Dimensionality reduction ($\ell_2$-to-$\ell_2$)

A (randomized) map from $\mathbb{R}^n$ to $\mathbb{R}^m$ that "preserves geometry" of vectors.
Dimensionality reduction ($\ell_2$-to-$\ell_2$)

A (randomized) map from $\mathbb{R}^n$ to $\mathbb{R}^m$ that “preserves geometry” of vectors.

A pre-processing step in many applications:

- Document classification tasks (Weinberger et al. '09, etc)
- SVMs (Paul et al. '14)
- k-means/k-medians (Makarychev, Makarychev, Razenshteyn '18)
- Nearest neighbors (Ailon, Chazelle '09, Har-Peled et al. '14, Wei '19)
- Numerical linear algebra (Clarkson and Woodruff '12, Nelson and Nguyen '14, etc.)

Our contribution:
Dimensionality reduction ($\ell_2$-to-$\ell_2$)

A (randomized) map from $\mathbb{R}^n$ to $\mathbb{R}^m$ that “preserves geometry” of vectors.

A pre-processing step in many applications:

- Document classification tasks (Weinberger et al. ’09, etc)
- SVMs (Paul et al. ’14)
- k-means/k-medians (Makarychev, Makarychev, Razenshteyn ’18)
- Nearest neighbors (Ailon, Chazelle ’09, Har-Peled et al. ’14, Wei ’19)
- Numerical linear algebra (Clarkson and Woodruff ’12, Nelson and Nguyen ’14, etc.)
Dimensionality reduction ($l_2$-to-$l_2$)

A (randomized) map from $\mathbb{R}^n$ to $\mathbb{R}^m$ that “preserves geometry” of vectors.

A pre-processing step in many applications:
- Document classification tasks (Weinberger et al. ’09, etc)
- SVMs (Paul et al. ’14)
- k-means/k-medians (Makarychev, Makarychev, Razenshteyn ’18)
- Nearest neighbors (Ailon, Chazelle ’09, Har-Peled et al. ’14, Wei ’19)
- Numerical linear algebra (Clarkson and Woodruff ’12, Nelson and Nguyen ’14, etc.)

**Our contribution:** Theoretical analysis of a state-of-the-art dimensionality reduction scheme on feature vectors. Could inform how to optimally set parameters in practice.
Feature hashing (Weinberger et al. ’09)

Use a hash function $h : \{1, \ldots, n\} \rightarrow \{1, \ldots, m\}$ on coordinates.
Feature hashing (Weinberger et al. ’09)

Use a **hash function** \( h : \{1, \ldots, n\} \rightarrow \{1, \ldots, m\} \) on coordinates.
Feature hashing (Weinberger et al. ’09)

Use a **hash function** $h : \{1, \ldots, n\} \rightarrow \{1, \ldots, m\}$ on coordinates.

How should collisions be handled?
Feature hashing (Weinberger et al. ’09)

Use a hash function $h : \{1, \ldots, n\} \rightarrow \{1, \ldots, m\}$ on coordinates.

Use random signs to handle collisions (unbiased estimator of $\ell_2^2$ norm).
Use many hash functions $h_1, h_2, \ldots, h_s$:

$$\{1, \ldots, n\} \rightarrow \{1, \ldots, m\}.$$  

Anti-correlate hash functions so $h_j(i) \neq h_k(i)$.

Use random signs to deal with collisions.

Scale the resulting vector by $\frac{1}{\sqrt{s}}$.

Sparse JL is a state-of-the-art sparse dimensionality reduction.

Central question:

How should the number of hash functions $s$ and dimension $m$ be set?
Use many hash functions $h_1, h_2, \ldots, h_s : \{1, \ldots, n\} \rightarrow \{1, \ldots, m\}$.
Use many hash functions $h_1, h_2, \ldots, h_s : \{1, \ldots, n\} \rightarrow \{1, \ldots, m\}$.

