LOSSLESS COMPRESSION & GENERALIZATION IN OVERPARAMETERIZED MODELS: THE CASE OF BOOSTING

Nikolaos Nikolaou
Department of Physics and Astronomy, University College London
Gower Street, London, WC1E 6BT, UK
n.nikolaou@ucl.ac.uk

ABSTRACT

Successful learning algorithms like DNNs, kernel methods or ensemble learning methods, have been known to produce models that exhibit good generalization despite being drawn from overparameterized model families. This observation has put in question the convex relationship between model complexity and generalization. We instead propose rethinking the relevant notion of model complexity for the purposes of assessing the complexity of models trained on a given dataset. Borrowing from information theory, we identify the optimal model one can train on a given dataset as one achieving its *lossless maximal compression*. In the noiseless dataset setting, it can be shown that such a model coincides with an *average margin maximizer* of the training data. Experimental results on gradient boosting confirm our observations and show that the minimal generalization error is attained in expectation by models achieving lossless maximal compression of the training data.

1 INTRODUCTION

Recent theoretical and empirical results have demonstrated that contrary to the traditionally-held belief that *overparameterized models*[1] are likely to overfit, in practice, models produced by contemporary learning algorithms like DNNs, kernel methods or ensemble learning methods, can exhibit good generalization despite being drawn from overparameterized model families Wyner et al. (2015); Belkin et al. (2019); Hastie et al. (2019), even without the use of explicit regularization Bühlmann & Hothorn (2007); Zhang et al. (2016); Kawaguchi et al. (2017). The related "*double descent*" behavior of generalization errors Belkin et al. (2019); Loog et al. (2020); Chen et al. (2020) has revealed situations in which increasing model complexity, causes generalization performance to first deteriorate and then improve (even to do so repeatedly), prompting us to rethink our understanding of the relation between generalization and model complexity.

We propose an alternative approach for reconciling these recent findings to the established theory: by *rethinking the notion of model complexity* that is useful in such a setting. The goal is to compare the *potential of overfitting* of models trained in a supervised manner on a given training dataset and for doing so we need to move beyond naive parameter counts[2] of the trained models or worst-case measures of the richness of their respective model family[3], like the VC-dimension Vapnik & Chervonenkis (2015) or the Rademacher complexity Bartlett & Mendelson (2002).

Taking an information-theoretic perspective to learning, inspired by the *information bottleneck* principle Tishby et al. (2000); Tishby & Zaslavsky (2015); Shwartz-Ziv & Tishby (2017), we treat the features & targets of the training data, as well as the outputs of the trained model as random

[1] Models whose degrees of freedom exceed the number of training datapoints.

[2] Although sparse models tend to be more resistant to overfitting, one can easily come up with counterexamples. Sparsity can make models more robust (in the sense that their output will not change much for small perturbations of their input) but it is neither a necessary, nor a sufficient condition for robustness.

[3] Successful learning algorithms should, after all, have inductive biases that steer them towards models that are robust, regardless of the capacity of the underlying model family.

variables. We then identify the ideal model as one that achieves *lossless maximal compression (LMC)* of the training dataset, i.e. extracts from the features all the useful information for predicting the target and no more. We show that in the *noiseless* binary classification setting, LMC models are equivalent to *average margin maximizers*, models known to exhibit good generalization guarantees Vapnik (1982); Schapire et al. (1998); Sokolić et al. (2017); Dziugaite & Roy (2017); Neyshabur et al. (2017); Wei et al. (2018). In the *noisy* case -when margin maximizers are known to perform poorly Kalai & Servedio (2003); Servedio (2003); Bootkrajang & Kabán (2013); Poggio et al. (2017)- we show that the aforementioned equivalence collapses, as LMC models also capture label uncertainty while margin maximizers are not even guaranteed to be lossless.

We posit that adequate overparameterization guarantees losslessness (in the noiseless case this means interpolation) and the inductive biases of successful machine learning algorithms guide them towards maximal compression, yielding LMC models. This explains their surprisingly good generalization properties Wyner et al. (2015); Belkin et al. (2019); Hastie et al. (2019); Muthukumar et al. (2020); Bartlett et al. (2020); Muthukumar et al. (2020). We support our observations with theoretical arguments and empirical evidence, using gradient boosting as an example and identify interesting directions for future work.

