PAC-Bayesian Bounds and Beyond : Self-Bounding Algorithms and New Perspectives on Generalization in Machine Learning

Paul Viallard

N°d'ordre NNT : 2022LYSES057

THÈSE de DOCTORAT
DE L'UNIVERSITÉ JEAN MONNET SAINT-ÉTIENNE

Membre de la COMUE de LYON

École Doctorale N° 488
Sciences Ingénierie Santé

Spécialité / discipline de doctorat :
Informatique

Soutenue publiquement le 07/12/2022, par :
Paul Viallard

PAC-Bayesian Bounds and Beyond: Self-Bounding Algorithms and New Perspectives on Generalization in Machine Learning

---

Devant le jury composé de :

Gribonval, Rémi	Directeur de recherche	ENS Lyon, Inria	Président
Canu, Stéphane	Professeur	INSA de Rouen	Rapporteur
Ralaivola, Liva	VP Research	Criteo AI Lab	Rapporteur
Tommasi, Marc	Professeur	Université de Lille, Inria	Examinateur
Habrard, Amaury	Professeur	Université de Saint-Etienne	Directeur
Germain, Pascal	Professeur adjoint	Université Laval	Co-encadrant
Morvant, Emilie	Maître de conférences	Université de Saint-Etienne	Co-encadrante

Acknowledgements

First of all, I would like to thank Liva and Stéphane for reviewing my thesis (in a short time). Moreover, I would like to thank Marc and Rémi for accepting to be, respectively, the examiner and the president; I am very honored to have you on my jury!

I would like to sincerely thank my supervisors, who were always here and patient with me; I believe that I was very lucky! Thank you, Amaury, for proposing this PhD subject to me and for always having a critical viewpoint on my work. In the meantime, I always felt free to work where and when I want; thank you. Thank you, Pascal, for your guidance when writing the papers. Our discussions were always interesting, and I learned a lot. Visiting Laval University has not been possible with the pandemic... I hope that it will be possible later! Thank you, Emilie, for being there literally every day for me. I really enjoyed talking about science and personal things with you. Thank you for showing me (for a long time) how to write a scientific paper and how to do presentations. Moreover, thank you for letting me the chance to perform lectures (which was hard but so interesting). Lastly, thanks, Emilie and Cécile, for making me discover Kung Fu (which I gave up but enjoyed a lot)! Thanks, Valentina and Rémi, for having contributed together. This thesis would not have looked like this without you.

This PhD journey has also been full of fun moments and scientific discussions among the permanent and non-permanent staff of the lab. Generally, thank you all[1] for the good mood in the lab that I really enjoyed during these years. More specifically, thanks, Léo and Jules, for introducing me to the "ASEC" parties that were awesome with you! I also want to thank the "ASEC" team (*i.e.*, Aubin, Damien, Jules, and, Léo) for the good nights. Unfortunately, the pandemic[2] stopped, for a long time, these good moments...

Moreover, this thesis marks the end of all those school years. I would have never bet on myself about writing a "book", especially in English! This could never have been done if I had stopped studying early; thanks, mom and dad, for your support during these years. More generally, I want to thank my family, who have all been supportive during these moments of hard work with the projects, the exams, and during the PhD.

I want to thank some special friends that have been important to me since primary,

[1] The list is too long to be enumerated!

[2] Thanks, Brrrmmaou, Titi, and Bibi, for sleeping next to me when I was working; it was motivating!

middle, high school, and the university. Thanks, Jules, for all these years: I would never have been the same without you! Thank you for making me discover the Loire gorges along with hiking since middle school. Our philosophical discussions in this area are so memorable! Your recklessness helped me a lot to push my limits, *e.g.*, during the urban and cave explorations. Thank you for accompanying me in my first climbing and paragliding! Thanks, Fleurian, for these school years: our hiking and climbing sessions were so cool (and usually with Jules)! Moreover, thank you for making me discover a new sport while writing my thesis: karting. I hope that there will be other road trips with the Dyane car! Thanks, Benoit, for sharing my passion for computer science and coding together for all these years. For instance, we coded an optimization algorithm in high school! Likewise, thanks, Angélique, for sharing your interest in computer science with me since high school. I hope that we will continue to hike or drink tea at the Oriol Tower! Thanks, Guillaume, for these wonderful student years spent with you since the first year of "licence"! We shared many parties, met our respective friends, and had good times. Moreover, thank you for doing sports with me, such as climbing, ice skating, and hiking (during classes). These good moments were accompanied by a lot of work: for you for being my project partner for almost five years! I also appreciated our first internship together at the same desk and our NeurIPS paper! It is funny where life leads us. I would like to thank Coralie for asking me for information about horse riding during my first internship. Since then, I have ridden again and learned a lot about horses! Your restlessness (and your desire to always do something) has been valuable, especially when avoiding thinking about the thesis.

Last but not least, thank you, Chloé, for your unwavering support over these past two years. It was precious, notably for the difficult moment like the submissions and the thesis writing! Thank you for removing (sometimes) my nose from the grindstone to enjoy life (with you). I really hope to continue horse riding, hiking, traveling, or having common projects.

The further backward you look, the further forward you can see.

Winston Churchill

Science is a bit like the joke about the drunk who is looking under a lamppost for a key that he has lost on the other side of the street, because that's where the light is. It has no other choice.

Noam Chomsky

Contents

Contents **6**

List of Figures **9**

List of Theorems **12**

List of Notations **15**

Preamble **17**

List of Publications **25**
International Conference 25
International Workshop 25
National Conference 25
Research Report 26

I Background 27

1 An introduction to Statistical Learning Theory **29**
1.1 Introduction 30
1.2 Hypothesis Selection 40
1.3 Generalization Bounds 43
1.4 Conclusion and Summary 53

2 The PAC-Bayesian Theory and the Majority Vote **55**
2.1 Introduction 56
2.2 PAC-Bayesian Majority Votes 56
2.3 PAC-Bayesian Bounds 64
2.4 Disintegrated PAC-Bayesian Bounds 75
2.5 Conclusion and Summary 78

II PAC-Bayesian Majority Vote: Theory and Self-bounding Algorithms 79

3 PAC-Bayesian Theory for the Robust Majority Vote **81**

3.1 Introduction . . . 82
3.2 Adversarially Robust Majority Vote . . . 83
3.3 Adversarially Robust PAC-Bayes . . . 87
3.4 Experimental Evaluation on Differentiable Decision Trees . . . 95
3.5 Conclusion and Summary . . . 98

4 Self-Bounding Algorithms for the Majority Vote 101
4.1 Introduction . . . 102
4.2 Setting . . . 103
4.3 State of the Art: PAC-Bayesian Bounds for the Majority Vote . . . 104
4.4 Contribution: Algorithms based on the PAC-Bayesian C-Bounds . . . 108
4.5 Experiments . . . 114
4.6 Conclusion and Summary . . . 123

5 Toward a Stochastic Majority Vote 125
5.1 Introduction . . . 126
5.2 The Stochastic Majority Vote . . . 129
5.3 From a PAC-Bayesian Bound to an Algorithm . . . 133
5.4 Experiments . . . 138
5.5 Conclusion and Summary . . . 141

III Beyond PAC-Bayesian Bounds: From Disintegration to Novel Bounds 147

6 On the Practical uses of the Disintegrated Bounds 149
6.1 Introduction . . . 150
6.2 Setting and PAC-Bayesian Bounds . . . 151
6.3 Disintegrated PAC-Bayesian Theorems . . . 152
6.4 The Disintegration in Action . . . 158
6.5 Experiments with Neural Networks . . . 162
6.6 Perspectives for the Majority Vote . . . 171
6.7 Summary and Conclusion . . . 172

7 Generalization Bounds with Complexity Measures 175
7.1 Introduction . . . 176
7.2 Preliminaries . . . 177
7.3 Integrating Arbitrary Complexities in Generalization Bounds . . . 179
7.4 Using Arbitrary Complexities in Practice . . . 184
7.5 Comparison with the Generalization Bounds of the Literature . . . 192
7.6 Conclusion and Summary . . . 197

IV Conclusion and Perspectives **199**

Conclusion and Perspectives **201**

V Appendix **205**

A Some Mathematical Tools **207**
A.1 JENSEN's Inequality . . . 207
A.2 MARKOV's Inequality . . . 207
A.3 2nd Order MARKOV's Inequality . . . 208
A.4 CHEBYSHEV-CANTELLI Inequality . . . 209
A.5 HÖLDER's Inequality . . . 209

B Appendix of Chapter 2 **213**

C Appendix of Chapter 3 **245**

D Appendix of Chapter 4 **265**

E Appendix of Chapter 5 **285**

F Appendix of Chapter 6 **303**

G Appendix of Chapter 7 **347**

References **367**

List of Figures

Preamble

1 Illustration of the Supervised Classification Setting in Machine Learning . . 18
2 Example of the Majority Vote's Prediction 20
3 Example of a Majority Vote with Three Voters 21

Chapter 1

1.1 Representration of Statistical Learning 30
1.2 Schematic Representation of a Data Distribution 32
1.3 Example of Decision Stump 33
1.4 Example of Decision Tree 34
1.5 Plot of the Exclusive Or Function 36
1.6 Examples of Linear Classifiers 37
1.7 Example of Neural Network 37
1.8 Details on the Neural Network of Figure 1.7 38
1.9 Plot of Different Losses 39
1.10 Representation of the Underfitting, Overfitting and the Generalization . . . 42
1.11 Illustration of the VC-dimension for Linear Classifiers in 2-dimension 46
1.12 Illustration of the Rademacher Complexity in Multi-class Classification . . . 47
1.13 Schematic Representation of the Uniform Stability 49
1.14 Schematic Representation of the Algorithmic Robustness 51

Chapter 2

2.1 Illustration of the Margin of the Majority Vote 59
2.2 Plots that Summarize the Relationship Between the Surrogates 65
2.3 Illustration of the Functions $\underline{\mathrm{kl}}$ and $\overline{\mathrm{kl}}$ 71
2.4 Illustration of the Tightness of $\underline{\mathrm{kl}}$ and $\overline{\mathrm{kl}}$ 72

Chapter 3

3.1 Illustration of the Adversarial Examples 82
3.2 Visualization of the impact of the TV term in Equation (3.9) 98
3.3 Visualization of the risk and bound values for "Defense=Attack" 99

Chapter 4

4.2 Comparison Between the Test Risks and the Bounds (1/6) 115
4.3 Comparison Between the Test Risks and the Bounds (2/6) 116
4.4 Comparison Between the Test Risks and the Bounds (3/6) 117
4.5 Comparison Between the Test Risks and the Bounds (4/6) 118
4.6 Comparison Between the Test Risks and the Bounds (5/6) 119
4.7 Comparison Between the Test Risks and the Bounds (6/6) 120

Chapter 5

5.1 Comparison of the Minimization of the Surrogates on Moons 126
5.2 Comparison of the Self-bounding Algorithms on Moons 128
5.3 Examples of Probability Density Functions for the Dirichlet Distribution . . 130
5.4 Plot of the Digamma Function and its Derivative 134
5.5 Plot of the Performance of Algorithms 5.1 and 5.2 on Moons (1/2) 140
5.6 Plot of the Performance of Algorithms 5.1 and 5.2 on Moons (2/2) 141
5.7 Comparison of the Risks and the Bound Values (1/4) 142
5.8 Comparison of the Risks and the Bound Values (2/4) 143
5.9 Comparison of the Risks and the Bound Values (3/4) 144
5.10 Comparison of the Risks and the Bound Values (4/4) 145

Chapter 6

6.1 Evolution of the Bound Values in Terms of the Split Ratio (1/2) 165
6.2 Evolution of the Bound Values in Terms of the Split Ratio (2/2) 166
6.3 Comparisons of the PAC-Bayesian Bounds and the Disintegrated Bounds . 167
6.4 Comparison of the Disintegrated Bounds and the Test Risks 168
6.5 Evolution of the Bound Values in Terms of the Variance σ^2 170

Chapter 7

7.1 Tightness of Equations (7.7) and (7.8) on MNIST 188
7.2 Tightness of Equations (7.7) and (7.8) on FashionMNIST 189
7.3 Influence of the Parameter α 190
7.4 Influence of the Depth/Width 191

List of Algorithms

Chapter 1

1.1 Example of Decision Stump 33
1.2 Example of Decision Tree 34
1.3 Empirical Risk Minimization 41
1.4 Structural Risk Minimization 42

Chapter 2

2.1 Compute $\overline{\mathrm{kl}}(q|\tau)$ *resp.* $\underline{\mathrm{kl}}(q|\tau)$ through the bisection method 72

Chapter 3

3.1 Average Adversarial Training with Guarantee 94

Chapter 4

4.1 Minimization of Equation (4.2) by Stochastic Gradient Descent 108
4.2 Minimization of Equation (4.3) by Stochastic Gradient Descent 111
4.3 Minimization of Equation (4.4) by Stochastic Gradient Descent 113

Chapter 5

5.1 Approximating the Stochastic Risk 131
5.2 Computing Exactly the Stochastic Majority Vote's Risk 133
5.3 Minimization of Theorem 5.3.1's Bound 137
5.4 Minimization of Corollary 5.3.1's Bound 137

Chapter 7

7.1 Stochastic MALA 185

List of Theorems

Chapter 1

1.1.1 Definition (Learning sample) . . . 31
1.1.2 Definition (Loss function) . . . 38
1.1.3 Definition (Empirical Risk) . . . 39
1.1.4 Definition (True Risk) . . . 40
1.3.1 Definition (PAC Generalization Bound) . . . 43
1.3.2 Definition (Uniform Convergence Bound) . . . 43
1.3.1 Theorem (Generalization Bound for Finite $\mathbb{H}$) . . . 44
1.3.3 Definition (Vapnik-Chervonenkis (VC) Dimension) . . . 45
1.3.2 Theorem (VC-Dimension-based Generalization Bounds) . . . 45
1.3.4 Definition (Rademacher Complexity) . . . 46
1.3.3 Theorem (Rademacher-complexity-based Generalization Bound) . . . 47
1.3.5 Definition (Algorithmic-dependent Generalization Bound) . . . 48
1.3.6 Definition (Uniform Stability) . . . 49
1.3.4 Theorem (Stability-based Bounds) . . . 50
1.3.7 Definition (Robustness) . . . 50
1.3.5 Theorem (Robustness-based Bounds) . . . 51
1.3.6 Theorem (PAC-Bayesian Bound of Maurer (2004)) . . . 52

Chapter 2

2.2.1 Definition (Majority Vote) . . . 56
2.2.2 Definition (Risk of the Majority Vote) . . . 58
2.2.3 Definition (Margin of the Majority Vote) . . . 58
2.2.4 Definition ($\frac{1}{2}$-Margin of the Majority Vote) . . . 59
2.2.5 Definition (Gibbs Risk) . . . 60
2.2.1 Theorem (Risk Upper Bound Based on the Gibbs Risk) . . . 61
2.2.6 Definition (Joint Error) . . . 61
2.2.2 Theorem (Risk Upper Bound Based on the Joint Error) . . . 62
2.2.7 Definition (Disagreement) . . . 62
2.2.3 Theorem (The C-Bound) . . . 63
2.2.4 Theorem (Relationship between Theorems 2.2.1 to 2.2.3) . . . 64
2.3.1 Definition (PAC-Bayesian Generalization Bound) . . . 66
2.3.1 Theorem (General PAC-Bayesian Bound of Germain *et al.* (2009)) . . . 67
2.3.2 Theorem (PAC-Bayesian Bound of McAllester (2003)) . . . 67

2.3.3 Theorem (PAC-Bayesian Bound of Catoni (2007)) 68
2.3.2 Definition (KL Divergence Between Two Bernoulli Distributions) 69
2.3.4 Theorem (PAC-Bayesian Bound of Seeger (2002)) 70
2.3.3 Definition (Inverting Functions of $\mathrm{kl}()$) 70
2.3.1 Proposition (Donsker-Varadhan Variational Representation) . . . 73
2.3.5 Theorem (General PAC-Bayesian Bound of Bégin *et al.* (2016)) . . . 74
2.4.1 Definition (Disintegrated PAC-Bayesian Generalization Bound) 75
2.4.1 Theorem (General Disintegrated Bound of Rivasplata *et al.* (2020)) 76
2.4.2 Theorem (Disintegrated Bound of Catoni (2007)) 76
2.4.3 Theorem (Disintegrated Bound of Blanchard and Fleuret (2007)) 77

Chapter 3

3.2.1 Definition (True/Empirical Adversarial Risk) 84
3.2.2 Definition (Averaged Adversarial Risk) 86
3.2.3 Definition (Averaged-Max Adversarial Risk) 87
3.3.1 Proposition (Relations Between the Averaged Adversarial Risks) 87
3.3.2 Proposition (Classical and Averaged Adversarial Risks) 88
3.3.1 Definition (Surrogates on the Averaged Adversarial Risks) 89
3.3.1 Theorem (Upper Bounds on the Surrogates) 89
3.3.2 Theorem (PAC-Bayesian Bound on $r_{\mathcal{E}}(\rho)$) 90
3.3.3 Theorem (PAC-Bayesian Bound on $a_{\mathcal{E}^n}(\rho)$) 90
3.3.1 Corollary (PAC-Bayesian Bound on $r_{\mathcal{E}}(\rho)$) 92
3.3.2 Corollary (PAC-Bayesian Bound on $a_{\mathcal{E}^n}(\rho)$) 92

Chapter 4

4.3.1 Theorem (PAC-Bayesian Bound Based on the Gibbs Risk) 104
4.3.2 Theorem (PAC-Bayesian Bound Based on the Joint Error) 105
4.3.3 Theorem (PAC-Bayesian C-Bound of Roy *et al.* (2016)) 106
4.3.4 Theorem (PAC-Bayesian C-Bound of Germain *et al.* (2015)) 107
4.3.5 Theorem (PAC-Bayesian C-Bound of Lacasse *et al.* (2006)) 107
4.4.1 Theorem (Reformulation of Lacasse *et al.*'s PAC-Bayesian C-Bound) 111

Chapter 5

5.2.1 Definition (Risks of the stochastic majority vote) 129
5.2.2 Definition (Dirichlet Distribution) 130
5.2.1 Lemma (Computation of the Stochastic Risk) 131
5.2.1 Corollary (Closed-form Solution of the Stochastic Risks) 132
5.3.1 Theorem (PAC-Bayesian Bound for Stochastic Majority Votes) 133

5.3.2 Theorem (PAC-Bayesian bound with data-dependent voters) 135
5.3.1 Corollary (PAC-Bayesian bound with data-dependent voters) 136

Chapter 6

2.3.5 Theorem (General PAC-Bayesian Bound of Bégin *et al.* (2016)) . . . 151
2.4.1 Definition (Disintegrated PAC-Bayesian Generalization Bound) 152
6.3.1 Theorem (General Disintegrated PAC-Bayes Bound) 153
6.3.1 Corollary (Extreme Cases of Theorem 6.3.1) 154
2.4.1 Theorem (General Disintegrated Bound of Rivasplata *et al.* (2020)) 155
6.3.2 Theorem (Parametrizable Disintegrated PAC-Bayes Bound) 156
6.3.1 Proposition (Optimal Bound of Theorem 6.3.2) 157
6.4.1 Corollary (Instantiation of Theorem 6.3.1 for Neural Networks) 159
6.4.2 Corollary (Instantiation of Known Bounds for Neural Networks) 159
6.4.3 Corollary (PAC-Bayesian Bound for Stochastic Neural Networks) 161
6.6.1 Corollary (Instantiation of Theorem 2.4.1 to Stochastic Majority Votes) 171
6.6.2 Corollary (Instantiation of Theorem 6.3.1 to Stochastic Majority Votes) 172

Chapter 7

2.4.1 Definition (Disintegrated PAC-Bayesian Generalization Bound) 178
2.4.1 Theorem (General Disintegrated Bound of Rivasplata *et al.* (2020)) 178
7.3.1 Definition (Generalization Bound with Complexity Measures) 180
7.3.1 Theorem (Generalization Bound with Complexity Measures) 182
7.3.1 Corollary (Practical Generalization Bound with Complexity Measures) . 182
7.5.1 Proposition (Set-theoretic view of Definition 7.3.1) 193
1.3.2 Definition (Uniform Convergence Bound) 193
7.5.2 Proposition (Set-theoretic View of Uniform Convergence Bounds) . . . 194
7.5.1 Corollary (Uniform Convergence Bound from Theorem 7.3.1) 194
1.3.5 Definition (Algorithmic-dependent Generalization Bound) 195
7.5.3 Proposition (Set-theoretic View of Algorithmic Dependent Bounds) . . 196
7.5.2 Corollary (Algorithmic-dependent Bound from Theorem 7.3.1) 196

Appendix

A.1.1 Theorem (Jensen's Inequality) 207
A.2.1 Theorem (Markov's Inequality) 207
A.3.1 Theorem (2nd Order Markov's Inequality) 208
A.4.1 Theorem (Chebyshev-Cantelli Inequality) 209
A.5.1 Lemma (Young's Inequality) 209
A.5.1 Theorem (Hölder's Inequality) 210

List of Notations

General

a	A scalar (integer or real)
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
$\mathbb{A}, \mathbb{a}$	A set
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_*$	The set of real numbers excluding 0
$\mathbb{R}_*^+$	The set of positive real numbers excluding 0
$\mathbb{N}$	The set of natural numbers
$\mathbb{N}_*$	The set of natural numbers excluding 0
$\mathrm{card}(\cdot)$	The cardinal of a set
a_i	i-th element of the vector $\mathbf{a}$

Statistical Learning Theory

$\mathbb{X}$	Set of d-dimensional inputs ($\subseteq \mathbb{R}^d$)
$\mathbb{Y}$	Set of labels
$\mathbf{x}$	A real-valued input $\mathbf{x} \in \mathbb{X}$
y	A label $y \in \mathbb{Y}$ associated to the input $\mathbf{x}$
$\mathcal{D}$	Unknown data distribution on $\mathbb{X} \times \mathbb{Y}$
$\mathcal{D}^m$	Unknown data distribution on the m-samples, *i.e.*, on $(\mathbb{X}\times\mathbb{Y})^m$
$\mathbb{S}$	Learning sample $\mathbb{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ drawn from $\mathcal{D}^m$
$\mathcal{S}$	Uniform distribution on $\mathbb{S}$
$\mathbb{T}$	Test set drawn from $\mathcal{D}^m$

$\mathcal{T}$	Uniform distribution on $\mathbb{T}$
$\mathbb{H}$	The set of hypotheses
h	A hypothesis $h \in \mathbb{H}$
$\ell(\cdot, \cdot)$	Loss function
$\mathrm{R}^{\ell}_{\mathcal{D}'}(h)$	Risk of the hypothesis $h \in \mathbb{H}$ *w.r.t.* the loss function $\ell()$ on $\mathcal{D}'$
$\mathrm{R}_{\mathcal{D}'}(h)$	Risk of the hypothesis $h \in \mathbb{H}$ *w.r.t.* the 0-1 loss on $\mathcal{D}'$

Probability Theory

$\mathbb{E}_{X\sim\mathcal{X}}[\cdot]$	The expectation *w.r.t.* the random variable $X \sim \mathcal{X}$
$\mathbb{P}_{X\sim\mathcal{X}}[\cdot]$	The probability *w.r.t.* the random variable $X \sim \mathcal{X}$
$\mathrm{I}[a]$	Indicator function; returns 1 if a is true and 0 otherwise
$\mathbb{M}(\mathbb{H})$	Set of Probability densities *w.r.t.* the reference measure on $\mathbb{H}$
ρ	Posterior distribution $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$
π	Prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$
$\mathrm{KL}(\rho\|\pi)$	Kullback-Leibler (KL) divergence between ρ and π
$\mathrm{D}_\alpha(\rho\|\pi)$	Rényi Divergence between ρ and π
$\mathrm{Uni}(\mathbb{A})$	Uniform distribution on $\mathbb{A}$
$\mathrm{Dir}(\boldsymbol{\alpha})$	Dirichlet distribution of parameters $\boldsymbol{\alpha} \in \mathbb{R}^+_*$

Majority Vote

$\mathrm{MV}_\rho(\cdot)$	The majority vote classifier
$m_\rho(\mathbf{x}, y)$	Majority vote's margin for the example $(\mathbf{x}, y) \in \mathbb{X}\times\mathbb{Y}$
$\widehat{m_\rho}(\mathbf{x}, y)$	$\frac{1}{2}$-Margin for the example $(\mathbf{x}, y) \in \mathbb{X}\times\mathbb{Y}$
$\mathrm{sign}(a)$	Sign function; returns $+1$ if $a \geq 0$ and -1 otherwise
$r_{\mathcal{D}'}(\rho)$	Gibbs risk on the distribution $\mathcal{D}'$ associated to the majority vote $\mathrm{MV}_\rho()$
$e_{\mathcal{D}'}(\rho)$	Joint Error on the distribution $\mathcal{D}'$ associated to the majority vote $\mathrm{MV}_\rho()$
$d_{\mathcal{D}'}(\rho)$	Disagreement on the distribution $\mathcal{D}'$ associated to the majority vote $\mathrm{MV}_\rho()$

Preamble

Introduction

Statistical Machine Learning is a subfield of artificial intelligence at the intersection of computer science, statistics, and optimization that consists of a set of learning methods that learn mathematical models[3] automatically to solve a task from a statistical perspective. We refer the reader to the textbook of Russell and Norvig (2020) for a general introduction to artificial intelligence and to Bishop (2007), Mohri *et al.* (2012), or Shalev-Shwartz and Ben-David (2014) for an introduction to machine learning.

Various tasks can be solved through these methods, such as image recognition, medical diagnosis, fraud detection, recommendation system, etc. These machine learning methods aim to find a model h belonging to a set $\mathbb{H}$ that solves a given task. These method assume that we have some data, *i.e.*, a set of examples, that are sufficiently representative of the task. Each example obtained from the task is generally composed of an input represented by some features and its corresponding output. Different types of output can be considered: the supervised regression setting uses real-valued output, while the supervised classification setting assumes that the outputs are categories (*a.k.a.* classes or labels). This thesis stands in the supervised classification setting. The supervised classification methods learn a model, called a classifier, that separates/classifies the inputs into different categories.[4] For instance, in Figure 1, we illustrate an image classification task: it consists in predicting if an image contains a horse or a cat. More precisely, the input is an image, and the label is either "cat" (the red image) or "horse" (the blue image). From the examples in the data, the learned model h (the black line) separates the red images from the blue images: the classifier correctly predicts all the images from the learning sample.

One way to assess if a model performs well on the examples is to compute the probability that the model misclassifies an example in the available data; this quantity is called the *empirical risk*. However, the machine learning model may learn by heart the examples with an empirical risk at nearly 0. In this case, the model may be completely inefficient on new examples from the task, *i.e.*, unseen data; we say that the model *overfits* the data. To characterize if a model performs well on unseen examples from

[3] The machine learning models are also referred to as hypotheses in statistical learning theory.

[4] When the outputs are absent from the examples, unsupervised learning methods learn models that group (*i.e.* cluster) together similar inputs.

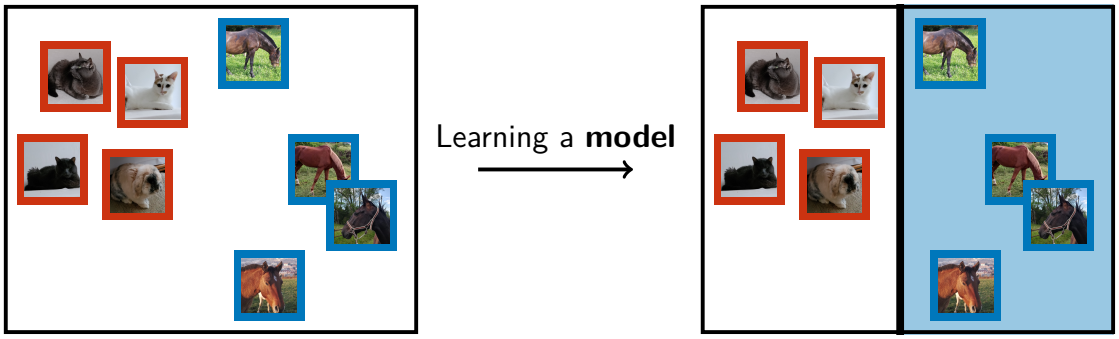


Figure 1. *Illustration of the supervised classification setting in machine learning. Given some labeled examples (i.e., the image and its associated category), a model $h \in \mathbb{H}$ is learned, and then, once learned, it can be used to classify, possibly new examples. The images in the blue area are classified as "horse" while the white area corresponds to the images classified as "cat".*

the task, we can define the notion of *true risk*. This notion is the probability that the model misclassifies an example from the task (represented by an unknown distribution). To assess if the empirical risk is representative of the true risk for a given model h, we are interested in the *generalization gap* defined as

$$\textit{Generalization Gap}(h) = \Big|\textit{True Risk}(h) - \textit{Empirical Risk}(h)\Big|.$$

The model overfits the data when the generalization gap is high (close to 1 in the worst case) while the empirical risk is close to 0. On the contrary, when the gap is close to 0, the empirical risk is a good approximation of the true risk. Hence, to obtain a model that performs well on a task, the generalization gap must be close to 0 and the empirical risk close to 0 as well. However, the gap is not computable because of the true risk since it relies on an unknown quantity: the underlying distribution of the task. Then, another strategy to assess the quality of the model is to consider computable upper bounds on the generalization gap called generalization bounds. The form of the first generalization bound, introduced by Vapnik and Chervonenkis (1968, 1971, 1974), has the following form:

$$\begin{aligned}&\textit{For all model } h \in \mathbb{H},\\ &\Big|\textit{True Risk}(h) - \textit{Empirical Risk}(h)\Big| \leq \textit{Generalization Bound}(\mathbb{H}).\end{aligned}$$

The generalization bounds are *probabilistic*, meaning that with high probability (over the examples sampled from the unknown distribution), the bound holds. They generally depend on the number of examples and a complexity term. This complexity term determines the potential of a model to overfit: the higher the complexity, the more

plausible the model overfits the data. Ideally, the upper bound decreases when the number of examples increases for a given finite fixed complexity term. After the seminal work of VAPNIK and CHERVONENKIS (1968, 1971, 1974), generalization bounds have been extended in several directions (see *e.g.*, MCALLESTER, 1998; BARTLETT and MENDELSON, 2002; BOUSQUET and ELISSEEFF, 2002). By rearranging the terms, a bound (that is computable) on the true risk (uncomputable) can be deduced for all model $h \in \mathbb{H}$:

$$\textit{True Risk}(h) \leq \textit{Empirical Risk}(h) + \textit{Generalization Bound}(\mathbb{H}). \tag{1}$$

This leads to a central point of this thesis: the possibility to derive algorithms minimizing a generalization bound, *e.g.*, the right hand size of Equation (1). Algorithms minimizing a generalization bound are called self-bounding algorithms (FREUND, 1998). The advantage of minimizing a generalization bound is the capacity to directly control or have at least an influence on the evolution of the true risk. In particular, as a side effect, the minimization of the generalization bound allows us to control the overfitting phenomenon better.

One particular type of generalization bounds in which we are specifically interested comes from the PAC-Bayesian framework (SHAWE-TAYLOR and WILLIAMSON, 1997; MCALLESTER, 1998). This framework assumes that each model $h \in \mathbb{H}$ is associated with a positive weight $\rho(h)$ that forms a probability distribution over $\mathbb{H}$ called the *posterior distribution* ρ. Based on this assumption, the PAC-Bayesian generalization bounds allow us to upper-bound the expected generalization gap; the form of the bounds is defined as

$$\underset{h \text{ sampled from } \rho}{\textit{Expectation}} \left[\left| \textit{True Risk}(h) - \textit{Empirical Risk}(h) \right| \right] \leq \textit{Generalization Bound}(\rho).$$

This framework allows to upper-bound the risk of a *stochastic* model which, for each input $\mathbf{x}$, *(i)* samples a new model $h \in \mathbb{H}$ from ρ and *(ii)* predicts the output of $\mathbf{x}$ with $h(\mathbf{x})$. Actually, the risk of the stochastic model can be linked to the risk of a model in which we are particularly interested in this thesis: the majority vote; we provide an overview of such a model in Figure 2. The majority vote has a long history in science: CONDORCET (1785) started to explore mathematically voting systems. Famous machine learning models can be seen as a majority vote, such as linear classifiers, Support Vector Machine (GRAEPEL *et al.*, 2005), k-Nearest Neighbors (BELLET *et al.*, 2014), or neural networks vote (KAWAGUCHI *et al.*, 2017; VIALLARD *et al.*, 2019). Some approaches that learn majority votes belong to the ensemble methods (DIETTERICH, 2000) aiming to combine supervised classifiers (called voters) to create an accurate model. For instance, bagging (BREIMAN, 1996), random forest (BREIMAN, 2001) and boosting (FREUND and SCHAPIRE, 1996) are famous examples of ensemble methods.

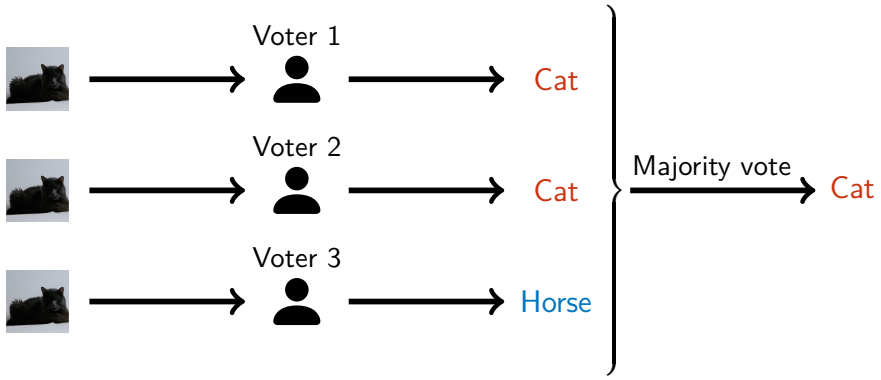


Figure 2. *Example of the majority vote's prediction: given an image, each voter outputs a label ("cat" or "horse"), and the majority vote gathers the results to output the majority label.*

In these methods, the voters' decisions are combined to obtain a better decision compared to the individual voters' decisions (which can be weak). An important notion when combining different classifiers is the notion of diversity (Dietterich, 2000; Kuncheva, 2014). Indeed, when voters are *weak* and perform a bit better than random as in boosting (Freund and Schapire, 1996), sufficiently diverse voters may improve the accuracy of the majority vote. We give in Figure 3 an example of the diversity's importance. The combination can be done in very different ways depending on the methods: in bagging (Breiman, 1996) and random forest (Breiman, 2001), the voters' predictions are only averaged while boosting (Freund and Schapire, 1996) performs a weighted average. In this thesis, we consider a convex combination of the voters where each voter h is associated with the weight $\rho(h)$ encoding its importance in the majority vote. More formally, in the binary classification setting, each model h (*i.e.*, voter) belonging to $\mathbb{H}$ predicts either the class -1 or the class $+1$. A weighted majority vote over the voters in $\mathbb{H}$ applied on a given input $\mathbf{x}$ is defined as:

$$\operatorname{sign}\left[\sum_{h\in\mathbb{H}} \rho(h)h(\mathbf{x})\right],$$

where $\operatorname{sign}[a] = -1$ if $a < 0$ and $\operatorname{sign}[a] = +1$ otherwise.

However, one drawback of the PAC-Bayesian theory is that it is not possible to bound the generalization gap of only one model h in $\mathbb{H}$. Hopefully, the *disintegrated* PAC-Bayesian bounds – introduced by Blanchard and Fleuret (2007) and Catoni (2007) – overcome this drawback. The bounds have the following form:

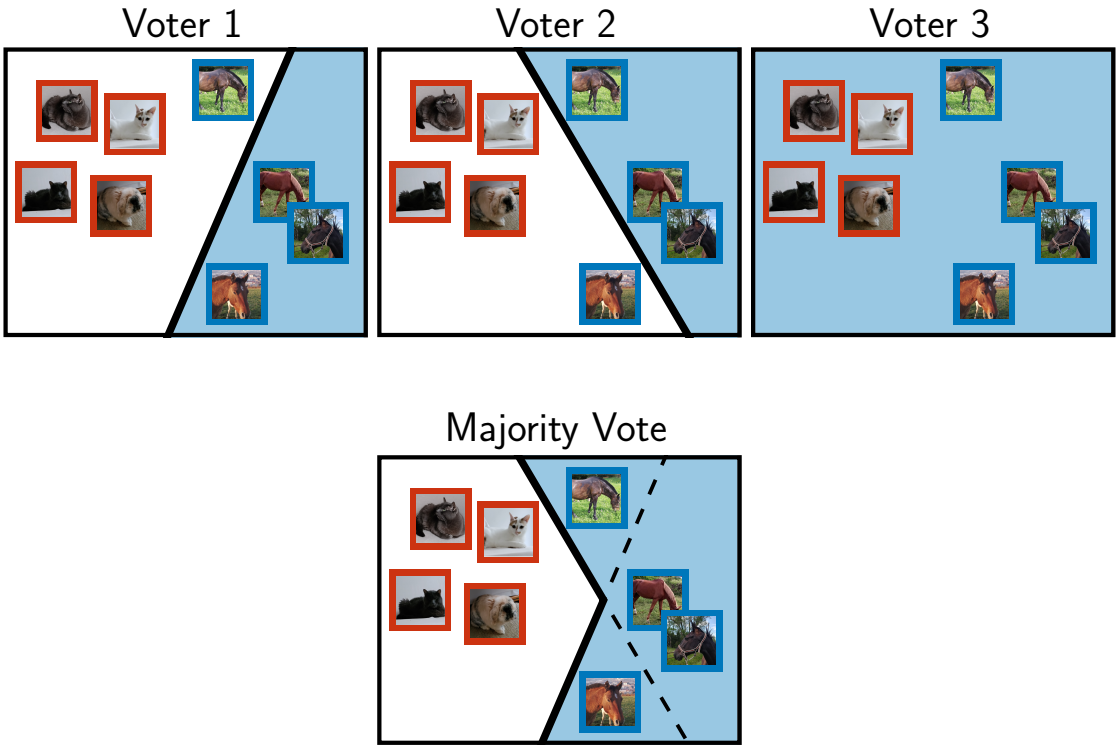


Figure 3. *Example of a majority vote with three voters on the classification horse/cat classification task presented before. Each of the three voters makes some mistakes in the data. However, when the majority rule combines the voters, the final vote classifies all the data correctly. It is mainly because the three voters are diverse: they do not make the same mistakes, while the combination corrects the individual errors.*

$$\text{With high probability over the model } h \text{ sampled from the posterior } \rho,$$
$$\Big|\text{True Risk}(h) - \text{Empirical Risk}(h)\Big| \leq \text{Generalization Bound}(\rho, h).$$

They allow us to obtain a bound on the generalization gap for a unique model h (sampled from the posterior distribution) that holds with high probability and which will serve as a basis for some contributions of this thesis.

Long Story Short

Motivations of this thesis. As discussed above, generalization bounds can be used to assess when machine learning models generalize, *i.e.*, when the empirical risk is repre-

sentative of the true risk. In this context, the PAC-Bayesian theory is adapted to upper-bound the generalization gap of models based on the majority vote or the stochastic classifier. However, in the PAC-Bayesian literature, few works propose to minimize a generalization bound to learn a machine learning model, such as, MASEGOSA *et al.* (2020). In the first series of contributions of this thesis, we develop new self-bounding algorithms for three settings. Firstly, we develop a new adversarial robustness setting tailored for the PAC-Bayesian and robustify majority votes after proving generalization bounds. Secondly, we minimize three particular PAC-Bayesian bounds on the majority vote's risk that was considered difficult to optimize (MASEGOSA *et al.*, 2020). Finally, we introduce a stochastic version of the majority vote, *i.e.*, where the weights are assumed to be sampled from a probability distribution. The stochastic majority vote allows one to derive guarantees on majority-vote-based models. However, upper-bounding the generalization gap of a single classifier with the PAC-Bayesian theory is tedious and generally applicable only to certain classifiers such as the majority vote (see *e.g.*, LANGFORD and SHAWE-TAYLOR, 2002; GERMAIN *et al.*, 2009; LETARTE *et al.*, 2019). In our second series of contributions, we propose to overcome this drawback by considering the notion of disintegrated PAC-Bayesian bounds. Such bounds are able to provide generalization bounds for a single model. By leveraging this framework, we provide new bounds that are easily optimizable and allow us to derive new self-bounding algorithms. In our last contribution, we make use of this framework to develop a general way for incorporating arbitrary complexity measures in generalization bounds.

Outline of this thesis. This thesis is composed of three parts.
Part I is dedicated to the introduction of the field of statistical learning theory and the PAC-Bayesian theory.

(i) Chapter 1 presents the general setting of this thesis. We introduce the notion of learning and solving a task with a statistical machine learning algorithm. Then, we introduce some machine learning models and some methods to learn them. Afterwards, we recall several classical generalization bounds, notably from VAPNIK and CHERVONENKIS (1974), that assess the quality of the obtained model for the chosen task.

(ii) In Chapter 2, we mainly recall some results from the PAC-Bayesian framework. After a reminder about the majority vote, we recall different PAC-Bayesian bounds, which will serve as a basis for deriving new results in Part II. We also remind the first *disintegrated* PAC-Bayesian bounds, which are useful when we are interested in one model sampled from the posterior distribution.

Based on the PAC-Bayesian theory, Part II deals with our first series of contributions focusing on the derivation of self-bounding algorithms (Freund, 1998) that minimize PAC-Bayesian bounds to obtain a majority vote with guarantees on the true risk.

(i) Chapter 3 stands in the adversarial robustness setting (Goodfellow *et al.*, 2015): the goal is to make the majority vote robust to small changes/perturbations in the input. This setting is in contrast with the classical setting in machine learning, where no perturbations are applied to the input. To the best of our knowledge, we are the first to *(i)* formalize the robustness setting in the PAC-Bayesian framework and *(ii)* assess the robustness of the majority vote with this framework. We also derive a self-bounding algorithm that minimizes our new generalization bounds.

(ii) In Chapter 4, we come back to the classical supervised classification setting. We introduce the minimization of PAC-Bayesian bounds on the majority vote risk's surrogate called the C-Bound (recalled in Chapter 2). Unlike the algorithms of the PAC-Bayesian literature, our learning algorithms better consider the voters' correlations.

(iii) However, the self-bounding algorithms (including ours) do not fully exploit the diversity of the voters in general. Hence, after introducing the *stochastic* majority vote in Chapter 5, we develop a self-bounding learning algorithm to minimize the risk. It allows us to optimize the expected risk directly without requiring the use of a surrogate of the majority vote's risk.

The PAC-Bayesian theory, as considered in Part II, has a major drawback. While the majority vote's generalization abilities can be analyzed through the PAC-Bayesian theory, it becomes more difficult to analyze the generalization of a single voter chosen randomly according to the weights. Thanks to the disintegrated bounds (recalled in Chapter 2), we present two contributions in Part III that introduce self-bounding algorithms to choose a *single* classifier.

(i) The disintegrated bounds of the literature are difficult to optimize (and to obtain self-bounding algorithms). Hence, in Chapter 6, we *(i)* derive new disintegrated PAC-Bayesian bounds (easier to optimize) and *(ii)* provide the first empirical study of self-bounding algorithms using these bounds. We also instantiate the bounds with neural networks and compare them with the PAC-Bayesian considered, *e.g.*, by Dziugaite and Roy (2017), Zhou *et al.* (2019), and Pérez-Ortiz *et al.* (2021).

(ii) Chapter 7 offers a new viewpoint on generalization bounds by leveraging the disintegrated PAC-Bayesian framework. To the best of our knowledge, Chapter 7

introduces – for the first time – a way to include arbitrary complexity measures in generalization bounds. This work is another step toward the practical use of generalization bounds since the users can now include their complexity measures.

Finally, Part IV presents some perspectives and future works. Note that, for the sake of completeness and clarity, we provide in Appendix all the proofs; we give a hyperlink to the proof for each theorem, corollary, and proposition. Moreover, in order to reproduce the experiments and the figures, we provide the different source codes developed in the context of this thesis at

`https://github.com/paulviallard/PhDThesis`.

Context of this thesis

This thesis was carried out in the Data Intelligence team of the laboratoire Hubert Curien: a joint research unit (UMR 5516) affiliated with the French National Center for Scientific Research (CNRS); the Institut d'Optique Graduate School, and the Université Jean Monnet in Saint-Etienne, France. The french ANR (Agence Nationale de la Recherche) financially supported this thesis through the project APRIORI (A PAC-Bayesian RepresentatIOn LeaRnIng Perspective) ANR-18-CE23-0015. This research project was also the subject of a collaboration with Université Laval, Québec city, Canada.

List of Publications

International Conference

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. Self-bounding Majority Vote Learning Algorithms by the Direct Minimization of a Tight PAC-Bayesian C-Bound. *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. (2021a).

Paul Viallard, Guillaume Vidot, Amaury Habrard, and Emilie Morvant. A PAC-Bayes Analysis of Adversarial Robustness. *Advances in Neural Information Processing Systems (NeurIPS)*. (2021d).

Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound. *Advances in Neural Information Processing Systems (NeurIPS)*. (2021).

International Workshop

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. Interpreting Neural Networks as Majority Votes through the PAC-Bayesian Theory. *NeurIPS 2019 Workshop on Machine Learning with Guarantees*. (2019).

National Conference

Paul Viallard, Rémi Emonet, Pascal Germain, Amaury Habrard, Emilie Morvant, and Valentina Zantedeschi. Intérêt des bornes désintégrées pour la généralisation avec des mesures de complexité. *Conférence sur l'Apprentissage automatique (CAp)*. (2022a).

Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound. *Conférence sur l'Apprentissage automatique (CAp)*. (2022).

Paul Viallard, Pascal Germain, and Emilie Morvant. Apprentissage de Vote de Majorité par Minimisation d'une C-Borne PAC-Bayésienne. *Conférence sur l'Apprentissage automatique (CAp)*. (2021b).

Paul Viallard, Pascal Germain, and Emilie Morvant. Dérandomisation des Bornes PAC-Bayésiennes. *Conférence sur l'Apprentissage automatique (CAp)*. (2021c).

Guillaume Vidot, Paul Viallard, and Emilie Morvant. Une Analyse PAC-Bayésienne de la Robustesse Adversariale. *Conférence sur l'Apprentissage automatique (CAp)*. (2021).

Paul Viallard, Rémi Emonet, Amaury Habrard, Emilie Morvant, and Pascal Germain. Théorie PAC-Bayésienne pour l'apprentissage en deux étapes de réseaux de neurones. *Conférence sur l'Apprentissage automatique (CAp)*. (2020).

Research Report

Paul Viallard, Rémi Emonet, Amaury Habrard, Emilie Morvant, and Valentina Zantedeschi. Generalization Bounds with Arbitrary Complexity Measures. *Submitted to ICLR 2023*. (2022b).

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A General Framework for the Disintegration of PAC-Bayesian Bounds. *Submitted to Machine Learning Journal*. (2022c).

Part I

Background

1 An introduction to Statistical Learning Theory

Contents

1.1 Introduction **30**
1.1.1 Representation of a Task 30
1.1.2 "Solving" a Task 32
1.2 Hypothesis Selection **40**
1.2.1 Empirical Risk Minimization 41
1.2.2 Structural Risk Minimization 41
1.3 Generalization Bounds **43**
1.3.1 Uniform Convergence Bounds 43
1.3.2 Algorithmic-dependent Generalization Bounds 48
1.3.3 PAC-Bayesian Generalization Bounds 52
1.4 Conclusion and Summary **53**

Abstract

This chapter provides an overview of statistical learning theory. We introduce the main concepts and notations used in this thesis focusing on supervised learning. Moreover, we review the main theoretical frameworks allowing one to derive some guarantees on the quality of the learning process.

1.1 Introduction

Statistical learning encompasses a set of statistical methods that aims to automatically solve a task with the help of a computer from data. For example, the task can consist in identifying a digit in an image. In practice, the task is represented by a database (*a.k.a.* dataset or learning sample) containing a finite number of examples. More precisely, one example is composed of an input and its corresponding output (*a.k.a.* label). For the digit recognition problem, one example is made up of an image (the input) with its corresponding digit (the output).

To solve the task, we have to design a model that outputs the correct label given the input. This model is usually a potentially complicated mathematical function. From a human perspective, designing by hand a mathematical function that solves the task can be tedious. Instead, in statistical learning, we aim to find *automatically* such a function with an algorithm. This algorithm is usually called *learning algorithm* since the process of finding a function boils down to *learning* a function that solves the task. The setting of statistical learning and the learning process is illustrated in Figure 1.1; it is defined more formally in the rest of the section.

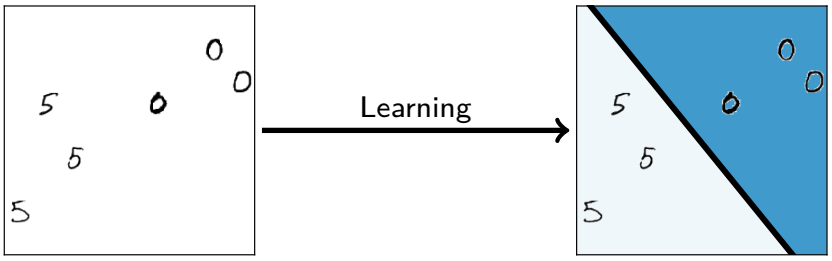


Figure 1.1. *Rough representation of statistical learning for the digit recognition task. On the left, each input (the handwritten digits) is represented in 2 dimensions. On the right, we illustrate the notion of learning, i.e., an algorithm finds a function (represented by the black line) predicting the digit 0 for the points belonging to the blue area and the digit 5 for those belonging to the white one. The black line is also called the decision boundary.*

1.1.1 Representation of a Task

This thesis studies the classification problem from a supervised learning perspective. More formally, we consider a set of d-dimensional inputs $\mathbb{X} \subseteq \mathbb{R}^d$ and a set of outputs $\mathbb{Y}$ (*a.k.a.* set of labels). In the binary classification setting, when $\mathbb{Y} = \{-1, +1\}$,

the input is classified either into the label -1 or $+1$. In the multi-class classification setting, when $\mathbb{Y} = \{1, 2, \dots, l\}$, the input is classified into $l \geq 2$ different labels. Given an input set $\mathbb{X}$ and a label set $\mathbb{Y}$, we assume the task to be represented by an *unknown* function $h^* : \mathbb{X} \to \mathbb{Y}$. The unknown function h^* can be stochastic, meaning that $h^*(\mathbf{x})$ involves randomness, and thus different outputs $y \in \mathbb{Y}$ are probable for the same input $\mathbf{x} \in \mathbb{X}$. Moreover, some inputs $\mathbf{x} \in \mathbb{X}$ are more representative of the task, *i.e.*, some inputs may be more probable than others. To take this probabilities into account, the function h^* is replaced with an *unknown* distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$; the distribution $\mathcal{D}$ represents the probability to sample a given input $\mathbf{x} \in \mathbb{X}$ and the output of the function h^*. More precisely, we assume that each couple $(\mathbf{x}, y) \in \mathbb{X}\times\mathbb{Y}$ (*a.k.a.* example) is a realization from this *unknown* data distribution $\mathcal{D}$ on the set $\mathbb{X}\times\mathbb{Y}$. Even if the distribution $\mathcal{D}$ is *unknown*, we usually assume that we have access to some examples that we hope to be sufficiently representative and sampled from $\mathcal{D}$. This set of examples is defined as follows.

Definition 1.1.1 (Learning sample)**.** We define as *learning sample* (or *training set*) a set of m random variables *independent and identically distributed* (*i.i.d.*) following the distribution $\mathcal{D}$. We have

$$\mathbb{S} \triangleq \bigcup_{i=1}^{m} \Big\{(\mathbf{x}_i, y_i)\Big\} \triangleq \Big\{(\mathbf{x}_i, y_i)\Big\}_{i=1}^{m} \subseteq (\mathbb{X} \times \mathbb{Y})^m,$$

$$\text{where} \quad \forall i \in \{1, \dots, m\}, \quad (\mathbf{x}_i, y_i) \sim \mathcal{D} \quad \text{and} \quad \mathbb{S} = \Big\{(\mathbf{x}_i, y_i)\Big\}_{i=1}^{m} \sim \mathcal{D}^m$$

The notation $\mathcal{D}^m$ denotes the distribution of m examples following $\mathcal{D}$:

$$\mathcal{D}^m\Big(\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}\Big) = \mathcal{D}^m\Big(\mathbb{S}\Big) \triangleq \prod_{i=1}^{m} \mathcal{D}\Big((\mathbf{x}_i, y_i)\Big).$$

In practice, the set of examples, organized in a dataset, can be created manually by a human and/or collected automatically by a computer. One famous dataset associated with the digit recognition task is the MNIST dataset from LeCun *et al.* (1998). For instance, the set of inputs $\mathbb{X}$ is the set of gray-scale images of size 28×28, and the set of labels $\mathbb{Y} = \{0, \dots, 9\}$ is the set of possible digits. The learning sample $\mathbb{S}$ is composed of $m = 60{,}000$ pairs of images (input) and digit (label). A schematic representation of this task is presented in Figure 1.2.

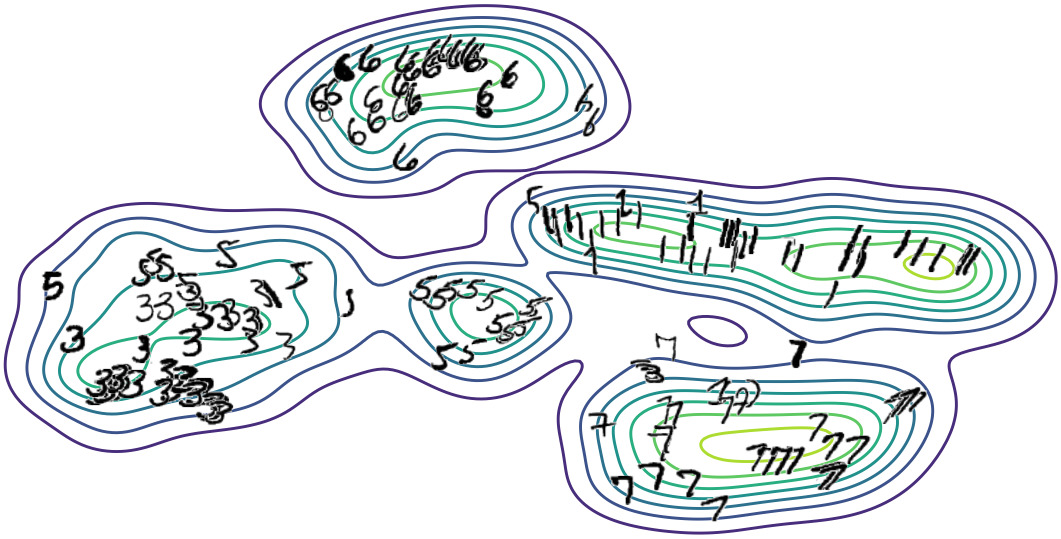


Figure 1.2. *Schematic representation of the distribution $\mathcal{D}$ for the digit recognition task. The density of the distribution $\mathcal{D}$ is schematized with the contour lines where the purple resp. green is a low resp. high-density region. The examples (i.e., the handwritten digits) are sampled in high-density areas. A subset of the learning sample $\mathbb{S} \sim \mathcal{D}^m$ is represented.*

Deducing a mathematical model that predicts, for all $i \in \{1, \ldots, m\}$, the output y_i for the input $\mathbf{x}_i$ is not necessarily trivial. To overcome this problem, statistical learning methods are developed to find *automatically* such a model that aims to solve the task by predicting the examples $(\mathbf{x}_i, y_i) \in \mathbb{X} \times \mathbb{Y}$ correctly.

1.1.2 "Solving" a Task

The model found automatically by the statistical learning method is actually a function $h : \mathbb{X} \to \mathbb{Y}$ (called hypothesis in statistical learning theory) that takes an input $\mathbf{x} \in \mathbb{X}$ and outputs $h(\mathbf{x}) \in \mathbb{Y}$. This hypothesis h is selected by a learning algorithm among different candidates in a *hypothesis set* $\mathbb{H}$ potentially infinite. As an illustration, we recall four types of hypotheses used in practice.

1.1.2.1 Examples of Hypotheses

The four types of hypotheses we recall are the decision stump (Iba and Langley, 1992), the decision tree (Breiman *et al.*, 1984), the linear classifier, and the neural network. More precisely, these hypotheses are called *classifier* since they classify an input $\mathbf{x} \in \mathbb{X}$, *i.e.*, assign a label to an input.

Decision Stump The decision stump (Iba and Langley, 1992) is a classifier composed of one decision rule of the form "Is $x_i \leq \tau$?", where x_i is the i-th component of the vector $\mathbf{x}$ and $\tau \in \mathbb{R}$ is a threshold. More precisely, this *decision rule* (implemented as an if-condition) assigns a label to the input $\mathbf{x}$. For example, we present in Algorithm 1.1 a decision stump that classifies the input $\mathbf{x}$ as $+1$ when $x_1 \leq 0.54$ and -1 otherwise. Its graphical representation and the associated decision are given in Figure 1.3.

Algorithm 1.1 Example of Decision Stump

Given: An input $\mathbf{x} \in \mathbb{R}^d$
if $x_1 \leq 0.54$ **then**
  return $+1$
else
  return -1

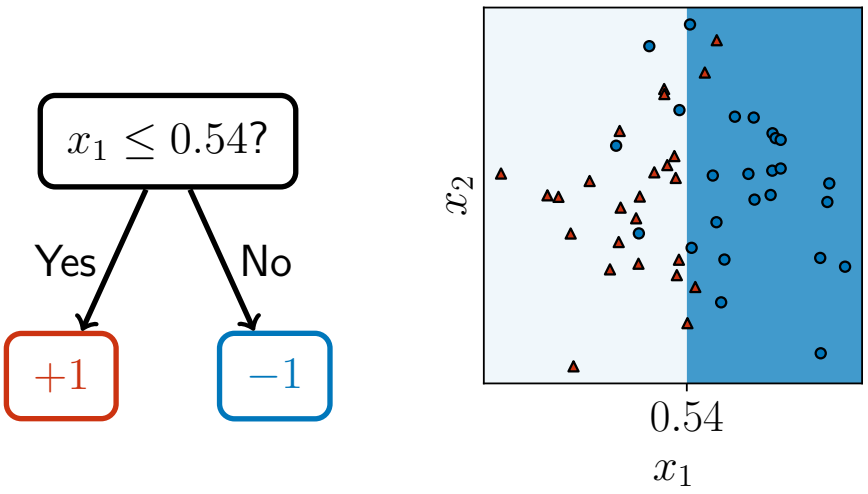


Figure 1.3. *Example of decision stump with the decision rule "Is $x_1 \leq 0.54$?". On the left, the decision of Algorithm 1.1 is shown with its binary tree representation. The right plot shows the decision boundary (in white and blue) of the given decision stump. Moreover, the red triangles resp. the blue dots are the examples in the learning sample $\mathbb{S}$ with the label $+1$ resp. -1.*

To classify the input $\mathbf{x} \in \mathbb{X}$, the decision stump returns a label (by executing its associated algorithm). This algorithm is composed of only one if-condition that can be interpreted as a binary tree. This simple classifier can be complexified by adding more if-conditions to fit the data better. Indeed, a decision tree can be constructed

by considering nested if-conditions. Hence, the decision stump[1] is a particular case of a decision tree (Breiman *et al.*, 1984).

Decision Trees The *decision tree* (Breiman *et al.*, 1984) is a classifier that can be seen as a succession of *decision rules* to classify the label of the input $\mathbf{x} \in \mathbb{R}^d$. Algorithm 1.2 is an example of algorithm associated to the decision tree of Figure 1.4.

Algorithm 1.2 Example of Decision Tree

Given: An input $\mathbf{x} \in \mathbb{R}^d$
if $x_1 \leq 0.59$ **then**
  if $x_2 \leq 0.63$ **then**
    return $+1$
  else
    return -1
else
  return -1

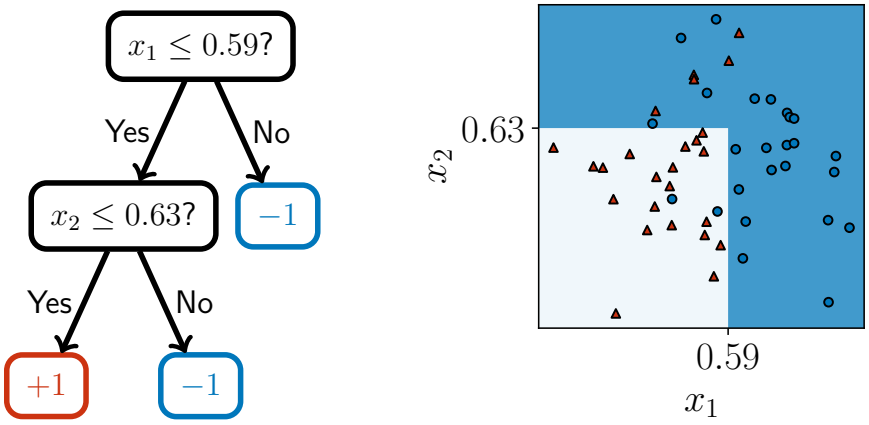


Figure 1.4. *Example of decision tree composed of two decision rules "Is $x_1 \leq 0.59$?" and "Is $x_2 \leq 0.63$?". The binary tree on the left is the graphical representation of the decision stump. The right plot is the stump's decision boundary on the 2-dimensional data in $\mathbb{S}$ (with the red triangles and the blue dots).*

Numerous algorithms have been developed to infer decision trees (see *e.g.*, Breiman *et al.*, 1984; Quinlan, 1986, 1993). One of the most popular is the CART algorithm (Breiman *et al.*, 1984): it is, *e.g.*, implemented in the well-known library

[1] Iba and Langley (1992) introduced the word "stump" for a one-level decision tree.

`scikit-learn` (Pedregosa *et al.*, 2011). This algorithm is summarized as follows. The algorithm tests the decision rules "Is $x_i \leq \tau$?" with different components x_i and thresholds τ. Given a new decision rule, the examples $(\mathbf{x}, y) \in \mathbb{S}$ are split into two groups: the ones that respect the rules and those that do not. Then, the algorithm selects the best decision rule according to a criterion and the two new groups. Finally, this step is repeated recursively to obtain the entire tree. In Part II, we learn a *majority vote* classifier which is a convex combination of the outputs of decision trees or decision stumps.

Linear classifier One of the most simple type of classifiers is the linear classifier. Given a function $\sigma : \mathbb{R} \to \mathbb{R}$, this type of classifier is defined as

$$h(\mathbf{x}) = \sigma\left(\sum_{i=1}^{d} w_i x_i + b\right),$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector and $b \in \mathbb{R}$ is the bias that both need to be learned. The function $\sigma()$ is called *activation* function notably in the context of neural networks (that we precise in the following). For instance, the Perceptron (McCulloch and Pitts, 1943; Rosenblatt, 1958) algorithm returns a linear classifier with the activation function $\sigma()$ called threshold function defined as $\sigma(x) = 1$ if $x \geq 0$ and $\sigma(x) = 0$ otherwise. Moreover, the well-known Support Vector Machine (SVM) algorithm introduced by Boser *et al.* (1992) and Cortes and Vapnik (1995) learns a linear classifier with the sign as activation function, *i.e.*, $\sigma(x) = \text{sign}(x) = +1$ if $a \geq 1$ and -1 otherwise. We provide two 2-dimensional cases when the activation function is the sign function in Figure 1.6.

Neural networks. The Perceptron was abandoned when Minsky and Papert (1972) show that this model cannot learn simple (boolean) functions like the exclusive or (see Figure 1.5). Hopefully, these restrictions are avoided when generalizing the Perceptron, and they regain interest in the 80s. Nevertheless, in the 2010s, these models have become popular when the neural network AlexNet (Krizhevsky *et al.*, 2012) won the computer vision challenge "ImageNet Large Scale Visual Recognition Challenge". The increase in popularity of such models has been helped by the existence of many programming frameworks such as `Torch7` (Collobert *et al.*, 2011), `Tensorflow` (Abadi *et al.*, 2015), `Theano` (Al-Rfou *et al.*, 2016), `JAX` (Bradbury *et al.*, 2018; Frostig *et al.*, 2018), or `PyTorch` (Paszke *et al.*, 2019). To learn more details about neural networks, we refer the reader to Goodfellow *et al.* (2016) for an extensive introduction. Actually, a neural network can be seen as a succession of linear classifiers: it is a composition of L linear classifiers $\mathbf{h}^{(1)}(), \ldots, \mathbf{h}^{(L)}()$ called layer or module, *i.e.*,

$$h = \mathbf{h}^{(L)} \circ \cdots \circ \mathbf{h}^{(2)} \circ \mathbf{h}^{(1)},$$

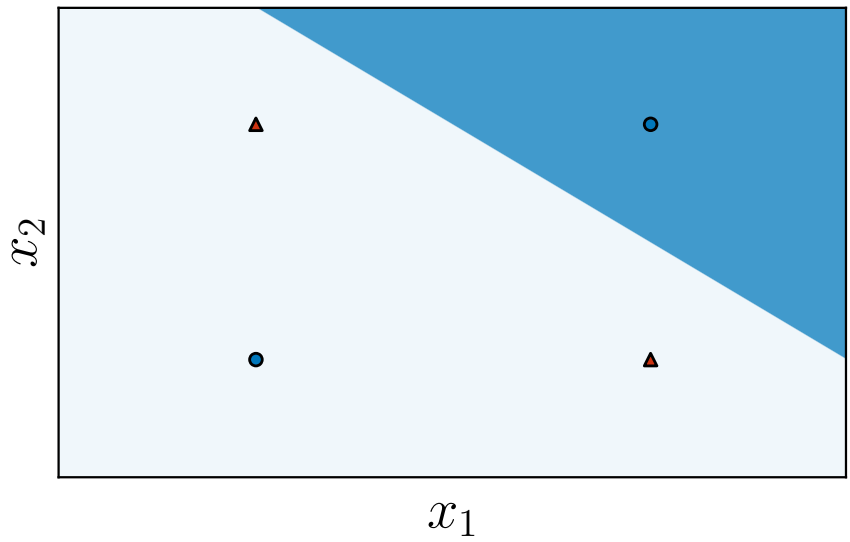


Figure 1.5. *Plot of the decision boundary of a linear classifier learned on the learning sample* $\mathbb{S}=\{([0,0]^\top,-1),([0,1]^\top,+1),([1,1]^\top,-1),([1,0]^\top,+1)\}$ *that represents the exclusive or function. As proved by* MINSKY *and* PAPERT *(1972), this model cannot predict correctly all the examples representing such a function.*

where $f \circ g$ is the composition of the function $f()$ with $g()$. More precisely, given an activation function $\boldsymbol{\sigma}^{(i)} : \mathbb{R}^{d^{(i)}} \to \mathbb{R}^{d^{(i)}}$, the i-th layer $\mathbf{h}^{(i)} : \mathbb{R}^{d^{(i-1)}} \to \mathbb{R}^{d^{(i)}}$ of the network is defined by

$$\mathbf{h}^{(i)}(\mathbf{x}) = \boldsymbol{\sigma}^{(i)}\left(\mathbf{W}\mathbf{x} + \mathbf{b}\right),$$

where the matrix $\mathbf{W} \in \mathbb{R}^{d^{(i)} \times d^{(i-1)}}$ and the vector $\mathbf{b} \in \mathbb{R}^{d^{(i)}}$ are respectively the weights and the bias parameterizing the classifier that need to be learned. Note that, to respect the definition of the hypothesis $h : \mathbb{R}^d \to \mathbb{R}$, we have $d^{(0)} = d$ and $d^{(L)} = 1$.
An example of neural network is given in Figure 1.7. We see in this example that the succession of linear classifiers offers better expressiveness, *i.e.*, its decision boundary is not restricted to lines. Interestingly, these models produce new features of the original data in each layer. Indeed, in Figure 1.8, the i-th layer transforms all inputs $\mathbf{x} \in \mathbb{X}$ into $\mathbf{h}^{(i)}(\ldots\mathbf{h}^{(1)}(\mathbf{x}))$ and these new inputs are classified with $\mathbf{h}^{(L)}(\ldots\mathbf{h}^{(i)}(\mathbf{x}))$. Hence, all these transformations can be interpreted as a new representation of these inputs. In the following, we denote by $\mathbf{w} \in \mathbb{R}^D$ the vector of the network's weights and biases concatenated all together; we have thus D weights/bias in the networks. Besides, when D is large compared to the number of data m, we say in this case that the model is over-parametrized, and when L is large, we consider that the network is "deep"[2]. This

[2]The term deep learning comes from the fact that the number of layers L is large.

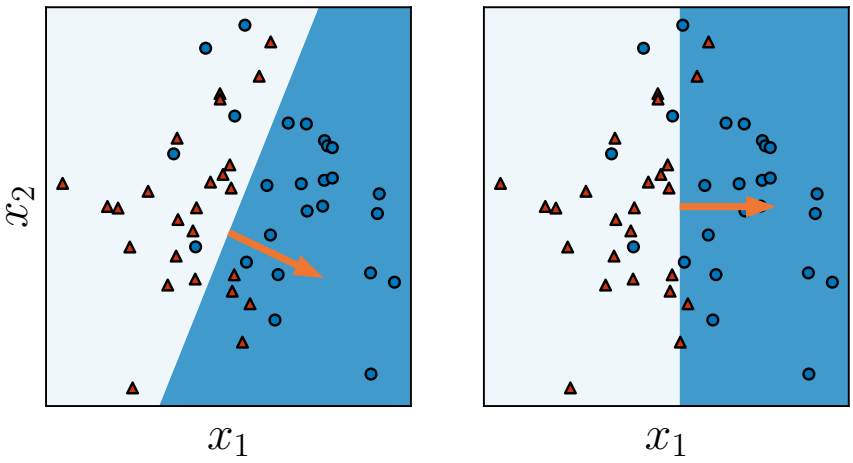


Figure 1.6. *The decision of two linear classifiers when the activation function is the* sign. *On the left plot, the weights are* $\mathbf{w} = [3.2, -1.4]^\top$ *and the bias is* $b = -1$. *On the right plot, the weights are* $\mathbf{w} = [1, 0]^\top$ *and the bias is* $b = -0.54$; *this linear classifier is equivalent to the decision stump in Figure 1.3.*

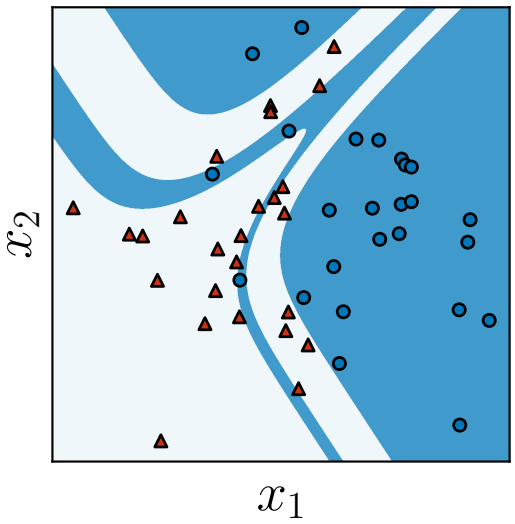


Figure 1.7. *The decision boundary of a neural network classifier of a with 5 layers and the dimensions* $d^{(1)}{=}d^{(2)}{=}d^{(3)}{=}d^{(4)}{=}2$ *and* $d^{(5)}{=}1$. *The activation functions* $\boldsymbol{\sigma}^{(1)}(), \boldsymbol{\sigma}^{(2)}(), \boldsymbol{\sigma}^{(3)}(), \boldsymbol{\sigma}^{(4)}()$ *are* $\tanh$ *applied element-wise and* $\boldsymbol{\sigma}^{(5)}()$ *is the threshold function.*

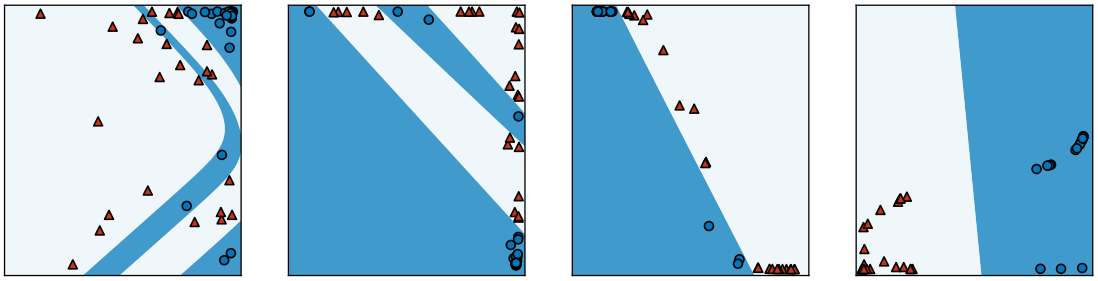


Figure 1.8. *Each plot corresponds to the representation of the i-layer (where $i = 1$ on the left plot and $i = 4$ on the right plot). On each plot, the original examples $(\mathbf{x}, y) \in \mathcal{S}$ (with the red triangles and the blue dots) are transformed and plotted by the i-th layer $(\mathbf{h}^{(1)} \circ \cdots \circ \mathbf{h}^{(i)})(\mathbf{x})$. Moreover, the decision boundary of the classifier $\mathbf{h}^{(i)} \circ \cdots \circ \mathbf{h}^{(L)}$ is given.*

model is studied in Part III when the neural network is over-parametrized and "deep".

1.1.2.2 Loss and Risk

We need to assess to which extent the learned hypothesis $h \in \mathbb{H}$ predicts correctly an example. This kind of measure is often called a *loss function*, defined as follows.

> **Definition 1.1.2** (Loss function)**.** A *loss* function is a function $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to [0, 1]$ that, given a hypothesis $h \in \mathbb{H}$, evaluates the quality of the prediction $h(\mathbf{x})$ compared to the true label $y \in \mathbb{Y}$. The lower the loss, the better the quality of the hypothesis.

For a classification task, the most natural loss function is the 01-loss defined by

$$\forall h \in \mathbb{H}, \quad \forall (\mathbf{x}, y) \in \mathbb{X} \times \mathbb{Y}, \quad \ell^{01}(h, (\mathbf{x}, y)) = \mathrm{I}[h(\mathbf{x}) \neq y] \triangleq \begin{cases} 1 & \text{if } h(\mathbf{x}) \neq y \\ 0 & \text{otherwise} \end{cases}.$$

The 01-loss returns 1 when the hypothesis $h \in \mathbb{H}$ misclassifies the example $(\mathbf{x}, y) \in \mathbb{X} \times \mathbb{Y}$ and 0 otherwise. However, since many learning methods are based on the gradient $\frac{\partial \ell}{\partial h}(h, (\mathbf{x}, y))$ to choose a hypothesis h that lowers the loss function, the 01-loss $\ell^{01}()$ is not practical. Its gradient $\frac{\partial \ell^{01}}{\partial h}(h, (\mathbf{x}, y)) = 0$ for all examples $(\mathbf{x}, y) \in \mathbb{X} \times \mathbb{Y}$, which does not indicate a descent direction (that lowers the loss). To address this issue, in practice, one uses relaxation of the 01-loss. In particular, some relaxed losses rely on a notion of *margin*: in binary classification the margin is defined as $m_h(\mathbf{x}, y) = yh(\mathbf{x})$ for all $h : \mathbb{X} \to [-1, +1]$. When the margin is positive, the hypothesis makes a correct

prediction. In other words, the margin is defined such that we have $\ell^{01}(h,(\mathbf{x},y)) = \mathrm{I}[h(\mathbf{x}) \neq y] = \mathrm{I}[m_h(\mathbf{x},y) \leq 0]$. Thanks to the margin and MARKOV's inequality (Theorems A.2.1 and A.3.1), one can prove two upper bounds of the 01-loss. These two upper bounds are defined by

$$\ell^{1\text{st}}(h,(\mathbf{x},y)) = 1 - m_h(\mathbf{x},y) \qquad \text{(LANGFORD and SHAWE-TAYLOR, 2002),}$$

$$\text{and} \quad \ell^{2\text{nd}}(h,(\mathbf{x},y)) = \Big[1 - m_h(\mathbf{x},y)\Big]^2 \qquad \text{(MASEGOSA } \textit{et al.}\text{, 2020).}$$

We plot in Figure 1.9 the 01-loss and the two upper bounds that we use in Chapters 3 and 4 to derive learning algorithms. We introduce these upper bounds with more details in Chapter 2.

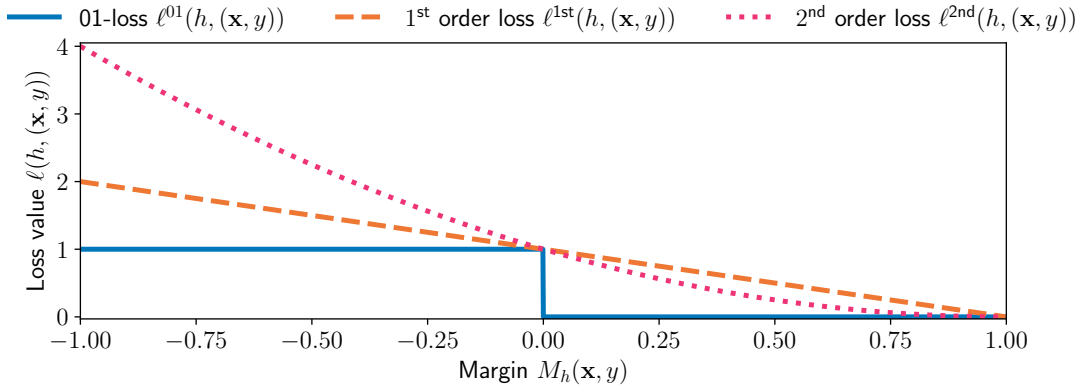


Figure 1.9. *Plot of the losses ($\ell^{01}()$, $\ell^{1st}()$, and $\ell^{2nd}()$) where the x-axis represents the value of the margin $m_h(\mathbf{x},y)$ and the y-axis represents the value of the losses $\ell(h,(\mathbf{x},y))$ for a given example $(\mathbf{x},y)$ and hypothesis h.*

While the loss $\ell(h,(\mathbf{x},y))$ is computed for an example $(\mathbf{x},y)$ and a hypothesis h, we are usually interested in the loss computed over all the examples $(\mathbf{x},y) \in \mathbb{S}$. The value of the loss averaged over the learning sample $\mathbb{S}$ is called *empirical risk* and represents to what extent the hypothesis h predicts correctly all the examples $(\mathbf{x},y) \in \mathbb{S}$. The empirical risk is defined as follows.

Definition 1.1.3 (Empirical Risk)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any learning sample $\mathbb{S} = \{(\mathbf{x}_i,y_i)\}_{i=1}^m \sim \mathcal{D}^m$, we define the empirical risk as

$$\mathrm{R}^{\ell}_{\mathcal{S}}(h) \triangleq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \ell(h,(\mathbf{x},y)) = \frac{1}{m}\sum_{i=1}^{m} \ell(h,(\mathbf{x}_i,y_i)),$$

where its associated uniform distribution $\mathcal{S}$ on $\mathbb{S}$ can be defined as $\mathcal{S}((\mathbf{x}_i, y_i)) \triangleq \frac{1}{m}$ for all $i \in \{1, \dots, m\}$. Moreover, we define the empirical risk with the 01-loss by

$$\mathsf{R}_{\mathcal{S}}(h) \triangleq \mathsf{R}_{\mathcal{S}}^{\ell^{01}}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{I}\left[h(\mathbf{x}_i) \neq y_i\right].$$

Since the distribution $\mathcal{D}$ is unknown, the performance of the hypothesis $h \in \mathbb{H}$ can be computed by means of the learning sample $\mathbb{S}$ (through the empirical risk). However, we are interested in the performance of h on the task, *i.e.*, over all the examples. The performance is thus defined through the true risk $\mathsf{R}_{\mathcal{D}}^{\ell}(h)$ defined as follows.

Definition 1.1.4 (True Risk)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any loss function $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0, 1]$, we define the true risk as

$$\mathsf{R}_{\mathcal{D}}^{\ell}(h) \triangleq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \ell(h, (\mathbf{x}, y)).$$

Moreover, we define the true risk with the 01-loss by

$$\mathsf{R}_{\mathcal{D}}(h) \triangleq \mathsf{R}_{\mathcal{D}}^{\ell^{01}}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}\left[h(\mathbf{x}) \neq y\right] = \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[h(\mathbf{x}) \neq y\right].$$

The true risk can be interpreted as the expected loss over all examples $(\mathbf{x}, y) \sim \mathcal{D}$. Actually, when the hypothesis h does not depend on $\mathbb{S}$, the true risk of h is the expected empirical risk. In other words, the empirical risk can be seen as an unbiased estimator of the true risk, *i.e.*, we have

$$\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathsf{R}_{\mathcal{S}}^{\ell}(h) = \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \frac{1}{m} \sum_{i=1}^{m} \ell(h, (\mathbf{x}_i, y_i)) = \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{(\mathbf{x}_i,y_i)\sim\mathcal{D}} \ell(h, (\mathbf{x}_i, y_i)) = \mathsf{R}_{\mathcal{D}}^{\ell}(h).$$

Besides, selecting a hypothesis with a low true risk boils down to finding a hypothesis with low empirical risk on average. However, since we only have access to one learning sample $\mathbb{S}$, the hypothesis selection cannot be performed on the distribution $\mathcal{D}$ but rather on the learning sample $\mathbb{S}$.

1.2 Hypothesis Selection

In order to obtain a hypothesis $h \in \mathbb{H}$ that solves the task, the hypothesis can be selected such that the empirical risk $\mathsf{R}_{\mathcal{S}}^{\ell}(h)$ is low. We recall two approaches from statistical learning: Empirical Risk Minimization (ERM) and Structural Risk Minimization (SRM).

1.2.1 Empirical Risk Minimization

Given the hypothesis set $\mathbb{H}$, a common approach in statistical learning is to minimize the empirical risk $\mathsf{R}_{\mathcal{S}}^{\ell}(h)$ with respect to the hypothesis $h \in \mathbb{H}$. This approach – known as *Empirical Risk Minimization* (ERM) – has been pioneered by VAPNIK and CHERVONENKIS (1968, 1971, 1974) in the 70s (see VAPNIK (1998) for an introduction). Formally, given a learning sample $\mathbb{S} \sim \mathcal{D}^m$, we select the hypothesis such that the empirical risk $\mathsf{R}_{\mathcal{S}}^{\ell}(h)$ is minimum; this approach is summarized in Algorithm 1.3.

Algorithm 1.3 Empirical Risk Minimization

Given: Learning sample $\mathbb{S}$, Hypothesis set $\mathbb{H}$, Loss function $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to [0, 1]$

$$h = \operatorname*{argmin}_{h' \in \mathbb{H}} \mathsf{R}_{\mathcal{S}}^{\ell}(h')$$

return hypothesis h

However, if the empirical risk $\mathsf{R}_{\mathcal{S}}^{\ell}(h)$ of a hypothesis $h \in \mathbb{H}$ is approximately 0, then h potentially overfits the data; see Figure 1.10. Roughly speaking, in the case of overfitting, we interpret that $h \in \mathbb{H}$ has learned by heart the examples in the learning sample. Such a phenomenon can arise when the hypothesis h is complex; the complexity is manifested differently for each type of hypothesis. To measure such complexity, a real-valued function can be defined $\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$. Given a learning sample $\mathbb{S}$ and a complexity function, the hypothesis $h \in \mathbb{H}$ is considered complex when its associated complexity measure $\mu(h, \mathbb{S})$ is large. In Chapter 7, we derive generalization bounds that integrate a user-specified complexity function.

1.2.2 Structural Risk Minimization

To avoid overfitting, we can select a hypothesis h with a small complexity measure $\mu(h, \mathbb{S})$. The *Structural Risk Minimization* approach – introduced by VAPNIK and CHERVONENKIS (1974) – minimizes a trade-off between the empirical risk $\mathsf{R}_{\mathcal{S}}^{\ell}(h)$ and the complexity measure $\mu(h, \mathbb{S})$. The complexity measures $\mu()$ used in this approach is constant over the hypothesis $h \in \mathbb{H}$ and the learning sample $\mathbb{S} \in (\mathbb{X} \times \mathbb{Y})^m$. In other words, if we denote by abuse of notation $\mu(\mathbb{H})$ the complexity measure associated to the hypothesis set $\mathbb{H}$, we have $\forall h \in \mathbb{H}, \forall \mathbb{S} \in (\mathbb{X} \times \mathbb{Y})^m, \mu(h, \mathbb{S}) = \mu(\mathbb{H})$. However, the complexity measure $\mu(\mathbb{H})$ may not be representative of overfitting for a single hypothesis since $\mu(\mathbb{H})$ is constant. Hence, the hypothesis set $\mathbb{H}$ is structured into countable and nested hypothesis subsets $\mathbb{H}_1 \subseteq \mathbb{H}_2 \subseteq \cdots \subseteq \mathbb{H}$. By doing so, the complexity measure may be increasing when the hypothesis set grows: we can have $\mu(\mathbb{H}_i) \leq \mu(\mathbb{H}_{i+1})$ when $\mathbb{H}_i \subseteq \mathbb{H}_{i+1}$. Then, thanks to these nested complexity measure, one can find a hypothesis h belonging to a set $\mathbb{H}_i$ with $i \in \mathbb{N}_*$ that minimizes the trade-off between the empirical risk $\mathsf{R}_{\mathcal{S}}^{\ell}(h)$ and the complexity measure $\mu(\mathbb{H}_i)$ of the set

$\mathbb{H}_i$. The minimization of the trade-off performed by the *Structural Risk Minimization* approach is summarized in Algorithm 1.4.

Algorithm 1.4 Structural Risk Minimization

Given: Learning sample $\mathcal{S}$, Hypothesis set $\mathbb{H}$

$$h = \underset{i\in\mathbb{N}_*,\ h'\in\mathbb{H}_i}{\operatorname{argmin}} \left[\mathsf{R}^{\ell}_{\mathcal{S}}(h') + \mu(\mathbb{H}_i)\right]$$

return hypothesis h

If the complexity measure $\mu(\mathbb{H}')$ of the subset $\mathbb{H}' \subseteq \mathbb{H}$ approximates well the difference $\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)$, SRM can be used to obtain a hypothesis $h \in \mathbb{H}'$ that *generalizes* well. In other words, a hypothesis $h \in \mathbb{H}'$ that *generalizes* has a small true risk $\mathsf{R}^{\ell}_{\mathcal{D}}(h)$ and empirical risk $\mathsf{R}^{\ell}_{\mathcal{S}}(h)$ This situation is represented in Figure 1.10. In other words, when h generalizes, the *generalization gap*[3] $\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)$ is close to 0 (that is $\mathsf{R}^{\ell}_{\mathcal{D}}(h) \approx \mathsf{R}^{\ell}_{\mathcal{S}}(h)$). However, since the difference $\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)$ is not computable because of the true risk $\mathsf{R}^{\ell}_{\mathcal{D}}(h)$, we have to bound it.

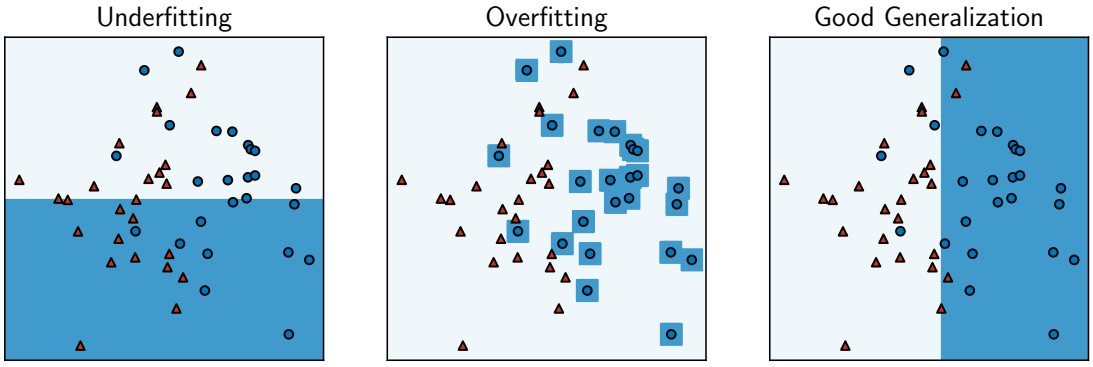


Figure 1.10. *Representation of the three situations (one per scatter plot) that can arise in statistical learning. The left figure corresponds to a hypothesis h that underfits the data: both the empirical risk $R_{\mathcal{S}}(h)$ and the true risk $R_{\mathcal{D}}(h)$ are high. In the middle figure, the hypothesis h overfits the data: the empirical risk $R_{\mathcal{S}}(h) = 0$ but the true risk $R_{\mathcal{D}}(h)$ is high. The left figure represents the perfect case: the difference between the true risk $R_{\mathcal{D}}(h)$ and the empirical risk $R_{\mathcal{S}}(h)$ is small as well as $R_{\mathcal{S}}(h)$.*

[3]We consider sometimes the generalization gap $\left|\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)\right|$ instead, but, we are mostly interested in upper bounding the true risk $\mathsf{R}^{\ell}_{\mathcal{D}}(h)$.

1.3 Generalization Bounds

To upper-bound the generalization gap $\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)$, we can provide a bound on the true risk $\mathsf{R}^{\ell}_{\mathcal{D}}(h)$ with high probability over the random choice of $\mathbb{S} \sim \mathcal{D}^m$. Such a bound is a *Probably Approximately Correct* (PAC) generalization bound (Valiant, 1984) and can be defined as follows.

Definition 1.3.1 (PAC Generalization Bound)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any data-dependent hypothesis set $\mathbb{H} = \{h_{\mathbb{S}}\}_{\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m}$ where $h_{\mathbb{S}}$ is a hypothesis dependent on the learning sample $\mathbb{S} \in (\mathbb{X}\times\mathbb{Y})^m$, a PAC generalization bound is defined by

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathsf{R}^{\ell}_{\mathcal{D}}(h_{\mathbb{S}}) \leq \Phi\right] \geq 1 - \delta.$$

With high Probability (with probability at least $1-\delta$), the true risk of the hypothesis $h_{\mathbb{S}}$ is *Approximately Correct*, *i.e.*, upper-bounded by Φ. Based on this definition, Valiant (1984) has brought a computational framework: the PAC-Learnability.[4] A data-dependent hypothesis set $\mathbb{H} = \{h_{\mathbb{S}}\}_{\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m}$ is PAC-Learnable if this high probability bound holds when the number of examples m is polynomial in $1/\delta$ and $1/\Phi$. In practice, to obtain such an upper bound Φ, we make use of *concentration inequalities* (see Boucheron *et al.* (2013) for an extensive introduction on the subject). These inequalities allow us to bound an expectation (the true risk in our case) with its empirical counterpart (the empirical risk). Depending on the concentration inequalities, one can obtain different frameworks, and so, different upper bounds Φ.

1.3.1 Uniform Convergence Bounds

The first framework introduced in the literature to obtain PAC generalization bounds is referred to as *uniform convergence* bounds (Vapnik and Chervonenkis, 1968, 1971). Given a hypothesis set $\mathbb{H}$ (not necessarily data-dependent), a uniform convergence bound holds for all hypotheses of $\mathbb{H}$. By doing so, the true risk $\mathsf{R}^{\ell}_{\mathcal{D}}(h_{\mathbb{S}})$ of a data-dependent hypothesis $h_{\mathbb{S}} \in \mathbb{H}$ is upper-bounded to obtain a PAC generalization bound. This type of bounds takes the following form.

Definition 1.3.2 (Uniform Convergence Bound)**.** Let $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$ be a loss function and $\phi : [0,1]^2 \to \mathbb{R}$ a generalization gap. A uniform convergence bound is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis

[4] Valiant won the Turing prize in 2010 for the definition of PAC-Learnability.

set $\mathbb{H}$, there exists a function $\Phi_{\mathrm{u}} : (0,1] \to \mathbb{R}$, such that for any $\delta \in (0,1]$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\sup_{h\in\mathbb{H}} \phi(\mathsf{R}^\ell_{\mathcal{D}}(h), \mathsf{R}^\ell_{\mathcal{S}}(h)) \leq \Phi_{\mathrm{u}}\Big(\delta\Big)\right] \geq 1-\delta, \tag{1.1}$$

where usually $\phi(\mathsf{R}^\ell_{\mathcal{D}}(h), \mathsf{R}^\ell_{\mathcal{S}}(h)) = \mathsf{R}^\ell_{\mathcal{D}}(h) - \mathsf{R}^\ell_{\mathcal{S}}(h)$.

A uniform convergence bound consists in bounding the supremum of the generalization gap over the hypotheses $h \in \mathbb{H}$ with the upper bound $\Phi_{\mathrm{u}}\Big(\delta\Big)$ (with probability at least $1-\delta$ over the learning sample $\mathbb{S} \sim \mathcal{D}^m$). In other words, the inequality boils down to bound the generalization gap of all hypotheses $h \in \mathbb{H}$ by an upper bound $\Phi_{\mathrm{u}}\Big(\delta\Big)$ on the largest generalization gap; the hypothesis associated to the largest generalization gap is considered as the "worst" hypothesis in $\mathbb{H}$. As we will see in the following, the upper bound $\Phi_{\mathrm{u}}\Big(\delta\Big)$ depends also on the number of examples m in $\mathbb{S}$; we have simplified $\Phi_{\mathrm{u}}\Big(\delta\Big)$ for readability. Hence, if $\lim_{m\to+\infty} \Phi_{\mathrm{u}}(\delta) = 0$, with probability at least $1-\delta$ over $\mathbb{S} \sim \mathcal{D}^m$, the empirical risk *converges uniformly* on $\mathbb{H}$ to the true risk $\mathsf{R}^\ell_{\mathcal{D}}(h)$. Because of the uniform convergence, as we see in the two examples below, the upper bound $\Phi_{\mathrm{u}}\Big(\delta\Big)$ depends on a *complexity* $\mu(\mathbb{H})$ that measures, in some sense, the performance of the worst hypothesis in $\mathbb{H}$.

1.3.1.1 Uniform Convergence Bounds for Finite Hypothesis Sets

The simplest complexity measure appearing in generalization bounds arises when considering a finite set $\mathbb{H}$ of hypotheses. We recall an instantiation of uniform convergence bounds (Definition 1.3.2) in the following theorem (see, *e.g.*, Mohri *et al.* (2012) for more details).

Theorem 1.3.1 (Generalization Bound for Finite $\mathbb{H}$)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any finite hypothesis set $\mathbb{H}$ ($\mathrm{card}(\mathbb{H}) < +\infty$), for any loss function $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\sup_{h\in\mathbb{H}} \Big[\mathsf{R}^\ell_{\mathcal{D}}(h) - \mathsf{R}^\ell_{\mathcal{S}}(h)\Big] \leq \underbrace{\sqrt{\frac{\ln \mathrm{card}(\mathbb{H}) + \ln\frac{1}{\delta}}{2m}}}_{\Phi_{\mathrm{u}}(\delta)}\right] \geq 1-\delta,$$

where $\mathrm{card}(\mathbb{H})$ is the cardinal of the set $\mathbb{H}$. Equivalently, with probability at least $1-\delta$ over $\mathbb{S} \sim \mathcal{D}^m$, we have

$$\forall h \in \mathbb{H}, \quad \mathsf{R}^\ell_{\mathcal{D}}(h) \leq \mathsf{R}^\ell_{\mathcal{S}}(h) + \sqrt{\frac{\ln \mathrm{card}(\mathbb{H}) + \ln\frac{1}{\delta}}{2m}}.$$

In this case, the complexity measure for a finite hypothesis set is defined by $\mu(\mathbb{H}) = \mathrm{card}(\mathbb{H})$. According to this complexity measure, the more hypotheses in $\mathbb{H}$, the more complex the hypothesis set $\mathbb{H}$ is. The computation of this complexity can be fast, making this generalization bound easy to evaluate and to obtain an upper bound on the generalization gap $\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)$. However, the bound does not converges when the cardinality of the set is large, *e.g.*, when $\mathrm{card}(\mathbb{H}) \geq e^m$ for $m \in \mathbb{N}$, the upper bound of Theorem 1.3.1 is $\sqrt{\frac{1}{2}} \leq \sqrt{\frac{1}{2m}(\ln \mathrm{card}(\mathbb{H}) + \ln \frac{1}{\delta})}$. Hence, there is a need to develop generalization bounds for hypothesis sets $\mathbb{H}$ with large $\mathrm{card}(\mathbb{H})$ or infinite hypothesis sets. Indeed, these types of set are usually considered in practice, *e.g.*, the set of linear classifiers.

1.3.1.2 VC-Dimension-based Generalization Bounds

When the loss is the 01-loss, a bound for infinite hypothesis sets is proposed based on a complexity measure $\mu()$ called the Vapnik-Chervonenkis (VC)-Dimension (Vapnik and Chervonenkis, 1968, 1971).

Definition 1.3.3 (Vapnik-Chervonenkis (VC) Dimension)**.** Given a hypothesis set $\mathbb{H}$ with hypotheses $h : \mathbb{X} \to \{-1, +1\}$ (for binary classification), the VC-dimension of the set $\mathbb{H}$ is defined as

$$\mathtt{vc}(\mathbb{H}) \triangleq \max \{m \ : \ \forall \mathbb{S} \in (\mathbb{X}\times\mathbb{Y})^m, \exists h \in \mathbb{H} \ \textit{s.t.}\ \mathsf{R}_{\mathcal{S}}(h) = 0\}\,,$$

where $\mathsf{R}_{\mathcal{S}}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}[h(\mathbf{x}) \neq y]$.

Put into words, the VC-Dimension $\mu(\mathbb{H}) = \mathtt{vc}(\mathbb{H})$ of a hypothesis set $\mathbb{H}$ is the maximum number of examples that a hypothesis h from $\mathbb{H}$ can perfectly fit in binary classification. Note that if the hypothesis set $\mathbb{H}$ is infinite, its associated VC-dimension $\mathtt{vc}(\mathbb{H})$ can be finite. For example, the VC-Dimension of the d-dimensional linear classifiers is $d+1$ (see Mohri *et al.* (2012, Example 3.12) for a proof) even if the set of linear classifiers $\mathbb{H}$ is infinite; we illustrate in Figure 1.11 the 2-dimensional case. Thanks to this complexity measure, we can prove the following generalization bound (Mohri *et al.*, 2012, Theorem 3.17, Corollaries 3.18 and 3.19).

Theorem 1.3.2 (VC-Dimension-based Generalization Bounds)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any set $\mathbb{H}$ with hypotheses $h : \mathbb{X} \to \{-1, +1\}$ and VC-Dimension $\mathtt{vc}(\mathbb{H})$, for any $\delta \in (0, 1]$, we have

$$\underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[\sup_{h\in\mathbb{H}}\left[\mathsf{R}_{\mathcal{D}}(h) - \mathsf{R}_{\mathcal{S}}(h)\right] \leq \underbrace{\sqrt{\frac{2\mathtt{vc}(\mathbb{H})\left(1 + \ln\frac{m}{\mathtt{vc}(\mathbb{H})}\right)}{m}} + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}}_{\Phi_{\mathrm{u}}(\delta)}\right] \geq 1-\delta.$$

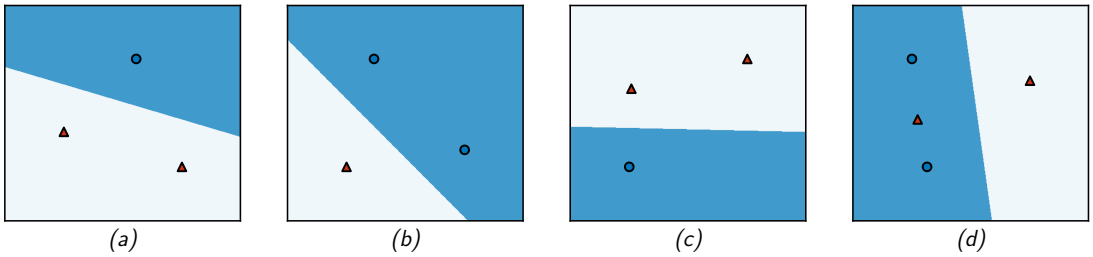


Figure 1.11. *Illustration of the VC-dimension for linear classifiers in* 2*-dimension. When* $m{=}3$*, a linear classifier can always perfectly fit the data, i.e.,* $R_{\mathcal{S}}(h) = 0$ *(as shown in the plot (a), (b), and (c)). However, when* $m{=}4$*, the empirical risk* $R_{\mathcal{S}}(h) \geq 0$ *for all linear classifiers (as illustrated in the plot (d)). Hence, in this case, the VC-Dimension is* $\mathtt{vc}(\mathbb{H}) = 3$*.*

Equivalently, with probability at least $1{-}\delta$ over $\mathbb{S} \sim \mathcal{D}^m$, we have

$$\forall h \in \mathbb{H}, \quad \mathsf{R}_{\mathcal{D}}(h) \leq \mathsf{R}_{\mathcal{S}}(h) + \sqrt{\frac{2\mathtt{vc}(\mathbb{H})\left(1 + \ln \frac{m}{\mathtt{vc}(\mathbb{H})}\right)}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

Hence, if we know the VC-dimension of the hypothesis set $\mathbb{H}$ of interest, the bound becomes easily computable to obtain an upper bound of the generalization gap. Furthermore, it is possible to prove that the ERM algorithm is consistent (Vapnik, 1998) thanks to this bound (see Proposition 4.1 Mohri *et al.*, 2012). When ERM is consistent, *(a)* the empirical risk $\mathsf{R}_{\mathcal{S}}(h)$ of the classifier h obtained by ERM converges in probability to $\inf_{h' \in \mathbb{H}} \mathsf{R}_{\mathcal{D}}(h')$, and *(b)* its true risk $\mathsf{R}_{\mathcal{S}}(h)$ converges in probability to $\inf_{h' \in \mathbb{H}} \mathsf{R}_{\mathcal{D}}(h')$ as well. Nevertheless, one drawback of the VC-Dimension is that it is only for binary classifiers and the 01-loss; there are several extensions for the multi-class setting, we introduce one of them: the Rademacher complexity (Bartlett and Mendelson, 2002).

1.3.1.3 Rademacher-complexity-based Generalization Bounds

The Rademacher complexity (Bartlett and Mendelson, 2002) of a hypothesis set $\mathbb{H}$ can be defined as follows.

Definition 1.3.4 (Rademacher Complexity)**.** Given a hypothesis set $\mathbb{H}$ and a loss function $\ell : \mathbb{H} \times (\mathbb{X}{\times}\mathbb{Y}) \to [0, 1]$, the Rademacher complexity $\mathtt{rad}(\mathbb{H})$ of the set $\mathbb{H}$

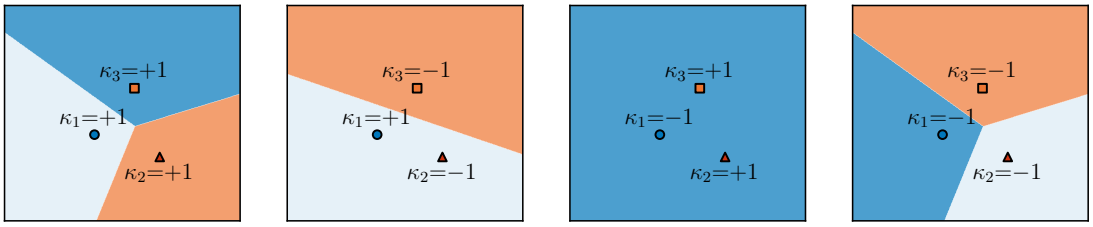


Figure 1.12. *Illustration of the Rademacher complexity* $\texttt{rad}(\mathbb{H})$ *in multi-class classification with* $\ell(h,(\mathbf{x},y)) = \mathrm{I}[h(\mathbf{x}) \neq y]$ *and* $m = 3$. *Given a learning sample* $\mathbb{S} \in (\mathbb{X} \times \mathbb{Y})^2$, *we show for each* $(\kappa_1, \kappa_2) \in \{-1,+1\}^2$ *the examples and a classifier* h *s.t.* $R^{\ell}_{\mathcal{S}}(h) = \sup_{h \in \mathbb{H}} \frac{1}{m} \sum_{i=1}^{m} \kappa_i \ell(h,(\mathbf{x}_i, y_i))$.

is defined as

$$\texttt{rad}(\mathbb{H}) = \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{\{\kappa_1,\ldots,\kappa_m\}\sim\mathcal{K}^m} \sup_{h\in\mathbb{H}} \frac{1}{m} \sum_{i=1}^{m} \kappa_i \ell(h,(\mathbf{x}_i,y_i)),$$

where $\mathcal{K}$ is the Rademacher distribution, *i.e.*, $\mathcal{K}(+1) = \mathcal{K}(-1) = \frac{1}{2}$.

Given a learning sample $\mathbb{S} \sim \mathcal{D}^m$ and the Rademacher variables $\{\kappa_1, \ldots, \kappa_m\} \sim \mathcal{K}^m$, the supremum is attained when the loss $\ell(h,(\mathbf{x}_i,y_i))$ is maximized (*resp.* minimized) when $\kappa_i{=}{+}1$ (*resp.* $\kappa_i{=}{-}1$). Then, the Rademacher complexity is the expected supremum over the learning sample and the Rademacher variables. For example, in binary classification, the Rademacher complexity measures the capacity of the hypotheses in $\mathbb{H}$ to learn examples with random labels. As illustrated in Figure 1.12 when $m = 3$, (multi-class) linear classifiers are able to fit any data points with random labels. In this case, the Rademacher complexity $\texttt{rad}(\mathbb{H}) = 1$, but hopefully when $m \to +\infty$ the Rademacher complexity of the linear classifiers tends to 0 at a rate of $O\left(\sqrt{\frac{1}{m}}\right)$ (see, *e.g.*, Kakade *et al.* (2008), for the binary classification case).
From McDiarmid's concentration inequality (McDiarmid, 1989), Bartlett and Mendelson (2002) derived the following generalization bound that depends on the Rademacher complexity.

Theorem 1.3.3 (Rademacher-complexity-based Generalization Bound)**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any loss function $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m} \left[\sup_{h\in\mathbb{H}} \left[\mathrm{R}^{\ell}_{\mathcal{D}}(h) - \mathrm{R}^{\ell}_{\mathbb{S}}(h) \right] \leq \underbrace{2\texttt{rad}(\mathbb{H}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}}_{\Phi_{\mathrm{u}}(\delta)} \right] \geq 1-\delta.$$

Equivalently, with probability at least $1-\delta$ over $\mathbb{S} \sim \mathcal{D}^m$, we have

$$\forall h \in \mathbb{H}, \quad \mathsf{R}^{\ell}_{\mathcal{D}}(h) \leq \mathsf{R}^{\ell}_{\mathcal{S}}(h) + 2\mathtt{rad}(\mathbb{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

Put into words, the generalization gap is upper-bounded by the Rademacher complexity of the set $\mathbb{H}$ along with a tight term when the number of examples m is large. As for the VC-Dimension-based bound, if we have an analytic expression of the complexity (or an upper bound), we can compute the generalization bound for all hypotheses h in $\mathbb{H}$. Since the bound holds for all $h \in \mathbb{H}$ (due to $\sup_{h\in\mathbb{H}}$), it can be seen as a *worst-case* analysis. Indeed, the same upper bound (*i.e.*, $\Phi_{\mathtt{u}}(\delta)$) holds for all $h \in \mathbb{H}$, including the best, but also the worst (with the largest generalization gap $\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)$). This *worst-case* analysis makes the derivation of non-vacuous bounds hard (*i.e.*, with $\Phi_{\mathtt{u}}(\delta) < 1$). Hence, other generalization bounds have been developed to avoid this worst-case analysis as we see in the next section.

1.3.2 Algorithmic-dependent Generalization Bounds

Let consider an algorithm that takes a learning sample $\mathbb{S} \in (\mathbb{X} \times \mathbb{Y})^m$ as input and outputs the hypothesis $h_{\mathbb{S}}$ belonging to the hypothesis set $\mathbb{H} = \{h_{\mathbb{S}}\}_{\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m}$. Thanks to this algorithm, one can derive generalization bounds for the hypothesis $h_{\mathbb{S}}$ given $\mathbb{S} \sim \mathcal{D}^m$ that has the following form.

Definition 1.3.5 (Algorithmic-dependent Generalization Bound)**.** Let $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$ be a loss function and $\phi : [0,1]^2 \to \mathbb{R}$ a generalization gap. An algorithmic-dependent generalization bound is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, there exists a function $\Phi_{\mathtt{a}} : (0,1] \to \mathbb{R}$, such that for any $\delta \in (0,1]$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\phi(\mathsf{R}^{\ell}_{\mathcal{D}}(h_{\mathbb{S}}), \mathsf{R}^{\ell}_{\mathcal{S}}(h_{\mathbb{S}})) \leq \Phi_{\mathtt{a}}(\delta)\right] \geq 1-\delta, \tag{1.2}$$

where usually $\phi(\mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell}_{\mathcal{S}}(h)) = \mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)$ and $h_{\mathbb{S}}$ is the hypothesis learned from an algorithm with $\mathbb{S} \sim \mathcal{D}^m$.

According to Definition 1.3.5, one can derive an upper bound $\Phi_{\mathtt{a}}(\delta)$ of the generalization gap $\mathsf{R}^{\ell}_{\mathcal{D}}(h_{\mathbb{S}}) - \mathsf{R}^{\ell}_{\mathcal{S}}(h_{\mathbb{S}})$ for the algorithmic-dependent hypothesis $h_{\mathbb{S}}$. Actually, to derive such an upper bound $\Phi_{\mathtt{a}}(\delta)$, a property of the considered learning algorithm must be considered; we recall in the following the algorithmic stability (Bousquet and Elisseeff, 2002) and the algorithmic robustness of Xu and Mannor (2010, 2012).

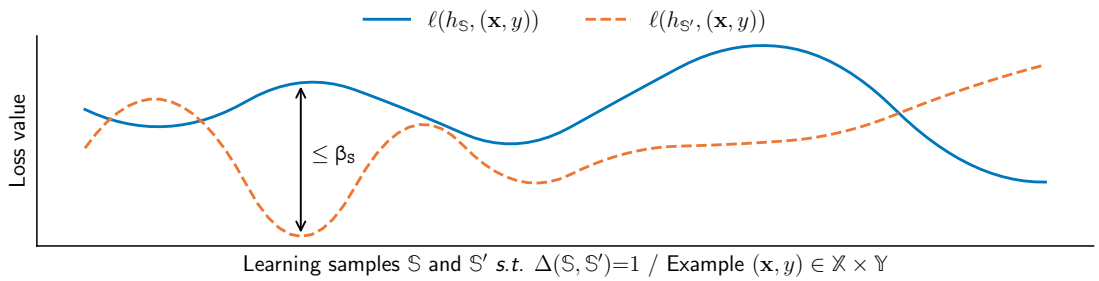


Figure 1.13. *Schematic representation of the notion of stability* β_s*. The two curves represents the losses* $\ell(h_{\mathbb{S}}, (\mathbf{x}, y))$ *and* $\ell(h_{\mathbb{S}'}, (\mathbf{x}, y))$ *(the y-axis) for different combinations of examples* $(\mathbf{x}, y) \in \mathbb{X} \times \mathbb{Y}$ *and learning samples* $\mathbb{S}$ *and* $\mathbb{S}'$ *that differs from one example (the x-axis).*

1.3.2.1 Stability-based Generalization Bounds

The first algorithmic property we recall is the uniform stability (Bousquet and Elisseeff, 2002) and is defined as follows.

Definition 1.3.6 (Uniform Stability)**.** Given the hypothesis set $\mathbb{H} = \{h_{\mathbb{S}}\}_{\mathbb{S} \in (\mathbb{X} \times \mathbb{Y})^m}$ and a loss function $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to [0, 1]$, an algorithm is β_s-uniformly stable if

$$\sup_{\substack{\mathbb{S}, \mathbb{S}' \in (\mathbb{X} \times \mathbb{Y})^m \\ \text{s.t. } \Delta(\mathbb{S}, \mathbb{S}')=1}} \sup_{(\mathbf{x}, y) \in \mathbb{X} \times \mathbb{Y}} \left| \ell(h_{\mathbb{S}}, (\mathbf{x}, y)) - \ell(h_{\mathbb{S}'}, (\mathbf{x}, y)) \right| \leq \beta_s,$$

where $\Delta(\mathbb{S}, \mathbb{S}') = \sum_{i=1}^{m} \mathrm{I}\left[(\mathbf{x}_i, y_i) \neq (\mathbf{x}'_i, y'_i)\right]$ is the Hamming distance between the learning samples $\mathbb{S}$ and $\mathbb{S}'$.

For two learning samples $\mathbb{S}$ and $\mathbb{S}'$ that differ from only one example (*i.e.*, that are very similar), the uniform stability β_s measures how much stable the algorithm is under small changes in the learning sample. To be stable, the difference of losses between the hypotheses $h_{\mathbb{S}}$ and $h_{\mathbb{S}'}$ must be small (and upper bounded by β_s) for all examples $(\mathbf{x}, y) \in \mathbb{X} \times \mathbb{Y}$. Typically, β_s can be seen as a complexity measure that depends on the number of examples m. When $\beta_s = O(\frac{1}{\sqrt{m}})$ or $\beta_s = O(\frac{1}{m})$, the algorithm becomes more stable as the number of examples in $\mathbb{S}$ increases. The notion of algorithmic stability is illustrated and summarized in Figure 1.13. From this notion of uniform stability and McDiarmid's inequality (McDiarmid, 1989), the following generalization bounds can be derived (Bousquet and Elisseeff, 2002).

Theorem 1.3.4 (Stability-based Bounds)**.** Given the hypothesis set $\mathbb{H} = \{h_{\mathbb{S}}\}_{\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m}$, for any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any loss function $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y}) \to [0,1]$, for any $\beta_{\mathbb{S}}$-uniformly stable algorithm, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathsf{R}^{\ell}_{\mathcal{D}}(h_{\mathbb{S}}) - \mathsf{R}^{\ell}_{\mathcal{S}}(h_{\mathbb{S}}) \leq 2\beta_{\mathbb{S}} + (4m\beta_{\mathbb{S}}+1)\sqrt{\frac{\ln\frac{1}{\delta}}{2m}}\right] \geq 1-\delta.$$

Equivalently, with probability at least $1-\delta$ over $\mathbb{S}\sim\mathcal{D}^m$, we have

$$\mathsf{R}^{\ell}_{\mathcal{D}}(h_{\mathbb{S}}) \leq \mathsf{R}^{\ell}_{\mathcal{S}}(h_{\mathbb{S}}) + 2\beta_{\mathbb{S}} + (4m\beta_{\mathbb{S}}+1)\sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

Note that, for this particular bound, if the algorithm is $O(1)$-uniformly stable, the above bound does not converge and is vacuous. In contrast, an $O(\frac{1}{\sqrt{m}})$-uniformly stable algorithm gives an $O(\frac{1}{\sqrt{m}})$ convergence rate. Recently, tighter generalization bounds have been improved: the bound converges with a $O(\frac{1}{m})$-uniformly stable algorithm (Feldman and Vondrák, 2018, 2019; Bousquet *et al.*, 2020). As for the uniform-convergence-based bounds, an analytical expression of the stability $\beta_{\mathbb{S}}$ has to be derived to compute the bound.

1.3.2.2 Robustness-based Generalization Bounds

Another learning algorithm property used to derive generalization bounds is the algorithmic robustness. This notion can be defined as follows.

Definition 1.3.7 (Robustness)**.** Given the hypothesis set $\mathbb{H} = \{h_{\mathbb{S}}\}_{\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m}$, a loss function $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y}) \to [0,1]$ and N disjoint sets *s.t.* $(\mathbb{X}\times\mathbb{Y}) = \bigcup_{i=1}^{N}\mathbb{Z}_i$, the algorithm is $(\{\mathbb{Z}_i\}_{i=1}^{N}, \beta_{\mathbb{R}})$-robust[5] if

$$\sup_{\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m}\sup_{i\in\{1,\ldots,N\}}\sup_{(\mathbf{x},y),(\mathbf{x}',y')\in\mathbb{Z}_i}\left|\ell(h_{\mathbb{S}},(\mathbf{x},y)) - \ell(h_{\mathbb{S}},(\mathbf{x}',y'))\right| \leq \beta_{\mathbb{R}}.$$

For each subset $\mathbb{Z}_i \subseteq \mathbb{X}\times\mathbb{Y}$, the difference of losses between two examples $(\mathbf{x},y)\in\mathbb{Z}_i$ and $(\mathbf{x}',y')\in\mathbb{Z}_i$ has to be upper-bounded by $\beta_{\mathbb{R}}$; for pedagogical purposes, we summarize in Figure 1.14 the algorithmic robustness. The following bound can be derived

[5] In the original definition, the robust parameter $\beta_{\mathbb{R}}$ can depend on the learning sample $\mathbb{S}\sim\mathcal{D}^m$. However, in the examples of Xu and Mannor (2012) for the classification setting, $\beta_{\mathbb{R}}$ depends only on the number of examples m.

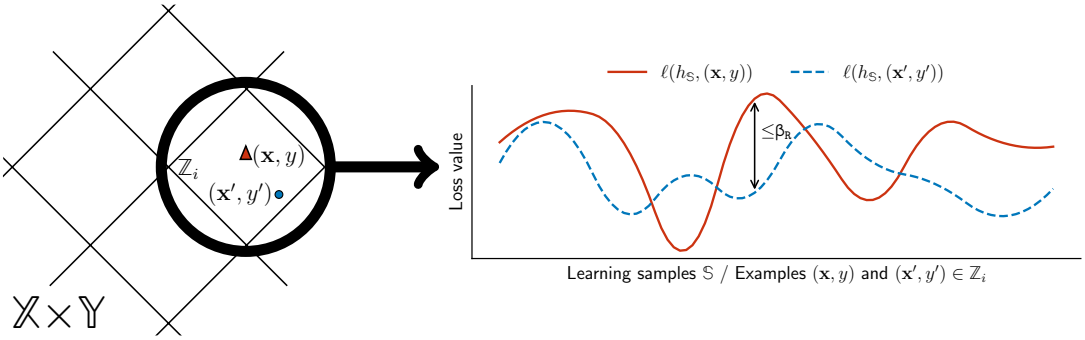


Figure 1.14. *Schematic representation of the notion of algorithmic robustness. On the left plot, an example of partition $\{\mathbb{Z}_i\}_{i=1}^N$ for the set $\mathbb{X} \times \mathbb{Y}$ is shown. Then, for a given subset $\mathbb{Z}_i$, we represent on the right plot the losses $\ell(h_{\mathbb{S}}, (\mathbf{x}, y))$ and $\ell(h_{\mathbb{S}}, (\mathbf{x}', y'))$ for any learning sample $\mathbb{S}$, and examples $(\mathbf{x}, y) \in \mathbb{Z}_i$ and $(\mathbf{x}', y') \in \mathbb{Z}_i$.*

based on this property and thanks to the Breteganolle-Huber-Carol concentration inequality (Vaart and Wellner, 1996, Proposition A.6.6).

Theorem 1.3.5 (Robustness-based Bounds)**.** Given the hypothesis set $\mathbb{H} = \{h_{\mathbb{S}}\}_{\mathbb{S} \in (\mathbb{X} \times \mathbb{Y})^m}$, for any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any loss function $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to [0, 1]$, for any $(\{\mathbb{Z}_i\}_{i=1}^N, \beta_{\mathtt{R}})$-robust algorithm, for any $\delta \in (0, 1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m} \left[\mathrm{R}_{\mathcal{D}}^{\ell}(h_{\mathbb{S}}) \leq \mathrm{R}_{\mathbb{S}}^{\ell}(h_{\mathbb{S}}) + \beta_{\mathtt{R}} + \sqrt{\frac{2N \ln 2 + 2 \ln \frac{1}{\delta}}{2m}} \right] \geq 1 - \delta.$$

Equivalently, with probability at least $1 - \delta$ over $\mathbb{S} \sim \mathcal{D}^m$, we have

$$\mathrm{R}_{\mathcal{D}}^{\ell}(h_{\mathbb{S}}) \leq \mathrm{R}_{\mathbb{S}}^{\ell}(h_{\mathbb{S}}) + \beta_{\mathtt{R}} + \sqrt{\frac{2N \ln 2 + 2 \ln \frac{1}{\delta}}{2m}}.$$

The bound is computable if we have an analytical expression of the algorithmic robustness parameter $\beta_{\mathtt{R}}$. Ideally, the robustness parameter $\beta_{\mathtt{R}}$ must depend on m for the bound to converge. Furthermore, we can remark that there is a trade-off to find between the number of disjoint subsets N and the robust upper bound $\beta_{\mathtt{R}}$. Indeed, the larger N is, the smaller we expect the parameter $\beta_{\mathtt{R}}$ to be. However, if $N \geq m$, the bound cannot be tighter than $\sqrt{\ln(2)}$, and hence, does not converge towards 0 when $m \to +\infty$.

The major drawback of the algorithmic-based bounds is that we have to derive the parameter $\beta_{\mathtt{S}}$ or $\beta_{\mathtt{R}}$ for each algorithm. Hence, such derivation can be tedious, and the

upper-bound $\Phi_{\mathsf{a}}(\delta)$ is constant over the learning sample $\mathbb{S} \sim \mathcal{D}^m$. Another kind of bounds – called *PAC-Bayesian Bounds* – does not have such drawbacks and is appealing because of its facility to derive learning algorithms from it, as we will see in Part II. An in-depth introduction of these bounds is done in Chapter 2, but we give a quick overview in the rest of the chapter.

1.3.3 PAC-Bayesian Generalization Bounds

PAC-Bayesian bounds were introduced notably by Shawe-Taylor and Williamson (1997) and McAllester (1999), but it has been improved since then (Seeger, 2002; Maurer, 2004; Catoni, 2007). The PAC-Bayesian bounds differ significantly from the ones of Sections 1.3.1 and 1.3.2: they require a probability distribution (denoted by ρ) on the hypothesis set $\mathbb{H}$. This distribution is used to assign a weight $\rho(h)$ to each hypothesis $h \in \mathbb{H}$. Hopefully, when the hypothesis h generalizes well, its weight $\rho(h)$ should be high. Thanks to this assumption, we can consider a *stochastic* hypothesis: given an input $\mathbf{x} \in \mathbb{X}$, its output is obtained by *(1)* sampling a hypothesis h from ρ, and by *(2)* computing the prediction $h(\mathbf{x})$. Actually, the PAC-Bayesian bounds allow us to upper-bound the true risk of *stochastic* hypotheses which is the *expected* true risk of a hypothesis sampled from a distribution ρ. More precisely, the PAC-Bayesian bounds study the *expected* generalization gap $\mathbb{E}_{h\sim\rho}\,\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathbb{E}_{h\sim\rho}\,\mathsf{R}^{\ell}_{\mathcal{S}}(h)$. We present one bound derived by Maurer (2004) in the following theorem.

Theorem 1.3.6 (PAC-Bayesian Bound of Maurer (2004))**.** For any distribution $\mathcal{D}$ over $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any loss function $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y}) \to [0,1]$, for any distribution π on $\mathbb{H}$ (defined *apriori*), for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\begin{array}{l}\text{For all distributions } \rho \text{ on } \mathbb{H},\\ \mathbb{E}_{h\sim\rho}\,\mathsf{R}^{\ell}_{\mathcal{D}}(h) \le \mathbb{E}_{h\sim\rho}\,\mathsf{R}^{\ell}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]}\end{array}\right] \ge 1-\delta,$$

where the Kullback–Leibler (KL) divergence is defined as $\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{h\sim\rho}\ln\frac{\rho(h)}{\pi(h)}$. Equivalently, with probability at least $1-\delta$ over $\mathbb{S}\sim\mathcal{D}^m$, we have

$$\forall \text{ distributions } \rho \text{ on } \mathbb{H}, \quad \mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h) \le \mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]}.$$

The bound of this theorem depends on the KL divergence between ρ and π, which measures the complexity of the hypothesis sampled from ρ. Indeed, given a distribution π selected *apriori* before having the learning sample $\mathbb{S}\sim\mathcal{D}^m$, the higher $\mathrm{KL}(\rho\|\pi)$,

the more different the two distributions ρ and π are. Hence, in some sense, the KL divergence captures how much the distribution ρ depends on the data $\mathbb{S}$.

1.4 Conclusion and Summary

In this chapter, we have seen an introduction to statistical learning. It includes an overview of algorithms to perform hypothesis selection: Empirical Risk Minimization and Structural Risk Minimization. Additionally, we recall some generalization bounds based on the uniform convergence (*e.g.*, with the VC-Dimension or the Rademacher complexity) or an algorithmic property (*e.g.*, the stability or the robustness) Moreover, we recall a generalization bound from a key theory in our contributions: the PAC-Bayesian theory. This is why we recall, with more details, in the next chapter the PAC-Bayesian theory.

The PAC-Bayesian Theory and the Majority Vote

2

Contents

2.1 Introduction **56**
2.2 PAC-Bayesian Majority Votes **56**
2.2.1 Definition 56
2.2.2 Upper Bounds on the Majority Vote's Risk 60
2.3 PAC-Bayesian Bounds **64**
2.3.1 General PAC-Bayesian Bound of Germain *et al.* (2009) 66
2.3.2 General PAC-Bayesian Bound of Bégin *et al.* (2016) 73
2.4 Disintegrated PAC-Bayesian Bounds **75**
2.4.1 General Disintegrated Bound of Rivasplata *et al.* (2020) 76
2.4.2 Disintegrated Bound of Catoni (2007) 76
2.4.3 Disintegrated Bound of Blanchard and Fleuret (2007) 77
2.5 Conclusion and Summary **78**

Abstract

In this chapter, we introduce, with more details, the PAC-Bayes theory that we outlined in Chapter 1. This theory allows us to upper-bound the risk of the *stochastic* classifier which samples, for each input, a hypothesis to predict the output. Moreover, the risk of the *stochastic* classifier can be linked to the *majority vote*'s risk; we remind in this chapter the *majority vote classifier* which can be seen as a weighted combination of hypotheses. However, when we want to consider only one hypothesis, the *disinstegrated* PAC-Bayesian theory becomes more adapted. Indeed, it upper-bounds the true risk of a *single* hypothesis associated with a high weight. Such generalization bounds are recalled as well.

2.1 Introduction

In this chapter, we details one statistical learning theory that is key in this thesis: the PAC-Bayesian theory. It is introduced by Shawe-Taylor and Williamson (1997) and McAllester (1998) for which we recall some bounds in Section 2.3. This theory assumes that each hypothesis $h \in \mathbb{H}$ is associated with a (positive) weight $\rho(h)$ that forms a probability distribution. Thanks to this assumption, the expected generalization gap is upper-bounded over $h \sim \rho$. The expected generalization gap allows to bound the true risk of the *stochastic* classifier; for each input $\mathbf{x}$, the model *(i)* samples a hypothesis $h \sim \rho$ and *(ii)* predicts the label with $h(\mathbf{x})$.

The stochastic classifier is related to a model in which we are particularly interested: the majority vote (see Section 2.2.2). For instance, majority votes are considered in boosting (Freund and Schapire, 1996) or bagging (Breiman, 1996). In this context, a hypothesis is called *voter*, and a weight (modeled by the probability distribution ρ) is used to define the importance of each voter. Then, the majority vote is defined as a weighted combination of all the hypotheses from $\mathbb{H}$.

A majority vote might bring no significant improvements when the voters are strong (*i.e.*, when their individual risks are small). Hence, considering a *single* hypothesis associated with a high weight $\rho(h)$ can be a better option. Such a classifier is obtained by sampling from the probability distribution ρ. After sampling, generalization guarantees on the classifier can be derived through the *disintegrated* PAC-Bayesian bounds introduced independently by Blanchard and Fleuret (2007) and Catoni (2007). This type of bounds is recalled in Section 2.4.

For the sake of completeness, we defer in Appendix B the proofs.

2.2 PAC-Bayesian Majority Votes

2.2.1 Definition

Given a set of hypotheses $\mathbb{H}$, which is called set of voters in this context, the goal of a majority vote learning algorithm is to find a probability distribution ρ on $\mathbb{H}$. The distribution ρ defines the weights of the voters, *i.e.*, the importance of each voter in the majority vote. The majority vote is defined in the following way.

Definition 2.2.1 (Majority Vote)**.** For any hypothesis set $\mathbb{H}$ with voters $h : \mathbb{X} \to [-1, +1]$, for a distribution ρ on $\mathbb{H}$, the majority vote in the binary

classification setting ($\mathbb{Y} = \{-1, +1\}$) is defined as

$$\forall \mathbf{x} \in \mathbb{X}, \quad \mathsf{MV}_\rho(\mathbf{x}) \triangleq \operatorname{sign}\left(\mathop{\mathbb{E}}_{h \sim \rho} h(\mathbf{x}) \right).$$

In the multi-class setting ($\mathbb{Y} = \{1, 2, \ldots, l\}$), for any hypothesis set $\mathbb{H}$ with voters $h : \mathbb{X} \to \mathbb{Y}$, the ρ-weighted majority vote is defined as

$$\forall \mathbf{x} \in \mathbb{X}, \quad \mathsf{MV}_\rho(\mathbf{x}) \triangleq \operatorname*{argmax}_{y' \in \mathbb{Y}} \mathop{\mathbb{P}}_{h \sim \rho} [h(\mathbf{x}) = y'] = \operatorname*{argmax}_{y' \in \mathbb{Y}} \mathop{\mathbb{E}}_{h \sim \rho} \mathrm{I}\,[h(\mathbf{x}) = y'] .$$

In the binary setting, the majority vote predicts as label the sign associated with the ρ-weighted average of the voters' outputs. In the multi-class setting, the majority vote predicts the label $y' \in \mathbb{Y}$ with the highest associated score $\mathbb{P}_{h \sim \rho}[h(\mathbf{x}) = y']$. When the hypothesis set $\mathbb{H}$ is composed of voters $h : \mathbb{X} \to \{-1, +1\}$ in the binary setting, the majority vote for the multi-class setting can be seen as a generalization. Indeed, we have $\operatorname{sign}\left(\mathbb{E}_{h \sim \rho}\, h(\mathbf{x})\right) = +1$ if $\mathbb{P}_{h \sim \rho}[h(\mathbf{x}) = +1] \geq \mathbb{P}_{h \sim \rho}[h(\mathbf{x}) = -1]$ and $\operatorname{sign}\left(\mathbb{E}_{h \sim \rho}\, h(\mathbf{x})\right) = -1$ otherwise.

While this definition of the majority vote classifier might appear a bit restrictive, it encompasses multiple widespread classifiers. For example, the Support Vector Machine (Cortes and Vapnik, 1995) can be implicitly expressed as a majority vote where each voter depends on one example (Graepel *et al.*, 2005). The well-known k-Nearest Neighbors (Cover and Hart, 1967) is a majority vote (Bellet *et al.*, 2014). Additionally, when the voters depend on the whole learning sample $\mathbb{S}$, neural networks can be seen as a majority vote (Kawaguchi *et al.*, 2017; Viallard *et al.*, 2019).

There are many approaches to learn a majority vote classifier based on ensemble methods. For example, the bagging (Breiman, 1996) method splits the learning sample and learns one voter with each subset. Then, the majority vote classifier averages the decision of the voters to take the final decision; random forest (Breiman, 2001) is, for instance, a particular bagging algorithm for decision trees. Moreover, the weights ρ can be learned greedily: this is the purpose of boosting algorithms such as Adaboost (Freund and Schapire, 1996). This algorithm has been improved in various directions, *e.g.*, for the multi-class setting (Schapire and Singer, 1998, 1999, 2000; Zhu *et al.*, 2009) or the ranking setting with RankBoost (Freund *et al.*, 1998, 2003). Boosting algorithms have also been generalized for differentiable loss functions in a method called gradient boosting (Friedman, 2001).

Given a distribution $\mathcal{D}'$ on $\mathbb{X} \times \mathbb{Y}$ (that encompasses the distributions $\mathcal{D}$ and $\mathcal{S}$), the learner wants to learn MV_ρ that commits as few errors as possible on $\mathcal{D}$. To reduce the number of errors of the majority vote MV_ρ on $\mathcal{D}'$, the learner aims to minimize the risk $\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho)$ under the 01-loss in the following definition.

Definition 2.2.2 (Risk of the Majority Vote)**.** For any distribution $\mathcal{D}'$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, the risk of the majority vote is defined as

$$\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho) \triangleq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \mathrm{I}\left[\mathrm{MV}_\rho(\mathbf{x}) \neq y\right] = \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'} \left[\mathrm{MV}_\rho(\mathbf{x}) \neq y\right].$$

Specifically, when $\mathcal{D}' = \mathcal{D}$, the risk $\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho)$=$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho)$ is the true risk of the majority vote and when $\mathcal{D}' = \mathcal{S}$ the risk $\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho)$=$\mathrm{R}_{\mathcal{S}}(\mathrm{MV}_\rho)$ is the empirical risk of the majority vote. In order to gain insight into the majority vote's decision, the margin captures how much the classifier makes errors. Indeed, the majority vote's risk can be expressed in terms of the margin defined in the following way.

Definition 2.2.3 (Margin of the Majority Vote)**.** For any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, the margin of the majority vote is defined as

$$m_\rho(\mathbf{x}, y) \triangleq \mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y\right] - \max_{y'\in\mathbb{Y}, y'\neq y} \mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y'\right].$$

The margin is positive if the score $\mathbb{P}_{h\sim\rho}\left[h(\mathbf{x}) = y\right]$ is higher than the score associated with the other labels $y' \neq y$ and is negative otherwise. It captures if an example $(\mathbf{x}, y) \sim \mathcal{D}'$ is misclassified. Indeed, thanks to the margin, the majority vote's risk $\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho)$ can be rewritten as follows:

$$\begin{aligned}\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho) &= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y\right] \leq \max_{y'\in\mathbb{Y}, y'\neq y} \mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y'\right]\right] \\ &= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[m_\rho(\mathbf{x}, y) \leq 0\right].\end{aligned}$$

Hence, the risk is the probability that one of the scores is higher than the one of the correct label. Unfortunately, deriving a learning algorithm to optimize the margin can be challenging since it is non-convex *w.r.t.* the posterior ρ because of the $\max$. To overcome this issue, Laviolette *et al.* (2017) propose to consider a convex lower bound of the true margin called $\frac{1}{2}$-margin[1].

[1]We multiply the $\frac{1}{2}$-margin of Laviolette *et al.* (2017) by two.

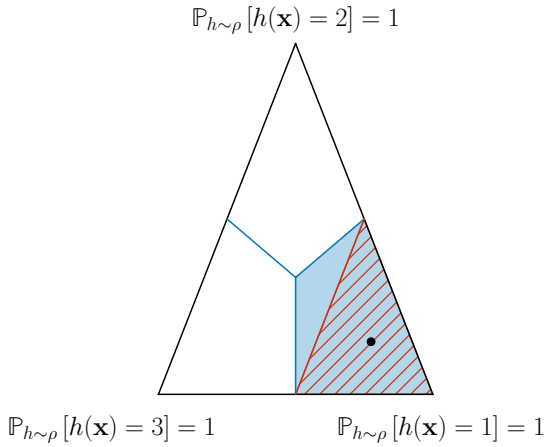


Figure 2.1. *Illustration of the majority vote's margin in the multi-class setting with 3 classes, i.e.,* $\mathbb{Y}=\{1,2,3\}$*. The triangle represents the ternary plot of the three scores where each triangle's vertex is the three possible maximum scores with* $\mathbb{P}_{h\sim\rho}[h(\mathbf{x})=i]=1$ *for all* $i\in\mathbb{Y}$*. The black dot represents the prediction of an example* $(\mathbf{x},y)\in\mathbb{X}\times\mathbb{Y}$ *by the majority vote: the scores are* $\mathbb{P}_{h\sim\rho}[h(\mathbf{x})=1]=0.7$ *and* $\mathbb{P}_{h\sim\rho}[h(\mathbf{x})=2]=\mathbb{P}_{h\sim\rho}[h(\mathbf{x})=3]=0.15$*. The blue area is where the majority vote predicts the label* y *for the input* $\mathbf{x}$ *and where the margin* $m_\rho(\mathbf{x},y)$ *is positive. Whereas the red area represents the predictions where* $\mathbb{P}_{h\sim\rho}[h(\mathbf{x})=1]\geq\frac{1}{2}$ *(i.e., where the* $\frac{1}{2}$*-margin* $\widehat{m_\rho}(\mathbf{x},y)$ *is positive).*

Definition 2.2.4 ($\frac{1}{2}$-Margin of the Majority Vote)**.** For any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, the $\frac{1}{2}$-margin is defined as

$$\widehat{m_\rho}(\mathbf{x},y)=2\left[\mathop{\mathbb{P}}_{h\sim\rho}[h(\mathbf{x})=y]-\frac{1}{2}\right].$$

When the score $\mathbb{P}_{h\sim\rho}[h(\mathbf{x})=y]$ exceeds $\frac{1}{2}$, the majority vote surely classifies correctly the example $(\mathbf{x},y)\sim\mathcal{D}'$. Hence, the idea of this margin is to compute the difference between the score and $\frac{1}{2}$: when the margin is positive, the example is correctly classified. We illustrate the difference between the two margins in Figure 2.1. In binary classification, the $\frac{1}{2}$-margin boils down to $\widehat{m_\rho}(\mathbf{x},y)=y\,\mathbb{E}_{h\sim\rho}\,h(\mathbf{x})=\mathbb{E}_{h\sim\rho}\,m_h(\mathbf{x},y)$ (which is the margin introduced in Section 1.1.2.2). From Definition 2.2.4, we can deduce the following upper bound on the majority vote's risk, *i.e.*, we have

$$\mathsf{R}_{\mathcal{D}'}(\mathsf{MV}_\rho)\leq\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\mathop{\mathbb{P}}_{h\sim\rho}[h(\mathbf{x})=y]\leq\frac{1}{2}\right]=\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\widehat{m_\rho}(\mathbf{x},y)\leq 0\right].$$

In general, the majority vote's risk $\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho)$ is not practical in learning algorithms because the gradient is zero (see Section 1.1.2.2). Hence, in the next section, we recall different upper bounds on the majority vote's risk considered as surrogate losses.

2.2.2 Upper Bounds on the Majority Vote's Risk

We recall three surrogates on the majority vote's risk introduced in the literature, namely, the Gibbs Risk (Langford and Shawe-Taylor, 2002; McAllester, 2003), the joint error (Lacasse *et al.*, 2006; Germain *et al.*, 2015; Masegosa *et al.*, 2020), and the C-Bound (Breiman, 2001; Lacasse *et al.*, 2006; Roy *et al.*, 2011). We make use of these surrogate in Part II to derive self-bounding learning algorithms.[2] More precisely, in Chapter 3, we leverage the Gibbs risk for the adversarially robust setting and, in Chapter 4, we learn majority vote classifiers through the C-Bound.

2.2.2.1 The Gibbs Risk

The first surrogate on the risk $\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho)$ introduced in the PAC-Bayesian literature is the *Gibbs risk* (Langford and Shawe-Taylor, 2002; McAllester, 2003). This risk gives the average performance of individual voters in the majority vote. It is defined in the following way.

Definition 2.2.5 (Gibbs Risk)**.** For any distribution $\mathcal{D}'$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, the *Gibbs risk* is defined as

$$r_{\mathcal{D}'}(\rho) \triangleq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{I}\left[h(\mathbf{x}) \neq y\right] = \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}',h\sim\rho} \left[h(\mathbf{x}) \neq y\right].$$

Put into words, the Gibbs risk is the ρ-weighted average of the voters' risk. Moreover, we can write the Gibbs Risk with respect to the $\frac{1}{2}$-margin. Indeed, we have

$$r_{\mathcal{D}'}(\rho) = \frac{1}{2}\left[1 - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x},y)\right]. \tag{2.1}$$

Langford and Shawe-Taylor (2002) show that the majority vote's risk is upper-bounded by twice the Gibbs risk. The inequality is recalled in the following theorem.

[2] A self-bounding algorithm, coined by Freund (1998), is a learning algorithm that comes with a generalization guarantee.

Theorem 2.2.1 (Risk Upper Bound Based on the Gibbs Risk)**.** For any distribution $\mathcal{D}'$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, we have

$$\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho) \leq 2\, r_{\mathcal{D}'}(\rho). \tag{2.2}$$

Proof. Deferred to Appendix B.1. ■

However, in ensemble methods where one wants to combine voters efficiently, the Gibbs risk appears to be an imprecise surrogate since the combination of voters might compensate for individual errors. Hence, other surrogates need to be taken into account.

2.2.2.2 Joint Error

The joint error, introduced by Lacasse *et al.* (2006), takes better the voters' correlation into account; this quantity is defined in the following way.

Definition 2.2.6 (Joint Error)**.** For any distribution $\mathcal{D}'$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, the *joint error* is defined as

$$\begin{aligned} e_{\mathcal{D}'}(\rho) &\triangleq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \mathop{\mathbb{E}}_{h\sim\rho} \mathop{\mathbb{E}}_{h'\sim\rho} \mathrm{I}\left[h(\mathbf{x}) \neq y\right] \mathrm{I}\left[h'(\mathbf{x}) \neq y\right] \\ &= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}',h\sim\rho,h'\sim\rho} \left[h(\mathbf{x}) \neq y, h'(\mathbf{x}) \neq y\right]. \end{aligned}$$

Similarly to Equation (2.1) for the Gibbs risk, we can reinterpret this surrogate with the $\frac{1}{2}$-margin. Indeed, from Germain *et al.* (2015), we have

$$\begin{aligned} e_{\mathcal{D}'}(\rho) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \left(\mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) \neq y\right]\right)^2 \\ &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \left(\frac{1}{2}\left[1 - \widehat{m_\rho}(\mathbf{x},y)\right]\right)^2 \\ &= \frac{1}{4}\left(1 - 2\mathop{\mathbb{E}}_{h\sim\rho} \widehat{m_\rho}(\mathbf{x},y) + \mathop{\mathbb{E}}_{h\sim\rho} \widehat{m_\rho}(\mathbf{x},y)^2\right). \end{aligned} \tag{2.3}$$

Recently, Masegosa *et al.* (2020) proposed to deal directly with the joint error to bound the majority vote's risk; Their bound is presented in the following theorem.

