Research Interests

Through a PhD at INSERT SCHOOL HERE, I would like to blur the lines between theory and applications by designing statistical methods/algorithms that address real-world problems while simultaneously achieving strong theoretical guarantees. In this regard, I am especially interested in **statistical learning theory**, but I am certainly open to other areas of statistics such as **Bayesian inference** and **time series analysis**.

Main Research Experience

REU: Uncertainty Quantification of Diffusion Coefficient Estimation
This past summer I was a Summer Scholars Fellow at Carnegie Mellon University, where I worked under the guidance of Prof. Jerry Wang (CMU) and in collaboration with two other fellows, Jennifer Zhang (MIT) and Courtney Francois (U Southern Mississippi). Our project explores the relationship between materials science and statistics.

Motivation
Particles move according to Brownian motion and are therefore well modeled by **random walks**. By regressing on random walk data, engineers obtain a **mean squared displacement** (MSD) value that serves as an estimate of the diffusion coefficient of a particle system. One issue with this approach is that random walks are inherently **autocorrelated** and **heteroskedastic**, which violates the **Gauss-Markov assumptions** needed for ordinary least squares regression.

Results
Addressing the Gauss-Markov violation consists of two parts: confirming that the violation is nonnegligible and applying corrections to construct a more reliable confidence interval. Nonnegligibility was confirmed via **Monte Carlo simulations**; by regressing on data in which the true MSD was known, we were able to show that a critical value of 2 corresponded to a **coverage probability** of about 0.25 instead of the expected 0.95. After using the **Newey-West estimate** of the covariance matrix of the random walk data, we were able to obtain proper coverage probabilities that matched what you would expect for the 68-95-99.7 rule. I will be presenting our poster at the **2025 Joint Math Meeting**.

My Contributions
In addition to significant work on our deliverable—the final poster is based on my draft—I discovered a generalization that strengthened our results. We obtained our results under the assumption that random walks are driven by **zero-mean IID** variables. Non-zero random variables have **drift**, which causes the MSD to be nonlinear. We can correct for the drift from non-zero variables with a **translation** at each time step. The translation is achieved by subtracting an estimate of the kth variable's population mean at the kth time step. That is, if a walk is driven by $W_n = X_1 + X_2 + \cdots + X_n$, then the expected value of $W_n' = (X_1 - \hat{\mu_1}) + (X_2 - \hat{\mu_2}) + \cdots + (X_n - \hat{\mu_n})$ is zero, which enables us to use our original results.

Senior Thesis: Learning Theory

Purpose
My home institution does not have a statistics department, so preparation for graduate study in statistics via coursework beyond what I have already studied (probability, machine learning foundations, and deep learning) isn't possible. In an effort to work around this issue, I proposed (and got approved for) a **year-long study** of advanced machine learning under the guidance of Prof. Tom O'Connell (Skidmore, Computer Science).

Plan of Study
After inspecting several graduate syllabi, I decided to use *Understanding Machine Learning: From Theory to Algorithms* by Shalev-Shwartz and Ben-David as my main reference. In particular, I am working through the **theory** section (chapters 2-7: Probably Appoximately Correct (PAC) Learnability, Uniform Convergence, VC-dimension, etc.), selected chapters from the **algorithms and models** sections (chapters 12-17, 21-23: Convex Learning, Stochastic Gradient Descent, Clustering, Online Learning, etc.), and some of the **advanced theory** (Rademacher complexities, PAC-Bayes, Multiclass learnability).

Deliverable
My deliverable will consist of formal notes—typed in LaTeX as if prepared for lecture—in which I rederive all the major proofs (especially details that are explicitly or implicitly "left to the reader"), Monte Carlo simulations of major theoretical results (e.g. the learnability of finite hypothesis classes), and implementations of various learning algorithms/paradigms. I will conclude the project with an

examination of current open problems from the **Conference of Learning Theory**. Progress is on going and is aperiodically updated on my website.

Goals

In addition to designing my own novel algorithms and methods, I'd like to make existing theoretical results more accessible to applied sciences. Consider, for instance, multiclass classification (MC). The PAC learnability of MC is a deceptively challenging problem, but has recently seen progress via list PAC learning [CITE].

The work in [CITE] draws from a variety of theoretical disciplines including graph theory, combinatorial learning theory, and topology and is therefore not necessarily accessible to those who may have a need for MC, like medical practioners or corporations. I want to be involved in the intermediary steps necessary to apply results like list learning to real-world problems. For instance, is it possible to use list learning to achieve robust tumor classification? Moreover, if it is possible, is it practical? That is, could we achieve sufficiently reliable results with a reasonably small sample size, or are we better off using existing methods like neural networks?

As for my long term goals, I intend to conduct research for the remainder of my career. A PhD will give me the training to be a research leader either in academia or in an industry research lab.

Fit

* this is where I will mention professors I want to work with *