ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Generalization Performances of Randomized Classifiers and Algorithms built on Data Dependent Distributions

Luca Oneto[1], Sandro Ridella[2], and Davide Anguita[1]

1 - DIBRIS - University of Genova
Via Opera Pia 13, I-16145 Genova - Italy

2 - DITEN - University of Genova
Via Opera Pia 11A, I-16145 Genova - Italy

Abstract. In this paper we prove that a randomized algorithm based on the data generating dependent prior and data dependent posterior Boltzmann distributions of Catoni (2007) is Differentially Private (DP) and shows better generalization properties than the Gibbs (randomized) classifier associated to the same distributions. For this purpose, we will develop a tight DP-based generalization bound, which improve over the current state-of-the-art Hoeffding-type bound.

1 Introduction

Differential Privacy (DP) addresses the apparently self-contradictory problem of keeping private the information about an individual observation while learning useful information about a population [1]. In particular, a procedure is DP if and only if its output is almost independent from any of the individual observations. In other words, considering a randomized algorithm, the probability of selecting a certain model should not change significantly if one individual is present or not. DP allowed to reach 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 randomized learning algorithm which shows DP properties also generalizes [2, 3, 4]. In this paper we will derive a tight DP-based generalization bound which improves over the state-of-the-art Hoeffding-type bound.

Then, we will show how to use this result together with the recent results of [5] in order to develop a DP algorithm, that we will call Catoni's inspired DP algorithm (CDP), which exploits the randomness introduced by the data dependent posterior distribution developed by [5]. In particular we will show that CDP possesses better generalization properties than a Gibbs (randomized) Classifier (GC) which exploits the Catoni's data generating dependent prior and data dependent posterior distributions (CGC), and consequently the associated Bayes Classifier (BC) [6, 7], even if they are both based on the same data dependent posterior distribution. In particular we will compare our results with the state-of-the-art ones based on PAC-Bayes [7]. Finally, we will show on a simple example the importance of the results obtained in this paper by also underlining some interesting properties which should be better investigated in the future.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

2 Differential Privacy and Generalization

Let us consider the binary classification problem where we have an input space $\mathcal{X}$ and an output space $\mathcal{Y} = \{-1, +1\}$. We indicate with $\mathfrak{P}_{\mathcal{X}}$, $\mathfrak{P}_{\mathcal{Y}}$, and $\mathfrak{P}_{\mathcal{Z}}$ respectively the distributions over $\mathcal{X}$, $\mathcal{Y}$, and the cartesian product between the input and the output space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. From $\mathcal{Z}$ we observe a series of n i.i.d. samples $s = \{z_1, \cdots, z_n\}$. $\boldsymbol{Z}$ is a random variable sampled from $\mathcal{Z}$ according to $\mathfrak{P}_{\mathcal{Z}}$. s is a dataset inside the space of all the possible datasets $\mathcal{S} = \mathcal{Z}^n$. $\mathfrak{P}_{\mathcal{S}}$ is the distribution of probability generated by $\mathfrak{P}_{\mathcal{Z}}$ over $\mathcal{S}$. Analogously to $\boldsymbol{Z}$, $\boldsymbol{S}$ is a random variable sampled from $\mathcal{S}$ according to $\mathfrak{P}_{\mathcal{S}}$. We denote with $\dot{s}$ the neighborhood dataset of s such that $s = \{z_1, \cdots, z_{i-1}, \dot{z}_i, z_{i+1}, \cdots, z_n\}$ where i may assume any value in $\{1, \cdots, n\}$ and $\dot{z}_i$ i.i.d. with z_i. We denote with $\breve{\mathcal{S}}$ a subset of the space of datasets $\mathcal{S}$: $\breve{\mathcal{S}} \subseteq \mathcal{S}$. Let us define with $f : \mathcal{X} \to [-1, 1]$ a function in a space $\mathcal{F}$ of all the possible functions and $\breve{\mathcal{F}} \subseteq \mathcal{F}$. A randomized algorithm $\mathscr{A} : \mathcal{S} \to \mathcal{F}$ maps a dataset $s \in \mathcal{S}$ in a function $f \in \mathcal{F}$ with nondeterministic rules that can be encapsulated in a distribution $\mathfrak{P}_{\mathscr{A}}$ over $\mathcal{F}$ given $s \in \mathcal{S}$. We also define an operator $\breve{D}$ which maps a function $f \in \mathcal{F}$ into a subset of all the possible datasets $\breve{\mathcal{S}}$. The accuracy of $f \in \mathcal{F}$ in representing $\mathfrak{P}_{\mathcal{Z}}$ is measured with reference to the hard loss function $\ell : \mathcal{F} \times \mathcal{Z} \to \{0, 1\}$ which counts the number of misclassified examples. Hence, we can define the true risk of f, namely generalization error, as $L(f) = \mathbb{E}_{\boldsymbol{Z}} \ell(f, \boldsymbol{Z})$. Since $\mathfrak{P}_{\mathcal{Z}}$ is unknown, $L(f)$ cannot be computed. Therefore, we have to resort to its empirical estimator, the empirical error $\widehat{L}_n^s(f) = {}^1\!/_n \sum_{i=1}^n \ell(f, z_i)$. Let us recall the definition of DP.

