Rigorous Statistical Methods for Trustworthy Guarantees in Fair Machine Learning and Beyond

Cyrus Cousins

Fall 2021

Research Statement

Introduction My research has always been broadly interdisciplinary, with a focus on introducing rigorous modern statistical techniques and finite-sample concentration-of-measure analyses into existing problem domains of real-world interest. I have worked as such in *computational biology* [6], *traditional machine learning* [11, 2, 7, 14, 13], *empirical game theory* [18, 21, 19, 20], *data science settings* [1, 15, 9, 5], and *Markov-chain Monte-Carlo analysis* [5, 13], though recently my focus has been *fair machine learning* [3, 10, 4].

In many cases, the state-of-the-art had not advanced beyond asymptotic Gaussian CLT, heuristic permutation tests, or extremely loose Hoeffding-union bounds. I improved existing techniques by introducing more sophisticated analyses, improving the rigour of asymptotic and approximate methods with finite-sample guarantees, and sharpening existing finite-sample bounds to nearly match asymptotic lower-bounds. To do so, I have introduced several new analytical tools, such as the *empirically centralized Rademacher average* [8], which I showed to close the asymptotic gap between upper and lower bounds for the uniform convergence of empirical means to their expectations, and the *inter-trace variance* [5], which similarly closes the gap for mean estimation from dependent samples (traces) drawn from Markov chains.

Overall, my goal has been to push these types of analyses as far as possible, matching lower bounds where appropriate, *without sacrificing rigorous finite-sample guarantees.* This has been rewarding and theoretically interesting, as it tells us which problems are actually hard (as opposed to which problems merely *appear difficult* due to analysis artifacts), and practically it has also been extremely fruitful, as many data science and sampling algorithms are limited by the amount of data that must be collected or processed. As a practical vignette, in [9] I showed that three estimators for the *betweenness centrality*[1] of all graph vertices, namely the RK, ABRA, and BP estimators, can be contrasted directly and sharply vis-à-vis their sample complexities, as measured both by Rademacher averages and (per-vertex) variances, whereas prior work showed only loose and mutually-incomparable upper-bounds on the sample complexity of each.[2] Historically, many consider VC-theoretic, entropy integral, and other learning-theoretic bounds to be elegant and interesting, but of little practical use, due to prohibitively large constant factors on sample complexities, and much of my work aims to challenge this viewpoint, both via lower-bound comparisons and practical experimentation.

In particular, a major theme that arises in my work is taking various minimax (knowledge-aware) sample complexity *lower bounds*, and showing knowledge-agnostic analyses and algorithms with matching or nearly matching sample complexity. Lacking a priori-knowledge of, e.g., variances, Rademacher averages, or (sharp bounds on) mixing times, these algorithms necessarily have dynamic sample-consumption (via progressive sampling), yet still they yield high-probability sample-complexity *upper-bounds*, matching (knowledge-aware) lower-bounds to within small constant factors and log-log terms (to account for early termination in progressive sampling schedules). This arises, for instance, in my work on *uniform convergence* bounds with *centralized Rademacher averages* [8], and also in variance-aware bounds over independent or rapidly-mixing Markovian processes [12, 5, 13]. In all cases, the interesting part of the work is showing that, even though we don't assume knowledge of key quantities in such minimax-lower bounds, we can still nearly match them, essentially by replacing dependence on a priori knowledge with appropriate data-dependent estimates. The main challenge

[1] Roughly speaking, the betweenness centrality of vertex v is the proportion of shortest paths in a graph that contain v.

[2] Ironically, it is the BP estimator, proposed years before RK or ABRA, which has the best statistical performance, but this was unclear without applying sharper statistical analysis techniques and contrasting lower and upper bounds.

that arises is that coarse application of weak concentration of measure bounds (i.e., Azuma-Hoeffding bounds) generally introduce *asymptotic gaps* in sample complexity, but in my work more careful analysis in many cases closes these gaps (up to log-log factors).

