Understanding Sparse JL for Feature Hashing

Meena Jagadeesan

Harvard University, Class of 2020

NeurIPS 2019

Dimensionality reduction (ℓ_2-to-ℓ_2)

A (randomized) map $\mathbb{R}^n \to \mathbb{R}^m$ (where $m \ll n$) that "preserves geometry".

Dimensionality reduction (ℓ_2-to-ℓ_2)

A (randomized) map $\mathbb{R}^n \to \mathbb{R}^m$ (where $m \ll n$) that "preserves geometry".

A pre-processing step in many applications:

- Document classification tasks (Weinberger et al. '09, etc)
- 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, etc.)

Dimensionality reduction (ℓ_2-to-ℓ_2)

A (randomized) map $\mathbb{R}^n \to \mathbb{R}^m$ (where $m \ll n$) that "preserves geometry".

A pre-processing step in many applications:

- Document classification tasks (Weinberger et al. '09, etc)
- 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, etc.)

Key question: What is the tradeoff between the dimension m, the projection time, and the performance in geometry preservation?

Dimensionality reduction (ℓ_2-to-ℓ_2)

A (randomized) map $\mathbb{R}^n \to \mathbb{R}^m$ (where $m \ll n$) that "preserves geometry".

A pre-processing step in many applications:

- Document classification tasks (Weinberger et al. '09, etc)
- 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, etc.)

Key question: What is the tradeoff between the dimension m, the projection time, and the performance in geometry preservation?

This paper: A theoretical analysis of this tradeoff for 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)

One standard dimensionality reduction scheme is feature hashing.

Feature hashing (Weinberger et al. '09)

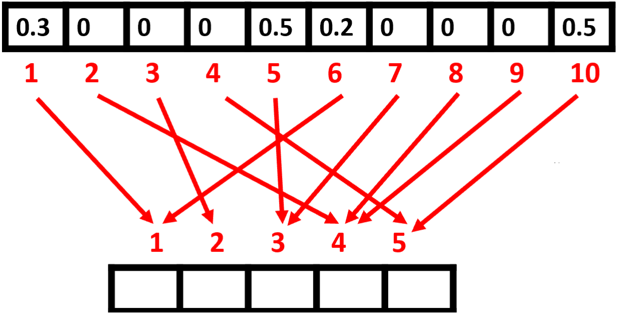
One standard dimensionality reduction scheme is feature hashing.

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

Feature hashing (Weinberger et al. '09)

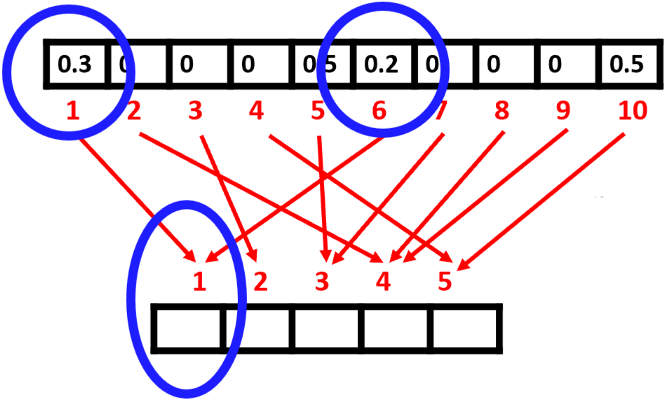
One standard dimensionality reduction scheme is feature hashing.

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



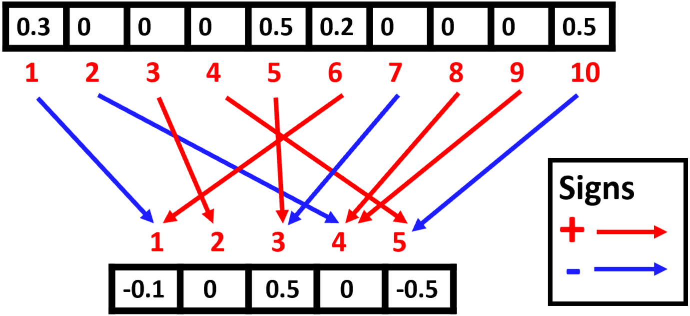
Feature hashing (Weinberger et al. '09)

