Approximate Near Neighbors for General Symmetric Norms

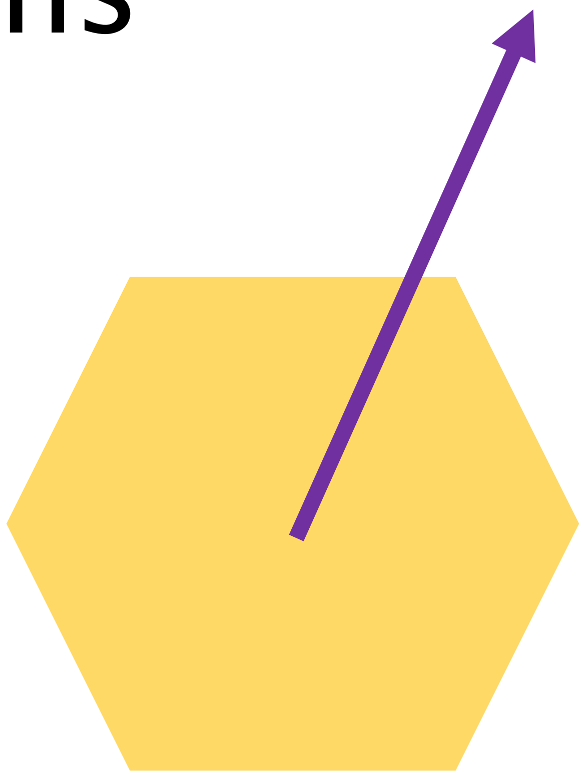
Alexandr Andoni (Columbia University)

Huy Nguyen (Northeastern)

Aleksandar Nikolov (University of Toronto)

Ilya Razenshteyn (MIT CSAIL)

Erik Waingarten (Columbia University)



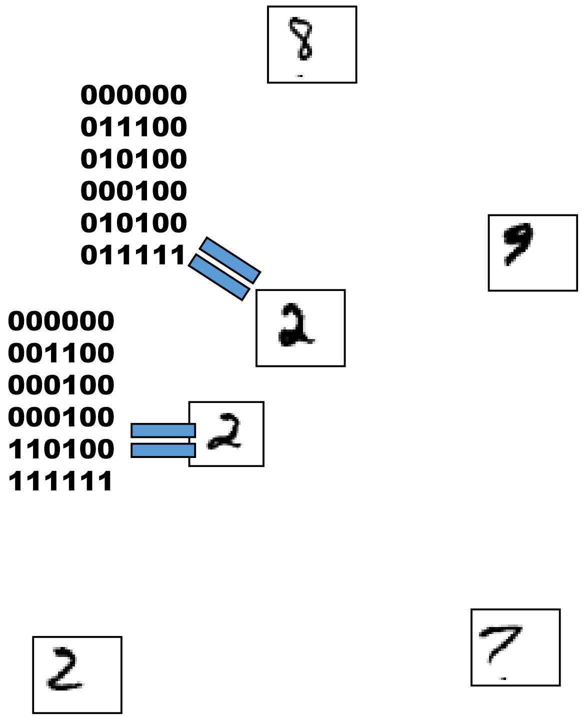
Motivation

Data → Feature vector space + distance function

Data analysis → Geometry / Linear Algebra / Optimization

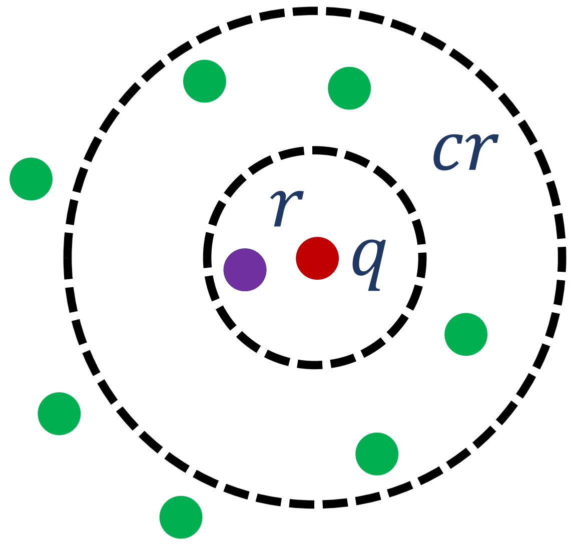
Similarity search → **Nearest Neighbor Search**

