Simple Average-case Lower Bounds for Approximate Near-Neighbor from Isoperimetric Inequalities

Yitong Yin
Nanjing University
Nearest Neighbor Search (NNS)

metric space \((X,\text{dist})\)

database
\[ y = (y_1, y_2, \ldots, y_n) \in X^n \]

query \( x \in X \)

output: database point \( y_i \) closest to the query point \( x \)

applications: database, pattern matching, machine learning, ...
Near Neighbor Problem
($\lambda$-NN)

metric space $(X, \text{dist})$

database

$y = (y_1, y_2, \ldots, y_n) \in X^n$

query $x \in X$

data structure

radius $\lambda$

preprocessing

$\lambda$-NN: answer “yes” if $\exists y_i$ that is $\leq \lambda$-close to $x$

“no” if all $y_i$ are $>\lambda$-faraway from $x$
Approximate Near Neighbor (ANN)

metric space \((X, \text{dist})\)

database
\[ y = (y_1, y_2, \ldots, y_n) \in X^n \]

query \(x \in X\)

\((\gamma, \lambda)\)-ANN:
- **answer “yes”** if \(\exists y_i\) that is \(\leq \lambda\)-close to \(x\)
- **“no”** if all \(y_i\) are \(> \gamma \lambda\)-faraway from \(x\)
  - arbitrary if otherwise
Approximate Near Neighbor (ANN)

metric space \((X, \text{dist})\)

database
\[ y = (y_1, y_2, \ldots, y_n) \in X^n \]

query \(x \in X\)

data structure

radius \(\lambda\)

preprocessing

approximation ratio \(\gamma \geq 1\)

Hamming space
\[ X = \{0, 1\}^d \]
\[ \text{dist}(x, z) = \|x - z\|_1 \]

Curse of dimensionality!

\[ 100 \log n < d < n^{o(1)} \]

Hamming distance
Cell-Probe Model

data structure problem:

\[ f : X \times Y \rightarrow Z \]

query \( x \in X \)

algorithm \( A \):

\[(\text{decision tree})\]

t adaptive cell-probes

table

protocol: the pair \((A, T)\)

\((s, w, t)\)-cell-probing scheme

\[ T : Y \rightarrow \Sigma^s \]

where \( \Sigma = \{0, 1\}^w \)

database

\[ y \in Y \]
# Near-Neighbor Lower Bounds

**Hamming space** \( X = \{0, 1\}^d \)

**database size** \( n \)

**time:** \( t \) cell-probes;

**space:** \( s \) cells, each of size \( w \) bits

<table>
<thead>
<tr>
<th>Approximate Near-Neighbor (ANN)</th>
<th>Randomized Exact Near-Neighbor (RENN)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deterministic</strong></td>
<td><strong>Randomized</strong></td>
</tr>
<tr>
<td>( t = \Omega\left(\frac{d}{\log s}\right) )</td>
<td>( t = O(1) ) for ( s = \text{poly}(n) )</td>
</tr>
<tr>
<td>[Miltersen et al. 1995]</td>
<td>[Chakrabarti Regev 2004]</td>
</tr>
<tr>
<td>[Liu 2004]</td>
<td>[Borodin Ostrovsky Rabani 1999]</td>
</tr>
<tr>
<td>[Pătraşcu Thorup 2006]</td>
<td>[Barkol Rabani 2000]</td>
</tr>
<tr>
<td>( t = \Omega\left(\frac{\log n}{\log \log n}\right) )</td>
<td>( t = \Omega\left(\frac{\log m}{\log \log \log n}\right) )</td>
</tr>
</tbody>
</table>

- \( t = \Omega\left(\frac{d}{\log n}\right) \)

- matches the highest known lower bounds for any data structure problems:
  - Polynomial Evaluation [Larsen’12], ball-inheritance (range reporting) [Grønlund, Larsen’16]
Why are data structure lower bounds so difficult?