Theorem 2.2.2 (Risk Upper Bound Based on the Joint Error)**.** For any distribution $\mathcal{D}'$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, we have

$$\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho) \leq 4e_{\mathcal{D}'}(\rho). \tag{2.4}$$

Proof. Deferred to Appendix B.2. ■

This inequality captures the fact that the voters need to be sufficiently diverse and commit errors on different points. However, when the joint error $e_{\mathcal{D}'}(\rho)$ exceeds $\frac{1}{4}$, the bound exceeds 1 and is uninformative. In some cases, these two upper bounds appear to be less tight than another one: the C-Bound (Breiman, 2001; Lacasse *et al.*, 2006).

2.2.2.3 The C-Bound

The C-Bound[3] (Breiman, 2001; Lacasse *et al.*, 2006) is another surrogate (and upper bound) of the majority vote's risk. It is derived from the Chebyshev-Cantelli Inequality (see Theorem A.4.1). It depends on the *Gibbs risk*, the *joint error*, and the *disagreement*. This latter is defined in the following way.

Definition 2.2.7 (Disagreement)**.** For any distribution $\mathcal{D}'$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, the *disagreement* is defined as

$$\begin{aligned} d_{\mathcal{D}'}(\rho) &\triangleq 2 \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \mathop{\mathbb{E}}_{h\sim\rho} \mathop{\mathbb{E}}_{h'\sim\rho} \mathrm{I}\left[h(\mathbf{x}) \neq y\right] \mathrm{I}\left[h'(\mathbf{x}) = y\right] \\ &= 2 \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}',h\sim\rho,h'\sim\rho} \left[h(\mathbf{x}) \neq y, h'(\mathbf{x}) = y\right]. \end{aligned}$$

The higher the disagreement, the more the voters do not perform the same prediction for a given $(\mathbf{x}, y) \sim \mathcal{D}'$. Similarly to the Gibbs risk and the joint error, we can express the disagreement with the margin. Indeed, we have

$$d_{\mathcal{D}'}(\rho) = \frac{1}{2}\left[1 - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y)^2\right]. \tag{2.5}$$

Interestingly, by developing Equation (2.3), we can relate the Gibbs risk and the joint error to the disagreement with

$$d_{\mathcal{D}'}(\rho) = 2\left[r_{\mathcal{D}'}(\rho) - e_{\mathcal{D}'}(\rho)\right]. \tag{2.6}$$

[3]The term "C-Bound" was introduced in the PAC-Bayesian literature by (Lacasse *et al.*, 2006).

This inequality tells us two facts: *(i)* the lower the Gibbs risk, the lower the disagreement, and *(ii)* the disagreement increases as the joint error decreases. The expression of the disagreement in binary classification can be simplified (with voters $h : \mathbb{X} \to \{-1, +1\}$). Indeed, given an example $(\mathbf{x}, y) \sim \mathcal{D}'$ with $y \in \{-1, +1\}$, we have

$$\begin{aligned}\mathop{\mathbb{P}}_{h\sim\rho,h'\sim\rho}[h(\mathbf{x}) \neq h'(\mathbf{x})] &= \mathop{\mathbb{P}}_{h\sim\rho,h'\sim\rho}[h(\mathbf{x}) = y, h'(\mathbf{x}) \neq y] + \mathop{\mathbb{P}}_{h\sim\rho,h'\sim\rho}[h(\mathbf{x}) \neq y, h'(\mathbf{x})=y] \\ &= 2 \mathop{\mathbb{P}}_{h\sim\rho,h'\sim\rho}[h(\mathbf{x}) \neq y, h'(\mathbf{x}) = y]\,,\end{aligned}$$

which gives $d_{\mathcal{D}'}(\rho) = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}',h\sim\rho,h'\sim\rho}[h(\mathbf{x}) \neq h'(\mathbf{x})]$.
Thanks to this quantity, we are now able to recall the C-Bound. Note that it was first introduced by Lacasse *et al.* (2006) for the PAC-Bayesian majority vote in binary classification. The generalization to the multi-class setting has been introduced by Laviolette *et al.* (2017, Theorem 2 and Corollary 1); we recall the C-Bound in the following theorem.

Theorem 2.2.3 (The C-Bound)**.** For any distribution $\mathcal{D}'$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, if

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) > 0 \iff r_{\mathcal{D}'}(\rho) < \tfrac{1}{2} \iff 2e_{\mathcal{D}'}(\rho) + d_{\mathcal{D}'}(\rho) < 1,$$

we have

$$\mathsf{R}_{\mathcal{D}'}(\mathsf{MV}_\rho) \leq 1 - \frac{\left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\widehat{m_\rho}(\mathbf{x}, y)\right]\right)^2}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x}, y)\right)^2} \tag{2.7}$$

$$= 1 - \frac{\left(1 - 2r_{\mathcal{D}'}(\rho)\right)^2}{1 - 2d_{\mathcal{D}'}(\rho)} \tag{2.8}$$

$$= 1 - \frac{\left(1 - \left[2e_{\mathcal{D}'}(\rho) + d_{\mathcal{D}'}(\rho)\right]\right)^2}{1 - 2d_{\mathcal{D}'}(\rho)} \tag{2.9}$$

$$= C_{\mathcal{D}'}(\rho).$$

Proof. Deferred to Appendix B.3. ■

This surrogate is expressed as a trade-off between the first and the second statistical moment of the $\frac{1}{2}$-margin $\widehat{m_\rho}(\mathbf{x}, y)$ (Equation (2.7)). Based on Equations (2.1) and (2.5), the trade-off can be expressed through the the disagreement and the Gibbs risk (see Equation (2.8)). Moreover, from Equation (2.6), we can also see the C-Bound

as a trade-off between the joint error and the disagreement in Equation (2.9).

The three surrogates of Theorems 2.2.1 to 2.2.3 can be easily compared. For instance, when $r_{\mathcal{D}'}(\rho) \leq d_{\mathcal{D}'}(\rho)$, the C-Bound $C_{\mathcal{D}'}(\rho)$ (Equation (2.7)) is tighter than $2r_{\mathcal{D}'}(\rho)$ (Equation (2.2)) and $4e_{\mathcal{D}'}(\rho)$ (Equation (2.4)). Hence, when the disagreement increases, the C-Bound appears to be a good trade-off between the Gibbs risk and the disagreement. More precisely, the main interest of the C-bound compared to Equation (2.4) is that when $e_{\mathcal{D}'}(\rho)$ is close to $\frac{1}{4}$, the C-Bound can be close to 0 depending on the value of the disagreement $d_{\mathcal{D}'}(\rho)$: the C-bound is then more precise. Moreover, it is important to notice that the C-Bound is tighter than $4e_{\mathcal{D}'}(\rho)$ for all cases. We summarize the relationships between the different surrogates in the next theorem and illustrate it in Figure 2.2; this relation is notably given by Germain *et al.* (2015) and Masegosa *et al.* (2020).

Theorem 2.2.4 (Relationship between Theorems 2.2.1 to 2.2.3)**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any voters set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, if $r_{\mathcal{D}'}(\rho) < \frac{1}{2}$ (*i.e.*, $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) > 0$), we have

(i) $\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho) \leq C_{\mathcal{D}'}(\rho) \leq 4e_{\mathcal{D}'}(\rho) \leq 2r_{\mathcal{D}'}(\rho)$, if $r_{\mathcal{D}'}(\rho) \leq d_{\mathcal{D}'}(\rho)$,

(ii) $\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho) \leq 2r_{\mathcal{D}'}(\rho) \leq C_{\mathcal{D}'}(\rho) \leq 4e_{\mathcal{D}'}(\rho)$, otherwise.

Proof. Deferred to Appendix B.4. ■

Since the distribution $\mathcal{D}$ is unknown, the true risk of the majority vote $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho)$ is not computable. Hence, one can minimize the empirical risk of the majority vote $\mathrm{R}_{\mathcal{S}}(\mathrm{MV}_\rho)$ through the Empirical Risk Minimization approach (see Algorithm 1.3). However, this minimization does not necessarily lead to a low true risk $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho)$ since overfitting can occur (see Section 1.2). To tackle this issue, one solution is to deal with the minimization of a generalization bound to get a self-bounding algorithm (Freund, 1998), *i.e.*, the minimization of the risk with generalization guarantees. We will see in the next section PAC-Bayesian generalization bounds that further allow us to upper-bound the true risk $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho)$ in Part II.

2.3 PAC-Bayesian Bounds

The PAC-Bayesian theory[4], introduced by Shawe-Taylor and Williamson (1997) and McAllester (1999), aims to provide PAC generalization bounds for Bayesian-like algorithms. Such Bayesian algorithms assume a probability distribution defined

[4]See (Guedj, 2019; Alquier, 2021) for recent on the PAC-Bayesian theory.

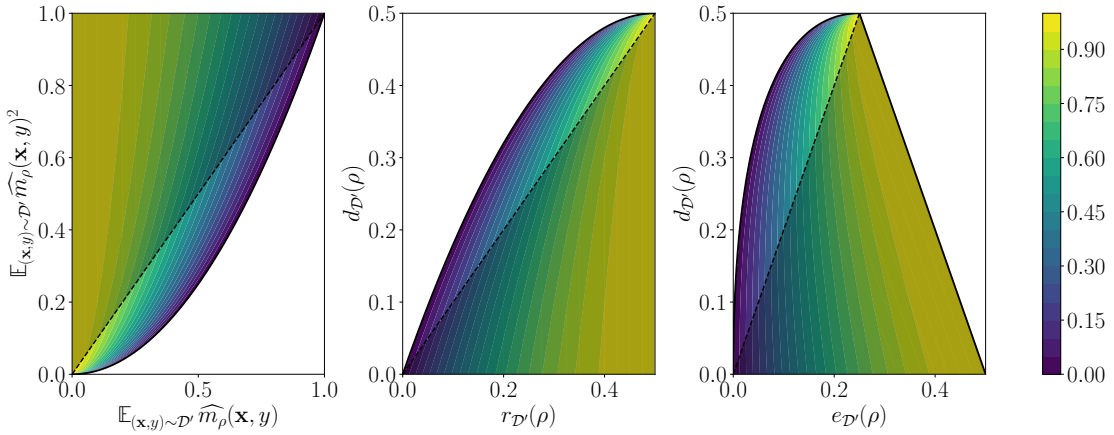


Figure 2.2. *Plots (from left to right) of the different C-Bounds in Equations* (2.7) *to* (2.9)*. For each plot, the darker area represents the cases where* $2r_{\mathcal{D}'}(\rho)$ *is tighter than the C-Bound* $C_{\mathcal{D}'}(\rho)$*. The dashed line represents the cases where* $4e_{\mathcal{D}'}(\rho)$ *matches the C-Bound* $C_{\mathcal{D}'}(\rho)$*.*

apriori on the hypothesis set $\mathbb{H}$, and thanks to Bayes' theorem, obtain an *a posteriori* probability distribution on $\mathbb{H}$ thanks to the learning sample; see Bishop (2007) for more details on this Bayesian inference procedure. Contrary to the classical Bayesian inference where the *posterior* distribution is proportional to the product of the *prior* distribution and the likelihood of the data, in PAC-Bayesian theory, an arbitrary *prior* distribution can be considered. Actually, the term "Bayesian" in the PAC-Bayesian theory comes from the fact that we usually upper-bound the *expected* generalization gap $\left|\mathbb{E}_{h\sim\rho_{\mathbb{S}}}\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathbb{E}_{h\sim\rho_{\mathbb{S}}}\mathsf{R}^{\ell}_{\mathcal{S}}(h)\right|$ where $h \in \mathbb{H}$ is sampled from a data-dependent distribution $\rho_{\mathbb{S}}$ called *posterior*. This theory has been extended to various settings such as transductive learning (Derbeko *et al.*, 2004; Bégin *et al.*, 2014), regression (Germain *et al.*, 2016; Shalaeva *et al.*, 2020), structured prediction (Laviolette *et al.*, 2017), domain adaptation (Germain *et al.*, 2020), or randomized learning (London, 2017).

In order to define a PAC-Bayesian bound more formally, we denote by $\rho_{\mathbb{S}}$ (or ρ) the *posterior* distribution on $\mathbb{H}$ and π the *prior* distribution. Each probability distribution ρ is defined through its probability density function $h \mapsto \rho(h)$ with respect to a reference measure[5] on $\mathbb{H}$; we denote by $\mathbb{M}(\mathbb{H})$ the set of probability density function on $\mathbb{H}$. Hence, the distribution $\rho_{\mathbb{S}} \in \mathbb{M}(\mathbb{H})$ is the Radon–Nikodym derivative of a probability measure *w.r.t.* the reference measure. We also denote by $\mathbb{M}^*(\mathbb{H}) \subseteq \mathbb{M}(\mathbb{H})$ the set of strictly positive probability densities on $\mathbb{H}$. Moreover, for convenience, we assume that the support of the posterior $\rho_{\mathbb{S}}$ is in the support of π, *i.e.*, if $\pi(h) = 0$ then $\rho_{\mathbb{S}}(h) = 0$

[5]For instance, if $\mathbb{H} = \mathbb{R}^d$, then the reference measure is the Lebesgue one.

(the absolute continuity); hence $\pi \in \mathbb{M}^*(\mathbb{H})$.

With the setting in place, we can now define the general form of a PAC-Bayesian bound in the following.

Definition 2.3.1 (PAC-Bayesian Generalization Bound)**.** Let $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$ be a loss function and $\phi : [0,1]^2 \to \mathbb{R}$ a generalization gap. A PAC-Bayesian bound is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, there exists a function $\Phi : \mathbb{M}(\mathbb{H})\times\mathbb{M}^*(\mathbb{H})\times(0,1] \to \mathbb{R}$, such that for any $\delta \in (0,1]$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\Big[\phi(\mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{S}}(h)) \leq \Phi\Big(\rho_{\mathbb{S}}, \pi, \delta\Big)\Big] \geq 1-\delta,$$

where *e.g.* $\phi(\mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{S}}(h)) = \Big|\mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big|$.

Definition 2.3.1 is a general definition in the sense that the expected generalization gap $\Big|\mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big|$ is not the only usable deviation between the true risk $\mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{D}}(h)$ and the empirical risk $\mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{S}}(h)$. For example, one can consider the one-sided difference $\mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{S}}(h)$ or the squared difference $\Big[\mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big]^2$. As we will see in this chapter, several gaps $\phi()$ have been considered in the literature. These gaps are upper-bounded with probability at least $1-\delta$ with a function $\Phi()$ that depends on δ and two probability distributions on $\mathbb{H}$. Usually, the lower the parameter $\delta \in (0,1]$, the higher the upper bound $\Phi()$, *i.e.*, the function $\Phi()$ is decreasing *w.r.t.* δ. Moreover, this upper bound $\Phi()$ depends on a data-dependent *posterior* distribution $\rho_{\mathbb{S}} \in \mathbb{M}(\mathbb{H})$ and a *prior* distribution $\pi \in \mathbb{M}^*(\mathbb{H})$. The prior distribution is data-free and can incorporate prior knowledge, *e.g.*, coming from an expert knowledge or an additional learning sample (Parrado-Hernández *et al.*, 2012; Dziugaite *et al.*, 2021). In the rest of the section, we present different instantiations of Definition 2.3.1, making an overview of the PAC-Bayesian bounds in the literature.

2.3.1 General PAC-Bayesian Bound of Germain *et al.* (2009)

There are different kinds of bounds in the PAC-Bayesian literature (*e.g.*, Seeger, 2002; McAllester, 2003; Catoni, 2007). Several existing bounds can be proved from a general theorem by Germain *et al.* (2009); it is recalled in the following theorem.

Theorem 2.3.1 (General PAC-Bayesian Bound of Germain *et al.* (2009))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any measurable function $\varphi : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \in \mathbb{M}(\mathbb{H}),\ \mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S}) \leq \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi} e^{\varphi(h',\mathbb{S}')}\right)\right] \geq 1-\delta,$$

where $\mathrm{KL}(\rho\|\pi) = \mathop{\mathbb{E}}_{h\sim\rho}\ln\frac{\rho(h)}{\pi(h)}$ is the Kullback-Leibler (KL) divergence between the distributions ρ and π.

Proof. Deferred to Appendix B.6. ■

Note that this bound holds for all posterior distributions $\rho \in \mathbb{M}(\mathbb{H})$, which includes notably the prior distribution π, or any data-dependent posterior $\rho_{\mathbb{S}}$. Furthermore, this bound is penalized by the KL divergence[6] between ρ and π. The closer the posterior ρ is to the prior π, the smaller the divergence and the bound. Moreover, the bound holds for a function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}$ that further captures a deviation between the true risk $\mathrm{R}^{\ell}_{\mathcal{D}}(h)$ and the empirical risk $\mathrm{R}^{\ell}_{\mathcal{S}}(h)$. For example, with $\varphi(h,\mathbb{S}) = m\phi(\mathrm{R}^{\ell}_{\mathcal{D}}(h), \mathrm{R}^{\ell}_{\mathcal{S}}(h))$ where $\phi()$ is convex, we are able to retrieve Definition 2.3.1. By setting this function $\phi()$ accordingly, one can retrieve some classical bounds that have been previously introduced in the literature (Seeger, 2002; McAllester, 2003; Catoni, 2007) as we show further.

2.3.1.1 McAllester-like bound

First of all, we can retrieve Theorem 1.3.6 presented in Chapter 1 which is a tighter version of the bound derived by McAllester (2003, Theorem 1). Indeed, by setting $\varphi(h,\mathbb{S}) = 2m\left[\mathrm{R}^{\ell}_{\mathcal{D}}(h) - \mathrm{R}^{\ell}_{\mathcal{S}}(h)\right]^2$ in Theorem 2.3.1, one can deduce the following result.

Theorem 2.3.2 (PAC-Bayesian Bound of McAllester (2003))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior $\pi \in \mathbb{M}^*(\mathbb{H})$, for any loss $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to [0,1]$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \in \mathbb{M}(\mathbb{H}), \left|\mathop{\mathbb{E}}_{h\sim\rho}\mathrm{R}^{\ell}_{\mathcal{D}}(h) - \mathop{\mathbb{E}}_{h\sim\rho}\mathrm{R}^{\ell}_{\mathcal{S}}(h)\right| \leq \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}}{\delta}\right]}\right] \geq 1-\delta. \tag{2.10}$$

[6] The principal properties of the KL divergence are given in Appendix B.5

Proof. Deferred to Appendix B.7. ■

According to Theorem 2.3.2, the gap $|\mathbb{E}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathbb{E}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{S}}(h)|$ tends to zero when the number of examples increase. Hence, the more examples we have, the closer the expected empirical risk $\mathbb{E}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{S}}(h)$ is from expected true empirical risk $\mathbb{E}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h)$ for all $\rho \in \mathbb{M}(\mathbb{H})$. Actually, bounding this gap gives an upper bound on the expected true risk. Indeed, with probability at least $1-\delta$ over the random choice of $\mathbb{S}\sim\mathcal{D}^m$, we have

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h) \leq \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m}\Big[\mathrm{KL}(\rho\|\pi) + \ln \tfrac{2\sqrt{m}}{\delta}\Big]} \tag{2.11}$$

$$\text{and} \quad \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h) \geq \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{S}}(h) - \sqrt{\frac{1}{2m}\Big[\mathrm{KL}(\rho\|\pi) + \ln \tfrac{2\sqrt{m}}{\delta}\Big]} \tag{2.12}$$

Put into words, the expected true risk $\mathbb{E}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h)$ can be lower and upper bounded by the expected empirical risk $\mathbb{E}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{S}}(h)$ and the bound $\sqrt{\frac{1}{2m}[\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta}]}$. When our objective is to obtain a low expected true risk $\mathbb{E}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h)$, one can obtain a distribution $\rho \in \mathbb{M}(\mathbb{H})$ minimizing the bound with the minimization problem

$$\min_{\rho\in\mathbb{M}(\mathbb{H})} \left\{ \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m}\Big[\mathrm{KL}(\rho\|\pi) + \ln \tfrac{2\sqrt{m}}{\delta}\Big]} \right\}.$$

Since the bound holds for all $\rho \in \mathbb{M}(\mathbb{H})$ (with high probability), it also holds for the optimal solution. The contributions in Part II rely on such a minimization problem to obtain models with a certified expected test risk. Nevertheless, as we will see further, the bound of Theorem 2.3.2 is not the tightest.

2.3.1.2 Catoni-like bound

Catoni (2007, Theorem 1.2.1) proposed a bound that can be tighter than the one of Theorem 2.3.2 by considering a parameter $c > 0$. His proposed bound can be retrieved from Theorem 2.3.1 by defining $\varphi(h, \mathbb{S}) = m\phi(\mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell}_{\mathcal{S}}(h))$ where the deviation is $\phi(\mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell}_{\mathcal{S}}(h)) = -\ln\Big(1 - (1-e^{-c})\,\mathsf{R}^{\ell}_{\mathcal{D}}(h)\Big) - c\mathsf{R}^{\ell}_{\mathcal{S}}(h)$. We state the bound in the following theorem.

Theorem 2.3.3 (PAC-Bayesian Bound of Catoni (2007))**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior $\pi \in \mathbb{M}^*(\mathbb{H})$, for any loss ℓ :

$\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to [0,1]$, for any $c>0$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}), -\ln\Big(1-\left[1-e^{-c}\right]\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)\Big) - c\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}\right]\right] \geq 1-\delta. \quad (2.13)$$

Proof. Deferred to Appendix B.8. ■

The result is difficult to interpret because of the gap $-\ln\Big(1-\left[1-e^{-c}\right]\mathbb{E}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)\Big) - c\,\mathbb{E}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)$ is not easy to analyze. However, rewriting it as an upper bound of the expected true risk $\mathbb{E}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)$ makes its interpretation easier. Indeed, with probability at least $1-\delta$ over the random choice of $\mathcal{S}\sim\mathcal{D}^m$, we have

$$\forall\rho\in\mathbb{M}(\mathbb{H}),\ \mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h) \leq \frac{1}{1-e^{-c}}\left[1-\exp\left(-c\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)-\frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}\right]\right)\right].$$

Put into words, the expected true risk $\mathbb{E}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)$ is upper-bounded by a trade-off, controlled by the parameter c, between the expected empirical risk $\mathbb{E}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)$ and the term $\frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}\right]$. This is in contrast with Equations (2.10) to (2.12) since this bound depends on a parameter $c>0$, which allows one to tune the tightness of the bound.
In practice, it is hard to set the parameter c since the bound holds with high probability on $\mathcal{S}\sim\mathcal{D}^m$ for all parameters $c>0$. Hence, it is not possible to condition c on $\mathcal{S}\sim\mathcal{D}^m$. To tackle this issue, one usually applies a union bound to get a bound holding for any c belonging to a countable set. Hopefully, as shown further, one can derive a bound that avoids this parameter and is potentially tighter.

2.3.1.3 Seeger-like bound

One of the tightest PAC-Bayesian bound (that avoids the parameter c) is the one proven by Seeger (2002). This bound depends on the KL divergence between two Bernoulli distributions. This function is defined in the following way.

Definition 2.3.2 (KL Divergence Between Two Bernoulli Distributions)**.** For any distribution $q\in[0,1]$ and $p\in[0,1]$, the small kl is defined as

$$\mathrm{kl}(q\|p) \triangleq \mathrm{KL}(\mathcal{B}(q)\|\mathcal{B}(p)) = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p},$$

where $\mathcal{B}(q)$ and $\mathcal{B}(p)$ are two Bernoulli distributions with bias q and p.

The first PAC-Bayesian theorem based on the divergence $\mathrm{kl}()$ was proved by SEEGER (2002). Few years later, by setting $\varphi(h, \mathbb{S}) = m\,\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))$, GERMAIN *et al.* (2009) retrieved an improved PAC-Bayesian generalization bound proved by MAURER (2004). MAURER's bound and stated in the following theorem.

Theorem 2.3.4 (PAC-Bayesian Bound of SEEGER (2002))**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior $\pi \in \mathbb{M}^*(\mathbb{H})$, for any loss $\ell : \mathbb{H}\times(\mathbb{X} \times \mathbb{Y})^m \to [0, 1]$, for any $\delta \in (0, 1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \in \mathbb{M}(\mathbb{H}),\ \mathrm{kl}\Big(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)\Big\|\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{S}}(h)\Big) \le \frac{1}{m}\Big[\mathrm{KL}(\rho\|\pi)+\ln\tfrac{2\sqrt{m}}{\delta}\Big]\right] \ge 1-\delta. \tag{2.14}$$

Proof. Deferred to Appendix B.9. ■

The bound of Theorem 2.3.4 consists in upper-bounding a deviation between $\mathbb{E}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)$ and $\mathbb{E}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{S}}(h)$. This deviation is hard to interpret. Thanks to the PINSKER's inequality (Theorem B.5.1), *i.e.*, $\forall (p, q) \in [0, 1]^2, 2(q - p)^2 \le \mathrm{kl}(q\|p)$, it can be shown that the bound of Theorem 2.3.2 is tighter than the one of Theorem 2.3.4. Interestingly, GERMAIN *et al.* (2009, Proposition 2.1) and LACASSE (2010, Proposition 6.2.2) related the bound of CATONI (Theorem 2.3.3) and Theorem 2.3.4 with the equality

$$\max_{c>0}\left\{-\ln\Big(1-\Big[1-e^{-c}\Big]p\Big)-cq\right\} = \mathrm{kl}(q\|p).$$

Put into words, given $p \in (0, 1]$ and $q \in [0, 1]$, the $\mathrm{kl}(q\|p)$ matches the function $-\ln(1-[1-e^{-c}]p)-cq$ with the optimal parameter $c \ge 0$. Expressed as it is, Equation (2.14) does not permit to upper or lower bound the expected true risk $\mathsf{R}^\ell_{\mathcal{D}}(h)$ contrary to Theorems 2.3.2 and 2.3.3. In order to rewrite Equation (2.14) (of Theorem 2.3.4), we define the inverting functions of $\mathrm{kl}()$ in the following way.

Definition 2.3.3 (Inverting Functions of $\mathrm{kl}()$)**.** Given $\tau \ge 0$, for any $q \in [0, 1]$, the inverting functions of the $\mathrm{kl}()$ are defined as

$$\overline{\mathrm{kl}}(q|\tau) \triangleq \max\Big\{p \in (0,1) \;\Big|\; \mathrm{kl}(q\|p) \le \tau\Big\},$$
$$\text{and}\quad \underline{\mathrm{kl}}(q|\tau) \triangleq \min\Big\{p \in (0,1) \;\Big|\; \mathrm{kl}(q\|p) \le \tau\Big\}.$$

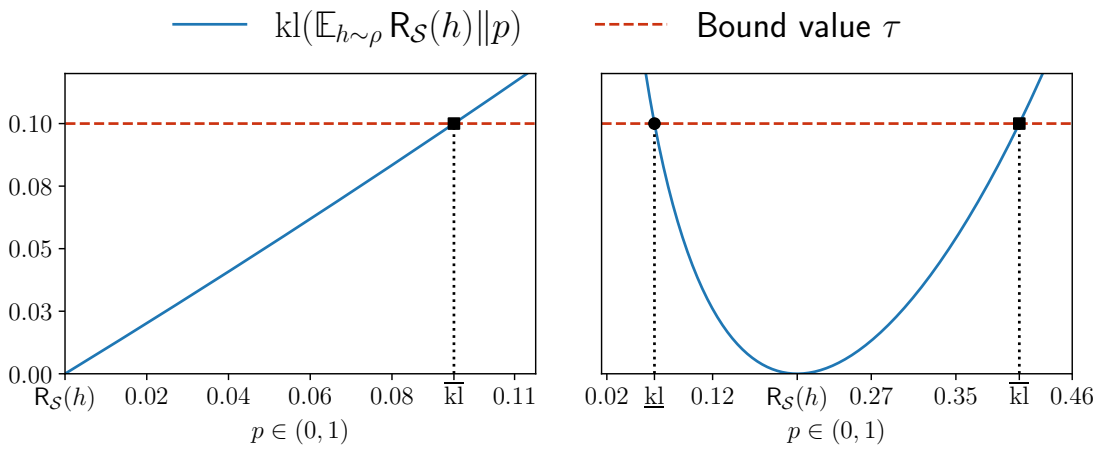


Figure 2.3. *Illustration of the inverting functions of* $\mathrm{kl}()$ *for* $q = R^{\ell}_{\mathcal{S}}(h)$ *s.t.* $R^{\ell}_{\mathcal{S}}(h) \in \{0.0, 0.2\}$ *in two different plots. We plot the curve of the function* $\mathrm{kl}(R^{\ell}_{\mathcal{S}}(h)\|\cdot)$*, the bound value* $\tau = 0.1$ *and the two inverting functions* $\underline{\mathrm{kl}}(R^{\ell}_{\mathcal{S}}(h)|\tau)$ *and* $\overline{\mathrm{kl}}(R^{\ell}_{\mathcal{S}}(h)|\tau)$ *(abbreviated* $\underline{\mathrm{kl}}$ *and* $\overline{\mathrm{kl}}$*). The dot (resp. the square) corresponds to the solution of the minimization (resp. maximization) problem associated with* $\underline{\mathrm{kl}}$ *(resp.* $\overline{\mathrm{kl}}$*).*

The function $\overline{\mathrm{kl}}()$ (*resp.* $\underline{\mathrm{kl}}()$) denotes the maximum (*resp.* minimum) value $p \in (0,1)$ such that the inequality $\mathrm{kl}(q\|p) \leq \tau$ holds. Figure 2.3 gives a graphical illustration of these inverting functions. The values associated with the inverting functions $\underline{\mathrm{kl}}()$ and $\overline{\mathrm{kl}}()$ can be approximated and easily computed from Pinsker's inequality (Theorem B.5.1). Indeed, we have

$$\overline{\mathrm{kl}}(q|\tau) \leq q + \sqrt{\frac{1}{2}\tau} \qquad \text{and} \qquad q - \sqrt{\frac{1}{2}\tau} \leq \underline{\mathrm{kl}}(q|\tau). \tag{2.15}$$

We present in Figure 2.4 an illustration of the tightness of this approximation.
To calculate $\underline{\mathrm{kl}}()$ and $\overline{\mathrm{kl}}()$ exactly, two optimization problems need to be solved; Reeb *et al.* (2018) proposed an algorithm based on the bisection method. We recall its pseudo-code in Algorithm 2.1. The principle of this algorithm is to iteratively refine the interval $[p_{\mathsf{min}}, p_{\mathsf{max}}]$ to which the bias $p \in (0,1]$ belongs. When the equality $\mathrm{kl}(q\|p) = \tau$ is attained or when the interval $[p_{\mathsf{min}}, p_{\mathsf{max}}]$ is small enough, the bias p is found.
Moreover, Reeb *et al.* (2018) found the expression of the derivatives with respect to q and τ, allowing them to derive bound minimization algorithms based on gradient descent; these derivatives are defined by

$$\frac{\partial \mathrm{k}(q|\psi)}{\partial q} = \frac{\ln \frac{1-q}{1-\mathrm{k}(q|\psi)} - \ln \frac{q}{\mathrm{k}(q|\psi)}}{\frac{1-q}{1-\mathrm{k}(q|\psi)} - \frac{q}{\mathrm{k}(q|\psi)}}, \text{ and } \frac{\partial \mathrm{k}(q|\psi)}{\partial \psi} = \frac{1}{\frac{1-q}{1-\mathrm{k}(q|\psi)} - \frac{q}{\mathrm{k}(q|\psi)}}, \tag{2.16}$$

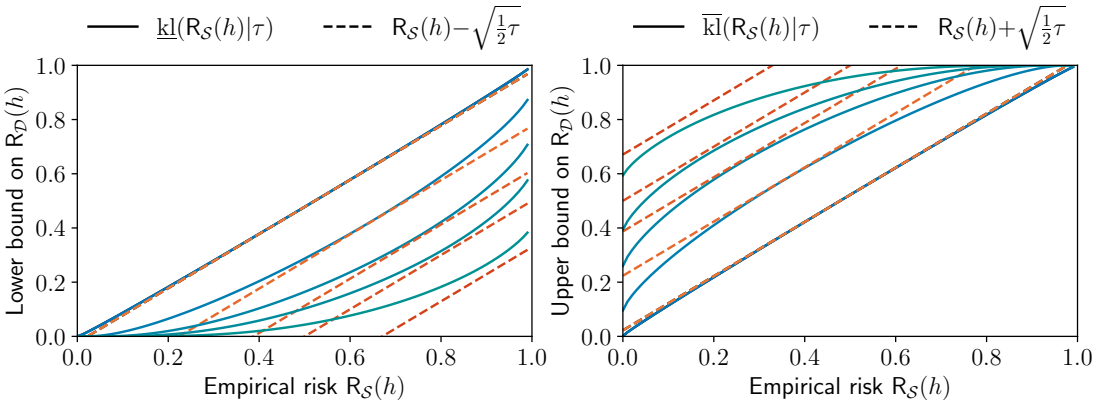


Figure 2.4. *Illustration of the tightness of $\underline{\mathrm{kl}}()$ and $\overline{\mathrm{kl}}()$. The line represents the functions $\underline{\mathrm{kl}}(\cdot|\tau)$ and $\overline{\mathrm{kl}}(\cdot|\tau)$ for different values of $\tau \in \{0.001, 0.1, 0.3, 0.5, 0.9\}$. On the left plot, we represent the function $\underline{\mathrm{kl}}(\cdot|\tau)$ while the right plot represents the function $\overline{\mathrm{kl}}(\cdot|\tau)$. In the left (resp. right) plot, the lower (resp. higher) the inverting function, the smaller (resp. larger) the value of τ. Moreover, for each inverting function, we plot with the dotted lines its approximation through* PINSKER*'s inequality.*

Algorithm 2.1 Compute $\overline{\mathrm{kl}}(q|\tau)$ *resp.* $\underline{\mathrm{kl}}(q|\tau)$ through the bisection method

Given: Bias $q \in [0,1]$ (the empirical risk), the bound value $\tau \geq 0$
Hyperparameters: tolerance ϵ, maximal number of iterations T_{max}
$p_{\mathsf{max}} \leftarrow 1$ and $p_{\mathsf{min}} \leftarrow q$ (resp. $p_{\mathsf{max}} \leftarrow q$ and $p_{\mathsf{min}} \leftarrow 0$)
for $t \leftarrow 1$ to T_{max} **do**
  $p = \frac{1}{2}\left[p_{\mathsf{min}} + p_{\mathsf{max}}\right]$
  if $\mathrm{kl}(q\|p) = \tau$ or $(p_{\mathsf{min}} - p_{\mathsf{max}}) < \epsilon$ **then return** p
  if $\mathrm{kl}(q\|p) > \tau$ **then** $p_{\mathsf{max}} = p$ (resp. $p_{\mathsf{min}} = p$)
  if $\mathrm{kl}(q\|p) < \tau$ **then** $p_{\mathsf{min}} = p$ (resp. $p_{\mathsf{max}} = p$)
return p

where $\mathrm{k}()$ is either $\underline{\mathrm{kl}}()$ or $\overline{\mathrm{kl}}()$.
Thanks to Definition 2.3.3, we can rewrite the bound of Theorem 2.3.4 to upper-bound the expected true risk $\mathbb{E}_{h\sim\rho}\mathrm{R}^{\ell}_{\mathcal{D}}(h)$. With probability at least $1-\delta$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$, we have for all $\rho \in \mathbb{M}(\mathbb{H})$

$$\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}^{\ell}_{\mathcal{D}}(h) \leq \overline{\mathrm{kl}}\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}^{\ell}_{\mathcal{S}}(h) \;\middle|\; \frac{1}{m}\Big[\mathrm{KL}(\rho\|\pi) + \ln \tfrac{2\sqrt{m}}{\delta}\Big] \right) \tag{2.17}$$

and

$$\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}^{\ell}_{\mathcal{D}}(h) \geq \underline{\mathrm{kl}}\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}^{\ell}_{\mathcal{S}}(h) \;\middle|\; \frac{1}{m}\Big[\mathrm{KL}(\rho\|\pi) + \ln \tfrac{2\sqrt{m}}{\delta}\Big] \right). \tag{2.18}$$

Thanks to the approximation in Equation (2.15), one can prove that the bound of

Theorem 2.3.4 is tighter than the one of Theorem 2.3.2. More precisely, if we apply Equation (2.15) to Equations (2.17) and (2.18), we retrieve Equations (2.11) and (2.12).

2.3.2 General PAC-Bayesian Bound of Bégin *et al.* (2016)

As shown previously, the KL divergence between the posterior distribution ρ and the prior distribution π is ubiquitous in PAC-Bayesian bounds. By considering this divergence, as illustrated in Theorem 2.3.1, the term $\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi}\exp\left[\varphi(h',\mathbb{S}')\right]$ appears inevitably. Indeed, the KL divergence $\mathrm{KL}(\rho\|\pi)$ between ρ and π can be expressed as a difference between two terms: $\mathbb{E}_{h\sim\rho}\varphi(h,\mathbb{S})$ and $\mathbb{E}_{h\sim\pi}\exp\left[\varphi(h,\mathbb{S})\right]$ thanks to Donsker and Varadhan (1976). This representation is actually used to obtain a *change of measure inequality* which quantifies how much two expectations (coming from two densities ρ and π) differ. This change of measure inequality and the expression of the KL divergence are in the following.

Proposition 2.3.1 (Donsker-Varadhan Variational Representation)**.** For any hypothesis set $\mathbb{H}$, for any distribution $\pi\in\mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any measurable function $\varphi:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m\to\mathbb{R}$ *s.t.* $\mathbb{E}_{h'\sim\pi}e^{\varphi(h',\mathbb{S})}<+\infty$ for all $\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m$, we have

$$\forall\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m,\quad\forall\rho\in\mathbb{M}(\mathbb{H}),\qquad\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})-\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}e^{\varphi(h,\mathbb{S})}\right)\leq\mathrm{KL}(\rho\|\pi)$$
$$\iff\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})\leq\mathrm{KL}(\rho\|\pi)+\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}e^{\varphi(h,\mathbb{S})}\right).$$

When the distribution ρ is defined as $\rho(h)=\pi(h)\dfrac{e^{\varphi(h,\mathbb{S})}}{\mathbb{E}_{h'\sim\pi}e^{\varphi(h',\mathbb{S})}}$, we have

$$\forall\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m,\qquad\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})-\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}e^{\varphi(h,\mathbb{S})}\right)=\mathrm{KL}(\rho\|\pi),$$
$$\iff\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})=\mathrm{KL}(\rho\|\pi)+\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}e^{\varphi(h,\mathbb{S})}\right).$$

Proof. Deferred to Appendix B.10. ■

As we can remark, this inequality resembles to the general bound of Germain *et al.* (2009) (in Theorem 2.3.1). Hence, the change of measure inequality appears indirectly in the PAC-Bayesian bounds making the constant term $\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi}e^{\varphi(h',\mathbb{S}')}$ arising in the bound. Other divergences can be considered to obtain different constant terms (that are further upper-bounded). For example, Ohnishi and Honorio (2021) prove

other change of measure inequalities for several divergences. Prior to this work, Bégin *et al.* (2016) derive a new general PAC-Bayesian bound with the Rényi divergence defined as $\mathrm{D}_\lambda(\rho\|\pi) = \frac{1}{\lambda-1}\ln\left(\mathbb{E}_{h\sim\pi}\left[\frac{\rho(h)}{\pi(h)}\right]^\lambda\right)$ for any $\lambda > 1$. Their bound is recalled in the following theorem.

Theorem 2.3.5 (General PAC-Bayesian Bound of Bégin *et al.* (2016))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}_*^+$, for any $\lambda > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}), \frac{\lambda}{\lambda-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})\right] \right.$$
$$\left. \leq \mathrm{D}_\lambda(\rho\|\pi) + \ln\left[\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right]\right] \geq 1-\delta.$$

Proof. Deferred to Appendix B.11. ■

Unlike Theorem 2.3.1, the function $(h,\mathbb{S}) \mapsto \frac{\lambda}{\lambda-1}\ln\left[\mathbb{E}_{h\sim\rho}\varphi(h,\mathbb{S})\right]$ represents the generalization gap and is upper-bounded by the Rényi divergence $\mathrm{D}_\lambda(\rho\|\pi)$ and a constant term $\ln[\frac{1}{\delta}\,\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{h'\sim\pi}\,\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}]$. At first sight, Theorem 2.3.5 seems very different from Theorem 2.3.1, however, they are actually related. Indeed, if we replace $\varphi(h,\mathbb{S})$ by $\exp(\frac{\lambda-1}{\lambda}\varphi(h,\mathbb{S}))$ and we apply Jensen's inequality (Theorem A.1.1) on the left-hand side, we obtain with probability at least $1-\delta$ over the random choice of $\mathbb{S}\sim\mathcal{D}^m$

$$\forall\rho\in\mathbb{M}(\mathbb{H}),\quad \mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S}) \leq \mathrm{D}_\lambda(\rho\|\pi) + \ln\left[\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}e^{\varphi(h',\mathbb{S}')}\right].$$

This bound is slightly looser than the one of Theorem 2.3.1: for any $\lambda > 1$ and for any distributions ρ and π, we have $\mathrm{KL}(\rho\|\pi) \leq \mathrm{D}_\lambda(\rho\|\pi)$ and $\lim_{\lambda\to1^+}\mathrm{D}_\lambda(\rho\|\pi) = \mathrm{KL}(\rho\|\pi)$ (Erven and Harremoës, 2014). As for the general bound in Theorem 2.3.1, setting the function $\varphi()$ and upper-bounding the term $\mathbb{E}_{\mathbb{S}\sim\mathcal{D}^m}\,\mathbb{E}_{h\sim\pi}\,\varphi(h,\mathbb{S})^{\frac{\lambda}{\lambda-1}}$ gives a computable bound. Namely, Theorem 2.3.5 allows one to obtain a McAllester, Catoni or a Seeger-like PAC-Bayesian bound based on the Rényi divergence. For the sake of completeness, we show in Appendix B.12 the proof of the three types of bounds based on Theorem 2.3.5. Compared to Theorems 2.3.2 to 2.3.4, only the KL divergence is replaced by the looser Rényi divergence.

Generally speaking, all the PAC-Bayesian bounds share a common property: they bound the *expectation* of $\varphi()$ *w.r.t.* the posterior ρ. However, it might be relevant to upper-bound $\varphi()$ for a *unique* hypothesis $h \sim \rho$: this is the purpose of the *disintegrated* bounds.

2.4 Disintegrated PAC-Bayesian Bounds

Deriving PAC-Bayesian guarantees for a unique hypothesis is tedious. Indeed, the PAC-Bayesian theory is tailored for bounding the *expected* true risk. Hence, additional derivations are needed to derive a PAC-Bayesian guarantee for a unique hypothesis. See Langford and Shawe-Taylor (2002), Langford (2005), and Germain *et al.* (2009) for some examples. To avoid this issue and to get a bound for a single hypothesis, another possible solution is to sample the hypothesis $h \in \mathbb{H}$ from the posterior distribution $\rho_{\mathbb{S}} \in \mathbb{M}(\mathbb{H})$. By doing so, the gap $\left|\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)\right|$ can be upper-bounded with a generalization bound. This bound is defined in the following way.

Definition 2.4.1 (Disintegrated PAC-Bayesian Generalization Bound)**.** Let $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$ be a loss function and $\phi : [0,1]^2 \to [0,1]$ a generalization gap. A *disintegrated* PAC-Bayesian bound is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, there exists a function $\Phi : \mathbb{M}(\mathbb{H}) \times \mathbb{M}^*(\mathbb{H}) \times (0,1] \to \mathbb{R}$ such that for any $\delta \in (0,1]$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \Big[\phi(\mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell}_{\mathcal{S}}(h)) \leq \Phi\Big(\rho_{\mathbb{S}}, \pi, \delta\Big)\Big] \geq 1 - \delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S}, \pi)$ is output by the deterministic algorithm A and $\phi()$ is, for example, $\phi(\mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell}_{\mathcal{S}}(h)) = \left|\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)\right|$.

Compared to the PAC-Bayesian bounds (see Definition 2.3.1), the expectation $\mathbb{E}_{h\sim\rho_{\mathbb{S}}}[\cdot]$ is moved outside the indicator function: this is the disintegration. Moreover, unlike the PAC-Bayesian bounds, the posterior $\rho_{\mathbb{S}}$ is obtained from an algorithm that depends on the prior $\pi \in \mathbb{M}^*(\mathbb{H})$ and the learning sample $\mathbb{S}$. This type of bounds has been introduced in two concurrent works, *i.e.*, Catoni (2007, Theorem 1.2.7) and Blanchard and Fleuret (2007). Moreover, there exists a general disintegrated bound as for the PAC-Bayesian bounds. We now present these three bounds by starting with the most general one.

2.4.1 General Disintegrated Bound of Rivasplata *et al.* (2020)

A general disintegrated PAC-Bayesian bound has been actually proposed very recently by Rivasplata *et al.* (2020, Theorem 1-(i)). This bound is presented below.

Theorem 2.4.1 (General Disintegrated Bound of Rivasplata *et al.* (2020))**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any measurable function $\varphi : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$, for any $\delta \in (0, 1]$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\varphi(h, \mathbb{S}) \leq \underbrace{\ln\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right] + \ln\left[\frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} \exp\left(\varphi(h', \mathbb{S}')\right)\right]}_{\Phi(\rho_{\mathbb{S}}, \pi, \delta)} \right] \geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S}, \pi)$ is output by the deterministic algorithm A.

Proof. Deferred to Appendix B.13. ■

Compared to the classical PAC-Bayesian bounds, this bound holds with high probability over the random choice of the learning sample $\mathbb{S} \sim \mathcal{D}^m$ *and* the hypothesis $h \sim \rho_{\mathbb{S}}$. Moreover, instead of depending on the KL divergence, the bound depends on the *disintegrated* KL divergence[7] $\ln \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}$. This term is the log ratio of the density of the prior $\pi(h)$ and the posterior distribution $\rho_{\mathbb{S}}(h)$ for the sampled hypothesis $h \sim \rho_{\mathbb{S}}$. Intuitively, the closer the posterior density $\rho_{\mathbb{S}}(h)$ to the prior density $\pi(h)$ for $h \sim \rho_{\mathbb{S}}$, the lower the disintegrated KL. As for Germain *et al.* (2009)'s general bound, we need to define $\varphi()$ and upper-bound the term $\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m} \mathbb{E}_{h'\sim\pi} \exp\left(\varphi(h', \mathbb{S}')\right)$ to obtain a bound that can be computed. Similarly to Theorems 2.3.1 and 2.3.5, this disintegrated bound generalizes other bounds such as the one derived by Catoni (2007, Theorem 1.2.7).

2.4.2 Disintegrated Bound of Catoni (2007)

The bound of Catoni (2007, Theorem 1.2.7), which is one of the first disintegrated bound in the literature, is recalled in the following theorem.

Theorem 2.4.2 (Disintegrated Bound of Catoni (2007))**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior $\pi \in \mathbb{M}^*(\mathbb{H})$, for any loss ℓ :

[7] The disintegration of the KL divergence has a slightly different meaning than the disintegration of the PAC-Bayesian bound: in the former case, the integration/expectation "is removed" from the divergence.

$\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to [0,1]$, for any $c > 0$, for any $\delta \in (0,1]$, for any algorithm $A\colon(\mathbb{X}\times\mathbb{Y})^m\times\mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),-\ln\left(1-\left[1-e^{-c}\right]\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)\right)-c\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{S}}(h)\right.$$
$$\left.\leq\frac{1}{m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{1}{\delta}\right]\right]\geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S},\pi)$ is output by the deterministic algorithm A.

Proof. Deferred to Appendix B.14. ■

After sampling the learning sample $\mathbb{S}\sim\mathcal{D}^m$ and the hypothesis $h\sim\rho_{\mathbb{S}}$, the obtained guarantee is similar to the one of Theorem 2.3.3: only the KL divergence is replaced by its disintegrated counterpart. Similar to the non-disintegrated PAC-Bayesian bound, other deviations between the true risk $\mathsf{R}^\ell_{\mathcal{D}}(h)$ and the empirical risk $\mathsf{R}^\ell_{\mathcal{S}}(\rho)$ can be considered. For instance, Blanchard and Fleuret (2007) proposed a bound on $\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))$.

2.4.3 Disintegrated Bound of Blanchard and Fleuret (2007)

The bound of Blanchard and Fleuret (2007) is based on another proof technique called *Occam's hammer*. For instance, they prove a Seeger (2002)-like disintegrated PAC-Bayesian bound from this framework; it is recalled in the following theorem.

Theorem 2.4.3 (Disintegrated Bound of Blanchard and Fleuret (2007))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi\in\mathbb{M}^*(\mathbb{H})$, for any loss $\ell:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})\to\{0,1\}$, for any $k>1$, for any $\delta\in(0,1]$, for any algorithm $A\colon(\mathbb{X}\times\mathbb{Y})^m\times\mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\mathrm{kl}_+(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))\leq\frac{1}{m}\left[\ln\frac{k+1}{\delta}+\left(1+\frac{1}{k}\right)\ln_+\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]\right]\geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S},\pi)$ is output by the deterministic algorithm A, the $\ln_+(x)=\max(\ln(x),0)$ and $\mathrm{kl}_+(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))=\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))$ if $\mathsf{R}^\ell_{\mathcal{S}}(h)<\mathsf{R}^\ell_{\mathcal{D}}(h)$ and 0 otherwise.

Proof. Deferred to Appendix B.15. ■

Similarly to CATONI (2007), this bound is parametrized: the tightness depends on the parameter $k > 1$. However, the optimal parameter depends on the log ratio $\ln_+ \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}$ and cannot be set in advance since we don't know the sampled $\mathbb{S} \sim \mathcal{D}^m$.

2.5 Conclusion and Summary

This chapter introduces generalization bounds from the PAC-Bayesian literature. These bounds allow deriving theoretical guarantees for some machine learning models, *e.g.*, the majority vote. They are further useful to derive practical learning algorithms guaranteeing that the model is not too sensitive to overfitting; see Part II. Indeed, in Chapter 3, we derive a new self-bounding learning algorithm that minimizes a PAC-Bayesian generalization bound for the adversarially robust setting. Roughly speaking, we derive surrogates similar to Equation (2.2) to obtain guarantees and derive learning algorithms that "robustify" the majority vote. Then, Chapter 4 introduces our contributions to minimizing of the PAC-Bayesian C-Bound, which aims to minimize the true risk of the majority vote. Lastly in Part II, we introduce the stochastic majority vote (where the distribution ρ follows a Dirichlet distribution) in Chapter 5. This majority vote allows using easily a PAC-Bayesian bound on the expected true risk to learn such a classifier.

However, the main drawback of the PAC-Bayesian generalization bounds is that we bound $\mathbb{E}_{h\sim\rho}\,\varphi(h, \mathbb{S})$ instead of $\varphi(h, \mathbb{S})$. In contrast, the disintegrated bounds allow us to bound the term $\varphi(h, \mathbb{S})$, which makes more sense if we want to deal with a unique hypothesis $h \sim \rho$. Part III shows the potential of such bounds for the analysis of the generalization of over-parametrized models. Indeed, Chapter 6 shows the first application of such bounds in practice, notably with over-parametrized models; this also leads to the derivation of new disintegrated bounds that are more appealing to optimization. Lastly, Chapter 7 introduces new perspectives based on these bounds since we can derive generalization bounds that do not depend on classical complexity measures such as the VC-Dimension or the Rademacher complexity (see Section 1.3).

Part II

PAC-Bayesian Majority Vote:
Theory and Self-bounding Algorithms

PAC-Bayesian Theory for the Robust Majority Vote

3

This chapter is based on the following paper

Paul Viallard, Guillaume Vidot, Amaury Habrard, and Emilie Morvant. A PAC-Bayes Analysis of Adversarial Robustness. *Advances in Neural Information Processing Systems (NeurIPS)*. (2021b)

Contents

3.1 Introduction **82**
3.2 Adversarially Robust Majority Vote **83**
3.2.1 Setting 83
3.2.2 Related Works 84
3.2.3 PAC-Bayesian Adversarial Risks 86
3.3 Adversarially Robust PAC-Bayes **87**
3.3.1 Relations Between the Adversarial Risks 87
3.3.2 PAC-Bayesian Bounds on the Adversarially Robust Majority Vote . . . 89
3.3.3 From the Bounds to an Algorithm 92
3.4 Experimental Evaluation on Differentiable Decision Trees **95**
3.4.1 Experiments 95
3.5 Conclusion and Summary **98**

Abstract

In this chapter, we derive the first general PAC-Bayesian generalization bounds for adversarial robustness, that estimate, how much the majority vote will be robust to imperceptible perturbations in the input. Instead of deriving a worst-case analysis of the risk of the majority vote over all the possible perturbations, we leverage the PAC-Bayesian framework recalled in Chapter 2 to bound the averaged risk on the perturbations. Our theoretically founded analysis has the advantage to provide general bounds *(i)* that are valid for any kind of adversarial attacks, *(ii)* that are tight, *(iii)* that can be directly minimized in a self-bounding algorithm to obtain a robust majority vote. We empirically show this robustness on different attacks.

3.1 Introduction

In this chapter, we first formalize in Section 3.2 the notion of majority vote for the adversarial robustness setting. To do so, we adapt the majority vote (recalled in Chapter 2 for supervised learning) by assuming that the inputs can be slightly modified/perturbed to fool the prediction of the majority vote, often in a malicious way; this setting is called *adversarial robustness*. The existence of such modified inputs, known as *adversarial examples* (Biggio *et al.*, 2013; Szegedy *et al.*, 2014) and illustrated in Figure 3.1.

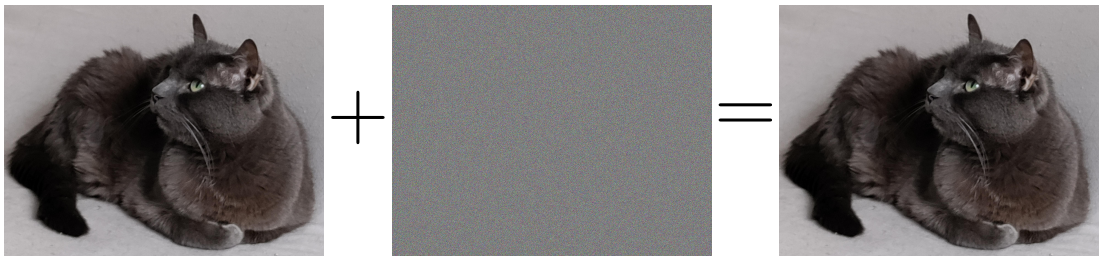


Figure 3.1. *On the left, the original image is predicted correctly by the classifier as "cat". The middle image corresponds to a perturbation/noise added to this original image. On the right, by applying the perturbation on the original image, the result is an image that looks identical to the original to the human eye, however, the prediction changes radically. This result image with the imperceptible perturbation is called an adversarial example.*

The models (*i.e.*, the majority vote in our case) must be robust to these small inputs' perturbations to better guarantees the user safety. Indeed, when machine learning models are applied to real problems, such as autonomous vehicles, the perturbations must not compromise the safety of the users. The perturbed examples is obtained from an *adversarial attack* that fools the considered model while the *adversarial defense* techniques enhance the adversarial robustness to make the attacks useless (see *e.g.*, Goodfellow *et al.*, 2015; Papernot *et al.*, 2016; Carlini and Wagner, 2017; Kurakin *et al.*, 2017; Zantedeschi *et al.*, 2017; Madry *et al.*, 2018). However, the majority votes and many other models lack guarantees on the robustness. To tackle this issue, we propose to formulate the adversarial robustness through the lens of the PAC-Bayesian theory recalled in Chapter 2; we call here our setting the *adversarially robust PAC-Bayes*.

The idea consists in considering an *averaged adversarial robustness risk* corresponding to the probability that the model misclassifies a perturbed example (this can be interpreted as an averaged risk over the perturbations). We also define an *averaged-max*

adversarial risk as the probability that there exists at least one perturbation that leads to a misclassification. These definitions, based on averaged quantities, have the advantage *(i)* of being suitable for the PAC-Bayesian framework and majority vote classifiers, and *(ii)* of being related to the classical adversarial robustness risk. Then, for each of our adversarial risks, we derive a PAC-Bayesian generalization bound that is valid to any kind of attack. From an algorithmic point of view, these bounds are directly minimizable to learn a majority vote robust in average to attacks. Since we directly minimize a generalization bound, our algorithms stand in the class of *self-bounding algorithms* (FREUND, 1998). We empirically illustrate that our framework is able to provide generalization guarantees with non-vacuous bounds for the adversarial risk while ensuring efficient protection to adversarial attacks.

Note that all the proofs of this chapter are deferred in Appendix C.

3.2 Adversarially Robust Majority Vote

3.2.1 Setting

We mainly adopt the setting of Chapter 2. We tackle *binary* classification tasks with the input space $\mathbb{X}=\mathbb{R}^d$ and the output/label space $\mathbb{Y}=\{-1,+1\}$. We assume that $\mathcal{D}$ is a fixed but unknown distribution on $\mathbb{X}\times\mathbb{Y}$. An example is denoted by $(\mathbf{x},y)\in\mathbb{X}\times\mathbb{Y}$. Let $\mathbb{S}=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$ be the learning sample of m examples *i.i.d.* sampled from $\mathcal{D}$; We denote the distribution of such m-sample by $\mathcal{D}^m$. Let $\mathbb{H}$ be a set of real-valued voters from $\mathbb{X}$ to $[-1,+1]$. Assuming the voters set $\mathbb{H}$ and a learning sample $\mathbb{S}$, our goal is to learn a well-performing ρ-weighted majority vote defined in Definition 2.2.1 by

$$\forall \mathbf{x}\in\mathbb{X},\quad \mathsf{MV}_\rho(\mathbf{x})=\operatorname{sign}\left[\mathop{\mathbb{E}}_{h\sim\rho} h(\mathbf{x})\right].$$

One wants to find a ρ-weighted majority vote that minimizes the true risk $\mathsf{R}_\mathcal{D}(\mathsf{MV}_\rho)$ on $\mathcal{D}$ defined in Definition 2.2.2 as

$$\mathsf{R}_\mathcal{D}(\mathsf{MV}_\rho)\triangleq\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x})\neq y\right].$$

However, in real-life applications, an imperceptible perturbation of the input can have a bad influence on the classification performance on unseen data (SZEGEDY *et al.*, 2014): the usual generalization guarantees do not stand anymore. Such an imperceptible perturbation can be modeled by a (relatively small) additive noise ϵ applied an input $\mathbf{x}$ leading to a perturbed input $\mathbf{x}+\epsilon$. Let $b>0$ and $\|\cdot\|$ be an arbitrary norm, the set of possible noises $\mathbb{B}$ is defined by[1]

$$\mathbb{B}=\Big\{\epsilon\in\mathbb{R}^d\;\Big|\;\|\epsilon\|\leq b\Big\}.$$

[1]The most used norms in the set of possible noises are the ℓ_1, ℓ_2 and ℓ_∞-norms.

The learner aims now to find an *adversarial robust* classifier that is robust in average to all noises in $\mathbb{B}$ over $(\mathbf{x}, y) \sim \mathcal{D}$. More formally, one wants to minimize the true adversarial risk $\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho)$ defined in the following definition.

Definition 3.2.1 (True/Empirical Adversarial Risk)**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any distribution ρ on $\mathbb{H}$, the *true adversarial risk* is defined as

$$\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \max_{\epsilon\in\mathbb{B}} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}+\epsilon) \neq y\right].$$

Since $\mathcal{D}$ is unknown, $\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho)$ cannot be directly computed, then one usually deals with the empirical adversarial risk defined as

$$\mathsf{A}_{\mathcal{S}}(\mathsf{MV}_\rho) = \frac{1}{m}\sum_{i=1}^{m} \max_{\epsilon\in\mathbb{B}} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}_i+\epsilon) \neq y_i\right].$$

In this chapter, our objective is to try to make the majority vote classifier MV_ρ robust to *adversarial attacks* that aim at finding an *adversarial example* $\mathbf{x} + \epsilon^*(\mathbf{x},y)$ to fool $\mathsf{MV}_\rho()$ for given example $(\mathbf{x}, y)$, where $\epsilon^*(\mathbf{x},y)$ is defined as

$$\epsilon^*(\mathbf{x},y) \in \operatorname*{argmax}_{\epsilon\in\mathbb{B}} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}+\epsilon) \neq y\right]. \tag{3.1}$$

In consequence, *adversarial defense* mechanisms often rely on the adversarial attacks by replacing the original examples $(\mathbf{x}, y)$ with the adversarial ones $(\mathbf{x} + \epsilon^*(\mathbf{x}, y), y)$ during the learning phase; this procedure is called adversarial training. Even if there are other defenses, as we will see later, adversarial training appears to be one of the most efficient defense mechanisms (Ren *et al.*, 2020). Optimizing Equation (3.1) is however intractable due to the non-convexity of MV_ρ induced by the sign function. The adversarial attacks of the existing frameworks in the literature (that we discuss in Section 3.2.2) aim at finding the optimal perturbation $\epsilon^*(\mathbf{x},y)$, but, in practice, one considers an approximation of this perturbation.

3.2.2 Related Works

Adversarial Attacks/Defenses. Numerous methods[2] exist to solve— or approximate —the optimization of Equation (3.1). Among them, the Fast Gradient Sign Method (fgsm) of Goodfellow *et al.* (2015) is an attack consisting in generating a noise ϵ in the direction of the gradient of the loss function with respect to the input $\mathbf{x}$. Kurakin *et al.* (2017) introduced ifgsm, an iterative version of fgsm: at

[2]The reader can refer to Ren *et al.* (2020) for a survey on adversarial attacks and defenses.

each iteration, one repeats FGSM and adds to $\mathbf{x}$ a noise, that is the sign of the gradient of the loss with respect to $\mathbf{x}$. Following the same principle as IFGSM, MADRY *et al.* (2018) proposed a method based on Projected Gradient Descent (PGD) that includes a random initialization of $\mathbf{x}$ before the optimization. Another technique known as the *Carlini and Wagner Attack* (CARLINI and WAGNER, 2017) aims at finding adversarial examples $\mathbf{x} + \epsilon^*(\mathbf{x},y)$ that are as close as possible to the original $\mathbf{x}$, *i.e.*, they want an attack being the most imperceptible as possible. However, producing such imperceptible perturbation leads to a high-running time in practice. Contrary to the most popular techniques that look for a model with a low adversarial robust risk, our work stands in another line of research where the idea is to relax this worst-case risk measure by considering an *averaged* adversarial robust risk over the noises instead of a $\max$-based formulation (see, *e.g.*, ZANTEDESCHI *et al.*, 2017; HENDRYCKS and DIETTERICH, 2019). Our averaged formulation is introduced in the Section 3.2.

Generalization Bounds for Adversarial Robustness. Recently, few generalization bounds for adversarial robustness have been introduced (*e.g.*, KHIM and LOH, 2018; COHEN *et al.*, 2019; MONTASSER *et al.*, 2019; PINOT *et al.*, 2019; SALMAN *et al.*, 2019; YIN *et al.*, 2019; MONTASSER *et al.*, 2020; PINOT *et al.*, 2022). KHIM and LOH, and YIN *et al.*'s results are Rademacher complexity-based bounds. The former makes use of a surrogate of the adversarial risk; the latter provides bounds in the specific case of neural networks and linear classifiers and involves an unavoidable polynomial dependence on the dimension of the input. MONTASSER *et al.* study robust PAC-learning for PAC-learnable classes with finite VC-dimension for unweighted majority votes that have been "robustified" with a boosting algorithm. However, their algorithm requires to consider all possible adversarial perturbations for each example, which is intractable in practice, and their bound also suffers from a large constant as indicated at the end of the MONTASSER *et al.* (Theorem 3.1 2019)'s proof. COHEN *et al.* provide bounds that estimate what is the minimum noise to get an adversarial example (in the case of perturbations expressed as Gaussian noise) while our results give the probability of being fooled by an adversarial example. SALMAN *et al.* leverage COHEN *et al.*'s method and adversarial training in order to get tighter bounds. Moreover, FARNIA *et al.* present margin-based bounds on the adversarial robust risk for specific neural networks and attacks (such as FGSM or PGD). While they made use of a classical PAC-Bayes bound, their result is not a PAC-Bayesian analysis and stands in the family of uniform-convergence bounds (see NAGARAJAN and KOLTER, 2019b, Ap. J for details). In this thesis, we provide PAC-Bayesian bounds for general models expressed as majority votes, their bounds are thus not directly comparable to ours.

3.2.3 PAC-Bayesian Adversarial Risks

Instead of looking for the noise from Equation (3.1) that maximizes the chance of fooling the algorithm, we propose to model the perturbation according to an example-dependent distribution. This example-dependent distribution is further used to define our new risks. First let us define $\mathcal{B}_{(x,y)}$ a distribution, on the set of possible noises $\mathbb{B}$, that is dependent on an example $(\mathbf{x}, y) \in \mathbb{X}\times\mathbb{Y}$. Then, we denote as $\mathcal{E}$ the distribution on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$ defined as

$$\mathcal{E}((\mathbf{x}, y), \epsilon) = \mathcal{D}(\mathbf{x}, y) \cdot \mathcal{B}_{(x,y)}(\epsilon),$$

which further permits to generate *perturbed examples*. For a given example $(\mathbf{x}_i, y_i)\sim\mathcal{D}$, we consider a set of n perturbations sampled from $\mathcal{B}_{(x_i,y_i)}$ denoted by $\mathfrak{C}_i=\{\epsilon_j^i\}_{j=1}^n$. Then we consider as a learning set the $m\times n$-sample $\widehat{\mathbb{S}} = \{((\mathbf{x}_i, y_i), \mathfrak{C}_i)\}_{i=1}^m \in (\mathbb{X}\times\mathbb{Y}\times\mathbb{B}^n)^m$. In other words, each $((\mathbf{x}_i, y_i), \mathfrak{C}_i) \in \widehat{\mathbb{S}}$ is sampled from a distribution that we denote by $\mathcal{E}^n$ such that

$$\mathcal{E}^n((\mathbf{x}_i, y_i), \mathfrak{C}_i) = \mathcal{D}(\mathbf{x}_i, y_i) \cdot \prod_{j=1}^{n} \mathcal{B}_{(x_i,y_i)}(\epsilon_j^i).$$

Furthermore, we denote as $(\mathcal{E}^n)^m$ the empirical distribution on the perturbed learning sample consisted of m examples and n perturbations for each example. Then, inspired by the works of Zantedeschi *et al.* (2017) and Hendrycks and Dietterich (2019), we define our *robustness averaged adversarial risk* as follows.

Definition 3.2.2 (Averaged Adversarial Risk)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any distribution ρ on $\mathbb{H}$, the averaged adversarial risk of MV_ρ is defined as

$$\begin{aligned} \mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) &= \mathop{\mathbb{P}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}}(\mathsf{MV}_\rho(\mathbf{x}+\epsilon) \neq y) \\ &= \mathop{\mathbb{E}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}+\epsilon) \neq y\right]. \end{aligned}$$

The empirical averaged adversarial risk is computed on a $m\times n$-sample $\widehat{\mathbb{S}} = \{((\mathbf{x}_i, y_i), \mathfrak{C}_i)\}_{i=1}^m$ is

$$\mathsf{R}_{\widehat{\mathcal{S}}}(\mathsf{MV}_\rho) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}_i+\epsilon_j^i) \neq y_i\right].$$

As we will show in Proposition 3.3.1, the risk $\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho)$ is considered optimistic regarding $\epsilon^*(\mathbf{x},y)$ of Equation (3.1). Indeed, instead of taking the ϵ maximizing the loss, the ϵ is drawn from a distribution. Hence, it can lead to a non-informative risk if

the ϵ are not informative enough to fool the classifier. To overcome this, we propose an extension that we refer as *averaged-max adversarial risk*. Note that we abuse the notation of the adversarial true risk defined in Definition 3.2.1.

Definition 3.2.3 (Averaged-Max Adversarial Risk)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any distribution ρ on $\mathbb{H}$, the averaged-max adversarial risk of MV_ρ is defined as

$$\mathsf{A}_{\mathcal{E}^n}(\mathsf{MV}_\rho) = \mathop{\mathbb{P}}_{((\mathbf{x},y),\mathfrak{C})\sim\mathcal{E}^n}\Big(\exists\,\epsilon\in\mathfrak{C}, \mathsf{MV}_\rho(\mathbf{x}+\epsilon)\neq y\Big).$$

The empirical averaged-max adversarial risk computed on a $m\times n$-sample $\widehat{\mathbb{S}} = \{((\mathbf{x}_i,y_i),\mathfrak{C}_i)\}_{i=1}^m$ is

$$\mathsf{A}_{\widehat{\mathcal{S}}}(\mathsf{MV}_\rho) = \frac{1}{m}\sum_{i=1}^m \max_{\epsilon\in\mathfrak{C}_i} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}_i+\epsilon)\neq y_i\right].$$

For an example $(\mathbf{x},y)\sim\mathcal{D}$, instead of checking if one perturbed example $\mathbf{x}+\epsilon$ is adversarial, we sample n perturbed examples $\mathbf{x}+\epsilon_1,\ldots,\mathbf{x}+\epsilon_n$ and we check if at least one example is adversarial.

3.3 Adversarially Robust PAC-Bayes

We show in Section 3.3.1 the relations between the different risks. Section 3.3.2 introduces the PAC-Bayesian bounds to assess the robustness of the majority vote.

3.3.1 Relations Between the Adversarial Risks

Proposition 3.3.1 below shows the intrinsic relationships between the classical adversarial risk $\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho)$ and our two relaxations $\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho)$ and $\mathsf{A}_{\mathcal{E}^n}(\mathsf{MV}_\rho)$. In particular, Proposition 3.3.1 shows that the larger number of perturbed examples n, the higher is the chance to get an adversarial example and then to be close to the adversarial risk $\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho)$.

Proposition 3.3.1 (Relations Between the Averaged Adversarial Risks)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any distribution ρ on $\mathbb{H}$, for any $(n,n')\in\mathbb{N}^2$, with $1\leq n'\leq n$, we have

$$\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) \;\leq\; \mathsf{A}_{\mathcal{E}^{n'}}(\mathsf{MV}_\rho) \;\leq\; \mathsf{A}_{\mathcal{E}^n}(\mathsf{MV}_\rho) \;\leq\; \mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho). \tag{3.2}$$

Proof. Deferred to Appendix C.1. ■

The left-hand side of Equation (3.2) confirms that the averaged adversarial risk $\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho)$ is optimistic regarding the classical $\mathrm{A}_{\mathcal{D}}(\mathrm{MV}_\rho)$. Proposition 3.3.2 estimates how close $\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho)$ can be to $\mathrm{A}_{\mathcal{D}}(\mathrm{MV}_\rho)$.

Proposition 3.3.2 (Classical and Averaged Adversarial Risks)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any distribution ρ on $\mathbb{H}$, we have

$$\mathrm{A}_{\mathcal{D}}(\mathrm{MV}_\rho) - \mathrm{TV}(\gamma\|\Gamma) \;\leq\; \mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho),$$

where Γ and γ are distributions on $\mathbb{X}\times\mathbb{Y}$ and $\mathrm{TV}(\gamma\|\Gamma) = \mathbb{E}_{(\mathbf{x}',y')\sim\Gamma}\,\frac{1}{2}\left|\frac{\gamma(\mathbf{x}',y')}{\Gamma(\mathbf{x}',y')}-1\right|$, is the Total Variation (TV) distance between γ and Γ.
The density $\Gamma(\mathbf{x}',y')$ corresponds to the probability of drawing a perturbed example $(\mathbf{x}',y') = (\mathbf{x}+\epsilon, y)$ with $((\mathbf{x},y),\epsilon)\sim\mathcal{E}$, *i.e.*, we have

$$\Gamma(\mathbf{x}',y') = \Pr_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}}\left[\mathbf{x}+\epsilon = \mathbf{x}',\; y = y'\right].$$

The density $\gamma(\mathbf{x}',y')$ is the probability to draw an adversarial example $(\mathbf{x}',y') = (\mathbf{x}+\epsilon^*(\mathbf{x},y), y)$ with $(\mathbf{x},y)\sim\mathcal{D}$, *i.e.*, we have

$$\gamma(\mathbf{x}',y') = \Pr_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbf{x}+\epsilon^*(\mathbf{x},y) = \mathbf{x}',\; y = y'\right].$$

Proof. Deferred to Appendix C.2. ■

Note that $\epsilon^*(\mathbf{x},y)$ depends on ρ, hence γ depends on ρ. From Proposition 3.3.2 and the distributions Γ and γ, the risks $\mathrm{A}_{\mathcal{D}}(\mathrm{MV}_\rho)$ and $\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho)$ can be rewritten as

$$\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho) = \Pr_{(\mathbf{x}',y')\sim\Gamma}\left[\mathrm{MV}_\rho(\mathbf{x}') \neq y'\right],$$
$$\text{and}\quad \mathrm{A}_{\mathcal{D}}(\mathrm{MV}_\rho) = \Pr_{(\mathbf{x}',y')\sim\gamma}\left[\mathrm{MV}_\rho(\mathbf{x}') \neq y'\right].$$

Finally, Propositions 3.3.1 and 3.3.2 relate the adversarial risk $\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho)$ to the "standard" adversarial risk $\mathrm{A}_{\mathcal{D}}(\mathrm{MV}_\rho)$. Indeed, from the two propositions we obtain

$$\mathrm{A}_{\mathcal{D}}(\mathrm{MV}_\rho) - \mathrm{TV}(\gamma\|\Gamma) \;\leq\; \mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho) \;\leq\; \mathrm{A}_{\mathcal{E}^n}(\mathrm{MV}_\rho) \;\leq\; \mathrm{A}_{\mathcal{D}}(\mathrm{MV}_\rho). \tag{3.3}$$

Hence, the smaller the TV distance $\mathrm{TV}(\gamma\|\Gamma)$, the closer the averaged adversarial risk $\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho)$ is from $\mathrm{A}_{\mathcal{D}}(\mathrm{MV}_\rho)$ and the more probable an example $((\mathbf{x},y),\epsilon)$ sampled

from $\mathcal{E}$ would be adversarial, *i.e.*, when our "averaged" adversarial example looks like a "specific" adversarial example. Moreover, Equation (3.3) justifies that the PAC-Bayesian point of view makes sense for adversarial learning with theoretical guarantees: the PAC-Bayesian guarantees we derive in the next section for our adversarial risks implies guarantees on the adversarial risk $\mathrm{A}_{\mathcal{D}}(\mathrm{MV}_\rho)$.

3.3.2 PAC-Bayesian Bounds on the Adversarially Robust Majority Vote

To derive PAC-Bayesian generalization bounds on the risk $\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho)$, respectively on $\mathrm{A}_{\mathcal{E}^n}(\mathrm{MV}_\rho)$, we consider one of the classical surrogates, *i.e.*, the Gibbs risk (Definition 2.2.5), defined below in Equation (3.4), respectively Equation (3.5).

Definition 3.3.1 (Surrogates on the Averaged Adversarial Risks)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathfrak{C}$, for any hypothesis set $\mathbb{H}$, for any ρ on $\mathbb{H}$,

$$r_{\mathcal{E}}(\rho) = \mathop{\mathbb{E}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}} \frac{1}{2}\left[1 - \mathop{\mathbb{E}}_{h\sim\rho} y h(\mathbf{x}+\epsilon)\right], \tag{3.4}$$

$$\text{and} \quad a_{\mathcal{E}^n}(\rho) = \mathop{\mathbb{E}}_{((\mathbf{x},y),\mathfrak{C})\sim\mathcal{E}^n} \frac{1}{2}\left[1 - \min_{\epsilon\in\mathfrak{C}} \left(y \mathop{\mathbb{E}}_{h\sim\rho} h(\mathbf{x}+\epsilon)\right)\right]. \tag{3.5}$$

Put into words, these surrogates are expressed as the expectation over ρ of the individual risks of the voters involved in $\mathbb{H}$. From an algorithmic perspective, $r_{\mathcal{E}}(\rho)$ and $a_{\mathcal{E}^n}(\rho)$ have the advantages *(i)* of being differentiable contrary to $\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho)$ and $\mathrm{A}_{\mathcal{E}^n}(\mathrm{MV}_\rho)$, and *(ii)* to upper-bound to $\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho)$ and $\mathrm{A}_{\mathcal{E}^n}(\mathrm{MV}_\rho)$ as follows.

Theorem 3.3.1 (Upper Bounds on the Surrogates)**.** For any distributions $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$ and ρ on $\mathbb{H}$, for any $n>1$, we have

$$\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho) \le 2r_{\mathcal{E}}(\rho), \qquad \text{and} \qquad \mathrm{A}_{\mathcal{E}^n}(\mathrm{MV}_\rho) \le 2a_{\mathcal{E}^n}(\rho).$$

Proof. Deferred to Appendix C.3. ■

This theorem implies that a generalization bound on $r_{\mathcal{E}}(\rho)$, resp $a_{\mathcal{E}^n}(\rho)$ leads to a generalization bound on $\mathrm{R}_{\mathcal{E}}(\mathrm{MV}_\rho)$, *resp.*, $\mathrm{A}_{\mathcal{E}^n}(\mathrm{MV}_\rho)$. Theorem 3.3.2 *resp.* Theorem 3.3.3 below presents our PAC-Bayesian generalization bounds for $r_{\mathcal{E}}(\rho)$ *resp.* $a_{\mathcal{E}^n}(\rho)$.

Theorem 3.3.2 (PAC-Bayesian Bound on $r_{\mathcal{E}}(\rho)$)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any set of voters $\mathbb{H}$, for any prior $\pi \in \mathbb{M}*(\mathbb{H})$ on $\mathbb{H}$, for any $n \in \mathbb{N}_*$, with probability at least $1-\delta$ over $\widehat{\mathbb{S}} \sim (\mathcal{E}^n)^m$, for all posteriors $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$, we have

$$\mathrm{kl}(r_{\widehat{\mathbb{S}}}(\rho)\|r_{\mathcal{E}}(\rho)) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{m+1}{\delta}\right], \tag{3.6}$$

$$\text{and} \quad r_{\mathcal{E}}(\rho) \leq r_{\widehat{\mathbb{S}}}(\rho) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{m+1}{\delta}\right]}, \tag{3.7}$$

$$\text{where} \quad r_{\widehat{\mathbb{S}}}(\rho) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{2}\left[1-y_i \mathop{\mathbb{E}}_{h\sim\rho} h(\mathbf{x}_i+\epsilon_j^i)\right].$$

Proof. Deferred to Appendix C.4. ■

It is important to mention that the empirical counterpart of $r_{\mathcal{E}}(\rho)$ is computed on $\widehat{\mathbb{S}}$ which is composed of non identically independently distributed samples, meaning that a "classical" proof technique is not applicable. The trick here is to make use of a result of Ralaivola *et al.* (2010) that provides a *chromatic PAC-Bayesian bound*, *i.e.*, a bound which supports non-independent data. Surprisingly, this theorem states bounds that do not depend on the number of perturbed examples n but only on the number of original examples m. The reason is that the n perturbed examples are inter-dependent (see the proof in Appendix). Note that Equation (3.6) is expressed as a Seeger (2002)'s bound and is tighter but less interpretable than Equation (3.7) expressed as a McAllester (1998)'s bound; these bounds involve the usual trade-off between the empirical risk $r_{\widehat{\mathbb{S}}}(\rho)$ and $\mathrm{KL}(\rho\|\pi)$.

We now state a generalization bound for $a_{\mathcal{E}^n}(\rho)$. Since this value involves a minimum term, we cannot use the same trick as for Theorem 3.3.2. To bypass this issue, we use the TV distance between two "artificial" distributions on $\mathfrak{C}_i$. Given $((\mathbf{x}_i, y_i), \mathfrak{C}_i) \in \widehat{\mathbb{S}}$, let Θ_i be an arbitrary distribution on $\mathfrak{C}_i$, and given $h \in \mathbb{H}$, let θ_i^h be a Dirac distribution on $\mathfrak{C}_i$ such that $\theta_i^h(\epsilon)=1$ if $\epsilon = \mathrm{argmax}_{\epsilon\in\mathfrak{C}_i}\frac{1}{2}\left[1-y_i h(\mathbf{x}_i+\epsilon)\right]$ (*i.e.*, if ϵ is maximizing the linear loss), and 0 otherwise.

Theorem 3.3.3 (PAC-Bayesian Bound on $a_{\mathcal{E}^n}(\rho)$)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any set of voters $\mathbb{H}$, for any prior $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any $n \in \mathbb{N}_*$, with probability at least $1-\delta$ over $\widehat{\mathbb{S}} \sim (\mathcal{E}^n)^m$, for all posteriors $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$, for all $i \in \{1,\ldots,m\}$, for all distributions Θ_i on $\mathfrak{C}_i$ independent from a voter $h \in \mathbb{H}$,

we have

$$a_{\mathcal{E}^n}(\rho) \leq \frac{1}{m} \mathop{\mathbb{E}}_{h\sim\rho} \sum_{i=1}^{m} \max_{\epsilon\in\mathfrak{C}_i} \frac{1}{2}\left(1-y_i h(\mathbf{x}_i+\epsilon)\right) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}}{\delta}\right]} \quad (3.8)$$

$$\leq a_{\widehat{\mathbb{S}}}(\rho) + \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{TV}(\theta_i^h\|\Theta_i) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}}{\delta}\right]}, \quad (3.9)$$

where the empirical risk $a_{\widehat{\mathbb{S}}}(\rho) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\Big[1-\min_{\epsilon\in\mathfrak{C}_i}\Big(y_i\,\mathbb{E}_{h\sim\rho}\,h(\mathbf{x}_i+\epsilon)\Big)\Big]$, and the TV distance $\mathrm{TV}(\theta\|\Theta) = \mathbb{E}_{\epsilon\sim\Theta}\,\frac{1}{2}\left|\left[\frac{\theta(\epsilon)}{\Theta(\epsilon)}\right]-1\right|$.

Proof. Deferred to Appendix C.5. ■

To minimize the true averaged-max risk $a_{\mathcal{E}^n}(\rho)$ from Equation (3.8), we have to minimize a trade-off between $\mathrm{KL}(\rho\|\pi)$ (*i.e.*, how much the posterior weights are close to the prior ones) and the empirical risk $\frac{1}{m}\,\mathbb{E}_{h\sim\rho}\sum_{i=1}^{m}\max_{\epsilon\in\mathfrak{C}_i}\frac{1}{2}\left(1-y_i h(\mathbf{x}_i+\epsilon)\right)$. However, to compute the empirical risk, the loss for each voter and each perturbation has to be calculated and can be time-consuming. With Equation (3.9), we propose an alternative, which can be efficiently optimized using $\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{h\sim\rho}\,\mathrm{TV}(\theta_i^h\|\Theta_i)$ and the empirical averaged-max risk $a_{\widehat{\mathbb{S}}}(\rho)$. Intuitively, Equation (3.9) can be seen as a trade-off between the empirical risk, which reflects the robustness of the majority vote, and two penalization terms: the KL term and the TV term. The KL-divergence $\mathrm{KL}(\rho\|\pi)$ controls how much the posterior ρ can differ from the prior ones π. While the TV term $\mathbb{E}_h\,\mathrm{TV}(\theta_i^h\|\Theta_i)$ controls the diversity of the voters, *i.e.*, the ability of the voters to be fooled on the same adversarial example. From an algorithmic view, an interesting behavior is that the bound of Equation (3.9) stands for all distributions Θ_i on $\mathfrak{C}_i$. This suggests that given $(\mathbf{x}_i, y_i)$, we want to find Θ_i minimizing $\mathbb{E}_{h\sim\rho}\,\mathrm{TV}(\theta_i^h\|\Theta_i)$. Ideally, this term tends to 0 when Θ_i is close[3] to θ_i^h and all voters have their loss maximized by the same perturbation $\epsilon\in\mathfrak{C}_i$.

To learn a well-performing majority vote, one solution is to minimize the right-hand side of the bounds, meaning that we would like to find a good trade-off between a low empirical risk $r_{\widehat{\mathbb{S}}}(\rho)$ or $a_{\widehat{\mathbb{S}}}(\rho)$ and a low divergence between the prior weights and the learned posterior ones $\mathrm{KL}(\rho\|\pi)$. However, the bounds of Equation (3.8) and Equation (3.6) are, in their form, not appealing for optimization. Firstly, Equation (3.6) is not directly optimizable since we upper-bound the $\mathrm{kl}()$ function between the empirical

[3]Since θ_i^h is a Dirac distribution, we have $\mathbb{E}_h\,\mathrm{TV}(\theta_i^h\|\Theta_i) = \frac{1}{2}\left[1-\mathbb{E}_h\,\Theta_i(\epsilon_h^*)+\mathbb{E}_h\sum_{\epsilon\neq\epsilon_h^*}\Theta_i(\epsilon)\right]$, with $\epsilon_h^* = \mathrm{argmax}_{\epsilon\in\mathfrak{C}_i}\,\frac{1}{2}\left[1-y_i h(\mathbf{x}_i+\epsilon)\right]$.

and true risk. To obtain an optimizable bound, we can use the $\overline{\mathrm{kl}}()$ function introduced in Definition 2.3.3. Secondly, from an algorithmic perspective, the prior π is fixed and cannot depend on the learning sample $\mathbb{S}$. To overcome this issue, we propose to use the union bound by considering T priors that can be selected *a posteriori* with $\mathbb{S}$; the two new bounds are presented in the following corollaries.

Corollary 3.3.1 (PAC-Bayesian Bound on $r_{\mathcal{E}}(\rho)$)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any set of voters $\mathbb{H}$, for any $T \in \mathbb{N}_*$, for any priors' set $\{\pi_1, \ldots, \pi_T\} \in \mathbb{M}^*(\mathbb{H})^T$, for any $n \in \mathbb{N}_*$, with probability at least $1-\delta$ over $\widehat{\mathbb{S}} \sim (\mathcal{E}^n)^m$, for all posteriors $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$, for any $\pi \in \{\pi_1, \ldots, \pi_T\} \in \mathbb{M}^*(\mathbb{H})^T$ we have

$$r_{\mathcal{E}}(\rho) \leq \overline{\mathrm{kl}}\left(r_{\widehat{\mathbb{S}}}(\rho) \,\middle|\, \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{T(m+1)}{\delta}\right]\right). \tag{3.10}$$

Proof. Deferred to Appendix C.6. ■

Corollary 3.3.2 (PAC-Bayesian Bound on $a_{\mathcal{E}^n}(\rho)$)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any set of voters $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $n \in \mathbb{N}_*$, with probability at least $1-\delta$ over $\widehat{\mathbb{S}} \sim (\mathcal{E}^n)^m$, for all posteriors $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$, for all $i \in \{1, \ldots, m\}$, for all distributions Θ_i on $\mathfrak{C}_i$ independent from a voter $h \in \mathbb{H}$, we have

$$a_{\mathcal{E}^n}(\rho) \leq a_{\widehat{\mathbb{S}}}(\rho) + \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{TV}(\theta_i^h\|\Theta_i) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2T\sqrt{m}}{\delta}\right]}. \tag{3.11}$$

Proof. Deferred to Appendix C.6. ■

Thanks to Corollaries 3.3.1 and 3.3.2, we now derive an algorithm that minimizes such bounds.

3.3.3 From the Bounds to an Algorithm

We are now able to derive a learning algorithm that minimizes either the bound in Equation (3.10) or Equation (3.11): it is a *self-bounding* algorithm (Freund, 1998). We consider a finite set of voters $\mathbb{H}$ that are differentiable and where each $h \in \mathbb{H}$ is parametrized by a weight vector $\mathbf{w}^h$. Inspired by Masegosa *et al.*, 2020, the voters of $\mathbb{H}$ and the data-dependent prior distribution π is learned from a first learning

set $\mathbb{S}'$ (independent from $\mathbb{S}$); this is a common approach in PAC-Bayes (Parrado-Hernández *et al.*, 2012; Lever *et al.*, 2013; Dziugaite and Roy, 2018; Dziugaite *et al.*, 2021). Then, the posterior distribution is learned from the second learning set $\mathbb{S}$ by minimizing the bounds of Corollaries 3.3.1 and 3.3.2. Concretely, we minimize an objective function that is approximated with a mini-batch $\mathbb{U} \subseteq \mathbb{S}$. The objective function to optimize Equation (3.10) *resp.* Equation (3.11) is defined as

$$G_{\mathcal{U}}(\rho) = \overline{\mathrm{kl}}\left(r_{\mathcal{U}}(\rho) \,\middle|\, \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{T(m+1)}{\delta}\right]\right),$$

$$\textit{resp.} \quad G_{\mathcal{U}}(\rho) = a_{\mathcal{U}}(\rho) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln \tfrac{2T\sqrt{m}}{\delta}\right]}.$$

The TV distance does not appear in the objective function, since we make the choice to set $n{=}1$, *i.e.*, we sample one noise per example. Indeed, when $n{=}1$, the value of the TV distance is 0. Note that, if we had $n > 1$ we would have to minimize it.

We propose now an adversarial training algorithm which is based on a two-step learning procedure presented in Algorithm 3.1. The first step of the algorithm aims at building the set of voters $\mathbb{H}$ and the associated prior π, the second is dedicated to the learning of the majority vote parameter ρ by minimizing the objective function associated to our bound. These steps are presented below.

Attacking the examples. The attacks in Algorithm 3.1 differ from the attack that generates the perturbed set $\widehat{\mathbb{S}}$ (to compute the bound). Indeed, at each iteration (in both steps), we attack an example with the current model while $\widehat{\mathbb{S}}$ is generated with the prior majority vote MV_π (the output of Step 1).

Step 1. Starting from an initial prior π_0 (*e.g.*, the uniform distribution) and an initial set of voters $\mathbb{H}_0$, where each voter h is parametrized by a weight vector $\mathbf{w}_0^h$, the objective of this step is to construct the hypothesis set $\mathbb{H}$ and the prior distribution π to give as input to Step 2 for minimizing the bound. To do so, at each epoch t of Step 1, we learn from $\mathbb{S}'$ an "intermediate" prior π_t on an "intermediate" hypothesis set $\mathbb{H}_t$ consisting of voters h parametrized by the weights $\mathbf{w}_t^h$; note that the optimization is done with respect to $\mathbf{w}_t{=}\{\mathbf{w}_t^h\}_{h\in\mathbb{H}_t}$. At each iteration of the optimizer, for each $(\mathbf{x}, y)$ of the current mini-batch $\mathbb{U}$, we attack the majority vote MV_{π_t} to obtain a perturbed example $\mathbf{x} + \epsilon$. Then, we perform a forward pass in the majority vote with the perturbed examples and update the weights $\mathbf{w}_t$ and the prior π_t according to the linear loss. To sum up, at the end of Step 1, the prior π and the hypothesis set $\mathbb{H}$ constructed for Step 2 are the ones associated to the best epoch $t^* \in \{1, \ldots, T'\}$ that permits to minimize $r_{\mathbb{S}_t}(\mathrm{MV}_{\pi_t})$, where $\mathbb{S}_t$ is the perturbed set obtained by attacking the majority vote MV_{π_t} with the examples of $\mathbb{S}$. Our selection of the prior π with $\mathbb{S}$ may seem like "cheating", but this remains a valid strategy since Equations (3.10) and (3.11) hold

for all prior $\pi \in \{\pi_1, \ldots, \pi_T\}$.

Step 2. Starting from the prior π on $\mathbb{H}$ and the learning set $\mathbb{S}$, we perform the same process as in Step 1 except that the considered objective function corresponds to the desired bound to optimize (denoted $G()$). Note that the "intermediate" priors do not depend on $\mathbb{S}$, since they are learned from $\mathbb{S}'$: the bounds are then valid.

Algorithm 3.1 Average Adversarial Training with Guarantee

Given: disjoint learning samples $\mathbb{S}$ and $\mathbb{S}'$, initial prior π_0 on $\mathbb{H}_0$ (with $\mathbf{w}_0$), the objective function $G()$

Hyperparameters: number of epochs T, T', the attack function

Step 1 – Prior and Voters' Set Construction

for $t \leftarrow 1$ to T' **do**
  $\pi_t \leftarrow \pi_{t-1}$ and $\mathbb{H}_t \leftarrow \mathbb{H}_{t-1}$ ($\mathbf{w}_t \leftarrow \mathbf{w}_{t-1}$)
  for all mini-batch $\mathbb{U} \subseteq \mathbb{S}'$ **do**
    $\mathbb{U} \leftarrow$ Attack MV_{π_t} with the examples $(\mathbf{x}, y)$ in the mini-batch $\mathbb{U}$
    Update π_t with $\nabla_{\pi_t} r_{\mathcal{U}}(\pi_t)$
    Update $\mathbf{w}_t$ with $\nabla_{\mathbf{w}_t} r_{\mathcal{U}}(\pi_t)$
  $\mathbb{S}_t \leftarrow$ Attack MV_{π_t} with the examples of $\mathbb{S}$
$(\pi, \mathbb{H}) \leftarrow (\pi_{t^*}, \mathbb{H}_{t^*})$ with $t^* \leftarrow \mathrm{argmin}_{t' \in \{1,\ldots,t\}} r_{\mathcal{S}_{t'}}(\pi_{t'})$

Step 2 – Bound Minimization

$\rho_0 \leftarrow \pi$
for $t \leftarrow 1$ to T **do**
  $\rho_t \leftarrow \rho_{t-1}$
  for all mini-batch $\mathbb{U} \subseteq \mathbb{S}$ **do**
    $\mathbb{U} \leftarrow$ Attack MV_{π} with the examples $(\mathbf{x}, y)$ in the mini-batch $\mathbb{U}$
    Update ρ_t with $\nabla_{\rho_t} G_{\mathcal{U}}(\rho_t)$
  $\mathbb{S}_t \leftarrow$ Attack MV_{π} with the examples of $\mathbb{S}$
$\rho \leftarrow \rho_{t^*}$ with $t^* \leftarrow \mathrm{argmin}_{t' \in \{1,\ldots,t\}} G_{\mathbb{S}_{t'}}(\rho_{t'})$

3.4 Experimental Evaluation on Differentiable Decision Trees

3.4.1 Experiments

In this section, we empirically illustrate that our PAC-Bayesian framework for adversarial robustness is able to provide generalization guarantees with non-vacuous bounds for the adversarial risk.

Setting. We stand in a white-box setting meaning that the attacker knows the voters set $\mathbb{H}$, the prior distribution π, and the posterior one ρ. The set of voters is composed of 25 differentiable decision trees (KONTSCHIEDER *et al.*, 2016); see Appendix C.7 for more details. We empirically study two attacks with the ℓ_2-norm and ℓ_∞-norm: the Projected Gradient Descent (PGD, MADRY *et al.* (2018)) and the iterative version of FGSM (IFGSM, KURAKIN *et al.* (2017)). We fix the number of iterations at $k{=}20$ and the step size at $\frac{b}{k}$ for PGD and IFGSM (where $b{=}1$ for ℓ_2-norm and $b{=}0.1$ for ℓ_∞-norm). One specificity of our setting is that we deal with the perturbation distribution $\mathcal{B}_{(x,y)}$. However, in order to obtain valid bounds, $\mathcal{B}_{(x,y)}$ must be defined *a priori*. Since the prior π is defined *a priori* as well, $\mathcal{B}_{(x,y)}$ can depend on π. Hence, $\mathcal{B}_{(x,y)}$ boils down to generating a perturbed example $(\mathbf{x}{+}\epsilon, y)$ by attacking the prior majority vote MV_π. Based on this fact, we propose PGD_U and $\mathrm{IFGSM}_\mathsf{U}$, two variants of PGD and IFGSM. To attack an example with PGD_U or $\mathrm{IFGSM}_\mathsf{U}$ we proceed with the following steps.

(i) We attack the prior majority vote MV_π with the attack PGD or IFGSM: we will obtain a first perturbation ϵ';

(ii) We sample n uniform noises $\zeta_1, \ldots, \zeta_n$ between -10^{-2} and $+10^{-2}$;

(iii) We set the i-th perturbation as $\epsilon_i = \epsilon' + \zeta_i$.

Note that, for PGD_U and $\mathrm{IFGSM}_\mathsf{U}$, after one attack we end up with $n{=}100$ perturbed examples. For Algorithm 3.1, when these attacks are used as a defense mechanism, we set $n{=}1$. This makes sound since our adversarial training is iterative, we do not need to sample numerous perturbations for each example: we sample a new perturbation each time the example is forwarded through the decision trees. We additionally consider a naive defense referred to as UNIF that only adds a noise uniformly such that the ℓ_p-norm of the added noise is lower than b.
We study the following scenarios of defense/attack: they correspond to all the pairs (Defense, Attack) belonging to the set $\{\text{—}, \mathrm{UNIF}, \mathrm{PGD}, \mathrm{IFGSM}\}{\times}\{\text{—}, \mathrm{PGD}, \mathrm{IFGSM}\}$ for the baseline, and $\{\text{—}, \mathrm{UNIF}, \mathrm{PGD}_\mathsf{U}, \mathrm{IFGSM}_\mathsf{U}\}{\times}\{\text{—}, \mathrm{PGD}_\mathsf{U}, \mathrm{IFGSM}_\mathsf{U}\}$, where "—" means that we do not defend, *i.e.*, the attack returns the original example (note that PGD_U

and IFGSM_U when "Attack without U" refers to PGD and IFGSM for computing the classical adversarial risk).

Datasets and algorithm description. We perform our experiment on six binary classification tasks from MNIST (LeCun *et al.*, 1998) (1vs7, 4vs9, 5vs6) and Fashion MNIST (Xiao *et al.*, 2017) (Coat vs Shirt, Sandal vs Ankle Boot, Top vs Pullover). We decompose the learning set into two disjoint subsets $\mathbb{S}'$ of around $7,000$ examples (to learn the prior and the voters) and $\mathbb{S}$ of exactly $5,000$ examples (to learn the posterior). We keep as test set $\mathbb{T}$ the original test set that contains around $2,000$ examples. Moreover, we need a perturbed test set, denoted by $\widehat{\mathbb{T}}$, to compute our averaged(-max) adversarial risks. Depending on the scenario, $\widehat{\mathbb{T}}$ is constructed from $\mathbb{T}$ by attacking the prior model MV_π with PGD_U or IFGSM_U with $n{=}100$. We run our Algorithm 3.1 for Equation (3.6) (Theorem 3.3.2), respectively Equation (3.9) (Theorem 3.3.3), and we compute our risk $\text{R}_{\widehat{\mathcal{T}}}(\text{MV}_\rho)$, respectively $\text{A}_{\widehat{\mathcal{T}}}(\text{MV}_\rho)$, the bound value and the usual adversarial risk associated to the model learned $\text{A}_{\mathcal{T}}(\text{MV}_\rho)$. Note that, during the evaluation of the bounds, we have to compute our relaxed adversarial risks $\text{R}_{\widehat{\mathcal{S}}}(\text{MV}_\rho)$ and $\text{A}_{\widehat{\mathcal{S}}}(\text{MV}_\rho)$ on $\mathbb{S}$. For Step 1, the initial prior P_0 is fixed to the uniform distribution, the initial set of voters $\mathbb{H}_0$ is constructed with weights initialized with Xavier Initializer (Glorot and Bengio, 2010) and bias initialized at 0. During Step 2, to optimize the bound, we fix the confidence parameter $\delta{=}0.05$, and we consider two settings for $\mathbb{H}$: the set $\mathbb{H}$ as it is output by Step 1, and the set $\mathbb{H}^{\text{SIGN}} = \{h'(\cdot) = \text{sign}(h(\cdot)) \,|\, h \in \mathbb{H}\}$ for which the theoretical results are still valid. Note that for all attacks on the majority votes with $\mathbb{H}^{\text{SIGN}}$, in order to be differentiable with respect to the input, we remove the $\text{sign}()$ function on the voters' outputs during the attacks. For the two steps, we use Adam optimizer (Kingma and Ba, 2015) for $T{=}T'{=}20$ epochs with a learning rate at 10^{-2} and a batch size at 64.

Analysis of the results. For the sake of readability, we exhibit the detailed results for one task (MNIST:1vs7) and all the pairs (Defense,Attack) with ℓ_2-norm in Table 3.1, and we report in Figure 3.2 the influence of the TV term in the bound of Theorem 3.3.3 (Equation (3.9)). The detailed results on the other tasks are reported in Appendix C.8. We provide in Figure 3.3 an overview of the results we obtained on all the tasks for the pairs (Defense,Attack) where "Defense=Attack" and with $\mathbb{H}^{\text{SIGN}}$.

First of all, from Table 3.1 the bounds of Theorem 3.3.2 are tighter than the ones of Theorem 3.3.3: this is an expected result since we showed that the averaged-max adversarial risk $\text{A}_{\mathcal{E}^n}(\text{MV}_\rho)$ is more pessimistic than its averaged counterpart $\text{R}_{\mathcal{E}}(\text{MV}_\rho)$. Note that the bound values of Equation (3.8) are tighter than the ones of Equation (3.9). This is expected since Equation (3.8) is a lower bound on Equation (3.9).

Table 3.1. *Test risks and bounds for MNIST:1vs7 with $n{=}100$ perturbations for all pairs (Defense,Attack) with the two voters' set $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$. The results in* ***bold*** *correspond to the best values between results for $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$. To quantify the gap between our risks and the classical definition we put in italic the risk of our models against the classical attacks: we replace PGD_U and IFGSM_U by PGD or IFGSM (i.e., we did not sample from the uniform distribution). Since Eq.* (3.9) *upper-bounds Eq.* (3.8) *thanks to the TV term, we compute the two bound values of Theorem 3.3.3.*

ℓ_2-norm		Algo. 3.1 with Eq. (3.6)						Algo. 3.1 with Eq. (3.9)							
$b=1$		Attack without U						Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.005*	*.005*	.005	.005	**.017**	.019	*.005*	*.005*	.005	.005	**.099**	0.100	**.099**	.100
—	PGD_U	***.245***	*.255*	**.263**	.276	.577	**.448**	*.315*	***.313***	**.325**	.326	**.801**	1.667	.684	**.515**
—	IFGSM_U	***.084***	*.086*	**.066**	.080	**.170**	.185	*.117*	***.113***	**.106**	.110	**.356**	1.431	.286	**.251**
UNIF	—	*.005*	*.005*	.005	.005	**.018**	.019	*.005*	*.005*	.005	.005	**.099**	0.100	**.099**	.100
UNIF	PGD_U	*.151*	***.146***	**.151**	.158	.355	**.292**	*.183*	***.178***	.190	**.189**	**.531**	1.620	.454	**.355**
UNIF	IFGSM_U	*.063*	***.061***	**.031**	.035	**.088**	.114	*.071*	***.070***	.056	**.054**	**.248**	1.405	.200	**.186**
PGD_U	—	***.006***	*.007*	**.006**	.007	**.023**	.024	***.006***	*.007*	**.006**	.007	**.102**	0.103	**.102**	.103
PGD_U	PGD_U	***.028***	*.030*	**.021**	.025	.065	**.064**	***.028***	*.029*	**.025**	.028	**.143**	1.389	.137	**.136**
PGD_U	IFGSM_U	***.021***	*.022*	**.013**	.016	**.043**	.045	*.022*	*.022*	**.018**	.019	**.125**	1.362	.121	**.119**
IFGSM_U	—	***.006***	*.007*	**.006**	.007	**.019**	.021	***.006***	*.007*	**.006**	.007	**.100**	0.102	**.100**	.102
IFGSM_U	PGD_U	***.040***	*.041*	**.033**	.035	**.086**	.094	*.040*	***.039***	.040	**.038**	**.184**	1.368	.166	**.163**
IFGSM_U	IFGSM_U	***.021***	*.022*	**.013**	.014	**.039**	.049	***.021***	*.022*	**.018**	.021	**.131**	1.329	**.122**	.123

Second, the bounds with $\mathbb{H}^{\text{SIGN}}$ are all informative (lower than 1) and give insightful guarantees for our models. For Theorem 3.3.3 (Equation (3.9)) with $\mathbb{H}$, while the risks are comparable to the risks obtained with $\mathbb{H}^{\text{SIGN}}$, the bound values are greater than 1, meaning that we have no more guarantee on the model learned. As we can observe in Figure 3.2, this is due to the TV term involved in the bound. Considering $\mathbb{H}^{\text{SIGN}}$ when optimizing $\mathsf{A}()$ helps to control the TV term. Even if the bounds are non-vacuous for Theorem 3.3.2 with $\mathbb{H}$, the best models with the best guarantees are obtained with $\mathbb{H}^{\text{SIGN}}$. This is confirmed by the columns $\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$ that are always worse than $\mathsf{R}_{\widehat{\mathcal{T}}}(\mathsf{MV}_\rho)$ and mostly worse than $\mathsf{A}_{\widehat{\mathcal{T}}}(\mathsf{MV}_\rho)$ with $\mathbb{H}^{\text{SIGN}}$. The performance obtained with $\mathbb{H}^{\text{SIGN}}$ can be explained by the fact that the sign "saturates" the output of the voters which makes the majority vote more robust to noises. Thus, we focus the rest of the analysis on results obtained with $\mathbb{H}^{\text{SIGN}}$.

Third, we observe that the naive defense UNIF is able to improve the risks $\mathsf{R}_{\widehat{\mathcal{T}}}(\mathsf{MV}_\rho)$ and $\mathsf{A}_{\widehat{\mathcal{T}}}(\mathsf{MV}_\rho)$, but the improvement with the defenses based on PGD_U and IFGSM_U is much more significant specifically against a PGD_U attack (up to 13 times better). We observe the same phenomenon for both bounds (Theorems 3.3.2 and 3.3.3). This is an interesting fact because this behavior confirms that we are able to learn models

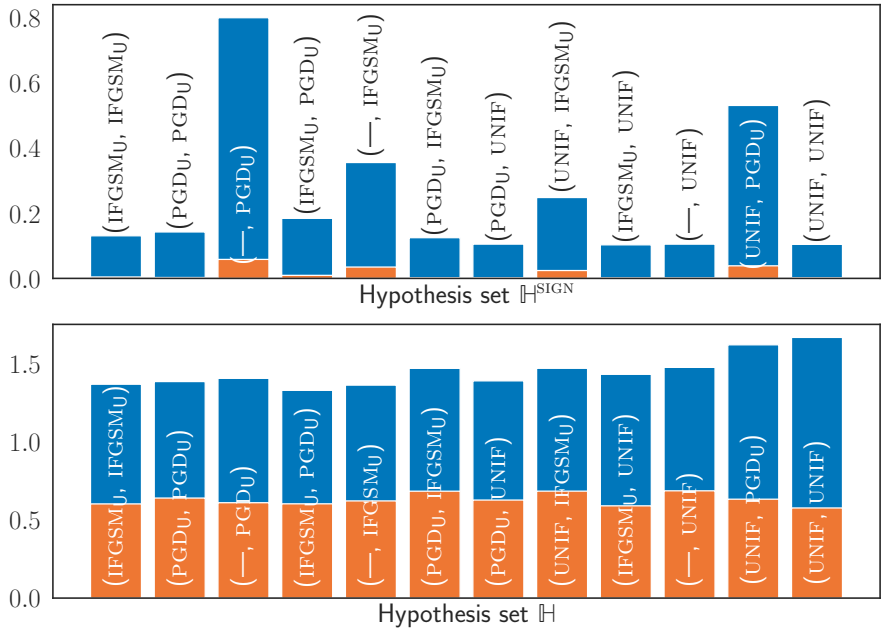


Figure 3.2. *Visualization of the impact of the TV term in Equation* (3.9)*. The top, respectively the bottom, bar plot show the bounds for the set of voters* $\mathbb{H}^{\text{SIGN}}$*, respectively* $\mathbb{H}$*. We plot the bounds for all the scenarios of Table 3.1 that use the* TV *distance, i.e., all except the pairs* $(\cdot, —)$*. In orange we represent the value of the TV term while in blue we represent all the remaining terms of the bound.*

that are robust against the attacks tested with theoretical guarantees.
Lastly, from Figure 3.3 and Table 3.1, it is important to notice that the gap between the classical risk and our relaxed risks is small, meaning that our relaxation are not too optimistic. Despite the pessimism of the classical risk $\mathrm{A}_{\mathcal{T}}(\mathrm{MV}_\rho)$, it remains consistent with our bounds, *i.e.*, it is lower than the bounds. In other words, in addition to giving upper bounds for our risks $\mathrm{R}_{\widehat{\mathcal{T}}}(\mathrm{MV}_\rho)$ and $\mathrm{A}_{\widehat{\mathcal{T}}}(\mathrm{MV}_\rho)$, our bounds give non-vacuous guarantees on the classical risks $\mathrm{A}_{\mathcal{T}}(\mathrm{MV}_\rho)$.

3.5 Conclusion and Summary

To the best of our knowledge, our work is the first one that studies adversarial robustness through the PAC-Bayesian theory for the ρ-weighted majority vote. We have started by formalizing a new adversarial robustness setting (for binary classification) with our new averaged risks. This formulation allowed us to derive PAC-Bayesian generalization bounds on the majority vote's adversarial risk. We illustrated the usefulness of this setting on the training of (differentiable) decision trees. The main objective

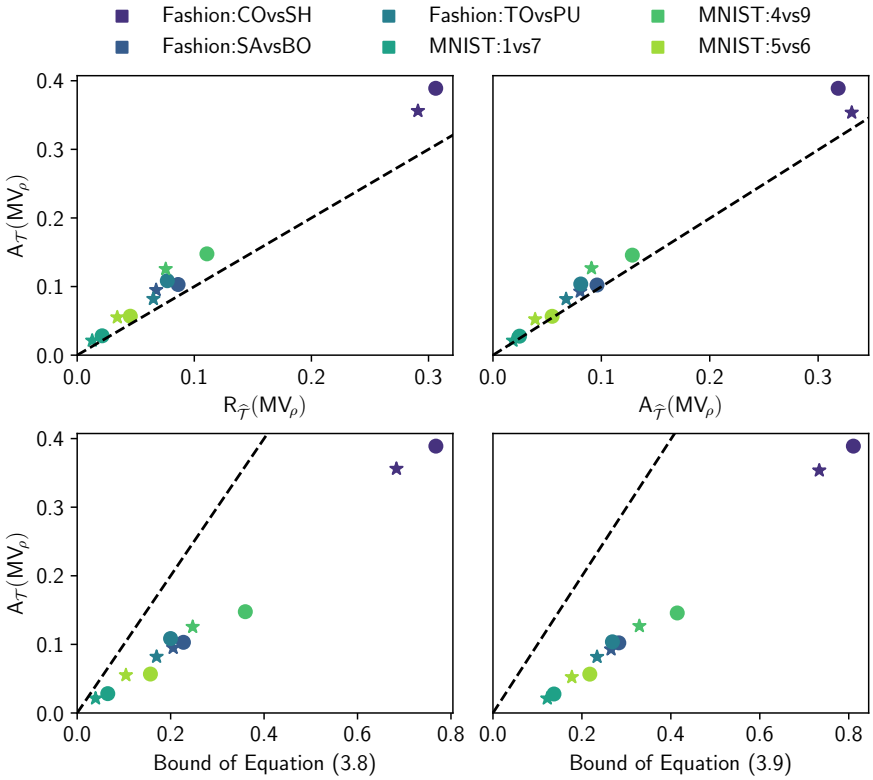


Figure 3.3. *Visualization of the risk and bound values for "Defense=Attack" when the set of voters is* $\mathbb{H}^{\text{SIGN}}$*. Results obtained with the* PGD_U*, respectively* IFGSM_U*, defense are represented by a star* ★*, respectively a circle* ● *(reminder:* $A_{\mathcal{T}}(MV_\rho)$ *is computed with a* PGD*, respectively* IFGSM*, attack). The dashed line corresponds to bisecting line* $y{=}x$*. For* $R_{\widehat{\mathcal{T}}}(MV_\rho)$ *and* $A_{\widehat{\mathcal{T}}}(MV_\rho)$*, the closer the datasets are to the bisecting line, the more accurate our relaxed risk is compared to the classical adversarial risk* $A_{\mathcal{T}}(MV_\rho)$*. For the bounds, the closer the datasets are to the bisecting line, the tighter the bound.*

of this work was to provide some theoretical guarantees for adversarial training. Our aim was not to improve directly the performance of the state of the art which would require a dedicated work.

One of the limitation of this work is that this PAC-Bayesian analysis holds for a majority vote only. One perspective is then to extend this work for other classifiers such as neural networks. To do so, we can leverage the disintegrated PAC-Bayesian generalization bounds (introduced in Section 2.4) to bound the adversarial risk of a *single classifier* belonging to $\mathbb{H}$ and sampled from ρ. Another perspective of this work is to continue analyzing the adversarial true risk of the majority vote. Indeed, the C-Bound (LACASSE *et al.*, 2006) or the joint error (MASEGOSA *et al.*, 2020) (introduced in Section 2.2.2) adapted to our averaged risks might be a better choice to obtain a self-bounding algorithm since the it is more precise than twice the Gibbs risk (see Theorem 2.2.4). Moreover, thanks to the $\frac{1}{2}$-margin of LAVIOLETTE *et al.* (2017) (recalled in Definition 2.2.4), the multi-class case can be considered.

In the next chapters, we consider the classical supervised setting, *i.e.*, where no inputs' perturbations are added. For instance, Chapters 4 and 5 focus on deriving new learning algorithms for the majority vote in the classical supervised setting. More precisely, we derive, in Chapter 4, the first self-bounding algorithms based on the minimization of the PAC-Bayesian C-Bound (*i.e.* PAC-Bayesian bounds on the C-Bound). Chapter 5 presents a algorithm to minimize the risk of a PAC-Bayesian stochastic majority vote (where the distribution ρ are sampled from another distribution).

Self-Bounding Algorithms for the Majority Vote

4

This chapter is based on the following paper

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. Self-bounding Majority Vote Learning Algorithms by the Direct Minimization of a Tight PAC-Bayesian C-Bound. *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. (2021a)

Contents

4.1 Introduction **102**
4.2 Setting **103**
4.3 State of the Art: PAC-Bayesian Bounds for the Majority Vote **104**
4.3.1 PAC-Bayesian Bound on the Gibbs Risk 104
4.3.2 PAC-Bayesian Bound on the Joint Error 105
4.3.3 PAC-Bayesian C-Bound of Roy *et al.* 105
4.3.4 PAC-Bayesian C-Bound of Germain *et al.* 106
4.3.5 PAC-Bayesian C-Bound of Lacasse *et al.* 107
4.4 Contribution: Algorithms based on the PAC-Bayesian C-Bounds **108**
4.4.1 Algorithm based on Equation (4.2) 108
4.4.2 Algorithm based on Equation (4.3) 110
4.4.3 Algorithm based on Theorem 4.3.5 111
4.5 Experiments **114**
4.5.1 Setting 114
4.5.2 Analysis of the Results 122
4.6 Conclusion and Summary **123**

Abstract

As we have seen in Chapter 2, the C-Bound is an insightful upper bound on the risk of a majority vote classifier. Learning algorithms in the literature minimize the empirical version of the C-Bound, instead of explicit PAC-Bayesian generalization bounds. In this chapter, we derive self-bounding majority vote learning algorithms to directly optimize PAC-Bayesian guarantees on the C-Bound. Our algorithms based on gradient descent are scalable and lead to accurate predictors paired with non-vacuous guarantees.

4.1 Introduction

In this chapter, we introduce new learning algorithms for the majority vote in the context of supervised classification. The goal of this algorithm is to minimize the true risk of the majority vote. To do so, one way to minimize such a risk is to minimize the empirical C-Bound (Breiman, 2001; Lacasse *et al.*, 2006) introduced in Section 2.2.2 and estimated on the learning sample $\mathbb{S}$. This bound has the advantage of involving the performance of the individual voters and the diversity in the voters' set. Indeed, these elements are important when one learns a combination (Dietterich, 2000; Kuncheva, 2014). A good majority vote is made up of voters that are "sufficiently diverse".

Previous algorithms have been developed to minimize the *empirical* C-Bound such as `MinCq` (Roy *et al.*, 2011), `P-MinCq` (Bellet *et al.*, 2014), `CqBoost` (Roy *et al.*, 2016), or `CB-Boost` (Bauvin *et al.*, 2020). Roy *et al.* (2011) first proposed `MinCq` which consist in minimizing a quadratic problem to learn a majority vote. `MinCq` considers a specific voters' set to regularize the minimization process; the algorithm `P-MinCq` generalizes `MinCq` by allowing prior distributions different from the uniform one. One drawback of `MinCq` and `P-MinCq` is that the optimization problem is not scalable to large datasets. Lately, Bauvin *et al.* (2020) proposed `CB-Boost` that minimizes in a boosting-based procedure with the advantage to be more scalable while obtaining sparser majority vote. However, since both `MinCq` and `CB-Boost` minimize the empirical C-Bound, the PAC-Bayesian generalization bound associated with their learned majority vote predictors can be vacuous. Note that `CB-Boost` has been proposed to improve another algorithm called `CqBoost` (Roy *et al.*, 2016). Despite being empirically efficient and justified by theoretical analyses based on the C-Bound, all these methods minimize the empirical C-Bound and not directly a PAC-Bayesian generalization bound on the C-Bound. This can lead to vacuous generalization bound values and, thus, to poor risk certificates. When it comes to deriving a learning algorithm that directly minimizes a PAC-Bayesian bound, it is mentioned in the literature that optimizing a PAC-Bayesian bound on the C-bound is not trivial (Lorenzen *et al.*, 2019; Masegosa *et al.*, 2020). This underlines the need for other majority vote learning algorithms based on the C-Bound, which motivates our contributions of Section 4.4.

We cover in this chapter three different PAC-Bayesian viewpoints on generalization bounds for the C-Bound (Seeger, 2002; McAllester, 2003; Lacasse *et al.*, 2006). We derive three algorithms from these three views to optimize generalization bounds on the C-Bound. By doing so, we achieve *self-bounding algorithms* (Freund, 1998): the predictor returned by the learner comes with a statistically valid risk upper bound. Importantly, our algorithms rely on fast gradient descent procedures. As far as we know, this is the first work that proposes both efficient algorithms for C-Bound

optimization and non-trivial risk-bound values.

We provide all the proofs in Appendix D for completeness.

4.2 Setting

We stand in supervised classification by following Chapter 2. In this context, let $\mathbb{X} \subseteq \mathbb{R}^d$ be a d-dimensional input space, and $\mathbb{Y}$ the label space defined by $\mathbb{Y} = \{-1, +1\}$ (in binary classification) or $\mathbb{Y} = \{1, 2, \ldots, l\}$ (in multi-class classification). We assume an unknown data distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$ and a learning sample $\mathbb{S}=\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ where each example $(\mathbf{x}_i, y_i)$ is drawn *i.i.d.* from $\mathcal{D}$; we denote by $\mathbb{S} \sim \mathcal{D}^m$ the random draw of such a sample. Given $\mathbb{H}$ a hypothesis set constituted by voters $h : \mathbb{X}\to\mathbb{Y}$, and $\mathbb{S}$, the learner aims to find a weighted combination of the voters from $\mathbb{H}$; a distribution models the weights on $\mathbb{H}$. To learn such a combination in the PAC-Bayesian framework, we assume a *prior* distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, and—after the observation of $\mathbb{S}$—we learn a *posterior* distribution $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$. More precisely, we aim to learn a well-performing classifier that is expressed as a *ρ-weighted majority vote* MV_ρ defined as

$$\forall \mathbf{x} \in \mathbb{X}, \quad \mathsf{MV}_\rho(\mathbf{x}) \triangleq \operatorname*{argmax}_{y' \in \mathbb{Y}} \mathop{\mathbb{P}}_{h\sim\rho} [h(\mathbf{x}) = y'] = \operatorname*{argmax}_{y' \in \mathbb{Y}} \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{I}\,[h(\mathbf{x}) = y']\,.$$

We thus want to learn MV_ρ that commits as few errors as possible on unseen data from $\mathcal{D}$, *i.e.*, that leads to a low true risk $\mathsf{R}_\mathcal{D}(\mathsf{MV}_\rho)$ under the 01-loss defined as

$$\mathsf{R}_\mathcal{D}(\mathsf{MV}_\rho) \triangleq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}\,[\mathsf{MV}_\rho(\mathbf{x}) \neq y] = \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}} [\mathsf{MV}_\rho(\mathbf{x}) \neq y]\,.$$

Since the majority vote's risk is not appealing for optimization (because its gradient is zero everywhere), some surrogates have been introduced (see Section 2.2.2). For instance, the Gibbs risk (Definition 2.2.5) is the average risk of the voters and is defined by

$$r_\mathcal{D}(\rho) \triangleq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{I}\,[h(\mathbf{x}) \neq y] = \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D},h\sim\rho} [h(\mathbf{x}) \neq y]\,.$$

Unlike the Gibbs risk, the disagreement (Definition 2.2.7) defined as

$$d_\mathcal{D}(\rho) \triangleq 2 \cdot \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{h\sim\rho} \mathop{\mathbb{E}}_{h'\sim\rho} \mathrm{I}\,[h(\mathbf{x}) \neq y]\,\mathrm{I}\,[h'(\mathbf{x}) = y]\,,$$

takes the diversity of the voters into account. Moreover, the joint error (Definition 2.2.6) can be seen as a trade-off between these two quantities. It is defined as

$$\begin{aligned} e_\mathcal{D}(\rho) &\triangleq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{h\sim\rho} \mathop{\mathbb{E}}_{h'\sim\rho} \mathrm{I}\left[h(\mathbf{x}) \neq y\right] \mathrm{I}\left[h'(\mathbf{x}) \neq y\right] \\ &= r_\mathcal{D}(\rho) - \frac{1}{2} d_\mathcal{D}(\rho). \end{aligned}$$

By combining these surrogates, one can prove an upper-bound on the majority vote true risk called the C-Bound (Theorem 2.2.3) and defined as

$$\begin{aligned} \mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) &\leq 1 - \frac{\left(1 - 2r_{\mathcal{D}}(\rho)\right)^2}{1 - 2d_{\mathcal{D}}(\rho)} \\ &= 1 - \frac{\Big(1 - \left[2e_{\mathcal{D}}(\rho) + d_{\mathcal{D}}(\rho)\right]\Big)^2}{1 - 2d_{\mathcal{D}}(\rho)}. \end{aligned}$$

However, these surrogates and the C-Bound are not computable because the distribution $\mathcal{D}$ is considered *unknown*. Hence, we need to use PAC-Bayesian generalization bounds in order to upper-bound the majority vote's true risk with a C-Bound based on the empirical counterparts of these surrogates. Combined with the C-Bound, the PAC-Bayesian theory offers a natural way to analyze the risk of the majority vote. The principal PAC-Bayesian bounds for the majority vote are recalled in the next section.

4.3 State of the Art: PAC-Bayesian Bounds for the Majority Vote

This section recalls different PAC-Bayesian bounds upper-bounding the majority vote's true risk. In particular, we remind two PAC-Bayesian bounds used based on two surrogates from Section 2.2.2: Gibbs risk $r_{\mathcal{D}}(\rho)$ and the joint error $e_{\mathcal{D}}(\rho)$. Additionally, we recall three PAC-Bayesian bound on the C-Bound, that we call *PAC-Bayesian C-Bound*, which is key in our contribution of this chapter. Note that the PAC-Bayesian C-Bounds were initially developed for the binary setting but the extension for the multi-class is direct with the $\frac{1}{2}$-margin of Laviolette *et al.* (2017).

4.3.1 PAC-Bayesian Bound on the Gibbs Risk

One PAC-Bayesian bound, originally derived by Germain *et al.* (2015), is based on the Gibbs risk $r_{\mathcal{D}}(\rho)$. It is recalled in the following theorem.

Theorem 4.3.1 (PAC-Bayesian Bound Based on the Gibbs Risk)**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any $\delta \in (0, 1]$, with probability at least $1-\delta$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$ we have

$$\begin{aligned} &\forall \rho \in \mathbb{M}(\mathbb{H}), \quad r_{\mathcal{D}}(\rho) \leq \overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right] \right), \\ &\text{and } \forall \rho \in \mathbb{M}(\mathbb{H}), \quad \mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq 2 \left[\overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right] \right) \right], \end{aligned} \tag{4.1}$$

where $\overline{\mathrm{kl}}(q|\tau) \triangleq \max \left\{ p \in (0,1) \;\middle|\; \mathrm{kl}(q\|p) \leq \tau \right\}$ (see Section 2.3.1.3).

Proof. Deferred to Appendix D.1. ■

However, since the Gibbs risk does not consider the voters' correlation, the majority votes obtained by minimizing this bound do not perform well in practice. Hence, other PAC-Bayesian bounds have therefore been derived to address this issue.

4.3.2 PAC-Bayesian Bound on the Joint Error

Another PAC-Bayesian bound based on the joint error $e_{\mathcal{D}}(\rho)$ can be derived (Germain *et al.*, 2015, Theorem 25). Compared to Theorem 4.3.1, the bound of Theorem 4.3.2 takes better into account the voters' diversity (see Section 2.2.2).

Theorem 4.3.2 (PAC-Bayesian Bound Based on the Joint Error)**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$ we have

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad e_{\mathcal{D}}(\rho) \leq \overline{\mathrm{kl}}\left(e_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right] \right),$$

$$\text{and} \quad \forall \rho \in \mathbb{M}(\mathbb{H}), \quad \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq 4 \left[\overline{\mathrm{kl}}\left(e_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right] \right) \right].$$

Proof. Deferred to Appendix D.2. ■

Note that a looser bound based on the $\mathrm{kl}()$ relaxation of Thiemann *et al.* (2017) is presented by Masegosa *et al.* (2020). However, from Theorem 2.2.4, we know that there is a tighter bound on the majority vote's risk: the C-Bound (Breiman, 2001; Lacasse *et al.*, 2006).

4.3.3 PAC-Bayesian C-Bound of Roy *et al.*

PAC-Bayesian bounds can be used jointly with the C-Bound to obtain a computable bound on the majority vote's true risk; we call such a bound a *PAC-Bayesian C-Bound*. The most intuitive and interpretable PAC-Bayesian C-Bound has been derived by Roy *et al.* (2016) and Laviolette *et al.* (2017). The first proof of this PAC-Bayesian bound has been developed by (Roy *et al.*, 2016) in the binary setting; it has

been extended to the multi-class setting by LAVIOLETTE *et al.* (2017, Theorem 3). It consists in upper-bounding separately the Gibbs risk $r_{\mathcal{D}}(\rho)$ and the disagreement $d_{\mathcal{D}}(\rho)$ with the MCALLESTER's PAC-Bayesian bound (Theorem 2.3.2). This intuitive PAC-Bayesian bound is recalled in the following theorem.

Theorem 4.3.3 (PAC-Bayesian C-Bound of ROY *et al.* (2016))**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$ we have for all $\rho \in \mathbb{M}(\mathbb{H})$

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq \underbrace{1 - \frac{\left(1 - 2\min\left[\frac{1}{2}, r_{\mathcal{S}}(\rho) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}\right]\right)^2}{1 - 2\max\left[0, d_{\mathcal{S}}(\rho) - \sqrt{\frac{1}{2m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}\right]}}_{\triangleq\ C^{\mathtt{M}}_{\mathcal{S}}(\rho)} . \tag{4.2}$$

Proof. Deferred to Appendix D.3. ■

While there is no algorithm that directly minimizes Theorem 4.3.3, this kind of interpretable bound can be seen as a justification of the optimization of $r_{\mathcal{S}}(\rho)$ and $d_{\mathcal{S}}(\rho)$ in the empirical C-Bound such as for `MinCq` (ROY *et al.*, 2011) or `CB-Boost` (BAUVIN *et al.*, 2020). In Section 4.4.1, we derive the first algorithm to directly minimize it.
However, this PAC-Bayesian C-Bound can have a severe disadvantage with a small m and a Gibbs risk close to $\frac{1}{2}$: even for a $\mathrm{KL}(\rho\|\pi)$ close to 0, the value of the PAC-Bayesian C-Bound will be close to 1. To overcome this drawback, one solution is to follow another tighter PAC-Bayesian bound, the one proposed by SEEGER (2002) (Theorem 2.3.4). Actually, we further recall two bounds based on this approach: the first one in Theorem 4.3.4 involves the Gibbs risk $r_{\mathcal{S}}(\rho)$ and the disagreement $d_{\mathcal{S}}(\rho)$ (as Theorem 4.3.3) and the second one in Theorem 4.3.5 involves the joint error $e_{\mathcal{S}}(\rho)$ and the disagreement $d_{\mathcal{S}}(\rho)$.

4.3.4 PAC-Bayesian C-Bound of Germain *et al.*

The PAC-Bayesian generalization bounds based on the SEEGER's approach are known to produce tighter bounds. As for Theorem 4.3.3, the result below bounds independently the Gibbs risk $r_{\mathcal{D}}(\rho)$ and the disagreement $d_{\mathcal{D}}(\rho)$; see the PAC-Bound 1 of GERMAIN *et al.* (2015). Note that GERMAIN *et al.* (2015) proved the bound for the binary setting but the multi-class setting is also handled. It is recalled in the following theorem.

Theorem 4.3.4 (PAC-Bayesian C-Bound of GERMAIN *et al.* (2015))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi\in\mathbb{M}^*(\mathbb{H})$, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over the random choice of $\mathbb{S}\sim\mathcal{D}^m$ we have for all $\rho\in\mathbb{M}(\mathbb{H})$

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho)\leq 1-\underbrace{\frac{\left(1-2\min\left[\frac{1}{2},\overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho)\;\middle|\;\frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}\right]\right)\right]\right)^2}{1-2\max\left[0,\underline{\mathrm{kl}}\left(d_{\mathcal{S}}(\rho)\;\middle|\;\frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}\right]\right)\right]}}_{\triangleq\; C^{\mathbb{S}}_{\mathcal{S}}(\rho)}. \tag{4.3}$$

Proof. Deferred to Appendix D.4. ■

Note that this PAC-Bayesian C-Bound is tighter than Equation (4.2) because the SEEGER's PAC-Bayesian bound is tighter than the one of MCALLESTER's one (see Section 2.3). However, one drawback of this PAC-Bayesian C-Bound is that the Gibbs risk $r_{\mathcal{D}}(\rho)$ and the disagreement $d_{\mathcal{D}}(\rho)$ are upper-bounded independently.

4.3.5 PAC-Bayesian C-Bound of Lacasse *et al.*

LACASSE *et al.* (2006) proposed to bound simultaneously the joint error $e_{\mathcal{D}}(\rho)$ and the disagreement $d_{\mathcal{D}}(\rho)$. Here, to compute the bound, we need to find the worst C-Bound value that can be obtained with a couple of joint error and disagreement denoted by (e,d) belonging to the set $\mathbb{A}_{\mathcal{S}}(\rho)$ that is defined by

$$\mathbb{A}_{\mathcal{S}}(\rho)=\left\{(e,d)\;\middle|\;\mathrm{kl}\left(e_{\mathcal{S}}(\rho),d_{\mathcal{S}}(\rho)\|e,d\right)\leq\frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{m}+m}{\delta}\right],\right.$$
$$\left.d\leq 2\sqrt{e}-2e\,,\;2e+d<1\right\},$$
$$\text{where}\quad \mathrm{kl}(q_1,q_2\|p_1,p_2)=q_1\ln\frac{q_1}{p_1}+q_2\ln\frac{q_2}{p_2}+(1-q_1-q_2)\ln\frac{1-q_1-q_2}{1-p_1-p_2}.$$

Based on $\mathbb{A}_{\mathcal{S}}(\rho)$, LACASSE *et al.* (2006) derive the following PAC-Bayesian C-Bound.

Theorem 4.3.5 (PAC-Bayesian C-Bound of LACASSE *et al.* (2006))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi\in\mathbb{M}^*(\mathbb{H})$,

for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$ we have for all $\rho \in \mathbb{M}(\mathbb{H})$

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq \sup_{(e,d)\in\mathbb{A}_{\mathcal{S}}(\rho)} \left[1 - \frac{\left(1-(2e+d)\right)^2}{1-2d}\right],$$

$$\text{where} \quad \mathbb{A}_{\mathcal{S}}(\rho) = \left\{(e,d) \;\middle|\; \mathrm{kl}\left(e_{\mathcal{S}}(\rho), d_{\mathcal{S}}(\rho) \| e, d\right) \leq \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln \tfrac{2\sqrt{m}+m}{\delta}\right], \right.$$
$$\left. d \leq 2\sqrt{e} - 2e\,,\; 2e+d < 1 \right\}.$$

Proof. Deferred to Appendix D.5. ■

This PAC-Bayesian C-Bound can be more challenging to compute: it requires to solve a (convex) optimization problem to obtain a bound value.

4.4 Contribution: Algorithms based on the PAC-Bayesian C-Bounds

In this section, we present three self-bounding algorithms minimizing directly the PAC-Bayesian C-Bounds introduced previously.

4.4.1 Algorithm based on Equation (4.2)

Algorithm 4.1 Minimization of Equation (4.2) by Stochastic Gradient Descent

Given: learning sample $\mathbb{S}$, prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, the objective function $G^{\mathtt{M}}_{\mathcal{S}}(\rho)$
Hyperparameters: number of iterations T
$\rho \leftarrow \pi$
for $t \leftarrow 1$ to T **do**
  for all mini-batches $\mathbb{U} \subseteq \mathbb{S}$ **do**
    $\rho \leftarrow$ Update ρ with $G^{\mathtt{M}}_{\mathcal{U}}(\rho)$ by gradient descent[1]
return ρ

[1] The update of ρ can be done with a vanilla gradient descent or with the update of another algorithm like Adam (Kingma and Ba, 2015) or COCOB (Orabona and Tommasi, 2017).

We derive in Algorithm 4.1 a method to directly minimize the PAC-Bayesian C-Bound of Theorem 4.3.3 by stochastic gradient descent. An important aspect of the optimization is that if $r_{\mathcal{S}}(\rho)+\sqrt{\frac{1}{2m}[\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}]} \geq \frac{1}{2}$, the gradient of the numerator in $C^{\mathtt{M}}_{\mathcal{S}}(\rho)$ with respect to ρ is 0 which makes the optimization impossible. Hence, we aim to minimize the following constraint optimization problem:

$$\min_{\rho\in\mathbb{M}(\mathbb{H})} \underbrace{\left\{1-\frac{\left(1-2\min\left[\frac{1}{2}, r_{\mathcal{S}}(\rho)+\sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}\right]}\right]\right)^2}{1-2\max\left[0, d_{\mathcal{S}}(\rho)-\sqrt{\frac{1}{2m}\left[2\,\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}\right]}\right]}\right\}}_{\triangleq\ C^{\mathtt{M}}_{\mathcal{S}}(\rho)}$$
$$\text{s.t}\quad r_{\mathcal{S}}(\rho)+\sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}\right]} \leq \frac{1}{2}.$$

From this formulation, we deduce a non-constrained optimization problem:

$$\min_{\rho\in\mathbb{M}(\mathbb{H})} \left\{ C^{\mathtt{M}}_{\mathcal{S}}(\rho) + \mathrm{B}\left[r_{\mathcal{S}}(\rho)+\sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}\right]}-\frac{1}{2}\right]\right\},$$

where $\mathrm{B}()$ is the barrier function defined as $\mathrm{B}(a)=0$ if $a\leq 0$ and $\mathrm{B}(a)=+\infty$ otherwise. Due to the nature of $\mathrm{B}()$, this problem is not suitable for optimization: the objective function will be infinite when $a>0$. To tackle this drawback, we replace $\mathrm{B}()$ by the approximation introduced by Kervadec *et al.* (2019) called the log-barrier extension and defined as

$$\mathrm{B}_\lambda(a) = \begin{cases} -\frac{1}{\lambda}\ln(-a), & \text{if } a \leq -\frac{1}{\lambda^2}, \\ \lambda a - \frac{1}{\lambda}\ln(\frac{1}{\lambda^2})+\frac{1}{\lambda}, & \text{otherwise.} \end{cases}$$

The parameter $\lambda \in \mathbb{R}^+_*$ parameterized the log-barrier extension $\mathrm{B}_\lambda()$. The function $\mathrm{B}_\lambda()$ tends to $\mathrm{B}()$ when λ tends to $+\infty$; we plot in Figure 4.1 these two functions. Compared to the standard log-barrier[2], the function $\mathrm{B}_\lambda()$ is differentiable even when the constraint is not satisfied, *i.e.*, when $a>0$. Thanks to $\mathrm{B}_\lambda()$, we can take the constraint $r_{\mathcal{S}}(\rho)+\sqrt{\frac{1}{2m}[\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}]} \leq \frac{1}{2}$ into account. Moreover, when the number of examples m is large, we estimate the PAC-Bayesian C-Bound $C^{\mathtt{M}}_{\mathcal{S}}(\rho)$ and the Gibbs risk $r_{\mathcal{S}}(\rho)$ with a mini-batch $\mathbb{U}\subseteq\mathbb{S}$. Concretely, our objective function that is minimized by stochastic gradient descent with Algorithm 4.1 is the following:

$$G^{\mathtt{M}}_{\mathcal{U}}(\rho) \;=\; C^{\mathtt{M}}_{\mathcal{U}}(\rho) + \mathrm{B}_\lambda\left[r_{\mathcal{U}}(\rho)+\sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}\right]}-\frac{1}{2}\right].$$

[2]The reader can refer to Boyd and Vandenberghe (2004) for an introduction of standard log-barrier and interior-point methods.

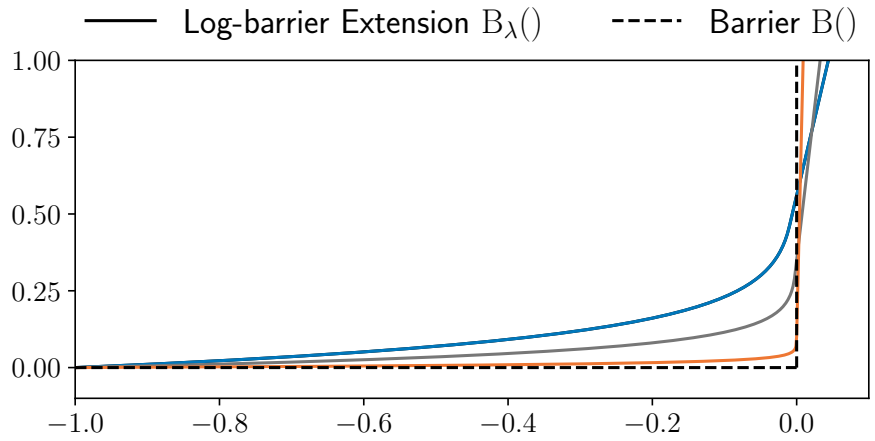


Figure 4.1. *Plot of the barrier function* $\mathrm{B}()$ *(with the dotted line) and the log-barrier extension* $\mathrm{B}_\lambda()$ *(plain lines). We plot the function with three parameters:* $\lambda \in \{10, 20, 100\}$. *The higher the parameter* λ*, the closer the function* $\mathrm{B}_\lambda()$ *to the barrier function* $\mathrm{B}()$. *Thus, the blue, gray and orange curves are respectively with the parameter* $\lambda = 10$, $\lambda = 20$ *and* $\lambda = 100$.

For a given λ, the optimizer will find a solution with a good trade-off between minimizing $C_{\mathcal{U}}^{\mathtt{M}}(\rho)$ and the log-barrier extension function $\mathrm{B}_\lambda()$. As shown in the experiments, minimizing the McAllester-based bound does not lead to the tightest bound. Indeed, such a bound is looser than Seeger-based bounds and leads to a looser PAC-Bayesian C-Bound.

4.4.2 Algorithm based on Equation (4.3)

In order to obtain better generalization guarantees, we should optimize the Seeger-based C-bound of Theorem 4.3.4. Hence to minimize such a PAC-Bayesian C-Bound, we seek to minimize the following optimization problem:

$$\min_{\rho \in \mathbb{M}(\mathbb{H})} \underbrace{\left\{ 1 - \frac{\left(1 - 2 \min\left[\frac{1}{2}, \overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{4\sqrt{m}}{\delta}\right]\right)\right]\right)^2}{1 - 2 \max\left[0, \underline{\mathrm{kl}}\left(d_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln \frac{4\sqrt{m}}{\delta}\right]\right)\right]} \right\}}_{\triangleq\; C_{\mathcal{S}}^{\mathtt{S}}(\rho)}$$
$$\text{s.t} \quad \overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{4\sqrt{m}}{\delta}\right]\right) \leq \frac{1}{2}.$$

For the same reasons as for deriving Algorithm 4.1, we propose to solve by stochastic gradient descent with a mini-batch $\mathbb{U} \subseteq \mathbb{S}$:

$$G^{\mathbb{S}}_{\mathcal{U}}(\rho) = C^{\mathbb{S}}_{\mathcal{U}}(\rho) + \mathrm{B}_{\lambda}\left[\overline{\mathrm{kl}}\left(r_{\mathcal{U}}(\rho) \;\middle|\; \tfrac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \tfrac{4\sqrt{m}}{\delta}\right]\right) - \frac{1}{2}\right].$$

The main challenge in optimizing it is to evaluate $\overline{\mathrm{kl}}$ or $\underline{\mathrm{kl}}$ and to compute their derivatives. The evaluation of $\overline{\mathrm{kl}}$ or $\underline{\mathrm{kl}}$ is done by Algorithm 2.1 proposed by Reeb *et al.* (2018). This method consists in refining iteratively an interval $[p_{\mathsf{min}}, p_{\mathsf{max}}]$ with $p \in [p_{\mathsf{min}}, p_{\mathsf{max}}]$ such that $\mathrm{kl}(q\|p) = \psi$. Moreover, to compute the derivatives with respect to the posterior ρ, we use the chain rule for differentiation with a deep learning framework (such as `PyTorch` (Paszke *et al.*, 2019)) and the derivatives in Equation (2.16). The global algorithm is summarized in Algorithm 4.2.

4.4.3 Algorithm based on Theorem 4.3.5

Theorem 4.3.5 jointly upper-bounds the joint error $e_{\mathcal{D}}(\rho)$ and the disagreement $d_{\mathcal{D}}(\rho)$; But as pointed out in Section 4.3.5 its optimization can be hard. To ease its manipulation, we derive below a C-Bound resulting of a reformulation of the constraints involved in the set $\mathbb{A}_{\mathcal{S}}(\rho)$.

