# Randomized Learning and Generalization of Fair and Private Classifiers: from PAC-Bayes to Stability and Differential Privacy

Luca Oneto[a,*], Michele Donini[b], Massimiliano Pontil[c,d], John Shawe-Taylor[d]

[a] *University of Pisa, Largo Bruno Pontecorvo 3, 56127, Pisa, Italy*
[b] *Amazon, 1800 9th Ave, 98101, Seattle (WA), USA*
[c] *Istituto Italiano di Teconologia, Via Enrico Melen 83, 16152, Genova, Italy*
[d] *University College London, UCL Malet Place, WC1E 6BT, London, United Kingdom*

**Abstract**

We address the problem of randomized learning and generalization of fair and private classifiers. From one side we want to ensure that sensitive information does not unfairly influence the outcome of a classifier. From the other side we have to learn from data while preserving the privacy of individual observations. We initially face this issue in the PAC-Bayes framework presenting an approach which trades off and bounds the risk and the fairness of the randomized (Gibbs) classifier. Our new approach is able to handle several different state-of-the-art fairness measures. For this purpose, we further develop the idea that the PAC-Bayes prior can be defined based on the data-generating distribution without actually knowing it. In particular, we define a prior and a posterior which give more weight to functions with good generalization and fairness properties. Furthermore, we will show that this randomized classifier possesses interesting stability properties using the algorithmic distribution stability theory. Finally, we will show that the new posterior can be exploited to define a randomized accurate and fair algorithm. Differential privacy theory will allow us to derive that the latter algorithm has interesting privacy preserving properties ensuring our threefold goal of good generalization, fairness, and privacy of the final model.

*Corresponding author

*Email addresses:* luca.oneto@unipi.it (Luca Oneto), donini@amazon.com (Michele Donini), massimiliano.pontil@iit.it (Massimiliano Pontil), j.shawe-taylor@ucl.ac.uk (John Shawe-Taylor)

*Keywords:* Algorithmic Fairness, Privacy Preserving Data Analysis, Generalization, Randomized Classifier, PAC-Bayes, Algorithmic (Distribution) Stability, Randomized Algorithm, Differential Privacy

## 1. Introduction

Randomized models and learning algorithms are nowadays becoming a trending research interest because of their effectiveness in many real world applications [1, 2]. From the Deep and Shallow Neural Networks [3, 4] to the Extreme Learning Machine [5] and the Ensemble Methods [6], randomness plays a crucial role in improving the effectiveness of a learning paradigm. The idea of using randomness to enrich the set of models [7], to improve the optimization techniques [8], or to improve the generalization capabilities of a model [9] has been a breakthrough which allowed to develop techniques such as Dropout [10] or Random Forest [11, 12].

At the same time, it is becoming increasingly important to construct models able to exhibit privacy and fairness properties, namely to ensure the ability to learn from data while preserving the privacy of individual observations and to ensure that sensitive information (e.g. knowledge about gender of an individual) does not unfairly influence the outcome of a learning algorithm.

The problem of learning from data while preserving the privacy of individual observations has a long history and spans over multiple disciplines [13–15]. One way to preserve privacy is to corrupt the learning procedure with noise without destroying the information that we want to extract. Differential Privacy (DP) is one of the most powerful tools in this context [15, 16]. DP addresses the problem of keeping private the information about an individual observation while learning useful information about a population. In particular, a procedure is DP if and only if its output is almost independent from any of the individual observations. In other words, the probability of a certain output should not change significantly if one individual is present or not, where the probabilities are taken over the noise introduced by the procedure. In the last years, DP has been deeply studied from a theoretical point of view [17–28] and exploited to develop new learning strategies for solving real world problems [29–36]. Another way to preserve privacy is to federate the learning procedure in order to keep the data decentralized and not distribute sensitive information [37–41]. Although this approach sounds plausible, it is

not supported by statistical or privacy guarantees if no additional privacy preserving layers are added [42, 43].

Another problem which lately received a lot of attention is algorithmic fairness [44–58]. The central question is how to enhance learning algorithms with fairness requirements, namely ensuring that sensitive information (e.g. knowledge about the ethnic group of an individual) does not 'unfairly' influence the outcome of a learning algorithm. For example if the learning problem is to decide whether a person should be offered a loan based on the previous credit card scores, we would like to build a model which does not unfairly use additional sensitive information such as race or sex. Several measures of fairness of a classifier have been studied in the literature [59, 60] like the Demographic Parity (DPa) [61], the Equal Odds (EOd) and the Equal Opportunity (EOp) [47], the Disparate Treatment, Impact, and Mistreatment [48], among others. Works on algorithmic fairness can be divided in three families. Methods in the first family modify a pre-trained classifier in order to increase its fairness properties while maintaining as much as possible the classification performance [47, 62–64]. Methods in the second family enforce fairness directly during the training step [45, 60, 65, 66]. The third family of methods implements fairness by modifying the data representation and then employs standard machine learning methods [45, 49, 52, 67–70]. All methods in the previous families have in common the goal of creating a fair model from scratch on the specific task at hand. This solution may work well in specific cases, but in a large number of real world applications it is common to perform a fine tuning over pre-trained models [71], keeping the internal representation fixed. Indeed, most modern machine learning frameworks (especially the deep learning ones) offer a set of pre-trained models that are distributed in so-called model zoos[1]. Unfortunately, fine tuning pre-trained models on novel previously unseen tasks could lead to unexpected unfairness behavior, even starting from an apparently fair model for previous tasks (e.g. discriminatory transfer [72], or negative legacy [73]), due to missing generalization guarantees concerning the fairness property of the model. For this reason many recent methods try to address the problem of learning a fair representation not just a fair model [49, 63, 74–81].

In this paper, we address the problem of randomized learning and generalization of fair and private classifiers. From one side we want to ensure that

[1]See for example the Caffe Model Zoo: `github.com/BVLC/caffe/wiki/Model-Zoo`

sensitive information does not unfairly influence the outcome of a classifier. From the other side we want to ensure to be able to learn from data while preserving the privacy of individual observations. We first face this issue in the PAC-Bayes (PB) framework and we present an approach which trades off and bounds the risk and the fairness of the Randomized (Gibbs) Classifier (RC), together with the Bayes Classifier (BC) which is its deterministic counterpart, measured with respect to different state-of-the-art fairness measures (e.g. EOp, EOd, and DPa). For this purpose, we exploit further the idea that the PB prior can be defined based on the data-generating distribution without actually needing to know it. In this sense, we define a prior and a posterior with the goal of giving more weight to functions with good generalization and fairness properties. Furthermore, we will show that this randomized classifier possesses interesting stability properties using the Algorithmic (Distribution) Stability (AS) theory. Finally, we will show that the new posterior introduced for building an accurate and fair RC can be exploited to define an accurate and fair Randomized Learning Algorithm (RLA). The latter will also show to possess interesting privacy preserving properties ensuring generalization, fairness, and privacy of the final model. DP theory will allow us to derive such results.

To the best of our knowledge, our approach is the first one that is able to face the problem of learning from data under fairness and privacy properties, backed up by three different theoretical frameworks. The only paper which addresses a similar problem is [82] – but with several differences with respect to ours. First, [82] is able to deal with a single notion of fairness, while our approach is able to deal with a large family of them, and also with different kind of sensitive attributes. Secondly, the theoretical analysis of [82] is based on classical statistical learning theory, characterized by loose constants and rates of convergence. Our work, instead, is backed up by three different state-of-the-art tools (PB, AS, and DP theories) and shows optimal constants and rates of convergence. Thirdly, the post processing technique proposed in [82] – as also stated by the authors – is suboptimal with respect to the in processing techniques (like the one we propose in this paper). Moreover, the method in [82] requires the knowledge of a subroutine that can optimally solve classification problems absent from fairness constraint and even of the protected attribute at test time. Our method instead does not require any of these constraints. Finally, [82] introduces privacy with a simple Laplacian mechanism of perturbation of the outputs of the non-private counterparts of the algorithms (previously developed by [47, 65]) while our method is

intrinsically fair and private by construction.

In order to better understand our results, let us clarify the difference between deterministic and randomized models and learning algorithms in the context of classification. A Deterministic Classifier (DC) assigns always the same label given an input, while an RC may assigns different labels to the same input if we repeat the labelling process. Analogously, a Deterministic Learning Algorithm (DLA) learns the same DC (or RC) if we keep the training set fixed, while an RLA may learn a different DC (or RC) even if we keep fixed the training set repeating the learning process. In order to estimate the generalization performance of an RC the PB theory is one of the sharpest analysis frameworks, since it can provide tight bounds on the risk of the RC and BC [83]. The RC chooses a classifier in the set of classifiers according to a posterior distribution each time a new sample has to be classified [84] while the BC takes the decision based on the expected value of the RC over the posterior distribution [83]. In particular, in the PB theory a prior distribution over the different classifiers must be defined before seeing the data, then, based on the available data, a posterior distribution can be chosen, and the risk of the associate RC and BC is computed, based on the empirical risk and the divergence between the prior and posterior distributions [85]. The PB theory bounds the risk of the RC [85], while the $\mathcal{C}$-bound bounds the error of the BC based on the properties of the RC [86]. The first result of this work is to derive a PB-based bound on the fairness (measured with the EOd or EOp or the DPa) of a RC model. Then we focus on the problem of choosing the right posterior and prior distributions since the divergence between prior and posterior distributions forms part of the bound. This choice is critical: in some cases this choice proves to be too generic and not suited for the particular problem [84], other times some data are kept apart from the learning process and exploited to derive a generally good prior [87, 88]. Consequently, in the first case the divergence term in the PB analysis can typically be large, while in the second case the bound tends to be loose since some data are wasted in order to design the prior. In order to address this issue in [89] a localized PB analysis is proposed, which uses a Boltzmann prior distribution defined in terms of the distribution that generated the data. Note that, since the prior depends on the distribution, the PB analysis is still valid because the prior is defined before observing the data [84, 89, 90]. By tuning the prior to the distribution, Catoni was able to remove the divergence term from the bound, hence significantly reducing the complexity penalty. Note that other approaches for removing the divergence

exist. One approach is to design a prior and a posterior such that they are aligned [83, 91, 92]. The second one is to design a so called expectation-prior which does not require any separate set of data to build a prior which will be probably close to the posterior [88]. Every approach has its own strengths and weaknesses but the approach of Catoni seems to be the most promising one [84, 89, 90] even if using Boltzmann distributions in some contexts can be seen as a limitation [84]. In fact, keeping the divergence term allowed many researchers to design new model selection methods and learning algorithms [93–95]. Nevertheless, in this work we exploit the idea of Catoni and we define a Boltzmann prior and posterior which give more weight to functions which exhibit good generalization and fairness properties and we show that it is possible to remove the divergence term from the bound on the risk and the fairness of the corresponding RC. Then we analyze the RC induced by the newly defined fair and accurate posterior through the use of the AS theory originally developed in [90, 96–105] and then further refined in [90, 98] to deal with the RC. AS allows to give an answer to a fundamental question in learning theory, namely what are the general conditions for predictivity. AS answers this question in a very intuitive way: if the algorithm selects similar hypothesis, even if the training data are (slightly) modified, then we can be confident that the learning algorithm is stable [97]. For RC the AS theory proves that, if the criteria used to define the posterior distribution based on the available data do not change too much when the training data are (slightly) modified, then the associated RC will have good generalization performance [106]. In this work we show that the newly posterior fair and accurate distribution inspired by the works of [84, 89, 90] has the AS property, which allows us to bound the risk and the fairness of the RC in a new way. By exploiting the $\mathcal{C}$-bound it is also possible to bound the risk of the associated BC [83, 86]. Finally, we will show how to use the newly introduced fair and accurate posterior in order to develop a RLA algorithm which exploits the same randomness introduced by this posterior distribution. In particular, we will show that this RLA may possess better generalization, fairness, and privacy properties than the RC which exploits the fair and accurate newly introduced posterior, even if they are both based on the same data dependent posterior distribution. For this purpose we will use the DP theory. DP addresses the problem of keeping private the information about an individual observation while learning useful information about a population [15]. In particular, a procedure is DP if and only if its output is almost independent from any of the individual observations. DP allowed to reach

Table 1: Abbreviations and Symbols

| Abbreviation | Description |
|---|---|
| DLA | Deterministic Learning Algorithms |
| RLA | Randomized Learning Algorithms |
| DC | Deterministic Classifier |
| RC | Randomized (Gibbs) Classifier |
| BC | Bayes Classifier |
| PB | PAC-Bayes |
| AS | Algorithmic (Distribution) Stability |
| DP | Differential Privacy |
| EOp | Equal Opportunity |
| EOd | Equal Odds |
| DPa | Demographic Parity |
| KLD | Kullback-Leibler Divergence |

a milestone result by connecting the field of privacy preserving data analysis and the generalization capability of a randomized learning algorithm. In particular DP allows to prove that a RLA which shows DP properties also generalizes [107–109], namely we can effectively bound the risk of the selected model. In this work, we will derive a DP-based bound on both the risk and the fairness of the model selected with the RLA which exploits the fair and accurate newly introduced posterior.