Definitions 1 ([1]). *$\mathscr{A}$ is ϵ-DP if $\forall f \in \mathcal{F}$ and $\forall s \in \mathcal{S}$ we have that $\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(s) = f\} \leq e^{\epsilon} \mathbb{P}_{\mathscr{A}}\{\mathscr{A}(\dot{s}) = f\}$.*

The milestone result of [2] shows that an ϵ-DP algorithm generalizes. In particular two main results are derived. The first one is very general and shows that if a function $\breve{D}(f)$ is defined for each element $f \in \mathcal{F}$ and the probability that $\boldsymbol{S} \in \breve{D}(f)$ is small, then the probability remains small if f is chosen based on $\boldsymbol{S}$ and $\mathscr{A}$. In other words the probability that $\boldsymbol{S} \in \breve{D}(\mathscr{A}(\boldsymbol{S}))$ remains small[1].

Theorem 1 ([2]). *Let $\mathscr{A}$ be an ϵ-DP. Let us suppose that $\mathbb{P}_{\boldsymbol{S}}\{\boldsymbol{S} \in \breve{D}(f)\} \leq \beta$, $\forall f \in \mathcal{F}$. Then, for $\epsilon \leq \sqrt{\ln({}^1\!/_\beta)/2n}$ we have that $\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{\boldsymbol{S} \in \breve{D}(\boldsymbol{F})\} \leq 3\sqrt{\beta}$.*

The second result, which builds upon Theorem 1, shows that the empirical error of a function chosen with an ϵ-DP algorithm is concentrated around its generalization error.

Corollary 1 ([2]). *Let $\mathscr{A}$ be an ϵ-DP, then for any $t > 0$, setting $\epsilon \leq t$ ensures that $\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{L(\boldsymbol{F}) \geq \widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) + t\} \leq 3e^{-nt^2}$.*

The limitation of Corollary 1 is the slow convergence rate $O({}^1\!/_{\sqrt{n}})$. When the empirical error is small we would like to retrieve a Chernoff-type result [8]. For this reason let us recall the following inequality.

Theorem 2 ([8]). *Given an $f \in \mathcal{F}$ it is possible to prove that $\mathbb{P}_{\boldsymbol{S}}\{L(f) \geq \widehat{L}_n^{\boldsymbol{S}}(f) + \sqrt{4L(f)}t\} \leq e^{-2nt^2}$.*

[1] From now on with a little abuse of notation we will identify $\boldsymbol{F} = \mathscr{A}(\boldsymbol{S})$.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

By combining Theorem 2 with Theorem 1 it is possible to prove the following result which improves the state-of-the-art one of Corollary 1.

Corollary 2. *Let $\mathscr{A}$ be an ϵ-DP, then for any $t > 0$, setting $\epsilon \leq t$ ensures that* $\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{L(\boldsymbol{F}) \geq \widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) + \sqrt{4L(\boldsymbol{F})t}\} \leq 3e^{-nt^2}$.

Proof. Let us consider Theorem 2. By setting in Theorem 2 $\breve{D}(f) = \{s \in \mathcal{S} : L(f) \geq \widehat{L}_n^{\boldsymbol{S}}(f) + \sqrt{4L(f)t}\}$ and $\beta = e^{-2nt^2}$ we have that for $\epsilon \leq t$ the statement of the theorem is proved.

Note that the rate of convergence of Corollary 2 can be faster with respect to the one of Corollary 1. In fact when $\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) = 0$ the convergence of the bound can reach $O(1/n)$. Corollary 2 can be further tightened by exploiting the exact confidence intervals for Binomial tails.

