This allows, for example, to not consider hypotheses with high generalization error. The CSDB show also how to estimate the size of these shells based on the histogram of the empirical errors.

DDWUB takes inspiration from several works in the field. The first idea, which is also a driver of the CSDB, is that during any learning procedure the hypotheses with high error will be never taken into account and consequently we should not pay the risk for those hypotheses [10]. The second idea is that since we do not know the true error of the hypotheses but just its empirical one, we should discard those hypotheses for which the estimated confidence intervals do not overlap [25] with the ones of the hypothesis of minimal training error. The third idea is that since there is no supporting theory for discarding the hypothesis with non-overlapping confidence intervals, we should weight differently the risk associated with each hypothesis based on their true error analogously to what is done in the field of multiple hypotheses testing [26]. The fourth idea is that other researchers have shown that a distribution-dependent weighting strategy can be performed without the actual knowledge of the distribution [27]. DDWUB combines all these ideas and improves both on the UB and the CSDB.

DDWUB applies to finite hypotheses spaces and surely more sophisticated techniques, such as Local Vapnik–Chervonenkis [28] or the Local Rademacher Complexity [10], can be employed and can sometimes result in tighter bounds. However, insight into finite classes remains quite useful [20,29]. Finite class analysis can be exploited for as a pedagogical tool. Finite class analysis can teach new directions in which to look for the development and evolution of more sophisticated bounds. Finite class analysis can be useful for model selection purposes (e.g., selecting the most suitable hypothesis space, or set of hyperparameters, or algorithm). Finite class analysis can be useful when the models are represented with limited number of bits because of the constants involved in the bounds.

The rest of the paper is organized as follows. Section 2 presents the DDWUB in a simplified setting. In Section 3 we present the DDWUB in a generalized setting, we derive the sufficient conditions which state when DDWUB improves over the UB, we will show that it is possible to find simple classes of functions which satisfy these conditions, and we will make the connection between our results and the ones of [25]. Section 4 reports a comparison between DDWUB and the UB by means of closed form results. Section 5 reports a comparison between DDWUB and the UB by means of an extensive set of numerical results. Section 6 compares DDWUB with CSDB by means of an extensive set of numerical results. Section 7 shows the applicability and the potentiality of DDWUB. Section 8 concludes the paper. In the Appendices known results, proof, and technicalities (See in Appendixs A–C) are reported for completeness.

2. Distribution-Dependent Weighted Union Bound: Simplified Setting

Let us consider Theorem 3 and the bound proposed by [16] recalled by Theorem A1 in Appendix A. Let us also suppose, for simplicity, that we are interested in upper bounding the generalization error of the empirical risk minimizer (Extensions will be discussed at the end of this section). In this setting it is possible to state our DDWUB.

Corollary 1. *If*

$$f_i(r_1, \cdots, r_m) = \frac{e^{-\gamma \max[\theta, r_i]}}{\sum_{j \in \mathcal{I}} e^{-\gamma \max[\theta, r_j]}}, \quad \forall i \in \mathcal{I},$$

with $\gamma \in [0, \infty)$ *and* $\theta \in [0, 1]$ *then the following bound holds*

$$\mathbb{P}\left\{ \max\left[0, \hat{R}_i - \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right] \leq R_i \leq \min\left[1, \hat{R}_i + \sqrt{\frac{\log\left(\frac{2}{\delta f_i(R_1, \cdots, R_m)}\right)}{2n}}\right] \ \forall i \in \mathcal{I} \right\} \geq 1 - \delta.$$

Corollary 1 is a direct consequence of Theorems 3 and A1.

The choice of the weights takes inspiration from the work of [27] which proposed, in the context of the PAC-Bayes theory, a distribution-dependent method for assigning an a priori distribution over a set of hypotheses to give a higher probability to the hypothesis with small generalization error. This method has been shown to possess interesting theoretical properties [30,31] and to be also quite effective in practical applications [32].

Since we are interested in choosing and bounding the generalization error of the empirical minimizer, let us define

$$i^* = \arg\min_{i \in \mathcal{I}} \hat{R}_i.$$

This approach is analogous to Page's criterion [33], which was designed as a process inspection scheme to detect deviations in average in only one direction (one-sided) in a stochastic process.

In Corollary 1, γ acts as a weighting factor. The larger is γ the larger are the weights of the risks associated with hypotheses with small empirical error and the smaller are the weights of the risks associated with hypotheses with large empirical error. For $\gamma \to \infty$ we have that (For simplicity, we assume in this statement that the empirical minimizer is unique.) $p_{i^*} \to 1$ and $p_i \to 0 \ \forall i \in \mathcal{I} \setminus i^*$. The smaller is γ the less is the difference between the weights of the risks. For $\gamma \to 0$ we have that $p_i = 1/m \ \forall i \in \mathcal{I}$.

In Corollary 1, θ, instead, acts as a protection against the fact that the empirical error is measured over a finite number of samples and, if the sample size is small, hypotheses with a small difference in the empirical error are indistinguishable. In other words, the weights depend on unknown parameters of the data generating distribution, then we will have to estimate them and since the number of sample is finite these estimates will not allow us to distinguish hypotheses which show similar empirical error. For this reasons, θ gives the same the weight to the risks associated with hypotheses with small empirical error.

The values of γ and θ must be set in a particular way to be sure that DDWUB improves over the UB. In particular

- Lemma 1 shows that to upper bound the generalization error of the empirical risk minimizer based on DDWUB of Corollary 1 we must solve an optimization problem;
- Lemmas 2 and 3 show that for particular values of γ the solution is unique and can be found by simply search for the fixed point of a simple function;
- Theorem 4 and Lemma 4 show that for particular values of θ it is possible to prove that DDWUB is tighter than, or in the worst case as tight as, the UB.

Thanks to Corollary 1 we can state the following lemma.

Lemma 1. *Under the same conditions of Corollary 1 if*

$$i^* = \arg\min_{i \in \mathcal{I}} \hat{R}_i,$$

then we can state that following bound holds

$$R_{i^*} \leq \max_{r_1,\cdots,r_m} \; r_{i^*}$$

$$s.t. \; \max\left[0, \hat{R}_i - \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right] \leq r_i \leq \min\left[1, \hat{R}_i + \sqrt{\frac{\log\left(\frac{2\sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, r_j]}}{\delta e^{-\gamma \max[\theta, r_i]}}\right)}{2n}}\right], \forall i \in \mathcal{I}.$$

Lemma 1 can be further simplified as follows.

Lemma 2. *Under the same conditions of Lemma 1 the following bound holds*

$$R_{i^*} \leq \max_{r_{i^*}} \quad r_{i^*}$$

$$s.t. \quad r_{i^*} \leq \min\left[1, \hat{R}_{i^*} + \sqrt{\frac{\log\left(\frac{2\sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, r_j]}}{\delta e^{-\gamma \max[\theta, r_{i^*}]}}\right)}{2n}}\right],$$

where $r_i = \max\left[0, \hat{R}_i - \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right] \; \forall i \in \mathcal{I} \setminus i^*.$

The proof can be found in Appendix B.

Please note that the optimization problem of Lemma 2 can be further simplified noting that for particular values of γ, the solution of the optimization problem is unique.

Lemma 3. *Under the same conditions of Lemma 2 if*

$$\gamma \leq 2\sqrt{n},$$

the solution of the optimization problem of Lemma 2 exists, it is unique, and it is the fixed point $r^*_{i^*}$ *of the following function of* r_{i^*}

$$r_{i^*} = \min\left[1, \hat{R}_{i^*} + \sqrt{\frac{\log\left(\frac{2\sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, r_j]}}{\delta e^{-\gamma \max[\theta, r_{i^*}]}}\right)}{2n}}\right]$$

where $r_i = \max\left[0, \hat{R}_i - \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right] \; \forall i \in \mathcal{I} \setminus i^*.$

The proof can be found in Appendix B.

Please note that to find the fixed point defined in Lemma 3 a simple bisection method can be applied.

For particular values of θ, it is possible to state that DDWUB is tighter than, or in the worst case as tight as, the UB.

Theorem 4. *Under the same conditions of Lemma 3 if*

$$\theta \geq \min\left[1, \hat{R}_{i^*} + \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right],$$

then

$$r^*_{i^*} \leq \min\left[1, \hat{R}_{i^*} + \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right].$$

The proof can be found in Appendix B.

The problem of the θ defined in Theorem 4 is that it is data-dependent since we do not know i^* before seeing the data. For this reason, the following lemma suggests a data independent threshold of θ which satisfies the conditions of Theorem 4.

Lemma 4. *Under the same conditions of Theorem 4*

$$\min\left[1, \hat{R}_{i^*} + \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right] \leq \min\left[1, \min[R_1, \cdots, R_m] + 2\sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right].$$

The proof can be found in Appendix B.

Lemma 4 provides us a method for finding a $\theta \geq \mathtt{U}\left(\hat{R}_{i^*}, \frac{\delta}{2m}\right)$ in a data independent way by setting

$$\theta = \min\left[1, \min[R_1, \cdots, R_m] + 2\sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right]. \tag{1}$$

By finding the $r^*_{i^*}$ for all possible values of θ and then by selecting the largest one which satisfies Equation (1) we have the results of our DDWUB.

Please note that the above-mentioned result easily extends to the whole supervised learning framework, until a bounded loss function is employed, since the inequality proposed by [16] cover this case.

Following the same argument described in this section it is possible to derive the DDWUB for lower bounding the generalization error of the empirical risk minimizer by simply setting in Theorem 3

$$f^q_i(R_1, \cdots, R_m) = \frac{1}{2}\frac{e^{-\gamma \min[\theta, 1-R_i]}}{\sum_{j\in\mathcal{I}} e^{-\gamma \min[\theta, 1-R_j]}}, \quad f^p_i(R_1, \cdots, R_m) = \frac{1}{2m}, \quad \forall i \in \mathcal{I}.$$

Finally, it is possible to derive the DDWUB for upper and lower bounding the generalization error of the empirical risk minimizer by simply setting in Theorem 3

$$f^q_i(R_1, \cdots, R_m) = \frac{1}{2}\frac{e^{-\gamma \min[\theta, 1-R_i]}}{\sum_{j\in\mathcal{I}} e^{-\gamma \min[\theta, 1-R_j]}}, \quad \forall i \in \mathcal{I},$$

$$f^p_i(R_1, \cdots, R_m) = \frac{1}{2}\frac{e^{-\gamma \max[\theta, R_i]}}{\sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, R_j]}}, \quad \forall i \in \mathcal{I}.$$

Example 1. *Before presenting DDWUB in the general setting we would like to show an application of DDWUB in the simplified setting. Let us consider the case when (More general examples can be*

derived, and we will do it later with both closed form and numerical results, but here we want to keep the presentation as simple as possible.)

$$\hat{R}_1 = \hat{R}_2 = 0,\ \hat{R}_3 = \hat{R}_4 = \cdots = \hat{R}_m = \nu, \quad \nu \in \left\{\frac{1}{n}, \frac{2}{n}, \cdots, 1\right\}.$$

Let us set $\gamma = 2\sqrt{n}$ *(see Lemma 3) and note that to upper bound the function with the smallest empirical error (i.e., the one corresponding to* $\hat{R}_1$*) we have that DDWUB states that*

$$\begin{cases} r_1 = \sqrt{\dfrac{\ln\left(\frac{2\sum_{i=1}^{m} e^{-2\sqrt{n}\max[\theta, r_i]}}{\delta e^{-2\sqrt{n}\max[\theta, r_1]}}\right)}{2n}} \\ r_2 = 0 \\ r_3 = r_4 = \cdots = \nu - \sqrt{\dfrac{\ln\left(\frac{2m}{\delta}\right)}{2n}} \\ \theta = \min\left[1, \min[r_1, \cdots, r_m] + 2\sqrt{\dfrac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right]. \end{cases}$$

Please note that for a finite but large enough value of n

$$\min[r_1, \cdots, r_m] = 0 \rightarrow \theta = 2\sqrt{\frac{\ln\left(\frac{2m}{\delta}\right)}{2n}}.$$

Thanks to the theory of DDWUB (see Lemma 4) we can state that

$$r_1^* \leq \theta.$$

Let us note that if

$$m < \frac{\delta e^{2n\left(\frac{\nu}{2}\right)^2}}{2},$$

then

$$r_3 = \cdots = r_m > \theta.$$

Then we can easily state that

$$\lim_{n\to\infty} \frac{2\sum_{i=1}^{m} e^{-2\sqrt{n}\max[\theta, r_i]}}{\delta e^{-2\sqrt{n}\max[\theta, r_1]}} = \frac{4}{\delta},$$

which means that all the hypothesis in the space with $\hat{R} \neq 0$*, if* $m < {}_{\delta e^{2n(\nu/2)^2}}/2$*, are not taken into account, asymptotically, in estimating the upper bound of the hypothesis with the smaller error with DDWUB.*

3. Distribution-Dependent Weighted Union Bound: General Setting

In this section, we will derive the sufficient conditions for stating that DDWUB is tighter than, or in the worst case as tight as, the UB.

In particular, as we have done in Section 2, we will start by supposing that we are just interested in upper bounding the generalization error of the empirical risk minimizer. Nevertheless, as pointed out in Section 2, DDWUB can be easily generalized also to the lower bounds, or to both lower and upper bounds, and to the general supervised setting with bounded loss functions but, in this work, we did not report all these extensions in order not to make the notation and the presentation over-complicated.

As noted in the introduction, the weights should not depend on the data, but they can depend on some parameters of the data generating distribution. For this reason, we define a set of functions $f_i : \mathbb{R}^m \rightarrow \mathbb{R}$ with $i \in \mathcal{I}$ such that

$$f_i(R_1, \cdots, R_m) \in (0, 1) \quad \forall i \in \mathcal{I}, \quad \sum_{i\in\mathcal{I}} f_i(R_1, \cdots, R_m) = 1.$$

In this setting DDWUB can be formulated as follows.

Corollary 2. *If* $\forall r_1, \cdots, r_m \in [0,1]$

$$f_i(r_1, \cdots, r_m) \in (0,1) \quad \forall i \in \mathcal{I}, \quad \sum_{i \in \mathcal{I}} f_i(r_1, \cdots, r_m) = 1,$$

then the following bound holds

$$\mathbb{P}\left\{ \mathtt{L}\left(\hat{R}_i, \frac{\delta}{2m} \right) \leq R_i \leq \mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(R_1, \cdots, R_m)}{2} \right) \ \forall i \in \mathcal{I} \right\} \geq 1 - \delta.$$

Corollary 2 is a direct consequence of Theorem 2.

To prove that DDWUB outperforms UB we will require some sufficient conditions that $\mathtt{L}$, $\mathtt{U}$, and f_i with $i \in \mathcal{I}$ must satisfy. Please note that from Corollary 2, it is possible to derive all the lower bounds of the generalization error of the hypotheses in the class since $\mathtt{L}\left(\hat{R}_i, \frac{\delta}{2m} \right)$ depends just on known quantities. For what concerns, instead, the upper bounds, the answer is not as easy.

In the rest of this section, we will show how to find the upper bound of the generalization error of a hypothesis chosen in $\mathcal{H}$ based on DDWUB (Corollary 2) and under which conditions these upper bounds are tighter than the one of UB (Theorem 1). For this purpose

- Lemma 5 will show that under certain conditions, the bound of Corollary 2 can be exploited to compute the upper bound of the generalization error of a hypothesis chosen in a class of possible ones based on the observation of a set of data, by solving a complex optimization problem;
- Lemma 6 will show the conditions under which the optimization problem of Lemma 5 can be simplified;
- Lemma 7 will show the conditions under which the solution of the optimization problem of Lemma 6 is unique;
- Theorem 5 will show the conditions under which the upper bound of the generalization error of Lemma 5 found with Lemma 7 is never looser than the one computed with the UB of Theorem 1. These conditions require the knowledge of a data-dependent threshold;
- Lemma 8 shows that it is possible to estimate this threshold of Theorem 5 in a data independent fashion.

Thanks to Corollary 2 we can state the following lemma.

Lemma 5. *Under the same conditions of Corollary 2, if* $\forall \hat{r} \in [0,1]$ *and* $\forall \delta \in (0,1)$

$$\mathtt{L}(\hat{r}, \delta) \in [0,1), \quad \mathtt{U}(\hat{r}, \delta) \in (0,1]$$

then the following bound holds with probability at least $(1 - \delta)$ *and* $\forall i \in \mathcal{I}$

$$R_i \leq \max_{r_1, \cdots, r_m} \quad r_i$$
$$s.t. \quad \mathtt{L}\left(\hat{R}_j, \frac{\delta}{2m} \right) \leq r_j \leq \mathtt{U}\left(\hat{R}_j, \frac{\delta f_j(r_1, \cdots, r_m)}{2} \right), \quad j \in \mathcal{I}.$$

The solution of the optimization problem of Lemma 5 is not trivial to be found and its properties are not easy to catch.

The following lemma helps us in simplifying the optimization problem of Lemma 5 under a quite natural condition: the upper bound of the generalization error of a hypothesis should decrease if the generalization error of one of the other hypotheses in the class increases.