The paper is organized as follows. Section 2 introduces the notation while Appendix A reports the state-of-the-art results needed for taking into account fairness and privacy issues for both RC (Appendix A.1) and/or RLA (Appendix A.2) in the framework of the PB (Appendix A.1.1), AS (Appendix A.1.2), and DP (Appendix A.2) theories. Section 3 presents our proposal by first defining a RC, in the PB theory framework, where the prior and the posterior give more weight to functions which exhibit good generalization and fairness properties and then by showing that this RC possesses interesting properties in the AS theory framework (Section 3.1). Then, in Section 3.2, we will show that the proposed posterior can be exploited to define an accurate and fair RLA which is shown to possess interesting privacy preserving properties ensuring generalization, fairness, and privacy of the final model. Appendix B reports the proofs not reported in Section 3. Finally, Section 4 concludes the paper. In order to improve the readability of our work, Tables 1 and 2 report, respectively, the abbreviation and the symbols used in the paper.

Table 2: Abbreviations and Symbols

| Symbols | Description |
|---|---|
| $\mathcal{X}$ | Input space |
| $\mathcal{S}$ | Group membership $\{1, \cdots, k\}$ |
| $\mathcal{Y}$ | Set of binary output labels $\{-1, +1\}$ |
| $\mathcal{Z}$ | $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$ |
| $n$ | Number of samples |
| $\mathcal{D}$ | $\mathcal{Z}^n$ |
| $\mathfrak{P}_{\mathcal{X}}$, $\mathfrak{P}_{\mathcal{S}}$, $\mathfrak{P}_{\mathcal{Y}}$, $\mathfrak{P}_{\mathcal{Z}}$ | Unknown probability distributions over $\mathcal{X}$, $\mathcal{S}$, $\mathcal{Y}$, and $\mathcal{Z}$ respectively |
| $\mathfrak{P}_{\mathcal{D}}$ | Distribution of probability induced by $\mathfrak{P}_{\mathcal{Z}}$ over $\mathcal{D}$ |
| $\boldsymbol{X}$, $\boldsymbol{S}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$, $\boldsymbol{D}$ | Random variables sampled from $\mathcal{X}$, $\mathcal{S}$, $\mathcal{Y}$, $\mathcal{Z}$, and $\mathcal{D}$ according to $\mathfrak{P}_{\mathcal{X}}$, $\mathfrak{P}_{\mathcal{S}}$, $\mathfrak{P}_{\mathcal{Y}}$, $\mathfrak{P}_{\mathcal{Z}}$, and $\mathfrak{P}_{\mathcal{D}}$ respectively |
| $x$, $s$, $y$, $z$, $d$ | Element in $\mathcal{X}$, $\mathcal{S}$, $\mathcal{Y}$, $\mathcal{Z}$, and $\mathcal{D}$ respectively |
| $\diamond$ | Placeholder for one of the following symbols $\{-, +, -\vee+\}$ |
| $d_{g,\diamond}$ | $\{(x, s, y) : (x, s, y) \in d, s = g, y = \diamond 1\}$ |
| $n_{g,\diamond}$ | $\lvert d_{g,\diamond}\rvert$ |
| $d^{\setminus i}$ | $d \setminus z_i$ |
| $d^i$ | $d^{\setminus i} \cup \dot{z}_i$ |
| $\dot{d}$ | $d^i$ where $i$ in $\{1, \cdots, n\}$ and $\dot{z}_i$ from $\mathcal{Z}$ according to $\mathfrak{P}_{\mathcal{Z}}$ |
| $\mathcal{H}$ | Set of DC |
| $h$ | DC in $\mathcal{H}$ |
| $\mathtt{Q}$ | Probability distribution over $\mathcal{H}$ |
| $G_{\mathtt{Q}}$ | RC |
| $B_{\mathtt{Q}}$ | BC |
| $\mathscr{A}$ | A DLA or a RLA that maps a dataset $d$ into an DC or a RC |
| $\mathfrak{P}_{\mathscr{A}}$ | Probability distribution that encapsulates non-deterministic rules behind the RLA |
| $\ell$ | A loss function such that $\ell : \mathcal{H} \times \mathcal{Z} \to [0, 1]$ |
| $R^{\ell}(\cdot)$ | Risk of a DC or an RC measured according to $\ell$ |
| $\widehat{R}^{\ell}_d(\cdot)$ | Empirical Risk of a DC or an RC measured according to $\ell$ over $d$ |
| `kl` | KLD function |
| `KL` | KLD |
| $F^{\ell}(\cdot)$ | Fariness of a DC or an RC measured according to $\ell$ |
| $\widehat{F}^{\ell}_d(\cdot)$ | Empirical Fairness of a DC or an RC measured according to $\ell$ over $d$ |

## 2. Preliminaries

Let us consider the binary classification problem [110] and the notation needed for taking into account fairness [45, 47] and privacy issues [15, 107, 111, 112] for both RLA and/or RC [113] in the framework of the PB [84, 90, 95], AS [90, 98], and DP [107, 112] theories [9]. Let $d = \{z_1, \cdots, z_n\} = \{(x_1, s_1, y_1), \cdots, (x_n, s_n, y_n)\}$ be a sequence of $n$ samples drawn independently from an unknown probability distribution $\mathfrak{P}_{\mathcal{Z}}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, where $\mathcal{Y} = \{-1, +1\}$ is the set of binary output labels, $\mathcal{S} = \{1, \cdots, k\}$ represents group membership, and $\mathcal{X}$ is the input space. We indicate with $\mathfrak{P}_{\mathcal{X}}$, $\mathfrak{P}_{\mathcal{S}}$, and $\mathfrak{P}_{\mathcal{Y}}$ respectively the distributions over $\mathcal{X}$, $\mathcal{S}$, and $\mathcal{Y}$. We indicate with $\boldsymbol{X}$, $\boldsymbol{S}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ the random variable sampled respectively from $\mathcal{X}$, $\mathcal{S}$, $\mathcal{Y}$, and $\mathcal{Z}$ according respectively to $\mathfrak{P}_{\mathcal{X}}$, $\mathfrak{P}_{\mathcal{S}}$, $\mathfrak{P}_{\mathcal{Y}}$, and $\mathfrak{P}_{\mathcal{Z}}$. In this new perspective, $d$ is a dataset inside the space of all the possible datasets $\mathcal{D} = \mathcal{Z}^n$, $\mathfrak{P}_{\mathcal{D}}$ is the distribution of probability generated by $\mathfrak{P}_{\mathcal{Z}}$ over $\mathcal{D}$ and $\boldsymbol{D}$ is a random variable sampled from $\mathcal{D}$ according to $\mathfrak{P}_{\mathcal{D}}$.

For every $g \in \mathcal{S}$ and operator $\diamond \in \{-, +, - \vee +\}$, we define the subset of training points negatively or positively labeled which belongs to the group $g$ as $d_{g,\diamond} = \{(x, s, y) : (x, s, y) \in d, s = g, y = \diamond 1\}$ where $n_{g,\diamond} = |d_{g,\diamond}|$, noting that $d_{g,-\vee+} = \{(x, s, y) : (x, s, y) \in d, s = g\}$.

We denote a series of auxiliary datasets with $d^{\setminus i} = d \setminus z_i$, with $d^i = d^{\setminus i} \cup \dot{z}_i$ and with $\dot{d} = d^i$ where $i$ may assume any value in $\{1, \cdots, n\}$ and $\dot{z}_i$ is sampled from $\mathcal{Z}$ according to $\mathfrak{P}_{\mathcal{Z}}$.

Let us consider a DC $h$ belonging to a set $\mathcal{H}$ of possible ones, whose functional form may $h : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$ or may not $h : \mathcal{X} \to \mathbb{R}$, consider the sensitive feature.

A RC, instead, draws an $h \in \mathcal{H}$, according to a probability distribution $\mathtt{Q}$ over $\mathcal{H}$, each time a label for an input $x \in \mathcal{X}$ is required. We will call $G_{\mathtt{Q}}$ this RC.

A DLA $\mathscr{A}$ maps a dataset $d$ into a classifier (which can be both a DC or a RC). A RLA, instead, maps a dataset into a classifier (which can be, again, both a DC or a RC but for the purpose of this paper it will be always a DC) with non-deterministic rules that can be encapsulated in a probability distribution $\mathfrak{P}_{\mathscr{A}}$ over the whole possible set of classifiers (in our case over $\mathcal{H}$) given the dataset at hand (in our case $d$).

The accuracy of $h \in \mathcal{H}$ in representing the unknown relation between input and output space is measured with reference to a prescribed $[0, 1]$-bounded loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0, 1]$. Hence, we can define the true risk

(or generalization error) of $h$, namely generalization error, as

$$R^{\ell}(h) = \mathbb{E}_{\boldsymbol{Z}}\{\ell(h, \boldsymbol{Z})\}, \tag{1}$$

Since $\mathfrak{P}_{\mathcal{Z}}$ is unknown, $R^{\ell}(h)$ cannot be computed. Therefore, we have to resort to its empirical estimator, the empirical error

$$\widehat{R}^{\ell}_{d}(h) = \frac{1}{|d|}\sum_{z\in d}\ell(h, z). \tag{2}$$

Also for $G_{\mathtt{Q}}$, we can define its risk together with its empirical counterpart, respectively

$$R^{\ell}(G_{\mathtt{Q}}) = \mathbb{E}_{h\sim\mathtt{Q}}\{R^{\ell}(h)\} \quad \text{and} \quad \widehat{R}^{\ell}_{d}(G_{\mathtt{Q}}) = \mathbb{E}_{h\sim\mathtt{Q}}\widehat{R}^{\ell}_{d}(h). \tag{3}$$

The deterministic counterpart of $G_{\mathtt{Q}}$, namely the BC $B_{\mathtt{Q}}$, is defined as

$$B_{\mathtt{Q}}(x(,s)) = \mathbb{E}_{h\sim\mathtt{Q}}h(x(,s)) \tag{4}$$

and, consequently, its true and empirical error are, respectively

$$R^{\ell}(B_{\mathtt{Q}}) = \mathbb{E}_{\boldsymbol{Z}}\ell(B_{\mathtt{Q}}, \boldsymbol{Z}) \quad \text{and} \quad \widehat{R}^{\ell}_{d}(B_{\mathtt{Q}}) = \frac{1}{|d|}\sum_{z\in d}\ell(B_{\mathtt{Q}}, z). \tag{5}$$

Let us recall, in the PB theory framework, the definitions of the Kullback-Leibler Divergence (KLD) function [114] of two real numbers in the interval $(0, 1)$ as

$$\mathtt{kl}[q||p] = q\ln\left[\frac{q}{p}\right] + [1-q]\ln\left[\frac{1-q}{1-p}\right], \tag{6}$$

and the KLD between two distributions [114] $\mathtt{Q}$ and $\mathtt{P}$ over $\mathcal{F}$ as $\mathtt{KL}[\mathtt{Q}||\mathtt{P}]$.