Algorithm 4.2 Minimization of Equation (4.3) by Stochastic Gradient Descent

Given: learning sample $\mathbb{S}$, prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, the objective function $G^{\mathbb{S}}_{\mathcal{S}}(\rho)$
Hyperparameters: number of iterations T
$\rho \leftarrow \pi$
for $t \leftarrow 1$ to T **do**
  for all mini-batches $\mathbb{U} \subseteq \mathbb{S}$ **do**
    Compute $G^{\mathbb{S}}_{\mathcal{U}}(\rho)$ using Algorithm 2.1
    $\rho \leftarrow$ Update ρ with $G^{\mathbb{S}}_{\mathcal{U}}(\rho)$ by gradient descent
return ρ

Theorem 4.4.1 (Reformulation of Lacasse *et al.*'s PAC-Bayesian C-Bound)**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the random

choice of $\mathbb{S} \sim \mathcal{D}^m$ we have for all $\rho \in \mathbb{M}(\mathbb{H})$

$$\mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq \sup_{(e,d)\in\mathbb{A}'_{\mathcal{S}}(\rho)} \left[1 - \frac{\left(1-(2e+d)\right)^2}{1-2d}\right], \tag{4.4}$$

$$\mathbb{A}'_{\mathcal{S}}(\rho) = \left\{(e,d) \;\middle|\; \mathrm{kl}\left(e_{\mathcal{S}}(\rho), d_{\mathcal{S}}(\rho) \| e, d\right) \leq \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}+m}{\delta}\right], \right.$$
$$\left. d \leq 2\sqrt{\min\left(e, \tfrac{1}{4}\right)} - 2e,\ d < \tfrac{1}{2}\right\}.$$

Proof. Deferred to Appendix D.6. ■

Theorem 4.4.1 suggests then the following constrained optimization problem:

$$\min_{\rho\in\mathbb{M}(\mathbb{H})} \left\{ \sup_{(e,d)\in\left[0,\frac{1}{2}\right]^2} \left(1 - \frac{\left[1-(2e+d)\right]^2}{1-2d}\right) \text{ s.t. } (e,d)\in\mathbb{A}'_{\mathcal{S}}(\rho) \right\} \text{ s.t. } 2e_{\mathcal{S}}(\rho)+d_{\mathcal{S}}(\rho)\leq 1,$$

Actually, we can rewrite this constrained optimization problem into an unconstrained one using the barrier function. We obtain

$$\min_{\rho\in\mathbb{M}(\mathbb{H})} \left\{ \max_{(e,d)\in\left[0,\frac{1}{2}\right]^2} \left(C^{\mathrm{L}}(e,d) - \mathrm{B}\left[d - 2\sqrt{\min\left(e,\tfrac{1}{4}\right)} - 2e\right] - \mathrm{B}\left[d - \tfrac{1}{2}\right] \right.\right.$$
$$\left. - \mathrm{B}\left[\mathrm{kl}\left(e_{\mathcal{S}}(\rho), d_{\mathcal{S}}(\rho)\|e,d\right) - \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}+m}{\delta}\right]\right] \right)$$
$$\left. + \mathrm{B}\left[2e_{\mathcal{S}}(\rho) + d_{\mathcal{S}}(\rho) - 1\right] \right\}, \tag{4.5}$$

where $C^{\mathrm{L}}(e,d) = 1 - \frac{(1-(2e+d))^2}{1-2d}$ if $d < \frac{1}{2}$, and $C^{\mathrm{L}}(e,d) = 1$ otherwise. However, this problem cannot be optimized directly by stochastic gradient descent. In this case, we have a min-max optimization problem, *i.e.*, for each descent step we need to find the couple (e,d) that maximizes the $C^{\mathrm{L}}(e,d)$ given the three constraints that define $\mathbb{A}'_{\mathcal{S}}(\rho)$ before updating the posterior distribution ρ.

First, to derive our optimization procedure, we focus on the inner maximization problem when $e_{\mathcal{S}}(\rho)$ and $d_{\mathcal{S}}(\rho)$ are fixed in order to find the optimal (e,d). However, the function $C^{\mathrm{L}}(e,d)$ we aim at maximizing is not concave for all $(e,d) \in \mathbb{R}^2$, implying

Algorithm 4.3 Minimization of Equation (4.4) by Stochastic Gradient Descent

Given: learning sample $\mathbb{S}$, prior $\pi \in \mathbb{M}^*(\mathbb{H})$, the objective function $G_{\mathcal{S}}^{e^*,d^*}(\rho)$
Hyperparameters: number of iterations T
$\rho \leftarrow \pi$
for $t \leftarrow 1$ to T **do**
  for all mini-batches $\mathbb{U} \subseteq \mathbb{S}$ **do**
    $(e^*, d^*) \leftarrow$MAXIMIZE-e-$d(e_{\mathcal{U}}(\rho), d_{\mathcal{U}}(\rho))$
    $\rho \leftarrow$ Update ρ with $G_{\mathcal{U}}^{e^*,d^*}(\rho)$ by gradient descent
return ρ

Given: learning sample $\mathbb{S}$, joint error $e_{\mathcal{S}}(\rho)$, disagreement $d_{\mathcal{S}}(\rho)$
Hyperparameters: tolerance ϵ
function MAXIMIZE-e-$d(e_{\mathcal{S}}(\rho), d_{\mathcal{S}}(\rho))$
  $\alpha_{\text{min}} = 0$ and $\alpha_{\text{max}} = 1$
  while $\alpha_{\text{max}} - \alpha_{\text{min}} > \epsilon$ **do**
    $\alpha = \frac{1}{2}(\alpha_{\text{min}} + \alpha_{\text{max}})$
    $(e, d) \leftarrow$ Solve Equation (4.6)
    if $C^{\text{L}}(e, d) \geq 1-\alpha$ **then** $\alpha_{\text{max}} \leftarrow \alpha$ **else** $\alpha_{\text{min}} \leftarrow \alpha$
  return (e, d)

that the implementation of its maximization can be hard[3]. Fortunately, $C^{\text{L}}(e, d)$ is quasi-concave (GERMAIN *et al.*, 2015) for $(e, d) \in [0, 1] \times [0, \frac{1}{2}]$. Then by definition of quasi-concavity, we have:

$$\forall \alpha \in [0,1], \quad \left\{ (e,d) \,\middle|\, 1 - \frac{\left[1-(2e+d)\right]^2}{1-2d} \geq 1-\alpha \right\}$$
$$\iff \quad \forall \alpha \in [0,1], \quad \left\{ (e,d) \,\middle|\, \alpha(1-2d) - \left[1-(2e+d)\right]^2 \geq 0 \right\}.$$

Hence, for any fixed $\alpha \in [0, 1]$ we can look for (e, d) that maximizes $C^{\text{L}}(e, d)$ and respects the constraints involved in $\mathbb{A}'_{\mathcal{S}}(\rho)$. This is equivalent to solving the following

[3]For example, when using `CVXPY` (DIAMOND and BOYD, 2016), that uses Disciplined Convex Programming (DCP (GRANT *et al.*, 2006)), the maximization of a non-concave function is not possible.

problem for a given $\alpha \in [0,1]$:

$$\begin{aligned} \max_{(e,d)\in[0,\frac{1}{2}]^2} \quad & \alpha(1-2d) - \Big[1-(2e+d)\Big]^2 \\ \text{s.t. } & d \leq 2\sqrt{\min\left(e,\tfrac{1}{4}\right)}-2e \\ \text{and } & \operatorname{kl}\Big(e_{\mathcal{S}}(\rho), d_{\mathcal{S}}(\rho) \| e, d\Big) \leq \frac{1}{m}\left[2\operatorname{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}+m}{\delta}\right]. \end{aligned} \tag{4.6}$$

In fact, we aim at finding $\alpha \in [0,1]$ such that the maximization of Equation (4.6) leads to $1-\alpha$ equals to the largest value of $C^{\mathrm{L}}(e,d)$ under the constraints. To do so, we make use of the "Bisection method for quasi-convex optimization" (Boyd and Vandenberghe, 2004) that is summarized in maximize-e-d in Algorithm 4.3. We denote by (e^*, d^*) the solution of Equation (4.6). Note that, in practice, the joint error and the disagreement is approximated through the mini-batch $\mathbb{U} \subseteq \mathbb{S}$. It remains then to solve the outer minimization problem that becomes:

$$\min_{\rho\in\mathbb{M}(\mathbb{H})} \Bigg\{ \mathrm{B}\left[2e_{\mathcal{S}}(\rho)+d_{\mathcal{S}}(\rho)-1\right] - \mathrm{B}\left[\operatorname{kl}\left(e_{\mathcal{S}}(\rho), d_{\mathcal{S}}(\rho)\|e^*, d^*\right) - \frac{1}{m}\left[2\operatorname{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}+m}{\delta}\right]\right] \Bigg\}.$$

To obtain a objective function that is suitable for stochastic gradient descent, we bring two modifications to the outer minimization problem: *(i)* we replace $\mathrm{B}()$ by the log-barrier extension $\mathrm{B}_\lambda()$ and *(ii)* we approximate the disagreement and the joint error with a mini-batch $\mathbb{U} \subseteq \mathbb{S}$. Hence, we obtain the following objective function:

$$\begin{aligned} G_{\mathcal{U}}^{e^*,d^*}(\rho) = & \,\mathrm{B}_\lambda\left[2e_{\mathcal{U}}(\rho)+d_{\mathcal{U}}(\rho)-1\right] \\ & - \mathrm{B}_\lambda\left[\operatorname{kl}\left(e_{\mathcal{U}}(\rho), d_{\mathcal{U}}(\rho)\|e^*, d^*\right) - \frac{1}{m}\left[2\operatorname{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}+m}{\delta}\right]\right]. \end{aligned}$$

The global method is summarized in Algorithm 4.3. As a side note, we mention that the classic Danskin Theorem (Danskin, 1966) used in min-max optimization theory is not applicable in our case since our objective function is not differentiable for all $(e,d) \in [0,\frac{1}{2}]^2$. We discuss this point in Appendix D.8.

4.5 Experiments

4.5.1 Setting

Our experiments have a two-fold objective: *(i)* assessing the guarantees given by the associated PAC-Bayesian bounds, and *(ii)* comparing the performance of the different

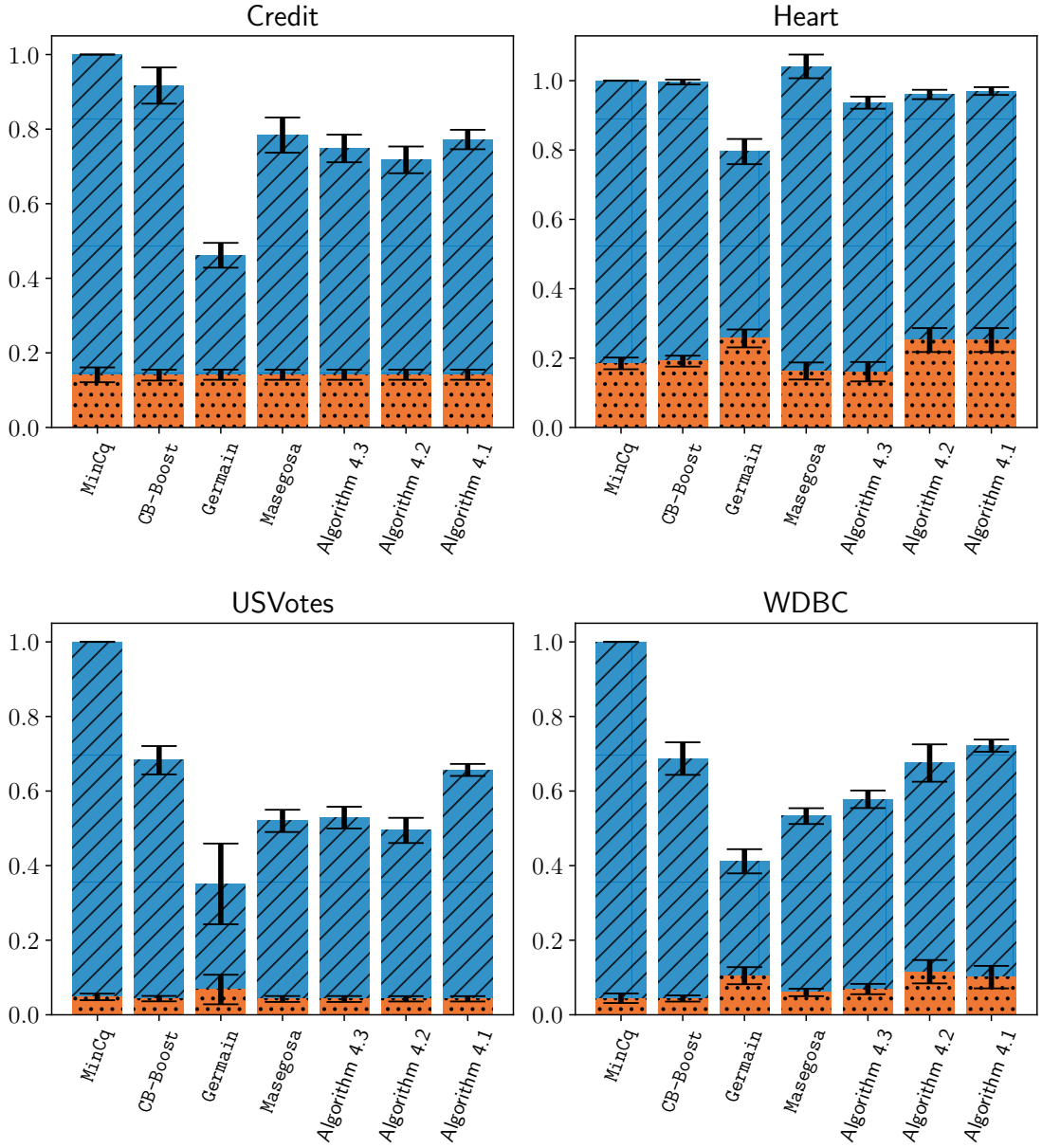


Figure 4.2. *Plot of a comparison between the test risks $R_T(MV_\rho)$ and the generalization bounds in the binary setting when the voters are decision stumps. For each algorithm, we represent the mean of the test risks in the orange bars and the bounds' mean in the blue bars. Additionally, the black lines are the standard deviations.*

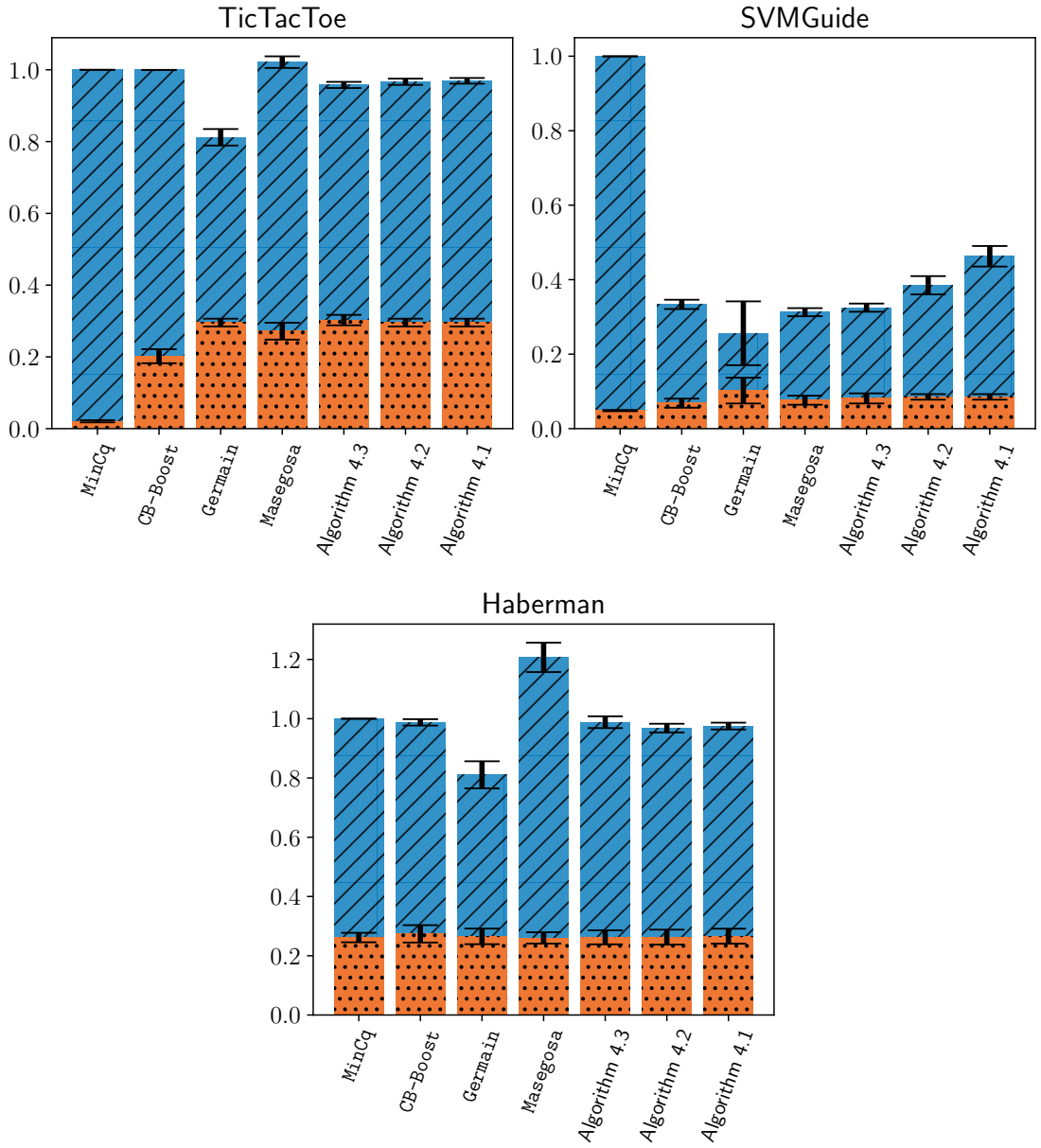


Figure 4.3. *Plot of a comparison between the test risks $R_{\mathcal{T}}(MV_{\rho})$ and the generalization bounds in the binary setting when the voters are decision stumps. For each algorithm, we represent the mean of the test risks in the orange bars and the bounds' mean in the blue bars. Additionally, the black lines are the standard deviations.*

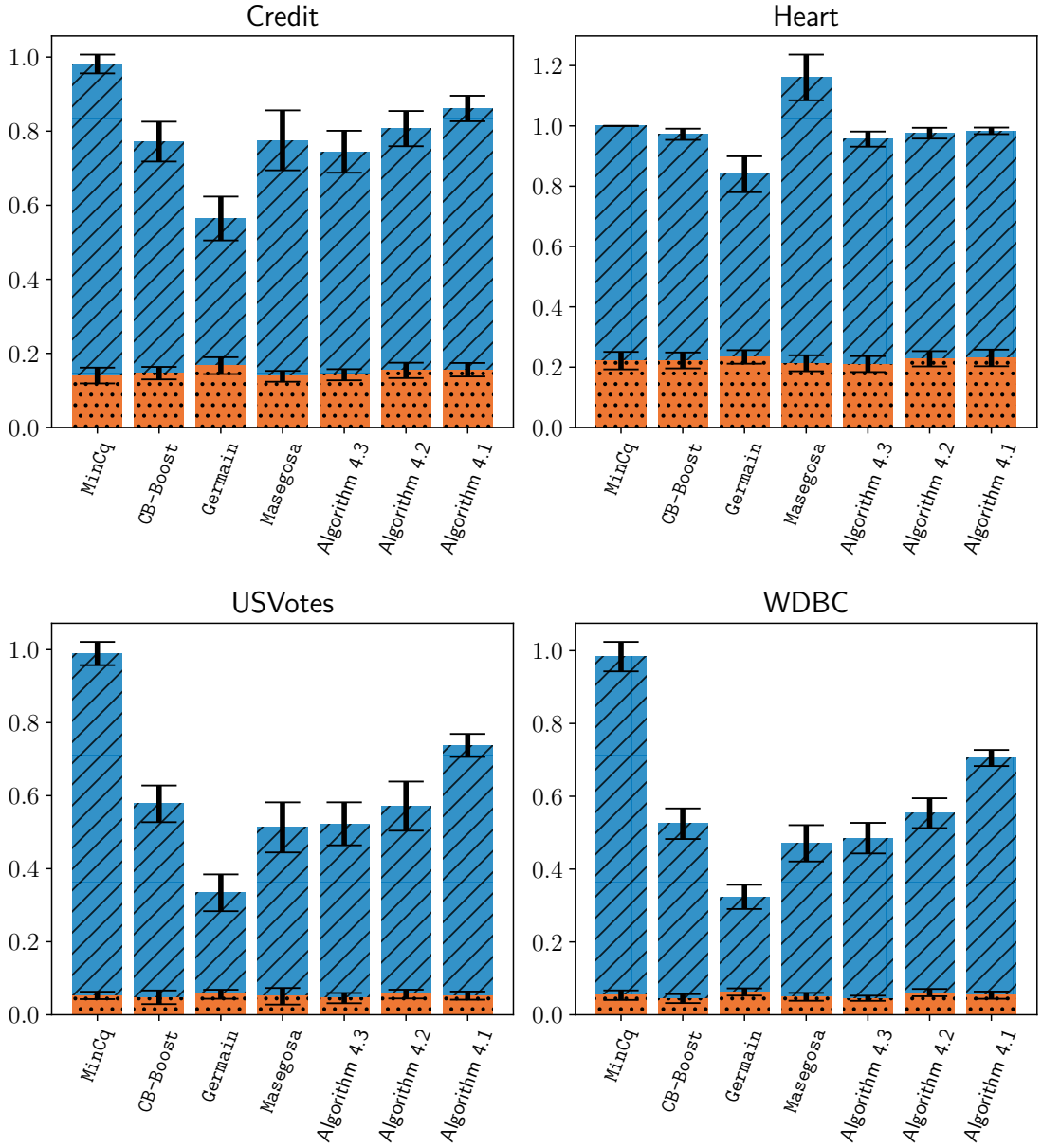


Figure 4.4. *Plot of a comparison between the test risks $R_{\mathcal{T}}(MV_{\rho})$ and the generalization bounds in the binary setting when the voters are decision trees. For each algorithm, we represent the mean of the test risks in the orange bars and the bounds' mean in the blue bars. Additionally, the black lines are the standard deviations.*

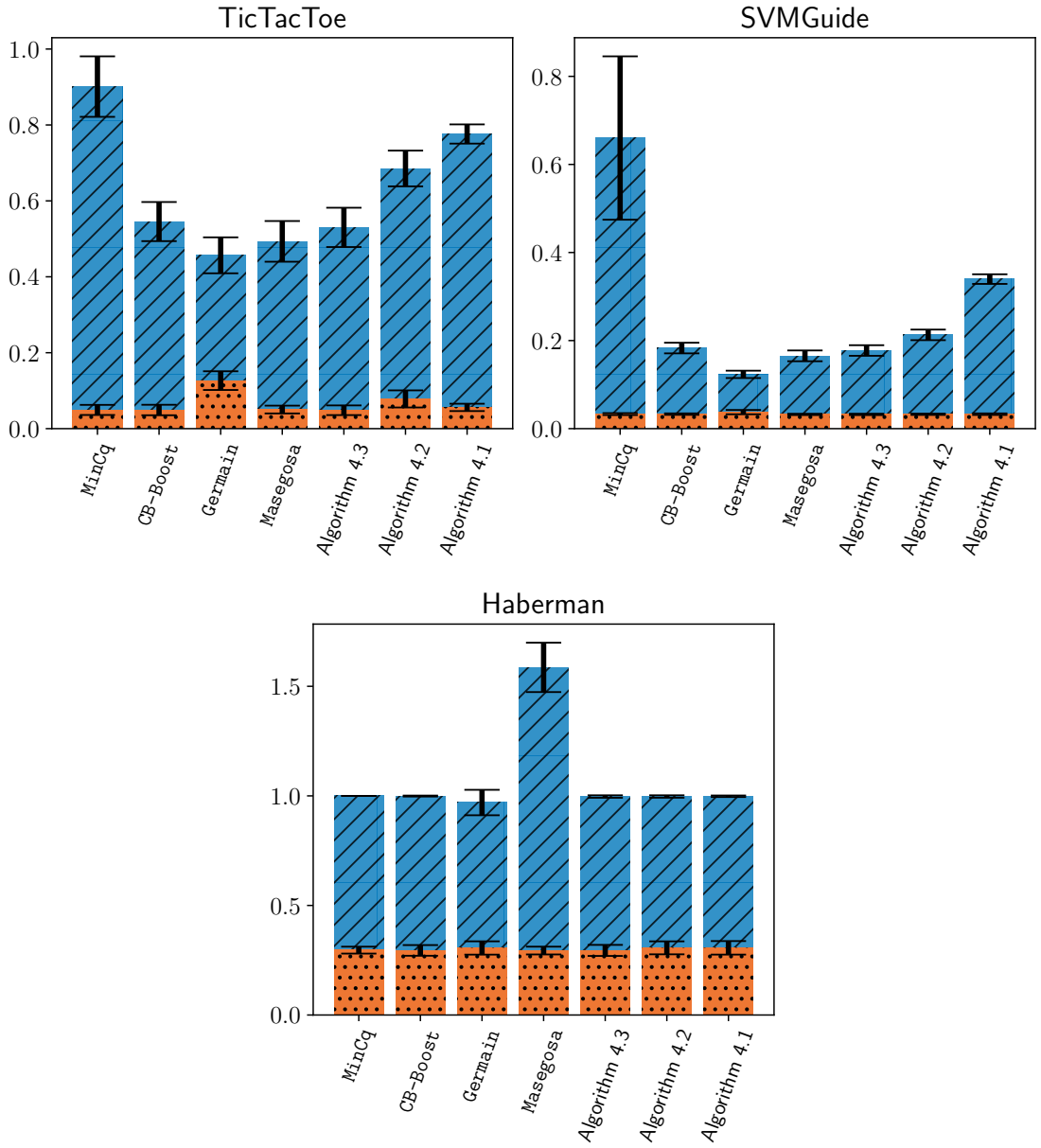


Figure 4.5. *Plot of a comparison between the test risks $R_T(MV_\rho)$ and the generalization bounds in the binary setting when the voters are decision trees. For each algorithm, we represent the mean of the test risks in the orange bars and the bounds' mean in the blue bars. Additionally, the black lines are the standard deviations.*

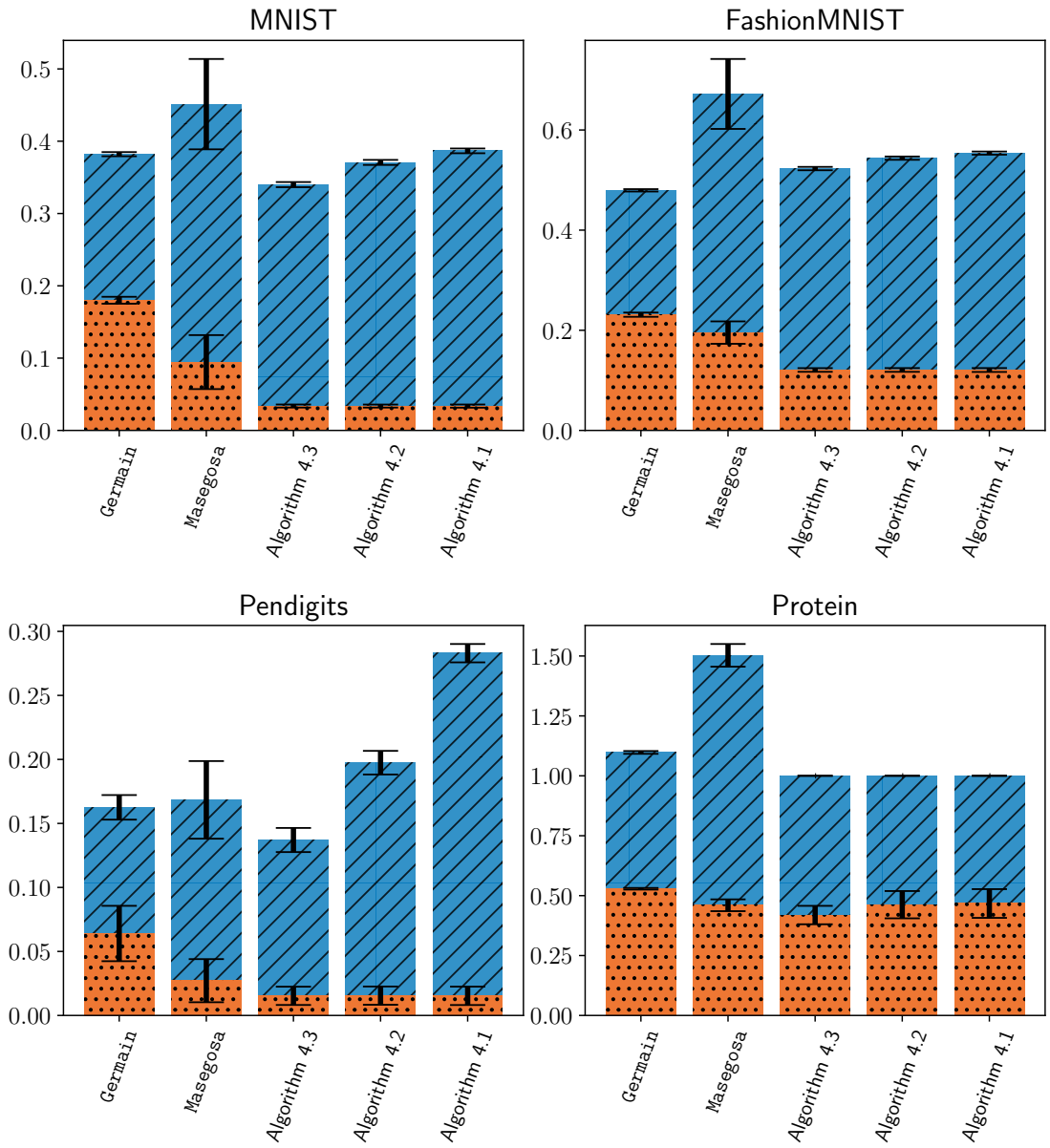


Figure 4.6. *Plot of a comparison between the test risks $R_{\mathcal{T}}(MV_{\rho})$ and the generalization bounds in the multi-class setting when the voters are decision trees. For each algorithm, we represent the mean of the test risks in the orange bars and the bounds' mean in the blue bars. Additionally, the black lines are the standard deviations.*

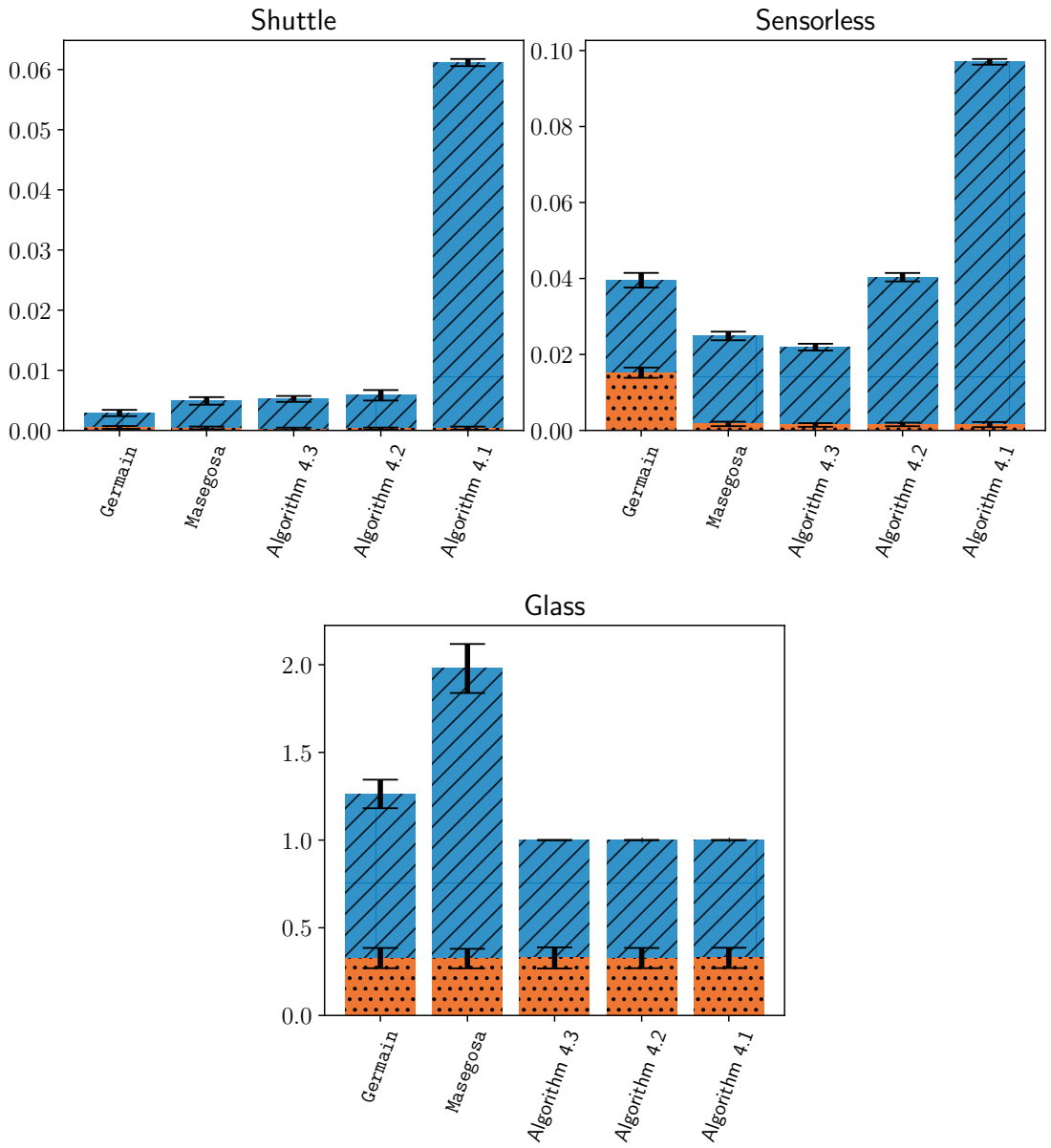


Figure 4.7. *Plot of a comparison between the test risks $R_{\mathcal{T}}(MV_\rho)$ and the generalization bounds in the multi-class setting when the voters are decision trees. For each algorithm, we represent the mean of the test risks in the orange bars and the bounds' mean in the blue bars. Additionally, the black lines are the standard deviations.*

C-bound-based algorithms in terms of risk optimization. We introduce the setting in the following and we report in Figures 4.2 to 4.7 the mean/standard deviation of the risks on the test set $\mathbb{T}$ and the bound values (with $\delta = 0.05$) for 10 runs; see also Appendix D.10.1 for more details. The setting of the experiments is as follows.

Dataset. We consider several binary and multi-class datasets like FashionMNIST (Xiao *et al.*, 2017), MNIST (LeCun *et al.*, 1998) and some coming from the UCI repository (Dua and Graff, 2017). For each different run, we keep the same number of examples in the test or the train set as in the original split. When there is no original split, we use 50% of data in the training set $\mathbb{S}$ and 50% in the test set (except for Sensorless where we have 15% in the test set because the original set is large).

Voters. In the binary setting, we consider either a set $\mathbb{H}$ of decision trees or decision stumps that is complemented: if $h \in \mathbb{H}$ then there is $h' \in \mathbb{H}$ *s.t.* $h'(\mathbf{x}) = -h(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{X}$. Concerning the multi-class setting, we only consider decision trees. Indeed, having decision stumps would have resulted in too many voters. Following Masegosa *et al.* (2020), the prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ is set as the uniform distribution. For the decision stumps, following Roy *et al.* (2011) and Bauvin *et al.* (2020), we use 10 decision stumps per feature. For the decision trees, we follow a general setting similar to the one of (Masegosa *et al.*, 2020). Moreover, 100 trees are learned with 50% of the training data (the remaining part serves to learn the posterior ρ). More precisely, for each tree $\sqrt{d}$ features of the d-dimensional input space are selected, and the trees are learned by using the Gini criterion until the leaves are pure.

Algorithms' parameters. To update of the posterior ρ in Algorithms 4.1 to 4.3 is done through the COCOB-Backprop optimizer (Orabona and Tommasi, 2017) (its parameter remains the default one). In the binary setting, we optimize for $T = 2,000$ iterations (by batch gradient descent), and in the multi-class setting, we set 20 epochs with a batch size of 64. Lastly, we consider the parameter $\lambda{=}100$ for log-barrier extension $\mathrm{B}_\lambda()$.

Comparisons. We compare the three algorithms proposed in this chapter to the following state-of-the-art PAC-Bayesian methods for majority vote learning.

- We compare with the algorithm proposed by Masegosa *et al.* (2020) that optimizes a PAC-Bayesian bound on $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq 4e_{\mathcal{D}}(\rho)$ (Theorem 2.2.2); see Theorem 9 and Appendix G of (Masegosa *et al.*, 2020) for a description of this algorithm that we denote by `Masegosa`. For Masegosa *et al.*'s algorithm, we kept the original parameters.

- Our algorithm to optimize the PAC-Bayesian bound on $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq 2r_{\mathcal{D}}(\rho)$ (Theorem 2.2.1) recalled in Theorem 4.3.1 and derived by (GERMAIN *et al.*, 2015, PAC-Bound 0). Even though (GERMAIN *et al.*, 2015) does not optimize the bound, we denote this algorithm by `Germain`. The algorithm is similar to Algorithm 4.2, but without the numerator of the C-Bound (*i.e.*, the disagreement); more details are given in Appendix D.9.

- In the binary setting only, we compare with `MinCq` (ROY *et al.*, 2011) and `CB-Boost` (BAUVIN *et al.*, 2020) that are based on the minimization of the empirical C-Bound $C_{\mathcal{S}}(\rho)$. Indeed, these algorithms are developed for this setting only. For comparison purposes and since `MinCq` and `CB-Boost` do not explicitly minimize a PAC-Bayesian bound, we report the bound values of Theorem 4.4.1 instantiated with the models learned; Moreover, for `MinCq`, we select the margin parameter among 20 values uniformly spaced between $[0, \frac{1}{2}]$ by 3-fold cross validation. For `CB-Boost`, which is based on a Boosting approach, we fix the maximal number of boosting iterations to 200.

4.5.2 Analysis of the Results

When comparing only the PAC-Bayesian C-Bounds, we observe in Figures 4.2 to 4.7 that Algorithm 4.1 provides the worst bound. Algorithm 4.3 provides usually tighter bounds than Algorithms 4.1 and 4.2 except for Harberman and USVotes. We believe that Algorithm 4.3 provides lower bounds than Algorithm 4.2 because the LACASSE *et al.*'s approach bounds simultaneously the joint error and the disagreement. Algorithm 4.3 appears then to be the best algorithm among our three self-bounding algorithms that minimize a PAC-Bayesian C-Bound. Moreover, Algorithm 4.3 gives usually the lowest true risks or it is comparable to the two other algorithms.

Compared to the baselines, `Germain` gives usually the lowest bounds among all the algorithms, but at the price of a large test risk. This clearly illustrates the limitation of considering *only* the Gibbs risk as an estimator of the majority vote risk: as discussed in Section 2.2.2, the Gibbs risk is an unfair estimator since an increase in the diversity between the voters can have a negative impact on the Gibbs risk.
Second, compared to MASEGOSA *et al.*'s approach, the results are comparable. This behavior was expected since minimizing the bound of MASEGOSA *et al.* (2020) or the PAC-Bayesian C-Bound boils down to minimize a trade-off between the risk and the disagreement. Third, in the binary setting, compared to empirical C-bound minimization algorithms, we see that Algorithm 4.3 outputs better results than `CB-Boost` and `MinCq` for which the difference is significative, and the bounds are close to 1 (*i.e.*, non-informative). Optimizing the risk bounds tends to provide better guarantees that justify that optimizing the empirical C-bound is often too optimistic (as done in

CB-Boost or MinCq); we provide in Appendix D.10.3 an illustration of the different solutions obtained from the algorithms.

Overall, from these experiments, our Algorithm 4.3 is the one that provides the best trade-off between having good performances in terms of risk optimization and ensuring good theoretical guarantees with informative bounds. Moreover, in Appendix D.10.2 we show that Algorithm 4.3 has a higher computation time than the others algorithms. This makes Algorithm 4.2 a good trade-off between ensuring good theoretical guarantees and having a low computation time.

4.6 Conclusion and Summary

This chapter presents learning algorithms that minimize the majority vote's risk with PAC-Bayesian generalization bounds based on the C-Bound. More precisely, we propose solving three optimization problems, each derived from an existing PAC-Bayesian bound. Our methods belong to the class of *self-bounding* learning algorithms (Freund, 1998): the learned predictor comes with a tight and statistically valid risk upper bound. Our experimental evaluation has confirmed the quality of the learned predictor and the tightness of the bounds with respect to state-of-the-art methods minimizing the C-Bound.

As we said before, no algorithm minimizes the *empirical* C-Bound in the multi-class setting. One of the reasons is that it is not easy to find a convex program like for MinCq or P-MinCq algorithm. Hopefully, thanks to the deep learning framework, minimizing the C-Bound is possible even without convexity through a stochastic gradient descent algorithm (as we show in Chapter 5 in the binary setting). Hence, in the future, we plan to explore more the minimization of the C-Bound.

However, one drawback of the PAC-Bayesian C-Bounds and the other ones of the literature is that the majority vote true risk is not directly minimized: we need surrogates such as the Gibbs risk, the disagreement, or the joint error. Indeed, the C-Bound is already an upper bound on the true risk, which makes the PAC-Bayesian C-Bound even looser, thus, these generalization bounds cannot be tight. To avoid this issue, we bound the *expected* true risk of the majority vote. This is done in the next chapter: we *(i)* introduce the *stochastic* majority vote that samples a distribution ρ for each prediction, and we *(ii)* provide guarantee on the true risk for this classifier. As we will see, our obtained (PAC-Bayesian) guarantee is differentiable, and we can derive self-bounding learning to learn such a classifier.

Toward a Stochastic Majority Vote

5

This chapter is based on the following paper

Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound. *Advances in Neural Information Processing Systems (NeurIPS)*. (2021)

Contents

5.1 Introduction **126**
5.2 The Stochastic Majority Vote **129**
5.2.1 Definitions 129
5.2.2 Approximation of the Stochastic Risk 131
5.2.3 Computing Exactly the Stochastic Risk 131
5.3 From a PAC-Bayesian Bound to an Algorithm **133**
5.3.1 A PAC-Bayesian Bound for Stochastic Majority Votes 133
5.3.2 A PAC-Bayesian Bound for Data-dependent Voters 135
5.3.3 Learning Algorithms for the Stochastic Majority Vote 136
5.4 Experiments **138**
5.4.1 Comparison Between the Computations of the Risk 139
5.4.2 Performance of Algorithms 5.3 and 5.4 139
5.5 Conclusion and Summary **141**

Abstract

We study a stochastic counterpart of the majority vote classifier called the *stochastic majority vote*, and study its generalization properties. Unlike Chapter 4, the posterior distribution associated with the majority vote is sampled from another probability distribution. While the stochastic majority vote holds for arbitrary distributions, we instantiate it with Dirichlet distributions: this allows to derive a closed-form and differentiable expression for the expected risk. Then, we derive self-bounding algorithms for stochastic majority vote, that benefit from tight generalization bounds when compared to self-bounding algorithms studied in Chapter 4.

5.1 Introduction

In Chapter 4, we considered some self-bounding algorithms (FREUND, 1998) that minimize a PAC-Bayesian bound on the majority vote true risk. Each PAC-Bayesian bound depends on a surrogate on the majority vote risk (see Section 2.2.2). Given any distribution ρ on an hypothesis set $\mathbb{H}$, we have seen three surrogates of the majority vote's risk in Chapter 2:

(i) Twice the Gibbs Risk $r_{\mathcal{S}}(\rho) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{S}} \mathbb{E}_{h\sim\rho} \mathrm{I}\left[h(\mathbf{x}) \neq y\right]$ (Theorem 2.2.1),

(ii) 4 times the joint error $e_{\mathcal{S}}(\rho) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{S}} \mathbb{E}_{h\sim\rho} \mathbb{E}_{h'\sim\rho} \mathrm{I}\left[h(\mathbf{x}) \neq y\right] \mathrm{I}\left[h'(\mathbf{x}) \neq y\right]$ (Theorem 2.2.2),

(iii) The C-Bound $C_{\mathcal{S}}(\rho) = 1 - \frac{(1-2r_{\mathcal{S}}(\rho))^2}{1-2d_{\mathcal{S}}(\rho)}$, where $d_{\mathcal{S}}(\rho)$ is the disagreement defined as $d_{\mathcal{S}}(\rho) = 2\,\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{S}} \mathbb{E}_{h\sim\rho} \mathbb{E}_{h'\sim\rho} \mathrm{I}\left[h(\mathbf{x}) \neq y\right] \mathrm{I}\left[h'(\mathbf{x}) = y\right]$ (Theorem 2.2.3).

Figure 5.1 illustrates the models obtained by the minimization of the empirical risk with respect to the three surrogates recalled above on a simple dataset (here, moons): only the empirical C-Bound fully leverage the diversity of the voters to obtain $\mathrm{R}_{\mathcal{S}}(\mathrm{MV}_\rho) = 0$.

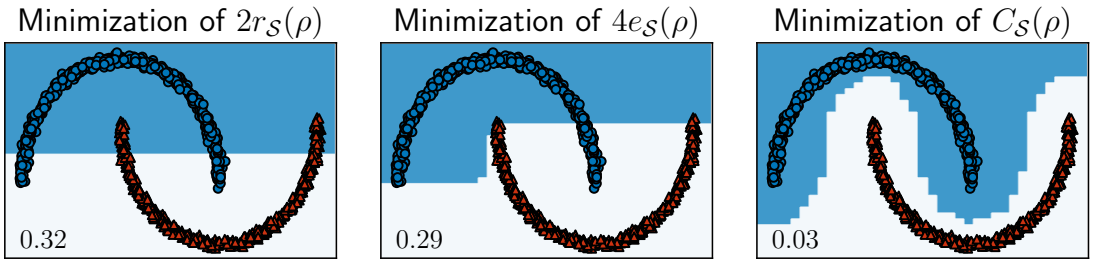


Figure 5.1. *Plot of the majority vote's decision boundary obtained by minimizing the upper bounds on the majority vote's empirical risk with the surrogates recalled in Theorems 2.2.1 to 2.2.3 for moons dataset with a learning sample size* $m = 1,000$. *For each plot, we print the surrogate value in the bottom left corner.*

As highlighted in Figure 5.1, the first two surrogates (Theorems 2.2.1 and 2.2.2) are not efficient to minimize the empirical majority vote's risk. In contrast, since the empirical C-Bound $C_{\mathcal{S}}(\rho)$ (Theorem 2.2.3) is a tighter upper bound on the majority vote's empirical risk (Theorem 2.2.4), the minimization of the empirical risk through the C-Bound is more accurate. Another precise surrogate on the majority vote's empirical risk is based on the *randomized majority vote*. This model, introduced by LACASSE *et al.* (2010), consists in defining a majority vote

$$\mathrm{MV}_\sigma(\mathbf{x}) \triangleq \operatorname*{argmax}_{y'\in\mathbb{Y}} \mathop{\mathbb{E}}_{h\sim\sigma} \mathrm{I}\left[h(\mathbf{x}) = y'\right],$$

where σ is constructed as follows: N voters $\mathbb{H}' = \{h_1, \ldots, h_N\}$ are sampled from ρ and a uniform posterior distribution σ is defined such that $\sigma(h) = \frac{1}{N}$ for all $h \in \mathbb{H}'$. Following LACASSE *et al.* (2010), the randomized majority vote's empirical risk is defined as

$$\begin{aligned}
&\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{S},\mathsf{MV}_\sigma\sim\rho^N}\left[\mathsf{MV}_\sigma(\mathbf{x}) \neq y\right] \\
&\quad\leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathbb{S}}\left[\sum_{j=\lceil\frac{N}{2}\rceil}^{N}\binom{N}{j}\left[\frac{1}{2}\left(1-\widehat{m_\rho}(\mathbf{x},y)\right)\right]^j\left[1-\frac{1}{2}\left(1-\widehat{m_\rho}(\mathbf{x},y)\right)\right]^{(N-j)}\right] \\
&\quad\triangleq b_{\mathbb{S}}^N(\rho),
\end{aligned}$$

where $\widehat{m_\rho}(\mathbf{x},y)$ is the $\frac{1}{2}$-margin defined in Definition 2.2.4. Given an example $(\mathbf{x},y)\sim\mathbb{S}$, the sum corresponds to the probability that at least $\frac{N}{2}$ voters make an error over N voters sampled from ρ. In fact, it is the complementary cumulative distribution function of the binomial distribution with parameter $\frac{1}{2}(1-\widehat{m_\rho}(\mathbf{x},y))$ and $\lceil\frac{N}{2}\rceil$ trials. Hence, $b_{\mathbb{S}}^N(\rho)$ is the expected complementary cumulative distribution function on the learning sample $\mathbb{S}$. Moreover, the *randomized* majority vote can be linked to the classical majority vote MV_ρ: the term $b_{\mathcal{S}}^N(\rho)$ is another surrogate on the majority vote empirical risks (LACASSE *et al.*, 2010). Indeed, we have

$$\mathsf{R}_{\mathcal{S}}(\mathsf{MV}_\rho) \leq 2b_{\mathcal{S}}^N(\rho). \tag{5.1}$$

Note that we provide a proof of this bound in Appendix E.1. Hopefully, the higher N, the better $b_{\mathcal{S}}^N(\rho)$ approximates the majority vote's empirical risk risk. As for the other surrogates (*i.e.*, the Gibbs risk, the joint error, and the C-Bound), when one wants to upper-bound the majority vote's true risk, PAC-Bayesian bounds are used. In the rest of the chapter, we denote by `Lacasse` our algorithm that minimizes a PAC-Bayesian bound depending on $b_{\mathcal{S}}^N(\rho)$ (see Appendix E.1). However, when the true risk is minimized through self-bounding algorithms, it gives even worse results on the moons dataset. This is illustrated on Figure 5.2 that plots the different self-bounding procedures.

As we can remark, except for the one labelled `Lacasse` and our new Algorithm 5.3, the minimization of the PAC-Bayesian bounds does not fully leverage the voters' correlations to obtain a *(i)* tight generalization bound and *(ii)* a small empirical risk. This is mainly due to the fact that they do not minimize directly the majority vote's risk (but surrogates instead). In contrast, as illustrated by LACASSE (2010)'s result, the randomized majority vote offers a way to obtain a tight generalization guarantee with a small empirical risk $\mathsf{R}_{\mathcal{S}}(\mathsf{MV}_\rho)$. However, the considered majority vote in LACASSE *et al.* (2010)'s approach is not the original majority vote MV_ρ, but a rather restricted form.

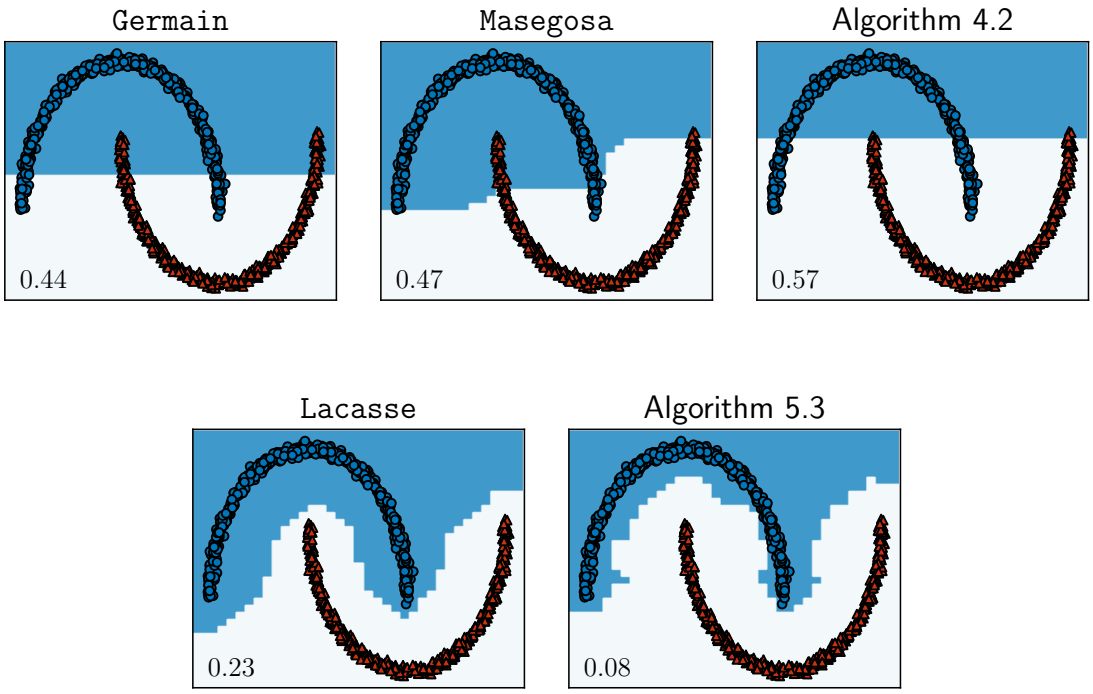


Figure 5.2. *Plot of the majority vote's decision boundary obtained by executing the self-bounding algorithms for moons dataset with* $m = 1,000$ *(and* $N = 100$ *for* `Lacasse`*). We represent in the bottom left of the plots, the value of each bound minimized with Algorithm 4.2, Algorithm 5.3 (and Algorithm 5.2),* `Germain` *(Algorithm D.1),* `Masegosa` *(*Masegosa et al.*, 2020, Appendix G), and* `Lacasse` *(Algorithm E.1).*

Hence, in this chapter, we introduce the *stochastic* majority vote that overcomes the drawbacks of the literature's methods: we provide tight generalization bounds on the classical majority vote MV_ρ. The *stochastic* majority vote is defined as follows: for each input $\mathbf{x}$, a majority vote MV_ρ is obtained by sampling the weights from another probability distribution.

This new majority vote is presented in more details in Section 5.2; its definition is given in Section 5.2.1. Moreover, we show in Section 5.2.2 how to approximate the risk and in Section 5.2.3 the exact computation. Along with the risk computation, we derive in Sections 5.3.1 and 5.3.2 two PAC-Bayesian bounds essential to derive self-bounding algorithms in Section 5.3.3. Section 5.4 provides a study of these algorithms. The proofs are deferred in Appendix E.

5.2 The Stochastic Majority Vote

5.2.1 Definitions

For the *stochastic* majority vote, we consider that its weights $\rho \in \mathbb{M}(\mathbb{H})$ are sampled from a distribution P called hyper-posterior[1]; we say that P is a hyper-posterior on $\mathbb{H}$. By considering hyper-posteriors, the majority votes become *stochastic*, *i.e.*, for each input $\mathbf{x} \in \mathbb{X}$ the weights ρ are sampled from the hyper-posterior P to obtain the prediction $\mathrm{MV}_\rho(\mathbf{x})$. The main advantage of considering a stochastic majority vote is that it allows to derive and to optimize PAC-Bayesian generalization bounds directly. The true risk and empirical risk of the proposed stochastic weighted majority vote take into account the risks of MV_ρ where ρ is sampled from the hyper-posterior P; the risk is defined in the following definition.

Definition 5.2.1 (Risks of the stochastic majority vote)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any for any voters' set $\mathbb{H}$, for any hyper-posterior distribution P on $\mathbb{H}$, the stochastic risks are defined as

$$\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) = \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}\left[m_\rho(\mathbf{x},y) \leq 0\right],$$

$$\text{and} \quad \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathrm{R}_{\mathcal{S}}(\mathrm{MV}_\rho) = \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}\left[m_\rho(\mathbf{x}_i,y_i) \leq 0\right].$$

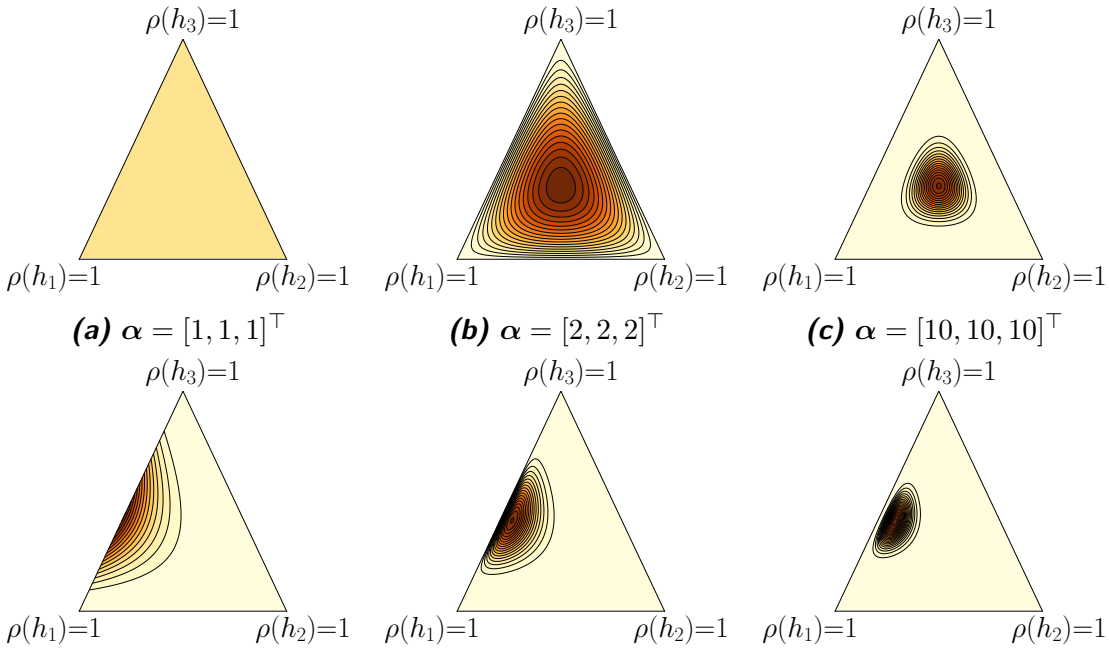
We use the $\frac{1}{2}$-margin of Laviolette *et al.* (2017) to upper-bound the risk of the stochastic majority vote. Indeed, we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}\left[\widehat{m_\rho}(\mathbf{x},y) \leq 0\right] = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}}(\mathbf{x},y), \tag{5.2}$$

$$\text{and} \quad \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathrm{R}_{\mathcal{S}}(\mathrm{MV}_\rho) \leq \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}\left[\widehat{m_\rho}(\mathbf{x}_i,y_i) \leq 0\right] = \frac{1}{m}\sum_{i=1}^{m} s_{\mathrm{P}}(\mathbf{x}_i,y_i). \tag{5.3}$$

where $s_{\mathrm{P}}(\mathbf{x},y) \triangleq \mathbb{E}_{\rho\sim\mathrm{P}} \mathrm{I}\left[\widehat{m_\rho}(\mathbf{x},y) \leq 0\right]$ is the *stochastic* risk. Note first that the inequality is attained in the binary setting (see Section 2.2.2). Additionally, the advantage of $s_{\mathrm{P}}(\mathbf{x},y)$ is that we are able to derive a closed form solution in Section 5.2.3. Actually, we introduce two ways to compute this risk: we can either *(i)* approximate it (*e.g.*, through Monte Carlo methods) or *(ii)* compute its closed form. In both cases, assumptions have to be made on the distribution P. When the distribution ρ is discrete, it lies in the $(\mathrm{card}(\mathbb{H})-1)$ dimensional probability simplex: $\mathbb{\Delta}^{(\mathrm{card}(\mathbb{H})-1)}$. Hence,

[1] The hyper-posteriors have been first used in the PAC-Bayesian theory by Pentina and Lampert (2014) in the context of Life-long learning problem in order to be able to consider different majority votes adapted to the different specific tasks seen during training.



(a) $\boldsymbol{\alpha} = [1, 1, 1]^\top$ ***(b)*** $\boldsymbol{\alpha} = [2, 2, 2]^\top$ ***(c)*** $\boldsymbol{\alpha} = [10, 10, 10]^\top$

(d) $\boldsymbol{\alpha} = [5, 1, 4]^\top$ ***(e)*** $\boldsymbol{\alpha} = [10, 2, 8]^\top$ ***(f)*** $\boldsymbol{\alpha} = [25, 5, 20]^\top$

Figure 5.3. *The figure shows the probability density function for different value of parameters $\boldsymbol{\alpha}$. More precisely, the "triangle" correspond to the 2-dimensional probability simplex $\mathbb{\Delta}^2$ where its extremities correspond to the extreme probability distributions. Hence, a point in the simplex is a linear combination of these extreme probability distributions.*

a natural choice for the hyper-posterior is the Dirichlet distribution; its probability density function is defined as follows.

Definition 5.2.2 (Dirichlet Distribution)**.** Let $n = \mathrm{card}(\mathbb{H})$ be the cardinality of a finite hypothesis set $\mathbb{H}$. Given the concentration parameters $\boldsymbol{\alpha} \in (\mathbb{R}_*^+)^n$, the Dirichlet Distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ is defined as

$$\rho \sim \mathrm{P} \iff (\rho(h_1), \ldots, \rho(h_n)) \sim \mathrm{Dir}(\boldsymbol{\alpha}),$$
$$\text{where} \quad \mathrm{P}(\rho) \triangleq \frac{1}{Z(\boldsymbol{\alpha})} \prod_{j=1}^{n} \Big[\rho(h_j)\Big]^{\alpha_j - 1} \propto \prod_{j=1}^{n} \Big[\rho(h_j)\Big]^{\alpha_j - 1}.$$

We provide in Figure 5.3 some examples of Dirichlet distributions. Notice that by taking $\boldsymbol{\alpha}$ as the vector of all ones, the distribution corresponds to a uniform distribution over the simplex $\mathbb{\Delta}^{(\mathrm{card}(\mathbb{H})-1)}$.

Under Dirichlet assumptions on the hyper-posterior distribution P, we propose in the next section an algorithm to approximate the stochastic risk $s_{\mathrm{P}}(\mathbf{x}, y)$. This algorithm is actually part of our self-bounding algorithm in Section 5.3.3.

5.2.2 Approximation of the Stochastic Risk

We now propose a Monte Carlo (MC) algorithm to compute $s_{\mathrm{P}}(\mathbf{x}, y)$ that is suited to speed up the optimization. For this optimization algorithm, we need to introduce a surrogate of the true risk to update $\boldsymbol{\alpha}$ by gradient descent as the gradients of the 01-loss are always zero. We make use of a *tempered* sigmoid loss $\mathrm{sig}_c(x) = \frac{1}{1+\exp(-cx)}$ with slope parameter $c \in \mathbb{R}^+$. Because of the surrogate, this optimization algorithm solves a relaxation of the original problem and not its exact form (Nesterov, 2005). The MC-based optimization algorithm is described in Algorithm 5.1.

Algorithm 5.1 Approximating the Stochastic Risk

Given: Dirichlet distribution $\mathrm{P} = \mathrm{Dir}(\boldsymbol{\alpha})$, learning sample $\mathbb{S}$
Hyperparameters: number of draws K
Draw a sample $\{\rho_k\}_{k=1}^K \sim \mathrm{P}^K = \mathrm{Dir}(\boldsymbol{\alpha})^K$
for all example $(\mathbf{x}_i, y_i) \in \mathbb{S}$ **do**

$$s_{\mathrm{P}}(\mathbf{x}_i, y_i) \approx \frac{1}{K} \sum_{k=1}^{K} \mathrm{sig}_c\left[-\widehat{m_{\rho_k}}(\mathbf{x}_i, y_i)\right]$$

return $\frac{1}{m}\sum_{i=1}^m s_{\mathrm{P}}(\mathbf{x}_i, y_i)$

This algorithm first samples K majority votes and computes an approximation of the stochastic risk $s_{\mathrm{P}}(\mathbf{x}_i, y_i)$ for each example $(\mathbf{x}_i, y_i) \in \mathbb{S}$ by an average. A drawback of Algorithm 5.1 is that it requires to sample K majority votes and predict all the examples in the learning sample $\mathbb{S}$. Hence, to overcome this issue, we derive a closed-form solution of the stochastic risk $s_{\mathrm{P}}(\mathbf{x}, y)$ in the next section.

5.2.3 Computing Exactly the Stochastic Risk

Under Dirichlet assumptions, a closed-form solution can be derived for the expected risk. The following lemma introduces this solution.

Lemma 5.2.1 (Computation of the Stochastic Risk)**.** For a given $(\mathbf{x}, y) \in \mathbb{X}\times\mathbb{Y}$, let

$$\mathbb{F}(\mathbf{x}, y) = \{j \ : \ h_j(\mathbf{x}) \neq y\} \quad \text{and} \quad \mathbb{T}(\mathbf{x}, y) = \{j \ : \ h_j(\mathbf{x}) = y\}$$

be respectively the set of indices of the voters that misclassify $(\mathbf{x}, y)$ and the set of indices of the voters that correctly classify $(\mathbf{x}, y)$. Then, the stochastic risk $s_{\mathrm{P}}(\mathbf{x}, y)$ can be rewritten as

$$s_{\mathrm{P}}(\mathbf{x}, y) = \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathrm{I}\left[\widehat{m_\rho}(\mathbf{x}, y) \leq 0\right] = I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right),$$

with $I_{0.5}()$ the regularized incomplete beta function evaluated at 0.5. It is defined as

$$I_{0.5}(a, b) \triangleq \frac{B_{0.5}(a, b)}{B_1(a, b)}, \quad \text{where} \quad B_t(a, b) \triangleq \int_0^t x^{a-1}(1-x)^{b-1}dx$$

is the incomplete beta function.

Proof. Deferred to Appendix E.3. ■

Lemma 5.2.1 tells us that the stochastic risk given an example $(\mathbf{x}, y)$ can be computed with a closed-form solution. In consequence, thanks to Lemma 5.2.1, we compute an upper-bound on the risk of the stochastic majority vote based on the stochastic risk. Indeed, we deduce the following corollary.

Corollary 5.2.1 (Closed-form Solution of the Stochastic Risks)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any learning sample $\mathbb{S} \sim \mathcal{D}^m$, for any finite hypothesis set $\mathbb{H}$, for any distribution $\mathrm{P} = \mathrm{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in (\mathbb{R}_*^+)^{\mathrm{card}(\mathbb{H})}$, we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}}(\mathbf{x}, y) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right),$$

$$\text{and} \quad \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathsf{R}_{\mathcal{S}}(\mathsf{MV}_\rho) \leq \frac{1}{m}\sum_{i=1}^{m} s_{\mathrm{P}}(\mathbf{x}_i, y_i) = \frac{1}{m}\sum_{i=1}^{m} I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x}_i,y_i)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x}_i,y_i)} \alpha_j\right).$$

Proof. Deferred to Appendix E.4. ■

From Corollary 5.2.1, we propose to compute directly the empirical stochastic risk. This is in contrast with Algorithm 5.1 that approximates the stochastic risk by Monte Carlo sampling. The computation is summarized in Algorithm 5.2. Note that we provide in 5.4.1 an empirical study showing in which regimes each algorithm is more efficient.

Algorithm 5.2 Computing Exactly the Stochastic Majority Vote's Risk

Given: Dirichlet distribution $\mathrm{P} = \mathrm{Dir}(\boldsymbol{\alpha})$, learning sample $\mathbb{S}$

for all example $(\mathbf{x}_i, y_i) \in \mathbb{S}$ **do**

$$s_{\mathrm{P}}(\mathbf{x}_i, y_i) = I_{0.5}\left(\sum_{j \in \mathbb{T}(\mathbf{x}_i, y_i)} \alpha_j, \sum_{j \in \mathbb{F}(\mathbf{x}_i, y_i)} \alpha_j\right)$$

return $\frac{1}{m}\sum_{i=1}^{m} s_{\mathrm{P}}(\mathbf{x}_i, y_i)$

Thanks to Algorithm 5.1 and Algorithm 5.2, we are now able to compute the compute the empirical stochastic risk. This is a key step to derive our self-bounding algorithm in Section 5.3.3 for the stochastic majority vote. Indeed, the PAC-Bayesian generalization bound that we derive in the next section requires the computation of the stochastic risk in order to be minimized.

5.3 From a PAC-Bayesian Bound to an Algorithm

We now derive PAC-Bayesian generalization bounds for our proposed stochastic majority vote. To do so, we upper-bound the true stochastic risk with a SEEGER-like PAC-Bayesian bound. More precisely, we propose in Section 5.3.1 a PAC-Bayesian bound for a stochastic majority vote with voters that do not depend on the learning sample $\mathbb{S}$ and in Section 5.3.2 we derive a PAC-Bayesian for data-dependent voters.

5.3.1 A PAC-Bayesian Bound for Stochastic Majority Votes

Before presenting our PAC-Bayesian bound for the stochastic majority vote, we consider that we have an *apriori* on the majority vote weights $\rho \in \mathbb{M}(\mathbb{H})$, *i.e.*, we assume a hyper-prior distribution Π over the voters' set $\mathbb{H}$. By doing so, we are able to derive a bound that depends on the KL divergence $\mathrm{KL}(\mathrm{P}\|\Pi)$ between the hyper-prior Π and the hyper-posterior P. Our bound is presented in the following theorem.

Theorem 5.3.1 (PAC-Bayesian Bound for Stochastic Majority Votes)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any finite hypothesis set $\mathbb{H}$, for any distribution $\Pi = \mathrm{Dir}(\boldsymbol{\beta})$ with $\boldsymbol{\beta} \in (\mathbb{R}_*^+)^{\mathrm{card}(\mathbb{H})}$, for any $\delta \in (0, 1]$, with probability at least $1-\delta$ over

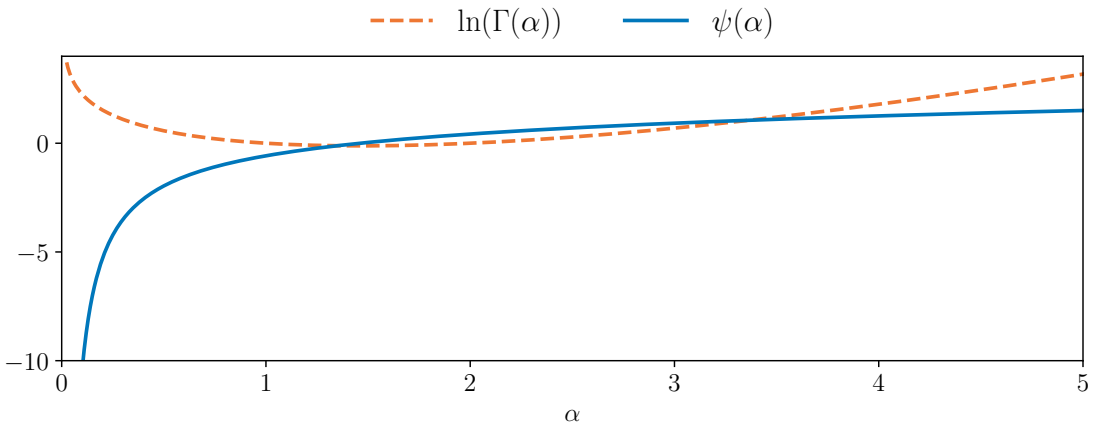


Figure 5.4. *Plot of the Digamma function $\psi(\cdot)$ in the plain blue curve and its derivative $\ln(\Gamma(\cdot))$ (i.e., the logarithm of the Gamma function $\Gamma(\cdot)$) in the dotted orange curve.*

the random choice of $\mathbb{S} \sim \mathcal{D}^m$, we have for all hyper-posterior P on $\mathbb{H}$

$$\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}}(\mathbf{x},y) \leq \overline{\mathrm{kl}}\left(\frac{1}{m}\sum_{i=1}^{m} s_{\mathrm{P}}(\mathbf{x}_i,y_i) \;\middle|\; \frac{\mathrm{KL}(\mathrm{P}\|\Pi)+\ln\frac{2\sqrt{m}}{\delta}}{m}\right),$$

$$\text{with}\quad \mathrm{KL}(\mathrm{P}\|\Pi) = \sum_{j=1}^{\mathrm{card}(\mathbb{H})} \ln[\Gamma(\beta_j)] - \ln\left[\Gamma\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\beta_j\right)\right] - \sum_{j=1}^{\mathrm{card}(\mathbb{H})} \ln[\Gamma(\alpha_j)] + \ln\left[\Gamma\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\alpha_j\right)\right] + \sum_{j=1}^{\mathrm{card}(\mathbb{H})} (\alpha_j-\beta_j)\left[\psi(\alpha_j)-\psi\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\alpha_j\right)\right],$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ is the Gamma function and the Digamma function $\Psi(\alpha)$ is defined as the derivative of $\ln[\Gamma(\alpha)]$; these two functions are plotted in Figure 5.4.

Proof. Deferred to Appendix E.5. ■

Note, while the bound shown in this theorem is based on SEEGER's form, our contribution does not restrict the choice of generalization bounds. Indeed, one can derive other versions of the bound based on MCALLESTER or CATONI's approach. As we will see in Section 5.3.3, we make use of Theorem 5.3.1 to derive a new learning algorithm for the stochastic majority vote.

As we have seen previously (notably in Section 2.3), the higher the KL divergence $\mathrm{KL}\left(\mathrm{P}\|\Pi\right)$, the more different the two distributions P and Π are. In this context, an increase of the Dirichlet parameters $\boldsymbol{\alpha}$ can result in the increase of the KL divergence and a concentration of the Dirichlet distribution (see Figure 5.3). Indeed, when the parameters $\boldsymbol{\alpha}$ increase, the stochastic majority vote risks tends to the risk of the majority vote with weight $\frac{\alpha_i}{\|\boldsymbol{\alpha}\|_1}$ for the voter i. However, at the same time, the parameters increasing, favour a large KL divergence: the bound is a trade-off between the risk's concentration and the KL divergence. For now, Theorem 5.3.1 has a major drawback: it assumes that the voters are not dependent on the learning sample $\mathbb{S}$. Hence, to overcome this issue, we derive in Section 5.3.2 a PAC-Bayesian bound that allows us to have data-dependent voters.

5.3.2 A PAC-Bayesian Bound for Data-dependent Voters

Importantly, Theorem 5.3.1 holds when the hyper-prior Π and the set of voters $\mathbb{H}$ are defined *apriori*, *i.e.*, they are independent from the data $\mathbb{S} \sim \mathcal{D}^m$. However, it is known that considering a data-dependent prior can lead to tighter PAC-Bayes bounds (Parrado-Hernández *et al.*, 2012; Dziugaite *et al.*, 2021). Following recent works on PAC-Bayesian bounds with data-dependent priors (Thiemann *et al.*, 2017; Mhammedi *et al.*, 2019), we derive a generalization bound that allows us to learn the voters from an additional set. More precisely, we consider two independent training sets $\mathbb{S}_1$ and $\mathbb{S}_2$ and we learn a set of voters on each training set (determining the set of voters $\mathbb{H}_1$ and $\mathbb{H}_2$). We refer to the hyper-prior distribution over $\mathbb{H}_1$ *resp.* over $\mathbb{H}_2$ as Π_1 *resp.* Π_2. In the same way, we can then define a hyper-posterior distribution per voters' set: P_1 and P_2. The following theorem shows that we can bound the risk of two combined stochastic majority votes, as long as their empirical risks are evaluated on the data split that was not used for learning their respective voters.

Theorem 5.3.2 (PAC-Bayesian bound with data-dependent voters)**.** Let Π_1 and P_1 be the hyper-prior and hyper-posterior distributions on $\mathbb{H}_1$ defined with $\mathbb{S}_1$, and Π_2 and P_2 the prior and posterior distributions on $\mathbb{H}_2$ defined with $\mathbb{S}_2$. For any $\lambda \in [0,1]$ and $\delta \in (0,1]$ with probability at least $1-\delta$ over samples $\mathbb{S}_1 \sim \mathcal{D}^{m_1}$ and $\mathbb{S}_2 \sim \mathcal{D}^{m_2}$, we have

$$\lambda \mathop{\mathbb{E}}_{\rho\sim \mathrm{P}_1} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) + (1-\lambda) \mathop{\mathbb{E}}_{\rho'\sim \mathrm{P}_2} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_{\rho'}) \leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\lambda s_{\mathrm{P}_1}(\mathbf{x},y) + (1-\lambda) s_{\mathrm{P}_2}(\mathbf{x},y)\right] \leq$$

$$\overline{\mathrm{kl}}\left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}_1} \frac{s_{\mathrm{P}_1}(\mathbf{x},y)}{\frac{1}{\lambda}} + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}_2} \frac{s_{\mathrm{P}_2}(\mathbf{x},y)}{\frac{1}{1-\lambda}} \;\middle|\; \frac{\mathrm{KL}(\mathrm{P}_1\|\Pi_1) + \ln\frac{4\sqrt{m}}{\delta}}{\frac{m}{\lambda}} + \frac{\mathrm{KL}(\mathrm{P}_2\|\Pi_2) + \ln\frac{4\sqrt{m'}}{\delta}}{\frac{m'}{1-\lambda}} \right].$$

Proof. Deferred to Appendix E.6. ■

Following Mhammedi *et al.* (2019) we set $\lambda = 0.5$ and we applied a 50%/50% split in the training data. In the case when the number of data points is odd, we evaluate the bound with $m_2 = \mathrm{card}(\mathbb{S}_2) - 1$ that still gives a correct bound (but simplifies the expression of the bound). In this case the bound is given in the following corollary.

Corollary 5.3.1 (PAC-Bayesian bound with data-dependent voters)**.** Let Π_1 and P_1 be the hyper-prior and hyper-posterior distributions on $\mathbb{H}_1$, and Π_2 and P_2 the prior and posterior distributions on $\mathbb{H}_2$. For any $\delta \in (0,1)$ with probability at least $1-\delta$ over samples $\mathbb{S}_1 \sim \mathcal{D}^{m_1}$ and $\mathbb{S}_2 \sim \mathcal{D}^{m_2}$, we have

$$\frac{1}{2}\left[\mathop{\mathbb{E}}_{\rho\sim \mathrm{P}_1} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) + \mathop{\mathbb{E}}_{\rho'\sim \mathrm{P}_2} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_{\rho'})\right] \leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \frac{1}{2}\left[s_{\mathrm{P}_1}(\mathbf{x},y) + s_{\mathrm{P}_2}(\mathbf{x},y)\right] \leq$$
$$\overline{\mathrm{kl}}\left[\frac{1}{2}\left(\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}_1} s_{\mathrm{P}_1}(\mathbf{x},y) + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}_2} s_{\mathrm{P}_2}(\mathbf{x},y)\right) \,\middle|\, \frac{\mathrm{KL}(\mathrm{P}_1\|\Pi_1) + \mathrm{KL}(\mathrm{P}_2\|\Pi_2) + 2\ln\frac{4\sqrt{m}}{\delta}}{m}\right],$$

where $m = 2\lfloor \frac{m_1+m_2}{2} \rfloor$ and $\lfloor\cdot\rfloor$ is the floor function.

As for Theorem 5.3.1, the true risk of two combined stochastic majority votes is upper-bounded by a PAC-Bayesian bound that depends on two terms. Indeed, it depends on the empirical risk of the two stochastic majority votes and two KL divergences between the hyper-priors and the hyper-posteriors. This bound is evaluated in practice (in Section 5.4) when considering the data-dependent voters.

5.3.3 Learning Algorithms for the Stochastic Majority Vote

As in Chapters 3 and 4, we derive a self-bounding algorithm (Freund, 1998) from Theorem 5.3.1 and Corollary 5.3.1. We based the derivation such an algorithm on the stochastic gradient descent. To do so, we consider mini-batches $\mathbb{U} \subseteq \mathbb{S}$ to optimize the generalization bounds. More precisely, the considered objective function for Theorem 5.3.1 is

$$G_{\mathcal{U}}(\mathrm{P}) \triangleq \overline{\mathrm{kl}}\left(\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{U}} s_{\mathrm{P}}(\mathbf{x},y) \,\middle|\, \frac{\mathrm{KL}(\mathrm{P}\|\Pi) + \ln\frac{2\sqrt{m}}{\delta}}{m}\right), \tag{5.4}$$

which is the upper bound applied on the mini-batch $\mathbb{U}$. To optimize such an objective function, we apply Algorithm 5.3 in conjunction with Algorithm 5.1 when we approximate the empirical stochastic risks or Algorithm 5.2 when we compute the risk exactly; the algorithm is summarized in the following algorithm.

Algorithm 5.3 Minimization of Theorem 5.3.1's Bound

Given: learning sample $\mathbb{S}$, hyper-prior distribution Π on $\mathbb{H}$
Hyperparameters: number of iterations T
$\mathrm{P} \leftarrow \Pi$
for $t \leftarrow 1$ to T **do**
  for all mini-batches $\mathbb{U} \subseteq \mathbb{S}$ **do**
    Compute the empirical stochastic risk $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{U}}\, s_{\mathrm{P}}(\mathbf{x},y)$
    with Algorithm 5.1 or Algorithm 5.2
    $\mathrm{P} \leftarrow$ Update P with $G_{\mathcal{U}}(\mathrm{P})$ by gradient descent
return P

For each iteration we compute the empirical risk on the mini-batch $\mathbb{U}$. To do so, the risk is either approximated with Algorithm 5.1 or computed exactly with Algorithm 5.2. Then, we compute the objective function and update the hyper-posterior P with a gradient descent algorithm. For the data-dependent voters version, similarly as Corollary 5.3.1 which relies on the presence of two learning samples $\mathbb{S}_1$ and $\mathbb{S}_2$, the objective function is defined as

$$G_{\mathcal{U}_1,\mathcal{U}_2}(\mathrm{P}) \triangleq \overline{\mathrm{kl}}\left[\frac{1}{2}\left(\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{U}_1} s_{\mathrm{P}_1}(\mathbf{x},y) + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{U}_2} s_{\mathrm{P}_2}(\mathbf{x},y)\right) \middle| \frac{\mathrm{KL}(\mathrm{P}_1\|\Pi_1) + \mathrm{KL}(\mathrm{P}_2\|\Pi_2) + 2\ln\frac{4\sqrt{m}}{\delta}}{m}\right]. \quad (5.5)$$

This objective function estimated through a mini-batch $\mathbb{U}_1$ from $\mathbb{S}_1$ and $\mathbb{U}_2$ from $\mathbb{S}_2$. Similarly as Equation (5.4), the objective function in Equation (5.5) estimates the upper-bound of Theorem 5.3.2. The algorithm considered to minimize such a bound is described in Algorithm 5.3.

Algorithm 5.4 Minimization of Corollary 5.3.1's Bound

Given: learning samples $\mathbb{S}_1$ and $\mathbb{S}_2$, hyper-priors Π_1 on $\mathbb{H}_1$ and Π_2 on $\mathbb{H}_2$
Hyperparameters: number of iterations T
$(\mathrm{P}_1, \mathrm{P}_2) \leftarrow (\Pi_1, \Pi_2)$
for $t \leftarrow 1$ to T **do**
  for all mini-batches $\mathbb{U}_1 \subseteq \mathbb{S}_1$ and $\mathbb{U}_2 \subseteq \mathbb{S}_2$ **do**
    Compute the empirical stochastic risks $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{U}_1}\, s_{\mathrm{P}_1}(\mathbf{x},y)$ and
    $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{U}_2}\, s_{\mathrm{P}_2}(\mathbf{x},y)$ with Algorithm 5.1 or Algorithm 5.2
    $(\mathrm{P}_1, \mathrm{P}_2) \leftarrow$ Update P_1 and P_2 with $G_{\mathcal{U}_1,\mathcal{U}_2}(\mathrm{P})$ by gradient descent
return $(\mathrm{P}_1, \mathrm{P}_2)$

For each iteration of the algorithm, the two stochastic risks are computed with Algorithm 5.1 or Algorithm 5.2. Then, thanks to these two values, we can compute the objective function $G_{\mathcal{U}_1,\mathcal{U}_2}(\mathrm{P})$ (Equation (5.5)). Finally, the hyper-posteriors P_1 and P_2 are updated through a gradient descent step.

Computing the derivatives. When the risk is computed from Algorithm 5.1, the sum is differentiable. However, since the risk is obtained by a Monte Carlo sampling, we use the implicit reparameterization trick (Figurnov *et al.*, 2018; Jankowiak and Obermeyer, 2018) to obtain the derivatives; it is directly implemented in the automatic differentiation framework such as `PyTorch` (Paszke *et al.*, 2019). Moreover, when the closed form solution (derived in Corollary 5.2.1) risk is computed through Algorithm 5.2, the risk depends on the function $I_{0.5}()$ which is differentiable as well (see Boik and Robinson-Cox, 1999).

In the following section, we evaluate Algorithms 5.3 and 5.4 and compared them with the algorithms of Chapter 4.

5.4 Experiments

In this section, we compare the generalization bounds and the test risks obtained with our algorithms and the ones in Chapter 4. We show that our algorithms allow us to derive generalization bounds that are tight and non-vacuous (*i.e.*, smaller than 1) with decision stumps and decision trees.

We consider as baselines the following PAC-Bayesian self-bounding algorithms:

(i) The algorithm `Germain` that minimizes the PAC-Bayesian of Germain *et al.* (2015, PAC-Bound 0) on $2r_{\mathcal{D}}(\rho)$ (Theorem 2.2.1),

(ii) The algorithm `Masegosa` of Masegosa *et al.* (2020) minimizing a PAC-Bayesian bound on $4e_{\mathcal{D}}(\rho)$ (Theorem 2.2.2),

(iii) Algorithm 4.2 minimizing the PAC-Bayesian C-Bound based on Seeger's approach (Theorem 4.3.4),

(iv) `Lacasse` (Lacasse *et al.*, 2010) minimizes Equation (5.1) for a majority vote that samples N voters from ρ.

The parameters of the baseline are the same as in Section 4.5. For `Lacasse` the number of voters drawn is set to $N{=}100$. As in Chapter 4 the generalization bounds are evaluated with $\delta = 0.05$ and the sigmoid's slope parameter c is set to 100 for Algorithm 5.2. Moreover, the values are averaged over 10 runs.

5.4.1 Comparison Between the Computations of the Risk

For this set of experiments, we optimize Theorem 5.3.1 with Algorithm 5.3, for $T = 2,000$ iterations with COCOB-Backprop optimizer (Orabona and Tommasi, 2017). We study the performance of our method on the binary classification moons dataset, with 2 features, 2 classes and $\mathcal{N}(0, 0.05)$ Gaussian noise, for which we draw m points for training, and $\mathrm{card}(\mathbb{T}) = 2,000$ points for testing.

Figure 5.5 reports a comparison of Algorithms 5.1 and 5.2 in terms of test risk $\mathbb{E}_{\rho\sim\mathrm{P}} \mathrm{R}_{\mathcal{T}}(\mathrm{MV}_\rho)$, PAC-Bayesian generalization bound and training time when the number of decision stumps increases (with $m = 2,000$). We observe that the test risks and bound values can degrade for higher values of decision stumps for all methods. This is due to the KL divergence increasing with the number of voters $\mathrm{card}(\mathbb{H})$, as highlighted in Appendix E.7, becoming a too strong regularization during training and making the bound looser. Moreover, when the number of decision stumps increases, Algorithm 5.2 can be quicker than Algorithm 5.1 especially when the Monte Carlo draws K is high compared to m.

We report in Figure 5.6 the evolution of the test risk $\mathbb{E}_{\rho\sim\mathrm{P}} \mathrm{R}_{\mathcal{T}}(\mathrm{MV}_\rho)$, PAC-Bayesian bounds and training time when m increases. When m is large enough, Algorithm 5.1 achieves comparable test risk and bound values compared to Algorithm 5.2 even for $K = 1$. Increasing the number of Monte Carlo draws K unsurprisingly allows to recover Algorithm 5.2's performance, and at lower computational cost for reasonable values of m and K.

5.4.2 Performance of Algorithms 5.3 and 5.4

We now compare Algorithms 5.3 and 5.4 on different datasets namely FashionMNIST (Xiao *et al.*, 2017), MNIST (LeCun *et al.*, 1998) and datasets coming from the UCI repository (Dua and Graff, 2017); the processing of the dataset is the same as in Section 4.5. More precisely, the same number of examples is kept in the test or the train set as in the original split. When no original split was proposed, we use 50% of data in the training set $\mathbb{S}$ and 50% in the test set (except for Sensorless where we have 15% in the test set). When making use of *data-independent voters*, we chose decision stumps as voters; When making use of *data-dependent voters*, we build decision trees as set of voters without bounding their maximal depth (unless stated otherwise). The voters are exactly the same as in Chapter 4.

We train the stochastic majority vote models by Stochastic Gradient Descent (SGD) using COCOB-Backprop (Orabona and Tommasi, 2017) with batch size equal to 64. We fix the number of epochs to 20 and for Algorithm 5.1 we fix $K = 10$ to increase randomness.

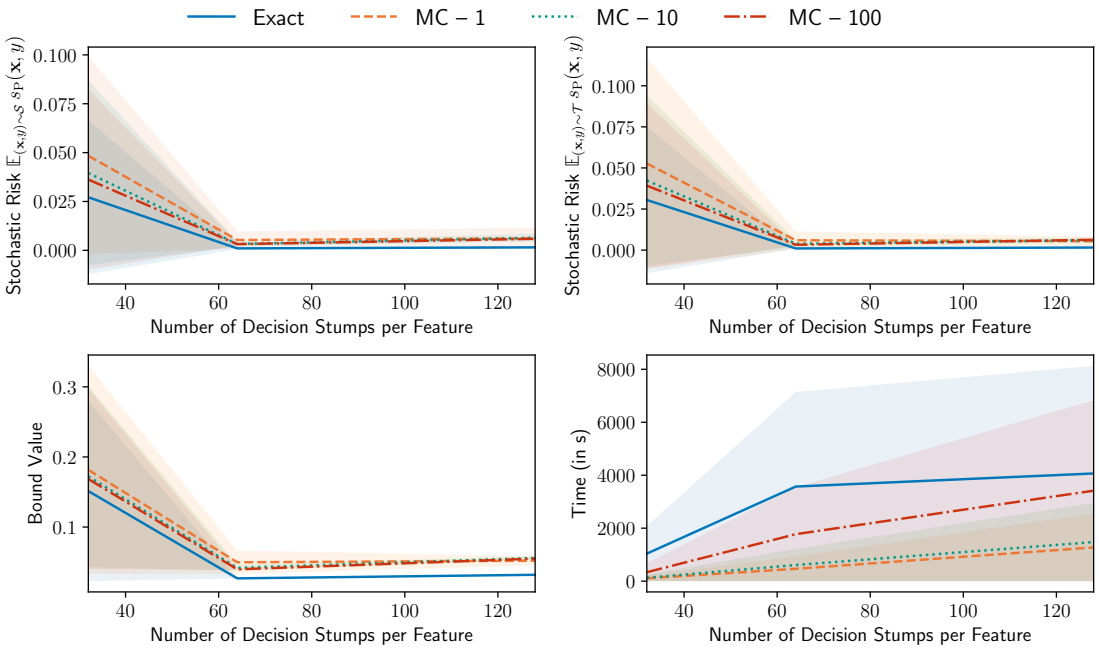


Figure 5.5. *Plot of the average performance on 10 runs of Algorithm 5.1 (with $K \in \{1, 10, 100\}$) and Algorithm 5.2 as a function of the number of decision stumps per feature with learning sample size $m = 2000$.*

We report the stochastic test risks $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{T}}\, s_{\mathrm{P}}(\mathbf{x}, y)$ for our algorithms, the test risk $\mathrm{R}_{\mathcal{T}}(\mathrm{MV}_\rho)$ for the others and the generalization bounds in Figures 5.7 to 5.10 (additional results are reported in Appendix E.7). More precisely, we compare the different self-bounding algorithms with Algorithm 5.3 on binary datasets in Figures 5.7 and 5.8, and on multi-class datasets with data-dependent voters and Algorithm 5.4 in Figures 5.9 and 5.10. First, we remark that Algorithms 5.3 and 5.4 have similar performance in terms of stochastic test risks and bound values. Moreover, note that we notice that the bounds obtained from Algorithms 5.3 and 5.4 are consistently non vacuous and tighter than those obtained with the other algorithms. While the risks between the methods are not comparable we remark that the stochastic test risk $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{T}}\, s_{\mathrm{P}}(\mathbf{x}, y)$ has similar values than the test risks $\mathrm{R}_{\mathcal{T}}(\mathrm{MV}_\rho)$. It means that our algorithms obtain similar test risk $\mathrm{R}_{\mathcal{T}}(\mathrm{MV}_\rho)$ in expectation (where $\rho \sim \mathrm{P}$). We believe that it is due to the fact that the stochastic risks depend on the $\frac{1}{2}$-margin of Laviolette *et al.* (2017). Indeed, even if our learning algorithms optimizes the 01-loss, it does not fully distinguish examples that are classified correctly or not.

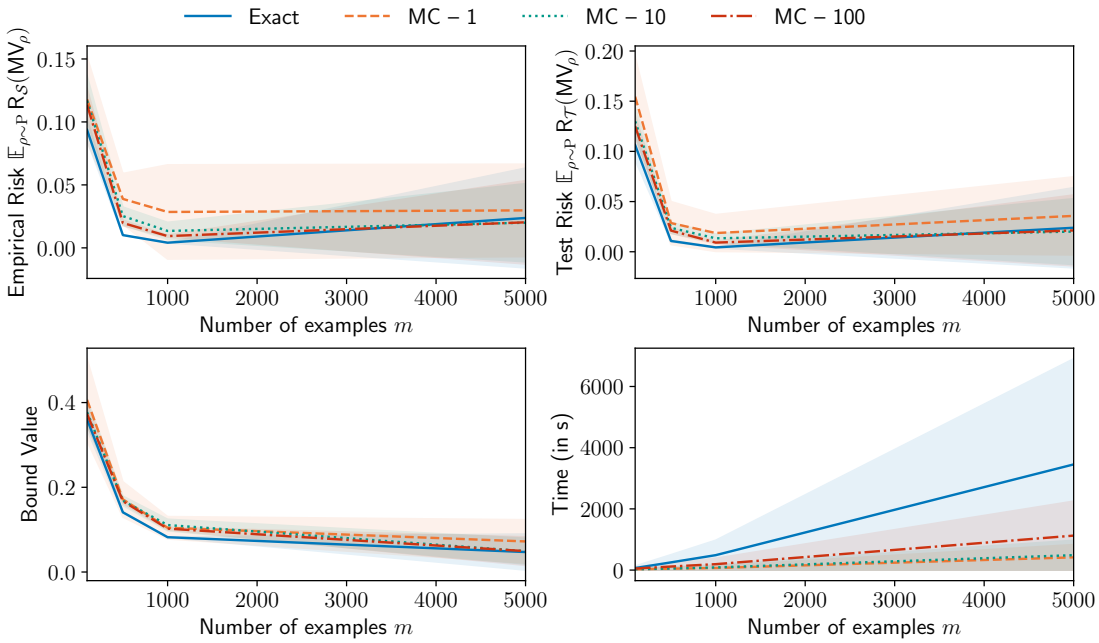


Figure 5.6. *Plot of the average performance on 10 runs of Algorithm 5.1 (with $K \in \{1, 10, 100\}$) and Algorithm 5.2 as a function of the learning sample size m with 32 decision stump per feature.*

5.5 Conclusion and Summary

In this chapter, we have studied a new type of majority vote: the *stochastic majority vote*. For each input $\mathbf{x} \in \mathcal{X}$, it samples a majority vote MV_ρ from a probability distribution called hyper-posterior P and output $\mathrm{MV}_\rho(\mathbf{x})$. When the hyper-posterior is a Dirichlet distribution, the stochastic risk can be either approximated or computed exactly. This allows us to derive a self-bounding algorithm for the stochastic majority vote. The experiments show that our learning algorithm provides a tight PAC-Bayesian generalization bound along with a small empirical risk.

One of the perspectives of this work is to consider the risk of the stochastic majority vote by doing without the $\frac{1}{2}$-margin of Laviolette *et al.* (2017). However, we may not find the closed-form solution of the stochastic majority vote's risk, in this context. In other words, the risk might be only approximated by Monte Carlo sampling. To avoid such a drawback we can make use of a different type of bounds: the *disintegrated* PAC-Bayesian bounds. Indeed, in Part III, we study the *disintegrated* bounds in more details. As recalled in Chapter 2, these bounds have been introduced by Blanchard and Fleuret (2007) and Catoni (2007) and have been rediscovered lately

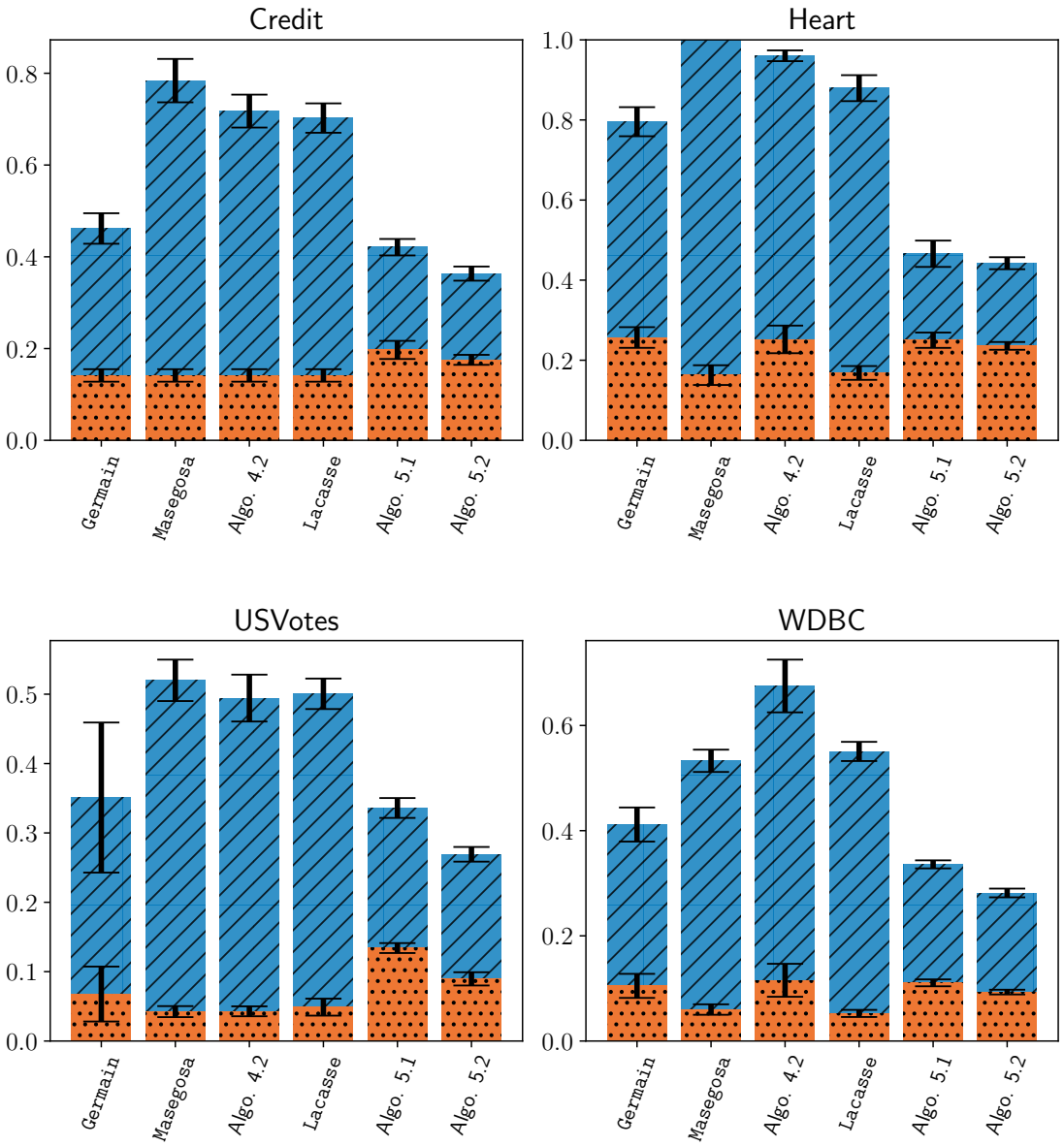


Figure 5.7. *Comparison in terms of test risks, stochastic test risks and bound values. We report in the dotted orange error bars, the means and standard deviations of the test risks and stochastic risks. Moreover, the hatched blue error bars represent the mean and the standard deviations of the PAC-Bayesian bounds. The values are average over 10 different runs.*

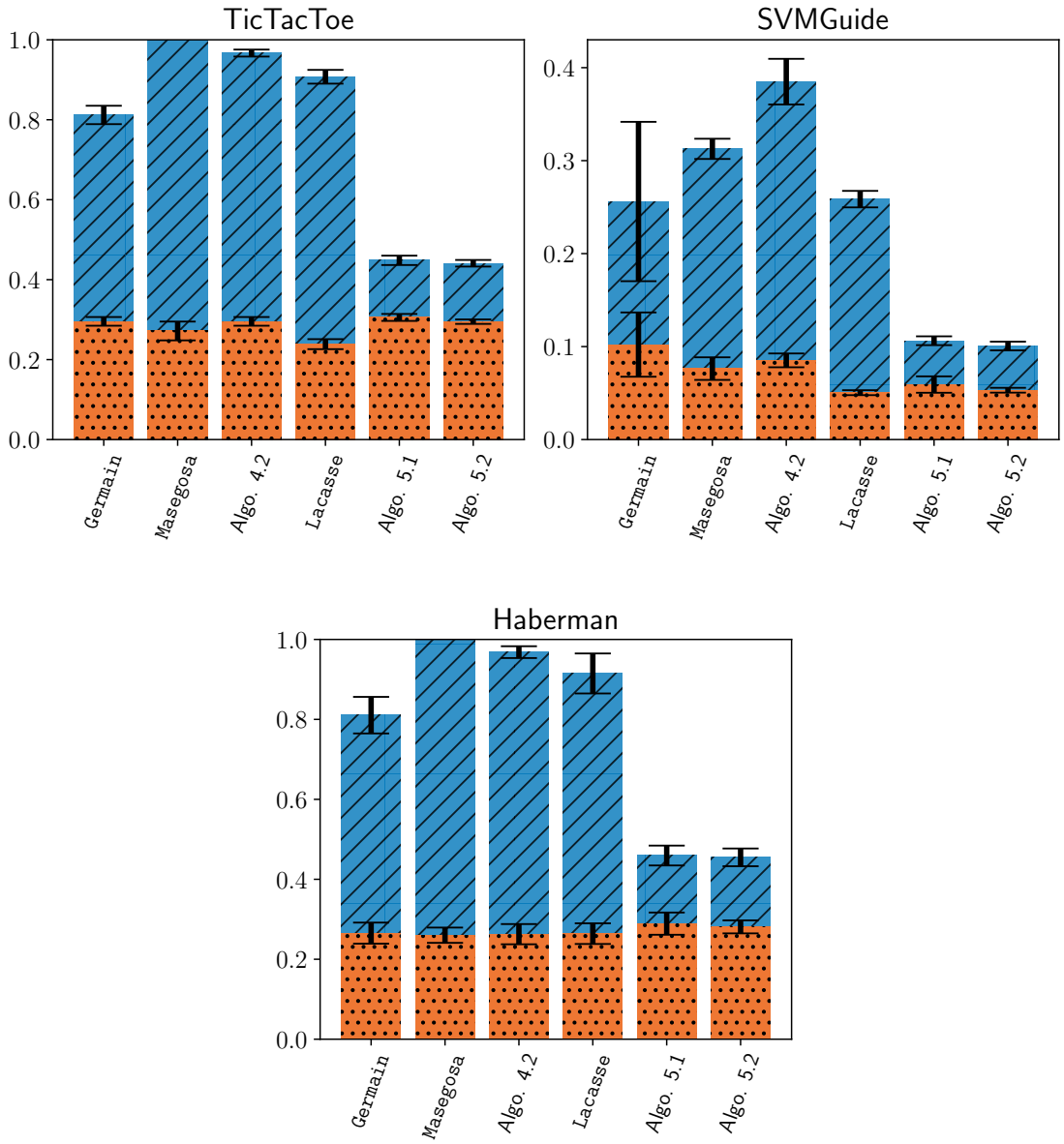


Figure 5.8. *Comparison in terms of test risks, stochastic test risks and bound values. We report in the dotted orange error bars, the means and standard deviations of the test risks and stochastic risks. Moreover, the hatched blue error bars represent the mean and the standard deviations of the PAC-Bayesian bounds. The values are average over 10 different runs.*

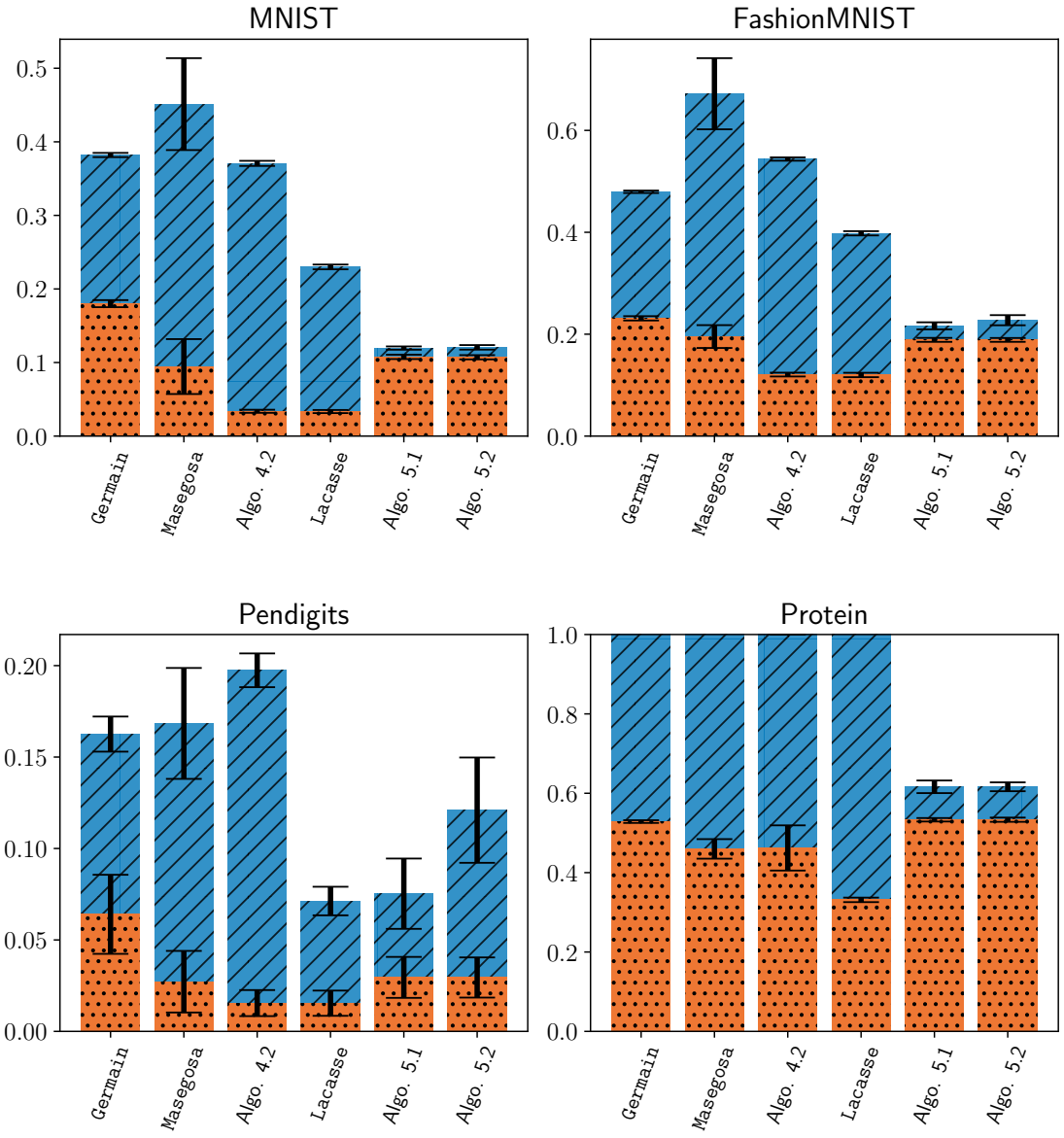


Figure 5.9. *Comparison in terms of test risks, stochastic test risks and bound values. We report in the dotted orange error bars, the means and standard deviations of the test risks and stochastic risks. Moreover, the hatched blue error bars represent the mean and the standard deviations of the PAC-Bayesian bounds. The values are average over 10 different runs.*

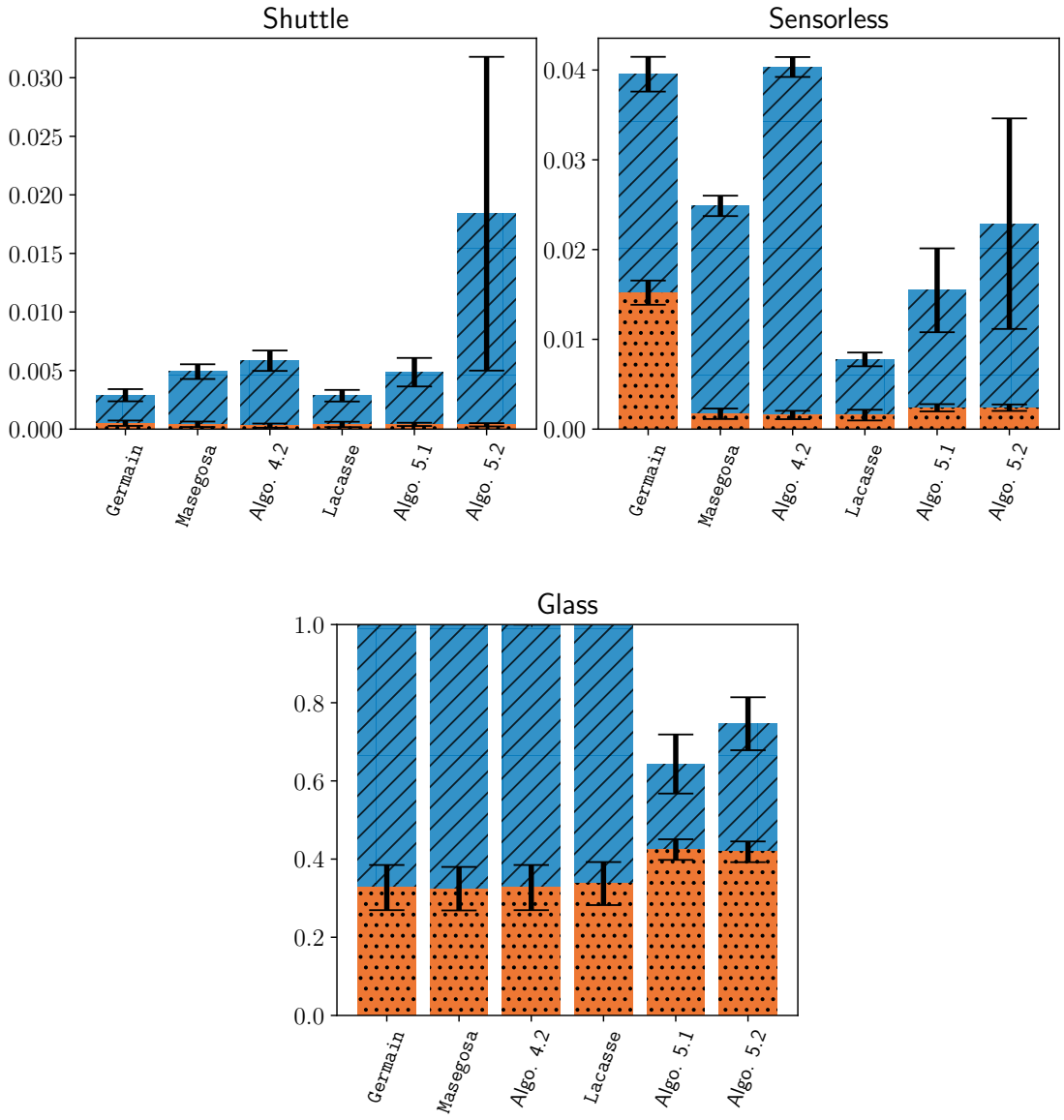


Figure 5.10. *Comparison in terms of test risks, stochastic test risks and bound values. We report in the dotted orange error bars, the means and standard deviations of the test risks and stochastic risks. Moreover, the hatched blue error bars represent the mean and the standard deviations of the PAC-Bayesian bounds. The values are average over 10 different runs.*

by Rivasplata *et al.* (2020). They allow us to derive a bound for a *single* voter or hypothesis from the hypothesis set $\mathbb{H}$. Chapter 6 introduces new *disintegrated* bounds based on the Rényi divergence (that can be used for the stochastic majority vote) and Chapter 7 presents generalization bounds that are not restrained to the KL divergence or the Rényi one but can depend on a complexity term defined by the users.

Part III

Beyond PAC-Bayesian Bounds:
From Disintegration to Novel Bounds

6 On the Practical uses of the Disintegrated Bounds

This chapter is based on the following paper

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A General Framework for the Disintegration of PAC-Bayesian Bounds. *Submitted to Machine Learning Journal*. (2022b)

Contents

6.1 Introduction **150**
6.2 Setting and PAC-Bayesian Bounds **151**
6.3 Disintegrated PAC-Bayesian Theorems **152**
6.3.1 Form of a Disintegrated PAC-Bayesian Bound 152
6.3.2 Disintegrated Bounds with the Rényi Divergence 153
6.4 The Disintegration in Action **158**
6.4.1 Specialization to Neural Network Classifiers 158
6.4.2 A Note About Stochastic Neural Networks 161
6.5 Experiments with Neural Networks **162**
6.5.1 Training Method 162
6.5.2 Optimization Procedure in Algorithms A and A_{prior} 163
6.5.3 Experimental Setting 163
6.5.4 Results 164
6.6 Perspectives for the Majority Vote **171**
6.7 Summary and Conclusion **172**

Abstract

PAC-Bayesian bounds are known to be tight and informative when studying the generalization ability of stochastic classifiers (see *e.g.*, Chapter 5). However, they require a loose and costly derandomization step when applied to some families of deterministic models such as neural networks. As an alternative to this step, we introduce new PAC-Bayesian generalization bounds that have the originality to provide *disintegrated* bounds, *i.e.*, they give guarantees over one *single* hypothesis instead of the usual *averaged* analysis. Our bounds are easily optimizable and can be used to design self-bounding algorithms. We illustrate this behavior on neural networks and show a significant practical improvement over the state-of-the-art framework.

6.1 Introduction

The PAC-Bayesian theory is a powerful framework for upper-bounding the true risk of stochastic models such as the stochastic majority vote (considered in Chapter 5). Remember that, in general, the stochastic model samples an hypothesis from the posterior distribution and predicts a label with the sampled hypothesis. However, the vast majority of machine learning methods nevertheless need guarantees on deterministic models (*i.e.* that are not stochastic). In this case, a *derandomization step* of the bound is required to get a bound on the deterministic model's risk. In general, the *derandomization step* consists in obtaining a bound on the risk of a deterministic model from a bound originally valid for stochastic models. Different forms of derandomization have been introduced in the literature for specific settings. Among them, Langford and Shawe-Taylor (2002) proposed a derandomization for Gaussian posteriors over linear classifiers: thanks to the Gaussian symmetry, a bound on the risk of the *maximum a posteriori* (deterministic) classifier is obtainable from the bound on the average risk of the stochastic classifier. Also relying on Gaussian posteriors, Letarte *et al.* (2019) derived a PAC-Bayesian bound for a very specific deterministic network architecture using sign functions as activations; this approach has been further extended by Biggs and Guedj (2021, 2022). Another line of works derandomizes neural networks (Neyshabur *et al.*, 2018; Nagarajan and Kolter, 2019b). While technically different, it starts from PAC-Bayesian guarantees on the stochastic classifier and uses an "output perturbation" bound to convert a guarantee from a random classifier to the mean classifier. The relative diversity and specificity of these works highlight nevertheless the lack of a general framework for the derandomization of classic PAC-Bayesian bounds.

This chapter focuses on another kind of derandomization through the *disintegration of the PAC-Bayesian bound*, proposed by Catoni (2007, Th.1.2.7) and Blanchard and Fleuret (2007); see Section 2.4. Despite their interest in derandomizing PAC-Bayesian bounds, these kinds of bounds have only received little study in the literature and have never been used in practice. Driven by machine learning practical purposes, our objective is thus twofold. We derive new tight and usable *disintegrated* PAC-Bayesian bounds *(i)* that directly derandomize any classifiers without any other additional step and with *almost* no impact on the guarantee, and *(ii)* that can be easily optimized to learn classifiers with strong guarantees. To achieve this objective, our contribution consists in providing a new general disintegration framework based on the Rényi divergence (in Theorem 6.3.1), allowing us to meet the practical goal of efficient learning. From a theoretical standpoint, due to the Rényi divergence term, our bound is expected to be looser than the one of Rivasplata *et al.* (2020, Th.1*(i)*) in which the divergence term is "disintegrated" but depends on the sampled hypothesis only. However, as we show in our experimental evaluation on neural networks, their

"disintegrated" term is, in practice, subject to high variance, making their bound harder to optimize. This variance arises because the sampled hypothesis does not influence our RÉNYI divergence term. Our bound has then the main advantage of leading to a more stable learning algorithm with better empirical results. In addition, we derive in Appendix F.9 new theoretical results based on the mutual information, giving different insights into disintegration procedures.

The rest of the chapter is organized as follows. Section 6.2 recalled the notations we follow. In Section 6.3, we derive our main contribution relying on *disintegrated* PAC-Bayesian bounds. Then, we illustrate the practical usefulness of this disintegration on deterministic neural networks in Section 6.5. Before concluding in Section 6.7, we discuss in Appendix F.9 another point of view of the disintegrated through an information-theoretic bound. For readability, we deferred the proofs of our theoretical results to Appendix F.

6.2 Setting and PAC-Bayesian Bounds

In this chapter, we consider supervised classification tasks as described in Chapter 2 with $\mathbb{X}$ the input space, $\mathbb{Y}$ the label set, and $\mathcal{D}$ an unknown data distribution on $\mathbb{X}\times\mathbb{Y}$. We consider a hypothesis set $\mathbb{H}$ of functions $h : \mathbb{X} \to \mathbb{Y}$. The learner aims to find $h \in \mathbb{H}$ that assigns a label y to an input $\mathbf{x}$ as accurately as possible. Given an example $(\mathbf{x}, y)$ and a hypothesis h, we assess the quality of the prediction of h with a loss function $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0, 1]$ evaluating to which extent the prediction is accurate. The learner wants to find the hypothesis h from $\mathbb{H}$ that minimizes the true risk $\mathrm{R}^{\ell}_{\mathcal{D}}(h)$. However, we cannot compute $\mathrm{R}^{\ell}_{\mathcal{D}}(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\, \ell(h, (\mathbf{x}, y))$ since the distribution $\mathcal{D}$ is unknown. We only have access to a learning sample $\mathbb{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ with the empirical risk defined as $\mathrm{R}^{\ell}_{\mathbb{S}}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, (\mathbf{x}_i, y_i))$.

We now recall BÉGIN *et al.* (2016)'s general bound at the heart of our contribution (and introduced in Section 2.3). This bound depends on the RÉNYI divergence between ρ and π defined as $\mathrm{D}_{\lambda}(\rho\|\pi) \triangleq \frac{1}{\lambda-1} \ln\left[\mathop{\mathbb{E}}_{h\sim\pi}\left[\frac{\rho(h)}{\pi(h)}\right]^{\lambda}\right]$ with parameter $\lambda > 1$.

Theorem 2.3.5 (General PAC-Bayesian Bound of BÉGIN *et al.* (2016))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any measurable function $\varphi : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}^+_*$, for any $\lambda > 1$, for any

$\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \in \mathbb{M}(\mathbb{H}), \frac{\lambda}{\lambda-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})\right] \right.$$
$$\left. \leq \mathrm{D}_\lambda(\rho\|\pi) + \ln\left[\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right]\right] \geq 1-\delta.$$

A key notion here is that the PAC-Bayesian bounds apply on the expectation over the risk of the individual classifiers in $\mathbb{H}$; this randomization is the risk of the stochastic classifier. A key issue for usual machine learning tasks is then the derandomization of the PAC-Bayesian bounds to obtain a guarantee for a deterministic classifier instead of a stochastic one (by removing the expectation on $\mathbb{H}$). In some cases, this derandomization results from the structure of the hypotheses, such as for stochastic linear classifiers that can be directly expressed as one deterministic linear classifier (Germain *et al.*, 2009). However, in other cases, the derandomization is much more complex and specific to the class of hypotheses, such as for neural networks (*e.g.*, Neyshabur *et al.* (2018), Nagarajan and Kolter (2019a, Ap. J), Biggs and Guedj (2022)).

The next section states our main contribution to this chapter: a general derandomization framework based on the Rényi divergence for disintegrating PAC-Bayesian bounds into a bound for a single hypothesis from $\mathbb{H}$.

6.3 Disintegrated PAC-Bayesian Theorems

6.3.1 Form of a Disintegrated PAC-Bayesian Bound

We recall now the main form of a disintegrated PAC-Bayesian bound used in this chapter.

Definition 2.4.1 (Disintegrated PAC-Bayesian Generalization Bound)**.** Let $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$ be a loss function and $\phi : [0,1]^2 \to [0,1]$ a generalization gap. A *disintegrated* PAC-Bayesian bound is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, there exists a function $\Phi : \mathbb{M}(\mathbb{H})\times\mathbb{M}^*(\mathbb{H})\times(0,1] \to \mathbb{R}$ such that for any $\delta \in (0,1]$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\phi(\mathrm{R}^\ell_{\mathcal{D}}(h), \mathrm{R}^\ell_{\mathcal{S}}(h)) \leq \Phi\Big(\rho_{\mathbb{S}}, \pi, \delta\Big)\right] \geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S}, \pi)$ is output by the deterministic algorithm A and $\phi()$ is, for example, $\phi(\mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell}_{\mathcal{S}}(h)) = \left|\mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathsf{R}^{\ell}_{\mathcal{S}}(h)\right|$.

More precisely, the posterior $\rho_{\mathbb{S}} \triangleq A(\mathbb{S}, \pi)$ is defined through a given *deterministic* algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$ chosen *apriori*. The algorithm *(i)* takes a learning sample $\mathbb{S} \in (\mathbb{X}\times\mathbb{Y})^m$ and a prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ as inputs, and *(ii)* outputs a *data-dependent* distribution $\rho_{\mathbb{S}} \triangleq A(\mathbb{S}, \pi)$. Concretely, this kind of generalization bound allows one to derandomize the usual PAC-Bayes bounds as follows. Instead of considering a bound holding for all the posterior distributions on $\mathbb{H}$ as usually done in PAC-Bayes (the "$\forall\rho$" in Theorem 2.3.5), we consider only the posterior distribution $\rho_{\mathbb{S}}$ obtained through a deterministic algorithm A taking the learning sample $\mathbb{S}$ and the prior π as inputs. Then, the above bound holds for a unique hypothesis $h\sim\rho_{\mathbb{S}}$ instead of the stochastic classifier: the individual risks are no longer averaged with respect to $\rho_{\mathbb{S}}$; this is the *PAC-Bayesian bound disintegration*. The dependence in probability on $\rho_{\mathbb{S}}$ means that the bound is valid with probability at least $1-\delta$ over the random choice of the learning sample $\mathbb{S}\sim\mathcal{D}^m$ and the hypothesis $h\sim\rho_{\mathbb{S}}$. Under this principle, we introduce in Theorems 6.3.1 and 6.3.2 below two new general disintegrated PAC-Bayesian bounds. A key asset of our results is that the bounds are instantiable to specific settings as for the "classical" PAC-Bayesian bounds (*e.g.*, with *i.i.d.*/non-*i.i.d.* data, unbounded losses, etc.). By instantiating such a bound, we obtain an easily optimizable bound, leading to a self-bounding algorithm (Freund, 1998) with theoretical guarantees. As an illustration of the usefulness of our results, we provide, in Section 6.4, such an instantiation for neural networks.

6.3.2 Disintegrated Bounds with the Rényi Divergence

6.3.2.1 Our Main Contribution: a General Disintegrated Bound

In the same spirit as Theorem 2.3.5 our main result stated in Theorem 6.3.1 is a general bound involving the Rényi divergence $\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)$ of order $\lambda > 1$.

Theorem 6.3.1 (General Disintegrated PAC-Bayes Bound)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}^*_+$, for any $\lambda > 1$, for any $\delta \in (0, 1]$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left(\frac{\lambda}{\lambda-1}\ln\left(\varphi(h,\mathbb{S})\right)\right.$$
$$\left.\leq \frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta} + \mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi) + \ln\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]\right) \geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S}, \pi)$ is output by the deterministic algorithm A.

Proof sketch (see Appendix F.1 for details). Recall that $\rho_{\mathbb{S}}$ is obtained with the algorithm $A(\mathbb{S}, \pi)$. Applying MARKOV's inequality (Theorem A.2.1) on $\varphi(h,\mathbb{S})$ with the random variable h and using HÖLDER's inequality (Theorem A.5.1) to introduce $\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)$, we have, with probability at least $1-\frac{\delta}{2}$ on $\mathbb{S}\sim\mathcal{D}^m$ and $h\sim\rho_{\mathbb{S}}$,

$$\begin{aligned}\frac{\lambda}{\lambda-1}\ln\left[\varphi(h,\mathbb{S})\right] &\leq \frac{\lambda}{\lambda-1}\ln\left[\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\varphi(h',\mathbb{S})\right]\\ &\leq \mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi) + \frac{\lambda}{\lambda-1}\ln\frac{2}{\delta} + \ln\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\right].\end{aligned}$$

By applying again MARKOV's inequality (Theorem A.2.1) on $\varphi(h,\mathbb{S})$ with the random variable $\mathbb{S}$, we have, with probability at least $1-\frac{\delta}{2}$ on $\mathbb{S}\sim\mathcal{D}^m$ and $h\sim\rho_{\mathbb{S}}$,

$$\ln\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\right] \leq \ln\left[\frac{2}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right].$$

Lastly, we combine the two bounds with a union bound argument. ■

As for the general classical PAC-Bayesian bounds (Theorem 2.3.5), the above theorem can be seen as the starting point of the derivation of generalization bounds depending on the choice of the function φ, as done in Corollary 6.4.1 in Section 6.4.1; this property makes it the main result of this chapter. In its proof, HÖLDER's inequality (Theorem A.5.1) is used differently than in the classic PAC-Bayes bound's proofs. Indeed, in the proof of BÉGIN *et al.* (2016, Th. 8), the change of measure based on HÖLDER's inequality is key for deriving a bound that holds for all posteriors ρ with high probability, while our bound holds for a unique posterior $\rho_{\mathbb{S}}$ dependent on the sample $\mathbb{S}$ and the prior π. In fact, we use HÖLDER's inequality to introduce a prior π independent from $\mathbb{S}$: a crucial point for our bound instantiated in Corollary 6.4.1. Compared to Theorem 2.3.5, our bound requires an additional term $\ln 2+\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta}$. However, by setting $\varphi(h,\mathbb{S}) = m\ \mathrm{kl}(\mathrm{R}^\ell_{\mathcal{S}}(h)\|\mathrm{R}^\ell_{\mathcal{D}}(h))$ and $\lambda{=}2$, the term $\ln\frac{8}{\delta^2}$ is multiplied by $\frac{1}{m}$, which turns out to be a reasonable cost to "derandomize" a bound into a disintegrated one. For instance, if $m = 5,000$ (a reasonable sample size) and $\delta = 0.05$, we have $\frac{1}{m}\ln\frac{8}{\delta^2} \approx 0.002$.
We instantiate below Theorem 6.3.1 for $\lambda{\to}1^+$ and $\lambda{\to}{+\infty}$ showing that the bound converges when $\lambda{\to}1^+$ and $\lambda{\to}{+\infty}$.

Corollary 6.3.1 (Extreme Cases of Theorem 6.3.1)**.** Under the assumptions of Theorem 6.3.1, when $\lambda \to 1^+$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left(\ln \varphi(h,\mathbb{S}) \leq \ln\frac{2}{\delta} + \ln\left[\mathop{\operatorname{esssup}}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}), h'\in\mathbb{H}} \varphi(h',\mathbb{S}')\right]\right) \geq 1-\delta,$$

when $\lambda \to +\infty$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left(\ln \varphi(h,\mathbb{S}) \leq \ln \mathop{\operatorname{esssup}}_{h'\in\mathbb{H}} \frac{\rho_{\mathbb{S}}(h')}{\pi(h')} + \ln\left[\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi} \varphi(h',\mathbb{S}')\right]\right) \geq 1-\delta,$$

where esssup is the essential supremum defined as the supremum on a set with non-zero probability measures, *i.e.*,

$$\mathop{\operatorname{esssup}}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}), h'\in\mathbb{H}} \varphi(h',\mathbb{S}') = \inf\left\{\tau\in\mathbb{R}, \mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\varphi(h,\mathbb{S}) > \tau\right] = 0\right\},$$

$$\text{and}\quad \mathop{\operatorname{esssup}}_{h'\in\mathbb{H}} \frac{\rho_{\mathbb{S}}(h')}{\pi(h')} = \inf\left\{\tau\in\mathbb{R}, \mathop{\mathbb{P}}_{h\sim\rho_{\mathbb{S}}}\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)} > \tau\right] = 0\right\}.$$

Proof. Deferred to Appendix F.2. ■

This corollary illustrates that the parameter λ controls the trade-off between the Rényi divergence $\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)$ and $\ln\left[\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi}\,\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right]$. Indeed, when $\lambda\to1^+$, the Rényi divergence vanishes ($\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)\to0$) while the other term converges toward $\ln\left[\operatorname{esssup}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}),h'\in\mathbb{H}}\varphi(h',\mathbb{S}')\right]$, roughly speaking the maximal value possible for the second term. On the other hand, when $\lambda\to+\infty$, the Rényi divergence increases and converges toward $\ln\operatorname{esssup}_{h'\in\mathbb{H}}\frac{\rho_{\mathbb{S}}(h')}{\pi(h')}$ and the other term decreases toward $\ln[\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi}\,\varphi(h',\mathbb{S}')]$.

6.3.2.2 Comparison with the Bound of Rivasplata *et al.* (2020)

For the sake of comparison, we recall in Theorem 2.4.1 the bound proposed by Rivasplata *et al.* (2020, Th.1*(i)*), that is more general than the bounds of Blanchard and Fleuret (2007) and Catoni (2007, Th.1.2.7).

Theorem 2.4.1 (General Disintegrated Bound of Rivasplata *et al.* (2020))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi\in\mathbb{M}^*(\mathbb{H})$, for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}$, for any $\delta\in(0,1]$,

for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\varphi(h,\mathbb{S}) \leq \underbrace{\ln\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right] + \ln\left[\frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} \exp\left(\varphi(h',\mathbb{S}')\right)\right]}_{\Phi(\rho_{\mathbb{S}},\pi,\delta)} \right] \geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S},\pi)$ is output by the deterministic algorithm A.

Note that the bound can be rewritten with the logarithm, *i.e.*, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\ln\left(\varphi(h,\mathbb{S})\right) \leq \ln\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right] + \ln\left[\frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} \varphi(h',\mathbb{S}')\right] \right] \geq 1-\delta.$$

The term $\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}$ (also involved in Blanchard and Fleuret (2007) and Catoni (2007)) can be seen as a "disintegrated[1] KL divergence" depending only on the sampled $h\sim\rho_{\mathbb{S}}$. In contrast, our bound involves the Rényi divergence $D_\lambda(\rho_{\mathbb{S}}\|\pi)$ between the prior π and the posterior $\rho_{\mathbb{S}}$, meaning our bound involves only one term that depends on the sample hypothesis (the risk): the divergence value is the same whatever the hypothesis. Our bound is expected to be looser because of the Rényi divergence (see Erven and Harremoës, 2014) and the dependence in δ (which is worse than in Theorem 2.4.1). Nevertheless, our divergence term is the main advantage of our bound. Indeed, as confirmed by our experiments (Section 6.5), our bound with $\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)$ makes the learning procedure (in our self-bounding algorithm) more stable and efficient compared to the optimization of Theorem 2.4.1's bound with $\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}$ that is subject to high variance.

6.3.2.3 A Parameterizable General Disintegrated Bound

In the PAC-Bayesian literature, parameterized bounds have been introduced (*e.g.*, Catoni (2007) and Thiemann *et al.* (2017)) to control the trade-off between the empirical risk and the divergence along with the additional term. For the sake of completeness, we now provide a parameterized version of our bound, enlarging its practical scope. We follow a similar approach to introduce a version of a disintegrated Rényi divergence-based bound that has the advantage of being parameterizable.

Theorem 6.3.2 (Parametrizable Disintegrated PAC-Bayes Bound)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$,

[1] We say that the KL divergence is "disintegrated" since the log term is not averaged in contrast to the KL divergence.

for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}_+^*$, for any $\delta \in (0,1]$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left(\forall\lambda{>}0,\ \ln\left(\varphi(h,\mathbb{S})\right) \le \ln\left[\frac{\lambda}{2}e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)} + \frac{8}{2\lambda\delta^3}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left[\varphi(h',\mathbb{S}')^2\right]\right]\right) \ge 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S},\pi)$ is output by the deterministic algorithm A.

Proof. Deferred to Appendix F.3. ■

Note that $e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)}$ is closely related to the χ^2-distance. Indeed we have: $\chi^2(\rho_{\mathbb{S}}\|\pi) \triangleq \mathbb{E}_{h\sim\pi}\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]^2 - 1 = e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)} - 1$. An asset of Theorem 6.3.2 is the parameter λ controlling the trade-off between the exponentiated RÉNYI divergence $e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)}$ and $\frac{1}{\delta^3}\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi}\varphi(h',\mathbb{S}')^2$. Our bound is valid for all $\lambda > 0$, thus, from a practical view, we can learn/tune the parameter λ to minimize the bound and control the possible numerical instability due to $e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)}$. Indeed, if $\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)$ is large, the numerical computation can lead to an infinite value due to finite precision arithmetic. It is important to notice that, like other parameterized bounds (*e.g.*, THIEMANN *et al.*, 2017), there exists a closed-form solution of the optimal parameter λ (for a fixed π and $\rho_{\mathbb{S}}$); the solution is derived in Proposition 6.3.1 and shows that the optimal bound of Theorem 6.3.2 corresponds to the bound of Theorem 6.3.1.

Proposition 6.3.1 (Optimal Bound of Theorem 6.3.2)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution π on $\mathbb{H}$, for any $\delta\in(0,1]$, for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}_+^*$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, let

$$\lambda^* = \mathop{\mathrm{argmin}}_{\lambda>0} \ln\left[\frac{\lambda}{2}e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)} + \frac{\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left[8\varphi(h',\mathbb{S}')^2\right]}{2\lambda\delta^3}\right],$$

then, we have

$$\overbrace{2\ln\left[\frac{\lambda^*}{2}e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)} + \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\frac{8\varphi(h',\mathbb{S}')^2}{2\lambda^*\delta^3}\right)\right]}^{\text{Theorem 6.3.2}}$$
$$= \underbrace{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi) + \ln\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\frac{8\varphi(h',\mathbb{S}')^2}{\delta^3}\right)\right]}_{\text{Theorem 6.3.1 with } \lambda = 2.},$$

where $\lambda^* = \sqrt{\dfrac{\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{h'\sim\pi}\left[8\varphi(h',\mathbb{S}')^2\right]}{\delta^3\exp(\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi))}}$.

Put into words: the optimal λ^* gives the same bound for Theorem 6.3.1 and Theorem 6.3.2.

Proof. Deferred to Appendix F.4. ■

6.4 The Disintegration in Action

So far, we have introduced theoretical results to derandomize PAC-Bayesian bounds through a disintegration approach. Indeed, the disintegration allows us to obtain a bound for a unique model sampled from $\rho_{\mathbb{S}}$ instead of having a bound on the averaged risk. This section proposes to illustrate the instantiation and usefulness of Theorem 6.3.1 on neural networks compared to the classical PAC-Bayesian bounds.

6.4.1 Specialization to Neural Network Classifiers

We aim to learn the weights of the Neural Networks (NN) leading to the lowest true risk $\mathrm{R}_{\mathcal{D}}(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\,\mathrm{I}\left[h(\mathbf{x}) \neq y\right]$. In other words, we consider that the hypothesis set $\mathbb{H}$ is a set of neural networks with different weights for a given architecture. Practitioners usually proceed by epochs and obtain one "intermediate" NN after each epoch. Then, they select the "intermediate" NN associated with the lowest validation risk. We propose translating this practice into our PAC-Bayesian setting by considering one prior per epoch. Given T epochs, we hence have T priors $\mathbb{P}=\{\pi_t\}_{t=1}^{T}$, where $\forall t\in\{1,\ldots,T\}, \pi_t = \mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_D)$ is a Gaussian distribution centered at $\mathbf{v}_t$ (the weight vector associated with the t-th "intermediate" NN) with a covariance matrix of $\sigma^2\mathbf{I}_D$ (where $\mathbf{I}_D$ is the $D\times D$-dimensional identity matrix). Assuming the T priors are learned from a set $\mathbb{S}_{\text{prior}}$ such that $\mathbb{S}_{\text{prior}}\bigcap\mathbb{S}=\emptyset$, then Corollaries 6.4.1 and 6.4.2 will guide us to learn a posterior $\rho_{\mathbb{S}}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_D)$ from the prior $\pi\in\mathbb{P}$ minimizing the empirical risk on $\mathbb{S}$ (we give more details on the procedure after the forthcoming corollaries). Note that considering Gaussian distributions has the advantage of simplifying the expression of the KL divergence and thus is commonly used in the PAC-Bayesian literature for neural networks (*e.g.*, Dziugaite and Roy, 2017; Letarte *et al.*, 2019; Zhou *et al.*, 2019).[2]

[2]This has been first studied in the context of linear classifiers (*e.g.*, Ambroladze *et al.*, 2006; Germain *et al.*, 2009, 2020). However, in this context the symmetry of the Gaussian distribution also ease the derandomization.

Corollary 6.4.1 below instantiates Theorem 6.3.1 to this neural networks setting. Then, for the sake of comparison, Corollary 6.4.2 instantiates other disintegrated bounds from the literature; more precisely, Equation (6.1) corresponds to Rivasplata *et al.* (2020)'s bound of Theorem 2.4.1, Equation (6.2) to Blanchard and Fleuret (2007)'s one, and Equation (6.3) to Catoni (2007)'s one.

Corollary 6.4.1 (Instantiation of Theorem 6.3.1 for Neural Networks)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any set $\mathbb{P}=\{\pi_1,\ldots,\pi_T\}$ of T priors on $\mathbb{H}$ where $\pi_t=\mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_D)$, for any algorithm $A:(\mathbb{X}\times\mathbb{Y})^m\times\mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, for any loss $\ell:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})\to[0,1]$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left(\forall\pi_t\in\mathbb{P},\ \mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))\leq\frac{1}{m}\left[\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{\sigma^2}+\ln\frac{16T\sqrt{m}}{\delta^3}\right]\right)\geq 1-\delta,$$

where $\mathrm{kl}(a\|b)=a\ln\frac{a}{b}+(1-a)\ln\frac{1-a}{1-b}$, $\rho_{\mathbb{S}}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_D)$, and the hypothesis $h\sim\rho_{\mathbb{S}}$ is parameterized by $\mathbf{w}+\boldsymbol{\epsilon}$.

Proof. Deferred to Appendix F.5. ■

Corollary 6.4.2 (Instantiation of Known Bounds for Neural Networks)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any set $\mathbb{P}=\{\pi_1,\ldots,\pi_T\}$ of T priors on $\mathbb{H}$ where $\pi_t=\mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_D)$, for any algorithm $A:(\mathbb{X}\times\mathbb{Y})^m\times\mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, for any loss $\ell:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})\to\{0,1\}$, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over the learning sample $\mathbb{S}\sim\mathcal{D}^m$ and the hypothesis $h\sim\rho_{\mathbb{S}}$ parameterized by $\mathbf{w}+\boldsymbol{\epsilon}$, we have $\forall\pi_t\in\mathbb{P}$

$$\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))\leq\frac{1}{m}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}+\ln\frac{2T\sqrt{m}}{\delta}\right],\tag{6.1}$$

$$\forall b\in\mathbb{b},\ \mathrm{kl}_+(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))\leq\frac{1}{m}\left[\frac{b+1}{b}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}\right]_+\!+\ln\frac{(b+1)T\,\mathrm{card}(\mathbb{b})}{\delta}\right],\tag{6.2}$$

$$\forall c\in\mathbb{c},\ \mathsf{R}^\ell_{\mathcal{D}}(h)\leq\frac{1-\exp\left(-c\mathsf{R}^\ell_{\mathcal{S}}(h)-\frac{1}{m}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}+\ln\frac{T\,\mathrm{card}(\mathbb{c})}{\delta}\right]\right)}{1-e^{-c}},\tag{6.3}$$

with $[x]_+=\max(x,0)$, and $\mathrm{kl}_+(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))=\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))$ if $\mathsf{R}^\ell_{\mathcal{S}}(h)<\mathsf{R}^\ell_{\mathcal{D}}(h)$ and 0 otherwise. Moreover, $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_D)$ is a Gaussian noise such that $\mathbf{w}+\boldsymbol{\epsilon}$ are

the weights of $h \sim \rho_{\mathbb{S}}$ with $\rho_{\mathbb{S}} = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_D)$, and $\mathbb{c}$, $\mathbb{b}$ are two sets of hyperparameters fixed a priori.

Proof. Deferred to Appendix F.6. ■

Since we aim to minimize the true risk $\mathrm{R}_{\mathcal{D}}(h)$, *i.e.*, we consider in practice the 01-loss $\ell(h,(\mathbf{x},y)) = \mathrm{I}\left[h(\mathbf{x}) \neq y\right]$. As the parameter λ of the Theorem 6.3.2, $c \in \mathbb{c}$ is a hyperparameter that controls a trade-off between the empirical risk $\mathrm{R}^{\ell}_{\mathcal{S}}(h)$ and $\frac{1}{m}\Big[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} + \ln\frac{T\,\mathrm{card}(\mathbb{c})}{\delta}\Big]$. Besides, the parameter $b \in \mathbb{b}$ controls the tightness of the bound. These parameters can generally be tuned to minimize the bound of Equation (6.2) and Equation (6.3); however, there is no closed-form solution for the expression of the minimum of these equations. Consequently, our experimental protocol requires minimizing the bounds by gradient descent for each $b \in \mathbb{b}$ or $c \in \mathbb{c}$ to learn the distribution $\rho_{\mathbb{S}}$ leading to the lowest bound value. To obtain a tight bound, the divergence between one prior $\pi_t \in \mathbb{P}$ and $\rho_{\mathbb{S}}$ must be low, *i.e.*, $\|\mathbf{w}-\mathbf{v}_t\|_2^2$ (or $\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2$) has to be small. One solution is to split the learning sample into 2 non-overlapping subsets $\mathbb{S}_{\mathsf{prior}}$ and $\mathbb{S}$, where $\mathbb{S}_{\mathsf{prior}}$ is used to learn the prior, while $\mathbb{S}$ is used both to learn the posterior and compute the bound. Hence, if we "pre-learn" a good enough prior $\pi_t \in \mathbb{P}$ from $\mathbb{S}_{\mathsf{prior}}$, then we can expect to have a low $\|\mathbf{w}-\mathbf{v}_t\|_2$.

