Random Projections & Applications To Dimensionality Reduction

Aditya Krishna Menon
(BSc. Advanced)

Supervisors:
Dr. Sanjay Chawla
Dr. Anastasios Viglas

High-dimensionality

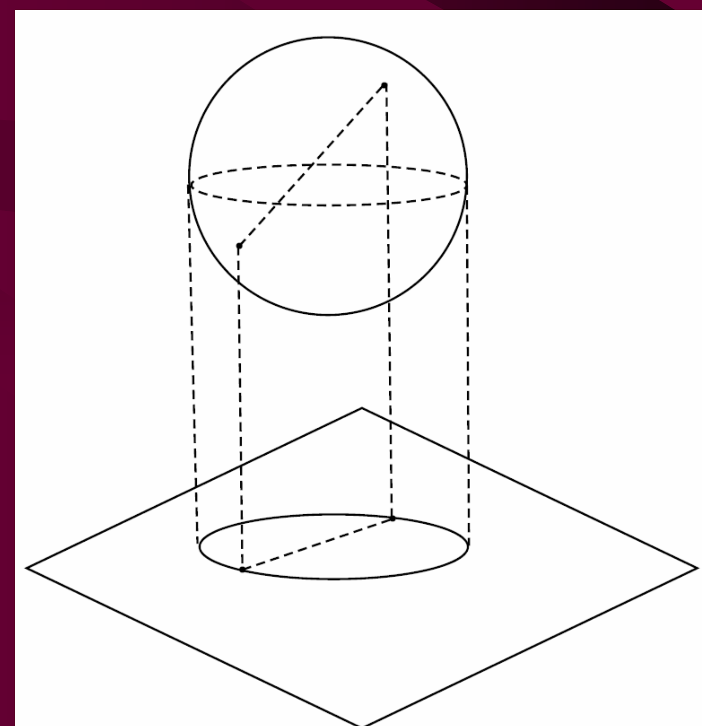
- Lots of data $\rightarrow$ objects/items with some attributes
  - i.e. high-dimensional points
  - $\Rightarrow$ Matrix
- Problem: number of dimensions usually quite large
  - Data analysis usually sensitive to this
    - e.g. Learning, clustering, searching, …
  - $\Rightarrow$ Analysis can become very expensive
- The ‘curse of dimensionality’
  - Add more attributes $\Rightarrow$ exponentially more time to analyze data

Solution?

- Reduce dimensions, but keep structure
  - i.e. map original data $\rightarrow$ lower dimensional space
  - Aim: do not distort original too much
  - ‘Dimensionality reduction’
- Easier to solve problems in new space
  - Not much distortion $\Rightarrow$ can relate solution to original space

Random projections

- Recent approach: random projections
- Idea: project data onto random lower dimensional space
  - Key: most distances (approx.) preserved
  - Matrix multiplication



Illustration

Original n points in d dimensions

$$A \xrightarrow{A.R} E$$

$n \times d$ → $n \times k$

New n points in k dimensions

R is some 'special' random matrix
e.g. Gaussian

Guarantee: With high probability, distances between points in *E* will be very close to distances between points in *A* [Johnson and Lindenstrauss]

Aims of my project

- Can we solve data-streaming problems efficiently, and accurately, using projections?
- Can we improve existing theory on 'interesting' properties random projections?
  - Preservation of dot-products
  - Guarantees on the reduced dimension

My contributions

- Application of projections to data streaming
- Novel result on preservation of dot-product
- Theoretical results on lowest dimension bounds

I: Streaming scenario

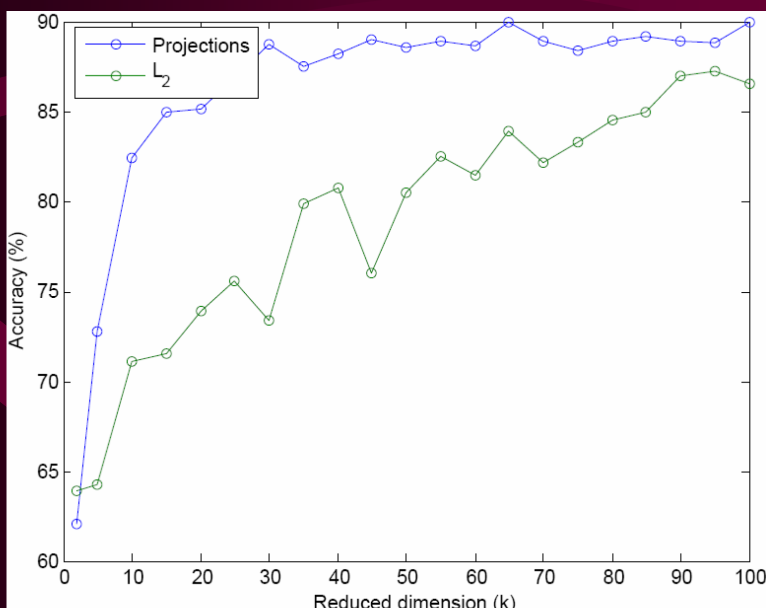
- Scenario: have a series of high-dimensional streams, updated asynchronously
  - i.e. Arbitrarily updated
- Want to query on distance / dot-product between streams
  - e.g. To cluster the streams at fixed point in time
- Problem: might be infeasible to instantiate the data
  - Or might be too expensive to work with high-dimensions
- Usual approach is to keep a *sketch*
  - Small space
  - Fast, accurate queries
- Aim: can we use projections to maintain a sketch?
  - Comparison to existing sketches?