2 AN INFORMATION-THEORETIC VIEW OF MODEL COMPLEXITY

We shall focus on the case of binary classification and we will treat the features X & targets Y of the training dataset S, as well as the outputs (scores) of the trained model F as random variables. Being a deterministic transformation of X, F cannot contain more information[4] than X. So, $H(F|X) = 0 \iff I(X;F) = H(F) \leq H(X)$. We shall now define the following properties:

Noiselessness: The dataset S is noiseless if and only if $H(Y|X) = 0$. Otherwise, S is noisy and $H(Y|X) > 0$.

A noiseless training dataset S is one in which no datapoints with the same feature vector have different labels. For such a dataset, there exists a model that can achieve zero empirical risk (training error), i.e. that can perfectly classify the training data. In other words, the features X, contain all information to perfectly describe the target Y.

Losslessness: The model F is lossless on the dataset S if and only if $I(F;Y) = I(Y;X)$. Otherwise, the model is lossy on S and $I(F;Y) < I(Y;X)$.

A lossless model F on a dataset S is one that captures all the information in features X that is relevant for describing the target Y[5]. If a model F is lossless on a training set S, its output can be used to describe the target Y with the only source of training error being the irreducible class overlap in the training set.

Maximal Compression: The model F is a maximal compressor of the dataset S if and only if $I(F;X) = I(F;Y)$. Otherwise, the model is undercompressed on S and $I(F;X) > I(F;Y)$.

A model F that is a maximal compressor of a training dataset S is one that only captures from the features X information relevant for describing the target Y. It does not necessarily capture *all* that information; this special case, merits a definition of its own given below.

Lossless Maximal Compression - (LMC): The model F is a lossless maximal compressor (LMC) of the training dataset S if and only if it is lossless on S and a maximal compressor on S.

A model F that is an LMC of a training dataset S is one that only captures from the features X all the information relevant for describing the target Y[6]. From an information-theoretic perspective, an LMC of S is the optimal classification model that can be constructed from S.

Average Margin Maximization: A model F is an average margin maximizer of a training dataset S if and only if there exists some invertible transformation g such that $g(F) = Y$.

[4]Information-theoretic quantities refer to estimates obtained on the training data (empirical distribution).

[5]Equivalently: the r.v. F is a *sufficient statistic* of the empirical distribution of the training data.

[6]Equivalently: the r.v. F is a *minimal sufficient statistic* of the empirical distribution of the training data.

In other words, an average margin maximizer would assign the same score s_+ to all positively labelled training examples and the same score s_- to all negatively labelled training examples.

From the above definitions the following propositions can easily be derived Nikolaou et al. (2020):

(I) The model F is an LMC of a training dataset S, if and only if I(F;X) = I(F;Y) = I(Y;X).

(II) Any LMC model on a noiseless dataset S is also an average margin maximizer on S and vice-versa.

(III) In the noisy case, the above equivalence no longer holds. A lossless model is one that also captures the uncertainty introduced by the ambiguous labelling of a feature vector i.e. $P(Y|X)$. So a lossless model should assign different scores F to examples corresponding to different $P(Y|X)$, even if they have the same label Y. An average margin maximizer, assigning only 2 values s_+ to positives & s_- to negatives (regardless of label uncertainty) would therefore not even necessarily be lossless, let alone a LMC on a noisy dataset.

3 Empirical Evidence

3.1 Experimental Setup

Boosting, a method that explicitly maximizes the margins of the training examples, can be shown empirically to also converge to LMC models on noiseless datasets. After lossless maximal compression is achieved, so is the minimal generalization error, as estimated by the error on the test set. To demonstrate this, we plot the *trajectory* of the boosting ensemble on the *entropy-normalized information plane*, $I(F;Y)/H(Y)$ vs. $I(F;X)/H(X)$ Tishby & Zaslavsky (2015); Shwartz-Ziv & Tishby (2017). For each boosting round t, $F = F_t$ denotes the random variable of which the ensemble's outputs are realizations.

The boosting ensemble consisted of a maximum of $T = 100$ decision trees (i.e. rounds of boosting) of maximal depth 6. No shrinkage of the updates or subsampling of the examples was performed (both are techniques to counter overfitting), and the exponential loss function was used (i.e. the loss minimized by AdaBoost). We performed no hyperparameter optimization. Plotting trajectories on the information plane follows Tishby & Zaslavsky (2015); Shwartz-Ziv & Tishby (2017). All information-theoretic quantities were estimated on the training data by first discretizing the features & model outputs in $b = 100$ equal-sized bins[7], then using maximum likelihood estimators. The joint r.v. X was then constructed by the discretized features $X_1, X_2, \ldots, X_d$ as $X = \sum_{i=1}^{d} X_i b^{i-1}$. We plot average results across 100 runs with different train-test splits (50%–50%) on the same original data. We also visualize the trajectories obtained by some random individual runs to showcase that although they can vary significantly from one another, they all follow the same general pattern[8].

