The Pick-to-Learn Algorithm: Empowering Compression for Tight Generalization Bounds & Improved Post-training Performance

Dario Paccagnan, Marco C. Campi, Simone Garatti



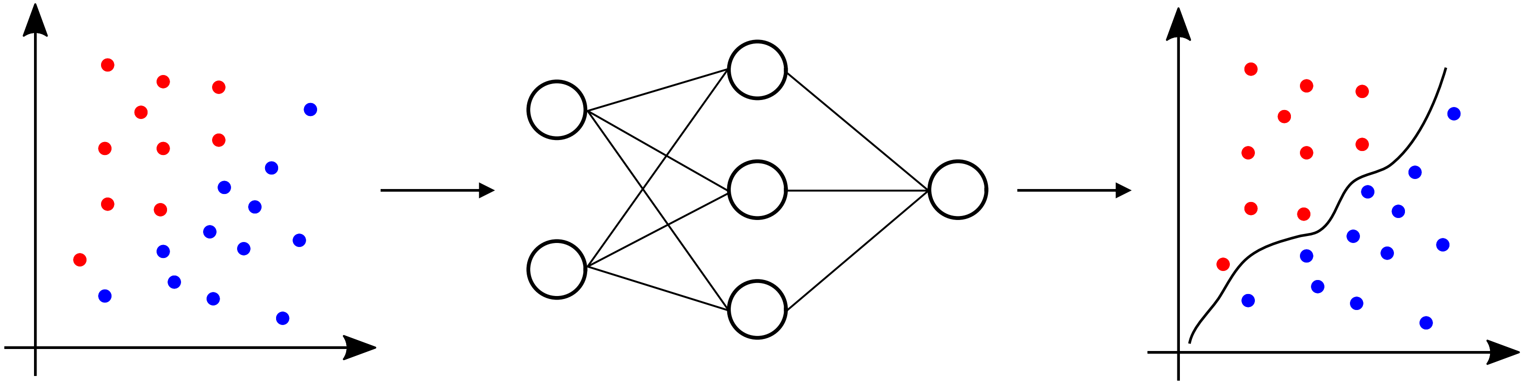
NeurIPS 2023 spotlight

Imperial College London

UNIVERSITY OF BRESCIA

POLITECNICO MILANO 1863

Motivation: ML algo with good performance & guarantees



- True generalization
- Bound on generalization

In this work we address both

Why?

Not just a theoretical exercise...

Normal

Heart attack

Generalization bounds: existing approaches & limitations

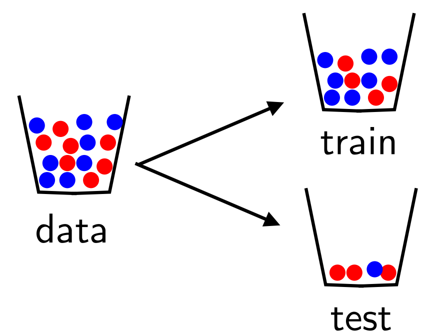
Does not use additional data:

- VC dimension [Vapnik & Chervonenkis, 1971]
- Radamacher complexity [Bartlett & Mendelson, 2001]
- Sharpness [Keskar et al., 2017]

+ use all data for training / **– loose bounds**

Uses additional data:

- Test-set bounds, e.g., [Chernhoff, 1952]
- PAC-Bayes [Dziugate & Roy, 2017] [Perez et al, 2021]*



– not use all data for training / **+ tighter bounds**

Can we break this barrier?

Yes! With (preferent) compression theory

arXiv:2301.12767v1 [cs.LG] 30 Jan 2023

Compression, Generalization and Learning

Marco C. Campi* Simone Garatti†

Abstract

A *compression function* is a map that slims down an observational set into a subset of reduced size, while preserving its informational content. In multiple applications, the condition that one new observation makes the compressed set change is interpreted that this observation brings in extra information and, in learning theory, this corresponds to misclassification, or misprediction. In this paper, we lay the foundations of a new theory that allows one to keep control on the probability of change of compression (called the "risk"). We identify conditions under which the cardinality of the compressed set is a consistent estimator for the risk (without any upper limit on the size of the compressed set) and prove unprecedentedly tight bounds to evaluate the risk under a generally applicable condition of *preference*. All results are usable in a fully *agnostic* setup, without requiring any *a priori* knowledge on the probability distribution of the observations. Not only these results offer a valid support to develop trust in observation-driven methodologies, they also play a fundamental role in learning techniques as a tool for hyper-parameter tuning.

Keywords: Compression Schemes, Statistical Risk, Statistical Learning Theory.

1 Introduction

Compression is an established topic in theoretical learning, and various generalization bounds have been proven for compression schemes.