An example


000000
011100
010100
000100
010100
011111
000000
001100
000100
000100
110100
111111

Approximate Near Neighbors (ANN)

- **Dataset:** n points in a metric space X (denoted by P)
- Approximation $c > 1$, distance threshold $r > 0$
- **Query:** $q \in X$ such that there is $p^* \in P$ with $d_X(q, p^*) \leq r$
- **Output:** $\tilde{p} \in P$ such that
$$d_X(q, \tilde{p}) \leq cr$$
- Parameters: space, query time



FAQ

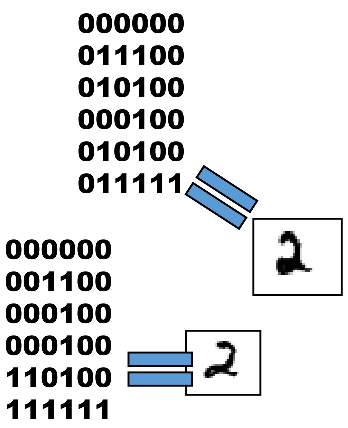
Focus of this talk

- **Q:** why approximation?
- **A:** the exact case is **hard** for the high-dimensional problem.
- **Q:** what does "high-dimensional" mean?
- **A:** dimension $d = \omega(\log n)$.
- **Q:** how is the dimension defined?
- **A:** a metric is typically defined on R^d; alternatively, doubling dimension, etc.

This talk: a metric on R^d, where $\omega(\log n) \leq d \leq n^{o(1)}$

Should depend on d as $d^{O(1)}$

Which distance function to use?



- A distance function
  - Must capture semantic similarity well
  - Must be algorithmically tractable
- E.g.: Hamming, Euclidean, Earth-mover distance…

Goal: classify metrics by the complexity of high-dimensional ANN

For theory: what's the relevant **property** of the metric

For practice: **universal** algorithm for ANN

- **Non-solution:** ANN for small doubling dimension
  - [Clarkson'99,Krauthgamer-Lee'04,Beygelzimer-Kakade-Langford'06]
  - ~2^k for doubling dimension k

Metric class: High-dimensional norms

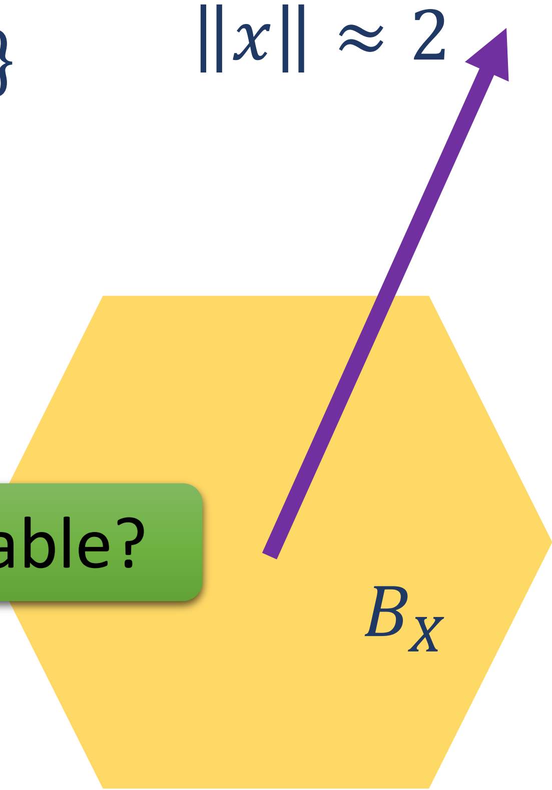
- Important case: X is a **normed space**
  - $d_X(x_1, x_2) = \|x_1 - x_2\|$, where $\|\cdot\|: R^d \to R_+$ is such that
    - $\|x\| = 0$ iff $x = 0$
    - $\|\alpha x\| = |\alpha| \|x\|$
    - $\|x_1 + x_2\| \leq \|x_1\| + \|x_2\|$
- Lots of tools (functional analysis)
  - E.g., can characterizate norms that allow efficient **sketching** (succinct summarization), which **implies** efficient ANN **[A, Krauthgamer, Razenshteyn 2015]**
- ANN with approximation $\mathrm{O}(\sqrt{d})$ is easy

Unit balls of norms

- A norm given by its unit ball $B_X = \{x \in R^d \,|\, \|x\| \leq 1\}$
- **Claim:** B_X is a symmetric convex body
- **Claim:** any such body can be a unit ball
  - $\|x\|_K = \inf\left\{t > 0 \,\middle|\, \frac{x}{t} \in K\right\}$

What property of a convex body makes ANN wrt it tractable?

