# Asymptotic Normality of Higher Order Turing Formulae

Jie Chang

Preprint no. 2022-05

**Abstract**

Higher order Turing formulae, denoted as $T_r$ for $r \in \mathbb{Z}^+$, are a powerful result allowing one to estimate the total probability associated with words from a random piece of writing, which have been observed exactly $r$ times in a random sample. In particular $T_r$ estimates the probability of seeing words not appearing in the sample. To perform statistical inference, e.g., constructing the asymptotic confidence intervals, the asymptotic properties of the higher Turing formulae need to be studied. In this thesis we extend the validity of the asymptotic normality beyond the previously proven cases by establishing a sufficient and necessary condition for the asymptotic normality of higher order Turing formulae when the underlying distribution is both fixed and changing. We also conduct simulation studies with the complete works of William Shakespeare and data generated from different underlying distributions to check the finite sample performance of the derived asymptotic confidence interval. Based on our theoretical results we also developed two methodologies for authorship detection with real twitter data analysis.

Department of Mathematics, UNC Charlotte, Charlotte, NC 28223