- Anti-correlate hash functions so $h_j(i) \neq h_k(i)$.
Use many hash functions $h_1, h_2, \ldots, h_s : \{1, \ldots, n\} \rightarrow \{1, \ldots, m\}$.

- Anti-correlate hash functions so $h_j(i) \neq h_k(i)$.

Use random signs to deal with collisions.
Use many hash functions $h_1, h_2, \ldots, h_s : \{1, \ldots, n\} \rightarrow \{1, \ldots, m\}$.

- Anti-correlate hash functions so $h_j(i) \neq h_k(i)$.

Use random signs to deal with collisions.

Scale the resulting vector by $\frac{1}{\sqrt{s}}$. 

Sparse Johnson-Lindenstrauss transform (KN '12)
Use many hash functions $h_1, h_2, \ldots, h_s : \{1, \ldots, n\} \rightarrow \{1, \ldots, m\}$.

- Anti-correlate hash functions so $h_j(i) \neq h_k(i)$.

Use random signs to deal with collisions.

Scale the resulting vector by $\frac{1}{\sqrt{s}}$.

**Sparse JL is a state-of-the-art sparse dimensionality reduction.**
Sparse Johnson-Lindenstrauss transform (KN ’12)

Use many hash functions $h_1, h_2, \ldots, h_s : \{1, \ldots, n\} \rightarrow \{1, \ldots, m\}$.
  - Anti-correlate hash functions so $h_j(i) \neq h_k(i)$.

Use random signs to deal with collisions.

Scale the resulting vector by $\frac{1}{\sqrt{s}}$.

**Sparse JL is a state-of-the-art sparse dimensionality reduction.**

**Central question**

*How should the # of hash functions $s$ and dimension $m$ be set?*
Intuition for our contribution
The function $v$ captures the performance of sparse JL on feature vectors.
Mathematical framework

Use a probability distribution $F$ over maps $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$.

What does it mean to “preserve geometry”?

For each $x, y \in \mathbb{R}^n$:

$$P_{f \in F} \left[ (1 - \epsilon) \| x - y \|_2 \leq \| f(x) - f(y) \|_2 \leq (1 + \epsilon) \| x - y \|_2 \right] > 1 - \delta,$$

for $\epsilon$ target error, $\delta$ target failure probability.

Focus on linear maps:

$$P_{f \in F} \left[ (1 - \epsilon) \| x \|_2 \leq \| f(x) \|_2 \leq (1 + \epsilon) \| x \|_2 \right] > 1 - \delta.$$
Use a probability distribution $\mathcal{F}$ over maps $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. 

What does it mean to "preserve geometry"?

For each $x, y \in \mathbb{R}^n$:

$$P_{f \in \mathcal{F}} \left[ (1 - \epsilon) \| x - y \|^2 \leq \| f(x) - f(y) \|^2 \leq (1 + \epsilon) \| x - y \|^2 \right] > 1 - \delta,$$

for $\epsilon$ target error, $\delta$ target failure probability.

Focus on linear maps:

$$P_{f \in \mathcal{F}} \left[ (1 - \epsilon) \| x \|^2 \leq \| f(x) \|^2 \leq (1 + \epsilon) \| x \|^2 \right] > 1 - \delta.$$
Mathematical framework

Use a probability distribution $\mathcal{F}$ over maps $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$.

What does it mean to “preserve geometry”? 

For each $x, y \in \mathbb{R}^n$:

$$P_{f \in \mathcal{F}} \left[ (1 - \epsilon) \| x - y \|_2 \leq \| f(x) - f(y) \|_2 \leq (1 + \epsilon) \| x - y \|_2 \right] > 1 - \delta,$$

for $\epsilon$ target error, $\delta$ target failure probability.

Focus on linear maps:

$$P_{f \in \mathcal{F}} \left[ (1 - \epsilon) \| x \|_2 \leq \| f(x) \|_2 \leq (1 + \epsilon) \| x \|_2 \right] > 1 - \delta.$$
Use a probability distribution $\mathcal{F}$ over maps $f : \mathbb{R}^n \to \mathbb{R}^m$.

