Open Problem: Information Complexity of VC Learning

Thomas Steinke CMI@THOMAS-STEINKE.NET
IBM Research – Almaden

Lydia Zakynthinou ZAKYNTHINOU.L@NORTHEASTERN.EDU
Khoury College of Computer Sciences, Northeastern University

Editors: Jacob Abernethy and Shivani Agarwal

1. Overview

A recent line of work has studied the information complexity of learning (Russo and Zou, 2016; Raginsky et al., 2016; Alabdulmohsin, 2016; Feldman and Steinke, 2018; Bassily et al., 2018; Dwork et al., 2015a; Rogers et al., 2016; Smith, 2017; Xu and Raginsky, 2017; Nachum and Yehudayoff, 2018; Nachum et al., 2018; Esposito et al., 2019; Bu et al., 2019, etc.). That is, we ask: *How much information does the output of a learner reveal about its input?* A natural example of a quantity that can be used to quantify the revealed information is the mutual information $I(A(Z); Z)$ between the output of an algorithm A and its input Z (consisting of i.i.d. samples from an unknown distribution).

Measuring the information complexity of a learning algorithm can be very informative, as it is related to several properties or guarantees that we might wish to establish. In particular, low information complexity entails generalization guarantees. That is, it implies that that the loss of the output hypothesis on the input dataset is close to its loss on the distribution (Russo and Zou, 2016; Xu and Raginsky, 2017). Conversely, overfitting entails high information complexity. Moreover, these generalization guarantees are robust to postprocessing, thanks to the data processing inequality.

The information complexity of an algorithm is also closely related to the study of privacy. Differential privacy (Dwork et al., 2006) has been widely-accepted as a robust guarantee that a hypothesis or model produced by a machine learning algorithm does not reveal sensitive personal information contained in the training data. Differential privacy is known to imply low mutual information (McGregor et al., 2010; Bun and Steinke, 2016), as well as strong generalization guarantees (Dwork et al., 2015b; Bassily et al., 2016; Jung et al., 2019).

The celebrated theory of uniform convergence (Vapnik and Chervonenkis, 1971) approaches learning from a different perspective by studying the complexity of hypothesis classes. A hypothesis class has the uniform convergence property if, with high probability over the drawing of an i.i.d. dataset, all hypotheses simultaneously generalize – i.e., their loss on the dataset and on the distribution are similar. In particular, hypothesis classes with bounded Vapnik-Chervonenkis (VC) dimension exhibit strong uniform convergence, which implies sample-efficient PAC learning and agnostic learning (Vapnik and Chervonenkis, 1971; Talagrand, 1994; Alon et al., 1997).

Uniform convergence (and VC dimension) make no reference to the learning algorithm itself, instead they only depend on its range. On the other hand, information complexity is a property of a particular algorithm, which also implies generalization. So it is natural to ask whether there is

a bridge between VC dimension and information complexity, despite the incompatibility of their definitions. The main question that arises is:

Do all classes with bounded VC dimension admit a learner with low information complexity?

Unfortunately, Bassily et al. (2018) showed that *any* proper and consistent learner for threshold functions on an unbounded domain must have unbounded mutual information (for worst-case distributions). In contrast, the VC dimension of threshold functions is 1, which implies strong uniform convergence bounds even for infinite domains. So, if we restrict our learning algorithms to the natural class of empirical risk minimizers (ERMs), and measure information complexity with respect to mutual information, this impossibility result gives a negative answer to the question.

Recently, Steinke and Zakynthinou (2020) proposed a more refined measure of information complexity via conditional mutual information (CMI), which, like mutual information, also implies generalization. The authors showed that thresholds admit a proper consistent learner with constant CMI. So, measuring the information complexity of an algorithm with respect to CMI allows us to overcome the impossibility result of Bassily et al. (2018).

In general, Steinke and Zakynthinou (2020) showed that any hypothesis class with VC dimension d admits an ERM with CMI $O(d \log n)$, where n is the size of the input sample. This gives a positive answer to our question. However, the generalization guarantees that one can retrieve for this learner via the $O(d \log n)$ CMI bound do not match the tight bounds guaranteed by uniform convergence.

