Bounds and Applications of Concentration of Measure in Fair Machine Learning and Data Science

Ph.D. Dissertation Summary

Cyrus Cousins

Brown University Department of Computer Science

Spring 2021

In this thesis, I introduce novel concentration-of-measure bounds for the *supremum deviation*, several *variance concepts*, and a family of *game-theoretic welfare functions*. It is divided into three parts, the first focusing on statistical methods and machine learning, followed by fair machine learning, and concluding with dependent statistical estimation tasks. In particular, I introduce *empirically centralized Rademacher averages* to probabilistically bound the deviations between the *empirical* and *true* means over a *family of functions*, with applications to *multiple comparisons problems* in statistical and scientific settings (smaller p-values and tighter confidence intervals), and to various supervised and unsupervised machine learning settings (reduced sample complexity and sharper generalization bounds). I then show applications of these bounds to various machine-learning, fair machine-learning, and data-science settings, with deep theoretical implications and impactful practical consequences. Parts I and II assume an *independently and identically distributed* (i.i.d.) setting, where we observe many statistically independent occurrences before drawing conclusions, but in closing, Part III extends some of my methods and themes to study non-i.i.d. mean-estimation problems. Naturally, some conclusions are weaker in these relaxed settings, but I find that many of the same data-dependent and variance-sensitivity themes apply, and give practical algorithms for realistic problems where the i.i.d. assumption is prohibitive.

Special care is taken throughout to connect complicated bounds and ideas to simple and intuitive concepts (e.g., relating tail bounds to central limit theorems, or advanced learning models to linear regression), while explaining and rigorously justifying all finite-sample probabilistic guarantees, as well as assumptions and proof techniques. Consequently, at the heart of this thesis is the marriage of simple ideas about intuitive random processes to sophisticated techniques and bounds from the theory of concentration of measure and probabilistic methods, applied in complicated settings to a variety of learning and statistical estimation problems. Topics are chosen both for their theoretical relevance and how well they illustrate or connect such ideas, as well as for their practical relevance and applicability to high-impact real-world problems.

Part I: Concentration of Measure and Uniform Convergence in Machine Learning The first part deals primarily with statistical estimation guarantees in standard machine-learning settings. Chapter 1 introduces the *empirically centralized Rademacher average*, which yields probabilistic data-dependent bounds on the *supremum deviation* (SD) of empirical means of functions in a family $\mathcal{F}$ from their expectations (i.e., $\sup_{f \in \mathcal{F}} \big|\hat{\mathbb{E}}[f] - \mathbb{E}[f]\big|$, the largest absolute difference between *empirical means* and *expectations*), with optimal dependence on the *supremum variance* and the function ranges. Such bounds are impactful in machine learning, as they bound the *generalization gap* between training and test error, and thus control overfitting and quantify the bias-complexity tradeoff.[1] More generally, such uniform convergence bounds have applications to *multiple comparisons problems* in statistical and scientific settings (smaller p-values and tighter confidence intervals), and to various supervised and unsupervised machine learning settings (reduced sample complexity and sharper generalization bounds). Empirical centralization yields *data-dependent* supremum deviation bounds that improve the dependence of *non-centralized* (standard) Rademacher averages on *raw variances* to *centralized variances*, thus matching (asymptotically) known lower-bounds for mean estimation. To compute the bounds in practice, I develop novel tightly-concentrated Monte-Carlo estimators for the empirical Rademacher average of the empirically-centralized family, and show novel concentration results for the empirical supremum variance. My experimental evaluation shows that these bounds greatly outperform their non-centralized counterparts, and are extremely practical, even at small sample sizes.

A major issue, ubiquitous in supervised learning settings, is the cost of obtaining labeled training data. The methods of Chapter 1 may be viewed as a mechanism to get more out of *limited labeled data*, e.g., by showing that far less of it is required than previously thought, but Chapter 2 takes a complimentary and

[1] I.e., the tension at the heart of machine learning between *better fit* and *increased selection bias* as model complexity increases.