Research Ethics As a young student, I largely thought researchers and educators were essentially ethically neutral parties, as the consequences of their research and teaching largely depend on what their students and readers do with their knowledge. With a greater appreciation for the harmful consequences of industrial and governmental application of modern technology (in particular machine learning) in recent years, and the simple fact that the deep understanding of scholars in technical fields put them in a unique position to commentate and contemplate the ethical dimensions of their work, I now view ethical considerations as a core component of scholarly responsibility. For these reasons, I have chosen to shift my focus to clearly ethical research, with immediate and obvious real-world social-good impact.

However, it is my position that this does not mean my work needs to be any less technically rigorous or academically interesting, as I have striven to bring the same rigor and techniques from my statistical learning theory research into the field of fair machine learning. This has opened up many interesting research questions, largely untreated by the theoretical community, and only heuristically addressed in the existing fairness literature. I see great value in placing fairness-sensitive analysis on firm theoretical ground, as without theoretical guarantees, it's quite difficult to determine if a putatively-fair method will produce subtly-biased or unfair results in new real-world applications.

Applications in Fair Machine Learning While on the face of it, a statistical-learning-theoretic perspective on fair machine learning may seem unorthodox, as problems of bias and fairness can be hard to precisely characterize mathematically. Indeed, much of the existing work has focused on showing practical issues with existing systems (w.r.t. both data bias and lack of appreciation for how naïve learning objectives may yield unfairness), and experimental support for new methods. However, I argue that this theoretical perspective is invaluable, as it both leads to stronger guarantees in the fairness domain, and it also pushes us to precisely define and explore what we want fairness to mean.

In my work, I seek to apply statistical learning tools to rigorously prove fairness guarantees in machine learning. Arguably, historic failures of fairness in machine learning have largely occurred due to lack of understanding and rigorous analysis in existing methods, so I believe putting fair learning methods on sturdy theoretical ground with techniques from learning theory is valuable not only as an academic exercise, but also as a paradigm to instill confidence in the fairness of learning methods. I see this as indispensable, as what is the history of machine learning, if not the endless cycle of developing new methods models, recognizing their potential to overfit, and analyzing the phenomenon, only to replace them with the next generation and begin anew? After all, many have argued [16, 17, 3], what good is showing that a model obeys a demographic parity constraint on the training set, if one can't also guarantee that (at least with w.h.p.) a similar fairness condition holds over the underlying distribution?

In fair machine learning, I argue that overfitting is manifest not only as *generalization error* on risk or objective values, but also in *overfitting to fairness*, wherein models *appear fair* in training, but fail to be so in reality. For example, fair training methods, e.g., optimizing (constrained) objectives sensitive to protected group membership, may yield *selection bias*, and thus fairness constraints are violated, true welfare is significantly lower than on the training set, or model performance is substantially worse for some groups. Consequently, although such methods may improve fairness, training performance and constraints overstate their efficacy, and statistical analysis is the tool that allows us to understand the degree to which fairness in training may degrade after deployment.

In this line of inquiry, I seek to show that it is computationally tractable to optimize a model w.r.t. some notion of fairness, and that the "empirical fairness" of a model is reflective of its "true fairness" over a (similarly distributed) population. From a welfare-theoretic perspective, this reduces to showing that a measure of group-wellbeing (i.e., minimax-optimality over the average loss among multiple groups) can be optimized efficiently (i.e., bounding *optimization error*), and also that the hypothesis class is sufficiently simple so as not to "overfit to fairness" (i.e., bounding *sampling* or *generalization error*). These goals closely align with bounding optimization and generalization error in standard machine learning settings, though here our objective function is not generally a simple average statistic, so we require tools beyond standard mean-bounds and uniform convergence.

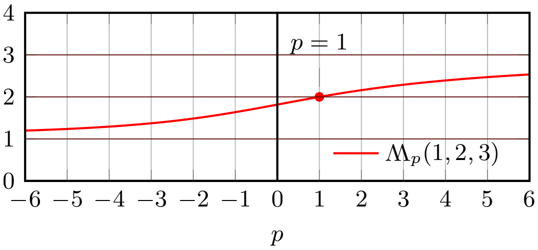

Figure 1: Power-mean malfare $\mathrm{M}_p(1, 2, 3)$, plotted as a function of p. Note $p \geq 1$ is required for *malfare functions*, but $p \leq 1$ are appropriate for use as *welfare functions*.