What does it mean to “preserve geometry”?

For each $x, y \in \mathbb{R}^n$:

$$\mathbb{P}_{f \in \mathcal{F}}[(1 - \epsilon) \|x - y\|_2 \leq \|f(x) - f(y)\|_2 \leq (1 + \epsilon) \|x - y\|_2] > 1 - \delta,$$

for $\epsilon$ target error, $\delta$ target failure probability.
Mathematical framework

Use a probability distribution $\mathcal{F}$ over maps $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$.

What does it mean to “preserve geometry”? 

For each $x, y \in \mathbb{R}^n$:

$$\mathbb{P}_{f \in \mathcal{F}}[(1 - \epsilon) \|x - y\|_2 \leq \|f(x) - f(y)\|_2 \leq (1 + \epsilon) \|x - y\|_2] > 1 - \delta,$$

for $\epsilon$ target error, $\delta$ target failure probability.

Focus on linear maps:

$$\mathbb{P}_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta.$$
Goal: \( P_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta. \)
Performance on feature vectors (Weinberger et al. ’09)

Goal: $\mathbb{P}_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta$.

Sometimes a much smaller $m$ works on feature vectors in practice than traditional theory on $\mathbb{R}^n$ suggests...
Performance on feature vectors (Weinberger et al. '09)

Goal: \( \mathbb{P}_{f \in \mathcal{F}}[ (1 - \epsilon) \| x \|_2 \leq \| f(x) \|_2 \leq (1 + \epsilon) \| x \|_2 ] > 1 - \delta. \)

Sometimes a much smaller \( m \) works on feature vectors in practice than traditional theory on \( \mathbb{R}^n \) suggests...

Consider vectors w/ small \( \ell_\infty \)-to-\( \ell_2 \) norm ratio:

\[
S_v = \{ x \in \mathbb{R}^n \mid \| x \|_\infty \leq v \| x \|_2 \}.
\]
Goal: \( \mathbb{P}_{f \in F} [(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta. \)

Sometimes a much smaller \( m \) works on feature vectors in practice than traditional theory on \( \mathbb{R}^n \) suggests...

Consider vectors w/ small \( \ell_\infty \)-to-\( \ell_2 \) norm ratio:

\[
S_v = \{ x \in \mathbb{R}^n \mid \|x\|_\infty \leq v \|x\|_2 \}.
\]

\( v(m, \epsilon, \delta, s) := \sup \text{ over } v \in [0, 1] \text{ s.t. sparse JL meets } \ell_2 \text{ goal on } x \in S_v. \)
Goal: $\mathbb{P}_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta$.

Sometimes a much smaller $m$ works on feature vectors in practice than traditional theory on $\mathbb{R}^n$ suggests...

Consider vectors w/ small $\ell_\infty$-to-$\ell_2$ norm ratio:

$$S_v = \{ x \in \mathbb{R}^n \mid \|x\|_\infty \leq v \|x\|_2 \}.$$  

$v(m, \epsilon, \delta, s) := \sup$ over $v \in [0, 1]$ s.t. sparse JL meets $\ell_2$ goal on $x \in S_v$.

**We give a tight theoretical analysis of the function $v(m, \epsilon, \delta, s)$, that could inform how to optimally set $s$ and $m$ in practice.**
Informal statement of main result

Goal: $\mathbb{P}_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta$.

$v(m, \epsilon, \delta, s) := \sup \forall v \in [0, 1]$ s.t. sparse JL meets $\ell_2$ goal on $x \in S_v$. 

Theorem (Informal)

Sparse JL has four regimes in terms of how it performs on norm preservation. For error $\epsilon$ and failure probability $\delta$, sparse JL with projected dimension $m$ and $s$ hash functions has performance $v(m, \epsilon, \delta, s)$ equal to:

- $1$ (full performance) High $m\sqrt{sB_1}$
- $m\sqrt{s\min(B_1, B_2)}$ (partial performance) Middle $m\sqrt{s\min(B_1, B_2)}$
- $0$ (poor performance) Small $m$.

where $B_1, B_2$ are functions of $m, \epsilon, \delta$. 