orthogonal approach. In Chapter 2, I adopt a *transductive learning* setting, wherein we have a vast set of points, a tiny fraction of which are labeled, and wish to learn to predict the remaining labels. I then describe a transductive learning algorithm that uses these data, alongside additional knowledge in the form of *weak labelers*, which are arbitrarily inaccurate predictors for either the target task or some related task. In particular, I show that accurate learning is possible with a small labeled training set when (a subset of) the family of weak labelers is somewhat well-aligned with the target task. The strategy here is to estimate and (probabilistically) bound appropriately-chosen statistics of the weak labelers using the supervised data, and then consider a *feasible set* of *possible labelings* for the *unlabeled data* that respect these statistics. When sufficient labeled data are available to well-estimate the chosen family of statistics, and the feasible label space is of low diameter, learning the *minimax-optimal* classifier over this feasible set then yields strong generalization guarantees. In particular, if the weak-labeler statistics give sufficient information about the target task, and we have sufficient labeled data to accurately constrain their statistics, then all *feasible labelings* are reasonably accurate, and if sufficiently many *unlabeled data* are available, then we can learn a complicated model; these conditions neatly factor the labeled and unlabeled sample complexities, and describe conditions under which weak-labelers are sufficient to supplement a small labeled training set.

Part II: Fairness with Aggregator Functions: Malfare, Welfare, and Fair-PAC Learning Chapter 3 begins with an axiomatic justification to the *power mean* family[2] of social welfare functions, which summarize societal wellbeing, while making tradeoffs between the needs of the overall population and of marginalized groups (c.f., *utilitarianism* vs *egalitarianism*). From the same axiomatization, starting with a measure of *discontentment* (loss) rather than *contentment* (utility), I derive the parallel concept of *malfare*. For linear welfare functions, malfare acts as the *negative welfare* of the *negative utility*, but malfare extends also to *nonlinear* welfare functions (e.g., the egalitarian welfare, or the geometric mean utility) that are undefined for negative utilities. This is crucial to fairness, as we require nonlinear power-mean malfare functions to specify fairness tradeoffs. These arguments are strongly grounded in the economic theory of *cardinal welfare*, but from them I show statistical estimation and learning guarantees more characteristic of the computer science literature. In particular, malfare is a natural target in machine learning problems where we *minimize* (negatively connoted) loss, rather than *maximize* (positively connoted) utility.

As an application, in Chapter 4, I cast a *streaming-media codec-selection* problem as a *fairness-sensitive* learning problem, wherein we seek to *efficiently select* a *small set* of media-encoders that can mutually satisfy a user-base with *diverse preferences* (e.g., quality vs bandwidth consumption). Optimizing welfare objectives lead to diverse codec choice, whereas without considering fairness and welfare objectives, it is easy to optimize only for a target demographic, or under invalid assumptions on users, making streaming inaccessible and inflexible. Fair codec selection an important *accessibility issue*, as these types of considerations ensure that audiovisual streaming and telecommunications services effectively serve populations that are often sidelined by digital services, including those with limited internet access, as well as those with various audiovisual perception conditions. I explore various *welfare* and *Pareto* optimality concepts, and how the bias-complexity tradeoff manifests in multivariate settings and with fairness issues.

From a more theoretical angle, in Chapter 5, I show *statistical estimation guarantees* for welfare and malfare, and from the *social planning problem*, develop a theory of fair machine learning, based on the *probably approximately correct* (PAC) learning framework, termed *fair-PAC* learning. A fair-PAC learner is an algorithm that learns an ε-δ malfare-optimal model with bounded sample complexity, for *any data distribution*, and for *any* (axiomatically justified) malfare concept. We show broad conditions under which, with appropriate modifications, many standard PAC-learners may be converted to fair-PAC learners. This places fair-PAC learning on firm theoretical ground, as it yields statistical, and in some cases computational, efficiency guarantees for many well-studied machine-learning models. Fair-PAC learning is also practically relevant, as it democratizes fair machine learning, by providing concrete training algorithms and rigorous generalization guarantees for these models.

Part III: Sample-Efficient Mean-Estimation with Dependent Sequential Data Finally, Part III extends the methods and themes of Parts I and II, wherein I assume i.i.d. data samples, into more general *weakly dependent* non-i.i.d. settings. As with the *data-dependent* guarantees using *empirically centralized*

[2] Given g groups, *sentiment vector* $\mathcal{S} \in \mathbb{R}^{g}_{0+}$, and *probability measure* $\boldsymbol{w} \in \mathbb{R}^{g}_{0+}$, the *p-power mean* for $p \geq 0$ is defined as $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \doteq \lim_{\rho \to p} \sqrt[\rho]{\sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i^{\rho}}$. Special cases include $p = 1$, where it reduces to the weighted arithmetic mean (ubiquitous in *risk-minimization*) and $p = \infty$ for $\boldsymbol{w} \succ \mathbf{0}$, where it reduces to the maximum (standard in *robust* or *minimax learning*).