In [3], the fairness concept is *malfare*, an analog of cardinal welfare used to measure overall disutility across a population of g groups, given some disutility vector $\mathcal{S} \in \mathbb{R}^g_{0+}$ where $\mathcal{S}_i$ denotes the dissatisfication of group i. Subject to appropriate axioms, malfare is restricted to the *p-power-mean* family (see figure 1)

$$\mathrm{M}_p(\mathcal{S}) \doteq \sqrt[p]{\frac{1}{g}\sum_{i=1}^{g} \mathcal{S}_i} \ , \tag{1}$$

for $p \geq 1$, or weightings thereof. In machine learning settings, I take $\mathcal{S}_i$ to be the *risk* (expected loss) of group i over some distribution $\mathcal{D}_i$, thus $\mathrm{M}_p(\mathcal{S})$ summarizes overall risk across the entire population. Here $p = 1$ yields the *arithmetic mean* risk (utilitarian), and taking $p \to \infty$ converges to the worst-case or *maximum* risk (egalitarian), thus p controls the societal concept of fairness, or in other words, the degree to which we favor improving a model for groups most negatively impacted by it, at the expense of overall performance.

A hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$ is *fair-PAC-learnable* w.r.t. some loss function ℓ if for any confidence parameters $\varepsilon > 0$, $\delta \in (0, 1)$, with bounded sample complexity $m(\varepsilon, \delta)$, an algorithm can, given any malfare function $\mathrm{M}(\cdot)$ and per-group instance distributions $\mathcal{D}_{1:g}$, learn a hypothesis $\hat{h} \in \mathcal{H}$ s.t. with probability at least $1 - \delta$, it holds

$$\mathrm{M}\left(i \mapsto \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\left[\ell(\hat{h}(x), y)\right]\right) \leq \varepsilon + \inf_{h^* \in \mathcal{H}} \mathrm{M}\left(i \mapsto \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\left[\ell(h^*(x), y)\right]\right) \ . \tag{2}$$

For $g = 1$ and the 0-1 loss, this reduces to standard *PAC-learnability*, as $\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\ell(h^*(x), y)\right]$ is the expected loss (risk) for group i on distribution $\mathcal{D}_i$.

I argue that this is a clean generalization of PAC-learning and risk minimization, and with the specific choices of axioms made in [3], we obtain an interesting class containing fair variants of many classic machine learning models, including but not limited to those with convex risk objectives exhibiting uniform convergence under mild regularity constraints (e.g., SVM, GLM, convex combinations of experts, etc.). Despite the apparent simplicity of the idea, it is often difficult to get comparable guarantees in other fair learning settings. For example, for *fairness-constrained* optimization, the sample complexity of identifying whether a hypothesis satisfies a fairness constraint on the underlying distribution can be unbounded, and for *welfare maximization*, important welfare functions like the Nash social welfare $\mathrm{W}_0(\mathcal{S}) \doteq \sqrt[g]{\prod_{i=1}^{g} \mathcal{S}_i}$ can be Lipschitz discontinuous in $\mathcal{S}$, potentially also resulting in unbounded sample complexity.

The Case of the Missing Data: On Fair Learning, Data Quality, and Robust Learning Guarantees In [14], I address the highly practical problem that real-world data often either lack or have only partial label information. In particular, I use various semisupervised methods to obtain partial label information, including leveraging statistics of *weak labelers* and a *small labeled training set*. I then establish a *feasible set* $\boldsymbol{Y}$ of possible labelings $\boldsymbol{y} \in \mathcal{Y}^m$ for an unlabeled training set $\boldsymbol{x} \in \mathcal{X}^m$, and cast risk minimization as a minimax optimization problem over the training data, i.e., w.r.t. some loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$, I pose