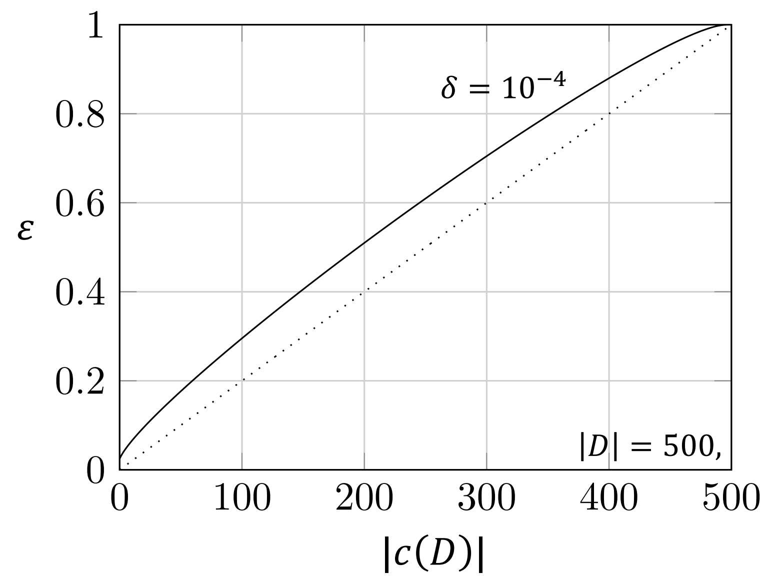
According to a definition introduced in [30], a compression scheme consists of i. a *compression function* c, which maps any list of observed examples $S = ((x_1, y_1), \dots, (x_N, y_N))$ (x_i is called an "instance" and y_i a "label") into a sub-list $\mathsf{c}(S)$, and ii. a *reconstruction function* ρ, which maps any list of examples S into a classifier $\rho(S)$. An important feature of a classifier is its *risk* and, in the context of compression schemes, one is interested in the risk associated to the classifier $\rho(\mathsf{c}(S))$. The concept of risk finds a natural definition in *statistical*

*Department of Information Engineering – University of Brescia, via Branze 38, 25123 Brescia, Italy. Email: marco.campi@unibs.it

†Dipartimento di Elettronica, Informazione e Bioingegneria – Politecnico di Milano, piazza L. da Vinci 32, 20133 Milano, Italy. Email: simone.garatti@polimi.it

1

$$\mathbb{P}^N\{\text{risk} \le \varepsilon(\delta, |c(D)|)\} \ge 1 - \delta$$



Challenge: ML algos do not have compression properties

[Bousquet et al, 2020] [Hanneke & Kantorovich, 2021]

Main result: P2L induces preferent compression

Goal: Given a black-box learning algo L, construct a meta-algorithm (P2L) around it to secure preferent compression

Input: dataset D; learning algorithm L(D), scoring function $s_h(z)$

Initialize: $T = \emptyset$, $h = h_0$, $z^* = \underset{z \in D \setminus T}{\operatorname{argmax}}\, s_h(z)$

While $\max_{z \in D \setminus T} s_h(z) >$ threshold **do**

$$T \leftarrow T \cup \{z^*\}$$

$$h \leftarrow L(T)$$

$$z^* \leftarrow \underset{z \in D \setminus T}{\operatorname{argmax}}\, s_h(z)$$

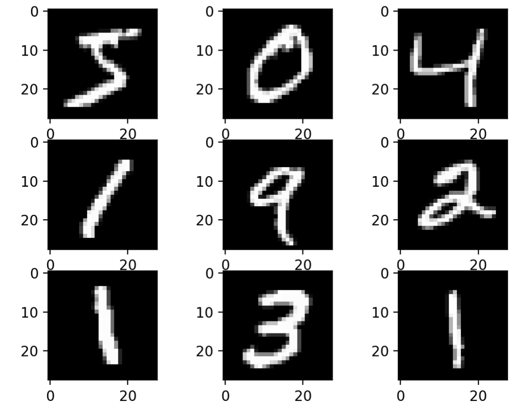
Theorem (informal): P2L is a preferent compression algorithm