• (Observed by [Miltersen et al. 1995]) An $\omega(\log n)$ cell-probe lower bound on polynomial space for any function in $\mathbf{P}$ would prove $\mathbf{P} \not\subseteq$ linear-time poly-size Boolean branching programs. (Solved in [Ajtai 1999])

• (Observed by [Brody, Larsen 2012]) Even non-adaptive data structures are circuits with arbitrary gates of depth 2:

\[
f : X \times Y \rightarrow Z
\]

Table cells:

Data $y$

\[
y_1 \quad y_2 \quad \cdots
\]

\[f(x,y)\]

\[t \text{ fan-in}\]

\[f(x',y)\]

\[\cdots\]

\[s \text{ fan-in & -out}\]

\[y_{n-1} \quad y_n\]
# Near-Neighbor Lower Bounds

Hamming space $X = \{0, 1\}^d$

- **database size:** $n$
- **time:** $t$ cell-probes;
- **space:** $s$ cells, each of $w$ bits

<table>
<thead>
<tr>
<th>Approximate Near-Neighbor (ANN)</th>
<th>Randomized Exact Near-Neighbor (RENN)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deterministic</strong></td>
<td><strong>Randomized</strong></td>
</tr>
<tr>
<td>$t = \Omega \left( \frac{d}{\log s} \right)$</td>
<td>$t = \Omega \left( \frac{d}{\log s} \right)$</td>
</tr>
<tr>
<td>[Miltersen et al. 1995]</td>
<td>[Borodin Ostrovsky Rabani 1999]</td>
</tr>
<tr>
<td>[Liu 2004]</td>
<td>[Barkol Rabani 2000]</td>
</tr>
<tr>
<td>$t = \Omega \left( \frac{d}{\log \frac{sw}{n}} \right)$</td>
<td>$t = \Omega \left( \frac{\log n}{\log \frac{sw}{n}} \right)$</td>
</tr>
<tr>
<td>$t = \Omega \left( \frac{d}{\log \frac{sw}{n d}} \right)$</td>
<td>$t = \Omega \left( \frac{d}{\log \frac{sw}{n}} \right)$</td>
</tr>
<tr>
<td>[Wang Y. 2014]</td>
<td>[Pătraşcu Thorup 2006]</td>
</tr>
</tbody>
</table>
Average-Case Lower Bounds

• **Hard distribution**: [Barkol Rabani 2000] [Liu 2004] [PTW’08 ’10]
  • database: \( y_1, ..., y_n \in \{0,1\}^d \) *i.i.d. uniform*
  • query: uniform and independent \( x \in \{0,1\}^d \)

• **Expected cell-probe complexity:**
  • \( E_{(x,y)}[# \text{ of cell-probes to resolve query } x \text{ on database } y] \)

• “Curse of dimensionality” should hold on average.

• In **data-dependent LSH** [Andoni Razenshtein 2015]: a key step is to solve the problem on random input.
## Average-Case Lower Bounds

**Hamming space** \( X = \{0, 1\}^d \)

**Database size**: \( n \)

**Time**: \( t \) cell-probes;  

**Space**: \( s \) cells, each of \( w \) bits

<table>
<thead>
<tr>
<th>Approximate Near-Neighbor (ANN)</th>
<th>Randomized Exact Near-Neighbor (RENN)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deterministic</strong></td>
<td><strong>Randomized</strong></td>
</tr>
<tr>
<td>( t = \Omega \left( \frac{d}{\log s} \right) )</td>
<td>( t = \Omega \left( \frac{d}{\log s} \right) )</td>
</tr>
<tr>
<td>[Miltersen et al. 1995]</td>
<td>[Borodin Ostrovsky Rabani 1999]</td>
</tr>
<tr>
<td>[Liu 2004]</td>
<td>[Barkol Rabani 2000]</td>
</tr>
<tr>
<td>( t = \Omega \left( \frac{d}{\log \frac{sw}{n^2}} \right) )</td>
<td>( t = \Omega \left( \frac{d}{\log \frac{sw}{n}} \right) )</td>
</tr>
<tr>
<td>( t = \Omega \left( \frac{d}{\log \frac{sw}{n^2}} \right) )</td>
<td>( t = \Omega \left( \frac{d}{\log \frac{sw}{n}} \right) )</td>
</tr>
<tr>
<td>[Wang Y. 2014]</td>
<td>[Pătraşcu Thorup 2006]</td>
</tr>
</tbody>
</table>
**Average-Case Lower Bounds**