The fairness of the model[2] $h \in \mathcal{H}$, instead, can be measured with respect to many notions of fairness [45, 47, 60]. For a deterministic model we can use the EOp constraint defined as $\text{EOp}^{\diamond}$ for $\diamond \in \{-, +\}$ as

$$\mathbb{P}_{\boldsymbol{Z}}\{h(\boldsymbol{X}(,\boldsymbol{S}))>0|\boldsymbol{S}=1, \boldsymbol{Y}=\diamond 1\} = \cdots = \mathbb{P}_{\boldsymbol{Z}}\{h(\boldsymbol{X}(,\boldsymbol{S}))>0|\boldsymbol{S}=k, \boldsymbol{Y}=\diamond 1\}, \tag{7}$$

[2]Remember that the functional form of the model may depend or not on the sensitive feature and so we will write $h(x(,s))$.

since we can define the EOp of the positively ($\text{EOp}^+$) or negatively ($\text{EOp}^-$) labeled samples, or the EOd constraint defined as the concurrent verification of the $\text{EOp}^+$ and $\text{EOp}^-$, or also the DPa constraint defined as

$$\mathbb{P}_{\boldsymbol{Z}}\{h(\boldsymbol{X}(,\boldsymbol{S}))>0|\boldsymbol{S}=1\} = \cdots = \mathbb{P}_{\boldsymbol{Z}}\{h(\boldsymbol{X}(,\boldsymbol{S}))>0|\boldsymbol{S}=k\}, \tag{8}$$

which is equivalent to the $\text{EOp}^{-\vee+}$. Since $h$, in general, will not be able to exactly fulfill the $\text{EOp}^\diamond$ constraint with $\diamond \in \{-,+\}$, nor the EOd constraints, nor the DPa constraints we define the Difference of $\text{EOp}^\diamond$, namely $\text{DEOp}^\diamond(h)$, with $\diamond \in \{-,+\}$ as

$$\frac{1}{k}\sum_{g\in\mathcal{S}} \left|\mathbb{P}_{\boldsymbol{Z}}\{h(\boldsymbol{X}(,\boldsymbol{S}))>0|\boldsymbol{S}=g,\boldsymbol{Y}=\diamond 1\} - \bar{P}(h)\right|, \tag{9}$$

where

$$\bar{P}(h) = \frac{1}{k}\sum_{g_2\in\mathcal{S}} \mathbb{P}_{\boldsymbol{Z}}\{h(\boldsymbol{X}(,\boldsymbol{S}))>0|\boldsymbol{S}=g_2,\boldsymbol{Y}=\diamond 1\}, \tag{10}$$

the Difference of EOd, namely $\text{DEOd}(h)$, which is defined as the average value between the $\text{DEOp}^+(h)$ and $\text{DEOp}^-(h)$, and the Difference of DPa, namely $\text{DDPa}(h)$, as

$$\frac{1}{k}\sum_{g_1\in\mathcal{S}} \left|\mathbb{P}_{\boldsymbol{Z}}\{h(\boldsymbol{X}(,\boldsymbol{S}))>0|\boldsymbol{S}=g_1\} - \frac{1}{k}\sum_{g_2\in\mathcal{S}} \mathbb{P}_{\boldsymbol{Z}}\{h(\boldsymbol{X}(,\boldsymbol{S}))>0|\boldsymbol{S}=g_2\}\right|. \tag{11}$$

Note that all these fairness measures can be reformulated as difference of risks [45, 60]. In particular, let us define the Hard loss function $\ell_H$, namely the function which detects a classification

$$\ell_H(h,z) = \mathbb{1}\{yh(x(,s)) \leq 0\}. \tag{12}$$

Then the $\text{EOp}^\diamond$ constraint for $\diamond \in \{-,+\}$ can be defined as[3]

$$R^{\ell_H}_{d_{1,\diamond}}(h) = \cdots = R^{\ell_H}_{d_{k,\diamond}}(h), \tag{13}$$

[3]Note that, $R^{\ell_H}_{d_{i,\diamond}}(h) = \mathbb{E}_{\boldsymbol{Z}}\{\ell(h,\boldsymbol{Z})|\boldsymbol{S}=i,\boldsymbol{Y}=\diamond 1\}$

and consequently the $\mathrm{DEOp}^{\diamond}(h)$ for $\diamond \in \{-,+\}$ can be defined as

$$\frac{1}{k}\sum_{g_1\in\mathcal{S}}\left|R^{\ell_H}_{d_{g_1,\diamond}}(h)-\sum_{g_2\in\mathcal{S}}R^{\ell_H}_{d_{g_2,\diamond}}(h)\right|. \tag{14}$$

Analogously it is possible to reformulate the EOd and the DPa together with the $\mathrm{DEOd}(h)$ and the $\mathrm{DDPa}(h)$. In order to simplify the presentation from now on we will focus on the $\mathrm{EOp}^{+}$ and the $\mathrm{DEOp}^{+}(h)$ which we will call respectively, for brevity, EOp and $F(h)$ and to the case when $k=2$ since the extension to the more general case is conceptually trivial but rather technical. Then the EOp constraint can be be defined respectively as

$$R^{\ell_H}_{d_{1,+}}(h)=R^{\ell_H}_{d_{2,+}}(h), \tag{15}$$

and the $F(h)$ can be be defined respectively as

$$F(h)=\left|R^{\ell_H}_{d_{1,+}}(h)-R^{\ell_H}_{d_{2,+}}(h)\right|. \tag{16}$$

As we did for the risk we can also define the empirical EOp constraint, namely $\widehat{\mathrm{EOp}}$

$$\widehat{R}^{\ell_H}_{d_{1,+}}(h)=\widehat{R}^{\ell_H}_{d_{2,+}}(h), \tag{17}$$

and the empirical $F(h)$, namely $\widehat{F}_d(h)$, as

$$\widehat{F}_d(h)=\left|\widehat{R}^{\ell_H}_{d_{1,+}}(h)-\widehat{R}^{\ell_H}_{d_{2,+}}(h)\right|. \tag{18}$$

Analogously, it is possible to define the EOp constraint and the $F(G_{\mathtt{Q}})$ for a RC respectively as

$$R^{\ell_H}_{d_{1,+}}(G_{\mathtt{Q}})=R^{\ell_H}_{d_{2,+}}(G_{\mathtt{Q}}) \quad \text{and} \quad F(G_{\mathtt{Q}})=\left|R^{\ell_H}_{d_{1,+}}(G_{\mathtt{Q}})-R^{\ell_H}_{d_{2,+}}(G_{\mathtt{Q}})\right|, \tag{19}$$

together with their empirical counterparts

$$\widehat{R}^{\ell_H}_{d_{1,+}}(G_{\mathtt{Q}})=\widehat{R}^{\ell_H}_{d_{2,+}}(G_{\mathtt{Q}}) \quad \text{and} \quad \widehat{F}_d(G_{\mathtt{Q}})=\left|\widehat{R}^{\ell_H}_{d_{1,+}}(G_{\mathtt{Q}})-\widehat{R}^{\ell_H}_{d_{2,+}}(G_{\mathtt{Q}})\right|. \tag{20}$$

Appendix A reports the state-of-the-art results needed for taking into account fairness and privacy issues for both RC (Appendix A.1) and/or RLA (Appendix A.2) in the framework of the PB (Appendix A.1.1), AS (Appendix A.1.2), and DP (Appendix A.2) theories.

## 3. Fair and Private Randomized Learning

In this section we will extend the state-of-the-art results on Randomized Learning and generalization (see Section 2 and Appendix A) to the problem of learning accurate and fair models without compromising the privacy of the individual observations. For this purpose, we further develop the idea that the PB prior can be defined based on the data-generating distribution without actually needing to know it. In particular, we define a prior and a posterior giving more weight to functions which exhibit good generalization and fairness properties. Furthermore, we will show that this RC possesses interesting stability properties using the AS theory. Finally, we will show that the new posterior introduced for building a randomized accurate and fair classifier can be exploited to define accurate and fair RLA. Exploiting DP theory, we will also show that the accurate and fair RLA possesses interesting privacy preserving properties ensuring generalization, fairness, and privacy of the final model. In order to improve the readability of this section, we have included a summary our results in Table 3.

### *3.1. Fair Randomized Classifiers*

With respect to what is described in Appendix A.1.1, our scope is not to simply fit the data minimizing the risk of the RC, but we require also the fairness of the solution (measured with respect to distance from the equal opportunity). In other words we want to simultaneously minimize the risk and the fairness of the RC, respectively the $R(G_{\mathtt{Q}_d})$ and the $F(G_{\mathtt{Q}_d})$. In order to achieve this goal, first we have to bound the $F(G_{\mathtt{Q}_d})$ analogously to what has been done with $R^\ell(G_{\mathtt{Q}_d})$ in Theorem 10; then we will have to define a $\mathtt{P}$ and an $\mathtt{Q}_d$ able to both reduce the risk, the fairness, and the KLD. Let us start with the first objective with the following theorem.

**Theorem 1.** *For any probability distribution* $\mathtt{P}$ *over* $\mathcal{H}$*, chosen before seeing* $d$*, we have that*

$$\mathbb{P}_{\boldsymbol{D}}\left\{\exists \mathtt{Q}_d : \left|\widehat{F}_{\boldsymbol{D}}(G_{\mathtt{Q}_{\boldsymbol{D}}}) - F(G_{\mathtt{Q}_{\boldsymbol{D}}})\right| \geq \sqrt{\frac{1}{2n_{1,+}}\left[\mathtt{KL}[\mathtt{Q}_{\boldsymbol{D}}||\mathtt{P}] + \ln\left(\frac{2\sqrt{n_{1,+}}}{\delta}\right)\right]} + \sqrt{\frac{1}{2n_{2,+}}\left[\mathtt{KL}[\mathtt{Q}_{\boldsymbol{D}}||\mathtt{P}] + \ln\left(\frac{2\sqrt{n_{2,+}}}{\delta}\right)\right]}\right\} \leq 2\delta. \tag{21}$$

| Posterior of Eq. (27) and Prior of Eq. (25)<br>$\gamma \to 0$ uniform probability over all functions, $\gamma \to \infty$ probability one of the risk minimizer<br>$\lambda = 0$ no fairness constraint, $\lambda = \infty$ strong fairness constraint | | | |
|---|---|---|---|
| Theory | PB | AS | DP |
| Algorithm / Classifier | RC & BC<br>Eqns. (27) & (25) | RC & BC<br>Eq. (27) | RLA<br>Eq. (27) |
| Property | Accurate<br>Fair | Accurate<br>Fair | Accurate<br>Fair<br>Private |
| Bounds Risk & Fairness | Theorem 3 | Theorem 6 | Theorem 9 |
| Convergence Risk | $O\left(\sqrt{\frac{\ln(\min(n_{1,+},n_{2,+}))}{\min(n_{1,+},n_{2,+})}}\right)$ | $O\left(\sqrt{\frac{1}{\min(n_{1,+},n_{2,+})}}\right)$ | $O\left(\sqrt{\frac{1}{n}}\right)$ |
| Convergence Fairness | $O\left(\sqrt{\frac{\ln(\min(n_{1,+},n_{2,+}))}{\min(n_{1,+},n_{2,+})}}\right)$ | $O\left(\sqrt{\frac{1}{\min(n_{1,+},n_{2,+})}}\right)$ | $O\left(\sqrt{\frac{1}{\min(n_{1,+},n_{2,+})}}\right)$ |
| $\gamma$ max speed | $O(\sqrt{\min[n_{1,+},n_{2,+}]})$ | Slower than $O(\sqrt{\min[n_{1,+},n_{2,+}]})$ | $O(\sqrt{\min[n_{1,+},n_{2,+}]})$ |
| Fairness & Risk Tension | $\gamma\lambda \leq$ $O(\sqrt{\min[n_{1,+},n_{2,+}]})$ | $\gamma\lambda <$ $O(\sqrt{\min[n_{1,+},n_{2,+}]})$ | $\gamma\lambda \leq$ $O(\sqrt{\min[n_{1,+},n_{2,+}]})$ |

Table 3: Summary of the results of Section 3.

The proof can be retrieved in Appendix B.1.