3.2 Experimental Results

In Figure 1 we present some example trajectories on noiseless datasets. Below follows a summary of our observations:

Boosting leads to lossless maximal compression: The boosting ensemble traces a trajectory on the information plane that leads to the LMC point and once it reaches it in never escapes.

Lossless maximal compression coincides with margin maximization: The image on the information plane of the models that minimize the margin coincides with the LMC point.

Lossless maximal compression coincides with maximal generalization: The point of the ensemble's trajectory corresponding to the minimal test error coincides –on average– with the LMC point on the information plane (and so does the margin maximization point).

Average trajectory shape: After the training error is minimized, the test error can be further

[7]Note that by discretizing the features we might convert an originally noiseless dataset into a noisy one. In the experiments included in this paper this did not happen for any dataset for the numbers of discretization bins chosen. So all results shown are on noiseless datasets.

[8]The full theoretical & experimental results along with a detailed description of the datasets used and the experimental setup can be found in Nikolaou et al. (2020). All datasets & code used in the experiments can be downloaded at `https://github.com/nnikolaou/margin_maximization_LMC`.

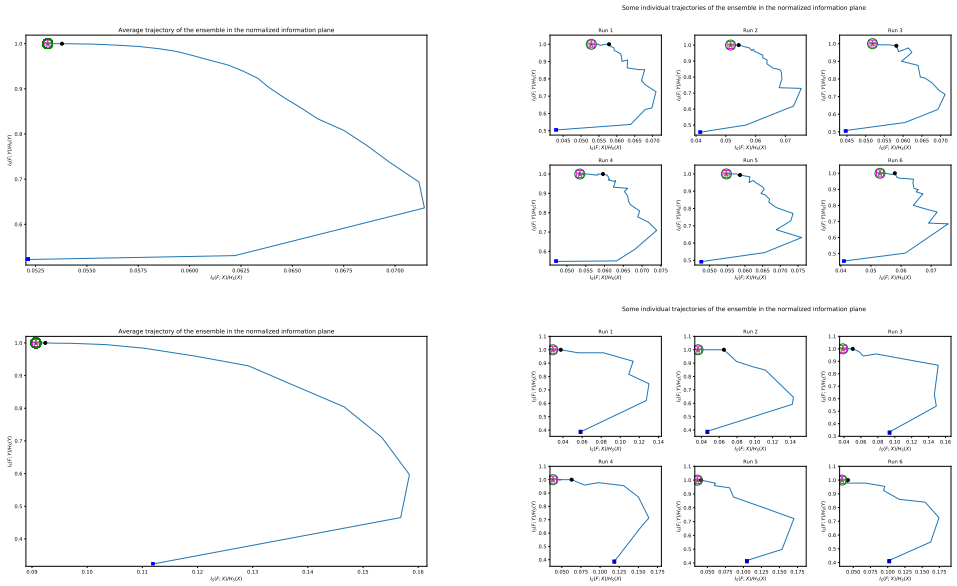

Figure 1: Trajectory of the boosting ensemble on the information plane. We highlight the point of the first model in the ensemble (blue rectangle), the point on which the training error is first minimized (full black circle), the point on which the test error is first minimized (magenta square), the point on which the margins are first maximized (hollow green circle) and the lossless maximal compression point (red star) on the [TOP] *musk* & [BOTTOM] *credit* datasets (both noiseless) from the UCI repository. LEFT: Average trajectory across 100 runs; RIGHT: Some individual trajectories.

decreased by training for more rounds –a known result in boosting, explained via margin theory Schapire et al. (1998). Here we give an information-theoretic interpretation. Training until training error minimization, amounts to achieving losslessness. Subsequent rounds result in travelling along the line of maximal $I(F;X)$ on the information plane, towards the LMC point. This compresses the model (relieves it of remaining information from X irrelevant for predicting Y), decreasing its effective complexity[9].

Training in boosting consists of 2 (typically distinct) phases: A similar behaviour was observed in Shwartz-Ziv & Tishby (2017) for the trajectories of the representations learned by DNNs. Following the terminology of Shwartz-Ziv & Tishby (2017), these are the *empirical risk minimization (ERM) phase*, when $I(F;Y)$ increases (the model better fits the training data) but typically so does $I(F;X)$ (the model uses more information from X) and the *compression phase*, when $I(F;X)$ decreases (the model uses increasingly less information from X, reducing its effective complexity), without decreasing $I(F;Y)$. The ERM phase is usually much shorter than the compression phase, as is the case with DNNs Shwartz-Ziv & Tishby (2017).