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



Feature hashing (Weinberger et al. '09)

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



Use **random signs** to handle collisions (unbiased estimator of ℓ_2^2 norm):

$$f(x)_i = \sum_{j \in h^{-1}(i)} \sigma_j x_j.$$

Sparse Johnson-Lindenstrauss transform (KN '12)

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

Sparse Johnson-Lindenstrauss transform (KN '12)

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

Use many (anti-correlated) hash fns $h_1, \ldots, h_s : \{1, \ldots, n\} \to \{1, \ldots, m\}$.
$\implies$ Each input coordinate is mapped to s output coordinates.

Sparse Johnson-Lindenstrauss transform (KN '12)

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

Use many (anti-correlated) hash fns $h_1, \ldots, h_s : \{1, \ldots, n\} \to \{1, \ldots, m\}$.
$\implies$ Each input coordinate is mapped to s output coordinates.

Use random signs to deal with collisions; scale the resulting vector by $\frac{1}{\sqrt{s}}$.

That is: $f(x)_i = \frac{1}{\sqrt{s}} \sum_{k=1}^{s} \left(\sum_{j \in h_k^{-1}(i)} \sigma_j^k x_j \right)$.

Sparse Johnson-Lindenstrauss transform (KN '12)

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

Use many (anti-correlated) hash fns $h_1, \ldots, h_s : \{1, \ldots, n\} \to \{1, \ldots, m\}$.
$\implies$ Each input coordinate is mapped to s output coordinates.

Use random signs to deal with collisions; scale the resulting vector by $\frac{1}{\sqrt{s}}$.

That is: $f(x)_i = \frac{1}{\sqrt{s}} \sum_{k=1}^{s} \left(\sum_{j \in h_k^{-1}(i)} \sigma_j^k x_j \right)$.

(Alternate view: a random sparse matrix w/ s nonzero entries per column.)

Sparse Johnson-Lindenstrauss transform (KN '12)

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

Use many (anti-correlated) hash fns $h_1, \ldots, h_s : \{1, \ldots, n\} \to \{1, \ldots, m\}$.
$\implies$ Each input coordinate is mapped to s output coordinates.

Use random signs to deal with collisions; scale the resulting vector by $\frac{1}{\sqrt{s}}$.

That is: $f(x)_i = \frac{1}{\sqrt{s}} \sum_{k=1}^{s} \left(\sum_{j \in h_k^{-1}(i)} \sigma_j^k x_j \right)$.

(Alternate view: a random sparse matrix w/ s nonzero entries per column.)

Higher s has intuitively better performance in preserving ℓ_2 norm, but projection time is $O(s \|x\|_0)$.

Sparse Johnson-Lindenstrauss transform (KN '12)

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

Use many (anti-correlated) hash fns $h_1, \ldots, h_s : \{1, \ldots, n\} \to \{1, \ldots, m\}$.
$\implies$ Each input coordinate is mapped to s output coordinates.

Use random signs to deal with collisions; scale the resulting vector by $\frac{1}{\sqrt{s}}$.

That is: $f(x)_i = \frac{1}{\sqrt{s}} \sum_{k=1}^{s} \left(\sum_{j \in h_k^{-1}(i)} \sigma_j^k x_j \right)$.

(Alternate view: a random sparse matrix w/ s nonzero entries per column.)

Higher s has intuitively better performance in preserving ℓ_2 norm, but projection time is $O(s \|x\|_0)$.

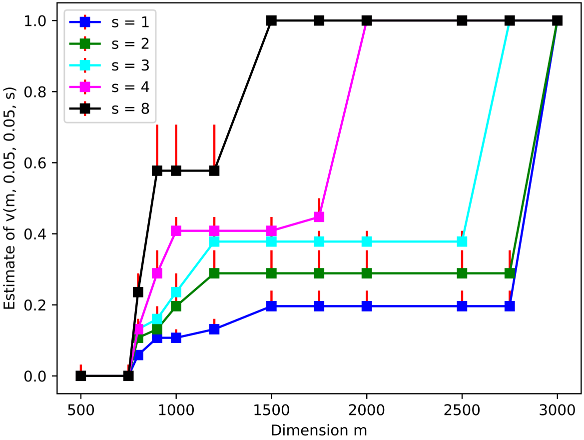
This work

Analysis of tradeoff for sparse JL between # of hash functions s, dimension m, and performance in ℓ_2-norm preservation.