After this first result, analogously to what is described in Appendix A.1.1, we have to define a $\mathtt{P}$ and a $\mathtt{Q}_d$ able to give more importance to functions with good accuracy and fairness. In particular, exploiting the idea developed in [45, 60], a good function should minimize the risk subject to fairness constraints such that

$$h^* : \arg\min_{h\in\mathcal{H}} \quad R^\ell(h) \qquad \text{s.t.} \quad F(h) \leq \Delta, \tag{22}$$

where $\Delta \in (0, 1]$ is necessary since for $\Delta = 0$ some problems may arise [62]. Equivalently, for a particular value of $\lambda \in [0, \infty)$ [115]

$$h^* : \arg\min_{h\in\mathcal{H}} \quad R^\ell(h) + \lambda F(h). \tag{23}$$

Consequently $\Delta$ and $\lambda$ regulate the trade off between accuracy and fairness of the solution. Note that, for small $\lambda$, or equivalently for large $\Delta$, we relax the

fairness constraint and we just care about minimizing the error. Contrarily, for large $\lambda$, or equivalently for small $\Delta$, we strongly enforce the fairness constraint giving less importance to the accuracy of the model. This tension is a classical result in fairness which shows that, in many cases, it is not possible to simultaneously achieve accuracy and fairness [62]. Consequently, as we will see later, $\lambda$ (or equivalently $\Delta$), cannot be arbitrarily set if we want to maintain certain generalization properties of the algorithm. In [45, 60] it is proved that, if the empirical counterpart of the above mentioned problem without the fairness constraint is consistent, it is also consistent the following fair empirical risk minimization problem

$$\widehat{h}^* : \arg\min_{h\in\mathcal{H}} \quad \widehat{R}_d^\ell(h) + \lambda \widehat{F}_d(h). \tag{24}$$

Then, following the ideas in [84, 89] we propose to use the following probability density function for $\mathtt{Q}_d$

$$\mathtt{q}_d(h) = Z_{\mathtt{Q}_d} \mathbb{e}^{-\gamma\left[\widehat{R}_d^\ell(h)+\lambda\widehat{F}_d(h)\right]}, \tag{25}$$

where

$$Z_{\mathtt{Q}_d}^{-1} = \int_{\mathcal{H}} \mathbb{e}^{-\gamma\left[\widehat{R}_d^\ell(h)+\lambda\widehat{F}_d(h)\right]} dh, \tag{26}$$

and consequently the following probability density function for $\mathtt{P}$

$$\mathtt{p}(h) = Z_{\mathtt{P}} \mathbb{e}^{-\gamma\left[R_d^\ell(h)+\lambda F(h)\right]}, \tag{27}$$

where

$$Z_{\mathtt{P}}^{-1} = \int_{\mathcal{H}} \mathbb{e}^{-\gamma\left[R_d^\ell(h)+\lambda F(h)\right]} dh. \tag{28}$$

Basically our posterior distribution weights more the optimal solution of the Problem (24) and exponentially less the other ones based on their distance, in terms of cost function, from the optimal one. The distribution of these weights is regulated by $\gamma$. The larger is $\gamma$ the more weight is associated to functions characterized by small error. As a consequence, our desiderata would be to have $\gamma$ as large as possible but this, as we will see later, will not be allowed if we want to maintain certain generalization properties of the algorithm.

Note that, the proposed algorithm does not necessarily require the knowledge of the sensitive attribute at test time [45, 60, 82] and this is a very important property of our method (in order to satisfy practical and/or legal requirements).

Note also that, from a computational point of view, the proposed model of Eq. (25) has the same applicability and computational efficiency of the one proposed in [84, 89]. In fact, in order to sample $h \in \mathcal{H}$ according to this particular $\mathtt{Q}_d$, there are two main cases. In the first case the cardinality of $\mathcal{H}$ is finite and reasonably small to compute exactly $\mathtt{q}_d(h)$. In the second case $\mathcal{H}$ contains too many functions (or even an infinite number), and consequently we have to resort to a subsampling of $\mathcal{H}$ via Monte Carlo estimation[4] in order to make the problem treatable and then compute $\mathtt{q}_d(h)$. Note that this last approach may produce numerical problems when $\gamma$ is large, but, as we will see later, $\gamma$ is never too big in order to keep the algorithm consistent, stable, and privacy preserving.

If the $\mathtt{Q}_d$ and $\mathtt{P}$ defined respectively in Eqns. (25) and (27) are exploited we can prove the following theorem.

**Theorem 2.** *Given the prior* $\mathtt{P}$ *and the posterior* $\mathtt{Q}_d$ *defined in Eqns.* (27) *and* (25)*, we can state that*

$$\mathbb{P}_{\boldsymbol{D}}\left\{\mathtt{KL}[\mathtt{Q}_{\boldsymbol{D}}||\mathtt{P}] \geq \mathtt{KL}_2(\delta, n, n_{1,+}, n_{2,+})\right\} \leq 6\delta, \tag{29}$$

*where*

$$\begin{aligned}
&\mathtt{KL}_2(\delta, n, n_{1,+}, n_{2,+}) = a^2 + 2a\sqrt{b} + b, \qquad (30)\\
&a = \gamma\left(\sqrt{\tfrac{1}{2n}} + \lambda\left(\sqrt{\tfrac{1}{2n_{1,+}}} + \sqrt{\tfrac{1}{2n_{2,+}}}\right)\right),\\
&b = 2\gamma\left(\sqrt{\tfrac{1}{2n}\ln\left(\tfrac{2\sqrt{n}}{\delta}\right)} + \lambda\left(\sqrt{\tfrac{1}{2n_{1,+}}\ln\left(\tfrac{2\sqrt{n_{1,+}}}{\delta}\right)} + \sqrt{\tfrac{1}{2n_{2,+}}\ln\left(\tfrac{2\sqrt{n_{2,+}}}{\delta}\right)}\right)\right).
\end{aligned}$$

The proof is reported in Appendix B.2.

By plugging the results of Theorem 2 into Theorems 10 and 1 it is possible to obtain a fully empirical bound on the risk and the fairness of the RC where the prior $\mathtt{P}$ and the posterior $\mathtt{Q}_d$ are defined respectively in Eqns. (27) and (25).

[4] And, in this case, a further problem would be to control the additional estimation gap, but this is in practice often negligible and out of the scope of this paper.

**Theorem 3.** *Given the prior* $\mathtt{P}$ *and the posterior* $\mathtt{Q}_d$ *defined in Eqns.* (27) *and* (25)*, we can simultaneously bound the risk and the fairness of the corresponding RC*

$$\mathbb{P}_{\boldsymbol{D}}\left\{\left|\widehat{R}^{\ell}_{\boldsymbol{D}}(G_{\mathtt{Q}_{\boldsymbol{D}}})-R^{\ell}(G_{\mathtt{Q}_{\boldsymbol{D}}})\right|\geq\sqrt{\frac{1}{2n}\left[\mathtt{KL}_2(\delta,n,n_{1,+},n_{2,+})+\ln\left(\frac{2\sqrt{n}}{\delta}\right)\right]}\right\}\leq 7\delta, \tag{31}$$

$$\mathbb{P}_{\boldsymbol{D}}\left\{\left|\widehat{F}_{\boldsymbol{D}}(G_{\mathtt{Q}_{\boldsymbol{D}}})-F(G_{\mathtt{Q}_{\boldsymbol{D}}})\right|\geq\sqrt{\frac{1}{2n_{1,+}}\left[\mathtt{KL}_2(\delta,n,n_{1,+},n_{2,+})+\ln\left(\frac{2\sqrt{n_{1,+}}}{\delta}\right)\right]}\right.$$
$$\left.+\sqrt{\frac{1}{2n_{2,+}}\left[\mathtt{KL}_2(\delta,n,n_{1,+},n_{2,+})+\ln\left(\frac{2\sqrt{n_{2,+}}}{\delta}\right)\right]}\right\}\leq 8\delta, \tag{32}$$

*using the same notation of Theorem 2.*

The prove is not reported since it comes trivially from the application of the union bound [116].

The final rate of the bound is $O(\sqrt{\ln(\min(n_{1,+},n_{2,+}))/\min(n_{1,+},n_{2,+})})$, which is optimal in the general case [84, 117] (see the state-of-the-art bound of Theorem 11) since we are simultaneously controlling the risk and the fairness based on three empirical estimators $\widehat{R}^{\ell}_{d}(G_{\mathtt{Q}_d})$, $\widehat{R}^{\ell}_{d_{1,+}}(G_{\mathtt{Q}_d})$, and $\widehat{R}^{\ell}_{d_{2,+}}(G_{\mathtt{Q}_d})$ which exploit respectively $n$, $n_{1,+}$, and $n_{2,+}$ samples (note also that $n \geq \max(n_{1,+},n_{2,+})$). In order to ensure the consistency of the bound, $\gamma$ can increase at maximum with a rate which is $O(\sqrt{\min(n_{1,+},n_{2,+})})$, that is again similar to what has been obtained in [84, 117] since we are exploiting the estimators mentioned above. Since larger $\gamma$ means more weight to function close to the $\widehat{f}^*$ and the less to the others, we would like that $\gamma$ is as large as possible so the maximum rate of increase of $\gamma$ is a very important parameter. Instead, for what concerns $\lambda$, it is important to note that this parameter regulates the trade-off between accuracy and fairness, and for this reason it is usually considered constant and depends on the particular application.

The RC generalization abilities can be also studied, both in terms of its risk and its fairness, using the AS theory, analogously to what has been done in Appendix A.1.2. Theorem 13 allows to bound the risk of a distribution stable algorithm. The following theorem allows to bound the fairness of a distribution stable algorithm.

**Theorem 4.** *If the criteria exploited for choosing a symmetric posterior distribution satisfy the Distribution Stability property described in Theorem 13, then we can state that*

$$\mathbb{P}_{\boldsymbol{D}}\left\{\left|\widehat{F}_{\boldsymbol{D}}(G_{\mathtt{Q}_{\boldsymbol{D}}})-F(G_{\mathtt{Q}_{\boldsymbol{D}}})\right|\right. \tag{33}$$
$$\left.\geq 4\beta_{\mathtt{Q}}^{\ell}+\left(4n_{1,+}\beta_{\mathtt{Q}}^{\ell}+1\right)\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n_{1,+}}}+\left(4n_{2,+}\beta_{\mathtt{Q}}^{\ell}+1\right)\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n_{2,+}}}\right\}\leq 4\delta.$$

The proof is not reported since analogous to the one of Appendix B.1.

Furthermore, if we use the criteria described in Eq. (25) for defining $\mathtt{Q}_d$ it is possible to also show the following result.

**Theorem 5.** *The criteria described in Eq.* (25) *for defining* $\mathtt{Q}_d$ *shows the Distribution Stability property, in fact*

$$\beta_{\mathtt{Q}}^{\ell}\leq\frac{2\gamma}{n}+\frac{2\lambda\gamma}{\min[n_{1,+},n_{2,+}]}. \tag{34}$$

The proof is reported in Appendix B.3.

Note that, the result of Theorem 5 reduces to the one of [112] (see Appendix A) when no fairness constraint is active (namely $\lambda=0$), while the stability decreases when $\lambda>0$. This phenomenon is somehow expected since the larger is $\lambda$ the more importance the constraint has over the final model.

Since Theorem 5 states that the criteria described in Eq. (25) for defining $\mathtt{Q}_d$ shows the Distribution Stability property, it is possible to obtain a fully empirical bound on the risk and the fairness of the corresponding RC.