$$\underbrace{\operatorname*{argmin}_{\theta \in \Theta}}_{\substack{\text{Model}\\\text{Parameters}}} \ \underbrace{\max_{\boldsymbol{y} \in \boldsymbol{Y}}}_{\substack{\text{Adversarial}\\\text{Labeling}}} \ \underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big)}_{\text{Empirical Risk}} \ , \tag{3}$$

wherein we seek to optimize model parameters $\theta \in \Theta$, while an adversary selects the worst possible feasible labeling $\boldsymbol{y}$ given θ. Standard convex optimization routines yield polynomial-time training efficiency guarantees, and I show generalization bounds via (empirical) Rademacher averages and robust minimax estimation bounds, thus addressing both the *computational* and *statistical* aspects of the learning problem.

Crucially, operating over feasible sets and producing minimax guarantees yields hard probabilistic guarantees under minimal assumptions. Such methods circumvent potential *bias issues* due to model misalignment in methods like label imputation, or the independence assumptions of, e.g., Dawid-Skene estimators, theoretical guarantees for both of which generally require difficult-to-validate assumptions.

Data-quality issues often compound unfairness, as majority groups are often well-studied, with copious high-quality data available, while marginalized or minority groups are *understudied*, and available data *lack in quality*. I'm currently working in the setting wherein only partial information is available on protected group membership. In particular, here data are triplets $(x, y, z) \in (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$, where $\mathcal{Z} \doteq \{1, 2, \ldots, g\}$ is a (WLOG finite) space of g *protected groups*. Given m training points, we observe covariates $\boldsymbol{x}_{1:m}$, and labels $\boldsymbol{y}_{1:m}$, *but not* group identities $\boldsymbol{z}_{1:m}$. Instead, we are given a *feasible set* $\boldsymbol{Z}$ of *distributions over* $\mathcal{Z}^m$, i.e., each $\boldsymbol{z} \in \boldsymbol{Z}$ is a right-stochastic matrix $\boldsymbol{z} \in [0, 1]^{m \times g}$, and $\boldsymbol{z}_{i,\cdot}$ (row i) is thus a *distribution* (weighting) over groups.[3] The task is then to perform (group-dependent) fair learning, with rigorous statistical guarantees.

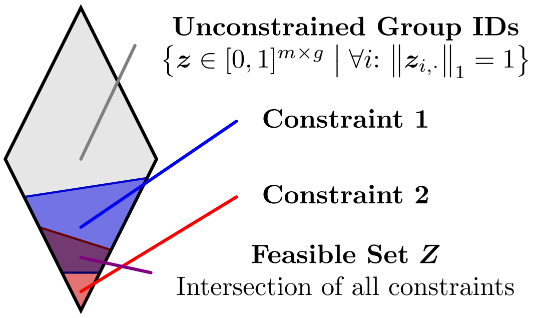
Unconstrained Group IDs
$\left\{\boldsymbol{z} \in [0,1]^{m \times g} \;\middle|\; \forall i\colon \left\|\boldsymbol{z}_{i,\cdot}\right\|_1 = 1\right\}$
Constraint 1
Constraint 2
Feasible Set $\boldsymbol{Z}$
Intersection of all constraints

Figure 2: A feasible set $\boldsymbol{Z}$ of group labelings is generated by the intersection of multiple constraints on the simplicial space of possible per-instance group labelings.

Concretely, in this setting, the *empirical malfare minimization* training objective is

$$\operatorname*{argmin}_{\theta \in \Theta} \max_{\boldsymbol{z} \in \boldsymbol{Z}} \mathrm{M}\left(j \mapsto \frac{1}{\sum_{i=1}^{m} \boldsymbol{z}_{i,j}} \sum_{i=1}^{m} \boldsymbol{z}_{i,j} \ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big) \right) \ , \tag{4}$$

which like (3) may be solved with convex programming when loss is convex in θ.[4] Finally, for fairness-constrained objectives, given fairness statistics $\boldsymbol{C}(\ldots)$ representing constraints on, e.g., ϵ-demographic-parity, ϵ-equality-of-opportunity, etc.,[5]the constrained training objective is (WLOG)

$$\operatorname*{argmin}_{\substack{\theta \in \Theta \text{ s.t.} \\ \forall \boldsymbol{z} \in \boldsymbol{Z}\colon\ \boldsymbol{C}(h_\theta(\cdot), \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \succeq 0}} \frac{1}{m} \sum_{i=1}^{m} \ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big) \ . \tag{5}$$