- **John's theorem:** any symmetric convex body is close to an ellipsoid (gives approximation $\sqrt{d}$)



Our result

Invariant under permutation of coordinates and changing signs

- If X is a **symmetric** normed space, and $d = n^{o(1)}$, can solve ANN with:
  - Approximation $\cancel{O(1)}$ $(\log \log n)^{O(1)}$
  - Space $n^{1+o(1)}$
  - Query time $n^{o(1)}$

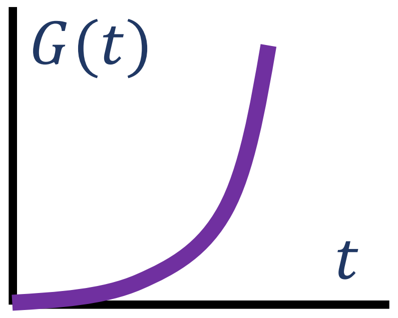
Examples

- Usual ℓ_p **norms** $\|x\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$
- **Top-k norm**: sum of k largest absolute values of coordinates
  - Interpolates between ℓ_1 and ℓ_∞
- **Orlicz norms**: a unit ball is

$$\{x \in R^d \,|\, \textstyle\sum_i G(|x_i|) \le 1\},$$

Where $G(\cdot)$ is convex and non-negative, and $G(0) = 0$.

  - Gives ℓ_p norms for $G(t) = t^p$



- k-support norm, box-Θ norm, K-functional (arise in probability and machine learning)

Related work: symmetric norms

- **[Blasiok, Braverman, Chestnut, Krauthgamer, Yang 2017]**: classification of symmetric norms according to their **streaming complexity**
  - Depends on how well the norm concentrates on the Euclidean ball
  - Unlike streaming, ANN is **always** tractable

Prior work: ANN

- Classically, focus on ℓ_1 (Manhattan) and ℓ_2 (Euclidean) norms

  captures many applications!

- Allow efficient algorithms based on **hashing**
  - Locality-Sensitive Hashing

    [Indyk, Motwani 1998] [Andoni, Indyk 2006]
  - Data-dependent LSH

    [A, Indyk, Nguyen, Razenshteyn 2014] [A, Razenshteyn 2015]
  - tight trade-off between space and query time

    [A, Laarhoven, Razensheyn, Waingarten 2017]
- Other norms: few results for ℓ_∞, general ℓ_p (will see later)

ANN for ℓ_∞

- ANN for d-dimensional ℓ_∞ **[Indyk 1998]:**
  - Space $d \cdot n^{1+\varepsilon}$
  - Query time $O(d \log n)$
  - Approximation $O_\epsilon(\log \log d)$

- **Idea:** recursively build a decision tree

- $O(\log \log d)$ approximation is **tight** for decision trees!