**Theorem 6.** *The risk and the fairness of the RC which uses as* $\mathtt{Q}_d$ *the prob-*

*ability density function defined in Eq.* (25) *can be bounded as follows*

$$\mathbb{P}_{\boldsymbol{D}}\left\{\left|R^{\ell}(G_{\mathbb{Q}_{\boldsymbol{D}}})-\widehat{R}^{\ell}_{\boldsymbol{D}}(G_{\mathbb{Q}_{\boldsymbol{D}}})\right| \right. \tag{35}$$

$$\left. \geq \frac{4\gamma}{n}+\frac{4\lambda\gamma}{\min[n_{1,+},n_{2,+}]}+\left(8\gamma+\frac{8n\lambda\gamma}{\min[n_{1,+},n_{2,+}]}+1\right)\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right\}\leq 2\delta,$$

$$\mathbb{P}_{\boldsymbol{D}}\left\{\left|\widehat{F}_{\boldsymbol{D}}(G_{\mathbb{Q}_{\boldsymbol{D}}})-F(G_{\mathbb{Q}_{\boldsymbol{D}}})\right| \right. \tag{36}$$

$$\geq \frac{8\gamma}{n}+\frac{8\lambda\gamma}{\min[n_{1,+},n_{2,+}]}+\left(\frac{8n_{1,+}\gamma}{n}+\frac{8n_{1,+}\lambda\gamma}{\min[n_{1,+},n_{2,+}]}+1\right)\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n_{1,+}}}$$

$$\left.+\left(\frac{8n_{2,+}\gamma}{n}+\frac{8n_{2,+}\lambda\gamma}{\min[n_{1,+},n_{2,+}]}\beta^{\ell}_{\mathbb{Q}}+1\right)\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n_{2,+}}}\right\}\leq 4\delta.$$

The proof consists in simply plugging the result of Theorem 5 in Theorems 13 and 4.

The final rate of the bound is $O(\sqrt{1/\min(n_{1,+},n_{2,+})})$, which is optimal in the general case since we are simultaneously controlling the risk and the fairness based on three empirical estimators which exploit respectively $n$, $n_{1,+}$, and $n_{2,+}$ samples and better than the one obtained with the PB theory. In order to ensure the consistency of the bound, $\gamma$ can increase with a rate slower than $O(\sqrt{\min(n_{1,+},n_{2,+})})$, which is a worse result than the one obtained with the PB theory since we would like $\gamma$ as large as possible.

### *3.2. Fair and Private Randomized Learning Algorithms*

In this section we will show that the posterior distribution defined in Eq. (25) can be exploited to derive a RLA which simultaneously possesses accuracy, fairness, and privacy properties thanks to the use of the DP theory.

Theorem 15 allows to bound the risk of an $\epsilon$-DP RLA, while the following results allow to bound its fairness.

**Theorem 7.** *Let $\mathscr{A}$ be an $\epsilon$-DP, then for any $t>0$ and for*

$$\epsilon\leq\sqrt{t^2-\frac{\ln(2)}{3\max[n_{1,+},n_{2,+}]}}, \tag{37}$$

*we can state that*

$$\mathbb{P}_{\mathscr{A},\boldsymbol{D}}\left\{\left|F^{\ell}(\mathscr{A}(\boldsymbol{D}))-\widehat{F}^{\ell}_{\boldsymbol{D}}(\mathscr{A}(\boldsymbol{D}))\right|\geq t\right\}\leq 6\sqrt{2}\mathrm{e}^{-\min[n_{1,+},n_{2,+}]t^2}. \tag{38}$$

The proof is not reported since analogous to the one of Appendix B.1.

At this point, analogously to what has been done in Appendix A.2, let us suppose that our RLA works in this particular way. $\mathscr{A}$ selects one single $h \in \mathcal{H}$ according to a distribution which depended on the data $\mathtt{Q}_d$. If we exploit Eq. (25) for defining $\mathtt{Q}_d$ we can state the following result.

**Theorem 8.** *Let us consider as $\mathscr{A}$ a RLA which, given a dataset $d$, selects a function $h \in \mathcal{H}$ according to the $\mathtt{Q}_d$ defined in Eq. (25). Then $\mathscr{A}$ is $(2\gamma/n + 2\gamma\lambda/\min[n_{1,+}, n_{2,+}])$-DP.*

The proof of this result is reported in Appendix B.4

Note that, the result of Theorem 5 reduces to the one of [112] (see Appendix A) when no fairness constraint is active (namely $\lambda = 0$), while the privacy decreases when $\lambda > 0$. This is an expected phenomenon, in fact, the larger is $\lambda$ the more importance the constraint has over the final model, and consequently the less is the ability of the algorithm to protect the single observation given the further data-dependent constraint.

Thanks to the result of Theorem 8 we can state that the RLA which selects a function $h \in \mathcal{H}$, given $d$, according to the $\mathtt{Q}_d$ defined in Eq. (25) is a privacy preserving RLA. Moreover, exploiting Theorems 15 and 7 we can also bound the risk and the fairness of the final model.

**Theorem 9.** *Let us consider as $\mathscr{A}$ a RLA which, given a dataset $d$, selects a function $h \in \mathcal{H}$ according to the $\mathtt{Q}_d$ defined in Eq. (25). Then $\mathscr{A}$ is $(2\gamma/n + 2\gamma\lambda/\min[n_{1,+}, n_{2,+}])$-DP and by setting*

$$\gamma \leq \frac{\sqrt{\frac{1}{n}\left[\ln\left(\frac{3\sqrt{2}}{2}\right) - \frac{\ln(2)}{3}\right]}}{\frac{2}{n} + \frac{2\lambda}{\min[n_{1,+}, n_{2,+}]}} \tag{39}$$

*we can state that*

$$\mathbb{P}_{\mathscr{A},\boldsymbol{D}}\left\{\left|R^{\ell}(\mathscr{A}(\boldsymbol{D})) - \widehat{R}^{\ell}_{\boldsymbol{D}}(\mathscr{A}(\boldsymbol{D}))\right| \geq \sqrt{\frac{\ln\left(\frac{3\sqrt{2}}{\delta}\right)}{n}}\right\} \leq \delta. \tag{40}$$

*While by setting*

$$\gamma \leq \frac{\sqrt{\frac{1}{\min[n_{1,+}, n_{2,+}]}\left[\ln\left(\frac{6\sqrt{2}}{2}\right) - \frac{\ln(2)}{3}\right]}}{\frac{2}{n} + \frac{2\lambda}{\min[n_{1,+}, n_{2,+}]}} \tag{41}$$

*we can state that*

$$\mathbb{P}_{\mathscr{A},\boldsymbol{D}}\left\{\left|F^{\ell}(\mathscr{A}(\boldsymbol{D}))-\widehat{F}^{\ell}_{\boldsymbol{D}}(\mathscr{A}(\boldsymbol{D}))\right|\geq\sqrt{\frac{\ln\left(\frac{6\sqrt{2}}{\delta}\right)}{\min[n_{1,+}+n_{2,+}]}}\right\}\leq\delta. \tag{42}$$

The proof consists in simply plugging the result of Theorem 8 in Theorems 15 and 7.

Note that, the bound on $\gamma$ of Eq. (41) in Theorem 9 clearly shows the tension between privacy, fairness, and accuracy of the model. In fact, if we suppose that the model which minimizes the error also satisfies the fairness constraint, the best option would be to choose $\gamma, \lambda = \infty$ since, as a consequence, this would results in the deterministic selection of the best model. Since in real world this will not happen, we will have to chose the best desired trade-off between accuracy and fairness using $\lambda \in (0, \infty)$ but still select $\gamma = \infty$ in order to have a deterministic selection of the model which best fits our fairness and accuracy desires. On the other hand, if we ignore the fairness constraint (namely $\lambda = 0$) we cannot select the model which minimizes the error (namely $\gamma = \infty$), since $\gamma$ should be small enough to protect the privacy of the individuals. Moreover, if we enforce the data-dependent fairness constraints (namely $\lambda > 0$), the more we impose this constraint (i.e. the larger is $\lambda$), the smaller should be $\gamma$, in order to still protect the privacy of the individuals.

Note that, the final rate of the bounds is $O(\sqrt{1/n})$ for the risk and $O(\sqrt{1/\min[n_{1,+}, n_{2,+}]})$ for the fairness, which is optimal in the general case since we are simultaneously controlling the risk and the fairness based on three empirical estimators which exploit respectively $n$, $n_{1,+}$, and $n_{2,+}$ samples and better than the ones obtained with the PB and AS theories. In order to ensure the consistency of the bound $\gamma$ can increase with a maximum rate of $O(\sqrt{\min[n_{1,+}, n_{2,+}]})$, which is better than the results obtained with the AS theory and equivalent to the one obtained with the PB theory since we would like $\gamma$ as large as possible. Moreover, the RLA shows, as mentioned above, privacy preserving properties.

In conclusion, we were able to derive a RLA which possesses interesting privacy preserving properties ensuring generalization, fairness, and privacy of the final model.

## 4. Discussion and Conclusions

In this paper we addressed the problem of randomized learning and generalization of fair and private classifiers where, using randomized learning algorithms and/or randomized classifiers, we want to simultaneously ensure that sensitive information does not unfairly influence the outcome of a classifier and to be able to learn from data while preserving the privacy of individual observations. We first faced this issue in the PAC-Bayes framework presenting a new approach trading off and bounding risk and fairness of the randomized classifier by further developing the idea that the PAC-Bayes prior can be defined based on the data-generating distribution without the need of knowing it explicitly. In particular, we defined a prior and a posterior with the purpose of giving more weight to functions which exhibit good generalization and fairness properties. Furthermore, we showed that this randomized classifier possesses interesting stability properties using the algorithmic distribution stability theory. Finally, we also showed that the newly proposed posterior can be exploited to define a randomized accurate and fair algorithm with interesting privacy preserving properties ensuring generalization, fairness, and privacy of the final model using the differential privacy theory.

In this paper we also discussed the advantages and the disadvantages of the different approaches exploiting the derived theoretical results which still leave open questions about the most practical and effective way to learn fair and private classifiers using randomized learning. For this reason, in the future, it would be interesting to try to translate these theoretical results into practical algorithms, in order to be able to compare them with other state-of-the-art approaches that have been developed in the literature.

## Appendix A. State-of-the-art results

In order to improve the readability of this appendix, a summary of the results can be found in Table A.4.

### *Appendix A.1. Randomized Classifiers and Generalization*

Thanks to the PB [83, 84, 86, 90, 93, 118, 119] and the AS [90, 96–98, 120] theories it is possible to bound the risk of a RC selected by a RLA or a DLA. Since the purpose of the paper is not to retrieve optimal and rates of convergence as the extension is trivial and rather technical (see e.g. [83,

| Posterior of Eq. (A.2) and Prior of Eq. (A.4)<br>$\gamma \to 0$ uniform probability over all functions, $\gamma \to \infty$ probability one of the risk minimizer | | | |
|---|---|---|---|
| Theory | PB | AS | DP |
| Algorithm/Classifier | RC & BC<br>Eqns. (A.2) & (A.4) | RC & BC<br>Eq. (A.4) | RLA<br>Eq. (A.4) |
| Property | Accurate | Accurate | Accurate<br>Private |
| Bounds Risk & Fairness | Theorem 11 | Theorem 14 | Theorem 16 |
| Convergence Risk | $O\left(\sqrt{\frac{\ln(n)}{n}}\right)$ | $O\left(\sqrt{\frac{1}{n}}\right)$ | $O\left(\sqrt{\frac{1}{n}}\right)$ |
| $\gamma$ max speed | $O(\sqrt{n})$ | Slower than $O(\sqrt{n})$ | $O(\sqrt{n})$ |

Table A.4: Summary of the results of Appendix A.

90, 120–122]) we will exploit basic bounds which are still rather tight and have optimal rate in the general case (better rates can be achieved only in the lucky case of zero empirical error).

*Appendix A.1.1. Randomized Classifiers and PAC-Bayes Theory*

The PB theory [83, 84, 86, 90, 93, 118, 119] is surely one of the most powerful tools for bounding the risk of a RC. In this section, we will recall the PB-based risk bounds[5].