Lemma 6. *Under the same conditions of Lemma 5, if* $\forall \hat{r}_i \in \{0, 1/n, \cdots, 1\}$, $\forall r_1, \cdots, r_m \in [0,1]$ *and* $\forall j \in \mathcal{I} \setminus i$ *and* $\forall r'_j, r''_j \in [0,1]$ *such that* $r'_j < r''_j$

$$\mathtt{U}\left(\hat{r}_i, \frac{\delta f_i(r_1, \cdots, r_{j-1}, r'_j, r_{j+1}, \cdots, r_m)}{2}\right) - \mathtt{U}\left(\hat{r}_i, \frac{\delta f_i(r_1, \cdots, r_{j-1}, r''_j, r_{j+1}, \cdots, r_m)}{2}\right) \geq 0,$$

where $i \in \mathcal{I}$, *then the optimization problem of Lemma 5 is equivalent to the following one*

$$\begin{aligned} R_i \leq \max_{r_i} \quad & r_i \\ s.t. \quad & r_i \leq \mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(r_1, \cdots, r_m)}{2}\right). \end{aligned}$$

where $r_j = \mathtt{L}\left(\hat{R}_j, \frac{\delta}{2m}\right)$, *with* $j \in \mathcal{I} \setminus i$.

In fact, the hypotheses of the lemma imply that to reach the maximum or r_i, one must reach the lower bounds of r_j with $j \in \mathcal{I} \setminus i$.

Even if the optimization problem of Lemma 6 is much simpler than the one of Lemma 5, the next result further simplifies it under another sufficient condition which ensure the existence and uniqueness of the solution: the upper bound of the generalization error of a hypothesis should not increase too fast it its generalization error decreases.

Lemma 7. *Under the same conditions of Lemma 6, if* $\forall \hat{r}_i \in \{0, 1/n, \cdots, 1\}$, $\forall r_1, \cdots, r_m \in [0,1]$, *and* $\forall r'_i, r''_i \in [0,1]$ *such that* $r'_i < r''_i$

$$\frac{\mathtt{U}\left(\hat{r}_i, \frac{\delta f_i(r_1, \cdots, r_{i-1}, r''_i, r_{i+1}, \cdots, r_m)}{2}\right) - \mathtt{U}\left(\hat{r}_i, \frac{\delta f_i(r_1, \cdots, r_{i-1}, r'_i, r_{i+1}, \cdots, r_m)}{2}\right)}{r''_i - r'_i} < 1,$$

then the solution r_i^* *of the optimization problem of Lemma 6 exists, it is unique, and it is the fixed point of the following function of* r_i

$$r_i = \mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(r_1, \cdots, r_m)}{2}\right),$$

where $r_j = \mathtt{L}\left(\hat{R}_j, \frac{\delta}{2m}\right)$ *with* $j \in \mathcal{I} \setminus i$.

In fact, the hypothesis of the lemma guarantees the uniqueness of the solution.

Algorithm 1 reports a simple pseudo-code for finding the fixed point of Lemma 7 based on the bisection method.

The next result introduces a parameter, more specifically a threshold, θ and states the condition over θ which states that the DDWUB improves over the UB.

Theorem 5. *Under the same conditions of Lemma 7, let us consider* $\theta \in [0,1]$ *and suppose that* $\forall r_1, \cdots, r_m \in [0,1]$ *and* $\forall r'_j, r''_j \in [0, \theta]$

$$f_j(r_1, \cdots, r_{j-1}, r'_j, r_{j+1}, \cdots, r_m) - f_j(r_1, \cdots, r_{j-1}, r''_j, r_{j+1}, \cdots, r_m) = 0,$$

with $j \in \mathcal{I}$. *Let us also suppose that if* $r_1, \cdots, r_m \in [0, \theta]$ *then* $\forall j \in \mathcal{I}$

$$f_j(r_1, \cdots, r_m) = \frac{1}{m}.$$

If

$$\theta \geq \mathtt{U}\left(\hat{R}_i, \frac{\delta}{2m}\right),$$

and if $\forall r_1, \cdots, r_m \in [0,1]$, $\forall j \in \mathcal{I} \setminus i$, $\forall r'_j, r''_j \in (\theta, 1]$ *such that* $r'_j < r''_j$

$$f_i(r_1, \cdots, r_{j-1}, r'_j, r_{j+1}, \cdots, r_m) - f_i(r_1, \cdots, r_{j-1}, r''_j, r_{j+1}, \cdots, r_m) < 0$$

then the following bound holds

$$r_i^* \leq \mathtt{U}\left(\hat{R}_i, \frac{\delta}{2m}\right).$$

The proof of Theorem 5 can be found in Appendix B.

Theorem 5 basically states that under particular conditions, the solution of the problem of Corollary 2 is never looser than the one of Theorem 1.

Unfortunately, we cannot set $\theta = \mathtt{U}\left(\hat{R}_i, \frac{\delta}{2m}\right)$ since this would be a data-dependent choice which will result in a data-dependent weighting strategy. The next lemma addresses this problem.

Algorithm 1: Algorithm for finding the fixed point of Lemma 7 based on the bisection method.

Input: m, $\hat{R}_i$ and $f_i(r_1, \cdots, r_m)$ with $i \in \mathcal{I}$, δ, n, and the precision ϵ
Output: r_i^*

$r_j = \mathtt{L}\left(\hat{R}_j, \frac{\delta}{2m}\right), \quad j \in \mathcal{I} \setminus i;$
$r_i^l = 0, r_i^u = 1;$
while $r_i^u - r_i^l > \epsilon$ **do**
  $r_i = \frac{r_i^u + r_i^l}{2};$
  if $r_i - \mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(r_1, \cdots, r_m)}{2}\right) \leq 0$ **then**
    $r_i^l = r_i;$
  else
    $r_i^u = r_i;$
$r_i^* = r_i;$

Lemma 8. *Under the same conditions of Theorem 5, if* $\forall \hat{r}, \hat{r}', \hat{r}'' \in \{0, 1/m, \cdots, 1\}$ *such that* $\hat{r}' < \hat{r}''$ *we have that*

$$\begin{aligned} &i^* = \arg\min_{i \in \mathcal{I}} \hat{R}_i, \\ &\mathtt{L}(\hat{r}', \delta) - \mathtt{L}(\hat{r}'', \delta) \leq 0, \\ &\exists \mathtt{L}^{-1} : \mathtt{L}^{-1}(\mathtt{L}(\hat{r}, \delta), \delta) \geq \hat{r}, \end{aligned}$$

then

$$\mathtt{U}\left(\hat{R}_{i^*}, \frac{\delta}{2m}\right) \leq \mathtt{U}\left(\mathtt{L}^{-1}\left(\min[R_1, \cdots, R_m], \frac{\delta}{2m}\right), \frac{\delta}{2m}\right).$$

The proof of Lemma 8 can be found in Appendix B.

Lemma 8 provides us a method for finding a $\theta \geq \mathtt{U}\left(\hat{R}_{i^*}, \frac{\delta}{2m}\right)$ in a data independent way by setting

$$\theta = \mathtt{U}\left(\mathtt{L}^{-1}\left(\min[R_1, \cdots, R_m], \frac{\delta}{2m}\right), \frac{\delta}{2m}\right).$$

Thanks to all these results we can provide a method for finding the fixed point of Lemma 7 but with data independent weighting strategy which satisfies the hypothesis of Theorem 5 and a data independent θ defined in Lemma 8.

Lemma 9. *Under the same conditions of Lemma 8, Algorithm 2 finds the fixed point of Lemma 7 but with a data independent weighting strategy which satisfies the hypothesis of Theorem 5 and a data independent θ defined in Lemma 8.*

The proof of Lemma 9 can be found in Appendix B.

Algorithm 2: Algorithm for finding the fixed point of Lemma 7 but with data independent weighting strategy which satisfies the hypothesis of Theorem 5 and a data independent θ defined in Lemma 8.

Input: m, $\hat{R}_i$ and $f_i(r_1, \cdots, r_m)$ with $i \in \mathcal{I}$, δ, n, and the precision ϵ
Output: r_i^*

$i^* = \arg\min_{i \in \mathcal{I}} \hat{R}_i$;
$r_j = \mathtt{L}\left(\hat{R}_j, n, \frac{\delta}{2m}\right), \quad j \in \mathcal{I} \setminus i^*$;
for $\theta \leftarrow 0$ **to** 1 ***by*** ϵ **do**
  $r_{i^*}^l = 0, r_{i^*}^u = 1$;
  while $r_{i^*}^u - r_{i^*}^l > \epsilon$ **do**
    $r_{i^*} = \frac{r_{i^*}^u + r_{i^*}^l}{2}$;
    if $r_{i^*} - \mathtt{U}\left(\hat{R}_{i^*}, n, \frac{\delta f_{i^*}(r_1, \cdots, r_m)}{2}\right) \leq 0$ **then**
      $r_{i^*}^l = r_{i^*}$;
    else
      $r_{i^*}^u = r_{i^*}$;
  if $\left|\theta - \mathtt{U}\left(\mathtt{L}^{-1}\left(\min[r_1, \cdots, r_m], n, \frac{\delta}{2m}\right), n, \frac{\delta}{2m}\right)\right| \leq \epsilon$ **then**
    $r_{i^*}^* = r_{i^*}$;

Please note that if we apply this general theory to Corollary 1 we obtain the same results of Section 2.

In the next section, instead, we apply the general theory to a the more complex case of when [15] is employed together with the same weights exploited in Corollary 1.

3.1. From Theory to Practice

In this section, we will exploit the same solution of Corollary 1 for the weights needed in DDWUB and we will show that is satisfies hypothesis of the Theorems, the Corollaries, and the Lemmas presented in the previous section. In particular we will set

$$f_i(R_1, \cdots, R_m) = \frac{e^{-\gamma \max[\theta, R_i]}}{\sum_{j \in \mathcal{I}} e^{-\gamma \max[\theta, R_j]}} \quad \forall i \in \mathcal{I},$$

where $\gamma \in [0, +\infty)$ and $\theta \in [0, 1]$ are finite constants which regulates the shape of the distribution of the weights.

The following lemma shows that these weighs satisfy the sufficient conditions which states that DDWUB outperforms, or in the worst case performs as, the UB.

Lemma 10. *If $\gamma \in [0, +\infty)$ is a finite constant and*

$$f_i(R_1, \cdots, R_m) = \frac{e^{-\gamma \max[\theta, R_i]}}{\sum_{j \in \mathcal{I}} e^{-\gamma \max[\theta, R_j]}} \quad \forall i \in \mathcal{I},$$

then the hypotheses of Corollary 2 and Theorem 5 are satisfied.

The proof of Lemma 10 can be found in Appendix B.

Please note that for what concerns the hypothesis of Lemmas 6 and 7, we cannot prove that condition holds without knowing the shape of the lower and upper bounds of the generalization error. For this reason, we exploit the generalization bounds of [15] which are the tightest ones in the settings of this paper [6]. This will allow us in Section 5 to compare the UB with the DDWUB with a set of numerical experiments.

To reach our goal let us define the Regularized Incomplete Beta Function $p = \mathtt{F}(r;a,b) = \mathtt{I}_r(a,b)$ and its inverse $r = \mathtt{F}^{-1}(p;a,b)$ with parameters specified by $a \in \mathbb{N}^*$ and $b \in \mathbb{N}^*$ for the corresponding values of r and probabilities in p

$$\mathtt{F}(r;a,b) = \frac{1}{\mathtt{B}(a,b)} \int_0^r t^{a-1}(1-t)^{b-1}dt, \quad \mathtt{F}^{-1}(p;a,b) = \{r : \mathtt{F}(r;a,b) = p\},$$

where

$$\mathtt{B}(a,b) = \mathtt{B}(b,a) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$$

is the Complete Beta Function.

The following lemma states the conditions under which all the hypotheses of Corollary 2, Lemmas 6 and 7, Theorem 5, and Lemma 8 are satisfied if, in Corollary 2, the lower and upper generalization bounds proposed by [15], recalled by Theorem A2 in the Appendix A, and the weights defined in Lemma 10 are exploited.

Lemma 11. *Let us exploit the lower and upper generalization bounds proposed by [15] in Corollary 2*

$$\mathtt{L}(\hat{R},\delta) = \begin{cases} \mathtt{F}^{-1}(\delta; n\hat{R}, n - n\hat{R} + 1) & \hat{R} \in \left\{\frac{1}{n}, \frac{2}{n}, \cdots, 1\right\} \\ 0 & \hat{R} = 0 \end{cases},$$

$$\mathtt{U}(\hat{R},\delta) = \begin{cases} \mathtt{F}^{-1}(1-\delta; n\hat{R} + 1, n - n\hat{R}) & \hat{R} \in \left\{0, \frac{1}{n}, \cdots, \frac{n-1}{n}\right\} \\ 1 & \hat{R} = 1 \end{cases},$$

together with the weights defined in Lemma 10. Then if

$$i^* = \arg\min_{i \in \mathcal{I}} \hat{R}_i,$$

$$\hat{R}_{i^*} \neq 1,$$

$$\gamma < \min_{\hat{r} \in \left\{0, \frac{1}{n}, \cdots, \frac{n-1}{n}\right\},\ p \in (0,1)} \frac{\left(\mathtt{U}\left(\hat{r}, \frac{\delta p}{2}\right)\right)^{n\hat{r}} \left(1 - \mathtt{U}\left(\hat{r}, \frac{\delta p}{2}\right)\right)^{n - n\hat{r} - 1}}{\mathtt{B}(n\hat{r} + 1, n - n\hat{r}) \frac{\delta}{2} p(1-p)},$$

the hypotheses of Corollary 2, Lemmas 6 and 7, Theorem 5, and Lemma 8 are satisfied. Moreover, note that

$$\mathtt{L}^{-1}(r,n,\delta) = \min\left\{\hat{r} : \hat{r} \in \left\{0, \frac{1}{n}, \cdots, 1\right\}, r \leq \mathtt{L}(\hat{r},\delta)\right\},$$

$$\min_{\hat{r} \in \left\{0, \frac{1}{n}, \cdots, \frac{n-1}{n}\right\},\ p \in (0,1)} \frac{\left(\mathtt{U}\left(\hat{r}, \frac{\delta p}{2}\right)\right)^{n\hat{r}} \left(1 - \mathtt{U}\left(\hat{r}, \frac{\delta p}{2}\right)\right)^{n - n\hat{r} - 1}}{\mathtt{B}(n\hat{r} + 1, n - n\hat{r}) \frac{\delta}{2} p(1-p)} \geq \frac{2\sqrt{n}}{\sqrt{2\pi} e^{\frac{7}{6}}}.$$

The proof of Lemma 11 can be found in Appendix B.

The case where $\hat{R}_{i^*} = 1$ is trivial since, in this case, we can safely state that $R_{i^*} \leq 1$.

Please note that the limit in the value of γ is $O(\sqrt{n})$ and this result is connected and in agreement with the consistency results derived in the PAC-Bayes [30] and in Algorithmic Stability [31] theories where the distribution-dependent prior of [27] is exploited.

3.2. Observation

Let us consider the case where the empirical errors of the hypotheses in the space have been sorted as follows

$$\hat{R}_1 \leq \hat{R}_2 \leq \cdots \leq \hat{R}_m,$$

and let us exploit the results of Section 3.1.