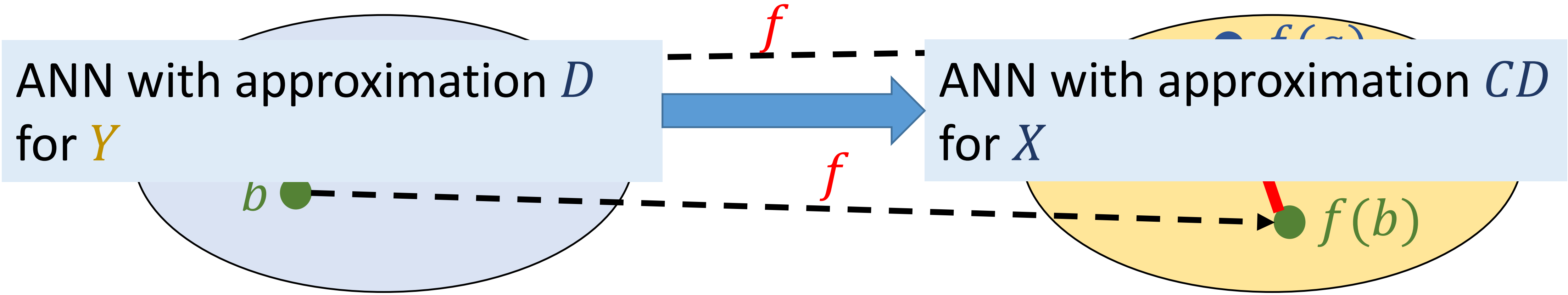
  [A, Croitoru, Patrascu 2008] [Kapralov, Panigrahy 2012]

Metric embeddings

- A map $f: X \to Y$ is an **embedding with distortion $\boldsymbol{C}$**, if for $a, b \in X$:

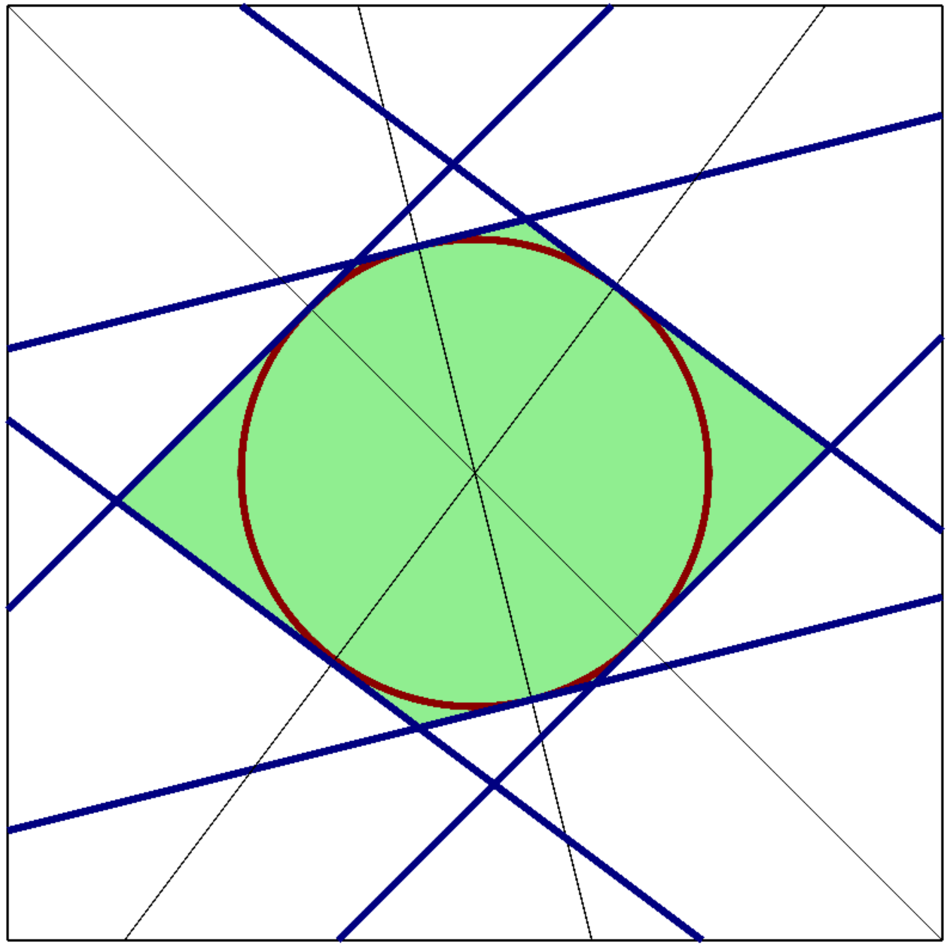
$$d_Y\big(f(a), f(b)\big)/C \leq d_X(a, b) \leq d_Y(f(a), f(b))$$

- Reductions for geometric problems



Embedding norms into ℓ_∞

- For a normed space X and $\varepsilon > 0$ there exists $f: X \to \ell_\infty^{d'}$ with
  $\|f(x)\|_\infty \in (1 \pm \varepsilon) \cdot \|x\|_X$
- **Idea:** $\|x\|_X \approx \max_{y \in N} |\langle x, y \rangle|$
  - Approximate the unit ball with a polytope



Can we use this embedding + ANN for ℓ_∞ to get ANN for any norm?
No, since $d' = 2^{\Omega(d)}$, even for ℓ_2.

