Language modeling, the task of estimating the probability of sequences of words, is an important component in many applications such as speech recognition and machine translation. While seemingly simple, the large vocabulary space and power law nature of language lead to a severe data sparsity problem, making parameter estimation challenging.

In this work, I present power low rank ensembles (PLRE), a framework for language modeling that consists of collections of low rank matrices and tensors. Our method can be understood as a generalization of n-gram modeling to non-integer n, and includes standard techniques such as absolute discounting, deleted-interpolation, and Kneser Ney smoothing as special cases. Our approach consistently outperforms state-of-the-art modified Kneser Ney baselines while preserving the computational advantages of n-gram models. In particular, unlike other recent advances such as neural language models, our method does not have any partition functions, thus enabling fast evaluation at test time.

This work received the best paper runner up award at EMNLP 2014 and is joint work with Avneesh Saluja, Chris Dyer, and Eric Xing.

Ankur Parikh is a PhD candidate at Carnegie Mellon University advised by Professor Eric Xing. He is passionate about research in machine learning, natural language processing (NLP), and computational biology. In particular, his thesis explores probabilistic modeling from the perspective of linear algebra and applications of these insights to design effective solutions for NLP tasks. Ankur has received a best paper runner up award at EMNLP 2014, a best paper in translational bioinformatics at ISMB 2011, and an NSF Graduate Fellowship in 2011. He previously graduated from Princeton University in 2009 with highest honors.