Let us set

$$\gamma = \frac{\sqrt{n}}{4} < \frac{2\sqrt{n}}{\sqrt{2\pi}e^{\frac{7}{6}}},$$

and suppose that

$$\hat{R}_1 \neq 1.$$

Then by using the UB (see Theorem 1) we can state that with probability at least $(1-\delta)$

$$R_1 \in \left[0, \mathtt{U}\left(\hat{R}_1, \frac{\delta}{2m}\right)\right],$$

while, by using DDWUB (see Lemma 11), we can state that with probability at least $(1-\delta)$

$$0 \leq R_1 \leq r_1^* = \max_{r_1} \quad r_1$$
$$\text{s.t.} \quad \begin{cases} r_1 = \mathtt{U}\left(\hat{R}_1, \frac{\delta f_1(r_1,\cdots,r_m)}{2}\right) \\ r_i = \mathtt{L}\left(\hat{R}_i, \frac{\delta}{2m}\right), \quad i \in \mathcal{I} \setminus 1 \\ f_i(r_1,\cdots,r_m) = \frac{e^{-\gamma \max[\theta, r_i]}}{\sum_{j=1}^{m} e^{-\gamma \max[\theta, r_j]}}, \quad i \in \mathcal{I} \\ \gamma = \frac{\sqrt{n}}{4} \\ \theta = \mathtt{U}\left(\mathtt{L}^{-1}\left(\min[r_1,\cdots,r_m], \frac{\delta}{2m}\right), \frac{\delta}{2m}\right) \end{cases}.$$

By looking at this last problem and by setting $r_1 = r_1^*$ for simplicity, we can observe some properties. The first one is that

$$f_i(r_1,\cdots,r_m) = \begin{cases} \frac{e^{-\gamma\theta}}{|\mathcal{J}|e^{-\gamma\theta} + \sum_{j=\mathcal{I}\setminus\mathcal{J}} e^{-\gamma r_j}} & i \in \mathcal{J} \\ \frac{e^{-\gamma r_i}}{|\mathcal{J}|e^{-\gamma\theta} + \sum_{j=\mathcal{I}\setminus\mathcal{J}} e^{-\gamma r_j}} & i \in \mathcal{I}\setminus\mathcal{J} \end{cases}, \quad \mathcal{J} = \{j : j \in \mathcal{I}, r_j \leq \theta\}.$$

The second one is that $\forall i \in \mathcal{I} \setminus \mathcal{J}$

$$\frac{e^{-\gamma r_i}}{|\mathcal{J}|e^{-\gamma\theta} + \sum_{j=\mathcal{I}\setminus\mathcal{J}} e^{-\gamma r_j}} \leq \frac{e^{-\gamma\theta}}{|\mathcal{J}|e^{-\gamma\theta} + \sum_{j=\mathcal{I}\setminus\mathcal{J}} e^{-\gamma r_j}}.$$

These properties state that DDWUB, with respect to the UB, can discard, or more properly reduce, the risk of the hypotheses h_i with $i \in \mathcal{I} \setminus \mathcal{J}$ when estimating the generalization error of the hypothesis h_1. As a first raw approximation, we can state that we must pay a risk only for the hypotheses h_i with $i \in \mathcal{J}$ obtaining

$$R_1 \in \left[0, \mathtt{U}\left(\hat{R}_1, \frac{\delta}{2|\mathcal{J}|}\right)\right],$$

where $|\mathcal{J}| \leq m$. A quite important aspect is then to understand some properties of θ. Let us recall that

$$\theta = \mathtt{U}\left(\mathtt{L}^{-1}\left(\min[r_1,\cdots,r_m], \frac{\delta}{2m}\right), \frac{\delta}{2m}\right),$$

but we can easily state that

$$\min[r_1, \cdots, r_m] = \min\left[\mathtt{U}\left(\hat{R}_1, \frac{\delta f_1(r_1, \cdots, r_m)}{2}\right), \mathtt{L}\left(\hat{R}_2, \frac{\delta}{2m}\right)\right].$$

Thanks to Theorem 5, we can state that $f_1(r_1, \cdots, r_m) \geq \frac{1}{m}$ and then the case where

$$\min\left[\mathtt{U}\left(\hat{R}_1, \frac{\delta f_1(r_1, \cdots, r_m)}{2}\right), \mathtt{L}\left(\hat{R}_2, \frac{\delta}{2m}\right)\right] = \mathtt{U}\left(\hat{R}_1, \frac{\delta f_1(r_1, \cdots, r_m)}{2}\right),$$

is quite unusual since it means that $\hat{R}_2 - \hat{R}_1$ is large, or, in other words, it means that our set of hypotheses contains just one hypothesis with small error and many hypotheses with high error. Instead, if

$$\min\left[\mathtt{U}\left(\hat{R}_1, \frac{\delta f_1(r_1, \cdots, r_m)}{2}\right), \mathtt{L}\left(\hat{R}_2, \frac{\delta}{2m}\right)\right] = \mathtt{L}\left(\hat{R}_2, \frac{\delta}{2m}\right),$$

then

$$\theta = \mathtt{L}\left(\hat{R}_2, \frac{\delta}{2m}\right),$$

which means that the threshold θ is obtained from the upper bound of the second-best hypothesis in the space, namely the hypothesis with the second smallest empirical error. To the best of our knowledge, this result is new since in the past many researchers have tried, with similar approaches but with no supporting theory, in proposing to clean the hypothesis space from the hypotheses with high empirical error. The basic idea of these methods is that it is reasonable to discard all the hypotheses such that the lower bound of their generalization error is greater than $\mathtt{U}\left(\hat{R}_1, \frac{\delta}{2m}\right)$, namely the upper bound of the generalization error of the hypothesis with the smallest empirical error, see for example [25]. Our theory states, instead, that we can smooth the risk due to the hypotheses such that the lower bound of their generalization error is greater than $\mathtt{U}\left(\hat{R}_2, n, \frac{\delta}{2m}\right)$, namely the upper bound of the generalization error of the hypothesis with the second smallest empirical error.

4. Closed Form Results

In this section, we will report some closed form results regarding the Lemma 11. These examples are useful for providing an idea of the effect of the DDWUB, an extensive comparison with numerical results is provided in Section 5.

Let us consider the case where

$$\hat{R}_1 = \hat{R}_2 = 0, \quad \hat{R}_3 = \cdots = \hat{R}_m = 1.$$

Let us set

$$\gamma = \frac{\sqrt{n}}{4} < \frac{2\sqrt{n}}{\sqrt{2\pi}e^{\frac{7}{6}}},$$

and note that

$$\arg\min_{i \in \mathcal{I}} \hat{R}_i = 1, \quad \hat{R}_1 \neq 1.$$

If we use the UB (see Theorem 1) with the [15] we obtain that with probability at least $(1 - \delta)$ the following bound holds

$$R_1 \in \left[0, 1 - \sqrt[n]{\frac{\delta}{2m}}\right].$$

If, instead, we use DDWUB (see Lemma 9 and Algorithm 2) we obtain that with probability at least $(1-\delta)$ the following bound holds

$$0 \leq R_1 \leq r_1^* = \max_r \quad r$$
$$\text{s.t.} \quad \begin{cases} r = 1 - \sqrt[n]{\frac{\delta f_1(r, r_2, \cdots, r_m)}{2}} \\ f_1(r, r_2, \cdots, r_m) = \frac{e^{-\gamma \max[\theta, r]}}{e^{-\gamma \max[\theta, r]} + \sum_{j=2}^{m} e^{-\gamma \max[\theta, r_j]}} \\ r_2 = 0 \\ r_3 = \cdots = r_m = \sqrt[n]{\frac{\delta}{2m}} \\ \theta = \mathtt{U}\left(\mathtt{L}^{-1}\left(\min[r, r_2, \cdots, r_m], \frac{\delta}{2m}\right), \frac{\delta}{2m}\right) \\ \gamma = \frac{\sqrt{n}}{4} \end{cases}.$$

Since we can surely state that

$$\min[r_1^*, \cdots, r_m] = 0,$$

then

$$\theta = 1 - \sqrt[n]{\frac{\delta}{2m}}.$$

Moreover, thanks to Theorem 5, we can state that

$$r_1^* \leq \theta.$$

Asymptotic Case

Please note that

$$f_1(r_1^*, r_2, \cdots, r_m) = \frac{1}{2 + (m-2)e^{\frac{\sqrt{n}}{4}\left(1 - 2\sqrt[n]{\frac{\delta}{2m}}\right)}},$$

moreover

$$\lim_{n \to \infty} \frac{1}{2 + (m-2)e^{\frac{\sqrt{n}}{4}\left(1 - 2\sqrt[n]{\frac{\delta}{2m}}\right)}} = \frac{1}{2}.$$

Consequently, at least asymptotically, DDWUB can discard entirely the risk due to the hypotheses with high empirical error.

Finite Sample Case

DDWUB obviously gives an advantage, with respect to the UB, until

$$r_3 = \cdots = r_m > \theta,$$

This means that we have an advantage in terms of the tightness of the estimated upper bound for R_1 until

$$\sqrt[n]{\frac{\delta}{2m}} > 1 - \sqrt[n]{\frac{\delta}{2m}},$$

and consequently until

$$m < \delta 2^{n-1}.$$

5. Numerical Results

This section is devoted to the numerical comparison between the UB (see Theorem 1) and the DDWUB (see Lemma 11).

In particular we will focus on upper bounding the generalization error of the hypothesis, in the set of possible ones, characterized by the smallest empirical error. The comparison between the

UB and the DDWUB will be made in different scenarios to better understand the advantages of the DDWUB.

Scenario A.

Scenario A is an optimistic scenario, the same of Section 4: the set of hypotheses contains few useful hypotheses (small empirical error) and a lot of useless hypotheses (high empirical error) such that

$$\hat{R}_1 = \hat{R}_2 = 0, \quad \hat{R}_3 = \cdots = \hat{R}_m = 1.$$

We set $\delta = 0.05$ and set the numerical precision of the algorithms to $\epsilon = 0.0001$.

Figure 1 reports the estimated generalization error upper bound of the hypothesis with the smallest empirical error estimated with the UB and the DDWUB, together with the percentage of improvement in three different sub-scenarios:

- Sub-scenario A.1 (Figure 1a): we set $m = 1000$ and we vary $n \in \{10, 20, 40, 100, 200, 400, 1000\}$;
- Sub-scenario A.2 (Figure 1b): we vary $m \in \{10, 100, 1000, 10000, 100000\}$ and we set $n = 40$;
- Sub-scenario A.3 (Figure 1c): we vary $m \in \{10, 100, 1000, 10000, 100000\}$ and we set $n = 100$.

Based on the results reported in Figure 1 we can observe that

- DDWUB is always tighter, or equivalent in the worst case, with respect to the UB;
- increasing the number of samples always increases the advantage of the DDWUB over the UB until all the risk of the hypothesis with largest empirical error is disregarded;
- increasing m increases the advantage of the DDWUB over the UB until a limit value for m: if too many useless hypotheses are present it is not able anymore to disregard their risk. Nevertheless, the larger is n the far is this value for m.

In a slightly less optimistic scenario when

$$\hat{R}_1 = \hat{R}_2 = 0, \quad \hat{R}_3 = \cdots = \hat{R}_m = \frac{1}{2},$$

which is the case of lot of charlatans and just a few good candidates in a hiring process [34], the results are reported in Figure 2a–c and do not change too much, apart from the limit of m when the DDWUB stops to improve over the UB which is obviously smaller.

Scenario B.

The second scenario is a more classical one when

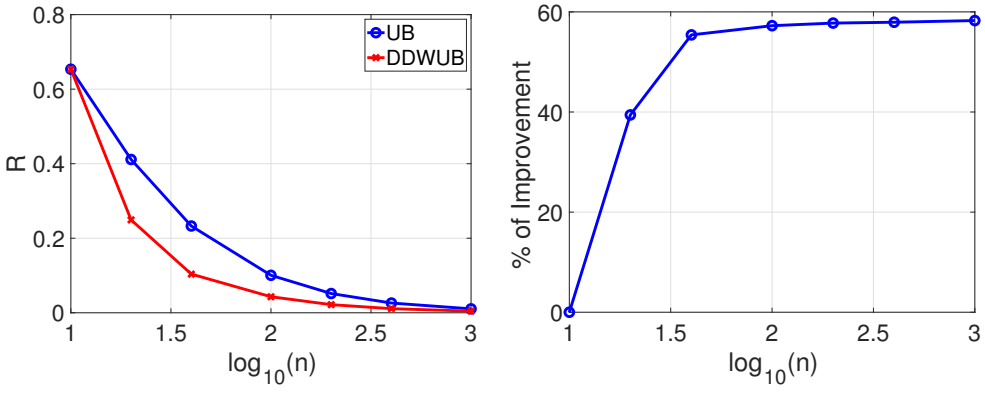
$$\hat{R}_1 = 0, \quad \hat{R}_2 = \frac{1}{n}, \quad \cdots, \quad \hat{R}_{n+1} = 1.$$

This case is exploited in many applications (e.g., the CSDB of [20], or the Structural Risk Minimization of [2], or the Structural Risk Minimization over data-dependent hierarchies of [19]).

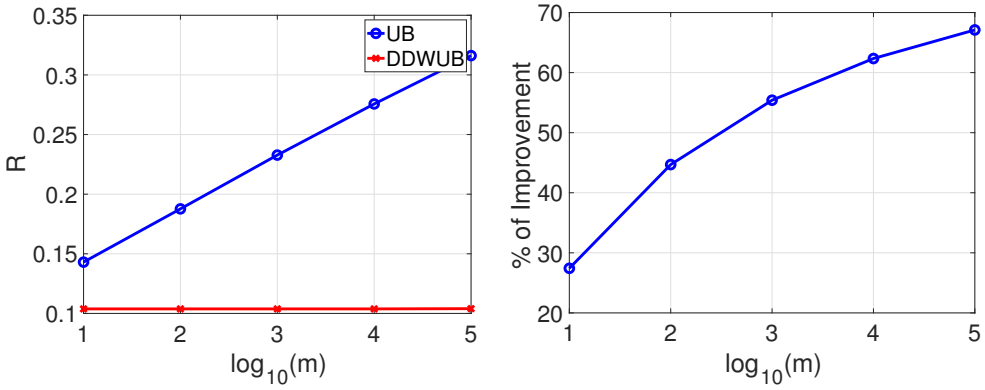
Also, in this scenario we set $\delta = 0.05$ and we set the numerical precision to $\epsilon = 0.0001$. Then we vary $n \in \{10, 20, 40, 100, 200, 400, 1000\}$.

Figure 3 reports the estimated generalization error upper bound of the hypothesis with the smallest empirical error estimated with the UB and the DDWUB, together with the percentage of improvement.

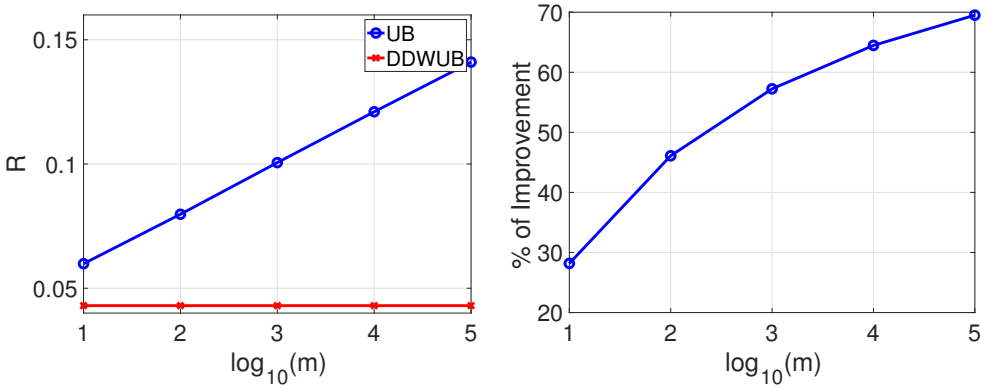
Figure 3 clearly shows the advantage of the DDWUB over the UB and the improvement in the advantage as soon as n increases.



(**a**) Sub-scenario A.1 ($m = 1000$).

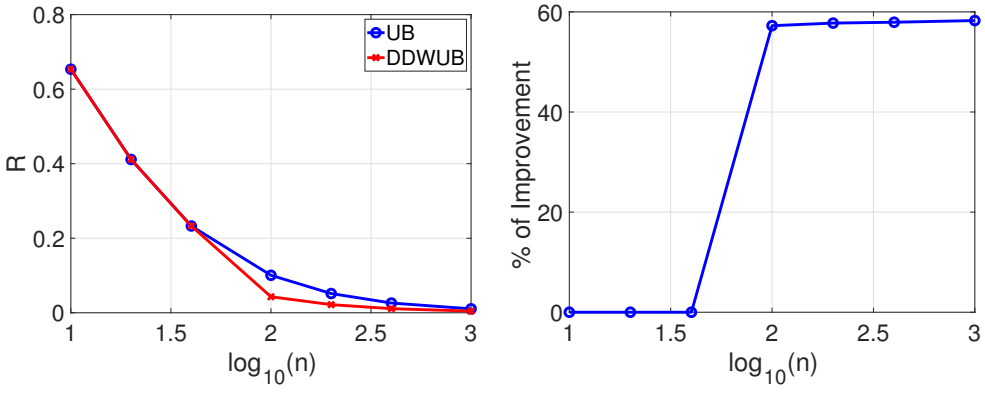


(**b**) Sub-scenario A.2 ($n = 30$).

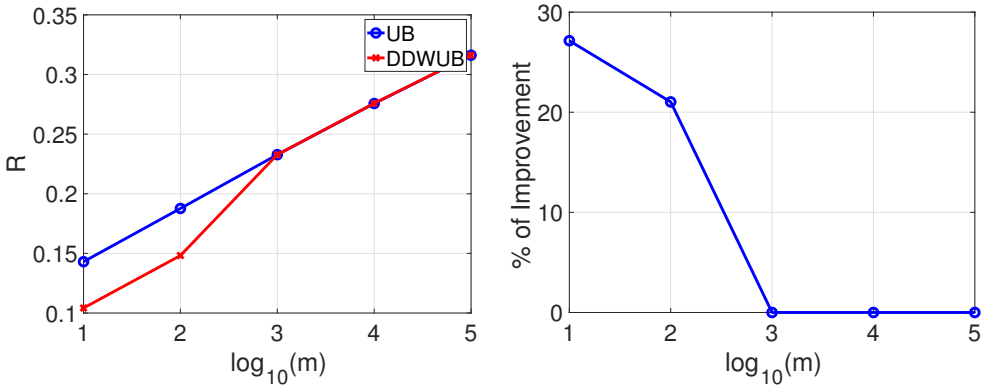


(**c**) Sub-scenario A.3 ($n = 100$).