Hamming space $X = \{0, 1\}^d$

database size: $n$

time: $t$ cell-probes; space: $s$ cells, each of $w$ bits

<table>
<thead>
<tr>
<th>Approximate Near-Neighbor (ANN)</th>
<th>Randomized Exact Near-Neighbor (RENN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic</td>
<td>Deterministic</td>
</tr>
<tr>
<td>$t = \Omega \left( \frac{d}{\log s} \right)$</td>
<td>$t = \Omega \left( \frac{d}{\log s} \right)$</td>
</tr>
<tr>
<td>[Miltersen et al. 1995]</td>
<td>[Borodin Ostrovsky Rabani 1999]</td>
</tr>
<tr>
<td>[Liu 2004]</td>
<td>[Barkol Rabani 2000]</td>
</tr>
<tr>
<td><strong>our result:</strong></td>
<td></td>
</tr>
<tr>
<td>$t = \Omega \left( \frac{d}{\log \frac{sw}{nd}} \right)$</td>
<td>$t = \Omega \left( \frac{\log n}{\log \frac{sw}{n}} \right)$</td>
</tr>
<tr>
<td></td>
<td>[Panigrahy Talwar Wieder 2008, 2010]</td>
</tr>
</tbody>
</table>
Metric Expansion

[Panigrahy Talwar Wieder 2010]

metric space \((X, \text{dist})\)

\(\lambda\)-neighborhood: \( \forall x \in X, \ N_\lambda(x) = \{ z \in X \mid \text{dist}(x, z) \leq \lambda \} \)
\( \forall A \subseteq X, \ N_\lambda(A) = \{ z \in X \mid \exists x \in A \text{ s.t. } \text{dist}(x, z) \leq \lambda \} \)

probability distribution \(\mu\) over \(X\)

- \(\lambda\)-neighborhoods are weakly independent under \(\mu\):
  \( \forall x \in X, \ \mu(N_\lambda(x)) < 0.99/n \)

- \(\lambda\)-neighborhoods are \((\Phi, \Psi)\)-expanding under \(\mu\):
  \( \forall A \subseteq X, \ \mu(A) \geq 1/\Phi \Rightarrow \mu(N_\lambda(A)) \geq 1 - 1/\Psi \)
Metric Expansion

[Panigrahy Talwar Wieder 2010]

metric space \((X, \text{dist})\) probability distribution \(\mu\) over \(X\)

- \(\lambda\)-neighborhoods are \((\Phi, \Psi)\)-expanding under \(\mu\):
  \[\forall A \subseteq X, \mu(A) \geq 1/\Phi \Rightarrow \mu(N_\lambda(A)) \geq 1-1/\Psi\]

vertex expansion, “blow-up” effect
Main Theorem:

For $(\gamma, \lambda)$-ANN in metric space $(X, \text{dist})$ where

- $\gamma \lambda$-neighborhoods are **weakly independent** under $\mu$:
  \[ \mu(N_{\gamma \lambda}(x)) < 0.99/n \text{ for } \forall x \in X \]

- $\lambda$-neighborhoods are $(\Phi, \Psi)$-expanding under $\mu$:
  \[ \forall A \subseteq X \text{ that } \mu(A) \geq 1/\Phi \Rightarrow \mu(N_{\lambda}(A)) \geq 1 - 1/\Psi \]

\forall deterministic algorithm that makes $t$ cell-probes in expectation on a table of size $s$ cells, each of $w$ bits (assuming $w + \log s < n / \log \Phi$), under the **input distribution**:

**database** $y = (y_1, y_2, ..., y_n)$ where $y_1, y_2, ..., y_n \sim \mu$, i.i.d.

**query** $x \sim \mu$, independently

\[ t = \Omega \left( \frac{\log \Phi}{\log \frac{sw}{n \log \Psi}} \right) \]
Main Theorem:

For \((\gamma, \lambda)\)-ANN in metric space \((X, \text{dist})\) where

- \(\gamma \lambda\)-neighborhoods are **weakly independent** under \(\mu\):
  \[
  \mu(N_{\gamma \lambda}(x)) < \frac{0.99}{n} \text{ for } \forall x \in X
  \]
- \(\lambda\)-neighborhoods are \((\Phi, \Psi)\)-expanding under \(\mu\):
  \[
  \forall A \subseteq X \text{ that } \mu(A) \geq \frac{1}{\Phi} \Rightarrow \mu(N_{\lambda}(A)) \geq 1 - \frac{1}{\Psi}
  \]

\(\forall\) **deterministic** algorithm that makes \(t\) cell-probes *in expectation* on a table of size \(s\) cells, each of \(w\) bits (assuming \(w + \log s < n / \log \Phi\)), under the **input distribution**:

- database \(y = (y_1, y_2, \ldots, y_n)\) where \(y_1, y_2, \ldots, y_n \sim \mu\), i.i.d.
- query \(x \sim \mu\), independently

\[
\text{\arr{green} \quad t = \Omega \left( \frac{\log \Phi}{\log \left( \frac{sw}{n \log \Psi} \right)} \right)}
\]
The Richness Lemma

\[ f(x, y) \]

\[ f : X \times Y \rightarrow \{0, 1\} \]

\[ x \in X \quad \text{cell-probing algorithm} \]

\[ t \log s \]

\[ \mu(A) \geq 2^{-O(t \log s)} \]

\[ \nu(B) \geq 2^{-O(t \log s + tw)} \]

\[ y \in Y \quad \text{table} \ (s \text{ cells, each of } w \text{ bits}) \]

distributions \( \mu \) over \( X \), \( \nu \) over \( Y \)

\( \alpha \)-dense: density of 1s \( \geq \alpha \) under \( \mu \times \nu \)

monochromatic 1-rectangle: \( A \times B \) with \( A \subseteq X \), \( B \subseteq Y \)

\( s.t. \ \forall (x, y) \in A \times B, f(x, y) = 1 \)

Richness lemma (Miltersen, Nisan, Safra, Wigderson, 1995)

\[ f \text{ is } 0.01\text{-dense under } \mu \times \nu \]

\( f \text{ has } (s, w, t)\text{-cell-probing scheme} \quad \Rightarrow \quad f \text{ has 1-rectangle } A \times B \text{ with } \]

\[ \left\{ \begin{array}{l}
\mu(A) \geq 2^{-O(t \log s)} \\
\nu(B) \geq 2^{-O(t \log s + tw)}
\end{array} \right.\]
A New Richness Lemma

\[ f : X \times Y \to \{0, 1\} \] distributions \( \mu \) over \( X \), \( \nu \) over \( Y \)

**Richness lemma** (Miltersen, Nisan, Safra, Wigderson, 1995)

\[ f \text{ is } 0.01\text{-dense under } \mu \times \nu \]
\[ f \text{ has } (s, w, t)\text{-cell-probing scheme} \]
\[ f \text{ has } 1\text{-rectangle } A \times B \text{ with} \]
\[ \mu(A) \geq 2^{-O(t \log s)} \]
\[ \nu(B) \geq 2^{-O(t \log s + tw)} \]

**New Richness lemma**

\[ f \text{ is } 0.01\text{-dense under } \mu \times \nu \]
\[ f \text{ has average-case } (s, w, t)\text{-cell-probing scheme} \]
\[ f \text{ has } 1\text{-rectangle } A \times B \text{ with} \]
\[ \forall \Delta \in [320000t, s], \]
\[ \mu(A) \geq 2^{-O(t \log (s/\Delta))} \]
\[ \nu(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} \]

when \( \Delta = O(t) \), it becomes the richness lemma (with slightly better bounds)
\( f : X \times Y \rightarrow \{0, 1\} \) distributions \( \mu \) over \( X \), \( \nu \) over \( Y \)