Optimizing (5) is often computationally difficult, however, for instance when $\boldsymbol{Z}$ is defined by linear constraints, $\boldsymbol{C}(\ldots)$ is a linear function of per-group risk values, and again group proportions are roughly constant, we can (rigorously) approximate $\boldsymbol{C}(\ldots)$ and bound the feasible set with a simple linear constraint set.

In each of (3), (4), and (5), generalization guarantees are possible; these depend both on the degree of overfitting in the model, and the degree of uncertainty in the feasible set $\boldsymbol{Z}$. Overfitting is a function of the amount of unlabeled data m, as well as the complexity of the model class parameterized by $\theta \in \Theta$ (as described by, e.g., a Rademacher average), whereas uncertainty in the feasible set depends on how it is generated (discussed in the sequel). For now, note that as $\boldsymbol{Z}$ contracts to a point, the adversary becomes powerless, and the bounds reflect this, reducing to standard (non-adversarial) guarantees as the diameter of $\boldsymbol{Z}$ or minimax gaps w.r.t. $\boldsymbol{Z}$, Θ and $\ell(\cdot, \cdot)$ tend to 0.

Concretely, this project was motivated by the fact that many datasets do not contain fine-grained gender or race information; for example, the `race` attribute of the classic `adult` dataset contains only {`White`, `Asian-Pac-Islander`, `Amer-Indian-Eskimo`(sic), `Other`, `Black`}, and without more detailed information, it is difficult to even study fairness questions. Suppose we are interested in a fine-grained notion

[3] We generally assume each individual is deterministically associated with a single group, but relaxing this to *weightings over groups* and applying *linear statistical constraints* generates a feasible set $\boldsymbol{Z}$ that is a convex H-polytope. This convex relaxation ensures that, e.g., the inner $\max_{\boldsymbol{z} \in \boldsymbol{Z}}$ operation of (4) is tractable, so long as the objective is concave in $\boldsymbol{z}$.

[4] Technically, due to the ratio $\frac{1}{\sum_{i=1}^{m} \boldsymbol{z}_{i,j}}$, this can be non-convex, but when $\boldsymbol{Z}$ is constrained such that $\frac{1}{\sum_{i=1}^{m} \boldsymbol{z}_{i,j}}$ is roughly constant for all $\boldsymbol{z} \in \boldsymbol{Z}$, $j \in 1, \ldots, g$, it can be approximated efficiently.

[5] Note that to ensure the constraints are satisfied on the underlying distribution (w.h.p.), a sharper constraint must be imposed on the training set, the value of which must itself be a function of the degree of overfitting. We thus consider only *inequality constraints*, as equality constraints are in general impossible to guarantee due to *estimation error*.

of race,[6] e.g., one that subcategorizes `Asian-Pac-Islander` into `South-Asian`, `East-Asian`, etc. The goal of this project is to learn fairly across these groups, and to do so reliably (with generalization guarantees), w.r.t. various fairness concepts, by constraining the set of feasible group memberships $\boldsymbol{z}$ to $\boldsymbol{Z}$, and learning robustly w.r.t. the uncertainty in $\boldsymbol{Z}$.

The given race groups do at least provide *individual-level constraints*, i.e., people listed as `Asian-Pac-Islander` could feasibly be placed into the `South-Asian` or `East-Asian` groups, but likely not the `Black` or `White` groups. Similarly, population demographics imply *probabilistic group-level constraints*, i.e., the proportion of individuals in the dataset belonging to each fine-grained racial group should approximately match the population proportions (w.h.p.), which establishes some dense global linear constraints on group membership, i.e., if such statistics are not respected for a putative group labeling $\boldsymbol{z}$, then $\boldsymbol{z}$ is infeasible.