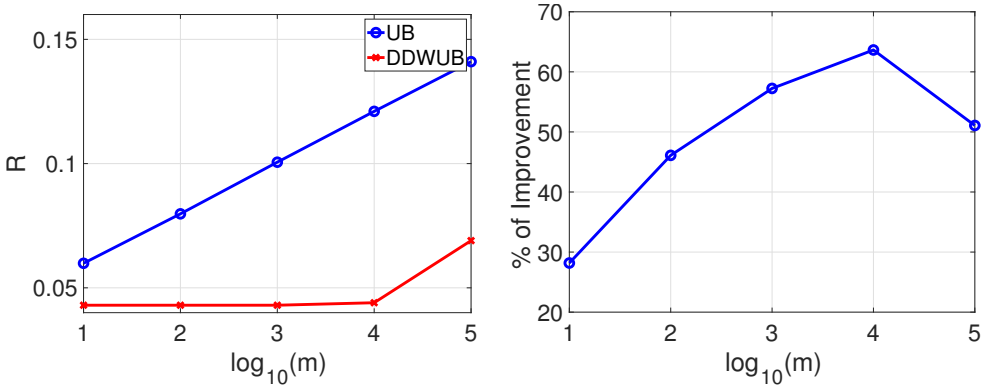
Figure 1. Scenario A ($\hat{R}_1 = \hat{R}_2 = 0$ and $\hat{R}_3 = \cdots = \hat{R}_m = 1$): upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the UB and the DDWUB together with the percentage of improvement.



(**a**) Sub-scenario A.1 ($m = 1000$)



(**b**) Sub-scenario A.2 ($n = 30$)



(**c**) Sub-scenario A.3 ($n = 100$)

Figure 2. Scenario A ($\hat{R}_1 = \hat{R}_2 = 0$ and $\hat{R}_3 = \cdots = \hat{R}_m = \frac{1}{2}$): upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the UB and the DDWUB together with the percentage of improvement.

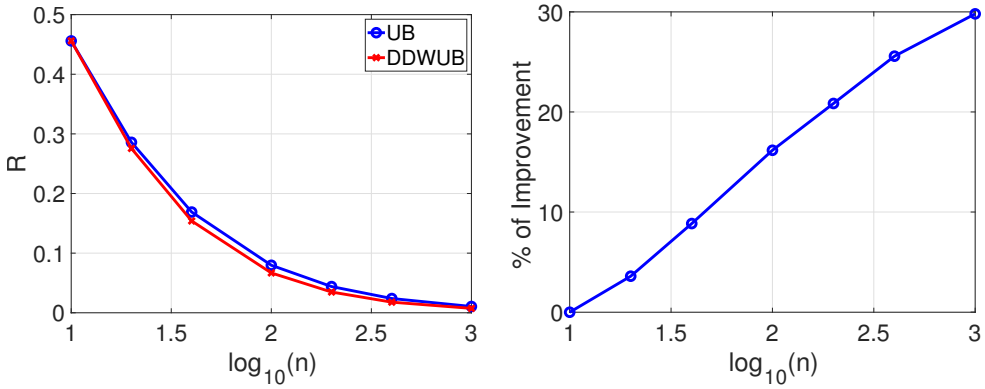


Figure 3. Scenario B: upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the UB and the DDWUB together with the percentage of improvement.

Scenario C.

The last scenario involves the unlucky case in which our set of m hypotheses is taken from all 2^n possible binary hypotheses over n data. Please note that the number of hypotheses with i errors in the 2^n possible binary hypotheses over n data are $\binom{n}{i}$. Then we force our m hypotheses to have at least one hypothesis for each possible value of the empirical error and consequently $m \geq n+1$. The remaining $m-n-1$ are taken from the 2^n possible binary hypotheses over n to approximate the distribution of the 2^n possible binary hypotheses. The result of this approach is that our set of hypotheses will be composed by $m = \sum_{j=0}^{n} \left\lceil \frac{\binom{n}{j}}{2^n} z \right\rceil$ hypotheses, with $z \in \mathbb{N}^*$, as follows

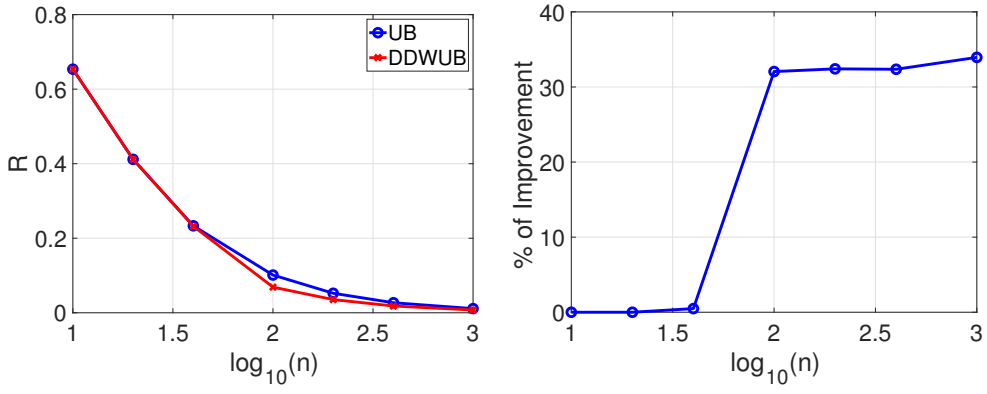
$$\begin{aligned}
&\hat{R}_1 = \cdots = \hat{R}_{\left\lceil \frac{\binom{n}{0}}{2^n} z \right\rceil} = 0,\\
&\hat{R}_{\left\lceil \frac{\binom{n}{0}}{2^n} z \right\rceil + 1} = \cdots = \hat{R}_{\left\lceil \frac{\binom{n}{0}}{2^n} z \right\rceil + \left\lceil \frac{\binom{n}{1}}{2^n} z \right\rceil} = \frac{1}{n},\\
&\quad \cdots\\
&\hat{R}_{\sum_{j=0}^{n-1} \left\lceil \frac{\binom{n}{j}}{2^n} z \right\rceil + 1} = \cdots = \hat{R}_{\sum_{j=0}^{n} \left\lceil \frac{\binom{n}{j}}{2^n} z \right\rceil} = 1.
\end{aligned}$$

Please note that for example, when $z = 1$ we obtain the Scenario B.

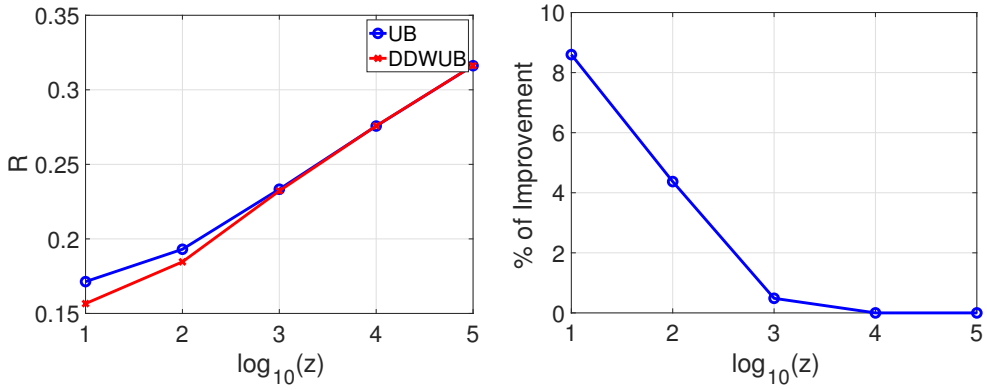
Also, in this scenario we set $\delta = 0.05$ and we set the numerical precision to $\epsilon = 0.0001$.

Figure 4a–c reports the estimated generalization error upper bound of the hypothesis with the smallest empirical error estimated with the UB and the DDWUB, together with the percentage of improvement in the same sub-scenarios of Scenario A.

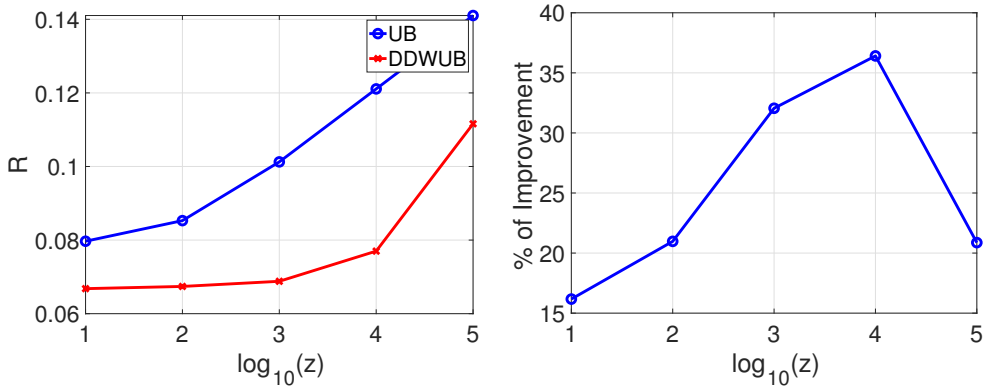
Figure 4 shows that even in this unlucky case, the DDWUB can remarkably outperform the UB.



(**a**) Sub-scenario C.1 ($z = 1000$)



(**b**) Sub-scenario C.2 ($n = 30$)



(**c**) Sub-scenario C.3 ($n = 100$)

Figure 4. Scenario C: upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the UB and the DDWUB together with the percentage of improvement.

5.1. The Importance of γ and θ

In this section, we would like to discuss the importance of γ and θ also by means of some numerical experiments supporting our claims.

Let us start with θ. From one side setting $\theta = 1$ would make the DDWUB degenerate in the UB eliminating all the benefits of using the DDWUB. From the other side setting $\theta = 0$, namely removing θ, is not possible because of the constraint on θ of Theorem 5 which does not allow us, in this case, to guarantee that the solution of the DDWUB always outperform, or in the worst case degenerates in, the UB. Hence, $\theta \in (0, 1)$ deals with the fact that we do not know the generalization error of our hypotheses, then we must estimate it, and consequently hypotheses with a small but close empirical error cannot be distinguished. Unfortunately, the constraint of Theorem 5 on θ is data-dependent and consequently we must resort to a suboptimal, but data independent, limitation on θ as reported in Lemma 8. The question which raises here is the practical difference of all these choices. For this reason, we consider the Scenario A previously defined with $\hat{R}_1 = 0.1$, $\hat{R}_2 = \nu$ and $\hat{R}_3 = \cdots = \hat{R}_m = 1$,

where we set $\delta = 0.05$, $n = 100$, and $m = 1000$. Then we vary $\nu \in \{0.1, 0.2, \cdots, 1\}$, and we reported in Figure 5 the comparison between the UB and the DDWUB with $\theta = 0$, $\theta = \hat{\theta}$ i.e., equal to the lower limit of the constraint of Theorem 5, $\theta = \hat{\hat{\theta}}$ i.e., equal to the lower limit of the constraint of Lemma 8, and $\theta = 1$.

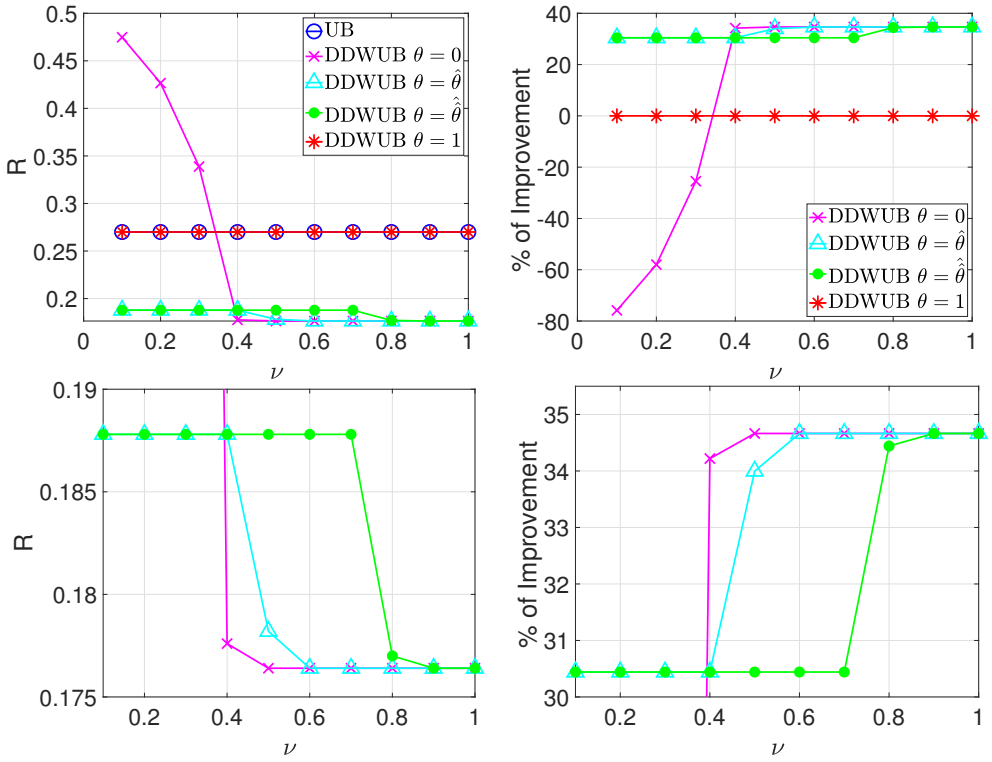


Figure 5. Scenario A ($\hat{R}_1 = 0.1$, $\hat{R}_2 = \nu$ and $\hat{R}_3 = \cdots = \hat{R}_m = 1$): upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the UB and the DDWUB with $\theta \in \{0, \hat{\theta}, \hat{\hat{\theta}}, 1\}$ together with the percentage of improvement when we set $\delta = 0.05$, $n = 100$, and $m = 1000$ and we vary $\nu \in \{0.1, 0.2, \cdots, 1\}$. The two figures above depict the whole range while the two figures below report a zoom on the most interesting parts.

From Figure 5 it is possible to derive some observations. Setting $\theta = 0$ in the DDWUB can result in worse estimates with respect to the ones of the UB since when $\hat{R}_2$ is close to $\hat{R}_1$ (small ν) we cannot distinguish between the first two hypotheses which are the ones with the lowest generalization error. As soon as ν grows the difference between $\hat{R}_1$ and $\hat{R}_2$ becomes statistical relevant and so setting $\theta = 0$ in the DDWUB results in better estimates with respect to the UB or even better that the ones of the DDWUB since θ in this particular scenario for large ν is useless. Setting $\theta = \hat{\theta}$ gives the best results and always outperform the UB while setting $\theta = \hat{\hat{\theta}}$ results in worse estimates but still better than the ones of the UB. Please note that for small and large ν, $\hat{\theta}$ and $\hat{\hat{\theta}}$ are equivalent while there is a middle range of ν for which $\hat{\theta}$ performs better $\hat{\hat{\theta}}$. The explanation of this phenomena can be derived using the observations of Section 3.2. The hypothesis that regulates $\hat{\theta}$ is the one corresponding to $\hat{R}_1$ (the one with the smallest empirical error). The hypothesis that regulates $\hat{\hat{\theta}}$, instead, is the one corresponding to $\hat{R}_2$ (the second-best hypothesis) if the distance between $\hat{R}_1$ and $\hat{R}_2$ is small, i.e., for small ν, while is the one corresponding to $\hat{R}_1$ if the distance between $\hat{R}_1$ and $\hat{R}_2$ is large. Consequently, when $\hat{R}_1$ and $\hat{R}_2$ are close it is indifferent to choose one or the other and the estimates of the DDWUB are almost equivalent. When, instead, the difference between $\hat{R}_1$ and $\hat{R}_2$ increases, $\hat{R}_2$, instead of $\hat{R}_1$, regulates θ and the DDWUB with $\hat{\theta}$ performs better than the DDWUB with $\hat{\hat{\theta}}$. Finally, when the distance between $\hat{R}_1$ and $\hat{R}_2$ is large, θ is regulated by $\hat{R}_1$ both for $\hat{\theta}$ and $\hat{\hat{\theta}}$ and consequently the corresponding estimates of the DDWUB are equivalent. Finally, setting $\theta = 1$ results in the UB.

Let us now consider γ. From one side setting $\gamma = 0$ would make the DDWUB degenerate in the UB eliminating all the benefits of using the DDWUB. From the other side setting $\gamma = \infty$ would result in splitting equally the confidence just over the hypotheses with estimated error less than θ,

not considering all the other hypotheses, which is our scope but unfortunately this is not possible because of the constraints of Lemma 7, which then result in the limitation over γ in Lemma 11. This is the reason we set, in the experiments, γ to the limits of what Lemma 11 allows. After that limit we do not know how the solution of the DDWUB behaves since we do not even know how to retrieve it. Setting a γ smaller than the maximum value allowed by Lemma 11 would diminish the performance of the DDWUB until the DDWUB degenerates in the UB. To support this statement we consider Scenario A with $\hat{R}_1 = \hat{R}_2 = 0$ and $\hat{R}_3 = \cdots = \hat{R}_m = 1$, then set $\delta = 0.05$, $n = 100$ and $m = 1000$, and we vary $\gamma \in \{10^{-5}\sqrt{n}, 10^{-4}\sqrt{n}, \cdots, 10^{-1}\sqrt{n}, \hat{\gamma}\}$, where $\hat{\gamma}$ is the limit defined in Lemma 11, and we reported the comparison between the DDWUB and the UB in Figure 6.