Understanding Sparse JL for Feature Hashing Meena Jagadeesan (Harvard University)
Informal statement of main result

Goal: $\mathbb{P}_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta$.

$v(m, \epsilon, \delta, s) := \sup_{\nu \in [0, 1]} \text{s.t. sparse JL meets } \ell_2 \text{ goal on } x \in S_\nu.$

Theorem (Informal)

Sparse JL has four regimes in terms of how it performs on norm preservation. For error $\epsilon$ and failure probability $\delta$, sparse JL with projected dimension $m$ and $s$ hash functions has performance $v(m, \epsilon, \delta, s)$ equal to:

\[
\begin{align*}
1 \text{ (full performance)} & \quad \text{High } m \\
\sqrt{s} B_1 \text{ (partial performance)} & \quad \text{Middle } m \\
\sqrt{s} \min(B_1, B_2) \text{ (partial performance)} & \quad \text{Middle } m \\
0 \text{ (poor performance)} & \quad \text{Small } m,
\end{align*}
\]

where $B_1, B_2$ are functions of $m, \epsilon, \delta$. 

Understanding Sparse JL for Feature Hashing
Meena Jagadeesan (Harvard University)
$v(m, \epsilon, \delta, s)$ on more synthetic data
$v(m, \epsilon, \delta, s)$ on more synthetic data
Sparse JL on News20 dataset

Sparse JL with $\geq 4$ hash functions can perform much better than feature hashing in practice.
Sparse JL on News20 dataset

Sparse JL with \( \geq 4 \) hash functions can perform much better than feature hashing in practice.
Sparse JL on News20 dataset

Sparse JL with $\geq 4$ hash functions can perform much better than feature hashing in practice.
Comparison to previous work

Goal: $\mathbb{P}_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta$.

$v(m, \epsilon, \delta, s) := \sup_{v \in [0, 1]}$ s.t. sparse JL meets $\ell_2$ goal on $x \in S_v$. 
Comparison to previous work

Goal: $P_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta$.

\[ \nu(m, \epsilon, \delta, s) := \sup \text{ over } \nu \in [0, 1] \text{ s.t. sparse JL meets } \ell_2 \text{ goal on } x \in S_{\nu}. \]

Bounds on $\nu$:

- $\nu(m, \epsilon, \delta, 1)$ understood (Weinberger et al ’09, …, Freksen et al. ’18)
- $\nu(m, \epsilon, \delta, s)$ lower bound for multiple hashing (Weinberger et al ’09)
Comparison to previous work

Goal: \( P_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta \).

\( \nu(m, \epsilon, \delta, s) := \sup \text{ over } \nu \in [0, 1] \text{ s.t. sparse JL meets } \ell_2 \text{ goal on } x \in S_v. \)

Bounds on \( \nu \):
- \( \nu(m, \epsilon, \delta, 1) \) understood (Weinberger et al '09,..., Freksen et al. '18)
- \( \nu(m, \epsilon, \delta, s) \) lower bound for multiple hashing (Weinberger et al '09)

Bounds for sparse JL on full space \( \mathbb{R}^n \):
- Can set \( m \approx \epsilon^{-2} \log(1/\delta), s \approx \epsilon^{-1} \log(1/\delta) \) (Kane and Nelson '12)
- Can set \( m \approx \min(2\epsilon^{-2}/\delta, \epsilon^{-2} \log(1/\delta)e^{\Theta(\epsilon^{-1} \log(1/\delta)/s)}) \) (Cohen '16)
Comparison to previous work

Goal: \( \mathbb{P}_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta \).