**New Richness lemma**

\( f \) is 0.01-dense under \( \mu \times \nu \) \( \rightarrow \forall \Delta \in [320000t, s], \)

\( f \) has average-case \((s, w, t)\)-cell-probing scheme under \( \mu \times \nu \)

\( \begin{align*}
\mu(A) &\geq 2^{-O(t \log (s/\Delta))} \\
\nu(B) &\geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)}
\end{align*} \)

metric space \((X, \text{dist})\), query \( x \in X \), database \( y = (y_1, \ldots, y_n) \in X^n \)

\((\gamma, \lambda)\)-ANN: \( f(x, y) = \bigwedge_{i=1}^{n} g(x, y_i) \)

where \( g(x, y_i) = \begin{cases} 
1 & \text{dist}(x, y_i) > \gamma \lambda \\
0 & \text{dist}(x, y_i) \leq \lambda \\
\star & \text{otherwise}
\end{cases} \)

Other examples: partial match, membership, range query, ...
**New Richness lemma**

\[ f \text{ is } 0.01\text{-dense under } \mu \times \nu \] \[ \Rightarrow \] \[ \forall \Delta \in [320000t,s], \quad f \text{ has } 1\text{-rectangle } A \times B \text{ with} \] \[ \begin{cases} \mu(A) \geq 2^{-O(t \log (s/\Delta))} \\ \nu(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} \end{cases} \]

- \( \gamma \lambda \)-neighborhoods are weakly independent under \( \mu \):
  \[ \mu(N_{\gamma \lambda}(x)) < 0.99/n \text{ for } \forall x \in X \]
  \[ \Rightarrow \] density of 0s in \( g \) is \( \leq 0.99/n \) under \( \mu \times \mu \)

- \( \lambda \)-neighborhoods are \((\Phi, \Psi)\)-expanding under \( \mu \):
  \[ \forall A \subseteq X, \mu(A) \geq 1/\Phi \Rightarrow \mu(N_{\lambda}(A)) \geq 1-1/\Psi \]
  \[ \Rightarrow \] \( g \) does not have 1-rectangle \( A \times C \) with \( \mu(A) > 1/\Phi \) and \( \mu(C) > 1/\Psi \)

- \( f \) does not have 1-rectangle \( A \times B \) with \( \mu(A) > 1/\Phi \) and \( \mu^n(B) > 1/\Psi^n \)

choose \( \Delta = O\left(\frac{n \log \Psi}{w}\right) \) so that \( \mu^n(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} > 1/\Psi^n \)

\[ 1/\Phi \geq \mu(A) \geq 2^{-O(t \log (s/\Delta))} \]

\[ t = \Omega\left(\frac{\log \Phi}{\log \frac{sw}{n \log \Psi}}\right) \]
New Richness lemma

\( f \) is 0.01-dense under \( \mu \times \nu \)

\( f \) has average-case \((s,w,t)\)-cell-probing scheme under \( \mu \times \nu \)

\[ \forall \Delta \in [320000t, s], \]
\[ f \text{ has 1-rectangle } A \times B \text{ with } \]
\[ \left\{ \begin{array}{l}
\mu(A) \geq 2^{-O(t \log (s/\Delta))} \\
\nu(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)}
\end{array} \right. \]

\[ \geq 0.0025 \text{-fraction (under } \nu \text{) of databases } y \in Y \text{ are “good”:} \]

s.t. \( \forall \text{ good database } y, \)

\[ \left\{ \begin{array}{l}
\geq 0.005 \text{-fraction of queries } x \in X \text{ are positive} \\
\text{avg. cell-probes for positive queries } \leq 80000t
\end{array} \right. \]

positive queries:

\[ T_y : \]

\[ \exists \Delta \text{ cells resolving } 2^{-O(t \log (s/\Delta))} \text{ fraction (under } \mu \text{) positive queries} \]
New Richness lemma

\[ f \text{ is } 0.01\text{-dense under } \mu \times \nu \]

\[ f \text{ has average-case } (s, w, t)\text{-cell-probing scheme under } \mu \times \nu \]

\[ \forall \Delta \in [320000t, s], \]