Training Method

The original training set is split in two distinct subsets: $\mathbb{S}_{\mathsf{prior}}$ and $\mathbb{S}$ (respectively of size m_{prior} and m, that can be different).
The training has two phases.
1) The prior distribution π is "pre-learned" with $\mathbb{S}_{\mathsf{prior}}$ and selected by early stopping, with $\mathbb{S}$ as validation set, using the algorithm A_{prior} (an arbitrary learning algorithm).
2) Given $\mathbb{S}$ and π, we learn the posterior $\rho_{\mathbb{S}}$ with the algorithm A (defined *a priori*).

At first sight, the selection of the prior weights with $\mathbb{S}$ by early stopping may appear to be "cheating". However, this procedure can be seen as: *(i)* first constructing $\mathbb{P}$ from the T "intermediate" NNs learned after each epoch from $\mathbb{S}_{\mathsf{prior}}$, then *(ii)* optimizing the bound with the prior that leads to the best risk on $\mathbb{S}$. This gives a statistically valid result: since Corollary 6.4.1 is valid for every $\pi_t \in \mathbb{P}$, we can select the one we want, in particular the one minimizing $\mathrm{R}^{\ell}_{\mathcal{S}}(h)$ for a sampled $h \sim \pi_t$. This heuristic makes sense: it allows us to detect if a prior is concentrated around hypotheses that potentially overfit the learning sample $\mathbb{S}_{\mathsf{prior}}$. Usually, practitioners consider this "best" prior as the final NN. In our case, the advantage is that we refine this "best" prior with $\mathbb{S}$ to learn the posterior $\rho_{\mathbb{S}}$. Note that Pérez-Ortiz *et al.* (2021) have already introduced tight generalization bounds with data-dependent priors for—non-derandomized—stochastic NNs. Nevertheless, our training method to learn the prior differs greatly since *(i)* we

learn T NNs (*i.e.*, T priors) instead of only one, *(ii)* we fix the variance of the Gaussian in the posterior $\rho_{\mathbb{S}}$. To the best of our knowledge, our training method for the prior is new.

6.4.2 A Note About Stochastic Neural Networks

Due to its stochastic nature, PAC-Bayesian theory has been explored to study stochastic NNs (*e.g.*, Dziugaite and Roy (2017, 2018), Zhou *et al.* (2019), and Pérez-Ortiz *et al.* (2021)). In Corollary 6.4.3 below, we instantiate the bound of Theorem 2.3.1 for stochastic NNs to empirically compare the stochastic and the deterministic NNs associated with the same prior and posterior distributions. We recall that, in this chapter, a deterministic NN is a *single* h sampled from the posterior distribution $\rho_{\mathbb{S}}{=}\mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_D)$ output by the algorithm A. Hence, for each example, the prediction is performed by the same deterministic NN: the one parameterized by the weights $\mathbf{w} + \boldsymbol{\epsilon} \in \mathbb{R}^D$. Conversely, the stochastic NN associated with a posterior distribution $\rho{=}\mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_D)$ predicts the label of a given example by *(i)* first sampling h according to ρ, and *(ii)* then returning the label predicted by h. Thus, the risk of the stochastic NN is the expected risk value $\mathbb{E}_{h\sim\rho}\,\mathsf{R}^{\ell}_{\mathcal{D}}(h)$, where the expectation is taken over *all* h sampled from ρ. We compute the empirical risk of the stochastic NN from a Monte Carlo approximation: *(i)* we sample K weight vectors, and *(ii)* we average the risk over the K associated NNs; we denote by ρ^K the distribution of such K-sample. In this context, we obtain the following PAC-Bayesian bound.

Corollary 6.4.3 (PAC-Bayesian Bound for Stochastic Neural Networks)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}{\times}\mathbb{Y}$, for any set $\mathbb{P} = \{\pi_1, \ldots, \pi_T\}$ of T priors on $\mathbb{H}$ where $\pi_t = \mathcal{N}(\mathbf{v}_t, \sigma^2\mathbf{I}_D)$, for any loss $\ell : \mathbb{H}{\times}(\mathbb{X}{\times}\mathbb{Y}){\to}\{0,1\}$, for any $\delta{\in}(0,1]$, with probability at least $1{-}\delta$ over $\mathbb{S}{\sim}\mathcal{D}^m$ and $\{h_1, \ldots, h_K\}{\sim}\rho^K$, we have simultaneously $\forall \pi_t {\in} \mathbb{P}$,

$$\mathrm{kl}\Big(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)\,\Big\|\,\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)\Big) \leq \frac{1}{m}\left[\frac{\|\mathbf{w}{-}\mathbf{v}_t\|_2^2}{2\sigma^2} + \ln\frac{4T\sqrt{m}}{\delta}\right], \tag{6.4}$$

$$\text{and}\quad \mathrm{kl}\left(\frac{1}{K}\sum_{i=1}^{K}\mathsf{R}_{\mathbb{S}}(h_i)\,\Big\|\,\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)\right) \leq \frac{1}{n}\ln\frac{4}{\delta}, \tag{6.5}$$

where $\rho = \mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_D)$ and the hypothesis h sampled from ρ is parameterized by $\mathbf{w} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_D)$.

Proof. Deferred to Appendix F.7. ■

This result shows two features that allow considering it as an adapted baseline for a fair comparison between disintegrated and classical PAC-Bayesian bounds, thus between deterministic and stochastic NNs. Firstly, it involves the same terms as Corollary 6.4.1.

Secondly, it is close to the bound of PÉREZ-ORTIZ *et al.* (2021, Sec. 6.2), since *(i)* we adapt the KL divergence to our setting (*i.e.*, $\mathrm{KL}(\rho\|\pi)=\frac{1}{2\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$), *(ii)* the bound holds for T priors thanks to a union bound argument.

6.5 Experiments with Neural Networks

In this section, we do not seek state-of-the-art performance; in fact, we have a three-fold objective: *(i)* we check if $50\%/50\%$ is a good choice for splitting the original train set into $(\mathbb{S}_{\text{prior}}, \mathbb{S})$ (which is the most common split in the PAC-Bayesian literature (GERMAIN *et al.*, 2009; PÉREZ-ORTIZ *et al.*, 2021)); *(ii)* we highlight that our disintegrated bound associated with the deterministic NN is tighter than the randomized bound associated with the stochastic NN (Corollary 6.4.3); *(iii)* we show that our disintegrated bound (Corollary 6.4.1) is tighter and more stable than the ones based on RIVASPLATA *et al.* (2020), BLANCHARD and FLEURET (2007) and CATONI (2007) (Corollary 6.4.2).

6.5.1 Training Method

We follow our Training Method (Section 6.4.1) in which we integrate the direct minimization of all the bounds. We refer as `ours` the training method based on the minimization of our bound in Corollary 6.4.1, as `rivasplata` the one based on Equation (6.1), as `blanchard` the one based on Equation (6.2), and as `catoni` the one based on Equation (6.3). `stochastic` denotes the PAC-Bayesian bound with the prior and posterior distributions obtained from `ours`. To optimize the bound with gradient descent, we replace the non-differentiable 0-1 loss $\mathrm{I}\left[h(\mathbf{x}) \neq y\right]$ with a surrogate: the bounded cross-entropy loss (DZIUGAITE and ROY, 2018). We made this replacement since cross-entropy minimization works well in practice for neural networks (GOODFELLOW *et al.*, 2016) and because this loss is bounded between 0 and 1, which is required for the $\mathrm{kl}()$ function. The cross-entropy is defined in a multi-class setting with $y \in \mathbb{Y}$ by $\ell(h,(\mathbf{x},y)) = -\frac{1}{Z}\ln\left[e^{-Z} + (1-2e^{-Z})h[y]\right] \in [0,1]$ where $h[y]$ is the y-th output of the NN; we set $Z{=}4$, the default parameter of DZIUGAITE and ROY (2018). That being said, to learn a good enough prior $\pi \in \mathbb{P}$ and the posterior $\rho_{\mathbb{S}}$, we run our Training Method with two stochastic gradient descent-based algorithms A_{prior} and A. Note that the randomness in the stochastic gradient descent algorithm is fixed to have deterministic algorithms. In phase **1)** algorithm A_{prior} learns from $\mathbb{S}_{\text{prior}}$ the T priors $\pi_1, \ldots, \pi_T \in \mathbb{P}$ (*i.e.*, during T epochs) by minimizing the bounded cross-entropy loss. In other words, at the end of the epoch t, the weights $\mathbf{w}_t$ of the classifier are used to define the prior $\pi_t = \mathcal{N}(\mathbf{w}_t, \sigma^2\mathbf{I}_D)$. Then, the best prior $\pi \in \mathbb{P}$ is selected by early stopping on $\mathbb{S}$. In phase **2)**, given $\mathbb{S}$ and π, algorithm A integrates the direct optimization of the bounds with the bounded cross-entropy loss.

6.5.2 Optimization[3] Procedure in Algorithms A and A_{prior}

Let $\boldsymbol{\omega}$ be the mean vector of a Gaussian distribution used as NN weights that we are optimizing. In algorithms A and A_{prior}, we use the Adam optimizer (Kingma and Ba, 2015), and we sample a noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_D)$ at each iteration of the optimizer. Then, we forward the examples of the mini-batch in the NN parameterized by the weights $\boldsymbol{\omega}+\boldsymbol{\epsilon}$, and we update $\boldsymbol{\omega}$ according to the bounded cross-entropy loss. Note that during phase **1)**, at the end of each epoch t, $\pi_t = \mathcal{N}(\boldsymbol{\omega}, \sigma^2 \mathbf{I}_D) = \mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_D)$ and finally at the end of phase **2)** we have $\rho_{\mathbb{S}} = \mathcal{N}(\boldsymbol{\omega}, \sigma^2 \mathbf{I}_D) = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_D)$.

6.5.3 Experimental Setting

6.5.3.1 Datasets

We perform our experimental study on three datasets: MNIST (LeCun *et al.*, 1998), Fashion-MNIST (Xiao *et al.*, 2017), and CIFAR-10 (Krizhevsky, 2009). We divide each original train set into two independent subsets $\mathbb{S}_{\text{prior}}$ of size m_{prior} and $\mathbb{S}$ of size m with varying split ratios defined as $\frac{m_{\text{prior}}}{m+m_{\text{prior}}} \in \{0, .1, .2, .3, .4, .5, .6, .7, .8, .9\}$. The test sets denoted by $\mathbb{T}$ remain the original ones.

6.5.3.2 Models

For the (Fashion-)MNIST datasets, we train a variant of the All Convolutional Network (Springenberg *et al.*, 2015). The model is a 3-hidden layers convolutional network with 96 channels. We use 5×5 convolutions with a padding of size 1 and a stride of size 1 everywhere except on the second convolution where we use a stride of size 2. We adopt the Leaky ReLU activation functions after each convolution. Lastly, we use a global average pooling of size 8×8 to obtain the desired output size. Furthermore, the weights are initialized with Xavier Normal initializer (Glorot and Bengio, 2010) while each bias of size l is initialized uniformly between $-1/\sqrt{l}$ and $1/\sqrt{l}$. For the CIFAR-10 dataset, we train a ResNet-20 network, *i.e.*, a ResNet network from He *et al.* (2016) with 20 layers. The weights are initialized with Kaiming Normal initializer (He *et al.*, 2015) and each bias of size l is initialized uniformly between $-1/\sqrt{l}$ and $1/\sqrt{l}$.

6.5.3.3 Optimization

For the (Fashion-)MNIST datasets, we learn the parameters of our prior distributions $\pi_1, \ldots, \pi_T$ by using Adam optimizer for $T = 10$ epochs with a learning rate of $\eta = 10^{-3}$ and a batch size of 32 (the other parameters of Adam are left by default). Moreover, the posterior distribution's parameters are learned for one epoch with the same batch size and optimizer (except that the learning rate is either $\eta = 10^{-4}$ or $\eta = 10^{-6}$). For

[3] The details of the optimization and the evaluation of the bounds are described in Appendix.

the CIFAR-10 dataset, the parameters of the priors $\pi_1, \ldots, \pi_T$ are learned for $T = 100$ epochs and the posterior distribution $\rho_{\mathbb{S}}$ for 10 epochs with a batch size of 32 by using Adam optimizer as well. Additionally, the learning rate to learn the prior for CIFAR-10 is $\eta = 10^{-2}$.

6.5.3.4 Bounds

We report the bounds' values with the 0-1 loss $\ell(h, (\mathbf{x}, y)) = \mathrm{I}\left[h(\mathbf{x}) \neq y\right]$. Moreover, for `blanchard`'s bounds, we define the set of hyperparameters $\mathbb{b}$ in the following way: $\mathbb{b}{=}\{b{\in}\mathbb{N} \mid b{=}\sqrt{x},\ (x{+}1) \leq 2\sqrt{m}\}$, *i.e.*, such that `blanchard`'s bounds can be tighter than `rivasplata`'s ones. We fixed the set of hyperparameters for `catoni` as $\mathbb{c}{=}\left\{10^k | k{\in}\{-3, -2, \ldots, +3\}\right\}$. We consider different values for the variance $\sigma^2{\in}\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ associated with the disintegrated KL divergence $\ln \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}$ equals to $\ln \frac{\rho_{\mathbb{S}}(h)}{\pi(h)} = \frac{1}{2\sigma^2}(\|\mathbf{w}{+}\boldsymbol{\epsilon}{-}\mathbf{v}_t\|_2^2 {-} \|\boldsymbol{\epsilon}\|_2^2)$, the "normal" $\mathrm{R\acute{E}NYI}$ divergence $\mathrm{D}_2(\rho\|\pi) = \frac{1}{\sigma^2}\|\mathbf{w}{-}\mathbf{v}_t\|_2^2$ and the KL divergence $\mathrm{KL}(\rho\|\pi) = \frac{1}{2\sigma^2}\|\mathbf{w}{-}\mathbf{v}_t\|_2^2$. For all the figures, the values are averaged over 400 deterministic NNs sampled from $\rho_{\mathbb{S}}$ (the standard deviation is small and presented in the Appendix). We additionally report as `stochastic` (Corollary 6.4.3) the randomized bound value and KL divergence $\mathrm{KL}(\rho\|\pi)$ associated with the model learned by `ours`, meaning that $K{=}400$ and that the test risk reported for `ours` also corresponds to the risk of the stochastic NN approximated with these 400 NNs.

6.5.4 Results

6.5.4.1 Analysis of the Influence of the Split Ratio Between $\mathbb{S}_{\text{prior}}$ and $\mathbb{S}$

Figures 6.1 and 6.2 study the evolution of the bound values after optimizing the bounds with our Training Method for different parameters. Specifically, the split ratio of the original train set varies from 0.1 to 0.9 (0.1 means that $m_{\text{prior}} = 0.1(m + m_{\text{prior}})$), for four variances values σ^2 and the two learning rates ($\eta = 10^{-6}$ and $\eta = 10^{-4}$). For the sake of readability, we present detailed results when the split ratio is 0 in Table F.1. We first remark that the behavior is different for the two learning rates. On the one hand, for $\eta = 10^{-6}$, the mean bound values are close to each other, which is not surprising since the disintegrated KL divergences and the $\mathrm{R\acute{E}NYI}$ divergences are close to zero (see Table F.2 to Table F.10). Moreover, for MNIST and Fashion-MNIST, there is a trade-off between learning a good prior with $\mathbb{S}_{\text{prior}}$ and minimizing a generalization bound with $\mathbb{S}$. In this case, the split ratio 0.5 appears to be a good choice to obtain a tight (disintegrated) PAC-Bayesian bound. This ratio is widely used in the PAC-Bayesian literature (see, *e.g.*, in the context of linear classifiers ($\mathrm{GERMAIN}$ *et al.*, 2009), majority votes ($\mathrm{ZANTEDESCHI}$ *et al.*, 2021), and neural networks ($\mathrm{LETARTE}$ *et al.*, 2019; $\mathrm{P\acute{E}REZ\text{-}ORTIZ}$ *et al.*, 2021)). On the other hand, when $\eta = 10^{-4}$, the

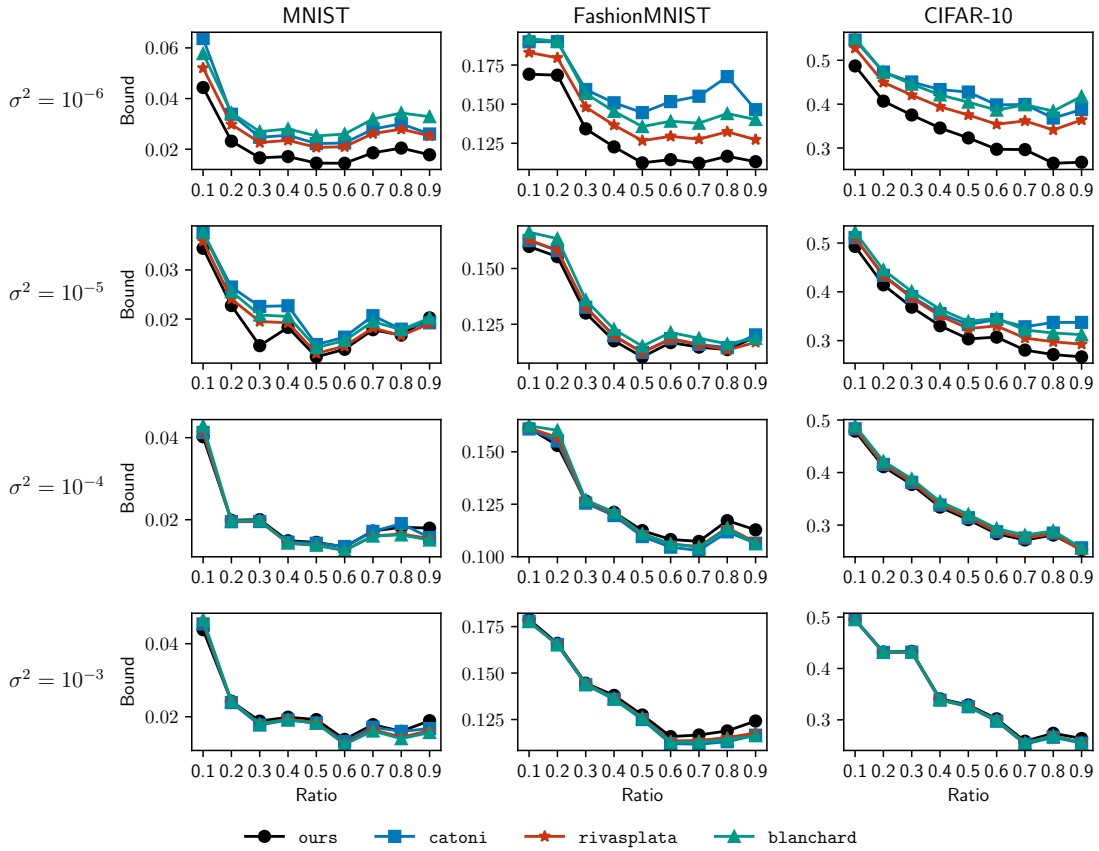


Figure 6.1. *Evolution of the bound values in terms of the split ratio. The x-axis represents the different split ratios, and the y-axis represents the bound values obtained after their optimization using our Training Method. Each row corresponds to a given variance σ^2, and each column corresponds to a dataset (MNIST, Fashion-MNIST, or CIFAR-10). In this figure, we consider a learning rate of $\eta = 10^{-6}$.*

mean bound values tend to increase when the split ratio increases as well for the bounds introduced in the literature (*i.e.*, for `blanchard`, `catoni`, and `rivasplata`), while the mean bound values of our bound remain low. Indeed, m decreases as long as the split ratio increases, which has the effect of increasing the bound value drastically when the disintegrated KL divergence is high. We further explain why the disintegrated KL divergence can become high for the disintegrated bounds of the literature. To do so, we now restrict our study to a split ratio of 0.5 in order to *(i)* compare the tightness of the bounds, *(ii)* understand why the disintegrated bounds of the literature diverge.

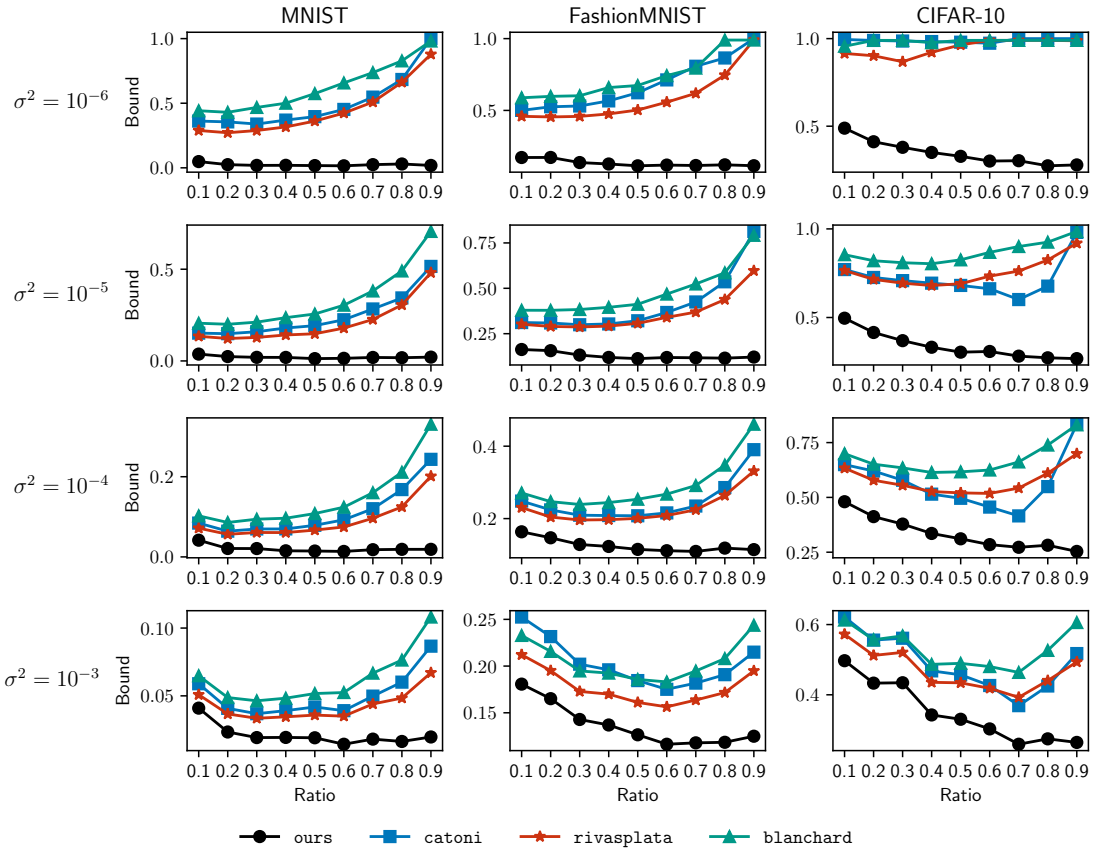


Figure 6.2. *Evolution of the bound values in terms of the split ratio. The x-axis represents the different split ratios, and the y-axis represents the bound values obtained after their optimization using our Training Method. Each row corresponds to a given variance σ^2, and each column corresponds to a dataset (MNIST, Fashion-MNIST, or CIFAR-10). In this figure, we consider a learning rate of $\eta = 10^{-4}$.*

6.5.4.2 Comparison Between Disintegrated and "Classic" Bounds

We first compare the "classic" PAC-Bayesian bound (Corollary 6.4.3) and our disintegrated PAC-Bayesian bound (Corollary 6.4.1). To do so, we fix the variance $\sigma^2{=}10^{-3}$ (along with the split ratio equals 0.5). We report in Figure 6.3, the mean bound values associated with `ours` (*i.e.*, the Training Method that minimizes our bound) and `stochastic` (we recall that `stochastic` is the PAC-Bayesian bound of Corollary 6.4.3 on the model learned by `ours`). Actually, `ours` leads to more precise bounds than the randomized `stochastic` even if the two empirical risks are the same and the KL divergence is smaller than the RÉNYI one. This imprecision is due to the non-avoidable sampling according to ρ done in the randomized PAC-Bayesian bound

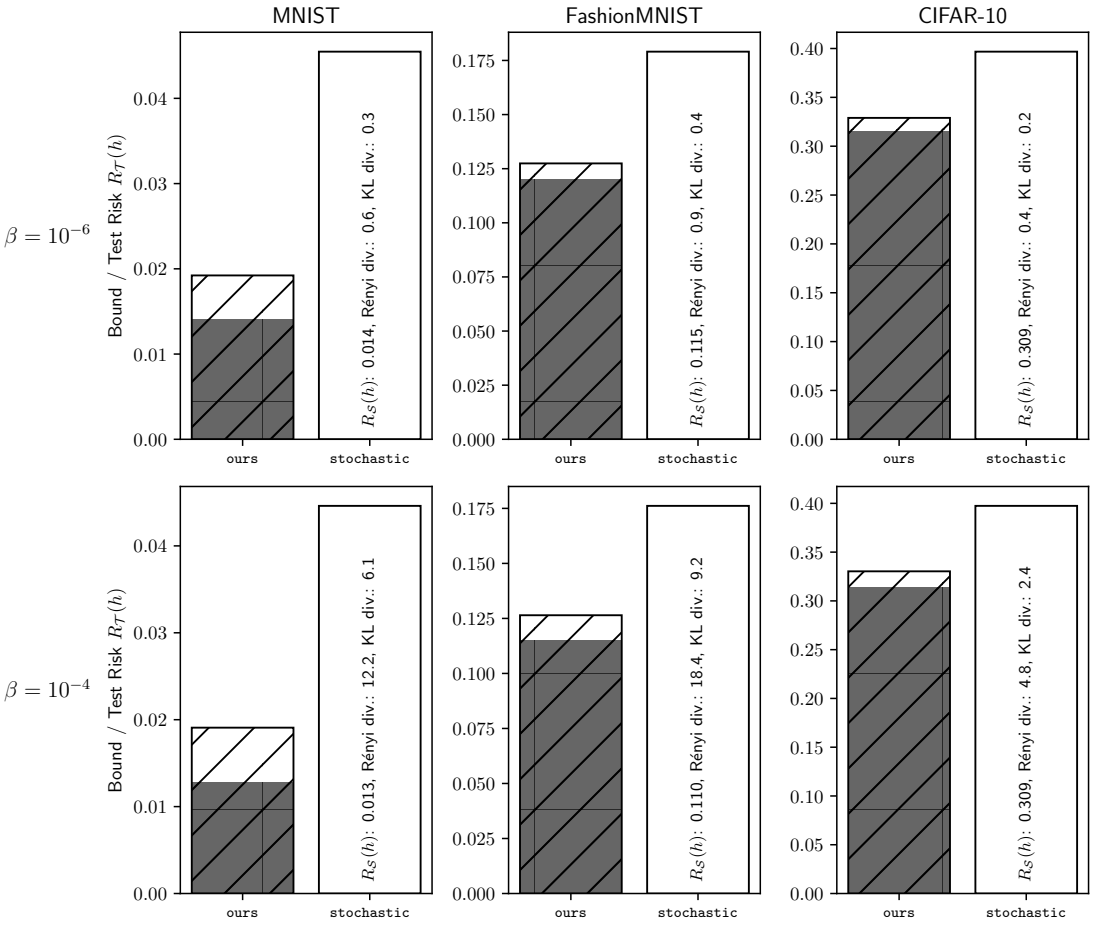


Figure 6.3. *The figure illustrates the values of the PAC-Bayes bound (Corollary 6.4.3) and the values of the disintegrated bound (Corollary 6.4.1) where the learning rate is $\eta = 10^{-4}$ or $\eta = 10^{-6}$ and the split ratio is 0.5. The y-axis shows the values of the bounds (the hatched bar for* `ours` *(Corollary 6.4.1) and the white bar for* `stochastic` *(Corollary 6.4.3)) and the test risks $R_{\mathcal{T}}(h)$ (grey shaded bar). We also report the values of the empirical risk $R^{\ell}_{\mathcal{S}}(h)$, the* RÉNYI *divergence (associated with* `ours`*' bound), and the KL divergence (associated with* `stochastic`*'s bound).*

of Corollary 6.4.3 (the higher K, the tighter the bound). Thus, using a disintegrated PAC-Bayesian bound avoids sampling many NNs to obtain a low risk. This confirms that our framework makes sense for practical purposes and has a great advantage in terms of time complexity when computing the bounds.

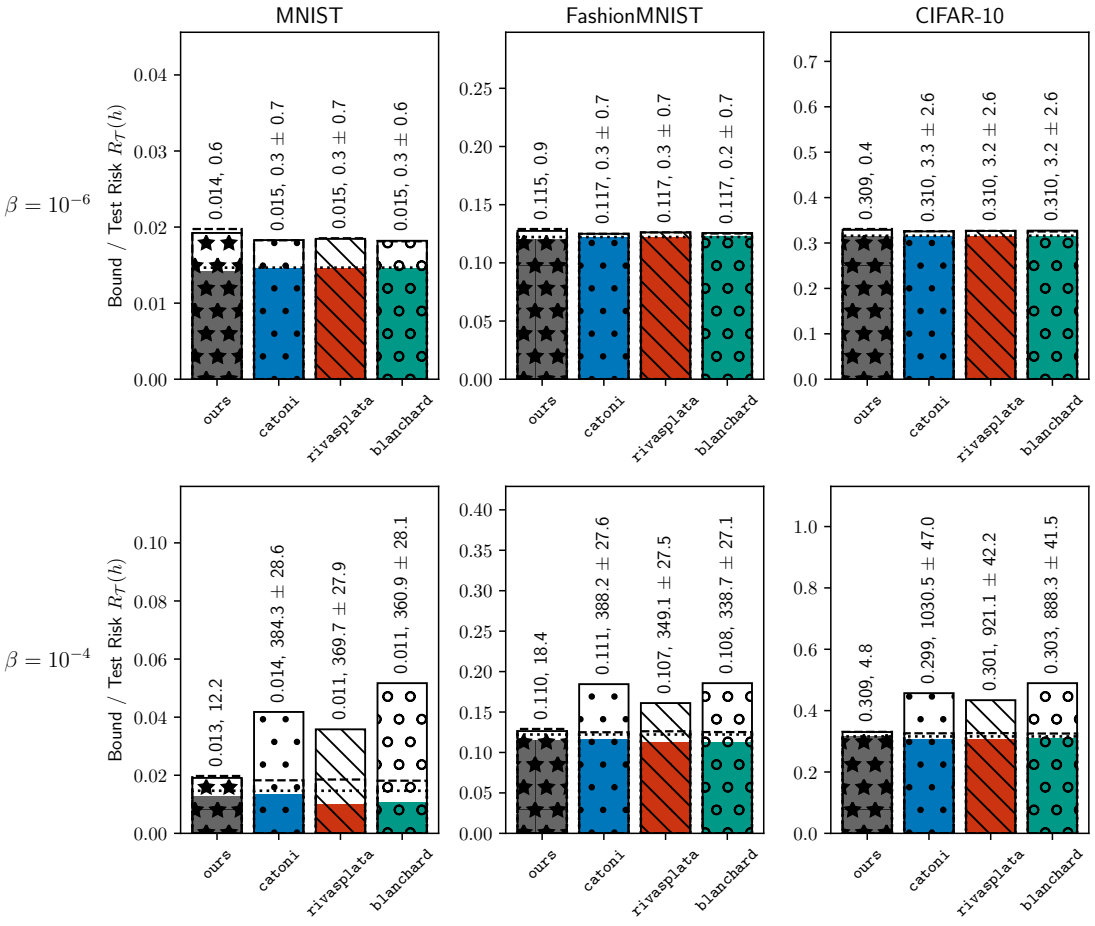


Figure 6.4. *This figure shows the value of the disintegrated bounds (the colored bars) and the test risks (the hatched bars) for Corollary 6.4.1 ("*`ours`*") and Corollary 6.4.2 ("*`catoni`*", "*`rivasplata`*" and "*`blanchard`*") in two different settings, i.e., with a learning rate of $\eta = 10^{-6}$ and $\eta = 10^{-4}$ and with split ratio of 0.5. We also plot the value of the bounds (the dashed lines) and the test risks (the dotted lines) before executing Step* **2)** *of our Training Method. The y-axis shows the values of the bounds and the test risks $R_{\mathcal{T}}(h)$. The empirical risk $R_{\mathcal{S}}(h)$ is presented above each bar. Moreover, the second value represents the mean value of the divergence (the standard deviations are also given for the disintegrated bounds of the literature).*

6.5.4.3 Analysis of the Tightness of the Disintegrated Bounds

We now compare the tightness of the different disintegrated PAC-Bayesian bounds (*i.e.*, our bound and the ones in the literature). We study, as before, the case where the split ratio is 0.5 and the variance $\sigma^2 = 10^{-3}$. We report in Figure 6.4

for `ours`, `rivasplata`, `blanchard` and `catoni`, the mean bounds values; the mean test risk $\mathrm{R}_{\mathcal{T}}(h)$ before (*i.e.*, with the prior π) and after applying Step **2)** (*i.e.*, with the posterior $\rho_{\mathbb{S}}$). Moreover, we report above the bars the mean train risks $\mathrm{R}_{\mathcal{S}}^{\ell}(h)$ and the mean/standard deviation divergence values obtained after Step **2)**, *i.e.*, the RÉNYI divergence $\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)=\frac{1}{\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$ for `ours` and the disintegrated KL divergence $\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}=\frac{1}{2\sigma^2}\left[\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2\right]$ for the others. First of all, we can remark that we observe two different behaviors for $\eta=10^{-4}$ and $\eta=10^{-6}$. For $\eta=10^{-6}$, the bound values are close to each other, as well as the empirical risks and the divergences (which are close to 0). In Figure 6.4, we observe that the bound values and the test risks are close to the one associated with the prior distribution because the divergence is close to 0. This is probably due to the fact that the learning rate is too small, implying that the bounds are not optimized. With a higher learning rate of $\eta=10^{-4}$, we observe that our bound remains tight while the disintegrated bounds of the literature are looser. Hopefully, our bound is improved after performing Step **2)** of our Training Method. However, for the bounds of the literature, the value of the disintegrated KL divergence is large, making the bounds looser after executing Step **2)**. We now investigate the reasons for the divergence of the bounds by looking at the influence of the variance σ^2.

6.5.4.4 Analysis of the Influence of the Variance σ^2

Given a split ratio of 0.5 and $\eta\in\{10^{-6},10^{-4}\}$, we report in Figure 6.5 the evolution of the bound values associated with `ours`, `rivasplata`, `blanchard`, and `catoni` when the variance varies from 10^{-6} to 10^{-3}. First of all, an important point is that `ours` behaves differently than `rivasplata`, `blanchard`, and `catoni`. Indeed, for both learning rates, when σ^2 decreases, the value of our bound remains low, while the others increase drastically due to the explosion of the disintegrated KL divergence term (see Table F.6 in Appendix for more details). Concretely, the disintegrated KL divergence in Corollary 6.4.2 involves the noise $\boldsymbol{\epsilon}$ through $\frac{1}{2\sigma^2}\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2$ compared to our divergence which is $\frac{1}{\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$ (without noise). Then, the sampled noise during the optimization procedure $\boldsymbol{\epsilon}$ influences the disintegrated KL divergence, making it prone to high variations during training (depending thus on σ^2). To illustrate the difference during the optimization, we focus on the objective function (detailed in Appendix) of Corollary 6.4.1 and Corollary 6.4.2 (Equation (6.1)). Roughly speaking, the divergence in Corollary 6.4.1 does not depend on the sampled hypothesis h (with weights $\boldsymbol{\omega}+\boldsymbol{\epsilon}$), while the divergence of Equation (6.1) does. In consequence, the derivatives are less dependent on h for Corollary 6 than for Equation (6.1). To be convinced of this, we propose to study the gradient with respect to the current mean vector $\boldsymbol{\omega}$. On the one hand, the gradient $\frac{\partial \mathrm{R}_{\mathcal{S}}^{\ell}(h)}{\partial\boldsymbol{\omega}}$ of the risk *w.r.t.* $\boldsymbol{\omega}$ is the same for both bounds (with the loss of DZIUGAITE and ROY (2018)); hence, the phenomenon cannot come from this derivative. On the other hand, the gradients of the divergence in Equation (6.1) and

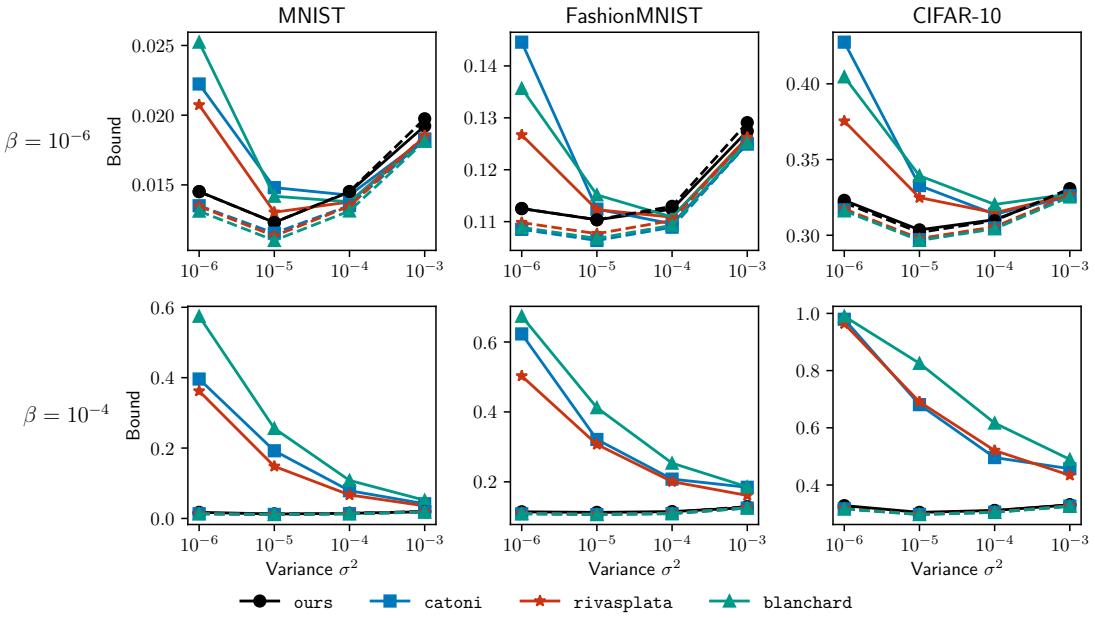


Figure 6.5. *We plot the evolution of the mean bound values (the plain lines) in terms of the variance σ^2 after optimizing the bounds with our Training Method. Moreover, we plot the mean bound values (the dashed lines) obtained before executing Step* **2)** *of our Training Method. The variance is represented on the x-axis, while the bound values are represented on the y-axis. Furthermore, each row corresponds to a given learning rate ($\eta = 10^{-6}$ or $\eta = 10^{-4}$), and each column corresponds to a dataset (either MNIST, FashionMNIST, or CIFAR-10). The split ratio considered is 0.5.*

Corollary 6.4.1 are respectively

$$\frac{\partial}{\partial \boldsymbol{\omega}}\left[\frac{1}{m}\left(\frac{\|\boldsymbol{\omega}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}\right)\right] = \frac{\partial}{\partial \boldsymbol{\omega}}\left[\frac{1}{m2\sigma^2}\|\boldsymbol{\omega}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2\right]$$
$$= \frac{1}{m\sigma^2}(\boldsymbol{\omega}+\boldsymbol{\epsilon}-\mathbf{v}_t) = \diamondsuit,$$

and
$$\frac{\partial}{\partial \boldsymbol{\omega}}\left[\frac{1}{m}\left(\frac{\|\boldsymbol{\omega}-\mathbf{v}_t\|_2^2}{\sigma^2}\right)\right] = \frac{\partial}{\partial \boldsymbol{\omega}}\left[\frac{1}{m\sigma^2}\|\boldsymbol{\omega}-\mathbf{v}_t\|_2^2\right]$$
$$= \frac{2}{m\sigma^2}(\boldsymbol{\omega}-\mathbf{v}_t) = \heartsuit.$$

From the two derivatives, we deduce that $\diamondsuit = \frac{1}{2}\heartsuit + \frac{1}{m\sigma^2}\boldsymbol{\epsilon}$. Hence, each gradient step involves a noise in the gradient of the disintegrated KL divergence $\frac{1}{m\sigma^2}\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \frac{1}{m}\mathbf{I}_D)$, which is high for a small m. This randomness causes the disintegrated KL divergence $\frac{1}{2\sigma^2}\|\boldsymbol{\omega}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2$ to be larger when σ^2 decreases since *(i)* the divergence is divided by $2m\sigma^2$ and *(ii)* the deviation between $\boldsymbol{\omega}$ and $\mathbf{v}_t$ increases. In conclusion, this

makes the objective function (*i.e.*, the bound) subjects to high variations during the optimization, implying higher final bound values. Thus, the RÉNYI divergence has a valuable asset over the disintegrated KL divergence since it does not depend on the sampled noise $\boldsymbol{\epsilon}$.

6.5.4.5 Take Home Message from the Experiments

To summarize, our experiments show that our disintegrated bound is, in practice, tighter than the ones in the literature. This tightness allows us to precisely bound the true risk $\mathrm{R}_{\mathcal{D}}(h)$ (or the test risk $\mathrm{R}_{\mathcal{T}}(h)$); thus, the model selection from the disintegrated bound is effective. Moreover, we show that our bound is more easily optimizable than the others. This is mainly due to the disintegrated KL divergence, which depends on the sampled hypothesis h with weights $\boldsymbol{\omega}+\boldsymbol{\epsilon}$. Indeed, the gradients of the disintegrated KL divergence with respect to $\boldsymbol{\omega}$ include the noise $\boldsymbol{\epsilon}$, making the gradient inaccurate (especially with a "high" learning rate and small variance σ^2).

6.6 Perspectives for the Majority Vote

Before concluding, we discuss some perspectives for the stochastic majority vote introduced in Chapter 5. Recall that the *stochastic* majority vote has its weights ρ sampled from the Dirichlet distribution P defined as $\mathrm{P}(\rho) \triangleq \frac{1}{Z(\boldsymbol{\alpha})} \prod_{j=1}^{\mathrm{card}(\mathbb{H})} \rho(h_j)^{\alpha_j - 1}$ (also called hyper-posterior). One drawback of the PAC-Bayesian approach for the *stochastic* majority vote in Chapter 5 is that we are not considering only one majority vote (with weights sampled from the hyper-posterior P). Moreover, based on the margin theory (initiated by SCHAPIRE *et al.*, 1998) and our work in Chapter 5, BIGGS *et al.* (2020) derived PAC-Bayesian bound for the expected majority vote. This latter work illustrates that a special care is unfortunately needed to obtain a bound for a unique majority vote through the PAC-Bayesian theory. Hopefully, thanks to the disintegrated PAC-Bayesian framework, we can bound the true risk of a single majority vote classifier. As an illustration we derive two bounds for such a classifier based on RIVASPLATA *et al.* (2020)'s bound (Theorem 2.4.1) and based on Theorem 6.3.1.

Corollary 6.6.1 (Instantiation of Theorem 2.4.1 to Stochastic Majority Votes)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any finite set of voters $\mathbb{H}$, for any hyper-prior distribution $\Pi = \mathrm{Dir}(\boldsymbol{\beta})$ on $\mathbb{H}$ with $\boldsymbol{\beta} \in (\mathbb{R}_*^+)^{\mathrm{card}(\mathbb{H})}$, for any loss $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$, for any $\delta \in (0,1]$, for any algorithm A that outputs a hyper-posterior given a learning sample and a hyper-prior, with probability at least $1-\delta$ over the learning sample $\mathbb{S}\sim\mathcal{D}^m$ and the posterior distribution $\rho\sim\mathrm{P}_{\mathbb{S}} = \mathrm{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in (\mathbb{R}_*^+)^{\mathrm{card}(\mathbb{H})}$

we have

$$\mathrm{kl}(\mathrm{R}^{\ell}_{\mathcal{S}}(\mathrm{MV}_{\rho})\|\mathrm{R}^{\ell}_{\mathcal{D}}(\mathrm{MV}_{\rho})) \leq \frac{1}{m}\left[\ln\frac{Z(\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})} + \sum_{j=1}^{\mathrm{card}(\mathbb{H})}(\alpha_j-\beta_j)\ln(\rho(h_j)) + \ln\frac{2\sqrt{m}}{\delta}\right],$$

where $\mathrm{P}_{\mathbb{S}} \triangleq A(\mathbb{S},\Pi)$ is output by the deterministic algorithm A.

Proof. Deferred to Appendix F.13 ■

Corollary 6.6.2 (Instantiation of Theorem 6.3.1 to Stochastic Majority Votes)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any finite set of voters $\mathbb{H}$, for any hyper-prior distribution $\Pi = \mathrm{Dir}(\boldsymbol{\beta})$ on $\mathbb{H}$ with $\boldsymbol{\beta} \in (\mathbb{R}^{+}_{*})^{\mathrm{card}(\mathbb{H})}$, for any loss $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y}) \to [0,1]$, for any $\lambda > 1$, for any $\delta \in (0,1]$, for any algorithm A that outputs a hyper-posterior given a learning sample and a hyper-prior, with probability at least $1-\delta$ over the learning sample $\mathbb{S}\sim\mathcal{D}^m$ and the posterior distribution $\rho\sim\mathrm{P}_{\mathbb{S}} = \mathrm{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in (\mathbb{R}^{+}_{*})^{\mathrm{card}(\mathbb{H})}$ we have

$$\mathrm{kl}(\mathrm{R}^{\ell}_{\mathcal{S}}(\mathrm{MV}_{\rho})\|\mathrm{R}^{\ell}_{\mathcal{D}}(\mathrm{MV}_{\rho})) \leq \frac{1}{m}\left[\frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta} + \ln\frac{Z(\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})} + \frac{1}{\lambda-1}\ln\frac{Z(\lambda\boldsymbol{\alpha}+(1-\lambda)\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})} + \ln(2\sqrt{m})\right],$$

where $\mathrm{P}_{\mathbb{S}} \triangleq A(\mathbb{S},\Pi)$ is output by the deterministic algorithm A.

Proof. Deferred to Appendix F.14 ■

These two bounds offer great perspectives to upper-bound the true risk of the majority vote. First, as we can remark, the two theorems holds for all (bounded) losses $\ell()$ including the 01-loss. Hence, the true risk $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_{\rho})$ can be upper-bounded directly without using the $\frac{1}{2}$-margin of Laviolette *et al.* (2017) as in Chapter 5. While our bounds require to sample a single majority vote from P, they might be tighter since it does not rely on margin bound (as Biggs *et al.*, 2020) and directly deal with the 01-loss.

6.7 Summary and Conclusion

We provide a new and general disintegrated PAC-Bayesian bound (Theorem 6.3.1) in the family of disintegrated PAC-Bayesian bounds (Definition 2.4.1), *i.e.*, when the de-

randomization step consists in *(i)* learning a posterior distribution $\rho_{\mathbb{S}}$ on the classifiers set (given an algorithm, a learning sample $\mathbb{S}$ and a prior distribution π) and *(ii)* sampling a hypothesis h from this posterior $\rho_{\mathbb{S}}$. While our bound can be looser than the ones of Blanchard and Fleuret (2007), Catoni (2007), and Rivasplata *et al.* (2020), it provides nice opportunities for learning deterministic classifiers. Indeed, our bound can be used not only to study the theoretical guarantees of deterministic classifiers but also to derive self-bounding algorithms (based on the bound optimization) that are more stable and efficient than the ones we obtain from the bounds of the literature. Concretely, the bounds of Blanchard and Fleuret (2007), Catoni (2007), and Rivasplata *et al.* (2020) depend on two terms related to the classifier drawn: the risk and the "disintegrated KL divergence", while in our bound the (Rényi) divergence term depends on the hypothesis set, implying that the divergence remains the same whatever which classifier is drawn. In this sense, our bound is more stable as the learning algorithm seeking to minimize the bound allows, in practice, to choose a better hypothesis than with the bounds of Blanchard and Fleuret (2007), Catoni (2007), and Rivasplata *et al.* (2020). We have illustrated the interest of our bound on neural networks and provides perspectives on the the stochastic majority vote classifier introduced in Chapter 5.

One future research direction related to this work is to develop new proof techniques to convert generalization bounds holding in expectation (see *e.g.*, Xu and Raginsky, 2017) to high-probability bounds. To do so, one can apply Markov's Inequality (Theorem A.2.1) and the Donsker-Varadhan Variational Representation of the KL divergence to obtain a high-probability bound with the bound in expectation as upper-bound. This work might be significant since the dependence in δ is, for now, only polynomial, while a logarithmic dependence is preferable. However, the limitation would be that the bound in expectation must consider a specific distribution (related to the Donsker-Varadhan Variational Representation).

In the next chapter, we leverage the disintegrated KL divergence and the bound of Rivasplata *et al.* (2020) to obtain generalization bounds with an arbitrary complexity measure. Hence, given a complexity measure defined by the user, we are able to give a generalization bound that holds with high probability over the learning sample and a hypothesis sampled from a complexity-measure-dependent distribution. Moreover, the bound with arbitrary complexity measures encompasses the generalization bounds of the literature and is recalled in Chapter 1.

Generalization Bounds with Complexity Measures

7

This chapter is based on the following paper

Paul Viallard, Rémi Emonet, Amaury Habrard, Emilie Morvant, and Valentina Zantedeschi. Generalization Bounds with Arbitrary Complexity Measures. *Submitted to ICLR 2023.* (2022a)

Contents

7.1 Introduction **176**
7.2 Preliminaries **177**
7.2.1 Setting 177
7.2.2 Reminder on Disintegrated PAC-Bayesian Bounds 178
7.3 Integrating Arbitrary Complexities in Generalization Bounds **179**
7.3.1 An Introduction to our Results 179
7.3.2 About the Gibbs Distribution 180
7.3.3 Our Main Results: Generalization Bound with Complexity Measures . 181
7.4 Using Arbitrary Complexities in Practice **184**
7.4.1 Sampling from the Gibbs Distribution 184
7.4.2 Experiments 186
7.5 Comparison with the Generalization Bounds of the Literature **192**
7.5.1 Bounds with Arbitrary Complexity Measures 193
7.5.2 Uniform Convergence Bounds 193
7.5.3 Algorithmic-Dependent Bounds 195
7.6 Conclusion and Summary **197**

Abstract

In statistical learning theory, generalization bounds usually involve a complexity measure that is constrained by the considered theoretical framework limiting the scope of such analysis. Among the measured mentioned in this thesis (Chapter 1) we can cite for example the VC-dimension, the stability constant of the uniform stability framework or the KL divergence used in the PAC-Bayesian framework studied in Part II. Recently, the empirical study of Jiang *et al.* (2019), made in the context of neural networks learning, has shown that *(i)* common complexity measures (such as the VC-dimension) do not necessarily correlate with the generalization gap, and that *(ii)* there exist *arbitrary* complexity measures that are better correlated with the generalization gap, but come without generalization guarantees. In this chapter, we propose to address the second point by presenting a general framework allowing one to derive some generalization bounds able to take into account a general complexity measures.

7.1 Introduction

As shown in Chapter 1 and notably in Section 1.3, statistical learning theory offers various theoretical frameworks to assess generalization by studying whether the empirical risk is representative of the true risk thanks to an upper bounding strategy of the generalization gap. While the generalization gap represents a deviation between the true risk and the empirical risk, an upper bound on this generalization gap is generally a function of mainly two quantities: *(i)* the size of the training sample and *(ii)* a complexity measure that captures how much the model overfits the data. There are some well-known complexity measures in the literature such as the VC-Dimension (Definition 1.3.3) or the Rademacher complexity (Definition 1.3.4) that quantify how much a hypothesis from the hypothesis set can overfit the data. Following a different framework, instead of considering the whole hypothesis set, algorithmic-dependent complexity measures propose to quantify how much the hypothesis obtained by a learning algorithm overfit the data: the uniform stability (Definition 1.3.6) or the robustness (Definition 1.3.7) parameters are two examples.

In order to study generalization capabilities, a recent line of works, mainly in the context of neural networks learning, is dedicated to the empirical study of different complexity measures to find those that correlate the most with the generalization gap (Jiang *et al.*, 2019; Dziugaite *et al.*, 2020; Jiang *et al.*, 2021). For example, given an arbitrary complexity measure, Jiang *et al.* estimate empirically the generalization gap and the complexity measure for over-parametrized models. They are able to rank the measure and the gap to obtain the Kendall's rank coefficient (Kendall, 1938) to evaluate how a measure reflects the generalization for a particular model. However, this correlation coefficient were criticized by Dziugaite *et al.* (2020) because they found that if the measures empirically correlate well *on average*, they are not robust to changes of the fixed parameters (such as the depth or the width of the model).

On the one hand, while these results are extremely important to understand generalization, notably for over-parametrized models, they remain incomplete since they are only empirical. On the other hand, as we can see in Section 1.3, the bounds in the literature are restrictive because the practitioner cannot integrate in a bound its own complexity measure. In other words, to the best our knowledge, there is no generalization bound involving arbitrary complexity measures that are found to be good proxies for the generalization gap. In this chapter, we aim to provide generalization bounds with arbitrary complexity measures that the practitioner can define. We believe that this direction is of important interest in advancing the understanding of generalization, as the generalization gap can be *provably* upper-bounded with a term that depends on a user-specified complexity measure. To get such generalization bounds, we leverage a disintegrated PAC-Bayesian bound (Theorem 2.4.1). Such a bound further allows

to derive theoretical guarantees for arbitrary complexity measures that depend on the sampled model, and on the learning sample, which is uncommon in statistical learning theory. Hence, our novel results provide theoretical foundation to the many regularization used in practice to perform model selection (*e.g.*, L2 regularization).

The rest of the chapter is organized as follows. In Section 7.2, we provide some preliminaries dedicated to this chapter. Then, we present the main contribution in Sections 7.3 and 7.5 before concluding in Section 7.6. As usual in this thesis, in order to improve the readability of the chapter, the proofs are deferred in Appendix G.

7.2 Preliminaries

We provide in this section some quick recaps on notions that have been introduced previously in this thesis to make the reading easier. We first remind some elements about supervised learning and the generalization gap. Then, we review quickly the disintegrated generalization bounds from which the contribution of this chapter are developed.

7.2.1 Setting

We follow the setting of Chapter 2 and Section 2.4 where we consider the supervised classification setting. We recall that $\mathbb{X}$ denotes the input space and $\mathbb{Y}$ the label space. Moreover, we consider that an example $(\mathbf{x}, y) \in \mathbb{X}\times\mathbb{Y}$ is sampled from an unknown data distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$. A learning sample $\mathbb{S}=\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ contains m examples drawn *i.i.d.* from $\mathcal{D}$; we denote the distribution of such an m-sample by $\mathcal{D}^m$. Let $\mathbb{H}$ be a potentially infinite set of hypotheses $h : \mathbb{X}\to\mathbb{Y}$ that associate a label belonging to $\mathbb{Y}$ given an input from $\mathbb{X}$.

Given a learning sample $\mathbb{S}$, we aim to find $h\in\mathbb{H}$ that minimizes the so-called true risk $\mathrm{R}_{\mathcal{D}}(h)=\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[h(x)\neq y\right]$. In practice, as the data distribution $\mathcal{D}$ is unknown, we estimate the true risk with its empirical counterpart, *i.e.*, the empirical risk $\mathrm{R}_{\mathbb{S}}(h) = \frac{1}{m}\sum_{i=1}^m \mathrm{I}\left[h(x_i)\neq y_i\right]$. We hereafter denote the generalization gap by $\phi : [0,1]^2 \to \mathbb{R}$, which is usually defined by $\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathbb{S}}(h)) = |\mathrm{R}_{\mathcal{D}}(h)-\mathrm{R}_{\mathbb{S}}(h)|$ that quantify how much the empirical risk is representative of the true risk. Since the generalization bound is not computable, a complexity measure can be used to capture the overfitting phenomenon.

To incorporate arbitrary complexities in the bounds, we leverage the disintegrated PAC-Bayesian bounds framework (recalled in Section 2.4), *i.e.*, we upper-bound the generalization gap for a hypothesis h sampled from $\rho_{\mathbb{S}} \in \mathbb{M}(\mathbb{H})$ with a function that depends on an *arbitrary* measure of complexity. To do so, we need to consider an *apriori* belief on the hypotheses in $\mathbb{H}$ that is modeled by a prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$.

Hence, we aim to learn, from $\mathbb{S}$ and π, a *posterior* distribution $\rho_{\mathbb{S}}$ to assign higher probability to the best hypotheses in $\mathbb{H}$; the hypothesis h is then sampled from $\rho_{\mathbb{S}}$ to obtain the guarantee with an arbitrary complexity measure.

7.2.2 Reminder on Disintegrated PAC-Bayesian Bounds

The disintegrated PAC-Bayesian bound (recalled in Section 2.4) has been introduced by Catoni (2007, Th 1.2.7) and Blanchard and Fleuret (2007, Prop 3.1). To the best of our knowledge, despite their interest, they have been little used in the literature and have only recently received renewed interest for deriving tight bounds in practice (see Chapter 6). Such bounds take the following form (introduced in Chapter 2).

Definition 2.4.1 (Disintegrated PAC-Bayesian Generalization Bound)**.** Let $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$ be a loss function and $\phi : [0,1]^2 \to [0,1]$ a generalization gap. A *disintegrated* PAC-Bayesian bound is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, there exists a function $\Phi : \mathbb{M}(\mathbb{H}) \times \mathbb{M}^*(\mathbb{H}) \times (0,1] \to \mathbb{R}$ such that for any $\delta \in (0,1]$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \Big[\phi(\mathsf{R}^\ell_{\mathcal{D}}(h), \mathsf{R}^\ell_{\mathcal{S}}(h)) \leq \Phi\Big(\rho_{\mathbb{S}}, \pi, \delta\Big)\Big] \geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S}, \pi)$ is output by the deterministic algorithm A and $\phi()$ is, for example, $\phi(\mathsf{R}^\ell_{\mathcal{D}}(h), \mathsf{R}^\ell_{\mathcal{S}}(h)) = \Big|\mathsf{R}^\ell_{\mathcal{D}}(h) - \mathsf{R}^\ell_{\mathcal{S}}(h)\Big|$.

Put into words, given a training set $\mathbb{S}$ sampled from $\mathcal{D}^m$, we can learn the distribution $\rho_{\mathbb{S}}$ from $\mathbb{S}$, and then sample the hypothesis h from $\rho_{\mathbb{S}}$ to get a bound with high probability (at least $1-\delta$) over the random choice of $\mathbb{S}$ and h. In this chapter, we mainly focus on Rivasplata *et al.* (2020)'s bound recalled below.

Theorem 2.4.1 (General Disintegrated Bound of Rivasplata *et al.* (2020))**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any measurable function $\varphi : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}$, for any $\delta \in (0,1]$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\varphi(h,\mathbb{S}) \leq \underbrace{\ln\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right] + \ln\left[\frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} \exp\left(\varphi(h',\mathbb{S}')\right)\right]}_{\Phi(\rho_{\mathbb{S}},\pi,\delta)}\right] \geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S}, \pi)$ is output by the deterministic algorithm A.

In this case, the function $\varphi(h, \mathbb{S}) = m\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h))$ is a deviation between the true risk $\mathrm{R}_{\mathcal{D}}(h)$ and the empirical risk $\mathrm{R}_{\mathcal{S}}(h)$. Moreover, the function $\Phi(\rho_{\mathbb{S}}, \pi, \delta)$ is constituted by 2 terms: *(i)* the *disintegrated* KL divergence $\ln \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}$ defining how much the prior and posterior distributions deviate for a single h, and *(ii)* the term $\ln\Big[\frac{1}{\delta}\, \mathbb{E}_{\mathbb{S}' \sim \mathcal{D}^m}\, \mathbb{E}_{h' \sim \pi} \exp\left(\varphi(h', \mathbb{S}')\right)\Big]$ is constant *w.r.t.* $h \in \mathbb{H}$ and $\mathbb{S} \in (\mathbb{X} \times \mathbb{Y})^m$ and usually upper-bounded to instantiate the bound. In the following we refer to the whole right-hand side of the bound, $\Phi()$, as the complexity measure. This is in slight contrast with the standard definition of complexity, where the term *(ii)* (related to δ and the sample size m) is not included. This additional term is, in fact, constant *w.r.t.* the hypothesis $h \sim \rho_{\mathbb{S}}$ and the learning sample $\mathbb{S} \sim \mathcal{D}^m$.

In the bound of Theorem 2.4.1, the complexity term $\Phi()$ depends on the disintegrated KL divergence and suffers from some drawbacks. Indeed, the complexity term is imposed by the framework and it can be subject to high variance in practice (see Chapter 6). However, it is important to notice that this disintegrated KL divergence has a clear advantage: it only depends on the hypothesis h and $\mathbb{S}$, instead of the whole hypothesis class (as it is often the case, *e.g.*, when using the KL divergence in PAC-Bayesian bounds, or the VC-dimension). This might implies a better correlation between the generalization gap and some complexity measures. In the next section, we leverage this disintegrated KL divergence to derive our main contribution: a general bound that involves arbitrary complexity measures.

7.3 Integrating Arbitrary Complexities in Generalization Bounds

We first begin with a short presentation of our result to give some preliminary intuitions and to introduce the notion of Gibbs distribution which is a key element of our contribution. We then provide formally our theoretical result in Section 7.3.3.

7.3.1 An Introduction to our Results

Let $\Phi_\mu(h, \mathbb{S}, \delta)$ be a real-valued function dependent on an additional function $\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$ that takes a hypothesis h, a learning sample $\mathbb{S}$, and the parameter δ as arguments. This function $\mu()$ parametrizes the complexity measure based on the data sample $\mathbb{S}$ and the model h, and thus allows us to introduce arbitrary complexity measures in the bound; we further denote the function $\mu()$ as *parametric function*. As an example, when $\mathbb{H}$ is a set of hypotheses $h_{\mathbf{w}}$ parameterized by some weights $\mathbf{w} \in \mathbb{R}^d$, we can fix $\mu(h_{\mathbf{w}}, \mathbb{S}) = \|\mathbf{w}\|$, for some norm $\|\cdot\|$. This means that $\mu(h_{\mathbf{w}}, \mathbb{S})$ can be set to the regularization term of the chosen objective function, so that the complexity,

hence the bound, will depend on it. This is not really new since for example uniform stability bounds allow one to consider such norms (see *e.g.*, KAKADE *et al.*, 2008). Here we want to illustrate that we simply incorporate some norms of parameters but as the reader may guess our framework allows one ton consider a broader family of complexity measures as we will see later. Given such a parametric function $\mu()$, the bound we derive in Theorem 7.3.1 takes the following form.

Definition 7.3.1 (Generalization Bound with Complexity Measures)**.** Let $\phi : [0,1]^2 \to \mathbb{R}$ be the generalization gap, $\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$ be a parametric function. A generalization bound with arbitrary complexity measures is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, there exists a real-valued function $\Phi_\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \times (0,1] \to \mathbb{R}$ such that for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m,\ h \sim \rho_{\mathbb{S}}} \left[\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_\mu(h, \mathbb{S}, \delta) \right] \geq 1 - \delta. \tag{7.1}$$

The main trick to obtain such a result is via a posterior distribution $\rho_{\mathbb{S}}$: we incorporate the function $\mu()$ by fixing the distribution $\rho_{\mathbb{S}}$ as a Gibbs distribution defined by

$$\rho_{\mathbb{S}}(h) \propto \exp\left[-\alpha \mathsf{R}_{\mathcal{S}}(h) - \mu(h, \mathbb{S}) \right], \quad \text{where} \quad \alpha \in \mathbb{R}^+. \tag{7.2}$$

This Gibbs distribution $\rho_{\mathbb{S}}$ is interesting from an optimization point of view: a hypothesis h is more likely to be sampled from it when the objective function $h \mapsto \mathsf{R}_{\mathcal{S}}(h) + \frac{1}{\alpha}\mu(h, \mathbb{S})$ is low for a given $\mathbb{S}$. For example, when $\mu(h, \mathbb{S}) = 0$, a hypothesis h is more likely to be sampled when its empirical risk $\mathsf{R}_{\mathcal{S}}(h)$ is low. Conversely, when $\mu()$ is non-constant, the function serves as a "regularizing term", so that a hypothesis is more likely to be sampled when the trade-off $\mathsf{R}_{\mathcal{S}}(h) + \frac{1}{\alpha}\mu(h, \mathbb{S})$ is low. In both cases, the higher α, the more the density of $\rho_{\mathbb{S}}$ is concentrated around those hypotheses that minimize the empirical risk. Moreover, it seems that Equation (7.2) is restrictive but it can actually represent any probability density functions. Indeed, let $\rho'_{\mathbb{S}}$ be a probability distribution on $\mathbb{H}$, *e.g.*, a Gaussian or a Laplace distribution, by setting $\mu(h, \mathbb{S}) = -\alpha \mathsf{R}_{\mathcal{S}}(h) - \ln \rho'_{\mathbb{S}}(h)$ we can retrieve the distribution $\rho'_{\mathbb{S}}$. Actually, the Gibbs distribution is well-known and studied in learning theory as discussed in the next section.

7.3.2 About the Gibbs Distribution

In this section, we would like to highlight two main lines of works that are related to our setting: *(i)* the usage of the Gibbs distribution in the "classical" PAC-Bayesian theory and *(ii)* the link between this distribution and optimization.

Gibbs distribution in the PAC-Bayesian theory. The Gibbs distribution has started to be studied in the PAC-Bayesian theory by CATONI (2004, 2007). Moreover, ALQUIER *et al.* (2016, Theorem 4.2 and 4.3) developed PAC-Bayesian generalization bounds with the Gibbs distribution considered in Equation (7.2) with $\mu(h, \mathbb{S}) = 0$ as posterior. However, their theorems analyze the expected true risk $\mathbb{E}_{h\sim\rho_{\mathbb{S}}} \mathsf{R}_{\mathcal{D}}(h)$ while we are interested in a *single* hypothesis h sampled from $\rho_{\mathbb{S}}$. Moreover, their bounds involve the non-computable and hypothesis-set-dependent KL divergence between the Gibbs distribution and a prior distribution. Hence, the computation of the KL divergence must be upper-bounded to allow one to instantiate this bound in practice. As we will see further, the bounds of Theorem 7.3.1 and Corollary 7.3.1 do not have this issue since they only require to know the expression of the density (up to the normalization constant) for $h \sim \rho_{\mathbb{S}}$ and $h' \sim \pi$.

Relationship between optimization and the Gibbs distribution. Given an objective function $f : \mathbb{R}^D \to \mathbb{R}$, the Stochastic Gradient LANGEVIN Dynamics (SGLD) algorithm (WELLING and TEH, 2011) learns some parameter $\mathbf{w} \in \mathbb{R}^D$ by running several iterations of the form

$$\mathbf{w}_t \longleftarrow \mathbf{w}_{t-1} - \eta\nabla f(\mathbf{w}) + \sqrt{\frac{2\eta}{\alpha}}\epsilon_t, \quad \text{where} \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D), \tag{7.3}$$

where $\mathbf{w}_t$ is the weight vector learned at the iteration $t \in \mathbb{N}$, the parameter η is the learning rate and α is the concentration parameter of the Gibbs distribution. This algorithm has an interesting feature: when the learning rate η tends to zero, the SGLD algorithm becomes a continuous-time process, called LANGEVIN diffusion, defined as the following stochastic differential equation in Equation (7.4). Indeed, Equation (7.3) can be seen as the EULER-MARUYAMA discretization (see *e.g.*, RAGINSKY *et al.*, 2017) of Equation (7.4) defined for $t \geq 0$ as

$$d\mathbf{w}(t) = -\nabla f(\mathbf{w}(t))dt + \sqrt{2\alpha}\mathbf{B}(t), \tag{7.4}$$

where $\mathbf{B}(t)$ is the Brownian motion in $\mathbb{R}^D$. Under some mild assumptions on the function f, CHIANG *et al.* (1987) show that the invariant distribution of the LANGEVIN diffusion is the Gibbs distribution proportional to $\exp(-\alpha f(\mathbf{w}))$.

7.3.3 Our Main Results: Generalization Bound with Complexity Measures

Thanks to the definition of $\rho_{\mathbb{S}}$, we now state our main result which consists in providing a bound on the generalization gap involving our parametric measure μ with respect to a posterior $\rho_{\mathbb{S}}(h) \propto \exp\left[-\alpha \mathsf{R}_{\mathcal{S}}(h) - \mu(h, \mathbb{S})\right]$.

Theorem 7.3.1 (Generalization Bound with Complexity Measures)**.** Let $\phi : [0,1]^2 \to \mathbb{R}$ be the generalization gap. For any $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any $\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,\ h'\sim\pi,\ h\sim\rho_{\mathbb{S}}} \left[\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \left[\alpha \mathsf{R}_{\mathcal{S}}(h') + \mu(h', \mathbb{S})\right] - \left[\alpha \mathsf{R}_{\mathcal{S}}(h) + \mu(h, \mathbb{S})\right] + \ln \frac{\pi(h')}{\pi(h)} + \ln \left(\frac{4}{\delta^2} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} \exp\left[\phi(\mathsf{R}_{\mathcal{D}}(h'), \mathsf{R}_{\mathcal{S}'}(h'))\right] \right) \right] \geq 1 - \delta.$$

Proof. Deferred to Appendix G.1. ■

This theorem is general since it depends only on the functions $\phi()$ and $\mu()$ that must be fixed by the practitioner. Given $\phi()$ and $\mu()$, we can note an element that can be surprising at the first reading: it appears indeed possible to sample hypotheses with a bad objective $\mathsf{R}_{\mathcal{S}}(h) + \frac{1}{\alpha}\mu(h, \mathbb{S})$ value and to obtain a tight generalization bound. However, by definition the Gibbs distribution $\rho_{\mathbb{S}}$, such a sampled hypothesis $h \sim \rho_{\mathbb{S}}$ is less probable since the density is higher when the objective is low. On the other hand, it is highly probable to sample a hypothesis h from $\rho_{\mathbb{S}}$ with a good objective $\mathsf{R}_{\mathcal{S}}(h) + \frac{1}{\alpha}\mu(h, \mathbb{S})$ value. Concerning the tightness of the bound, it may appear loose. However to get a bound that converges when m increases, it is sufficient to fix $\phi()$ as a function of m such as $\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) = 2m[\mathsf{R}_{\mathcal{D}}(h) - \mathsf{R}_{\mathcal{S}}(h)]^2$ or $\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) = m\,\mathrm{kl}[\mathsf{R}_{\mathcal{S}}(h) \| \mathsf{R}_{\mathcal{D}}(h)]$. Then, the tightness of the bound depends on the generalization gap $\phi()$, the parametric function $\mu()$ and α. The remaining challenge is to upper-bound $\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m} \mathbb{E}_{h'\sim\pi} \exp[\phi(\mathsf{R}_{\mathcal{D}}(h'), \mathsf{R}_{\mathcal{S}'}(h'))]$ and $\mathbb{E}_{h'\sim\pi} \ln \frac{\pi(h')}{\pi(h)}$ to get a practical bound. As an illustration we provide in the next corollary an instantiation of Theorem 7.3.1 with the generalization gap $\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) = 2m[\mathsf{R}_{\mathcal{D}}(h) - \mathsf{R}_{\mathcal{S}}(h)]^2$ and $\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) = m\,\mathrm{kl}[\mathsf{R}_{\mathcal{S}}(h) \| \mathsf{R}_{\mathcal{D}}(h)]$ when π is a uniform distribution on a bounded hypothesis set $\mathbb{H}$.

Corollary 7.3.1 (Practical Generalization Bound with Complexity Measures)**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any bounded hypothesis set $\mathbb{H}$, given the uniform prior distribution π on $\mathbb{H}$, for any $\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$, for any $\delta \in (0,1]$, with

probability at least $1-\delta$ over $\mathbb{S} \sim \mathcal{D}^m$, $h' \sim \pi$, $h \sim \rho_{\mathbb{S}}$ we have

$$\mathrm{kl}\left[\mathsf{R}_{\mathcal{S}}(h) \| \mathsf{R}_{\mathcal{D}}(h)\right] \leq \frac{1}{m}\left[\left[\alpha \mathsf{R}_{\mathcal{S}}(h') + \mu(h', \mathbb{S})\right] - \left[\alpha \mathsf{R}_{\mathcal{S}}(h) + \mu(h, \mathbb{S})\right] + \frac{8\sqrt{m}}{\delta^2}\right]_+, \tag{7.5}$$

$$\left|\mathsf{R}_{\mathcal{D}}(h) - \mathsf{R}_{\mathcal{S}}(h)\right| \leq \sqrt{\frac{1}{2m}\left[\left[\alpha \mathsf{R}_{\mathcal{S}}(h') + \mu(h', \mathbb{S})\right] - \left[\alpha \mathsf{R}_{\mathcal{S}}(h) + \mu(h, \mathbb{S})\right] + \frac{8\sqrt{m}}{\delta^2}\right]_+}, \tag{7.6}$$

where $[a]_+ = \max(0, a)$, and $\rho_{\mathbb{S}}$ is the Gibbs distribution defined by Equation (7.2).

Proof. Deferred to Appendix G.2. ■

Interestingly, Corollary 7.3.1 gives a bound on $|\mathsf{R}_{\mathcal{D}}(h) - \mathsf{R}_{\mathcal{S}}(h)|$ and $\mathrm{kl}\left[\mathsf{R}_{\mathcal{S}}(h) \| \mathsf{R}_{\mathcal{D}}(h)\right]$ where all terms except $\mathsf{R}_{\mathcal{D}}(h)$ are computable. To compute Equations (7.5) and (7.6) we can rearrange the terms to obtain a generalization bound on the true risk $\mathsf{R}_{\mathcal{D}}(h)$. Indeed, we have respectively

$$\mathsf{R}_{\mathcal{D}}(h) \leq \overline{\mathrm{kl}}\left(\mathsf{R}_{\mathcal{S}}(h) \;\middle|\; \frac{1}{m}\left[\left[\alpha \mathsf{R}_{\mathcal{S}}(h') + \mu(h', \mathbb{S})\right] - \left[\alpha \mathsf{R}_{\mathcal{S}}(h) + \mu(h, \mathbb{S})\right] + \frac{8\sqrt{m}}{\delta^2}\right]_+\right), \tag{7.7}$$

$$\text{and } \mathsf{R}_{\mathcal{D}}(h) \leq \mathsf{R}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m}\left[\left[\alpha \mathsf{R}_{\mathcal{S}}(h') + \mu(h', \mathbb{S})\right] - \left[\alpha \mathsf{R}_{\mathcal{S}}(h) + \mu(h, \mathbb{S})\right] + \frac{8\sqrt{m}}{\delta^2}\right]_+}, \tag{7.8}$$

where $\overline{\mathrm{kl}}(q|\tau) = \max\left\{p \in (0,1) \;\middle|\; \mathrm{kl}(q\|p) \leq \tau\right\}$ (see Definition 2.3.3). These bounds are used in Section 7.4 to illustrate the generalization guarantees with different values of the parametric function $\mu()$ and α. Moreover, as illustrated in the experiments, Equation (7.7) is a tighter bound on the true risk $\mathsf{R}_{\mathcal{D}}(h)$ than Equation (7.8): we can prove it formally with Pinsker's inequality (see all Equation (2.15)).

Nevertheless, the complexity measures $\Phi_\mu(h, \mathbb{S}, \delta)$ of Equations (7.5) and (7.6) enjoys asymptotic convergence for $m \to +\infty$. However, as mentioned previously, the complexity also depends on α and $\mu()$: the convergence rate can then be degraded by $[\alpha \mathsf{R}_{\mathcal{S}}(h') + \mu(h', \mathbb{S})] - [\alpha \mathsf{R}_{\mathcal{S}}(h) + \mu(h, \mathbb{S})]$ which may be potentially large. For example, for all $(h, \mathbb{S}) \in \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m$, if the empirical risk $\mathsf{R}_{\mathcal{S}}(h')$ is large (which is often the case when h' is sampled from a uniform prior on $\mathbb{H}$), $\alpha = m$, and $\mu(h, \mathbb{S}) = 0$, then

the complexity measure, simplified into $\Phi_\mu(h, \mathbb{S}, \delta) = \left[[\mathrm{R}_\mathcal{S}(h') - \mathrm{R}_\mathcal{S}(h)] + \frac{1}{m} \ln \frac{2\sqrt{m}}{\delta}\right]_+$ for $\phi(\mathrm{R}_\mathcal{D}(h), \mathrm{R}_\mathcal{S}(h)) = \mathrm{kl}[\mathrm{R}_\mathcal{S}(h) \| \mathrm{R}_\mathcal{D}(h)]$, will be large. We thus have to set α and $\mu()$ such that *(i)* $\rho_\mathbb{S}$ allows to sample a hypothesis h associated with a low objective function $h \mapsto \mathrm{R}_\mathcal{S}(h) + \frac{1}{\alpha}\mu(h, \mathbb{S})$ and *(ii)* the complexity measure $\Phi_\mu(h, \mathbb{S}, \delta)$ is tight, resulting in a meaningful bound. For example, with $\alpha = \sqrt{m}$ and $\mu(h, \mathbb{S}) = 0$, the distribution $\rho_\mathbb{S}$ will be less concentrated around the minimizers of the empirical risk, but the complexity measure will be tighter compared to the previous example: $\left[\frac{1}{\sqrt{m}}[\mathrm{R}_\mathcal{S}(h') - \mathrm{R}_\mathcal{S}(h)] + \frac{1}{m} \ln \frac{2\sqrt{m}}{\delta}\right]_+$.

The tightness of the bounds can be potentially improved since we choose a uniform distribution for the prior π. To obtain better bounds, data-dependent priors have been heavily used in the PAC-Bayesian literature (see *e.g.*, Parrado-Hernández *et al.*, 2012; Dziugaite *et al.*, 2021; Pérez-Ortiz *et al.*, 2021). However, we think that the uniform distribution helps to better understand the generalization phenomenon. Indeed, the hypothesis h sampled from the uniform distribution π has a high chance of underfitting. Hence, if the hypothesis $h \sim \rho_\mathbb{S}$ has its associated bound value that is tight, we can easily interpret this hypothesis generalizes well. On the other hand, if a generalization bound with a data-dependent prior π is tight, it means that a posterior $\rho_\mathbb{S}$ (not very far from the prior) allows us to generalize well. In this case, we do not know if the generalization capability is essentially due to the choice of a good prior π or if it comes mainly from the learned posterior $\rho_\mathbb{S}$.

7.4 Using Arbitrary Complexities in Practice

The bound of Corollary 7.3.1 is not directly usable in its current form: the remaining challenge is to sample h from the Gibbs distribution $\rho_\mathbb{S}$ defined in Equation (7.2); we address the sampling issue in Section 7.4.1. Then, we make use of the proposed solution to assess our bound in practice. Section 7.4.2.1 introduces our experimental setting and Section 7.4.2.2 gives an overview of the tightness of the bound. Additionally, we give an overview of the influence of α and the number of parameters in Sections 7.4.2.3 and 7.4.2.4.

7.4.1 Sampling from the Gibbs Distribution

Sampling from the Gibbs distribution of Equation (7.2) is a hard task: naively, it requires to evaluate the function $h \mapsto -\alpha \mathrm{R}_\mathcal{S}(h) - \mu(h, \mathbb{S})$ for all $h \in \mathbb{H}$, which is intractable when $\mathbb{H}$ is infinite or even large. We tackle this issue for over-parameterized models, which we later consider in Section 7.4.2.1 in an empirical study of our bound. Let us consider a set $\mathbb{H}$ of hypotheses $h_\mathbf{w}$ parameterized by $\mathbf{w} \in \mathbb{R}^D$, and a tractable

Algorithm 7.1 Stochastic MALA

1: **Input:** Learning set $\mathbb{S}$, weights $\mathbf{w}$, function $\mu()$, loss function $\ell()$
2: **Hyperparameters:** Number of iterations T, learning rate η, parameter α
3: **for** $t \leftarrow 1 \dots T$ **do**
4: $\quad \mathbb{U} \leftarrow$ Sample (without replacement) a mini-batch from $\mathbb{S}$
5: $\quad \mathbf{w}' \leftarrow$ Sample from the distribution $\mathrm{P}^{\mathbf{w}}_{\mathbb{U}}$
6: $\quad \tau \leftarrow \min\left(1, \frac{\rho_{\mathbb{U}}(\mathbf{w}')\mathrm{P}^{\mathbf{w}'}_{\mathbb{U}}(\mathbf{w})}{\rho_{\mathbb{U}}(\mathbf{w})\mathrm{P}^{\mathbf{w}}_{\mathbb{U}}(\mathbf{w}')}\right)$
7: $\quad u \leftarrow$ Sample from the distribution $\mathrm{Uni}(0,1)$
8: $\quad$ **if** $u \leq \tau$ **then**
9: $\quad\quad \mathbf{w} \leftarrow \mathbf{w}'$
10: **return** $\mathbf{w}$

distribution denoted $\mathrm{P}^{\mathbf{w}}_{\mathbb{U}}$ (*e.g.*, a Gaussian distribution) such that its density approximates the density of $\rho_{\mathbb{S}}$. In this setting, to learn such an auxiliary distribution, we propose in Algorithm 7.1 a stochastic version of the Metropolis Adjusted Langevin Algorithm (MALA, Besag (1994))[1]. Its objective is to generate samples from $\rho_{\mathbb{S}}$ by iteratively refining the auxiliary distribution, that we define as

$$\mathrm{P}^{\mathbf{w}}_{\mathbb{U}} = \mathcal{N}\left(\mathbf{w} - \eta\nabla\Big[\mathrm{R}^{\ell}_{\mathcal{U}}(\mathbf{w}) + \frac{1}{\alpha}\,\mu(\mathbf{w},\mathbb{U})\Big],\ \tfrac{2\eta}{\alpha}\mathbf{I}\right), \tag{7.9}$$

where $\mathrm{R}^{\ell}_{\mathcal{U}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{U}}\,\ell(h_{\mathbf{w}},(\mathbf{x},y))$ is the empirical risk on the mini-batch $\mathbb{U} \subseteq \mathbb{S}$, and $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$ is a loss function. Concretely, we initialize the parameters $\mathbf{w}$ of the model as the output of an optimization algorithm (Vanilla SGD in our case). Then, we refine them as follows: at each iteration, given the current weights $\mathbf{w}$ and a mini-batch $\mathbb{U} \subseteq \mathbb{S}$ (Line 4), we sample a candidate vector $\mathbf{w}'$ (Line 5) according to the distribution $\mathrm{P}^{\mathbf{w}}_{\mathbb{U}}$; then (Line 6 to 9) we decide to reject or accept the new candidate to become our current weights $\mathbf{w}$, depending on its ratio $\tau = \min\left(1, \frac{\rho_{\mathbb{U}}(\mathbf{w}')\mathrm{P}^{\mathbf{w}'}_{\mathbb{U}}(\mathbf{w})}{\rho_{\mathbb{U}}(\mathbf{w})\mathrm{P}^{\mathbf{w}}_{\mathbb{U}}(\mathbf{w}')}\right)$ being larger than a control value u sampled from the uniform distribution on $[0,1]$. Under the mild assumption that $\rho_{\mathbb{S}}$ is absolute continuous *w.r.t.* $\mathrm{P}^{\mathbf{w}}_{\mathbb{S}}$ (see Chib and Greenberg, 1995, for details), when the number of iterations tends to infinity and when $\mathbb{U}=\mathbb{S}$, the returned $\mathbf{w}$ is sampled according to $\rho_{\mathbb{S}}$ (Smith and Roberts, 1993). Note that this assumption requires that the tractable distribution $\mathrm{P}^{\mathbf{w}}_{\mathbb{S}}$ has a strictly positive density when the density of $\rho_{\mathbb{S}}$ is strictly positive as well (see Chib and Greenberg, 1995).

[1] See Chib and Greenberg (1995) for an introduction on Metropolis-Hastings Algo on which MALA is based.

7.4.2 Experiments

7.4.2.1 Experimental Setting

In this section, we investigate the tightness of our bound of Equations (7.7) and (7.8) on the MNIST (LeCun *et al.*, 1998) and FashionMNIST (Xiao *et al.*, 2017) datasets. We keep the original learning set as $\mathbb{S}$ and the original test set $\mathbb{T}$ to estimate the true risk, that we refer to as test risk $\mathrm{R}_{\mathcal{T}}(h)$.

Model. We use a "Convolutional Network in Network" (Lin *et al.*, 2013) similarly to Jiang *et al.* (2019) and Dziugaite *et al.* (2020) , that consists of several modules of 3 convolutional layers each followed by a leaky ReLU activation function (its negative slope is set to 10^{-2}). The depth of the network L is the number of convolutional layers, and the width H is the number of channels of each convolution. In addition, for each layer i, we denote its weights by $\mathbf{w}_i$. More precisely, the modules of this model can be described as follows. A module takes two parameters as argument: the number of input channels H_{in} and the number of output channels H_{out} and applies consecutively three convolutional layers (each followed by a leaky ReLU activation function). The first layer is composed of a $3{\times}3$ kernel (where the stride *resp.* padding is set to 2 *resp.* 1) with H_{in} channels as input and H_{out} as output. The two other layers have a $1{\times}1$ kernel with H_{out} channels as input and output. Then, the network is constructed as follows: *(a)* we have a module where $H_{\text{out}} = H$ and H_{in} is the number of channels in the input *(b)* we have $(L/3) - 1$ modules with $H_{\text{in}} = H_{\text{out}} = H$ and *(c)* we have a convolutional layer with a $1{\times}1$ kernel with $\mathrm{card}(\mathbb{Y})$ channels as output followed by a leaky ReLU activation and an average pooling layer. In the experiments, we consider $L{\in}\{9, 12, 15\}$ and $H{\in}\{128, 256\}$. Furthermore, we initialize the network with the weights $\mathbf{w}^0{\in}\mathbb{R}$ obtained the uniform Kaiming He initializer (He *et al.*, 2015). The set $\mathbb{H}$ corresponds to the hypotheses $h_{\mathbf{w}}$ that can be obtained from this initialization (and we clamp the weights during the optimization in the initialization's interval).

Arbitrary complexity measures. We study different complexity measures parametrized by different functions $\mu()$ from Jiang *et al.* (2019, Sec. C)[2]. Indeed, we consider the

[2]Note we consider a subset of the functions studied by Jiang *et al.*: we select those that are optimizable.

6 following parametric functions $\mu()$:

$$\textsc{Dist\_Fro}(h_{\mathbf{w}}) = \sum_{i=1}^{L} \|\mathbf{w}_i - \mathbf{w}_i^0\|_2, \quad \text{and} \quad \textsc{Dist\_L}_2(h_{\mathbf{w}}) = \|\mathbf{w} - \mathbf{w}^0\|_2,$$

$$\text{and} \quad \textsc{Param\_Norm}(h_{\mathbf{w}}) = \sum_{i=1}^{L} \|\mathbf{w}_i\|_2^2, \quad \text{and} \quad \textsc{Path\_Norm}(h_{\mathbf{w}}) = \sum_{i=1}^{\text{card}(\mathbb{Y})} h_{\mathbf{w}^2}(\mathbf{1})[i],$$

$$\text{and} \quad \textsc{Sum\_Fro}(h_{\mathbf{w}}) = L\left(\prod_{i=1}^{L} \|\mathbf{w}_i\|_2^2\right)^{\frac{1}{L}}, \quad \text{and} \quad \textsc{Zero}(h_{\mathbf{w}}) = 0.$$

We define the considered measures with α taken among 5 values uniformly spaced between $[\sqrt{m}, m]$. Note that we analyze other parametric functions that depends on the learning sample $\mathbb{S}$. However, since the results are similar, we decided to defer the results in Appendix G.

Bound optimization. To compute our bound in Equations (7.7) and (7.8), we aim to minimize the objective function $\mathbf{w} \mapsto \mathrm{R}_{\mathcal{U}}^{\ell}(\mathbf{w}) + \frac{1}{\alpha}\mu(\mathbf{w}, \mathbb{U})$, via Algorithm 7.1. We set the loss function to the bounded cross entropy from Dziugaite and Roy (2018): $\ell(h, (\mathbf{x}, y)) = -\frac{1}{4}\ln(e^{-4} + (1 - 2e^{-4})h[y])$, where $h[y]$ is the probability assigned to label y by h. The advantage of Dziugaite and Roy (2018)'s cross-entropy is that it lies in $\ell(h, (\mathbf{x}, y)) \in [0, 1]$, whereas the classical cross-entropy is unbounded. Indeed, taking into account the classical cross-entropy when optimizing the objective function would lead to focus too much on the risk minimization, while we want to take into account $\frac{1}{\alpha}\mu(\mathbf{w}, \mathbb{U})$. We initialize the weights $\mathbf{w} \in \mathbb{R}^D$ to the solution found by optimizing the objective function with a Vanilla SGD (with 10 epochs, a learning rate of 10^{-1}, and a batch size of 64). Given these initial parameters $\mathbf{w}$, we execute Algorithm 7.1 for 1 epoch with a mini-batch of size 64, where $\eta = 10^{-4}$.

7.4.2.2 Tightness of the Bounds

For each parametric function $\mu()$, we report in Figures 7.1 and 7.2, the test risks $\mathrm{R}_{\mathcal{T}}(h)$ and the values of the tightest bound (*w.r.t.* α) associated to Equations (7.7) and (7.8) for different parameters (depth L, width H). First of all, we can remark that certain empirical risks are high. This is due to the sampling of the hypothesis h from the distribution $\rho_{\mathbb{S}}$: the hypothesis does not necessarily minimizes the objective function $h \mapsto \mathrm{R}_{\mathcal{S}}(h) + \frac{1}{\alpha}\mu(h, \mathbb{S})$. We can nevertheless observe that the bounds' values are higher when the empirical risk $\mathrm{R}_{\mathcal{T}}(h)$ is low. This can be explained by the fact that $[\alpha \mathrm{R}_{\mathcal{S}}(h') + \mu(h', \mathbb{S})] - [\alpha \mathrm{R}_{\mathcal{S}}(h) + \mu(h, \mathbb{S})]$ is large in this case. When the empirical risks are a bit higher, the bounds become tighter for certain parametric function such as $\textsc{Dist\_L}_2$, $\textsc{Sum\_Fro}$. This confirms that there is an interest to use a parametric function that captures information on the model during the training phase.

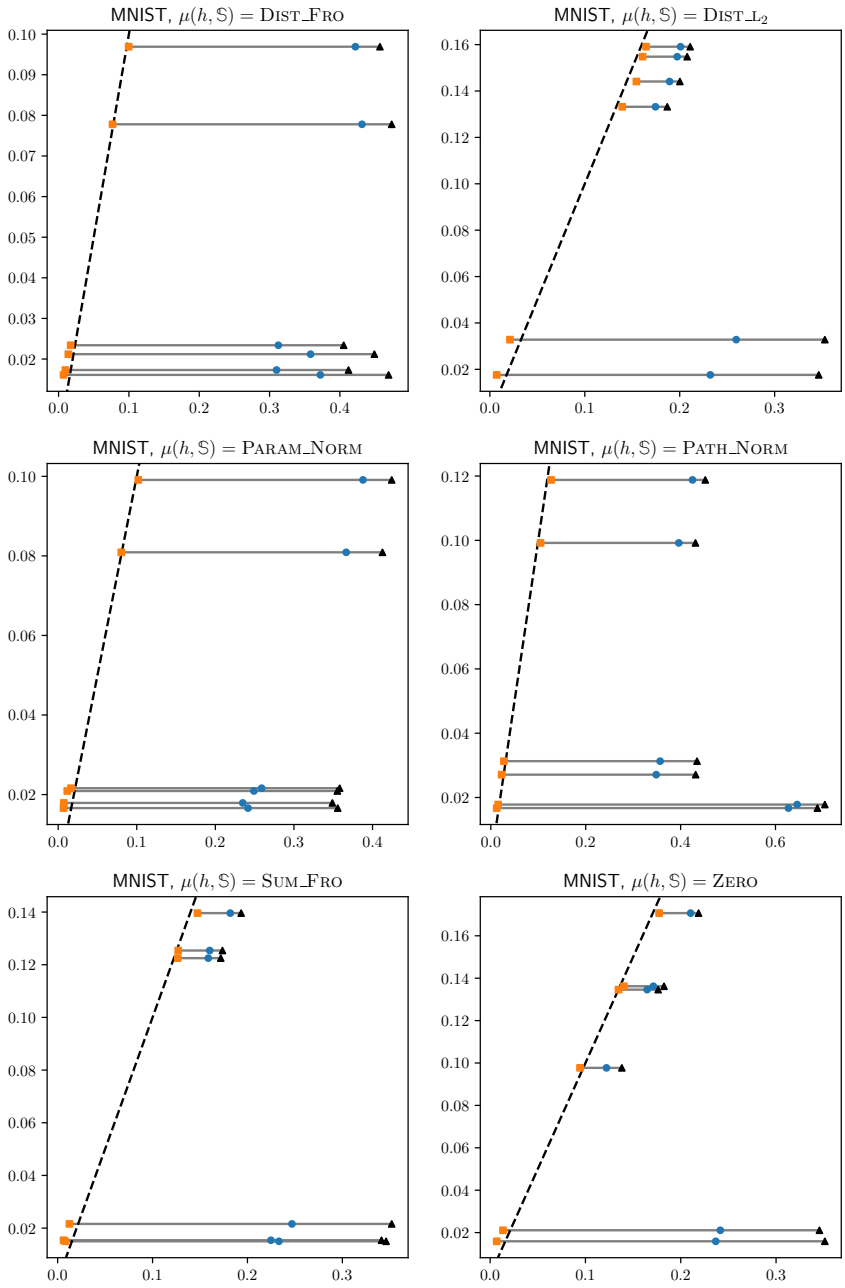


Figure 7.1. *Scatter plot given a parametric function $\mu(h, \mathbb{S})$, where each segment represents a neural network $h_{\mathbf{w}}$ learned with a given α, width H and depth L. For each segment, there is a corresponding orange square and a blue circle. The orange squares corresponds to the empirical risk $R_{\mathcal{S}}(h)$ (x-axis) and test risk $R_{\mathcal{T}}(h)$ (y-axis). The blue circle resp. the black triangle represents Equation (7.7) resp. Equation (7.8) in the x-axis and the test risk $R_{\mathcal{T}}(h)$ in the y-axis. The dashed line is the identity function.*

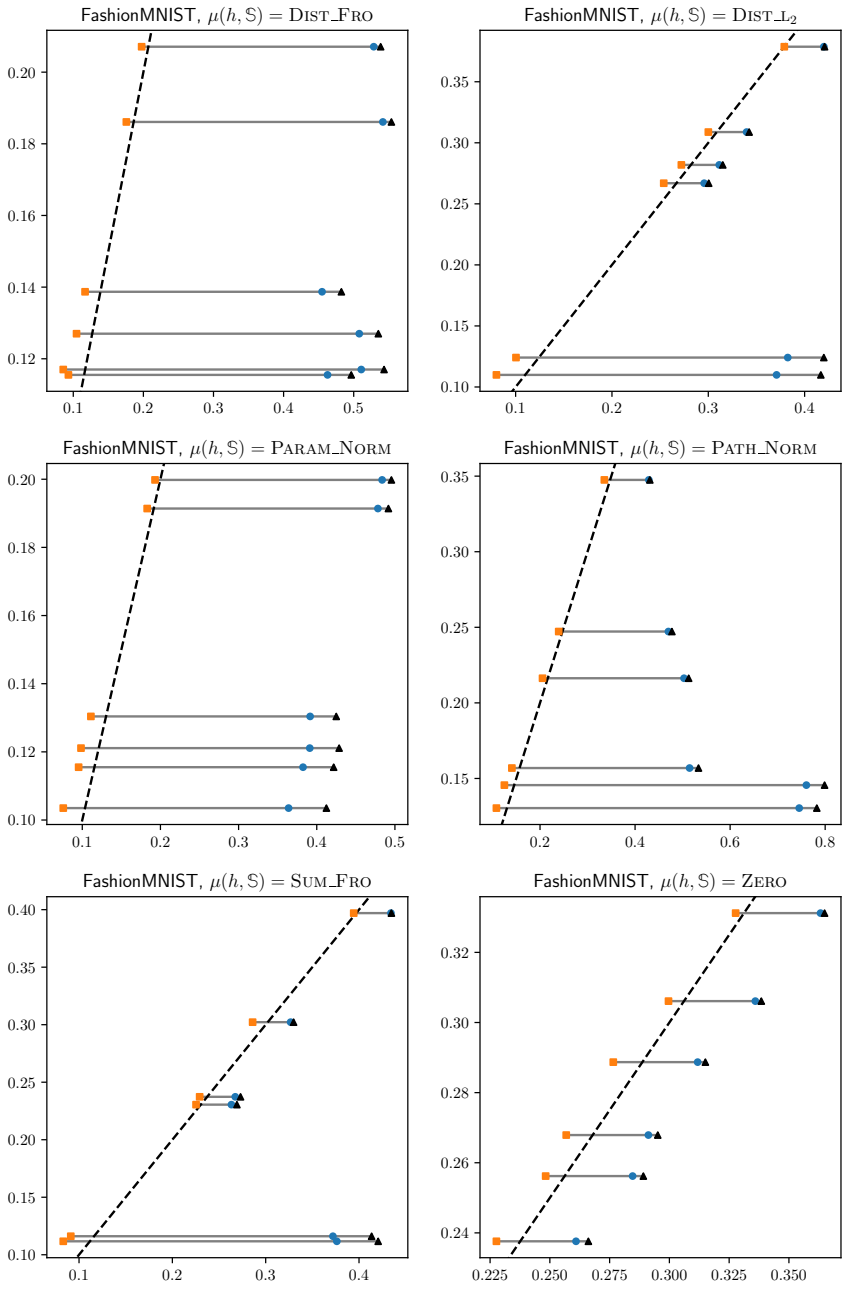


Figure 7.2. *Scatter plot given a parametric function $\mu(h, \mathbb{S})$, where each segment represents a neural network $h_{\mathbf{w}}$ learned with a given α, width H and depth L. For each segment, there is a corresponding orange square and a blue circle. The orange squares corresponds to the empirical risk $R_{\mathcal{S}}(h)$ (x-axis) and test risk $R_{\mathcal{T}}(h)$ (y-axis). The blue circle resp. the black triangle represents Equation (7.7) resp. Equation (7.8) in the x-axis and the test risk $R_{\mathcal{T}}(h)$ in the y-axis. The dashed line is the identity function.*

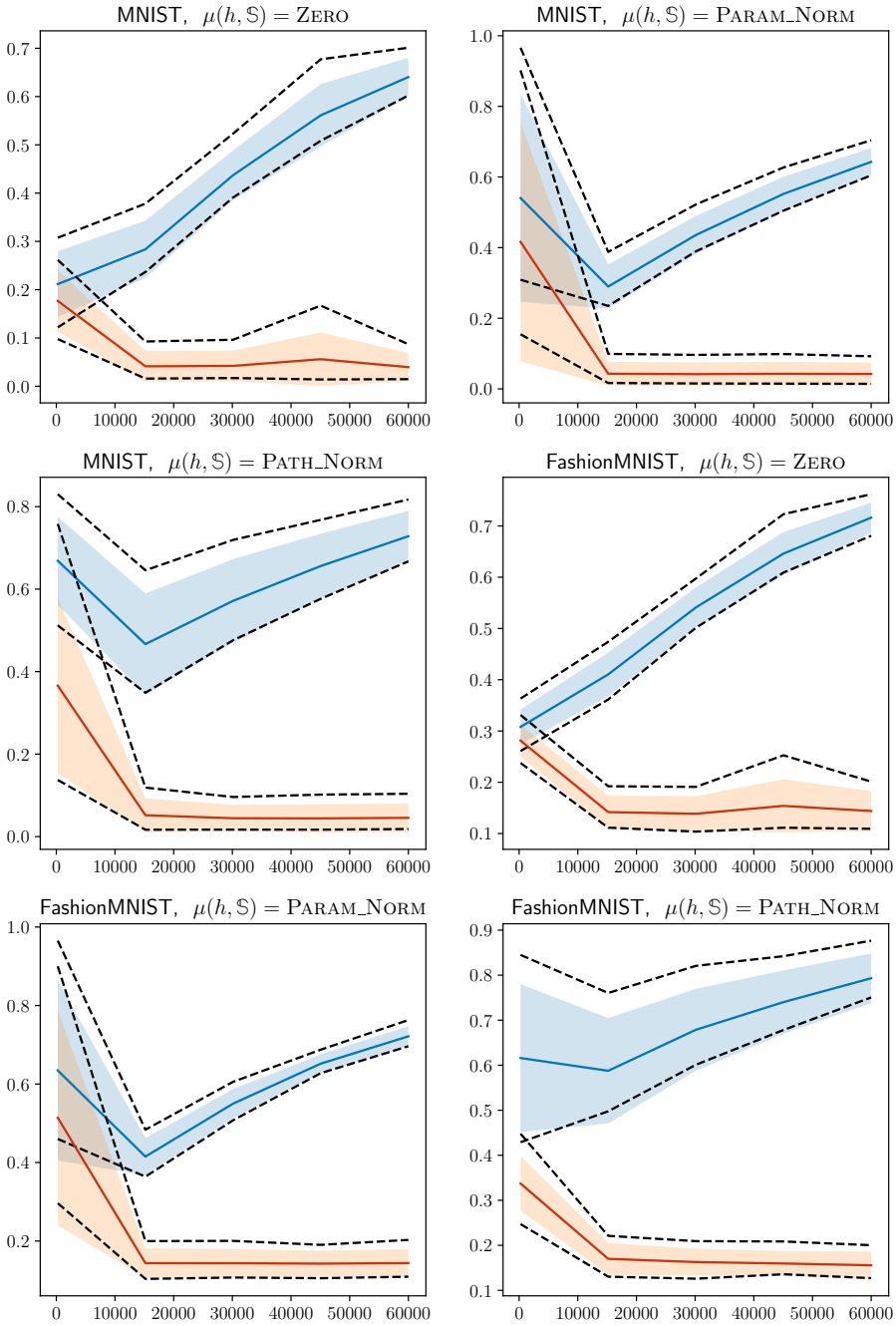


Figure 7.3. *Influence of the parameter α (in the x-axis) for three parametric functions:* ZERO, PARAM_NORM, *and* PATH_NORM *for MNIST and FashionMNIST. The bound values are represented in blue and the test risk in red. The two (solid) lines are the mean values computed on the depths and widths; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.*

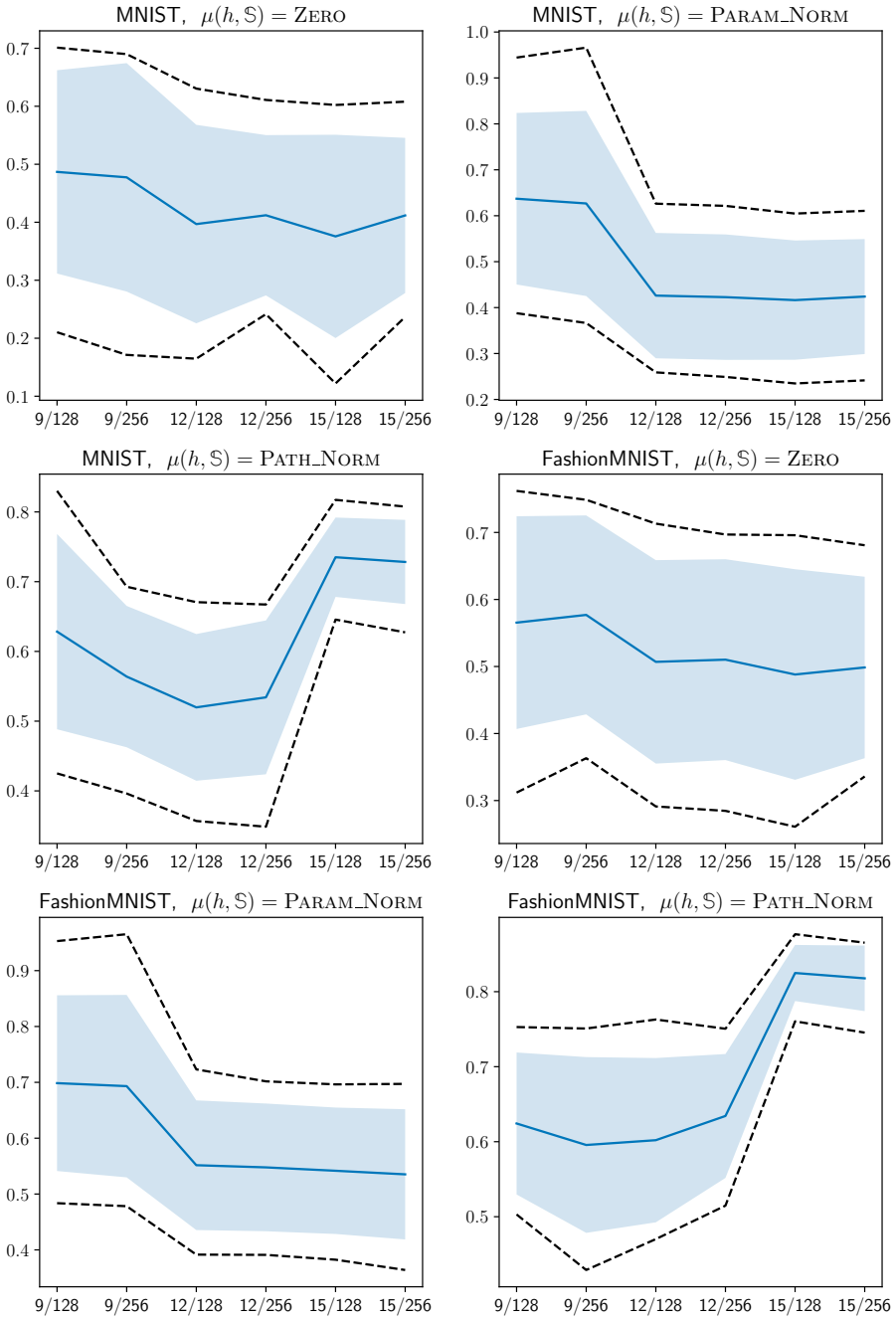


Figure 7.4. *Influence of the depth and the width (in the x-axis as "depth/width") for three parametric functions:* ZERO, PARAM_NORM, *and* PATH_NORM *for MNIST and FashionMNIST. The (solid) lines are the mean values computed on the different values of α; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.*

7.4.2.3 Influence of the Parameter α

We analyze the influence of the parameter α in Equation (7.7). To do so, we plot an overview of the evolution for the bounds and the test risks $\mathrm{R}_{\mathcal{T}}(h)$; the details are reported in Appendix G. For each parameter α, we plot in Figure 7.3 the mean, the standard deviation, the minimum and the maximum for the different parameters (depth and width). In general, the bound increases when the α tends to m but the test risks $\mathrm{R}_{\mathcal{T}}(h)$ are less prone to variations. Indeed, the higher the parameter α, the more concentrated around the minimizers the hypothesis will be sampled. On the contrary, for a small α (*e.g.*, $\alpha = \sqrt{m}$), the Gibbs distribution defined in Equation (7.2) is less concentrated making the test risks potentially high with a tighter generalization bound.

7.4.2.4 Influence of the Depth/Width

In Figure 7.4, we show an overview of the evolution of Equation (7.7) with respect to the depth and the width. More precisely, we report the mean, the standard deviation, the minimum and the maximum values for three parametric functions (ZERO, PARAM_NORM, and PATH_NORM).

Interestingly, the evolution of the bounds highly depends on the chosen parametric function $\mu()$. For instance, the bound increases with PATH_NORM when the depth and the width increase. This is in contrast with PARAM_NORM that decreases when the number of parameters increases. This shows the interest of our framework: considering a user-specified complexity measure $\Phi_{\mu}()$ can help to understand the generalization of over-parameterized models (that are sampled from $\rho_{\mathbb{S}}$).

7.5 Comparison with the Generalization Bounds of the Literature

In this section, we theoretically compare generalization bounds with arbitrary complexity measures compared to literature's bounds. We prove that our bound generalizes the uniform-convergence and the algorithmic-dependent bounds. Additionally, we show that the algorithmic-dependent bounds can be tighter than uniform-convergence bounds. To do so, we propose in Propositions 7.5.1 to 7.5.3 a reinterpretation of the high probability bounds in terms of sets. Additionally, we prove in Corollaries 7.5.1 and 7.5.2 two special cases of our bound in Theorem 7.3.1 that generalizes the two types of bounds.

7.5.1 Bounds with Arbitrary Complexity Measures

In order to compare our framework with the uniform-convergence and the algorithmic-dependent bounds, we translate Definition 7.3.1 into the following set-theoretic result.

Proposition 7.5.1 (Set-theoretic view of Definition 7.3.1)**.** Let $\phi : [0,1]^2 \to \mathbb{R}$ be a generalization gap and assume that there exists a function $\Phi_\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \times (0,1] \to \mathbb{R}$ fulfilling Definition 7.3.1. Under these conditions, with $\mathbb{Z}_{\mathtt{d}} = \Big\{(h,\mathbb{S}) \in \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m : \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_\mu(h,\mathbb{S},\delta)\Big\}$, and $\mathbb{P}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}[(h,\mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}] \geq 1-\delta$, we have

$$\begin{aligned} \text{Equation (7.1)} &\iff \forall (h,\mathbb{S}) \in \mathbb{Z}_{\mathtt{d}},\ \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_\mu(h,\mathbb{S},\delta) \\ &\iff \sup_{(h,\mathbb{S})\in\mathbb{Z}_{\mathtt{d}}} \Big\{\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) - \Phi_\mu(h,\mathbb{S},\delta)\Big\} \leq 0. \end{aligned}$$

Proof. Deferred to Appendix G.3. ■

For a given confidence δ, with probability at least $1-\delta$, the bound is then valid for all $(h,\mathbb{S})$ belonging to a (reduced) set $\mathbb{Z}_{\mathtt{d}} \subseteq \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m$. In other words, the bound always holds for a given hypothesis and learning sample $(h,\mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}$, and its value depends on these h and $\mathbb{S}$. The generality of our framework can thus generalize both uniform convergence and algorithmic dependent bounds as we see in the rest of this section.

7.5.2 Uniform Convergence Bounds

Uniform-convergence-based bounds were the first type of generalization bounds to be introduced, notably in Vapnik and Chervonenkis (1971) using the VC-dimension as complexity. Other bounds were later developed based on the Gaussian/Rademacher complexity (Bartlett and Mendelson, 2002) instead. We recall the definition of this type of bounds encountered in Chapter 1.

Definition 1.3.2 (Uniform Convergence Bound)**.** Let $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to [0,1]$ be a loss function and $\phi : [0,1]^2 \to \mathbb{R}$ a generalization gap. A uniform convergence bound is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, there exists a function $\Phi_{\mathtt{u}} : (0,1] \to \mathbb{R}$, such that for any $\delta \in (0,1]$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\sup_{h\in\mathbb{H}} \phi(\mathsf{R}^\ell_{\mathcal{D}}(h), \mathsf{R}^\ell_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{u}}\Big(\delta\Big)\right] \geq 1-\delta, \tag{1.1}$$

where usually $\phi(\mathsf{R}^\ell_{\mathcal{D}}(h), \mathsf{R}^\ell_{\mathcal{S}}(h)) = \mathsf{R}^\ell_{\mathcal{D}}(h) - \mathsf{R}^\ell_{\mathcal{S}}(h)$.

Remember that this definition encompasses different complexity measures, such as $\Phi_{\mathrm{u}}(\delta)=\mathtt{rad}(\mathbb{H})+\sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}$ in Theorem 1.3.3, or $\Phi_{\mathrm{u}}(\delta)=\sqrt{\frac{1}{m}2\mathtt{vc}(\mathbb{H})\ln\frac{em}{\mathtt{vc}(\mathbb{H})}}+\sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}$ described in Theorem 1.3.2. For ease of comparison, we refine and reinterpret this type of bounds in a set-theoretic manner as follows. This result has been originally remarked by Nagarajan and Kolter (2019b) (but not proved).

Proposition 7.5.2 (Set-theoretic View of Uniform Convergence Bounds)**.** Let $\phi : [0,1]^2\to\mathbb{R}$ be a generalization gap and assume that there exists a function $\Phi_{\mathrm{u}} : (0,1] \to \mathbb{R}$ fulfilling Definition 1.3.2. Under these conditions, with $\mathbb{Z}_{\mathrm{u}} = \left\{\mathbb{S} \in (\mathbb{X}\times\mathbb{Y})^m : \forall h \in \mathbb{H},\, \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathrm{u}}(\delta)\right\}$, and $\mathbb{P}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathbb{S}\in\mathbb{Z}_{\mathrm{u}}\right]\geq 1-\delta$, we have

$$\begin{aligned}\text{Equation (1.1)} &\iff \forall \mathbb{S} \in \mathbb{Z}_{\mathrm{u}},\ \forall h\in\mathbb{H},\, \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathrm{u}}(\delta)\\ &\iff \sup_{\mathbb{S}\in\mathbb{Z}_{\mathrm{u}}}\sup_{h\in\mathbb{H}}\left\{\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h))\right\}\leq\Phi_{\mathrm{u}}(\delta).\end{aligned}$$

Proof. Deferred to Appendix G.4. ■

Proposition 7.5.2 is, in fact, a reinterpretation of PAC generalization bounds by identifying the subset $\mathbb{Z}_{\mathrm{u}}\subseteq(\mathbb{X}\times\mathbb{Y})^m$ for which the upper bound $\Phi_{\mathrm{u}}(\delta)$ is valid. This highlights their worst-case nature: given a confidence δ, the generalization gap $\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h))$ is upper-bounded by a complexity measure $\Phi_{\mathrm{u}}(\delta)$ for all $(h,\mathbb{S}) \in \mathbb{H}\times\mathbb{Z}_{\mathrm{u}}$. To get a bound holding with probability at least $1-\delta$, since the complexity $\Phi_{\mathrm{u}}(\delta)$ does not depend on h or $\mathbb{S}$, the complexity has to upper-bound the worst hypothesis $h \in \mathbb{H}$ and the worst learning sample $\mathbb{S}\in\mathbb{Z}_{\mathrm{u}}$. As a consequence, $\Phi_{\mathrm{u}}(\delta)$ is lower-bounded by $\sup_{\mathbb{S}\in\mathbb{Z}_{\mathrm{u}}}\sup_{h\in\mathbb{H}}\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h))$. As we have seen in Proposition 7.5.1, our bound is more permissive than the uniform convergence bounds since the upper bound can depend on the learning sample $\mathbb{S}$ and the hypothesis h. Hence, this dependence on $\mathbb{S}$ and h allows us to retrieve the uniform convergence bounds with our framework. Indeed, from Theorem 7.3.1, we can obtain the following generalization bound.

Corollary 7.5.1 (Uniform Convergence Bound from Theorem 7.3.1)**.** Let $\phi : [0,1]^2\to\mathbb{R}$ be the generalization gap and assume that there exists a function $\Phi_{\mathrm{u}} : (0,1] \to \mathbb{R}$ fulfilling Definition 1.3.2 such that $\Phi_{\mathrm{u}}(\delta) \geq \ln\left[\frac{4}{\delta^2}\,\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{h'\sim\pi}\exp\left(\phi(\mathsf{R}_{\mathcal{D}}(h'), \mathsf{R}_{\mathcal{S}'}(h'))\right)\right]$. For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathrm{u}}(\delta)\right] \geq 1-\delta.$$

Proof. Deferred to Appendix G.5. ■

Note that, to prove Corollary 7.5.1, we require an additional assumption: a lower-bound on $\Phi_{\mathtt{u}}(\delta)$. When the generalization gap is $\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h))=2m[\mathsf{R}_{\mathcal{D}}(h)-\mathsf{R}_{\mathcal{S}}(h)]^2$ or $\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h))=\mathrm{kl}[\mathsf{R}_{\mathcal{S}}(h)\|\mathsf{R}_{\mathcal{D}}(h)]$, the lower bound is $\ln\frac{8\sqrt{m}}{\delta^2}$ (see Corollary 7.3.1) which is low enough to be a worst-case upper-bound. To sum up, our framework is general enough to retrieve classical uniform convergence bounds (under the mild assumption) such that the ones based on the Rademacher complexity (Definition 1.3.4) or the VC-Dimension (Definition 1.3.3). In practice, the sampling involved in the bound of Corollary 7.5.1 is not necessary: the bound holds for all hypothesis $h \in \mathbb{H}$ with high probability. More precisely, for $\Phi_{\mathtt{u}}(h,\mathbb{S},\delta) = \Phi_{\mathtt{u}}(\delta)$, the set $\mathbb{Z}_{\mathtt{u}}$ in Proposition 7.5.2 can be seen as a subset of $\mathbb{Z}_{\mathtt{d}}$ since

$$\left\{\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m \;\middle|\; \forall h\in\mathbb{H},\ (\mathbb{S},h)\in\mathbb{Z}_{\mathtt{d}}\right\} = \mathbb{Z}_{\mathtt{u}}.$$

Hence, for a well-behaved learning sample $\mathbb{S}$, *i.e.*, for $\mathbb{S}\in\mathbb{Z}_{\mathtt{u}}$, we have that

$$\begin{aligned}\mathop{\mathbb{E}}_{h\in\rho_{\mathbb{S}}} \mathrm{I}\left[\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h)) \le \Phi_{\mathtt{u}}(\delta)\right] &= \mathrm{I}\left[\sup_{h\in\mathbb{H}}\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h)) \le \Phi_{\mathtt{u}}(\delta)\right]\\ &= 1,\end{aligned}$$

which gives a valid bound for all $h \in \mathbb{H}$ without sampling. This generality of this framework does not apply uniquely to these type of bounds. We can obtain a result similar for the algorithmic-dependent bounds that can be tighter than the uniform-convergence-based bounds.

7.5.3 Algorithmic-Dependent Bounds

The upper bound $\Phi_{\mathtt{u}}(\delta)$ can generally be improved by considering algorithmic-dependent bounds (Bousquet and Elisseeff, 2002; Xu and Mannor, 2012). In this case, only the output $h_{\mathbb{S}}$ of a learning algorithm given $\mathbb{S}$ is studied: we only bound the generalization gap $\phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}),\mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}}))$ specific to $h_{\mathbb{S}}$ (here, $\mathbb{H}=\{h_{\mathbb{S}}\}_{\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m}$). The definition of such bounds encountered in Chapter 1 is recalled in the following.

Definition 1.3.5 (Algorithmic-dependent Generalization Bound)**.** Let $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y}) \to [0,1]$ be a loss function and $\phi : [0,1]^2\to\mathbb{R}$ a generalization gap. An algorithmic-dependent generalization bound is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, there exists a function $\Phi_{\mathtt{a}} : (0,1]\to\mathbb{R}$, such that for any $\delta\in(0,1]$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\phi(\mathsf{R}^{\ell}_{\mathcal{D}}(h_{\mathbb{S}}),\mathsf{R}^{\ell}_{\mathcal{S}}(h_{\mathbb{S}})) \le \Phi_{\mathtt{a}}(\delta)\right] \ge 1-\delta, \tag{1.2}$$

where usually $\phi(\mathsf{R}^\ell_{\mathcal{D}}(h), \mathsf{R}^\ell_{\mathcal{S}}(h)) = \mathsf{R}^\ell_{\mathcal{D}}(h) - \mathsf{R}^\ell_{\mathcal{S}}(h)$ and $h_{\mathbb{S}}$ is the hypothesis learned from an algorithm with $\mathbb{S} \sim \mathcal{D}^m$.

Similarly to the uniform convergence bounds, these bounds can be reformulated through a similar set-theoretic lens stated in the following proposition.

Proposition 7.5.3 (Set-theoretic View of Algorithmic Dependent Bounds)**.** Let $\phi : [0,1]^2 \to \mathbb{R}$ be a generalization gap and assume that there exists a function $\Phi_{\mathtt{a}} : (0,1] \to \mathbb{R}$ fulfilling Definition 1.3.5. Under these conditions, with $\mathbb{Z}_{\mathtt{a}} = \left\{\mathbb{S} \in (\mathbb{X}\times\mathbb{Y})^m : \phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}), \mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}})) \leq \Phi_{\mathtt{a}}(\delta)\right\}$ and $\mathbb{P}_{\mathbb{S}\sim\mathcal{D}^m}[\mathbb{S} \in \mathbb{Z}_{\mathtt{a}}] \geq 1-\delta$, we have

$$\begin{aligned} \text{Equation (1.2)} &\iff \forall \mathbb{S} \in \mathbb{Z}_{\mathtt{a}}, \phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}), \mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}})) \leq \Phi_{\mathtt{a}}(\delta) \\ &\iff \sup_{\mathbb{S}\in\mathbb{Z}_{\mathtt{a}}} \phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}), \mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}})) \leq \Phi_{\mathtt{a}}(\delta). \end{aligned}$$

Proof. Deferred to Appendix G.6. ■

Since the upper bound $\Phi_{\mathtt{a}}(\delta)$ is at least $\sup_{\mathbb{S}\in\mathbb{Z}_{\mathtt{a}}} \phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}), \mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}}))$, this result has the potential to lead to tighter guarantees than the uniform convergence ones. For example, when $\mathbb{H} = \{h_{\mathbb{S}}\}_{\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m}$ is an algorithmic-dependent hypothesis set and $\mathbb{Z}_{\mathtt{a}} \subseteq \mathbb{Z}_{\mathtt{u}}$. The complexity measure $\Phi_{\mathtt{a}}(\delta)$ can potentially be smaller than $\Phi_{\mathtt{u}}(\delta)$ since we have the inequality

$$\begin{aligned} \sup_{\mathbb{S}\in\mathbb{Z}_{\mathtt{a}}} \phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}), \mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}})) &\leq \sup_{\mathbb{S}\in\mathbb{Z}_{\mathtt{u}}} \phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}), \mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}})) \\ &\leq \sup_{\mathbb{S}\in\mathbb{Z}_{\mathtt{u}}} \sup_{h\in\mathbb{H}} \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \\ &\leq \Phi_{\mathtt{u}}(\delta). \end{aligned}$$

Even though these type of bounds can be tighter, it is still not as permissive as our framework. Indeed, the upper bound $\Phi_{\mathtt{a}}(\delta)$ is a constant *w.r.t.* the hypothesis and the learning sample (like the uniform convergence bounds). Hence, since our bound can depend on the learning sample $\mathbb{S}$ and the hypthesis h, we retrieve the algorithmic-dependent bounds illustrating the generality of our framework (similarly to Corollary 7.5.1). The result is in the following corollary.

Corollary 7.5.2 (Algorithmic-dependent Bound from Theorem 7.3.1)**.** Let $\phi : [0,1]^2 \to \mathbb{R}$ be the generalization gap and assume that there exists a function $\Phi_{\mathtt{a}} : (0,1] \to \mathbb{R}$ fulfilling Definition 1.3.5 such that $\Phi_{\mathtt{a}}(\delta) \geq$

$\ln\left[\frac{4}{\delta^2}\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{h'\sim\pi}\exp\left(\phi(\mathsf{R}_{\mathcal{D}}(h'),\mathsf{R}_{\mathcal{S}'}(h'))\right)\right]$. For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\rho_{\mathcal{S}}}\left[\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h))\leq\Phi_{\mathtt{a}}(\delta)\right]\geq 1-\delta.$$

Proof. Deferred to Appendix G.7. ■

Compared to the bounds of Definition 1.3.5, Corollary 7.5.2 still involves the expectation over the hypotheses. Hopefully, the bound holds with high probability for the data-dependent hypothesis $h_{\mathcal{S}}$ (see Proposition 7.5.3). Hence, when using Corollary 7.5.2's bound the sampling is not necessary since we can consider the bound only for the hypothesis of interest, *i.e.*, $h_{\mathcal{S}}$ for all $\mathcal{S}\in\mathbb{Z}_{\mathtt{a}}$ which holds with high probability. In other words, when $\Phi_\mu(h,\mathcal{S},\delta)=\Phi_{\mathtt{a}}(\delta)$, the set $\mathbb{Z}_{\mathtt{a}}$ in Proposition 7.5.3 can be seen as a subset of $\mathbb{Z}_{\mathtt{d}}$ since

$$\left\{\mathcal{S}\in(\mathbb{X}\times\mathbb{Y})^m\;\middle|\;(\mathcal{S},h_{\mathcal{S}})\in\mathbb{Z}_{\mathtt{d}}\right\}=\mathbb{Z}_{\mathtt{a}}.$$

In summary, the framework proposed in this chapter is powerful enough to cover uniform-convergence-based bound and algorithm-dependent bounds with the integration of a complexity measure. To the best of our knowledge, this has not been identified before and we think this is something novel.

7.6 Conclusion and Summary

In this chapter, we provide a novel generalization bound that involves arbitrary complexity measures, unlike classical learning theory frameworks (for which the complexity is imposed by the framework itself). These measures incorporate a data and model dependent function, that allow us to generalize the previous framework introduced in the literature (see Section 1.3). Importantly, to the best of our knowledge, our framework provides for the first time theoretical guarantees for the many arbitrary complexity measures used in practice in machine learning, *e.g.*, for regularization purposes.

The limitation of this work is clearly that the hypothesis is obtained from a distribution difficult to use, namely the Gibbs distribution. Indeed, one sampling from the Gibbs distribution is performed by an algorithm such as Algorithm 7.1. Hopefully, the generality of this framework allows to avoid the sampling if we consider uniform-convergence-type bounds as in Corollary 7.5.1. We can easily imagine continuing in

this research direction. For instance, we can try to get rid of the supremum *w.r.t.* to the hypothesis set $\mathbb{H}$ in the Rademacher complexity (Definition 1.3.4) since it it hard to compute.
We hope that our results foster research in the topic and the development of new complexity measures for specific neural network architectures and for specific learning tasks. Indeed, we believe that this work paves the way to new research directions that tries to bridge the statistical learning theory and the practice. Indeed, finding a good complexity measures becomes a practical matter since any complexity measure can be integrated in our framework.

In general, this thesis explores the disintegrated bounds in order to explain better the generalization of over-parameterized models that is largely misunderstood. We believe that this type of bounds is not the only promising type of bounds that can explain the generalization phenomenon. In Part IV, we give an idea of research direction to explore other type of generalization bounds.

Part IV

Conclusion and Perspectives

Conclusion and Perspectives

Conclusion

This thesis mainly derives self-bounding algorithms that learn a model minimizing a (disintegrated) PAC-Bayesian generalization bound. This type of algorithm has received little attention in the machine learning literature and we propose some contributions in various contexts.

Indeed, Part II is dedicated to deriving self-bounding algorithms in the context of majority vote classifiers. In Chapters 3 and 4, we derived self-bounding algorithms for the majority vote classifier in two different settings: the adversarial robustness and the classical supervised setting. More precisely, Chapter 3's self-bounding algorithms robustify the majority votes against small perturbations. While Chapter 4 minimizes the majority vote's true risk through the PAC-Bayesian C-Bounds considered as challenging to optimize (Lorenzen *et al.*, 2019; Masegosa *et al.*, 2020). However, as shown in Chapter 5, the majority vote's self-bounding algorithms considered, *e.g.*, in Chapter 4, do not minimize tight generalization bounds on the true risk, even for simple tasks. Hence, to overcome this drawback, Chapter 5 introduces the stochastic majority vote, which samples a majority vote for each prediction. Considering such a majority vote allows us to obtain tight generalization bounds. Additionally, we derive a self-bounding algorithm that directly minimizes the risk of the stochastic majority vote in this context. However, the risk of a stochastic model is the expected risk of the hypotheses, which requires certain assumptions to be computed while we may be only interested in assessing the behavior of only one hypothesis in some situations. Hence, to overcome this drawback, we consider in Part III the *disintegrated* PAC-Bayesian bounds. Chapter 6 provides new bounds based on the Rényi divergence that are more easily optimizable (for self-bounding algorithms) than the ones of the literature (*i.e.*, Blanchard and Fleuret, 2007; Catoni, 2007; Rivasplata *et al.*, 2020). Even though Rivasplata *et al.* (2020)'s bound is not easily optimizable, it is a starting point to derive new generalizations bounds. Indeed, in the last contribution (Chapter 7), we leverage Rivasplata *et al.* (2020)'s disintegrated framework to derive generalization bounds with arbitrary complexity measures. Such work is fundamental in statistical learning theory: to the best of our knowledge, we are the first to provide generalization bounds that integrate complexity measures that can be defined by the user. This work allows the machine learning community to consider new generalization bounds by defining a new complexity measure. Hence, new works can focus on developing new complexity measures to understand better the generalization phenomenon.

Perspectives

We present several perspectives following the contributions of this thesis.

Perspectives on the Adversarial Robustness Setting

As recalled in Chapter 3, in the adversarial robustness setting, we aim to make the model robust to small perturbations in the input. Indeed, we must ensure that the model does not radically change its prediction for a slight change in the input. To do so, we consider that the model's output must not change in a ball of a given radius. This new constraint on the input actually creates a new unknown data distribution that is close, in some sense, to the original unknown data distribution.

On the other hand, the transfer learning/domain adaptation[3] consider two unknown data distributions: a source (*i.e.*, the original) and a target (*i.e.*, a new) distribution. In this setting, the model learned to solve a task (represented by the source distribution) is adapted to solve a new task (represented by the target distribution). In some transfer learning scenarios, we assume that we have access to the labels and the inputs obtained from the target distribution, while in unsupervised domain adaptation, only the inputs are considered. In these two settings, the true risk on the target distribution can be upper-bounded with a generalization bound (see *e.g.*, Ben-David *et al.*, 2010; Galanti *et al.*, 2016; McNamara and Balcan, 2017; Germain *et al.*, 2020).

Besides, it is known that domain adaptation and adversarial robustness are related: unlabeled examples (considered in domain adaptation) can be used to improve the adversarial robustness (Alayrac *et al.*, 2019; Carmon *et al.*, 2019; Deng *et al.*, 2021). As a perspective, we propose to investigate the link between these two settings from a theoretical viewpoint. First, we could explore the connection between the original distribution and the new data distribution induced by the adversarial robustness that can be respectively seen as a source and a target distribution in transfer learning. Then, this connection may help to leverage transfer learning/domain adaptation generalization bounds to obtain guarantees for the adversarial robustness setting. The new guarantees might serve to get self-bounding algorithms that *(i)* detect out-of-distribution examples[4] and *(ii)* robustify machine learning models.

Extending the Majority Vote

In Part II, we consider that the set of voters in the PAC-Bayesian majority vote is fixed. Hence, only the weights of the majority vote are adapted to fit the examples.

[3] We refer the reader to Redko *et al.* (2019, 2020) for an introduction on domain adaption.

[4] The examples that are not probable in a given distribution are called out-of-distribution examples.

Alternatively, in the (Gradient) Boosting framework (Freund and Schapire, 1996; Friedman, 2001), the voters are greedily learned one by one. Moreover, in bagging (Breiman, 1996) and random forest (Breiman, 2001), no weights are learned while the models are learned separately. For the Support Vector Machine (Graepel *et al.*, 2005) that can be interpreted as a majority vote, the voters are fixed before learning the weights by choosing a kernel. As we can remark in these approaches, the voters and the weights are not learned together. This appears as a limitation since learning the weights and the voters in an end-to-end way can offer a better accurate majority vote. Hence, one bottleneck has to be overcome: deriving differentiable voters such as differentiable decision stumps. By doing so, we may improve the voters' diversity while limiting the voters' complexity. Moreover, the disintegrated PAC-Bayesian framework (developed, *e.g.*, in Part III) may be leveraged to derive generalization guarantees for majority votes that depend on the full learning sample. Again, new generalization bounds can be further used to derive self-bounding algorithms.

Self-bounding and Optimization Algorithms

The optimization algorithms are key to obtain a good classifier in self-bounding algorithms. Specifically in Chapters 4 and 5, we use an optimization algorithm that tune automatically the learning rate, namely COCOB (Orabona and Tommasi, 2017). This approach, belonging to the parameter-free algorithms[5], is interesting in machine because it has a clear advantage: there is no need to tune the learning rate. Hence, the parameter-free algorithms could facilitate the use of machine learning approaches for practitioners. However, we believe that more hyper-parameters can be tuned automatically in parameter-free optimization algorithms such as the batch size, which offers interesting research perspectives.

One idea to derive new parameter-free algorithms is to take inspiration from the federated learning setting.[6] It considers different clients that learn collaboratively in a machine learning model; each client has its own learning sample and does not necessarily share it. For instance, to learn the model, each client has its own local model and runs an optimization algorithm, to obtain new weights. The new weights of each local model are aggregated to obtain a global model finally, without exchanging the data; see, *e.g.*, the FedAvg algorithm (see McMahan *et al.*, 2017). A modification must be made to FedAvg to obtain a parameter-free algorithm since each client runs an algorithm with different values of hyper-parameters. The aggregation of the weights can take different forms, such as a convex combination. With this latter type of aggregation, the PAC-Bayesian theory might be helpful to obtain convergence guarantees.

[5]We refer the reader to the ICML 2020 tutorial on Parameter-free online optimization for more details on the parameter-free algorithms.

[6]Federated learning is a sub-field of machine learning; see Kairouz *et al.* (2021) for a survey.

Towards a New Type of Generalization Bounds

The PAC-Bayesian theory considers that the data and the models are respectively sampled from two probability distributions: the unknown distribution and the posterior distribution. While it can be convenient to derive generalization guarantees on a single model sampled from the posterior distribution, we are usually interested in a model that is not necessarily sampled.
Instead of considering the posterior distribution on the models, one could consider a distribution on the label set conditioned on the input. If this distribution is somehow learned from the learning sample, it can be seen as a machine learning model. Thanks to this distribution, we could derive generalization bounds on the expected loss when the labels are sampled from the new data-dependent distribution (associated with a classifier). We can for example hope to obtain a bound dependent on the mutual information between the predictions and the labels. Roughly speaking, mutual information measures how much information on the labels is contained in the predictions. Hence, it can be seen as a complexity measure of the data-dependent distribution (representing the classifier). For instance, this quantity has been considered in the information bottleneck framework of Tishby *et al.* (2000). Besides, the training of neural networks has been studied through this framework (see Shwartz-Ziv and Tishby, 2017).

Part V

Appendix

Some Mathematical Tools

A

A.1 Jensen's Inequality

Theorem A.1.1 (Jensen's Inequality)**.** Let $X \in \mathbb{X}$ a random variable following a probability distribution $\mathcal{X}$ with $f : \mathbb{X} \to \mathbb{R}$ a measurable convex function, we have

$$f\left(\mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right) \leq \mathop{\mathbb{E}}_{X\sim\mathcal{X}}\Big[f\left(X\right)\Big].$$

Proof. Since $f()$ is a convex function, the following inequality holds, *i.e.*, we have

$$\forall X' \in \mathbb{X}, \quad a\left(X' - \mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right) \leq f(X') - f\left(\mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right),$$

where a is the tangent's slope. By taking the expectation to both sides of the inequality, we have

$$\underbrace{a\left(\mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X] - \mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right)}_{=0} \leq \mathop{\mathbb{E}}_{X\sim\mathcal{X}}[f(X)] - f\left(\mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right).$$

Hence, by rearranging the terms, we prove the claimed result. ■

A.2 Markov's Inequality

Theorem A.2.1 (Markov's Inequality)**.** Let $X \in \mathbb{X}$ a non-negative random variable following a probability distribution $\mathcal{X}$ and $\tau > 0$, we have

$$\mathop{\mathbb{P}}_{X\sim\mathcal{X}}[X \geq \tau] \leq \frac{\mathbb{E}_{X\sim\mathcal{X}}[X]}{\tau}.$$

Proof. First of all, remark that we have the following inequality for any $X \in \mathbb{X}$

$$\tau \, \mathrm{I}[X \geq \tau] \;\leq\; X \, \mathrm{I}[X \geq \tau] \;\leq\; X. \tag{A.1}$$

Indeed, on the one hand, if $X < \tau$, $\mathrm{I}[X \geq \tau] = 0$, the inequality holds trivially. On the other hand, if $X \geq \tau$, $\mathrm{I}[X \geq \tau] = 1$ and the inequality becomes $\tau \leq X$, which is true. By taking the expectation of Equation (A.1), we have

$$\mathop{\mathbb{E}}_{X \sim \mathcal{X}} \Big[\tau \, \mathrm{I}[X \geq \tau] \Big] \leq \mathop{\mathbb{E}}_{X \sim \mathcal{X}} \Big[X \Big].$$

From the fact that the expectation of a constant is the constant and by definition of the probability, we have

$$\tau \mathop{\mathbb{P}}_{X \sim \mathcal{X}} [X \geq \tau] \leq \mathop{\mathbb{E}}_{X \sim \mathcal{X}} \Big[X \Big] \quad \Longleftrightarrow \quad \mathop{\mathbb{P}}_{X \sim \mathcal{X}} [X \geq \tau] \leq \frac{\mathbb{E}_{X \sim \mathcal{X}} [X]}{\tau},$$

which is the desired result. ■

A.3 2nd Order Markov's Inequality

Theorem A.3.1 (2nd Order MARKOV's Inequality)**.** Let X a non-negative random variable following a probability distribution $\mathcal{X}$ and $\tau > 0$, we have

$$\mathop{\mathbb{P}}_{X \sim \mathcal{X}} [X \geq \tau] \leq \frac{\mathbb{E}_{X \sim \mathcal{X}} [X^2]}{\tau^2}.$$

Proof. We apply MARKOV's inequality (Theorem A.2.1) to have

$$\mathop{\mathbb{P}}_{X \sim \mathcal{X}} \Big[X^2 \geq \tau^2 \Big] \leq \frac{\mathbb{E}_{X \sim \mathcal{X}} [X^2]}{\tau^2}.$$

Moreover, since $\mathrm{I}[X \geq \tau] = \mathrm{I}[X^2 \geq \tau^2]$, we have

$$\mathop{\mathbb{P}}_{X \sim \mathcal{X}} [X \geq \tau] = \mathop{\mathbb{P}}_{X \sim \mathcal{X}} \Big[X^2 \geq \tau^2 \Big],$$

which proves the desired result. ■

A.4 Chebyshev-Cantelli Inequality

Theorem A.4.1 (Chebyshev-Cantelli Inequality)**.** Let X a random variable following a probability distribution $\mathcal{X}$ and $\tau > 0$, we have

$$\mathop{\mathbb{P}}_{X\sim\mathcal{X}}\left[X - \mathop{\mathbb{E}}_{X'\sim\mathcal{X}} X' \geq \tau\right] \leq \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\mathbb{V}_{X'\sim\mathcal{X}}\, X' + \tau^2}.$$

Proof. First of all, remark that we have

$$\begin{aligned}\mathop{\mathbb{P}}_{X\sim\mathcal{X}}\left[X - \mathop{\mathbb{E}}_{X'\sim\mathcal{X}} X' \geq \tau\right] &= \mathop{\mathbb{P}}_{X\sim\mathcal{X}}\left[X - \mathop{\mathbb{E}}_{X'\sim\mathcal{X}} X' + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau} \geq \tau + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau}\right] \\ &\leq \mathop{\mathbb{P}}_{X\sim\mathcal{X}}\left[\left[X - \mathop{\mathbb{E}}_{X'\sim\mathcal{X}} X' + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau}\right]^2 \geq \left[\tau + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau}\right]^2\right], \end{aligned} \tag{A.2}$$

where $\mathbb{V}_{X\sim\mathcal{X}}\, X$ is the variance of the random variable $X \sim \mathcal{X}$. From Equation (A.2) and Markov's Inequality (Theorem A.2.1), we can deduce that

$$\begin{aligned}\mathop{\mathbb{P}}_{X\sim\mathcal{X}}\left[X - \mathop{\mathbb{E}}_{X'\sim\mathcal{X}} X' \geq \tau\right] &\leq \frac{\mathbb{E}_{X\sim\mathcal{X}}\left[X - \mathbb{E}_{X'\sim\mathcal{X}}\, X' + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau}\right]^2}{\left[\tau + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau}\right]^2} \\ &= \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\mathbb{V}_{X'\sim\mathcal{X}}\, X' + \tau^2}.\end{aligned}$$

■

A.5 Hölder's Inequality

In order to prove Hölder's inequality, we first prove the following lemma.

Lemma A.5.1 (Young's Inequality)**.** For any $\alpha > 1$ and $\beta > 1$ such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, for any $a \geq 0$ and $b \geq 0$, we have

$$ab \leq \frac{a^\alpha}{\alpha} + \frac{b^\beta}{\beta}.$$

Proof. We first develop $\ln [ab]$ and we apply JENSEN's inequality (Theorem A.1.1) since the logarithm is concave and $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Indeed, we have

$$\ln [ab] = \ln a + \ln b = \frac{\alpha}{\alpha} \ln a + \frac{\beta}{\beta} \ln b = \frac{\ln a^\alpha}{\alpha} + \frac{\ln b^\beta}{\beta} \leq \ln \left[\frac{a^\alpha}{\alpha} + \frac{b^\beta}{\beta} \right].$$

Then, we take the exponential to both sides of the inequality and we are done. ■

We are now ready to prove HÖLDER's inequality.

