

# Lecture 10: Sketching S3: Nearest Neighbor Search

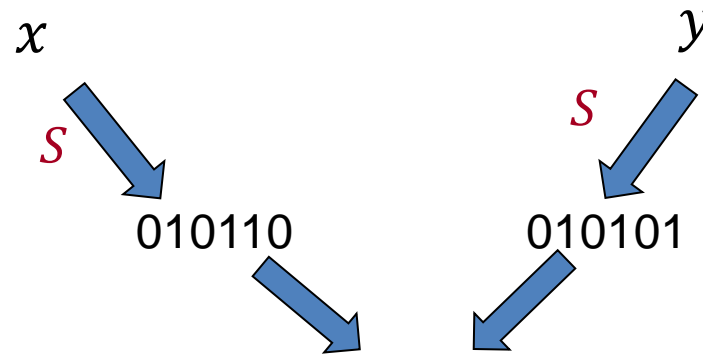


# Plan

- PS2 due yesterday, 7pm
- Sketching
- Nearest Neighbor Search
- Scriber?

# Sketching

- $S: \mathbb{R}^d \rightarrow$  short bit-strings
  - given  $S(x)$  and  $S(y)$ , should be able to estimate some function of  $x$  and  $y$
  - With 90% success probability ( $\delta = 0.1$ )
- $\ell_2, \ell_1$  norm:  $O(\epsilon^{-2})$  words
- Decision version: given  $r$  in advance...
- **Lemma:**  $\ell_2, \ell_1$  norm:  $O(\epsilon^{-2})$  bits



Distinguish between

$$\|x - y\| \leq r$$

$$\|x - y\| > (1 + \epsilon)r$$



# Sketching: decision version

- Consider Hamming space:  $x, y \in \{0,1\}^d$
- **Lemma:** for any  $r > 0$ , can achieve  $O(1/\epsilon^2)$  -bit sketch.
- [on blackboard]

# Conclusion

- Dimension reduction:
  - [Johnson-Lindenstrauss'84]:
    - a random linear projection into  $k$  dimensions
    - preserves  $\|x - y\|_2$ , up to  $1 + \epsilon$  approximation
    - with probability  $\geq 1 - e^{-\Omega(\epsilon^2 k)}$
  - Random linear projection:
    - Can be  $Gx$  where  $G$  is Gaussian, or  $\pm 1$  entry-wise
  - Hence: preserves distance between  $n$  points as long as  $k = \Theta\left(\frac{1}{\epsilon^2} \log n\right)$
  - Can do faster than  $O(dk)$  time
    - Using Fast Fourier Transform
- In  $\ell_1$ : no dimension reduction
  - But can do sketching
  - Using  $p$ -stable distributions (Cauchy for  $p = 1$ )
- Sketching: decision version, constant  $\delta = 0.1$ :
  - For  $\ell_1, \ell_2$ , can do with  $O\left(\frac{1}{\epsilon^2}\right)$  bits!

# Section 3:

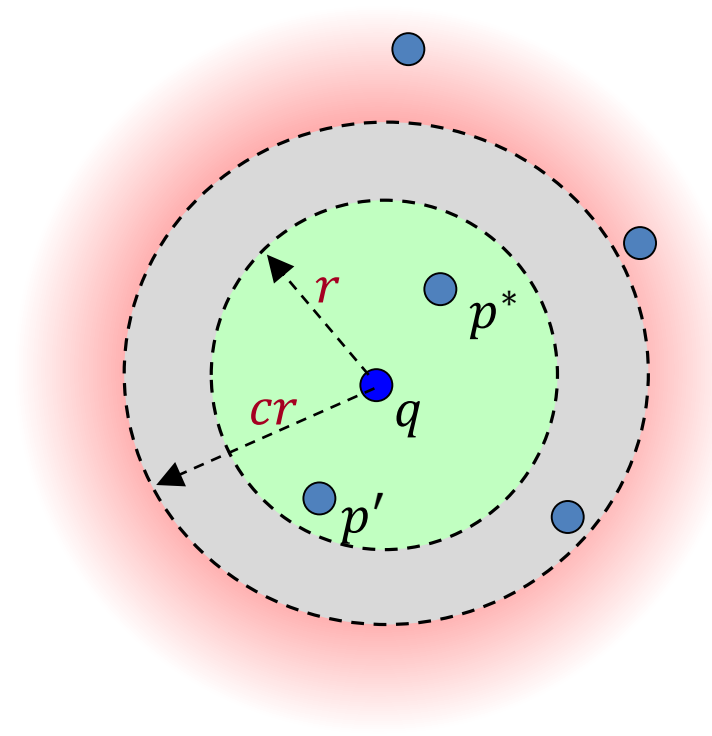
# Nearest Neighbor Search

# Approximate NNS

*c*-approximate

*r*-near neighbor: given a query point  $q$

- assuming there is a point  $p^*$  within distance  $r$ ,
- report a point  $p' \in D$  s.t.  $\|p' - q\| \leq cr$



# NNS: approach 1

- Boosted sketch:
  - Let  $S$  = sketch for the decision version (90% success probability)
  - new sketch  $W$  :
    - keep  $k = O(\log n)$  copies of  $S$
    - estimator is the majority answer of the  $k$  estimators
  - Sketch size:  $O(\epsilon^{-2} \log n)$  bits
  - Success probability:  $1 - n^{-2}$  (Chernoff)
- Preprocess: compute sketches  $W(p)$  for all the points  $p \in D$
- Query: compute sketch  $W(q)$ , and compute distance to all points using sketch
- Time: improved from  $O(nd)$  to  $O(n\epsilon^{-2} \log n)$



# NNS: approach 2

- Query time below  $n$  ?
- **Theorem [KOR98]:**  $O(d\epsilon^{-2}\log n)$  query time and  $n^{O(1/\epsilon^2)}$  space for  $1 + \epsilon$  approximation.
- Proof:
  - Note that  $W(q)$  has  $w = O(\epsilon^{-2} \log n)$  bits
  - Only  $2^w$  possible sketches!
  - Store an answer for each of  $2^w = n^{O(\epsilon^{-2})}$  possible inputs
- In general:
  - if a distance has constant-size sketch, admits a poly-space NNS data structure!
- Space closer to linear?
  - approach 3 will require more specialized sketches...