Rademacher averages, the goal here is to get the *strongest guarantees* under the *weakest assumptions*; in particular, this means my algorithms must be sensitive to structure found *in the data*. While notions of *approximate independence* can be difficult to rigorously bound from data, I find that appropriate *variance concepts* are easy to bound, and are sufficient to obtain asymptotically near-optimal sample-complexity guarantees for *mean estimation* in two dependent settings. In particular, I first examine *block databases*, where the key assumption is that *contiguous blocks of records* can be accessed nearly as efficiently as *individual records*, and second, I examine ergodic *Markov chains*, where each step of the chain is a *memoryless* random variable (i.e., conditionally independent from its history given the previous step), and it is as easy to collect a *trace of dependent* (often correlated) samples as it is to collect a pair of near-independent samples.

In both cases, I give *data-dependent* algorithms for ε-δ mean-estimation that avoid *worst-case* sample-complexity behavior. In particular, by averaging not just (approximately) independent samples,[3] but instead *all of the dependent samples* (available at no extra cost), I find that often the *variance* of the mean estimate decreases; particularly so when the data are less dependent than indicated by *a priori* structural assumptions. This decrease in variance implies more rapid convergence of the empirical mean by various central limit theorems, though this work undertakes the significantly more challenging endeavor of showing commensurate improvement to *finite-sample* convergence rates. Furthermore, as I do not assume *a priori* knowledge of the appropriate variance concepts, sufficient sample sizes are not known *a priori*, and thus I employ a *progressive sampling strategy* to avoid drawing too many samples. The details of variance-estimation and necessary multiple-comparisons corrections entailed are subtle, but intuitively, this acts as a "guess and check" method, wherein we optimistically select an initial small sample size, which we use to estimate variances and means, and we iteratively increase it until a sufficiently large sample has been drawn so as to provide the desired guarantee. Surprisingly, despite being variance-oblivious, these strategies are asymptotically optimal, up to log-log factors, in both the block-database and Markov-chain settings. In both cases, *a priori* guarantees on the amount of dependence do appear in our bounds, but only transiently, and as the additive error is taken to 0, terms involving only variance, which is estimated entirely from the data, come to dominate. This calls into question the importance of difficult-to-bound quantities measuring the degree of dependence.

This Work, as a Whole Read separately, each part represents a significant advancement in its respective field; the first in *statistical learning theory*, with an emphasis on *algorithms* and *generalization guarantees* requiring less *labeled data* than previous methods; the second in the *axiomatic philosophy*, *practice*, and *statistical learning guarantees* of *fair machine learning*; and the third in showing that themes and bounds from *sampling problems* in standard i.i.d. settings translate well into non-i.i.d. settings. Special attention is paid to keep each part readable outside of the greater context of this work, however the reader will better appreciate thematic connections, applications, and technical synergy when they are considered as a whole.

To make this concrete, I note that the methods of Part I and Part III are not mutually incompatible, opening the door to strong *uniform convergence bounds* in *non-i.i.d.* settings, and furthermore, the *fair learning setting* of Part II obviously benefits from the statistical bounds of Part I, but similar analysis is certainly possible in non-i.i.d. settings. Indeed, in a connected and interdependent world, it may be the case that practical fair systems need to consider the intricacies of non-i.i.d. learning, and ultimately the importance of philosophically grounded and statistically rigorous fair-learning systems, operating on real-world data with all the messy dependence structures that may entail, just may be exactly what is needed, not only to bring new deep and interesting problems to the computer science community, but also to solve problems of algorithmic and natural injustice and unfairness in the world at large.