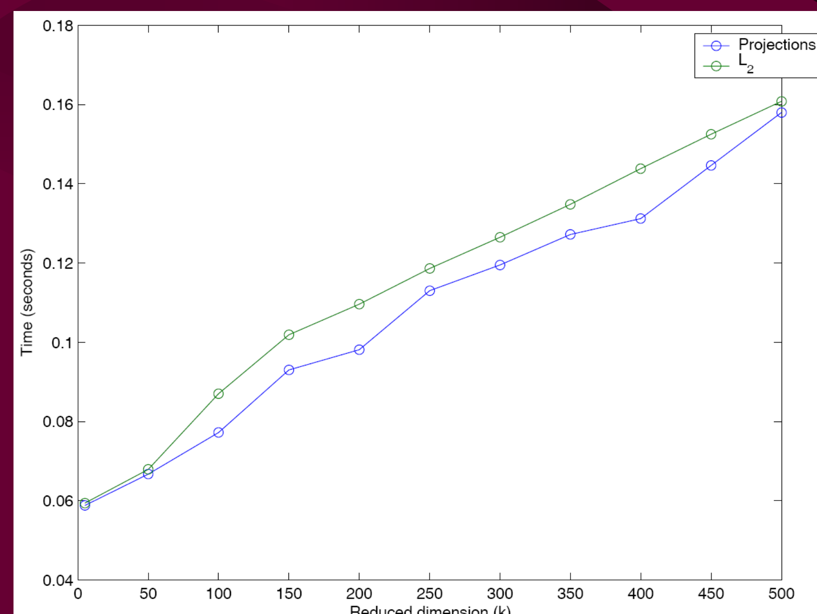
My work on streams

- Showed we can efficiently use projections to keep sketch
  - Can quickly make incremental updates to sketch
    - As if you did a projection each time!
  - Guarantee: preserves Euclidean distances among streams
- Generalization of [Indyk]
  - Related to a special case of a random projection
- Comparison
  - As accurate than [Indyk]
  - Faster than [Indyk]
    - 2/3rds sparse matrix [Achlioptas]

Experiments

- Use projections to allow *k*-means clustering of high-dimensional ($d = 10^4$) streams
- Results
  - At least as accurate than [Indyk]
  - Marginally quicker





II: Dot-product

- Dot-product is quite a useful quantity
  - e.g. For cosine similarity
- On average, projections preserve dot-products
  - But typically large variance
  - Not an easy problem
    - *"Inner product estimation is a difficult problem in the communication complexity setting captured by the small space constraint of the data stream model"* [Muthukrishnan]
- Question: can we derive bounds on the error?

My work on dot-products

- Result: derived new bound on error incurred in dot-product after random projection
  - High-probability upper bound on the error
  - Complements existing work on dot-product preservation
    - My bound based on distance error and lengths of vectors
    - Existing results based on reduced dimension and lengths of vectors

III: Lowest dimension bounds

- Projections give bounds on reduced dimension
  - ‘If I want 10% error in my distances, what is the lowest dimension I can project to’?
- [Achlioptas]’ bounds are most popular
  - But quite conservative [Lin and Gunopulos]
- Aim: try to improve results on bounds for reduced dimension
  - Look at when bound is not meaningful
  - Better special cases?

My work on bounds

- Results:
  - Theorem on analysis of applicability of [Achlioptas]' bound
    - NASC conditions for it to be 'meaningless'
      - Points exponential in number of dimensions
  - Stronger result for data from Gaussian distribution
    - Error restriction

Conclusion and future work

- Random projections are an exciting new technique
  - Applications to dimensionality reduction and algorithms
  - Worthwhile studying properties
- My contributions
  - Proposed application to data-streams
  - Novel result on preservation of dot-product
  - Improved theoretical analysis on bounds
- Future work
  - [Li et. al]'s matrix and data-streams
  - Lower bound analysis
  - Guarantees for projections in other problems e.g. circuit fault diagnosis

References

- [Achlioptas] Dimitris Achlioptas. 2001. Database-friendly random projections. In *PODS '01: Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 274–281, New York, NY, USA. ACM Press.
- [Indyk] Piotr Indyk. 2006. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323.
- [Johnson and Lindenstrauss] W.B. Johnson and J. Lindenstrauss. 1984. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, pages 189–206, Providence, RI, USA. American Mathematical Society

References

- [Li et al.] Ping Li, Trevor J. Hastie, and Kenneth W. Church. 2006. Very sparse random projections. In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 287–296, New York, NY, USA. ACM Press.
- [Lin and Gunopulos] Jessica Lin and Dimitrios Gunopulos. 2003. Dimensionality reduction by random projection and latent semantic indexing. Unpublished. In Proceedings Of The Text Mining Workshop at the 3rd International SIAM Conference On Data Mining.
- [Muthukrishnan] *Data Streams: Algorithms And Applications*. Now Publishers, 2005.