Early stopping does not improve generalization in gradient boosting: As long as losslessness can be achieved, additional boosting rounds do not hurt generalization. Once the model reaches the LMC point on the information plane, it never escapes it. This suggests that early stopping with boosting is unnecessary for improving generalization. This result agrees with recent observations in boosting Wyner et al. (2015), DNNs Belkin et al. (2019) and also linear models trained under general margin losses Soudry et al. (2017).

Consistency across datasets, hyperparameter & discretization settings: These observations hold across different datasets, hyperparameter settings and entropy estimation choices.

Margin maximization as a built-in regularization mechanism: Regularization methods like subsampling or shrinkage are not the main reason why boosting regularizes. Their contribution is small compared to the algorithm's built-in regularization mechanism: margin maximization, which as we saw amounts to lossless maximal compression of the training dataset. DNNs have also been

[9]Holds for average trajectories. Single runs include steps that both increase $I(F;X)$ & decrease $I(F;Y)$.

observed to perform implicit regularization and additional regularization control (e.g. dropout or batch normalization) not to be the main contributor to their good generalization Zhang et al. (2016); Shwartz-Ziv & Tishby (2017); Kawaguchi et al. (2017).

4 DISCUSSION

We characterized from an information theoretic perspective, models trained on a given training set w.r.t. the information they capture from it. We identified an ideal model trained on a given dataset as its lossless maximal compressor (LMC): one capturing all the information from the features relevant for predicting the target and no more. We established that an LMC is –in the case of noiseless classification– equivalent to an average margin maximizer of the dataset. In the noisy case the above equivalence collapses and average margin maximizers are suboptimal from an information-theoretic point.

Our experiments on gradient boosting, demonstrate that indeed, margin maximization amounts to lossless maximal compression on noiseless data. The evolution of the model constructed by boosting, traces a trajectory on the information plane that leads to the LMC point which also coincides with the point of margin maximization and the point on average exhibiting the best generalization.

This work gives an information-theoretic interpretation of margin maximization and provides us with a principled way to define model complexity for the purposes of generalization, thus shedding more light on the success of methods like gradient boosting and identifying situations in which they would underperform. It also opens various directions for future work. For instance, exploring how these concepts can be applied in model selection or to inform learning algorithm design to more efficiently traverse the information plane to reach the LMC point. It would also be of interest to identify the analogue of the LMC in learning tasks other than classification, like ranking or regression.

ACKNOWLEDGEMENTS

This project was partially supported by the EPSRC Doctoral Prize Fellowship at the University of Manchester and the EU Horizon 2020 research & innovation programme [grant No 758892, ExoAI]. N. Nikolaou also graciously acknowledges the support of the NVIDIA Corporation's GPU grant.

REFERENCES

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Jakramate Bootkrajang and Ata Kabán. Boosting in the presence of label noise. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 82–91. AUAI Press, 2013.

Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, pp. 477–505, 2007.

Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*, 2020.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Adam Kalai and Rocco A Servedio. Boosting in the presence of noise. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pp. 195–205. ACM, 2003.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.

Marco Loog, Tom Viering, Alexander Mey, Jesse H Krijthe, and David MJ Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626, 2020.

V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020. doi: 10.1109/JSAIT.2020.2984716.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5943–5952, 2017.

Nikolaos Nikolaou, Henry Reeve, and Gavin Brown. Margin maximization as lossless maximal compression. *arXiv preprint arXiv:2001.10318*, 2020.

Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, and Hrushikesh Mhaskar. Theory of deep learning iii: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.

Robert E Schapire, Yoav Freund, Peter Bartlett, Wee Sun Lee, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

Rocco A Servedio. Smooth boosting and learning with malicious noise. *Journal of Machine Learning Research*, 4(Sep):633–648, 2003.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel RD Rodrigues. Generalization error of deep neural networks: Role of classification margin and data structure. In *Sampling Theory and Applications (SampTA), 2017 International Conference on*, pp. 147–151. IEEE, 2017.

Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *Information Theory Workshop (ITW), 2015 IEEE*, pp. 1–5. IEEE, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Vladimir Vapnik. Estimation of dependences based on empirical data. *Springer Series in Statistics*, 1982.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pp. 11–30. Springer, 2015.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. On the margin theory of feedforward neural networks. *arXiv preprint arXiv:1810.05369*, 2018.

Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *arXiv preprint arXiv:1504.07676*, 2015.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.