# Speed and accuracy are not enough! Trustworthy machine learning

**Shiva Kaul**
Computer Science Department
Carnegie Mellon University
https://www.cs.cmu.edu/~skkaul/

**Abstract**

Classical linear/shallow learning is relatively easy to analyze and understand, but the power of deep learning is often desirable. I am developing a hybrid approach in order to obtain learning algorithms that are both trustworthy and accurate. My research has mostly focused on learning from corrupted or inconsistent training data ('agnostic learning'). Recently, I, as well as independent researchers, have found these same techniques could help make algorithms more fair.

Learning algorithms are primarily evaluated by how much time they take to run, and by the accuracy of their resulting decisions. By these criteria, deep learning has largely surpassed classical methods such as SVM and kernel methods. However, as algorithms are used for more critical and sensitive decisions — like whether a drug should be administered to a patient, or a loan applicant should be approved, or an autonomous vehicle should stop for pedestrians — we desire properties such as the following:

- resilience to corrupted or inconsistent training data,
- robustness to adversarial manipulation of test data, and
- fairness, accountability, and/or transparency of the resulting decisions.

I consider a learning algorithm "trustworthy" if it has these properties[1]. The first two properties involve adversarial reactions to the algorithm which may invalidate the initial training assumptions; the last involves unforeseen consequences of using the algorithm. These properties can't be ensured by treating the algorithm as a 'black box' and observing its performance on more data. We gain trust in the algorithm by understanding and analyzing it.

Assessing learning algorithms in terms of trustworthiness, along the traditional criteria of speed and accuracy, establishes a tradeoff between simplicity and power. Classical linear/shallow learning tends to be more trustworthy but slower or less accurate. Deep learning is relatively opaque and complex, despite a rapidly developing theory. In the context of the first property ("resilience to corrupted or inconsistent training data"), also known as agnostic learning, the conceptual simplicity/power tradeoff is quantitatively formulated and studied. The members of my thesis committee — Avrim Blum, Maria-Florina Balcan, Geoffrey Gordon, and Varun Kanade — are experts in this field. The central problem is called 'agnostically learning halfspaces" — learning, from corrupted or inconsistent training data, a classifier which is as accurate as the best linear classifier. Even if there is a linear classifier which is nearly consistent with the data, it may be computationally intractable (i.e. NP-hard) to actually find it. Using nonlinear classifiers (a technique called 'improper learning') can circumvent this computational intractability, but may require an overwhelming amount of data, or may introduce other computational difficulties. To make substantial progress in agnostic learning, we need to figure out how to utilize nonlinear classifiers without a concomitant explosion in complexity.

For more background on these tradeoffs, especially the tradeoff between linear and nonlinear classifiers in the context of agnostic learning, see my full thesis proposal (Kaul 2016). Here, I will briefly highlight my progress on taming these tradeoffs, and speculate on its relevance to other aspects of trustworthiness, particularly fairness.

## Depth without distress

My thesis features a new learning algorithm, called the sequence of averages (SoA), which returns a new kind of classifier, called smooth lists of halfspaces. They are both just a few lines of code (see the figure below). Though they are deep, they eschew much of the complexity of existing approaches. Smooth lists of halfspaces are nonlinear, but do not perform intermediate feature extraction (mapping the input vector to another vector). SoA is iterative, but does not involve parameters optimized by backpropagation.

Due to their simplicity, the classifier and algorithm have useful properties which aren't (provably) shared by their forebears. Smooth lists of halfspaces do not use more data (in the worst case) than linear classifiers, as quantified by Rademacher complexity bounds ((Kaul 2017) theorems 1 and 2). This alleviates concerns about overfitting that usually arise with improper learning. Though SoA does not update parameters like a typical iterative algorithm, it can be analyzed as a concrete dynamical system. Using techniques from this area, particularly the stable manifold theorem, we have shown the algorithm (or at least an arbitrarily minor smoothing thereof) doesn't get prematurely stuck: it can always

[1] I am not entirely attached to this terminology. As an alternative, in a forthcoming talk at NIPS, Moritz Hardt collectively refers to such properties as "safety beyond security".

For $t = 1, \ldots, T$:
  $w_t \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i y_i$
  $w_t \leftarrow \beta_t (w_t / ||w_t||)$
  $y_i \leftarrow e^{-|\langle w_t, x_i \rangle|} y_i$
Return $w_1, \ldots, w_T$

For $t = 1, \ldots, T$:
  With probability $1 - e^{-|\langle w_t, x \rangle|}$, return $\mathrm{sgn}(\langle w_t, x \rangle)$
Return $-1$ or $1$ uniformly at random.