Theorem A.5.1 (HÖLDER's Inequality)**.** For any measurable function $f()$ and $g()$, for any $\alpha > 1$ and $\beta > 1$ such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, we have

$$\mathop{\mathbb{E}}_{X \sim \mathcal{X}} |f(X)g(X)| \leq \left[\mathop{\mathbb{E}}_{X \sim \mathcal{X}} |f(X)|^\alpha \right]^{\frac{1}{\alpha}} \left[\mathop{\mathbb{E}}_{X \sim \mathcal{X}} |g(X)|^\beta \right]^{\frac{1}{\beta}}.$$

Proof. For convenience of notation, let $\|f\|_\alpha = [\mathbb{E}_{X \sim \mathcal{X}} |f(X)|^\alpha]^{\frac{1}{\alpha}}$ and $\|g\|_\beta = \left[\mathbb{E}_{X \sim \mathcal{X}} |g(X)|^\beta\right]^{\frac{1}{\beta}}$. If $\|f\|_\alpha = 0$ or $\|g\|_\beta = 0$, then $\mathbb{E}_{X \sim \mathcal{X}} |f(X)g(X)| = 0$, hence, the inequality holds in this case. Then for $\|f\|_\alpha > 0$ and $\|g\|_\beta > 0$, we upper-bound the term $\frac{|f(X)g(X)|}{\|f\|_\alpha \|g\|_\beta}$ with YOUNG's inequality (Lemma A.5.1), *i.e.*, we have

$$\frac{|f(X)g(X)|}{\|f\|_\alpha \|g\|_\beta} \leq \frac{|f(X)|^\alpha}{\alpha \|f\|_\alpha^\alpha} + \frac{|f(X)|^\beta}{\beta \|f\|_\beta^\beta}.$$

By taking the expectation *w.r.t.* $X \sim \mathcal{X}$, we have

$$\begin{aligned}
\frac{\mathbb{E}_{X \sim \mathcal{X}} |f(X)g(X)|}{\|f\|_\alpha \|g\|_\beta} &\leq \frac{\mathbb{E}_{X \sim \mathcal{X}} |f(X)|^\alpha}{\alpha \|f\|_\alpha^\alpha} + \frac{\mathbb{E}_{X \sim \mathcal{X}} |f(X)|^\beta}{\beta \|f\|_\beta^\beta} \\
&= \frac{\|f\|_\alpha^\alpha}{\alpha \|f\|_\alpha^\alpha} + \frac{\|f\|_\beta^\beta}{\beta \|f\|_\beta^\beta} \\
&= \frac{1}{\alpha} + \frac{1}{\beta} \\
&= 1.
\end{aligned}$$

This concludes the proof since

$$\frac{\mathbb{E}_{X\sim\mathcal{X}}\,|f(X)g(X)|}{\|f\|_\alpha\|g\|_\beta} \leq 1 \iff \mathop{\mathbb{E}}_{X\sim\mathcal{X}}|f(X)g(X)| \leq \|f\|_\alpha\|g\|_\beta.$$

■

Appendix of Chapter 2

B

B.1 Proof of Theorem 2.2.1

Theorem 2.2.1 (Risk Upper Bound Based on the Gibbs Risk)**.** For any distribution $\mathcal{D}'$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, we have

$$\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho) \leq 2\, r_{\mathcal{D}'}(\rho). \tag{2.2}$$

Proof. The proof is given by Germain *et al.* (2015). First of all, remark that

$$\mathrm{I}\left[\mathrm{MV}_\rho(\mathbf{x}) \neq y\right] \leq \mathrm{I}\left[\widehat{m_\rho}(\mathbf{x}, y) \leq 0\right].$$

Hence, by taking the expectation, we have

$$\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho) \leq \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\widehat{m_\rho}(\mathbf{x}, y) \leq 0\right].$$

From Markov's inequality (Theorem A.2.1), we have

$$\begin{aligned}
\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\widehat{m_\rho}(\mathbf{x}, y) \leq 0\right] &= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[1 - \widehat{m_\rho}(\mathbf{x}, y) \geq 1\right] \\
&= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[1 - 2\left[\mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y\right] - \frac{1}{2}\right] \geq 1\right] \\
&= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[1 - \left[\mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y\right]\right] \geq \frac{1}{2}\right] \\
&\leq 2 \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[1 - \mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y\right]\right] \\
&= 2 r_{\mathcal{D}'}(\rho).
\end{aligned}$$

■

B.2 Proof of Theorem 2.2.2

Theorem 2.2.2 (Risk Upper Bound Based on the Joint Error)**.** For any distribution $\mathcal{D}'$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, we have

$$\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho) \leq 4e_{\mathcal{D}'}(\rho). \tag{2.4}$$

Proof. The proof is given by Masegosa *et al.* (2020). First of all, remark that

$$\mathrm{I}\left[\mathrm{MV}_\rho(\mathbf{x}) \neq y\right] \leq \mathrm{I}\left[\widehat{m_\rho}(\mathbf{x}, y) \leq 0\right].$$

Hence, by taking the expectation, we have

$$\mathrm{R}_{\mathcal{D}'}(\mathrm{MV}_\rho) \leq \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\widehat{m_\rho}(\mathbf{x}, y) \leq 0\right].$$

From Markov's inequality (Theorem A.3.1), we have

$$\begin{aligned}
\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\widehat{m_\rho}(\mathbf{x}, y) \leq 0\right] &= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[1 - \widehat{m_\rho}(\mathbf{x}, y) \geq 0\right] \\
&= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[1 - 2\left[\mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y\right] - \frac{1}{2}\right] \geq 1\right] \\
&= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[1 - \left[\mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y\right]\right] \geq \frac{1}{2}\right] \\
&\leq 4 \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(1 - \mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y\right]\right)^2 \\
&= 4e_{\mathcal{D}'}(\rho).
\end{aligned}$$

■

B.3 Proof of Theorem 2.2.3

Theorem 2.2.3 (The C-Bound)**.** For any distribution $\mathcal{D}'$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, if

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) > 0 \iff r_{\mathcal{D}'}(\rho) < \tfrac{1}{2} \iff 2e_{\mathcal{D}'}(\rho) + d_{\mathcal{D}'}(\rho) < 1,$$

we have

$$\begin{aligned}
\mathsf{R}_{\mathcal{D}'}(\mathsf{MV}_\rho) &\leq 1 - \frac{\left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\widehat{m_\rho}(\mathbf{x},y)\right]\right)^2}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right)^2} && (2.7)\\
&= 1 - \frac{\left(1-2r_{\mathcal{D}'}(\rho)\right)^2}{1-2d_{\mathcal{D}'}(\rho)} && (2.8)\\
&= 1 - \frac{\left(1-\left[2e_{\mathcal{D}'}(\rho)+d_{\mathcal{D}'}(\rho)\right]\right)^2}{1-2d_{\mathcal{D}'}(\rho)} && (2.9)\\
&= C_{\mathcal{D}'}(\rho).
\end{aligned}$$

Proof. To prove Equation (2.7), we start from the definition of $\mathsf{R}_{\mathcal{D}'}(\mathsf{MV}_\rho)$ to have

$$\begin{aligned}
\mathsf{R}_{\mathcal{D}'}(\mathsf{MV}_\rho) &\leq \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\leq 0\right)\\
&= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(-\widehat{m_\rho}(\mathbf{x},y) + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y) \geq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\right)\\
&\leq \frac{\mathbb{V}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right)}{\mathbb{V}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right) + \left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\right)^2}\\
&= \frac{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)^2 - \left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\right)^2}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)^2 - \left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\right)^2 + \left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\right)^2}\\
&= \frac{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)^2 - \left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\right)^2}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)^2}\\
&= 1 - \frac{\left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\right)^2}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)^2},
\end{aligned}$$

where the second inequality comes from Chebyshev-Cantelli's inequality (Theorem A.4.1) and $\mathbb{V}_{A\sim\mathcal{A}}(A)$ is the variance of the random variable $A\sim\mathcal{A}$. Equation (2.9) is obtained by rewriting Equation (2.7) with Equations (2.1) and (2.5). Indeed, we have

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y) = 1-2r_{\mathcal{D}'}(\rho) \quad \text{and} \quad \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)^2 = 1-2d_{\mathcal{D}'}(\rho),$$

which gives Equation (2.8).

Moreover, thanks to Equation (2.6), we can rewrite the Gibbs risk as

$$r_{\mathcal{D}'}(\rho) = \left[e_{\mathcal{D}'}(\rho) + \frac{1}{2} d_{\mathcal{D}'}(\rho) \right],$$

which allows us to obtain Equation (2.9) by rewriting Equation (2.8). ■

B.4 Proof of Theorem 2.2.4

Theorem 2.2.4 (Relationship between Theorems 2.2.1 to 2.2.3)**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any voters set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, if $r_{\mathcal{D}'}(\rho) < \frac{1}{2}$ (*i.e.*, $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) > 0$), we have

(i) $\mathsf{R}_{\mathcal{D}'}(\mathsf{MV}_\rho) \leq C_{\mathcal{D}'}(\rho) \leq 4e_{\mathcal{D}'}(\rho) \leq 2r_{\mathcal{D}'}(\rho)$, if $r_{\mathcal{D}'}(\rho) \leq d_{\mathcal{D}'}(\rho)$,

(ii) $\mathsf{R}_{\mathcal{D}'}(\mathsf{MV}_\rho) \leq 2r_{\mathcal{D}'}(\rho) \leq C_{\mathcal{D}'}(\rho) \leq 4e_{\mathcal{D}'}(\rho)$, otherwise.

Proof. We first prove that $C_{\mathcal{D}'}(\rho) \leq 2r_{\mathcal{D}'}(\rho)$ is equivalent to $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) \geq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'} [\widehat{m_\rho}(\mathbf{x}, y)]^2$, *i.e.*, we have

$$\begin{aligned}
& C_{\mathcal{D}'}(\rho) \leq 2r_{\mathcal{D}'}(\rho) \\
\iff & 1 - \frac{\left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y)\right)^2}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'} \left(\widehat{m_\rho}(\mathbf{x}, y)\right)^2} \leq 1 - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) \\
\iff & \frac{\left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y)\right)^2}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'} \left(\widehat{m_\rho}(\mathbf{x}, y)\right)^2} \geq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) \\
\iff & \left(\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) \right)^2 \geq \left(\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) \right) \left(\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \left(\widehat{m_\rho}(\mathbf{x}, y)\right)^2 \right) \\
\iff & \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) \geq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} [\widehat{m_\rho}(\mathbf{x}, y)]^2 .
\end{aligned}$$

Then, we prove that $4e_{\mathcal{D}'}(\rho) \leq 2r_{\mathcal{D}'}(\rho)$ is equivalent to $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'} \widehat{m_\rho}(\mathbf{x}, y) \geq$

$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\widehat{m_\rho}(\mathbf{x},y)\right]^2$, *i.e.*, we have

$$\begin{aligned}
&4e_{\mathcal{D}'}(\rho) \leq 2r_{\mathcal{D}'}(\rho)\\
\iff &1 - 2\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y) + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right)^2 \leq 1 - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\\
\iff &2\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y) - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right)^2 \geq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\\
\iff &\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y) \geq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\widehat{m_\rho}(\mathbf{x},y)\right]^2.
\end{aligned}$$

Additionally, we prove that $C_{\mathcal{D}'}(\rho) \leq 4e_{\mathcal{D}'}(\rho)$, *i.e.*, we have

$$\begin{aligned}
&C_{\mathcal{D}'}(\rho) \leq 4e_{\mathcal{D}'}(\rho)\\
\iff &1 - \frac{\left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\right)^2}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right)^2} \leq 1 - 2\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y) + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right)^2\\
\iff &\frac{\left(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y)\right)^2}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right)^2} \geq 2\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y) - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right)^2\\
\iff &\left(\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\widehat{m_\rho}(\mathbf{x},y) - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right)^2\right)^2 \geq 0.
\end{aligned}$$

Finally, by merging the three equivalence, we obtain the claimed result. ■

B.5 About the KL Divergence

B.5.1 Basic Properties

The KL divergence has the following properties:

(1) It is positive, *i.e.*, we have $\mathrm{KL}(\rho\|\pi) \geq 0$ for all $\rho \in \mathbb{M}(\mathbb{H})$ and $\pi \in \mathbb{M}^*(\mathbb{H})$.

(2) We have $\mathrm{KL}(\pi\|\pi) = 0$ for all $\pi \in \mathbb{M}(\mathbb{H})$.

(3) In general, it is not symmetric: $\mathrm{KL}(\rho\|\pi) \neq \mathrm{KL}(\pi\|\rho)$ for all $\rho, \pi \in \mathbb{M}(\mathbb{H})$.

Proof. We prove the points *(1)*, *(2)* and *(3)* separately.
Concerning *(1)*. Since $-\ln()$ is convex, we have from JENSEN's inequality (Theorem A.1.1)

$$\mathrm{KL}(\rho\|\pi) = \mathop{\mathbb{E}}_{h\sim\rho}\ln\frac{\rho(h)}{\pi(h)} = \mathop{\mathbb{E}}_{h\sim\rho}\left[-\ln\frac{\pi(h)}{\rho(h)}\right] \geq -\ln\left[\mathop{\mathbb{E}}_{h\sim\rho}\frac{\pi(h)}{\rho(h)}\right] = 0.$$

Concerning *(2)*. The property follows directly by developing the KL divergence, *i.e.*, we have

$$\mathrm{KL}(\pi\|\pi) = \mathop{\mathbb{E}}_{h\sim\pi} \ln \frac{\pi(h)}{\pi(h)} = \mathop{\mathbb{E}}_{h\sim\pi} \ln(1) = 0.$$

Concerning *(3)*. For example, we have $\mathrm{KL}(\mathcal{B}(0.1)\|\mathcal{B}(0.5)) \neq \mathrm{KL}(\mathcal{B}(0.5)\|\mathcal{B}(0.1))$, where $\mathcal{B}(p)$ is a Bernoulli distribution with bias p. ■

Moreover, since we have (from l'Hôpital's rule) $\lim_{x\to 0^+} x \ln x = 0$, we adopt several conventions. Indeed, we consider that *(i)* $\rho(h) \ln \frac{\rho(h)}{\pi(h)} = 0$ whenever $\rho(h) = 0$ and $\pi(h) \geq 0$ and *(ii)* if $\pi(h) = 0$ and $\rho(h) > 0$, $\rho(h) \ln \frac{\rho(h)}{\pi(h)} = +\infty$ (which implies that $\mathrm{KL}(\rho\|\pi) = +\infty$).

B.5.2 Joint Convexity

The following proposition shows that the KL divergence is jointly convex.

Proposition B.5.1. For any pairs $(\rho_1, \rho_2) \in \mathbb{M}(\mathbb{H})^2$ and $(\pi_1, \pi_2) \in \mathbb{M}^*(\mathbb{H})^2$ of distributions, we have for all $\lambda \in [0, 1]$

$$\mathrm{KL}(\lambda\rho_1 + (1-\lambda)\rho_2\|\lambda\pi_1 + (1-\lambda)\pi_2) \leq \lambda\,\mathrm{KL}(\rho_1\|\pi_1) + (1-\lambda)\,\mathrm{KL}(\rho_2\|\pi_2).$$

In order to provide a proof for Proposition B.5.1. We need the log-sum inequality; a proof (based on Cover and Thomas (2006, Theorem 2.7.1)) is given bellow.

Lemma B.5.1 (Log-sum inequality)**.** For any strictly positive reals $p_1, \ldots, p_n \geq 0$ and $q_1, \ldots, q_n \geq 0$, we have

$$\left[\sum_{i=1}^n q_i\right] \ln \frac{\sum_{i=1}^n q_i}{\sum_{i=1}^n p_i} \leq \sum_{i=1}^n q_i \ln \frac{q_i}{p_i}.$$

Proof. First of all, note that $f(x) = x \ln x$ is convex *w.r.t.* $x \in \mathbb{R}_*^+$ since $\frac{\partial f}{\partial x}(x) = \frac{1}{x}$

for all $x \in \mathbb{R}_*^+$. Then, from JENSEN's inequality (Theorem A.1.1) we have

$$f\left(\sum_{i=1}^n \frac{p_i}{\|\mathbf{p}\|_1}\frac{q_i}{p_i}\right) \leq \sum_{i=1}^n \frac{p_i}{\|\mathbf{p}\|_1} f\left(\frac{q_i}{p_i}\right)$$
$$\iff \quad \|\mathbf{p}\|_1 f\left(\sum_{i=1}^n \frac{p_i}{\|\mathbf{p}\|_1}\frac{q_i}{p_i}\right) \leq \|\mathbf{p}\|_1 \sum_{i=1}^n \frac{p_i}{\|\mathbf{p}\|_1} f\left(\frac{q_i}{p_i}\right),$$

where $\|\mathbf{p}\|_1 = \sum_{i=1}^n p_i$. Then, we can develop right-hand side of the inequality, *i.e.*, we have

$$\|\mathbf{p}\|_1 f\left(\sum_{i=1}^n \frac{p_i}{\|\mathbf{p}\|_1}\frac{q_i}{p_i}\right) = \|\mathbf{p}\|_1 f\left(\frac{\|\mathbf{q}\|_1}{\|\mathbf{p}\|_1}\right) = \|\mathbf{q}\|_1 \ln \frac{\|\mathbf{q}\|_1}{\|\mathbf{p}\|_1} = \left[\sum_{i=1}^n q_i\right] \ln \frac{\sum_{i=1}^n q_i}{\sum_{i=1}^n p_i},$$

where $\|\mathbf{q}\|_1 = \sum_{i=1}^n q_i$. Similarly for the left-hand side, we have

$$\|\mathbf{p}\|_1 \sum_{i=1}^n \frac{p_i}{\|\mathbf{p}\|_1} f\left(\frac{q_i}{p_i}\right) = \sum_{i=1}^n p_i f\left(\frac{q_i}{p_i}\right) = \sum_{i=1}^n q_i \ln \frac{q_i}{p_i}.$$

■

We are now able to prove Proposition B.5.1 based on the proof of COVER and THOMAS (2006, Theorem 2.7.2).

Proof of Proposition B.5.1. From the log-sum inequality (Lemma B.5.1) with $q_1 = \lambda\rho_1(h)$, $q_2 = (1-\lambda)\rho_2(h)$, $p_1 = \lambda\pi_1(h)$, $p_2 = (1-\lambda)\pi_2(h)$, we have

$$[\lambda\rho_1(h) + (1-\lambda)\rho_2(h)] \ln \left[\frac{\lambda\rho_1(h) + (1-\lambda)\rho_2(h)}{\lambda\pi_1(h) + (1-\lambda)\pi_2(h)}\right]$$
$$\leq \lambda\rho_1(h) \ln \frac{\rho_1(h)}{\pi_1(h)} + (1-\lambda)\rho_2(h) \ln \frac{\rho_2(h)}{\pi_2(h)}.$$

Hence, by integrating over all h, gives us the desired result. ■

B.5.3 Pinsker's Inequality

We present a special case of PINSKER's inequality when we deal with two Bernoulli distributions. The presented proof is due to WU (2020) (see also CANONNE (2022)).

Theorem B.5.1 (Pinsker's inequality)**.** For any $p \in [0,1]$ and $q \in [0,1]$, we have

$$2(q-p)^2 \leq \mathrm{kl}(q\|p),$$

where $\mathrm{kl}(q\|p) \triangleq q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}$.

Proof. For any $p \in \{0,1\}$ and $q \in \{0,1\}$, we can easily verify that the inequality holds. Then, with $p \in (0,1)$ and $q \in (0,1)$, we have from the fundamental theorem of calculus

$$\mathrm{kl}(q\|p) = f(q) - f(p) = \int_q^p \frac{\partial f}{\partial x}(x)dx,$$

where $f(x) = q\ln x + (1-q)\ln(1-x)$. Hence, we have

$$\int_q^p \frac{\partial f}{\partial x}(x)dx = \int_q^p \frac{(q-x)}{(1-x)x}dx \leq 4\int_q^p (q-x)dx = 4\frac{1}{2}(q-p)^2 = 2(q-p)^2.$$

■

B.6 Proof of Theorem 2.3.1

Theorem 2.3.1 (General PAC-Bayesian Bound of Germain *et al.* (2009))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any measurable function $\varphi : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \in \mathbb{M}(\mathbb{H}),\ \mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S}) \leq \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi} e^{\varphi(h',\mathbb{S}')}\right)\right] \geq 1-\delta,$$

where $\mathrm{KL}(\rho\|\pi) = \mathop{\mathbb{E}}_{h\sim\rho}\ln\frac{\rho(h)}{\pi(h)}$ is the Kullback-Leibler (KL) divergence between the distributions ρ and π.

Proof. We start by developing $\mathbb{E}_{h\sim\rho}\varphi(h,\mathbb{S})$, *i.e.*, we obtain

$$\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S}) &= \mathop{\mathbb{E}}_{h\sim\rho}\ln\left[\exp(\varphi(h,\mathbb{S}))\right] \\
&= \mathop{\mathbb{E}}_{h\sim\rho}\ln\left[\frac{\rho(h)}{\pi(h)}\frac{\pi(h)}{\rho(h)}\exp(\varphi(h,\mathbb{S}))\right] \\
&= \mathop{\mathbb{E}}_{h\sim\rho}\ln\frac{\rho(h)}{\pi(h)} + \mathop{\mathbb{E}}_{h'\sim\rho}\ln\left[\frac{\pi(h')}{\rho(h')}\exp(\varphi(h',\mathbb{S}))\right] \\
&= \mathrm{KL}(\rho\|\pi) + \mathop{\mathbb{E}}_{h'\sim\rho}\ln\left[\frac{\pi(h')}{\rho(h')}\exp(\varphi(h',\mathbb{S}))\right].
\end{aligned}$$

Since $\ln$ is concave, we can apply JENSEN's inequality (Theorem A.1.1) on the right-most term to obtain

$$\begin{aligned}\mathop{\mathbb{E}}_{h\sim\rho}\ln\frac{\rho(h)}{\pi(h)}+\mathop{\mathbb{E}}_{h'\sim\rho}\ln\left[\frac{\pi(h')}{\rho(h')}\exp(\varphi(h',\mathbb{S}))\right]&\leq\mathop{\mathbb{E}}_{h'\sim\rho}\ln\left[\frac{\pi(h')}{\rho(h')}\exp(\varphi(h',\mathbb{S}))\right]\\&=\ln\left[\mathop{\mathbb{E}}_{h'\sim\pi}\exp(\varphi(h',\mathbb{S}))\right].\end{aligned}$$

Hence, we can deduce the following inequality:

$$\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})\leq\mathrm{KL}(\rho\|\pi)+\ln\left[\mathop{\mathbb{E}}_{h'\sim\pi}\exp(\varphi(h',\mathbb{S}))\right]. \tag{B.1}$$

Since the exponential function $\exp(a)$ is positive for all $a\in\mathbb{R}$, thus, the term $\mathbb{E}_{h'\sim\pi}\exp(\varphi(h',\mathbb{S}))$ is positive for all $\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m$ as well. We can apply MARKOV's inequality (Theorem A.2.1) to have

$$\begin{aligned}&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\exp(\varphi(h',\mathbb{S}))\leq\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\exp(\varphi(h',\mathbb{S}'))\right]\geq1-\delta\\
\iff&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\exp(\varphi(h',\mathbb{S}))\right)\leq\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\exp(\varphi(h',\mathbb{S}'))\right)\right]\geq1-\delta\\
\iff&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),\ \mathrm{KL}(\rho\|\pi)+\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\exp(\varphi(h',\mathbb{S}))\right)\right.\\
&\qquad\left.\leq\mathrm{KL}(\rho\|\pi)+\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\exp(\varphi(h',\mathbb{S}'))\right)\right]\geq1-\delta\end{aligned} \tag{B.2}$$

By combining Equations (B.1) and (B.2), we can deduce the claimed result. ■

B.7 Proof of Theorem 2.3.2

Theorem 2.3.2 (PAC-Bayesian Bound of MCALLESTER (2003))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior $\pi\in\mathbb{M}^*(\mathbb{H})$, for any loss $\ell:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m\to[0,1]$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),\left|\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)-\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathbb{S}}(h)\right|\leq\sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi)+\ln\tfrac{2\sqrt{m}}{\delta}\right]}\right]\geq1-\delta. \tag{2.10}$$

Proof. It is a direct consequence of PINSKER's inequality (Theorem B.5.1), *i.e.*, we have

$$2\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h) - \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{S}}(h)\right)^2 \leq \mathrm{kl}\Big(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h) \Big\| \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big)\,,$$

and from Theorem 2.3.4 by rearranging the terms. ■

B.8 Proof of Theorem 2.3.3

Theorem 2.3.3 (PAC-Bayesian Bound of CATONI (2007))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior $\pi\in\mathbb{M}^*(\mathbb{H})$, for any loss $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to [0,1]$, for any $c>0$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}), -\ln\Big(1-\Big[1-e^{-c}\Big]\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)\Big) - c\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}\right]\right] \geq 1-\delta. \quad (2.13)$$

Proof. We apply Theorem 2.3.1 with $\varphi(h,\mathcal{S}) = m\Big[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h)) - c\mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big]$, where $F(\mathsf{R}^{\ell}_{\mathcal{D}}(h)) \triangleq -\ln\Big(1-\mathsf{R}^{\ell}_{\mathcal{D}}(h)[1-e^{-c}]\Big)$. We have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),\ \mathop{\mathbb{E}}_{h\sim\rho}\Big[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h)) - c\mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big] \leq \frac{1}{m}\Big[\mathrm{KL}(\rho\|\pi)+\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi} e^{m\left[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h'))-c\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right]}\right)\Big]\right] \geq 1-\delta.$$

Since the distribution π on $\mathbb{H}$ does not depend on the $\mathcal{S}'\sim\mathcal{D}^m$, we can exchange the two expectations (with FUBINI's theorem), *i.e.*, we have

$$\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi} e^{m\left[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h'))-c\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right]} = \mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m} e^{m\left[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h'))-c\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right]}.$$

Then, from Lemma B.16.2, we have

$$\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} e^{m\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h))-c\mathsf{R}^\ell_{\mathbb{S}}(h)\right]} \leq 1 \Longrightarrow \ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} e^{m\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h'))-c\mathsf{R}^\ell_{\mathcal{S}'}(h')\right]}\right) \leq \ln\frac{1}{\delta}.$$

The function $F(x)$ is convex, since its second derivative is $\frac{\partial^2 F}{\partial x^2}(x)=\frac{(e^c-1)^2}{(x-e^c(x-1))^2}\geq 0$. In this case, easily conclude that $F(p)-cq$ is jointly convex in q and p. Hence, from Jensen's inequality (Theorem A.1.1) we have for all $\rho\in\mathbb{M}(\mathbb{H})$

$$F\left(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)\right) - c\left[\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathbb{S}}(h)\right] \leq \mathop{\mathbb{E}}_{h\sim\rho}\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h)) - c\mathsf{R}^\ell_{\mathbb{S}}(h)\right].$$

Hence, this gives the bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),\ F\left(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)\right) - c\left[\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathbb{S}}(h)\right] \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}\right]\right] \geq 1-\delta,$$

and by rearranging the terms we obtain the desired result. ■

B.9 Proof of Theorem 2.3.4

Theorem 2.3.4 (PAC-Bayesian Bound of Seeger (2002))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior $\pi\in\mathbb{M}^*(\mathbb{H})$, for any loss $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m\to[0,1]$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),\ \mathrm{kl}\left(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)\middle\|\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{S}}(h)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\tfrac{2\sqrt{m}}{\delta}\right]\right] \geq 1-\delta. \tag{2.14}$$

Proof. We can apply Theorem 2.3.1 with $\varphi(h,\mathbb{S}) = m\,\mathrm{kl}\left(\mathbb{E}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)\|\,\mathbb{E}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{S}}(h)\right)$,

i.e., we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),\ \mathop{\mathbb{E}}_{h\sim\rho}\mathrm{kl}\Big(\mathsf{R}^\ell_{\mathcal{D}}(h)\|\mathsf{R}^\ell_{\mathcal{S}}(h)\Big)\right.$$
$$\left.\leq\frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}e^{m\,\mathrm{kl}\left(\mathsf{R}^\ell_{\mathcal{D}}(h')\|\mathsf{R}^\ell_{\mathcal{S}'}(h')\right)}\right)\right]\right]\geq 1-\delta.$$

Since the distribution π on $\mathbb{H}$ does not depend on the $\mathbb{S}'\sim\mathcal{D}^m$, we can exchange the two expectations (with Fubini's theorem), *i.e.*, we have

$$\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}e^{m\,\mathrm{kl}\left(\mathsf{R}^\ell_{\mathcal{D}}(h')\|\mathsf{R}^\ell_{\mathcal{S}'}(h')\right)}=\mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}e^{m\,\mathrm{kl}\left(\mathsf{R}^\ell_{\mathcal{D}}(h')\|\mathsf{R}^\ell_{\mathcal{S}'}(h')\right)}.$$

Then, from Lemma B.16.1, we have

$$\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}e^{m\,\mathrm{kl}\left(\mathsf{R}^\ell_{\mathcal{D}}(h')\|\mathsf{R}^\ell_{\mathcal{S}'}(h')\right)}\leq 2\sqrt{m}$$
$$\implies\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}e^{m\,\mathrm{kl}\left(\mathsf{R}^\ell_{\mathcal{D}}(h')\|\mathsf{R}^\ell_{\mathcal{S}'}(h')\right)}\right)\leq\ln\frac{2\sqrt{m}}{\delta}.$$

Finally, thanks to the joint convexity of the KL divergence (Proposition B.5.1), $\mathrm{kl}(q\|p)$ is jointly convex in q and p. Hence, we have from Jensen's inequality (Theorem A.1.1), for all $\rho\in\mathbb{M}(\mathbb{H})$

$$\mathrm{kl}\left(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)\Big\|\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{S}}(h)\right)\leq\mathop{\mathbb{E}}_{h\sim\rho}\mathrm{kl}\Big(\mathsf{R}^\ell_{\mathcal{D}}(h)\|\mathsf{R}^\ell_{\mathcal{S}}(h)\Big).$$

■

B.10 Proof of Proposition 2.3.1

Proposition 2.3.1 (Donsker-Varadhan Variational Representation)**.** For any hypothesis set $\mathbb{H}$, for any distribution $\pi\in\mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any measurable function $\varphi:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m\to\mathbb{R}$ *s.t.* $\mathbb{E}_{h'\sim\pi}e^{\varphi(h',\mathbb{S})}<+\infty$ for all $\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m$, we have

$$\forall\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m,\quad\forall\rho\in\mathbb{M}(\mathbb{H}),\qquad\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})-\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}e^{\varphi(h,\mathbb{S})}\right)\leq\mathrm{KL}(\rho\|\pi)$$
$$\iff\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})\leq\mathrm{KL}(\rho\|\pi)+\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}e^{\varphi(h,\mathbb{S})}\right).$$

When the distribution ρ is defined as $\rho(h) = \pi(h)\frac{e^{\varphi(h,\mathbb{S})}}{\mathbb{E}_{h'\sim\pi} e^{\varphi(h',\mathbb{S})}}$, we have

$$\forall \mathbb{S} \in (\mathbb{X}\times\mathbb{Y})^m, \qquad \mathop{\mathbb{E}}_{h\sim\rho} \varphi(h,\mathbb{S}) - \ln\left(\mathop{\mathbb{E}}_{h\sim\pi} e^{\varphi(h,\mathbb{S})}\right) = \mathrm{KL}(\rho\|\pi),$$
$$\iff \mathop{\mathbb{E}}_{h\sim\rho} \varphi(h,\mathbb{S}) = \mathrm{KL}(\rho\|\pi) + \ln\left(\mathop{\mathbb{E}}_{h\sim\pi} e^{\varphi(h,\mathbb{S})}\right).$$

Proof. Let $\rho'(h) = \pi(h)\frac{e^{\varphi(h,\mathbb{S})}}{\mathbb{E}_{h'\sim\pi} e^{\varphi(h',\mathbb{S})}}$, then, we develop the term $\mathrm{KL}(\rho\|\rho')$. We have

$$\begin{aligned}
\mathrm{KL}(\rho\|\rho') &= \mathop{\mathbb{E}}_{h\sim\rho} \ln\left(\frac{\rho(h)}{\rho'(h)}\right) \\
&= \mathop{\mathbb{E}}_{h\sim\rho} \ln\left(\frac{\rho(h)}{\pi(h)e^{\varphi(h,\mathbb{S})}} \mathop{\mathbb{E}}_{h\sim\rho} e^{\varphi(h,\mathbb{S})}\right) \\
&= \mathop{\mathbb{E}}_{h\sim\rho} \ln\left(\frac{\rho(h)}{\pi(h)}\right) - \mathop{\mathbb{E}}_{h\sim\pi} \varphi(h,\mathbb{S}) + \ln\left(\mathop{\mathbb{E}}_{h\sim\pi} e^{\varphi(h,\mathbb{S})}\right).
\end{aligned}$$

Hence, by setting $\rho = \rho'$, we have $\mathrm{KL}(\rho\|\rho') = 0$ which leads to the desired result by rearranging the terms. ■

B.11 Proof of Theorem 2.3.5

Theorem 2.3.5 (General PAC-Bayesian Bound of Bégin *et al.* (2016))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}_*^+$, for any $\lambda > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}), \frac{\lambda}{\lambda-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})\right]\right.$$
$$\left.\leq \mathrm{D}_\lambda(\rho\|\pi) + \ln\left[\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right]\right] \geq 1-\delta.$$

Proof. We start by developing $\frac{\lambda}{\lambda-1}\ln\left[\mathbb{E}_{h\sim\rho}\,\varphi(h,\mathbb{S})\right]$, *i.e.*, we have for all $\rho\in\mathbb{M}(\mathbb{H})$

$$\frac{\lambda}{\lambda-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\rho}\varphi(h,\mathbb{S})\right]=\frac{\lambda}{\lambda-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\rho}\frac{\rho(h)}{\pi(h)}\frac{\pi(h)}{\rho(h)}\varphi(h,\mathbb{S})\right]=\frac{\lambda}{\lambda-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\pi}\frac{\rho(h)}{\pi(h)}\varphi(h,\mathbb{S})\right]. \tag{B.3}$$

We apply Hölder's inequality (Theorem A.5.1) to have

$$\mathop{\mathbb{E}}_{h\sim\pi}\frac{\rho(h)}{\pi(h)}\varphi(h,\mathbb{S})\le\left[\mathop{\mathbb{E}}_{h\sim\pi}\left[\frac{\rho(h)}{\pi(h)}\right]^{\lambda}\right]^{\frac{1}{\lambda}}\left[\mathop{\mathbb{E}}_{h\sim\pi}\varphi(h,\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right]^{\frac{\lambda-1}{\lambda}}.$$

By taking the logarithm (since both sides are positive) and multiplying by $\frac{\lambda}{\lambda-1}$ both sides of the inequality, we have

$$\begin{aligned}\frac{\lambda}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}\frac{\rho(h)}{\pi(h)}\varphi(h,\mathbb{S})\right)&\le\frac{1}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}\left[\frac{\rho(h)}{\pi(h)}\right]^{\lambda}\right)+\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\\&=\mathrm{D}_\lambda(\rho\|\pi)+\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right).\end{aligned} \tag{B.4}$$

Since the function $\varphi:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m\to\mathbb{R}_*^+$ is positive and $\frac{\lambda}{\lambda-1}>0$, we can deduce that $\mathbb{E}_{h'\sim\pi}\,\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}>0$ for all $\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m$. Then, we can apply Markov's inequality (Theorem A.2.1), we have

$$\begin{aligned}&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\le\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right]\ge1-\delta\\
\iff&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\le\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]\ge1-\delta\\
\iff&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),\ \mathrm{D}_\lambda(\rho\|\pi)+\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi'}\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\right.\\
&\qquad\left.\le\mathrm{D}_\lambda(\rho\|\pi)+\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]\ge1-\delta\end{aligned} \tag{B.5}$$

By combining Equations (B.3) to (B.5), we can deduce the claimed result. ■

B.12 About the Bounds Derived From Theorem 2.3.5

In this section, we provide a SEEGER, MCALLESTER and CATONI-like PAC-Bayesian generalization bound.

B.12.1 McAllester-like Bound

Based on Corollary B.12.2, we can obtain the following MCALLESTER-like bound.

Corollary B.12.1. For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any loss $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to [0,1]$, for any $\lambda > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\forall \rho \in \mathbb{M}(\mathbb{H}), \left|\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^\ell_{\mathcal{D}}(h) - \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^\ell_{\mathcal{S}}(h)\right| \leq \sqrt{\frac{1}{2m}\Big[\mathrm{D}_\lambda(\rho\|\pi) + \ln \tfrac{2\sqrt{m}}{\delta}\Big]}\right] \geq 1-\delta.$$

Proof. It is a direct consequence of PINSKER's inequality (Theorem B.5.1), *i.e.*, we have

$$2\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^\ell_{\mathcal{D}}(h) - \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^\ell_{\mathcal{S}}(h)\right)^2 \leq \mathrm{kl}\Big(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^\ell_{\mathcal{D}}(h) \Big\| \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^\ell_{\mathcal{S}}(h)\Big),$$

and from Corollary B.12.2 by rearranging the terms. ■

B.12.2 Seeger-like Bound

The SEEGER-like obtained from Theorem 2.3.5 is the following.

Corollary B.12.2. For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any loss $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to [0,1]$, for any $\lambda > 1$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\forall \rho \in \mathbb{M}(\mathbb{H}),\ \mathrm{kl}\Big(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^\ell_{\mathcal{D}}(h) \Big\| \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^\ell_{\mathcal{S}}(h)\Big) \leq \frac{1}{m}\Big[\mathrm{D}_\lambda(\rho\|\pi) + \ln \tfrac{2\sqrt{m}}{\delta}\Big]\right] \geq 1-\delta.$$

Proof. We apply Theorem 2.3.5 with $\varphi(h, \mathbb{S}) = \frac{\lambda-1}{\lambda} m \operatorname{kl}\left[\mathbb{E}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h) \| \mathbb{E}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{S}}(h)\right]$, *i.e.*, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \in \mathbb{M}(\mathbb{H}),\ \frac{\lambda}{\lambda-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\rho} e^{m\frac{\lambda-1}{\lambda}\operatorname{kl}\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h)\|\mathsf{R}^{\ell}_{\mathcal{S}}(h)\right)}\right] \leq \mathrm{D}_{\lambda}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi} e^{m\operatorname{kl}\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h')\|\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right)}\right)\right] \geq 1-\delta.$$

Since the distribution π on $\mathbb{H}$ does not depend on the $\mathbb{S}' \sim \mathcal{D}^m$, we can exchange the two expectations (with Fubini's theorem), *i.e.*, we have

$$\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi} e^{m\operatorname{kl}\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h')\|\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right)} = \mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} e^{m\operatorname{kl}\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h')\|\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right)}.$$

Then, from Lemma B.16.1, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} e^{m\operatorname{kl}\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h)\|\mathsf{R}^{\ell}_{\mathcal{S}}(h)\right)} \leq 2\sqrt{m} \\
\implies &\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} e^{m\operatorname{kl}\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h')\|\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right)}\right) \leq \ln\frac{2\sqrt{m}}{\delta}.
\end{aligned}$$

Thanks to the joint convexity of the KL divergence (Proposition B.5.1), the function $q, p \mapsto \exp\left(\frac{\lambda-1}{\lambda} m \operatorname{kl}(q\|p)\right)$ is jointly convex in q and p by composition (see, *e.g.*, Boyd and Vandenberghe (2004)). Hence, we have from Jensen's inequality (Theorem A.1.1), for all $\rho \in \mathbb{M}(\mathbb{H})$

$$\begin{aligned}
&e^{m\frac{\lambda-1}{\lambda}\operatorname{kl}(\mathbb{E}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)\|\,\mathbb{E}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h))} \leq \mathop{\mathbb{E}}_{h\sim\rho} e^{m\frac{\lambda-1}{\lambda}\operatorname{kl}\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h)\|\mathsf{R}^{\ell}_{\mathcal{S}}(h)\right)} \\
\iff &m\operatorname{kl}\left[\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)\|\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)\right] \leq \frac{\lambda}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h\sim\rho} e^{m\frac{\lambda-1}{\lambda}\operatorname{kl}\left(\mathsf{R}^{\ell}_{\mathcal{D}}(h)\|\mathsf{R}^{\ell}_{\mathcal{S}}(h)\right)}\right).
\end{aligned}$$

Finally, by rearranging the terms, we have the stated result. ■

B.12.3 Catoni-like Bound

The derivation of a Catoni (2007)-like generalization bound based on Theorem 2.3.5 is the following.

Corollary B.12.3. For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any loss $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to [0,1]$, for any $c > 0$,

for any $\lambda > 1$, for any $\delta \in (0, 1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}), -\ln\Big(1-\Big[1-e^{-c}\Big]\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)\Big) - c\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{S}}(h) \leq \frac{1}{m}\left[\mathrm{D}_\lambda(\rho\|\pi)+\ln\frac{1}{\delta}\right]\right] \geq 1-\delta.$$

Proof. We apply Theorem 2.3.5 with $\varphi(h,\mathbb{S}) = m\frac{\lambda-1}{\lambda}\Big[F(\mathsf{R}^\ell_{\mathcal{D}}(h)) - c\mathsf{R}^\ell_{\mathcal{S}}(h)\Big]$, where $F(\mathsf{R}^\ell_{\mathcal{D}}(h)) \triangleq -\ln\Big(1-\mathsf{R}^\ell_{\mathcal{D}}(h)[1-e^{-c}]\Big)$. We have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),\ \frac{\lambda}{\lambda-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\rho}e^{m\frac{\lambda-1}{\lambda}F(\mathsf{R}^\ell_{\mathcal{D}}(h))-c\mathsf{R}^\ell_{\mathcal{S}}(h)}\right] \leq \mathrm{D}_\lambda(\rho\|\pi)+\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}e^{m\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h'))-c\mathsf{R}^\ell_{\mathcal{S}'}(h')\right]}\right)\right] \geq 1-\delta.$$

Since the distribution π on $\mathbb{H}$ does not depend on the $\mathbb{S}' \sim \mathcal{D}^m$, we can exchange the two expectations (with Fubini's theorem), *i.e.*, we have

$$\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}e^{m\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h'))-c\mathsf{R}^\ell_{\mathcal{S}'}(h')\right]} = \mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}e^{m\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h'))-c\mathsf{R}^\ell_{\mathcal{S}'}(h')\right]}.$$

Then, from Lemma B.16.2, we have

$$\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}e^{m\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h))-c\mathsf{R}^\ell_{\mathcal{S}}(h)\right]} \leq 1 \Longrightarrow \ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}e^{m\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h'))-c\mathsf{R}^\ell_{\mathcal{S}'}(h')\right]}\right) \leq \ln\frac{1}{\delta}.$$

The function $F(x)$ is convex, since its second derivative is $\frac{\partial^2 F}{\partial x^2}(x) = \frac{(e^c-1)^2}{(x-e^c(x-1))^2} \geq 0$. Hence, we conclude that $F(p) - cq$ is jointly convex in q and p. Moreover, we can deduce that the function $q, p \mapsto \exp\left(m\frac{\lambda-1}{\lambda}\Big[F(\mathsf{R}^\ell_{\mathcal{D}}(h)) - c\mathsf{R}^\ell_{\mathcal{S}}(h)\Big]\right)$ is jointly convex in q and p by composition (see, *e.g.*, Boyd and Vandenberghe (2004)). Hence, from Jensen's inequality (Theorem A.1.1) we have for all $\rho \in \mathbb{M}(\mathbb{H})$

$$e^{m\frac{\lambda-1}{\lambda}\left[F(\mathbb{E}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h))-c\,\mathbb{E}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{S}}(h)\right]} \leq \mathop{\mathbb{E}}_{h\sim\rho}e^{m\frac{\lambda-1}{\lambda}\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h))-c\mathsf{R}^\ell_{\mathcal{S}}(h)\right]}$$

$$\iff m\left[F\left(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{D}}(h)\right) - c\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^\ell_{\mathcal{S}}(h)\right] \leq \frac{\lambda}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h\sim\rho}e^{m\frac{\lambda-1}{\lambda}\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h))-c\mathsf{R}^\ell_{\mathcal{S}}(h)\right]}\right).$$

Finally, by rearranging the terms, we have the stated result. ■

B.13 Proof of Theorem 2.4.1

Theorem 2.4.1 (General Disintegrated Bound of RIVASPLATA *et al.* (2020))**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any measurable function $\varphi : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$, for any $\delta \in (0, 1]$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\varphi(h, \mathbb{S}) \leq \underbrace{\ln\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right] + \ln\left[\frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} \exp\left(\varphi(h', \mathbb{S}')\right)\right]}_{\Phi(\rho_{\mathbb{S}}, \pi, \delta)} \right] \geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S}, \pi)$ is output by the deterministic algorithm A.

Proof. Note that $\exp\left[\varphi(h, \mathbb{S}) - \ln \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]$ is a non-negative random variable. Thus, we can apply MARKOV's inequality (Theorem A.2.1) to obtain

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\exp\left(\varphi(h, \mathbb{S}) - \ln \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right) \right.$$
$$\left. \leq \frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}} \exp\left(\varphi(h', \mathbb{S}') - \ln \frac{\rho_{\mathbb{S}'}(h')}{\pi(h')}\right) \right] \geq 1 - \delta.$$

Hence, by rearranging the terms, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\exp\left(\varphi(h, \mathbb{S}) - \ln \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right) \leq \frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}} \frac{\pi(h')}{\rho_{\mathbb{S}'}(h')} e^{\varphi(h', \mathbb{S}')} \right] \geq 1-\delta$$
$$\iff \mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\exp\left(\varphi(h, \mathbb{S}) - \ln \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right) \leq \frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} e^{\varphi(h', \mathbb{S}')} \right] \geq 1-\delta.$$

Since both sides of the inequality are strictly positive, we can apply the logarithm, *i.e.*, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\varphi(h, \mathbb{S}) \leq \ln \frac{\rho_{\mathbb{S}}(h)}{\pi(h)} + \ln\left(\frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} e^{\varphi(h', \mathbb{S}')}\right) \right] \geq 1 - \delta,$$

which is the desired result. ■

B.14 Proof of Theorem 2.4.2

Theorem 2.4.2 (Disintegrated Bound of Catoni (2007))**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis $\mathbb{H}$, for any prior $\pi \in \mathbb{M}^*(\mathbb{H})$, for any loss $\ell : \mathbb{H}\times(\mathbb{X} \times \mathbb{Y})^m \to [0,1]$, for any $c > 0$, for any $\delta \in (0,1]$, for any algorithm $A\colon(\mathbb{X}\times\mathbb{Y})^m\times\mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),-\ln\Big(1-\Big[1-e^{-c}\Big]\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)\Big)-c\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)\right.$$
$$\left.\leq\frac{1}{m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\frac{1}{\delta}\right]\right]\geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S},\pi)$ is output by the deterministic algorithm A.

Proof. We apply Theorem 2.4.1 with $\varphi(h,\mathbb{S}) = m\Big[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h)) - c\mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big]$, where $F(\mathsf{R}^{\ell}_{\mathcal{D}}(h)) \triangleq -\ln\Big(1-\mathsf{R}^{\ell}_{\mathcal{D}}(h)[1-e^{-c}]\Big)$. We have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall\rho\in\mathbb{M}(\mathbb{H}),\ \mathop{\mathbb{E}}_{h\sim\rho}\Big[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h))-c\mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big]\right.$$
$$\left.\leq\frac{1}{m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}+\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}e^{m\left[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h'))-c\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right]}\right)\right]\right]\geq 1-\delta.$$

Since the distribution π on $\mathbb{H}$ does not depend on the $\mathbb{S}'\sim\mathcal{D}^m$, we can exchange the two expectations (with Fubini's theorem), *i.e.*, we have

$$\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}e^{m\left[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h'))-c\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right]}=\mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}e^{m\left[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h'))-c\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right]}.$$

Then, from Lemma B.16.2, we have

$$\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}e^{m\left[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h))-c\mathsf{R}^{\ell}_{\mathcal{S}}(h)\right]}\leq 1\Longrightarrow\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}e^{m\left[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h'))-c\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\right]}\right)\leq\ln\frac{1}{\delta}.$$

The function $F(x)$ is convex, since its second derivative is $\frac{\partial^2 F}{\partial x^2}(x)=\frac{(e^c-1)^2}{(x-e^c(x-1))^2}\geq 0$. In this case, easily conclude that $F(p) - cq$ is jointly convex in q and p. Hence,

from JENSEN's inequality (Theorem A.1.1) we have for all $\rho \in \mathbb{M}(\mathbb{H})$

$$F\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h)\right) - c\left[\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathbb{S}}(h)\right] \leq \mathop{\mathbb{E}}_{h\sim\rho}\left[F(\mathsf{R}^{\ell}_{\mathcal{D}}(h)) - c\mathsf{R}^{\ell}_{\mathbb{S}}(h)\right].$$

Hence, this gives the bound

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \rho \in \mathbb{M}(\mathbb{H}),\ F\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h)\right) - c\left[\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathbb{S}}(h)\right] \leq \frac{1}{m}\left[\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)} + \ln\frac{1}{\delta}\right]\right] \geq 1-\delta,$$

and by rearranging the terms we obtain the desired result. ■

B.15 Proof of Theorem 2.4.3

The proof of Theorem 2.4.3 relies on another theorem from BLANCHARD and FLEURET (2007, Theorem 2.4) called OCCAM's Hammer.

Lemma B.15.1 (OCCAM's Hammer)**.** Given a measurable function measurable function $\varphi : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}$, assume that for any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\pi}\left[\varphi(h,\mathbb{S}) \geq \Phi(\delta)\right] \leq \delta,$$

where $\Phi : (0,1] \to \mathbb{R}$ is a decreasing function.
It implies that for any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution π on $\mathbb{H}$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\varphi(h,\mathbb{S}) \geq \Phi'\left(\delta f\left(\Theta(h)^{-1}\right)\right)\right] \leq \delta,$$

where $\Theta(h) \triangleq \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}$ and $\Phi'(\delta) \triangleq \Phi\Big(\max\Big(\delta, 1\Big)\Big)$ and $f : \mathbb{R}^+ \to \mathbb{R}^+_*$ a measurable increasing function *s.t.*

$$\int_{y>0} y^{-2} f(y)\, dy \leq 1.$$

Proof. The proof consists in upper-bounding the probability

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\varphi(h,\mathbb{S}) \geq \left(\delta f\left(\Theta(h)^{-1}\right)\right)\right].$$

We rewrite this term as

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\varphi(h,\mathbb{S}) \geq \Phi'\Big(\delta f\left(\Theta(h)^{-1}\right)\Big)\right] \\
&\qquad = \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\Big(\delta f\left(\Theta(h)^{-1}\right)\Big)\right] \\
&\qquad = \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\frac{\pi(h)}{\rho_{\mathbb{S}}(h)} \mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\Big(\delta f\left(\Theta(h)^{-1}\right)\Big)\right] \\
&\qquad = \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi} \frac{\rho_{\mathbb{S}}(h)}{\pi(h)} \mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\Big(\delta f\left(\Theta(h)^{-1}\right)\Big)\right] \\
&\qquad = \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi} \Theta(h) \mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\Big(\delta f\left(\Theta(h)^{-1}\right)\Big)\right].
\end{aligned}$$

We can actually express the term $\Theta(h)$ in a form of integral. Indeed, we have

$$\Theta(h) = \int_{\Theta(h)^{-1}}^{+\infty} y^{-2} dy = \int_{y>0} y^{-2}\,\mathrm{I}\left[y \geq \Theta(h)^{-1}\right] dy.$$

Then, thanks to Fubini's theorem, we can rewrite the probability as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi} \Theta(h)\,\mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\left(\delta f\left(\Theta(h)^{-1}\right)\right)\right] \\
&\quad = \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi} \left[\int_{y>0} y^{-2}\,\mathrm{I}\left[y \geq \Theta(h)^{-1}\right] dy\right] \mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\left(\delta f\left(\Theta(h)^{-1}\right)\right)\right] \\
&\quad = \int_{y>0} y^{-2} \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi} \mathrm{I}\left[y \geq \Theta(h)^{-1}\right] \mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\left(\delta f\left(\Theta(h)^{-1}\right)\right)\right] dy.
\end{aligned}$$

Since $f()$ is increasing, *i.e.*, for all $y > 0$ *s.t.* $\Theta(h)^{-1} < y$, we have $f(\Theta(h)^{-1}) \leq f(y)$. Moreover, since $\Phi()$ is decreasing, we have $\Phi'(f(y)) \leq \Phi'(f(\Theta(h)^{-1}))$. This implies that

$$\mathrm{I}\left[y \geq \Theta(h)^{-1}\right] \mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\left(\delta f\left(\Theta(h)^{-1}\right)\right)\right] \leq \mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\left(\delta f\left(y\right)\right)\right].$$

Based on this inequality, we have

$$\begin{aligned}
&\int_{y>0} y^{-2} \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\pi} \mathrm{I}\left[y \geq \Theta(h)^{-1}\right] \mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\left(\delta f\left(\Theta(h)^{-1}\right)\right)\right] dy \\
&\quad \leq \int_{y>0} y^{-2} \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\pi} \mathrm{I}\left[\varphi(h,\mathbb{S}) \geq \Phi'\left(\delta f\left(y\right)\right)\right] dy \\
&\quad = \int_{y>0} y^{-2} \mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\pi} \left[\varphi(h,\mathbb{S}) \geq \Phi'\left(\delta f\left(y\right)\right)\right] dy.
\end{aligned}$$

By assumption, we have $\int_{y>0} y^{-2} f\left(y\right) dy \leq 1$ and $\mathbb{P}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\pi}\left[\varphi(h,\mathbb{S}) \geq \Phi(\delta)\right] \leq \delta$, which leads to

$$\int_{y>0} y^{-2} \mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\pi} \left[\varphi(h,\mathbb{S}) \geq \Phi'\left(\delta f\left(y\right)\right)\right] dy \leq \int_{y>0} y^{-2} \delta f\left(y\right) dy \leq \delta.$$

■

This theorem is further used in addition to the two following lemmas. We first one known as Chernoff's bound is originally due to Chernoff (1952) but the proof is from Langford (2005, Lemma 3.6).

Lemma B.15.2 (Chernoff's bound)**.** For any

$$\mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{B}(p)^m} \left[\sum_{i=1}^{m} X_i \leq k\right] \leq e^{-m\,\mathrm{kl}_+\left(\frac{k}{m}\|p\right)},$$

where $\mathrm{kl}_+\left(\frac{k}{m}\|p\right) = \mathrm{kl}(\frac{k}{m}\|p)$ if $\frac{k}{m} < p$ and 0 otherwise.

Proof. First remark that we have

$$\mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{B}(p)^m} \left[\sum_{i=1}^{m} X_i \leq k\right] = \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{B}(p)^m} \left[e^{-m\lambda\frac{1}{m}\sum_{i=1}^{m} X_i} \geq e^{-m\lambda\frac{k}{m}}\right].$$

Then, from Markov's inequality (Theorem A.2.1), we have

$$\mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{B}(p)^m} \left[e^{-m\lambda\frac{1}{m}\sum_{i=1}^{m} X_i} \geq e^{-m\lambda\frac{k}{m}}\right] \leq \frac{\mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^m}\, e^{-\lambda\sum_{i=1}^{m} X_i}}{e^{-\lambda k}}.$$

Then, using the fact that $X_1, \ldots, X_m$ are *i.i.d.* and from the expression of the

moment generating function of the Bernoulli distribution $\mathcal{B}(p)$, we have

$$\frac{\mathbb{E}_{\mathbf{X}\sim\mathcal{B}(p)^m}\, e^{-\lambda\sum_{i=1}^m X_i}}{e^{-\lambda k}} = e^{\lambda k}\left[\mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{B}(p)} e^{-\lambda X}\right]^m = e^{\lambda k}\left[pe^{\lambda}+(1-p)\right]^m.$$

Actually, we can find the optimal value λ^*, which is

$$\lambda^* = \ln\left[p\left(1-\frac{k}{m}\right)\right] - \ln\left[\frac{k}{m}(1-p)\right],$$

for all $p > \frac{k}{m}$. Finally, setting $\lambda = \lambda^*$, we obtain

$$e^{\lambda k}\left[pe^{\lambda}+(1-p)\right]^m = e^{-m\,\mathrm{kl}_+\left(\frac{k}{m}\|p\right)}.$$

■

Chernoff's bound is actually used to prove the test set bound of Langford (2005, Theorem 3.3 and Corollary 3.7). We prove, with more details, his theorem in the following lemma.

Lemma B.15.3. For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any hypothesis $h\in\mathbb{H}$, for any loss $\ell:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})\to\{0,1\}$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathrm{kl}_+(\mathsf{R}^\ell_{\mathbb{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h)) \geq \frac{\ln\frac{1}{\delta}}{m}\right] \leq \delta.$$

Proof. **Step 1.** First of all, we prove that

$$\mathop{\mathbb{P}}_{Y_1,\ldots,Y_m\sim\mathcal{B}(p)^m}\left[\mathop{\mathbb{P}}_{X_1,\ldots,X_m\sim\mathcal{B}(p)^m}\left[\sum_{i=1}^m X_i \leq \sum_{i=1}^m Y_i\right] \leq \delta\right] \leq \delta. \tag{B.6}$$

To do so, let $k^* = \max\left\{k\in\{0,\ldots,m\} \;\middle|\; \mathbb{P}_{X_1,\ldots,X_m\sim\mathcal{B}(p)^m}\left[\sum_{i=1}^m X_i \leq k\right] \leq \delta\right\}$. Then, we have

$$\begin{aligned}
&\mathop{\mathbb{P}}_{Y_1,\ldots,Y_m\sim\mathcal{B}(p)^m}\left[\mathop{\mathbb{P}}_{X_1,\ldots,X_m\sim\mathcal{B}(p)^m}\left[\sum_{i=1}^m X_i \leq \sum_{i=1}^m Y_i\right] \leq \delta\right] \\
&= \mathop{\mathbb{E}}_{Y_1,\ldots,Y_m\sim\mathcal{B}(p)^m} \mathrm{I}\left[\mathop{\mathbb{P}}_{X_1,\ldots,X_m\sim\mathcal{B}(p)^m}\left[\sum_{i=1}^m X_i \leq \sum_{i=1}^m Y_i\right] \leq \delta\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=0}^{m} \mathop{\mathbb{P}}_{X_1,\dots,X_m \sim \mathcal{B}(p)^m} \left[\sum_{i=1}^{m} X_i = k\right] \cdot \mathrm{I}\left[\mathop{\mathbb{P}}_{X_1,\dots,X_m \sim \mathcal{B}(p)^m} \left[\sum_{i=1}^{m} X_i \leq \sum_{i=1}^{m} Y_i\right] \leq \delta\right] \\
&= \sum_{k=0}^{k^*} \mathop{\mathbb{P}}_{X_1,\dots,X_m \sim \mathcal{B}(p)^m} \left[\sum_{i=1}^{m} X_i = k\right] \\
&= \mathop{\mathbb{P}}_{X_1,\dots,X_m \sim \mathcal{B}(p)^m} \left[\sum_{i=1}^{m} X_i \leq k^*\right] \\
&\leq \delta.
\end{aligned}
$$

Step 2. From Equation (B.6), we can deduce that

$$
\mathop{\mathbb{P}}_{\mathsf{R}_{\mathcal{S}}^{\ell}(h) \sim \mathcal{B}(\mathsf{R}_{\mathcal{D}}^{\ell}(h))^m} \left[\mathop{\mathbb{P}}_{\mathbf{X} \sim \mathcal{B}(\mathsf{R}_{\mathcal{D}}^{\ell}(h))^m} \left[\frac{1}{m}\sum_{i=1}^{m} X_i \leq \mathsf{R}_{\mathcal{S}}^{\ell}(h)\right] \leq \delta\right] \leq \delta,
$$

where $\mathsf{R}_{\mathcal{S}}^{\ell}(h) \sim \mathcal{B}(\mathsf{R}_{\mathcal{D}}^{\ell}(h))^m$ is a slight abuse of notations since $m\mathsf{R}_{\mathcal{S}}^{\ell}(h)$ is the sum of the successes. From Chernoff's inequality (Lemma B.15.2), we can deduce that

$$
\mathop{\mathbb{P}}_{\mathcal{S} \sim \mathcal{D}^m} \left[\exp\Big(-m\,\mathrm{kl}_+\left[\mathsf{R}_{\mathcal{S}}^{\ell}(h) \| \mathsf{R}_{\mathcal{D}}^{\ell}(h)\right]\Big) \leq \delta\right] \leq \delta.
$$

Finally, by rearranging the terms, we obtain the desired result. ■

Finally, we are able to prove Lemma B.15.1 using Lemmas B.15.1 and B.15.2.

Theorem 2.4.3 (Disintegrated Bound of Blanchard and Fleuret (2007))**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any loss $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to \{0,1\}$, for any $k > 1$, for any $\delta \in (0,1]$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, we have

$$
\mathop{\mathbb{P}}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \rho_{\mathcal{S}}} \left[\mathrm{kl}_+(\mathsf{R}_{\mathcal{S}}^{\ell}(h) \| \mathsf{R}_{\mathcal{D}}^{\ell}(h)) \leq \frac{1}{m}\left[\ln \frac{k+1}{\delta} + \left(1 + \frac{1}{k}\right) \ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)}\right]\right] \geq 1 - \delta,
$$

where $\rho_{\mathcal{S}} \triangleq A(\mathcal{S}, \pi)$ is output by the deterministic algorithm A, the $\ln_+(x) = \max(\ln(x), 0)$ and $\mathrm{kl}_+(\mathsf{R}_{\mathcal{S}}^{\ell}(h) \| \mathsf{R}_{\mathcal{D}}^{\ell}(h)) = \mathrm{kl}(\mathsf{R}_{\mathcal{S}}^{\ell}(h) \| \mathsf{R}_{\mathcal{D}}^{\ell}(h))$ if $\mathsf{R}_{\mathcal{S}}^{\ell}(h) < \mathsf{R}_{\mathcal{D}}^{\ell}(h)$ and 0 otherwise.

Proof. The proof consist of applying Occam's Hammer (Lemma B.15.1). To do so, given $k > 1$, let $f : \mathbb{R}^+ \to \mathbb{R}^+_*$ the function defined as

$$
f(y) \triangleq \frac{1}{k+1} \min\left(y^{1+\frac{1}{k}}, 1\right)
$$

Indeed, $f()$ is increasing and we have

$$\begin{aligned}
\int_{y>0} y^{-2} f(y)\, dy &= \int_0^1 y^{-2} f(y)\, dy + \int_1^{+\infty} y^{-2} f(y)\, dy \\
&= \frac{1}{k+1}\left[\int_0^1 y^{-2} \cdot y^{1+\frac{1}{k}} dy + \int_1^{+\infty} y^{-2} dy\right] \\
&= \frac{1}{k+1}\left[\int_0^1 y^{-2} \cdot y^{1+\frac{1}{k}} dy + 1\right] \\
&= \frac{1}{k+1}\left[\int_0^1 y^{\frac{1}{k}-1} dy + 1\right] \\
&= \frac{1}{k+1}\left[k \cdot \left[1^{\frac{1}{k}} - 0^{\frac{1}{k}}\right] + 1\right] \\
&= 1.
\end{aligned}$$

Moreover, note that with Lemma B.15.3, we have $\varphi(h, \mathbb{S}) = \mathrm{kl}_+(\mathsf{R}^{\ell}_{\mathcal{S}}(h) \| \mathsf{R}^{\ell}_{\mathcal{D}}(h))$ and $\Phi(\delta) = \frac{\ln \frac{1}{\delta}}{m}$. Hence, we apply OCCAM's Hammer (Lemma B.15.1) to obtain

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\varphi(h, \mathbb{S}) \leq \Phi'\left(\delta f\left(\Theta(h)^{-1}\right)\right)\right] \geq 1 - \delta,$$

To obtain the final bound, we upper-bound the term $\Phi'(\delta f(\Theta(h)^{-1}))$, *i.e.*, we have

$$\begin{aligned}
\Phi'(\delta f(\Theta(h)^{-1})) &= -\frac{1}{m} \ln\left(\min(\delta f(\Theta(h)^{-1}), 1)\right) \\
&= \frac{1}{m} \max\left(\ln\left(\frac{1}{\delta f(\Theta(h)^{-1})}\right), 0\right) \\
&\leq \frac{1}{m} \ln\frac{1}{\delta} + \frac{1}{m} \max\left(-\ln\left(f(\Theta(h)^{-1})\right), 0\right) \\
&= \frac{1}{m} \ln\frac{1}{\delta} + \frac{1}{m} \max\left(-\ln\left(\frac{1}{k+1}\min(\Theta(h)^{-(1+\frac{1}{k})}, 1)\right), 0\right) \\
&\leq \frac{1}{m} \ln\frac{k+1}{\delta} + \frac{1}{m} \max\left(-\ln\left(\min(\Theta(h)^{-(1+\frac{1}{k})}, 1)\right), 0\right) \\
&= \frac{1}{m} \ln\frac{k+1}{\delta} + \frac{1}{m} \ln_+\left(\Theta(h)^{1+\frac{1}{k}}\right) \\
&= \frac{1}{m}\left[\ln\frac{k+1}{\delta} + \left(1 + \frac{1}{k}\right) \ln_+ \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right].
\end{aligned}$$

■

B.16 Proof of Lemmas B.16.1 and B.16.2

In this section, we prove the lemmas necessary to prove Theorems 2.3.3, 2.3.4 and 2.4.2 and Corollary B.12.2. The proof of Lemma B.16.1 is due to Maurer (2004) and Lemma B.16.2 was proven by Germain *et al.* (2009).

Lemma B.16.1. For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any loss $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to [0,1]$, we have

$$\forall h \in \mathbb{H}, \quad \mathop{\mathbb{E}}_{\mathcal{S} \sim \mathcal{D}^m} \exp\Big[m \operatorname{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h) \| \mathsf{R}^\ell_{\mathcal{D}}(h))\Big] \leq 2\sqrt{m}.$$

Lemma B.16.2. For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any loss $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to [0,1]$, we have

$$\forall h \in \mathbb{H}, \quad \mathop{\mathbb{E}}_{\mathcal{S} \sim \mathcal{D}^m} \exp\Big[m \Big[F(\mathsf{R}^\ell_{\mathcal{D}}(h)) - c\mathsf{R}^\ell_{\mathcal{S}}(h)\Big]\Big] \leq 1,$$

where $F(\mathsf{R}^\ell_{\mathcal{D}}(h)) \triangleq -\ln\Big(1 - \mathsf{R}^\ell_{\mathcal{D}}(h)\Big[1 - e^{-c}\Big]\Big)$.

However, before giving the proofs, we need to prove two lemmas.

Lemma B.16.3. For any $m \in \mathbb{N}_*$, any point $\mathbf{x} \in [0,1]^m$ can be written as a convex combination of the extremes points $\boldsymbol{\eta} \in \{0,1\}^m$, *i.e.*, we have

$$\forall \mathbf{x} \in [0,1]^m, \quad \mathbf{x} = \sum_{\boldsymbol{\eta} \in \{0,1\}^m} \left[\prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 0}} (1 - x_i) \prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 1}} x_i \right] \boldsymbol{\eta},$$

where $\sum_{\boldsymbol{\eta} \in \{0,1\}^m} \left[\prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 0}} (1 - x_i) \prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 1}} x_i \right] = 1.$

Proof. We prove this fact by induction.
For $m = 1$, we can easily prove the claim, *i.e.*, we have

$$\forall \mathbf{x} \in [0,1], \quad \mathbf{x} = x_1 \cdot 1 + (1 - x_1) \cdot 0, \quad \text{and} \quad (x_1) + (1 - x_1) = 1.$$

For $m > 1$, we assume that the claim is true for a particular m (from our induction

hypothesis) and we prove the equality for $m+1$.

$$\forall \mathbf{x} \in [0,1]^{m+1}, \sum_{\boldsymbol{\eta} \in \{0,1\}^{m+1}} \left[\prod_{\substack{i \in \{1,\dots,m+1\} \\ \text{s.t. } \eta_i = 0}} (1-x_i) \prod_{\substack{i \in \{1,\dots,m+1\} \\ \text{s.t. } \eta_i = 1}} x_i \right] \boldsymbol{\eta},$$

$$= \sum_{\boldsymbol{\eta} \in \{0,1\}^{m+1}} \left[\prod_{\substack{i \in \{1,\dots,m+1\} \\ \text{s.t. } \eta_i = 0}} (1-x_i) \prod_{\substack{i \in \{1,\dots,m+1\} \\ \text{s.t. } \eta_i = 1}} x_i \right] [\eta_1, \dots, \eta_{m+1}]^\top$$

$$= \sum_{\boldsymbol{\eta} \in \{0,1\}^{m}} \left[\prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 0}} (1-x_i) \prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 1}} x_i \right] x_{m+1} [\eta_1, \dots, 1]^\top$$

$$+ \sum_{\boldsymbol{\eta} \in \{0,1\}^{m}} \left[\prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 0}} (1-x_i) \prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 1}} x_i \right] (1-x_{m+1}) [\eta_1, \dots, 0]^\top .$$

For any $\mathbf{x} \in [0,1]^{m+1}$, its $m+1$-th component is

$$\underbrace{\left(\sum_{\boldsymbol{\eta} \in \{0,1\}^{m}} \left[\prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 0}} (1-x_i) \prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 1}} x_i \right] \right)}_{=1 \text{ by the induction hypothesis}} (x_{m+1} \cdot 1 + (1-x_{m+1}) \cdot 0) = x_{m+1}.$$

Moreover, from the 1-st to the m-th component , we have

$$\sum_{\boldsymbol{\eta} \in \{0,1\}^{m}} \left[\prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 0}} (1-x_i) \prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 1}} x_i \right] x_{m+1} \cdot \boldsymbol{\eta}$$

$$+ \sum_{\boldsymbol{\eta} \in \{0,1\}^{m}} \left[\prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 0}} (1-x_i) \prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 1}} x_i \right] (1-x_{m+1}) \cdot \boldsymbol{\eta}$$

$$= \sum_{\boldsymbol{\eta} \in \{0,1\}^{m}} \left[\prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 0}} (1-x_i) \prod_{\substack{i \in \{1,\dots,m\} \\ \text{s.t. } \eta_i = 1}} x_i \right] [x_{m+1} + (1-x_{m+1})] \cdot \boldsymbol{\eta}$$

$$= \sum_{\boldsymbol{\eta}\in\{0,1\}^m} \left[\prod_{\substack{i\in\{1,\dots,m\} \\ \text{s.t. } \eta_i=0}} (1-x_i) \prod_{\substack{i\in\{1,\dots,m\} \\ \text{s.t. } \eta_i=1}} x_i \right] \cdot \boldsymbol{\eta}$$
$$= [x_1, \dots, x_m]^\top,$$

where the last equality holds by the induction hypothesis. Finally, we prove that it sums to $\sum_{\boldsymbol{\eta}\in\{0,1\}^m} \left[\prod_{\substack{i\in\{1,\dots,m\} \\ \text{s.t. } \eta_i=0}} (1-x_i) \prod_{\substack{i\in\{1,\dots,m\} \\ \text{s.t. } \eta_i=1}} x_i \right] = 1$, *i.e.*, we have

$$\begin{aligned}
\forall \mathbf{x} \in [0,1]^{m+1}, \quad & \sum_{\boldsymbol{\eta}\in\{0,1\}^{m+1}} \left[\prod_{\substack{i\in\{1,\dots,m+1\} \\ \text{s.t. } \eta_i=0}} (1-x_i) \prod_{\substack{i\in\{1,\dots,m+1\} \\ \text{s.t. } \eta_i=1}} x_i \right] \\
&= \sum_{\boldsymbol{\eta}\in\{0,1\}^{m}} \left[\prod_{\substack{i\in\{1,\dots,m\} \\ \text{s.t. } \eta_i=0}} (1-x_i) \prod_{\substack{i\in\{1,\dots,m\} \\ \text{s.t. } \eta_i=1}} x_i \right] x_{m+1} \\
&+ \sum_{\boldsymbol{\eta}\in\{0,1\}^{m}} \left[\prod_{\substack{i\in\{1,\dots,m\} \\ \text{s.t. } \eta_i=0}} (1-x_i) \prod_{\substack{i\in\{1,\dots,m\} \\ \text{s.t. } \eta_i=1}} x_i \right] (1-x_{m+1}) \\
&= 1,
\end{aligned}$$

by the induction hypothesis. ■

The second lemma that we have to prove is the following.

Lemma B.16.4. Let $X \sim \mathcal{X}$ be a random variable such that $X \in [0,1]$ and X' be a random variable following a Bernoulli distribution of parameter $\mathbb{E}_{X\sim\mathcal{X}}[X]$, *i.e.*, $X' \sim \mathcal{B}(\mathbb{E}_{X\sim\mathcal{X}}[X])$. We define as $\mathbf{X} \sim \mathcal{X}^m$ (*resp.* $\mathbf{X}' \sim \mathcal{B}(\mathbb{E}_{X\sim\mathcal{X}}[X])^m$) the m independent copies of $X \sim \mathcal{X}$ (*resp.* $X' \sim \mathcal{B}(\mathbb{E}_{X\sim\mathcal{X}}[X])$).
If $f : [0,1]^m \to \mathbb{R}$ is a convex function and permutation symmetric, we have

$$\mathbb{E}_{\mathbf{X}\sim\mathcal{X}^m}[F(\mathbf{X})] \leq \mathbb{E}_{\mathbf{X}'\sim\mathcal{B}(\mathbb{E}_{X\sim\mathcal{X}}[X])^m}[F(\mathbf{X}')].$$

Proof. From Lemma B.16.3, we have

$$\mathbf{x} = \sum_{\boldsymbol{\eta}\in\{0,1\}^m} \left[\prod_{\substack{i\in\{1,\dots,m\}\\ \eta_i=0}} (1-x_i) \prod_{\substack{i\in\{1,\dots,m\}\\ \eta_i=1}} x_i \right] \boldsymbol{\eta}. \tag{B.7}$$

Hence from Equation (B.7) and from JENSEN's inequality (Theorem A.1.1), we have

$$\begin{aligned} F(\mathbf{x}) &= f\left(\sum_{\boldsymbol{\eta}\in\{0,1\}^m} \left[\prod_{\substack{i\in\{1,\dots,m\}\\ \eta_i=0}} (1-x_i) \prod_{\substack{i\in\{1,\dots,m\}\\ \eta_i=1}} x_i \right] \boldsymbol{\eta} \right) \\ &\leq \sum_{\boldsymbol{\eta}\in\{0,1\}^m} \left[\prod_{\substack{i\in\{1,\dots,m\}\\ \eta_i=0}} (1-x_i) \prod_{\substack{i\in\{1,\dots,m\}\\ \eta_i=1}} x_i \right] F(\boldsymbol{\eta}). \end{aligned}$$

Taking the expectation gives us

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{X}^m} F(\mathbf{X}) &\leq \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{X}^m} \left(\sum_{\boldsymbol{\eta}\in\{0,1\}^m} \left[\prod_{\substack{i\in\{1,\dots,m\}\\ \eta_i=0}} (1-X_i) \prod_{\substack{i\in\{1,\dots,m\}\\ \eta_i=1}} X_i \right] F(\boldsymbol{\eta}) \right) \\ &= \sum_{\boldsymbol{\eta}\in\{0,1\}^m} \left[\prod_{\substack{i\in\{1,\dots,m\}\\ \eta_i=0}} \left(1 - \mathop{\mathbb{E}}_{X_i\sim\mathcal{X}}[X_i]\right) \prod_{\substack{i\in\{1,\dots,m\}\\ \eta_i=1}} \mathop{\mathbb{E}}_{X_i\sim\mathcal{X}}[X_i] \right] F(\boldsymbol{\eta}) \\ &= \sum_{\boldsymbol{\eta}\in\{0,1\}^m} \left[\left(1 - \mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right)^{\mathrm{card}(\{i\,:\,\eta_i=0\})} \left(\mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right)^{\mathrm{card}(\{i\,:\,\eta_i=1\})} \right] F(\boldsymbol{\eta}) \\ &= \sum_{k=0}^{m} \binom{m}{k} \left(1 - \mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right)^{m-k} \left(\mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right)^{k} F(\underbrace{1,\dots,1}_{k \text{ times}}, \underbrace{0,\dots,0}_{m-k \text{ times}}) \\ &= \mathop{\mathbb{E}}_{\mathbf{X}'\sim\mathcal{B}(\mathbb{E}_{X\sim\mathcal{X}}[X])^m} F(\mathbf{X}'). \end{aligned}$$

■

We are now ready to prove Lemma B.16.1.

Proof of Lemma B.16.1. Since the KL divergence is jointly convex so is the function $\exp\left[m\,\mathrm{kl}(\cdot\|\mathsf{R}^\ell_{\mathcal{D}}(h))\right]$ (see, *e.g.*, BOYD and VANDENBERGHE (2004, Section 3.2.4)). Then, we can apply Lemma B.16.4 to have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m} \exp\left[m\,\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))\right] \\
&\quad\le \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{B}(\mathsf{R}^\ell_{\mathcal{D}}(h))^m} \exp\left[m\,\mathrm{kl}\left(\frac{1}{m}\sum_{i=1}^m X_i\|\mathsf{R}^\ell_{\mathcal{D}}(h)\right)\right] \\
&\quad= \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{B}(\mathsf{R}^\ell_{\mathcal{D}}(h))^m} \left[\frac{\frac{1}{m}\sum_{i=1}^m X_i}{\mathsf{R}^\ell_{\mathcal{D}}(h)}\right]^{\sum_{i=1}^m X_i} \left[\frac{1-\frac{1}{m}\sum_{i=1}^m X_i}{1-\mathsf{R}^\ell_{\mathcal{D}}(h)}\right]^{m-\sum_{i=1}^m X_i} \\
&\quad= \sum_{k=0}^m \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{B}(\mathsf{R}^\ell_{\mathcal{D}}(h))^m}\left[\sum_{i=1}^m X_i = k\right] \left[\frac{\frac{k}{m}}{\mathsf{R}^\ell_{\mathcal{D}}(h)}\right]^k \left[\frac{1-\frac{k}{m}}{1-\mathsf{R}^\ell_{\mathcal{D}}(h)}\right]^{m-k} \\
&\quad= \sum_{k=0}^m \binom{m}{k} \left(1-\mathsf{R}^\ell_{\mathcal{D}}(h)\right)^{m-k} \left(\mathsf{R}^\ell_{\mathcal{D}}(h)\right)^k \left[\frac{\frac{k}{m}}{\mathsf{R}^\ell_{\mathcal{D}}(h)}\right]^k \left[\frac{1-\frac{k}{m}}{1-\mathsf{R}^\ell_{\mathcal{D}}(h)}\right]^{m-k} \\
&\quad= \sum_{k=0}^m \binom{m}{k} \left[\frac{k}{m}\right]^k \left[1-\frac{k}{m}\right]^{m-k}.
\end{aligned}$$

Finally, MAURER (2004) proves that $\sum_{k=0}^m \binom{m}{k}\left[\frac{k}{m}\right]^k\left[1-\frac{k}{m}\right]^{m-k} \le 2\sqrt{m}$ for $m \ge 8$ and GERMAIN *et al.* (2015) verify computationally that the inequality holds also for $m \in \{1,\ldots,7\}$. ■

We can prove Lemma B.16.2.

Proof of Lemma B.16.2. We have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m} \exp\left(m\left[F(\mathsf{R}^\ell_{\mathcal{D}}(h)) - c\mathsf{R}^\ell_{\mathcal{S}}(h)\right]\right) \\
&\qquad\le \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{B}(\mathsf{R}^\ell_{\mathcal{D}}(h))^m} \exp\left(mF(\mathsf{R}^\ell_{\mathcal{D}}(h)) - c\left(\sum_{i=1}^m X_i\right)\right) \qquad \text{(B.8)} \\
&\qquad= \sum_{k=0}^m \mathop{\mathbb{P}}_{\mathbf{X}\sim\mathcal{B}(\mathsf{R}^\ell_{\mathcal{D}}(h))^m}\left[\sum_{i=1}^m X_i = k\right] \exp\left(mF(\mathsf{R}^\ell_{\mathcal{D}}(h)) - ck\right) \\
&\qquad= \sum_{k=0}^m \binom{m}{k} \mathsf{R}^\ell_{\mathcal{D}}(h)^k (1-\mathsf{R}^\ell_{\mathcal{D}}(h))^{m-k} \exp\left(mF(\mathsf{R}^\ell_{\mathcal{D}}(h)) - ck\right) \\
&\qquad= e^{mF(\mathsf{R}^\ell_{\mathcal{D}}(h))} \sum_{k=0}^m \binom{m}{k} (\mathsf{R}^\ell_{\mathcal{D}}(h)e^{-c})^k (1-\mathsf{R}^\ell_{\mathcal{D}}(h))^{m-k} \\
&\qquad= e^{mF(\mathsf{R}^\ell_{\mathcal{D}}(h))} \left(1-\mathsf{R}^\ell_{\mathcal{D}}(h)\left[1-e^{-c}\right]\right)^m \qquad \text{(B.9)}
\end{aligned}$$

$$= \left(1 - \mathsf{R}^{\ell}_{\mathcal{D}}(h)\left[1 - e^{-c}\right]\right)^{-m}\left(1 - \mathsf{R}^{\ell}_{\mathcal{D}}(h)\left[1 - e^{-c}\right]\right)^{m} \tag{B.10}$$
$$= 1,$$

where we apply Lemma B.16.4 to obtain Equation (B.8), with the Binomial theorem we have Equation (B.9), and we deduce Equation (B.10) by definition of $F(\mathsf{R}^{\ell}_{\mathcal{D}}(h))$. ■

Appendix of Chapter 3

C

C.1 Proof of Proposition 3.3.1

Proposition 3.3.1 (Relations Between the Averaged Adversarial Risks)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any distribution ρ on $\mathbb{H}$, for any $(n, n') \in \mathbb{N}^2$, with $1 \leq n' \leq n$, we have

$$\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) \;\leq\; \mathsf{A}_{\mathcal{E}^{n'}}(\mathsf{MV}_\rho) \;\leq\; \mathsf{A}_{\mathcal{E}^{n}}(\mathsf{MV}_\rho) \;\leq\; \mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho). \tag{3.2}$$

Proof. First, we prove $\mathsf{A}_{\mathcal{E}^1}(\mathsf{MV}_\rho)$=$\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho)$. We have

$$\begin{aligned}
\mathsf{A}_{\mathcal{E}^1}(\mathsf{MV}_\rho) &= 1 - \mathop{\mathbb{P}}_{((\mathbf{x},y),\mathfrak{C})\sim\mathcal{E}^1} (\forall \epsilon \in \mathfrak{C}, \mathsf{MV}_\rho(\mathbf{x}+\epsilon) = y) \\
&= 1 - \mathop{\mathbb{P}}_{((\mathbf{x},y),\mathfrak{C})\sim\mathcal{E}^1} (\forall \epsilon \in \{\epsilon_1\}, \mathsf{MV}_\rho(\mathbf{x}+\epsilon) = y) \\
&= 1 - \mathop{\mathbb{P}}_{((\mathbf{x},y),\mathfrak{C})\sim\mathcal{E}^1} (\mathsf{MV}_\rho(\mathbf{x}+\epsilon_1) = y) = \mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho).
\end{aligned}$$

Then, we prove the inequality $\mathsf{A}_{\mathcal{E}^{n'}}(\mathsf{MV}_\rho) \leq \mathsf{A}_{\mathcal{E}^{n}}(\mathsf{MV}_\rho)$ from the fact that the indicator function $\mathrm{I}[\cdot]$ is upper-bounded by 1. Indeed, from Definition 3.2.3 we have

$$\begin{aligned}
1 - \mathsf{A}_{\mathcal{E}^n}(\mathsf{MV}_\rho) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{\mathfrak{C}\sim\mathcal{B}^n_{(x,y)}} \mathrm{I}[\forall \epsilon \in \mathfrak{C}, \mathsf{MV}_\rho(\mathbf{x}+\epsilon) = y] \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\prod_{i=1}^{n} \mathop{\mathbb{E}}_{\epsilon_i\sim\mathcal{B}_{(x,y)}} \mathrm{I}[\mathsf{MV}_\rho(\mathbf{x}+\epsilon_i) = y] \right] \\
&\leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\prod_{i=1}^{n'} \mathop{\mathbb{E}}_{\epsilon_i\sim\mathcal{B}_{(x,y)}} \mathrm{I}[\mathsf{MV}_\rho(\mathbf{x}+\epsilon_i) = y] \right] \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{\mathfrak{C}'\sim\mathcal{B}^{n'}_{(x,y)}} \mathrm{I}[\forall \epsilon \in \mathfrak{C}', \mathsf{MV}_\rho(\mathbf{x}+\epsilon) = y] \\
&= 1 - \mathsf{A}_{\mathcal{E}^{n'}}(\mathsf{MV}_\rho).
\end{aligned}$$

Lastly, to prove the right-most inequality, we have to use the fact that the expectation over the set $\mathbb{B}$ is bounded by the maximum over the set $\mathbb{B}$. We have

$$\mathsf{A}_{\mathcal{E}^n}(\mathsf{MV}_\rho) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{\epsilon_1\sim\mathcal{B}_{(x,y)}} \cdots \mathop{\mathbb{E}}_{\epsilon_n\sim\mathcal{B}_{(x,y)}} \mathrm{I}[\exists \epsilon \in \{\epsilon_1, \ldots, \epsilon_n\}, \mathsf{MV}_\rho(\mathbf{x}+\epsilon) \neq y]$$

$$
\begin{aligned}
&\leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \max_{\epsilon_1\in\mathbb{B}} \dots \max_{\epsilon_n\in\mathbb{B}} \mathrm{I}\left[\exists \epsilon \in \{\epsilon_1,\dots,\epsilon_n\}, \mathsf{MV}_\rho(\mathbf{x}+\epsilon) \neq y\right] \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \max_{\epsilon_1\in\mathbb{B}} \dots \max_{\epsilon_{n-1}\in\mathbb{B}} \mathrm{I}\left[\exists \epsilon \in \{\epsilon_1,\dots,\epsilon^*\}, \mathsf{MV}_\rho(\mathbf{x}+\epsilon) \neq y\right] \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}+\epsilon^*) \neq y\right] \\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \max_{\epsilon\in\mathbb{B}} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}+\epsilon) \neq y\right] \;=\; \mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho).
\end{aligned}
$$

Merging the three equations proves the claim. ■

C.2 Proof of Proposition 3.3.2

In this section, we provide the proof of Proposition 3.3.2 that relies on Lemmas C.2.1 and C.2.2 which are also described and proved. Lemma C.2.1 shows that $\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho)$ equals $\mathsf{R}_{\Gamma}(\mathsf{MV}_\rho)$.

Lemma C.2.1. For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$ and its associated distribution Γ, for any posterior ρ on $\mathbb{H}$, we have

$$
\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) = \Pr_{(\mathbf{x}+\epsilon,y)\sim\Gamma}\left[\mathsf{MV}_\rho(\mathbf{x}+\epsilon)\neq y\right] = \mathsf{R}_{\Gamma}(\mathsf{MV}_\rho).
$$

Proof. Starting from the averaged risk $\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) = \mathbb{E}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}+\epsilon)\neq y\right]$, we have

$$
\begin{aligned}
\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) &= \mathop{\mathbb{E}}_{(\mathbf{x}'+\epsilon',y')\sim\Gamma} \frac{1}{\Gamma(\mathbf{x}'+\epsilon',y')} \left[\Pr_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}} \left[\mathsf{MV}_\rho(\mathbf{x}+\epsilon)\neq y, \mathbf{x}'+\epsilon'=\mathbf{x}+\epsilon, y'=y\right]\right] \\
&= \mathop{\mathbb{E}}_{(\mathbf{x}'+\epsilon',y')\sim\Gamma} \frac{1}{\Gamma(\mathbf{x}'+\epsilon',y')} \left[\mathop{\mathbb{E}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}} \mathrm{I}[\mathsf{MV}_\rho(\mathbf{x}+\epsilon)\neq y]\, \mathrm{I}[\mathbf{x}'+\epsilon'=\mathbf{x}+\epsilon, y'=y]\right].
\end{aligned}
$$

In other words, the double expectation only rearranges the terms of the original expectation: given an example $(\mathbf{x}'+\epsilon',y')$, we gather probabilities such that $\mathsf{MV}_\rho(\mathbf{x}+\epsilon)\neq y$ with $(\mathbf{x}+\epsilon,y)=(\mathbf{x}'+\epsilon',y')$ in the inner expectation, while integrating over all couple $(\mathbf{x}'+\epsilon', y') \in \mathbb{X}\times\mathbb{Y}$ in the outer expectation. Then, from the fact

that when $\mathbf{x}'+\epsilon'=\mathbf{x}+\epsilon$ and $y'=y$, $\mathrm{I}[\mathsf{MV}_\rho(\mathbf{x}+\epsilon)\neq y] = \mathrm{I}[\mathsf{MV}_\rho(\mathbf{x}'+\epsilon')\neq y']$, we have

$$\begin{aligned}\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) &= \mathop{\mathbb{E}}_{(\mathbf{x}'+\epsilon',y')\sim\Gamma} \frac{1}{\Gamma(\mathbf{x}'+\epsilon',y')} \left[\mathop{\mathbb{E}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}} \mathrm{I}[\mathsf{MV}_\rho(\mathbf{x}'+\epsilon')\neq y'] \mathrm{I}[\mathbf{x}'+\epsilon'=\mathbf{x}+\epsilon, y'=y] \right] \\ &= \mathop{\mathbb{E}}_{(\mathbf{x}'+\epsilon',y')\sim\Gamma} \frac{1}{\Gamma(\mathbf{x}'+\epsilon',y')} \left[\mathrm{I}[\mathsf{MV}_\rho(\mathbf{x}'+\epsilon')\neq y'] \mathop{\mathbb{E}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}} \mathrm{I}[\mathbf{x}'+\epsilon'=\mathbf{x}+\epsilon, y'=y] \right].\end{aligned}$$

Finally, by definition of $\Gamma(\mathbf{x}'+\epsilon',y')$, we can deduce that

$$\begin{aligned}\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) &= \mathop{\mathbb{E}}_{(\mathbf{x}'+\epsilon',y')\sim\Gamma} \frac{1}{\Gamma(\mathbf{x}'+\epsilon',y')} [\mathrm{I}[\mathsf{MV}_\rho(\mathbf{x}'+\epsilon')\neq y'] \, \Gamma(\mathbf{x}'+\epsilon',y')] \\ &= \mathop{\mathbb{E}}_{(\mathbf{x}'+\epsilon',y')\sim\Gamma} \mathrm{I}[\mathsf{MV}_\rho(\mathbf{x}'+\epsilon')\neq y'] = \mathsf{R}_\Gamma(\mathsf{MV}_\rho).\end{aligned}$$

■

Similarly, Lemma C.2.2 shows that $\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho)$ is equivalent to $\mathsf{R}_\gamma(\mathsf{MV}_\rho)$.

Lemma C.2.2. For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$ and its associated distribution γ, for any posterior ρ on $\mathbb{H}$, we have

$$\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho) = \Pr_{(\mathbf{x}+\epsilon,y)\sim\gamma}[\mathsf{MV}_\rho(\mathbf{x}+\epsilon)\neq y] = \mathsf{R}_\gamma(\mathsf{MV}_\rho).$$

Proof. The proof is similar to the one of Lemma C.2.1. Indeed, starting from the definition of $\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}[\mathsf{MV}_\rho(\mathbf{x}+\epsilon^*(\mathbf{x},y)) \neq y]$, we have

$$\begin{aligned}&\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho) \\ &= \mathop{\mathbb{E}}_{(\mathbf{x}'+\epsilon',y')\sim\gamma} \frac{1}{\gamma(\mathbf{x}'+\epsilon',y')} \left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}+\epsilon^*(\mathbf{x},y)) \neq y\right] \mathrm{I}[\mathbf{x}'+\epsilon'=\mathbf{x}+\epsilon^*(\mathbf{x},y), y'=y] \right] \\ &= \mathop{\mathbb{E}}_{(\mathbf{x}'+\epsilon',y')\sim\gamma} \frac{1}{\gamma(\mathbf{x}'+\epsilon',y')} \left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}'+\epsilon') \neq y'\right] \mathrm{I}[\mathbf{x}'+\epsilon'=\mathbf{x}+\epsilon^*(\mathbf{x},y), y'=y] \right].\end{aligned}$$

Finally, by definition of $\gamma(\mathbf{x}'+\epsilon', y')$, we can deduce that

$$\begin{aligned}\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho) &= \mathop{\mathbb{E}}_{(\mathbf{x}'+\epsilon',y')\sim\gamma} \frac{1}{\gamma(\mathbf{x}'+\epsilon',y')} \left[\mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}'+\epsilon') \neq y'\right] \gamma(\mathbf{x}'+\epsilon',y')\right] \\ &= \mathop{\mathbb{E}}_{(\mathbf{x}'+\epsilon',y')\sim\gamma} \mathrm{I}\left[\mathsf{MV}_\rho(\mathbf{x}'+\epsilon')\neq y'\right] = \mathsf{R}_\gamma(\mathsf{MV}_\rho).\end{aligned}$$

■

We can now prove Proposition 3.3.2.

Proposition 3.3.2 (Classical and Averaged Adversarial Risks)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X} \times \mathbb{Y}) \times \mathbb{B}$, for any distribution ρ on $\mathbb{H}$, we have

$$\mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho) - \mathrm{TV}(\gamma \| \Gamma) \;\leq\; \mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho),$$

where Γ and γ are distributions on $\mathbb{X}\times\mathbb{Y}$ and $\mathrm{TV}(\gamma\|\Gamma) = \mathbb{E}_{(\mathbf{x}',y')\sim\Gamma} \frac{1}{2}\left|\frac{\gamma(\mathbf{x}',y')}{\Gamma(\mathbf{x}',y')} - 1\right|$, is the Total Variation (TV) distance between γ and Γ.
The density $\Gamma(\mathbf{x}', y')$ corresponds to the probability of drawing a perturbed example $(\mathbf{x}', y') = (\mathbf{x}+\epsilon, y)$ with $((\mathbf{x}, y), \epsilon)\sim\mathcal{E}$, *i.e.*, we have

$$\Gamma(\mathbf{x}', y') = \Pr_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}} [\mathbf{x}+\epsilon = \mathbf{x}',\ y = y'] .$$

The density $\gamma(\mathbf{x}', y')$ is the probability to draw an adversarial example $(\mathbf{x}', y') = (\mathbf{x}+\epsilon^*(\mathbf{x},y), y)$ with $(\mathbf{x}, y)\sim\mathcal{D}$, *i.e.*, we have

$$\gamma(\mathbf{x}', y') = \Pr_{(\mathbf{x},y)\sim\mathcal{D}} [\mathbf{x}+\epsilon^*(\mathbf{x}, y) = \mathbf{x}',\ y = y'] .$$

Proof. From Lemmas C.2.1 and C.2.2, we have

$$\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) = \mathsf{R}_{\Gamma}(\mathsf{MV}_\rho), \quad \text{and} \quad \mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho) = \mathsf{R}_{\gamma}(\mathsf{MV}_\rho).$$

Then, we apply Lemma 4 of Ohnishi and Honorio (2021), we have

$$\mathsf{R}_{\gamma}(\mathsf{MV}_\rho) \leq \mathrm{TV}(\gamma\|\Gamma) + \mathsf{R}_{\Gamma}(\mathsf{MV}_\rho) \quad \Longleftrightarrow \quad \mathsf{A}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq \mathrm{TV}(\gamma\|\Gamma) + \mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho).$$

■

C.3 Proof of Theorem 3.3.1

Theorem 3.3.1 (Upper Bounds on the Surrogates)**.** For any distributions $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$ and ρ on $\mathbb{H}$, for any $n>1$, we have

$$\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) \leq 2r_{\mathcal{E}}(\rho), \qquad \text{and} \qquad \mathsf{A}_{\mathcal{E}^n}(\mathsf{MV}_\rho) \leq 2a_{\mathcal{E}^n}(\rho).$$

Proof. By the definition of the majority vote and from Markov's inequality (The-

orem A.2.1), we have

$$\begin{aligned}\frac{1}{2}\mathsf{R}_{\mathcal{E}}(\mathsf{MV}_\rho) &= \frac{1}{2}\mathop{\mathbb{P}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}}\left(y\mathop{\mathbb{E}}_{h\sim\rho}h(\mathbf{x}+\epsilon)\leq 0\right)\\ &= \frac{1}{2}\mathop{\mathbb{P}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}}\left(1-y\mathop{\mathbb{E}}_{h\sim\rho}h(\mathbf{x}+\epsilon)\geq 1\right)\\ &\leq \mathop{\mathbb{E}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}}\frac{1}{2}\Big[1-y\mathop{\mathbb{E}}_{h\sim\rho}h(\mathbf{x}+\epsilon)\Big]\\ &= r_{\mathcal{E}}(\rho).\end{aligned}$$

Similarly we have

$$\begin{aligned}\frac{1}{2}\mathsf{A}_{\mathcal{E}^n}(\mathsf{MV}_\rho) &= \frac{1}{2}\mathop{\mathbb{P}}_{((\mathbf{x},y),\mathfrak{C})\sim\mathcal{E}^n}\left(\exists\epsilon\in\mathfrak{C}, y\mathop{\mathbb{E}}_{h\sim\rho}h(\mathbf{x}+\epsilon)\leq 0\right)\\ &= \frac{1}{2}\mathop{\mathbb{P}}_{((\mathbf{x},y),\mathfrak{C})\sim\mathcal{E}^n}\left(\min_{\epsilon\in\mathfrak{C}}\left(y\mathop{\mathbb{E}}_{h\sim\rho}h(\mathbf{x}+\epsilon)\right)\leq 0\right)\\ &= \frac{1}{2}\mathop{\mathbb{P}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}}\left(1-\min_{\epsilon\in\mathfrak{C}}\left(y\mathop{\mathbb{E}}_{h\sim\rho}h(\mathbf{x}+\epsilon)\right)\geq 1\right)\\ &\leq \mathop{\mathbb{E}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}}\frac{1}{2}\Big[1-\min_{\epsilon\in\mathfrak{C}}\left(y\mathop{\mathbb{E}}_{h\sim\rho}h(\mathbf{x}+\epsilon)\right)\Big]\\ &= a_{\mathcal{E}^n}(\rho).\end{aligned}$$

■

C.4 Proof of Theorem 3.3.2

Theorem 3.3.2 (PAC-Bayesian Bound on $r_{\mathcal{E}}(\rho)$)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any set of voters $\mathbb{H}$, for any prior $\pi\in\mathbb{M}*(\mathbb{H})$ on $\mathbb{H}$, for any $n\in\mathbb{N}_*$, with probability at least $1-\delta$ over $\widehat{\mathbb{S}}\sim(\mathcal{E}^n)^m$, for all posteriors $\rho\in\mathbb{M}(\mathbb{H})$ on $\mathbb{H}$, we have

$$\mathrm{kl}(r_{\widehat{\mathbb{S}}}(\rho)\|r_{\mathcal{E}}(\rho))\leq\frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{m+1}{\delta}\right], \tag{3.6}$$

and

$$r_{\mathcal{E}}(\rho)\leq r_{\widehat{\mathbb{S}}}(\rho)+\sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{m+1}{\delta}\right]}, \tag{3.7}$$

where

$$r_{\widehat{\mathbb{S}}}(\rho)=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{2}\left[1-y_i\mathop{\mathbb{E}}_{h\sim\rho}h(\mathbf{x}_i+\epsilon_j^i)\right].$$

Proof. Let $G{=}(V,E)$ be the graph representing the dependencies between the random variables where *(i)* the set of vertices is $V{=}\widehat{\mathbb{S}}$, *(ii)* the set of edges E is defined such that $(((\mathbf{x},y),\epsilon),((\mathbf{x}',y'),\epsilon')) \notin E \Leftrightarrow x \neq \mathbf{x}'$. Then, applying Theorem 8 of Ralaivola *et al.* (2010) with our notations gives

$$\mathrm{kl}(r_{\widehat{\mathbb{S}}}(\rho)\|r_{\mathcal{E}}(\rho)) \leq \frac{\chi(G)}{mn}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{mn+\chi(G)}{\delta\chi(G)}\right],$$

where $\chi(G)$ is the fractional chromatic number of G. From a property of Scheinerman and Ullman (2011), we have

$$c(G) \leq \chi(G) \leq \Delta(G)+1,$$

where $c(G)$ is the order of the largest clique in G and $\Delta(G)$ is the maximum degree of a vertex in G. By construction of G, $c(G){=}n$ and $\Delta(G){=}n{-}1$. Thus, $\chi(G){=}n$ and rearranging the terms proves Equation (3.6). Finally, by applying Pinsker's inequality (*i.e.*, $|a{-}b|{\leq}\sqrt{\frac{1}{2}\,\mathrm{kl}(a\|b)}$), we obtain Equation (3.7). ■

C.5 Proof of Theorem 3.3.3

Theorem 3.3.3 (PAC-Bayesian Bound on $a_{\mathcal{E}^n}(\rho)$)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}{\times}\mathbb{Y}){\times}\mathbb{B}$, for any set of voters $\mathbb{H}$, for any prior $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any $n \in \mathbb{N}_*$, with probability at least $1{-}\delta$ over $\widehat{\mathbb{S}} \sim (\mathcal{E}^n)^m$, for all posteriors $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$, for all $i \in \{1,\ldots,m\}$, for all distributions Θ_i on $\mathfrak{C}_i$ independent from a voter $h \in \mathbb{H}$, we have

$$a_{\mathcal{E}^n}(\rho) \leq \frac{1}{m}\mathop{\mathbb{E}}_{h\sim\rho}\sum_{i=1}^{m}\max_{\epsilon\in\mathfrak{C}_i}\frac{1}{2}\left(1{-}y_i h(\mathbf{x}_i{+}\epsilon)\right) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}}{\delta}\right]} \quad (3.8)$$

$$\leq a_{\widehat{\mathbb{S}}}(\rho) + \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim\rho}\mathrm{TV}(\theta_i^h\|\Theta_i) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}}{\delta}\right]}, \quad (3.9)$$

where the empirical risk $a_{\widehat{\mathbb{S}}}(\rho) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left[1{-}\min_{\epsilon\in\mathfrak{C}_i}\left(y_i\,\mathbb{E}_{h\sim\rho}\,h(\mathbf{x}_i{+}\epsilon)\right)\right]$, and the TV distance $\mathrm{TV}(\theta\|\Theta) = \mathbb{E}_{\epsilon\sim\Theta}\,\frac{1}{2}\left|\left[\frac{\theta(\epsilon)}{\Theta(\epsilon)}\right]-1\right|$.

Proof. Let $L_{h,(\mathbf{x},y),\epsilon}{=}\frac{1}{2}\left[1{-}yh(\mathbf{x}+\epsilon)\right]$ for the sake of readability. Given $h \in \mathbb{H}$, the losses $\max_{\epsilon\in\mathfrak{C}_1}L_{h,(\mathbf{x}_1,y_1),\epsilon}$, … $\max_{\epsilon\in\mathfrak{C}_1}L_{h,(\mathbf{x}_m,y_m),\epsilon}$ are *i.i.d.*. Hence, we can

apply Theorem 20 of Germain *et al.* (2015) and Pinsker's inequality, *i.e.*, the inequality $|q-p| \leq \sqrt{\frac{1}{2}\operatorname{kl}(q\|p)}$ (Theorem B.5.1) to obtain

$$\mathop{\mathbb{E}}_{h\sim\rho}\mathop{\mathbb{E}}_{(\mathbf{x},y),\mathfrak{C})\sim\mathcal{E}^n}\max_{\epsilon\in\mathfrak{C}} L_{h,(\mathbf{x},y),\epsilon} \leq \mathop{\mathbb{E}}_{h\sim\rho}\frac{1}{m}\sum_{i=1}^{m}\max_{\epsilon\in\mathfrak{C}_i} L_{h,(\mathbf{x}_i,y_i),\epsilon} + \sqrt{\frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{m}}{\delta}}{2m}}.$$

Then, we lower-bound the left-hand side of the inequality with $a_{\mathcal{E}^n}(\rho)$, we have

$$a_{\mathcal{E}^n}(\rho) \leq \mathop{\mathbb{E}}_{h\sim\rho}\mathop{\mathbb{E}}_{((\mathbf{x},y),\mathfrak{C})\sim\mathcal{E}^n}\max_{\epsilon\in\mathfrak{C}} L_{h,(\mathbf{x},y),\epsilon}.$$

Finally, from the definition of θ_i^h, and from Lemma 4 of Ohnishi and Honorio (2021), we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\rho}\frac{1}{m}\sum_{i=1}^{m}\max_{\epsilon\in\mathfrak{C}_i} L_{h,(\mathbf{x}_i,y_i),\epsilon} &= \mathop{\mathbb{E}}_{h\sim\rho}\frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\epsilon\sim\theta_i^h} L_{h,(\mathbf{x}_i,y_i),\epsilon} \\
&\leq \mathop{\mathbb{E}}_{h\sim\rho}\frac{1}{m}\sum_{i=1}^{m}\mathrm{TV}(\theta_i^h\|\Theta_i) + \mathop{\mathbb{E}}_{h\sim\rho}\frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\epsilon\sim\Theta_i} L_{h,(\mathbf{x}_i,y_i),\epsilon} \\
&= \mathop{\mathbb{E}}_{h\sim\rho}\frac{1}{m}\sum_{i=1}^{m}\mathrm{TV}(\theta_i^h\|\Theta_i) + \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\epsilon\sim\Theta_i}\mathop{\mathbb{E}}_{h\sim\rho} L_{h,(\mathbf{x}_i,y_i),\epsilon} \\
&\leq \mathop{\mathbb{E}}_{h\sim\rho}\frac{1}{m}\sum_{i=1}^{m}\mathrm{TV}(\theta_i^h\|\Theta_i) + a_{\widehat{\mathbb{S}}}(\rho).
\end{aligned}$$

■

C.6 Proof of Corollaries 3.3.1 and 3.3.2

We start to prove Corollary 3.3.1.

Corollary 3.3.1 (PAC-Bayesian Bound on $r_{\mathcal{E}}(\rho)$)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any set of voters $\mathbb{H}$, for any $T \in \mathbb{N}_*$, for any priors' set $\{\pi_1,\ldots,\pi_T\} \in \mathbb{M}^*(\mathbb{H})^T$, for any $n \in \mathbb{N}_*$, with probability at least $1-\delta$ over $\widehat{\mathbb{S}} \sim (\mathcal{E}^n)^m$, for all posteriors $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$, for any $\pi \in \{\pi_1,\ldots,\pi_T\} \in \mathbb{M}^*(\mathbb{H})^T$ we have

$$r_{\mathcal{E}}(\rho) \leq \overline{\operatorname{kl}}\left(r_{\widehat{\mathbb{S}}}(\rho)\,\middle|\,\frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{T(m+1)}{\delta}\right]\right). \tag{3.10}$$

Proof. Let $\mathcal{E}_1, \ldots, \mathcal{E}_T$ be T distributions defined as $\mathcal{E}_1 = \mathcal{D}(\mathbf{x}, y)\mathcal{B}^1_{(x,y)}(\epsilon)$, $\ldots$, $\mathcal{E}_T = \mathcal{D}(\mathbf{x}, y)\mathcal{B}^T_{(x,y)}(\epsilon)$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$ where each distribution $\mathcal{B}^t_{(x,y)}$ depends on the example $(\mathbf{x}, y)$ and possibly on the fixed prior π_t. Then, for all distributions $\mathcal{E}_t$, we can derive a bound on the risk $r_{\mathcal{E}_t}(\rho)$ which holds with probability at least $1 - \frac{\delta}{T}$, we have

$$\Pr_{\widehat{\mathbb{S}}_t\sim(\mathcal{E}_t^n)^m}\left[\forall\rho,\ \mathrm{kl}(r_{\widehat{\mathbb{S}}_t}(\rho)\|r_{\mathcal{E}_t}(\rho)) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi_t) + \ln\frac{T(m+1)}{\delta}\right]\right] \geq 1-\tfrac{\delta}{T}.$$

Then, from a union bound argument, we have

$$\Pr_{\widehat{\mathbb{S}}_1\sim(\mathcal{E}_1^n)^m,\ldots,\widehat{\mathbb{S}}_T\sim(\mathcal{E}_T^n)^m}\left[\forall\rho,\ \mathrm{kl}(r_{\widehat{\mathbb{S}}_1}(\rho)\|r_{\mathcal{E}_1}(\rho)) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi_t) + \ln\frac{T(m+1)}{\delta}\right],\right.$$
$$\left.\ldots,\ \text{and}\ \mathrm{kl}(r_{\widehat{\mathbb{S}}_T}(\rho)\|r_{\mathcal{E}_T}(\rho)) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi_T) + \ln\frac{T(m+1)}{\delta}\right]\right] \geq 1-\delta.$$

Hence, we have

$$\mathrm{kl}\left(r_{\widehat{\mathbb{S}}}(\rho)\|r_{\mathcal{E}}(\rho)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{T(m+1)}{\delta}\right],$$

where $\mathcal{B}_{(x,y)}$ can be dependent on the selected prior π. From Definition 2.3.3, we can obtain the claimed result. ■

We can prove Corollary 3.3.2 similarly to Corollary 3.3.1.

Corollary 3.3.2 (PAC-Bayesian Bound on $a_{\mathcal{E}^n}(\rho)$)**.** For any distribution $\mathcal{E}$ on $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$, for any set of voters $\mathbb{H}$, for any prior π on $\mathbb{H}$, for any $n \in \mathbb{N}_*$, with probability at least $1-\delta$ over $\widehat{\mathbb{S}} \sim (\mathcal{E}^n)^m$, for all posteriors $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$, for all $i \in \{1, \ldots, m\}$, for all distributions Θ_i on $\mathfrak{C}_i$ independent from a voter $h \in \mathbb{H}$, we have

$$a_{\mathcal{E}^n}(\rho) \leq a_{\widehat{\mathbb{S}}}(\rho) + \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim\rho}\mathrm{TV}(\theta_i^h\|\Theta_i) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2T\sqrt{m}}{\delta}\right]}. \qquad (3.11)$$

Proof. From a union bound argument, we obtain the claimed result. ■

C.7 About the (Differentiable) Decision Trees

In this section, we introduce the differentiable decision trees, *i.e.*, the voters of our majority vote. Note that we adapt the model of KONTSCHIEDER *et al.* (2016) in order to fit with our framework: a voter must output a real between -1 and $+1$. An example of such a tree is represented in Figure C.1.

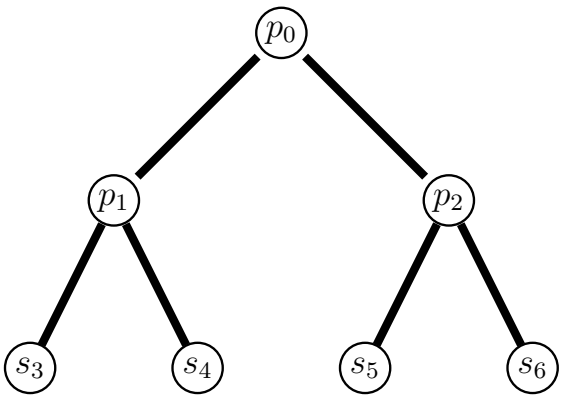


Figure C.1. *Representation of a (differentiable) decision tree of depth $l = 2$; The root is the node 0 and the leafs are 4; 5; 6 and 7. The probability $p_i(\mathbf{x})$ (respectively $1-p_i(\mathbf{x})$) to go left (respectively right) at the node i is represented by p_i (we omitted the dependence on $\mathbf{x}$ for simplicity). Similarly, the predicted label (a "score" between -1 and $+1$) at the leaf i is represented by s_i.*

This differentiable decision tree is stochastic by nature: at each node i of the tree, we continue recursively to the left sub-tree with a probability of $p_i(\mathbf{x})$ and to the right sub-tree with a probability of $1-p_i(\mathbf{x})$; When we attain a leaf j, the tree predicts the label s_j. Precisely, the probability $p_i(\mathbf{x})$ is constructed by *(i)* selecting randomly 50% of the input features $\mathbf{x}$ and applying a random mask $M_i \in \mathbb{R}^d$ on $\mathbf{x}$ (where the k-th entry of the mask is 1 if the k-th feature is selected and 0 otherwise), by *(ii)* multiplying this quantity by a learned weight vector $v_i \in \mathbb{R}^d$, and by *(iii)* applying a sigmoid function to output a probability. Indeed, we have

$$p_i(\mathbf{x}) = \sigma\Big(\langle v_i, M_i \odot x \rangle\Big),$$

where $\sigma(a) = [1 + e^{-a}]^{-1}$ is the sigmoid function; $\langle a, b \rangle$ is the dot product between the vector a and b and $a \odot b$ is the elementwise product between the vector a and b. Moreover, s_i is obtained by learning a parameter $u_i \in \mathbb{R}$ and applying a $\tanh$ function, *i.e.*, we have

$$s_i = \tanh\Big(u_i\Big).$$

Finally, instead of having a stochastic voter, h will output the expected label predicted by the tree (see KONTSCHIEDER *et al.* (2016) for more details). It can be computed by $h(\mathbf{x}) = f(\mathbf{x}, 0, 0)$ with

$$f(\mathbf{x}, i, l') = \begin{cases} s_i & \text{if } l' = l \\ p_i(\mathbf{x}) f(\mathbf{x}, 2i{+}1, l'{+}1) + (1 - p_i(\mathbf{x})) f(\mathbf{x}, 2i{+}2, l'{+}1) & \text{otherwise} \end{cases} .$$

C.8 Additional Experimental Results

In this section, we present the detailed results for the 6 tasks (3 on MNIST and 3 on Fashion MNIST) on which we perform experiments that show the test risks and the bounds for the different scenarios of (Defense, Attack). We train all the models using the same parameters as described in Section 3.4.1. Table C.1 and Appendix C.8 complement Table 3.1 to present the results for all the tasks when using the ℓ_2-norm with $b = 1$ (the maximum noise allowed by the norm). Then, we run again the same experiment but we use the ℓ_∞-norm with $b = 0.1$ and exhibit the results in Appendix C.8 and Table C.6. For the experiments on the 5 other tasks using the ℓ_2-norm, we have a similar behavior than MNIST:1vs7. Indeed, using the attacks PGD_U and IFGSM_U as defense mechanism allows to obtain better risks and also tighter bounds compared to the bounds obtained with a defense based on UNIF (which is a naive defense). For the experiments on the 6 tasks using the ℓ_∞-norm, the trend is the same as with the ℓ_2-norm, *i.e.*, the appropriate defense leads to better risks and bounds.

We also run experiments that do not rely on the PAC-Bayesian framework. In other words, we train the models following only Step 1 of our adversarial training procedure (*i.e.*, Algorithm 3.1) using classical attacks (PGD or IFGSM): we refer to this experiment as a baseline. In our cases, it means learning a majority vote $\text{MV}_{\pi'}$ that follows a distribution π'. As a reminder, the studied scenarios for the baseline are all the pairs $(\text{Defense}, \text{Attack})$ belonging to the set $\{—, \text{UNIF}, \text{PGD}, \text{IFGSM}\} \times \{—, \text{PGD}, \text{IFGSM}\}$. We report the results in Table C.8 and Table C.9. With this experiment, we are now able to compare our defense based on PGD_U or IFGSM_U and a classical defense based on PGD and IFGSM. Hence, considering the test risks $\text{A}_\mathcal{T}(\text{MV}_\rho)$ (columns "Attack without U" of Tables 3.1 to C.6) and $\text{A}_\mathcal{T}(\text{MV}_{\pi'})$ (in Tables C.8 and C.9) , we observe similar results between the baseline and our framework.

Table C.1. *Test risks and bounds for 2 tasks of MNIST with n=100 perturbations for all pairs (Defense,Attack) with the two voters' set $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$. The results in* ***bold*** *correspond to the best values between results for $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$. To quantify the gap between our risks and the classical definition we put in italic the risk of our models against the classical attacks: we replace PGD_U and IFGSM_U by PGD or IFGSM (i.e., we did not sample from the uniform distribution). Since Eq.* (3.9) *upperbounds Eq.* (3.8) *thanks to the TV term, we compute the two bound values of Theorem 3.3.3.*

ℓ_2-norm		Algo. 3.1 with Eq. (3.6)						Algo. 3.1 with Eq. (3.9)							
$b=1$		Attack without U						Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.015*	*.015*	.015	.015	**0.060**	.067	*.015*	*.015*	.015	.015	**0.129**	0.135	**0.129**	.135
—	PGD_U	*.632*	***.628***	**.520**	.526	1.059	**.847**	*.672*	***.641***	**.683**	.684	**1.718**	2.405	1.392	**.962**
—	IFGSM_U	*.447*	***.443***	**.157**	.166	**0.387**	.572	*.461*	***.451***	**.337**	.345	**1.137**	2.090	0.776	**.669**
UNIF	—	*.024*	*.024*	.024	.024	**0.073**	.083	*.024*	*.024*	.024	.024	**0.140**	0.148	**0.140**	.148
UNIF	PGD_U	*.646*	***.619***	**.486**	.500	1.016	**.809**	*.649*	***.626***	**.648**	.650	**1.646**	2.417	1.338	**.915**
UNIF	IFGSM_U	*.442*	*.442*	**.128**	.139	**0.316**	.528	*.442*	*.442*	**.281**	.293	**0.907**	2.118	0.633	**.617**
PGD_U	—	***.024***	*.025*	**.024**	.025	**0.094**	.101	***.024***	*.025*	**.024**	.025	**0.158**	0.163	**0.158**	.163
PGD_U	PGD_U	*.148*	***.135***	.111	**.103**	0.360	**.355**	*.146*	***.136***	.129	**.120**	**0.442**	2.062	0.414	**.403**
PGD_U	IFGSM_U	*.104*	***.103***	.072	.072	0.277	.277	*.102*	*.102*	.090	**.084**	**0.358**	1.954	0.335	**.328**
IFGSM_U	—	*.027*	***.025***	.027	**.025**	**0.080**	.091	*.027*	***.025***	.027	**.025**	**0.146**	0.154	**0.146**	.154
IFGSM_U	PGD_U	*.188*	***.178***	**.111**	.119	**0.383**	.405	*.190*	***.178***	**.126**	.134	**0.501**	2.063	0.454	.454
IFGSM_U	IFGSM_U	*.126*	***.115***	.076	**.070**	**0.248**	.290	*.127*	***.115***	.091	**.085**	**0.371**	1.918	**0.329**	.342

(a) *MNIST 4vs9*

ℓ_2-norm		Algo. 3.1 with Eq. (3.6)						Algo. 3.1 with Eq. (3.9)							
$b=1$		Attack without U						Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.015*	*.015*	.015	.015	**.043**	.045	*.015*	*.015*	.015	.015	**.117**	0.118	**.117**	.118
—	PGD_U	*.279*	***.271***	**.232**	.234	.600	**.453**	*.284*	***.274***	.284	.284	**.829**	1.929	.724	**.524**
—	IFGSM_U	*.143*	***.137***	**.089**	.090	**.204**	.227	*.144*	***.139***	**.125**	.127	**.422**	1.662	.337	**.293**
UNIF	—	*.017*	*.017*	.017	.017	**.054**	.055	*.017*	*.017*	.017	.017	**.124**	0.125	**.124**	.125
UNIF	PGD_U	*.219*	***.201***	**.172**	.177	.433	**.350**	*.219*	***.209***	**.217**	.218	**.671**	1.810	.565	**.419**
UNIF	IFGSM_U	*.122*	*.122*	**.052**	.055	**.119**	.181	***.122***	*.123*	**.077**	.082	**.307**	1.554	**.242**	.248
PGD_U	—	***.013***	*.015*	**.013**	.015	.061	.061	***.013***	*.015*	**.013**	.015	**.131**	0.130	.131	**.130**
PGD_U	PGD_U	*.057*	*.057*	.045	**.041**	**.157**	.160	*.057*	*.057*	.055	**.045**	**.227**	1.536	.218	.218
PGD_U	IFGSM_U	*.043*	*.043*	**.027**	.031	**.114**	.119	***.042***	*.043*	.037	**.035**	**.187**	1.433	**.179**	.181
IFGSM_U	—	*.014*	***.012***	.014	**.012**	.057	.057	*.014*	***.013***	.014	**.013**	.128	**0.127**	.128	**.127**
IFGSM_U	PGD_U	*.077*	***.072***	.054	**.043**	**.170**	.174	*.076*	***.075***	.055	**.052**	**.252**	1.510	**.233**	.236
IFGSM_U	IFGSM_U	*.055*	***.048***	.034	**.030**	**.105**	.121	*.052*	***.051***	.039	**.032**	**.191**	1.379	**.177**	.185

(b) *MNIST 5vs6*

Table C.2. *Test risks and bounds for 3 tasks Fashion MNIST with n=100 perturbations for all pairs (Defense,Attack) with the two voters' set $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$. The results in* ***bold*** *correspond to the best values between results for $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$. To quantify the gap between our risks and the classical definition we put in italic the risk of our models against the classical attacks: we replace PGD_U and IFGSM_U by PGD or IFGSM (i.e., we did not sample from the uniform distribution). Since Eq.* (3.9) *upperbounds Eq.* (3.8) *thanks to the TV term, we compute the two bound values of Theorem 3.3.3.*

ℓ_2-norm		Algo. 3.1 with Eq. (3.6)						Algo. 3.1 with Eq. (3.9)							
$b=1$		Attack without U						Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.021*	***.020***	.021	**.020**	**0.060**	0.070	*.019*	*.019*	**.019**	**.019**	**0.130**	0.139	**0.130**	0.139
—	PGD_U	*.695*	***.650***	**.494**	.568	**1.042**	1.090	***.677***	*.686*	**.588**	.674	**1.326**	2.307	1.152	**1.082**
—	IFGSM_U	*.451*	*.451*	**.269**	.328	**0.585**	0.731	***.405***	*.438*	**.295**	.381	**0.878**	1.971	**0.730**	0.746
UNIF	—	*.071*	*.071*	.071	.071	**0.185**	0.191	*.071*	*.071*	.071	.071	**0.236**	0.241	**0.236**	0.241
UNIF	PGD_U	***.423***	*.477*	**.418**	.425	0.957	**0.755**	*.486*	*.486*	.513	.513	**1.372**	2.173	1.151	**0.869**
UNIF	IFGSM_U	***.326***	*.331*	.105	.105	**0.273**	0.422	*.333*	***.331***	.144	**.142**	**0.496**	1.642	**0.397**	0.504
PGD_U	—	*.034*	***.032***	.034	**.032**	**0.094**	0.114	*.034*	***.032***	.034	**.032**	**0.158**	0.174	**0.158**	0.174
PGD_U	PGD_U	***.103***	*.115*	**.086**	.091	**0.227**	0.289	***.102***	*.115*	**.096**	.101	**0.299**	1.985	**0.283**	0.338
PGD_U	IFGSM_U	***.092***	*.099*	**.073**	.076	**0.195**	0.248	***.092***	*.099*	.082	.082	**0.266**	1.914	**0.253**	0.299
IFGSM_U	—	***.028***	*.030*	**.028**	.030	**0.091**	0.105	***.027***	*.030*	**.027**	.030	**0.155**	0.166	**0.155**	0.166
IFGSM_U	PGD_U	*.115*	***.114***	.085	.085	**0.254**	0.287	***.112***	*.114*	**.096**	.101	**0.331**	2.026	**0.313**	0.337
IFGSM_U	IFGSM_U	***.095***	*.097*	**.067**	.068	**0.206**	0.232	***.093***	*.097*	**.080**	.081	**0.282**	1.927	**0.266**	0.285

(a) *Fashion MNIST Sandall vs Ankle Boot*

ℓ_2-norm		Algo. 3.1 with Eq. (3.6)						Algo. 3.1 with Eq. (3.9)							
$b=1$		Attack without U						Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.038*	***.037***	.038	**.037**	**.088**	.091	*.038*	***.037***	.038	**.037**	**.153**	0.155	**.153**	.155
—	PGD_U	*.292*	***.248***	.233	**.112**	.452	**.363**	*.289*	***.272***	.287	**.246**	**.578**	1.314	.525	**.479**
—	IFGSM_U	*.194*	***.154***	.132	**.075**	.300	**.262**	*.193*	***.181***	.176	**.148**	**.423**	1.103	.376	**.359**
UNIF	—	*.039*	*.039*	**.039**	**.039**	**.091**	.093	*.041*	***.039***	.041	**.039**	**.155**	0.157	**.155**	.157
UNIF	PGD_U	*.240*	***.220***	**.099**	.117	.346	**.332**	*.250*	***.231***	.250	**.245**	**.553**	1.228	.490	**.443**
UNIF	IFGSM_U	*.177*	***.171***	**.070**	.078	**.228**	.247	*.197*	***.185***	.186	**.164**	**.445**	1.046	.371	**.346**
PGD_U	—	*.045*	***.044***	.045	**.044**	.108	**.105**	*.046*	***.045***	.046	**.045**	.172	**0.167**	.172	**.167**
PGD_U	PGD_U	*.108*	***.100***	**.077**	.082	**.203**	.211	*.104*	***.100***	**.081**	.087	**.279**	1.118	.269	**.264**
PGD_U	IFGSM_U	*.094*	***.086***	.071	**.069**	**.184**	.186	*.090*	***.086***	.076	**.073**	**.257**	1.015	.248	**.241**
IFGSM_U	—	***.041***	*.043*	**.041**	.043	**.094**	.101	***.039***	*.042*	**.039**	.042	**.158**	0.163	**.158**	.163
IFGSM_U	PGD_U	***.106***	*.114*	**.078**	.092	**.220**	.226	***.109***	*.113*	**.084**	.095	**.293**	1.052	.279	**.275**
IFGSM_U	IFGSM_U	***.082***	*.087*	**.065**	.072	**.171**	.176	***.082***	*.089*	**.068**	.078	**.247**	0.927	.234	**.232**

(b) *Fashion MNIST Top vs Pullover*

ℓ_2-norm $b=1$		Algo. 3.1 with Eq. (3.6) Attack without U						Algo. 3.1 with Eq. (3.9) Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widetilde{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widetilde{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.122*	*.122*	.122	.122	**0.276**	0.286	*.122*	*.122*	.122	.122	**0.318**	0.328	**0.318**	0.328
—	PGD_{U}	*.744*	***.738***	**.674**	.689	1.386	**1.066**	*.745*	***.740***	**.767**	.768	**1.773**	2.386	1.576	**1.180**
—	IFGSM_{U}	*.652*	***.646***	**.454**	.474	0.947	**0.887**	*.659*	***.648***	**.618**	.632	**1.597**	2.214	1.276	**0.992**
UNIF	—	*.204*	*.204*	.204	.204	0.444	0.444	*.204*	*.204*	.204	.204	**0.475**	0.476	**0.475**	0.476
UNIF	PGD_{U}	*.750*	***.714***	.682	**.671**	1.350	**1.069**	*.750*	***.719***	.752	**.749**	**1.732**	2.063	1.524	**1.189**
UNIF	IFGSM_{U}	*.605*	***.575***	**.423**	.431	0.871	**0.866**	*.605*	***.578***	.530	**.526**	**1.304**	1.860	1.091	**0.956**
PGD_{U}	—	*.168*	***.165***	.168	**.165**	**0.423**	0.428	*.167*	***.165***	.167	**.165**	0.463	**0.461**	0.463	**0.460**
PGD_{U}	PGD_{U}	***.389***	*.402*	**.306**	.369	0.768	**0.719**	***.390***	*.402*	**.319**	.403	**0.847**	2.354	0.810	**0.755**
PGD_{U}	IFGSM_{U}	***.361***	*.368*	**.298**	.324	0.693	**0.672**	***.362***	*.368*	**.320**	.361	**0.799**	2.258	0.754	**0.707**
IFGSM_{U}	—	***.150***	*.163*	**.150**	.163	**0.424**	0.428	***.149***	*.163*	**.149**	.163	**0.458**	0.461	**0.458**	0.461
IFGSM_{U}	PGD_{U}	***.391***	*.428*	.347	**.292**	0.778	**0.757**	***.390***	*.426*	.371	**.298**	**0.856**	2.327	0.820	**0.791**
IFGSM_{U}	IFGSM_{U}	***.356***	*.382*	.291	**.273**	**0.685**	0.689	***.354***	*.382*	.331	**.278**	**0.772**	2.218	0.734	**0.723**

(a) *Fashion MNIST Coat vs Shirt*

Table C.4. *Test risks and bounds for 3 tasks of MNIST with $n{=}100$ perturbations for all pairs (Defense,Attack) with the two voters' set $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$. The results in* ***bold*** *correspond to the best values between results for $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$. To quantify the gap between our risks and the classical definition we put in italic the risk of our models against the classical attacks: we replace PGD_U and IFGSM_U by PGD or IFGSM (i.e., we did not sample from the uniform distribution). Since Eq.* (3.9) *upperbounds Eq.* (3.8) *thanks to the TV term, we compute the two bound values of Theorem 3.3.3.*

ℓ_∞-norm $b=0.1$		Algo. 3.1 with Eq. (3.6)						Algo. 3.1 with Eq. (3.9)							
		Attack without U						Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widehat{\mathbb{I}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widehat{\mathbb{I}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.005*	*.005*	.005	.005	**.017**	.019	*.005*	*.005*	.005	.005	**0.099**	0.100	**.099**	.100
—	PGD_U	*.454*	*.454*	**.375**	.384	.770	**.638**	*.492*	***.484***	.480	**.476**	**1.127**	2.031	.946	**.716**
—	IFGSM_U	*.428*	***.423***	**.350**	.361	.727	**.610**	*.474*	***.465***	.448	**.443**	**1.061**	2.008	.886	**.686**
UNIF	—	*.004*	*.004*	.004	.004	**.018**	.019	*.004*	*.004*	.004	.004	**0.099**	0.100	**.099**	.100
UNIF	PGD_U	***.487***	*.491*	**.369**	.392	.779	**.667**	*.512*	***.507***	**.484**	.487	**1.179**	2.083	.972	**.739**
UNIF	IFGSM_U	***.436***	*.442*	**.325**	.337	.664	**.598**	*.466*	***.459***	.417	.417	**1.023**	1.959	.841	**.671**
PGD_U	—	*.006*	*.006*	.006	.006	.024	.024	***.005***	*.006*	**.005**	.006	0.103	0.103	.103	.103
PGD_U	PGD_U	***.018***	*.020*	**.013**	.016	**.046**	.050	***.018***	*.020*	**.015**	.020	**0.127**	1.461	**.122**	.123
PGD_U	IFGSM_U	***.020***	*.021*	**.012**	.016	**.048**	.054	***.019***	*.021*	**.015**	.020	**0.130**	1.455	**.125**	.127
IFGSM_U	—	***.006***	*.007*	**.006**	.007	**.023**	.024	***.006***	*.007*	**.006**	.007	**0.102**	0.103	**.102**	.103
IFGSM_U	PGD_U	***.018***	*.019*	.016	.016	**.046**	.051	***.018***	*.019*	**.018**	.019	**0.126**	1.489	**.122**	.124
IFGSM_U	IFGSM_U	*.020*	*.020*	**.015**	.016	**.050**	.055	*.020*	*.020*	.020	**.019**	**0.131**	1.481	**.126**	.127

(a) *MNIST 1 vs 7*

ℓ_∞-norm $b=0.1$		Algo. 3.1 with Eq. (3.6)						Algo. 3.1 with Eq. (3.9)							
		Attack without U						Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widehat{\mathbb{I}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widehat{\mathbb{I}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.015*	*.015*	.015	.015	**0.060**	0.067	*.015*	*.015*	.015	.015	**0.129**	0.135	**0.129**	0.135
—	PGD_U	***.929***	*.930*	**.651**	.662	1.367	**1.125**	***.920***	*.925*	**.874**	.880	**2.213**	2.661	1.792	**1.266**
—	IFGSM_U	*.935*	*.935*	**.601**	.609	1.243	**1.088**	***.926***	*.928*	**.800**	.806	**2.047**	2.615	1.649	**1.224**
UNIF	—	*.017*	*.017*	.017	.017	**0.062**	0.072	*.017*	*.017*	.017	.017	**0.131**	0.139	**0.131**	0.139
UNIF	PGD_U	*.895*	*.895*	**.615**	.623	1.302	**1.078**	***.884***	*.888*	**.815**	.818	**2.035**	2.722	1.670	**1.208**
UNIF	IFGSM_U	*.898*	*.898*	**.516**	.528	1.112	**1.027**	***.884***	*.890*	**.697**	.706	**1.875**	2.658	1.497	**1.153**
PGD_U	—	*.039*	***.037***	.039	**.037**	**0.093**	0.094	*.039*	***.037***	.039	**.037**	**0.156**	0.157	**0.156**	0.157
PGD_U	PGD_U	***.108***	*.109*	.090	.090	**0.200**	0.209	***.108***	*.109*	**.110**	.112	**0.337**	1.874	0.290	**0.271**
PGD_U	IFGSM_U	***.121***	*.124*	**.101**	.103	**0.229**	0.235	***.121***	*.124*	.126	**.125**	**0.378**	1.890	0.326	**0.297**
IFGSM_U	—	*.046*	***.044***	.046	**.044**	**0.102**	0.119	*.046*	***.044***	.046	**.044**	**0.164**	0.178	**0.164**	0.178
IFGSM_U	PGD_U	*.105*	***.093***	.091	**.078**	**0.203**	0.214	*.105*	***.093***	.108	**.089**	**0.321**	1.810	0.286	**0.269**
IFGSM_U	IFGSM_U	*.119*	***.095***	.102	**.080**	**0.220**	0.229	*.119*	***.095***	.122	**.090**	**0.357**	1.821	0.309	**0.283**

(b) *MNIST 4 vs 9*

ℓ_∞-norm $b = 0.1$		Algo. 3.1 with Eq. (3.6) Attack without u						Algo. 3.1 with Eq. (3.9) Attack without u							
		$A_{\mathcal{T}}(MV_\rho)$		$R_{\widehat{\mathbb{H}}}(MV_\rho)$		Th. 3.3.2		$A_{\mathcal{T}}(MV_\rho)$		$A_{\widehat{\mathbb{H}}}(MV_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.015*	*.015*	.015	.015	**.043**	.045	*.015*	*.015*	.015	.015	**0.117**	0.118	**0.117**	.118
—	PGD_U	*.500*	***.499***	**.387**	.390	.923	**.744**	*.502*	***.500***	**.474**	.475	**1.361**	2.275	1.146	**.830**
—	IFGSM_U	*.519*	***.505***	**.395**	.398	.915	**.762**	***.514***	*.516*	.481	.481	**1.335**	2.283	1.129	**.847**
UNIF	—	*.015*	*.015*	.015	.015	**.052**	.053	*.015*	*.015*	.015	.015	**0.123**	0.124	**0.123**	.124
UNIF	PGD_U	***.529***	*.544*	**.388**	.393	.925	**.761**	***.517***	*.532*	**.481**	.482	**1.342**	2.349	1.137	**.848**
UNIF	IFGSM_U	***.536***	*.544*	**.372**	.379	.881	**.774**	***.523***	*.544*	**.451**	.456	**1.268**	2.348	1.077	**.857**
PGD_U	—	*.015*	***.014***	.015	**.014**	**.060**	.064	*.015*	***.014***	.015	**.014**	**0.130**	0.133	**0.130**	.133
PGD_U	PGD_U	***.055***	*.058*	**.037**	.039	**.131**	.143	***.056***	*.057*	.046	.046	**0.219**	1.619	**0.202**	.204
PGD_U	IFGSM_U	***.061***	*.065*	**.040**	.043	**.146**	.154	***.059***	*.062*	.050	**.046**	**0.232**	1.626	0.216	**.214**
IFGSM_U	—	*.019*	***.014***	.019	**.014**	.069	**.064**	*.018*	***.014***	.018	**.014**	0.136	**0.132**	0.136	**.132**
IFGSM_U	PGD_U	*.061*	*.061*	**.040**	.050	.143	**.142**	*.061*	*.061*	**.045**	.061	**0.218**	1.694	0.208	**.205**
IFGSM_U	IFGSM_U	***.066***	*.069*	**.044**	.054	.154	**.152**	***.065***	*.069*	**.048**	.068	**0.228**	1.708	0.216	**.214**

(a) *MNIST 5 vs 6*

Table C.6. *Test risks and bounds for 3 tasks of Fashion MNIST with $n{=}100$ perturbations for all pairs (Defense,Attack) with the two voters' set $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$. The results in* ***bold*** *correspond to the best values between results for $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$. To quantify the gap between our risks and the classical definition we put in italic the risk of our models against the classical attacks: we replace PGD_U and IFGSM_U by PGD or IFGSM (i.e., we did not sample from the uniform distribution). Since Eq.* (3.9) *upperbounds Eq.* (3.8) *thanks to the TV term, we compute the two bound values of Theorem 3.3.3.*

ℓ_∞-norm $b=0.1$		Algo. 3.1 with Eq. (3.6)						Algo. 3.1 with Eq. (3.9)							
		Attack without U						Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.021*	***.020***	.021	**.020**	**0.060**	0.070	*.019*	*.019*	.019	.019	**0.130**	0.139	**0.130**	0.139
—	PGD_U	*.951*	***.944***	**.606**	.719	**1.275**	1.333	*.935*	***.920***	**.762**	.864	**1.617**	2.503	1.421	**1.317**
—	IFGSM_U	*.957*	***.947***	**.588**	.718	**1.231**	1.336	*.950*	*.950*	**.734**	.851	**1.587**	2.495	1.395	**1.316**
UNIF	—	***.076***	*.077*	**.076**	.077	**0.178**	0.184	***.076***	*.077*	**.076**	.077	**0.230**	0.235	**0.230**	0.235
UNIF	PGD_U	*.964*	***.961***	**.714**	.719	1.496	**1.265**	*.966*	***.963***	**.853**	.859	**2.098**	2.417	1.785	**1.416**
UNIF	IFGSM_U	*.978*	***.976***	**.627**	.632	1.306	**1.259**	*.979*	*.979*	**.758**	.762	**1.914**	2.422	1.597	**1.396**
PGD_U	—	*.041*	***.040***	.041	**.040**	0.114	**0.111**	*.041*	***.040***	.041	**.040**	0.173	**0.171**	0.173	**0.171**
PGD_U	PGD_U	*.098*	***.097***	.089	**.086**	**0.207**	0.210	*.099*	***.097***	.101	**.100**	**0.306**	1.826	0.281	**0.267**
PGD_U	IFGSM_U	*.113*	***.112***	.105	**.101**	**0.244**	0.246	*.115*	***.112***	.120	**.113**	**0.353**	1.853	0.321	**0.302**
IFGSM_U	—	***.045***	*.047*	**.045**	.047	**0.131**	0.137	***.045***	*.047*	**.045**	.047	**0.188**	0.194	**0.188**	0.194
IFGSM_U	PGD_U	***.100***	*.102*	.089	**.085**	**0.203**	0.232	***.100***	*.102*	**.102**	**.102**	**0.298**	1.645	**0.274**	0.287
IFGSM_U	IFGSM_U	***.112***	*.116*	.099	**.096**	**0.232**	0.260	***.112***	*.116*	.114	**.112**	**0.328**	1.687	**0.301**	0.313

(a) *Fashion MNIST Sandall vs Ankle Boot*

ℓ_∞-norm $b=0.1$		Algo. 3.1 with Eq. (3.6)						Algo. 3.1 with Eq. (3.9)							
		Attack without U						Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.038*	***.037***	.038	**.037**	**.088**	.091	*.038*	***.037***	.038	**.037**	**0.153**	0.155	**0.153**	.155
—	PGD_U	*.596*	***.515***	.477	**.218**	.844	**.662**	*.590*	***.576***	.570	**.502**	**1.049**	1.924	0.948	**.857**
—	IFGSM_U	*.723*	***.623***	.573	**.257**	.971	**.751**	*.716*	***.695***	.678	**.598**	**1.189**	2.031	1.080	**.980**
UNIF	—	*.032*	*.032*	.032	.032	**.083**	.085	***.032***	*.033*	**.032**	.033	**0.149**	0.151	**0.149**	.151
UNIF	PGD_U	***.438***	*.439*	.356	**.245**	.813	**.563**	*.435*	*.435*	.423	**.312**	**1.082**	1.867	0.959	**.688**
UNIF	IFGSM_U	***.546***	*.547*	.453	**.325**	.974	**.690**	***.544***	*.547*	.530	**.409**	**1.266**	2.009	1.128	**.823**
PGD_U	—	***.048***	*.053*	**.048**	.053	**.115**	.130	***.048***	*.053*	**.048**	.053	**0.177**	0.188	**0.177**	.188
PGD_U	PGD_U	***.102***	*.116*	**.089**	.099	**.205**	.223	***.102***	*.116*	**.096**	.115	**0.282**	1.323	**0.266**	.278
PGD_U	IFGSM_U	***.120***	*.135*	**.102**	.115	**.237**	.255	***.120***	*.135*	**.109**	.133	**0.318**	1.380	**0.299**	.309
IFGSM_U	—	*.051*	***.045***	.051	**.045**	.120	**.115**	*.051*	***.045***	.051	**.045**	0.179	**0.175**	0.179	**.175**
IFGSM_U	PGD_U	*.106*	***.094***	.091	**.085**	.211	**.193**	*.106*	***.094***	.102	**.097**	**0.292**	1.488	0.273	**.252**
IFGSM_U	IFGSM_U	*.120*	***.111***	**.101**	.102	.239	**.218**	*.119*	***.111***	.113	.113	**0.322**	1.546	0.299	**.277**

(b) *Fashion MNIST Top vs Pullover*

ℓ_∞-norm		Algo. 3.1 with Eq. (3.6)						Algo. 3.1 with Eq. (3.9)							
$b = 0.1$		Attack without U						Attack without U							
		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{R}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.2		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_\rho)$		$\mathsf{A}_{\widehat{\mathbb{T}}}(\mathsf{MV}_\rho)$		Th. 3.3.3 - Eq. (3.9)		Th. 3.3.3 - Eq. (3.8)	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	*.122*	*.122*	.122	.122	**0.276**	0.286	*.122*	*.122*	.122	.122	**0.318**	0.328	**0.318**	0.328
—	PGD_U	***.884***	*.887*	**.781**	.795	1.579	**1.268**	***.882***	*.886*	**.864**	.872	**2.020**	2.640	1.803	**1.390**
—	IFGSM_U	***.901***	*.902*	**.756**	.774	1.558	**1.272**	***.901***	*.902*	**.865**	.876	**2.032**	2.651	1.795	**1.393**
UNIF	—	*.166*	*.166*	.166	.166	**0.352**	0.357	.166	*.166*	*.166*	.166	**0.389**	0.394	**0.389**	0.394
UNIF	PGD_U	***.911***	*.914*	**.796**	.798	1.402	**1.326**	***.913***	*.914*	.896	**.888**	**1.934**	2.325	1.713	**1.447**
UNIF	IFGSM_U	***.935***	*.937*	**.787**	.798	1.392	**1.350**	***.934***	*.936*	.887	**.882**	**1.905**	2.378	1.693	**1.469**
PGD_U	—	*.163*	***.162***	.163	**.162**	**0.386**	0.395	*.163*	***.162***	.163	**.162**	**0.419**	0.430	**0.419**	0.430
PGD_U	PGD_U	***.394***	*.396*	.359	**.329**	0.764	**0.673**	***.394***	*.396*	.403	**.394**	**0.954**	2.321	0.865	**0.726**
PGD_U	IFGSM_U	***.475***	*.480*	.442	**.410**	0.910	**0.769**	***.477***	*.480*	.487	**.472**	**1.121**	2.411	1.020	**0.826**
IFGSM_U	—	***.167***	*.168*	**.167**	.168	0.411	**0.395**	***.167***	*.168*	**.167**	.168	0.445	**0.429**	0.445	**0.429**
IFGSM_U	PGD_U	*.396*	***.373***	.359	**.293**	0.772	**0.641**	*.396*	***.373***	.405	**.328**	**0.970**	2.368	0.877	**0.692**
IFGSM_U	IFGSM_U	*.465*	***.428***	.424	**.334**	0.891	**0.705**	*.465*	***.429***	.470	**.372**	**1.090**	2.425	0.995	**0.758**

(a) *Fashion MNIST Coat vs Shirt*

Table C.8. *Test risks for 6 tasks of MNIST and Fashion MNIST datasets for all pairs (Defense,Attack) with the two voters' set $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$ using ℓ_2-norm. The results of these tables are computed considering defenses of the literature, i.e., adversarial training using* PGD *or* IFGSM*. We also add an adversarial training using* UNIF *for the completeness of comparison between this baseline defense and our algorithm. The results in* ***bold*** *correspond to the best values between results for $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$.*

ℓ_2-norm, $b=1$		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.005	.005
—	PGD	**.326**	.327
—	IFGSM	.122	**.121**
UNIF	—	.005	.005
UNIF	PGD	.191	**.190**
UNIF	IFGSM	**.071**	.072
PGD	—	.007	.007
PGD	PGD	.027	**.026**
PGD	IFGSM	.022	**.021**
IFGSM	—	**.005**	.006
IFGSM	PGD	.041	**.035**
IFGSM	IFGSM	.021	.021

(a) *MNIST 1 vs 7*

ℓ_2-norm, $b=1$		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.015	.015
—	PGD	.692	.692
—	IFGSM	.464	**.462**
UNIF	—	.024	.024
UNIF	PGD	.653	.653
UNIF	IFGSM	.441	**.438**
PGD	—	**.024**	.027
PGD	PGD	**.136**	.138
PGD	IFGSM	**.097**	.102
IFGSM	—	**.022**	.027
IFGSM	PGD	**.166**	.186
IFGSM	IFGSM	**.113**	.124

(b) *MNIST 4 vs 9*

ℓ_2-norm, $b=1$		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.015	.015
—	PGD	.283	.283
—	IFGSM	.144	.144
UNIF	—	.017	.017
UNIF	PGD	.220	**.219**
UNIF	IFGSM	.122	.122
PGD	—	.014	**.013**
PGD	PGD	.056	**.055**
PGD	IFGSM	.045	**.041**
IFGSM	—	**.013**	.014
IFGSM	PGD	.077	**.070**
IFGSM	IFGSM	.053	**.047**

(c) *MNIST 5 vs 6*

ℓ_2-norm, $b=1$		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.019	.019
—	PGD	.709	**.708**
—	IFGSM	.426	**.414**
UNIF	—	**.071**	.072
UNIF	PGD	.531	.531
UNIF	IFGSM	.331	**.329**
PGD	—	**.034**	.036
PGD	PGD	.107	**.103**
PGD	IFGSM	.091	**.087**
IFGSM	—	.031	**.029**
IFGSM	PGD	.125	**.108**
IFGSM	IFGSM	.104	**.090**

(d) *Fashion MNIST Sandall vs Ankle Boot*

ℓ_2-norm, $b=1$		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.038	.038
—	PGD	.286	**.285**
—	IFGSM	.188	**.186**
UNIF	—	.041	**.039**
UNIF	PGD	.249	**.248**
UNIF	IFGSM	.197	**.192**
PGD	—	**.043**	.045
PGD	PGD	**.102**	.117
PGD	IFGSM	**.090**	.094
IFGSM	—	**.038**	.040
IFGSM	PGD	.120	**.106**
IFGSM	IFGSM	.092	**.080**

(e) *Fashion MNIST Top vs Pullover*

ℓ_2-norm, $b=1$		$\mathsf{A}_{\mathcal{T}}(\mathsf{MV}_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.122	.122
—	PGD	.768	**.767**
—	IFGSM	.683	**.680**
UNIF	—	.204	.204
UNIF	PGD	**.753**	.754
UNIF	IFGSM	.607	**.606**
PGD	—	.182	**.178**
PGD	PGD	.453	**.412**
PGD	IFGSM	.408	**.379**
IFGSM	—	.148	**.146**
IFGSM	PGD	**.405**	.411
IFGSM	IFGSM	.369	**.364**

(f) *Fashion MNIST Coat vs Shirt*

Table C.9. *Test risks for 6 tasks of MNIST and Fashion MNIST datasets for all pairs (Defense,Attack) with the two voters' set $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$ using ℓ_∞-norm. The results of these tables are computed considering defenses of the literature, i.e., adversarial training using* PGD *or* IFGSM*. We also add an adversarial training using* UNIF *for the completeness of comparison between this baseline defense and our algorithm. The results in* ***bold*** *correspond to the best values between results for $\mathbb{H}$ and $\mathbb{H}^{\text{SIGN}}$.*

ℓ_∞-norm, $b = 0.1$		$A_{\mathcal{T}}(MV_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.005	.005
—	PGD	.499	**.498**
—	IFGSM	**.479**	.480
UNIF	—	.004	.004
UNIF	PGD	.516	**.515**
UNIF	IFGSM	.467	.467
PGD	—	**.006**	.007
PGD	PGD	.019	.019
PGD	IFGSM	.021	.021
IFGSM	—	.007	.007
IFGSM	PGD	**.017**	.018
IFGSM	IFGSM	**.019**	.020

(a) *MNIST 1 vs 7*

ℓ_∞-norm, $b = 0.1$		$A_{\mathcal{T}}(MV_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.015	.015
—	PGD	.921	.921
—	IFGSM	.923	.923
UNIF	—	.017	.017
UNIF	PGD	.877	**.876**
UNIF	IFGSM	.877	.877
PGD	—	.041	**.040**
PGD	PGD	**.108**	.109
PGD	IFGSM	**.122**	.123
IFGSM	—	.057	**.044**
IFGSM	PGD	.109	**.101**
IFGSM	IFGSM	.119	**.108**

(b) *MNIST 4 vs 9*

ℓ_∞-norm, $b = 0.1$		$A_{\mathcal{T}}(MV_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.015	.015
—	PGD	.498	.498
—	IFGSM	.511	**.510**
UNIF	—	.015	.015
UNIF	PGD	.512	**.511**
UNIF	IFGSM	.511	.511
PGD	—	.014	.014
PGD	PGD	.065	**.058**
PGD	IFGSM	.068	**.065**
IFGSM	—	.018	**.017**
IFGSM	PGD	**.061**	.063
IFGSM	IFGSM	**.069**	.071

(c) *MNIST 5 vs 6*

ℓ_∞-norm, $b = 0.1$		$A_{\mathcal{T}}(MV_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.019	.019
—	PGD	.938	.938
—	IFGSM	**.948**	.949
UNIF	—	**.076**	.077
UNIF	PGD	.970	**.969**
UNIF	IFGSM	.981	.981
PGD	—	.041	**.040**
PGD	PGD	.098	**.097**
PGD	IFGSM	.115	**.111**
IFGSM	—	.112	**.047**
IFGSM	PGD	**.045**	.100
IFGSM	IFGSM	**.101**	.114

(d) *Fashion MNIST Sandall vs Ankell Boot*

ℓ_∞-norm, $b = 0.1$		$A_{\mathcal{T}}(MV_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.038	.038
—	PGD	**.574**	.577
—	IFGSM	.700	**.696**
UNIF	—	**.032**	.033
UNIF	PGD	**.428**	.435
UNIF	IFGSM	**.540**	.550
PGD	—	**.047**	.049
PGD	PGD	.101	**.097**
PGD	IFGSM	.118	**.112**
IFGSM	—	.049	**.048**
IFGSM	PGD	.100	**.090**
IFGSM	IFGSM	.112	**.108**

(e) *Fashion MNIST Top vs Pullover*

ℓ_∞-norm, $b = 0.1$		$A_{\mathcal{T}}(MV_{\pi'})$	
Defense	Attack	$\mathbb{H}^{\text{SIGN}}$	$\mathbb{H}$
—	—	.122	.122
—	PGD	.879	.879
—	IFGSM	.898	.898
UNIF	—	.166	.166
UNIF	PGD	.913	**.911**
UNIF	IFGSM	.934	**.933**
PGD	—	**.164**	.167
PGD	PGD	.398	**.395**
PGD	IFGSM	**.479**	.481
IFGSM	—	**.163**	.169
IFGSM	PGD	**.356**	.391
IFGSM	IFGSM	**.422**	.461

(f) *Fashion MNIST Coat vs Shirt*

Appendix of Chapter 4

D

D.1 Proof of Theorem 4.3.1

Theorem 4.3.1 (PAC-Bayesian Bound Based on the Gibbs Risk)**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any $\delta \in (0, 1]$, with probability at least $1-\delta$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$ we have

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad r_{\mathcal{D}}(\rho) \leq \overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta}\right]\right),$$

$$\text{and } \forall \rho \in \mathbb{M}(\mathbb{H}), \quad \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq 2\left[\overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta}\right]\right)\right], \quad (4.1)$$

where $\overline{\mathrm{kl}}(q|\tau) \triangleq \max\left\{p \in (0,1) \;\middle|\; \mathrm{kl}(q\|p) \leq \tau\right\}$ (see Section 2.3.1.3).