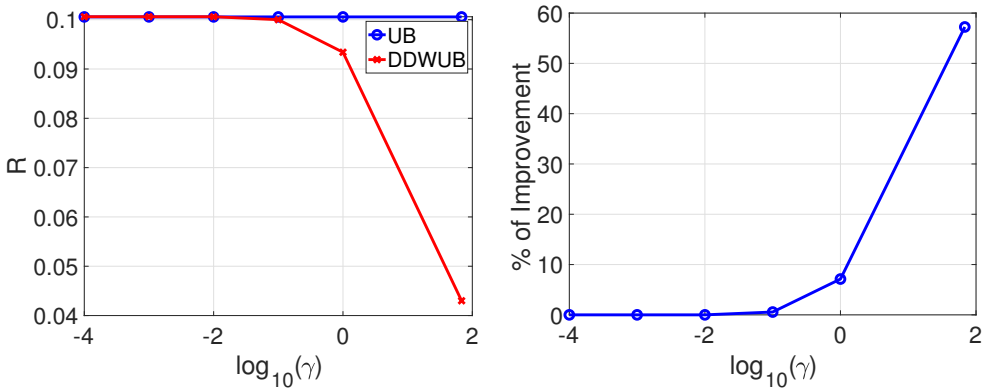


Figure 6. Scenario A ($\hat{R}_1 = \hat{R}_2 = 0$ and $\hat{R}_3 = \cdots = \hat{R}_m = 1$): upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the UB and the DDWUB together with the percentage of improvement when we set $n = 100$ and $m = 1000$ and we vary $\gamma \in \{10^{-5}\sqrt{n}, 10^{-4}\sqrt{n}, \cdots, 10^{-1}\sqrt{n}, \hat{\gamma}\}$, where $\hat{\gamma}$ is the limit defined in Lemma 11.

From Figure 6 it is possible to clearly observe that the maximum improvement is achieved when γ is maximum and consequently when $\gamma = \hat{\gamma}$. The same result can be observed in all the other scenarios.

6. What About the Computable Shell Decomposition Bounds?

In this section, we will show that the DDWUB also improves over the CSDB. Before starting the comparison, we must recall this milestone result.

Theorem 6 ([20]). *The following bounds hold*

$$\mathbb{P}\left\{R_i \leq \max\left\{r : r \in [0,1], \mathtt{kl}(\hat{R}_i||r) \leq \frac{\hat{\mathfrak{s}}(\lceil\lceil r\rceil\rceil, n, \delta) + \ln\left(\frac{4n}{\delta}\right)}{n}\right\} \forall i \in \mathcal{I}\right\} \geq 1-\delta,$$

where $\lceil\lceil r\rceil\rceil = {}^{\max[1,\lceil rn\rceil]}/_{n} \in \{{}^{1}/_{n}, \cdots, {}^{n}/_{n}\}$ *and if* $\lceil\lceil r\rceil\rceil = {}^{k}/_{n}$ *then* $r \in [{}^{k-1}/_{n}, {}^{k}/_{m}]$, $\mathtt{kl}(q||p) = q\ln({}^{q}/_{p}) + (1-q)\ln({}^{1-q}/_{1-p})$ *is the Kullback–Leibler divergence, and*

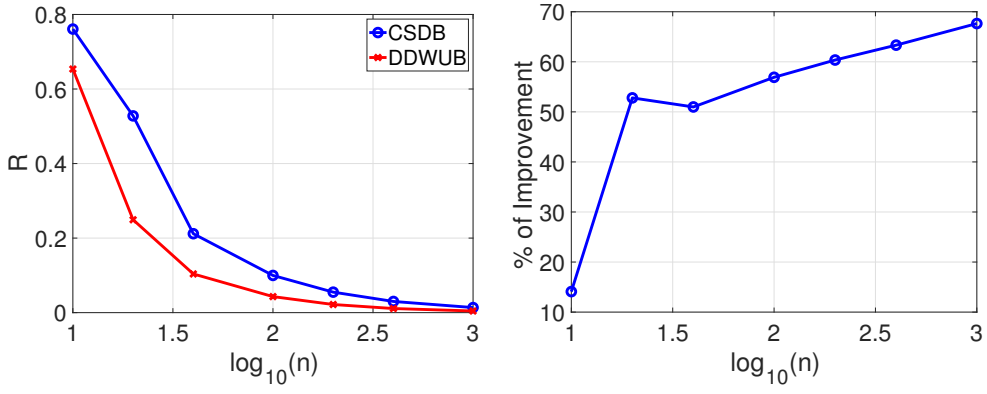
$$\hat{\mathfrak{s}}\left(\frac{k}{n}, n, \delta\right) = \ln\left(\max\left[1, 2\left|\left\{h_i : i \in \mathcal{I}, \left|\hat{R}_i - \frac{k}{n}\right| \leq \frac{1}{n} + \sqrt{\frac{\ln\left(\frac{16n^2}{\delta}\right)}{2n-1}}\right\}\right|\right]\right).$$

Let us consider the same scenario of Section 4.

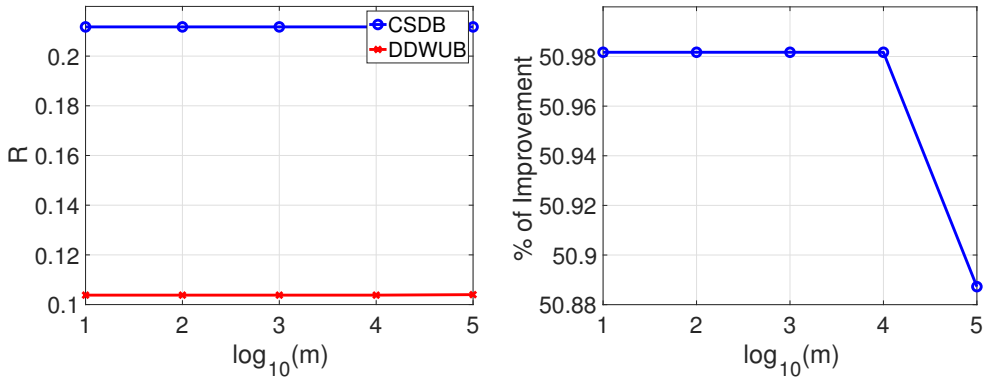
The DDWUB is able, at least asymptotically, to discard all risk associated with the hypotheses with high empirical error obtaining that for n large enough, the rate of convergence of the bound on R_1 is $O({}^{1}/_{n})$.

Instead, if we use the CSDB of Theorem 6 we get that for n large enough, the rate of convergence of the bound on R_1 is $O({}^{\ln(n)}/_{n})$. Moreover, note that $O({}^{\ln(n)}/_{n})$ is also the fastest possible rate of convergence for Theorem 6.

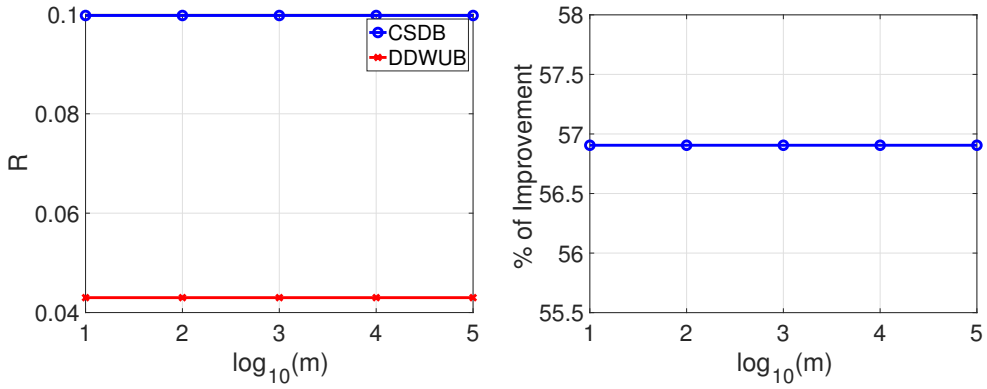
If instead of checking the asymptotic behavior of the DDWUB and the CSDB we check their finite sample behavior by means of numerical experiments as in Section 5, we can derive other interesting observations. Figures 7–10 (and the associated sub-figures) report the same comparison of Figures 1–4 in Section 5 but, instead of comparing the UB with the DDWUB, here we compare the CSDB with the DDWUB.



(**a**) Sub-scenario A.1 ($m = 1000$).

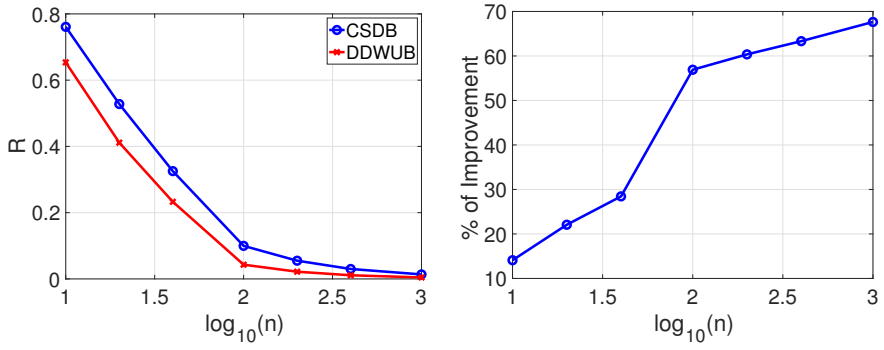


(**b**) Sub-scenario A.2 ($n = 30$).

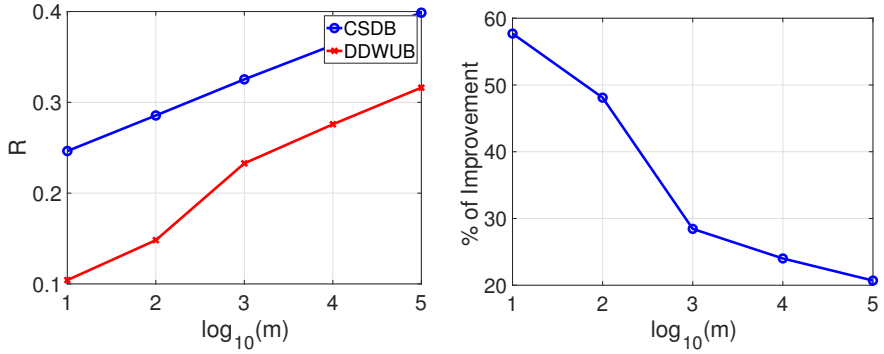


(**c**) Sub-scenario A.3 ($n = 100$).

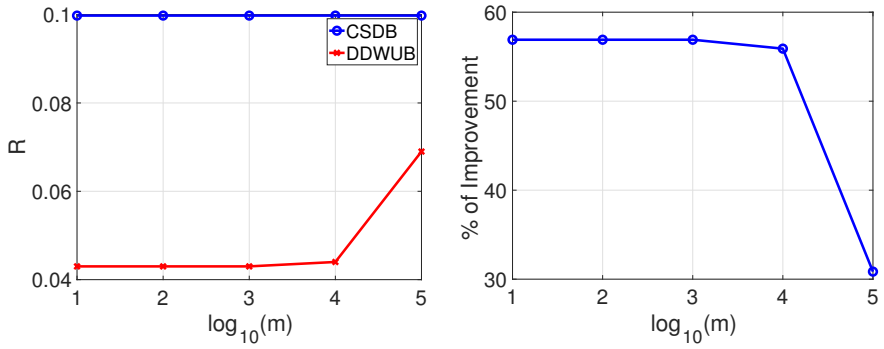
Figure 7. Scenario A ($\hat{R}_1 = \hat{R}_2 = 0$ and $\hat{R}_3 = \cdots = \hat{R}_m = 1$): upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the CSDB and the DDWUB together with the percentage of improvement.



(**a**) Sub-scenario A.1 ($m = 1000$)



(**b**) Sub-scenario A.2 ($n = 30$)



(**c**) Sub-scenario A.3 ($n = 100$)

Figure 8. Scenario A ($\hat{R}_1 = \hat{R}_2 = 0$ and $\hat{R}_3 = \cdots = \hat{R}_m = \frac{1}{2}$): upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the CSDB and the DDWUB together with the percentage of improvement.

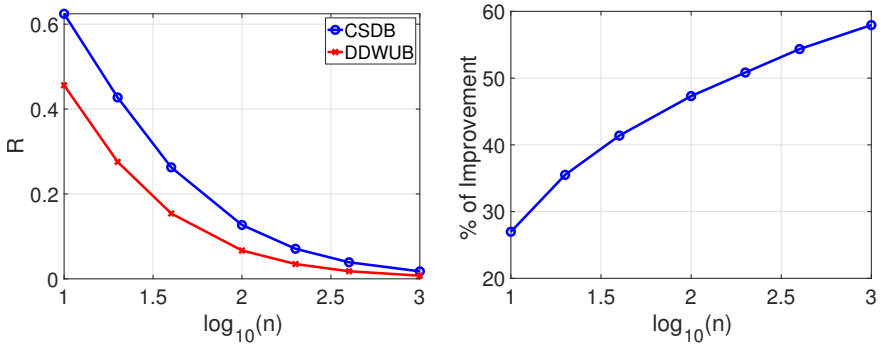
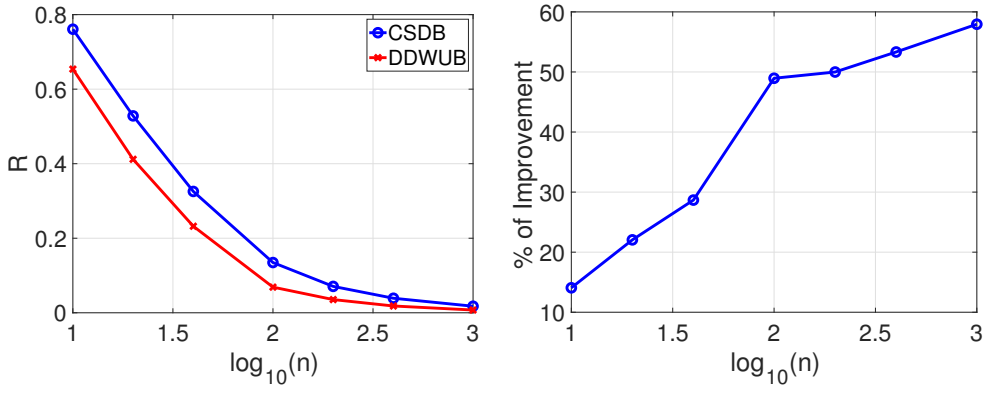
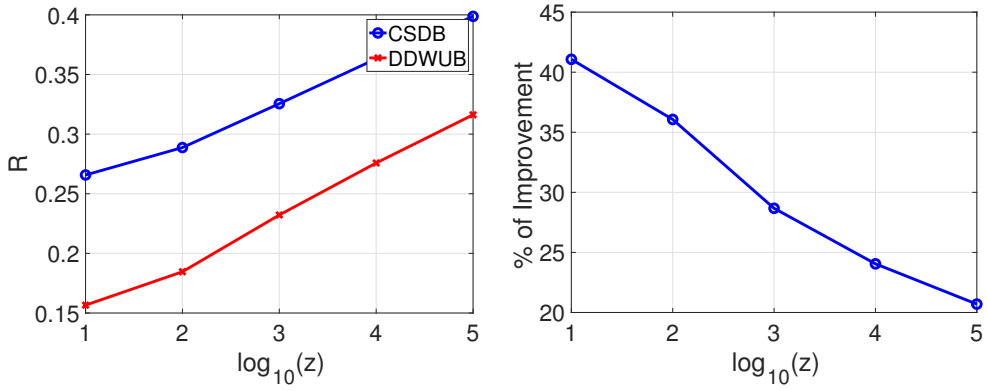


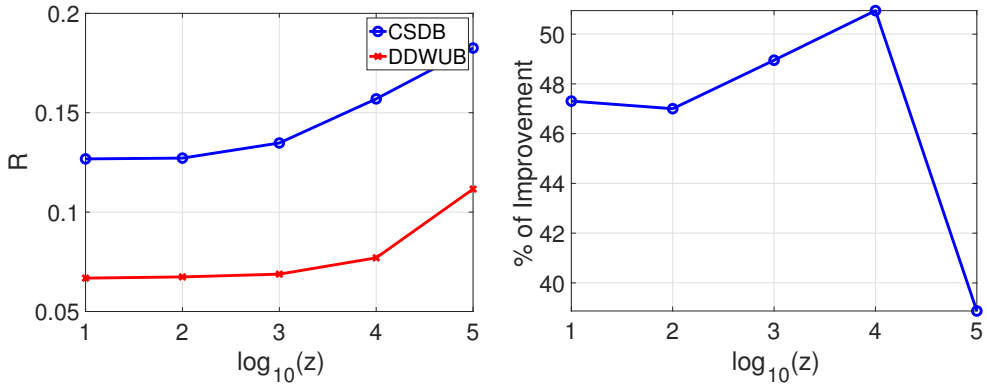
Figure 9. Scenario B: upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the CSDB and the DDWUB together with the percentage of improvement.