\[ f \text{ has } 1\text{-rectangle } A \times B \text{ with } \]

\[ \left\{ \begin{array}{l}
\mu(A) \geq 2^{-O(t \log (s/\Delta))} \\
\nu(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)}
\end{array} \right. \]

\[ \geq 0.0025\text{-fraction (under } \nu \text{) of databases } y \in Y \text{ are "good"} : \]

s.t. \( \forall \text{ good database } y, \)

\[ \exists \Delta \text{ cells resolving } 2^{-O(t \log (s/\Delta))} \text{ fraction (under } \mu \text{) positive queries} \]

\[ T_y \]

\[ \{w \text{ bits} \}

\[ \leq \left( \frac{s}{\Delta} \right) 2^{\Delta w} = 2^{O(\Delta \log \frac{s}{\Delta} + \Delta w)} \text{ possibilities} \]

\[ \text{good } y \longleftrightarrow \omega \]

\[ B : \]

\[ \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} \text{ fraction (under } \nu \text{) good } y \longrightarrow \text{ the same } \omega \]

\[ \text{cell-probe model: once } \omega \text{ is fixed,} \]

\[ A : \]

\[ \text{the set of positive queries resolved by } \omega \text{ is fixed} \]
\( f : X \times Y \rightarrow \{0, 1\} \) distributions \( \mu \) over \( X \), \( \nu \) over \( Y \)

**New Richness lemma**

\( f \) is 0.01-dense under \( \mu \times \nu \) \( \iff \forall \Delta \in [320000t, s] \),

\( f \) has average-case \((s, w, t)\)-cell-probing scheme under \( \mu \times \nu \)

\[
\begin{align*}
\mu(A) & \geq 2^{-O(t \log (s/\Delta))} \\
\nu(B) & \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)}
\end{align*}
\]
Main Theorem:

For \((\gamma, \lambda)\)-ANN in metric space \((X, \text{dist})\) where

- \(\gamma \lambda\)-neighborhoods are \textbf{weakly independent} under \(\mu\):
  \[ \mu(N_{\gamma \lambda}(x)) < 0.99/n \text{ for } \forall x \in X \]

- \(\lambda\)-neighborhoods are \((\Phi, \Psi)\)-expanding under \(\mu\):
  \[ \forall A \subseteq X \text{ that } \mu(A) \geq 1/\Phi \Rightarrow \mu(N_{\lambda}(A)) \geq 1 - 1/\Psi \]

\(\forall \text{ deterministic} \) algorithm that makes \(t\) cell-probes \textit{in expectation} on a table of size \(s\) cells, each of \(w\) bits (assuming \(w + \log s < n / \log \Phi\)), under the \textit{input distribution}:

- database \(y = (y_1, y_2, \ldots, y_n)\) where \(y_1, y_2, \ldots, y_n \sim \mu, \text{ i.i.d.}\)
- query \(x \sim \mu, \text{ independently}\)

\[ t = \Omega \left( \frac{\log \Phi}{\log \frac{sw}{n \log \Psi}} \right) \]
# Average-Case Lower Bounds

**Hamming space** \( X = \{0, 1\}^d \)  
**database size:** \( n \)  
**time:** \( t \) cell-probes;  
**space:** \( s \) cells, each of \( w \) bits  
- **database:** \( y_1, \ldots, y_n \in \{0, 1\}^d \) i.i.d. uniform  
- **query:** uniform and independent \( x \in \{0, 1\}^d \)

## Approximate Near-Neighbor (ANN)

<table>
<thead>
<tr>
<th>Deterministic</th>
<th>Randomized</th>
<th>Randomized Exact Near-Neighbor (RENN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t = \Omega\left(\frac{d}{\log s}\right) )</td>
<td>( t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right) )</td>
<td>( t = \Omega\left(\frac{d}{\log s}\right) )</td>
</tr>
<tr>
<td>[Liu 2004]</td>
<td></td>
<td>[Barkol Rabani 2000]</td>
</tr>
</tbody>
</table>

**our result:**  
\( t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right) \)
Thank you!