These constraints leave a large feasible set $\boldsymbol{Z}$, under which minimax learning guarantees may be vacuous, since, e.g., the wealthiest $X\%$ could be placed into one `Asian-Pac-Islander` subgroup, and the most-impoverished remainder into another, or vice versa, while still respecting proportionality constraints. However, additional constraints can alleviate such issues. For instance, covariates like `native-country` can be highly correlated with race, and may thus be used, not by, e.g., assuming all members from country X have race Z, but instead by assuming training-sample proportions should approximately match known statistics on the race of immigrants from each country, and constraining $\boldsymbol{Z}$ appropriately. The process of generating a feasible set $\boldsymbol{Z}$ of group labelings is illustrated in figure 2. More sophisticated methods are indeed possible, employing, e.g., *weak labelers*, *clustering*, and *inter-feature correlation statistics* — such methods are the subject of ongoing work. Let it suffice to say for now that when a group of statistical constraints yields a feasible set $\boldsymbol{Z}$ of small ℓ_1 diameter, i.e., $\frac{1}{2m}\sup_{\boldsymbol{z},\boldsymbol{z}'\in\boldsymbol{Z}}\left\|\boldsymbol{z}-\boldsymbol{z}'\right\|_1 \approx 0$, then robust learning methods can effectively learn fairly, despite incomplete group-membership information.

My Work as a Coherent Research Agenda One can argue, and not baselessly, that the study of fair methods in machine learning is the most real-world impactful research area in computer science today. I've made the case that rigorous theoretical justification of such methods is crucial to any such theory. Still, to claim that the work of machine learning theorists even comes close to spanning the breadth of settings considered by machine learning practitioners would be absurdly naïve. Despite my passion for theoretical machine learning, I appreciate that strong assumptions made in order to develop said theory are at best vague approximations of the assumptions we can safely make or verify in practice, e.g., the i.i.d. assumption, lack of distribution shift, boundedness, etc., and furthermore the availability of labeled data is generally a prerequisite for the theory, and with insufficient data come no guarantees.

In ongoing work, I strive to relax the assumptions made by classical statistical analyses of machine learning problems, letting the balance between what is *needed in practice* and what is *feasible to bound in theory* serve as the lodestone to point to problems I meaningfully contribute to. For example, in addition to my work on semisupervised methods with partial group membership information, I've worked on adapting supervised fair learning guarantees to *reinforcement learning* settings [4], and I'm currently working on adapting them to temporally-dependent settings (e.g., learning from time series or autoregressive data).

The theme running through all of my work is the importance placed on rigorous guarantees for domain-appropriate properties in various probabilistic processes, whether it be for (fair) machine learning tasks, *sampling problems* (mean estimation guarantees), or in *data science* (Nash equilibria, frequent itemsets, betweenness centrality, etc.) settings. In particular, almost all of my papers rigorously show *finite-sample probabilistic guarantees* in some probabilistic setting, and overall my work illustrates that while such guarantees may induce significant proof complexity, often they asymptotically match or nearly match known lower bounds for sample complexity and approximate central-limit-theorem bounds. I argue that this additional analytical overhead is a price well-worth paying, as finite-sample guarantees are a necessary component in understanding the behavior and failure modes of machine learning and data science systems. As we see machine learning increasingly employed in our day-to-day lives, and witness the catastrophic consequences of failures of such systems, the importance of rigorous analysis of these methods to *diagnose*, *understand*, and *avoid* such negative outcomes becomes increasingly apparent.

[6] Let us for now put aside the issues inherent in the socially-constructed definition of race itself, as the goal is to learn while considering predefined racial groups fairly.