(**a**) Sub-scenario C.1 ($z = 1000$)



(**b**) Sub-scenario C.2 ($n = 30$)



(**c**) Sub-scenario C.3 ($n = 100$)

Figure 10. Scenario C: upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the CSDB and the DDWUB together with the percentage of improvement.

As it can be clearly seen from the results the DDWUB performs consistently better than the CSDB.

7. Improving the Computable Shell Decomposition Bounds

The purpose of this section is to demonstrate that the DDWUB can be exploited to improve known results in Statistical Learning Theory. In particular, this section we will show that DDWUB can be exploited to improve the CSDB.

The proof of the CSDB of Theorem 6 relies on two main results, reported in the next theorem, combined with the UB.

Theorem 7 ([20]). *The following bounds hold*

$$\mathbb{P}\left\{\mathtt{kl}(\hat{R}_i||R_i) \leq \frac{\ln(|\mathcal{H}|)+\ln\left(\frac{2}{\delta}\right)}{n}\ \forall i \in \mathcal{I}\right\} \geq 1-\delta,$$
$$\mathbb{P}\left\{\mathtt{S}\left(\frac{k}{n},n\right) \leq \hat{\mathtt{S}}\left(\frac{k}{n},n,2n\delta\right)\right\} \geq 1-\delta,$$

where

$$\mathtt{S}\left(\frac{k}{n},n\right) = \ln\left(\max\left[1,2\left|\left\{h_i : i \in \mathcal{I}, R_i \in \left[\frac{k-1}{n},\frac{k}{n}\right]\right\}\right|\right]\right)$$

and $\hat{\mathtt{S}}\left(\frac{k}{n},n,\delta\right)$ *is defined as in Theorem 6.*

By splitting the hypotheses in shells based on their generalization error, namely $\mathcal{H}_r = \{h_i : i \in \mathcal{I}, R_i \in [{}^{k-1}/_n, {}^{k}/_n]\}$ with $r \in \{{}^{1}/_n, {}^{2}/_n, \cdots, 1\}$, by combining the two probabilistic bounds of Theorem 7, by using the UB, and by considering the worst case scenario, the result of Theorem 6 is derived. Consequently, [20] consider $2n$ probabilistic bounds, two for each of one the shells, and spread the confidence (risk) equally over them.

We propose, instead, to use the same approach of the DDWUB in the CSDB. Instead of spreading the confidence equally over the $2n$ probabilistic bounds, we spread the confidence over them based on the maximum generalization error of the function in each of the n shells to which the bounds refer. The results is reported in the following lemma.

Lemma 12. *The following bound holds*

$$\mathbb{P}\left\{R_i \leq \max\left\{r : r \in [0,1], \mathtt{kl}(\hat{R}_i||r) \leq \frac{\hat{\mathtt{S}}(\lceil\lceil r\rceil\rceil, n, np(\lceil\lceil r\rceil\rceil)\delta)+\ln\left(\frac{4}{\delta p(\lceil\lceil r\rceil\rceil)}\right)}{n}\right\}\ \forall i \in \mathcal{I}\right\} \geq 1-\delta,$$

where

$$p(r) = \frac{e^{-\frac{nr}{\ln(n)}}}{\sum_{i=1}^{n} e^{-\frac{i}{\ln(n)}}}.$$

The proof is the simple application of the concepts behind the DDWUB. θ is not needed since for each one of the shells we exactly know by definition the maximum generalization error of the functions inside it. γ is set to ${}^{n}/_{\ln(n)}$ and the reason is the following one. The rate of convergence of CSDB is $O(\sqrt{{}^{\ln(n)}/_n})$ in the general case and $O({}^{\ln(n)}/_n)$ when $\hat{R}_i = 0$. Let us study instead the rate of convergence of the bound of Lemma 12. Thanks to the Geometric Series we can state that

$$p(r) = \frac{e^{-\frac{nr}{\ln(n)}}}{\sum_{i=1}^{n} e^{-\frac{i}{\ln(n)}}} = e^{-\frac{nr}{\ln(n)}} \frac{1-e^{\frac{1}{\ln(n)}}}{1-e^{\frac{n}{\ln(n)}}}$$

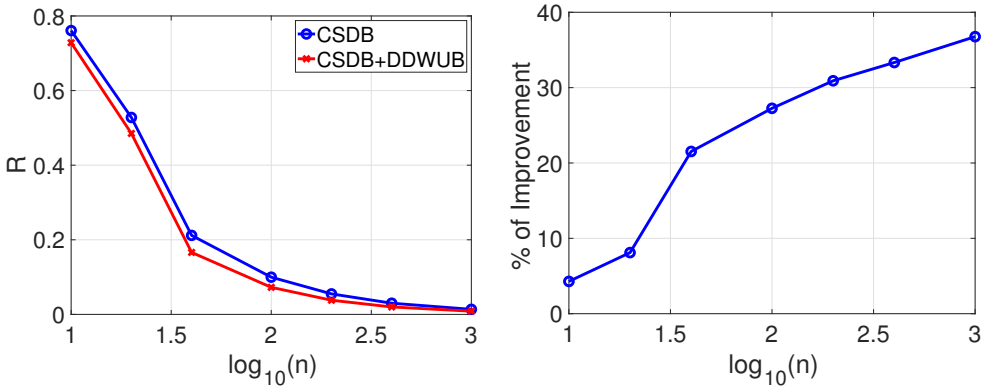
For n large enough we can state that

$$\ln\left(\frac{1}{p(r)}\right) \approx \frac{nr}{\ln(n)} + \ln(\ln(n))$$

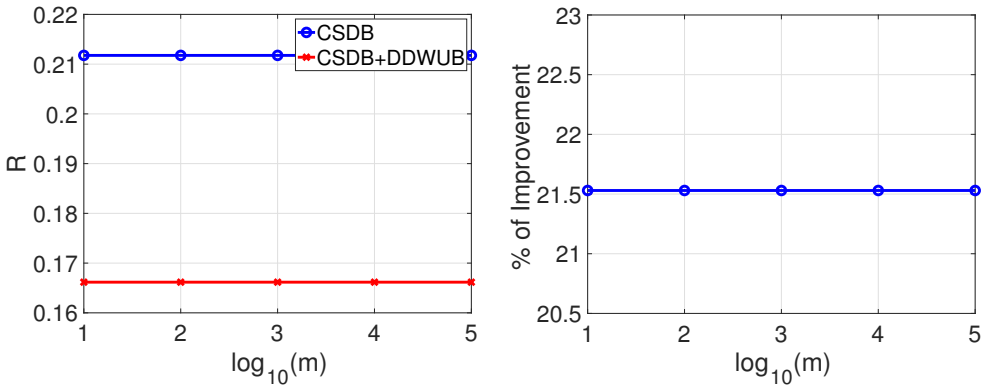
Consequently, the rate of convergence of the bound of Lemma 12 is $O(\sqrt{{}^{\ln(\ln(n))}/_n})$ in the general case and $O({}^{\ln(\ln(n))}/_n)$ when $\hat{R}_i = 0$. This means that for $\gamma = {}^{n}/_{\ln(n)}$ the rate of convergence of the bound of Lemma 12 of better than the one of CSDB.

It could be possible to improve the bound with a different values of γ and θ or to prove that for particular values of γ and θ, Lemma 12 is always better than the CSDB but this is beyond of the scope in this paper.

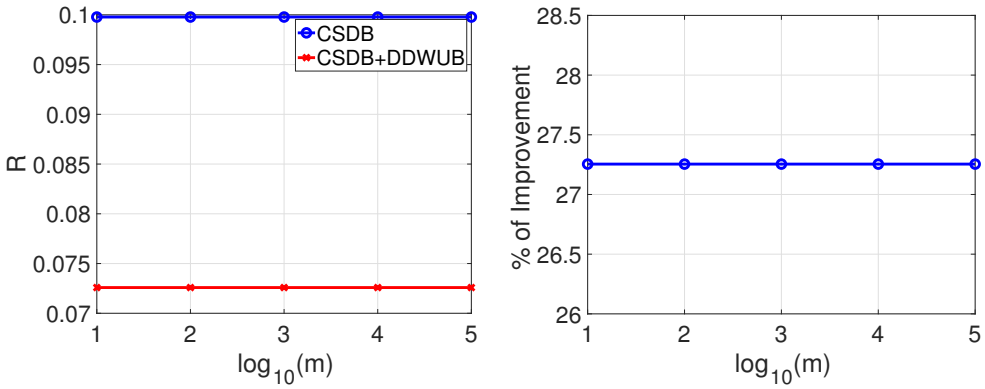
If, instead, we compare the finite sample behavior of the CSDB and the Lemma 12 (that we will call CSDB+DDWUB) by means of numerical experiments as in Section 5, we can observe the possible benefit of using DDWUB in CSDB, a well-known result of Statistical Learning Theory. Figures 11–14 (and the associated sub-figures) report the same comparison of Figures 1–4 in Section 5 but, instead of comparing the UB with the DDWUB, here we compare the CSDB with the CSDB+DDWUB.



(a) Sub-scenario A.1 ($m = 1000$).

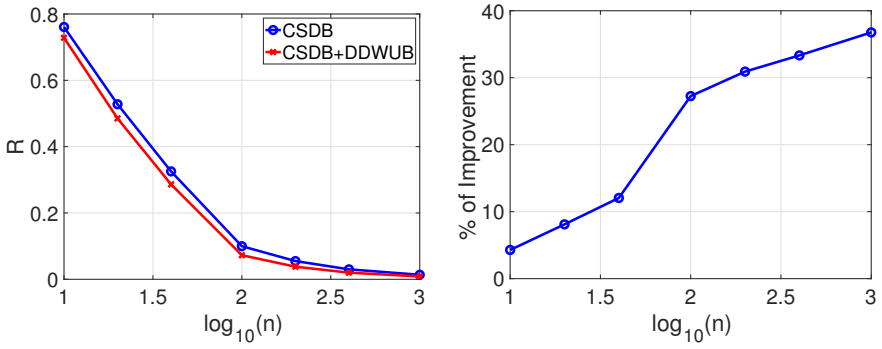


(b) Sub-scenario A.2 ($n = 30$).

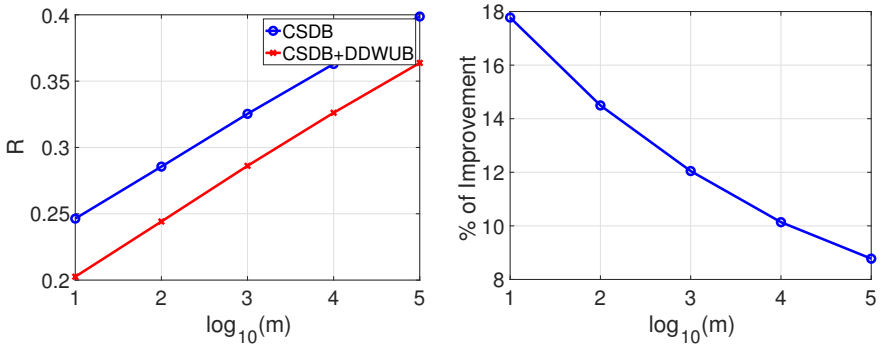


(c) Sub-scenario A.3 ($n = 100$).

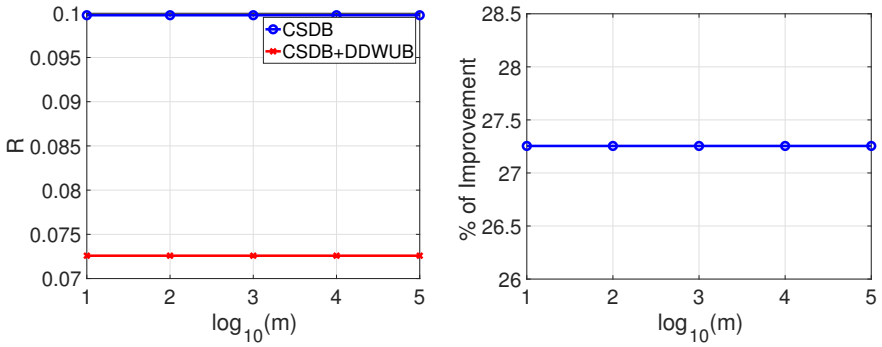
Figure 11. Scenario A ($\hat{R}_1 = \hat{R}_2 = 0$ and $\hat{R}_3 = \cdots = \hat{R}_m = 1$): upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the CSDB and the CSDB+DDWUB together with the percentage of improvement.



(**a**) Sub-scenario A.1 ($m = 1000$)



(**b**) Sub-scenario A.2 ($n = 30$)



(**c**) Sub-scenario A.3 ($n = 100$)

Figure 12. Scenario A ($\hat{R}_1 = \hat{R}_2 = 0$ and $\hat{R}_3 = \cdots = \hat{R}_m = \frac{1}{2}$): upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the CSDB and the CSDB+DDWUB together with the percentage of improvement.

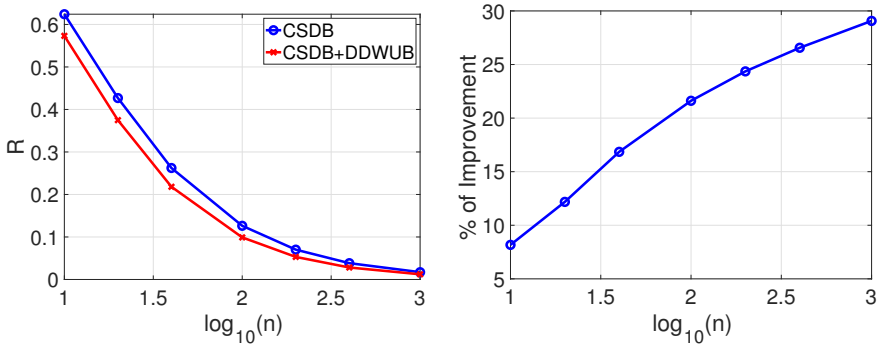
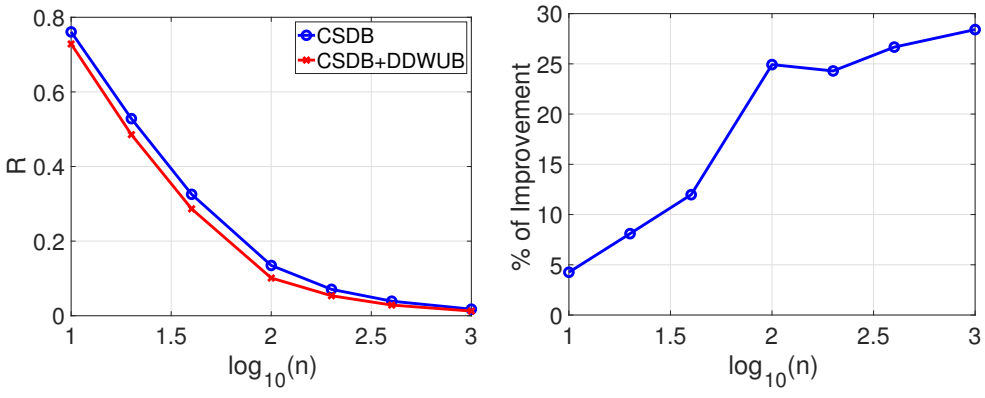
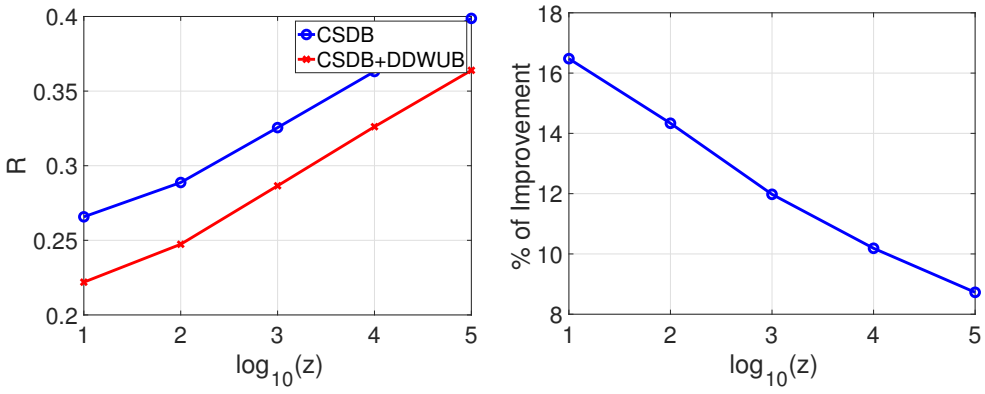


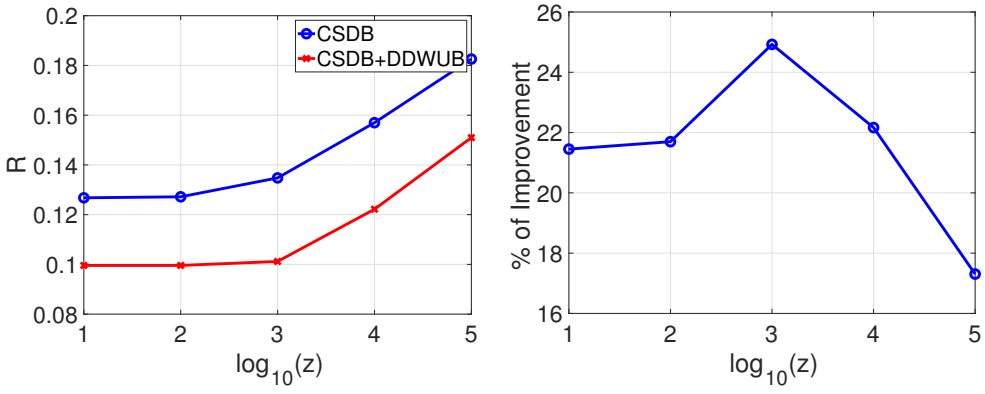
Figure 13. Scenario B: upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the CSDB and the CSDB+DDWUB together with the percentage of improvement.