Theorem 3 ([9])**.** *Given an $f \in \mathcal{F}$ it is possible to prove that $\mathbb{P}_{\boldsymbol{S}}\{L(f) \geq \mathtt{Q}[1-\delta; n\widehat{L}_n^{\boldsymbol{S}}(f)+1, n-n\widehat{L}_n^{\boldsymbol{S}}(f)]\} \leq \delta$ where $\mathtt{Q}[p; v, w]$ is the p-th quantile from a Beta distribution with shape parameters v and w.*

By combining Theorem 3 with Theorem 1, analogously to what has been done for Corollary 2, it is possible to prove the following result which improves the one of Corollary 2.

Corollary 3. *Let $\mathscr{A}$ be an ϵ-DP, then for any $t > 0$, setting $\epsilon \leq \sqrt{\ln(1/\delta)/2n}$ ensures that $\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{L(\boldsymbol{F}) \geq \mathtt{Q}[1-\delta; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})+1, n-n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})]\} \leq 3\sqrt{\delta}$.*

3 Catoni's inspired DP algorithm

In order to define the CDP and compare its generalization performance with the one of the CGC, we need to recall some preliminary additional definitions. A GC draws an $f \in \mathcal{F}$, according to a probability distribution $\mathtt{Q}$ over $\mathcal{F}$, each time a label for an input $x \in \mathcal{X}$ is required. For the GC, that we will call $G_{\mathtt{Q}}$, we can define its risk together with its empirical counterpart [10], respectively $L(G_{\mathtt{Q}}) = \mathbb{E}_{f\sim\mathtt{Q}}\{L(f)\}$ and $\widehat{L}_n^s(G_{\mathtt{Q}}) = \mathbb{E}_{f\sim\mathtt{Q}}\widehat{L}_n^s(f)$. Let us recall the definitions of $\xi(n)$ [11], which is a function such that $\sqrt{n} \leq \xi(n) \leq 2\sqrt{n}$, and the Kullback-Leibler Divergence $\mathtt{kl}[q||p] = q\ln[q/p] + [1-q]\ln[1-q/1-p]$. Finally let us recall the data generating dependent prior and data dependent posterior probability distributions introduced by [5]. In particular, the density functions associated to $\mathtt{P}$ and $\mathtt{Q}$, respectively $p(f)$ and $q(f)$, are

$$p(f) = c_p e^{-\gamma L(f)}, \quad q(f) = c_q e^{-\gamma \widehat{L}_n^{\boldsymbol{S}}(f)}, \tag{1}$$

where $\gamma \in [0, \infty)$, while c_p and c_q are normalization terms. Catoni's prior gives more weight to functions with low generalization error while Catoni's posterior gives more weight to functions with low empirical error. γ regulates the decay of these weights with the increase of the generalization or empirical error.

Let us recall the state-of-the-art PAC-Bayes based bounds over the generalization error of the CGC.

Theorem 4 ([10])**.** *Given $\mathtt{P}$ and $\mathtt{Q}$ defined in Eq.* (1) *it is possible to state that* $\mathbb{P}\{\mathtt{kl}[\widehat{L}_n^{\boldsymbol{S}}(G_Q)||L(G_Q)] \geq \gamma^2/2n^2 + \gamma/n\sqrt{2\ln(2\xi_n/\delta)} + \ln(2\xi_n/\delta)/n\} \leq \delta$.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

Note that the rate of convergence of the bound of Theorem 4 ranges from $O(\ln(n)/n)$ when $\widehat{L}_n^{\boldsymbol{S}}(G_Q) = 0$ to $O(\sqrt{\ln(n)/n})$ [10].

After bounding the generalization error of the CGC, we will bound the generalization error of CDP based on the results of the Section 2. First we have to prove that CDP is private.

Theorem 5 (CDP Algorithm)**.** *Let us consider as $\mathscr{A}$ a randomized algorithm which, given a dataset s, selects a function $f \in \mathcal{F}$ based on the density function $q(f) = c_q e^{-\gamma \widehat{L}_n^s(f)}$ where $1/c_q = \int_{\mathcal{F}} e^{-\gamma \widehat{L}_n^s(f)} df$. $\mathscr{A}$ is $2\gamma/n$-DP.*