In Conclusion Part I introduces new statistical techniques for variance-sensitive uniform convergence bounds and generalization guarantees in machine learning. The centralization strategy of Chapter 1 achieves asymptotically-optimal bound convergence rates, and the Monte-Carlo estimation procedure yields sharp bounds with both *centralized* and *non-centralized* (standard) empirical Rademacher averages. This has broad implications in machine learning, data science, and statistical settings, and can also be used as a component statistical method, mutatis mutandis, in all subsequent chapters. Chapter 2 then continues this theme of efficient use of data by introducing an algorithm that augments a small amount of labeled data with the output of *weak labelers*, which are assumed to be *machine learning models* associated with *correlated tasks*. I show both *computational guarantees* on efficient learnability and statistical guarantees on generalization

[3] In particular, samples drawn from *independently sampled* blocks in the database setting, or samples with a "sufficiently large" number of steps between them (where "sufficiently large" depends on chain-specific analysis) in the Markov chain setting.

bounds involving both the number of labeled and unlabeled samples.

Overall, the work of Part I may be interpreted as novel methods to reduce the burden of acquiring large amounts of labeled data to train sophisticated machine learning models. Indeed, with the ever-expanding available computational power available (e.g., as characterized by Moore's law, and more recently with learning on GPUs, FPGAs, and specialized hardware) we have trained more complicated models (in particular, deeper neural networks), but datasets have had to grow commensurately to support training such models without overfitting. However, as computation becomes cheaper, the economic costs of collecting larger datasets do not always scale similarly, leaving learning from small datasets a problem that is both practically important and theoretically challenging. In fact, this setting is particularly impactful, as poor performance and analysis in the small sample setting can be a fairness issue, in the sense that by nature of their size and visibility, more data may be available on majority groups than minority or understudied groups, which contributes to marginalized groups being often poorly-served by machine learning systems.

Following the pure statistical methods of Part I, in Part II, I define *malfare* parallel to *welfare*, and show that subject to several intuitive and basic *axioms of cardinal welfare*, all welfare and malfare functions are *power means*. I then argue that fair machine learning methods should seek to *minimize malfare*, as this naturally generalizes *loss minimization* to multiple groups. I next combine in Chapter 4 the uniform convergence guarantees of Part I with the fairness setting of Part II, by introducing the *fair codec selection problem* and providing an algorithm to solve it. Finally, in Chapter 5, I define a concept of *fair-PAC-learnability* in terms of malfare minimization, and characterize both necessary and sufficient conditions for various flavors of fair learnability. Following Valiant's introduction of classical PAC-learning in order to rigorously characterize machine learning problems in the lexicon of computer science, I define malfare minimization as the target of fair-PAC learning, with axiomatically justified fairness characteristics, the precise details of which also lead to interesting theoretical computer science and statistical estimation questions. In particular, I show a hierarchy of PAC-learning and fair-PAC-learning settings, and am excited to see additional connections drawn in future work.

Finally, in Part III, I leave the i.i.d. setting behind, with two dependent *mean-estimation* settings of great practical and theoretical interest. Chapter 6 describes a common real-world problem in databases, where a user wants to estimate the *mean of a function*, without waiting for the database to iterate over all records. Motivated by the real-world performance characteristics of modern database systems, I show that, while in the worst case, sampling entire blocks may not improve over sampling individual records, my algorithm is able to detect independence within blocks automatically, and adapts accordingly. Chapter 7 adapts this idea of *algorithmically adapting* to independence detected within the data, and applies it to sample from Markov chains. Such settings are both of great interest to the theoretical computer science community, as increasingly attention is placed on *randomized algorithms*, and also to practitioners in diverse fields, from *significance testing* in the sciences to *machine learning* and *artificial intelligence*, wherein such methods are vital to efficiently reason under partial knowledge in an uncertain world.

The central theme running through all of my work is the importance placed on rigorous guarantees for domain-appropriate properties in various probabilistic processes, most notably in *machine learning* (generalization bounds), *sampling* (mean estimation guarantees), and *data science* (frequent itemsets, equilibria in empirical game theory, etc.). In particular, almost all of my publications rigorously show *finite-sample probabilistic guarantees*, 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 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 avoid negative outcomes becomes increasingly apparent.

In particular, while issues like *sample complexity*, *generalization bounds*, and *non-i.i.d. estimation guarantees* may seem less important to a broader audience than the *social good implications* of *fair machine learning* problems, this thesis presents the argument that it is exactly this sort of analytical approach that will allow us confidently learn from data. Indeed, by rigorously characterizing fair learnability and studying its limitations, we may not only avoid the biased mistakes and harmful outcomes that plague current machine learning systems, but also better understand, and perhaps avoid, the problems of the future.