\( v(m, \epsilon, \delta, s) := \sup \text{ over } v \in [0, 1] \text{ s.t. } \text{sparse JL meets } \ell_2 \text{ goal on } x \in S_v. \)

Bounds on \( v \):
- \( v(m, \epsilon, \delta, 1) \) understood (Weinberger et al. ’09,…, Freksen et al. ’18)
- \( v(m, \epsilon, \delta, s) \) lower bound for multiple hashing (Weinberger et al. ’09)

Bounds for sparse JL on full space \( \mathbb{R}^n \):
- Can set \( m \approx \epsilon^{-2} \log(1/\delta), s \approx \epsilon^{-1} \log(1/\delta) \) (Kane and Nelson ’12)
- Can set \( m \approx \min(2\epsilon^{-2}/\delta, \epsilon^{-2} \log(1/\delta) e^{\Theta(\epsilon^{-1} \log(1/\delta)/s)}) \) (Cohen ’16)

This work

**Tight bounds on** \( v(m, \epsilon, \delta, s) \) **for a general** \( s > 1 \) **for sparse JL.**
Comparison to previous work

Goal: \( \mathbb{P}_{f \in \mathcal{F}}[(1 - \epsilon) \|x\|_2 \leq \|f(x)\|_2 \leq (1 + \epsilon) \|x\|_2] > 1 - \delta \).

\[ v(m, \epsilon, \delta, s) := \sup \text{ over } v \in [0, 1] \text{ s.t. sparse JL meets } \ell_2 \text{ goal on } x \in S_v. \]

Bounds on \( v \):
- \( v(m, \epsilon, \delta, 1) \) understood (Weinberger et al ’09, ..., Freksen et al. ’18)
- \( v(m, \epsilon, \delta, s) \) lower bound for multiple hashing (Weinberger et al ’09)

Bounds for sparse JL on full space \( \mathbb{R}^n \):
- Can set \( m \approx \epsilon^{-2} \log(1/\delta), s \approx \epsilon^{-1} \log(1/\delta) \) (Kane and Nelson ’12)
- Can set \( m \approx \min(2\epsilon^{-2}/\delta, \epsilon^{-2} \log(1/\delta) e^{\Theta(\epsilon^{-1} \log(1/\delta)/s)}) \) (Cohen ’16)

This work

**Tight bounds on** \( v(m, \epsilon, \delta, s) \) **for a general** \( s > 1 \) **for sparse JL.**

\[ \Rightarrow \text{ Characterization of sparse JL performance in terms of } \epsilon, \delta, \text{ and } \ell_\infty\text{-to-}\ell_2 \text{ norm ratio for a general } \# \text{ of hash functions } s \]
Main result

Theorem

Under mild conditions, \( v(m, \epsilon, \delta, s) \) is equal to \( f'(m, \epsilon, \ln(1/\delta), s) \), where \( f'(m, \epsilon, p, s) \) is defined to be:

\[
\begin{cases}
1 & \text{if } m \geq \min \left(2\epsilon^{-2}e^{p}, \epsilon^{-2}pe^{\Theta\left(\max\left(1, \frac{pe^{-1}}{s}\right)\right)}\right) \\
\Theta\left(\sqrt{\epsilon s} \frac{\ln\left(\frac{me^{2}}{p}\right)}{\sqrt{p}}\right) & \text{else, if } m \geq \max\left(\Theta(\epsilon^{-2} p), s \cdot e^{\Theta\left(\max\left(1, \frac{pe^{-1}}{s}\right)\right)}\right) \\
\Theta\left(\sqrt{\epsilon s} \min\left(\frac{\ln\left(\frac{me}{p}\right)}{p}, \sqrt{\ln\left(\frac{me^{2}}{p}\right)}\right)\right) & \text{and } m \leq \epsilon^{-2}e^{\Theta(p)} \\
0 & \text{else, if } m \geq \Theta(\epsilon^{-2} p) \\
\end{cases}
\]

and \( m \leq \Theta(\epsilon^{-2} p) \).
Tight analysis of $v(m, \epsilon, \delta, s)$ for uniform sparse JL for a general $s$. Could inform how to optimally set parameters in practice.

