Govinda Kamath

govinda-kamath.github.io | Google Scholar | Linkedin | Github

CURRENT POSITION

MICROSOFT RESEARCH

Post-doctoral researcher
Aug 2019-present | Cambridge, MA
Research on knowledge distillation, bandit based rank-1 models, connections between sequence alignment and compression.

EDUCATION

STANFORD UNIVERSITY

PhD in Electrical Engineering
May 2019 | Stanford, CA
Adviser : David Tse
Thesis: Almost linear time algorithms for problems from Computational Genomics
CGPA: 4.0 / 4.0

INDIAN INSTITUTE OF SCIENCE

ME in Telecommunication Engineering
May 2012 | Bangalore, India

NATIONAL INSTITUTE OF TECHNOLOGY, KARNATAKA

B.Tech in Electrical and Communication Engineering
May 2010 | Surathkal, India

PHD INTERNSHIPS

Human Longevity Inc • Microsoft Research • Google X • Applied Protocol Research (blockchain startup)

COURSEWORK

Advanced Machine Learning
Advanced Statistics
Applied Statistics
Data Structures and Algorithms
Randomised Algorithms
Theory of Probability
Detection and Estimation Theory
Information Theory
Signal Processing

LANGUAGES

Python • C/C++ • MATLAB • Shell • R

LINKS

Github • Scholar • Website • Linkedin

SHORT BIO

I work at the intersection of machine learning, algorithms and statistics. My work has mainly involved designing and deploying scalable machine learning algorithms to solve real problems, and showing their goodness.

PROJECTS

NEAREST NEIGHBOURS, K-MEANS IN ALMOST LINEAR TIME

We designed and implemented a framework which computes nearest neighbors to n points in d dimensions in $O(n \log d)$ time and also runs each iteration of Lloyd's k-means algorithm in $O(nk \log d)$ time.

LOW-RANK MODELS FOR SEQUENCE ALIGNMENT

We considered a class of algorithms for sequence alignment - seed and extend algorithms, and drew connections to low rank models. We derived spectral estimators for such models and then used the framework from above to speed up its computation.

SPEEDING UP K-MEDOID CLUSTERING

We designed an algorithm to find the median in high-dimensions of n points in $O(n \log n)$ time improving over state-of-art $O(n^2)$ methods to speed up k-medoid clustering.

KNOWLEDGE DISTILLATION & SEMIPARAMETRIC INFERENCE

We studied the problem of knowledge distillation - where it has been observed that training a small model using logits soft class labels generated using a large model trained from the data does better than directly training the small model using the data. We cast knowledge distillation as a semiparametric inference problem and derive several new guarantees for the prediction error of standard distillation.

PREDICTING OFF TARGET CUTS OF CRISPR-CAS9 GUIDES

We train a triplet model to embed CRISPR-CAS9 guide sequences and targets into a common euclidean space where the closer the guides are to the targets, the more likely the guide is to cut the target.

BACTERIAL GENOME ASSEMBLY

We considered the problem of genome assembly from long reads and came up with a graph algorithm to resolve all resolvable repeats while not resolving those that could not be. We wrote a bacterial genome assembler in C++ and Python which is used by biologists.

HAPLOTYPE ASSEMBLY AND STOCHASTIC BLOCK MODELS

We posed haplotype assembly as a stochastic block model and came up with a spectral algorithm to solve it and proved that the algorithms are theoretically optimal.

SELECTED PUBLICATIONS AND PREPRINTS

- Knowledge Distillation as Semiparametric Inference (ICLR 2021)
- Adaptive Learning of Rank-One Models for Efficient Pairwise Sequence Alignment (NeurIPS 2020)
- Spectral Jaccard Similarity: A new approach to estimating pairwise sequence alignments (RECOMB 2020, Cell Patterns 2020)
- Adaptive monte-carlo optimization (Preprint)
- crispr2vec: Machine Learning Model Predicts Off-Target Cuts of CRISPR systems (Preprint)
- Medoids in almost linear time via multi-armed bandits (AISTATS 2018)
- Community recovery in graphs with locality (ICML 2016)
- HINGE: Long-Read Assembly Achieves Optimal Repeat Resolution (Genome Research 2017)
- Valid post-clustering differential analysis for single-cell RNA-Seq (RECOMB 2019, Cell Systems 2019)