Convex Hull and Related Problems in Data Streams

Subhash Suri

Computer Science Department
UC Santa Barbara

Joint work with John Hershberger
Stream of Points

- Imagine a stream of $N$ points in 2D plane. (Sensors, astrophysical data, spatial database.)
- Summarize the stream of points, preserving important data characteristics.
- Summary depends on the application. We are interested in extremal properties.

- Examples: Diameter, width, or extent in any direction.
Geometric Problems

- Is data set $A$ surrounded by data set $B$?
- Are the two sets linearly separable?

- Minimum distance between the sets $A$ and $B$?
- Many other natural geometric problems of this kind.
Convex Hull

- CH is a natural summary for extremal geometric problems.
- Diameter, width, separation, containment, distance queries efficiently answered from CH.

**Question:** With space for $r \ll N$ points, how best to represent the convex hull of the stream?
Related Work

- Feigenbaum-Kannan-Zhang: Diameter, sliding window.
- Indyk: Clustering.
- Cormode-Muthukrishnan: Radial Histogram.
- Robertson: Approximation of convex bodies.
- Muthukrishnan Survey.
Uniform Directional Sampling

- Maintain extremal points in $r$ directions: $\frac{2\pi j}{r}$.
- Convex hull of sampled extrema is the approximation.
- The region of error is the ring of “uncertainty triangles.”
Update Efficiency

- FKZ check new point for each of $r$ directions, $O(r)$ time.
- By maintaining CH dynamically, one can update CH in $O(\log r)$ amortized time.

* If $p \in CH$, skip. Else, compute tangents from $p$ to CH. They tell us directions in which $p$ is extremal.
* Add to CH only if it’s extremal in at least one direction, possibly deleting other points.
Error Bound

- Worst-case height of uncertainty triangles is $\Theta(D/r)$.

- Diameter error is additive $O(D/r^2)$.

- But error of $\Theta(D/r)$ in other directions can be too large. For example, Width $\ll$ Diameter.
Adaptive Sampling

- Start with extremal points in \( r \) uniform directions, \( \frac{2\pi j}{r} \).
- Adaptively sample in \( r \) additional directions.

Adaptive sampling reduces error to \( O(D/r^2) \).
Sampling Strategy

- Let $\theta(e)$ be angle interval of edge $e$; $\ell(e)$ be “length” of $e$. Set $\theta_0 = \frac{2\pi}{r}$. $P$ is perimeter of CH.

- Define weight of $e$ as: $w(e) = \frac{\ell(e)}{P/r} + \log \left( \frac{\theta(e)}{\theta_0} \right)$.

- Adaptive Sampling: Repeatedly, pick $e$ with $w(e) > 1$, and bisect its angle range.
Analysis

- Initially, $\sum_e w(e) = \sum_e \left( \frac{\ell(e)}{P/r} + \log \left( \frac{\theta(e)}{\theta_0} \right) \right) = r$.

- Length defined so that $\ell(e) = \ell(e') + \ell(e'')$, and $\ell(e')/\ell(e'')$ is ratio between the true lengths of $e'$ and $e''$.

- Each refinement shrinks sum of positive weights by $>1$. So at most $r$ new samples.

- If a CH edge has length $L$, then its uncertainty triangle has height at most $L\theta_0/2r(L/P)$, which is $O((D/r^2))$. 
Results

- We can maintain a CH using $2r$ points whose error from true CH is $\Theta(D/r^2)$.

- The approximate convex hull suffices to compute diameter, width, directional-extent, containment, separation, etc. upto an additive error of $O(D/r^2)$. 
Width

- Width of a set can be arbitrarily smaller than $D$, so the previous error is potentially unbounded.

- Additional analysis shows that the error in our width approximation is $O(W/r + D/2^\Theta(r))$. 
Conclusions

- Adaptive sampling improves error from $D/r$ to $D/r^2$.
- How to maintain CH in a Sliding Window model?
- Which CG problems are not amenable to data stream model?