Intuition for this paper

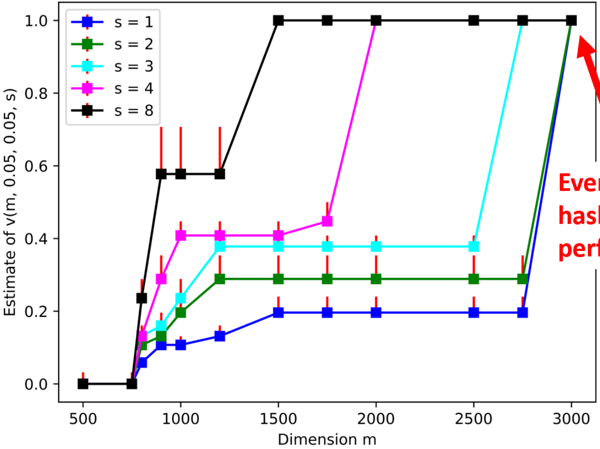
Intuition for this paper

The function v captures the performance of sparse JL on feature vectors.



Intuition for this paper

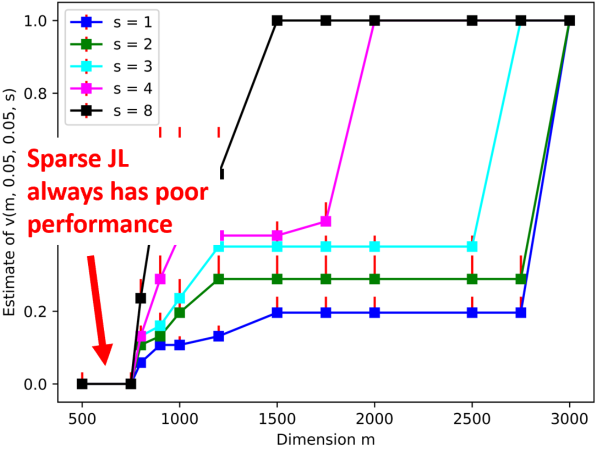
The function v captures the performance of sparse JL on feature vectors.



Even feature hashing has full performance

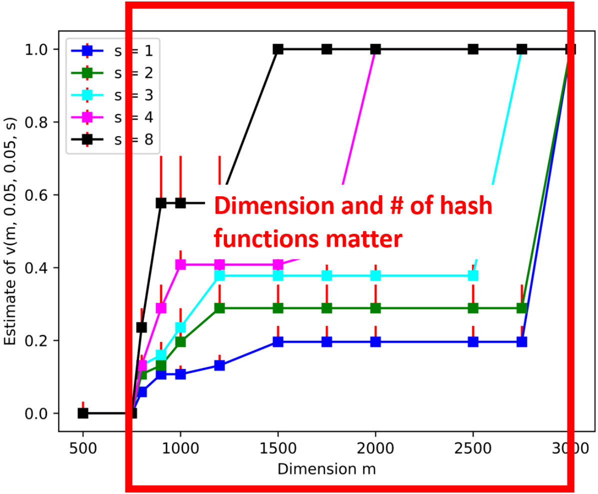
Intuition for this paper

The function v captures the performance of sparse JL on feature vectors.



Intuition for this paper

The function v captures the performance of sparse JL on feature vectors.



Traditional mathematical framework

Consider a probability distribution $\mathcal{F}$ over linear maps $f : \mathbb{R}^n \to \mathbb{R}^m$.

Traditional mathematical framework

Consider a probability distribution $\mathcal{F}$ over linear maps $f : \mathbb{R}^n \to \mathbb{R}^m$.

Geometry-preserving condition. For *each* $x \in \mathbb{R}^n$:

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

for ϵ target error, δ target failure probability.