(**a**) Sub-scenario C.1 ($z = 1000$)



(**b**) Sub-scenario C.2 ($n = 30$)



(**c**) Sub-scenario C.3 ($n = 100$)

Figure 14. Scenario C: upper bound of the generalization error of the hypothesis with the smallest empirical error computed with the CSDB and the CSDB+DDWUB together with the percentage of improvement.

As it can be clearly seen from the results the CSDB+DDWUB performs consistently better than the CSDB.

8. Conclusions and Discussion

In this work we derived, for an arbitrary finite hypothesis space, a fully empirical new upper bound on the generalization error of the hypothesis of minimal training error. As noted in the paper, although we presented just the upper bound, our result can be easily generalized also to lower or both upper and lower bounds.

In particular we depicted a quite general framework under which it is possible to improve the Union Bound with a distribution depended weighting strategy associated with the risk of each choice. Then we stated the conditions under which the proposed Distribution-Dependent Weighed Union Bound is always tighter than the one based on the Union Bound. We showed that these conditions

are quite easy to satisfy. By means of both closed form and numerical results we demonstrated that Distribution-Dependent Weighed Union Bound is consistently tighter than the Union Bound in different scenarios. Finally, we showed that the Distribution-Dependent Weighed Union Bound is also able to improve over the Computable Shell Decomposition Bound, another quite powerful distribution-dependent Union Bound.

The results of this work are quite promising and pave the way toward many different future improvements. One is surely to derive a class of weighting strategies which satisfies behind the Distribution-Dependent Weighed Union Bound. Another one is to understand how to exploit the results of this work for improving all the results in Statistical Learning Theory where the Union Bound is employed. A final one, and probably the most important one, is how to extend and apply our results to the infinite-dimensional hypothesis spaces. This extension, which is obviously not trivial, has multiple alternatives which can be speculated. The first, and naive, approach would be to plug our approach into the compression bound which already deals with the infinite-dimensional case, exploiting the concept of compression, and exploits naively the union bound. The second approach, less intuitive, would be to split the hypothesis spaces in a finite number of shells, estimate the size of each shell with classical bounds based on the Vapnik–Chervonenkis or Rademacher Complexity theories and then exploit our Distribution-Dependent Weighed Union Bound to pay the price of the choice of one of the shells. The last, and more challenging, approach would be to plug our weighting strategy directly in the derivation of the Vapnik–Chervonenkis or Rademacher Complexity bases bounds shrinking then the measures of complexity as alternative to the localization approaches.

Author Contributions: Conceptualization, L.O. and S.R.; Formal analysis, L.O.; Investigation, L.O. and S.R.; Methodology, L.O.; Software, L.O.; Writing—original draft, L.O.; Writing—review & editing, L.O. and S.R. Authors have equally contributed to this work. All authors have read and agreed to the published version of the manuscript.

Funding: No funding to declare.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: See Appendix D.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Known Results

Known bounds and results exploited in the paper are included in this section.

Theorem A1 ([16]). *The following bounds hold*

$$\mathbb{P}\left\{R \geq \hat{R} - \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right\} \geq 1-\delta, \quad \mathbb{P}\left\{R \leq \hat{R} + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right\} \geq 1-\delta.$$

Theorem A2 ([15]). *The following bounds hold*

$$\mathbb{P}\left\{R \geq \begin{cases} \mathtt{F}^{-1}(\delta; n\hat{R}, n-n\hat{R}+1) & \hat{R} \in \left\{\frac{1}{n}, \frac{2}{n}, \cdots, 1\right\} \\ 0 & \hat{R}=0 \end{cases}\right\} \geq 1-\delta,$$

$$\mathbb{P}\left\{R \leq \begin{cases} \mathtt{F}^{-1}(1-\delta; n\hat{R}+1, n-n\hat{R}) & \hat{R} \in \left\{0, \frac{1}{n}, \cdots, \frac{n-1}{n}\right\} \\ 1 & \hat{R}=1 \end{cases}\right\} \geq 1-\delta.$$

Theorem A3 ([35]). *If $n \in \mathbb{N}^*$, the following bounds hold*

$$\sqrt{2\pi n} n^n e^{-n} \leq n! \leq \sqrt{2\pi n} n^n e^{-n} e^{\frac{1}{12n}}.$$

Theorem A4 ([6]). *If $n \in \mathbb{N}^*$, $p \in [0,1]$, $q \in \{0, 1/n, 2/n, \cdots, 1\}$, and $\hat{R} < p$, the following bound holds*

$$\mathtt{F}(p; nq, n-nq+1) \leq e^{-n\mathtt{kl}(q||p)}$$

where

$$\mathtt{F}(p; nq, n-nq+1) = \sum_{i=0}^{nq} \binom{n}{i} p^i (1-p)^{n-i},$$

is the beta cumulative density function, and

$$\mathtt{kl}(q||p) = q \ln\left(\frac{q}{p}\right) + (1-q)\ln\left(\frac{1-q}{1-p}\right)$$

is the Kullback–Leibler divergence.

Appendix B. Proofs

Proofs of our results are included in this section.

Proof. (Lemma 2) Please note that under the assumptions of the lemma, $\forall r_1, \cdots, r_m \in [0,1]$, and $\forall i \in \mathcal{I} \setminus i^*$ and $\forall r'_k, r''_k \in [0,1]$ such that $r'_k < r''_k$

$$\sqrt{\frac{\log\left(\frac{2\sum_{j\in\mathcal{I}\setminus k} e^{-\gamma \max[\theta, r_j]} + e^{-\gamma \max[\theta, r'_k]}}{\delta e^{-\gamma \max[\theta, r_i]}}\right)}{2n}} - \sqrt{\frac{\log\left(\frac{2\sum_{j\in\mathcal{I}\setminus k} e^{-\gamma \max[\theta, r_j]} + e^{-\gamma \max[\theta, r''_k]}}{\delta e^{-\gamma \max[\theta, r_i]}}\right)}{2n}} \geq 0,$$

then the statement of the lemma is proved. □

Proof. (Lemma 3) Please note that under the assumptions of the lemma

$$0 < \min\left[1, \hat{R}_i + \sqrt{\frac{\log\left(\frac{2\sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, r_j]}}{\delta e^{-\gamma \max[\theta, r_{i^*}]}}\right)}{2n}}\right] \leq 1.$$

Note also that if

$$\gamma \leq 2\sqrt{n}$$
$$p_{i^*} = \frac{e^{-\gamma r_{i^*}}}{\sum_{j\in\mathcal{I}} e^{-\gamma r_j}}$$

then

$$\frac{\partial \sqrt{\frac{\log\left(\frac{2\sum_{j\in\mathcal{I}} e^{-\gamma r_j}}{\delta e^{-\gamma r_{i^*}}}\right)}{2n}}}{\partial r_{i^*}} = \frac{\gamma p_{i^*}(1-p_{i^*})}{4np_{i^*}\sqrt{\frac{\ln\left(\frac{2}{\delta p_{i^*}}\right)}{2n}}} < \frac{(1-p_{i^*})}{2\sqrt{\frac{\ln\left(\frac{2}{\delta p_{i^*}}\right)}{2}}} < 1,$$

then the statement of the lemma is proved. □

Proof. (Theorem 4) Let us define

$$\vartheta = \min\left[1, \hat{R}_{i^*} + \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right].$$

Let us suppose that

$$r_i \leq \vartheta, \quad \forall i \in \mathcal{I} \setminus i^*.$$

If we set

$$r^*_{i^*} = \vartheta,$$

we have, thanks to the hypothesis of the theorem that

$$\frac{e^{-\gamma \max[\theta, r_i]}}{\sum_{j \in \mathcal{I}} e^{-\gamma \max[\theta, r_j]}} = \frac{1}{m}, \quad \forall i \in \mathcal{I},$$

then $r^*_{i^*}$ is a fixed point and since it is unique by Lemma 3 we have that $r^*_{i^*} \leq \vartheta$.

Let us suppose now that

$$r_i \leq \vartheta, \quad \forall i \in \mathcal{I} \setminus \{i^*, k\}, \quad r_k > \vartheta, \quad k \in \mathcal{I} \setminus i^*.$$

Thanks to the hypothesis of the theorem, we can state that

$$\frac{e^{-\gamma \max[\theta, r_{i^*}]}}{\sum_{j \in \mathcal{I} \setminus \{i^*, k\}} e^{-\gamma \max[\theta, r_j]} + e^{-\gamma \max[\theta, r_{i^*}]} + e^{-\gamma \max[\theta, r_k]}} > \frac{e^{-\gamma \max[\theta, r_{i^*}]}}{\sum_{j \in \mathcal{I} \setminus i^*} e^{-\gamma \theta} + e^{-\gamma \max[\theta, r_{i^*}]}},$$

and, consequently, we can also state that

$$\begin{aligned} r_{i^*} &= \min \left[1, \hat{R}_{i^*} + \sqrt{\frac{\log\left(\frac{2 \sum_{j \in \mathcal{I} \setminus \{i^*, k\}} e^{-\gamma \max[\theta, r_j]} + e^{-\gamma \max[\theta, r_{i^*}]} + e^{-\gamma \max[\theta, r_k]}}{\delta e^{-\gamma \max[\theta, r_{i^*}]}} \right)}{2n}} \right] \\ &\leq \min \left[1, \hat{R}_{i^*} + \sqrt{\frac{\log\left(\frac{2 \sum_{j \in \mathcal{I} \setminus i^*} e^{-\gamma \theta} + e^{-\gamma \max[\theta, r_{i^*}]}}{\delta e^{-\gamma \max[\theta, r_{i^*}]}} \right)}{2n}} \right]. \end{aligned}$$

Consequently, by exploiting the same reasoning exploited before, $r^*_{i^*} \leq \vartheta$.

By induction, the statement of the theorem is proved. □

Proof. (Lemma 4) Please note that

$$\min[R_1, \cdots, R_m] \geq \max \left[0, \min\left[\hat{R}_1, \cdots, \hat{R}_m\right] - \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}} \right].$$

Then we can state that

$$\hat{R}_{i^*} \leq \min[R_1, \cdots, R_m] + \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}},$$

and consequently, the statement of the theorem is proved. □

Proof. (Theorem 5) Let us define

$$\vartheta = \mho\left(\hat{R}_i, \frac{\delta}{2m}\right).$$

Let us suppose that

$$r_j \leq \vartheta, \quad \forall j \in \mathcal{I} \setminus i.$$

If we set

$$r^*_i = \vartheta,$$

we have, thanks to the hypothesis of the theorem that

$$f_1(r_1, \cdots, r_{i-1}, r^*_i, r_{i+1}, \cdots, r_m) = \cdots = f_m(r_1, \cdots, r_{i-1}, r^*_i, r_{i+1}, \cdots, r_m) = \frac{1}{m},$$

then r^*_i is a fixed point and since it is unique by Lemma 7 we have that $r^*_i \leq \vartheta$.

Let us suppose now that

$$r_j \leq \vartheta, \quad \forall j \in \mathcal{I} \setminus \{i,k\}, \quad r_k > \vartheta, \quad k \in \mathcal{I} \setminus i.$$

Thanks to the hypothesis of the theorem, we can state that

$$f_i(\vartheta, \cdots, \vartheta, r_i, \vartheta, \cdots, \vartheta, r_k, \vartheta, \cdots, \vartheta) > f_i(\vartheta, \cdots, \vartheta, r_i, \vartheta, \cdots, \vartheta),$$

and, consequently, we can also state that

$$\begin{aligned} r_i &= \mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(\vartheta, \cdots, \vartheta, r_i, \vartheta, \cdots, \vartheta, r_k, \vartheta, \cdots, \vartheta)}{2}\right) \\ &\leq \mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(\vartheta, \cdots, \vartheta, r_i, \vartheta, \cdots, \vartheta)}{2}\right). \end{aligned}$$

Consequently, by exploiting the same reasoning exploited before, $r_i^* \leq \vartheta$.

By induction, the statement of the theorem is proved. □

Proof. (Lemma 8) Please note that

$$\min[R_1, \cdots, R_m] \geq L\left(\min\left[\hat{R}_1, \cdots, \hat{R}_m\right], \frac{\delta}{2m}\right).$$

Then we can state that

$$\hat{R}_{i^*} \leq \mathtt{L}^{-1}\left(\min[R_1, \cdots, R_m], \frac{\delta}{2m}\right),$$

and consequently, the statement of the theorem is proved. □

Proof. (Lemma 9) To prove the theorem we first just have to note that for a fixed θ, the problem can be solved with Algorithm 1 and that its solution is $r_{i^*}^*(\theta)$ and $r_j(\theta) = \mathtt{L}\left(\hat{R}_j, \frac{\delta}{2m}\right)$ with $j \in \mathcal{I} \setminus i^*$. Then, searching for the largest fixed point, varying θ, such that

$$\theta = U\left(L^{-1}\left(\min[r_1(\theta), \cdots, r_{i^*-1}(\theta), r_{i^*}^*(\theta), r_{i^*+1}(\theta), \cdots, r_m(\theta)], \frac{\delta}{2m}\right), \frac{\delta}{2m}\right)$$

gives the proposed algorithm. □

Proof. (Lemma 10) For what concerns Corollary 2, we must prove that $\forall r_1, \cdots, r_m \in [0,1]$

$$f_i(r_1, \cdots, r_m) \in (0,1) \quad \forall i \in \mathcal{I}, \quad \sum_{i \in \mathcal{I}} f_i(r_1, \cdots, r_m) = 1.$$

These two properties are trivially provable by definition.

For what concerns Theorem 5, instead, we must prove that for $\theta \in [0,1]$, $\forall r_1, \cdots, r_m \in [0,1]$, and $\forall r_j', r_j'' \in [0,\theta]$

$$f_j(r_1, \cdots, r_{j-1}, r_j', r_{j+1}, \cdots, r_m) - f_j(r_1, \cdots, r_{j-1}, r_j'', r_{j+1}, \cdots, r_m) = 0,$$

with $j \in \mathcal{I}$. Moreover, we must prove that $\forall r_1, \cdots, r_m \in [0,\theta]$ and $\forall j \in \mathcal{I}$

$$f_j(r_1, \cdots, r_m) = \frac{1}{m}.$$

These properties are trivially provable by definition. Then we must prove that $\forall r_1, \cdots, r_m \in [0,1]$, $\forall j \in \mathcal{I} \setminus i, \forall r_j', r_j'' \in (\theta, 1]$ such that $r_j' < r_j''$

$$f_i(r_1, \cdots, r_{j-1}, r_j', r_{j+1}, \cdots, r_m) - f_i(r_1, \cdots, r_{j-1}, r_j'', r_{j+1}, \cdots, r_m) < 0.$$

Then, let us note that

$$\frac{e^{-\gamma \max[\theta,r_i]}}{e^{-\gamma \max[\theta,r''_j]}+\sum_{k\in\mathcal{I}\setminus j} e^{-\gamma \max[\theta,r_k]}} - \frac{e^{-\gamma \max[\theta,r_i]}}{e^{-\gamma \max[\theta,r'_j]}+\sum_{k\in\mathcal{I}\setminus j} e^{-\gamma \max[\theta,r_k]}}$$
$$> \frac{e^{-\gamma \max[\theta,r_i]}}{e^{-\gamma \max[\theta,r'_j]}+\sum_{k\in\mathcal{I}\setminus j} e^{-\gamma \max[\theta,r_k]}} - \frac{e^{-\gamma \max[\theta,r_i]}}{e^{-\gamma \max[\theta,r'_j]}+\sum_{k\in\mathcal{I}\setminus j} e^{-\gamma \max[\theta,r_k]}} = 0,$$

consequently, the property is proven. □

Proof. (Lemma 11) For what concerns Corollary 2 and Theorem 5, the proof is a trivial consequence of Lemma 10.