The refined strategy

What	Where	Dimension

Bypass non-embeddability into low-dimensional ℓ_∞
by allowing a more complicated host space, which is still tractable

ℓ_p-direct product of metric spaces

- For metrics M_1, M_2, …, M_t, define $\bigoplus_{l_p} M_i$ as follows:
  - The ground set is $M_1 \times M_2 \times \cdots \times M_t$
  - The distance is:
  $d\big((x_1, x_2, \dots, x_t), (y_1, y_2, \dots, y_t)\big) = \|(d(x_1, y_1), d(x_2, y_2), \dots, d(x_t, y_t)\|_p$
- **Example**: $\bigoplus_{\ell_p} \ell_q$ (cascaded norms)
- Our **host space:** $\bigoplus_{\ell_\infty} \bigoplus_{\ell_1} X_{ij}$, where X_{ij} is top-k_{ij} norm on R^d
  - Outer dimension is of size $d^{O(1)}$
  - Inner dimension is of size d

Our algorithm

1) Embed any symmetric norm into $\bigoplus_{\ell_\infty} \bigoplus_{\ell_1} X_{ij}$

2) Solve ANN for $\bigoplus_{\ell_\infty} \bigoplus_{\ell_1} X_{ij}$

- Prior work on ANN via product spaces:
  - Frechet distance **[Indyk 2002]**
  - Edit distance **[Indyk 2004]**
  - Ulam distance **[A, Indyk, Krauthgamer 2009]**

ANN for $\bigoplus_{\ell_\infty} \bigoplus_{\ell_1} X_{ij}$

- **[Indyk 2002]**, **[A 2009]**:

Metrics $M_1, M_2, \ldots, M_t$ admit data structures for c-ANN

⇒

Direct-product $\bigoplus_{\ell_p} M_i$ admits $O(c \cdot \log\log n)$-ANN with almost the same time and space

  - A powerful generalization of ANN for l_∞ **[Indyk 1998]**
  - Eg, implies ANN for general l_p

- Enough to solve ANN for X_{ij} (top-k norms)!

ANN for top-k norms

- As hard as both ℓ_1 and ℓ_∞
- **Idea:** embed a top-k norm into $\ell_\infty^{d'}$ and use **[Indyk 1998]**
  - approximation: $\text{distortion } \times O(\log\log d')$
- **Hurdle:** ℓ_1 requires $2^{\Omega(d)}$-dimensional ℓ_∞
- **Solution:** use randomized embeddings

Embedding top-k norm into ℓ_∞

- First case: $k = d$ (that is, ℓ_1)
- Embedding (uses **min-stability** of exponential distribution):
  - Sample i.i.d. $u_1, u_2, \dots, u_d \sim \mathrm{Exp}(1)$
  - Embed $f: (x_1, x_2, \dots, x_d) \mapsto \left(\frac{x_1}{u_1}, \frac{x_2}{u_2}, \dots, \frac{x_d}{u_d}\right)$
- $\Pr[\|f(x)\|_\infty \leq t]$
- Constant distortion
- **General $\boldsymbol{k}$:** sample $u_i \sim \max\left(\frac{1}{k}, \mathrm{Exp}(1)\right)$
- Similarly, Orlicz norms: $u_i \sim \boldsymbol{\mathcal{D}}$

Where are we?

next

1) Embed any symmetric norm into $\bigoplus_{\ell_\infty} \bigoplus_{\ell_1} X_{ij}$