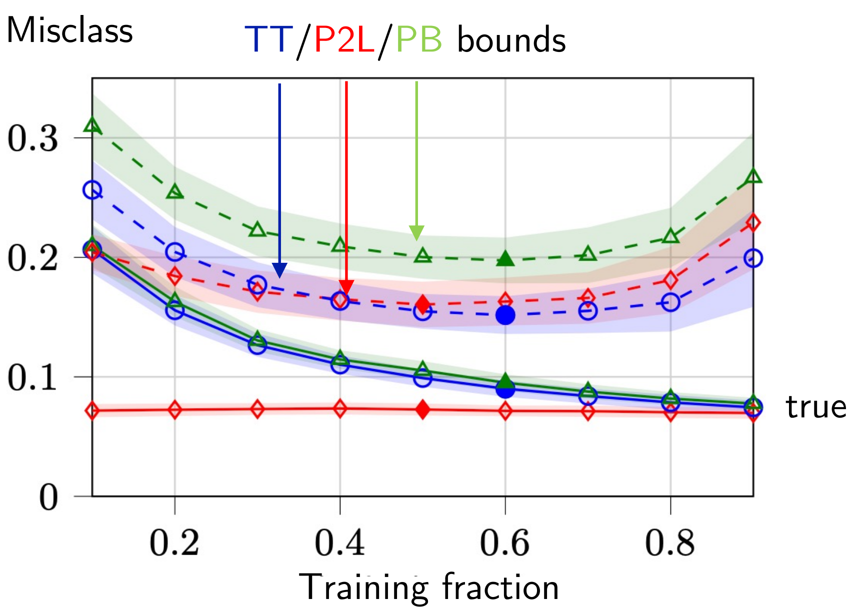
Hopes: - P2L compresses "a lot" $\Longrightarrow$ good bound on generalization
- P2L does not change the "nature" of L $\Longrightarrow$ good generalization

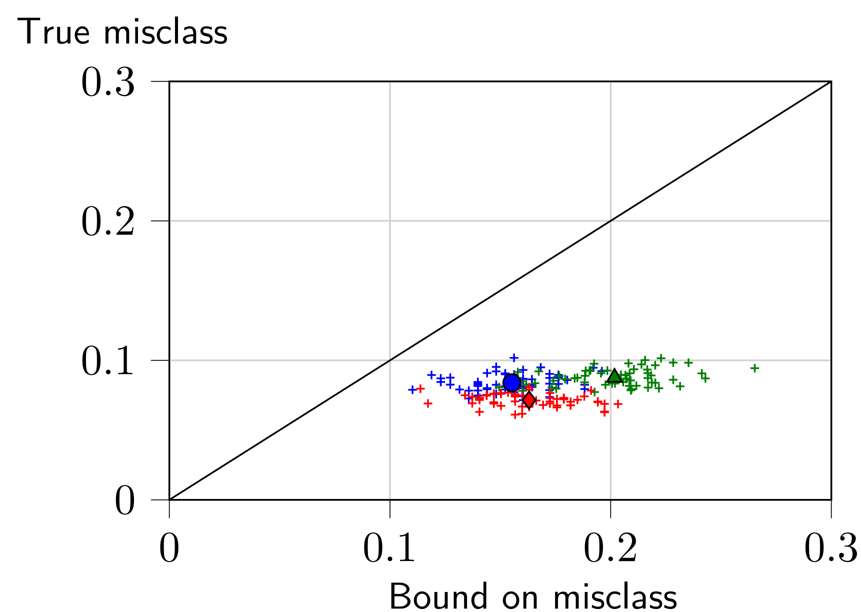
Experiments: MNIST classification

Experiment: binary MNIST, $N = 1000$

Comparison: P2L, Train+Test (TT), Pac-Bayes (PB)
care about both true gen *and* bound







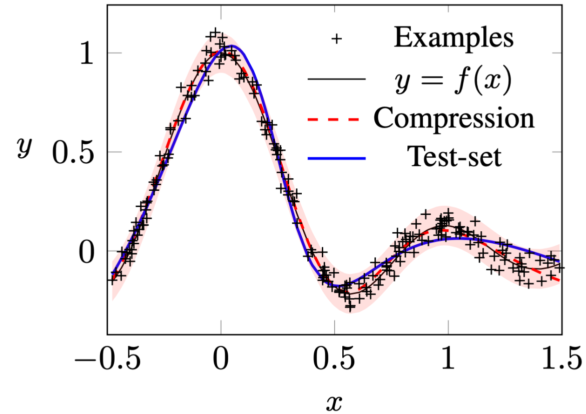
Take home: P2L superior to PB, comparable TT bound *but* better true misclass!

