Open Problem: How fast can a multiclass test set be overfit?

Vitaly Feldman
Google Research

Roy Frostig
Google Research

Moritz Hardt
University of California, Berkeley

Editors: Alina Beygelzimer and Daniel Hsu

1. Background

Several machine learning benchmarks have shown surprising longevity, such as the ILSVRC 2012 image classification benchmark based on the ImageNet database (Russakovsky et al., 2015). Even though it contains only $50{,}000$ data points, hundreds of results have been reported on this test set. Large-scale hyperparameter tuning and experimental trials across numerous studies likely add thousands of queries to the test data. Despite this excessive data reuse, recent replication studies (Recht et al., 2018, 2019; Yadav and Bottou, 2019) have shown that, while there are significant discrepancies in test results, the best performing models transfer rather gracefully to a newly produced test set collected from the same source according to the same protocol.

To maintain statistical validity, what matters is not only the number of times that a test (or holdout) set has been accessed, but also how it is accessed. Modern machine learning practice is *adaptive* in its nature. Prior information about a model's performance on the test set inevitably influences future modeling choices and hyperparameter settings. Adaptive behavior, in principle, can have a radical effect on generalization.

Standard concentration bounds teach us to expect a maximum error of $O(\sqrt{\log(k)/n})$ when estimating the means of k non-adaptively chosen bounded functions on a data set of size n. However, this upper bound sharply deteriorates to $O(\sqrt{k/n})$ for adaptively chosen functions, an exponential loss in k. Moreover, there exists a sequence of adaptively chosen functions, what we will call an *attack*, that causes an estimation error of $\Omega(\sqrt{k/n})$ (Dwork et al., 2014).

This suggests that, in principle, an analyst can overfit substantially to a test set after issuing relatively few queries to it. Powerful results in *adaptive data analysis* provide sophisticated holdout mechanisms that guarantee better error bounds through noise addition (Dwork et al., 2015) and limited feedback mechanisms (Blum and Hardt, 2015). However, the standard holdout method remains widely used in practice, ranging from machine learning benchmarks and data science competitions to validating scientific research and testing products during development. If the pessimistic bound were indicative of performance in practice, the holdout method would likely be much less useful than it is.

It seems evident that additional factors prevent this worst-case overfitting from happening in practice. In Feldman et al. (2019), the authors demonstrate (theoretically and empirically) that,

in multiclass classification problems, the large number of classes makes it substantially harder to overfit due to test set reuse. To state the known results more formally, we introduce some notation. A classifier is a mapping $f\colon X \to Y$, where $Y = [m] = \{1, \ldots, m\}$ is a discrete set consisting of m classes and X is the data domain. A data set of size n is a tuple $S \in (X \times Y)^n$ consisting of n labeled examples $(x_i, y_i)_{i \in [n]}$, where we assume each point is drawn independently from a fixed underlying population. In our model, we assume that a data analyst can *query* the data set by specifying a classifier $f\colon X \to Y$ and observing its accuracy $\mathsf{acc}_S(f)$ on the data set S, which is simply the fraction of points that are correctly labeled $f(x_i) = y_i$. We denote by $\mathsf{acc}(f) = \Pr\{f(x) = y\}$ the accuracy of f over the underlying population from which (x, y) are drawn. Proceeding in k rounds, the analyst is allowed to specify a function in each round and observe its accuracy on the data set. The function chosen at a round t may depend on all previously revealed information. The analyst builds up a sequence of adaptively chosen functions $f_1, \ldots, f_k$ in this manner.

We are interested in the largest value that $\mathsf{acc}_S(f_t) - \mathsf{acc}(f_t)$ can attain over all $1 \leq t \leq k$. Our theoretical analysis focuses on the worst case setting where an analyst has no prior knowledge (or, equivalently, has a uniform prior) over the correct label of each point in the test set. In this setting, the highest expected accuracy achievable on the unknown distribution is $1/m$. In effect, we analyze the expected advantage of the analyst over random guesses.

In reality, an analyst typically has substantial prior knowledge about the labels and starts out with a far stronger classifier than one that predicts at random. Using domain information, models, and training data, there are many conceivable ways to label many points with high accuracy and to pare down the set of labels for points the remaining points. Indeed, the experiments in Feldman et al. (2019) explore a few techniques for reducing label uncertainty given a good baseline classifier. After incorporating all prior information, there is usually still a large set of points for which there remains high uncertainty over the correct label. Effectively, to translate the theoretical bounds to a practical context, it is useful to think of the dataset size n as the number of point that are hard to classify, and to think of the class count m as a number of (roughly equally likely) candidate labels for those points.

Two bounds on the achievable bias in terms of the number of queries k, the number of data points n, and the number of classes m are relevant to the open problem.

Theorem 1.1 (informal) *There is a distribution P over examples labeled by m classes such that any algorithm that makes at most k accuracy queries to a dataset $S \sim P^n$ must satisfy with high probability*

$$\max_{1 \leq t \leq k} \mathsf{acc}_S(f_t) = \frac{1}{m} + O\left(\max\left\{\sqrt{\frac{k \log n}{nm}}, \frac{k \log n}{n}\right\}\right).$$

This bound has two regimes that emerge from the concentration properties of the binomial distribution. The more important regime for our discussion is when $k = \tilde{O}(n/m)$ for which the bound is $\tilde{O}(\sqrt{k/(nm)})$. In other words, achieving a particular bias requires $O(m)$ more queries than it would in the binary case.

On the other hand, we describe a query strategy that achieves the following bound on the bias.

