

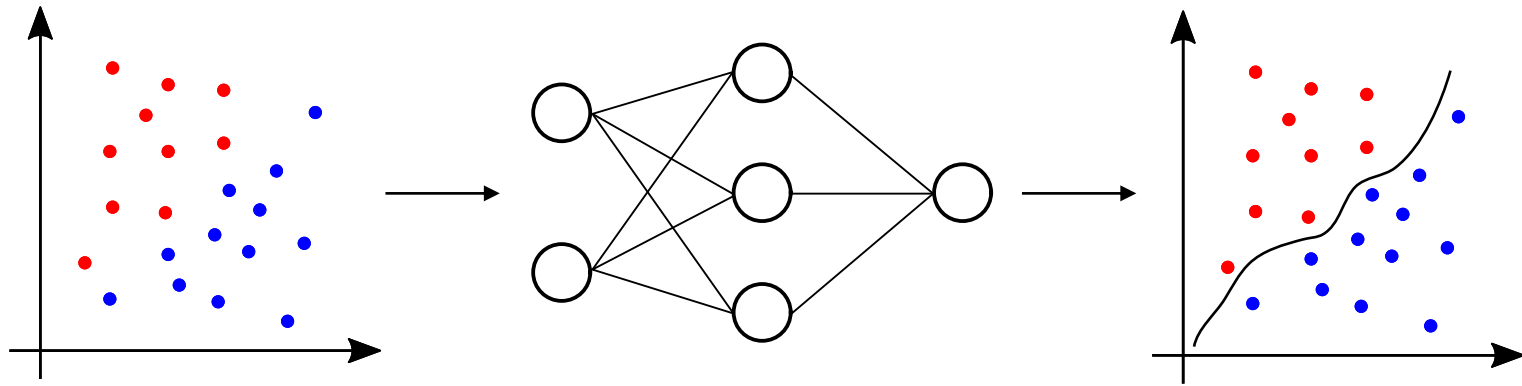
The Pick-to-Learn Algorithm: Empowering Compression for Tight Generalization Bounds & Improved Post-training Performance

Dario Paccagnan, Marco C. Campi, Simone Garatti



NeurIPS 2023 spotlight

Motivation: ML algo with good performance & guarantees



- True generalization
- Bound on generalization

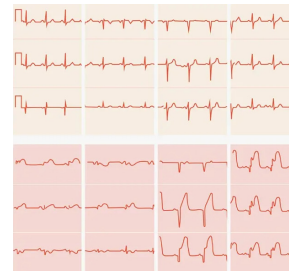
In this work we address both

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

ON THE UNIFORM CONVERGENCE OF RELATIVE FREQUENCIES OF EVENTS TO THEIR PROBABILITIES

V. N. VAPNIK AND A. YA. CHERVONENKIS

(Translated by B. Seckler)

ON LARGE-BATCH TRAINING FOR DEEP LEARNING: GENERALIZATION GAP AND SHARP MINIMA

Nishit Shrikar Keskar
Northwestern University
Evanston, IL 60208
keskar.nishit@northwestern.edu

Dheeravams Mudigere
Intel Corporation
Bangalore, India
@dheeravams.mudigere@intel.com

Jung Taek Lee
Northwestern University
Evanston, IL 60208
jtaek@northwestern.edu

Mihail Simjanov
Intel Corporation
Santa Clara, CA 95054
mihail.simjanov@intel.com

Ping Tak Peter Tang
Intel Corporation
Santa Clara, CA 95054
peter.tang@intel.com

ABSTRACT

The stochastic gradient descent (SGD) method and its variants are algorithms of choice for many Deep Learning tasks. These methods operate in a small batch regime where a fraction of the training data, say 32-512 data points, is sampled to compute an approximation to the gradient. It has been observed in practice that when using a larger batch there is a degradation in the quality of the model, as measured by its ability to generalize. We investigate the cause for this generalization drop in the large batch regime and present numerical evidence that supports the view that large batch methods tend to converge to sharp minimizers of the training and testing functions—and as is well known, sharp minima lead to poorer generalization. In contrast, small batch methods consistently converge to flat minimizers, and our experiments support a commonly held view that this is due to the inherent noise in the gradient estimation. We discuss several strategies to attempt to help large batch methods elude this generalization gap.