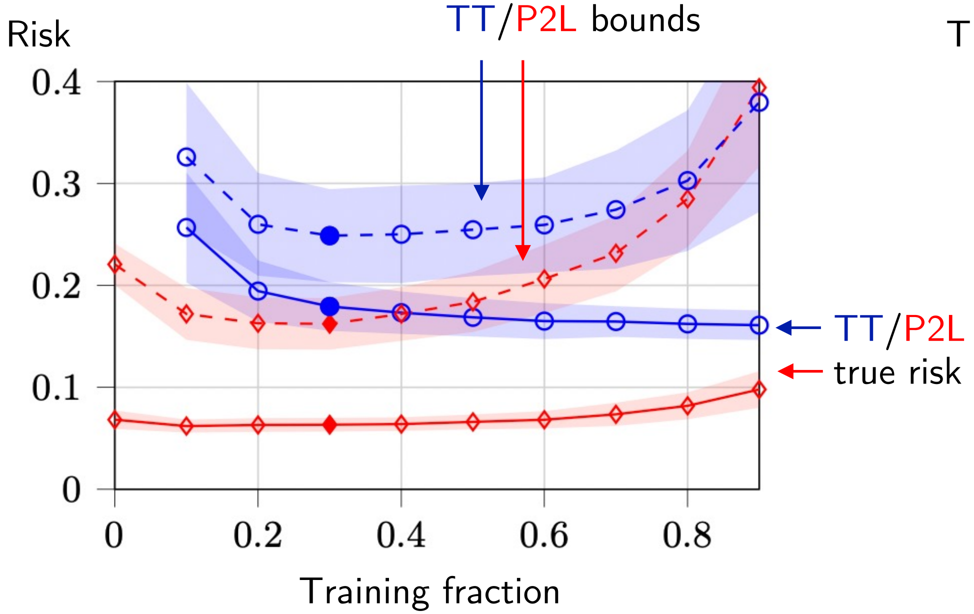
[Dziugate & Roy, 2017] [Perez et al, 2021]

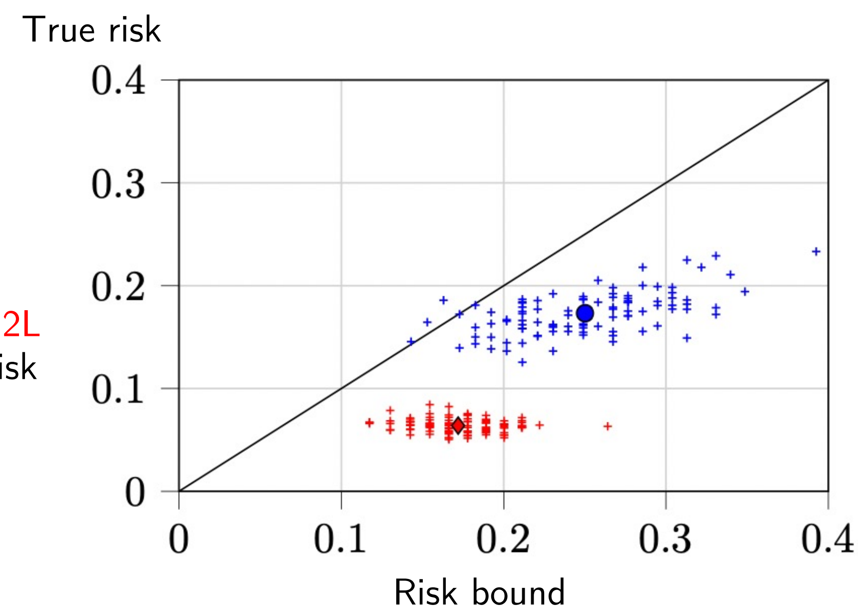
Experiments: regression

Experiment: noisy $\sin(2.5\pi x)/2.5\pi x$, $N = 200$

Comparison: Train+Test-set (TT) vs P2L
care about both perf *and* bound







Take home: P2L beat TT barrier, i.e., good bound *and* true risk!

The Pick-to-Learn Algorithm: Empowering Compression for Tight Generalization Bounds and Improved Post-training Performance

Dario Paccagnan
Dept. of Computing
Imperial College London
d.paccagnan@imperial.ac.uk

Marco C. Campi
Dip. di Ingegnerira dell'Informazione
Università di Brescia
marco.campi@unibs.it

Simone Garatti
Dip. di Elettronica Informazione e Bioingegneria
Politecnico di Milano
simone.garatti@polimi.it

Abstract

Generalization bounds are valuable both for theory and applications. On the one hand, they shed light on the mechanisms that underpin the learning processes; on the other, they certify how well a learned model performs against unseen inputs. In this work we build upon a recent breakthrough in *compression theory* (Campi & Garatti, 2023) to develop a new framework yielding tight generalization bounds of wide practical applicability. The core idea is to embed any given learning algorithm into a suitably-constructed meta-algorithm (here called Pick-to-Learn, P2L) in order to instill desirable compression properties. When applied to the MNIST classification dataset and to a synthetic regression problem, P2L not only attains generalization bounds that compare favorably with the state of the art (test-set and PAC-Bayes bounds), but it also learns models with better post-training performance.

https://openreview.net/forum?id=40L3viVWQN