**Theorem 10** ([84])**.** *For any probability distribution* $\mathtt{P}$ *over* $\mathcal{H}$*, chosen before seeing* $d$ *we can state that*

$$\mathbb{P}_{\boldsymbol{D}}\left\{\exists \mathtt{Q}_d : \left|\widehat{R}^{\ell}_{\boldsymbol{D}}(G_{\mathtt{Q}_{\boldsymbol{D}}}) - R^{\ell}(G_{\mathtt{Q}_{\boldsymbol{D}}})\right| \geq \sqrt{\frac{1}{2n}\left[\mathtt{KL}[\mathtt{Q}_{\boldsymbol{D}}||\mathtt{P}] + \ln\left(\frac{2\sqrt{n}}{\delta}\right)\right]}\right\} \leq \delta. \tag{A.1}$$

The main problem of the PB theory regards the choice of $\mathtt{P}$ and $\mathtt{Q}_d$. $\mathtt{Q}_d$ should fit our observations, but, at the same time, $\mathtt{Q}_d$ should be close to $\mathtt{P}$, in order the minimize the KLD. The milestone result of [89], later extended by [84, 90], proposes to use a Boltzmann prior distribution $\mathtt{P}$ which depends on the data generating distribution $\mathfrak{P}_{\mathcal{Z}}$. In particular, let us suppose that

[5]For showing explicitly the dependency of $\mathtt{Q}$ from $d$ and for clarity, we will indicate $\mathtt{Q}_d$

the density function associated to $\mathtt{P}$ is

$$\mathtt{p}(h) = Z_{\mathtt{P}} \mathtt{e}^{-\gamma R^{\ell}(h)}, \tag{A.2}$$

where $\gamma \in [0, \infty)$ and $Z_{\mathtt{P}}$ is a normalization term such that

$$Z_{\mathtt{P}}^{-1} = \int_{\mathcal{H}} \mathtt{e}^{-\gamma R^{\ell}(h)} dh. \tag{A.3}$$

Basically, this distribution gives more importance to functions that possess small risk. If we choose as posterior $\mathtt{Q}_d$ a distribution which gives more importance to functions with small empirical risk with the following density function

$$\mathtt{q}_d(h) = Z_{\mathtt{Q}_d} \mathtt{e}^{-\gamma \widehat{R}_d^{\ell}(h)}, \tag{A.4}$$

where $\gamma \in [0, \infty)$ and $Z_{\mathtt{Q}_d}$ is a normalization term such that

$$Z_{\mathtt{Q}_d}^{-1} = \int_{\mathcal{H}} \mathtt{e}^{-\gamma \widehat{R}_d^{\ell}(h)} dh, \tag{A.5}$$

it can be proved that this theorem, built on the result of Theorem 10, holds.

**Theorem 11** ([84])**.** *Given the prior* $\mathtt{P}$ *and the posterior* $\mathtt{Q}_d$ *defined in Eqns.* (A.2) *and* (A.4)*, we can state that*

$$\mathbb{P}_{\boldsymbol{D}} \left\{ \mathtt{KL}[\mathtt{Q}_{\boldsymbol{D}} || \mathtt{P}] \geq \mathtt{KL}_1(\gamma, \delta, n) \right\} \leq 2\delta, \tag{A.6}$$

*where*

$$\mathtt{KL}_1(\gamma, \delta, n) \doteq \frac{\gamma^2}{n} + \gamma \sqrt{\frac{2}{n} \ln \left( \frac{2\sqrt{n}}{\delta} \right)}. \tag{A.7}$$

*Consequently, we have that*

$$\mathbb{P}_{\boldsymbol{D}} \left\{ \left| \widehat{R}_{\boldsymbol{D}}^{\ell}(G_{\mathtt{Q}_{\boldsymbol{D}}}) - R^{\ell}(G_{\mathtt{Q}_{\boldsymbol{D}}}) \right| \geq \sqrt{\frac{1}{2n} \left[ \mathtt{KL}_1(\gamma, \delta, n) + \ln \left( \frac{2\sqrt{n}}{\delta} \right) \right]} \right\} \leq 3\delta. \tag{A.8}$$

Finally, let us also recall that it is possible to bound the risk of the BC in terms of the risk of the RC.

**Theorem 12** ([83, 86])**.** *It is possible to state that*

$$R^{\ell}(B_{\mathtt{Q}_d}) \leq 2R^{\ell}(G_{\mathtt{Q}_d}). \tag{A.9}$$

*Appendix A.1.2. Randomized Classifiers and Algorithmic Distribution Stability Theory*

In this section we recall how the AS theory can be exploited for bounding the risk of a RC. We will assume in this section, analogously to [90, 96], that the $\mathtt{Q}_d$ does not depend on the order of the elements in the training set (i.e. it is symmetric), that all functions are measurable and that all sets are countable. Under these assumptions we can recall the following risk bound.

**Theorem 13** ([90])**.** *If the criteria exploited for choosing a symmetric posterior distribution $\mathtt{Q}_d$ and $\mathtt{Q}_{d^{\setminus i}}$ satisfy the Distribution Stability property*

$$\left|\mathbb{E}_{h\sim\mathtt{Q}_d}\left\{\ell\left(h,\cdot\right)\right\}-\mathbb{E}_{h\sim\mathtt{Q}_{d^{\setminus i}}}\left\{\ell\left(h,\cdot\right)\right\}\right|_\infty\leq\beta_{\mathtt{Q}}^{\ell},\quad\forall d\in\mathcal{D},\forall i\in\{1,\cdots,n\},\tag{A.10}$$

*where $\beta_{\mathtt{Q}}^{\ell}$ is a constant that goes to zero at least as $O(1/n)$, then we can state that*

$$\mathbb{P}_{\boldsymbol{D}}\left\{\left|R^{\ell}(G_{\mathtt{Q}_{\boldsymbol{D}}})-\widehat{R}_d^{\ell}(G_{\mathtt{Q}_{\boldsymbol{D}}})\right|\geq 2\beta_{\mathtt{Q}}^{\ell}+\left(4n\beta_{\mathtt{Q}}^{\ell}+1\right)\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right\}\leq 2\delta.\tag{A.11}$$

Furthermore, if we use the criteria described in Eq. (A.4) for defining $\mathtt{Q}_d$ is is possible to also recall the following result.

**Theorem 14** ([90])**.** *The criteria described in Eq.* (A.4) *for defining $\mathtt{Q}_d$ shows the Distribution Stability property, in fact*

$$\beta_{\mathtt{Q}}^{\ell}\leq\frac{2\gamma}{n}.\tag{A.12}$$

*Consequently if we exploit Eq.* (A.4) *for defining $\mathtt{Q}_d$ we can state that*

$$\mathbb{P}_{\boldsymbol{D}}\left\{\left|R^{\ell}(G_{\mathtt{Q}_{\boldsymbol{D}}})-\widehat{R}_d^{\ell}(G_{\mathtt{Q}_{\boldsymbol{D}}})\right|\geq\frac{4\gamma}{n}+(8\gamma+1)\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right\}\leq 2\delta.\tag{A.13}$$

*Appendix A.2. Randomized Learning Algorithms and Generalization: Differential Privacy Theory*

Thanks to the DP [15, 107, 111, 112] theory it is possible to bound the true risk of both a DC or a RC chosen by a RLA. Analogously to what

stated in Appendix A.1, we will not focus on optimal constants and rates (see e.g. [112] for the technical details on this) and moreover, as stated in Section 2, we will deal just with the case of RLA which selects DC since the optimal way of bounding the generalization ability of RC selected with RLA is again the PB or the AS theories.

In order to recall the DP-based bound of the risk of a function selected by a RLA we first have to recall the notion of DP.

**Definitions 1** ([15, 112]). *A RLA $\mathscr{A}$ is $\epsilon$-DP if and only if*

$$\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(d) = h\} \leq \mathbb{e}^{\epsilon}\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(\dot{d}) = h\}, \quad \forall h \in \mathcal{F}, \forall d \in \mathcal{D}, \tag{A.14}$$

*where $\mathbb{e}$ is the Nepero's number and $\epsilon \geq 0$ is a constant.*

Basically, this definition states that the smaller (large $\epsilon$) is the ability to understand if a sample in the dataset has been changed based on $h$ the more private is the RLA, namely the RLA is able to preserve the privacy of the individual samples in the dataset.

Given this definition we can state the following DP-based risk bound.

**Theorem 15** ([107, 112]). *Let $\mathscr{A}$ be an $\epsilon$-DP, then for any $t > 0$ and for*

$$\epsilon \leq \sqrt{t^2 - \frac{\ln(2)}{3n}}, \tag{A.15}$$

*we can state that*

$$\mathbb{P}_{\mathscr{A},\boldsymbol{D}}\left\{\left|R^{\ell}(\mathscr{A}(\boldsymbol{D})) - \widehat{R}^{\ell}_{\boldsymbol{D}}(\mathscr{A}(\boldsymbol{D}))\right| \geq t\right\} \leq 3\sqrt{2}\mathbb{e}^{-nt^2}. \tag{A.16}$$

Let us now suppose that the RLA works in this particular way. $\mathscr{A}$ selects one single $h \in \mathcal{H}$ according to a distribution which depended on the data $\mathtt{Q}_d$. If we exploit Eq. (A.4) for defining $\mathtt{Q}_d$ we can state the following result.

**Theorem 16** ([112]). *Let us consider as $\mathscr{A}$ a RLA which, given a dataset $d$, selects a function $h \in \mathcal{H}$ according to the $\mathtt{Q}_d$ defined in Eq.* (A.4)*. Then $\mathscr{A}$ is $2\gamma/n$-DP. Consequently if we set*

$$\gamma \leq \frac{1}{2}\sqrt{n\left[\ln\left(\frac{3\sqrt{2}}{\delta}\right) - \frac{\ln(2)}{3}\right]}, \tag{A.17}$$

*we can state that*

$$\mathbb{P}_{\mathscr{A},\boldsymbol{D}}\left\{\left|R^{\ell}(\mathscr{A}(\boldsymbol{D}))-\widehat{R}^{\ell}_{\boldsymbol{D}}(\mathscr{A}(\boldsymbol{D}))\right|\geq\sqrt{\frac{\ln\left(\frac{3\sqrt{2}}{\delta}\right)}{n}}\right\}\leq\delta. \tag{A.18}$$

## Appendix B. Proofs

### *Appendix B.1. Proof of Theorem 1*

*Proof.* In order to prove our statement we have to note that, thanks to the reverse triangle inequality, we have that

$$\begin{aligned}
&\left|\widehat{F}_d(G_{\mathtt{Q}_d})-F(G_{\mathtt{Q}_d})\right| \\
&=\left|\left|\widehat{R}^{\ell_H}_{d_{1,+}}(G_{\mathtt{Q}_d})-\widehat{R}^{\ell_H}_{d_{2,+}}(G_{\mathtt{Q}_d})\right|-\left|R^{\ell_H}_{d_{1,+}}(G_{\mathtt{Q}_d})-R^{\ell_H}_{d_{2,+}}(G_{\mathtt{Q}_d})\right|\right| \\
&\leq\left|\widehat{R}^{\ell_H}_{d_{1,+}}(G_{\mathtt{Q}_d})-R^{\ell_H}_{d_{1,+}}(G_{\mathtt{Q}_d})\right|+\left|\widehat{R}^{\ell_H}_{d_{2,+}}(G_{\mathtt{Q}_d})-R^{\ell_H}_{d_{2,+}}(G_{\mathtt{Q}_d})\right|,
\end{aligned} \tag{B.1}$$

and by exploiting the Theorem 10 and the union bound [116] the statement of the theorem is proved. □