1 INTRODUCTION

Deep Learning has emerged as one of the cornerstones of large-scale machine learning. Deep Learning models are used for achieving state-of-the-art results on a wide variety of tasks including computer vision, natural language processing and reinforcement learning; see (Bengio et al., 2016) and the references therein. The problem of training these networks is one of non-convex optimization. Mathematically, this can be represented as

$$\min_{\theta} f(\theta) = \frac{1}{M} \sum_{i=1}^M f_i(\theta), \quad (1)$$

where f_i is a loss function for data point $i \in \{1, 2, \dots, M\}$ which captures the deviation of the model prediction from the data, and θ is the vector of weights being optimized. The process of optimizing this function is also called training of the network. Stochastic Gradient Descent (SGD) (Bottou, 1998; Sutskever et al., 2013) and its variants are often used for training deep networks.

Rademacher and Gaussian Complexities: Risk Bounds and Structural Results

Peter L. Bartlett
Shahar Mendelson
Research School of Information Sciences and Engineering
Australian National University
Canberra 0200, Australia

PETER.BARTLETT@ANU.EDU.AU
SHAHAR@CS.LANL.EDU.AU

Computing Nonconvex Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data

Gintar Karlinas Deigoite
Department of Engineering
University of Cambridge

David M. Roy
Department of Statistical Sciences
University of Toronto

Abstract

One of the defining properties of deep learning is that models are chosen to have many more parameters than available training data. In light of this capacity for overfitting, it is remarkable that simple algorithms like SGD reliably return solutions with low test error. One roadblock to explaining these phenomena in terms of explicit regularization, structural properties of the solution, and/or estimates of the data, is that many learning bounds are quantitatively useless when applied to networks trained by SGD in this “deep learning” regime. Logically, in order to explain generalization, we need nonconvex bounds. We return to an idea by Langford and Caron (2015), who used PAC-Bayes bounds to compute nonconvex numerical bounds on generalization error for stochastic one-layer two-hidden-unit neural networks via a sensitivity analysis. By optimizing the PAC-Bayes bound directly, we are able to extend their approach and obtain nonconvex generalization bounds for deep stochastic neural network classifiers with millions of parameters trained on only tens of thousands of examples. We compare our findings to recent and old work on flat minima and SGD-based explanation of generalization.

1 INTRODUCTION

By optimizing a PAC-Bayes bound, we show that it is possible to compute nonconvex numerical bounds on the generalization error of deep stochastic neural networks with millions of parameters, despite the training data sets being one or more orders of magnitude smaller than the number of parameters. To our knowledge, these are the first explicit and nonconvex numerical bounds proposed for trained neural networks in the modern deep learning regime where the number of network parameters exceeds the number of training examples.

The bounds we compute are data dependent, incorporating millions of hypotheses optimized essentially to minimize a large region in weight space with low average empirical error around the solution obtained by stochastic gradient descent (SGD). The data dependence is essential: indeed, the VC dimension of neural networks is typically bounded below by the number of parameters, and so one needs a many training data as parameters before (convex) PAC bounds are nonvacuous, i.e., before the generalization error falls below 1. To put this in concrete terms, on MNIST having over 70 hidden units in a fully connected that layer yields success PAC bounds.