Traditional mathematical framework

Consider a probability distribution $\mathcal{F}$ over linear maps $f : \mathbb{R}^n \to \mathbb{R}^m$.

Geometry-preserving condition. For *each* $x \in \mathbb{R}^n$:

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

for ϵ target error, δ target failure probability.

(Can apply to *differences* $x = x_1 - x_2$ since f is linear.)

Traditional mathematical framework

Consider a probability distribution $\mathcal{F}$ over linear maps $f : \mathbb{R}^n \to \mathbb{R}^m$.

Geometry-preserving condition. For *each* $x \in \mathbb{R}^n$:

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

for ϵ target error, δ target failure probability.

(Can apply to *differences* $x = x_1 - x_2$ since f is linear.)

Sparse JL can sometimes perform much better in practice on feature vectors than traditional theory on $\mathbb{R}^n$ suggests...

Performance on feature vectors (Weinberger et al. '09)

Consider vectors w/ small ℓ_∞-to-ℓ_2 norm ratio:

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

Performance on feature vectors (Weinberger et al. '09)

Consider vectors w/ small ℓ_∞-to-ℓ_2 norm ratio:

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

Let $\mathcal{F}_{s,m}$ be the distribution given by sparse JL w/ parameters m and s.

Performance on feature vectors (Weinberger et al. '09)

Consider vectors w/ small ℓ_∞-to-ℓ_2 norm ratio:

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

Let $\mathcal{F}_{s,m}$ be the distribution given by sparse JL w/ parameters m and s.

Definition

$v(m, \epsilon, \delta, s)$ is the supremum over $v \in [0, 1]$ such that:

$\mathbb{P}_{f \in \mathcal{F}_{s,m}}[\|f(x)\|_2 \in (1 \pm \epsilon) \|x\|_2] > 1 - \delta$ holds for each $x \in S_v$.

Performance on feature vectors (Weinberger et al. '09)

Consider vectors w/ small ℓ_∞-to-ℓ_2 norm ratio:

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

Let $\mathcal{F}_{s,m}$ be the distribution given by sparse JL w/ parameters m and s.

Definition

$v(m, \epsilon, \delta, s)$ is the supremum over $v \in [0, 1]$ such that:
$\mathbb{P}_{f \in \mathcal{F}_{s,m}}[\|f(x)\|_2 \in (1 \pm \epsilon) \|x\|_2] > 1 - \delta$ holds for each $x \in S_v$.

- $v(m, \epsilon, \delta, s) = 0 \implies$ poor performance
- $v(m, \epsilon, \delta, s) = 1 \implies$ full performance
- $v(m, \epsilon, \delta, s) \in (0, 1) \implies$ good performance on $x \in S_{v(m,\epsilon,\delta,s)}$

Performance on feature vectors (Weinberger et al. '09)

Consider vectors w/ small ℓ_∞-to-ℓ_2 norm ratio:

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

Let $\mathcal{F}_{s,m}$ be the distribution given by sparse JL w/ parameters m and s.

Definition

$v(m, \epsilon, \delta, s)$ is the supremum over $v \in [0, 1]$ such that:
$\mathbb{P}_{f \in \mathcal{F}_{s,m}}[\|f(x)\|_2 \in (1 \pm \epsilon) \|x\|_2] > 1 - \delta$ holds for each $x \in S_v$.

- $v(m, \epsilon, \delta, s) = 0 \implies$ poor performance
- $v(m, \epsilon, \delta, s) = 1 \implies$ full performance
- $v(m, \epsilon, \delta, s) \in (0, 1) \implies$ good performance on $x \in S_{v(m,\epsilon,\delta,s)}$

We give a tight theoretical analysis of the function $v(m, \epsilon, \delta, s)$.

Informal statement of main result

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

$v(m, \epsilon, \delta, s) :=$ sup over $v \in [0, 1]$ s.t. sparse JL meets ℓ_2 goal on $x \in S_v$.

Informal statement of main result

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

$v(m, \epsilon, \delta, s) :=$ sup over $v \in [0,1]$ s.t. sparse JL meets ℓ_2 goal on $x \in S_v$.