We conjecture that it is possible to attain CMI $O(d)$ instead, that is, to prove that for any class with VC dimension d there exists a learner (perhaps not necessarily an ERM), which has CMI $O(d)$. Proving such a bound would entail tight generalization bounds for VC classes via CMI.

2. Formal definitions and problem statement

We consider the standard setting of statistical learning (Valiant, 1984; Haussler, 1992; Kearns et al., 1994). There is an unknown probability distribution $\mathcal{D}$ over some known set $\mathcal{Z}$. We have access to a sample $Z \in \mathcal{Z}^n$ consisting of n independent draws from $\mathcal{D}$. There is a function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ and our goal is to find some $w_* \in \mathcal{W}$ that approximately minimizes $\ell(w_*, \mathcal{D}) := \mathop{\mathbb{E}}_{Z' \leftarrow \mathcal{D}} [\ell(w_*, Z')]$.

However, we cannot evaluate the true loss (a.k.a. "population loss" or "risk") $\ell(w_*, \mathcal{D})$ since the distribution $\mathcal{D}$ is unknown. Instead we can compute the empirical loss (a.k.a. "empirical risk") $\ell(w_*, Z) := \frac{1}{n} \sum_{i=1}^{n} \ell(w_*, Z_i)$ using the sample Z, which also leads to the natural learning strategy of "Empirical Risk Minimization (ERM)" – i.e., $w_* = \arg\min_{w \in \mathcal{W}} \ell(w, Z)$. In the supervised machine learning setting, $\mathcal{W}$ is a class of functions $w : \mathcal{X} \to \mathcal{Y}$ (e.g., $\mathcal{Y} = \{0, 1\}$ for binary classification). We consider the 0-1 loss $\ell : \mathcal{W} \times (\mathcal{X} \times \mathcal{Y}) \to \{0, 1\}$, $\ell(h, (x, y)) = 0 \Leftrightarrow h(x) = y$.

2.1. Uniform Convergence

The classical theory of uniform convergence (Vapnik and Chervonenkis, 1971) approaches the problem of generalization in this setting by studying the class of functions $\mathcal{H} := \{\ell(w, \cdot) : w \in \mathcal{W}\}$. VC dimension is a property of a hypothesis class, which implies uniform convergence:

Definition 1 (Vapnik-Chervonenkis dimension, Vapnik and Chervonenkis (1971)) *Let $\mathcal{W}$ be a class of functions $w : \mathcal{X} \to \{0, 1\}$. The VC dimension of $\mathcal{W}$ is the largest natural number d such that there exist $x_1, \cdots, x_d \in \mathcal{X}$ and $w_1, \cdots, w_{2^d} \in \mathcal{W}$ such that, for each $j, k \in [2^d]$ with $j \neq k$, there exists some $i \in [d]$ such that $w_j(x_i) \neq w_k(x_i)$.*

Bounded VC dimension implies that, for a sample of adequate size (depending linearly on the VC dimension), the true and empirical errors of *all* hypotheses will be close.

Theorem 2 (Talagrand (1994); Blumer et al. (1989)) *There exist positive constants c_1, c_2 such that the following holds. For every class $\mathcal{W}$ of functions $w : \mathcal{X} \to \{0,1\}$ of VC dimension $d \geq 1$, for every $n \geq 2d$, and for every distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$, we have*

$$\mathop{\mathbb{E}}_{Z \leftarrow \mathcal{D}^n} \left[\sup_{w \in \mathcal{W}} |\ell(w, \mathcal{D}) - \ell(w, Z)| \right] \leq c_1 \cdot \sqrt{\frac{d}{n}}$$

and

$$\mathop{\mathbb{E}}_{Z \leftarrow \mathcal{D}^n} \left[\sup_{w \in \mathcal{W} : \ell(w,Z)=0} \ell(w, \mathcal{D}) \right] \leq c_2 \cdot \frac{d}{n} \cdot \log\left(\frac{n}{d}\right).$$

2.2. Information Complexity of Learning via CMI

Intuitively, the problem with measuring information complexity via mutual information is that the information content of a single sample is infinite if the distribution is continuous. Thus $I(A(Z); Z)$ is infinite if A reveals even a single input datum. Thus mutual information easily becomes unbounded even when there is no issue for generalization.