### *Appendix B.2. Proof of Theorem 2*

*Proof.* The proof consists in noting that:

$$\begin{aligned}
&\mathtt{KL}[\mathtt{Q}||\mathtt{P}] \\
&=\mathbb{E}_{h\sim\mathtt{Q}_d}\gamma\left(R^{\ell}(h)-\widehat{R}^{\ell}_d(h)+\lambda\left(F(h)-\widehat{F}_d(f)\right)\right)-\ln\left(\frac{Z_{\mathtt{P}}}{Z_{\mathtt{Q}_d}}\right) \\
&=\gamma\left(R^{\ell}(G_{\mathtt{Q}_d})-\widehat{R}^{\ell}_d(G_{\mathtt{Q}_d})+\lambda\left(F(G_{\mathtt{Q}_d})-\widehat{F}_d(G_{\mathtt{Q}_d})\right)\right) \\
&\quad-\ln\left(\int_{\mathcal{H}}\mathtt{p}(h)\mathtt{e}^{-\gamma R^{\ell}(h)-\widehat{R}^{\ell}_d(h)+\lambda\left(F(h)-\widehat{F}_d(h)\right)}dh\right) \\
&\leq\gamma\left(R^{\ell}(G_{\mathtt{Q}_d})-\widehat{R}^{\ell}_d(G_{\mathtt{Q}_d})+\lambda\left(F(G_{\mathtt{Q}_d})-\widehat{F}_d(G_{\mathtt{Q}_d})\right)\right) \\
&\quad+\gamma\left(L^{\ell}(G_{\mathtt{P}})-\widehat{L}^{\ell}(G_{\mathtt{P}})+\lambda\left(F(G_{\mathtt{P}})-\widehat{F}_d(G_{\mathtt{P}})\right)\right),
\end{aligned} \tag{B.2}$$

where the last step follows from the Jensen's inequality [123]. By exploiting this last result, Theorems 10 and 1, and the union bound [116], we have that

the following inequality holds with probability at least $(1-6\delta)$

$$\begin{aligned}
&\mathtt{KL}[\mathtt{Q}_d||\mathtt{P}] \\
&\leq \gamma \left( \sqrt{\frac{\mathtt{KL}[\mathtt{Q}_d||\mathtt{P}] + \ln\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}} \right. \\
&\quad + \lambda \left( \sqrt{\frac{\mathtt{KL}[\mathtt{Q}_d||\mathtt{P}] + \ln\left(\frac{2\sqrt{n_{1,+}}}{\delta}\right)}{2n_{1,+}}} + \sqrt{\frac{\mathtt{KL}[\mathtt{Q}_d||\mathtt{P}] + \ln\left(\frac{2\sqrt{n_{2,+}}}{\delta}\right)}{2n_{2,+}}} \right) \Bigg) \\
&\quad + \gamma \left( \sqrt{\frac{\ln\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}} + \lambda \left( \sqrt{\frac{\ln\left(\frac{2\sqrt{n_{1,+}}}{\delta}\right)}{2n_{1,+}}} + \sqrt{\frac{\ln\left(\frac{2\sqrt{n_{2,+}}}{\delta}\right)}{2n_{2,+}}} \right) \right). \qquad \text{(B.3)}
\end{aligned}$$

The statement of the theorem is obtained by solving with respect to $\mathtt{KL}[\mathtt{Q}_d||\mathtt{P}]$. □

*Appendix B.3. Proof of Theorem 5*

*Proof.* Let us note that the probability density function associated to $\mathtt{Q}_d$ is

$$\mathtt{q}_d(h) = \frac{\mathtt{e}^{-\gamma\left[\frac{1}{|d|}\sum_{z\in d}\ell(h,z) + \lambda\left|\frac{1}{|d_{1,+}|}\sum_{z\in d_{1,+}}\ell_H(h,z) - \frac{1}{|d_{2,+}|}\sum_{z\in d_{2,+}}\ell_H(h,z)\right|\right]}}{\int_{\mathcal{H}} \mathtt{e}^{-\gamma\left[\frac{1}{|d|}\sum_{z\in d}\ell(h,z) + \lambda\left|\frac{1}{|d_{1,+}|}\sum_{z\in d_{1,+}}\ell_H(h,z) - \frac{1}{|d_{2,+}|}\sum_{z\in d_{2,+}}\ell_H(h,z)\right|\right]} dh}, \qquad \text{(B.4)}$$

while the one associated to $\mathtt{Q}_{d^{\setminus i}}$ is

$$\mathtt{q}_{d^{\setminus i}}(h) = \frac{\mathtt{e}^{-\gamma\left[\frac{1}{|d|}\sum_{z\in d^{\setminus i}}\ell(h,z) + \lambda\left|\frac{1}{|d_{1,+}|}\sum_{z\in d^{\setminus i}_{1,+}}\ell_H(h,z) - \frac{1}{|d_{2,+}|}\sum_{z\in d^{\setminus i}_{2,+}}\ell_H(h,z)\right|\right]}}{\int_{\mathcal{H}} \mathtt{e}^{-\gamma\left[\frac{1}{|d|}\sum_{z\in d^{\setminus i}}\ell(h,z) + \lambda\left|\frac{1}{|d_{1,+}|}\sum_{z\in d^{\setminus i}_{1,+}}\ell_H(h,z) - \frac{1}{|d_{2,+}|}\sum_{z\in d^{\setminus i}_{2,+}}\ell_H(h,z)\right|\right]} dh}. \qquad \text{(B.5)}$$

Consequently:

$$\begin{aligned}
\beta^{\ell}_{\mathtt{Q}} &= \left| \mathbb{E}_{h\sim\mathtt{Q}_d}\left\{\ell\left(h,\cdot\right)\right\} - \mathbb{E}_{h\sim\mathtt{Q}_{d^{\setminus i}}}\left\{\ell\left(h,\cdot\right)\right\} \right|_{\infty} && \text{(B.6)} \\
&\leq \frac{2\gamma}{n} + \frac{2\lambda\gamma}{\min[n_{1,+}, n_{2,+}]}. && \text{(B.7)}
\end{aligned}$$

The last upper bound is retrieved, with few technical steps, by substituting Eqns. (B.4) and (B.5) in Eq. (B.6) and then by adding and subtracting the missing term, therefore the statement is proved. □

*Appendix B.4. Proof of Theorem 8*

*Proof.* In order to prove the statement of the theorem let us note that

$$\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(d) = f\} \tag{B.8}$$
$$= \frac{\mathbb{e}^{-\gamma\left[\frac{1}{|d|}\sum_{z\in d}\ell(h,z)+\lambda\left|\frac{1}{|d_{1,+}|}\sum_{z\in d_{1,+}}\ell_H(h,z)-\frac{1}{|d_{2,+}|}\sum_{z\in d_{2,+}}\ell_H(h,z)\right|\right]}}{\int_{\mathcal{H}}\mathbb{e}^{-\gamma\left[\frac{1}{|d|}\sum_{z\in d}\ell(h,z)+\lambda\left|\frac{1}{|d_{1,+}|}\sum_{z\in d_{1,+}}\ell_H(h,z)-\frac{1}{|d_{2,+}|}\sum_{z\in d_{2,+}}\ell_H(h,z)\right|\right]}dh},$$
$$\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(\dot{d}) = f\} \tag{B.9}$$
$$= \frac{\mathbb{e}^{-\gamma\left[\frac{1}{|d^i|}\sum_{z\in d^i}\ell(h,z)+\lambda\left|\frac{1}{|d^i_{1,+}|}\sum_{z\in d^i_{1,+}}\ell_H(h,z)-\frac{1}{|d^i_{2,+}|}\sum_{z\in d^i_{2,+}}\ell_H(h,z)\right|\right]}}{\int_{\mathcal{H}}\mathbb{e}^{-\gamma\left[\frac{1}{|d^i|}\sum_{z\in d^i}\ell(h,z)+\lambda\left|\frac{1}{|d^i_{1,+}|}\sum_{z\in d^i_{1,+}}\ell_H(h,z)-\frac{1}{|d^i_{2,+}|}\sum_{z\in d^i_{2,+}}\ell_H(h,z)\right|\right]}dh}.$$

where $i$ may assume any value in $\{1, \cdots, n\}$. Then let us exploit Definition 1

$$\mathbb{e}^{\epsilon} = \frac{\mathbb{P}\{\mathscr{A}(s) = f\}}{\mathbb{P}\{\mathscr{A}(\dot{s}) = f\}} \tag{B.10}$$
$$\leq \mathbb{e}^{\frac{2\gamma}{n}+\frac{2\gamma\lambda}{\min[n_{1,+},n_{2,+}]}} \tag{B.11}$$

The last upper bound is retrieved, with few technical steps, by substituting Eqns. (B.8) and (B.9) in Eq. (B.10). □

## References


[1] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems, The Journal of Machine Learning Research 15 (1) (2014) 3133–3181.

[2] M. Wainberg, B. Alipanahi, B. J. Frey, Are random forests truly the best classifiers?, The Journal of Machine Learning Research 17 (1) (2016) 3837–3841.

[3] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[4] J. Schmidhuber, Deep learning in neural networks: An overview, Neural networks 61 (2015) 85–117.

[5] G. B. Huang, D. H. Wang, Y. Lan, Extreme learning machines: a survey, International Journal of Machine Learning and Cybernetics 2 (2) (2011) 107–122.

[6] T. G. Dietterich, Ensemble learning, in: The handbook of brain theory and neural networks, 2002.

[7] R. E. Schapire, Y. Freund, Boosting: Foundations and algorithms, MIT press, 2012.

[8] L. Bottou, Stochastic learning, in: Advanced lectures on machine learning, 2004.

[9] L. Oneto, Model Selection and Error Estimation in a Nutshell, Springer, 2019.

[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research 15 (1) (2014) 1929–1958.

[11] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[12] R. Blaser, P. Fryzlewicz, Random rotation ensembles, Journal of Machine Learning Research 17 (4) (2015) 1–15.

[13] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, Y. Theodoridis, State-of-the-art in privacy preserving data mining, ACM Sigmod Record 33 (1) (2004) 50–57.

[14] S. Greengard, Privacy matters, Commun. ACM 51 (9) (2008) 17–18.

[15] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Foundations and Trends in Theoretical Computer Science 9 (3-4) (2014) 1–277.

[16] C. Dwork, Differential privacy: A survey of results, in: International Conference on Theory and Applications of Models of Computation, 2008.

[17] C. Dwork, J. Lei, Differential privacy and robust statistics, in: Annual ACM Symposium on Theory of computing, 2009, pp. 371–380.

[18] C. Dwork, G. N. Rothblum, S. Vadhan, Boosting and differential privacy, in: IEEE Annual Symposium on Foundations of Computer Science, 2010.

[19] O. Williams, F. McSherry, Probabilistic inference and differential privacy, in: Neural Information Processing Systems, 2010.

[20] K. Chaudhuri, D. Hsu, Sample complexity bounds for differentially private learning., in: Conference on Learning Theory, 2011.

[21] J. Lei, Differentially private m-estimators, in: Neural Information Processing Systems, 2011.

[22] S. Song, K. Chaudhuri, A. D. Sarwate, Stochastic gradient descent with differentially private updates, in: IEEE Global Conference on Signal and Information Processing, 2013.

[23] P. Jain, A. G. Thakurta, (near) dimension independent risk bounds for differentially private learning, in: International Conference on Machine Learning, 2014.

[24] S. Oh, P. Viswanath, The composition theorem for differential privacy, in: International Conference on Machine Learning, 2015.

[25] P. Kairouz, S. Oh, P. Viswanath, Secure multi-party differential privacy, in: Neural Information Processing Systems, 2015.

[26] M. J. Kusner, J. Gardner, R. Garnett, K. Weinberger, Differentially private bayesian optimization, in: International Conference on Machine Learning, 2015.

[27] T. Steinke, J. Ullman, Interactive fingerprinting codes and the hardness of preventing false discovery, in: Conference on Learning Theory, 2015.

[28] R. Rogers, S. Vadhan, H. Lim, M. Gaboardi, Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing, in: International Conference on Machine Learning, 2016.

[29] A. Friedman, A. Schuster, Data mining with differential privacy, in: ACM international conference on Knowledge discovery and data mining, 2010.

[30] P. Jain, P. Kothari, A. Thakurta, Differentially private online learning., in: Conference on Learning Theory, 2012.

[31] A. Smith, A. Thakurta, Differentially private feature selection via stability arguments, and the robustness of the lasso, in: Conference on Learning Theory, 2013.

[32] P. Jain, A. Thakurta, Differentially private learning with kernels., in: International Conference on Machine Learning, 2013.

[33] K. Chaudhuri, S. A. Vinterbo, A stability-based validation procedure for differentially private machine learning, in: Neural Information Processing Systems, 2013.

[34] K. Chaudhuri, D. J. Hsu, S. Song, The large margin mechanism for differentially private maximization, in: Neural Information Processing Systems, 2014.

[35] A. Blum, M. Hardt, The ladder: A reliable leaderboard for machine learning competitions, in: International Conference on Machine Learning, 2015.