Theorem 1.2 (Point-wise attack) *For sufficiently large n and $n \geq k \geq k_{\min} = O(m \log m)$ there is an attack that uses k accuracy queries and, on any dataset S, outputs f such that*

$$\mathsf{acc}_S(f) = \frac{1}{m} + \Omega\left(\sqrt{\frac{k}{nm^2}}\right).$$

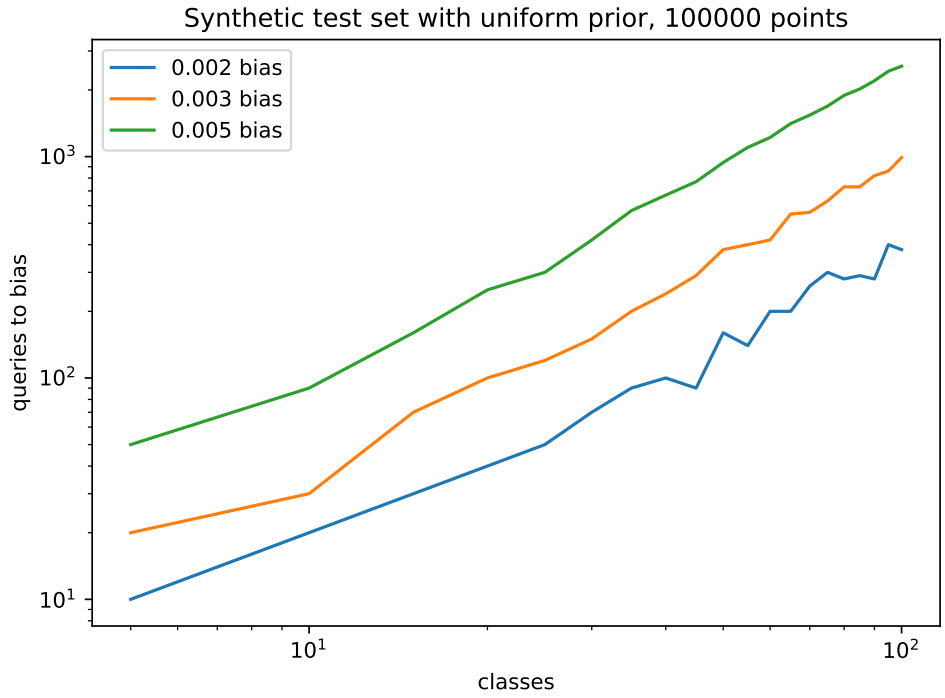

Figure 1: The number of queries at which a fixed advantage over $1/m$ is first attained, while the number of class labels m varies, on a randomly generated test set of size 100,000. The endpoints of the curves form slopes (under the log-log axis scaling) of roughly 1.2 (for the 0.002 bias curve) and 1.3 (for the other curves).

The algorithm underlying Theorem 1.2 outputs a classifier that computes a weighted plurality of the labels that comprise its queries, with weights determined by the per-query accuracies observed. Such an attack is rather natural, in that it resembles boosting and other common techniques for model aggregation. In addition, this attack is computationally efficient and we prove that it is optimal within a broad class of attacks that we call *point-wise*. Roughly speaking, such an attack predicts a label independently for each data point rather than reasoning jointly over the labels of multiple points in the test set.

2. Open problem

As can be seen from the description of the bounds, there is a quadratic gap in the dependence of the bias on the number of classes. Specifically, our upper bound suggests that the number of queries needed to achieve certain bias grows linearly with the number of classes m. In contrast, our best attack can only guarantee much worse growth: k needs to scale as m^2 to achieve the same bias. The open problem is to close this gap or at least improve either of the bounds. More concretely,

Open Problem 2.1 *Show whether there exists an algorithm that, for any distribution P over labeled examples with m classes, submits k accuracy queries to the test set S consisting of n examples*

sampled i.i.d. from P, and produces f such that

$$\mathop{\mathbf{E}}_{S\sim P^n}[\mathsf{acc}_S(f)] = \frac{1}{m} + \tilde{\Omega}\left(\sqrt{\frac{k}{nm}}\right).$$

Naturally, this bound can only hold under additional restrictions on the range of parameters. It would be sufficient to cover the regime where $k \leq n/m$ since beyond that regime tight upper and lower bounds are given in Feldman et al. (2019). From a more practical point of view, one should also restrict the attention to efficient algorithms. Still, the answer is not known even without this restriction. The answer is also not known for an even stronger class of attack algorithms that have access to points (but not labels) in the dataset S. An attack of this type is described in Feldman et al. (2019). For any S and $k = \Omega(m \log m)$ it outputs f such that:

$$\mathsf{acc}_S(f) = \min\left\{1, \frac{1}{m} + \Omega\left(\frac{k\log(k/m)}{n\log m}\right)\right\}.$$

Figure 1 is based on a simulation of the attack underlying Theorem 1.2. It shows the number of queries at which a fixed advantage over $1/m$ is first attained, while the number of class labels m varies, on a randomly generated test set of size 100,000. The endpoints of the curves in the figure form lines of slope greater than 1 on a log-log scale. This might suggest that, to attain a fixed bias using the attack underlying Theorem 1.2, the number of queries k must indeed grow super-linearly with m.

References

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. *CoRR*, abs/1502.04585, 2015. URL http://arxiv.org/abs/1502.04585.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *CoRR*, abs/1411.2664, 2014. Extended abstract in STOC 2015.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375. URL http://www.sciencemag.org/content/349/6248/636.abstract.

Vitaly Feldman, Roy Frostig, and Moritz Hardt. The advantages of multiple classes for reducing overfitting from test set reuse. In *ICML*, 2019. To appear.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? *CoRR*, abs/1806.00451, 2018. Study was extended to ImageNet in subsequent unpublished work (to appear shortly). Personal communication by the authors.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? *CoRR*, abs/1902.10811, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Chhavi Yadav and Léon Bottou. Cold case: The lost MNIST digits. 2019.