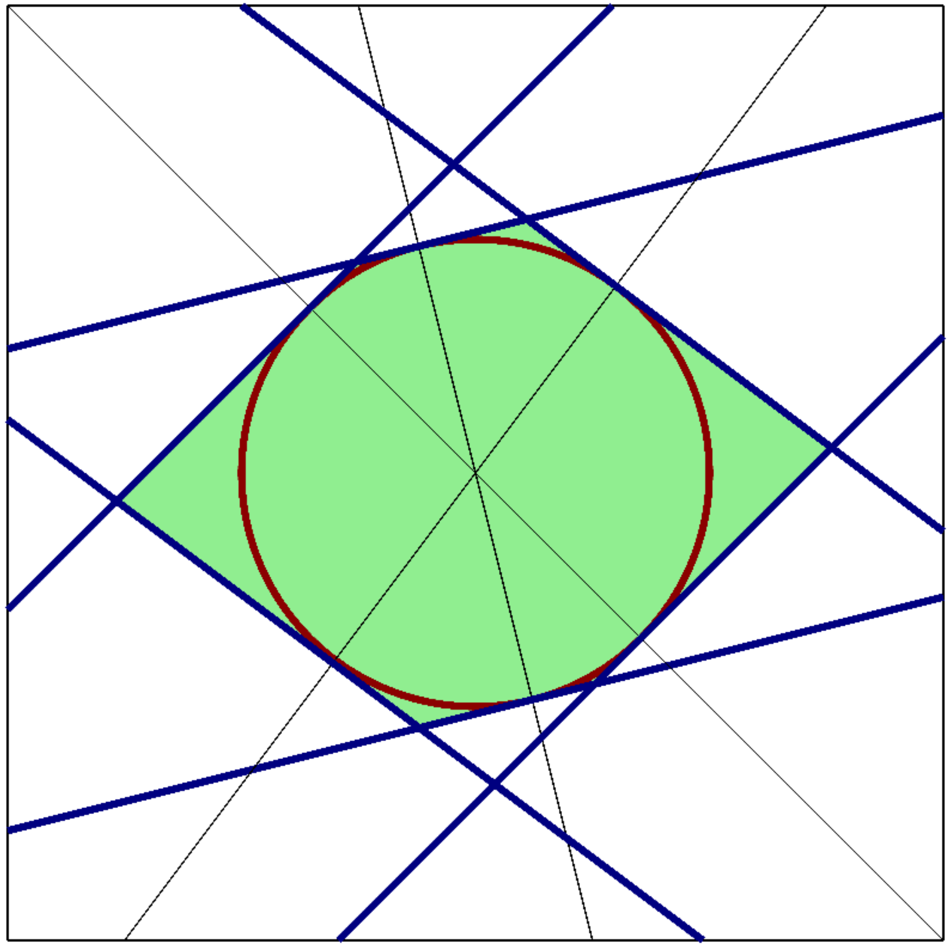
of polynomial dimension

where X_{ij} is R^d equipped with a top-k_{ij} norm

✓ 2) Solve ANN for $\bigoplus_{\ell_\infty} \bigoplus_{\ell_1} X_{ij}$

Embedding any norm into ℓ_∞

- **Thm:** for a normed space X and $\varepsilon > 0$ there exists $f: X \to \ell_\infty^{d'}$ with
$$\|f(x)\|_\infty \in (1 \pm \varepsilon) \cdot \|x\|_X$$
- **Idea**: $\|x\|_X \approx \max_{y \in N} |\langle x, y \rangle|$
- N is an ***ε*-net** of the unit ball $\boldsymbol{B}_{X^*}$ of the **dual norm**
  - $\|y\|_{X^*} = \sup_{\|x\|_X \leq 1} \langle x, y \rangle$
  - Can set $d' = (1/\varepsilon)^{O(d)}$
- Then: $\|x\|_X = \sup_{y \in B_{X^*}} \langle x, y \rangle \approx \max_{y \in N} \langle x, y \rangle$