[36] Y. Wang, Y. X. Wang, A. Singh, Differentially private subspace clustering, in: Neural Information Processing Systems, 2015.

[37] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, D. Bacon, Federated learning: Strategies for improving communication efficiency, arXiv preprint arXiv:1610.05492.

[38] V. Smith, C. K. Chiang, M. Sanjabi, A. S. Talwalkar, Federated multi-task learning, in: Neural Information Processing Systems, 2017.

[39] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, Communication-efficient learning of deep networks from decentralized data, arXiv preprint arXiv:1602.05629.

[40] J. Konečnỳ, H. B. McMahan, D. Ramage, P. Richtárik, Federated optimization: Distributed machine learning for on-device intelligence, arXiv preprint arXiv:1610.02527.

[41] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, K. Seth, Practical secure aggregation for federated learning on user-held data, arXiv preprint arXiv:1611.04482.

[42] R. C. Geyer, T. Klein, M. Nabi, Differentially private federated learning: A client level perspective, arXiv preprint arXiv:1712.07557.

[43] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How to backdoor federated learning, arXiv preprint arXiv:1807.00459.

[44] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2018.

[45] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, M. Pontil, Empirical risk minimization under fairness constraints, in: Advances in Neural Information Processing Systems, 2018.

[46] C. Dwork, N. Immorlica, A. T. Kalai, M. D. M. Leiserson, Decoupled classifiers for group-fair and efficient machine learning, in: Conference on Fairness, Accountability and Transparency, 2018.

[47] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Advances in neural information processing systems, 2016.

[48] M. B. Zafar, I. Valera, M. Gomez Rodriguez, K. P. Gummadi, Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment, in: International Conference on World Wide Web, 2017.

[49] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: International Conference on Machine Learning, 2013.

[50] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, Avoiding discrimination through causal reasoning, in: Neural Information Processing Systems, 2017.

[51] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: Neural Information Processing Systems, 2017.

[52] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, K. R. Varshney, Optimized pre-processing for discrimination prevention, in: Neural Information Processing Systems, 2017.

[53] M. Joseph, M. Kearns, J. H. Morgenstern, A. Roth, Fairness in learning: Classic and contextual bandits, in: Neural Information Processing Systems, 2016.

[54] F. Chierichetti, R. Kumar, S. Lattanzi, S. Vassilvitskii, Fair clustering through fairlets, in: Neural Information Processing Systems, 2017.

[55] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, A. Roth, Fair learning in markovian environments, in: Conference on Fairness, Accountability, and Transparency in Machine Learning, 2016.

[56] S. Yao, B. Huang, Beyond parity: Fairness objectives for collaborative filtering, in: Neural Information Processing Systems, 2017.

[57] K. Lum, J. Johndrow, A statistical framework for fair predictive algorithms, arXiv preprint arXiv:1610.08077.

[58] I. Zliobaite, On the relation between accuracy and fairness in binary classification, arXiv preprint arXiv:1505.05723.

[59] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, K. P. Gummadi, Fairness constraints: A flexible approach for fair classification, Journal of Machine Learning Research 20 (75) (2019) 1–42.

[60] L. Oneto, M. Donini, M. Pontil, General fair empirical risk minimization, arXiv preprint arXiv:1901.10080.

[61] T. Calders, F. Kamiran, M. Pechenizkiy, Building classifiers with independency constraints, in: IEEE international conference on Data mining, 2009.

[62] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, K. Q. Weinberger, On fairness and calibration, in: Advances in Neural Information Processing Systems, 2017.

[63] A. Beutel, J. Chen, Z. Zhao, E. H. Chi, Data decisions and theoretical implications when adversarially learning fair representations, in:

Conference on Fairness, Accountability, and Transparency in Machine Learning, 2017.

[64] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: International Conference on Knowledge Discovery and Data Mining, 2015.

[65] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, H. Wallach, A reductions approach to fair classification, in: International Conference on Machine Learning, 2018.

[66] A. Cotter, M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, S. You, Training well-generalizing classifiers for fairness metrics and other data-dependent constraints, arXiv preprint arXiv:1807.00028.

[67] J. Adebayo, L. Kagal, Iterative orthogonal feature projection for diagnosing bias in black-box models, in: Conference on Fairness, Accountability, and Transparency in Machine Learning, 2016.

[68] F. Kamiran, T. Calders, Classifying without discriminating, in: International Conference on Computer, Control and Communication, 2009.

[69] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (1) (2012) 1–33.

[70] F. Kamiran, T. Calders, Classification with no discrimination by preferential sampling, in: Machine Learning Conference, 2010.

[71] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: International conference on machine learning, 2014.

[72] C. Lan, J. Huan, Discriminatory transfer, arXiv preprint arXiv:1707.00780.

[73] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-aware classifier with prejudice remover regularizer, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2012.

[74] L. Oneto, M. Donini, A. Maurer, M. Pontil, Learning fair and transferable representations, arXiv preprint arXiv:1901.10080.

[75] H. Edwards, A. Storkey, Censoring representations with an adversary, arXiv preprint arXiv:1511.05897.

[76] C. Louizos, K. Swersky, Y. Li, M. Welling, R. Zemel, The variational fair autoencoder, arXiv preprint arXiv:1511.00830.

[77] D. Madras, E. Creager, T. Pitassi, R. Zemel, Learning adversarially fair and transferable representations, arXiv preprint arXiv:1802.06309.

[78] D. McNamara, C. S. Ong, R. C. Williamson, Provably fair representations, arXiv preprint arXiv:1710.04394.

[79] D. McNamara, C. S. Ong, B. Williamson, Costs and benefits of fair representation learning, in: AAAI Conference on Artificial Intelligence, Ethics and Society, 2019.

[80] Y. Wang, T. Koike-Akino, D. Erdogmus, Invariant representations from adversarially censored autoencoders, arXiv preprint arXiv:1805.08097.

[81] F. Johansson, U. Shalit, D. Sontag, Learning representations for counterfactual inference, in: International conference on machine learning, 2016.

[82] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, J. Ullman, Differentially private fair learning, in: International Conference on Machine Learning, 2019.

[83] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, J. F. Roy, Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm, The Journal of Machine Learning Research 16 (1) (2015) 787–860.

[84] G. Lever, F. Laviolette, J. Shawe-Taylor, Tighter pac-bayes bounds through distribution-dependent priors, Theoretical Computer Science 473 (2013) 4–28.

[85] D. A. McAllester, Some pac-bayesian theorems, in: Computational learning theory, 1998.

[86] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, N. Usunier, Pac-bayes bounds for the risk of the majority vote and the variance of the gibbs classifier, in: Neural Information Processing Systems, 2006.

[87] J. Langford, Tutorial on practical prediction theory for classification, Journal of machine learning research 6 (2005) 273–306.

[88] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, S. Sun, Pac-bayes bounds with data dependent priors, The Journal of Machine Learning Research 13 (1) (2012) 3507–3531.

[89] O. Catoni, PAC-Bayesian Supervised Classification, Institute of Mathematical Statistics, 2007.

[90] L. Oneto, D. Anguita, S. Ridella, Pac-bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis, Pattern Recognition Letters 80 (2016) 200–207.

[91] J. F. Roy, M. Marchand, F. Laviolette, From pac-bayes bounds to quadratic programs for majority votes, in: International Conference on Machine Learning, 2011.

[92] P. Germain, A. Lacoste, M. Marchand, S. Shanian, F. Laviolette, A pac-bayes sample-compression approach to kernel methods, in: International Conference on Machine Learning, 2011.

[93] J. Shawe-Taylor, J. Langford, Pac-bayes & margins, Neural information processing systems.

[94] A. Ambroladze, E. Parrado-Hernández, J. Shawe-Taylor, Tighter pac-bayes bounds, Advances in neural information processing systems.

[95] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, Pac-bayesian learning of linear classifiers, in: International Conference on Machine Learning, 2009.

[96] O. Bousquet, A. Elisseeff, Stability and generalization, Journal of machine learning research 2 (2002) 499–526.

[97] T. Poggio, R. Rifkin, S. Mukherjee, P. Niyogi, General conditions for predictivity in learning theory, Nature 428 (6981) (2004) 419.

[98] A. Elisseeff, T. Evgeniou, M. Pontil, Stability of randomized learning algorithms, Journal of Machine Learning Research 6 (2005) 55–79.

[99] W. H. Rogers, T. J. Wagner, A finite sample distribution-free performance bound for local discrimination rules, The Annals of Statistics (1978) 506–514.

[100] L. Devroye, T. Wagner, Distribution-free inequalities for the deleted and holdout error estimates, IEEE Transactions on Information Theory 25 (2) (1979) 202–207.

[101] L. Devroye, L. Györfi, G. Lugosi, A probabilistic theory of pattern recognition, Springer, 1996.

[102] M. Kearns, D. Ron, Algorithmic stability and sanity-check bounds for leave-one-out cross-validation, Neural Computation 11 (6) (1999) 1427–1453.

[103] S. Mukherjee, P. Niyogi, T. Poggio, R. Rifkin, Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization, Advances in Computational Mathematics 25 (1) (2006) 161–193.

[104] S. Shalev-Shwartz, O. Shamir, N. Srebro, K. Sridharan, Learnability, stability and uniform convergence, The Journal of Machine Learning Research 11 (2010) 2635–2670.

[105] L. Oneto, A. Ghio, S. Ridella, D. Anguita, Fully empirical and data-dependent stability-based bounds, IEEE Transactions on Cybernetics 45 (9) (2015) 1913–1926.

[106] O. Rivasplata, C. Szepesvari, J. S. Shawe-Taylor, E. Parrado-Hernandez, S. Sun, Pac-bayes bounds for stable algorithms with instance-dependent priors, in: Neural Information Processing Systems, 2018.

[107] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, A. Roth, Preserving statistical validity in adaptive data analysis, in: Annual ACM Symposium on Theory of Computing, 2015.

[108] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, A. Roth, Generalization in adaptive data analysis and holdout reuse, in: Neural Information Processing Systems, 2015.

[109] Y. X. Wang, J. Lei, S. E. Fienberg, Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle, The Journal of Machine Learning Research 17 (183) (2016) 1–40.

[110] V. N. Vapnik, Statistical learning theory, Wiley New York, 1998.

[111] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, A. Roth, The reusable holdout: Preserving validity in adaptive data analysis, Science 349 (6248) (2015) 636–638.

[112] L. Oneto, S. Ridella, D. Anguita, Differential privacy and generalization: Sharper bounds with applications, Pattern Recognition Letters 89 (2017) 31–38.

[113] L. Oneto, F. Cipollini, S. Ridella, D. Anguita, Randomized learning: Generalization performance of old and new theoretically grounded algorithms, Neurocomputing 298 (2018) 21–33.

[114] A. B. Tsybakov, Introduction to nonparametric estimation, Springer Science & Business Media, 2008.

[115] L. Oneto, S. Ridella, D. Anguita, Tikhonov, ivanov and morozov regularization for support vector machine learning, Machine Learning 103 (1) (2015) 103–136.

[116] C. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze 8 (1936) 3–62.

[117] D. McAllester, A pac-bayesian tutorial with a dropout bound, arXiv preprint arXiv:1307.2118.

[118] M. Seeger, Pac-bayesian generalisation error bounds for gaussian process classification, The Journal of Machine Learning Research 3 (2002) 233–269.

[119] I. O. Tolstikhin, Y. Seldin, Pac-bayes-empirical-bernstein inequality, in: Neural Information Processing Systems, 2013.

[120] A. Maurer, A second-order look at stability and generalization, in: Conference on Learning Theory, 2017.

[121] M. Younsi, Proof of a combinatorial conjecture coming from the pac-bayesian machine learning theory, arXiv preprint arXiv:1209.0824.

[122] L. Bégin, P. Germain, F. Laviolette, J. F. Roy, Pac-bayesian bounds based on the rényi divergence, in: Artificial Intelligence and Statistics, 2016.

[123] J. L. W. V. Jensen, Sur les fonctions convexes et les inégalités entre les valeurs moyennes, Acta mathematica 30 (1906) 175–193.