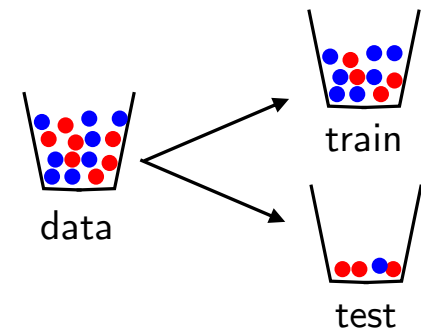
Even so, we are operating far from the worst case: observed generalization cannot be explained in terms the regularizing effect of the size of the neural network alone. This is an old observation, and one that attracted considerable theoretical attention two decades ago. Bartlett (1997; 1999) showed that, in large (infinite) neural networks, when the learned weights are small in magnitude, the flat-sharpening dimension is more important than the VC dimension for characterizing generalization. In particular, Bartlett established classification error bounds in terms of the empirical margin and the flat-sharpening dimension, and then gave flat-sharpening bounds for neural networks in terms of the magnitude of the weights and the depth of the network alone. Improved bound bounds were obtained using Rademacher and Gaussian complexity by Bartlett and Mendelson (2002) and Koltchinskii and Panchenko (2002).

These nonconvex bounds are the foundation of our current understanding of neural network generalization. It is widely accepted that these bounds explain observed generalization, at least “qualitatively”, and/or that the weights are explicitly regularized. Indeed, more work by Neyshabter, Tomczak, and Srebro (2014) goes both

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

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

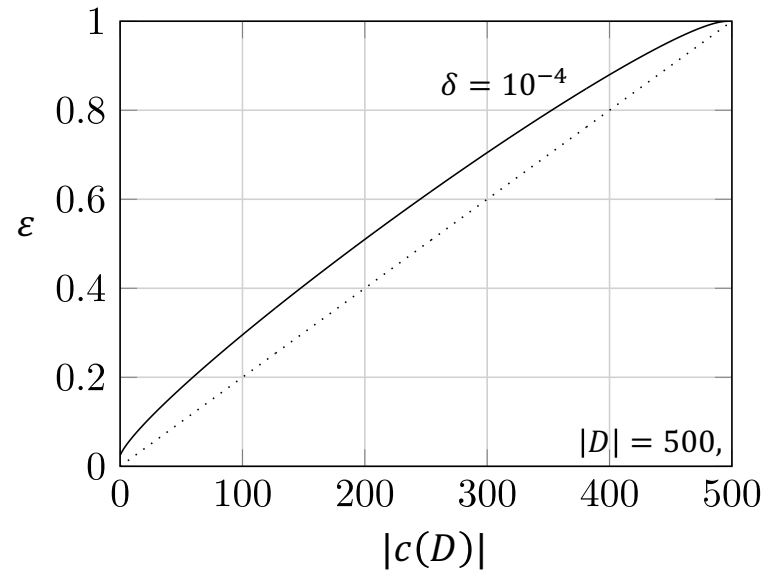
Compression is an established topic in theoretical learning, and various generalization bounds have been proven for compression schemes.

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 $c(S)$, 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 $\rho(c(S))$. 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