Steinke and Zakynthinou (2020) avoid this issue by moving to conditional mutual information (CMI). Intuitively, this "normalizes" the information content of each datum to one bit by conditioning on a superset of the training sample. We state the formal definition:

Definition 3 (Conditional Mutual Information (CMI) of an Algorithm) *Let $A : \mathcal{Z}^n \to \mathcal{W}$ be a randomized or deterministic algorithm. Let $\mathcal{D}$ be a probability distribution on $\mathcal{Z}$ and let $\tilde{Z} \in \mathcal{Z}^{n \times 2}$ consist of $2n$ samples drawn independently from $\mathcal{D}$. Let $S \in \{0,1\}^n$ be uniformly random. Assume S, $\tilde{Z}$ and A are independent. Define $\tilde{Z}_S \in \mathcal{Z}^n$ by $(\tilde{Z}_S)_i = \tilde{Z}_{i,S_i+1}$ for all $i \in [n]$ – that is, $\tilde{Z}_S$ is the subset of $\tilde{Z}$ indexed by S.*

The conditional mutual information (CMI) of A with respect to $\mathcal{D}$ *is*

$$\mathsf{CMI}_{\mathcal{D}}(A) := I(A(\tilde{Z}_S); S | \tilde{Z}).$$

The key property of CMI is, of course, that it implies generalization.

Theorem 4 (Steinke and Zakynthinou 2020, Cor. 5.2, Thm. 5.7) *Let $A : \mathcal{Z}^n \to \mathcal{W}$, $\ell : \mathcal{W} \times \mathcal{Z} \to [0,1]$, and let $\mathcal{D}$ be a distribution on $\mathcal{Z}$. Then*

$$\mathop{\mathbb{E}}_{Z \leftarrow \mathcal{D}^n, A} [\ell(A(Z), Z) - \ell(A(Z), \mathcal{D})] \leq \sqrt{\frac{2}{n} \cdot \mathsf{CMI}_{\mathcal{D}}(A)}$$

Furthermore, if $\mathop{\mathbb{E}}_{Z \leftarrow \mathcal{D}^n, A} [\ell(A(Z), Z)] = 0$, then

$$\mathop{\mathbb{E}}_{Z \leftarrow \mathcal{D}^n, A} [\ell(A(Z), \mathcal{D})] \leq \frac{\mathsf{CMI}_{\mathcal{D}}(A)}{n \cdot \log 2}.$$

For any class with bounded VC dimension, there exists an ERM with low CMI.

Theorem 5 (Steinke and Zakynthinou 2020, Thm. 4.12) *Let $\mathcal{Z} = \mathcal{X} \times \{0,1\}$ and let $\mathcal{W} = \{w : \mathcal{X} \to \{0,1\}\}$ be a hypothesis class with VC dimension d. Then, there exists an empirical risk minimizer $A : \mathcal{Z}^n \to \mathcal{W}$ such that $\mathsf{CMI}(A) \leq d \log n + 2$.*

2.3. Problem Statement

Theorem 5 has a $\log n$ term in the CMI bound. We conjecture that this can be removed.

We can combine Theorems 4 and 5 to obtain a generalization bound for VC classes. However, compared to Theorem 2, this is suboptimal by the $\log n$ factor.

Conjecture 6 *There exists an absolute constant c such that the following holds. For every class $\mathcal{W}$ of functions $w : \mathcal{X} \to \{0,1\}$ of VC dimension $d \geq 1$ and for every $n \geq d$, there exists an algorithm $A : (\mathcal{X} \times \{0,1\})^n \to \mathcal{W}$ such that $\mathsf{CMI}_{\mathcal{D}}(A) \leq c \cdot d$ for every distribution $\mathcal{D}$ and*

$$\forall z \in (\mathcal{X} \times \{0,1\})^n \quad \mathop{\mathbb{E}}_{A}[\ell(A(z), z)] \leq \inf_{w \in \mathcal{W}} \ell(w, z) + c \cdot \sqrt{\frac{d}{n}}.$$

The error we permit in Conjecture 6 corresponds to the error of uniform convergence for a worst-case distribution (Talagrand, 1994). In other words, the empirical error is of the same order as the generalization error.