Figure 1: The learning algorithm (left) operates upon data $\{x_i, y_i\}_{i=1}^m$ for $T$ iterations, and computes a sequence of averages. It uses a sequence of positive step sizes $\{\beta_t\}_{t=1}^T$. On each iteration, it computes, rescales (by Euclidean norm $||\cdot||$), and stores the average of all the data. It reduces the weight of data having high inner product with the average. The weight reduction corresponds to a passing probability in the smooth list of halfspaces (right), which operates upon an input $x$. A stored average $w_t$ is used to classify an input if they have high inner product; otherwise, the input is passed to the next average.

monotonically improve the classifier ((Kaul 2017) theorem 3). Similar convergence proofs for nonconvex optimization algorithms require much more stringent assumptions about the underlying data or objective (Lee et al. 2016).

Nonlinear smooth lists of halfspaces grant the algorithm just a bit of flexibility beyond linear classifiers. Nevertheless, it is enough to experimentally achieve state-of-the-art results on various problems originating in computational learning theory. These involve fitting boolean functions called juntas, including the notoriously challenging parity (aka 'XOR') function. These results are evidence that SoA avoids the typical compromises associated with nonlinear classifiers.

In future work, we hope to exhibit data which may be (weakly) fit by SoA, but provably cannot be fit by any efficient linear classifier. This would show that SoA is strictly more powerful than efficient linear classifiers. We also believe that replacing linear classifiers with smooth lists could make them more robust to adversarial manipulation of test data. Finally, we hope our techniques may be used to help make algorithms more fair, as described in the next section.

## Fairness and computation

Unlike speed or accuracy, there is no obvious definition of fairness for a learning algorithm. Many definitions have recently been proposed. Rather than converging to a single 'consensus' definition, researchers are analyzing the relationships among them, and examining the tradeoffs required to fulfill them. Many of these tradeoffs are surprisingly harsh, suggesting that simple notions of fairness cannot coincide (Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2017). Similarly, an interesting connection is developing between agnostic learning and fairness: two recent works have demonstrated that simultaneously ensuring fairness for many groups is algorithmically equivalent to agnostic learning (Kearns et al. 2017; Hébert-Johnson et al. 2017). Considering the difficulty of agnostic learning, these should be interpreted as further negative results.

I believe these results portend a deeper interplay between computational learning theory and fairness. Techniques for coping with corrupted or inconsistent data may help guarantee fairness. Sacrificing simplicity in favor of power will become even less defensible. This will affect not just the techniques used to guarantee fairness, but also the notions of fairness themselves. Useful definitions will have to strike a balance between ethical/moral significance and computational feasibility.

My submission to AIES illustrates how lessons from computational learning theory, and my thesis research in particular, may carry over to fairness (Kaul and Gordon 2017). From an algorithmic perspective, directly optimizing outcomes (e.g. 'accept' or 'reject') is hard, so most learning algorithms instead optimize corresponding real-valued quantities (e.g. 'high score' or 'low score'). I propose definitions of equal opportunity in terms of such real-valued quantities. I think these definitions are more ethically/morally significant in addition to being computationally expedient. I show that a linear classifier based on averaging — much like a single step of the SoA algorithm above — fulfills these definitions of equal opportunity more effectively than popular classifiers such as SVM.

In future work, I hope to develop algorithms which guarantee fairness to many groups, but avoid the aforementioned difficulties encountered with other definitions.

## References


Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*.

Hébert-Johnson, Ú.; Kim, M. P.; Reingold, O.; and Rothblum, G. N. 2017. Calibration for the (Computationally-Identifiable) Masses. *ArXiv e-prints*.

Kaul, S., and Gordon, G. 2017. Margins and opportunity. Submitted to AAAI/AIES 2018.

Kaul, S. 2016. Fast agnostic classification. `http://www.cs.cmu.edu/~skkaul/proposal.pdf`.

Kaul, S. 2017. Depth without distress. Submitted to ALT 2018.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2017. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *ArXiv e-prints*.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lee, J. D.; Simchowitz, M.; Jordan, M. I.; and Recht, B. 2016. Gradient descent only converges to minimizers. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, 1246–1257.