Data-Dependent Analysis of Learning Algorithms

Petra Philips

A thesis submitted for the degree of Doctor of Philosophy
at The Australian National University

May 2005

Typeset in Computer Modern by TEX and LATEX 2ε.

Except where otherwise indicated, this thesis is my own original work.

The results in this thesis were produced under the supervision of Shahar Mendelson and Bob Williamson, and partly in collaboration with Peter Bartlett. The main contribution of this thesis are two related parts. The main technical results in the first part on random subclass bounds appeared as a journal paper with Shahar Mendelson [1], and an earlier conference paper [2]. The results were discussed with my supervisors Shahar Mendelson and Bob Williamson, who gave me advice and direction. The results on the data-dependent estimation of localized complexities for the Empirical Risk Minimization algorithm appeared as part of a conference paper with Peter Bartlett and Shahar Mendelson [3], and the optimality results are work in progress and contained in an unpublished manuscript with Peter Bartlett and Shahar Mendelson [4]. This second part of the thesis is based on intensive discussions and technical advice from Shahar Mendelson and Peter Bartlett.

List of Publications:

[1] S. Mendelson and P. Philips. On the importance of small coordinate projections. *Journal of Machine Learning Research*, 5:219–238, 2004.

[2] S. Mendelson and P. Philips. Random subclass bounds. In B. Schölkopf and M. Warmuth, editors, *Proceedings of the 16th Annual Conference on Learning Theory, COLT 2003*, pages 329–343. Springer, 2003.

[3] P. L. Bartlett, S. Mendelson, and P. Philips. Local complexities for empirical risk minimization. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004*, pages 270–284. Springer, 2004.

[4] P. L. Bartlett, S. Mendelson, and P. Philips. Work in progress, 2005.

Petra Philips
8 May 2005

Ce n'est pas une image juste, c'est juste une image.

(Jean-Luc Godard)

Acknowledgements

First of all, I would like to express my special gratitude to both of my supervisors Bob Williamson and Shahar Mendelson for the continuous support they provided me with while working on this thesis. Bob Williamson introduced me to statistical learning theory, and was the inspiration for and the reason that I came to the ANU. I would like to thank Bob for his trust and understanding, and his valuable guidance, support, and advice throughout the years. I will also never forget the generosity of Bob, Angharad, and Myvanmy Williamson in being such lovely hosts during my first visit to Australia, many years before starting this Ph.D. project.

A most special thank you goes to Shahar Mendelson, who is an outstanding teacher, mentor, and friend. I feel privileged to have had the opportunity to work so closely with him, and I am very grateful for his constant support, availability, patience, for his intensive teaching, comments, constructive critics and discussions. His enthusiasm and sharpness encouraged, motivated, and inspired me immensely, and I treasure much our many friendly conversations sparkling with humour and wit.

I would also like to record my debt of gratitude to Peter Bartlett for being my advisor, and for giving me the unforgettable opportunity to work and learn from him while visiting UC Berkeley. My special thanks to Olivier Bousquet, Ralf Herbrich, Gábor Lugosi, and Bernhard Schölkopf for hosting me kindheartedly while visiting their institutes. Thank you to all of them for stimulating technical and very pleasant personal discussions.

I was lucky to have charming and helpful colleagues and motivating partners for general discussions, among them Olivier Buffet, Cheng Soon Ong, Gunnar Rätsch, Alex Smola, Vishy Vishvanathan. Thanks especially to Omri Guttman, Evan Greensmith, and Tim Sears for a great time while sharing an office (and heaps of chocolates).

There are many other people who inspired me professionally at some point throughout these years. They are Shai Ben-David, Nicolò Cesa-Bianchi, Stephane Boucheron, André Elisseeff, Matthias Hein, Vladimir Koltchinskii, Risi Kondor, John Langford, Phil Long, Ulrike von Luxburg, Shie Mannor, Mario Marchand, Sayan Mukherjee, Dmitry Panchenko, Alexander Rakhlin, Matthias Seeger, Karim Seghouane, John Shawe-Taylor, Alexandre Tsybakov, Manfred Warmuth, Tong Zhang, and Joel Zinn.

Thanks to Cheng Soon Ong, and especially to Alexander Rakhlin, for proof-reading background parts of this thesis.

My very heartfelt thank you goes to my friends who shared a lot of laughter, secrets, mishaps, debates, ideas, an extraordinary house, terrace, and garden, cooking evenings,

weddings, but who, above all, proved their loyalty, trustworthiness, and their support in times of disappointment, personal pain, and despair — especially in the last weeks of work on this thesis. Lisa Batten, Bina D'Costa, Dave Kilham, Emily Kilham, Tao Kong, Stephanie Lee, and Torsten Juelich were lovely house-mates. Bina and Dave are wonderful friends, whose kindness, integrity, and open-mindedness give them a special place in my heart. The "whole Kilhams" are magicians of delicious dinners and were a warm and welcoming family and the kindest of hosts. Emily became very fast from "Dave's sister" an exceptionally dear, delicate, and warmhearted friend. Thanks to Steph for her caring, enthusiastic, and lively presence, for animated discussions, and for the daily challenge of climbing Canberra's Black Mountain. Thanks to Bina and Dave, and also to my long-time friends in Europe Ina Ambela and Frank Abegg, for establishing that I am perfectly qualified to be bridesmaid in both a Bengoli-Australian and a Greek-German wedding – I still feel touched by the great trust and affection they showed me. Thanks to Lilach Zac for patience and advice when I needed it, and to Uwe Zimmer for numerous movie nights in his "film club".

And finally, deepest felt thanks to my parents for their unconditional love.

Abstract

This thesis studies the generalization ability of machine learning algorithms in a statistical setting. It focuses on the data-dependent analysis of the generalization performance of learning algorithms in order to make full use of the potential of the actual training sample from which these algorithms learn.

First, we propose an extension of the standard framework for the derivation of generalization bounds for algorithms taking their hypotheses from random classes of functions. This approach is motivated by the fact that the function produced by a learning algorithm based on a random sample of data depends on this sample and is therefore a random function. Such an approach avoids the detour of the worst-case uniform bounds as done in the standard approach. We show that the mechanism which allows one to obtain generalization bounds for random classes in our framework is based on a "small complexity" of certain random coordinate projections. We demonstrate how this notion of complexity relates to learnability and how one can explore geometric properties of these projections in order to derive estimates of rates of convergence and good confidence interval estimates for the expected risk. We then demonstrate the generality of our new approach by presenting a range of examples, among them the algorithm-dependent compression schemes and the data-dependent luckiness frameworks, which fall into our random subclass framework.

Second, we study in more detail generalization bounds for a specific algorithm which is of central importance in learning theory, namely the Empirical Risk Minimization algorithm (ERM). Recent results show that one can significantly improve the high-probability estimates for the convergence rates for empirical minimizers by a direct analysis of the ERM algorithm. These results are based on a new localized notion of complexity of subsets of hypothesis functions with identical expected errors and are therefore dependent on the underlying unknown distribution. We investigate the extent to which one can estimate these high-probability convergence rates in a data-dependent manner. We provide an algorithm which computes a data-dependent upper bound for the expected error of empirical minimizers in terms of the "complexity" of data-dependent local subsets. These subsets are sets of functions of empirical errors of a given range and can be determined based solely on empirical data. We then show that recent direct estimates, which are essentially sharp estimates on the high-probability convergence rate for the ERM algorithm, can not be recovered universally from empirical data.

Contents