Emily Diana

Title: *Balanced Filtering via Non-Disclosive Proxies*

Abstract: We study the problem of collecting a sample of data that is balanced with respect to sensitive groups when sensitive features are not available at collection time. To do this, we adopt a fairness pipeline perspective, in which a learner can use a small set of labeled data to train a proxy function that can later be used for this filtering task. Our proxy function is represented as a decision tree that first maps a sample to one of a finite set of proxy groups and then includes the sample in the dataset with an acceptance probability that is a function only of the proxy group. Importantly, we require that the proxy itself not reveal too much about the sensitive group membership of any individual sample, (i.e. that it be sufficiently non-disclosive). We present an algorithm that, under modest assumptions, can find such a proxy in a sample- and oracle-efficient manner and present an empirical evaluation of our algorithm.

Elvis Dohmatob

Title: *Tradeoffs between test error and robustness in different learning regimes for two-layer neural networks*

Abstract: Neural networks are known to be highly sensitive to adversarial examples. These may arise due to different factors, such as random initialization, or spurious correlations in the data distribution. To better understand these factors, we provide a precise study of the adversarial robustness in different scenarios, from initialization to the end of training in different regimes, as well as intermediate scenarios, where initialization still plays a role due to “lazy” training. We consider over-parameterized networks in high dimensions with quadratic targets and infinite samples. Our analysis allows us to identify new tradeoffs between test error and robustness, whereby robustness can only get worse when test error improves, and vice versa.

This is joint work with Alberto Bietti. ArXiv preprint: https://arxiv.org/abs/2203.11864

Shiwali Mohan

Title: *Collaborative Human-AI Systems*

Abstract: With the recent successes of AI and ML methods, the excitement about intelligent systems greatly enhancing human lives has reached a fever pitch. However, there is a deep chasm between algorithmic advances and systems that address human problems. This chasm is particularly perceptible as we build solutions for social good - health, sustainability, education etc. AI system development often overlooks humans who arguably are the most complex piece of the puzzle. Often, 'thinking about the human' is left to the design teams who have to overcome the impossible challenge of retrofitting such systems to the human use cases. My research explores an alternative pathway - how do we put 'thinking about the human' - at the center of intelligent system development. The talk will summarize various elements of a 'human-aware' system design process. This process is grounded in my experience in developing intelligent systems for preventive healthcare, sustainable transportation, collaborative robots, augmented reality learning etc. The talk will highlight insights and learning from these domains with an eye towards impact evaluation of these systems.

Esther Rolf

Title: *Incorporating Intent, Impact, and Context for Beneficial Machine Learning*

Abstract: In this talk, I will present an overview of my dissertation research on the context-aware design of machine learning systems that interface with individuals, our environment, and our societies. I will discuss (i) contrasting algorithmic impact with desired intent, (ii) contextualizing learning algorithms through structures in input data, and (iii) advancing machine learning with remotely sensed data as precise applications of intent- and context-aware design in practice.