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



Challenge: ML algos do not have compression properties

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

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

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

$h \leftarrow L(T)$

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

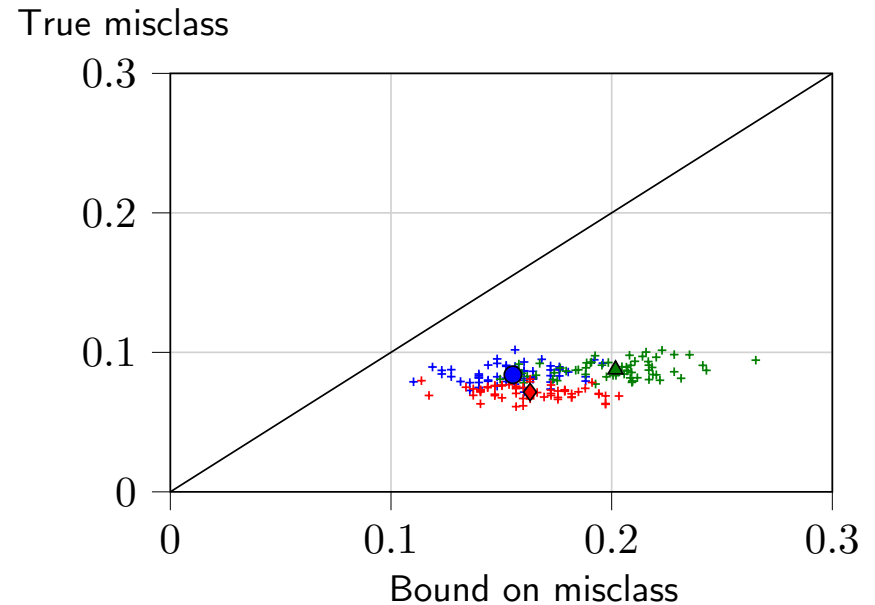
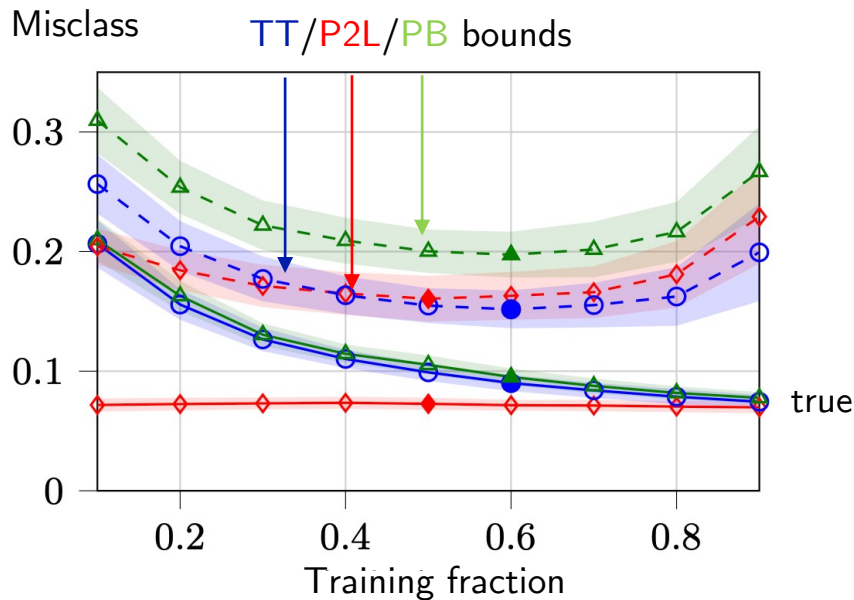
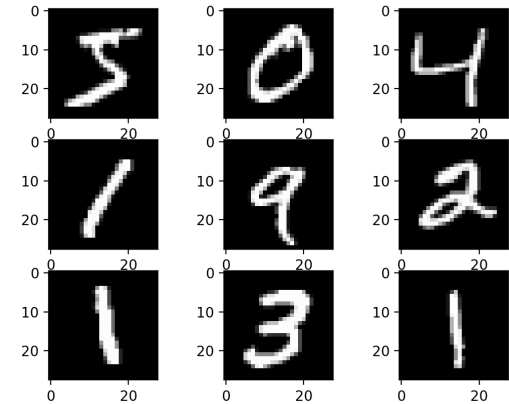
Theorem (informal): P2L is a preferent compression algorithm

Hopes: - P2L compresses “a lot” \Rightarrow good bound on generalization
- P2L does not change the “nature” of L \Rightarrow 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