Characterization of sparse JL performance in terms of $\epsilon$, $\delta$, and $\ell_\infty$-to-$\ell_2$ norm ratio for a general $\#$ of hash functions $s$.

Evaluation on real-world and synthetic data (sparse JL can perform much better than feature hashing).

Thank you!
PROOF OF MAIN RESULT
Sparse JL as a sparse random projection (KN ’12)

A distribution over $m \times n$ matrices with $s$ nonzero entries per column.

Uniform: Mildly correlate hash functions so $h_j(i) \neq h_k(i)$.

Example (Uniform Sparse JL)
Uniformly choose $s$ nonzero entries in each column; i.i.d signs for nonzero entries.

Block:
Take $h_i: \{1, \ldots, n\} \rightarrow \{(m/s)(i-1) + 1, \ldots, (m/s)i\}$

Example (Block Sparse JL)
Choose one nonzero coordinate per $m/s$-length block per column; i.i.d signs for nonzero entries.

Sparse JL distributions are state-of-the-art sparse random projections.
Sparse JL as a sparse random projection (KN ’12)

\[ A_{s,m,n} \text{ a distribution over } m \times n \text{ matrices w/ } s \text{ nonzero entries per column} \]
Sparse JL as a sparse random projection (KN ’12)

\( \mathcal{A}_{s,m,n} \) a distribution over \( m \times n \) matrices w/ \( s \) nonzero entries per column

**Uniform**: Mildly correlate hash functions so \( h_j(i) \neq h_k(i) \).

**Example (Uniform Sparse JL)**

Uniformly choose \( s \) nonzero entries in each column; i.i.d signs for nonzero entries.

**Block**: Take \( h_i : \{1, \ldots, n\} \rightarrow \{(m/s)(i-1)+1, \ldots, (m/s)(i)\} \)

**Example (Block Sparse JL)**

Choose one nonzero coordinate per \( m/s \)-length block per column; i.i.d signs for nonzero entries.
Sparse JL as a sparse random projection (KN ’12)

\( A_{s,m,n} \) a distribution over \( m \times n \) matrices w/ \( s \) nonzero entries per column

*Uniform*: Mildly correlate hash functions so \( h_j(i) \neq h_k(i) \).

**Example (Uniform Sparse JL)**

Uniformly choose \( s \) nonzero entries in each column; i.i.d signs for nonzero entries.

**Block**: Take \( h_i : \{1, \ldots, n\} \to \{(m/s)(i - 1) + 1, \ldots, (m/s)(i)\} \)

**Example (Block Sparse JL)**

Choose one nonzero coordinate per \( m/s \)-length block per column; i.i.d signs for nonzero entries.

Sparse JL distributions are state-of-the-art sparse random projections.
High-level approach of our analysis

$(r, i)$th coordinate is $\eta_{r,i}\sigma_{r,i}/\sqrt{s}$, where $\eta_{r,i} \in \{0, 1\}$, $\sigma_{r,i} \in \{-1, 1\}$
High-level approach of our analysis

$(r, i)$th coordinate is $\eta_{r,i}\sigma_{r,i}/\sqrt{s}$, where $\eta_{r,i} \in \{0, 1\}$, $\sigma_{r,i} \in \{-1, 1\}$

Analyze moments of “error” rv $\|Ax\|_2^2 - 1$ for $\|x\|_2 = 1$: 
High-level approach of our analysis

$(r, i)$th coordinate is $\eta_{r,i} \sigma_{r,i} / \sqrt{s}$, where $\eta_{r,i} \in \{0, 1\}$, $\sigma_{r,i} \in \{-1, 1\}$

Analyze moments of “error” rv $\|Ax\|^2_2 - 1$ for $\|x\|^2_2 = 1$:

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j$$

This random variable has been repeatedly analyzed in the literature.
High-level approach of our analysis

\((r, i)\)th coordinate is \(\eta_{r,i} \sigma_{r,i}/\sqrt{s}\), where \(\eta_{r,i} \in \{0, 1\}\), \(\sigma_{r,i} \in \{-1, 1\}\)

Analyze moments of “error” rv \(\|Ax\|_2^2 - 1\) for \(\|x\|_2 = 1\):

\[ R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j \]

This random variable has been repeatedly analyzed in the literature.