Proof. Let us apply the definition of ϵ-DP of Theorem 1:

$$\frac{\mathbb{P}\{\mathscr{A}(s)=f\}}{\mathbb{P}\{\mathscr{A}(\dot{s})=f\}} = \frac{e^{-\frac{\gamma}{n}\sum_{i=1}^n \ell(f,z_i)}}{\sum_{f_1\in\mathcal{F}} e^{-\frac{\gamma}{n}\sum_{i=1}^n \ell(f_1,z_i)}} \frac{\sum_{f_1\in\mathcal{F}} e^{-\frac{\gamma}{n}\left(\sum_{i=1,i\neq j}^n \ell(f_1,z_i)+\ell(f_1,\dot{z}_j)\right)}}{e^{-\frac{\gamma}{n}\left(\sum_{i=1,i\neq j}^n \ell(f,z_i)+\ell(f,\dot{z}_j)\right)}} \tag{2}$$

$$\leq \frac{e^0}{\sum_{f_1\in\mathcal{F}} e^{-\frac{\gamma}{n}\sum_{i=1,i\neq j}^n \ell(f_1,z_i)} e^{-\frac{\gamma}{n}}} \frac{\sum_{f_1\in\mathcal{F}} e^{-\frac{\gamma}{n}\sum_{i=1,i\neq j}^n \ell(f_1,z_i)} e^0}{e^{-\frac{\gamma}{n}}} = e^{\frac{2\gamma}{n}}.$$

In order to better understand the result of this section let us clarify the difference between the CDP and CGC. The training phase of the two algorithms is exactly the same: s is sampled and $q(f)$ is computed as defined in Eq. (1) [10, 7]. The forward phase instead is quite different. For CDP once the function $f \in \mathcal{F}$ is chosen based on the $\mathtt{Q}$ of Catoni, namely $f^C \sim \mathtt{Q}$, each time a new sample is sampled from $\mathfrak{P}_{\mathcal{X}}$, namely $x \sim \mathfrak{P}_{\mathcal{X}}$, one has to apply the same f^C to x in order to obtain the label $y = f^C(x)$ [10]. For CGC instead, as also explained at the beginning of this section, each time a new sample $x \sim \mathfrak{P}_{\mathcal{X}}$ has to be labelled, one has to sample a new $f^C \sim \mathtt{Q}$ and then obtain the label $y = f^C(x)$.

At this point we can show that the empirical error of the function chosen with CDP is tightly concentrated around its true expectation.

Theorem 6. *Given the CDP, for any $t > 0$, setting $\gamma \leq n/2\sqrt{\ln(1/\delta)/2n}$ ensures that $\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{L(\boldsymbol{F}) \geq \mathtt{Q}[1-\delta; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) + 1, n - n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})]\} \leq 3\sqrt{\delta}$.*

Proof. The proof consists in noting that CDP is $2\gamma/n$-DP as proven in Theorem 5 and then applying Corollary 3.

Note that Theorem 6 tightly connects DP and the generalization capabilities of the CDP. Unfortunately, it is not possible to set independently γ and the confidence of our statement. In order to be able to set the confidence level we need to reformulate Theorem 6 by fixing γ based on our desired confidence level. Since many γ are allowed we will choose the one that gives the tighter bounds.

Corollary 4. *Given the CDP, which is $2\gamma/n$-DP, if we set $\gamma = 1/2\sqrt{n\ln(3/\delta)}$ we can state that $\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{L(\boldsymbol{F}) \geq \mathtt{Q}[1-\delta^2/9; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) + 1, n - n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})]\} \leq \delta$.*

In order to compare the CDP and CGC we have to set $\gamma = 1/2\sqrt{n\ln(3/\delta)}$ in Corollary 4.

Corollary 5. *Given $\mathtt{P}$ and $\mathtt{Q}$ defined in Eq. (1) with $\gamma = 1/2\sqrt{n\ln(3/\delta)}$ it is possible to state that $\mathbb{P}\{\mathtt{kl}[\widehat{L}_n^{\boldsymbol{S}}(G_Q)||L(G_Q)] \geq \ln(3/\delta)/8n + \sqrt{\ln(3/\delta)/4n}\sqrt{2\ln(\xi_n/\delta)} + \ln(\xi_n/\delta)/n\} \leq 2\delta$.*

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

The result of Corollaries 4 and 5 give us many interesting insights over the learning process of the CDP. In particular, DP shows that in order to obtain tight bounds we need γ to grow as $O(\sqrt{n})$. This means that if we have more data in the training set, DP suggests that we can give more trust to the functions with low empirical error as one can note from Eq. (1). Another way of looking at the same property is from a DP point of view; the more data we have the less noise is needed to preserve privacy. Finally, note that the CDP has better generalization properties than the CGC. In particular, in the same conditions of Corollary 4 where $\gamma = {}^1\!/_2\sqrt{n\ln({}^3\!/_\delta)}$, the bound over the generalization error of CDP of Corollary 5 converges as $O({}^1\!/_{\sqrt{n}})$ in the general case and as $O({}^1\!/_n)$ when the empirical error or variance are small enough. Instead, the CGC PAC-Bayes based generalization bound converges as $O(\sqrt{{}^{\ln(n)}\!/_n})$ in the general case and as $O({}^{\ln(n)}\!/_n)$ when the empirical error or variance are small enough.