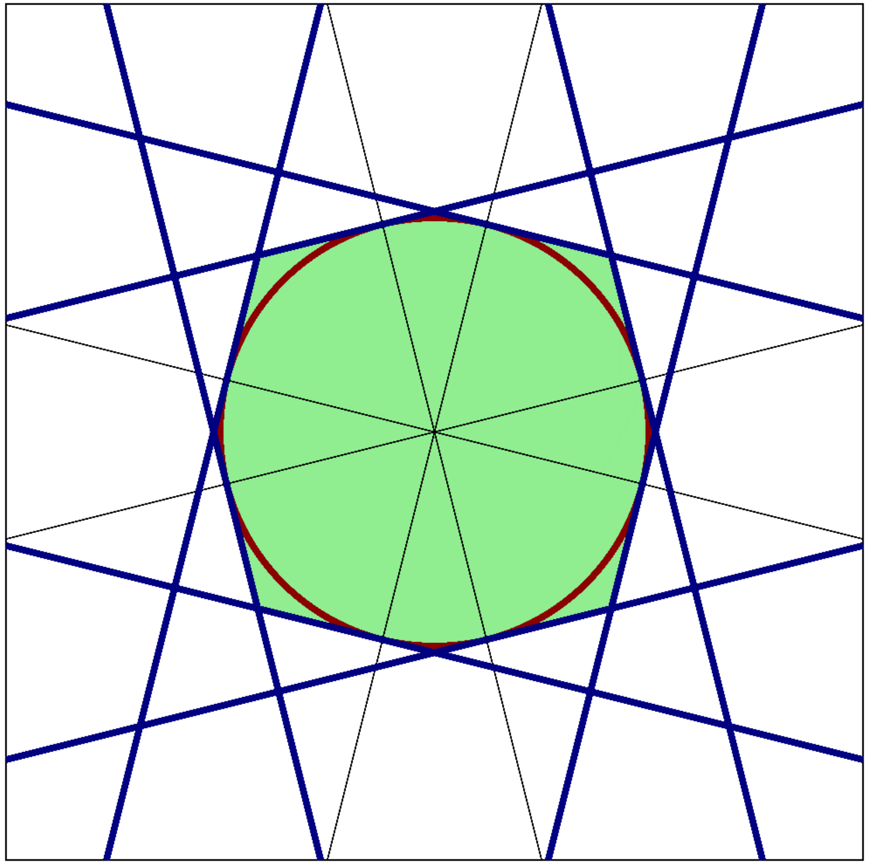


Better embedding for symmetric norm

- **Idea:** find a small net up to a symmetry
- Notation: $y_{\pi,\sigma}$ is y with permuted and flipped coordinates (acc to π and σ)
- Suppose that N is such that $\left\{y_{\pi,\sigma} \middle| y \in \overline{N}, \pi, \sigma\right\}$ is an ε-net of B_{X^*}
  - $\overline{N}$ is an ε-net of $B_{X^*} \cap \{y_1 \geq y_2 \geq \cdots \geq y_d \geq 0\}$
- Then, $\|x\|_X \approx \sup_{y \in \overline{N}, \pi, \sigma} \langle x, y_{\pi,\sigma} \rangle = \sup_{y \in \overline{N}} \sup_{\pi,\sigma} \langle x, y_{\pi,\sigma} \rangle$
- Claim: $\sup_{\pi,\sigma} \langle x, y_{\pi,\sigma} \rangle$ is a weighted sum of top-k norms of x
  - Hence, an embedding into $\bigoplus_{\ell_\infty} \bigoplus_{\ell_1} \mathrm{top}{-}k_{ij}$

New goal: find a small ε-net of $B_{X^*} \cap \{y_1 \geq y_2 \geq \cdots \geq y_d \geq 0\}$