But... existing bounds are limited to \(s = 1\) (Freksen et al., etc.)
High-level approach of our analysis

$(r, i)$th coordinate is $\eta_{r,i}\sigma_{r,i}/\sqrt{s}$, where $\eta_{r,i} \in \{0, 1\}$, $\sigma_{r,i} \in \{-1, 1\}$

Analyze moments of “error” rv $\|Ax\|_2^2 - 1$ for $\|x\|_2 = 1$:

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^m \eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j$$

This random variable has been repeatedly analyzed in the literature.

But... existing bounds are limited to $s = 1$ (Freksen et al., etc.) or limited to $\nu = 1$ (Kane and Nelson ’12, Cohen et al. ’18, etc.).
High-level approach of our analysis

$(r, i)$th coordinate is $\eta_{r,i} \sigma_{r,i} / \sqrt{s}$, where $\eta_{r,i} \in \{0,1\}$, $\sigma_{r,i} \in \{-1,1\}$

Analyze moments of “error” rv $\|Ax\|_2^2 - 1$ for $\|x\|_2 = 1$:

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j$$

This random variable has been repeatedly analyzed in the literature.

But... existing bounds are limited to $s = 1$ (Freksen et al., etc.) or limited to $v = 1$ (Kane and Nelson ’12, Cohen et al. ’18, etc.).

Need tight bounds on $\mathbb{E}[R(x_1, \ldots, x_n)^p]$ on $S_v$ at every threshold $v$. 
Bounding moments of $R(x_1, \ldots, x_n)$ at every threshold $v$

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j$$

Complexities: $\eta_{r,i}$ are correlated, sum has $\Theta(mn^2)$ terms

Issues with existing approaches:
1. Not clear how to generalize combinatorics of (Kane and Nelson '12, Freksen et al. '18, etc.)
2. Existing non-combinatorial approaches not sufficiently tight (Cohen et al. '18, Cohen '16, etc.)

We use a non-combinatorial approach with Rademacher-specific bounds.
Bounding moments of $R(x_1, \ldots, x_n)$ at every threshold $v$

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \alpha_{r,j} x_i x_j$$

Complexities: $\eta_{r,i}$ are correlated,
Bounding moments of $R(x_1, \ldots, x_n)$ at every threshold $v$

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j$$

Complexities: $\eta_{r,i}$ are correlated, sum has $\Theta(mn^2)$ terms
Bounding moments of $R(x_1, \ldots, x_n)$ at every threshold $\nu$

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j$$

Complexities: $\eta_{r,i}$ are correlated, sum has $\Theta(mn^2)$ terms

Issues with existing approaches:
Bounding moments of $R(x_1, \ldots, x_n)$ at every threshold $v$

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j$$

Complexities: $\eta_{r,i}$ are correlated, sum has $\Theta(mn^2)$ terms

Issues with existing approaches:

1. Not clear how to generalize combinatorics of (Kane and Nelson ’12, Freksen et al. ’18, etc.)
Bounding moments of $R(x_1, \ldots, x_n)$ at every threshold $v$

\[
R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j
\]

Complexities: $\eta_{r,i}$ are correlated, sum has $\Theta(mn^2)$ terms

Issues with existing approaches:

1. Not clear how to generalize combinatorics of (Kane and Nelson '12, Freksen et al. '18, etc.)
2. Existing non-combinatorial approaches not sufficiently tight (Cohen et al. '18, Cohen '16, etc.)
Bounding moments of $R(x_1, \ldots, x_n)$ at every threshold $v$

$$R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{1 \leq i \neq j \leq n} \sum_{r=1}^{m} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j$$