References

[1] Carsten Binnig, Fuat Basık, Benedetto Buratti, Ugur Cetintemel, Yeounoh Chung, Andrew Crotty, Cyrus Cousins, Dylan Ebert, Philipp Eichmann, Alex Galakatos, Benjamin Hättasch, Amir Ilkhechi, Tim Kraska, Zeyuan Shang, Isabella Tromba, Arif Usta, Prasetya Utama, Eli Upfal, Linnan Wang, Nathaniel Weir, Robert Zeleznik, and Emanuel Zgraggen. Towards interactive data exploration. In *Real-Time Business Intelligence and Analytics*, pages 177–190. Springer, 2017.

[2] Carsten Binnig, Benedetto Buratti, Yeounoh Chung, Cyrus Cousins, Tim Kraska, Zeyuan Shang, Eli Upfal, Robert Zeleznik, and Emanuel Zgraggen. Towards interactive curation & automatic tuning of ML pipelines. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, pages 1–4, 2018.

[3] Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. In *Advances in Neural Information Processing Systems*, 2021.

[4] Cyrus Cousins, Kavosh Asadi, and Michael Littman. Fair E^3: Efficient welfare-centric fair reinforcement learning. In *Under review; preprint available upon request*, 2021.

[5] Cyrus Cousins, Shahrzad Haddadan, and Eli Upfal. Making mean-estimation more efficient using an MCMC trace variance approach: DynaMITE. *arXiv:2011.11129*, 2020.

[6] Cyrus Cousins, Chirstopher M. Pietras, and Donna K. Slonim. Scalable FRaC variants: Anomaly detection for precision medicine. In *International Parallel and Distributed Processing Symposium Workshops*, pages 253–262. IEEE, 2017.

[7] Cyrus Cousins and Matteo Riondato. CaDET: interpretable parametric conditional density estimation with decision trees and forests. *Machine Learning*, 108(8-9):1613–1634, 2019.

[8] Cyrus Cousins and Matteo Riondato. Sharp uniform convergence bounds through empirical centralization. In *Advances in Neural Information Processing Systems*, 2020.

[9] Cyrus Cousins, Matteo Riondato, and Chloe Wohlgemuth. BAVARIAN: Betweenness centrality approximation with variance-aware Rademacher averages. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2021.

[10] Cyrus Cousins, Clayton Sanford, and Eli Upfal. Welfare-optimal codec selection with uniform convergence bounds. In *Under review; available as Chapter 4 of my dissertation*, 2021.

[11] Cyrus Cousins and Eli Upfal. The k-nearest representatives classifier: A distance-based classifier with strong generalization bounds. In *4th International Conference on Data Science and Advanced Analytics*, pages 1–10. IEEE, 2017.

[12] Cyrus Cousins, Eli Upfal, and Larry Rudolph. BlockSample: Speeding-up data sampling by analyzing blocks of consecutive records. In *Under review; preprint available upon request*, 2021.

[13] Shahrzad Haddadan*, Yue Zhuang*, Cyrus Cousins*, and Eli Upfal. Fast doubly-adaptive MCMC to estimate the Gibbs partition function with weak mixing time bounds. In *Advances in Neural Information Processing Systems*, 2021.

[14] Alessio Mazzetto*, Cyrus Cousins*, Dylan Sam, Stephen H. Bach, and Eli Upfal. Adversarial multiclass learning under weak supervision with performance guarantees. In *International Conference on Machine Learning*, pages 7534–7543. PMLR, 2021.

[15] Leonardo Pellegrina, Cyrus Cousins, Fabio Vandin, and Matteo Riondato. MCRapper: Monte-Carlo Rademacher averages for POSET families and approximate pattern mining. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2165–2174, 2020.

[16] Guy N Rothblum and Gal Yona. Probably approximately metric-fair learning. *arXiv:1803.03242*, 5(2), 2018.

[17] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

[18] Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Learning simulation-based games from data. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2019.

[19] Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Improved algorithms for learning equilibria in simulation-based games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 79–87, 2020.

[20] Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Learning competitive equilibria in noisy combinatorial markets. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 2021.

[21] Enrique Areyan Viqueira, Cyrus Cousins, Yasser Mohammad, and Amy Greenwald. Empirical mechanism design: Designing mechanisms from data. In *Uncertainty in Artificial Intelligence*, pages 1094–1104. PMLR, 2020.