Conjecture 6 covers the so-called agnostic setting. It may be easier to prove a result for the realizable setting, where we assume that a consistent hypothesis exists:

Conjecture 7 *There exist absolute constants c and c' such that the following holds. For every class $\mathcal{W}$ of functions $w : \mathcal{X} \to \{0,1\}$ of VC dimension $d \geq 1$ and for every $n \geq d$, there exists a randomized or deterministic algorithm $A : (\mathcal{X} \times \{0,1\})^n \to \mathcal{W}$ such that $\mathsf{CMI}_{\mathcal{D}}(A) \leq c \cdot d$ for every distribution $\mathcal{D}$ and, for every $z \in (\mathcal{X} \times \{0,1\})^n$, if there exists $w \in \mathcal{W}$ such that $\ell(w, z) = 0$, then*

$$\mathop{\mathbb{E}}_{A}[\ell(A(z), z)] \leq c' \cdot \frac{d}{n}.$$

Conjecture 8 *Conjecture 7 holds with $c' = 0$.*

Steinke and Zakynthinou (2020) prove Conjecture 8 for the special case of threshold functions on the real line.

The conjectures are stated for proper learners. However, proving them with improper learners would be interesting. (That is, we permit the algorithm to output a function $w : \mathcal{X} \to \{0,1\}$ that is not in the class $\mathcal{W}$.) Hanneke (2016) shows that there exist improper learners that attain optimal generalization error in the realizable setting (i.e., avoiding the logarithmic term in the second part of Theorem 2). This may be a starting point for a proof of Conjecture 7 or 8.

3. Bounty

For a proof of Conjecture 6, we offer a US$500 prize and, for a negative resolution, US$300. For a proof of Conjecture 8, we offer a US$250 prize. For a negative resolution to Conjecture 7, we offer a US$200 prize. For a proof of Conjecture 7 and a disproof of Conjecture 8, there is a prize of US$100. If only one of of Conjectures 7 and 8 is resolved and the other remains open, we offer a US$50 prize. For a proof of Conjecture 6 or Conjecture 8 that only provides an improper learner, we will provide a half prize, but, if this is accompanied by an impossibility result for the proper case, the full prize amount will be awarded. For a proof of Conjecture 6 for the special case of thresholds on the real line, we offer a prize of US$20 or, for an improper learner, US$15 (20% more if the algorithm is efficient and an additional 20% more if it is deterministic).

References

Ibrahim Alabdulmohsin. Uniform generalization, concentration, and adaptive learning. *arXiv preprint arXiv:1608.06072*, 2016.

Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 1046–1059, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4132-5. doi: 10.1145/2897518.2897566. URL http://doi.acm.org/10.1145/2897518.2897566.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 25–55. PMLR, 07–09 Apr 2018. URL http://proceedings.mlr.press/v83/bassily18a.html.

Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, October 1989. ISSN 0004-5411. doi: 10.1145/76359.76371. URL https://doi.org/10.1145/76359.76371.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 587–591. IEEE, 2019.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-32731-2, 978-3-540-32731-8. doi: 10.1007/11681878_14. URL http://dx.doi.org/10.1007/11681878_14.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126. ACM, 2015b.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. A new approach to adaptive data analysis and learning via maximal leakage. *arXiv preprint arXiv:1903.01777*, 2019.

Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 535–544. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/feldman18a.html.

Steve Hanneke. The optimal sample complexity of pac learning. *The Journal of Machine Learning Research*, 17(1):1319–1333, 2016.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. *arXiv preprint arXiv:1909.03577*, 2019.

Michael J Kearns, Umesh Virkumar Vazirani, and Umesh Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 81–90. IEEE, 2010.

Ido Nachum and Amir Yehudayoff. Average-case information complexity of learning. *arXiv preprint arXiv:1811.09923*, 2018.

Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. A direct sum result for the information complexity of learning. *arXiv preprint arXiv:1804.05474*, 2018.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. *2016 IEEE Information Theory Workshop (ITW)*, pages 26–30, 2016.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–494. IEEE, 2016.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/russo16.html.

Adam Smith. Information, privacy and stability in adaptive data analysis. *arXiv preprint arXiv:1706.00820*, 2017.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information, 2020. URL https://arxiv.org/abs/2001.09122.

Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76, 1994.

Leslie G Valiant. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM, 1984.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, January 1971. doi: 10.1137/1116025. URL https://doi.org/10.1137/1116025.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.