Complexities: $\eta_{r,i}$ are correlated, sum has $\Theta(mn^2)$ terms

Issues with existing approaches:

1. Not clear how to generalize combinatorics of (Kane and Nelson ’12, Freksen et al. ’18, etc.)

2. Existing non-combinatorial approaches not sufficiently tight (Cohen et. al ’18, Cohen ’16, etc.)

We use a non-combinatorial approach with Rademacher-specific bounds.
Overview of our approach

\[ R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} Z_r(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} \left( \sum_{1 \leq i \neq j \leq n} \eta_r, i \eta_r, j \sigma_r, i \sigma_r, j x_i x_j \right) . \]
Overview of our approach

\[ R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} Z_r(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} \left( \sum_{1 \leq i \neq j \leq n} \eta_r, i \eta_r, j \sigma_r, i \sigma_r, j x_i x_j \right). \]

Lower bound: \( \mathbb{E}[Z_r(x_1, \ldots, x_n)^q] = \mathbb{E}_\eta[\mathbb{E}_\sigma[Z_r(x_1, \ldots, x_n)^q]] \)
Overview of our approach

\[ R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} Z_r(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} \left( \sum_{1 \leq i \neq j \leq n} \eta_r, i \eta_r, j \sigma_r, i \sigma_r, j x_i x_j \right) . \]

Lower bound: \( E[Z_r(x_1, \ldots, x_n)^q] = E_\eta[E_\sigma[Z_r(x_1, \ldots, x_n)^q]] \)

1. Suffices to pick “worst” vector in each \( S_v \)
2. View \( Z_r(v, \ldots, v, 0, \ldots, 0) \) as a quadratic form of \( \pm 1 \) rvs
   Use known quadratic form moments bounds (Latała ’99)
3. Take expectation over \( \eta_r, i \); carefully combine over \( r \in [m] \)
Overview of our approach

\[ R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} Z_r(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} \left( \sum_{1 \leq i \neq j \leq n} \eta_r, i \eta_r, j \sigma_r, i \sigma_r, j x_i x_j \right). \]

Lower bound: \( \mathbb{E}[Z_r(x_1, \ldots, x_n)^q] = \mathbb{E}_\eta[\mathbb{E}_\sigma[Z_r(x_1, \ldots, x_n)^q]] \)

1. Suffices to pick "worst" vector in each \( S_v \)
2. View \( Z_r(v, \ldots, v, 0, \ldots, 0) \) as a quadratic form of \( \pm 1 \) rvs
   Use known quadratic form moments bounds (Latała ’99)
3. Take expectation over \( \eta_r, i \); carefully combine over \( r \in [m] \)

Upper bound: need to consider \( R(x_1, \ldots, x_n) \) for every \( x \in S_v \)
Overview of our approach

\[ R(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} Z_r(x_1, \ldots, x_n) = \frac{1}{s} \sum_{r=1}^{m} \left( \sum_{1 \leq i \neq j \leq n} \eta_{r,i} \eta_{r,j} \sigma_{r,i} \sigma_{r,j} x_i x_j \right). \]

Lower bound: \( \mathbb{E}[Z_r(x_1, \ldots, x_n)^q] = \mathbb{E}_{\eta}[\mathbb{E}_{\sigma}[Z_r(x_1, \ldots, x_n)^q]] \)

1. Suffices to pick “worst” vector in each \( S_v \)
2. View \( Z_r(v, \ldots, v, 0, \ldots, 0) \) as a quadratic form of \( \pm 1 \) rvs
   Use known quadratic form moments bounds (Latała ’99)
3. Take expectation over \( \eta_{r,i} \); carefully combine over \( r \in [m] \)

Upper bound: need to consider \( R(x_1, \ldots, x_n) \) for every \( x \in S_v \)

1. Create **tractable** versions of estimates in (Latała ’97, ’99)
   Structure of \( Z_r(x_1, \ldots, x_n) \) is helpful
2. Combine over \( r \in [m] \) using (Latała ’97)