Illustration



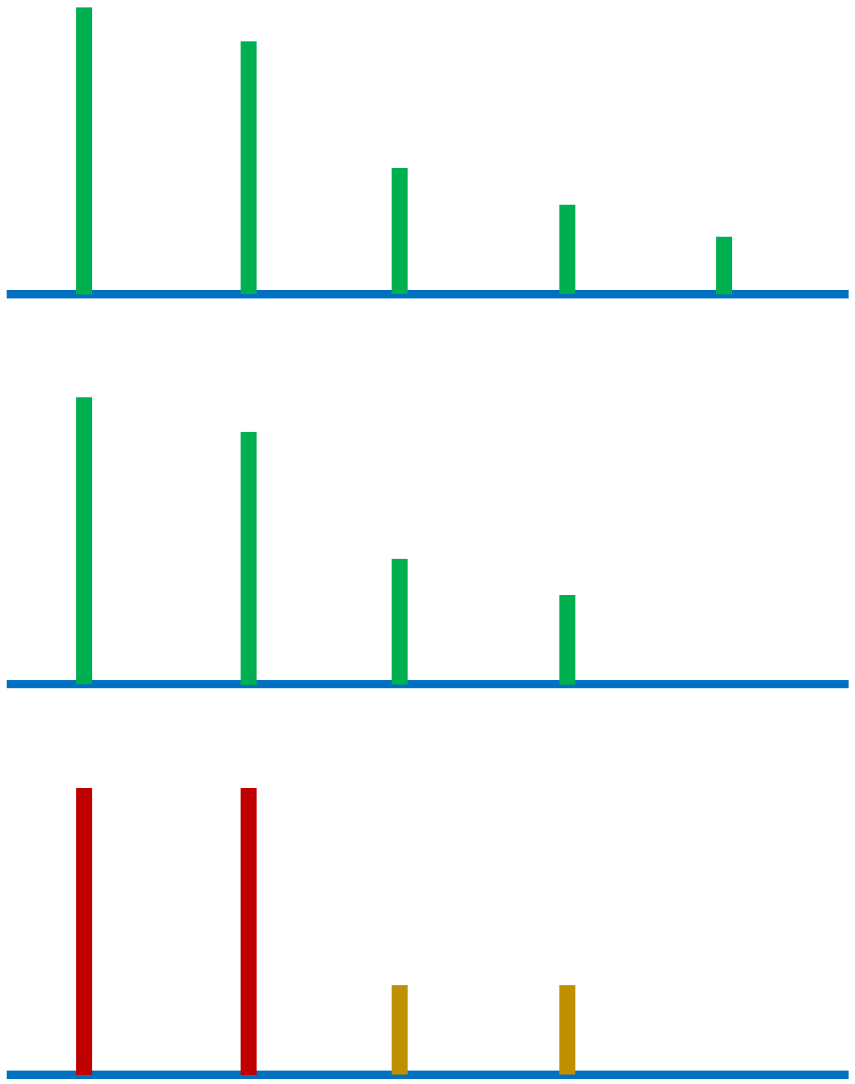
Small nets

New goal: find a small ε-net of $B_{X^*} \cap \{y_1 \geq y_2 \geq \cdots \geq y_d \geq 0\}$

- **Lemma:** can get of size $d^{O_\varepsilon(1)}$
- Will see a weaker bound of $d^{O_\varepsilon(\log d)}$, still non-trivial
  - Volume bound fails
  - Instead, a simple explicit construction

Small nets: continued

- Approximate $y \in B_{X^*}$ with $y_1 \geq y_2 \geq \cdots \geq y_d \geq 0$
- Zero small y_i's
- Round coordinates to a power of $(1 + \varepsilon)$
- $O_\varepsilon(\log d)$ scales
- Only cardinality of each scale matters
- $d^{O_\varepsilon(\log d)}$ vectors total
- Can be improved to $d^{O_\varepsilon(1)}$ by two more tricks



Summary

1) Embed any symmetric norm into $\bigoplus_{\ell_\infty} \bigoplus_{\ell_1} X_{ij}$,
($d^{O(1)}$-dimensional product space of top-k norms)

2) Solve ANN for $\bigoplus_{\ell_\infty} \bigoplus_{\ell_1} X_{ij}$

- reduce the ANN problem on the product space to ANN for the top-k norm
- use truncated exponential random variables to embed the top-k norm into ℓ_∞ and use a known ANN data structure there

An open question

- Improve approximation from $(\log\log n)^{O(1)}$ to $\mathrm{O}(\log\log d)$
  - Beyond $\log\log d$ is hard due to ℓ_∞
  - Need to bypass ANN for product spaces
  - **Randomized** embedding into low-dimensional ℓ_∞ for **any symmetric** norm?

General norms?

- Approximation $O(\sqrt{d})$ via embedding into ℓ_2
- **Symmetric** norms: by embedding into a **universal** $d^{O(1)}$-dimensional space

Is there an efficient ANN algorithm for general high-dimensional norms with approximation $d^{o(1)}$?

- Stronger hardness results?
- ~~Implied by: there is a family of spectral expanders that embed with distortion $O(1)$ into **some** $\log^{O(1)} n$-dimensional norm, where n is the number of nodes~~

[Naor 2017]