For what concerns Lemma 6, we have to prove that $\forall \hat{r}_i \in \{0, 1/m, \cdots, 1\}$, $\forall r_1, \cdots, r_m \in [0,1]$, and $\forall j \in \mathcal{I} \setminus i$ and $\forall r'_j, r''_j \in [0,1]$ such that $r'_j < r''_j$

$$\mathtt{U}\left(\hat{r}_i, \frac{\delta f_i(r_1,\cdots,r_{j-1},r'_j,r_{j+1},\cdots,r_m)}{2}\right) - \mathtt{U}\left(\hat{r}_i, \frac{\delta f_i(r_1,\cdots,r_{j-1},r''_j,r_{j+1},\cdots,r_m)}{2}\right) \geq 0,$$

where $i \in \mathcal{I}$. Let us define

$$g_i(r_1,\cdots,r_m) = \frac{e^{-\gamma r_i}}{\sum_{j\in\mathcal{I}} e^{-\gamma r_j}}, \quad i \in \mathcal{I},$$

and note that

$$\frac{\partial g_i(r_1,\cdots,r_m)}{\partial r_j} = \gamma g_i(r_1,\cdots,r_m) g_j(r_1,\cdots,r_m), \quad i \in \mathcal{I},\ j \in \mathcal{I}\setminus i.$$

Combining this property with Proposition A1 we obtain the result.

For what concerns Lemma 7, we must prove that $\forall r'_i, r''_i \in [0,1]$ such that $r'_i < r''_i$

$$\frac{\mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(r_1,\cdots,r''_i,\cdots,r_m)}{2}\right) - \mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(r_1,\cdots,r'_i,\cdots,r_m)}{2}\right)}{r''_i - r'_i} < 1,$$
$$\mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(r_1,\cdots,r_{i-1},0,r_{i+1},\cdots,r_m)}{2}\right) > 0,$$
$$\mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(r_1,\cdots,r_{i-1},1,r_{i+1},\cdots,r_m)}{2}\right) < 1.$$

The second and the third properties are trivially true if $\hat{R}_i \neq 1$. For the first, instead, note that

$$\frac{\partial g_i(r_1,\cdots,r_m)}{\partial r_i} = -\gamma g_i(r_1,\cdots,r_m)(1 - g_i(r_1,\cdots,r_m)), \quad i \in \mathcal{I}.$$

Then we can state, thanks also to Proposition A1 that

$$\frac{\mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(r_1,\cdots,r''_i,\cdots,r_m)}{2}\right) - \mathtt{U}\left(\hat{R}_i, \frac{\delta f_i(r_1,\cdots,r'_i,\cdots,r_m)}{2}\right)}{r''_i - r'_i}$$
$$\leq \max_{\hat{r}\in\left\{0,\frac{1}{n},\cdots,\frac{n-1}{n}\right\},\ p\in(0,1)} \frac{\mathtt{B}(n\hat{r}+1, n-n\hat{r})}{\left(\mathtt{U}\left(\hat{r},\frac{\delta p}{2}\right)\right)^{n\hat{r}}\left(1-\mathtt{U}\left(\hat{r},\frac{\delta p}{2}\right)\right)^{n-n\hat{r}-1}} \frac{\delta}{2}\gamma p(1-p)$$

Since this quantity must be strictly smaller than one we have that

$$\gamma < \min_{\hat{r}\in\left\{0,\frac{1}{n},\cdots,\frac{n-1}{n}\right\},\ p\in(0,1)} \frac{\left(\mathtt{U}\left(\hat{r},\frac{\delta p}{2}\right)\right)^{n\hat{r}}\left(1-\mathtt{U}\left(\hat{r},\frac{\delta p}{2}\right)\right)^{n-n\hat{r}-1}}{\mathtt{B}(n\hat{r}+1, n-n\hat{r})\frac{\delta}{2}p(1-p)},$$

Note also that by denoting

$$\hat{s} = n\hat{r}, \quad U = U\left(\hat{r}, \frac{\delta p}{2}\right),$$

then, thanks to Theorems A3 and A4 we can state that

$$
\begin{aligned}
&B(\hat{s}+1,n-\hat{s})\frac{1}{\mathtt{U}^{\hat{s}}(1-\mathtt{U})^{n-\hat{s}-1}}p=\frac{\hat{s}!(n-\hat{s}-1)!}{n!}\frac{1}{\mathtt{U}^{\hat{s}}(1-\mathtt{U})^{n-\hat{s}-1}}p\\
&\leq\frac{\sqrt{2\pi\hat{s}}\hat{s}^{\hat{s}}e^{-\hat{s}}e^{\frac{1}{12\hat{s}}}\sqrt{2\pi(n-\hat{s}-1)}(n-\hat{s}-1)^{n-\hat{s}-1}e^{-(n-\hat{s}-1)}e^{\frac{1}{12(n-\hat{s}-1)}}}{\sqrt{2\pi n}n^{n}e^{-n}}\frac{1}{\mathtt{U}^{\hat{s}}(1-\mathtt{U})^{n-\hat{s}-1}}p\\
&=\sqrt{2\pi}e^{1+\frac{1}{12\hat{s}}+\frac{1}{12(n-\hat{s}-1)}}\frac{\sqrt{\hat{s}}\hat{s}^{\hat{s}}\sqrt{(n-\hat{s}-1)}(n-\hat{s}-1)^{n-\hat{s}-1}}{\sqrt{n}n^{n}}\frac{1}{\mathtt{U}^{\hat{s}}(1-\mathtt{U})^{n-\hat{s}-1}}p\\
&\leq\sqrt{2\pi}e^{\frac{7}{6}}\frac{\sqrt{\hat{s}}\sqrt{(n-\hat{s})}\hat{s}^{\hat{s}}(n-\hat{s})^{n-\hat{s}}}{\sqrt{n}n^{n}}\frac{1}{\mathtt{U}^{\hat{s}}(1-\mathtt{U})^{n-\hat{s}}}\frac{1-\mathtt{U}}{n-\hat{s}}p\\
&=\sqrt{2\pi}e^{\frac{7}{6}}\frac{\sqrt{\hat{s}}\sqrt{(n-\hat{s})}\hat{s}^{\hat{s}}(n-\hat{s})^{n-\hat{s}}}{\sqrt{n}n^{n}}\frac{1}{\mathtt{U}^{\hat{s}}(1-\mathtt{U})^{n-\hat{s}}}\frac{1-\mathtt{U}}{n\left(1-\frac{\hat{s}}{n}\right)}p\\
&\leq\sqrt{2\pi}e^{\frac{7}{6}}\frac{\sqrt{\hat{s}}\sqrt{(n-\hat{s})}\hat{s}^{\hat{s}}(n-\hat{s})^{n-\hat{s}}}{\sqrt{n}n^{n}}\frac{1}{\mathtt{U}^{\hat{s}}(1-\mathtt{U})^{n-\hat{s}}}\frac{1}{n}p\\
&=\sqrt{2\pi}e^{\frac{7}{6}}\frac{\sqrt{\hat{s}(n-\hat{s})}}{\sqrt{n}}\frac{1}{e^{-n\left[\frac{\hat{s}}{n}\ln\left(\frac{\frac{\hat{s}}{n}}{\mathtt{U}}\right)+\frac{n-\hat{s}}{n}\ln\left(\frac{1-\frac{\hat{s}}{n}}{1-\mathtt{U}}\right)\right]}}\frac{1}{n}p\\
&\leq\sqrt{2\pi}e^{\frac{7}{6}}\frac{\sqrt{\hat{s}(n-\hat{s})}}{\sqrt{n}}\frac{1}{\sum_{i=0}^{\hat{s}}\binom{n}{i}\mathtt{U}^{i}(1-\mathtt{U})^{n-i}}\frac{1}{n}p\\
&=\sqrt{2\pi}e^{\frac{7}{6}}\frac{\sqrt{\frac{\hat{s}}{n}\left(1-\frac{\hat{s}}{n}\right)}}{\sqrt{n}}\frac{2}{\delta}\\
&\leq\sqrt{2\pi}e^{\frac{7}{6}}\frac{1}{\sqrt{n}}\frac{1}{\delta},
\end{aligned}
$$

and, consequently

$$
\min_{\hat{r}\in\left\{0,\frac{1}{n},\cdots,\frac{n-1}{n}\right\},\ p\in(0,1)}\frac{\left(\mathtt{U}\left(\hat{r},\frac{\delta p}{2}\right)\right)^{n\hat{r}}\left(1-\mathtt{U}\left(\hat{r},\frac{\delta p}{2}\right)\right)^{n-n\hat{r}-1}}{\mathtt{B}(n\hat{r}+1,n-n\hat{r})\frac{\delta}{2}p(1-p)}\geq\frac{2\sqrt{n}}{\sqrt{2\pi}e^{\frac{7}{6}}}.
$$

Finally, for what concerns Lemma 8, we must prove that $\forall\hat{r},\hat{r}',\hat{r}''\in\{0,1/m,\cdots,1\}$ such that $\hat{r}'<\hat{r}''$, we have that

$$
\begin{aligned}
&\mathtt{L}(\hat{r}',\delta)-\mathtt{L}(\hat{r}'',\delta)\leq 0,\\
&\exists\mathtt{L}^{-1}:\mathtt{L}^{-1}(\mathtt{L}(\hat{r},n,\delta),\delta)\geq\hat{r}.
\end{aligned}
$$

Exploiting Proposition A1 the first property can be easily derived. The second property is true by definition of $\mathtt{L}^{-1}(r,\delta)$. □

Appendix C. Technicalities

Technicalities of the paper are included in this section.

Proposition A1 (for Theorem A2)**.** *Let us define*

$$
\begin{aligned}
\mathtt{L}(\hat{R},n,\delta)&=\begin{cases}\mathtt{F}^{-1}(\delta;n\hat{R},n-n\hat{R}+1) & \hat{R}\in\left\{\frac{1}{n},\frac{2}{n},\cdots,1\right\}\\ 0 & \hat{R}=0\end{cases},\\
\mathtt{U}(\hat{R},n,\delta)&=\begin{cases}\mathtt{F}^{-1}(1-\delta;n\hat{R}+1,n-n\hat{R}) & \hat{R}\in\left\{0,\frac{1}{n},\cdots,\frac{n-1}{n}\right\}\\ 1 & \hat{R}=1\end{cases},
\end{aligned}
$$

then

$$
\mathtt{L}(\hat{R},n,\delta)\in[0,1],\quad\mathtt{U}(\hat{R},n,\delta)\in[0,1],
$$

and

$$\frac{\partial \mathtt{L}(\hat{R},\delta)}{\partial \delta} = \begin{cases} \frac{\mathtt{B}(n\hat{R},n-n\hat{R}+1)}{(\mathtt{L}(\hat{R},\delta))^{n\hat{R}-1}(1-\mathtt{L}(\hat{R},\delta))^{n-n\hat{R}}} & \hat{R} \in \left\{\frac{1}{n},\frac{2}{n},\cdots,1\right\} \\ 0 & \hat{R}=0 \end{cases} \in [0,\infty),$$

$$\frac{\partial \mathtt{U}(\hat{R},\delta)}{\partial \delta} = \begin{cases} \frac{-\mathtt{B}(n\hat{R}+1,n-n\hat{R})}{(\mathtt{U}(\hat{R},\delta))^{n\hat{R}}(1-\mathtt{U}(\hat{R},\delta))^{n-n\hat{R}-1}} & \hat{R} \in \left\{0,\frac{1}{n},\cdots,\frac{n-1}{n}\right\} \\ 0 & \hat{R}=1 \end{cases} \in (-\infty,0].$$

The derivation of Proposition A1 is trivial.

Appendix D. Code

All Matlab codes used in this paper are available by request to the corresponding author.

References

1. Kearns, M.J.; Vazirani, U.V. *An Introduction to Computational Learning Theory*; MIT Press: Cambridge, MA, USA, 1994.
2. Vapnik, V.N. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
3. Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2001.
4. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
5. Bartlett, P.L.; Boucheron, S.; Lugosi, G. Model selection and error estimation. *Mach. Learn.* **2002**, *48*, 85–113. [CrossRef]
6. Langford, J. Tutorial on practical prediction theory for classification. *J. Mach. Learn. Res.* **2005**, *6*, 273–306.
7. Oneto, L. Model Selection and Error Estimation Without the Agonizing Pain. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1252. [CrossRef]
8. Langford, J. Quantitatively Tight Sample Complexity Bounds. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsyivania, PA, USA, 2002.
9. Bartlett, P.L.; Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **2002**, *3*, 463–482.
10. Bartlett, P.L.; Bousquet, O.; Mendelson, S. Local Rademacher complexities. *Ann. Stat.* **2005**, *33*, 1497–1537. [CrossRef]
11. Bousquet, O.; Elisseeff, A. Stability and generalization. *J. Mach. Learn. Res.* **2002**, *2*, 499–526.
12. Floyd, S.; Warmuth, M. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Mach. Learn.* **1995**, *21*, 269–304. [CrossRef]
13. McAllester, D.A. Some PAC-Bayesian theorems. *Mach. Learn.* **1999**, *37*, 355–363. [CrossRef]
14. Dwork, C.; Feldman, V.; Hardt, M.; Pitassi, T.; Reingold, O.; Roth, A.L. Preserving Statistical Validity in Adaptive Data Analysis. In Proceedings of the ACM Symposium on Theory of Computing, Portland, OR, USA, 14–17 June 2015.
15. Clopper, C.J.; Pearson, E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **1934**, *26*, 404–413. [CrossRef]
16. Hoeffding, W. Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.* **1963**, *58*, 13–30. [CrossRef]
17. Hotz, T.; Kelma, F.; Wieditz, J. Non-asymptotic confidence sets for circular means. *Entropy* **2016**, *18*, 375. [CrossRef]
18. Mukherjee, S.; Niyogi, P.; Poggio, T.; Rifkin, R. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.* **2006**, *25*, 161–193. [CrossRef]
19. Shawe-Taylor, J.; Bartlett, P.L.; Williamson, R.C.; Anthony, M. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theory* **1998**, *44*, 1926–1940. [CrossRef]
20. Langford, J.; McAllester, D. Computable shell decomposition bounds. *J. Mach. Learn. Res.* **2004**, *5*, 529–547.
21. Tikhonov, A.N.; Arsenin, V.I.A.; John, F. *Solutions of Ill-Posed Problems*; Winston: Washington, DC, USA, 1977.
22. Schölkopf, B.; Smola, A.J.; Bach, F. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2002.
23. Schuster, T.; Kaltenbacher, B.; Hofmann, B.; Kazimierski, K.S. *Regularization Methods in Banach Spaces*; Walter de Gruyter: Berlin, Germany, 2012.
24. Oneto, L.; Ridella, S.; Anguita, D. Improving the Union Bound: A Distribution Dependent Approach. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Brugge, Belgium, 2–4 October 2020.
25. Maron, O.; Moore, A.W. The racing algorithm: Model selection for lazy learners. *Artif. Intell. Rev.* **1997**, *11*, 193–225. [CrossRef]
26. Roeder, K.; Wasserman, L. Genome-wide significance levels and weighted hypothesis testing. *Stat. Sci. Rev. J. Inst. Math. Stat.* **2009**, *24*, 398. [CrossRef]
27. Catoni, O. *PAC-Bayesian Supervised Classification*; Institute of Mathematical Statistics: Beachwood, OH, USA, 2007.
28. Oneto, L.; Anguita, D.; Ridella, S. A local Vapnik-Chervonenkis complexity. *Neural Netw.* **2016**, *82*, 62–75. [CrossRef]
29. Polato, M.; Lauriola, I.; Aiolli, F. A novel boolean kernels family for categorical data. *Entropy* **2018**, *20*, 444. [CrossRef]
30. Lever, G.; Laviolette, F.; Shawe-Taylor, J. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theor. Comput. Sci.* **2013**, *473*, 4–28. [CrossRef]

31. Oneto, L.; Anguita, D.; Ridella, S. PAC-Bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis. *Pattern Recognit. Lett.* **2016**, *80*, 200–207. [CrossRef]
32. Oneto, L.; Cipollini, F.; Ridella, S.; Anguita, D. Randomized Learning: Generalization Performance of Old and New Theoretically Grounded Algorithms. *Neurocomputing* **2018**, *298*, 21–33. [CrossRef]
33. Page, E.S. Continuous inspection schemes. *Biometrika* **1954**, *41*, 100–115. [CrossRef]
34. Jensen, D.D.; Cohen, P.R. Multiple comparisons in induction algorithms. *Mach. Learn.* **2000**, *38*, 309–338. [CrossRef]
35. Robbins, H. A remark on Stirling's formula. *Am. Math. Mon.* **1955**, *62*, 26–29. [CrossRef]