Proof. We can apply Theorem 2.3.4 with the loss $\ell(h, (\mathbf{x}, y)) = \mathrm{I}\left[h(\mathbf{x}) \neq y\right]$ to obtain

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad r_{\mathcal{D}}(\rho) \leq \overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta}\right]\right).$$

Then, from Theorem 2.2.1, we obtain

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq 2\left[\overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta}\right]\right)\right].$$

■

D.2 Proof of Theorem 4.3.2

Theorem 4.3.2 (PAC-Bayesian Bound Based on the Joint Error)**.** For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any

$\delta \in (0,1]$, with probability at least $1-\delta$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$ we have

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad e_{\mathcal{D}}(\rho) \leq \overline{\mathrm{kl}}\left(e_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right] \right),$$

$$\text{and} \quad \forall \rho \in \mathbb{M}(\mathbb{H}), \quad \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq 4 \left[\overline{\mathrm{kl}}\left(e_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right] \right) \right].$$

Proof. We consider the hypothesis set $\mathbb{H}^2 = \mathbb{H}\times\mathbb{H}$ with the distributions ρ^2 (*resp.* π^2) defined as $\rho^2((h,h')) = \rho(h)\rho(h')$ (*resp.* $\pi^2((h,h')) = \pi(h)\pi(h')$) on $\mathbb{H}^2$. We can apply Theorem 2.3.4 with the loss $\ell((h,h'),(\mathbf{x},y)) = \mathrm{I}\left[h(\mathbf{x}) \neq y\right] \mathrm{I}\left[h'(\mathbf{x}) \neq y\right]$ to obtain

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad e_{\mathcal{D}}(\rho) \leq \overline{\mathrm{kl}}\left(e_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho^2\|\pi^2) + \ln \frac{2\sqrt{m}}{\delta} \right] \right).$$

Then, from Theorem 2.2.2 and Lemma D.7.1, we obtain

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq 4 \left[\overline{\mathrm{kl}}\left(e_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right] \right) \right].$$

■

D.3 Proof of Theorem 4.3.3

Theorem 4.3.3 (PAC-Bayesian C-Bound of Roy *et al.* (2016))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$ we have for all $\rho \in \mathbb{M}(\mathbb{H})$

$$\mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq \underbrace{1 - \frac{\left(1 - 2\min\left[\frac{1}{2}, r_{\mathcal{S}}(\rho) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}\right]\right)^2}{1 - 2\max\left[0, d_{\mathcal{S}}(\rho) - \sqrt{\frac{1}{2m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}\right]}}_{\triangleq\; C^{\mathtt{M}}_{\mathcal{S}}(\rho)}. \tag{4.2}$$

Proof. First, remark that from Theorem 2.2.3, we have (even for $r_{\mathcal{D}}(\rho) \geq \frac{1}{2}$)

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq 1 - \frac{\left(1 - 2\min\left[\frac{1}{2}, r_{\mathcal{D}}(\rho)\right]\right)^2}{1 - 2\max\left[0, d_{\mathcal{D}}(\rho)\right]}.$$

We upper-bound $r_{\mathcal{D}}(\rho)$ by applying Theorem 2.3.2 with $\frac{\delta}{2}$ and the loss $\ell(h,(\mathbf{x},y)) = \mathrm{I}\left[h(\mathbf{x}) \neq y\right]$, to obtain

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad r_{\mathcal{D}}(\rho) \leq r_{\mathcal{S}}(\rho) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\tfrac{4\sqrt{m}}{\delta}\right]}.$$

To lower-bound $d_{\mathcal{D}}(\rho)$, we have to consider the hypothesis set $\mathbb{H}^2 = \mathbb{H}\times\mathbb{H}$ with the distributions ρ^2 (*resp.* π^2) defined as $\rho^2((h,h')) = \rho(h)\rho(h')$ (*resp.* $\pi^2((h,h')) = \pi(h)\pi(h')$) on $\mathbb{H}^2$. Then, we obtain the lower-bound by applying Theorem 2.3.4 with $\frac{\delta}{2}$ and the loss $\ell((h,h'),(\mathbf{x},y)) = 2\,\mathrm{I}\left[h(\mathbf{x}) \neq y\right]\mathrm{I}\left[h'(\mathbf{x}) = y\right]$. We have

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad d_{\mathcal{D}}(\rho) \geq d_{\mathcal{S}}(\rho) - \sqrt{\frac{1}{2m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\tfrac{4\sqrt{m}}{\delta}\right]},$$

where $\mathrm{KL}(\rho^2\|\pi^2) = 2\,\mathrm{KL}(\rho\|\pi)$ from Lemma D.7.1. ■

D.4 Proof of Theorem 4.3.4

Theorem 4.3.4 (PAC-Bayesian C-Bound of Germain *et al.* (2015))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$ we have for all $\rho \in \mathbb{M}(\mathbb{H})$

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq \underbrace{1 - \frac{\left(1 - 2\min\left[\frac{1}{2}, \overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]\right)\right]\right)^2}{1 - 2\max\left[0, \underline{\mathrm{kl}}\left(d_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]\right)\right]}}_{\triangleq\; C^{\mathrm{S}}_{\mathcal{S}}(\rho)}. \tag{4.3}$$

Proof. First, remark that from Theorem 2.2.3, we have (even for $r_{\mathcal{D}}(\rho) \geq \frac{1}{2}$)

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq 1 - \frac{\left(1 - 2\min\left[\frac{1}{2}, r_{\mathcal{D}}(\rho)\right]\right)^2}{1 - 2\max\left[0, d_{\mathcal{D}}(\rho)\right]}.$$

We upper-bound $r_{\mathcal{D}}(\rho)$ by applying Theorem 2.3.4 with $\frac{\delta}{2}$ and the loss $\ell(h,(\mathbf{x},y)) = \mathrm{I}\left[h(\mathbf{x}) \neq y\right]$, to obtain

$$\forall \rho \in \mathbb{M}(\mathcal{H}), \quad r_{\mathcal{D}}(\rho) \leq \overline{\mathrm{kl}}\left(r_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]\right).$$

To lower-bound $d_{\mathcal{D}}(\rho)$, we have to consider the hypothesis set $\mathcal{H}^2 = \mathcal{H}\times\mathcal{H}$ with the distributions ρ^2 (*resp.* π^2) defined as $\rho^2((h,h')) = \rho(h)\rho(h')$ (*resp.* $\pi^2((h,h')) = \pi(h)\pi(h')$) on $\mathcal{H}^2$. Then, we obtain the lower-bound by applying Theorem 2.3.4 with $\frac{\delta}{2}$ and the loss $\ell((h,h'),(\mathbf{x},y)) = 2\,\mathrm{I}\left[h(\mathbf{x}) \neq y\right]\mathrm{I}\left[h'(\mathbf{x}) = y\right]$. We have

$$\forall \rho \in \mathbb{M}(\mathcal{H}), \quad d_{\mathcal{D}}(\rho) \geq \underline{\mathrm{kl}}\left(d_{\mathcal{S}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho^2\|\pi^2) + \ln\frac{4\sqrt{m}}{\delta}\right]\right),$$

where $\mathrm{KL}(\rho^2\|\pi^2) = 2\,\mathrm{KL}(\rho\|\pi)$ from Lemma D.7.1. ■

D.5 Proof of Theorem 4.3.5

Theorem 4.3.5 (PAC-Bayesian C-Bound of Lacasse *et al.* (2006))**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathcal{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathcal{H})$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$ we have for all $\rho \in \mathbb{M}(\mathcal{H})$

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq \sup_{(e,d)\in\mathbb{A}_{\mathcal{S}}(\rho)} \left[1 - \frac{(1-(2e+d))^2}{1-2d}\right],$$

$$\text{where} \quad \mathbb{A}_{\mathcal{S}}(\rho) = \left\{(e,d) \;\middle|\; \mathrm{kl}\left(e_{\mathcal{S}}(\rho), d_{\mathcal{S}}(\rho)\|e,d\right) \leq \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}+m}{\delta}\right], \right.$$
$$\left. d \leq 2\sqrt{e} - 2e\,,\; 2e+d < 1\right\}.$$

Proof. First of all, we need to prove the following PAC-Bayesian bound:

$$\mathrm{kl}\left(e_{\mathcal{S}}(\rho), d_{\mathcal{S}}(\rho) \| e_{\mathcal{D}}(\rho), d_{\mathcal{D}}(\rho)\right) \leq \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln \tfrac{2\sqrt{m}+m}{\delta}\right]. \tag{D.1}$$

We apply Theorem 2.3.1 with $\varphi(h, \mathbb{S}) = m\,\mathrm{kl}\left(\mathsf{R}^{\ell}_{\mathcal{S}}(h), \mathsf{R}^{\ell'}_{\mathcal{S}}(h) \| \mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell'}_{\mathcal{D}}(h)\right)$ to obtain with probability at least $1-\delta$

$$\begin{aligned}
&\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{kl}\left(\mathsf{R}^{\ell}_{\mathcal{S}}(h), \mathsf{R}^{\ell'}_{\mathcal{S}}(h) \| \mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell'}_{\mathcal{D}}(h)\right) \\
&\qquad \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} e^{m\,\mathrm{kl}\left(\mathsf{R}^{\ell}_{\mathcal{S}}(h), \mathsf{R}^{\ell'}_{\mathcal{S}}(h) \| \mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell'}_{\mathcal{D}}(h)\right)}\right)\right],
\end{aligned} \tag{D.2}$$

where $\ell' : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$ and $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$. Moreover, Younsi (2012) and Younsi and Lacasse (2020) proves that

$$\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \exp\left[m\,\mathrm{kl}\left(\mathsf{R}^{\ell}_{\mathcal{S}}(h), \mathsf{R}^{\ell'}_{\mathcal{S}}(h) \| \mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell'}_{\mathcal{D}}(h)\right)\right] \leq 2\sqrt{m} + m.$$

Hence, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} e^{m\,\mathrm{kl}\left(\mathsf{R}^{\ell}_{\mathcal{S}}(h), \mathsf{R}^{\ell'}_{\mathcal{S}}(h) \| \mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell'}_{\mathcal{D}}(h)\right)} &= \mathop{\mathbb{E}}_{h'\sim\pi} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} e^{m\,\mathrm{kl}\left(\mathsf{R}^{\ell}_{\mathcal{S}}(h), \mathsf{R}^{\ell'}_{\mathcal{S}}(h) \| \mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell'}_{\mathcal{D}}(h)\right)} \\
&\leq 2\sqrt{m} + m.
\end{aligned} \tag{D.3}$$

Moreover, from Jensen's inequality (Theorem A.1.1), we have

$$\begin{aligned}
&\mathrm{kl}\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{S}}(h), \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell'}_{\mathcal{S}}(h) \Big\| \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}^{\ell'}_{\mathcal{D}}(h)\right) \\
&\qquad \leq \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{kl}\left(\mathsf{R}^{\ell}_{\mathcal{S}}(h), \mathsf{R}^{\ell'}_{\mathcal{S}}(h) \| \mathsf{R}^{\ell}_{\mathcal{D}}(h), \mathsf{R}^{\ell'}_{\mathcal{D}}(h)\right).
\end{aligned} \tag{D.4}$$

By combining Equations (D.2) to (D.4), we obtain Equation (D.1). Lastly the PAC-Bayesian bound in $\mathbb{A}_{\mathcal{S}}(\rho)$ is obtained by instantiating Equation (D.1) (and applying Lemma D.7.1) with

$$\begin{aligned}
\ell((h,h'),(\mathbf{x},y)) &= \mathrm{I}\left[h(\mathbf{x}) \neq y\right] \mathrm{I}\left[h'(\mathbf{x}) \neq y\right], \\
\text{and} \quad \ell'((h,h'),(\mathbf{x},y)) &= 2\,\mathrm{I}\left[h(\mathbf{x}) \neq y\right] \mathrm{I}\left[h'(\mathbf{x}) = y\right],
\end{aligned}$$

and the distributions ρ^2 and π^2 on $\mathbb{H}^2$.

Remark that the constraints involves in the set $\mathbb{A}_{\mathcal{S}}(\rho)$ do not remove the possible couple of joint error and disagreement. Indeed, since a variance is not negative,

we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left(\widehat{m_\rho}(\mathbf{x},y)\right)^2 \geq \left(\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}\left[\widehat{m_\rho}(\mathbf{x},y)\right]\right)^2\\
\iff & d_{\mathcal{D}'}(\rho) \leq 2r_{\mathcal{D}'}(\rho)(1-r_{\mathcal{D}'}(\rho))\\
\iff & d_{\mathcal{D}'}(\rho) \leq 2\left[\sqrt{e_{\mathcal{D}'}(\rho)} - e_{\mathcal{D}'}(\rho)\right].
\end{aligned}$$

Moreover, from Theorem 2.2.3, we know that $2e_{\mathcal{D}'}(\rho)+d_{\mathcal{D}'}(\rho) < 1$.

We can now prove that, under the constraints involved in $\mathbb{A}_{\mathcal{S}}(\rho)$, we still have a valid bound on $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho)$. To do so, we consider two cases.

Case 1: If for all $(e,d) \in \mathbb{A}_{\mathcal{S}}(\rho)$ we have $2e+d < 1$.
In this case $(e_{\mathcal{D}}(\rho), d_{\mathcal{D}}(\rho)) \in \mathbb{A}_{\mathcal{S}}(\rho)$ and $2e_{\mathcal{D}}(\rho)+d_{\mathcal{D}}(\rho) < 1$, thus, Theorem 2.2.3 holds with probability at least $1-\delta$. In other words, we have $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq 1 - \frac{[1-(2e_{\mathcal{D}}(\rho)+d_{\mathcal{D}}(\rho))]^2}{1-2d_{\mathcal{D}}(\rho)} \leq \sup_{(e,d)\in\mathbb{A}'_{\mathcal{S}}(\rho)} C^{\mathrm{L}}(e,d)$ with probability at least $1-\delta$.
Case 2: If there exists $(e,d) \in \mathbb{A}_{\mathcal{S}}(\rho)$ such that $2e+d=1$.
We have $\sup_{(e,d)\in\mathbb{A}'_{\mathcal{S}}(\rho)} C^{\mathrm{L}}(e,d) = 1$ that is a valid bound on $\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho)$. ■

D.6 Proof of Theorem 4.4.1

Theorem 4.4.1 (Reformulation of Lacasse *et al.*'s PAC-Bayesian C-Bound)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$ we have for all $\rho \in \mathbb{M}(\mathbb{H})$

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}_\rho) \leq \sup_{(e,d)\in\mathbb{A}'_{\mathcal{S}}(\rho)} \left[1 - \frac{(1-(2e+d))^2}{1-2d}\right], \tag{4.4}$$

$$\mathbb{A}'_{\mathcal{S}}(\rho) = \Bigg\{(e,d) \;\Bigg|\; \mathrm{kl}\left(e_{\mathcal{S}}(\rho), d_{\mathcal{S}}(\rho) \| e, d\right) \leq \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{m}+m}{\delta}\right],$$
$$d \leq 2\sqrt{\min\left(e, \tfrac{1}{4}\right)} - 2e,\ d < \tfrac{1}{2}\Bigg\}.$$

Proof. Beforehand, we explain how we fixed the constraints involved in $\mathbb{A}'_{\mathcal{S}}(\rho)$. Compared to $\mathbb{A}_{\mathcal{S}}(\rho)$, we add another constraint : $d < \frac{1}{2}$. Hence, we have the

constraints $d \leq 2\sqrt{e}-2e$, $d \leq 1-2e$, and $d < \frac{1}{2}$. We remark that when $e \leq \frac{1}{4}$, we have $2\sqrt{e}-2e \leq 1-2e$. Then, we merge $d \leq 2\sqrt{e}-2e$ and $d \leq 1-2e$ into $d \leq 2\sqrt{\min\left(e, \frac{1}{4}\right)}-2e$. Indeed, we have

$$d \leq 2\sqrt{\min(e, \tfrac{1}{4})}-2e \iff \begin{cases} d \leq 2\sqrt{e} - 2e & \text{if } e \leq \frac{1}{4}, \\ d < 1-2e & \text{if } e \geq \frac{1}{4}. \end{cases}$$

We prove now that under the constraints involved in $\mathbb{A}'_{\mathcal{S}}(\rho)$, we still have a valid bound on $\mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho)$. To do so, we consider two cases.

Case 1: If for all $(e,d) \in \mathbb{A}'_{\mathcal{S}}(\rho)$ we have $2e+d < 1$.
In this case $(e_{\mathcal{D}}(\rho), d_{\mathcal{D}}(\rho)) \in \mathbb{A}'_{\mathcal{S}}(\rho)$ and $2e_{\mathcal{D}}(\rho)+d_{\mathcal{D}}(\rho) < 1$, thus, Theorem 2.2.3 holds with probability at least $1-\delta$. In other words, we have $\mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq 1 - \frac{[1-(2e_{\mathcal{D}}(\rho)+d_{\mathcal{D}}(\rho))]^2}{1-2d_{\mathcal{D}}(\rho)} \leq \sup_{(e,d)\in\mathbb{A}'_{\mathcal{S}}(\rho)} C^{\mathrm{L}}(e,d)$ with probability at least $1-\delta$.
Case 2: If there exists $(e,d) \in \mathbb{A}'_{\mathcal{S}}(\rho)$ such that $2e+d=1$.
We have $\sup_{(e,d)\in\mathbb{A}'_{\mathcal{S}}(\rho)} C^{\mathrm{L}}(e,d) = 1$ that is a valid bound on $\mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho)$. ■

D.7 Proof of Lemma D.7.1

Lemma D.7.1. Let the distribution ρ^2 and π^2 on $\mathbb{H}^2 = \mathbb{H} \times \mathbb{H}$ defined as

$$\rho^2((h,h')) = \rho(h)\rho(h') \quad \text{and} \quad \pi^2((h,h')) = \pi(h)\pi(h').$$

The KL divergence between ρ^2 and π^2 can be expressed *w.r.t.* ρ and π. We have

$$\mathrm{KL}(\rho^2\|\pi^2) = 2\,\mathrm{KL}(\rho\|\pi).$$

Proof. We develop the term $\mathrm{KL}(\rho^2\|\pi^2)$, *i.e.*, we have

$$\begin{aligned} \mathrm{KL}(\rho^2\|\pi^2) &= \mathop{\mathbb{E}}_{(h,h')\sim\rho^2} \ln \frac{\rho^2((h,h'))}{\pi^2((h,h'))} \\ &= \mathop{\mathbb{E}}_{h\sim\rho} \mathop{\mathbb{E}}_{h'\sim\rho} \ln \frac{\rho(h)\rho(h')}{\pi(h)\pi(h')} \\ &= \mathop{\mathbb{E}}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)} + \mathop{\mathbb{E}}_{h'\sim\rho} \ln \frac{\rho(h')}{\pi(h')} \\ &= 2\,\mathrm{KL}(\rho\|\pi). \end{aligned}$$

■

D.8 About Danskin's Theorem

As mentioned in the context of the justification of the function MAXIMIZE-e-d in Algorithm 4.3, we now discuss the possible application of Danskin's Theorem (DANSKIN, 1966, Section I). The statement of the theorem is as follows.

Theorem D.8.1 (Danskin's Theorem)**.** Let $\mathbb{A} \subset \mathbb{R}^a$ be a compact set and $\phi : \mathbb{R}^b \times \mathbb{A} \to \mathbb{R}$ s.t. for all $\mathbf{a} \in \mathbb{A}$, we have that ϕ is continuously differentiable, then $\Phi(\mathbf{x}) = \max_{\mathbf{a}\in\mathbb{A}} \phi(\mathbf{x}, \mathbf{a})$ is directionally differentiable with directional derivatives

$$\Phi'(\mathbf{x}, \mathbf{d}) = \max_{\mathbf{a}\in\mathbb{A}^*} \langle \mathbf{d}, \nabla_{\mathbf{x}}\phi(\mathbf{x}, \mathbf{a}) \rangle ,$$

where $\mathbb{A}^* = \{\mathbf{a}^* \mid \phi(\mathbf{x}, \mathbf{a}^*) = \max_{\mathbf{a}\in\mathbb{A}} \phi(\mathbf{x}, \mathbf{a})\}$ and $\langle \cdot, \cdot \rangle$ is the dot product.

To optimize a problem $\min_{\mathbf{x}\in\mathbb{R}^b} \Phi(\mathbf{x})$ with $\Phi(\mathbf{x}) = \max_{\mathbf{a}\in\mathbb{A}} \phi(\mathbf{x}, \mathbf{a})$, this theorem tells us that under several assumptions, if we know a maximizer $\mathbf{a} \in \mathbb{A}$, then, we have an analytical expression of the directional derivatives of $\Phi(\mathbf{x})$. Thus, from this theorem, we also know a gradient to minimize the problem $\min_{\mathbf{x}\in\mathbb{R}^b} \Phi(\mathbf{x})$; this is expressed in the following corollary.

Corollary D.8.1 (MADRY *et al.*, 2018)**.** Assuming that the conditions of Theorem D.8.1 are fulfilled and let $\mathbf{a}^* \in \mathbb{A}^*$ be a maximizer of ϕ. If $\mathbf{d} = \nabla_{\mathbf{x}}\phi(\mathbf{x}, \mathbf{a}^*)$ with $|\mathbf{d}|_2^2 > 0$ then $-\mathbf{d}$ is a descent direction for $\Phi(\mathbf{x})$, *i.e.*, $\Phi'(\mathbf{x}, \mathbf{d}) > 0$.

Proof. By definition of the directional derivative $\Phi'(\mathbf{x}, \mathbf{d})$, we have:

$$\begin{aligned}\Phi'(\mathbf{x}, \mathbf{d}) &= \max_{\mathbf{a}\in\mathbb{A}^*} \langle \mathbf{d}, \nabla_{\mathbf{x}}\phi(\mathbf{x}, \mathbf{a}) \rangle \\ &= \max_{\mathbf{a}\in\mathbb{A}^*} \langle \nabla_{\mathbf{x}}\phi(\mathbf{x}, \mathbf{a}^*), \nabla_{\mathbf{x}}\phi(\mathbf{x}, \mathbf{a}) \rangle \geq |\nabla_{\mathbf{x}}\phi(\mathbf{x}, \mathbf{a}^*)|_2^2 > 0.\end{aligned}$$

■

Then, for each iteration of the min/max problem optimization, we can *(i)* optimize the inner maximization problem, *(ii)* fix the maximizer $\mathbf{a}^* \in \mathbb{A}$ and apply a gradient descent step with the derivative $\nabla_{\mathbf{x}}\phi(\mathbf{x}, \mathbf{a}^*)$. However, as we mentioned, the assumptions are not fulfilled in our case to apply Theorem D.8.1 since our inner objective in Equation (4.5) or its approximation is not differentiable everywhere in the compact set $[0, \frac{1}{2}]^2$. However, we never encounter problematic cases and this strategy is thus valid

for optimizing our proposed approximation. In practice, we have found that it is indeed an efficient and sound strategy.

D.9 Optimizing the Bound of Theorem 4.3.1

To minimize the bound of Theorem 4.3.1, the objective function is defined as

$$G^{\mathrm{R}}_{\mathcal{U}}(\rho) = 2\left[\overline{\mathrm{kl}}\left(r_{\mathcal{U}}(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]\right)\right].$$

We derive an algorithm (denoted as `Germain` in the setting description of the experiments of Section 4.5) similar to Algorithm 4.2. The algorithm is described in Algorithm D.1 below.

Algorithm D.1 Minimization of Equation (4.1) by Stochastic Gradient Descent

Given: learning sample $\mathbb{S}$, prior distribution π on $\mathbb{H}$, the objective function $G^{\mathrm{R}}_{\mathcal{U}}(\rho)$
Hyperparameters: number of iterations T
$\rho \leftarrow \pi$
for $t \leftarrow 1$ to T **do**
  for all mini-batches $\mathbb{U} \subseteq \mathbb{S}$ **do**
    $\rho \leftarrow$ Update ρ with $G^{\mathrm{R}}_{\mathcal{U}}(\rho)$ by gradient descent

D.10 Additional Experiments

D.10.1 Details for Figures 4.2 to 4.7

We provide in Tables D.1 to D.3 the details of the results in Figures 4.2 to 4.7.

D.10.2 Experiments on the Computation Time

Figure D.1 introduces a comparison of the computation time for the different algorithms. We consider the moons binary dataset and perform, in fact, two experiments on majority votes with decision stumps.

(i) We fix 128 stumps per feature and vary the number $m \in \{100, 500, 1000, 5000\}$ of examples in the dataset.

(ii) We fix the number of examples $m = 5000$ and vary the number of stumps per feature in $\{32, 64, 128\}$.

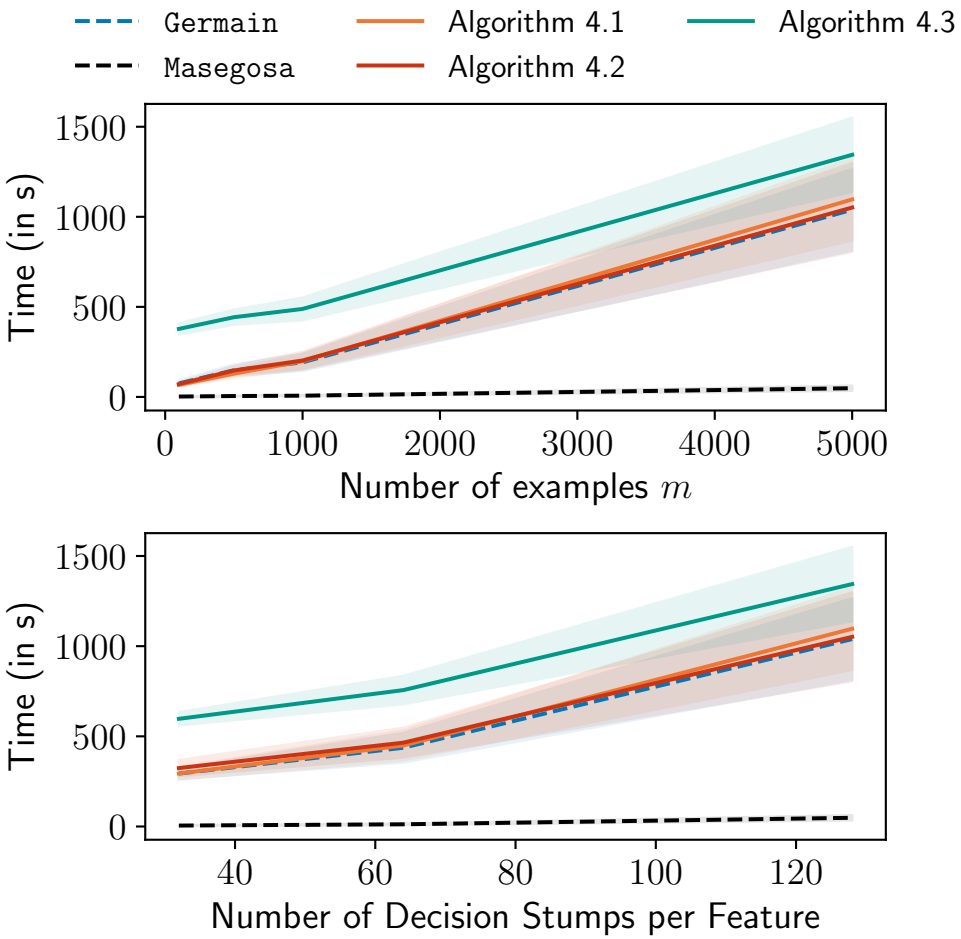


Figure D.1. *In the top plot represents the evolution of the computation time (in second) with respect to the number of examples m in moons dataset. The bottom plot is the evolution of the computation time in function of the number of decision stumps per feature. For each curve, we plot the mean computation time (the plain line) and the standard deviation (the shadow) over 10 runs.*

D.10.3 Details on the Empirical Joint Error and Disagreement

We report in Figures D.2 to D.7, the empirical joint error and disagreement obtained on the different datasets. These figures illustrate that the solutions found by Algorithm 4.3, `Masegosa` and `CB-Boost` are similar while `MinCq` and `Germain` provide usually very different solutions.

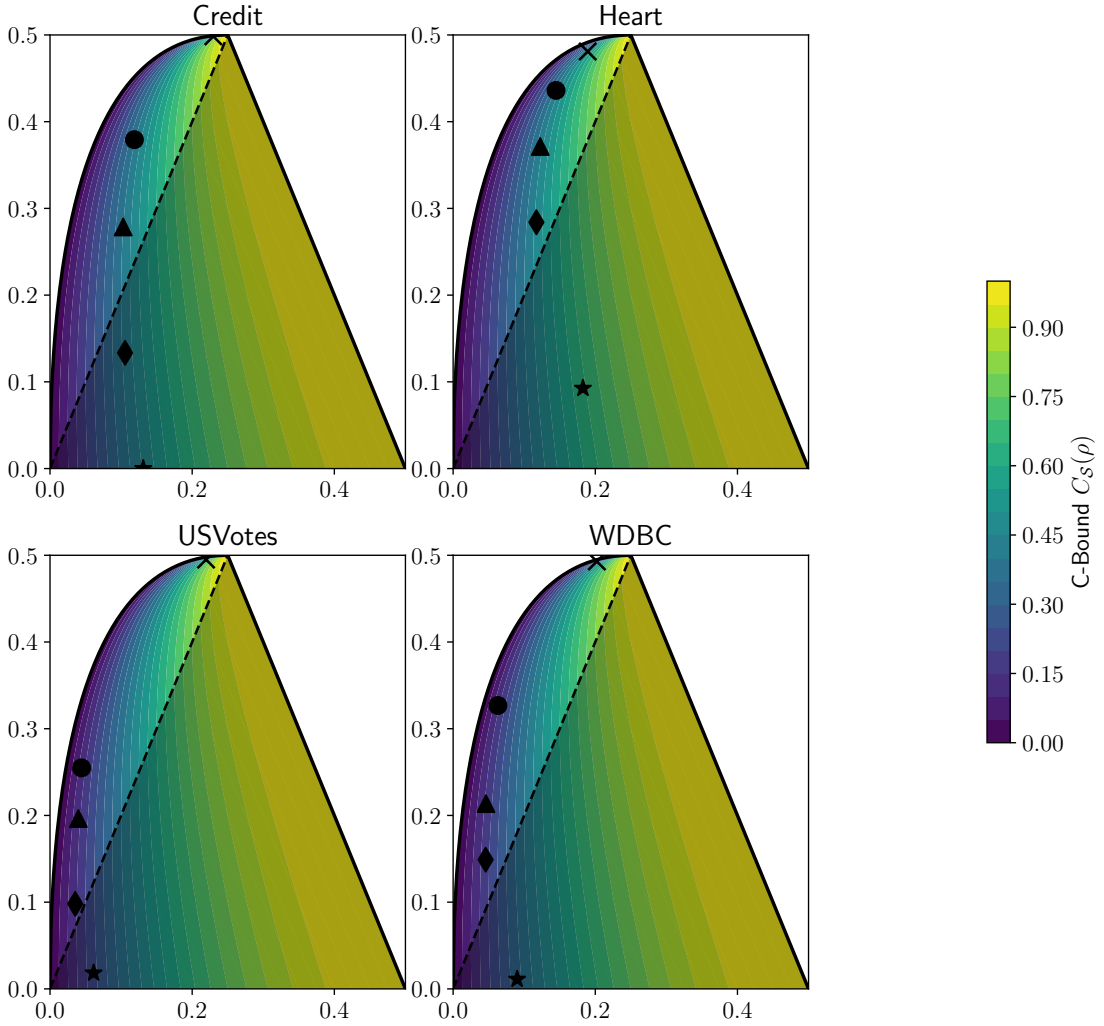


Figure D.2. *Representation of all the possible values of the empirical C-Bound $C_S(\rho)$ in function of the disagreement $d_S(\rho)$ (y-axis) and joint error $e_S(\rho)$ (x-axis). We report the values obtained with the decision stumps in the binary setting by Algorithm 4.3 (⧫),* `Masegosa` *(▲),* `Germain` *(★),* `CB-Boost` *(●), and* `MinCq` *(×). The disagreement and the joint are averaged over the 10 runs.*

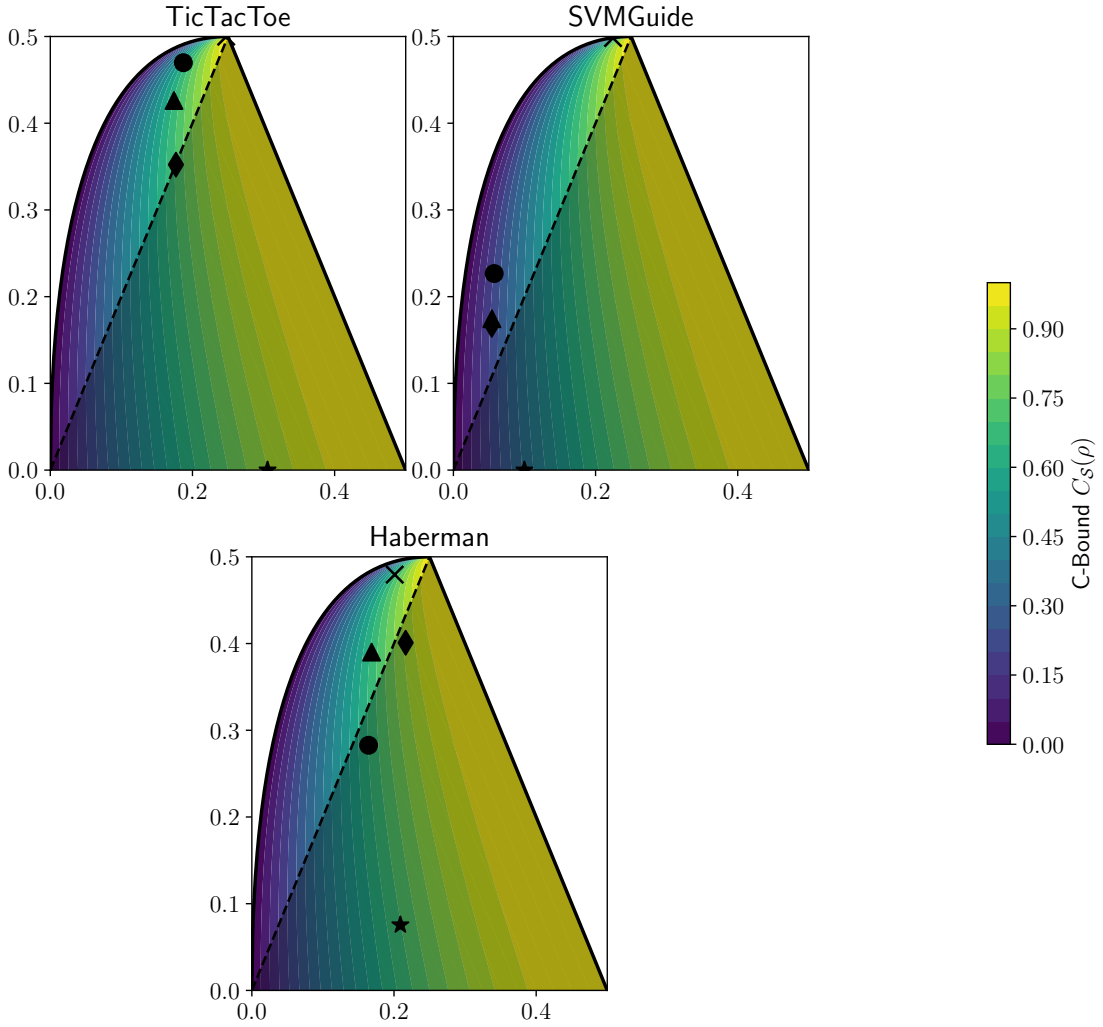


Figure D.3. *Representation of all the possible values of the empirical C-Bound $C_S(\rho)$ in function of the disagreement $d_S(\rho)$ (y-axis) and joint error $e_S(\rho)$ (x-axis). We report the values obtained with the decision stumps in the binary setting by Algorithm 4.3 (⧫),* `Masegosa` *(▲),* `Germain` *(★),* `CB-Boost` *(•), and* `MinCq` *(×). The disagreement and the joint are averaged over the 10 runs.*

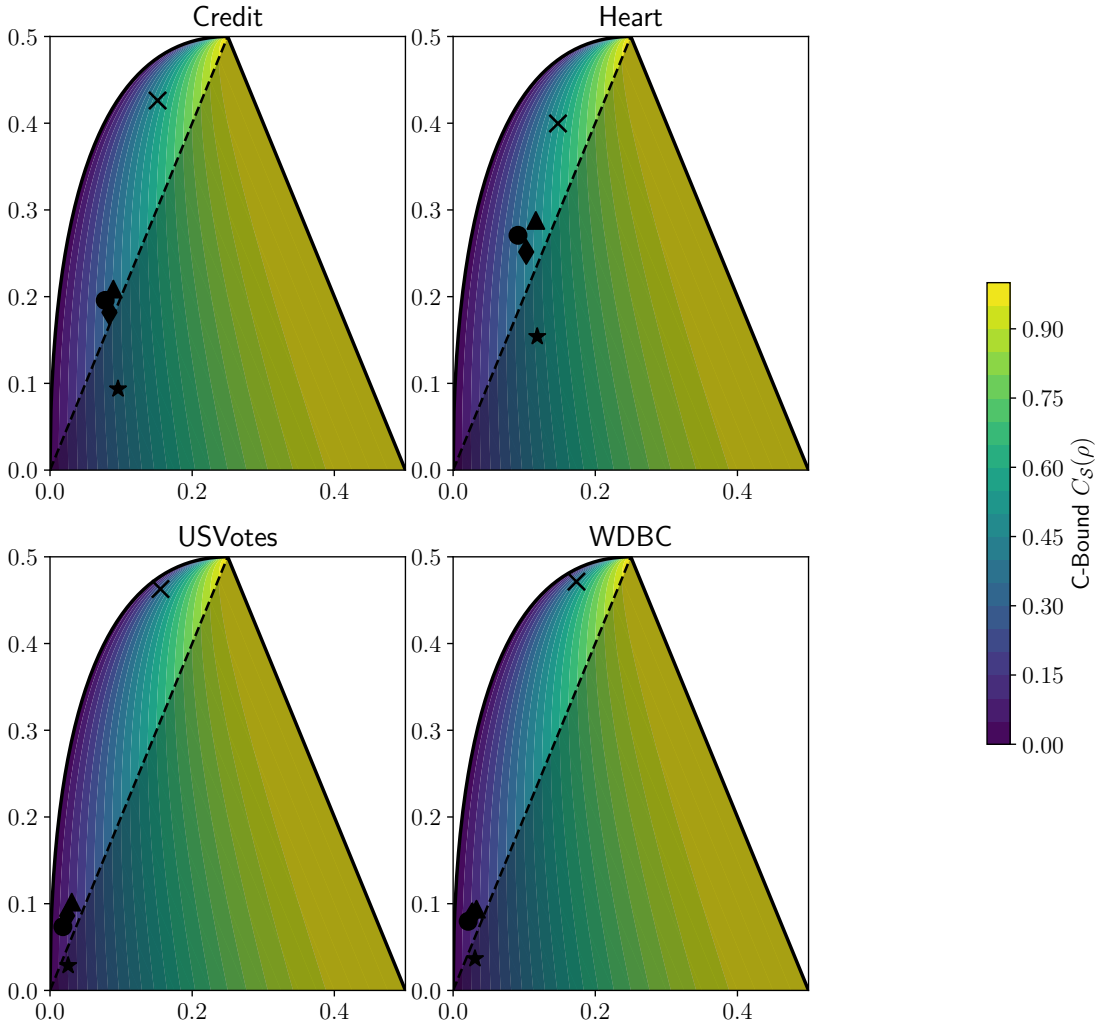


Figure D.4. *Representation of all the possible values of the empirical C-Bound $C_S(\rho)$ in function of the disagreement $d_S(\rho)$ (y-axis) and joint error $e_S(\rho)$ (x-axis). We report the values obtained with the decision trees in the binary setting by Algorithm 4.3 (⧫), `Masegosa` (▲), `Germain` (★), `CB-Boost` (•), and `MinCq` (×). The disagreement and the joint are averaged over the 10 runs.*

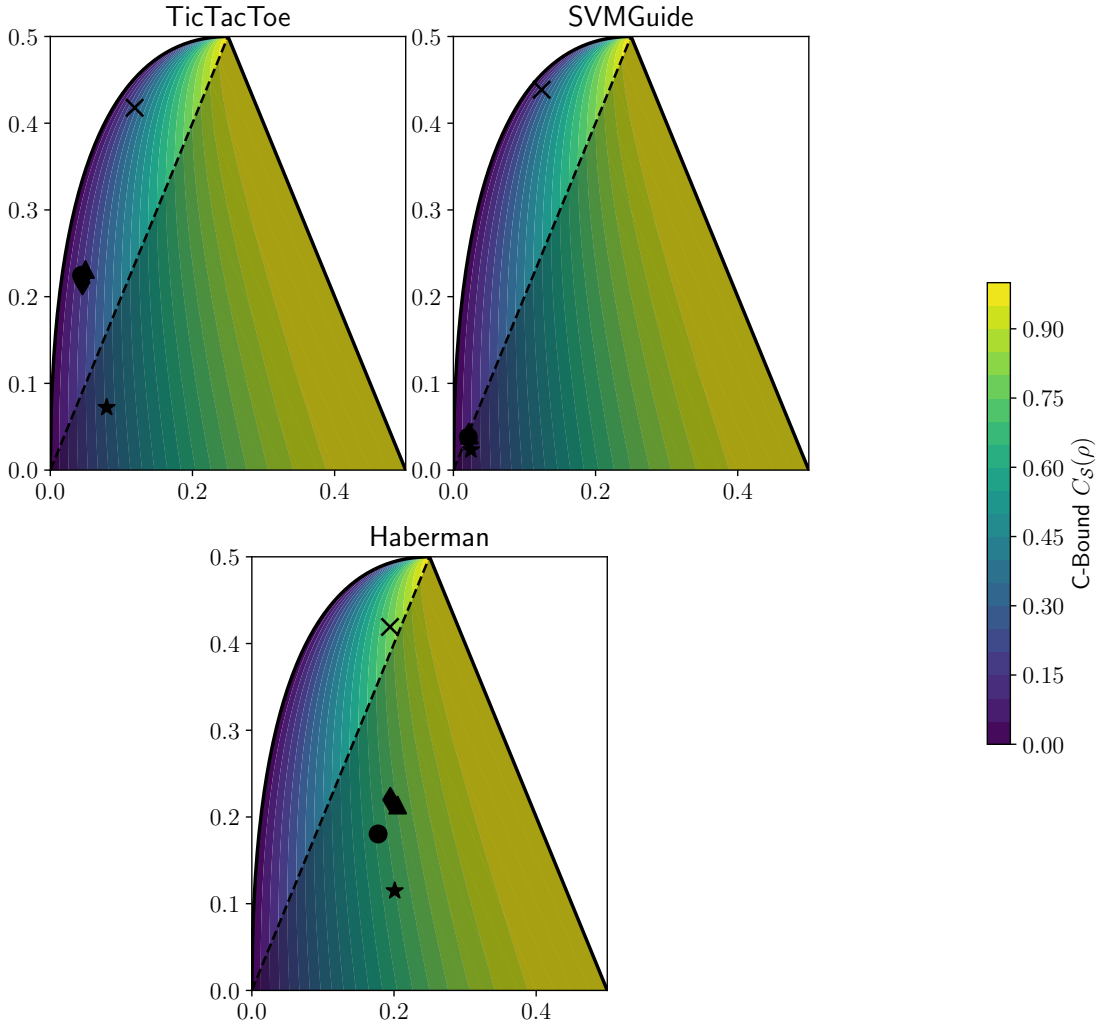


Figure D.5. *Representation of all the possible values of the empirical C-Bound $C_S(\rho)$ in function of the disagreement $d_S(\rho)$ (y-axis) and joint error $e_S(\rho)$ (x-axis). We report the values obtained with the decision trees in the binary setting by Algorithm 4.3 (⧫),* `Masegosa` *(▲),* `Germain` *(★),* `CB-Boost` *(●), and* `MinCq` *(×). The disagreement and the joint are averaged over the 10 runs.*

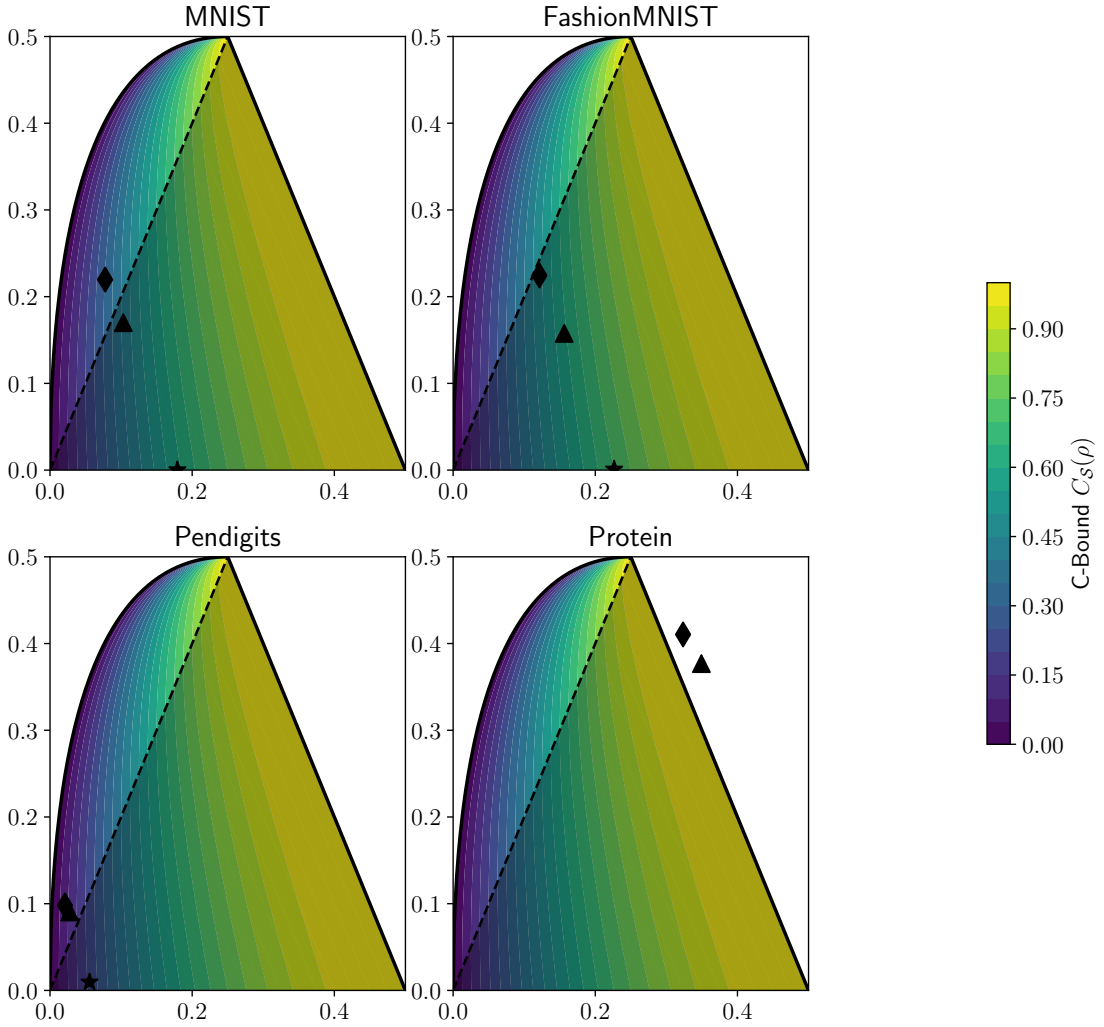


Figure D.6. *Representation of all the possible values of the empirical C-Bound $C_S(\rho)$ in function of the disagreement $d_S(\rho)$ (y-axis) and joint error $e_S(\rho)$ (x-axis). We report the values obtained with the decision trees in the multi-class setting by Algorithm 4.3 (♦),* `Masegosa` *(▲),* `Germain` *(★). The disagreement and the joint are averaged over the 10 runs.*

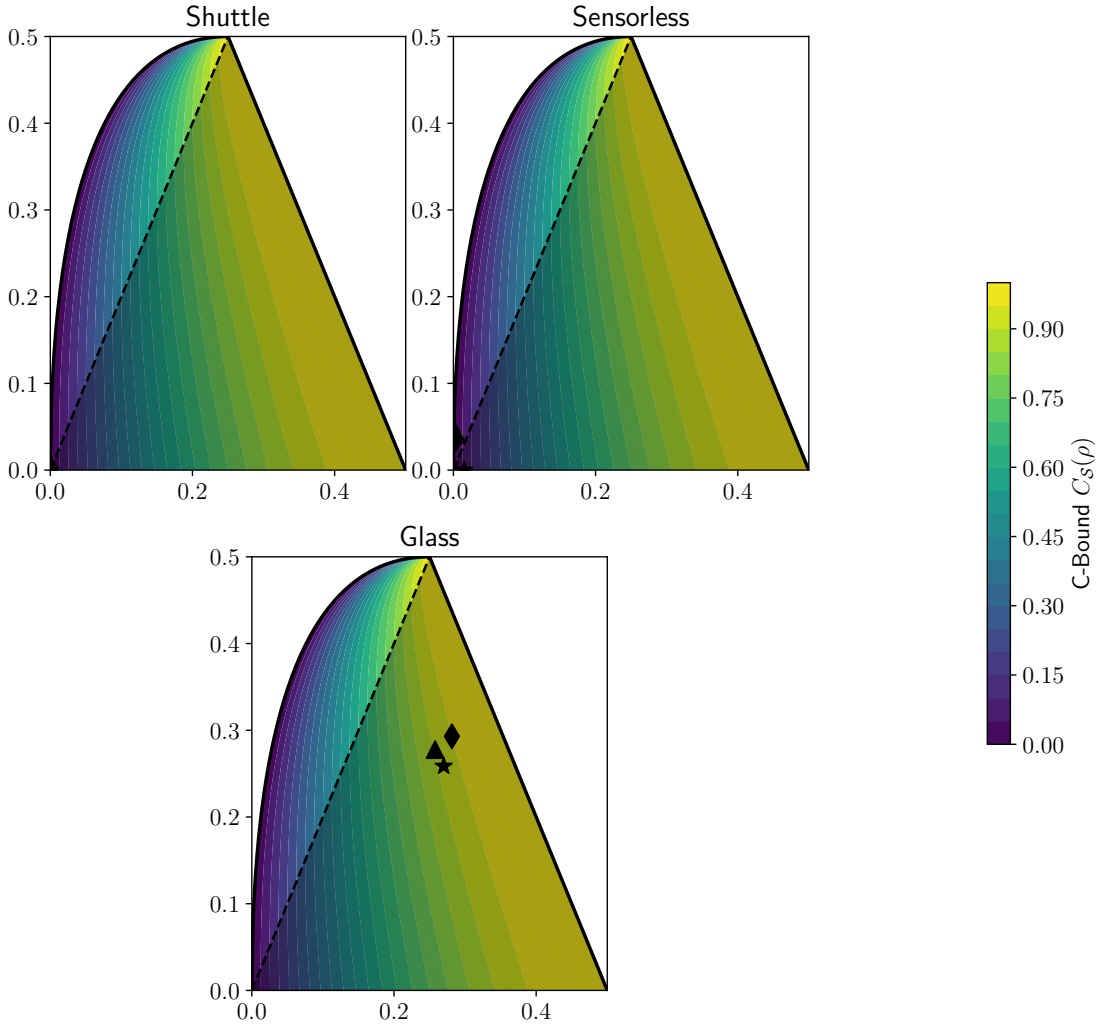


Figure D.7. *Representation of all the possible values of the empirical C-Bound $C_S(\rho)$ in function of the disagreement $d_S(\rho)$ (y-axis) and joint error $e_S(\rho)$ (x-axis). We report the values obtained with the decision trees in the multi-class setting by Algorithm 4.3 (♦),* `Masegosa` *(▲),* `Germain` *(★). The disagreement and the joint are averaged over the 10 runs.*

Table D.1. *Comparison of the true risks $R_{\mathcal{T}}(MV_\rho)$ and bound values obtained for each algorithm over 10 runs when the voters are decision stumps in the binary setting. More precisely, we report the mean $\pm$ the standard deviation. "Bound" is the mean value of the bound that is optimized, excepted for* `MinCq` *and* `CB-Boost` *for which we report the bound obtained with Theorem 4.4.1 instantiated with the majority vote learned. Results in* **bold** *are the couple ($R_{\mathcal{T}}(MV_\rho)$, Bound) associated to* **the lowest risk** *value. Italic and underlined results are the couple ($R_{\mathcal{T}}(MV_\rho)$, Bound) associated respectively to the lowest bound value and the second lowest bound values.*

	Algorithm 4.1		Algorithm 4.2		Algorithm 4.3		`Masegosa`	
	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound
Credit	.141 ± .014	.772 ± .026	.141 ± .014	.718 ± .036	.141 ± .014	.748 ± .037	.141 ± .014	.784 ± .047
Heart	.252 ± .034	.970 ± .011	.252 ± .034	.960 ± .013	**.161 ± .028**	**.937 ± .017**	.163 ± .025	1.041 ± .034
USVotes	.043 ± .007	.657 ± .016	.043 ± .007	.494 ± .034	**.042 ± .008**	**.529 ± .029**	**.042 ± .008**	**.520 ± .030**
WDBC	.101 ± .030	.722 ± .016	.115 ± .031	.675 ± .050	.069 ± .014	.578 ± .024	.060 ± .010	.533 ± .021
TicTacToe	.296 ± .011	.969 ± .008	.296 ± .011	.967 ± .009	.303 ± .015	.958 ± .009	.272 ± .024	1.021 ± .016
SVMGuide	.085 ± .007	.463 ± .028	.085 ± .007	.385 ± .025	.081 ± .014	.325 ± .011	.076 ± .012	.313 ± .011
Haberman	.266 ± .025	.975 ± .012	.263 ± .025	.968 ± .015	.262 ± .024	.988 ± .020	**.260 ± .019**	**1.207 ± .050**

	`Germain`		`CB-Boost`		`MinCq`	
	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound
Credit	*.141 ± .014*	*.462 ± .033*	**.140 ± .015**	**.917 ± .049**	.141 ± .020	1.000 ± .000
Heart	*.257 ± .026*	*.796 ± .036*	.191 ± .016	.996 ± .007	.185 ± .017	1.000 ± .000
USVotes	*.068 ± .040*	*.351 ± .108*	.043 ± .007	.683 ± .038	.048 ± .009	1.000 ± .000
WDBC	*.105 ± .023*	*.412 ± .032*	**.044 ± .008**	**.687 ± .044**	.044 ± .013	1.000 ± .000
TicTacToe	*.296 ± .011*	*.812 ± .023*	.202 ± .020	1.000 ± .000	**.020 ± .003**	**1.000 ± .000**
SVMGuide	*.102 ± .035*	*.256 ± .086*	.068 ± .012	.334 ± .012	**.048 ± .002**	**1.000 ± .000**
Haberman	*.265 ± .026*	*.811 ± .046*	.274 ± .029	.988 ± .011	.261 ± .016	1.000 ± .000

Table D.2. *Comparison of the true risks $R_{\mathcal{T}}(MV_\rho)$ and bound values obtained for each algorithm over 10 runs when the voters are decision trees in the binary setting. More precisely, we report the mean $\pm$ the standard deviation. "Bound" is the mean value of the bound that is optimized, excepted for* `MinCq` *and* `CB-Boost` *for which we report the bound obtained with Theorem 4.4.1 instantiated with the majority vote learned. Results in* **bold** *are the couple ($R_{\mathcal{T}}(MV_\rho)$, Bound) associated to* **the lowest risk** *value. Italic and underlined results are the couple ($R_{\mathcal{T}}(MV_\rho)$, Bound) associated respectively to the lowest bound value and the second lowest bound values.*

	Algorithm 4.1		Algorithm 4.2		Algorithm 4.3		`Masegosa`	
	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound
Credit	.156 ± .018	.861 ± .034	.154 ± .021	.807 ± .048	.142 ± .015	.744 ± .057	**.138 ± .015**	**.775 ± .081**
Heart	.230 ± .027	.984 ± .011	.228 ± .025	.976 ± .018	**.210 ± .026**	**.956 ± .025**	.213 ± .026	1.161 ± .076
USVotes	.053 ± .011	.738 ± .031	.057 ± .012	.571 ± .067	**.046 ± .014**	**.523 ± .059**	.051 ± .023	.513 ± .069
WDBC	.054 ± .010	.705 ± .022	.061 ± .010	.554 ± .041	.045 ± .007	.485 ± .042	.049 ± .011	.471 ± .050
TicTacToe	.056 ± .010	.776 ± .025	.078 ± .023	.685 ± .047	**.048 ± .013**	**.530 ± .052**	.050 ± .010	.493 ± .054
SVMGuide	.033 ± .002	.340 ± .011	.033 ± .001	.213 ± .012	.032 ± .002	.177 ± .012	**.032 ± .002**	**.165 ± .012**
Haberman	.307 ± .031	.998 ± .004	.307 ± .030	.997 ± .006	.295 ± .025	.997 ± .006	**.294 ± .018**	**1.586 ± .113**

	`Germain`		`CB-Boost`		`MinCq`	
	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound
Credit	*.167 ± .023*	*.564 ± .059*	.147 ± .017	.772 ± .054	.140 ± .021	.981 ± .025
Heart	*.234 ± .022*	*.840 ± .059*	.222 ± .026	.972 ± .019	.222 ± .029	1.000 ± .000
USVotes	*.056 ± .012*	*.334 ± .050*	.048 ± .019	.578 ± .050	.053 ± .010	.989 ± .032
WDBC	*.063 ± .010*	*.324 ± .033*	**.044 ± .012**	**.525 ± .042**	.054 ± .013	.984 ± .041
TicTacToe	*.127 ± .025*	*.457 ± .047*	.049 ± .014	.545 ± .052	.049 ± .013	.901 ± .080
SVMGuide	*.038 ± .005*	*.123 ± .008*	.033 ± .002	.183 ± .012	.033 ± .002	.660 ± .185
Haberman	*.306 ± .030*	*.970 ± .058*	.295 ± .024	.999 ± .002	.296 ± .016	1.000 ± .000

Table D.3. *Comparison of the true risks $R_{\mathcal{T}}(MV_\rho)$ and bound values obtained for each algorithm over 10 runs when the voters are decision trees in the multi-class setting. More precisely, we report the mean $\pm$ the standard deviation. Results in* **bold** *are the couple ($R_{\mathcal{T}}(MV_\rho)$, Bound) associated to* **the lowest risk** *value. Italic and underlined results are the couple ($R_{\mathcal{T}}(MV_\rho)$, Bound) associated respectively to the lowest bound value and the second lowest bound values.*

	Algorithm 4.1		Algorithm 4.2		Algorithm 4.3	
	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound
MNIST	**.034 ± .002**	**.387 ± .003**	.034 ± .002	.371 ± .003	*.034 ± .002*	*.340 ± .004*
FashionMNIST	**.121 ± .004**	**.554 ± .003**	.121 ± .004	.544 ± .003	.121 ± .004	.523 ± .003
Pendigits	.015 ± .007	.283 ± .007	.015 ± .007	.197 ± .009	***.015 ± .007***	***.137 ± .009***
Protein	*.467 ± .060*	*1.000 ± .000*	*.462 ± .057*	*1.000 ± .000*	***.419 ± .038***	***1.000 ± .000***
Shuttle	.000 ± .000	.061 ± .001	.000 ± .000	.006 ± .001	**.000 ± .000**	**.005 ± .000**
Sensorless	.002 ± .001	.097 ± .001	.002 ± .000	.040 ± .001	***.001 ± .000***	***.022 ± .001***
Glass	.328 ± .058	1.000 ± .000	.327 ± .058	1.000 ± .000	*.328 ± .060*	*1.000 ± .000*

	`Masegosa`		`Germain`	
	$R_{\mathcal{T}}(MV_\rho)$	Bound	$R_{\mathcal{T}}(MV_\rho)$	Bound
MNIST	.095 ± .037	.451 ± .062	.180 ± .005	.382 ± .003
FashionMNIST	.195 ± .022	.672 ± .070	*.231 ± .004*	*.479 ± .002*
Pendigits	.027 ± .017	.168 ± .030	.064 ± .022	.163 ± .010
Protein	.460 ± .025	1.503 ± .047	.529 ± .003	1.098 ± .006
Shuttle	.000 ± .000	.005 ± .001	*.001 ± .000*	*.003 ± .001*
Sensorless	.002 ± .001	.025 ± .001	.015 ± .001	.040 ± .002
Glass	**.324 ± .056**	**1.978 ± .140**	.327 ± .058	1.263 ± .081

Appendix of Chapter 5

E

E.1 About Equation (5.1)

For the sake of completeness, we prove the bound $\mathsf{R}_{\mathcal{D}'}(\mathsf{MV}_\rho) \leq 2b^N_{\mathcal{D}'}(\rho)$ in the following lemma.

Lemma E.1.1. For any distribution $\mathcal{D}'$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution ρ on $\mathbb{H}$, for any $N \in \mathbb{N}_*$, we have

$$\mathsf{R}_{\mathcal{D}'}(\mathsf{MV}_\rho) \leq 2b^N_{\mathcal{D}'}(\rho)$$
$$= 2 \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \left[\sum_{j=\lceil \frac{N}{2} \rceil}^{N} \binom{N}{j} \left[\frac{1}{2}\left(1-\widehat{m_\rho}(\mathbf{x},y)\right)\right]^j \left[1-\frac{1}{2}\left(1-\widehat{m_\rho}(\mathbf{x},y)\right)\right]^{(N-j)} \right].$$

Proof. The proof is based on Lacasse *et al.* (2010). Note that for a given example $(\mathbf{x}, y) \sim \mathcal{D}'$ *s.t.* $\widehat{m_\rho}(\mathbf{x}, y) = 0$, we have

$$\underbrace{\sum_{j=\lceil \frac{N}{2} \rceil}^{N} \binom{N}{j} \left[\frac{1}{2}\left(1-\widehat{m_\rho}(\mathbf{x},y)\right)\right]^j \left[1-\frac{1}{2}\left(1-\widehat{m_\rho}(\mathbf{x},y)\right)\right]^{(N-j)}}_{\triangleq \heartsuit^N_\rho(\mathbf{x},y)} \geq \frac{1}{2}.$$

Moreover, $\heartsuit^N_\rho(\mathbf{x}, y)$ is monotonically decreasing in $\widehat{m_\rho}(\mathbf{x}, y)$. From these two properties, we have for a given example $(\mathbf{x}, y) \sim \mathcal{D}'$ *s.t.* $\widehat{m_\rho}(\mathbf{x}, y) \leq 0$

$$\underbrace{\mathrm{I}\left[\widehat{m_\rho}(\mathbf{x},y) \leq 0\right]}_{=1} \leq 2\heartsuit^N_\rho(\mathbf{x},y),$$

and for a given example $(\mathbf{x}, y) \sim \mathcal{D}'$ *s.t.* $\widehat{m_\rho}(\mathbf{x}, y) > 0$

$$\underbrace{\mathrm{I}\left[\widehat{m_\rho}(\mathbf{x},y) \leq 0\right]}_{=0} \leq 2\heartsuit^N_\rho(\mathbf{x},y).$$

Hence, we can deduce that

$$\mathsf{R}_{\mathcal{D}'}(\mathsf{MV}_\rho) \leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \mathrm{I}\left[\widehat{m_\rho}(\mathbf{x},y) \leq 0\right] \leq 2 \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}'} \heartsuit_\rho^N(\mathbf{x},y) = 2b_{\mathcal{D}'}^N(\rho).$$

■

Based on Lemma E.1.1, we are now able to prove a PAC-Bayesian bound on the majority vote's true risk based on the surrogate $b_{\mathcal{D}'}^N(\rho)$. Note that Lacasse *et al.* (2010) prove a Catoni-like bound while our work is based on a Seeger-like one. Our bound avoids using a parameter $c \geq 0$ required in Catoni's bound. We derive the bound in the following theorem.

Theorem E.1.1. For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$ we have

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq 2\left[\overline{\mathrm{kl}}\left(b_{\mathcal{S}}^N(\rho) \;\middle|\; \frac{1}{m}\left[N\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]\right)\right]. \quad \text{(E.1)}$$

Proof. First, note that we have by definition

$$b_{\mathcal{D}}^N(\rho) = \mathop{\mathbb{P}}_{\mathsf{MV}_\sigma\sim\rho^N}\left[\widehat{m_\sigma}(\mathbf{x},y) \leq 0\right].$$

We apply Theorem 2.3.4 with the loss $\ell(h,(\mathbf{x},y)) = \mathrm{I}\left[\widehat{m_\sigma}(\mathbf{x},y) \leq 0\right]$, the posterior distribution ρ^N and the prior distribution π^N to have

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad b_{\mathcal{D}}^N(\rho) \leq \overline{\mathrm{kl}}\left(b_{\mathcal{S}}^N(\rho) \;\middle|\; \frac{1}{m}\left[\mathrm{KL}(\rho^N\|\pi^N) + \ln\frac{2\sqrt{m}}{\delta}\right]\right),$$

where $\mathrm{KL}(\rho^N\|\pi^N) = N\,\mathrm{KL}(\rho\|\pi)$. Then, from Lemma E.1.1, we obtain

$$\forall \rho \in \mathbb{M}(\mathbb{H}), \quad \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq 2\left[\overline{\mathrm{kl}}\left(b_{\mathcal{S}}^N(\rho) \;\middle|\; \frac{1}{m}\left[N\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]\right)\right].$$

■

From Theorem E.1.1, we can deduce a self-bounding algorithm denoted as Lacasse to minimize the majority vote's true risk. The objective function is defined as

$$G_{\mathcal{U}}^{\mathtt{L}}(\rho) = 2\left[\overline{\mathrm{kl}}\left(b_{\mathcal{U}}^N(\rho) \;\middle|\; \frac{1}{m}\left[N\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]\right)\right],$$

where $b_{\mathbb{S}}^{N}(\rho)$ is approximated through a mini-batch $\mathbb{U} \subseteq \mathbb{S}$ (and $b_{\mathcal{U}}^{N}(\rho)$ is computed instead). The algorithm denoted as `Lacasse` is described in Algorithm E.1 below.

Algorithm E.1 Minimization of Equation (E.1) by Stochastic Gradient Descent

Given: learning sample $\mathbb{S}$, prior distribution π on $\mathbb{H}$, the objective function $G_{\mathcal{U}}^{\mathrm{L}}(\rho)$
Hyperparameters: number of iterations T
$\rho \leftarrow \pi$
for $t \leftarrow 1$ to T **do**
  for all mini-batches $\mathbb{U} \subseteq \mathbb{S}$ **do**
    $\rho \leftarrow$ Update ρ with $G_{\mathcal{U}}^{\mathrm{L}}(\rho)$ by gradient descent

E.2 Aggregation Property of the Dirichlet Distributions

Lemma E.2.1. Let $\rho \sim \mathrm{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in (\mathbb{R}_*^+)^K$ and $K = \mathrm{card}(\mathbb{H})$. Then, the random variable associated to the random variable ρ where we sum the entries i and j follows a Dirichlet distribution with parameters $\boldsymbol{\alpha}' = [\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_K]^\top \in (\mathbb{R}_*^+)^{K-1}$, *i.e.*, we have

$$\Big(\rho(h_1), \ldots, \rho(h_i) + \rho(h_j), \ldots, \rho(h_K)\Big) \sim \mathrm{Dir}(\boldsymbol{\alpha}').$$

Proof. Without loss of generality let $i = 1$ and $j = 2$, then, first remark that we have

$$\begin{aligned}
&\int_0^{\rho(h_1)+\rho(h_2)} [x]^{\alpha_1-1} \, [\rho(h_1)+\rho(h_2)-x]^{\alpha_2-1} \, dx \\
&= \int_0^1 [(\rho(h_1)+\rho(h_2))x']^{\alpha_1-1} \, [(\rho(h_1)+\rho(h_2))(1-x')]^{\alpha_2-1} \, (\rho(h_1)+\rho(h_2)) dx' \\
&= [\rho(h_1)+\rho(h_2)]^{\alpha_1+\alpha_2-1} \int_0^1 [x']^{\alpha_1-1} \, [1-x']^{\alpha_2-1} \, dx' \\
&\propto [\rho(h_1)+\rho(h_2)]^{\alpha_1+\alpha_2-1} . \qquad \text{(E.2)}
\end{aligned}$$

Then, from Equation (E.2), the probability density function of $\Big(\rho(h_1), \ldots, \rho(h_i) +$

$\rho(h_j), \dots, \rho(h_K)\Big)$ can be rewritten (up to the normalization constant) as

$$\left(\int_0^{\rho(h_1)+\rho(h_2)} [x]^{\alpha_1-1} [\rho(h_1)+\rho(h_2)-x]^{\alpha_2-1}\, dx\right)\left(\prod_{i=3}^{K} [\rho(h_i)]^{\alpha_i-1}\right)$$
$$\propto [\rho(h_1)+\rho(h_2)]^{\alpha_1+\alpha_2-1} \prod_{i=3}^{K} [\rho(h_i)]^{\alpha_i-1},$$

which proves the desired result. ■

E.3 Proof of Lemma 5.2.1

Lemma 5.2.1 (Computation of the Stochastic Risk)**.** For a given $(\mathbf{x}, y) \in \mathbb{X}\times\mathbb{Y}$, let

$$\mathbb{F}(\mathbf{x}, y) = \{j \ : \ h_j(\mathbf{x}) \neq y\} \quad \text{and} \quad \mathbb{T}(\mathbf{x}, y) = \{j \ : \ h_j(\mathbf{x}) = y\}$$

be respectively the set of indices of the voters that misclassify $(\mathbf{x}, y)$ and the set of indices of the voters that correctly classify $(\mathbf{x}, y)$. Then, the stochastic risk $s_{\mathrm{P}}(\mathbf{x}, y)$ can be rewritten as

$$s_{\mathrm{P}}(\mathbf{x}, y) = \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathrm{I}\left[\widehat{m_\rho}(\mathbf{x}, y) \leq 0\right] = I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right),$$

with $I_{0.5}()$ the regularized incomplete beta function evaluated at 0.5. It is defined as

$$I_{0.5}(a, b) \triangleq \frac{B_{0.5}(a, b)}{B_1(a, b)}, \quad \text{where} \quad B_t(a, b) \triangleq \int_0^t x^{a-1}(1-x)^{b-1}dx$$

is the incomplete beta function.

Proof. Given an example $(\mathbf{x}, y) \in \mathbb{X} \times \mathbb{Y}$, by definition of the set $\mathbb{F}(\mathbf{x}, y)$ and

$\mathbb{T}(\mathbf{x}, y)$, we have

$$\begin{aligned}\mathop{\mathbb{P}}_{h\sim\rho}[h(\mathbf{x}) \neq y] &= \sum_{j=1}^{n} \rho(h_j)\, \mathrm{I}\left[h_j(\mathbf{x}) \neq y\right] \\ &= \sum_{j\in\mathbb{F}(\mathbf{x},y)} \rho(h_j), \\ \text{and} \quad \mathop{\mathbb{P}}_{h\sim\rho}[h(\mathbf{x}) = y] &= \sum_{j=1}^{n} \rho(h_j)\, \mathrm{I}\left[h_j(\mathbf{x}) = y\right] \\ &= \sum_{j\in\mathbb{T}(\mathbf{x},y)} \rho(h_j).\end{aligned}$$

Moreover, by definition of the Dirichlet distribution (Definition 5.2.2), we have

$$\rho \sim \mathrm{P} \iff (\rho(h_1), \ldots, \rho(h_n)) \sim \mathrm{Dir}(\boldsymbol{\alpha}).$$

Then, we use the aggregation property of the Dirichlet distributions (Lemma E.2.1) to obtain

$$\begin{aligned}&\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \rho(h_j), \sum_{j\in\mathbb{F}(\mathbf{x},y)} \rho(h_j)\right) \sim \mathrm{Dir}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right) \\ \iff &\left(\mathop{\mathbb{P}}_{h\sim\rho}[h(\mathbf{x}) = y], \mathop{\mathbb{P}}_{h\sim\rho}[h(\mathbf{x}) \neq y]\right) \sim \mathrm{Dir}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right).\end{aligned}$$

Thus, $(\mathbb{P}_{h\sim\rho}[h(\mathbf{x}) = y], \mathbb{P}_{h\sim\rho}[h(\mathbf{x}) \neq y])$ follows a bivariate Dirichlet distribution *a.k.a.* Beta distribution. We have

$$\mathop{\mathbb{P}}_{h\sim\rho}[h(\mathbf{x}) = y] \sim \mathrm{Beta}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right).$$

Finally, notice that the expected error is related to the cumulative probability function of the random variable $\mathbb{P}_{h\sim\rho}[h(\mathbf{x}) = y]$ which is the regularized incomplete beta function $I_p : \mathbb{R}_*^+ \times \mathbb{R}_*^+ \to [0, 1]$:

$$\mathbb{P}\left[\mathop{\mathbb{P}}_{h\sim\rho}[h(\mathbf{x}) = y] \leq 0.5\right] = I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right).$$

Then, we have

$$\begin{aligned} s_{\mathrm{P}}(\mathbf{x}, y) &= \mathop{\mathbb{P}}_{\rho\sim\mathrm{P}} \mathrm{I}\left[\widehat{m_\rho}(\mathbf{x}, y) \leq 0\right] \\ &= \mathbb{P}\left[\mathop{\mathbb{P}}_{h\sim\rho}\left[h(\mathbf{x}) = y\right] \leq 0.5\right] \\ &= I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right). \end{aligned}$$

■

E.4 Proof of Corollary 5.2.1

Corollary 5.2.1 (Closed-form Solution of the Stochastic Risks)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any learning sample $\mathcal{S} \sim \mathcal{D}^m$, for any finite hypothesis set $\mathbb{H}$, for any distribution $\mathrm{P} = \mathrm{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in (\mathbb{R}_*^+)^{\mathrm{card}(\mathbb{H})}$, we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}}(\mathbf{x}, y) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right),$$

and

$$\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathsf{R}_{\mathcal{S}}(\mathsf{MV}_\rho) \leq \frac{1}{m}\sum_{i=1}^{m} s_{\mathrm{P}}(\mathbf{x}_i, y_i) = \frac{1}{m}\sum_{i=1}^{m} I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x}_i,y_i)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x}_i,y_i)} \alpha_j\right).$$

Proof. From Equation (5.2) and Lemma 5.2.1, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) &\leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}}(\mathbf{x}, y) \\ &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right). \end{aligned}$$

Similarly, from Equation (5.3) and Lemma 5.2.1, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathsf{R}_{\mathcal{S}}(\mathsf{MV}_\rho) &\leq \frac{1}{m}\sum_{i=1}^{m} s_{\mathrm{P}}(\mathbf{x}_i, y_i) \\ &= \frac{1}{m}\sum_{i=1}^{m} I_{0.5}\left(\sum_{j\in\mathbb{T}(\mathbf{x},y)} \alpha_j, \sum_{j\in\mathbb{F}(\mathbf{x},y)} \alpha_j\right). \end{aligned}$$

■

E.5 Proof of Theorem 5.3.1

Theorem 5.3.1 (PAC-Bayesian Bound for Stochastic Majority Votes)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any finite hypothesis set $\mathbb{H}$, for any distribution $\Pi = \mathrm{Dir}(\boldsymbol{\beta})$ with $\boldsymbol{\beta} \in (\mathbb{R}_*^+)^{\mathrm{card}(\mathbb{H})}$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$, we have for all hyper-posterior P on $\mathbb{H}$

$$\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) \leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}}(\mathbf{x},y) \leq \overline{\mathrm{kl}}\left(\frac{1}{m}\sum_{i=1}^{m} s_{\mathrm{P}}(\mathbf{x}_i,y_i) \;\middle|\; \frac{\mathrm{KL}(\mathrm{P}\|\Pi) + \ln\frac{2\sqrt{m}}{\delta}}{m}\right),$$

$$\text{with}\quad \mathrm{KL}(\mathrm{P}\|\Pi) = \sum_{j=1}^{\mathrm{card}(\mathbb{H})} \ln[\Gamma(\beta_j)] - \ln\left[\Gamma\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\beta_j\right)\right] - \sum_{j=1}^{\mathrm{card}(\mathbb{H})} \ln[\Gamma(\alpha_j)]$$
$$+ \ln\left[\Gamma\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\alpha_j\right)\right] + \sum_{j=1}^{\mathrm{card}(\mathbb{H})} (\alpha_j-\beta_j)\left[\psi(\alpha_j) - \psi\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\alpha_j\right)\right],$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ is the Gamma function and the Digamma function $\Psi(\alpha)$ is defined as the derivative of $\ln[\Gamma(\alpha)]$; these two functions are plotted in Figure 5.4.

Proof. This is a direct application of Theorem 2.3.4 (and Definition 2.3.3). Indeed, we apply Theorem 2.3.4 with the loss $\ell : \mathbb{M}(\mathbb{H}) \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$ defined by $\ell(\rho,(\mathbf{x},y)) = \mathrm{I}\left[\widehat{m_\rho}(\mathbf{x},y) \leq 0\right]$.

Moreover, the closed-form solution of the KL divergence is

$$
\begin{aligned}
\mathrm{KL}(\Pi\|\mathrm{P}) &= \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \ln\left(\frac{\mathrm{P}(\rho)}{\Pi(\rho)}\right) \\
&= \mathop{\mathbb{E}}_{\rho\sim\mathrm{Dir}(\boldsymbol{\alpha})} \ln\left(\frac{Z(\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})}\frac{\prod_{j=1}^{\mathrm{card}(\mathbb{H})}[\rho(h_j)]^{\alpha_j-1}}{\prod_{j=1}^{\mathrm{card}(\mathbb{H})}[\rho(h_j)]^{\beta_j-1}}\right) \\
&= \mathop{\mathbb{E}}_{\rho\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\ln Z(\boldsymbol{\beta}) - \ln Z(\boldsymbol{\alpha}) + \ln\left(\prod_{j=1}^{\mathrm{card}(\mathbb{H})}[\rho(h_j)]^{\alpha_j-\beta_j}\right)\right] \\
&= \mathop{\mathbb{E}}_{\rho\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left[\ln Z(\boldsymbol{\beta}) - \ln Z(\boldsymbol{\alpha}) + \sum_{j=1}^{\mathrm{card}(\mathbb{H})}(\alpha_j-\beta_j)\ln[\rho(h_j)]\right] \\
&= \ln Z(\boldsymbol{\beta}) - \ln Z(\boldsymbol{\alpha}) + \sum_{j=1}^{\mathrm{card}(\mathbb{H})}(\alpha_j-\beta_j)\mathop{\mathbb{E}}_{\rho\sim\mathrm{Dir}(\boldsymbol{\alpha})}\ln[\rho(h_j)] \\
&= \sum_{j=1}^{\mathrm{card}(\mathbb{H})}\ln[\Gamma(\beta_j)] - \ln\left[\Gamma\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\beta_j\right)\right] \\
&\quad - \sum_{j=1}^{\mathrm{card}(\mathbb{H})}\ln[\Gamma(\alpha_j)] + \ln\left[\Gamma\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\alpha_j\right)\right] \\
&\quad + \sum_{j=1}^{\mathrm{card}(\mathbb{H})}(\alpha_j-\beta_j)\left[\psi(\alpha_j) - \psi\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\alpha_j\right)\right].
\end{aligned}
$$

The last equality follows by definition of Dirichlet's geometric mean

$$
\mathop{\mathbb{E}}_{\rho\sim\mathrm{P}}\ln[\rho(h_j)] = \psi(\alpha_j) - \psi\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\alpha_j\right).
$$

and the normalization constant

$$
\ln(Z(\boldsymbol{\alpha})) = \sum_{j=1}^{\mathrm{card}(\mathbb{H})}\ln[\Gamma(\alpha_j)] - \ln\left[\Gamma\left(\sum_{j=1}^{\mathrm{card}(\mathbb{H})}\alpha_j\right)\right].
$$

■

E.6 Proof of Theorem 5.3.2

Theorem 5.3.2 (PAC-Bayesian bound with data-dependent voters)**.** Let Π_1 and P_1 be the hyper-prior and hyper-posterior distributions on $\mathbb{H}_1$ defined with $\mathbb{S}_1$, and Π_2 and P_2 the prior and posterior distributions on $\mathbb{H}_2$ defined with $\mathbb{S}_2$. For any $\lambda \in [0,1]$ and $\delta \in (0,1]$ with probability at least $1-\delta$ over samples $\mathbb{S}_1 \sim \mathcal{D}^{m_1}$ and $\mathbb{S}_2 \sim \mathcal{D}^{m_2}$, we have

$$\lambda \mathop{\mathbb{E}}_{\rho\sim \mathrm{P}_1} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_\rho) + (1-\lambda) \mathop{\mathbb{E}}_{\rho'\sim \mathrm{P}_2} \mathsf{R}_{\mathcal{D}}(\mathsf{MV}_{\rho'}) \leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [\lambda s_{\mathrm{P}_1}(\mathbf{x},y) + (1-\lambda) s_{\mathrm{P}_2}(\mathbf{x},y)] \leq$$
$$\overline{\mathrm{kl}}\left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}_1} \frac{s_{\mathrm{P}_1}(\mathbf{x},y)}{\frac{1}{\lambda}} + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}_2} \frac{s_{\mathrm{P}_2}(\mathbf{x},y)}{\frac{1}{1-\lambda}} \;\middle|\; \frac{\mathrm{KL}(\mathrm{P}_1\|\Pi_1) + \ln\frac{4\sqrt{m}}{\delta}}{\frac{m}{\lambda}} + \frac{\mathrm{KL}(\mathrm{P}_2\|\Pi_2) + \ln\frac{4\sqrt{m'}}{\delta}}{\frac{m'}{1-\lambda}} \right].$$

Proof. From the joint convexity of $\mathrm{kl}()$, we have for any $\lambda \in [0,1]$

$$\mathrm{kl}\left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \frac{s_{\mathrm{P}}(\mathbf{x},y)}{\frac{1}{\lambda}} + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}'} \frac{s_{\mathrm{P}'}(\mathbf{x},y)}{\frac{1}{1-\lambda}} \;\middle|\; \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \frac{s_{\mathrm{P}}(\mathbf{x},y)}{\frac{1}{\lambda}} + \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \frac{s_{\mathrm{P}'}(\mathbf{x},y)}{\frac{1}{1-\lambda}} \right]$$
$$\leq \lambda\, \mathrm{kl}\left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} s_{\mathrm{P}}(\mathbf{x},y) \,\middle\|\, \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}}(\mathbf{x},y) \right] + (1-\lambda)\, \mathrm{kl}\left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}'} s_{\mathrm{P}'}(\mathbf{x},y) \,\middle\|\, \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}'}(\mathbf{x},y) \right].$$

For all hyper-prior Π on $\mathbb{H}$, we have with probability at least $1-\frac{\delta}{2}$ on the random choice $\mathbb{S}\sim\mathcal{D}^m$ for all hyper-posterior P on $\mathbb{H}$

$$\mathrm{kl}\left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} s_{\mathrm{P}}(\mathbf{x},y) \,\middle\|\, \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}}(\mathbf{x},y) \right] \leq \frac{\mathrm{KL}(\mathrm{P},\Pi) + \ln\frac{4\sqrt{m}}{\delta}}{m}$$
$$\iff \forall \lambda \in [0,1], \quad \lambda\, \mathrm{kl}\left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} s_{\mathrm{P}}(\mathbf{x},y) \,\middle\|\, \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}}(\mathbf{x},y) \right] \leq \lambda \frac{\mathrm{KL}(\mathrm{P},\Pi) + \ln\frac{4\sqrt{m}}{\delta}}{m}. \tag{E.3}$$

Similarly, for all hyper-prior Π' on $\mathbb{H}'$, we have with probability at least $1-\frac{\delta}{2}$ on the random choice $\mathbb{S}'\sim\mathcal{D}^m$ for all hyper-posterior P' on $\mathbb{H}$

$$\mathrm{kl}\left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}'} s_{\mathrm{P}'}(\mathbf{x},y) \,\middle\|\, \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}'}(\mathbf{x},y) \right] \leq \frac{\mathrm{KL}(\mathrm{P}',\Pi') + \ln\frac{4\sqrt{m}}{\delta}}{m} \quad \iff$$
$$\forall \lambda \in [0,1], (1-\lambda)\, \mathrm{kl}\left[\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}'} s_{\mathrm{P}'}(\mathbf{x},y) \,\middle\|\, \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} s_{\mathrm{P}'}(\mathbf{x},y) \right] \leq (1-\lambda) \frac{\mathrm{KL}(\mathrm{P}',\Pi') + \ln\frac{4\sqrt{m'}}{\delta}}{m'}. \tag{E.4}$$

Combining Equation (E.3) and Equation (E.4) using the union bound, we obtain the desired result with $1 - \delta$ probability. ■

E.7 Additional Results

E.7.1 Choice of the prior

In the other experiments, we fixed the hyper-prior distribution Π (parameterized by $\boldsymbol{\beta}$) to the uniform, *i.e.* $\forall j \in \{1, \ldots, \mathrm{card}(\mathbb{H})\}, \beta_j = 1$. This choice was to make the comparison with the baselines as fair as possible, as their prior was also fixed to the uniform. However, we can bias the sparsity of the posterior, or conversely its concentration, by choosing a different value for the prior distribution parameters. In some cases, tuning the prior parameters allows to obtain better performance, as reported in Figures E.1 to E.4. As in Section 5.4, the hypothesis sets $\mathbb{H}_1$ *resp.* $\mathbb{H}_2$ are composed of 50 decision trees learned with $\mathbb{S}_1$ *resp.* $\mathbb{S}_2$ with no limit on the depth. In general, these results suggest that the choice of prior distribution has a high impact on the learned model's performance and tuning its concentration parameters would be a viable option for improving the results.

E.7.2 Impact of voter strength

We report a study on the impact of voter strength on the learned models. More precisely, we provide results for additional datasets as well as the study of the expected strength of a voter as a function of the tree maximal depth. The hypothesis sets $\mathbb{H}_1$ *resp.* $\mathbb{H}_2$ are composed of 50 decision trees learned with $\mathbb{S}_1$ *resp.* $\mathbb{S}_2$; the prior's parameters are set to $\beta_j = 1$ for all $j \in \{1, \ldots, \mathrm{card}(\mathbb{H})\}$. The maximal depth is a values belonging to the set $\{1, 2, 4, 8, 16\}$, *i.e.*, the maximal depth varies from 1 to 16. We report in Figures E.5 to E.8 the stochastic test risks $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{T}}\, s_{\mathrm{P}}(\mathbf{x}, y)$ or the test risk $\mathrm{R}_{\mathcal{T}}(\mathrm{MV}_\rho)$, their corresponding empirical risks and the bound values. We can see that limiting the maximal depth is an effective way for controlling the strength of the voters. Indeed, the general trend tells us that increasing the strength of the voters generally yields more powerful ensembles for all methods: the risks and the bounds are decreasing or stay constant when the tree maximal depth is increasing.

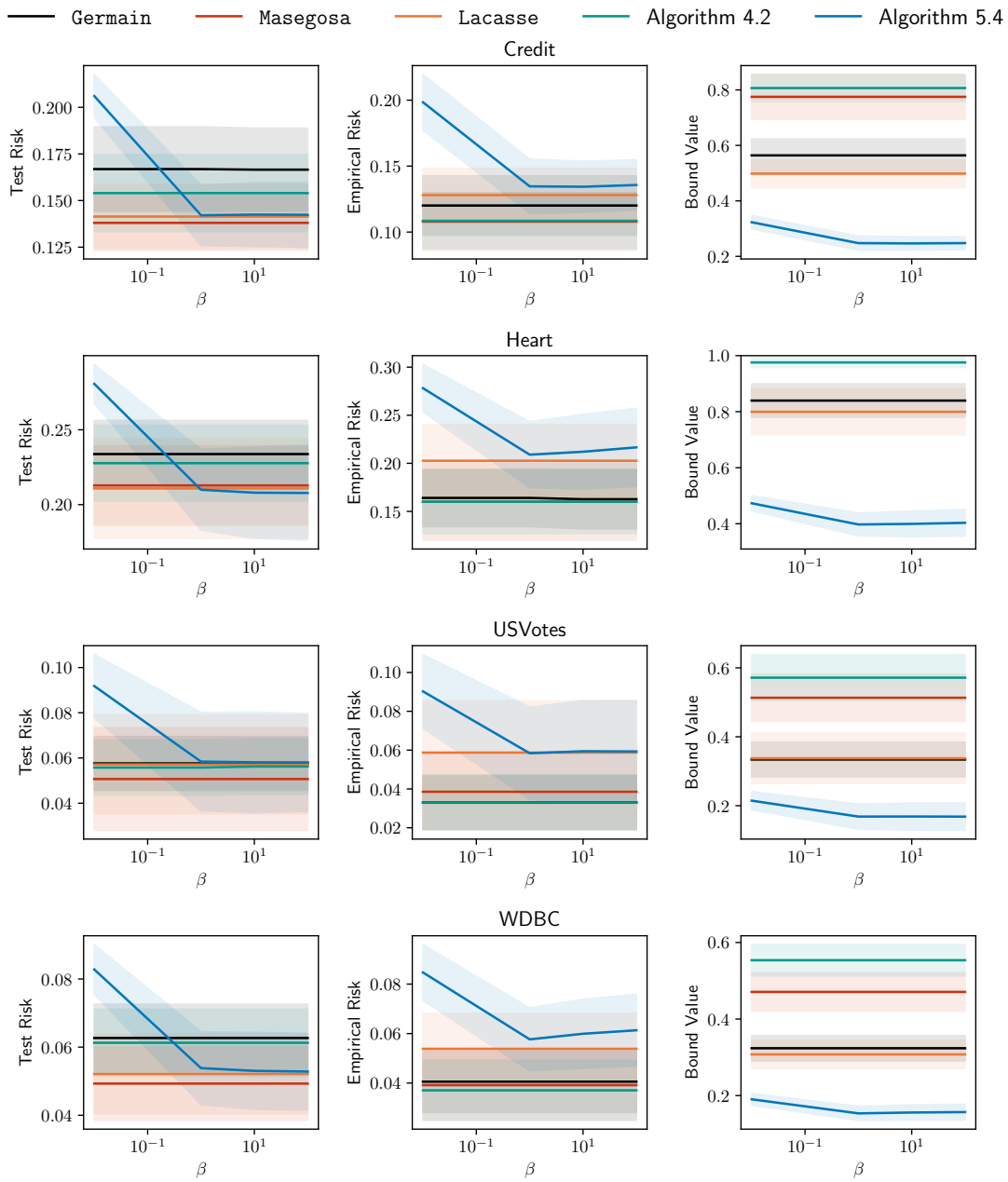


Figure E.1. *Plot of the impact of the prior β on the performance of the stochastic majority vote. More precisely, the x-axis represents the value of all the parameters β_j with $j \in \{1, \ldots, \mathrm{card}(\mathbb{H})\}$ and the y-axis are the values of the test risks, the empirical risks or the bound values. The mean (plain lines) and the standard deviations (shadows) are obtained for all values on 10 runs.*

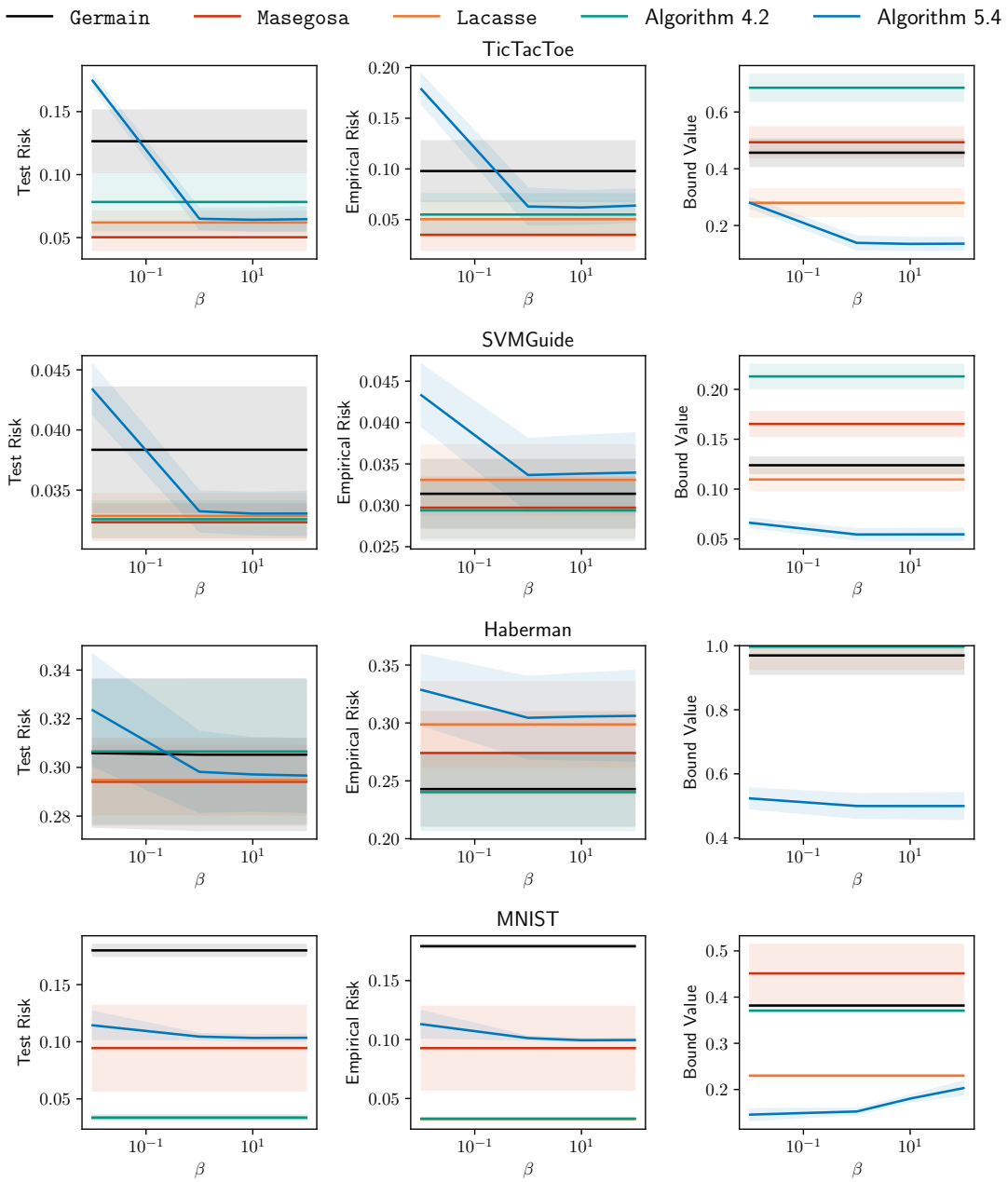


Figure E.2. *Plot of the impact of the prior β on the performance of the stochastic majority vote. More precisely, the x-axis represents the value of all the parameters β_j with $j \in \{1, \ldots, \mathrm{card}(\mathbb{H})\}$ and the y-axis are the values of the test risks, the empirical risks or the bound values. The mean (plain lines) and the standard deviations (shadows) are obtained for all values on 10 runs.*

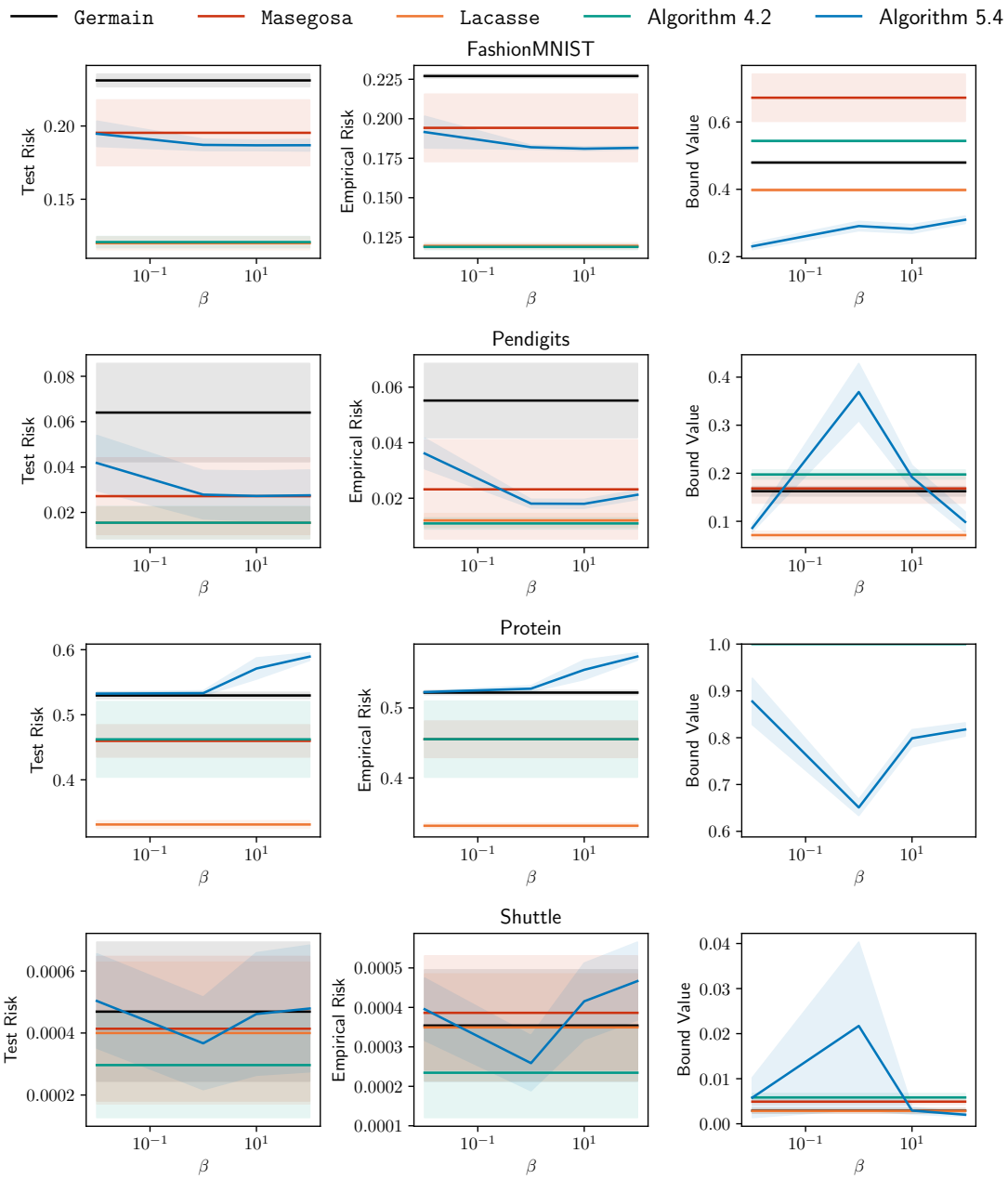


Figure E.3. *Plot of the impact of the prior β on the performance of the stochastic majority vote. More precisely, the x-axis represents the value of all the parameters β_j with $j \in \{1, \ldots, \mathrm{card}(\mathbb{H})\}$ and the y-axis are the values of the test risks, the empirical risks or the bound values. The mean (plain lines) and the standard deviations (shadows) are obtained for all values on 10 runs.*

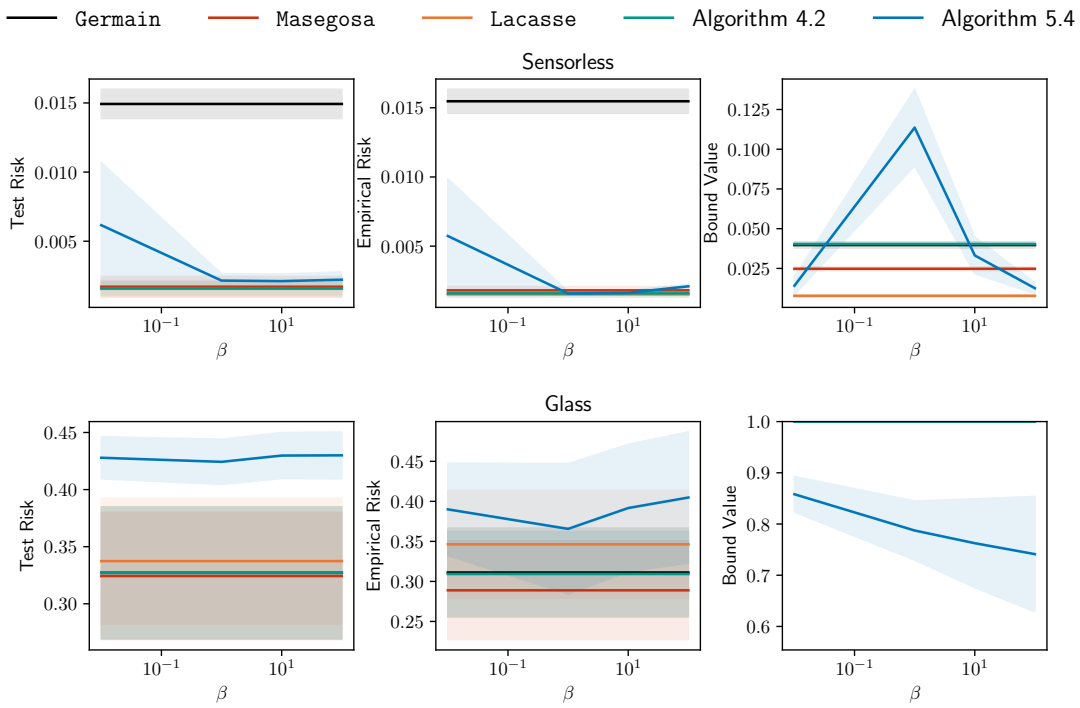


Figure E.4. *Plot of the impact of the prior β on the performance of the stochastic majority vote. More precisely, the x-axis represents the value of all the parameters β_j with $j \in \{1, \ldots, \mathrm{card}(\mathbb{H})\}$ and the y-axis are the values of the test risks, the empirical risks or the bound values. The mean (plain lines) and the standard deviations (shadows) are obtained for all values on 10 runs.*

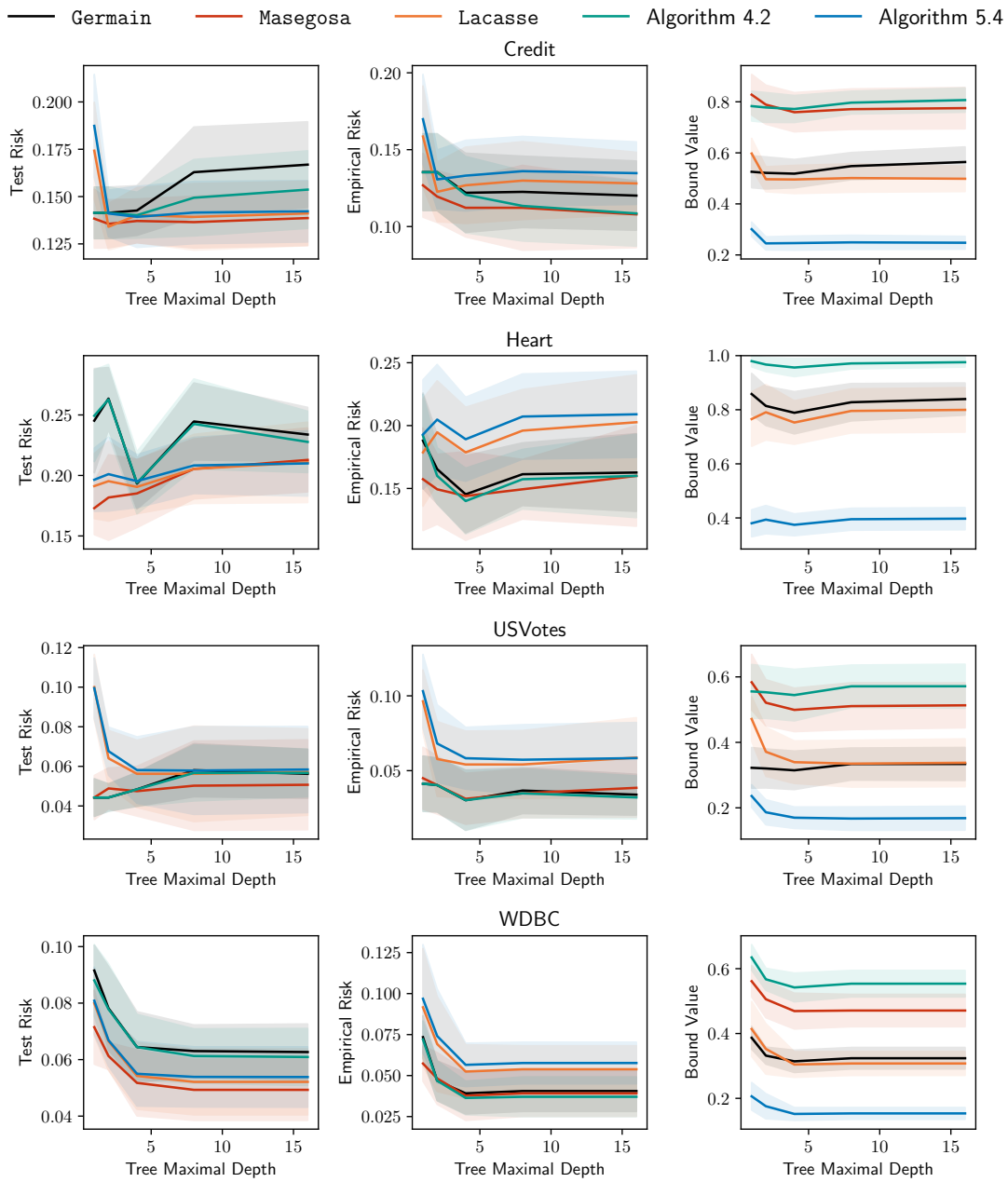


Figure E.5. *Plot of the impact of tree maximal depth on the performance of the stochastic majority vote. More precisely, the x-axis represents the value of the tree maximal depth and the y-axis are the values of the test risks, the empirical risks or the bound values. The mean (plain lines) and the standard deviations (shadows) are obtained for all values on 10 runs.*

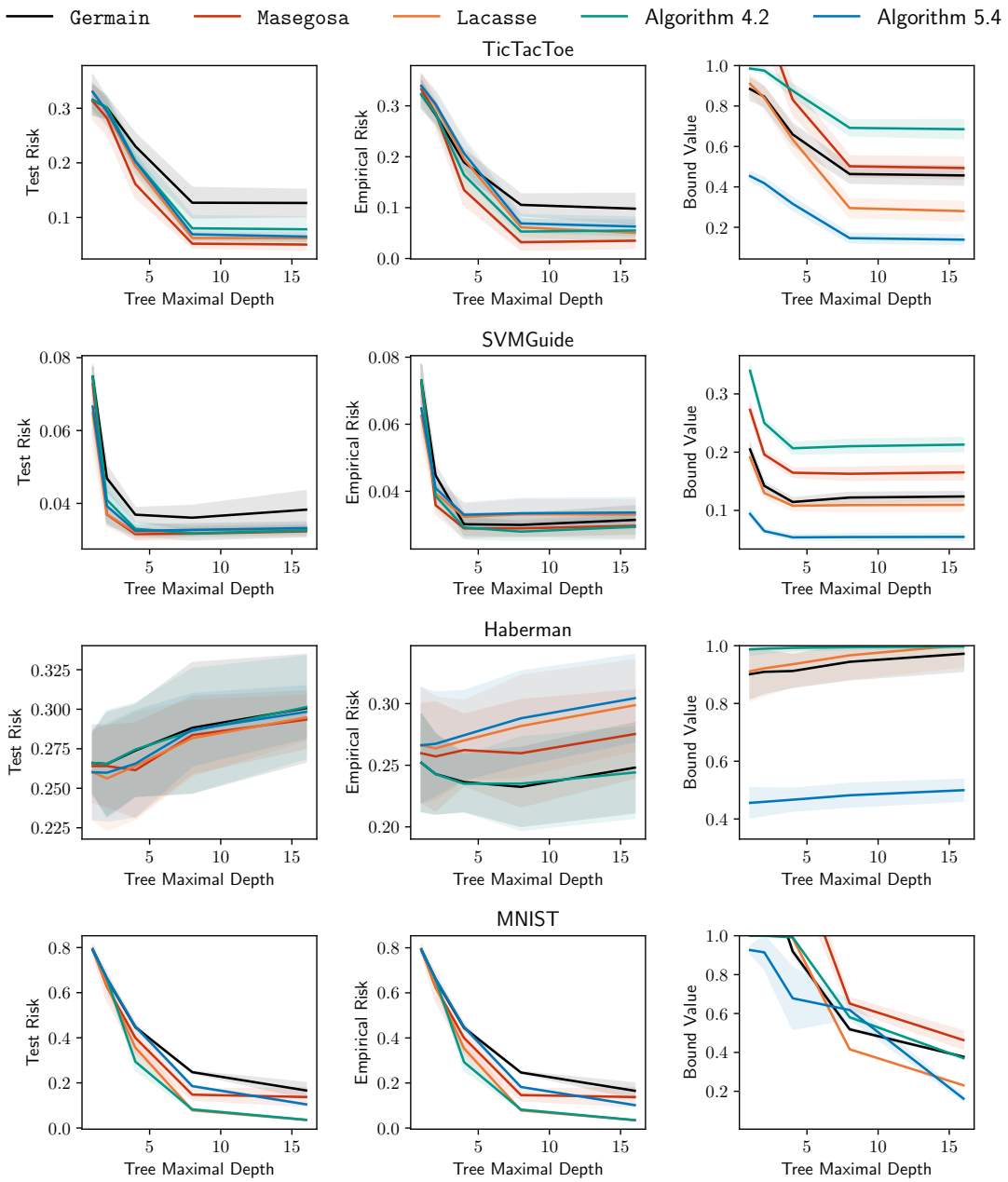


Figure E.6. *Plot of the impact of tree maximal depth on the performance of the stochastic majority vote. More precisely, the x-axis represents the value of the tree maximal depth and the y-axis are the values of the test risks, the empirical risks or the bound values. The mean (plain lines) and the standard deviations (shadows) are obtained for all values on 10 runs.*

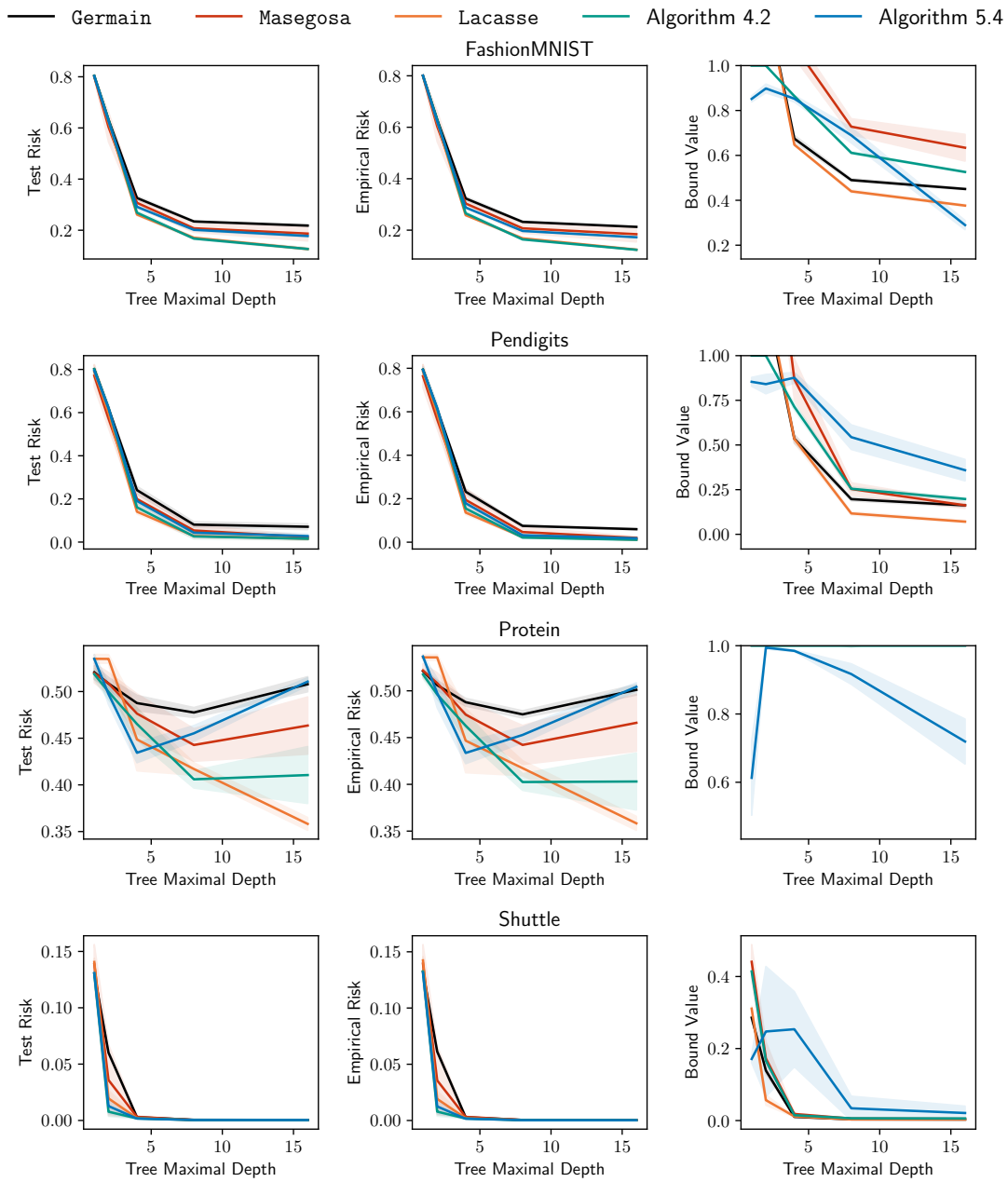


Figure E.7. *Plot of the impact of tree maximal depth on the performance of the stochastic majority vote. More precisely, the x-axis represents the value of the tree maximal depth and the y-axis are the values of the test risks, the empirical risks or the bound values. The mean (plain lines) and the standard deviations (shadows) are obtained for all values on 10 runs.*

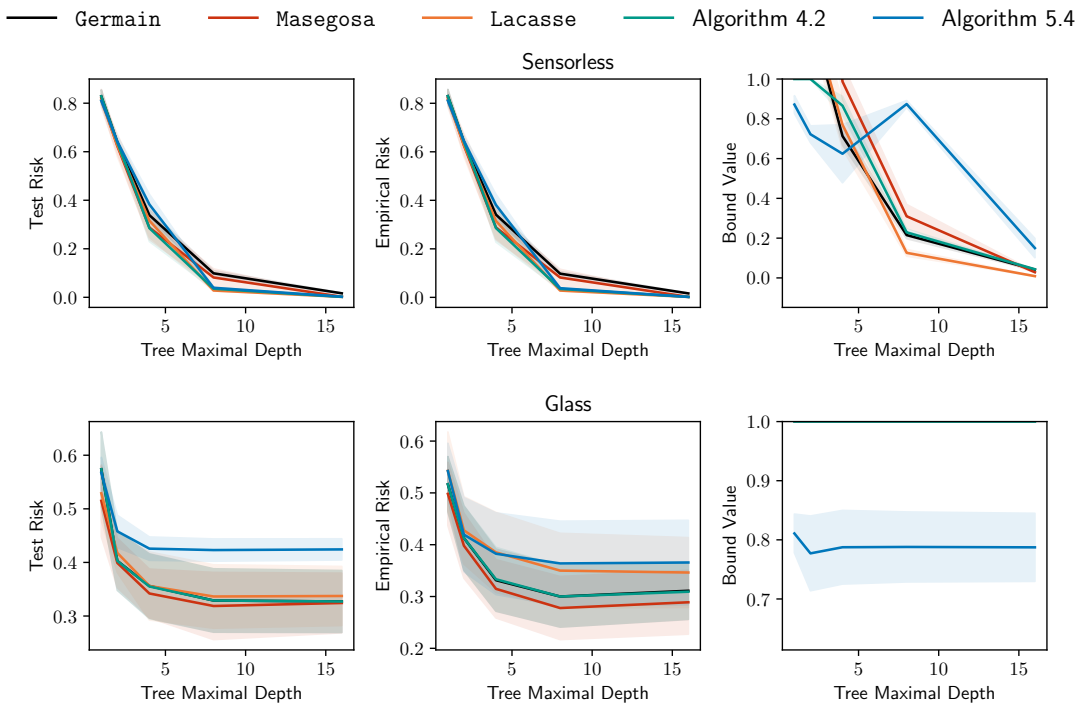


Figure E.8. *Plot of the impact of tree maximal depth on the performance of the stochastic majority vote. More precisely, the x-axis represents the value of the tree maximal depth and the y-axis are the values of the test risks, the empirical risks or the bound values. The mean (plain lines) and the standard deviations (shadows) are obtained for all values on 10 runs.*

Appendix of Chapter 6

F

F.1 Proof of Theorem 6.3.1

Theorem 6.3.1 (General Disintegrated PAC-Bayes Bound)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}_+^*$, for any $\lambda > 1$, for any $\delta \in (0,1]$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left(\frac{\lambda}{\lambda-1}\ln\left(\varphi(h,\mathbb{S})\right)\right.$$
$$\left.\leq \frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta} + \mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi) + \ln\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]\right) \geq 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S},\pi)$ is output by the deterministic algorithm A.

Proof. For any sample $\mathbb{S} \in (\mathbb{X}\times\mathbb{Y})^m$, prior $\pi \in \mathbb{M}^*(\mathbb{H})$ and deterministic algorithm A fixed apriori, let $\rho_{\mathbb{S}} = A(\mathbb{S},\pi)$ the distribution obtained from the algorithm A. Note that $\varphi(h,\mathbb{S})$ is a strictly-positive random variable. Hence, from Markov's inequality (Theorem A.2.1), we have

$$\mathop{\mathbb{P}}_{h\sim\rho_{\mathbb{S}}}\left[\varphi(h,\mathbb{S}) \leq \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\varphi(h',\mathbb{S})\right] \geq 1-\tfrac{\delta}{2}$$
$$\iff \mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\mathrm{I}\left[\varphi(h,\mathbb{S}) \leq \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\varphi(h',\mathbb{S})\right] \geq 1-\tfrac{\delta}{2}.$$

Taking the expectation over $\mathbb{S}\sim\mathcal{D}^m$ to both sides of the inequality gives

$$\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\mathrm{I}\left[\varphi(h,\mathbb{S}) \leq \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\varphi(h',\mathbb{S})\right] \geq 1-\tfrac{\delta}{2}$$
$$\iff \mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\varphi(h,\mathbb{S}) \leq \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\varphi(h',\mathbb{S})\right] \geq 1-\tfrac{\delta}{2}.$$

Since both sides of the inequality are strictly positive, we can take the logarithm

and multiply by $\frac{\lambda}{\lambda-1} > 0$ to obtain

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\frac{\lambda}{\lambda-1}\ln\left(\varphi(h,\mathbb{S})\right) \leq \frac{\lambda}{\lambda-1}\ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\varphi(h',\mathbb{S})\right)\right] \geq 1-\tfrac{\delta}{2}.$$

We develop the right-hand side of the inequality and take the expectation of the hypothesis over the prior distribution π. We have for all prior $\pi \in \mathbb{M}^*(\mathbb{H})$,

$$\begin{aligned}\frac{\lambda}{\lambda-1}\ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\varphi(h',\mathbb{S})\right) &= \frac{\lambda}{\lambda-1}\ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\frac{\rho_{\mathbb{S}}(h')\pi(h')}{\pi(h')\rho_{\mathbb{S}}(h')}\varphi(h',\mathbb{S})\right)\\ &= \frac{\lambda}{\lambda-1}\ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\pi}\frac{\rho_{\mathbb{S}}(h')}{\pi(h')}\varphi(h',\mathbb{S})\right),\end{aligned}$$

Remark that $\frac{1}{r}+\frac{1}{s}=1$ with $r=\lambda$ and $s=\frac{\lambda}{\lambda-1}$. Hence, we can apply HÖLDER's inequality (Theorem A.5.1):

$$\mathop{\mathbb{E}}_{h'\sim\pi}\frac{\rho_{\mathbb{S}}(h')}{\pi(h')}\varphi(h',\mathbb{S}) \leq \left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}}(h')}{\pi(h')}\right]^{\lambda}\right)\right]^{\frac{1}{\lambda}}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}.$$

Then, since both sides of the inequality are strictly positive, we take the logarithm; add $\ln(\frac{2}{\delta})$ and multiply by $\frac{\lambda}{\lambda-1} > 0$ to both sides of the inequality, to obtain

$$\begin{aligned}&\frac{\lambda}{\lambda-1}\ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\pi}\frac{\rho_{\mathbb{S}}(h')}{\pi(h')}\varphi(h',\mathbb{S})\right)\\ \leq&\frac{\lambda}{\lambda-1}\ln\left(\frac{2}{\delta}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}}(h')}{\pi(h')}\right]^{\lambda}\right)\right]^{\frac{1}{\lambda}}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\\ =&\frac{1}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}}(h')}{\pi(h')}\right]^{\lambda}\right)\right)+\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta}+\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\right)\\ =&\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)+\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta}+\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\right).\end{aligned}$$

From this inequality, we can deduce that

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\Bigg[\forall\pi\in\mathbb{M}^*(\mathbb{H}), &\frac{\lambda}{\lambda-1}\ln\left(\varphi(h,\mathbb{S})\right) \leq \mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)\\ &+\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta}+\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\right)\Bigg] \geq 1-\tfrac{\delta}{2}. \quad \text{(F.1)}\end{aligned}$$

Given a prior $\pi \in \mathbb{M}^*(\mathbb{H})$, note that $\mathbb{E}_{h'\sim\pi}\,\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}$ is a strictly positive random variable. Hence, we apply MARKOV's inequality (Theorem A.2.1) to have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right] \ge 1-\tfrac{\delta}{2}.$$

Since the inequality does not depend on the random variable $h\sim\rho_{\mathbb{S}}$, we have

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]\\
&= \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathrm{I}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]\\
&= \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\mathrm{I}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]\\
&= \mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right].
\end{aligned}$$

Since both sides of the inequality are strictly positive, we take the logarithm to both sides of the inequality, and we add $\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta}$ to have

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right] \ge 1-\tfrac{\delta}{2} \iff\\
&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta} + \ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\right) \le \frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta}\right.\\
&\qquad\left. + \ln\left(\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right)\right] \ge 1-\tfrac{\delta}{2}. \qquad \text{(F.2)}
\end{aligned}$$

Combining Equations (F.1) and (F.2) with a union bound gives us the desired result. ■

F.2 Proof of Corollary 6.3.1

Corollary 6.3.1 (Extreme Cases of Theorem 6.3.1)**.** Under the assumptions of Theorem 6.3.1, when $\lambda\to 1^+$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left(\ln\varphi(h,\mathbb{S}) \le \ln\frac{2}{\delta} + \ln\left[\operatorname*{esssup}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}),h'\in\mathbb{H}}\varphi(h',\mathbb{S}')\right]\right) \ge 1-\delta,$$

when $\lambda \to +\infty$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left(\ln\varphi(h,\mathbb{S}) \leq \ln \operatorname*{esssup}_{h'\in\mathbb{H}} \frac{\rho_{\mathbb{S}}(h')}{\pi(h')} + \ln\left[\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S}')\right]\right) \geq 1-\delta,$$

where esssup is the essential supremum defined as the supremum on a set with non-zero probability measures, *i.e.*,

$$\operatorname*{esssup}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}),h'\in\mathbb{H}} \varphi(h',\mathbb{S}') = \inf\left\{\tau\in\mathbb{R}, \mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\varphi(h,\mathbb{S}) > \tau\right] = 0\right\},$$

$$\text{and} \quad \operatorname*{esssup}_{h'\in\mathbb{H}} \frac{\rho_{\mathbb{S}}(h')}{\pi(h')} = \inf\left\{\tau\in\mathbb{R}, \mathop{\mathbb{P}}_{h\sim\rho_{\mathbb{S}}}\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)} > \tau\right] = 0\right\}.$$

Proof. Starting from Theorem 6.3.1 and rearranging, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\ln\left(\varphi(h,\mathbb{S})\right) \leq \frac{2\lambda-1}{\lambda}\ln\frac{2}{\delta} + \frac{\lambda-1}{\lambda}\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi) + \ln\left(\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\right] \geq 1-\delta.$$

Then, we will prove the case when $\lambda \to 1$ and $\lambda \to +\infty$ separately.

When $\lambda \to 1$. We have $\lim_{\lambda\to1^+}\frac{2\lambda-1}{\lambda}\ln\frac{2}{\delta} = \ln\frac{2}{\delta}$ and $\lim_{\lambda\to1^+}\frac{\lambda-1}{\lambda}\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi) = 0$. Furthermore, note that

$$\|\varphi\|_{\frac{\lambda}{\lambda-1}} = \left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(|\varphi(h',\mathbb{S}')|^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}} = \left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}$$

is the $L^{\frac{\lambda}{\lambda-1}}$-norm of the function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}_+^*$, where $\lim_{\lambda\to1}\|\varphi\|_{\frac{\lambda}{\lambda-1}} = \lim_{\lambda'\to+\infty}\|\varphi\|_{\lambda'}$ (since we have $\lim_{\lambda\to1^+}\frac{\lambda}{\lambda-1} = (\lim_{\lambda\to1}\lambda)(\lim_{\lambda\to1}\frac{1}{\lambda-1}) = +\infty$). Then, it is well known that

$$\|\varphi\|_\infty = \lim_{\lambda'\to+\infty}\|\varphi\|_{\lambda'} = \operatorname*{esssup}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}),h'\in\mathbb{H}}\varphi(h',\mathbb{S}').$$

Hence, we have

$$\begin{aligned}
&\lim_{\lambda\to 1}\ln\left(\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\\
=&\ln\left(\lim_{\lambda\to 1}\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\\
=&\ln\left(\lim_{\lambda\to 1}\|\varphi\|_{\frac{\lambda}{\lambda-1}}\right)=\ln\left(\lim_{\lambda'\to+\infty}\|\varphi\|_{\lambda'}\right)\\
=&\ln\left(\|\varphi\|_\infty\right)=\ln\left(\operatorname*{esssup}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}),h'\in\mathbb{H}}\varphi(h',\mathbb{S}')\right).
\end{aligned}$$

Finally, we can deduce that

$$\begin{aligned}
&\lim_{\lambda\to 1}\left[\frac{2\lambda-1}{\lambda}\ln\frac{2}{\delta}+\frac{\lambda-1}{\lambda}\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)+\ln\left(\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\right]\\
=&\ln\frac{2}{\delta}+\ln\left[\operatorname*{esssup}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}),h'\in\mathbb{H}}\varphi(h',\mathbb{S}')\right].
\end{aligned}$$

When $\lambda\to+\infty$. First, we have $\lim_{\lambda\to+\infty}\frac{2\lambda-1}{\lambda}\ln\frac{2}{\delta}=\ln\frac{2}{\delta}\left[2-\lim_{\lambda\to+\infty}\frac{1}{\lambda}\right]=2\ln\frac{2}{\delta}=\ln\frac{4}{\delta^2}$ and $\lim_{\lambda\to+\infty}\|\varphi\|_{\frac{\lambda}{\lambda-1}}=\lim_{\lambda'\to 1}\|\varphi\|_{\lambda'}=\|\varphi\|_1$ (since $\lim_{\lambda\to+\infty}\frac{\lambda}{\lambda-1}=\lim_{\lambda\to+\infty}\frac{1}{1-\frac{1}{\lambda}}=1$). Hence, we have

$$\begin{aligned}
&\lim_{\lambda\to+\infty}\ln\left(\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\\
=&\ln\left(\lim_{\lambda\to+\infty}\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\\
=&\ln\left(\lim_{\lambda\to+\infty}\|\varphi\|_{\frac{\lambda}{\lambda-1}}\right)=\ln\left(\lim_{\lambda'\to 1}\|\varphi\|_{\lambda'}\right)\\
=&\ln\left(\|\varphi\|_1\right)=\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S}')\right).
\end{aligned}$$

Moreover, by rearranging the terms in $\frac{\lambda-1}{\lambda}\,\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)$, we have

$$\begin{aligned}\frac{\lambda-1}{\lambda}\,\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi) &= \frac{1}{\lambda}\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]^\lambda\right)\right) = \ln\left(\left[\mathop{\mathbb{E}}_{h\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]^\lambda\right)\right]^{\frac{1}{\lambda}}\right)\\ &= \ln\left(\left[\mathop{\mathbb{E}}_{h\sim\pi}\left(\gamma(h)^\lambda\right)\right]^{\frac{1}{\lambda}}\right) = \ln(\|\gamma\|_\lambda)\,,\end{aligned}$$

where $\|\gamma\|_\lambda$ is the L^λ-norm of the function γ defined as $\gamma(h) = \frac{\rho_{\mathbb{S}}(h)}{\pi(h)}$. We have

$$\begin{aligned}\lim_{\lambda\to+\infty}\frac{\lambda-1}{\lambda}\,\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi) &= \lim_{\lambda\to+\infty}\ln(\|\gamma\|_\lambda) = \ln\left(\lim_{\lambda\to+\infty}\|\gamma\|_\lambda\right)\\ &= \ln\left(\|\gamma\|_\infty\right) = \ln\left(\operatorname*{esssup}_{h\in\mathbb{H}}\gamma(h)\right) = \ln\left(\operatorname*{esssup}_{h\in\mathbb{H}}\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right).\end{aligned}$$

Finally, we can deduce that

$$\begin{aligned}&\lim_{\lambda\to+\infty}\left[\frac{2\lambda-1}{\lambda}\ln\frac{2}{\delta} + \frac{\lambda-1}{\lambda}\,\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi) + \ln\left(\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\right]\\ &= \ln\operatorname*{esssup}_{h'\in\mathbb{H}}\frac{\rho_{\mathbb{S}}(h')}{\pi(h')} + \ln\left[\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S}')\right].\end{aligned}$$

■

F.3 Proof of Theorem 6.3.2

For the sake of completeness, we first prove an upper bound on $\sqrt{ab}$ (see, *e.g.*, Thiemann *et al.*, 2017).

Lemma F.3.1. For any $a > 0, b > 0$, we have

$$\sqrt{\tfrac{a}{b}} = \operatorname*{argmin}_{\lambda>0}\left(\frac{a}{\lambda}+\lambda b\right), \text{ and } \ 2\sqrt{ab} = \min_{\lambda>0}\left(\frac{a}{\lambda}+\lambda b\right),$$
$$\text{and } \ \forall\lambda>0, \sqrt{ab} \le \frac{1}{2}\left(\frac{a}{\lambda}+\lambda b\right).$$

Proof. Let $f(\lambda) = \left(\frac{a}{\lambda} + \lambda b\right)$. The first derivative of f *w.r.t.* λ is

$$\frac{\partial f}{\partial \lambda}(\lambda) = \left(b - \frac{a}{\lambda^2}\right).$$

Moreover, from the derivative we can deduce that we have $\frac{\partial f}{\partial \lambda}(\lambda) < 0 \iff \lambda \in (0, \sqrt{\frac{a}{b}})$, and $\frac{\partial f}{\partial \lambda}(\lambda) > 0 \iff \lambda > \sqrt{\frac{a}{b}}$ and $\frac{\partial f}{\partial \lambda}(\lambda) = 0 \iff \lambda = \sqrt{\frac{a}{b}}$. It implies that the function is strictly decreasing on $\lambda \in (0, \sqrt{\frac{a}{b}})$, strictly increasing for $\lambda > \sqrt{\frac{a}{b}}$ and admit a unique minimum at $\lambda^* = \sqrt{\frac{a}{b}}$. Additionally, $f(\lambda^*) = 2\sqrt{ab}$ which proves the claim. ■

We can now prove Theorem 6.3.2 with Lemma F.3.1.

Theorem 6.3.2 (Parametrizable Disintegrated PAC-Bayes Bound)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}_+^*$, for any $\delta \in (0, 1]$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left(\forall \lambda>0,\ \ln\left(\varphi(h,\mathbb{S})\right) \le \ln\left[\frac{\lambda}{2} e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)} + \frac{8}{2\lambda\delta^3} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi}\left[\varphi(h',\mathbb{S}')^2\right]\right]\right) \ge 1-\delta,$$

where $\rho_{\mathbb{S}} \triangleq A(\mathbb{S}, \pi)$ is output by the deterministic algorithm A.

