# Scientific Abstract:

# Learning Over Large Alphabets

Consider a finite sample from a large discrete alphabet. Making inference and predictions in this setup is a fundamental problem in machine learning and statistics, with many important applications. In this proposal we focus on estimating the underlying distribution in the large alphabet regime. That is, in cases where the alphabet size is comparable to (or larger than) the sample size.

Large alphabet estimation holds several basic challenges. For example, many symbols typically do not appear in the sample. Hence, estimating these unseen symbols seems almost impossible with no prior assumptions. This is known as the *missing mass problem*. Estimating the missing mass dates back to the early work of Laplace. In his work, Laplace suggested adding a single count to all the symbols in the alphabet (including unobserved symbols). Then, the missing mass estimate is simply the empirical distribution of all the symbols with a single count. Many years after Laplace, a major milestone was established in the work of Good and Turing. The Good-Turing (GT) framework suggests that unseen symbols are assigned a probability proportional to the number of events with a single appearance in the sample. This approach introduced a significant improvement compared to known estimators at the time. Both Laplace and Good-Turing hold several advantages and caveats. The Laplace estimator is optimal in a Bayesian sense, under a uniform prior. It also assumes a known alphabet size and it works well in cases where the alphabet is relatively small. On the other hand, the GT estimator does not assume that the alphabet size is known, and performs significantly better in the large alphabet regime.

Our preliminary results focus on missing mass estimation to achieve improved performance over large alphabets. Here, our objective is to minimize the maximal risk (in terms of MSE) over all probability distributions of a countable alphabet. As a first step towards this goal, we introduce a generalized Laplace estimator. Our proposed estimator significantly improves upon currently known methods, especially in large alphabets. Incorporation of our estimator into a hybrid estimator with the GT estimator yields further improvements, as it utilizes the advantages of both schemes.

While the missing mass problem is a major challenge on its own, our ultimate goal is the complete underlying distribution. Going forward, we aim at generalizing our results to hybrid estimators for the probability of symbols that appear k times in the sample (for $k \geq 0$). This challenge is far from trivial, as it requires adjustments to risk measures that are more typical for distribution estimation in the large alphabet regime. Our preliminary results indicate that such a hybrid framework would introduce a significant improvement to large alphabet distribution estimation and would hopefully impact many fields.