4 Discussion

In order to fully understand the result retrieved in this paper let us consider a simple example. In particular, a dataset is created, consisting of $n \in 2 \cdot \{1, \cdots, 150\}$ samples in a bidimensional input space: ${}^n\!/_2$ are equally spaced on a circle of radius 1 and center $(-c, -c)$ while the others ${}^n\!/_2$ are equally spaced on a circle of radius 1 and center (c, c). We choose $c \in \{{}^1\!/_2, 1\}$, $\gamma \in 10^{\{-2,-1.9,\cdots,3\}}$ and $\delta = 0.05$. We choose, as hypothesis space $\mathcal{F}$, all the possible linear separators in the input space. In this scenario, we tested the different results presented in this paper in order to underline some properties. In Figures 1(a) and 1(d) we reported the estimated generalization error based on Theorem 4 when $n = 100$ by varying γ together with the value of γ defined by the DP based on Corollary 4: $\gamma_{\text{DP}} = {}^1\!/_2\sqrt{n\ln({}^3\!/_\delta)}$. We define with γ^*_{GC} the γ which minimizes the estimated generalization error based on Theorem 4. In Figures 1(b) and 1(e) we reported the estimated generalization error of the CDP (Corollary 4) and CGC (Corollary 5) together with the percentage of improvement of Corollary 4 over Corollary 5 by varying n. Finally in Figures 1(c) and 1(f) we reported the estimated generalization error of the CDP (Corollary 4) and Theorem 4 where $\gamma = \gamma^*_{\text{GC}}$ by varying n. From the reported results it is possible to derive two main observations. The first one is that $\gamma_{\text{DP}} \approx \gamma^*_{\text{GC}}$, which means that the γ defined by the DP is very close to the γ which minimizes the estimated generalization error of the GC based on Theorem 4. This means that we can avoid the optimization of γ studied in [12]. Because of space constraints we cannot report all the experiments but also varying δ, n or the data distribution the property still holds. For this reason we think that in the future it will be important to study this property for better understanding this quite interesting result. The second observation is that the generalization properties of CDP are much better than the one of the CGC since the estimated generalization error of CDP can be almost 40% better than the one of the CGC, beside improving the rate of convergence as described in Section 3. This is again something that needs to be better investigated in the future on real world datasets in order to test, in practice, the properties of CDP and CGC against other state of the art algorithms.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

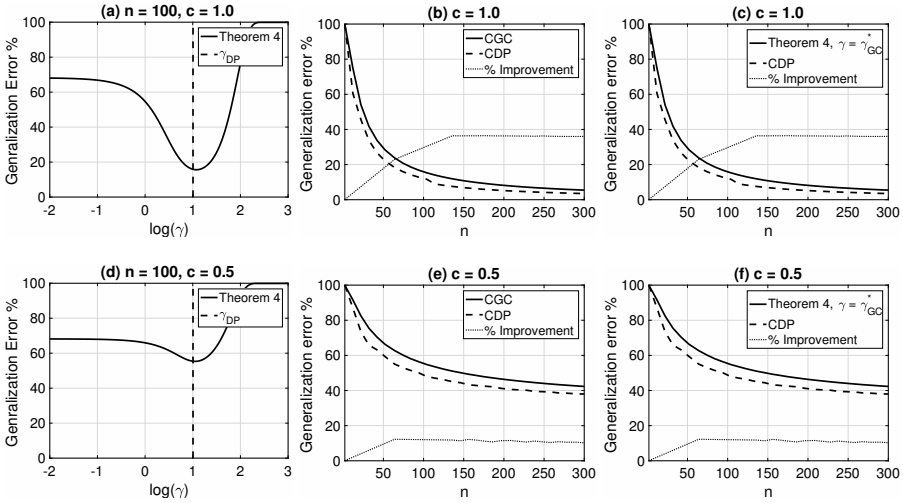

Fig. 1: Performance of the different bounds over the simple example.

References

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

[2] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. In *Symposium on Theory of Computing*, 2015.

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

[4] Y. X. Wang, J. Lei, and 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):1–40, 2016.

[5] O. Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics, 2007.

[6] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and 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(4):787–860, 2015.

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

[8] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[9] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

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

[11] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. Pac-bayesian learning of linear classifiers. In *International Conference on Machine Learning*, 2009.

[12] L. Oneto, S. Ridella, and D. Anguita. Tuning the distribution dependent prior in the pac-bayes framework based on empirical data. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.