Proof. The proof is similar to the one of Theorem 6.3.1. Since $\varphi(h,\mathbb{S})$ is a strictly positive random variable, from Markov's inequality (Theorem A.2.1), we have

$$\begin{aligned} &\mathop{\mathbb{P}}_{h\sim\rho_{\mathbb{S}}}\left[\varphi(h,\mathbb{S}) \le \frac{2}{\delta} \mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}} \varphi(h',\mathbb{S})\right] \ge 1 - \tfrac{\delta}{2} \\ \iff &\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \mathrm{I}\left[\varphi(h,\mathbb{S}) \le \frac{2}{\delta} \mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}} \varphi(h',\mathbb{S})\right] \ge 1 - \tfrac{\delta}{2}. \end{aligned}$$

Taking the expectation over $\mathbb{S} \sim \mathcal{D}^m$ to both sides of the inequality gives

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \mathrm{I}\left[\varphi(h,\mathbb{S}) \le \frac{2}{\delta} \mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}} \varphi(h',\mathbb{S})\right] \ge 1 - \tfrac{\delta}{2} \\ \iff &\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\varphi(h,\mathbb{S}) \le \frac{2}{\delta} \mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}} \varphi(h',\mathbb{S})\right] \ge 1 - \tfrac{\delta}{2}. \end{aligned}$$

Using Lemma F.3.1 with $a = \frac{4}{\delta^2}\varphi(h',\mathbb{S})^2$ and $b = \frac{\rho_{\mathbb{S}}(h')^2}{\pi(h')^2}$, we have for all prior $\pi \in \mathbb{M}^*(\mathbb{H})$

$$\forall \lambda > 0, \quad \frac{2}{\delta} \mathop{\mathbb{E}}_{h' \sim \rho_{\mathbb{S}}} \varphi(h',\mathbb{S}) = \mathop{\mathbb{E}}_{h' \sim \pi} \sqrt{\frac{\rho_{\mathbb{S}}(h')^2}{\pi(h')^2} \frac{4}{\delta^2} \varphi(h',\mathbb{S})^2}$$
$$\leq \frac{1}{2}\left[\lambda \mathop{\mathbb{E}}_{h' \sim \pi}\left(\frac{\rho_{\mathbb{S}}(h')}{\pi(h')}\right)^2 + \frac{4}{\lambda\delta^2} \mathop{\mathbb{E}}_{h' \sim \pi}\left(\varphi(h',\mathbb{S})^2\right)\right].$$

Then, since both sides of the inequality are strictly positive, we take the logarithm to obtain

$$\forall \lambda > 0, \ln\left(\frac{2}{\delta} \mathop{\mathbb{E}}_{h' \sim \rho_{\mathbb{S}}} \varphi(h',\mathbb{S})\right) \leq \ln\left(\frac{1}{2}\left[\lambda \mathop{\mathbb{E}}_{h' \sim \pi}\left(\frac{\rho_{\mathbb{S}}(h')}{\pi(h')}\right)^2 + \frac{4}{\lambda\delta^2} \mathop{\mathbb{E}}_{h' \sim \pi}\left(\varphi(h',\mathbb{S})^2\right)\right]\right)$$
$$= \ln\left(\frac{1}{2}\left[\lambda \exp(\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)) + \frac{4}{\lambda\delta^2} \mathop{\mathbb{E}}_{h' \sim \pi}\left(\varphi(h',\mathbb{S})^2\right)\right]\right).$$

Hence, we can deduce that

$$\mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m, h \sim \rho_{\mathbb{S}}}\Bigg[\forall \pi \in \mathbb{M}^*(\mathbb{H}),\ \forall \lambda > 0, \ln\left(\varphi(h,\mathbb{S})\right)$$
$$\leq \ln\left(\frac{1}{2}\left[\lambda e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)} + \frac{4}{\lambda\delta^2} \mathop{\mathbb{E}}_{h' \sim \pi}\left(\varphi(h',\mathbb{S})^2\right)\right]\right)\Bigg] \geq 1 - \tfrac{\delta}{2}. \tag{F.3}$$

Given a prior $\pi \in \mathbb{M}^*(\mathbb{H})$, note that $\mathbb{E}_{h' \sim \pi} \varphi(h',\mathbb{S})^2$ is a strictly-positive random variable. Hence, we apply Markov's inequality (Theorem A.2.1):

$$\mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m}\left[\mathop{\mathbb{E}}_{h' \sim \pi} \varphi(h',\mathbb{S})^2 \leq \frac{2}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{h' \sim \pi} \varphi(h',\mathbb{S}')^2\right] \geq 1 - \tfrac{\delta}{2}.$$

Since the inequality does not depend on the random variable $h \sim \rho_{\mathbb{S}}$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m}\left[\mathop{\mathbb{E}}_{h' \sim \pi}\left(\varphi(h',\mathbb{S})^2\right) \leq \frac{2}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{h' \sim \pi}\left(\varphi(h',\mathbb{S}')^2\right)\right]$$
$$= \mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m, h \sim \rho_{\mathbb{S}}}\left[\mathop{\mathbb{E}}_{h' \sim \pi}\left(\varphi(h',\mathbb{S})^2\right) \leq \frac{2}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{h' \sim \pi}\left(\varphi(h',\mathbb{S}')^2\right)\right].$$

Additionally, note that multiplying by $\frac{4}{2\lambda\delta^2} > 0$, adding $\frac{\lambda}{2}\exp(\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi))$, and taking the logarithm to both sides of the inequality results in the same indicator

function. Indeed,

$$\begin{aligned}
&\mathrm{I}\left[\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^2\right) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^2\right)\right]\\
=\ &\mathrm{I}\left[\forall\lambda>0, \tfrac{4}{2\lambda\delta^2}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^2\right) \le \tfrac{8}{2\lambda\delta^3}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^2\right)\right]\\
=\ &\mathrm{I}\left[\forall\lambda>0, \ln\left(\tfrac{\lambda}{2}\exp(\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi))+\tfrac{4}{2\lambda\delta^2}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^2\right)\right)\right.\\
&\qquad\left.\le \ln\left(\tfrac{\lambda}{2}\exp(\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi))+\tfrac{8}{2\lambda\delta^3}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^2\right)\right)\right].
\end{aligned}$$

Hence, we can deduce that

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\forall\lambda{>}0,\ \ln\left(\frac{1}{2}\left[\lambda\exp(\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi))+\frac{4}{\lambda\delta^2}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^2\right)\right]\right)\right.\\
&\left.\le \ln\left(\frac{1}{2}\left[\lambda\exp(\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi))+\frac{8}{\lambda\delta^3}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^2\right)\right]\right)\right] \ge 1-\tfrac{\delta}{2}. \qquad \text{(F.4)}
\end{aligned}$$

Combining Equations (F.3) and (F.4) with a union bound gives us the desired result. ■

F.4 Proof of Proposition 6.3.1

Proposition 6.3.1 (Optimal Bound of Theorem 6.3.2)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution π on $\mathbb{H}$, for any $\delta\in(0,1]$, for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m\to\mathbb{R}_+^*$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m\times\mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, let

$$\lambda^* = \mathop{\mathrm{argmin}}_{\lambda>0}\ln\left[\frac{\lambda}{2}e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)}+\frac{\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left[8\varphi(h',\mathbb{S}')^2\right]}{2\lambda\delta^3}\right],$$

then, we have

$$\begin{aligned}
&\overbrace{2\ln\left[\frac{\lambda^*}{2}e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)}+\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\frac{8\varphi(h',\mathbb{S}')^2}{2\lambda^*\delta^3}\right)\right]}^{\text{Theorem 6.3.2}}\\
&= \underbrace{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)+\ln\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\frac{8\varphi(h',\mathbb{S}')^2}{\delta^3}\right)\right]}_{\text{Theorem 6.3.1 with } \lambda=2.},
\end{aligned}$$

where $\lambda^* = \sqrt{\frac{\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi}\left[8\varphi(h',\mathbb{S}')^2\right]}{\delta^3\exp(\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi))}}$.
Put into words: the optimal λ^* gives the same bound for Theorem 6.3.1 and Theorem 6.3.2.

Proof. We consider the right-hand side of the inequality of Theorem 6.3.2 (which is strictly positive): we have

$$\ln\left[\frac{\lambda}{2}e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)}+\frac{8}{2\lambda\delta^3}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left[\varphi(h',\mathbb{S}')^2\right]\right]. \tag{F.5}$$

Since $\ln$ is a strictly increasing function, we have

$$\begin{aligned}&\min_{\lambda>0}\left\{\ln\left[\frac{\lambda}{2}e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)}+\frac{8}{2\lambda\delta^3}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left[\varphi(h',\mathbb{S}')^2\right]\right]\right\}\\ =&\ln\left[\min_{\lambda>0}\left\{\frac{\lambda}{2}e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)}+\frac{8}{2\lambda\delta^3}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left[\varphi(h',\mathbb{S}')^2\right]\right\}\right].\end{aligned}$$

Then, we apply Lemma F.3.1 by taking $a=\frac{8}{2\delta^3}\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi}\left[\varphi(h',\mathbb{S}')^2\right]$ and $b=\frac{1}{2}e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)}$ to obtain $\lambda^*=\sqrt{\frac{a}{b}}=\sqrt{\frac{\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi}[8\varphi(h',\mathbb{S}')^2]}{\delta^3\exp(\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi))}}$. Finally, by substituting λ^* into Equation (F.5), we obtain

$$\begin{aligned}&\ln\left[\frac{\lambda^*}{2}e^{\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)}+\frac{8}{2\lambda^*\delta^3}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left[\varphi(h',\mathbb{S}')^2\right]\right]\\ =&\frac{1}{2}\left(\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi)+\ln\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\frac{8\varphi(h',\mathbb{S}')^2}{\delta^3}\right)\right]\right),\end{aligned}$$

which is the desired result. ■

F.5 Proof of Corollary 6.4.1

We introduce Theorem F.5.1 which takes into account a set of priors $\mathbb{P}$ while Theorem 6.3.1 handles a unique prior π.

Theorem F.5.1. For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any priors set $\mathbb{P}=\{\pi_t\}_{t=1}^T$ of T prior $\pi\in\mathbb{M}^*(\mathbb{H})$, for any measurable function $\varphi:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m\to\mathbb{R}_+^*$, for any $\lambda>1$, for any $\delta\in(0,1]$, for any algorithm A :

$(\mathbb{X}\times\mathbb{Y})^m\times\mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\forall\pi_t\in\mathbb{P},\,\frac{\lambda}{\lambda-1}\ln\left(\varphi(h,\mathbb{S})\right)\leq \mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)+\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta}\right.$$
$$\left.+\ln\frac{2T}{\delta}+\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right)\right]\geq 1-\delta,$$

where $\rho_{\mathbb{S}}\triangleq A(\mathbb{S},\pi)$ is output by the deterministic algorithm A.

Proof. The proof is mainly the same as Theorem 6.3.1. Indeed, we first derive the same equation as Equation (F.1), we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\forall\pi\in\mathbb{M}^*(\mathbb{H}),\,\frac{\lambda}{\lambda-1}\ln(\varphi(h,\mathbb{S}))\leq \mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)\right.$$
$$\left.+\frac{\lambda}{\lambda-1}\ln\frac{2}{\delta}+\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right)\right)\right]\geq 1-\tfrac{\delta}{2}.$$

Then, we apply MARKOV's inequality (as in Theorem 6.3.1) T times with the T priors π_t belonging to $\mathbb{P}$, however, we set the confidence to $\frac{\delta}{2T}$ instead of $\frac{\delta}{2}$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi_t}\left[\varphi(h',\mathbb{S})^{\frac{\lambda}{\lambda-1}}\right]\right)\right.$$
$$\left.\leq\ln\frac{2T}{\delta}+\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi_t}\left[\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right]\right)\right]\geq 1-\tfrac{\delta}{2T}.$$

Finally, combining the $T+1$ bounds with a union bound gives us the desired result. ■

We now prove Corollary 6.4.1 from Theorem F.5.1.

Corollary 6.4.1 (Instantiation of Theorem 6.3.1 for Neural Networks)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any set $\mathbb{P}=\{\pi_1,\ldots,\pi_T\}$ of T priors on $\mathbb{H}$ where $\pi_t=\mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_D)$, for any algorithm $A:(\mathbb{X}\times\mathbb{Y})^m\times\mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, for any loss $\ell:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})\to[0,1]$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left(\forall\pi_t\in\mathbb{P},\;\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))\leq\frac{1}{m}\left[\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{\sigma^2}+\ln\frac{16T\sqrt{m}}{\delta^3}\right]\right)\geq 1-\delta,$$

where $\mathrm{kl}(a\|b) = a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$, $\rho_{\mathbb{S}} = \mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_D)$, and the hypothesis $h \sim \rho_{\mathbb{S}}$ is parameterized by $\mathbf{w}+\boldsymbol{\epsilon}$.

Proof. We instantiate Theorem F.5.1 with $\varphi(h,\mathbb{S}) = \exp\Big[\frac{\lambda-1}{\lambda} m\,\mathrm{kl}(\mathsf{R}^{\ell}_{\mathcal{S}}(h)\|\mathsf{R}^{\ell}_{\mathcal{D}}(h))\Big]$ and $\lambda = 2$. We have with probability at least $1-\delta$ over $\mathbb{S}\sim\mathcal{D}^m$ and $h\sim\rho_{\mathbb{S}}$, for all prior $\pi_t \in \mathbb{P}$

$$\mathrm{kl}(\mathsf{R}^{\ell}_{\mathcal{S}}(h)\|\mathsf{R}^{\ell}_{\mathcal{D}}(h)) \leq \frac{1}{m}\left[\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi_t) + \ln\left(\frac{8T}{\delta^3}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi_t} e^{m\,\mathrm{kl}(\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\|\mathsf{R}^{\ell}_{\mathcal{D}}(h'))}\right)\right].$$

From Maurer (2004) we upper-bound $\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi_t} e^{m\,\mathrm{kl}(\mathsf{R}^{\ell}_{\mathcal{S}'}(h')\|\mathsf{R}^{\ell}_{\mathcal{D}}(h'))}$ by $2\sqrt{m}$ for each prior π_t (Lemma B.16.1). Hence, we have, for all prior $\pi_t\in\mathbb{P}$

$$\mathrm{kl}(\mathsf{R}^{\ell}_{\mathcal{S}}(h)\|\mathsf{R}^{\ell}_{\mathcal{D}}(h)) \leq \frac{1}{m}\left[\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi_t) + \ln\left(\frac{16T\sqrt{m}}{\delta^3}\right)\right].$$

Additionally, the Rényi divergence $\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi_t)$ between two multivariate Gaussians $\rho_{\mathbb{S}}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_D)$ and $\pi_t=\mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_D)$ is well known: its closed-form solution is $\mathrm{D}_2(\rho_{\mathbb{S}}\|\pi_t)=\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{\sigma^2}$ (see, for example, (Gil *et al.*, 2013)). ■

F.6 Proof of Corollary 6.4.2

We first prove the following lemma in order to prove Corollary 6.4.2.

Lemma F.6.1. If $\rho_{\mathbb{S}} = \mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_D)$ and $\pi = \mathcal{N}(\mathbf{v},\sigma^2\mathbf{I}_D)$, we have

$$\ln\frac{\rho_{\mathbb{S}}(h)}{\pi(h)} = \frac{1}{2\sigma^2}\Big[\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2\Big],$$

where $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_D)$ is a Gaussian noise such that $\mathbf{w}+\boldsymbol{\epsilon}$ are the weights of $h\sim\rho_{\mathbb{S}}$ with $\rho_{\mathbb{S}}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_D)$.

Proof. The probability density functions of $\rho_{\mathbb{S}}$ and π for $h\sim\rho_{\mathbb{S}}$ (with the weights

$\mathbf{w}+\boldsymbol{\epsilon}$) can be rewritten as

$$\rho_{\mathbb{S}}(h) = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^D \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{w}\|_2^2\right) = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^D \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{\epsilon}\|_2^2\right)$$

and $\quad \pi(h) = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^D \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}\|_2^2\right).$

We can derive a closed-form expression of $\ln\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]$. Indeed, we have

$$\begin{aligned}
\ln\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right] &= \ln\left[\rho_{\mathbb{S}}(h)\right] - \ln\left[\pi(h)\right] \\
&= \ln\left(\left[\frac{1}{\sigma\sqrt{2\pi}}\right]^D \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{\epsilon}\|_2^2\right)\right) \\
&\quad - \ln\left(\left[\frac{1}{\sigma\sqrt{2\pi}}\right]^D \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}\|_2^2\right)\right) \\
&= -\frac{1}{2\sigma^2}\|\boldsymbol{\epsilon}\|_2^2 + \frac{1}{2\sigma^2}\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}\|_2^2 = \frac{1}{2\sigma^2}\Big[\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2\Big].
\end{aligned}$$

■

We can now prove Corollary 6.4.2.

Corollary 6.4.2 (Instantiation of Known Bounds for Neural Networks)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any set $\mathbb{P} = \{\pi_1, \ldots, \pi_T\}$ of T priors on $\mathbb{H}$ where $\pi_t = \mathcal{N}(\mathbf{v}_t, \sigma^2\mathbf{I}_D)$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, for any loss $\ell : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y}) \to \{0,1\}$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the learning sample $\mathbb{S}\sim\mathcal{D}^m$ and the hypothesis $h\sim\rho_{\mathbb{S}}$ parameterized by $\mathbf{w}+\boldsymbol{\epsilon}$, we have $\forall \pi_t \in \mathbb{P}$

$$\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h)) \leq \frac{1}{m}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} + \ln\frac{2T\sqrt{m}}{\delta}\right], \tag{6.1}$$

$$\forall b \in \mathbb{b}, \mathrm{kl}_+(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h)) \leq \frac{1}{m}\left[\frac{b+1}{b}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}\right]_+ + \ln\frac{(b+1)T\,\mathrm{card}(\mathbb{b})}{\delta}\right], \tag{6.2}$$

$$\forall c \in \mathbb{c},\ \mathsf{R}^\ell_{\mathcal{D}}(h) \leq \frac{1 - \exp\left(-c\mathsf{R}^\ell_{\mathcal{S}}(h) - \frac{1}{m}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} + \ln\frac{T\,\mathrm{card}(\mathbb{c})}{\delta}\right]\right)}{1-e^{-c}}, \tag{6.3}$$

with $[x]_+ = \max(x, 0)$, and $\mathrm{kl}_+(\mathsf{R}^\ell_{\mathcal{S}}(h) \| \mathsf{R}^\ell_{\mathcal{D}}(h)) = \mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h) \| \mathsf{R}^\ell_{\mathcal{D}}(h))$ if $\mathsf{R}^\ell_{\mathcal{S}}(h) < \mathsf{R}^\ell_{\mathcal{D}}(h)$ and 0 otherwise. Moreover, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_D)$ is a Gaussian noise such that $\mathbf{w} + \boldsymbol{\epsilon}$ are the weights of $h \sim \rho_{\mathbb{S}}$ with $\rho_{\mathbb{S}} = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_D)$, and $\mathbb{c}$, $\mathbb{b}$ are two sets of hyperparameters fixed a priori.

Proof. We will prove the three bounds separately.

Equation (6.1). We instantiate Theorem 1*(i)* of Rivasplata *et al.* (2020) (proved in Theorem 2.4.1) with $\varphi(h, \mathbb{S}) = \exp\Big[m \, \mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h) \| \mathsf{R}^\ell_{\mathcal{D}}(h))\Big]$, however, we apply the theorem T times for each prior $\pi_t \in \mathbb{P}$ (with a confidence $\frac{\delta}{T}$ instead of δ). Hence, for each prior $\pi_t \in \mathbb{P}$, we have with probability at least $1 - \frac{\delta}{T}$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$ and $h \sim \rho_{\mathbb{S}}$

$$\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h) \| \mathsf{R}^\ell_{\mathcal{D}}(h)) \leq \frac{1}{m} \left[\ln \left[\frac{\rho_{\mathbb{S}}(h)}{\pi_t(h)} \right] + \ln \left(\frac{T}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{h' \sim \pi} e^{m \, \mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}'}(h') \| \mathsf{R}^\ell_{\mathcal{D}}(h'))} \right) \right].$$

From Maurer (2004), we upper-bound $\mathbb{E}_{\mathbb{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \pi_t} e^{m \, \mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}'}(h') \| \mathsf{R}^\ell_{\mathcal{D}}(h'))}$ by $2\sqrt{m}$ (Lemma B.16.1) and using Lemma F.6.1 we rewrite the disintegrated KL divergence. Finally, a union bound argument gives us the claim.

Equation (6.2). We apply $T \operatorname{card}(\mathbb{b})$ times Proposition 3.1 of Blanchard and Fleuret (2007) (proved in Theorem 2.4.3) with a confidence $\frac{\delta}{T \operatorname{card}(\mathbb{b})}$ instead of δ. For each prior $\pi_t \in \mathbb{P}$ and hyperparameters $b \in \mathbb{b}$, we have with probability at least $1 - \frac{\delta}{T \operatorname{card}(\mathbb{b})}$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$ and $h \sim \rho_{\mathbb{S}}$

$$\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h) \| \mathsf{R}^\ell_{\mathcal{D}}(h)) \leq \frac{1}{m} \left[\frac{b+1}{b} \left[\ln \frac{\rho_{\mathbb{S}}(h)}{\pi_t(h)} \right]_+ + \ln \left(\frac{T \operatorname{card}(\mathbb{b})(b+1)}{\delta} \right) \right].$$

From Lemma F.6.1 and a union bound argument, we obtain the claim.

Equation (6.3). We apply $T \operatorname{card}(\mathbb{c})$ times Theorem 1.2.7 of Catoni (2007) (proved in Theorem 2.4.2) with a confidence $\frac{\delta}{T \operatorname{card}(\mathbb{c})}$ instead of δ. For each prior $\pi_t \in \mathbb{P}$ and hyperparameter $c \in \mathbb{c}$, we have with probability at least $1 - \frac{\delta}{T \operatorname{card}(\mathbb{c})}$ over the random choice of $\mathbb{S} \sim \mathcal{D}^m$ and $h \sim \rho_{\mathbb{S}}$

$$\mathsf{R}^\ell_{\mathcal{D}}(h) \leq \frac{1}{1 - e^{-c}} \left[1 - \exp \left(-c \mathsf{R}^\ell_{\mathcal{S}}(h) - \frac{1}{m} \left[\ln \left[\frac{\rho_{\mathbb{S}}(h)}{\pi_t(h)} \right] + \ln \frac{T \operatorname{card}(\mathbb{c})}{\delta} \right] \right) \right].$$

From Lemma F.6.1 and a union bound argument, we obtain the claim. ∎

F.7 Proof of Corollary 6.4.3

Corollary 6.4.3 (PAC-Bayesian Bound for Stochastic Neural Networks)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any set $\mathbb{P}=\{\pi_1,\ldots,\pi_T\}$ of T priors on $\mathbb{H}$ where $\pi_t=\mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_D)$, for any loss $\ell:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})\to\{0,1\}$, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over $\mathbb{S}\sim\mathcal{D}^m$ and $\{h_1,\ldots,h_K\}\sim\rho^K$, we have simultaneously $\forall\pi_t\in\mathbb{P}$,

$$\mathrm{kl}\Big(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big\|\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)\Big)\leq\frac{1}{m}\left[\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{2\sigma^2}+\ln\frac{4T\sqrt{m}}{\delta}\right],\tag{6.4}$$

$$\text{and}\quad \mathrm{kl}\left(\frac{1}{K}\sum_{i=1}^{K}\mathsf{R}_{\mathbb{S}}(h_i)\Big\|\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)\right)\leq\frac{1}{n}\ln\frac{4}{\delta},\tag{6.5}$$

where $\rho=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_D)$ and the hypothesis h sampled from ρ is parameterized by $\mathbf{w}+\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_D)$.

Proof. We instantiate Theorem 2.3.4 and apply Jensen's inequality (Theorem A.1.1) on the left-hand side of the inequation for each prior $\pi_t=\mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_D)$ with the posterior $\rho=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_D)$ with a confidence $\frac{\delta}{2T}$ instead of δ. Indeed, for each prior π_t, with probability at least $1-\frac{\delta}{2T}$ over the random choice of $\mathbb{S}\sim\mathcal{D}^m$, we have for all posterior ρ on $\mathbb{H}$,

$$\mathrm{kl}\Big(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big\|\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)\Big)\leq\frac{1}{m}\Big[\mathrm{KL}(\rho\|\pi_t)+\ln\tfrac{4T\sqrt{m}}{\delta}\Big].$$

Note that the closed-form solution of the KL divergence between the Gaussian distributions ρ and π_t is well known, we have $\mathrm{KL}(\rho\|\pi_t)=\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{2\sigma^2}$. Then, by applying a union bound argument over the T bounds obtained with the T priors π_t, we have with probability at least $1-\frac{\delta}{2}$ over the random choice of $\mathbb{S}\sim\mathcal{D}^m$, for all prior $\pi_t\in\mathbb{P}$, for all posterior ρ

$$\mathrm{kl}\Big(\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{S}}(h)\Big\|\mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}^{\ell}_{\mathcal{D}}(h)\Big)\leq\frac{1}{m}\Big[\tfrac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{2\sigma^2}+\ln\tfrac{4T\sqrt{m}}{\delta}\Big].\quad\text{(Equation (6.4))}$$

Additionally, we obtained Equation (6.5) by a direct application the Theorem 2.2 of Dziugaite and Roy (2017) (with confidence $\frac{\delta}{2}$ instead of δ). Finally, from a union bound of the two bounds in Equations (6.4) and (6.5) gives the result. ■

F.8 Evaluation and Minimization of the Bounds of Corollaries 6.4.1 to 6.4.3

We optimize and evaluate the bounds of the corollaries (except Equation (6.3)) thanks to the inverting functions of $\mathrm{kl}()$ defined in Definition 2.3.3. Indeed, for the different corollaries, the PAC-Bayesian generalization bounds become

$$\mathsf{R}_{\mathcal{D}}^{\ell}(h) \leq \underbrace{\overline{\mathrm{kl}}\left(\mathsf{R}_{\mathcal{S}}^{\ell}(h) \;\middle|\; \frac{1}{m}\left[\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{\sigma^2} + \ln\frac{16T\sqrt{m}}{\delta^3}\right]\right)}_{\text{Corollary 6.4.1}},$$

$$\mathsf{R}_{\mathcal{D}}^{\ell}(h) \leq \underbrace{\overline{\mathrm{kl}}\left(\mathsf{R}_{\mathcal{S}}^{\ell}(h) \;\middle|\; \frac{1}{m}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} + \ln\frac{2T\sqrt{m}}{\delta}\right]\right)}_{\text{Equation (6.1)}},$$

$$\mathsf{R}_{\mathcal{D}}^{\ell}(h) \leq \underbrace{\overline{\mathrm{kl}}\left(\mathsf{R}_{\mathcal{S}}^{\ell}(h) \;\middle|\; \frac{1}{m}\left[\frac{b+1}{b}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}\right]_+ + \ln\frac{(b+1)T\,\mathrm{card}(\mathbb{b})}{\delta}\right]\right)}_{\text{Equation (6.2)}},$$

$$\text{and} \quad \mathop{\mathbb{E}}_{h\sim\rho}\mathsf{R}_{\mathcal{D}}^{\ell}(h) \leq \underbrace{\overline{\mathrm{kl}}\left(\spadesuit \;\middle|\; \frac{1}{m}\left[\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{2\sigma^2} + \ln\frac{4T\sqrt{m}}{\delta}\right]\right)}_{\text{Corollary 6.4.3}},$$

$$\text{where} \quad \spadesuit = \overline{\mathrm{kl}}\left(\frac{1}{K}\sum_{i=1}^{K}\mathsf{R}_{\mathbb{S}}(h_i) \;\middle|\; \frac{1}{K}\ln\frac{4}{\delta}\right).$$

Based on these bounds, we can deduce some objective functions that is approximated on a mini-batch $\mathbb{U} \subseteq \mathbb{S}$. Indeed, at each iteration in phase **2)**, after sampling the noise $\boldsymbol{\epsilon}$, the algorithm updates the weights $\boldsymbol{\omega}$ (*i.e.*, the hypothesis h) by optimizing

$$\underbrace{\overline{\mathrm{kl}}\left(\mathsf{R}_{\mathcal{U}}^{\ell}(h) \middle| \frac{1}{m}\left[\frac{\|\boldsymbol{\omega}-\mathbf{v}_t\|_2^2}{\sigma^2} + \ln\frac{16T\sqrt{m}}{\delta^3}\right]\right)}_{\text{Objective function for Corollary 6.4.1}},$$

$$\underbrace{\overline{\mathrm{kl}}\left(\mathsf{R}_{\mathcal{U}}^{\ell}(h) \middle| \frac{1}{m}\left[\frac{\|\boldsymbol{\omega}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} + \ln\frac{2T\sqrt{m}}{\delta}\right]\right)}_{\text{Objective function for Equation (6.1)}},$$

$$\underbrace{\overline{\mathrm{kl}}\left(\mathsf{R}_{\mathcal{U}}^{\ell}(h) \;\middle|\; \frac{1}{m}\left[\frac{b+1}{b}\left[\frac{\|\boldsymbol{\omega}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}\right]_+ + \ln\frac{(b+1)T\,\mathrm{card}(\mathbb{b})}{\delta}\right]\right)}_{\text{Objective function for Equation (6.2)}},$$

where the loss $\ell()$ is the bounded cross-entropy loss of Dziugaite and Roy (2018), *i.e.*, $\ell(h,(\mathbf{x},y)) = -\frac{1}{Z}\ln\Big[e^{-Z} + (1-2e^{-Z})h[y]\Big]$.

Concerning the optimization of the hyperparameters $c \in \mathbb{c}$ and $b \in \mathbb{b}$ for Equations (6.2) and (6.3), we *(i)* initialize $b \in \mathbb{b}$ or $c \in \mathbb{c}$ with the one that performs best on the first mini-batch and *(ii)* optimize by gradient descent the hyperparameter. To evaluate Equations (6.2) and (6.3), we take $b \in \mathbb{b}$ and $c \in \mathbb{c}$ that leads to the tightest bound.

F.9 Disintegrated Information-theoretic Bounds

We discuss in this section another interpretation of the disintegration procedure through Theorems F.9.1 and F.9.2 below. Actually, the RÉNYI divergence between π and ρ is sensitive to the choice of the learning sample $\mathbb{S}$: when the posterior ρ learned from $\mathbb{S}$ differs greatly from the prior π the divergence is high. To avoid such a behavior, we consider mutual information which is a measure of dependence between the random variables $\mathbb{S} \in (\mathbb{X}\times\mathbb{Y})^m$ and $h \in \mathbb{H}$. More formally, the mutual information is defined as

$$\mathrm{MI}(h;\mathbb{S}) = \min_{\pi\in\mathbb{M}^*(\mathbb{H})} \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathrm{KL}(\rho_{\mathbb{S}}\|\pi).$$

From this quantity, we can derive the generalization bound introduced in the following theorem.

Theorem F.9.1. For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any measurable function $\varphi : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y})^m \to [1, +\infty[$, for any $\delta \in (0, 1]$, for any deterministic algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\ln \varphi(h,\mathbb{S}) \leq \frac{1}{\delta} \left[\mathrm{MI}(h;\mathbb{S}) + \ln \left(\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\pi^*} \varphi(h,\mathbb{S}) \right) \right] \right] \geq 1 - \delta,$$

where π^* is defined such that $\pi^*(h) = \mathbb{E}_{\mathbb{S}\sim\mathcal{D}^m} \rho_{\mathbb{S}'}(h)$.

Proof. Deferred to Appendix F.11. ■

As for the disintegrated bounds introduced in Section 6.3, the bound on $\ln \varphi(h,\mathbb{S})$ depends on mainly two terms: a term (*i.e.*, $\mathrm{MI}(h;\mathbb{S})$) that measures the dependence of $h \in \mathbb{H}$ on the learning sample $\mathbb{S}$ and $\ln\left(\mathbb{E}_{\mathbb{S}\sim\mathcal{D}^m} \mathbb{E}_{h\sim\pi^*} \varphi(h,\mathbb{S})\right)$ that must be upper-bounded to obtain a computable bound. However, the bound has a polynomial dependence of δ, *i.e.*, we have $\frac{1}{\delta}$ instead of $\ln\frac{1}{\delta}$. To improve such dependence, we consider Sibson's mutual information (VERDÚ, 2015). It involves an expectation over the learning samples of a given size m and is defined for a given $\lambda > 1$ by

$$\mathrm{MI}_\lambda(h;\mathbb{S}) \triangleq \min_{\pi\in\mathbb{M}^*(\mathbb{H})} \frac{1}{\lambda-1} \ln \left[\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\pi} \left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)} \right]^\lambda \right].$$

The higher $\mathrm{MI}_\lambda(h;\mathbb{S})$, the higher the correlation is, meaning that the sampling of h is highly dependent on the choice of $\mathbb{S}$. This measure has two interesting properties: it generalizes the mutual information (Verdú, 2015), and it can be related to the Rényi divergence. Indeed, let $\rho(h,\mathbb{S}){=}\rho_{\mathbb{S}}(h)\mathcal{D}^m(\mathbb{S})$, *resp.* $\pi(h,\mathbb{S}){=}\pi(h)\mathcal{D}^m(\mathbb{S})$, be the probability of sampling both $\mathbb{S}{\sim}\mathcal{D}^m$ and $h{\sim}\rho_{\mathbb{S}}$, *resp.* $\mathbb{S}{\sim}\mathcal{D}^m$ and $h{\sim}\pi$. Then we can write:

$$\begin{aligned}\mathrm{MI}_\lambda(h;\mathbb{S}) &= \min_{\pi\in\mathbb{M}^*(\mathbb{H})} \frac{1}{\lambda{-}1}\ln\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi}\left[\frac{\rho_{\mathbb{S}}(h)\mathcal{D}^m(\mathbb{S})}{\pi(h)\mathcal{D}^m(\mathbb{S})}\right]^\lambda\right]\\ &= \min_{\pi\in\mathbb{M}^*(\mathbb{H})} \mathrm{D}_\lambda(\rho\|\pi). \end{aligned} \tag{F.6}$$

From Verdú (2015) the optimal prior π^* minimizing Equation (F.6) is a *distribution-dependent* prior:

$$\pi^*(h) = \frac{\left[\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\,\rho_{\mathbb{S}'}(h)^\lambda\right]^{\frac{1}{\lambda}}}{\mathbb{E}_{h'\sim\pi}\,\frac{1}{\pi(h')}\left[\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\,\rho_{\mathbb{S}'}(h')^\lambda\right]^{\frac{1}{\lambda}}}.$$

This leads to an *Information-Theoretic generalization bound*.

Theorem F.9.2 (Disintegrated Information-Theoretic Bound)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}{\times}\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any measurable function $\varphi : \mathbb{H}{\times}(\mathbb{X}{\times}\mathbb{Y})^m{\to}\mathbb{R}_+^*$, for any $\lambda{>}1$, for any $\delta \in (0,1]$, for any algorithm $A : (\mathbb{X}{\times}\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}){\to}\mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left(\frac{\lambda}{\lambda{-}1}\ln(\varphi(h,\mathbb{S})) \le \mathrm{MI}_\lambda(h';\mathbb{S}'){+}\ln\left[\frac{1}{\delta^{\frac{\lambda}{\lambda-1}}}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi^*}\left[\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right]\right]\right) \ge 1{-}\delta.$$

Proof. Deferred to Appendix F.11. ■

We can remark that Theorem F.9.2 is tighter than Theorem F.9.1. For example, when we instantiate Theorem F.9.1 with $\varphi(h,\mathbb{S}) = \exp\left[m\,\mathrm{kl}(\mathsf{R}_{\mathcal{S}}^\ell(h)\|\mathsf{R}_{\mathcal{D}}^\ell(h))\right]$, the bound will be multiplied by $\frac{1}{\delta m}$, while the bound of Theorem F.9.2 is only multiplied by $\frac{1}{m}$ (but we add the term $\frac{1}{m}\ln\frac{1}{\delta}$ to the bound which is small even for small m).
For the sake of comparison, we introduce the following corollary of Theorem F.9.2.

Corollary F.9.1. Under the assumptions of Theorem F.9.2, when $\lambda \to 1^+$, with probability at least $1-\delta$ we have

$$\ln \varphi(h,\mathbb{S}) \leq \ln \frac{1}{\delta} + \ln \left[\operatorname*{esssup}_{\mathbb{S}' \in (\mathbb{X}\times\mathbb{Y}), h' \in \mathbb{H}} \varphi(h',\mathbb{S}') \right].$$

When $\lambda \to +\infty$, with probability at least $1-\delta$ we have

$$\ln \varphi(h,\mathbb{S}) \leq \ln \left(\operatorname*{esssup}_{\mathbb{S}\in\mathbb{S}, h\in\mathbb{H}} \frac{\rho_{\mathbb{S}}(h)}{\pi^*(h)} \right) + \ln \left[\frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} \varphi(h',\mathbb{S}') \right].$$

As for Theorem 6.3.1, this corollary illustrate a trade-off introduced by λ between the Sibson's mutual information $\mathrm{MI}_\lambda(h';\mathbb{S}')$ and the term $\ln\Big(\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\, \mathbb{E}_{h'\sim\pi} \left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\Big)$.

Furthermore, Esposito *et al.* (2020, Cor.4) introduced a bound involving Sibson's mutual information. Their bound holds with probability at least $1-\delta$ over $\mathbb{S} \sim \mathcal{D}^m$ and $h \sim \rho_{\mathbb{S}}$:

$$2(\mathsf{R}^\ell_{\mathcal{S}}(h) - \mathsf{R}^\ell_{\mathcal{D}}(h))^2 \leq \tfrac{1}{m} \left[\mathrm{MI}_\lambda(h';\mathbb{S}') + \ln \tfrac{2}{\delta^{\frac{\lambda}{\lambda-1}}} \right]. \tag{F.7}$$

Hence, we compare Equation (F.7) with the equations of the following corollary.

Corollary F.9.2. For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any $\lambda > 1$, for any $\delta \in (0,1]$, for any algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$, with probability at least $1-\delta$ over $\mathbb{S} \sim \mathcal{D}^m$ and $h \sim \rho_{\mathbb{S}}$, we have

$$\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h) \| \mathsf{R}^\ell_{\mathcal{D}}(h)) \leq \tfrac{1}{m} \left[\mathrm{MI}_\lambda(h';\mathbb{S}') + \ln \tfrac{2\sqrt{m}}{\delta^{\frac{\lambda}{\lambda-1}}} \right] \tag{F.8}$$

$$\text{and} \quad 2(\mathsf{R}^\ell_{\mathcal{S}}(h) - \mathsf{R}^\ell_{\mathcal{D}}(h))^2 \leq \tfrac{1}{m} \left[\mathrm{MI}_\lambda(h';\mathbb{S}') + \ln \tfrac{2\sqrt{m}}{\delta^{\frac{\lambda}{\lambda-1}}} \right]. \tag{F.9}$$

Proof. First of all, we instantiate Theorem F.9.2 with the function $\varphi(h,\mathbb{S}) = \exp\Big[\frac{\lambda-1}{\lambda} m\, \mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h) \| \mathsf{R}^\ell_{\mathcal{D}}(h))\Big]$, we have (by rearranging the terms)

$$\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h) \| \mathsf{R}^\ell_{\mathcal{D}}(h)) \leq \frac{1}{m} \left[\mathrm{MI}_\lambda(h';\mathbb{S}') + \ln \left(\tfrac{1}{\delta^{\frac{\lambda}{\lambda-1}}} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} e^{m\, \mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}'}(h') \| \mathsf{R}^\ell_{\mathcal{D}}(h'))} \right) \right].$$

Then, from Maurer (2004), we upper-bound $\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\, \mathbb{E}_{h'\sim\pi}\, e^{m\, \mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}'}(h') \| \mathsf{R}^\ell_{\mathcal{D}}(h'))}$ by $2\sqrt{m}$ (Lemma B.16.1) to obtain Equation (F.8). Finally, to obtain Equa-

tion (F.9), we apply PINSKER's inequality (Theorem B.5.1), *i.e.*, we have the inequality $2(\mathsf{R}^\ell_{\mathcal{S}}(h)-\mathsf{R}^\ell_{\mathcal{D}}(h))^2 \leq \mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))$ on Equation (F.8). ■

Equation (F.9) is slightly looser than Equation (F.7) since it involves an extra term of $\frac{1}{m}\ln\sqrt{m}$. However, Equation (F.8) is tighter than Equation (F.7) when we have $\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(h)\|\mathsf{R}^\ell_{\mathcal{D}}(h))-2(\mathsf{R}^\ell_{\mathcal{S}}(h)-\mathsf{R}^\ell_{\mathcal{D}}(h))^2 \geq \frac{1}{m}\ln\sqrt{m}$ (which becomes more frequent as m grows). Moreover, from a theoretical view, Theorem F.9.2 brings a different philosophy than the disintegrated PAC-Bayes bounds. Indeed, in Theorems 6.3.1 and 6.3.2, given $\mathbb{S}$, the RÉNYI divergence $\mathrm{D}_\lambda(\rho_{\mathbb{S}}\|\pi)$ suggests that the learned posterior $\rho_{\mathbb{S}}$ should be close enough to the prior π to get a low bound. While in Theorem F.9.2, the Sibson's mutual information $\mathrm{MI}_\lambda(h';\mathbb{S}')$ suggests that the random variable h has to be *not too much correlated* to $\mathbb{S}$. However, the bound of Theorem F.9.2 is not computable in practice due notably to the sample expectation over the unknown distribution $\mathcal{D}$ in $\mathrm{MI}_\lambda()$. An exciting line of future works could be to study how we can make use of Theorem F.9.2 in practice.

F.10 Proof of Theorem F.9.1

In order to prove Theorem F.9.1, we need to prove Lemma F.10.1.

Lemma F.10.1. For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to [1,+\infty[$, for any $\delta\in(0,1]$, for any deterministic algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\forall\pi\in\mathbb{M}^*(\mathbb{H}),\ \ln\varphi(h,\mathbb{S}) \leq \frac{1}{\delta}\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathrm{KL}(\rho_{\mathbb{S}}\|\pi) + \ln\left(\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi}\varphi(h,\mathbb{S})\right)\right]\right] \geq 1-\delta.$$

Proof. By developing $\mathbb{E}_{\mathbb{S}\sim\mathcal{D}^m}\mathbb{E}_{h\sim\rho_{\mathbb{S}}}\ln\varphi(h,\mathbb{S})$, we have for all prior $\pi\in\mathbb{M}^*(\mathbb{H})$

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\ln\varphi(h,\mathbb{S}) &= \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\ln\left[\frac{\rho_{\mathbb{S}}(h)\pi(h)}{\pi(h)\rho_{\mathbb{S}}(h)}\varphi(h,\mathbb{S})\right]\\ &= \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\ln\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]+\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\ln\left[\frac{\pi(h)}{\rho_{\mathbb{S}}(h)}\varphi(h,\mathbb{S})\right]\\ &= \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathrm{KL}(\rho_{\mathbb{S}}\|\pi)+\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\ln\left[\frac{\pi(h)}{\rho_{\mathbb{S}}(h)}\varphi(h,\mathbb{S})\right].\end{aligned}$$

From Jensen's inequality (Theorem A.1.1), we have for all prior $\pi\in\mathbb{M}^*(\mathbb{H})$

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathrm{KL}(\rho_{\mathbb{S}}\|\pi)+\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\ln\left[\frac{\pi(h)}{\rho_{\mathbb{S}}(h)}\varphi(h,\mathbb{S})\right]\\ \leq&\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathrm{KL}(\rho_{\mathbb{S}}\|\pi)+\ln\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}}\frac{\pi(h)}{\rho_{\mathbb{S}}(h)}\varphi(h,\mathbb{S})\right]\\ =&\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathrm{KL}(\rho_{\mathbb{S}}\|\pi)+\ln\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi}\varphi(h,\mathbb{S})\right].\end{aligned}\tag{F.10}$$

Since we assume in this case that $\varphi(h,\mathbb{S})\geq 1$ for all $h\in\mathbb{H}$ and $\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m$, we have $\ln\varphi(h,\mathbb{S})\geq 0$; we can apply Markov's inequality (Theorem A.2.1) to obtain

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\ln\varphi(h,\mathbb{S})\leq\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\ln\varphi(h,\mathbb{S})\right]\geq 1-\delta.\tag{F.11}$$

Then, from Equations (F.10) and (F.11), we can deduce the stated result. ■

We are now ready to prove Theorem F.9.1.

Theorem F.9.1. For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any measurable function $\varphi:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m\to[1,+\infty[$, for any $\delta\in(0,1]$, for any deterministic algorithm $A:(\mathbb{X}\times\mathbb{Y})^m\times\mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\ln\varphi(h,\mathbb{S})\leq\frac{1}{\delta}\left[\mathrm{MI}(h;\mathbb{S})+\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi^*}\varphi(h,\mathbb{S})\right)\right]\right]\geq 1-\delta,$$

where π^* is defined such that $\pi^*(h)=\mathbb{E}_{\mathbb{S}\sim\mathcal{D}^m}\rho_{\mathbb{S}'}(h)$.

Proof. Note that the mutual information is $\mathrm{MI}(h;\mathbb{S})=\min_{\pi\in\mathbb{M}^*(\mathbb{H})}\mathbb{E}_{\mathbb{S}\sim\mathcal{D}^m}\mathrm{KL}(\rho_{\mathbb{S}}\|\pi)$. Hence, to prove Theorem F.9.1, we have to instantiate Lemma F.10.1 with the

optimal prior, *i.e.*, the prior π which minimizes $\mathbb{E}_{\mathbb{S}\sim\mathcal{D}^m} \mathrm{KL}(\rho_{\mathbb{S}}\|\pi)$. The optimal prior is well-known (see, *e.g.*, Catoni, 2007; Lever *et al.*, 2013): for the sake of completeness, we derive it. First, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathrm{KL}(\rho_{\mathbb{S}}\|\pi) &= \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \ln \frac{\rho_{\mathbb{S}}(h)}{\pi(h)} \\
&= \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \ln \left[\frac{\rho_{\mathbb{S}}(h)[\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m} \rho_{\mathbb{S}'}(h)]}{\pi(h)[\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m} \rho_{\mathbb{S}'}(h)]} \right] \\
&= \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \ln \left[\frac{\rho_{\mathbb{S}}(h)}{\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m} \rho_{\mathbb{S}'}(h)} \right] + \mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \ln \left[\frac{\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m} \rho_{\mathbb{S}'}(h)}{\pi(h)} \right].
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathop{\mathrm{argmin}}_{\pi\in\mathbb{M}^*(\mathbb{H})} \mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathrm{KL}(\rho_{\mathbb{S}}\|\pi) &= \mathop{\mathrm{argmin}}_{\pi\in\mathbb{M}^*(\mathbb{H})} \left[\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \ln \left[\frac{\rho_{\mathbb{S}}(h)}{\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m} \rho_{\mathbb{S}'}(h)} \right] \right. \\
&\qquad \left. + \mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \ln \left[\frac{\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m} \rho_{\mathbb{S}'}(h)}{\pi(h)} \right] \right] \\
&= \mathop{\mathrm{argmin}}_{\pi\in\mathbb{M}^*(\mathbb{H})} \left[\mathop{\mathbb{E}}_{h\sim\rho_{\mathbb{S}}} \ln \left[\frac{\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m} \rho_{\mathbb{S}'}(h)}{\pi(h)} \right] \right] = \pi^*,
\end{aligned}$$

where $\pi^*(h) = \mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m} \rho_{\mathbb{S}'}(h)$. Note that π^* is defined from the data distribution $\mathcal{D}$, hence, π^* is a valid prior when instantiating Lemma F.10.1 with π^*. Then, we have with probability at least $1-\delta$ over $\mathbb{S}\sim\mathcal{D}^m$ and $h\sim\rho_{\mathbb{S}}$

$$\begin{aligned}
\ln \varphi(h,\mathbb{S}) &\leq \frac{1}{\delta} \left[\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathrm{KL}(\rho_{\mathbb{S}}\|\pi^*) + \ln \left(\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\pi^*} \varphi(h,\mathbb{S}) \right) \right] \\
&= \frac{1}{\delta} \left[\mathrm{MI}(h;\mathbb{S}) + \ln \left(\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim\pi^*} \varphi(h,\mathbb{S}) \right) \right].
\end{aligned}$$

■

F.11 Proof of Theorem F.9.2

We first introduce Lemma F.11.1 in order to prove Theorem F.9.2.

Lemma F.11.1. For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution π on $\mathbb{H}$, for any measurable function $\varphi : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m$, for any $\lambda > 1$, for any $\delta\in(0,1]$, for any deterministic algorithm $A : (\mathbb{X}\times\mathbb{Y})^m \times \mathbb{M}^*(\mathbb{H}) \to \mathbb{M}(\mathbb{H})$,

we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\Bigg[\forall\pi\in\mathbb{M}^*(\mathbb{H}),\frac{\lambda}{\lambda-1}\ln(\varphi(h,\mathbb{S}))\leq \mathrm{D}_\lambda(\rho\|\pi) + \ln\Big(\frac{1}{\delta^{\frac{\lambda}{\lambda-1}}}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\Big(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\Big)\Big)\Bigg]\geq 1-\delta.$$

where $\rho(h,\mathbb{S})=\rho_{\mathbb{S}}(h)\mathcal{D}^m(\mathbb{S})$; $\pi(h,\mathbb{S})=\pi(h)\mathcal{D}^m(\mathbb{S})$.

Proof. Note that $\varphi(h,\mathbb{S})$ is a non-negative random variable. From Markov's inequality (Theorem A.2.1), we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\Big[\varphi(h,\mathbb{S})\leq\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}}}\varphi(h',\mathbb{S}')\Big]\geq 1-\delta.$$

Then, since both sides of the inequality are strictly positive, we take the logarithm to both sides of the equality and multiply by $\frac{\lambda}{\lambda-1}>0$ to obtain

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\left[\frac{\lambda}{\lambda-1}\ln\left(\varphi(h,\mathbb{S})\right)\leq\frac{\lambda}{\lambda-1}\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}'}}\varphi(h',\mathbb{S}')\right)\right]\geq 1-\delta.$$

We develop the right-hand side of the inequality in the indicator function and make the expectation of the hypothesis over the distribution π appear. We have for all priors $\pi\in\mathbb{M}^*(\mathbb{H})$,

$$\begin{aligned}
&\frac{\lambda}{\lambda-1}\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}'}}\varphi(h',\mathbb{S}')\right)\\
=&\frac{\lambda}{\lambda-1}\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}'}}\frac{\rho_{\mathbb{S}'}(h')}{\pi(h')}\frac{\pi(h')}{\rho_{\mathbb{S}'}(h')}\varphi(h',\mathbb{S}')\right)\\
=&\frac{\lambda}{\lambda-1}\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\frac{\rho_{\mathbb{S}'}(h')}{\pi(h')}\varphi(h',\mathbb{S}')\right).
\end{aligned}$$

Then, since $\frac{1}{r}+\frac{1}{s}=1$ where $r=\lambda$ and $s=\frac{\lambda}{\lambda-1}$. Hence, Hölder's inequality (Theorem A.5.1) gives

$$\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}'}}\varphi(h',\mathbb{S}')\leq\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}'}(h')}{\pi(h')}\right]^{\lambda}\right)\right]^{\frac{1}{\lambda}}\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\Big(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\Big)\right]^{\frac{\lambda-1}{\lambda}}.$$

Since both sides of the inequality are positive, we take the logarithm. Moreover, we add $\ln(\frac{1}{\delta})$, and we multiply by $\frac{\lambda}{\lambda-1} > 0$ to both sides of the inequality. We have

$$\begin{aligned}
&\frac{\lambda}{\lambda-1}\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\rho_{\mathbb{S}'}}\varphi(h',\mathbb{S}')\right)\\
\leq&\frac{\lambda}{\lambda-1}\ln\left(\frac{1}{\delta}\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}'}(h')}{\pi(h')}\right]^{\lambda}\right)\right]^{\frac{1}{\lambda}}\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\\
=&\frac{1}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}'}(h')}{\pi(h')}\right]^{\lambda}\right)\right)+\ln\left(\frac{1}{\delta^{\frac{\lambda}{\lambda-1}}}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right).
\end{aligned}$$

Hence, we can deduce that

$$\begin{aligned}
\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,h\sim\rho_{\mathbb{S}}}\Bigg[\forall\pi\in\mathbb{M}^*(\mathbb{H}),\frac{\lambda}{\lambda-1}\ln(\varphi(h,\mathbb{S}))\leq\frac{1}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}'}(h')}{\pi(h')}\right]^{\lambda}\right)\right)\\
+\ln\left(\frac{1}{\delta^{\frac{\lambda}{\lambda-1}}}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right)\Bigg]\geq 1-\delta,
\end{aligned}$$

where by definition we have $\mathrm{D}_\lambda(\rho\|\pi)=\frac{1}{\lambda-1}\ln\left(\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}'}(h')}{\pi(h')}\right]^{\lambda}\right)\right)$. ■

From Lemma F.11.1, we prove Theorem F.9.2.

Theorem F.9.2 (Disintegrated Information-Theoretic Bound)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any measurable function $\varphi:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m\to\mathbb{R}_+^*$, for any $\lambda>1$, for any $\delta\in(0,1]$, for any algorithm $A:(\mathbb{X}\times\mathbb{Y})^m\times\mathbb{M}^*(\mathbb{H})\to\mathbb{M}(\mathbb{H})$, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m,\\ h\sim\rho_{\mathbb{S}}}}\left(\frac{\lambda}{\lambda-1}\ln(\varphi(h,\mathbb{S}))\leq\mathrm{MI}_\lambda(h';\mathbb{S}')+\ln\left[\frac{1}{\delta^{\frac{\lambda}{\lambda-1}}}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi^*}\left[\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right]\right]\right)\geq 1-\delta.$$

Proof. Sibson's mutual information is $\mathrm{MI}_\lambda(h;\mathbb{S})=\min_{\pi\in\mathbb{M}^*(\mathbb{H})}\mathrm{D}_\lambda(\rho\|\pi)$. Hence, in order to prove Theorem F.9.2, we have to instantiate Lemma F.11.1 with the optimal prior, *i.e.*, the prior π which minimizes $\mathrm{D}_\lambda(\rho\|\pi)$. Actually, this optimal prior has a closed-form solution (Verdú, 2015). For the sake of completeness,

we derive it. First, we have

$$
\begin{aligned}
&\mathrm{D}_\lambda(\rho\|\pi)\\
=&\frac{1}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi}\left(\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]^\lambda\right)\right)\\
=&\frac{1}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}}(h)^\lambda\right)\right]\left(\pi(h)^{-\lambda}\right)\right)\\
=&\frac{1}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}}(h)^\lambda\right)\right]\left(\pi(h)^{-\lambda}\right)\left[\frac{\mathbb{E}_{h'\sim\pi}\frac{1}{\pi(h')}\left[\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}'}(h')^\lambda\right)\right]^{\frac{1}{\lambda}}}{\mathbb{E}_{h'\sim\pi}\frac{1}{\pi(h')}\left[\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}'}(h')^\lambda\right)\right]^{\frac{1}{\lambda}}}\right]^\lambda\right)\\
=&\frac{\lambda}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\frac{1}{\pi(h')}\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}'}(h')^\lambda\right)\right]^{\frac{1}{\lambda}}\right)\\
&+\frac{1}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h\sim\pi}\frac{1}{\pi(h)^\lambda}\left[\frac{\left[\mathbb{E}_{\mathbb{S}\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}}(h)^\lambda\right)\right]^{\frac{1}{\lambda}}}{\mathbb{E}_{h'\sim\pi}\frac{1}{\pi(h')}\left[\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}'}(h')^\lambda\right)\right]^{\frac{1}{\lambda}}}\right]^\lambda\right)\\
=&\frac{\lambda}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\frac{1}{\pi(h')}\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}'}(h')^\lambda\right)\right]^{\frac{1}{\lambda}}\right)+\mathrm{D}_\lambda(\pi^*\|\pi),
\end{aligned}
$$

where $\pi^*(h)=\left[\frac{\left[\mathbb{E}_{\mathbb{S}\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}}(h)^\lambda\right)\right]^{\frac{1}{\lambda}}}{\mathbb{E}_{h'\sim\pi}\frac{1}{\pi(h')}\left[\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}'}(h')^\lambda\right)\right]^{\frac{1}{\lambda}}}\right]$.
From these equalities and using the fact that $\mathrm{D}_\lambda(\pi^*\|\pi)$ is minimal (*i.e.*, equal to zero) when $\pi^*=\pi$, we can deduce that

$$
\begin{aligned}
&\operatorname*{argmin}_{\pi\in\mathbb{M}^*(\mathbb{H})}\mathrm{D}_\lambda(\rho\|\pi)\\
=&\operatorname*{argmin}_{\pi\in\mathbb{M}^*(\mathbb{H})}\left[\frac{\lambda}{\lambda-1}\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\frac{1}{\pi(h')}\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\left(\rho_{\mathbb{S}'}(h')^\lambda\right)\right]^{\frac{1}{\lambda}}\right)+\mathrm{D}_\lambda(\pi^*\|\pi)\right]\\
=&\operatorname*{argmin}_{\pi\in\mathbb{M}^*(\mathbb{H})}\mathrm{D}_\lambda(\pi^*\|\pi)=\pi^*.
\end{aligned}
$$

Note that π^* is defined from the data distribution $\mathcal{D}$, hence, π^* is a valid prior when instantiating Lemma F.11.1 with π^*. Then, we have with probability at least $1-\delta$ over $\mathbb{S}\sim\mathcal{D}^m$ and $h\sim\rho_{\mathbb{S}}$

$$
\begin{aligned}
\frac{\lambda}{\lambda-1}\ln(\varphi(h,\mathbb{S}))&\le\mathrm{D}_\lambda(\rho\|\pi^*)+\ln\left(\tfrac{1}{\delta^{\frac{\lambda}{\lambda-1}}}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right)\\
&=\mathrm{MI}_\lambda(h';\mathbb{S}')+\ln\left(\tfrac{1}{\delta^{\frac{\lambda}{\lambda-1}}}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right).
\end{aligned}
$$

where $\pi^*(h,\mathbb{S}) = \pi^*(h)\mathcal{D}^m(\mathbb{S})$. ■

F.12 Proof of Corollary F.9.1

Corollary F.9.1. Under the assumptions of Theorem F.9.2, when $\lambda \to 1^+$, with probability at least $1-\delta$ we have

$$\ln \varphi(h,\mathbb{S}) \leq \ln \frac{1}{\delta} + \ln \left[\operatorname*{esssup}_{\mathbb{S}' \in (\mathbb{X}\times\mathbb{Y}), h' \in \mathbb{H}} \varphi(h',\mathbb{S}') \right].$$

When $\lambda \to +\infty$, with probability at least $1-\delta$ we have

$$\ln \varphi(h,\mathbb{S}) \leq \ln \left(\operatorname*{esssup}_{\mathbb{S}\in\mathbb{S}, h\in\mathbb{H}} \frac{\rho_{\mathbb{S}}(h)}{\pi^*(h)} \right) + \ln \left[\frac{1}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} \varphi(h',\mathbb{S}') \right].$$

Proof. The proof is similar to Corollary 6.3.1. Starting from Theorem F.9.2 and rearranging, we have

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m \\ h\sim\rho_{\mathbb{S}}}} \left[\ln(\varphi(h,\mathbb{S})) \leq \frac{\lambda-1}{\lambda} \operatorname{MI}_\lambda(h';\mathbb{S}') + \ln \frac{1}{\delta} + \ln \left(\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi^*} \left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}} \right) \right]^{\frac{\lambda-1}{\lambda}} \right) \right] \geq 1-\delta,$$

Then, we will prove separately the case when $\lambda \to 1$ and $\lambda \to +\infty$.

When $\lambda \to 1$. We have $\lim_{\lambda\to 1^+} \frac{\lambda-1}{\lambda} \operatorname{MI}_\lambda(h';\mathbb{S}') = 0$. Furthermore, note that

$$\|\varphi\|_{\frac{\lambda}{\lambda-1}} = \left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi^*} \left(|\varphi(h',\mathbb{S}')|^{\frac{\lambda}{\lambda-1}} \right) \right]^{\frac{\lambda-1}{\lambda}} = \left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi^*} \left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}} \right) \right]^{\frac{\lambda-1}{\lambda}}$$

is the $L^{\frac{\lambda}{\lambda-1}}$-norm of the function $\varphi : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}_+^*$, where $\lim_{\lambda\to 1} \|\varphi\|_{\frac{\lambda}{\lambda-1}} = \lim_{\lambda'\to+\infty} \|\varphi\|_{\lambda'}$ (since we have $\lim_{\lambda\to 1^+} \frac{\lambda}{\lambda-1} = (\lim_{\lambda\to 1} \lambda)(\lim_{\lambda\to 1} \frac{1}{\lambda-1}) = +\infty$). Then, it is well known that

$$\|\varphi\|_\infty = \lim_{\lambda'\to+\infty} \|\varphi\|_{\lambda'} = \operatorname*{esssup}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}), h'\in\mathbb{H}} \varphi(h',\mathbb{S}').$$

Hence, we have

$$\begin{aligned}
&\lim_{\lambda\to 1}\ln\left(\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi^*}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\\
=&\ln\left(\lim_{\lambda\to 1}\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi^*}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\\
=&\ln\left(\lim_{\lambda\to 1}\|\varphi\|_{\frac{\lambda}{\lambda-1}}\right)=\ln\left(\lim_{\lambda'\to+\infty}\|\varphi\|_{\lambda'}\right)\\
=&\ln\left(\|\varphi\|_\infty\right)=\ln\left(\operatorname*{esssup}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}),h'\in\mathbb{H}}\varphi(h',\mathbb{S}')\right).
\end{aligned}$$

Finally, we can deduce that

$$\begin{aligned}
&\lim_{\lambda\to 1}\left[\frac{\lambda-1}{\lambda}\,\mathrm{MI}_\lambda(h';\mathbb{S}')+\ \ln\frac{1}{\delta}+\ln\left(\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi^*}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\right]\\
=&\ln\frac{1}{\delta}+\ln\left[\operatorname*{esssup}_{\mathbb{S}'\in(\mathbb{X}\times\mathbb{Y}),h'\in\mathbb{H}}\varphi(h',\mathbb{S}')\right].
\end{aligned}$$

When $\lambda\to+\infty$. First, we have $\lim_{\lambda\to+\infty}\|\varphi\|_{\frac{\lambda}{\lambda-1}}=\lim_{\lambda'\to 1}\|\varphi\|_{\lambda'}=\|\varphi\|_1$ Hence, we have

$$\begin{aligned}
&\lim_{\lambda\to+\infty}\ln\left(\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi^*}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\\
=&\ln\left(\lim_{\lambda\to+\infty}\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi^*}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\\
=&\ln\left(\lim_{\lambda\to+\infty}\|\varphi\|_{\frac{\lambda}{\lambda-1}}\right)=\ln\left(\lim_{\lambda'\to 1}\|\varphi\|_{\lambda'}\right)\\
=&\ln\left(\|\varphi\|_1\right)=\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi^*}\varphi(h',\mathbb{S}')\right).
\end{aligned}$$

Moreover, by rearranging the terms in $\frac{\lambda-1}{\lambda}\,\mathrm{MI}_\lambda(h';\mathbb{S}')$, we have

$$\begin{aligned}\frac{\lambda-1}{\lambda}\,\mathrm{MI}_\lambda(h';\mathbb{S}') &= \frac{1}{\lambda}\ln\left(\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi^*}\left(\left[\frac{\rho_{\mathbb{S}}(h)}{\pi^*(h)}\right]^\lambda\right)\right)\\ &= \ln\left(\left[\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi^*}\left(\left[\frac{\rho_{\mathbb{S}}(h)}{\pi^*(h)}\right]^\lambda\right)\right]^{\frac{1}{\lambda}}\right)\\ &= \ln\left(\left[\mathop{\mathbb{E}}_{h\sim\pi^*}\left(\gamma(h)^\lambda\right)\right]^{\frac{1}{\lambda}}\right) = \ln(\|\gamma\|_\lambda)\,,\end{aligned}$$

where $\|\gamma\|_\lambda$ is the L^λ-norm of the function γ defined as $\gamma(h)=\frac{\rho_{\mathbb{S}}(h)}{\pi^*(h)}$. We have

$$\begin{aligned}\lim_{\lambda\to+\infty}\frac{\lambda-1}{\lambda}\,\mathrm{MI}_\lambda(h';\mathbb{S}') &= \lim_{\lambda\to+\infty}\ln(\|\gamma\|_\lambda) = \ln\left(\lim_{\lambda\to+\infty}\|\gamma\|_\lambda\right)\\ &= \ln\left(\|\gamma\|_\infty\right) = \ln\left(\operatorname*{esssup}_{\mathbb{S}\in\mathbb{S},h\in\mathbb{H}}\gamma(h)\right) = \ln\left(\operatorname*{esssup}_{\mathbb{S}\in\mathbb{S},h\in\mathbb{H}}\frac{\rho_{\mathbb{S}}(h)}{\pi^*(h)}\right).\end{aligned}$$

Finally, we can deduce that

$$\begin{aligned}&\lim_{\lambda\to1}\left[\frac{\lambda-1}{\lambda}\,\mathrm{MI}_\lambda(h';\mathbb{S}') + \ln\frac{1}{\delta} + \ln\left(\left[\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi^*}\left(\varphi(h',\mathbb{S}')^{\frac{\lambda}{\lambda-1}}\right)\right]^{\frac{\lambda-1}{\lambda}}\right)\right]\\ =\;& \ln\left(\operatorname*{esssup}_{\mathbb{S}\in\mathbb{S},h\in\mathbb{H}}\frac{\rho_{\mathbb{S}}(h)}{\pi^*(h)}\right) + \ln\left[\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\varphi(h',\mathbb{S}')\right].\end{aligned}$$

■

F.13 Proof of Corollary 6.6.1

Corollary 6.6.1 (Instantiation of Theorem 2.4.1 to Stochastic Majority Votes)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any finite set of voters $\mathbb{H}$, for any hyper-prior distribution $\Pi=\mathrm{Dir}(\boldsymbol{\beta})$ on $\mathbb{H}$ with $\boldsymbol{\beta}\in(\mathbb{R}_*^+)^{\mathrm{card}(\mathbb{H})}$, for any loss $\ell:\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})\to[0,1]$, for any $\delta\in(0,1]$, for any algorithm A that outputs a hyper-posterior given a learning sample and a hyper-prior, with probability at least $1-\delta$ over the learning sample $\mathbb{S}\sim\mathcal{D}^m$ and the posterior distribution $\rho\sim\mathrm{P}_{\mathbb{S}}=\mathrm{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha}\in(\mathbb{R}_*^+)^{\mathrm{card}(\mathbb{H})}$ we have

$$\mathrm{kl}(\mathsf{R}^\ell_{\mathcal{S}}(\mathsf{MV}_\rho)\|\mathsf{R}^\ell_{\mathcal{D}}(\mathsf{MV}_\rho)) \le \frac{1}{m}\left[\ln\frac{Z(\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})} + \sum_{j=1}^{\mathrm{card}(\mathbb{H})}(\alpha_j-\beta_j)\ln(\rho(h_j)) + \ln\frac{2\sqrt{m}}{\delta}\right],$$

where $\mathrm{P}_{\mathbb{S}} \triangleq A(\mathbb{S}, \Pi)$ is output by the deterministic algorithm A.

Proof. We apply Theorem 2.4.1 with $\phi(\rho, \mathbb{S}) = m\,\mathrm{kl}(\mathsf{R}^{\ell}_{\mathcal{S}}(\mathsf{MV}_{\rho}) \| \mathsf{R}^{\ell}_{\mathcal{D}}(\mathsf{MV}_{\rho}))$ to obtain with probability at least $1-\delta$ over the learning sample $\mathbb{S} \sim \mathcal{D}^m$ and the posterior distribution $\rho \sim \mathrm{P}$, we have

$$\mathrm{kl}(\mathsf{R}^{\ell}_{\mathcal{S}}(\mathsf{MV}_{\rho}) \| \mathsf{R}^{\ell}_{\mathcal{D}}(\mathsf{MV}_{\rho})) \leq \frac{1}{m}\left[\ln \frac{\mathrm{P}(\rho)}{\Pi(\rho)} + \ln\left[\frac{1}{\delta} \mathop{\mathbb{E}}_{\rho' \sim \Pi} e^{m\,\mathrm{kl}(\mathsf{R}^{\ell}_{\mathcal{S}}(\mathsf{MV}_{\rho'}) \| \mathsf{R}^{\ell}_{\mathcal{D}}(\mathsf{MV}_{\rho'}))}\right]\right]. \quad \text{(F.12)}$$

Moreover, the closed form solution of the disintegrated KL divergence $\ln \frac{\mathrm{P}(\rho)}{\Pi(\rho)}$ is

$$\begin{aligned}
\ln \frac{\mathrm{P}(\rho)}{\Pi(\rho)} &= \ln(\mathrm{P}(\rho)) - \ln(\Pi(\rho)) \\
&= \ln\left(\frac{1}{Z(\boldsymbol{\alpha})} \prod_{j=1}^{\mathrm{card}(\mathbb{H})} \Big[\rho(h_j)\Big]^{\alpha_j - 1}\right) - \ln\left(\frac{1}{Z(\boldsymbol{\beta})} \prod_{j=1}^{\mathrm{card}(\mathbb{H})} \Big[\rho(h_j)\Big]^{\beta_j - 1}\right) \\
&= \ln \frac{Z(\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})} + \sum_{j=1}^{\mathrm{card}(\mathbb{H})} (\alpha_j - 1) \ln(\rho(h_j)) - \sum_{j=1}^{\mathrm{card}(\mathbb{H})} (\beta_j - 1) \ln(\rho(h_j)) \\
&= \ln \frac{Z(\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})} + \sum_{j=1}^{\mathrm{card}(\mathbb{H})} (\alpha_j - \beta_j) \ln(\rho(h_j)). \qquad \text{(F.13)}
\end{aligned}$$

Additionally, from Lemma B.16.1, we have

$$\mathop{\mathbb{E}}_{\rho' \sim \Pi} e^{m\,\mathrm{kl}(\mathsf{R}^{\ell}_{\mathcal{S}}(\mathsf{MV}_{\rho'}) \| \mathsf{R}^{\ell}_{\mathcal{D}}(\mathsf{MV}_{\rho'}))} \leq 2\sqrt{m}. \qquad \text{(F.14)}$$

Lastly, by merging Equations (F.12) to (F.14) we obtain the claim. ■

F.14 Proof of Corollary 6.6.2

Corollary 6.6.2 (Instantiation of Theorem 6.3.1 to Stochastic Majority Votes)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any finite set of voters $\mathbb{H}$, for any hyper-prior distribution $\Pi = \mathrm{Dir}(\boldsymbol{\beta})$ on $\mathbb{H}$ with $\boldsymbol{\beta} \in (\mathbb{R}^{+}_{*})^{\mathrm{card}(\mathbb{H})}$, for any loss $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$, for any $\lambda > 1$, for any $\delta \in (0,1]$, for any algorithm A that outputs a hyper-posterior given a learning sample and a hyper-prior, with probability at least $1-\delta$ over the learning sample $\mathbb{S}\sim\mathcal{D}^m$ and the posterior distribution $\rho\sim\mathrm{P}_{\mathbb{S}} = \mathrm{Dir}(\boldsymbol{\alpha})$

with $\boldsymbol{\alpha} \in (\mathbb{R}_*^+)^{\mathrm{card}(\mathbb{H})}$ we have

$$\mathrm{kl}(\mathsf{R}_{\mathcal{S}}^{\ell}(\mathsf{MV}_\rho)\|\mathsf{R}_{\mathcal{D}}^{\ell}(\mathsf{MV}_\rho)) \leq \frac{1}{m}\left[\frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta} + \ln\frac{Z(\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})} + \frac{1}{\lambda-1}\ln\frac{Z(\lambda\boldsymbol{\alpha}+(1-\lambda)\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})} + \ln(2\sqrt{m})\right],$$

where $\mathrm{P}_{\mathbb{S}} \triangleq A(\mathbb{S}, \Pi)$ is output by the deterministic algorithm A.

Proof. We apply Theorem 6.3.1 with $\phi(\rho, \mathbb{S}) = \exp\left[\frac{\lambda-1}{\lambda} m\, \mathrm{kl}(\mathsf{R}_{\mathcal{S}}^{\ell}(\mathsf{MV}_\rho)\|\mathsf{R}_{\mathcal{D}}^{\ell}(\mathsf{MV}_\rho))\right]$ to obtain with probability at least $1-\delta$ over the learning sample $\mathbb{S} \sim \mathcal{D}^m$ and the posterior distribution $\rho \sim \mathrm{P}$, we have

$$\mathrm{kl}(\mathsf{R}_{\mathcal{S}}^{\ell}(\mathsf{MV}_\rho)\|\mathsf{R}_{\mathcal{D}}^{\ell}(\mathsf{MV}_\rho)) \leq \frac{1}{m}\left[\frac{2\lambda-1}{\lambda-1}\ln\frac{2}{\delta} + \mathrm{D}_\lambda(\mathrm{P}\|\Pi) + \ln\left(\mathop{\mathbb{E}}_{\rho'\sim\Pi} e^{m\,\mathrm{kl}(\mathsf{R}_{\mathcal{S}}^{\ell}(\mathsf{MV}_{\rho'})\|\mathsf{R}_{\mathcal{D}}^{\ell}(\mathsf{MV}_{\rho'}))}\right)\right]. \tag{F.15}$$

Moreover, the closed form solution of the RÉNYI divergence $\mathrm{D}_\lambda(\mathrm{P}\|\Pi)$ (GIL *et al.*,

2013) is given by

$$\begin{aligned}
\mathrm{D}_\lambda(\mathrm{P}\|\Pi) &= \frac{1}{\lambda-1}\ln\left(\int_{\mathbb{M}(\mathbb{H})}\mathrm{P}(\rho)^\lambda\Pi(\rho)^{1-\lambda}d\xi(\rho)\right)\\
&= \frac{1}{\lambda-1}\ln\left(\int_{\mathbb{M}(\mathbb{H})}\frac{1}{Z(\boldsymbol{\alpha})^\lambda}\prod_{j=1}^{\mathrm{card}(\mathbb{H})}(\rho(h_j))^{\lambda(\alpha_j-1)}\right.\\
&\qquad\left.\frac{1}{Z(\boldsymbol{\beta})^{1-\lambda}}\prod_{j=1}^{\mathrm{card}(\mathbb{H})}(\rho(h_j))^{(1-\lambda)(\beta_j-1)}d\xi(\rho)\right)\\
&= \frac{1}{\lambda-1}\ln\left(\frac{Z(\boldsymbol{\beta})^{\lambda-1}}{Z(\boldsymbol{\alpha})^\lambda}\right)\\
&\quad+\frac{1}{\lambda-1}\ln\left(\int_{\mathbb{M}(\mathbb{H})}\prod_{i=1}^{\mathrm{card}(\mathbb{H})}(\rho(h_j))^{\lambda\alpha_j+(1-\lambda)\beta_j-1}d\xi(\rho)\right)\\
&= \frac{1}{\lambda-1}\ln\left(\frac{Z(\boldsymbol{\beta})^{\lambda-1}}{Z(\boldsymbol{\alpha})^\lambda}\right)+\ln Z(\lambda\boldsymbol{\alpha}+(1-\lambda)\boldsymbol{\beta})\\
&= \frac{1}{\lambda-1}\ln\left(\frac{Z(\boldsymbol{\beta})^{\lambda-1}}{Z(\boldsymbol{\alpha})^{\lambda-1+1}}\right)+\ln Z(\lambda\boldsymbol{\alpha}+(1-\lambda)\boldsymbol{\beta})\\
&= \ln\frac{Z(\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})}+\frac{1}{\lambda-1}\ln\frac{Z(\lambda\boldsymbol{\alpha}+(1-\lambda)\boldsymbol{\beta})}{Z(\boldsymbol{\alpha})},
\end{aligned}\tag{F.16}$$

where ξ is the reference measure on $\mathbb{M}(\mathbb{H})$. Additionally, from Lemma B.16.1, we have

$$\mathop{\mathbb{E}}_{\rho'\sim\Pi} e^{m\,\mathrm{kl}(\mathsf{R}^\ell_S(\mathsf{MV}_{\rho'})\|\mathsf{R}^\ell_{\mathcal{D}}(\mathsf{MV}_{\rho'}))}\leq 2\sqrt{m}.\tag{F.17}$$

Lastly, by merging Equations (F.15) to (F.17) we obtain the claim. ■

F.15 Details of the Results

Table F.2 to Table F.10 report empirical results for split ratios going from 0.0 to 0.9. Table F.11 to Table F.13 report the performances of the prior before applying Step **2)**.

For the split 0.0, since Step **1)** is skipped, the prior distribution π is only initialized as introduced in Section 6.5.3.2. Note that in this case, $T = 1$ since we have only one prior. To do the same number of epochs compared to the other splits, we perform 11 epochs (instead of 1) for MNIST and Fashion-MNIST and 110 epochs (instead of 10) for CIFAR-10 during Step **2)**. The other parameters are not changed.

Table F.1. *Comparison of* ours, rivasplata, blanchard *and* catoni *based on the disintegrated bounds, and* stochastic *based on the randomized bounds learned with two learning rates* $\eta \in \{10^{-4}, 10^{-6}\}$ *and different variances* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. *We report the test risk (*$R_{\mathcal{T}}(h)$*), the bound value (Bnd), the empirical risk (*$R_{\mathcal{S}}(h)$*), and the divergence (Div) associated with each bound (the* RÉNYI *divergence for* ours, *the KL divergence for* stochastic, *and the disintegrated KL divergence for* rivasplata, blanchard *and* catoni*). More precisely, we report the mean* $\pm$ *the standard deviation for* 400 *neural networks sampled from* $\rho_{\mathbb{S}}$ *for* ours, rivasplata, blanchard, *and* catoni. *We consider, in this figure, that the split ratio is* 0.0.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.901 ± .002	.908 ± .002	.901 ± .002	.005	.897 ± .013	.904 ± .012	.897 ± .012	.009	.898 ± .017	.905 ± .016	.898 ± .016	.027	.902 ± .015	.908 ± .014	.901 ± .015	.671
	blanchard	.901 ± .002	.926 ± .002	.901 ± .002	122.846 ± 15.952	.897 ± .013	.912 ± .012	.897 ± .013	39.350 ± 8.999	.898 ± .017	.907 ± .016	.898 ± .017	13.023 ± 4.818	.901 ± .015	.907 ± .014	.901 ± .014	3.041 ± 2.459
	catoni	.901 ± .002	.926 ± .003	.901 ± .002	121.860 ± 15.930	.897 ± .013	.909 ± .012	.897 ± .013	38.552 ± 8.872	.898 ± .017	.905 ± .016	.898 ± .017	12.474 ± 4.774	.901 ± .014	.906 ± .013	.901 ± .014	3.088 ± 2.379
	rivasplata	.901 ± .002	.920 ± .002	.901 ± .002	123.301 ± 15.941	.896 ± .014	.908 ± .012	.896 ± .013	39.195 ± 8.959	.897 ± .017	.905 ± .016	.897 ± .017	12.827 ± 4.858	.902 ± .015	.907 ± .014	.901 ± .015	3.232 ± 2.454
	stochastic	—	.944	—	.002	—	.941	—	.004	—	.941	—	.014	—	.944	—	.336
Fashion	ours	.970 ± .028	.972 ± .025	.970 ± .027	.016	.944 ± .038	.949 ± .035	.944 ± .037	.046	.910 ± .027	.917 ± .026	.910 ± .027	.140	.901 ± .026	.909 ± .025	.901 ± .026	1.255
	blanchard	.970 ± .029	.978 ± .019	.970 ± .028	122.508 ± 16.085	.942 ± .038	.952 ± .032	.943 ± .038	39.957 ± 8.610	.910 ± .031	.919 ± .029	.910 ± .031	12.649 ± 4.846	.899 ± .028	.905 ± .027	.899 ± .028	3.206 ± 2.566
	catoni	.970 ± .028	.983 ± .017	.970 ± .027	122.364 ± 15.860	.945 ± .038	.954 ± .036	.945 ± .037	38.555 ± 8.873	.912 ± .032	.919 ± .031	.912 ± .032	12.167 ± 4.762	.899 ± .027	.905 ± .026	.899 ± .027	3.122 ± 2.392
	rivasplata	.970 ± .028	.977 ± .021	.971 ± .027	123.328 ± 15.929	.943 ± .038	.950 ± .033	.943 ± .038	39.300 ± 8.991	.908 ± .031	.916 ± .029	.908 ± .031	12.627 ± 4.890	.899 ± .028	.905 ± .027	.899 ± .028	3.591 ± 2.610
	stochastic	—	.990	—	.008	—	.975	—	.023	—	.950	—	.070	—	.944	—	.627
CIFAR-10	ours	.899 ± .000	.907 ± .000	.899 ± .000	3.113	.896 ± .002	.914 ± .002	.894 ± .002	107.797	.826 ± .011	.885 ± .009	.825 ± .010	76.475	.786 ± .019	.851 ± .015	.788 ± .018	714.351
	blanchard	.899 ± .000	.940 ± .001	.898 ± .000	314.983 ± 26.377	.888 ± .004	.927 ± .002	.885 ± .003	28.250 ± 25.255	.823 ± .010	.885 ± .008	.822 ± .010	422.401 ± 29.323	.798 ± .019	.856 ± .015	.799 ± .018	292.706 ± 25.318
	catoni	.899 ± .000	.941 ± .000	.898 ± .000	285.415 ± 25.085	.894 ± .002	.930 ± .004	.892 ± .002	169.713 ± 19.543	.857 ± .010	.915 ± .009	.856 ± .010	273.554 ± 23.212	.815 ± .019	.864 ± .017	.816 ± .018	209.069 ± 21.230
	rivasplata	.899 ± .001	.930 ± .001	.898 ± .000	362.070 ± 28.420	.864 ± .004	.933 ± .002	.862 ± .004	1568.007 ± 55.492	.748 ± .010	.837 ± .007	.750 ± .009	1219.178 ± 49.610	.769 ± .018	.828 ± .015	.771 ± .017	526.068 ± 33.837
	stochastic	—	.942	—	1.557	—	.945	—	53.898	—	.914	—	38.237	—	.884	—	357.175
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.901 ± .002	.909 ± .002	.901 ± .002	3.767	.896 ± .014	.904 ± .013	.896 ± .014	.835	.898 ± .016	.905 ± .015	.898 ± .016	1.062	.901 ± .015	.909 ± .014	.901 ± .015	6.022
	blanchard	.900 ± .003	.990 ± .000	.900 ± .003	12004.196 ± 152.632	.894 ± .017	.986 ± .006	.894 ± .016	3837.785 ± 93.560	.888 ± .021	.957 ± .013	.888 ± .020	1221.198 ± 49.920	.898 ± .015	.939 ± .012	.897 ± .015	391.343 ± 28.182
	catoni	.900 ± .003	.997 ± .002	.900 ± .003	5694.194 ± 102.906	.889 ± .020	.967 ± .012	.889 ± .019	3331.617 ± 78.945	.879 ± .025	.941 ± .016	.880 ± .025	1481.726 ± 53.973	.888 ± .023	.937 ± .015	.888 ± .023	567.893 ± 33.441
	rivasplata	.900 ± .004	.990 ± .000	.900 ± .003	1199.818 ± 152.557	.892 ± .017	.970 ± .009	.892 ± .016	3846.699 ± 84.643	.886 ± .020	.940 ± .015	.886 ± .020	1224.463 ± 49.970	.897 ± .018	.928 ± .015	.897 ± .018	393.757 ± 29.158
	stochastic	—	.944	—	1.884	—	.940	—	.417	—	.941	—	.531	—	.944	—	3.011
Fashion	ours	.977 ± .024	.979 ± .021	.977 ± .023	3.926	.947 ± .038	.951 ± .035	.947 ± .038	1.623	.907 ± .030	.914 ± .029	.907 ± .030	2.947	.900 ± .026	.910 ± .025	.900 ± .026	15.978
	blanchard	.984 ± .015	.990 ± .000	.984 ± .015	12019.121 ± 166.251	.912 ± .029	.988 ± .004	.911 ± .029	3846.861 ± 84.568	.883 ± .029	.953 ± .019	.883 ± .029	1232.645 ± 5.285	.403 ± .041	.648 ± .038	.399 ± .041	3853.231 ± 87.867
	catoni	.983 ± .018	1.000 ± .000	.983 ± .017	5654.642 ± 114.040	.903 ± .021	.985 ± .012	.902 ± .021	4354.538 ± 94.427	.751 ± .033	.867 ± .023	.750 ± .033	2702.652 ± 76.863	.504 ± .041	.673 ± .037	.502 ± .041	3172.609 ± 78.698
	rivasplata	.983 ± .016	.990 ± .000	.983 ± .016	11976.720 ± 165.964	.905 ± .023	.975 ± .007	.905 ± .023	3855.872 ± 84.676	.855 ± .035	.916 ± .027	.855 ± .035	125.110 ± 51.837	.365 ± .032	.559 ± .032	.359 ± .033	4823.725 ± 103.813
	stochastic	—	.990	—	1.963	—	.977	—	.812	—	.948	—	1.473	—	.944	—	7.989
CIFAR-10	ours	.899 ± .000	.915 ± .000	.899 ± .000	63.416	.890 ± .003	.932 ± .003	.886 ± .003	68.353	.786 ± .011	.888 ± .008	.787 ± .010	2072.610	.769 ± .017	.859 ± .013	.770 ± .017	1406.824
	blanchard	.869 ± .002	.990 ± .000	.866 ± .001	27237.938 ± 251.770	.813 ± .004	.990 ± .000	.812 ± .003	12052.733 ± 159.732	.697 ± .011	.920 ± .005	.700 ± .009	5137.799 ± 103.680	.674 ± .020	.861 ± .014	.675 ± .020	2814.450 ± 76.004
	catoni	.928 ± .001	1.000 ± .000	.925 ± .001	2145276.795 ± 2095.160	.821 ± .002	1.000 ± .000	.821 ± .002	375019.277 ± 896.780	.689 ± .011	.870 ± .007	.692 ± .010	5292.535 ± 106.380	.629 ± .019	.805 ± .015	.628 ± .019	4159.131 ± 96.763
	rivasplata	.867 ± .002	.990 ± .000	.864 ± .001	35956.152 ± 268.304	.812 ± .004	.976 ± .001	.811 ± .003	12135.134 ± 157.621	.698 ± .010	.874 ± .006	.701 ± .009	5191.665 ± 102.712	.677 ± .020	.819 ± .015	.678 ± .019	2839.514 ± 81.432
	stochastic	—	.947	—	31.708	—	.954	—	34.176	—	.908	—	1036.305	—	.886	—	703.412

Table F.2. *Comparison of* `ours`, `rivasplata`, `blanchard` *and* `catoni` *based on the disintegrated bounds, and* `stochastic` *based on the randomized bounds learned with two learning rates* $\eta \in \{10^{-4}, 10^{-6}\}$ *and different variances* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. *We report the test risk (*$R_{\mathcal{T}}(h)$*), the bound value (Bnd), the empirical risk (*$R_{\mathcal{S}}(h)$*), and the divergence (Div) associated with each bound (the* RÉNYI *divergence for* `ours`, *the KL divergence for* `stochastic`, *and the disintegrated KL divergence for* `rivasplata`, `blanchard` *and* `catoni`*). More precisely, we report the mean* $\pm$ *the standard deviation for* 400 *neural networks sampled from* $\rho_{\mathbb{S}}$ *for* `ours`, `rivasplata`, `blanchard`, *and* `catoni`. *We consider, in this table, that the split ratio is* 0.1.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.035 ± .000	.044 ± .000	.039 ± .000	.622	.024 ± .000	.034 ± .000	.029 ± .000	2.122	.029 ± .002	.040 ± .002	.034 ± .002	12.754	.034 ± .004	.044 ± .004	.038 ± .004	7.303
	blanchard	.034 ± .000	.058 ± .002	.038 ± .000	99.876 ± 14.858	.024 ± .000	.038 ± .001	.030 ± .000	21.775 ± 6.848	.034 ± .002	.043 ± .002	.038 ± .002	3.949 ± 2.877	.039 ± .005	.047 ± .005	.043 ± .005	.590 ± 1.085
	catoni	.035 ± .000	.064 ± .001	.039 ± .000	119.663 ± 15.854	.024 ± .000	.038 ± .001	.030 ± .000	26.277 ± 7.490	.033 ± .002	.041 ± .002	.037 ± .002	4.067 ± 2.882	.038 ± .005	.045 ± .005	.042 ± .004	.759 ± 1.217
	rivasplata	.034 ± .000	.052 ± .001	.038 ± .000	104.880 ± 15.268	.024 ± .000	.036 ± .001	.029 ± .000	23.007 ± 7.187	.033 ± .002	.042 ± .002	.037 ± .002	4.116 ± 2.845	.038 ± .005	.046 ± .004	.042 ± .004	.775 ± 1.231
	stochastic	—	.080	—	.311	—	.067	—	1.061	—	.074	—	6.377	—	.079	—	3.651
Fashion	ours	.166 ± .001	.169 ± .000	.159 ± .000	.580	.157 ± .001	.160 ± .001	.150 ± .001	2.128	.160 ± .002	.161 ± .003	.151 ± .002	3.503	.176 ± .006	.179 ± .006	.168 ± .005	1.268
	blanchard	.165 ± .001	.192 ± .002	.159 ± .000	96.822 ± 14.116	.157 ± .001	.166 ± .002	.150 ± .001	21.592 ± 6.681	.163 ± .003	.162 ± .003	.153 ± .003	3.846 ± 2.660	.178 ± .005	.178 ± .005	.170 ± .005	.463 ± .954
	catoni	.165 ± .001	.190 ± .003	.159 ± .000	119.927 ± 15.938	.157 ± .001	.163 ± .002	.150 ± .001	26.363 ± 7.355	.162 ± .003	.161 ± .003	.152 ± .003	4.152 ± 2.945	.177 ± .006	.178 ± .006	.169 ± .006	.548 ± 1.032
	rivasplata	.165 ± .001	.183 ± .002	.158 ± .000	101.954 ± 14.463	.157 ± .001	.163 ± .002	.150 ± .001	23.098 ± 6.977	.162 ± .003	.161 ± .003	.153 ± .003	3.852 ± 2.798	.177 ± .006	.177 ± .006	.169 ± .006	.516 ± .985
	stochastic	—	.227	—	.290	—	.216	—	1.064	—	.218	—	1.751	—	.237	—	.634
CIFAR-10	ours	.479 ± .000	.487 ± .000	.472 ± .000	.052	.479 ± .000	.493 ± .000	.477 ± .000	.065	.458 ± .001	.479 ± .000	.463 ± .000	.299	.480 ± .002	.495 ± .001	.480 ± .001	.793
	blanchard	.479 ± .000	.550 ± .003	.472 ± .000	27.644 ± 22.868	.479 ± .000	.522 ± .003	.477 ± .000	85.476 ± 12.781	.458 ± .001	.489 ± .003	.463 ± .000	24.608 ± 7.136	.481 ± .002	.495 ± .002	.480 ± .001	5.093 ± 3.299
	catoni	.479 ± .000	.546 ± .005	.472 ± .000	269.855 ± 22.883	.479 ± .000	.511 ± .003	.477 ± .000	85.113 ± 12.806	.458 ± .001	.483 ± .002	.463 ± .000	25.453 ± 7.155	.480 ± .002	.495 ± .001	.480 ± .001	5.468 ± 3.315
	rivasplata	.479 ± .000	.528 ± .002	.472 ± .000	27.588 ± 22.859	.479 ± .000	.511 ± .002	.477 ± .000	85.745 ± 13.357	.458 ± .001	.484 ± .002	.463 ± .001	25.051 ± 7.005	.481 ± .002	.494 ± .001	.480 ± .001	5.155 ± 3.260
	stochastic	—	.558	—	.026	—	.564	—	.032	—	.550	—	.150	—	.566	—	.397
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.035 ± .000	.048 ± .000	.039 ± .000	35.348	.024 ± .000	.037 ± .001	.029 ± .000	3.753	.022 ± .001	.042 ± .001	.027 ± .001	153.773	.025 ± .002	.041 ± .002	.029 ± .002	97.840
	blanchard	.032 ± .000	.442 ± .003	.036 ± .000	1181.482 ± 14.449	.022 ± .000	.206 ± .003	.027 ± .000	3851.110 ± 84.274	.019 ± .001	.102 ± .002	.023 ± .001	1306.371 ± 51.396	.024 ± .002	.065 ± .003	.027 ± .002	411.772 ± 29.458
	catoni	.035 ± .000	.362 ± .003	.039 ± .000	11925.734 ± 145.511	.024 ± .000	.152 ± .002	.029 ± .000	3841.248 ± 84.033	.027 ± .002	.084 ± .002	.032 ± .001	1235.287 ± 49.454	.027 ± .002	.059 ± .003	.030 ± .002	403.300 ± 28.587
	rivasplata	.030 ± .000	.289 ± .002	.034 ± .000	12022.576 ± 151.157	.021 ± .000	.134 ± .002	.026 ± .000	3912.803 ± 85.146	.018 ± .000	.072 ± .001	.022 ± .000	1348.169 ± 53.400	.023 ± .002	.051 ± .002	.026 ± .001	424.971 ± 29.301
	stochastic	—	.084	—	17.674	—	.069	—	15.376	—	.072	—	76.887	—	.072	—	48.920
Fashion	ours	.166 ± .001	.172 ± .000	.159 ± .000	13.084	.157 ± .001	.163 ± .001	.150 ± .001	16.513	.159 ± .002	.164 ± .002	.149 ± .002	2.344	.176 ± .005	.181 ± .005	.168 ± .005	11.331
	blanchard	.160 ± .001	.588 ± .003	.153 ± .000	1089.829 ± 137.125	.150 ± .001	.379 ± .003	.141 ± .001	3744.491 ± 83.656	.155 ± .002	.271 ± .003	.145 ± .002	1221.062 ± 49.548	.173 ± .005	.233 ± .006	.165 ± .004	369.721 ± 27.211
	catoni	.165 ± .001	.500 ± .003	.159 ± .000	11954.591 ± 141.463	.156 ± .001	.311 ± .002	.148 ± .001	3826.848 ± 86.111	.158 ± .002	.248 ± .003	.148 ± .002	1226.282 ± 5.332	.174 ± .005	.252 ± .006	.166 ± .004	393.542 ± 27.890
	rivasplata	.158 ± .001	.459 ± .002	.151 ± .000	11541.128 ± 14.706	.149 ± .001	.302 ± .002	.140 ± .001	3878.145 ± 85.782	.154 ± .002	.230 ± .002	.144 ± .001	1244.035 ± 49.268	.172 ± .005	.212 ± .005	.164 ± .004	378.990 ± 27.559
	stochastic	—	.229	—	6.542	—	.219	—	8.257	—	.219	—	1.172	—	.239	—	5.666
CIFAR-10	ours	.479 ± .000	.489 ± .000	.472 ± .000	4.882	.479 ± .000	.496 ± .000	.477 ± .000	9.273	.458 ± .001	.480 ± .000	.463 ± .000	4.988	.480 ± .002	.497 ± .001	.479 ± .001	8.681
	blanchard	.479 ± .000	.957 ± .001	.471 ± .000	22201.935 ± 218.369	.479 ± .000	.854 ± .002	.477 ± .000	8777.551 ± 125.716	.457 ± .001	.699 ± .003	.461 ± .000	2758.075 ± 77.155	.474 ± .001	.613 ± .003	.472 ± .001	903.948 ± 4.742
	catoni	.479 ± .000	.995 ± .000	.471 ± .000	26347.736 ± 225.908	.479 ± .000	.771 ± .002	.477 ± .000	8566.272 ± 124.834	.455 ± .001	.650 ± .002	.459 ± .000	3117.566 ± 75.178	.468 ± .001	.621 ± .001	.466 ± .001	1481.520 ± 52.533
	rivasplata	.479 ± .000	.915 ± .001	.471 ± .000	29489.241 ± 241.010	.479 ± .000	.765 ± .002	.477 ± .000	867.264 ± 126.038	.456 ± .001	.633 ± .002	.460 ± .000	2776.052 ± 72.901	.472 ± .001	.572 ± .002	.470 ± .001	937.091 ± 42.116
	stochastic	—	.559	—	2.441	—	.566	—	4.637	—	.551	—	2.494	—	.567	—	4.340

Table F.3. *Comparison of* `ours`, `rivasplata`, `blanchard` *and* `catoni` *based on the disintegrated bounds, and* `stochastic` *based on the randomized bounds learned with two learning rates* $\eta \in \{10^{-4}, 10^{-6}\}$ *and different variances* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. *We report the test risk* ($R_{\mathcal{T}}(h)$)*, the bound value (Bnd), the empirical risk* ($R_{\mathcal{S}}(h)$)*, and the divergence (Div) associated with each bound (the* RÉNYI *divergence for* `ours`*, the KL divergence for* `stochastic`*, and the disintegrated KL divergence for* `rivasplata`, `blanchard` *and* `catoni`*). More precisely, we report the mean* $\pm$ *the standard deviation for* 400 *neural networks sampled from* $\rho_{\mathcal{S}}$ *for* `ours`, `rivasplata`, `blanchard`, *and* `catoni`*. We consider, in this table, that the split ratio is* 0.2.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.016 ± .000	.023 ± .000	.019 ± .000	.336	.015 ± .000	.023 ± .000	.019 ± .000	.748	.014 ± .001	.020 ± .001	.016 ± .000	2.096	.019 ± .002	.024 ± .002	.020 ± .002	2.244
	blanchard	.016 ± .000	.034 ± .001	.019 ± .000	97.590 ± 14.260	.015 ± .000	.026 ± .001	.019 ± .000	21.153 ± 6.514	.015 ± .001	.020 ± .001	.016 ± .001	3.362 ± 2.569	.020 ± .002	.024 ± .002	.021 ± .002	.371 ± .875
	catoni	.016 ± .000	.034 ± .001	.019 ± .000	116.744 ± 15.447	.015 ± .000	.027 ± .002	.019 ± .000	24.135 ± 7.075	.015 ± .001	.020 ± .001	.016 ± .001	3.352 ± 2.667	.020 ± .002	.024 ± .002	.021 ± .002	.410 ± .890
	rivasplata	.016 ± .000	.030 ± .001	.019 ± .000	101.334 ± 14.728	.015 ± .000	.024 ± .001	.019 ± .000	21.663 ± 6.603	.015 ± .001	.020 ± .001	.016 ± .001	3.409 ± 2.666	.020 ± .002	.024 ± .002	.021 ± .002	.446 ± .927
	stochastic	—	.052	—	.168	—	.051	—	.374	—	.047	—	1.048	—	.053	—	1.122
Fashion	ours	.165 ± .002	.169 ± .001	.157 ± .001	4.811	.148 ± .003	.155 ± .002	.143 ± .002	1.856	.145 ± .005	.153 ± .006	.139 ± .005	15.453	.160 ± .005	.166 ± .005	.155 ± .005	1.633
	blanchard	.163 ± .002	.190 ± .003	.155 ± .001	96.264 ± 14.472	.152 ± .003	.163 ± .003	.147 ± .003	21.099 ± 6.507	.155 ± .007	.160 ± .007	.151 ± .007	3.929 ± 2.841	.163 ± .006	.165 ± .006	.158 ± .006	.340 ± .885
	catoni	.163 ± .002	.190 ± .004	.156 ± .001	121.542 ± 16.499	.150 ± .002	.158 ± .003	.144 ± .002	27.241 ± 7.318	.151 ± .006	.155 ± .006	.146 ± .006	5.120 ± 3.150	.162 ± .005	.165 ± .005	.157 ± .005	.444 ± .968
	rivasplata	.161 ± .001	.180 ± .002	.153 ± .001	106.403 ± 14.044	.150 ± .002	.158 ± .003	.145 ± .003	23.134 ± 7.064	.153 ± .006	.157 ± .006	.148 ± .007	4.439 ± 2.924	.162 ± .006	.165 ± .005	.157 ± .005	.417 ± .928
	stochastic	—	.226	—	2.405	—	.210	—	5.428	—	.207	—	7.727	—	.223	—	.816
CIFAR-10	ours	.390 ± .000	.407 ± .000	.391 ± .000	.040	.404 ± .000	.414 ± .000	.398 ± .000	.070	.396 ± .001	.411 ± .000	.395 ± .000	.155	.416 ± .002	.432 ± .001	.415 ± .001	.970
	blanchard	.390 ± .000	.473 ± .004	.391 ± .000	271.616 ± 23.555	.404 ± .000	.445 ± .003	.398 ± .000	84.868 ± 13.050	.396 ± .001	.422 ± .003	.395 ± .000	23.962 ± 7.208	.416 ± .002	.432 ± .002	.416 ± .001	4.496 ± 3.018
	catoni	.390 ± .000	.473 ± .006	.391 ± .000	27.502 ± 23.371	.404 ± .000	.434 ± .003	.398 ± .000	84.848 ± 12.992	.396 ± .001	.415 ± .002	.395 ± .000	24.505 ± 6.942	.416 ± .002	.431 ± .001	.415 ± .001	4.859 ± 3.176
	rivasplata	.390 ± .000	.450 ± .002	.391 ± .000	271.700 ± 23.586	.403 ± .000	.433 ± .002	.398 ± .000	85.027 ± 13.047	.396 ± .001	.416 ± .002	.395 ± .000	23.955 ± 7.093	.416 ± .002	.431 ± .002	.416 ± .001	4.610 ± 3.084
	stochastic	—	.477	—	.020	—	.485	—	.035	—	.482	—	.077	—	.503	—	.485
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.016 ± .000	.025 ± .000	.019 ± .000	14.490	.015 ± .000	.024 ± .000	.019 ± .000	8.583	.014 ± .000	.021 ± .001	.016 ± .000	13.055	.016 ± .001	.023 ± .001	.017 ± .001	25.556
	blanchard	.016 ± .000	.430 ± .004	.018 ± .000	11405.062 ± 153.554	.014 ± .000	.200 ± .003	.018 ± .000	3799.912 ± 89.585	.013 ± .000	.086 ± .002	.014 ± .000	1187.859 ± 48.700	.015 ± .001	.049 ± .002	.016 ± .001	38.983 ± 27.857
	catoni	.016 ± .000	.355 ± .002	.019 ± .000	11954.106 ± 15.709	.015 ± .000	.149 ± .003	.019 ± .000	3828.342 ± 83.937	.014 ± .001	.064 ± .002	.016 ± .001	1218.708 ± 48.514	.017 ± .001	.041 ± .002	.018 ± .001	389.726 ± 29.076
	rivasplata	.015 ± .000	.272 ± .002	.018 ± .000	1173.953 ± 149.364	.013 ± .000	.122 ± .002	.017 ± .000	3691.345 ± 82.512	.012 ± .000	.056 ± .001	.013 ± .000	1206.615 ± 5.381	.015 ± .001	.037 ± .001	.015 ± .001	391.881 ± 28.344
	stochastic	—	.053	—	7.245	—	.052	—	4.292	—	.048	—	6.528	—	.051	—	12.778
Fashion	ours	.165 ± .002	.172 ± .001	.157 ± .001	23.705	.141 ± .002	.156 ± .002	.137 ± .002	52.736	.131 ± .003	.147 ± .003	.126 ± .003	7.515	.156 ± .004	.165 ± .004	.151 ± .003	16.954
	blanchard	.136 ± .001	.598 ± .003	.130 ± .001	11334.327 ± 145.083	.125 ± .001	.379 ± .003	.121 ± .001	3998.068 ± 88.992	.124 ± .001	.247 ± .003	.117 ± .001	126.184 ± 48.814	.152 ± .003	.216 ± .004	.147 ± .003	364.531 ± 28.029
	catoni	.162 ± .001	.525 ± .004	.154 ± .001	11965.668 ± 152.681	.141 ± .002	.309 ± .003	.137 ± .002	384.802 ± 84.123	.132 ± .003	.224 ± .004	.127 ± .002	1239.918 ± 49.594	.155 ± .004	.232 ± .005	.150 ± .004	394.607 ± 28.146
	rivasplata	.131 ± .001	.455 ± .002	.127 ± .001	1193.209 ± 155.390	.123 ± .001	.290 ± .002	.119 ± .001	4005.169 ± 89.793	.123 ± .001	.204 ± .002	.116 ± .001	1294.726 ± 49.874	.152 ± .004	.195 ± .004	.146 ± .003	378.905 ± 27.422
	stochastic	—	.228	—	11.853	—	.209	—	26.368	—	.198	—	35.258	—	.221	—	8.477
CIFAR-10	ours	.390 ± .000	.411 ± .000	.391 ± .000	13.286	.404 ± .000	.415 ± .000	.398 ± .000	3.305	.396 ± .001	.412 ± .000	.395 ± .000	3.136	.415 ± .001	.433 ± .001	.415 ± .001	6.064
	blanchard	.389 ± .000	.990 ± .000	.391 ± .000	75424.764 ± 397.521	.403 ± .000	.820 ± .002	.397 ± .000	8815.324 ± 126.764	.395 ± .001	.651 ± .003	.394 ± .000	2738.066 ± 75.053	.408 ± .001	.557 ± .003	.405 ± .001	918.500 ± 42.347
	catoni	.390 ± .000	.990 ± .000	.391 ± .000	26434.787 ± 228.500	.403 ± .000	.726 ± .003	.397 ± .000	8651.380 ± 126.473	.394 ± .001	.620 ± .002	.393 ± .000	4178.302 ± 9.315	.401 ± .001	.556 ± .001	.396 ± .001	1462.235 ± 55.526
	rivasplata	.389 ± .000	.902 ± .001	.391 ± .000	31497.669 ± 249.683	.403 ± .000	.715 ± .002	.397 ± .000	8707.893 ± 133.239	.394 ± .001	.578 ± .003	.393 ± .000	2741.257 ± 74.942	.405 ± .001	.512 ± .002	.402 ± .001	967.818 ± 43.629
	stochastic	—	.480	—	6.643	—	.486	—	1.653	—	.483	—	1.568	—	.503	—	3.032

Table F.4. *Comparison of* ours, rivasplata, blanchard *and* catoni *based on the disintegrated bounds, and* stochastic *based on the randomized bounds learned with two learning rates* $\eta \in \{10^{-4}, 10^{-6}\}$ *and different variances* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. *We report the test risk (*$R_{\mathcal{T}}(h)$*), the bound value (Bnd), the empirical risk (*$R_{\mathcal{S}}(h)$*), and the divergence (Div) associated with each bound (the* RÉNYI *divergence for* ours, *the KL divergence for* stochastic, *and the disintegrated KL divergence for* rivasplata, blanchard *and* catoni*). More precisely, we report the mean* $\pm$ *the standard deviation for* 400 *neural networks sampled from* $\rho_{\mathcal{S}}$ *for* ours, rivasplata, blanchard, *and* catoni. *We consider, in this table, that the split ratio is* 0.3.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.012 ± .000	.017 ± .000	.013 ± .000	.181	.009 ± .000	.015 ± .000	.011 ± .000	.155	.012 ± .000	.020 ± .000	.016 ± .000	1.655	.013 ± .001	.019 ± .001	.015 ± .001	.615
	blanchard	.012 ± .000	.027 ± .001	.013 ± .000	93.915 ± 14.109	.012 ± .000	.021 ± .001	.014 ± .000	19.292 ± 6.037	.012 ± .000	.020 ± .001	.016 ± .000	3.023 ± 2.430	.014 ± .001	.018 ± .001	.015 ± .001	.368 ± .831
	catoni	.012 ± .000	.025 ± .001	.013 ± .000	113.574 ± 15.436	.012 ± .000	.023 ± .002	.014 ± .000	22.347 ± 6.877	.012 ± .000	.020 ± .001	.016 ± .000	2.918 ± 2.341	.013 ± .001	.018 ± .001	.015 ± .001	.336 ± .807
	rivasplata	.012 ± .000	.023 ± .001	.013 ± .000	96.392 ± 14.300	.012 ± .000	.020 ± .001	.014 ± .000	19.905 ± 6.254	.012 ± .000	.020 ± .001	.016 ± .000	2.931 ± 2.446	.013 ± .001	.018 ± .001	.015 ± .001	.355 ± .813
	stochastic	—	.042	—	.091	—	.039	—	.077	—	.047	—	.827	—	.045	—	.308
Fashion	ours	.126 ± .000	.134 ± .000	.124 ± .000	.328	.126 ± .001	.130 ± .001	.119 ± .001	1.692	.122 ± .002	.126 ± .002	.115 ± .002	4.617	.139 ± .005	.145 ± .005	.133 ± .005	2.425
	blanchard	.126 ± .000	.157 ± .003	.124 ± .000	88.034 ± 13.485	.126 ± .001	.136 ± .002	.120 ± .001	18.852 ± 6.115	.124 ± .002	.127 ± .002	.118 ± .002	3.014 ± 2.395	.142 ± .006	.144 ± .006	.137 ± .006	.370 ± .819
	catoni	.126 ± .000	.159 ± .004	.124 ± .000	114.259 ± 15.300	.126 ± .001	.133 ± .002	.120 ± .001	22.607 ± 6.871	.124 ± .002	.126 ± .002	.118 ± .002	3.100 ± 2.513	.141 ± .006	.144 ± .006	.136 ± .006	.390 ± .898
	rivasplata	.126 ± .000	.148 ± .002	.124 ± .000	93.107 ± 13.630	.126 ± .001	.133 ± .002	.120 ± .001	19.724 ± 6.320	.124 ± .002	.126 ± .002	.118 ± .002	2.980 ± 2.451	.142 ± .006	.144 ± .006	.136 ± .006	.371 ± .869
	stochastic	—	.187	—	.164	—	.182	—	.846	—	.178	—	2.309	—	.199	—	1.212
CIFAR-10	ours	.369 ± .000	.375 ± .000	.358 ± .000	.028	.351 ± .000	.368 ± .000	.352 ± .000	.041	.359 ± .001	.377 ± .000	.360 ± .000	.183	.419 ± .001	.433 ± .001	.416 ± .001	.759
	blanchard	.369 ± .000	.446 ± .004	.358 ± .000	269.789 ± 22.724	.351 ± .000	.401 ± .004	.352 ± .000	84.113 ± 12.530	.359 ± .001	.388 ± .003	.360 ± .000	22.878 ± 6.728	.419 ± .001	.432 ± .003	.416 ± .001	4.089 ± 2.818
	catoni	.369 ± .000	.450 ± .007	.358 ± .000	269.843 ± 24.225	.351 ± .000	.390 ± .004	.352 ± .000	84.500 ± 12.608	.359 ± .001	.381 ± .002	.360 ± .000	23.567 ± 7.181	.419 ± .001	.432 ± .001	.416 ± .001	4.285 ± 2.942
	rivasplata	.369 ± .000	.421 ± .003	.358 ± .000	27.224 ± 24.187	.351 ± .000	.388 ± .002	.352 ± .000	84.250 ± 13.274	.359 ± .001	.382 ± .002	.360 ± .000	23.053 ± 6.724	.419 ± .001	.431 ± .002	.416 ± .001	4.141 ± 2.985
	stochastic	—	.445	—	.014	—	.438	—	.020	—	.447	—	.092	—	.504	—	.380
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.012 ± .000	.019 ± .000	.013 ± .000	24.837	.012 ± .000	.020 ± .000	.014 ± .000	12.358	.012 ± .000	.021 ± .000	.015 ± .000	13.908	.013 ± .001	.019 ± .001	.014 ± .001	16.179
	blanchard	.012 ± .000	.467 ± .004	.013 ± .000	11819.223 ± 154.992	.011 ± .000	.211 ± .003	.014 ± .000	3808.981 ± 86.014	.010 ± .000	.094 ± .003	.014 ± .000	121.397 ± 51.944	.012 ± .001	.046 ± .002	.013 ± .001	372.832 ± 26.602
	catoni	.012 ± .000	.339 ± .002	.013 ± .000	1196.394 ± 15.704	.012 ± .000	.159 ± .003	.014 ± .000	3838.459 ± 88.155	.012 ± .000	.070 ± .002	.016 ± .000	1218.505 ± 51.783	.013 ± .001	.037 ± .001	.014 ± .001	386.824 ± 28.233
	rivasplata	.012 ± .000	.289 ± .003	.013 ± .000	1191.037 ± 152.759	.011 ± .000	.128 ± .002	.014 ± .000	3768.785 ± 9.947	.010 ± .000	.061 ± .001	.013 ± .000	1231.638 ± 49.362	.011 ± .001	.033 ± .001	.012 ± .000	382.225 ± 28.481
	stochastic	—	.044	—	12.418	—	.046	—	6.179	—	.047	—	6.954	—	.045	—	8.089
Fashion	ours	.126 ± .000	.137 ± .000	.124 ± .000	12.401	.125 ± .001	.132 ± .001	.119 ± .001	14.631	.120 ± .002	.128 ± .002	.113 ± .001	26.499	.133 ± .003	.143 ± .003	.127 ± .003	23.702
	blanchard	.123 ± .000	.602 ± .003	.121 ± .000	10558.872 ± 139.107	.119 ± .001	.383 ± .004	.112 ± .001	3893.091 ± 86.176	.113 ± .001	.239 ± .003	.106 ± .001	1204.211 ± 5.815	.132 ± .003	.195 ± .004	.125 ± .003	362.146 ± 27.801
	catoni	.126 ± .000	.531 ± .004	.124 ± .000	11966.223 ± 148.195	.125 ± .001	.299 ± .003	.118 ± .001	3829.806 ± 85.864	.119 ± .002	.209 ± .002	.113 ± .001	1225.310 ± 48.090	.134 ± .004	.202 ± .005	.127 ± .003	395.243 ± 29.182
	rivasplata	.123 ± .000	.458 ± .003	.120 ± .000	11209.156 ± 143.319	.118 ± .001	.287 ± .002	.111 ± .001	3815.804 ± 85.091	.112 ± .001	.196 ± .002	.105 ± .001	126.956 ± 49.255	.130 ± .003	.173 ± .004	.124 ± .003	376.904 ± 27.549
	stochastic	—	.189	—	6.200	—	.184	—	7.316	—	.179	—	13.250	—	.195	—	11.851
CIFAR-10	ours	.369 ± .000	.379 ± .000	.358 ± .000	11.657	.351 ± .000	.369 ± .000	.352 ± .000	2.267	.359 ± .001	.378 ± .000	.360 ± .000	2.616	.418 ± .001	.434 ± .001	.415 ± .001	5.675
	blanchard	.369 ± .000	.990 ± .000	.358 ± .000	40152.974 ± 291.721	.351 ± .000	.809 ± .003	.351 ± .000	8753.816 ± 136.801	.358 ± .001	.635 ± .004	.359 ± .000	2728.436 ± 73.835	.412 ± .001	.568 ± .004	.407 ± .001	91.026 ± 44.096
	catoni	.369 ± .000	.986 ± .000	.358 ± .000	24477.984 ± 223.367	.351 ± .000	.708 ± .003	.351 ± .000	8463.452 ± 135.001	.357 ± .001	.578 ± .002	.357 ± .000	3401.221 ± 84.878	.405 ± .001	.561 ± .002	.399 ± .001	1354.100 ± 51.315
	rivasplata	.369 ± .000	.868 ± .001	.358 ± .000	24424.968 ± 223.601	.351 ± .000	.694 ± .002	.351 ± .000	8665.339 ± 136.361	.358 ± .001	.555 ± .003	.358 ± .000	274.651 ± 74.784	.409 ± .001	.521 ± .003	.403 ± .001	955.211 ± 44.609
	stochastic	—	.448	—	5.829	—	.439	—	1.134	—	.448	—	1.308	—	.504	—	2.838

Table F.5. *Comparison of* `ours`, `rivasplata`, `blanchard` *and* `catoni` *based on the disintegrated bounds, and* `stochastic` *based on the randomized bounds learned with two learning rates* $\eta \in \{10^{-4}, 10^{-6}\}$ *and different variances* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. *We report the test risk* ($R_{\mathcal{T}}(h)$)*, the bound value (Bnd), the empirical risk* ($R_{\mathcal{S}}(h)$)*, and the divergence (Div) associated with each bound (the* RÉNYI *divergence for* `ours`*, the KL divergence for* `stochastic`*, and the disintegrated KL divergence for* `rivasplata`, `blanchard` *and* `catoni`*). More precisely, we report the mean* $\pm$ *the standard deviation for* 400 *neural networks sampled from* $\rho_{\mathbb{S}}$ *for* `ours`, `rivasplata`, `blanchard`, *and* `catoni`*. We consider, in this table, that the split ratio is* 0.4.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.010 ± .000	.017 ± .000	.013 ± .000	.194	.012 ± .000	.018 ± .000	.014 ± .000	.138	.009 ± .000	.015 ± .000	.011 ± .000	.235	.014 ± .001	.020 ± .001	.015 ± .001	1.111
	blanchard	.010 ± .000	.028 ± .001	.013 ± .000	88.323 ± 13.740	.012 ± .000	.021 ± .001	.014 ± .000	16.792 ± 5.702	.009 ± .000	.014 ± .001	.011 ± .000	2.449 ± 2.313	.014 ± .001	.019 ± .001	.016 ± .001	.244 ± .765
	catoni	.010 ± .000	.026 ± .001	.013 ± .000	109.202 ± 15.634	.012 ± .000	.023 ± .002	.014 ± .000	19.918 ± 6.526	.009 ± .000	.015 ± .001	.011 ± .000	2.486 ± 2.362	.014 ± .001	.019 ± .001	.016 ± .001	.298 ± .762
	rivasplata	.010 ± .000	.024 ± .001	.013 ± .000	91.872 ± 14.470	.012 ± .000	.019 ± .001	.014 ± .000	17.002 ± 5.882	.009 ± .000	.014 ± .000	.011 ± .000	2.529 ± 2.251	.014 ± .001	.019 ± .001	.016 ± .001	.308 ± .778
	stochastic	—	.043	—	.097	—	.044	—	.069	—	.039	—	.117	—	.047	—	.555
Fashion	ours	.118 ± .001	.123 ± .000	.112 ± .000	.269	.113 ± .001	.118 ± .001	.107 ± .001	.743	.117 ± .002	.121 ± .002	.110 ± .002	2.600	.131 ± .004	.138 ± .004	.126 ± .004	1.229
	blanchard	.118 ± .001	.145 ± .003	.112 ± .000	82.403 ± 13.230	.113 ± .001	.123 ± .002	.107 ± .001	16.836 ± 5.583	.119 ± .002	.121 ± .003	.112 ± .003	2.641 ± 2.369	.133 ± .004	.136 ± .004	.128 ± .004	.297 ± .731
	catoni	.118 ± .001	.151 ± .004	.112 ± .000	109.988 ± 15.347	.113 ± .001	.120 ± .002	.107 ± .001	19.889 ± 6.689	.118 ± .002	.120 ± .003	.112 ± .003	2.615 ± 2.234	.132 ± .004	.136 ± .004	.128 ± .004	.300 ± .811
	rivasplata	.118 ± .001	.137 ± .002	.112 ± .000	87.804 ± 13.640	.113 ± .001	.120 ± .002	.107 ± .001	17.491 ± 6.144	.118 ± .002	.121 ± .003	.112 ± .003	2.549 ± 2.175	.133 ± .005	.137 ± .004	.128 ± .004	.322 ± .794
	stochastic	—	.174	—	.135	—	.168	—	.372	—	.172	—	1.300	—	.191	—	.615
CIFAR-10	ours	.334 ± .000	.346 ± .000	.328 ± .000	.025	.322 ± .000	.331 ± .000	.313 ± .000	.050	.323 ± .001	.334 ± .000	.316 ± .000	.160	.333 ± .001	.341 ± .001	.323 ± .001	.461
	blanchard	.334 ± .000	.421 ± .004	.328 ± .000	269.875 ± 23.982	.322 ± .000	.364 ± .004	.313 ± .000	83.082 ± 13.029	.323 ± .001	.345 ± .004	.316 ± .000	21.614 ± 6.670	.333 ± .001	.340 ± .002	.323 ± .001	3.630 ± 2.750
	catoni	.334 ± .000	.433 ± .008	.328 ± .000	27.270 ± 24.201	.322 ± .000	.355 ± .005	.313 ± .000	84.148 ± 13.578	.323 ± .001	.338 ± .002	.316 ± .000	22.547 ± 6.801	.333 ± .001	.338 ± .001	.323 ± .001	3.831 ± 2.801
	rivasplata	.334 ± .000	.394 ± .003	.328 ± .000	27.133 ± 24.109	.322 ± .000	.351 ± .003	.313 ± .000	83.438 ± 13.033	.323 ± .001	.339 ± .002	.316 ± .000	21.688 ± 6.718	.333 ± .001	.339 ± .002	.323 ± .001	3.667 ± 2.757
	stochastic	—	.414	—	.013	—	.399	—	.025	—	.403	—	.080	—	.409	—	.230
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.010 ± .000	.019 ± .000	.013 ± .000	23.992	.012 ± .000	.019 ± .000	.014 ± .000	7.767	.009 ± .000	.015 ± .000	.011 ± .000	3.165	.012 ± .001	.019 ± .001	.013 ± .001	18.413
	blanchard	.010 ± .000	.500 ± .004	.013 ± .000	1123.328 ± 151.115	.012 ± .000	.236 ± .004	.014 ± .000	381.840 ± 94.218	.009 ± .000	.096 ± .003	.011 ± .000	1184.214 ± 47.208	.011 ± .001	.048 ± .002	.012 ± .001	363.194 ± 26.547
	catoni	.010 ± .000	.369 ± .003	.013 ± .000	1191.598 ± 154.180	.012 ± .000	.180 ± .003	.014 ± .000	3826.581 ± 85.362	.009 ± .000	.070 ± .002	.011 ± .000	1217.723 ± 49.984	.012 ± .001	.039 ± .002	.014 ± .001	384.476 ± 29.126
	rivasplata	.010 ± .000	.316 ± .003	.013 ± .000	11557.703 ± 151.498	.012 ± .000	.142 ± .002	.014 ± .000	3751.391 ± 84.542	.009 ± .000	.061 ± .002	.011 ± .000	1172.156 ± 46.933	.010 ± .001	.035 ± .001	.012 ± .001	373.003 ± 27.844
	stochastic	—	.045	—	11.996	—	.045	—	3.884	—	.040	—	1.583	—	.045	—	9.207
Fashion	ours	.118 ± .000	.127 ± .000	.112 ± .000	17.987	.113 ± .001	.119 ± .001	.107 ± .001	6.361	.114 ± .002	.123 ± .002	.107 ± .002	22.582	.125 ± .003	.137 ± .003	.122 ± .003	16.872
	blanchard	.115 ± .001	.659 ± .004	.110 ± .000	11835.780 ± 161.816	.110 ± .001	.395 ± .004	.104 ± .000	3828.562 ± 94.279	.108 ± .001	.244 ± .004	.102 ± .001	1185.882 ± 5.575	.123 ± .003	.192 ± .004	.119 ± .002	346.265 ± 27.827
	catoni	.118 ± .001	.566 ± .004	.112 ± .000	11921.114 ± 153.739	.113 ± .001	.304 ± .003	.107 ± .000	3822.647 ± 85.225	.114 ± .002	.208 ± .003	.107 ± .002	1217.879 ± 52.353	.125 ± .003	.196 ± .004	.121 ± .002	388.473 ± 29.475
	rivasplata	.114 ± .000	.476 ± .003	.109 ± .000	11206.239 ± 149.549	.110 ± .001	.292 ± .003	.103 ± .000	3745.930 ± 84.367	.106 ± .001	.197 ± .003	.101 ± .001	1229.005 ± 51.052	.122 ± .003	.170 ± .004	.118 ± .003	361.652 ± 28.452
	stochastic	—	.177	—	8.994	—	.169	—	3.180	—	.172	—	11.291	—	.189	—	8.436
CIFAR-10	ours	.334 ± .000	.350 ± .000	.328 ± .000	12.067	.322 ± .000	.332 ± .000	.313 ± .000	4.172	.323 ± .001	.336 ± .000	.316 ± .000	3.382	.332 ± .001	.343 ± .001	.322 ± .001	6.855
	blanchard	.334 ± .000	.977 ± .001	.328 ± .000	28565.558 ± 245.568	.322 ± .000	.803 ± .003	.313 ± .000	8479.553 ± 126.804	.321 ± .001	.614 ± .004	.315 ± .000	2727.786 ± 7.572	.327 ± .001	.487 ± .004	.317 ± .001	887.578 ± 42.449
	catoni	.334 ± .000	.983 ± .000	.328 ± .000	24136.528 ± 211.963	.322 ± .000	.694 ± .004	.313 ± .000	7928.671 ± 122.159	.320 ± .001	.515 ± .002	.314 ± .000	237.703 ± 65.952	.323 ± .001	.468 ± .002	.312 ± .001	1157.073 ± 47.283
	rivasplata	.334 ± .000	.922 ± .001	.328 ± .000	33282.032 ± 246.654	.322 ± .000	.680 ± .003	.312 ± .000	8493.458 ± 128.894	.320 ± .001	.527 ± .003	.314 ± .000	2739.108 ± 7.556	.325 ± .001	.436 ± .003	.314 ± .001	91.066 ± 43.389
	stochastic	—	.417	—	6.033	—	.400	—	2.086	—	.403	—	1.691	—	.410	—	3.427

Table F.6. *Comparison of* `ours`, `rivasplata`, `blanchard` *and* `catoni` *based on the disintegrated bounds, and* `stochastic` *based on the randomized bounds learned with two learning rates* $\eta \in \{10^{-4}, 10^{-6}\}$ *and different variances* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$*. We report the test risk (*$R_{\mathcal{T}}(h)$*), the bound value (Bnd), the empirical risk (*$R_{\mathcal{S}}(h)$*), and the divergence (Div) associated with each bound (the* RÉNYI *divergence for* `ours`*, the KL divergence for* `stochastic`*, and the disintegrated KL divergence for* `rivasplata`, `blanchard` *and* `catoni`*). More precisely, we report the mean* $\pm$ *the standard deviation for* 400 *neural networks sampled from* $\rho_{\mathbb{S}}$ *for* `ours`, `rivasplata`, `blanchard`, *and* `catoni`*. We consider, in this table, that the split ratio is* 0.5*.*

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.008 ± .000	.015 ± .000	.010 ± .000	.084	.006 ± .000	.012 ± .000	.009 ± .000	.053	.008 ± .000	.014 ± .000	.010 ± .000	.179	.014 ± .001	.019 ± .001	.014 ± .001	.576
	blanchard	.008 ± .000	.025 ± .001	.010 ± .000	81.167 ± 12.801	.006 ± .000	.014 ± .001	.009 ± .000	15.518 ± 5.438	.009 ± .000	.014 ± .001	.010 ± .000	2.140 ± 2.072	.015 ± .001	.018 ± .001	.015 ± .001	.284 ± .649
	catoni	.008 ± .000	.022 ± .001	.010 ± .000	104.063 ± 14.662	.006 ± .000	.015 ± .000	.009 ± .000	17.676 ± 5.963	.008 ± .000	.014 ± .001	.010 ± .000	2.152 ± 2.085	.015 ± .001	.018 ± .001	.015 ± .001	.252 ± .680
	rivasplata	.008 ± .000	.021 ± .001	.010 ± .000	84.581 ± 13.035	.006 ± .000	.013 ± .001	.009 ± .000	15.545 ± 5.594	.008 ± .000	.014 ± .000	.010 ± .000	2.185 ± 1.992	.015 ± .001	.018 ± .001	.015 ± .001	.276 ± .693
	stochastic	—	.039	—	.042	—	.035	—	.026	—	.038	—	.090	—	.045	—	.288
Fashion	ours	.106 ± .000	.113 ± .000	.101 ± .000	.133	.104 ± .001	.110 ± .000	.099 ± .000	.327	.108 ± .002	.112 ± .001	.101 ± .001	.903	.120 ± .004	.127 ± .003	.115 ± .003	.868
	blanchard	.106 ± .000	.136 ± .003	.101 ± .000	77.573 ± 12.564	.104 ± .001	.115 ± .003	.099 ± .000	15.278 ± 5.599	.109 ± .002	.111 ± .002	.102 ± .001	2.153 ± 2.081	.122 ± .004	.126 ± .004	.117 ± .004	.248 ± .715
	catoni	.106 ± .000	.145 ± .005	.101 ± .000	104.356 ± 14.712	.104 ± .001	.112 ± .002	.099 ± .000	17.566 ± 5.996	.109 ± .002	.110 ± .001	.102 ± .001	2.217 ± 2.084	.122 ± .004	.125 ± .004	.117 ± .004	.262 ± .699
	rivasplata	.106 ± .000	.127 ± .002	.101 ± .000	82.150 ± 12.955	.104 ± .001	.112 ± .001	.099 ± .000	15.509 ± 5.629	.109 ± .002	.111 ± .001	.102 ± .001	2.178 ± 2.060	.122 ± .004	.126 ± .004	.117 ± .004	.264 ± .704
	stochastic	—	.162	—	.066	—	.159	—	.164	—	.162	—	.451	—	.179	—	.434
CIFAR-10	ours	.312 ± .000	.323 ± .000	.304 ± .000	.027	.281 ± .000	.304 ± .000	.285 ± .000	.035	.298 ± .001	.310 ± .000	.291 ± .000	.101	.315 ± .001	.329 ± .001	.309 ± .001	.368
	blanchard	.312 ± .000	.405 ± .004	.304 ± .000	268.149 ± 22.835	.281 ± .000	.339 ± .004	.285 ± .000	8.690 ± 12.628	.298 ± .001	.320 ± .004	.291 ± .000	19.648 ± 6.249	.315 ± .001	.327 ± .003	.310 ± .001	3.213 ± 2.590
	catoni	.312 ± .000	.428 ± .009	.304 ± .000	269.415 ± 22.884	.281 ± .000	.333 ± .005	.285 ± .000	83.414 ± 13.018	.298 ± .001	.314 ± .003	.291 ± .000	2.711 ± 6.481	.315 ± .001	.326 ± .001	.310 ± .001	3.273 ± 2.597
	rivasplata	.312 ± .000	.375 ± .003	.304 ± .000	268.589 ± 22.845	.281 ± .000	.325 ± .003	.285 ± .000	81.532 ± 12.712	.298 ± .001	.315 ± .002	.291 ± .000	19.813 ± 6.288	.315 ± .001	.327 ± .002	.310 ± .001	3.233 ± 2.599
	stochastic	—	.391	—	.013	—	.370	—	.017	—	.377	—	.050	—	.397	—	.184
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.008 ± .000	.017 ± .000	.010 ± .000	29.993	.006 ± .000	.013 ± .000	.009 ± .000	3.162	.008 ± .000	.015 ± .000	.010 ± .000	1.418	.013 ± .001	.019 ± .001	.013 ± .001	12.231
	blanchard	.008 ± .000	.574 ± .005	.010 ± .000	11894.556 ± 155.958	.006 ± .000	.256 ± .004	.009 ± .000	3826.515 ± 86.973	.008 ± .000	.108 ± .003	.010 ± .000	1184.777 ± 48.158	.010 ± .001	.052 ± .002	.011 ± .000	36.865 ± 28.054
	catoni	.008 ± .000	.396 ± .003	.010 ± .000	11986.455 ± 15.722	.006 ± .000	.192 ± .002	.009 ± .000	3824.971 ± 85.072	.008 ± .000	.079 ± .002	.010 ± .000	1213.611 ± 48.751	.013 ± .001	.042 ± .002	.014 ± .001	384.275 ± 28.556
	rivasplata	.008 ± .000	.362 ± .003	.010 ± .000	11905.971 ± 15.609	.006 ± .000	.148 ± .003	.009 ± .000	377.259 ± 84.127	.008 ± .000	.067 ± .002	.010 ± .000	118.841 ± 5.043	.010 ± .001	.036 ± .001	.011 ± .000	369.675 ± 27.947
	stochastic	—	.041	—	14.996	—	.035	—	1.581	—	.039	—	.709	—	.045	—	6.116
Fashion	ours	.106 ± .000	.114 ± .000	.101 ± .000	6.310	.103 ± .001	.113 ± .000	.099 ± .000	9.312	.106 ± .002	.115 ± .001	.100 ± .001	14.924	.115 ± .003	.126 ± .003	.110 ± .002	18.364
	blanchard	.105 ± .000	.674 ± .004	.101 ± .000	10795.464 ± 143.426	.102 ± .000	.412 ± .004	.098 ± .000	3685.940 ± 82.481	.103 ± .001	.253 ± .004	.097 ± .001	1178.401 ± 48.359	.113 ± .002	.186 ± .004	.108 ± .002	338.697 ± 27.104
	catoni	.106 ± .000	.623 ± .005	.101 ± .000	11971.564 ± 15.589	.104 ± .001	.321 ± .004	.099 ± .000	3825.370 ± 87.728	.107 ± .002	.208 ± .003	.100 ± .001	1214.976 ± 48.846	.116 ± .003	.184 ± .004	.111 ± .003	388.197 ± 27.580
	rivasplata	.105 ± .000	.503 ± .003	.100 ± .000	11139.304 ± 15.540	.102 ± .000	.307 ± .003	.097 ± .000	381.075 ± 87.924	.102 ± .001	.201 ± .003	.096 ± .001	1201.832 ± 48.877	.112 ± .002	.161 ± .003	.107 ± .002	349.146 ± 27.482
	stochastic	—	.163	—	3.155	—	.161	—	4.656	—	.163	—	7.462	—	.176	—	9.182
CIFAR-10	ours	.312 ± .000	.328 ± .000	.304 ± .000	12.006	.281 ± .000	.304 ± .000	.285 ± .000	1.802	.297 ± .001	.311 ± .000	.291 ± .000	2.056	.314 ± .001	.330 ± .001	.309 ± .001	4.782
	blanchard	.312 ± .000	.990 ± .000	.304 ± .000	48007.471 ± 31.730	.280 ± .000	.825 ± .003	.284 ± .000	8824.774 ± 134.331	.296 ± .001	.617 ± .004	.290 ± .000	2723.775 ± 66.832	.309 ± .001	.490 ± .004	.303 ± .001	888.277 ± 41.530
	catoni	.312 ± .000	.980 ± .000	.304 ± .000	21278.808 ± 207.839	.280 ± .000	.681 ± .004	.284 ± .000	6951.932 ± 118.540	.296 ± .001	.496 ± .003	.290 ± .000	2145.470 ± 6.045	.305 ± .001	.457 ± .002	.299 ± .001	103.494 ± 47.021
	rivasplata	.312 ± .000	.964 ± .001	.304 ± .000	42834.626 ± 284.116	.280 ± .000	.690 ± .003	.284 ± .000	8675.531 ± 136.658	.296 ± .001	.521 ± .003	.290 ± .000	2718.415 ± 66.664	.307 ± .001	.434 ± .003	.301 ± .001	921.068 ± 42.158
	stochastic	—	.394	—	6.003	—	.371	—	.901	—	.378	—	1.028	—	.397	—	2.391

Table F.7. *Comparison of* `ours`, `rivasplata`, `blanchard` *and* `catoni` *based on the disintegrated bounds, and* `stochastic` *based on the randomized bounds learned with two learning rates* $\eta \in \{10^{-4}, 10^{-6}\}$ *and different variances* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. *We report the test risk (*$R_{\mathcal{T}}(h)$*), the bound value (Bnd), the empirical risk (*$R_{\mathcal{S}}(h)$*), and the divergence (Div) associated with each bound (the* RÉNYI *divergence for* `ours`, *the KL divergence for* `stochastic`, *and the disintegrated KL divergence for* `rivasplata`, `blanchard` *and* `catoni`*). More precisely, we report the mean* $\pm$ *the standard deviation for* 400 *neural networks sampled from* $\rho_{\mathcal{S}}$ *for* `ours`, `rivasplata`, `blanchard`, *and* `catoni`. *We consider, in this table, that the split ratio is* 0.6.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.008 ± .000	.014 ± .000	.010 ± .000	.040	.007 ± .000	.014 ± .000	.009 ± .000	.068	.008 ± .000	.013 ± .000	.009 ± .000	.092	.008 ± .000	.014 ± .001	.009 ± .000	.128
	blanchard	.008 ± .000	.026 ± .002	.010 ± .000	75.043 ± 11.586	.007 ± .000	.016 ± .001	.009 ± .000	13.220 ± 4.956	.008 ± .000	.012 ± .001	.009 ± .000	1.774 ± 1.772	.008 ± .000	.012 ± .001	.009 ± .000	.190 ± .594
	catoni	.008 ± .000	.022 ± .001	.010 ± .000	96.561 ± 13.980	.007 ± .000	.016 ± .000	.009 ± .000	15.107 ± 5.370	.008 ± .000	.013 ± .001	.009 ± .000	1.835 ± 1.837	.008 ± .000	.013 ± .000	.009 ± .000	.219 ± .619
	rivasplata	.008 ± .000	.021 ± .001	.010 ± .000	76.898 ± 12.301	.007 ± .000	.014 ± .001	.009 ± .000	13.370 ± 4.931	.008 ± .000	.013 ± .000	.009 ± .000	1.695 ± 1.741	.008 ± .000	.013 ± .001	.009 ± .000	.183 ± .580
	stochastic	—	.038	—	.020	—	.037	—	.034	—	.037	—	.046	—	.037	—	.064
Fashion	ours	.109 ± .000	.115 ± .000	.102 ± .000	.128	.114 ± .001	.117 ± .001	.104 ± .001	.436	.101 ± .001	.108 ± .001	.096 ± .001	.452	.110 ± .003	.116 ± .003	.103 ± .003	.438
	blanchard	.109 ± .000	.139 ± .003	.102 ± .000	7.878 ± 11.599	.114 ± .001	.121 ± .003	.104 ± .001	13.041 ± 5.012	.102 ± .001	.106 ± .002	.096 ± .001	1.840 ± 1.864	.111 ± .003	.113 ± .003	.104 ± .002	.184 ± .600
	catoni	.109 ± .000	.152 ± .006	.102 ± .000	96.732 ± 13.464	.114 ± .001	.119 ± .002	.104 ± .001	15.103 ± 5.363	.102 ± .001	.105 ± .001	.096 ± .001	1.825 ± 1.886	.111 ± .003	.112 ± .003	.104 ± .003	.224 ± .610
	rivasplata	.109 ± .000	.129 ± .002	.102 ± .000	75.029 ± 11.918	.114 ± .001	.118 ± .002	.104 ± .001	13.495 ± 5.112	.102 ± .001	.106 ± .001	.096 ± .001	1.798 ± 1.859	.111 ± .003	.114 ± .003	.104 ± .002	.219 ± .610
	stochastic	—	.164	—	.064	—	.167	—	.218	—	.157	—	.226	—	.165	—	.219
CIFAR-10	ours	.277 ± .000	.297 ± .000	.276 ± .000	.021	.288 ± .000	.307 ± .000	.286 ± .000	.027	.273 ± .001	.284 ± .000	.263 ± .000	.079	.281 ± .001	.302 ± .001	.281 ± .001	.227
	blanchard	.277 ± .000	.386 ± .005	.276 ± .000	262.952 ± 24.385	.288 ± .000	.346 ± .005	.286 ± .000	76.609 ± 12.923	.273 ± .001	.293 ± .004	.263 ± .000	17.724 ± 6.241	.281 ± .001	.299 ± .002	.281 ± .001	2.580 ± 2.299
	catoni	.277 ± .000	.398 ± .001	.276 ± .000	268.083 ± 24.567	.288 ± .000	.343 ± .007	.286 ± .000	82.887 ± 13.493	.273 ± .001	.287 ± .003	.263 ± .000	18.978 ± 6.437	.281 ± .001	.297 ± .001	.281 ± .001	2.661 ± 2.317
	rivasplata	.277 ± .000	.354 ± .004	.276 ± .000	263.581 ± 24.435	.288 ± .000	.330 ± .003	.286 ± .000	77.488 ± 12.464	.273 ± .001	.288 ± .002	.263 ± .000	17.704 ± 5.927	.281 ± .001	.299 ± .002	.281 ± .001	2.619 ± 2.297
	stochastic	—	.363	—	.010	—	.374	—	.014	—	.349	—	.040	—	.368	—	.113
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.008 ± .000	.016 ± .000	.010 ± .000	9.520	.007 ± .000	.014 ± .000	.009 ± .000	3.594	.008 ± .000	.014 ± .000	.009 ± .000	1.877	.008 ± .000	.014 ± .001	.009 ± .000	6.589
	blanchard	.008 ± .000	.657 ± .005	.010 ± .000	1209.158 ± 157.539	.007 ± .000	.304 ± .005	.009 ± .000	3795.285 ± 88.141	.008 ± .000	.124 ± .004	.009 ± .000	1183.704 ± 5.113	.007 ± .000	.052 ± .002	.009 ± .000	347.860 ± 25.275
	catoni	.008 ± .000	.452 ± .004	.010 ± .000	12032.708 ± 157.184	.007 ± .000	.225 ± .003	.009 ± .000	3834.246 ± 89.809	.008 ± .000	.093 ± .003	.009 ± .000	1225.575 ± 51.027	.007 ± .000	.039 ± .002	.008 ± .000	39.374 ± 26.987
	rivasplata	.008 ± .000	.423 ± .004	.010 ± .000	11943.688 ± 156.365	.007 ± .000	.179 ± .003	.009 ± .000	3787.407 ± 87.968	.008 ± .000	.075 ± .002	.009 ± .000	1173.457 ± 49.846	.007 ± .000	.035 ± .002	.008 ± .000	348.717 ± 26.495
	stochastic	—	.039	—	4.760	—	.038	—	1.797	—	.037	—	.938	—	.038	—	3.294
Fashion	ours	.109 ± .000	.119 ± .000	.102 ± .000	16.776	.114 ± .001	.119 ± .001	.104 ± .001	7.869	.101 ± .001	.111 ± .001	.095 ± .001	14.224	.109 ± .002	.116 ± .002	.101 ± .002	9.187
	blanchard	.108 ± .000	.743 ± .004	.101 ± .000	11048.501 ± 146.969	.112 ± .001	.468 ± .005	.101 ± .001	3798.865 ± 87.270	.099 ± .001	.268 ± .005	.093 ± .001	1144.740 ± 49.199	.106 ± .002	.183 ± .004	.099 ± .002	328.466 ± 24.435
	catoni	.109 ± .000	.712 ± .005	.102 ± .000	1191.096 ± 15.212	.114 ± .001	.367 ± .005	.104 ± .001	3831.104 ± 88.371	.101 ± .001	.216 ± .003	.095 ± .001	1221.392 ± 5.970	.108 ± .002	.175 ± .003	.101 ± .002	386.528 ± 26.498
	rivasplata	.108 ± .000	.557 ± .003	.101 ± .000	11148.085 ± 145.818	.111 ± .001	.340 ± .003	.100 ± .001	3757.976 ± 83.965	.098 ± .001	.209 ± .003	.092 ± .001	1176.081 ± 49.829	.106 ± .002	.156 ± .003	.098 ± .002	34.716 ± 24.874
	stochastic	—	.168	—	8.388	—	.168	—	3.935	—	.159	—	7.112	—	.165	—	4.594
CIFAR-10	ours	.277 ± .000	.301 ± .000	.276 ± .000	8.466	.288 ± .000	.308 ± .000	.286 ± .000	2.415	.273 ± .001	.285 ± .000	.263 ± .000	2.256	.280 ± .001	.303 ± .001	.280 ± .001	2.747
	blanchard	.277 ± .000	.990 ± .000	.276 ± .000	58878.209 ± 356.845	.288 ± .000	.868 ± .003	.286 ± .000	8858.838 ± 134.545	.272 ± .001	.625 ± .005	.262 ± .000	2709.659 ± 76.197	.278 ± .001	.480 ± .005	.277 ± .001	86.940 ± 43.864
	catoni	.277 ± .000	.974 ± .000	.276 ± .000	17581.286 ± 185.476	.288 ± .000	.662 ± .005	.286 ± .000	5118.582 ± 105.636	.272 ± .001	.456 ± .003	.262 ± .000	1548.107 ± 58.565	.277 ± .001	.426 ± .002	.274 ± .001	783.103 ± 41.593
	rivasplata	.277 ± .000	.990 ± .000	.276 ± .000	82459.214 ± 398.763	.288 ± .000	.733 ± .003	.286 ± .000	8674.850 ± 13.468	.272 ± .001	.518 ± .004	.262 ± .000	2709.173 ± 77.205	.277 ± .001	.418 ± .004	.275 ± .001	874.307 ± 44.089
	stochastic	—	.366	—	4.233	—	.374	—	1.207	—	.350	—	1.128	—	.369	—	1.374