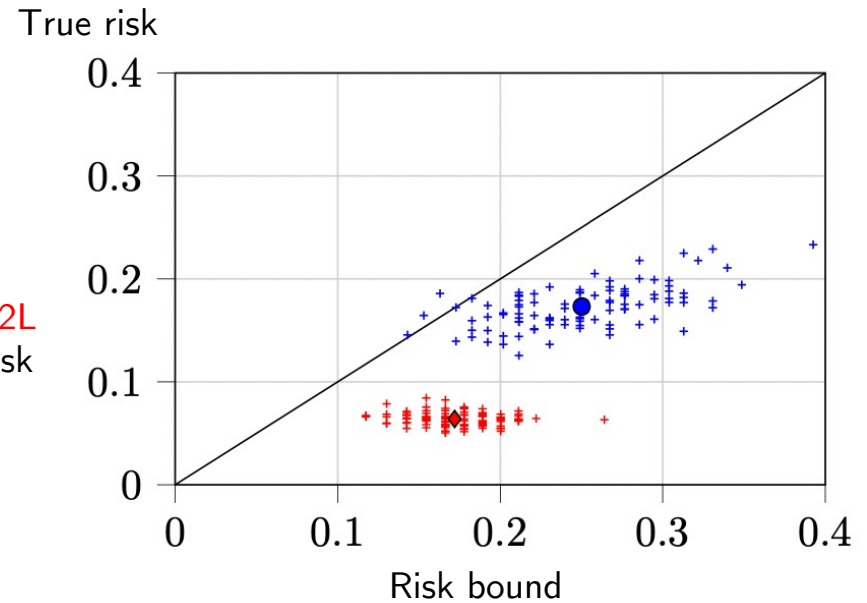
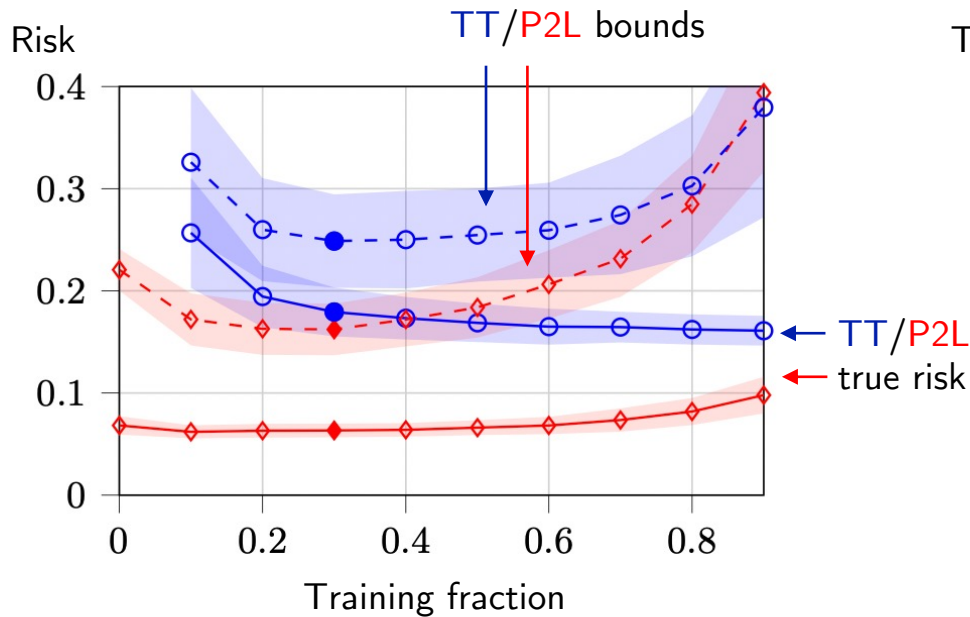
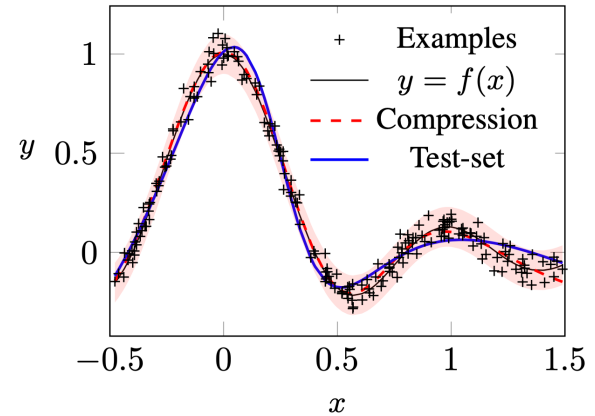


Take home: P2L superior to PB, comparable TT bound *but* better true misclass!

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 Ingegneria 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>