Table F.8. *Comparison of* ours, rivasplata, blanchard *and* catoni *based on the disintegrated bounds, and* stochastic *based on the randomized bounds learned with two learning rates* $\eta \in \{10^{-4}, 10^{-6}\}$ *and different variances* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. *We report the test risk (*$R_{\mathcal{T}}(h)$*), the bound value (Bnd), the empirical risk (*$R_{\mathcal{S}}(h)$*), and the divergence (Div) associated with each bound (the* RÉNYI *divergence for* ours, *the KL divergence for* stochastic, *and the disintegrated KL divergence for* rivasplata, blanchard *and* catoni*). More precisely, we report the mean* $\pm$ *the standard deviation for* 400 *neural networks sampled from* $\rho_{\mathbb{S}}$ *for* ours, rivasplata, blanchard, *and* catoni. *We consider, in this table, that the split ratio is* 0.7.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.011 ± .000	.019 ± .000	.013 ± .000	.047	.010 ± .000	.018 ± .000	.012 ± .000	.125	.010 ± .000	.017 ± .000	.012 ± .000	.116	.010 ± .001	.018 ± .001	.012 ± .001	.132
	blanchard	.011 ± .000	.032 ± .002	.013 ± .000	65.017 ± 11.099	.010 ± .000	.019 ± .001	.012 ± .000	1.819 ± 4.995	.010 ± .000	.016 ± .001	.012 ± .000	1.551 ± 1.635	.010 ± .001	.016 ± .001	.012 ± .001	.115 ± .560
	catoni	.011 ± .000	.028 ± .001	.013 ± .000	84.529 ± 13.023	.010 ± .000	.021 ± .000	.012 ± .000	11.910 ± 5.053	.010 ± .000	.017 ± .001	.012 ± .000	1.228 ± 1.637	.010 ± .001	.017 ± .001	.012 ± .001	.173 ± .512
	rivasplata	.011 ± .000	.026 ± .001	.013 ± .000	68.055 ± 11.606	.010 ± .000	.018 ± .001	.012 ± .000	1.637 ± 4.962	.010 ± .000	.016 ± .000	.012 ± .000	1.408 ± 1.639	.010 ± .001	.016 ± .001	.012 ± .001	.160 ± .529
	stochastic	—	.044	—	.023	—	.043	—	.062	—	.042	—	.058	—	.043	—	.066
Fashion	ours	.099 ± .000	.112 ± .000	.098 ± .000	.067	.107 ± .001	.115 ± .001	.100 ± .001	.542	.098 ± .002	.107 ± .001	.093 ± .001	.353	.108 ± .003	.117 ± .002	.102 ± .002	.312
	blanchard	.099 ± .000	.138 ± .004	.098 ± .000	61.733 ± 1.862	.107 ± .001	.119 ± .003	.101 ± .001	1.651 ± 4.230	.099 ± .001	.104 ± .002	.094 ± .001	1.342 ± 1.664	.108 ± .003	.113 ± .003	.103 ± .002	.143 ± .534
	catoni	.099 ± .000	.155 ± .007	.098 ± .000	83.929 ± 12.212	.107 ± .001	.116 ± .003	.101 ± .001	11.543 ± 4.870	.099 ± .002	.103 ± .002	.094 ± .001	1.437 ± 1.594	.108 ± .003	.112 ± .003	.103 ± .002	.153 ± .545
	rivasplata	.099 ± .000	.128 ± .002	.098 ± .000	65.737 ± 11.733	.107 ± .001	.116 ± .002	.101 ± .001	1.958 ± 4.794	.099 ± .002	.105 ± .002	.094 ± .001	1.491 ± 1.618	.108 ± .003	.114 ± .003	.103 ± .002	.155 ± .546
	stochastic	—	.161	—	.034	—	.164	—	.271	—	.155	—	.177	—	.166	—	.156
CIFAR-10	ours	.277 ± .000	.296 ± .000	.272 ± .000	.016	.266 ± .000	.281 ± .000	.257 ± .000	.022	.253 ± .001	.272 ± .000	.248 ± .000	.069	.236 ± .001	.258 ± .001	.235 ± .001	.118
	blanchard	.277 ± .000	.399 ± .006	.272 ± .000	257.371 ± 23.327	.266 ± .000	.322 ± .005	.257 ± .000	7.190 ± 11.685	.253 ± .001	.281 ± .005	.248 ± .000	15.214 ± 5.838	.236 ± .001	.255 ± .002	.235 ± .001	2.223 ± 2.016
	catoni	.277 ± .000	.399 ± .002	.272 ± .000	269.048 ± 23.489	.266 ± .000	.328 ± .009	.257 ± .000	81.217 ± 13.476	.253 ± .001	.275 ± .004	.248 ± .000	16.576 ± 6.087	.236 ± .001	.253 ± .002	.235 ± .001	2.248 ± 2.082
	rivasplata	.277 ± .000	.362 ± .004	.272 ± .000	258.993 ± 23.725	.266 ± .000	.305 ± .004	.257 ± .000	72.737 ± 12.750	.253 ± .001	.275 ± .003	.248 ± .000	15.342 ± 5.948	.236 ± .001	.255 ± .002	.235 ± .001	2.220 ± 2.180
	stochastic	—	.362	—	.008	—	.345	—	.011	—	.336	—	.034	—	.322	—	.059
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.011 ± .000	.025 ± .000	.013 ± .000	45.094	.010 ± .000	.019 ± .000	.012 ± .000	7.479	.010 ± .000	.018 ± .000	.012 ± .000	5.269	.010 ± .000	.018 ± .001	.011 ± .001	6.510
	blanchard	.011 ± .000	.737 ± .004	.013 ± .000	11285.050 ± 147.363	.010 ± .000	.381 ± .006	.012 ± .000	3785.071 ± 85.889	.010 ± .000	.160 ± .005	.011 ± .000	1181.043 ± 46.219	.009 ± .000	.067 ± .003	.011 ± .001	34.267 ± 26.244
	catoni	.011 ± .000	.547 ± .004	.013 ± .000	11965.668 ± 153.481	.010 ± .000	.283 ± .004	.012 ± .000	3811.642 ± 88.111	.010 ± .000	.120 ± .004	.011 ± .000	1212.373 ± 48.835	.009 ± .000	.050 ± .002	.010 ± .000	383.387 ± 27.059
	rivasplata	.011 ± .000	.509 ± .004	.013 ± .000	11555.623 ± 15.287	.010 ± .000	.226 ± .004	.012 ± .000	3695.054 ± 9.289	.009 ± .000	.096 ± .003	.011 ± .000	1171.892 ± 47.812	.009 ± .000	.044 ± .002	.010 ± .000	343.025 ± 25.804
	stochastic	—	.050	—	22.547	—	.044	—	3.740	—	.043	—	2.634	—	.043	—	3.255
Fashion	ours	.099 ± .000	.116 ± .000	.098 ± .000	11.922	.107 ± .001	.117 ± .001	.101 ± .001	6.556	.097 ± .001	.109 ± .001	.092 ± .001	9.235	.105 ± .002	.118 ± .002	.100 ± .002	1.362
	blanchard	.098 ± .000	.795 ± .004	.098 ± .000	10179.790 ± 138.889	.101 ± .001	.524 ± .006	.096 ± .001	3752.748 ± 9.952	.095 ± .001	.291 ± .005	.090 ± .001	1091.018 ± 47.577	.104 ± .002	.195 ± .005	.098 ± .002	309.857 ± 24.422
	catoni	.099 ± .000	.808 ± .002	.098 ± .000	11999.071 ± 158.418	.107 ± .001	.425 ± .006	.100 ± .001	3817.800 ± 91.674	.098 ± .001	.235 ± .004	.093 ± .001	1216.042 ± 5.641	.106 ± .002	.182 ± .004	.101 ± .002	376.493 ± 27.018
	rivasplata	.098 ± .000	.619 ± .004	.097 ± .000	10768.160 ± 146.634	.099 ± .001	.369 ± .004	.094 ± .001	3565.270 ± 88.164	.094 ± .001	.224 ± .004	.089 ± .001	1137.876 ± 48.421	.103 ± .002	.164 ± .003	.097 ± .002	318.512 ± 24.741
	stochastic	—	.164	—	5.961	—	.166	—	3.278	—	.156	—	4.618	—	.166	—	5.181
CIFAR-10	ours	.277 ± .000	.303 ± .000	.272 ± .000	12.803	.266 ± .000	.282 ± .000	.257 ± .000	2.312	.253 ± .001	.272 ± .000	.248 ± .000	1.641	.236 ± .001	.259 ± .001	.235 ± .001	1.929
	blanchard	.277 ± .000	.990 ± .000	.272 ± .000	2577.092 ± 236.075	.266 ± .000	.901 ± .003	.257 ± .000	8788.732 ± 134.680	.253 ± .001	.662 ± .005	.247 ± .000	2683.054 ± 73.139	.235 ± .001	.464 ± .006	.233 ± .001	85.586 ± 41.917
	catoni	.277 ± .000	1.000 ± .000	.272 ± .000	177807.417 ± 546.892	.266 ± .000	.601 ± .005	.257 ± .000	331.757 ± 83.561	.253 ± .001	.416 ± .003	.247 ± .000	85.973 ± 4.961	.234 ± .001	.369 ± .003	.233 ± .001	485.863 ± 31.335
	rivasplata	.277 ± .000	.990 ± .000	.272 ± .000	48522.489 ± 309.735	.266 ± .000	.762 ± .003	.257 ± .000	850.968 ± 131.507	.252 ± .001	.542 ± .004	.247 ± .000	2696.074 ± 73.062	.234 ± .001	.393 ± .004	.232 ± .001	858.936 ± 41.972
	stochastic	—	.366	—	6.401	—	.346	—	1.156	—	.336	—	.821	—	.322	—	.965

Table F.9. *Comparison of* `ours`, `rivasplata`, `blanchard` *and* `catoni` *based on the disintegrated bounds, and* `stochastic` *based on the randomized bounds learned with two learning rates* $\eta \in \{10^{-4}, 10^{-6}\}$ *and different variances* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. *We report the test risk* ($R_{\mathcal{T}}(h)$)*, the bound value (Bnd), the empirical risk* ($R_{\mathcal{S}}(h)$)*, and the divergence (Div) associated with each bound (the* RÉNYI *divergence for* `ours`*, the KL divergence for* `stochastic`*, and the disintegrated KL divergence for* `rivasplata`, `blanchard` *and* `catoni`*). More precisely, we report the mean* $\pm$ *the standard deviation for* 400 *neural networks sampled from* $\rho_{\mathcal{S}}$ *for* `ours`, `rivasplata`, `blanchard`, *and* `catoni`. *We consider, in this table, that the split ratio is* 0.8.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.011 ± .000	.020 ± .000	.013 ± .000	.064	.008 ± .000	.017 ± .000	.010 ± .000	.050	.011 ± .000	.018 ± .000	.011 ± .000	.112	.010 ± .001	.016 ± .001	.009 ± .001	.073
	blanchard	.011 ± .000	.034 ± .003	.013 ± .000	49.248 ± 1.541	.008 ± .000	.018 ± .001	.010 ± .000	8.031 ± 3.654	.011 ± .000	.016 ± .001	.011 ± .000	.810 ± 1.248	.010 ± .001	.014 ± .001	.010 ± .001	.102 ± .448
	catoni	.011 ± .000	.030 ± .002	.013 ± .000	66.244 ± 11.961	.008 ± .000	.018 ± .001	.010 ± .000	8.685 ± 3.987	.011 ± .000	.019 ± .001	.011 ± .000	1.011 ± 1.283	.010 ± .001	.016 ± .001	.010 ± .001	.131 ± .422
	rivasplata	.011 ± .000	.028 ± .002	.013 ± .000	5.344 ± 1.600	.008 ± .000	.017 ± .001	.010 ± .000	7.757 ± 4.187	.011 ± .000	.017 ± .001	.011 ± .000	.861 ± 1.361	.010 ± .001	.014 ± .001	.010 ± .001	.090 ± .460
	stochastic	—	.046	—	.032	—	.041	—	.025	—	.043	—	.056	—	.040	—	.037
Fashion	ours	.103 ± .000	.117 ± .000	.099 ± .000	.068	.098 ± .001	.114 ± .001	.096 ± .001	.178	.104 ± .001	.117 ± .002	.099 ± .002	.587	.107 ± .004	.119 ± .004	.101 ± .003	.328
	blanchard	.103 ± .000	.144 ± .004	.099 ± .000	5.069 ± 9.537	.098 ± .001	.116 ± .003	.096 ± .001	8.105 ± 3.874	.104 ± .001	.113 ± .002	.100 ± .002	.990 ± 1.435	.109 ± .004	.114 ± .004	.102 ± .004	.102 ± .444
	catoni	.103 ± .000	.168 ± .009	.099 ± .000	66.761 ± 1.939	.098 ± .001	.115 ± .004	.096 ± .001	8.698 ± 3.974	.104 ± .001	.112 ± .002	.100 ± .002	.934 ± 1.413	.109 ± .004	.113 ± .004	.102 ± .004	.100 ± .457
	rivasplata	.103 ± .000	.132 ± .003	.099 ± .000	52.096 ± 1.745	.098 ± .001	.113 ± .002	.096 ± .001	7.820 ± 4.154	.104 ± .001	.114 ± .002	.100 ± .002	.939 ± 1.417	.108 ± .004	.115 ± .004	.102 ± .004	.100 ± .464
	stochastic	—	.165	—	.034	—	.162	—	.089	—	.166	—	.294	—	.168	—	.164
CIFAR-10	ours	.249 ± .000	.265 ± .000	.237 ± .000	.014	.247 ± .000	.271 ± .000	.243 ± .000	.018	.259 ± .001	.281 ± .001	.252 ± .001	.055	.249 ± .001	.274 ± .001	.245 ± .001	.072
	blanchard	.249 ± .000	.384 ± .007	.237 ± .000	24.108 ± 22.114	.247 ± .000	.316 ± .006	.243 ± .000	59.096 ± 1.459	.259 ± .001	.289 ± .006	.252 ± .001	11.804 ± 5.001	.249 ± .001	.269 ± .003	.245 ± .001	1.578 ± 1.705
	catoni	.249 ± .000	.368 ± .003	.237 ± .000	27.284 ± 23.618	.247 ± .000	.337 ± .012	.243 ± .000	73.833 ± 11.692	.259 ± .001	.285 ± .005	.252 ± .001	12.808 ± 5.089	.249 ± .001	.266 ± .002	.245 ± .001	1.635 ± 1.773
	rivasplata	.249 ± .000	.341 ± .005	.237 ± .000	244.258 ± 22.339	.247 ± .000	.298 ± .004	.243 ± .000	61.907 ± 11.135	.259 ± .001	.283 ± .003	.252 ± .001	11.818 ± 4.923	.249 ± .001	.269 ± .002	.245 ± .001	1.629 ± 1.763
	stochastic	—	.328	—	.007	—	.334	—	.009	—	.344	—	.028	—	.337	—	.036
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.011 ± .000	.030 ± .000	.013 ± .000	53.875	.008 ± .000	.018 ± .000	.010 ± .000	4.369	.011 ± .000	.019 ± .000	.011 ± .000	5.063	.009 ± .001	.016 ± .001	.009 ± .001	4.854
	blanchard	.011 ± .000	.828 ± .004	.013 ± .000	10014.066 ± 14.769	.008 ± .000	.491 ± .007	.010 ± .000	3707.758 ± 86.461	.011 ± .000	.211 ± .007	.011 ± .000	1151.660 ± 47.238	.009 ± .001	.076 ± .005	.009 ± .001	303.402 ± 25.072
	catoni	.011 ± .000	.684 ± .004	.013 ± .000	12238.359 ± 158.595	.008 ± .000	.343 ± .005	.010 ± .000	3834.114 ± 88.516	.011 ± .000	.168 ± .006	.011 ± .000	121.777 ± 48.460	.009 ± .001	.060 ± .003	.008 ± .001	356.740 ± 25.649
	rivasplata	.011 ± .000	.662 ± .005	.013 ± .000	1207.265 ± 161.842	.008 ± .000	.305 ± .005	.010 ± .000	3785.930 ± 87.976	.011 ± .000	.125 ± .004	.010 ± .000	1141.437 ± 46.910	.009 ± .001	.048 ± .002	.008 ± .000	305.573 ± 23.629
	stochastic	—	.055	—	26.937	—	.042	—	2.185	—	.044	—	2.532	—	.040	—	2.427
Fashion	ours	.102 ± .000	.121 ± .000	.099 ± .000	1.120	.098 ± .001	.115 ± .001	.096 ± .001	3.956	.102 ± .001	.118 ± .002	.098 ± .001	7.830	.103 ± .003	.118 ± .003	.097 ± .002	8.797
	blanchard	.101 ± .000	.990 ± .000	.098 ± .000	27936.970 ± 235.840	.096 ± .001	.585 ± .007	.094 ± .001	321.105 ± 81.006	.098 ± .001	.348 ± .007	.094 ± .001	1045.641 ± 44.087	.101 ± .002	.208 ± .006	.095 ± .002	273.641 ± 24.046
	catoni	.103 ± .000	.865 ± .002	.099 ± .000	12143.837 ± 161.857	.098 ± .001	.536 ± .008	.096 ± .001	3802.871 ± 87.750	.103 ± .001	.286 ± .006	.098 ± .001	1202.907 ± 47.928	.105 ± .003	.191 ± .005	.098 ± .003	354.246 ± 25.507
	rivasplata	.102 ± .000	.746 ± .004	.098 ± .000	11305.448 ± 149.693	.096 ± .001	.438 ± .005	.093 ± .001	3458.977 ± 83.715	.097 ± .001	.264 ± .004	.093 ± .001	1101.567 ± 44.816	.099 ± .002	.172 ± .004	.094 ± .002	285.588 ± 24.451
	stochastic	—	.168	—	5.060	—	.163	—	1.978	—	.166	—	3.915	—	.166	—	4.399
CIFAR-10	ours	.249 ± .000	.274 ± .000	.237 ± .000	14.083	.247 ± .000	.273 ± .000	.243 ± .000	1.770	.259 ± .001	.282 ± .001	.252 ± .001	1.098	.248 ± .001	.275 ± .001	.245 ± .001	1.461
	blanchard	.249 ± .000	.990 ± .000	.237 ± .000	26575.507 ± 218.278	.247 ± .000	.925 ± .002	.243 ± .000	7135.143 ± 117.030	.259 ± .001	.739 ± .006	.251 ± .001	2581.211 ± 74.799	.247 ± .001	.526 ± .007	.243 ± .001	831.790 ± 4.592
	catoni	.249 ± .000	1.000 ± .000	.237 ± .000	154168.585 ± 539.590	.247 ± .000	.677 ± .008	.243 ± .000	3148.174 ± 83.069	.259 ± .001	.549 ± .006	.252 ± .001	1735.530 ± 57.888	.248 ± .001	.425 ± .005	.244 ± .001	675.780 ± 38.306
	rivasplata	.249 ± .000	.990 ± .000	.237 ± .000	35062.089 ± 246.257	.247 ± .000	.824 ± .003	.243 ± .000	8092.236 ± 125.162	.259 ± .001	.610 ± .005	.251 ± .001	2652.857 ± 75.369	.247 ± .001	.441 ± .005	.242 ± .001	84.056 ± 4.952
	stochastic	—	.334	—	7.041	—	.335	—	.885	—	.345	—	.549	—	.337	—	.731

Table F.10. *Comparison of* `ours`, `rivasplata`, `blanchard` *and* `catoni` *based on the disintegrated bounds, and* `stochastic` *based on the randomized bounds learned with two learning rates* $\eta \in \{10^{-4}, 10^{-6}\}$ *and different variances* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. *We report the test risk (*$R_{\mathcal{T}}(h)$*), the bound value (Bnd), the empirical risk (*$R_{\mathcal{S}}(h)$*), and the divergence (Div) associated with each bound (the* RÉNYI *divergence for* `ours`, *the KL divergence for* `stochastic`, *and the disintegrated KL divergence for* `rivasplata`, `blanchard` *and* `catoni`*). More precisely, we report the mean* $\pm$ *the standard deviation for* 400 *neural networks sampled from* $\rho_{\mathcal{S}}$ *for* `ours`, `rivasplata`, `blanchard`, *and* `catoni`. *We consider, in this table, that the split ratio is* 0.9.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.008 ± .000	.018 ± .000	.008 ± .000	.029	.011 ± .000	.020 ± .000	.010 ± .000	.052	.009 ± .000	.018 ± .001	.009 ± .000	.059	.008 ± .000	.019 ± .001	.009 ± .001	.023
	blanchard	.008 ± .000	.033 ± .004	.009 ± .000	35.446 ± 8.610	.011 ± .000	.020 ± .002	.010 ± .000	4.933 ± 2.958	.009 ± .000	.015 ± .001	.009 ± .000	.490 ± .960	.008 ± .001	.016 ± .001	.009 ± .001	.059 ± .299
	catoni	.008 ± .000	.026 ± .002	.009 ± .000	41.267 ± 9.234	.011 ± .000	.019 ± .001	.010 ± .000	4.564 ± 3.263	.009 ± .000	.016 ± .001	.009 ± .000	.581 ± .989	.008 ± .001	.017 ± .001	.009 ± .001	.078 ± .320
	rivasplata	.008 ± .000	.025 ± .002	.009 ± .000	35.856 ± 8.648	.011 ± .000	.019 ± .001	.010 ± .000	4.620 ± 2.983	.009 ± .000	.015 ± .001	.009 ± .000	.448 ± 1.045	.008 ± .000	.016 ± .001	.009 ± .001	.041 ± .330
	stochastic	—	.041	—	.014	—	.045	—	.026	—	.042	—	.030	—	.043	—	.012
Fashion	ours	.094 ± .000	.113 ± .000	.089 ± .000	.029	.091 ± .001	.119 ± .001	.095 ± .001	.107	.092 ± .002	.113 ± .001	.089 ± .001	.097	.103 ± .003	.124 ± .003	.099 ± .003	.045
	blanchard	.094 ± .000	.140 ± .006	.089 ± .000	32.563 ± 8.007	.091 ± .001	.119 ± .004	.095 ± .001	4.567 ± 2.912	.092 ± .002	.106 ± .002	.089 ± .001	.468 ± 1.101	.104 ± .003	.116 ± .003	.099 ± .003	.063 ± .300
	catoni	.094 ± .000	.146 ± .002	.089 ± .000	4.355 ± 9.121	.091 ± .001	.120 ± .005	.095 ± .001	4.895 ± 3.064	.092 ± .002	.106 ± .002	.089 ± .001	.473 ± 1.052	.103 ± .003	.117 ± .003	.099 ± .003	.079 ± .319
	rivasplata	.094 ± .000	.127 ± .004	.089 ± .000	33.175 ± 8.710	.091 ± .001	.117 ± .002	.095 ± .001	4.774 ± 3.003	.092 ± .002	.107 ± .002	.089 ± .001	.479 ± .924	.103 ± .003	.118 ± .003	.099 ± .002	.045 ± .330
	stochastic	—	.159	—	.015	—	.166	—	.053	—	.159	—	.048	—	.172	—	.023
CIFAR-10	ours	.231 ± .000	.268 ± .000	.228 ± .000	.011	.235 ± .000	.267 ± .000	.227 ± .000	.009	.218 ± .001	.253 ± .001	.214 ± .001	.024	.231 ± .001	.264 ± .002	.224 ± .002	.036
	blanchard	.231 ± .000	.418 ± .010	.228 ± .000	193.922 ± 19.216	.235 ± .000	.312 ± .009	.227 ± .000	39.705 ± 8.929	.218 ± .000	.256 ± .007	.214 ± .001	6.919 ± 3.722	.231 ± .001	.255 ± .003	.224 ± .002	.878 ± 1.248
	catoni	.231 ± .000	.388 ± .005	.228 ± .000	255.538 ± 22.306	.235 ± .000	.337 ± .003	.227 ± .000	53.736 ± 1.302	.218 ± .000	.257 ± .007	.214 ± .001	7.060 ± 3.626	.231 ± .001	.255 ± .003	.224 ± .002	.857 ± 1.264
	rivasplata	.231 ± .000	.364 ± .007	.228 ± .000	202.026 ± 19.688	.235 ± .000	.293 ± .006	.227 ± .000	42.458 ± 9.250	.218 ± .001	.251 ± .004	.214 ± .001	6.780 ± 3.575	.231 ± .001	.256 ± .002	.224 ± .002	.854 ± 1.275
	stochastic	—	.328	—	.005	—	.327	—	.005	—	.312	—	.012	—	.324	—	.018
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$\eta = 10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	ours	.008 ± .000	.018 ± .000	.009 ± .000	2.107	.011 ± .000	.021 ± .000	.010 ± .000	1.329	.008 ± .000	.019 ± .001	.008 ± .000	3.598	.008 ± .001	.020 ± .001	.009 ± .001	4.216
	blanchard	.008 ± .000	.982 ± .001	.008 ± .000	11722.999 ± 157.452	.011 ± .000	.706 ± .008	.010 ± .000	3475.807 ± 77.708	.008 ± .000	.331 ± .011	.008 ± .000	1076.767 ± 46.059	.008 ± .000	.108 ± .008	.009 ± .001	242.819 ± 23.775
	catoni	.008 ± .000	1.000 ± .000	.008 ± .000	60838.120 ± 346.289	.011 ± .000	.515 ± .007	.010 ± .000	3728.586 ± 86.803	.008 ± .000	.243 ± .007	.008 ± .000	1166.491 ± 48.086	.008 ± .000	.087 ± .006	.009 ± .001	277.823 ± 25.431
	rivasplata	.008 ± .000	.879 ± .003	.008 ± .000	12257.175 ± 152.738	.010 ± .000	.481 ± .007	.010 ± .000	3602.529 ± 78.717	.008 ± .000	.201 ± .007	.008 ± .000	1126.882 ± 47.430	.008 ± .001	.067 ± .004	.009 ± .001	242.366 ± 22.337
	stochastic	—	.042	—	1.053	—	.045	—	.664	—	.042	—	1.799	—	.043	—	2.108
Fashion	ours	.094 ± .000	.115 ± .000	.089 ± .000	2.501	.091 ± .001	.121 ± .001	.095 ± .001	2.925	.092 ± .002	.114 ± .001	.088 ± .001	3.069	.102 ± .002	.125 ± .003	.098 ± .002	3.159
	blanchard	.094 ± .000	.990 ± .000	.089 ± .000	19455.864 ± 19.460	.089 ± .001	.792 ± .007	.093 ± .001	3402.546 ± 86.590	.090 ± .001	.461 ± .010	.087 ± .001	1002.861 ± 44.393	.102 ± .002	.244 ± .009	.098 ± .002	206.177 ± 2.051
	catoni	.094 ± .000	1.000 ± .000	.089 ± .000	60888.029 ± 346.501	.091 ± .001	.813 ± .012	.095 ± .001	3756.375 ± 9.419	.092 ± .002	.390 ± .010	.089 ± .001	1161.884 ± 52.073	.103 ± .003	.215 ± .007	.099 ± .002	277.284 ± 25.479
	rivasplata	.094 ± .000	.990 ± .000	.089 ± .000	27137.315 ± 227.934	.088 ± .001	.597 ± .007	.093 ± .001	3371.321 ± 86.352	.090 ± .001	.331 ± .007	.086 ± .001	1003.481 ± 48.362	.101 ± .002	.195 ± .006	.097 ± .002	207.442 ± 21.896
	stochastic	—	.160	—	1.250	—	.167	—	1.463	—	.160	—	1.535	—	.172	—	1.579
CIFAR-10	ours	.231 ± .000	.279 ± .000	.228 ± .000	12.925	.235 ± .000	.268 ± .000	.227 ± .000	1.371	.218 ± .001	.254 ± .001	.214 ± .001	.715	.231 ± .001	.264 ± .002	.224 ± .002	1.019
	blanchard	.231 ± .000	.990 ± .000	.228 ± .000	26032.808 ± 222.475	.235 ± .000	.986 ± .001	.227 ± .000	6875.633 ± 112.137	.217 ± .000	.831 ± .006	.214 ± .001	2292.053 ± 68.347	.230 ± .001	.606 ± .010	.222 ± .001	76.644 ± 39.246
	catoni	.231 ± .000	1.000 ± .000	.228 ± .000	17684.651 ± 576.711	.235 ± .000	.980 ± .000	.227 ± .000	8265.727 ± 123.941	.218 ± .000	.834 ± .011	.214 ± .001	2664.069 ± 73.915	.231 ± .001	.517 ± .009	.224 ± .002	85.593 ± 41.022
	rivasplata	.231 ± .000	.988 ± .001	.228 ± .000	14284.846 ± 169.166	.235 ± .000	.919 ± .002	.227 ± .000	7121.350 ± 114.645	.217 ± .000	.699 ± .006	.213 ± .001	2502.412 ± 68.728	.229 ± .001	.494 ± .007	.221 ± .001	776.237 ± 39.540
	stochastic	—	.335	—	6.462	—	.328	—	.685	—	.313	—	.358	—	.324	—	.510

Table F.11. *Comparison of the bound values before performing Step* **2)** *of our Training Method for* `ours`, `rivasplata`, `blanchard` *and* `catoni`. *More precisely, for each split and each variance* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, *we report the mean* $\pm$ *the standard deviation (for* 400 *neural networks sampled from* π*) of the test risk (*$R_{\mathcal{T}}(h)$*), the empirical risk (*$R_{\mathcal{S}}(h)$*), and the value of the bounds of Corollaries 6.4.1 and 6.4.2. We consider in this table that the dataset is MNIST.*

	Split	$R_{\mathcal{T}}(h)$	$R_{\mathcal{S}}(h)$	Cor. 6.4.1	Eq. (6.1)	Eq. (6.2)	Eq. (6.3)
$\sigma^2 = 10^{-6}$	.0	.901 ± .002	.901 ± .002	.908 ± .002	.906 ± .002	.905 ± .002	.906 ± .002
	.1	.035 ± .000	.039 ± .000	.045 ± .000	.043 ± .000	.043 ± .000	.042 ± .000
	.2	.016 ± .000	.019 ± .000	.023 ± .000	.022 ± .000	.022 ± .000	.022 ± .000
	.3	.012 ± .000	.013 ± .000	.017 ± .000	.016 ± .000	.015 ± .000	.015 ± .000
	.4	.010 ± .000	.013 ± .000	.017 ± .000	.016 ± .000	.016 ± .000	.016 ± .000
	.5	.008 ± .000	.010 ± .000	.015 ± .000	.013 ± .000	.013 ± .000	.014 ± .000
	.6	.008 ± .000	.010 ± .000	.014 ± .000	.013 ± .000	.013 ± .000	.014 ± .000
	.7	.011 ± .000	.013 ± .000	.019 ± .000	.017 ± .000	.017 ± .000	.018 ± .000
	.8	.011 ± .000	.013 ± .000	.020 ± .000	.018 ± .000	.018 ± .000	.020 ± .000
	.9	.008 ± .000	.009 ± .000	.018 ± .000	.015 ± .000	.014 ± .000	.015 ± .000
$\sigma^2 = 10^{-5}$	.0	.897 ± .013	.897 ± .012	.904 ± .012	.902 ± .012	.902 ± .012	.903 ± .012
	.1	.024 ± .000	.030 ± .001	.035 ± .001	.034 ± .001	.033 ± .001	.033 ± .001
	.2	.015 ± .000	.019 ± .000	.023 ± .000	.022 ± .000	.021 ± .000	.021 ± .000
	.3	.009 ± .000	.011 ± .000	.015 ± .000	.014 ± .000	.013 ± .000	.013 ± .000
	.4	.012 ± .000	.014 ± .000	.018 ± .000	.017 ± .000	.017 ± .000	.017 ± .000
	.5	.006 ± .000	.009 ± .000	.012 ± .000	.011 ± .000	.011 ± .000	.012 ± .000
	.6	.007 ± .000	.009 ± .000	.014 ± .000	.013 ± .000	.012 ± .000	.013 ± .000
	.7	.010 ± .000	.012 ± .000	.018 ± .000	.016 ± .000	.016 ± .000	.017 ± .000
	.8	.008 ± .000	.010 ± .000	.017 ± .000	.015 ± .000	.014 ± .000	.017 ± .000
	.9	.011 ± .000	.010 ± .000	.020 ± .000	.017 ± .000	.017 ± .000	.018 ± .000

	Split	$R_{\mathcal{T}}(h)$	$R_{\mathcal{S}}(h)$	Cor. 6.4.1	Eq. (6.1)	Eq. (6.2)	Eq. (6.3)
$\sigma^2 = 10^{-4}$	.0	.898 ± .017	.898 ± .017	.905 ± .016	.903 ± .016	.902 ± .016	.903 ± .016
	.1	.035 ± .003	.039 ± .002	.045 ± .002	.044 ± .002	.043 ± .002	.043 ± .002
	.2	.015 ± .001	.016 ± .001	.020 ± .001	.019 ± .001	.019 ± .001	.019 ± .001
	.3	.012 ± .000	.016 ± .000	.020 ± .001	.019 ± .001	.019 ± .001	.019 ± .001
	.4	.009 ± .000	.011 ± .000	.015 ± .000	.014 ± .000	.014 ± .000	.014 ± .000
	.5	.008 ± .000	.010 ± .000	.015 ± .000	.013 ± .000	.013 ± .000	.014 ± .000
	.6	.008 ± .000	.009 ± .000	.013 ± .000	.012 ± .000	.012 ± .000	.013 ± .000
	.7	.010 ± .000	.012 ± .000	.017 ± .000	.016 ± .000	.015 ± .000	.016 ± .000
	.8	.011 ± .000	.011 ± .000	.018 ± .000	.016 ± .000	.016 ± .000	.018 ± .000
	.9	.009 ± .000	.009 ± .000	.018 ± .001	.015 ± .001	.015 ± .001	.016 ± .001
$\sigma^2 = 10^{-3}$	.0	.903 ± .014	.902 ± .014	.909 ± .013	.907 ± .013	.907 ± .013	.907 ± .013
	.1	.041 ± .005	.045 ± .005	.050 ± .005	.049 ± .005	.048 ± .005	.048 ± .005
	.2	.020 ± .002	.022 ± .002	.026 ± .002	.025 ± .002	.025 ± .002	.024 ± .002
	.3	.014 ± .001	.015 ± .001	.019 ± .001	.018 ± .001	.018 ± .001	.018 ± .001
	.4	.015 ± .001	.016 ± .001	.021 ± .001	.020 ± .001	.019 ± .001	.019 ± .001
	.5	.015 ± .001	.015 ± .001	.020 ± .001	.019 ± .001	.018 ± .001	.018 ± .001
	.6	.008 ± .000	.010 ± .000	.014 ± .001	.013 ± .001	.012 ± .000	.013 ± .000
	.7	.010 ± .001	.012 ± .001	.018 ± .001	.016 ± .001	.016 ± .001	.017 ± .001
	.8	.010 ± .001	.010 ± .001	.016 ± .001	.014 ± .001	.014 ± .001	.016 ± .001
	.9	.008 ± .000	.009 ± .001	.019 ± .001	.016 ± .001	.015 ± .001	.017 ± .001

Table F.12. *Comparison of the bound values before performing Step* **2)** *of our Training Method for* `ours`, `rivasplata`, `blanchard` *and* `catoni`. *More precisely, for each split and each variance* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, *we report the mean* $\pm$ *the standard deviation (for* 400 *neural networks sampled from* π*) of the test risk (*$R_{\mathcal{T}}(h)$*), the empirical risk (*$R_{\mathcal{S}}(h)$*), and the value of the bounds of Corollaries 6.4.1 and 6.4.2. We consider in this table that the dataset is Fashion-MNIST.*

	Split	$R_{\mathcal{T}}(h)$	$R_{\mathcal{S}}(h)$	Cor. 6.4.1	Eq. (6.1)	Eq. (6.2)	Eq. (6.3)
$\sigma^2 = 10^{-6}$	.0	.970 ± .028	.970 ± .027	.972 ± .025	.971 ± .025	.971 ± .026	.972 ± .026
	.1	.166 ± .001	.159 ± .000	.169 ± .000	.167 ± .000	.166 ± .000	.167 ± .000
	.2	.168 ± .002	.160 ± .001	.170 ± .001	.168 ± .001	.167 ± .001	.168 ± .001
	.3	.126 ± .000	.124 ± .000	.134 ± .000	.132 ± .000	.131 ± .000	.131 ± .000
	.4	.118 ± .001	.112 ± .000	.123 ± .000	.120 ± .000	.119 ± .000	.119 ± .000
	.5	.106 ± .000	.101 ± .000	.113 ± .000	.110 ± .000	.109 ± .000	.109 ± .000
	.6	.109 ± .000	.102 ± .000	.115 ± .000	.112 ± .000	.110 ± .000	.110 ± .000
	.7	.099 ± .000	.098 ± .000	.112 ± .000	.109 ± .000	.108 ± .000	.107 ± .000
	.8	.103 ± .000	.099 ± .000	.117 ± .000	.112 ± .000	.111 ± .000	.110 ± .000
	.9	.094 ± .000	.089 ± .000	.113 ± .000	.107 ± .000	.105 ± .000	.106 ± .000
$\sigma^2 = 10^{-5}$	.0	.945 ± .038	.945 ± .037	.949 ± .035	.948 ± .035	.948 ± .036	.948 ± .036
	.1	.158 ± .001	.151 ± .001	.161 ± .001	.159 ± .001	.158 ± .001	.159 ± .001
	.2	.157 ± .003	.151 ± .003	.162 ± .003	.159 ± .003	.158 ± .003	.159 ± .003
	.3	.126 ± .001	.121 ± .001	.131 ± .001	.128 ± .001	.127 ± .001	.128 ± .001
	.4	.114 ± .001	.107 ± .001	.118 ± .001	.115 ± .001	.114 ± .001	.114 ± .001
	.5	.104 ± .001	.099 ± .000	.110 ± .000	.108 ± .000	.107 ± .000	.106 ± .000
	.6	.115 ± .001	.104 ± .001	.117 ± .001	.114 ± .001	.113 ± .001	.112 ± .001
	.7	.107 ± .001	.101 ± .001	.115 ± .001	.111 ± .001	.110 ± .001	.109 ± .001
	.8	.098 ± .001	.096 ± .001	.114 ± .001	.109 ± .001	.108 ± .001	.107 ± .001
	.9	.091 ± .001	.095 ± .001	.119 ± .001	.113 ± .001	.111 ± .001	.112 ± .001

	Split	$R_{\mathcal{T}}(h)$	$R_{\mathcal{S}}(h)$	Cor. 6.4.1	Eq. (6.1)	Eq. (6.2)	Eq. (6.3)
$\sigma^2 = 10^{-4}$	.0	.912 ± .027	.912 ± .027	.918 ± .026	.916 ± .027	.916 ± .027	.916 ± .026
	.1	.164 ± .003	.154 ± .003	.164 ± .003	.162 ± .003	.161 ± .003	.162 ± .004
	.2	.164 ± .009	.160 ± .009	.170 ± .010	.168 ± .010	.167 ± .010	.168 ± .010
	.3	.125 ± .002	.119 ± .002	.129 ± .002	.126 ± .002	.126 ± .002	.126 ± .002
	.4	.119 ± .003	.113 ± .003	.124 ± .003	.121 ± .003	.120 ± .003	.120 ± .003
	.5	.109 ± .002	.102 ± .001	.113 ± .001	.110 ± .001	.109 ± .001	.109 ± .001
	.6	.102 ± .001	.096 ± .001	.109 ± .001	.105 ± .001	.105 ± .001	.104 ± .001
	.7	.099 ± .002	.094 ± .001	.108 ± .001	.104 ± .001	.103 ± .001	.102 ± .001
	.8	.104 ± .001	.100 ± .002	.118 ± .002	.113 ± .002	.112 ± .002	.111 ± .002
	.9	.092 ± .002	.089 ± .001	.113 ± .001	.107 ± .001	.105 ± .001	.106 ± .001
$\sigma^2 = 10^{-3}$	.0	.899 ± .026	.899 ± .027	.906 ± .026	.904 ± .026	.904 ± .026	.905 ± .025
	.1	.178 ± .006	.170 ± .006	.181 ± .006	.178 ± .006	.177 ± .006	.179 ± .006
	.2	.164 ± .006	.159 ± .006	.169 ± .006	.167 ± .006	.166 ± .006	.167 ± .006
	.3	.143 ± .007	.138 ± .007	.148 ± .007	.146 ± .007	.145 ± .007	.145 ± .007
	.4	.133 ± .005	.129 ± .005	.140 ± .005	.137 ± .005	.137 ± .005	.137 ± .005
	.5	.122 ± .004	.117 ± .004	.129 ± .004	.126 ± .004	.125 ± .004	.125 ± .004
	.6	.111 ± .003	.104 ± .003	.117 ± .003	.114 ± .003	.113 ± .003	.112 ± .003
	.7	.109 ± .003	.103 ± .003	.118 ± .003	.114 ± .003	.113 ± .003	.112 ± .003
	.8	.108 ± .004	.102 ± .004	.120 ± .004	.115 ± .004	.114 ± .004	.113 ± .004
	.9	.103 ± .003	.099 ± .002	.124 ± .003	.118 ± .003	.116 ± .003	.116 ± .003

Table F.13. *Comparison of the bound values before performing Step* **2)** *of our Training Method for* `ours`, `rivasplata`, `blanchard` *and* `catoni`. *More precisely, for each split and each variance* $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, *we report the mean* $\pm$ *the standard deviation (for* 400 *neural networks sampled from* π*) of the test risk (*$R_{\mathcal{T}}(h)$*), the empirical risk (*$R_{\mathcal{S}}(h)$*), and the value of the bounds of Corollaries 6.4.1 and 6.4.2. We consider in this table that the dataset is CIFAR-10.*

	Split	$R_{\mathcal{T}}(h)$	$R_{\mathcal{S}}(h)$	Cor. 6.4.1	Eq. (6.1)	Eq. (6.2)	Eq. (6.3)
$\sigma^2 = 10^{-6}$	.0	.899 ± .000	.899 ± .000	.906 ± .000	.904 ± .000	.903 ± .000	.904 ± .000
	.1	.476 ± .000	.470 ± .000	.486 ± .000	.482 ± .000	.481 ± .000	.485 ± .000
	.2	.390 ± .000	.389 ± .000	.406 ± .000	.402 ± .000	.401 ± .000	.404 ± .000
	.3	.370 ± .000	.358 ± .000	.374 ± .000	.371 ± .000	.370 ± .000	.372 ± .000
	.4	.334 ± .000	.328 ± .000	.346 ± .000	.342 ± .000	.341 ± .000	.342 ± .000
	.5	.307 ± .000	.302 ± .000	.321 ± .000	.317 ± .000	.316 ± .000	.317 ± .000
	.6	.274 ± .000	.276 ± .000	.297 ± .000	.293 ± .000	.291 ± .000	.291 ± .000
	.7	.275 ± .000	.272 ± .000	.296 ± .000	.290 ± .000	.289 ± .000	.288 ± .000
	.8	.249 ± .000	.237 ± .000	.265 ± .000	.259 ± .000	.257 ± .000	.256 ± .000
	.9	.227 ± .000	.230 ± .000	.269 ± .000	.260 ± .000	.258 ± .000	.258 ± .000
$\sigma^2 = 10^{-5}$	.0	.899 ± .001	.899 ± .000	.906 ± .000	.904 ± .000	.904 ± .000	.904 ± .000
	.1	.476 ± .000	.478 ± .000	.494 ± .000	.490 ± .000	.489 ± .000	.493 ± .000
	.2	.403 ± .000	.398 ± .000	.414 ± .000	.410 ± .000	.409 ± .000	.412 ± .000
	.3	.349 ± .000	.350 ± .000	.367 ± .000	.363 ± .000	.362 ± .000	.364 ± .000
	.4	.322 ± .000	.313 ± .000	.330 ± .000	.327 ± .000	.326 ± .000	.327 ± .000
	.5	.281 ± .000	.283 ± .000	.302 ± .000	.298 ± .000	.297 ± .000	.297 ± .000
	.6	.290 ± .000	.286 ± .000	.307 ± .000	.303 ± .000	.301 ± .000	.301 ± .000
	.7	.266 ± .000	.257 ± .000	.281 ± .000	.276 ± .000	.274 ± .000	.274 ± .000
	.8	.247 ± .000	.243 ± .000	.271 ± .000	.265 ± .000	.263 ± .000	.262 ± .000
	.9	.236 ± .000	.227 ± .000	.266 ± .000	.257 ± .000	.255 ± .000	.255 ± .000

	Split	$R_{\mathcal{T}}(h)$	$R_{\mathcal{S}}(h)$	Cor. 6.4.1	Eq. (6.1)	Eq. (6.2)	Eq. (6.3)
$\sigma^2 = 10^{-4}$	.0	.900 ± .004	.900 ± .003	.907 ± .003	.905 ± .003	.905 ± .003	.905 ± .003
	.1	.458 ± .001	.464 ± .001	.479 ± .001	.476 ± .001	.475 ± .001	.478 ± .001
	.2	.395 ± .001	.396 ± .000	.412 ± .000	.409 ± .000	.408 ± .000	.411 ± .000
	.3	.361 ± .001	.361 ± .000	.378 ± .000	.375 ± .000	.373 ± .000	.376 ± .000
	.4	.323 ± .001	.316 ± .000	.334 ± .000	.330 ± .000	.329 ± .000	.331 ± .000
	.5	.296 ± .001	.291 ± .000	.310 ± .000	.306 ± .000	.304 ± .000	.305 ± .000
	.6	.271 ± .001	.263 ± .000	.284 ± .000	.279 ± .000	.278 ± .000	.278 ± .000
	.7	.253 ± .001	.246 ± .000	.270 ± .000	.265 ± .000	.263 ± .000	.262 ± .000
	.8	.259 ± .001	.252 ± .001	.281 ± .001	.275 ± .001	.273 ± .001	.272 ± .001
	.9	.217 ± .000	.216 ± .001	.255 ± .001	.246 ± .001	.243 ± .001	.244 ± .001
$\sigma^2 = 10^{-3}$	.0	.905 ± .012	.904 ± .012	.911 ± .011	.909 ± .011	.909 ± .011	.909 ± .011
	.1	.479 ± .002	.480 ± .001	.496 ± .001	.493 ± .001	.491 ± .001	.495 ± .001
	.2	.415 ± .002	.415 ± .001	.432 ± .001	.428 ± .001	.427 ± .001	.430 ± .001
	.3	.417 ± .001	.416 ± .001	.434 ± .001	.430 ± .001	.429 ± .001	.431 ± .001
	.4	.333 ± .001	.323 ± .001	.341 ± .001	.337 ± .001	.336 ± .001	.338 ± .001
	.5	.316 ± .001	.311 ± .001	.331 ± .001	.327 ± .001	.325 ± .001	.326 ± .001
	.6	.280 ± .001	.281 ± .001	.302 ± .001	.298 ± .001	.296 ± .001	.296 ± .001
	.7	.239 ± .001	.234 ± .001	.257 ± .001	.252 ± .001	.250 ± .001	.250 ± .001
	.8	.249 ± .001	.245 ± .001	.274 ± .001	.268 ± .001	.266 ± .001	.264 ± .001
	.9	.233 ± .001	.232 ± .002	.272 ± .002	.263 ± .002	.260 ± .002	.260 ± .002

Appendix of Chapter 7

G

G.1 Proof of Theorem 7.3.1

Theorem 7.3.1 (Generalization Bound with Complexity Measures)**.** Let $\phi : [0,1]^2 \to \mathbb{R}$ be the generalization gap. For any $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any $\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,\ h'\sim\pi,\ h\sim\rho_{\mathbb{S}}} \left[\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \left[\alpha \mathsf{R}_{\mathcal{S}}(h') + \mu(h', \mathbb{S})\right] - \left[\alpha \mathsf{R}_{\mathcal{S}}(h) + \mu(h, \mathbb{S})\right] + \ln \frac{\pi(h')}{\pi(h)} + \ln \left(\frac{4}{\delta^2} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} \exp\left[\phi(\mathsf{R}_{\mathcal{D}}(h'), \mathsf{R}_{\mathcal{S}'}(h'))\right] \right) \right] \geq 1 - \delta.$$

Proof. First of all, we denote as $Z = \int_{\mathbb{H}} \exp\left[-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})\right] d\xi(h')$, the normalization constant of the Gibbs distribution $\rho_{\mathbb{S}}$ and ξ the reference measure on $\mathbb{H}$. Moreover, we have

$$\rho_{\mathbb{S}}(h) = \frac{1}{Z} \exp\left[-\alpha \mathsf{R}_{\mathcal{S}}(h) - \mu(h, \mathbb{S})\right] \propto \exp\left[-\alpha \mathsf{R}_{\mathcal{S}}(h) - \mu(h, \mathbb{S})\right].$$

We apply Theorem 2.4.1 with $\frac{\delta}{2}$ instead of δ and with the function $\varphi(h, \mathbb{S}) = \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h))$ to obtain with probability at least $1 - \frac{\delta}{2}$ over $\mathbb{S} \sim \mathcal{D}^m$ and $h \sim \rho_{\mathbb{S}}$

$$\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \ln \left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)} \right] + \ln \left[\frac{2}{\delta} \mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} e^{\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}'}(h'))} \right].$$

We develop the term $\ln\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right]$ in Theorem 2.4.1. We have

$$\begin{aligned}
\ln\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)}\right] &= \ln\left(\frac{\exp\left[-\alpha \mathsf{R}_{\mathcal{S}}(h) - \mu(h, \mathbb{S})\right]}{Z}\frac{1}{\pi(h)}\right) \\
&= \ln\left(\exp\left[-\alpha \mathsf{R}_{\mathcal{S}}(h) - \mu(h, \mathbb{S})\right]\right) \\
&\quad - \ln\left(\pi(h)\int_{\mathbb{H}} \exp\left[-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})\right] d\xi(h')\right) \\
&= -\alpha \mathsf{R}_{\mathcal{S}}(h) - \mu(h, \mathbb{S}) \\
&\quad - \ln\left(\pi(h)\int_{\mathbb{H}} \frac{\pi(h')}{\pi(h')} \exp\left[-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})\right] d\xi(h')\right) \\
&= -\alpha \mathsf{R}_{\mathcal{S}}(h) - \mu(h, \mathbb{S}) - \ln\left(\mathop{\mathbb{E}}_{h'\sim\pi} \frac{\pi(h)}{\pi(h')} e^{-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})}\right).
\end{aligned}$$

Hence, we obtain the following

$$\begin{aligned}
\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\Bigg[\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \ln\Bigg[\frac{2}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi} e^{\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}'}(h'))}\Bigg] \\
- \alpha \mathsf{R}_{\mathcal{S}}(h) - \mu(h, \mathbb{S}) - \ln\left(\mathbb{E}_{h'\sim\pi} \tfrac{\pi(h)}{\pi(h')} e^{-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})}\right)\Bigg] \geq 1 - \frac{\delta}{2}. \quad \text{(G.1)}
\end{aligned}$$

We can now upper-bound the term $-\ln\left(\mathbb{E}_{h'\sim\pi} \frac{\pi(h)}{\pi(h')} e^{-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})}\right)$. To do so, since $\frac{\pi(h)}{\pi(h')} e^{-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})} > 0$ for all $h \in \mathbb{H}$ and $\mathbb{S} \in (\mathbb{X} \times \mathbb{Y})^m$, we apply MARKOV's inequality (Theorem A.2.1) to obtain for all $h \in \mathbb{H}$ and $\mathbb{S} \in (\mathbb{X} \times \mathbb{Y})^m$ with probability at least $1 - \frac{\delta}{2}$ over $h' \sim \pi$

$$\begin{aligned}
&\frac{\pi(h)}{\pi(h')} e^{-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})} \leq \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\pi}\left(\frac{\pi(h)}{\pi(h')} e^{-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})}\right) \\
\iff\ &-\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi}\left[\frac{\pi(h)}{\pi(h')} e^{-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})}\right]\right) \leq \ln\frac{2}{\delta} - \ln\left(\frac{\pi(h)}{\pi(h')} e^{-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})}\right).
\end{aligned}$$

Moreover, by simplifying the right-hand side of the inequality, we have

$$-\ln\left(\frac{\pi(h)}{\pi(h')} e^{-\alpha \mathsf{R}_{\mathcal{S}}(h') - \mu(h', \mathbb{S})}\right) = \ln\frac{\pi(h')}{\pi(h)} + \alpha \mathsf{R}_{\mathcal{S}}(h') + \mu(h', \mathbb{S}).$$

Hence, we obtain the following inequality:

$$\mathop{\mathbb{P}}_{h'\sim\pi}\left[-\ln\left(\frac{\pi(h)}{\pi(h')}e^{-\alpha \mathsf{R}_{\mathcal{S}}(h')-\mu(h',\mathbb{S})}\right)\leq \ln\frac{2}{\delta}+\ln\frac{\pi(h')}{\pi(h)}+\alpha \mathsf{R}_{\mathcal{S}}(h')+\mu(h',\mathbb{S})\right]\geq 1-\frac{\delta}{2}. \tag{G.2}$$

By using an union bound on Equations (G.1) and (G.2) an rearranging the terms, we obtain the claimed result. ■

G.2 Proof of Corollary 7.3.1

Corollary 7.3.1 (Practical Generalization Bound with Complexity Measures)**.** For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any bounded hypothesis set $\mathbb{H}$, given the uniform prior distribution π on $\mathbb{H}$, for any $\mu : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m\to\mathbb{R}$, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over $\mathbb{S}\sim\mathcal{D}^m$, $h'\sim\pi$, $h\sim\rho_{\mathbb{S}}$ we have

$$\mathrm{kl}\left[\mathsf{R}_{\mathcal{S}}(h)\|\mathsf{R}_{\mathcal{D}}(h)\right]\leq\frac{1}{m}\left[\left[\alpha\mathsf{R}_{\mathcal{S}}(h')+\mu(h',\mathbb{S})\right]-\left[\alpha\mathsf{R}_{\mathcal{S}}(h)+\mu(h,\mathbb{S})\right]+\frac{8\sqrt{m}}{\delta^2}\right]_+, \tag{7.5}$$

$$\left|\mathsf{R}_{\mathcal{D}}(h)-\mathsf{R}_{\mathcal{S}}(h)\right|\leq\sqrt{\frac{1}{2m}\left[\left[\alpha\mathsf{R}_{\mathcal{S}}(h')+\mu(h',\mathbb{S})\right]-\left[\alpha\mathsf{R}_{\mathcal{S}}(h)+\mu(h,\mathbb{S})\right]+\frac{8\sqrt{m}}{\delta^2}\right]_+}, \tag{7.6}$$

where $[a]_+=\max(0,a)$, and $\rho_{\mathbb{S}}$ is the Gibbs distribution defined by Equation (7.2).

Proof. Since π is the uniform distribution we have: $\mathbb{E}_{h'\sim\pi}\ln\frac{\pi(h)}{\pi(h')}=0$. We instantiate Th. 7.3.1 with $\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h))=m\,\mathrm{kl}\left[\mathsf{R}_{\mathcal{S}}(h)\|\mathsf{R}_{\mathcal{D}}(h)\right]$. It remains to upper-bound $\mathbb{E}_{\mathbb{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\pi}\exp\left(m\,\mathrm{kl}\left[\mathsf{R}_{\mathcal{S}'}(h')\|\mathsf{R}_{\mathcal{D}}(h')\right]\right)$. We have

$$\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}e^{m\,\mathrm{kl}[\mathsf{R}_{\mathcal{S}'}(h')\|\mathsf{R}_{\mathcal{D}}(h')]}=\mathop{\mathbb{E}}_{h'\sim\pi}\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}e^{m\,\mathrm{kl}[\mathsf{R}_{\mathcal{S}'}(h')\|\mathsf{R}_{\mathcal{D}}(h')]} \tag{G.3}$$

$$\mathop{\mathbb{E}}_{\mathbb{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}e^{m\,\mathrm{kl}[\mathsf{R}_{\mathcal{S}'}(h')\|\mathsf{R}_{\mathcal{D}}(h')]}\leq 2\sqrt{m}, \tag{G.4}$$

where Equation (G.3) is due to Fubini's theorem (*i.e.*, we can exchange the two expectations), and Equation (G.4) is due to Maurer (2004) (see Lemma B.16.1). By rearranging the terms, with probability at least $1-\delta$ over $\mathbb{S}\sim\mathcal{D}^m$, $h\sim\rho_{\mathbb{S}}$, and

$h' \sim \pi$ we have

$$\mathrm{kl}\left[\mathsf{R}_{\mathcal{S}}(h) \| \mathsf{R}_{\mathcal{D}}(h)\right] \leq \frac{1}{m}\left[\left[\alpha \mathsf{R}_{\mathcal{S}}(h') + \mu(h', \mathbb{S})\right] - \left[\alpha \mathsf{R}_{\mathcal{S}}(h) + \mu(h, \mathbb{S})\right] + \ln \frac{8\sqrt{m}}{\delta^2}\right].$$

Hence, by definition of $[a]_+$, we can deduce Equation (7.5). From PINSKER's inequality (Theorem B.5.1), we have

$$2(\mathsf{R}_{\mathcal{D}}(h) - \mathsf{R}_{\mathcal{S}}(h))^2 \leq \mathrm{kl}\left[\mathsf{R}_{\mathcal{S}}(h) \| \mathsf{R}_{\mathcal{D}}(h)\right].$$

Hence, thanks to this inequality and by rearranging the terms, we obtain Equation (7.6). ■

G.3 Proof of Proposition 7.5.1

Proposition 7.5.1 (Set-theoretic view of Definition 7.3.1)**.** Let $\phi : [0,1]^2 \to \mathbb{R}$ be a generalization gap and assume that there exists a function $\Phi_\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \times (0,1] \to \mathbb{R}$ fulfilling Definition 7.3.1. Under these conditions, with $\mathbb{Z}_{\mathtt{d}} = \left\{(h, \mathbb{S}) \in \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m : \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_\mu(h, \mathbb{S}, \delta)\right\}$, and $\mathbb{P}_{\mathbb{S} \sim \mathcal{D}^m, h \sim \rho_{\mathbb{S}}}\left[(h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}\right] \geq 1 - \delta$, we have

$$\begin{aligned}\text{Equation (7.1)} &\iff \forall (h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}},\ \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_\mu(h, \mathbb{S}, \delta) \\ &\iff \sup_{(h,\mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}} \left\{\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) - \Phi_\mu(h, \mathbb{S}, \delta)\right\} \leq 0.\end{aligned}$$

Proof. First of all, by definition of the supremum, we have

$$\begin{aligned}\forall \delta \in (0,1], \quad &\forall (h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}, \quad \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_\mu(h, \mathbb{S}, \delta) \\ &\forall (h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}, \quad \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) - \Phi_\mu(h, \mathbb{S}, \delta) \leq 0 \\ &\iff \sup_{(h,\mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}} \left\{\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) - \Phi_\mu(h, \mathbb{S}, \delta)\right\} \leq 0.\end{aligned}$$

It remains to prove that

$$\underbrace{\text{Equation (7.1)}}_{\text{(A)}} \iff \underbrace{\forall (h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}, \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_\mu(h, \mathbb{S}, \delta)}_{\text{(B)}}$$

to complete the proof.

Step 1 ((A) $\Rightarrow$ (B)). Assuming that (A) holds, by definition of $\mathbb{Z}_{\mathtt{d}}$, we have

$$\begin{aligned} &\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mu}(h, \mathbb{S}, \delta)\right] \geq 1-\delta \\ \iff &\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[(h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}\right] \geq 1-\delta, \end{aligned}$$

since $\mathrm{I}\left[\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mu}(h, \mathbb{S}, \delta)\right] = \mathrm{I}\left[(h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}\right]$.
Additionally, by definition of $\mathbb{Z}_{\mathtt{d}}$, we know that

$$\forall (h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}, \quad \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mu}(h, \mathbb{S}, \delta),$$

where $\mathbb{P}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[(h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}\right] \geq 1-\delta$.
Step 2 ((A) $\Leftarrow$ (B)). Note that from the definition of $\mathbb{Z}_{\mathtt{d}}$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[(h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}\right] \geq 1-\delta.$$

Additionally, we can deduce that

$$\begin{aligned} &\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[(h, \mathbb{S}) \in \mathbb{Z}_{\mathtt{d}}\right] \geq 1-\delta \\ \iff &\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mu}(h, \mathbb{S}, \delta)\right] \geq 1-\delta. \end{aligned}$$

■

G.4 Proof of Proposition 7.5.2

Proposition 7.5.2 (Set-theoretic View of Uniform Convergence Bounds)**.** Let $\phi : [0,1]^2 \to \mathbb{R}$ be a generalization gap and assume that there exists a function $\Phi_{\mathtt{u}} : (0,1] \to \mathbb{R}$ fulfilling Definition 1.3.2. Under these conditions, with $\mathbb{Z}_{\mathtt{u}} = \left\{\mathbb{S} \in (\mathbb{X}\times\mathbb{Y})^m : \forall h \in \mathbb{H},\ \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{u}}(\delta)\right\}$, and $\mathbb{P}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathbb{S}\in\mathbb{Z}_{\mathtt{u}}\right] \geq 1-\delta$, we have

$$\begin{aligned} \text{Equation (1.1)} &\iff \forall \mathbb{S} \in \mathbb{Z}_{\mathtt{u}},\ \forall h \in \mathbb{H},\ \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{u}}(\delta) \\ &\iff \sup_{\mathbb{S}\in\mathbb{Z}_{\mathtt{u}}} \sup_{h\in\mathbb{H}} \left\{\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h))\right\} \leq \Phi_{\mathtt{u}}(\delta). \end{aligned}$$

Proof. First of all, by definition of the supremum, we have

$$\begin{aligned}\forall \delta \in (0,1], \quad \forall \mathbb{S} \in \mathbb{Z}_{\mathtt{u}},\ \forall h \in \mathbb{H},\ \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{u}}(\delta) \\ \iff \sup_{\mathbb{S}\in\mathbb{Z}_{\mathtt{u}}} \sup_{h\in\mathbb{H}} \left\{\phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h))\right\} \leq \Phi_{\mathtt{u}}(\delta).\end{aligned}$$

It remains to prove that

$$\underbrace{\text{Equation (1.1)}}_{\text{(A)}} \iff \underbrace{\forall \mathbb{S} \in \mathbb{Z}_{\mathtt{u}},\ \forall h \in \mathbb{H},\ \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{u}}(\delta)}_{\text{(B)}}$$

to complete the proof.
Step 1 ((A) $\Rightarrow$ (B)). Assuming that (A) holds, by definition of $\mathbb{Z}_{\mathtt{u}}$, we have

$$\begin{aligned}&\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall h \in \mathbb{H},\ \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{u}}(\delta)\right] \geq 1-\delta \\ \iff &\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathbb{S} \in \mathbb{Z}_{\mathtt{u}}\right] \geq 1-\delta,\end{aligned}$$

since $\mathrm{I}\left[\forall h \in \mathbb{H},\ \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{u}}(\delta)\right] = \mathrm{I}\left[\mathbb{S} \in \mathbb{Z}_{\mathtt{u}}\right]$.
Additionally, by definition of $\mathbb{Z}_{\mathtt{u}}$, we know that

$$\forall \mathbb{S} \in \mathbb{Z}_{\mathtt{u}},\ \forall h \in \mathbb{H}, \quad \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{u}}(\delta),$$

where $\mathbb{P}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathbb{S} \in \mathbb{Z}_{\mathtt{u}}\right] \geq 1-\delta$.
Step 2 ((A) $\Leftarrow$ (B)). Note that from the definition of $\mathbb{Z}_{\mathtt{u}}$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathbb{S} \in \mathbb{Z}_{\mathtt{u}}\right] \geq 1-\delta.$$

Additionally, we can deduce that

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathbb{S} \in \mathbb{Z}_{\mathtt{u}}\right] \geq 1-\delta \iff \mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall h \in \mathbb{H},\ \phi(\mathsf{R}_{\mathcal{D}}(h), \mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{u}}(\delta)\right] \geq 1-\delta.$$

■

G.5 Proof of Corollary 7.5.1

Corollary 7.5.1 (Uniform Convergence Bound from Theorem 7.3.1)**.** Let $\phi : [0,1]^2 \to \mathbb{R}$ be the generalization gap and assume that there exists a function $\Phi_{\mathtt{u}} : (0,1] \to \mathbb{R}$ fulfilling Definition 1.3.2 such that $\Phi_{\mathtt{u}}(\delta) \geq$

$\ln\left[\frac{4}{\delta^2}\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{h'\sim\pi}\exp\left(\phi(\mathsf{R}_{\mathcal{D}}(h'),\mathsf{R}_{\mathcal{S}'}(h'))\right)\right]$. For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any $\delta\in(0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\rho_{\mathcal{S}}}\left[\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h))\leq\Phi_{\mathtt{u}}(\delta)\right]\geq 1-\delta.$$

Proof. Let the parametric function $\mu()$ defined as

$$\forall(h,\mathcal{S})\in\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m,\quad \mu(h,\mathcal{S})=-\alpha\mathsf{R}_{\mathcal{S}}(h)-\ln\pi(h)+\Phi_{\mathtt{u}}\Big(\delta\Big).$$

Given the definition of $\rho_{\mathcal{S}}$ (with the parametric function $\mu()$ defined above), we can deduce from Theorem 7.3.1 that

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\ h'\sim\pi,h\sim\rho_{\mathcal{S}}}\left[\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h))\leq\ln\left(\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi}\exp\left[\phi(\mathsf{R}_{\mathcal{D}}(g),\mathsf{R}_{\mathcal{S}'}(g))\right]\right)\right]$$

$$=\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\ h\sim\rho_{\mathcal{S}}}\left[\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h))\leq\ln\left(\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi}\exp\left[\phi(\mathsf{R}_{\mathcal{D}}(g),\mathsf{R}_{\mathcal{S}'}(g))\right]\right)\right]\geq 1-\delta.$$

Note that the equality holds since $h'\sim\pi$ does not appear in the bound. If the assumption $\Phi_{\mathtt{u}}(\delta)\geq\ln\left[\frac{4}{\delta^2}\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{g\sim\pi}\exp\left(\phi(\mathsf{R}_{\mathcal{D}}(g),\mathsf{R}_{\mathcal{S}'}(g))\right)\right]$ is satisfied, then, we can deduce Corollary 7.5.1. ■

G.6 Proof of Proposition 7.5.3

Proposition 7.5.3 (Set-theoretic View of Algorithmic Dependent Bounds)**.** Let $\phi:[0,1]^2\to\mathbb{R}$ be a generalization gap and assume that there exists a function $\Phi_{\mathtt{a}}:(0,1]\to\mathbb{R}$ fulfilling Definition 1.3.5. Under these conditions, with $\mathbb{Z}_{\mathtt{a}}=\left\{\mathbb{S}\in(\mathbb{X}\times\mathbb{Y})^m:\phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}),\mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}}))\leq\Phi_{\mathtt{a}}(\delta)\right\}$ and $\mathbb{P}_{\mathbb{S}\sim\mathcal{D}^m}[\mathbb{S}\in\mathbb{Z}_{\mathtt{a}}]\geq 1-\delta$, we have

$$\begin{aligned}\text{Equation (1.2)}&\iff\forall\mathbb{S}\in\mathbb{Z}_{\mathtt{a}},\phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}),\mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}}))\leq\Phi_{\mathtt{a}}(\delta)\\&\iff\sup_{\mathbb{S}\in\mathbb{Z}_{\mathtt{a}}}\phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}),\mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}}))\leq\Phi_{\mathtt{a}}(\delta).\end{aligned}$$

Proof. First of all, by definition of the supremum, we have

$$\forall \delta \in (0,1], \quad \forall \mathbb{S} \in \mathbb{Z}_{\mathsf{a}}, \phi(h_{\mathbb{S}}, \mathbb{S}) \leq \Phi_{\mathsf{a}}(\delta) \iff \sup_{\mathbb{S} \in \mathbb{Z}_{\mathsf{a}}} \phi(h_{\mathbb{S}}, \mathbb{S}) \leq \Phi_{\mathsf{a}}(\delta).$$

It remains to prove that

$$\underbrace{\text{Equation (1.2)}}_{\text{(A)}} \iff \underbrace{\forall \mathbb{S} \in \mathbb{Z}_{\mathsf{a}}, \phi(\mathsf{R}_{\mathcal{D}}(h_{\mathbb{S}}), \mathsf{R}_{\mathcal{S}}(h_{\mathbb{S}})) \leq \Phi_{\mathsf{a}}(\delta)}_{\text{(B)}}$$

to complete the proof.
Step 1 ((A) $\Rightarrow$ (B)). Assuming that (A) holds, by definition of $\mathbb{Z}_{\mathsf{a}}$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m} \left[\phi(h_{\mathbb{S}}, \mathbb{S}) \leq \Phi_{\mathsf{a}}(\delta) \right] \geq 1 - \delta \iff \mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m} \left[\mathbb{S} \in \mathbb{Z}_{\mathsf{a}} \right] \geq 1 - \delta.$$

since $\mathrm{I}\left[\phi(h_{\mathbb{S}}, \mathbb{S}) \leq \Phi_{\mathsf{a}}(\delta)\right] = \mathrm{I}\left[\mathbb{S} \in \mathbb{Z}_{\mathsf{a}}\right]$.
Additionally, by definition of $\mathbb{Z}_{\mathsf{a}}$, we know that

$$\forall \mathbb{S} \in \mathbb{Z}_{\mathsf{a}}, \quad \phi(h_{\mathbb{S}}, \mathbb{S}) \leq \Phi_{\mathsf{a}}(\delta), \quad \text{where} \quad \mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m} \left[\mathbb{S} \in \mathbb{Z}_{\mathsf{a}} \right] \geq 1 - \delta.$$

Step 2 ((A) $\Leftarrow$ (B)). Note that from the definition of $\mathbb{Z}_{\mathsf{a}}$ we have

$$\mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m} \left[\mathbb{S} \in \mathbb{Z}_{\mathsf{a}} \right] \geq 1 - \delta.$$

Additionally, we can deduce that

$$\mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m} \left[\mathbb{S} \in \mathbb{Z}_{\mathsf{a}} \right] \geq 1 - \delta \iff \mathop{\mathbb{P}}_{\mathbb{S} \sim \mathcal{D}^m} \left[\phi(h_{\mathbb{S}}, \mathbb{S}) \leq \Phi_{\mathsf{a}}(\delta) \right] \geq 1 - \delta.$$

■

G.7 Proof of Corollary 7.5.2

Corollary 7.5.2 (Algorithmic-dependent Bound from Theorem 7.3.1)**.** Let $\phi : [0,1]^2 \to \mathbb{R}$ be the generalization gap and assume that there exists a function $\Phi_{\mathsf{a}} : (0,1] \to \mathbb{R}$ fulfilling Definition 1.3.5 such that $\Phi_{\mathsf{a}}(\delta) \geq \ln\left[\frac{4}{\delta^2} \mathbb{E}_{\mathbb{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \pi} \exp\left(\phi(\mathsf{R}_{\mathcal{D}}(h'), \mathsf{R}_{\mathcal{S}'}(h'))\right)\right]$. For any $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hy-

pothesis set $\mathbb{H}$, for any $\delta \in (0,1]$, we have

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}}\left[\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{a}}(\delta)\right] \geq 1-\delta.$$

Proof. Let the parametric function $\mu()$ defined as

$$\forall (h,\mathcal{S}) \in \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m, \quad \mu(h,\mathcal{S}) = -\alpha\mathsf{R}_{\mathcal{S}}(h) - \ln \pi(h) + \Phi_{\mathtt{a}}\Big(\delta\Big).$$

Given the definition of $\rho_{\mathbb{S}}$ (with the parametric function $\mu()$ defined above), we can deduce from Theorem 7.3.1 that

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\ h'\sim\pi, h\sim\rho_{\mathbb{S}}}\left[\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h)) \leq \ln\left(\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi}\exp\left[\phi(\mathsf{R}_{\mathcal{D}}(g),\mathsf{R}_{\mathcal{S}'}(g))\right]\right)\right]$$

$$= \mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\ h\sim\rho_{\mathbb{S}}}\left[\phi(\mathsf{R}_{\mathcal{D}}(h),\mathsf{R}_{\mathcal{S}}(h)) \leq \ln\left(\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi}\exp\left[\phi(\mathsf{R}_{\mathcal{D}}(g),\mathsf{R}_{\mathcal{S}'}(g))\right]\right)\right] \geq 1-\delta.$$

Note that the equality holds since $h' \sim \pi$ does not appear in the bound. If the assumption $\Phi_{\mathtt{a}}(\delta) \geq \ln\Big[\frac{4}{\delta^2}\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{g\sim\pi}\exp\left(\phi(\mathsf{R}_{\mathcal{D}}(g),\mathsf{R}_{\mathcal{S}'}(g))\right)\Big]$ is satisfied, then, we can deduce Corollary 7.5.2. ■

G.8 Details on the Experiments

In this section, we introduce additional figures concerning the tightness, the influence of α, and the influence of the number of parameters. Additionally, we provide more experiments with data-dependent complexity measures that we present in Appendix G.8.1.

G.8.1 Data-dependent Complexity Measures $\Phi_\mu(h, \mathbb{S}, \delta)$

As we have pointed out in the paper, the parametric function $\mu()$ depends on the learning sample $\mathbb{S}$. We illustrate this dependence with other parametric functions defined as

$$\begin{aligned}
\textsc{Dist_Fro}{-}\textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}) &= \textsc{Dist_Fro}(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}),\\
\textsc{Dist_L}_2{-}\textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}) &= \textsc{Dist_L}_2(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}),\\
\textsc{Param_Norm}{-}\textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}) &= \textsc{Param_Norm}(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}),\\
\textsc{Path_Norm}{-}\textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}) &= \textsc{Path_Norm}(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}),\\
\textsc{Sum_Fro}{-}\textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}) &= \textsc{Sum_Fro}(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}),\\
\textsc{Zero}{-}\textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}) &= \textsc{Zero}{-}\textsc{Aug}(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathbb{S}),
\end{aligned}$$

where

$$\textsc{Aug}(h, \mathbb{S}) = -\frac{1}{2}\mathsf{R}_{\mathcal{S}}(h) + \frac{1}{2}\mathsf{R}_{\widehat{\mathcal{S}}}(h),$$

and $\widehat{\mathbb{S}}$ is a data-augmented learning sample. More precisely, we apply to each example $(\mathbf{x}, y) \in \mathbb{S}$ *(a)* a random rotation (with a maximum angle set to $20°$) and *(b)* a random translation (with a maximum of 3 translated pixels per dimension).

G.8.2 Tightness of the Bounds

Figures G.1 and G.2 report the tightness of the bounds for the data-dependent complexity measures introduced in Appendix G.8.1.

G.8.3 Influence of the Parameter α

Figures G.3 to G.6 shows the influence of the parameter α for all parametric functions.

G.8.4 Influence of the Depth/Width

Figures G.7 to G.10 shows the influence of the depth and the width for all parametric functions.

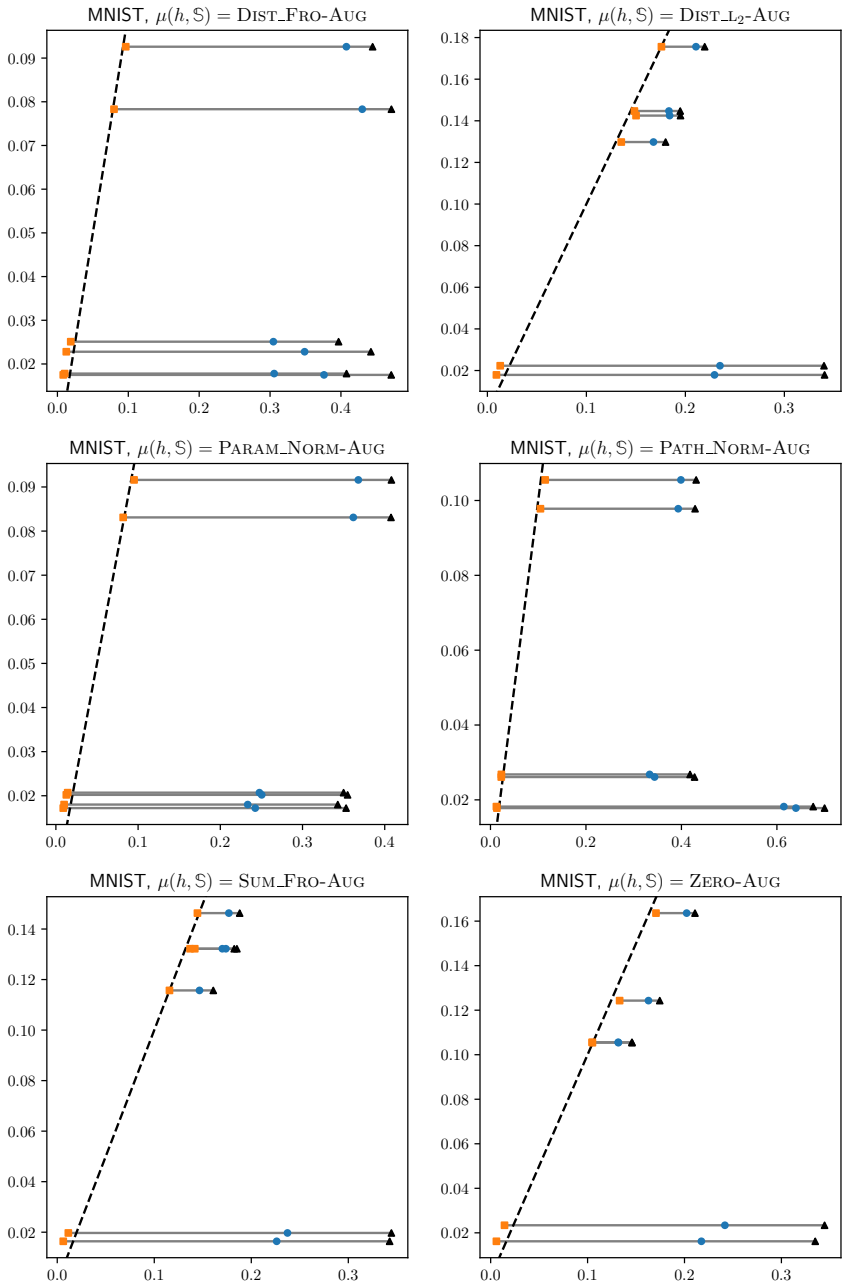


Figure G.1. *Scatter plot given a parametric function $\mu(h, \mathbb{S})$, where each segment represents a neural network $h_{\mathbf{w}}$ learned with a given α, width H and depth L. For each segment, there is a corresponding orange square and a blue circle. The orange squares corresponds to the empirical risk $R_{\mathcal{S}}(h)$ (x-axis) and test risk $R_{\mathcal{T}}(h)$ (y-axis). The blue circle resp. the black triangle represents Equation (7.7) resp. Equation (7.8) in the x-axis and the test risk $R_{\mathcal{T}}(h)$ in the y-axis. The dashed line is the identity function.*

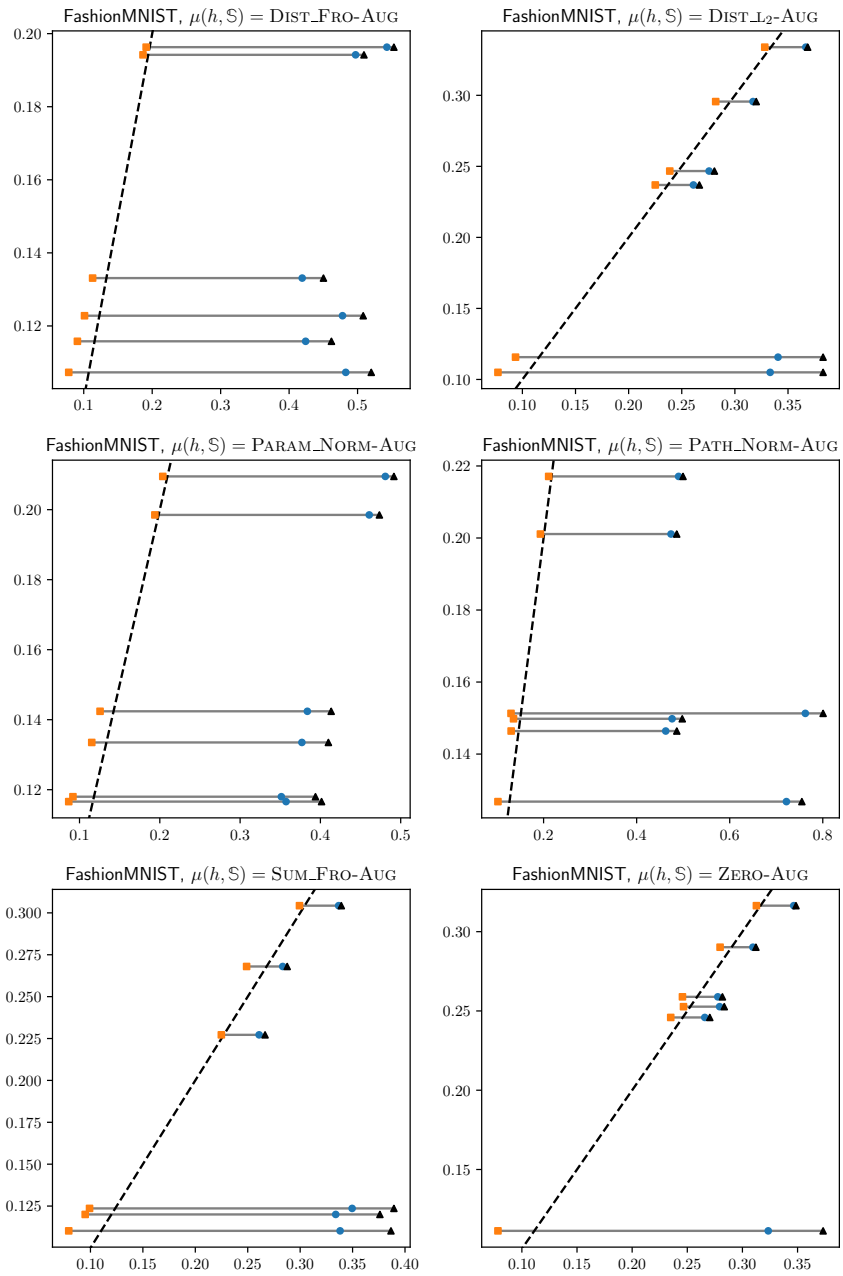


Figure G.2. *Scatter plot given a parametric function $\mu(h, \mathbb{S})$, where each segment represents a neural network $h_{\mathbf{w}}$ learned with a given α, width H and depth L. For each segment, there is a corresponding orange square and a blue circle. The orange squares corresponds to the empirical risk $R_{\mathcal{S}}(h)$ (x-axis) and test risk $R_{\mathcal{T}}(h)$ (y-axis). The blue circle resp. the black triangle represents Equation (7.7) resp. Equation (7.8) in the x-axis and the test risk $R_{\mathcal{T}}(h)$ in the y-axis. The dashed line is the identity function.*

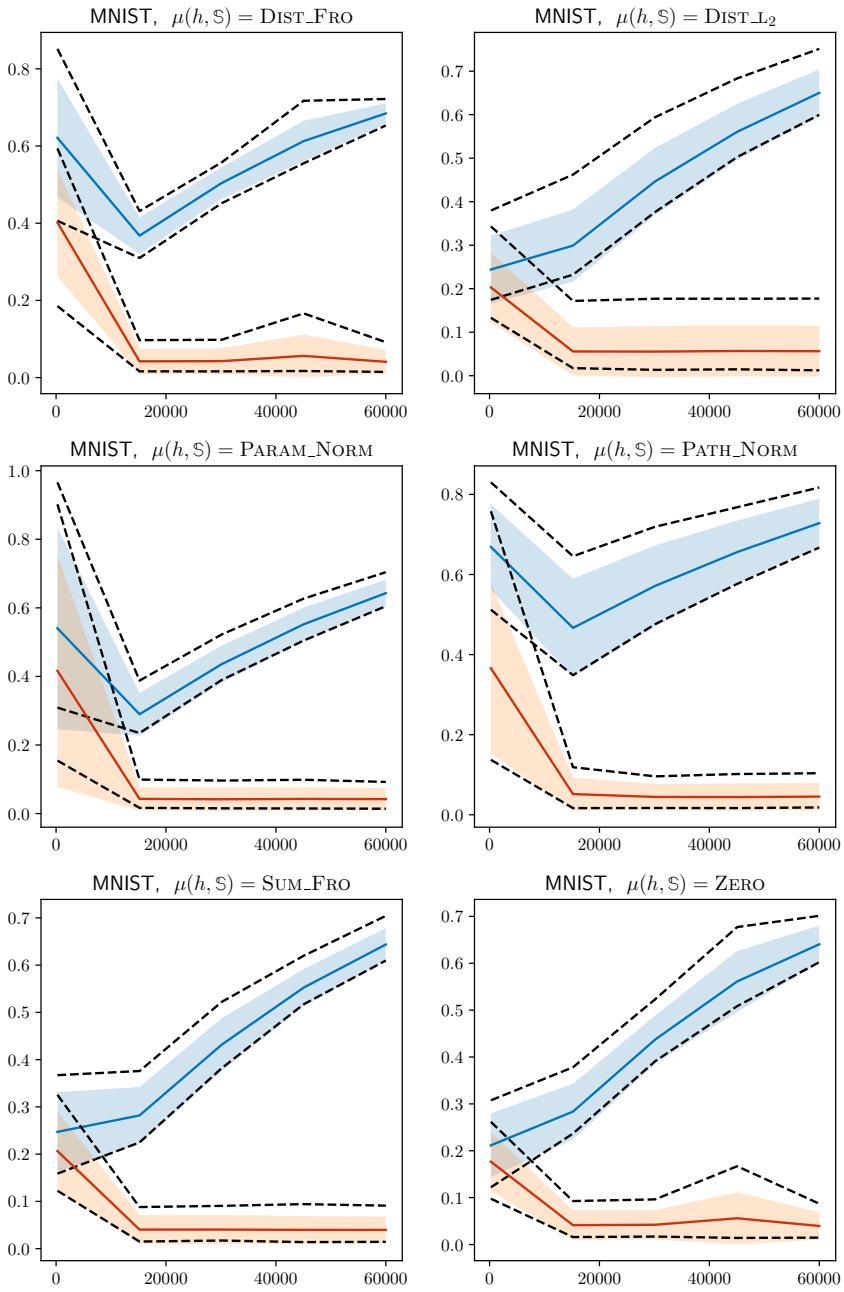


Figure G.3. *Influence of the parameter α in the x-axis. The bound values are represented in blue and the test risk in red. The two (solid) lines are the mean values computed on the depths and widths; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.*

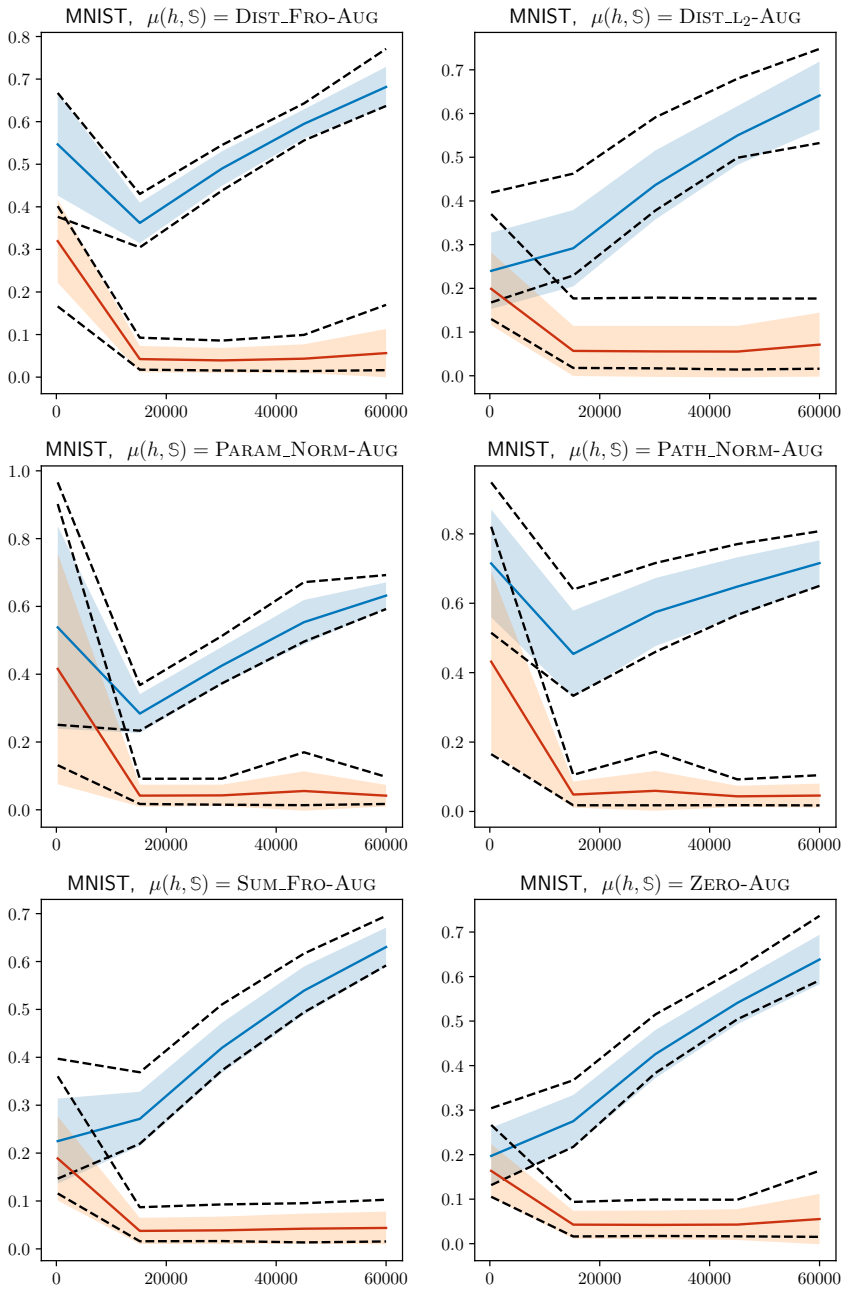


Figure G.4. *Influence of the parameter α in the x-axis. The bound values are represented in blue and the test risk in red. The two (solid) lines are the mean values computed on the depths and widths; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.*

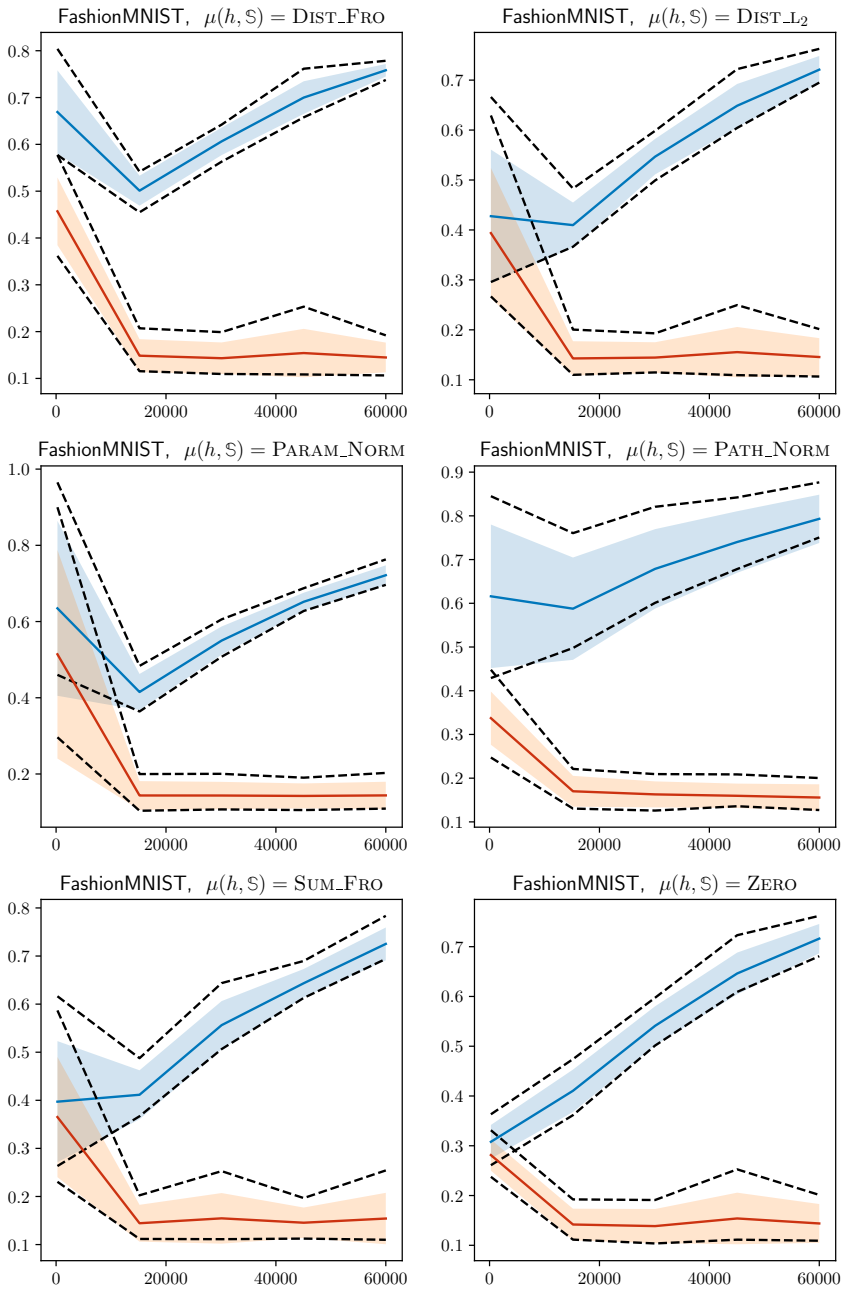


Figure G.5. *Influence of the parameter α in the x-axis. The bound values are represented in blue and the test risk in red. The two (solid) lines are the mean values computed on the depths and widths; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.*

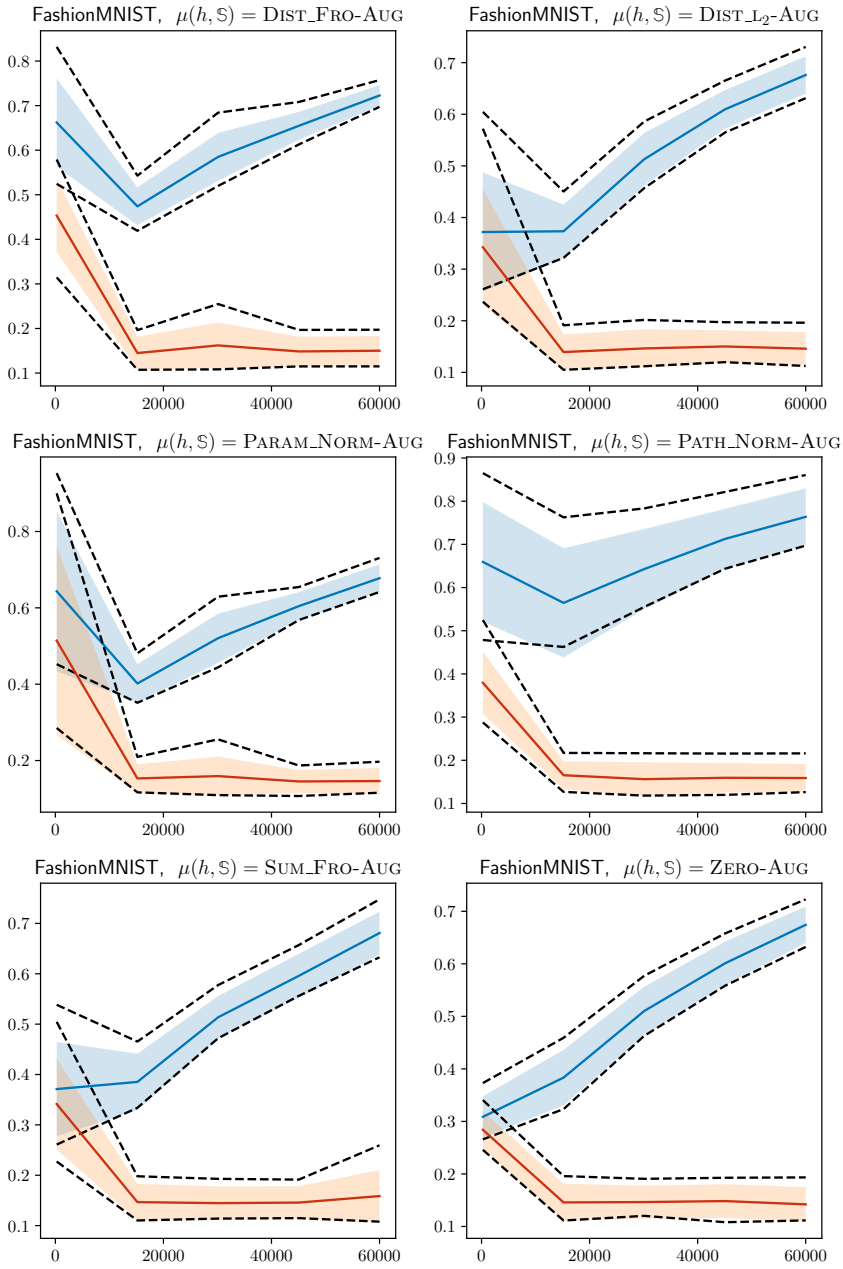


Figure G.6. *Influence of the parameter α in the x-axis. The bound values are represented in blue and the test risk in red. The two (solid) lines are the mean values computed on the depths and widths; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.*

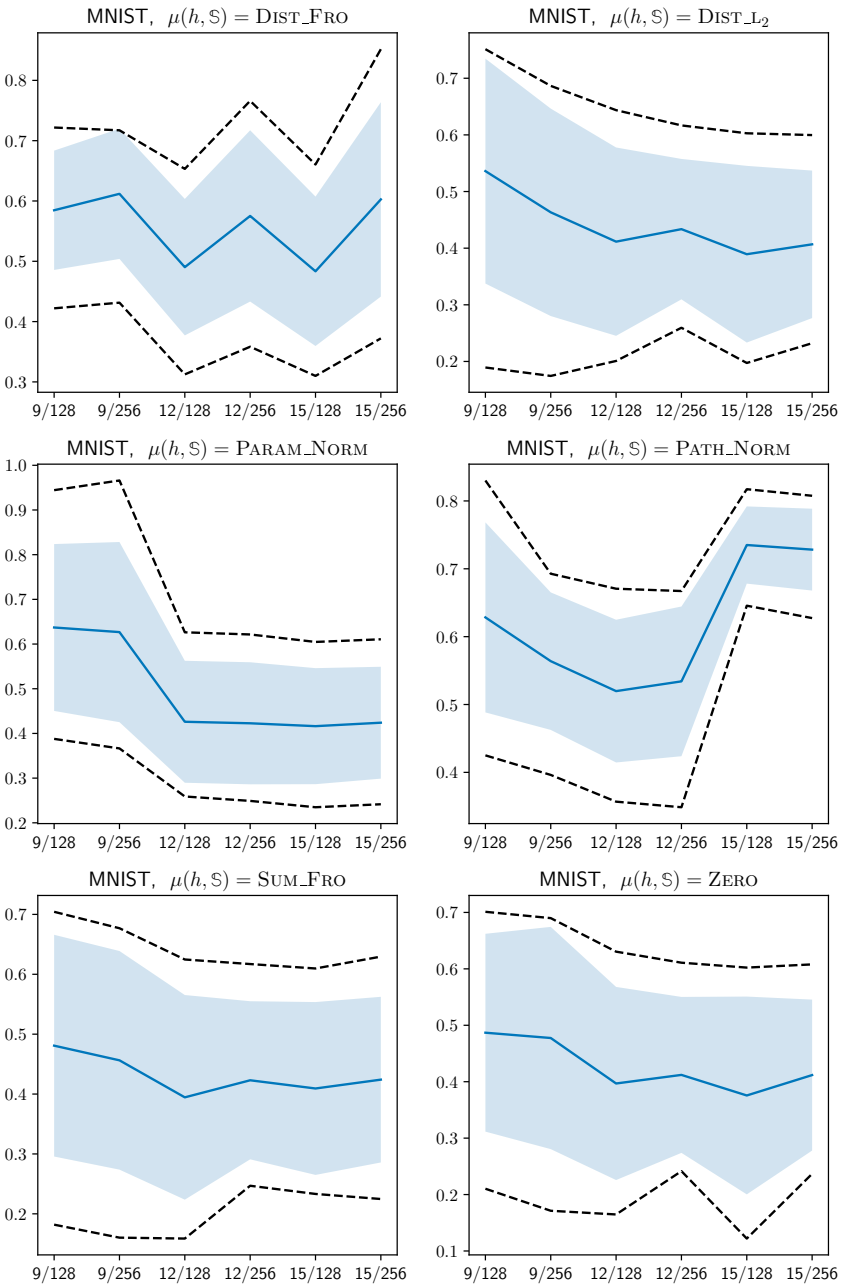


Figure G.7. *Influence of the depth and the width in the x-axis as "depth/width". The (solid) lines are the mean values computed on the different values of α; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.*

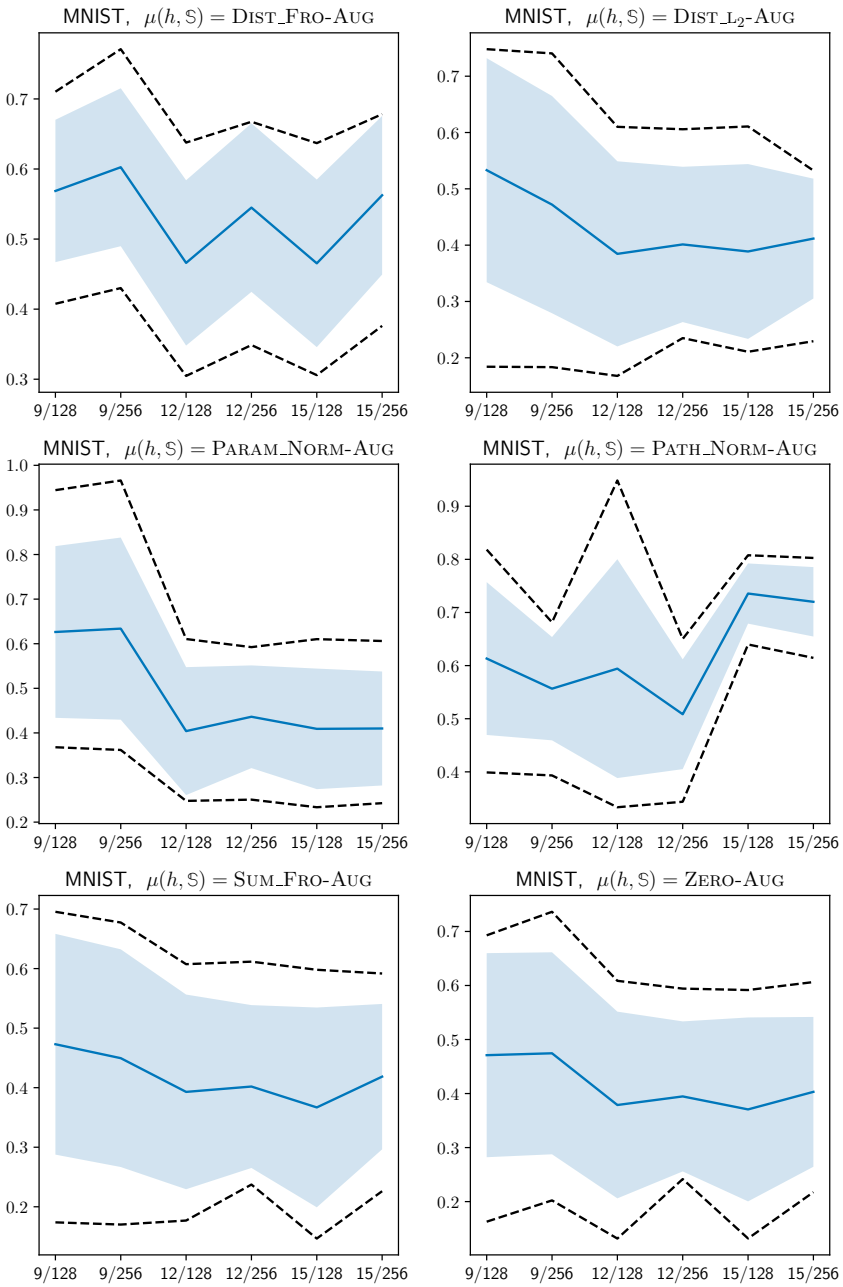


Figure G.8. *Influence of the depth and the width in the x-axis as "depth/width". The (solid) lines are the mean values computed on the different values of α; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.*

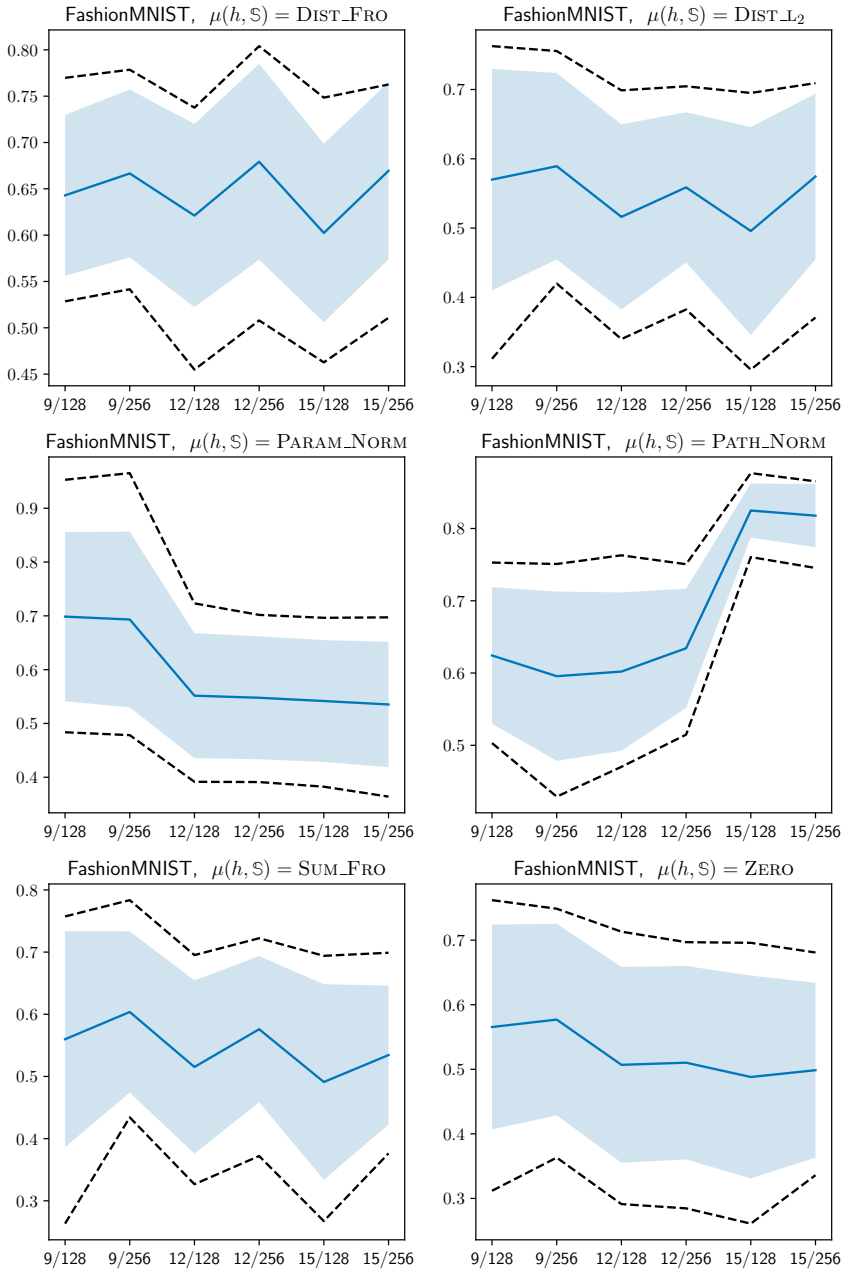


Figure G.9. *Influence of the depth and the width in the x-axis as "depth/width". The (solid) lines are the mean values computed on the different values of α; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.*

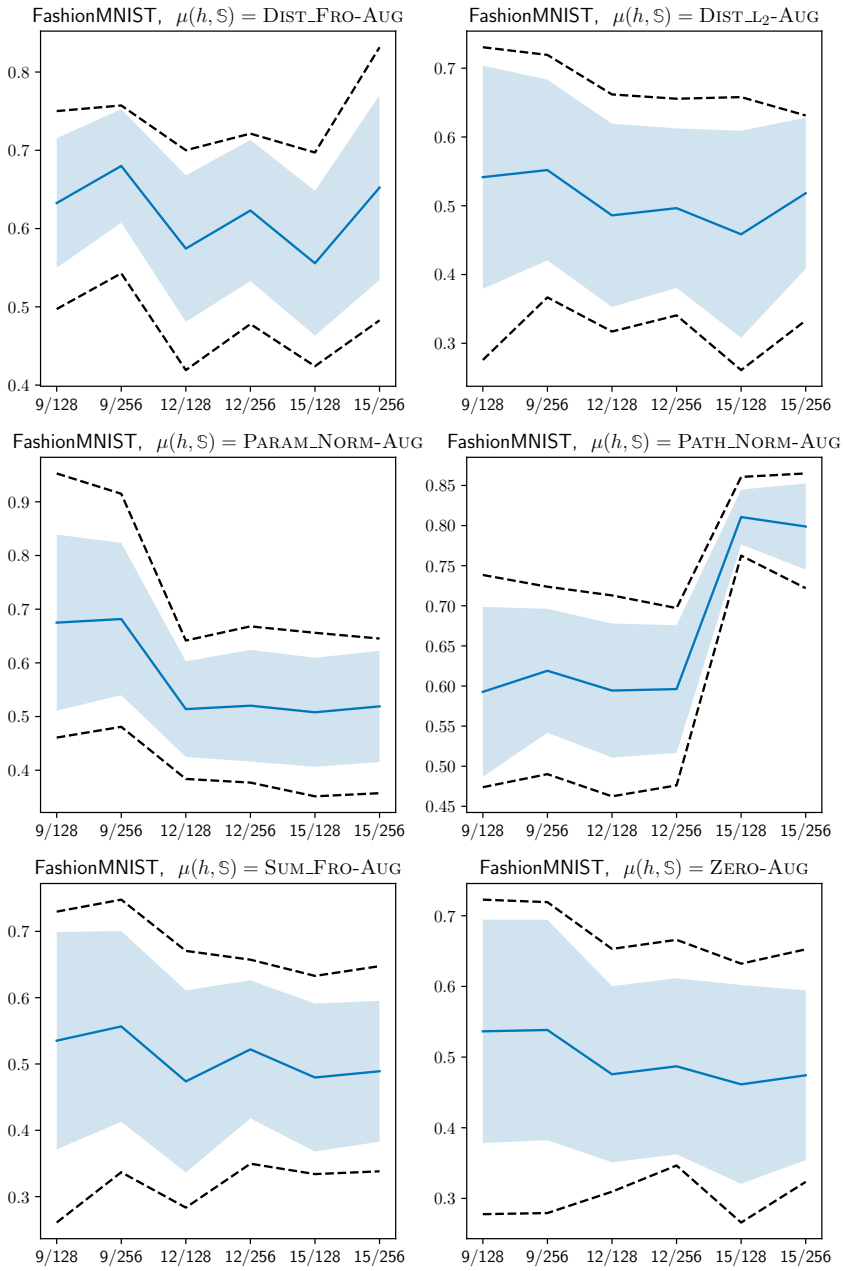


Figure G.10. *Influence of the depth and the width in the x-axis as "depth/width". The (solid) lines are the mean values computed on the different values of α; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.*

References

Martín Abadi *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015)
—— **Cited on page 35**.

Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are Labels Required for Improving Adversarial Robustness? *NeurIPS*. (2019)
—— **Cited on page 202**.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *CoRR*. abs/2110.11216. (2021)
—— **Cited on page 64**.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*. (2016)
—— **Cited on page 181**.

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes Bounds. *NIPS*. (2006)
—— **Cited on page 158**.

Peter Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*. (2002)
—— **Cited on pages 19, 46, 47, 193**.

Baptiste Bauvin, Cécile Capponi, Jean-Francis Roy, and François Laviolette. Fast greedy C-bound minimization with guarantees. *Machine Learning*. (2020)
—— **Cited on pages 102, 106, 121, 122**.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian Theory for Transductive Learning. *AISTATS*. (2014)
—— **Cited on page 65**.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian Bounds based on the Rényi Divergence. *AISTATS*. (2016)
—— **Cited on pages 73, 74, 151, 154, 225**.

Aurélien Bellet, Amaury Habrard, Emilie Morvant, and Marc Sebban. Learning A Priori Constrained Weighted Majority Votes. *Machine Learning*. (2014)
—— **Cited on pages 19, 57, 102**.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*. (2010)
—— **Cited on page 202**.

Julian Besag. Comments on "Representations of knowledge in complex systems" by U. Grenander and MI Miller. *Journal of the Royal Statistical Society, Series B*. (1994)
—— **Cited on page 185**.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion Attacks against Machine Learning at Test Time. *ECML PKDD*. (2013)
—— **Cited on page 82**.

Felix Biggs and Benjamin Guedj. Differentiable PAC-Bayes Objectives with Partially Aggregated Neural Networks. *Entropy*. (2021)
—— **Cited on page 150**.

Felix Biggs and Benjamin Guedj. On Margins and Derandomisation in PAC-Bayes. *AISTATS*. (2022)
—— **Cited on pages 150, 152**.

Felix Biggs, Valentina Zantedeschi, and Benjamin Guedj. On Margins and Generalisation for Voting Classifiers. *NeurIPS*. (2020)
—— **Cited on pages 171, 172**.

Christopher Bishop. Pattern recognition and machine learning (5th Edition). *Information science and statistics*. *Springer*. (2007)
—— **Cited on pages 17, 65**.

Gilles Blanchard and François Fleuret. Occam's Hammer. *COLT*. (2007)
——— **Cited on pages 20, 56, 75, 77, 141, 150, 155, 156, 159, 162, 173, 178, 201, 232, 236, 316**.

Robert Boik and James Robinson-Cox. Derivatives of the incomplete beta function. *Journal of Statistical Software*. (1999)
——— **Cited on page 138**.

Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. *COLT*. (1992)
——— **Cited on page 35**.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration Inequalities - A Nonasymptotic Theory of Independence. *Oxford University Press*. (2013)
——— **Cited on page 43**.

Olivier Bousquet and André Elisseeff. Stability and Generalization. *Journal of Machine Learning Research*. (2002)
——— **Cited on pages 19, 48, 49, 195**.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper Bounds for Uniformly Stable Algorithms. *COLT*. (2020)
——— **Cited on page 50**.

Stephen Boyd and Lieven Vandenberghe. Convex Optimization. *Cambridge University Press*. (2004)
——— **Cited on pages 109, 114, 228, 229, 242**.

James Bradbury *et al.* JAX: composable transformations of Python+NumPy programs. (2018)
——— **Cited on page 35**.

Leo Breiman. Bagging Predictors. *Machine Learning*. (1996)
——— **Cited on pages 19, 20, 56, 57, 203**.

Leo Breiman. Random Forests. *Machine Learning*. (2001)
——— **Cited on pages 19, 20, 57, 60, 62, 102, 105, 203**.

Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. Classification and Regression Trees. *Wadsworth*. (1984)
—— **Cited on pages 32, 34**.

Clément Canonne. A short note on an inequality between KL and TV. *CoRR*. abs/2202.07198. (2022)
—— **Cited on page 219**.

Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. *IEEE Symposium on Security and Privacy*. (2017)
—— **Cited on pages 82, 85**.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John Duchi, and Percy Liang. Unlabeled Data Improves Adversarial Robustness. *NeurIPS*. (2019)
—— **Cited on page 202**.

Olivier Catoni. Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001. *Springer Science & Business Media*. (2004)
—— **Cited on page 181**.

Olivier Catoni. Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. *CoRR*. abs/0712.0248. (2007)
—— **Cited on pages 20, 52, 56, 66–68, 70, 74–76, 78, 134, 141, 150, 155, 156, 159, 162, 173, 178, 181, 201, 222, 227, 228, 231, 286, 316, 324**.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*. (1952)
—— **Cited on pages 234, 235**.

Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*. (1987)
—— **Cited on page 181**.

Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings Algorithm. *The american statistician*. (1995)
—— **Cited on page 185**.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. *ICML*. (2019)
—— **Cited on page 85**.

Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. *NIPS 2011 BigLearn Workshop*. (2011)
—— **Cited on page 35**.

Nicolas de Condorcet. Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix. *Imprimerie Royale*. (1785)
—— **Cited on page 19**.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*. (1995)
—— **Cited on pages 35, 57**.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. (1967)
—— **Cited on page 57**.

Thomas Cover and Joy Thomas. Elements of Information Theory (2nd edition). *Wiley*. (2006)
—— **Cited on pages 218, 219**.

John Danskin. The Theory of Max-Min, with Applications. *SIAM Journal on Applied Mathematics*. (1966)
—— **Cited on pages 114, 272**.

Zhun Deng, Linjun Zhang, Amirata Ghorbani, and James Zou. Improving Adversarial Robustness via Unlabeled Out-of-Domain Data. *AISTATS*. (2021)
—— **Cited on page 202**.

Philip Derbeko, Ran El-Yaniv, and Ron Meir. Explicit Learning Curves for Transduction and Application to Clustering and Compression Algorithms. *Journal of Artificial Intelligence Research*. (2004)
—— **Cited on page 65**.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*. (2016)
—— **Cited on page 113**.

Thomas Dietterich. Ensemble methods in machine learning. *International workshop on multiple classifier systems*. (2000)
—— **Cited on pages 19, 20, 102**.

Monroe Donsker and Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time - III. *Communications on pure and applied Mathematics*. (1976)
—— **Cited on page 73**.

Dheeru Dua and Casey Graff. UCI Machine Learning Repository. (2017). url: http://archive.ics.uci.edu/ml
—— **Cited on pages 121, 139**.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. *NeurIPS*. (2020)
—— **Cited on pages 176, 186**.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in PAC-Bayes. *AISTATS*. (2021)
—— **Cited on pages 66, 93, 135, 184**.

Gintare Karolina Dziugaite and Daniel Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. *UAI*. (2017)
—— **Cited on pages 23, 158, 161, 317**.

Gintare Karolina Dziugaite and Daniel Roy. Data-dependent PAC-Bayes priors via differential privacy. *NeurIPS*. (2018)
—— **Cited on pages 93, 161, 162, 169, 187, 318**.

Tim van Erven and Peter Harremoës. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*. (2014)
—— **Cited on pages 74, 156**.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Robust Generalization via α-Mutual Information. *International Zurich Seminar on Information and Communication*. (2020)
—— **Cited on page 321**.

Farzan Farnia, Jesse Zhang, and David Tse. Generalizable Adversarial Training via Spectral Normalization. *ICLR*. (2019)
—— **Cited on page 85**.

Vitaly Feldman and Jan Vondrák. Generalization Bounds for Uniformly Stable Algorithms. *NeurIPS*. (2018)
—— **Cited on page 50**.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. *COLT*. (2019)
—— **Cited on page 50**.

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit Reparameterization Gradients. *NeurIPS*. (2018)
—— **Cited on page 138**.

Yoav Freund. Self Bounding Learning Algorithms. *COLT*. (1998)
—— **Cited on pages 19, 23, 60, 64, 83, 92, 102, 123, 126, 136, 153**.

Yoav Freund, Raj Iyer, Robert Schapire, and Yoram Singer. An Efficient Boosting Algorithm for Combining Preferences. *ICML*. (1998)
—— **Cited on page 57**.

Yoav Freund, Raj Iyer, Robert Schapire, and Yoram Singer. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*. (2003)
—— **Cited on page 57**.

Yoav Freund and Robert Schapire. Experiments with a New Boosting Algorithm. *ICML*. (1996)
—— **Cited on pages 19, 20, 56, 57, 203**.

Jerome Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*. (2001)
—— **Cited on pages 57, 203**.

Roy Frostig, Matthew James Johnson, and Chris Leary. Compiling machine learning programs via high-level tracing. *Systems for Machine Learning*. (2018)
—— **Cited on page 35**.

Tomer Galanti, Lior Wolf, and Tamir Hazan. A theoretical framework for deep transfer learning. *Information and Inference: A Journal of the IMA*. (2016)
—— **Cited on page 202**.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian Theory Meets Bayesian Inference. *NIPS*. (2016)
—— **Cited on page 65**.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. PAC-Bayes and domain adaptation. *Neurocomputing*. (2020)
—— **Cited on pages 65, 158, 202**.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian Learning of Linear Classifiers. *ICML*. (2009)
—— **Cited on pages 22, 66, 67, 70, 73, 75, 76, 152, 158, 162, 164, 220, 238**.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm. *Journal of Machine Learning Research*. (2015)
—— **Cited on pages 60, 61, 64, 104–107, 113, 122, 138, 213, 242, 251, 267**.

Manuel Gil, Fady Alajaji, and Tamás Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*. (2013)
—— **Cited on pages 314, 332**.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *AISTATS*. (2010)
—— **Cited on pages 96, 163**.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. *Adaptive computation and machine learning*. *MIT Press*. (2016)
—— **Cited on pages 35, 162**.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *ICLR*. (2015)
—— **Cited on pages 23, 82, 84**.

Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. PAC-Bayesian Compression Bounds on the Prediction Error of Learning Algorithms for Classification. *Machine Learning*. (2005)
—— **Cited on pages 19, 57, 203**.

Michael Grant, Stephen Boyd, and Yinyu Ye. Disciplined Convex Programming. *Global optimization*. (2006)
—— **Cited on page 113**.

Benjamin Guedj. A Primer on PAC-Bayesian Learning. *CoRR*. abs/1901.05353. (2019)
—— **Cited on page 64**.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *IEEE ICCV*. (2015)
—— **Cited on pages 163, 186**.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CVPR*. (2016)
—— **Cited on page 163**.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *ICLR*. (2019)
—— **Cited on pages 85, 86**.

Wayne Iba and Pat Langley. Induction of One-Level Decision Trees. *Ninth International Workshop on Machine Learning*. (1992)
—— **Cited on pages 32–34**.

Martin Jankowiak and Fritz Obermeyer. Pathwise Derivatives Beyond the Reparameterization Trick. *ICML*. (2018)
—— **Cited on page 138**.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic Generalization Measures and Where to Find Them. *ICLR*. (2019)
—— **Cited on pages 175, 176, 186**.

Yiding Jiang *et al.* Methods and Analysis of The First Competition in Predicting Generalization of Deep Learning. *NeurIPS 2020 Competition and Demonstration Track*. (2021)
—— **Cited on page 176**.

Peter Kairouz *et al.* Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*. (2021)
—— **Cited on page 203**.

Sham Kakade, Karthik Sridharan, and Ambuj Tewari. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. *NIPS*. (2008)
—— **Cited on pages 47, 180**.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in Deep Learning. *CoRR*. abs/1710.05468. (2017)
—— **Cited on pages 19, 57**.

Maurice Kendall. A new measure of rank correlation. *Biometrika*. (1938)
—— **Cited on page 176**.

Hoel Kervadec, Jose Dolz, Jing Yuan, Christian Desrosiers, Eric Granger, and Ismail Ben Ayed. Constrained Deep Networks: Lagrangian Optimization via Log-Barrier Extensions. *CoRR*. abs/1904.04205. (2019)
—— **Cited on page 109**.

Justin Khim and Po-Ling Loh. Adversarial Risk Bounds for Binary Classification via Function Transformation. *CoRR*. abs/1810.09519. (2018)
—— **Cited on page 85**.

Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *ICLR*. (2015)
—— **Cited on pages 96, 108, 163**.

Peter Kontschieder, Madalina Fiterau, Antonio Criminisi, and Samuel Rota Bulò. Deep Neural Decision Forests. *IJCAI*. (2016)
—— **Cited on pages 95, 253, 254**.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. MA thesis. University of Toronto, (2009)
—— **Cited on page 163**.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*. (2012)
—— **Cited on page 35**.

Ludmila Kuncheva. Combining pattern classifiers: methods and algorithms. *John Wiley & Sons*. (2014)
——— **Cited on pages 20, 102**.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Machine Learning at Scale. *ICLR*. (2017)
——— **Cited on pages 82, 84, 95**.

Alexandre Lacasse. Bornes PAC-Bayes et algorithmes d'apprentissage. PhD thesis. (2010)
——— **Cited on pages 70, 127**.

Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier. *NIPS*. (2006)
——— **Cited on pages 60–63, 100, 102, 105, 107, 111, 122, 268, 270**.

Alexandre Lacasse, François Laviolette, Mario Marchand, and Francis Turgeon-Boutin. Learning with Randomized Majority Votes. *ECML PKDD*. (2010)
——— **Cited on pages 126, 127, 138, 285, 286**.

John Langford. Tutorial on Practical Prediction Theory for Classification. *Journal of Machine Learning Research*. (2005)
——— **Cited on pages 75, 234, 235**.

John Langford and John Shawe-Taylor. PAC-Bayes & Margins. *NIPS*. (2002)
——— **Cited on pages 22, 39, 60, 75, 150**.

François Laviolette, Emilie Morvant, Liva Ralaivola, and Jean-Francis Roy. Risk upper bounds for general ensemble methods with an application to multiclass classification. *Neurocomputing*. (2017)
——— **Cited on pages 58, 63, 65, 100, 104–106, 129, 140, 141, 172**.

Yann LeCun, Corinna Cortes, and Christopher Burges. THE MNIST DATASET of handwritten digits. (1998). URL: `http://yann.lecun.com/exdb/mnist/`
——— **Cited on pages 31, 96, 121, 139, 163, 186**.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and François Laviolette. Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks. *NeurIPS*. (2019)
——— **Cited on pages 22, 150, 158, 164**.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*. (2013)
——— **Cited on pages 93, 324**.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *CoRR*. abs/1312.4400. (2013)
——— **Cited on page 186**.

Ben London. A PAC-Bayesian Analysis of Randomized Learning with Application to Stochastic Gradient Descent. *NIPS*. (2017)
——— **Cited on page 65**.

Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. On PAC-Bayesian bounds for random forests. *Machine Learning*. (2019)
——— **Cited on pages 102, 201**.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *ICLR*. (2018)
——— **Cited on pages 82, 85, 95, 272**.

Andrés Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Second Order PAC-Bayesian Bounds for the Weighted Majority Vote. *NeurIPS*. (2020)
——— **Cited on pages 22, 39, 60, 61, 64, 92, 100, 102, 105, 121, 122, 128, 138, 201, 214**.

Andreas Maurer. A Note on the PAC Bayesian Theorem. *CoRR*. cs.LG/0411099. (2004)
——— **Cited on pages 52, 70, 238, 242, 314, 316, 321, 349**.

David McAllester. Some PAC-Bayesian Theorems. *COLT*. (1998)
——— **Cited on pages 19, 56, 90, 134**.

David McAllester. Some PAC-Bayesian Theorems. *Machine Learning*. (1999)
——— **Cited on pages 52, 64**.

David McAllester. PAC-Bayesian Stochastic Model Selection. *Machine Learning*. (2003)
——— **Cited on pages 60, 66, 67, 74, 102, 106, 107, 110, 221, 227**.

Warren McCulloch and Walter Pitts. A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*. (1943)
——— **Cited on page 35**.

Colin McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*. (1989)
——— **Cited on pages 47, 49**.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. *AISTATS*. (2017)
——— **Cited on page 203**.

Daniel McNamara and Maria-Florina Balcan. Risk Bounds for Transferring Representations With and Without Fine-Tuning. *ICML*. (2017)
——— **Cited on page 202**.

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes Un-Expected Bernstein Inequality. *NeurIPS*. (2019)
——— **Cited on pages 135, 136**.

Marvin Minsky and Seymour Papert. Perceptrons: an introduction to computational geometry. *The MIT Press*. (1972)
——— **Cited on pages 35, 36**.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. *Adaptive computation and machine learning*. *MIT Press*. (2012)
——— **Cited on pages 17, 44–46**.

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC Classes are Adversarially Robustly Learnable, but Only Improperly. *COLT*. (2019)
——— **Cited on page 85**.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Reducing Adversarially Robust Learning to Non-Robust PAC Learning. *NeurIPS*. (2020)
—— **Cited on page 85**.

Vaishnavh Nagarajan and Zico Kolter. Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience. *ICLR*. (2019a)
—— **Cited on page 152**.

Vaishnavh Nagarajan and Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *NeurIPS*. (2019b)
—— **Cited on pages 85, 150, 194**.

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*. (2005)
—— **Cited on page 131**.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. *ICLR*. (2018)
—— **Cited on pages 150, 152**.

Yuki Ohnishi and Jean Honorio. Novel Change of Measure Inequalities with Applications to PAC-Bayesian Bounds and Monte Carlo Estimation. *AISTATS*. (2021)
—— **Cited on pages 73, 248, 251**.

Francesco Orabona and Tatiana Tommasi. Training Deep Networks without Learning Rates Through Coin Betting. *NIPS*. (2017)
—— **Cited on pages 108, 121, 139, 203**.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The Limitations of Deep Learning in Adversarial Settings. *IEEE EuroS&P*. (2016)
—— **Cited on page 82**.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-Bayes Bounds with Data Dependent Priors. *Journal of Machine Learning Research*. (2012)
—— **Cited on pages 66, 93, 135, 184**.

Adam Paszke *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *NeurIPS*. (2019)
——— **Cited on pages 35, 111, 138**.

Fabian Pedregosa *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. (2011)
——— **Cited on page 35**.

Anastasia Pentina and Christoph H. Lampert. A PAC-Bayesian bound for Lifelong Learning. *ICML*. (2014)
——— **Cited on page 129**.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter Risk Certificates for Neural Networks. *Journal of Machine Learning Research*. (2021)
——— **Cited on pages 23, 160–162, 164, 184**.

Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization. *NeurIPS*. (2019)
——— **Cited on page 85**.

Rafael Pinot, Laurent Meunier, Florian Yger, Cédric Gouy-Pailler, Yann Chevaleyre, and Jamal Atif. On the robustness of randomized classifiers to adversarial examples. *Machine Learning*. (2022)
——— **Cited on page 85**.

Ross Quinlan. Induction of Decision Trees. *Machine Learning*. (1986)
——— **Cited on page 34**.

Ross Quinlan. C4.5: Programs for Machine Learning. *Morgan Kaufmann*. (1993)
——— **Cited on page 34**.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. *COLT*. (2017)
——— **Cited on page 181**.

Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing

Processes. *Journal of Machine Learning Research*. (2010)
——— **Cited on pages 90, 250**.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. Advances in Domain Adaptation Theory. *Elsevier*. (2019)
——— **Cited on page 202**.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory. *CoRR*. abs/2004.11829. (2020)
——— **Cited on page 202**.

David Reeb, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch. Learning Gaussian Processes by Minimizing PAC-Bayesian Generalization Bounds. *NeurIPS*. (2018)
——— **Cited on pages 71, 111**.

Kui Ren, Tianhang Zheng, and Xue Liu. Adversarial Attacks and Defenses in Deep Learning. *Engineering*. (2020)
——— **Cited on page 84**.

Rami Al-Rfou *et al.* Theano: A Python framework for fast computation of mathematical expressions. *CoRR*. abs/1605.02688. (2016)
——— **Cited on page 35**.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes Analysis Beyond the Usual Bounds. *Advances in Neural Information Processing System (NeurIPS)*. (2020)
——— **Cited on pages 76, 146, 150, 155, 159, 162, 171, 173, 178, 201, 230, 316**.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*. (1958)
——— **Cited on page 35**.

Jean-Francis Roy, François Laviolette, and Mario Marchand. From PAC-Bayes Bounds to Quadratic Programs for Majority Votes. *ICML*. (2011)
——— **Cited on pages 60, 102, 106, 121, 122**.

Jean-Francis Roy, Mario Marchand, and François Laviolette. A Column Generation Bound Minimization Approach with PAC-Bayesian Generalization Guar-

antees. *AISTATS*. (2016)
——— **Cited on pages 102, 105, 106, 266**.

Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern Approach (4th Edition). *Pearson*. (2020)
——— **Cited on page 17**.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. *NeurIPS*. (2019)
——— **Cited on page 85**.

Robert Schapire, Yoav Freund, Peter Barlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*. (1998)
——— **Cited on page 171**.

Robert Schapire and Yoram Singer. Improved Boosting Algorithms using Confidence-Rated Predictions. *COLT*. (1998)
——— **Cited on page 57**.

Robert Schapire and Yoram Singer. Improved Boosting Algorithms Using Confidence-rated Predictions. *Machine Learning*. (1999)
——— **Cited on page 57**.

Robert Schapire and Yoram Singer. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*. (2000)
——— **Cited on page 57**.

Edward Scheinerman and Daniel Ullman. Fractional Graph Theory: A Rational Approach to the Theory of Graphs. *Courier Corporation*. (2011)
——— **Cited on page 250**.

Matthias Seeger. PAC-Bayesian Generalisation Error Bounds for Gaussian Process Classification. *Journal of Machine Learning Research*. (2002)
——— **Cited on pages 52, 66, 67, 69, 70, 74, 77, 90, 102, 106, 107, 110, 133, 134, 138, 223, 227, 286**.

Vera Shalaeva, Alireza Fakhrizadeh Esfahani, Pascal Germain, and Mihály Petreczky. Improved PAC-Bayesian Bounds for Linear Regression. *AAAI*.

(2020)
—— **Cited on page 65**.

Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning - From Theory to Algorithms. *Cambridge University Press*. (2014)
—— **Cited on page 17**.

John Shawe-Taylor and Robert Williamson. A PAC Analysis of a Bayesian Estimator. *COLT*. (1997)
—— **Cited on pages 19, 52, 56, 64**.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the Black Box of Deep Neural Networks via Information. *CoRR*. abs/1703.00810. (2017)
—— **Cited on page 204**.

Adrian Smith and Gareth Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B (Methodological)*. (1993)
—— **Cited on page 185**.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net. *ICLR*. (2015)
—— **Cited on page 163**.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *ICLR*. (2014)
—— **Cited on pages 82, 83**.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A Strongly Quasiconvex PAC-Bayesian Bound. *ALT*. (2017)
—— **Cited on pages 105, 135, 156, 157, 308**.

Naftali Tishby, Fernando Pereira, and William Bialek. The information bottleneck method. *CoRR*. physics/0004057. (2000)
—— **Cited on page 204**.

Aad van der Vaart and Jon Wellner. Weak convergence and empirical processes. *Springer series in statistics*. *Springer*. (1996)
—— **Cited on page 51**.

Leslie Valiant. A Theory of the Learnable. *Communications of the ACM*. (1984)
—— **Cited on page 43**.

Vladimir Vapnik. Statistical Learning Theory. *A Wiley-Interscience publication*. *Wiley*. (1998)
—— **Cited on pages 41, 46**.

Vladimir Vapnik and Alexey Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Doklady Akademii Nauk USSR*. (1968)
—— **Cited on pages 18, 19, 41, 43, 45**.

Vladimir Vapnik and Alexey Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*. (1971)
—— **Cited on pages 18, 19, 41, 43, 45, 193**.

Vladimir Vapnik and Alexey Chervonenkis. Theory of Pattern Recognition. *Nauka, Moscow*. (1974)
—— **Cited on pages 18, 19, 22, 41**.

Sergio Verdú. α-mutual information. *Information Theory and Applications Workshop*. (2015)
—— **Cited on pages 319, 320, 326**.

Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. *ICML*. (2011)
—— **Cited on page 181**.

Yihong Wu. Lecture notes on: Information-theoretic methods for high-dimensional statistics. (2020)
—— **Cited on page 219**.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR*. abs/1708.07747. (2017)
—— **Cited on pages 96, 121, 139, 163, 186**.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *NIPS*. (2017)
—— **Cited on page 173**.

Huan Xu and Shie Mannor. Robustness and Generalization. *COLT*. (2010)
—— **Cited on page 48**.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine Learning*. (2012)
—— **Cited on pages 48, 50, 195**.

Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher Complexity for Adversarially Robust Generalization. *ICML*. (2019)
—— **Cited on page 85**.

Malik Younsi. Proof of a Combinatorial Conjecture Coming from the PAC-Bayesian Machine Learning Theory. *CoRR*. abs/1209.0824. (2012)
—— **Cited on page 269**.

Malik Younsi and Alexandre Lacasse. A combinatorial conjecture from PAC-Bayesian machine learning. *CoRR*. abs/2006.01387. (2020)
—— **Cited on page 269**.

Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient Defenses Against Adversarial Attacks. *ACM Workshop on Artificial Intelligence and Security, AISec@CCS*. (2017)
—— **Cited on pages 82, 85, 86**.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan Adams, and Peter Orbanz. Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach. *ICLR*. (2019)
—— **Cited on pages 23, 158, 161**.

Ji Zhu, Hui Zou, Saharon Rosset, and Trevor Hastie. Multi-class AdaBoost. *Statistics and its Interface*. (2009)
—— **Cited on page 57**.

Abstract. In machine learning, a model is learned from data to solve a task automatically. In the supervised classification setting, the model aims to predict the label associated with an input. The model is learned using a limited number of examples, each consisting of an input and its associated label. However, the model's performance on the examples, computed by the empirical risk, does not necessarily reflect the performance on the task, which is represented by the true risk. Moreover, since it is not computable, the true risk is upper-bounded by a generalization bound that mainly depends on two quantities: the empirical risk and a complexity measure. One way to learn a model is to minimize a bound by a type of algorithm called self-bounding. PAC-Bayesian bounds are well suited to the derivation of this type of algorithm. In this context, the first contribution consists in developing self-bounding algorithms that minimize PAC-Bayesian bounds to learn majority votes. If these bounds are well adapted to majority votes, their use for other models becomes less natural. To overcome this difficulty, a second contribution focuses on the disintegrated PAC-Bayesian bounds that are natural for more general models. In this framework, we provide the first empirical study of these bounds. In a third contribution, we derive bounds that allow us to incorporate complexity measures defined by the user.

Keywords. Machine Learning, Generalization, PAC-Bayesian Bound, Disintegrated PAC-Bayesian Bound, Self-Bounding Algorithm, Majority Vote, Neural Network, Complexity Measure.

Résumé. En apprentissage automatique, un modèle est appris à partir de données pour résoudre une tâche de manière automatique. Dans le cadre de la classification supervisée, le modèle vise à prédire la classe associée à une entrée. Le modèle est appris à l'aide d'un nombre limité d'exemples, chacun étant constitué d'une entrée et de sa classe associée. Cependant, la performance du modèle sur les exemples, calculée par le risque empirique, ne reflète pas nécessairement la performance sur la tâche qui est représentée par le risque réel. De plus, n'étant pas calculable, le risque réel est majoré pour obtenir une borne en généralisation qui dépend principalement de deux quantités : le risque empirique et une mesure de complexité. Une façon d'apprendre un modèle est de minimiser une borne par un type d'algorithme appelé auto-certifié (ou auto-limitatif). Les bornes PAC-Bayésiennes sont bien adaptées à la dérivation de ce type d'algorithmes. Dans ce contexte, la première contribution consiste à développer des algorithmes auto-certifiés qui minimisent des bornes PAC-Bayésiennes pour apprendre des votes de majorité. Si ces bornes sont bien adaptées aux votes de majorité, leur utilisation pour d'autres modèles devient moins naturelle. Pour pallier cette difficulté, une seconde contribution se concentre sur les bornes PAC-Bayésiennes désintégrées qui sont naturelles pour des modèles plus généraux. Dans ce cadre, nous apportons la première étude empirique de ces bornes. Dans une troisième contribution, nous dérivons des bornes permettant d'incorporer des mesures de complexité pouvant être définies par l'utilisateur.

Mot-clés. Apprentissage Automatique, Généralisation, Borne PAC-Bayésienne, Borne PAC-Bayésienne Désintégrée, Algorithme Auto-certifié, Algorithme Auto-limitatif, Vote de Majorité, Réseau de Neurones, Mesure de Complexité.