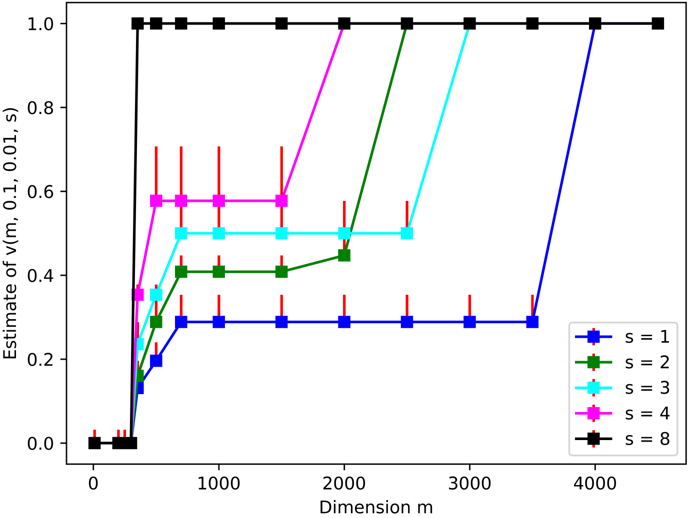
Theorem (Informal)

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

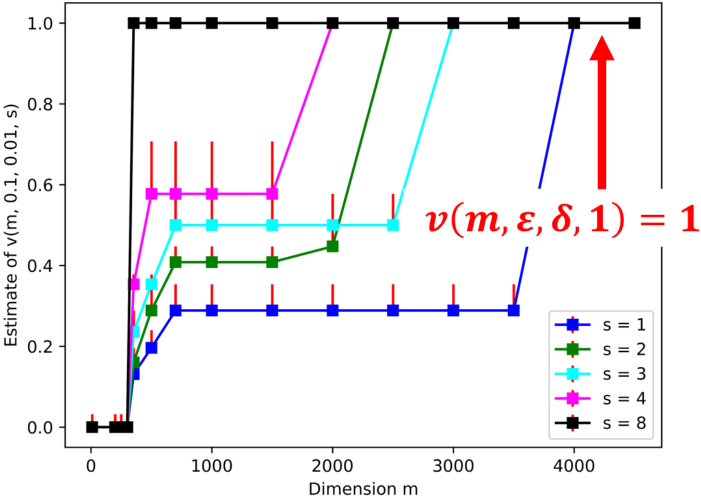
$$\begin{cases} 1 \textit{ (full performance)} & \textit{High } m \\ \sqrt{s} B_1 \textit{ (partial performance)} & \textit{Middle } m \\ \sqrt{s} \min(B_1, B_2) \textit{ (partial performance)} & \textit{Middle } m \\ 0 \textit{ (poor performance)} & \textit{Small } m, \end{cases}$$

where B_1, B_2 are functions of m, ϵ, δ.

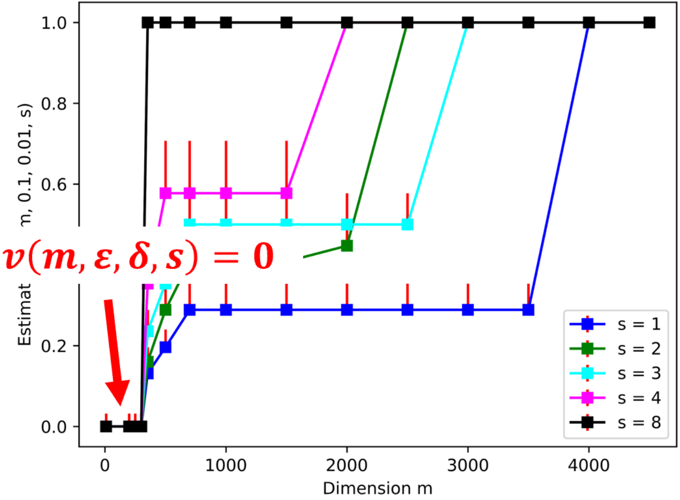
$v(m, \epsilon, \delta, s)$ on more synthetic data



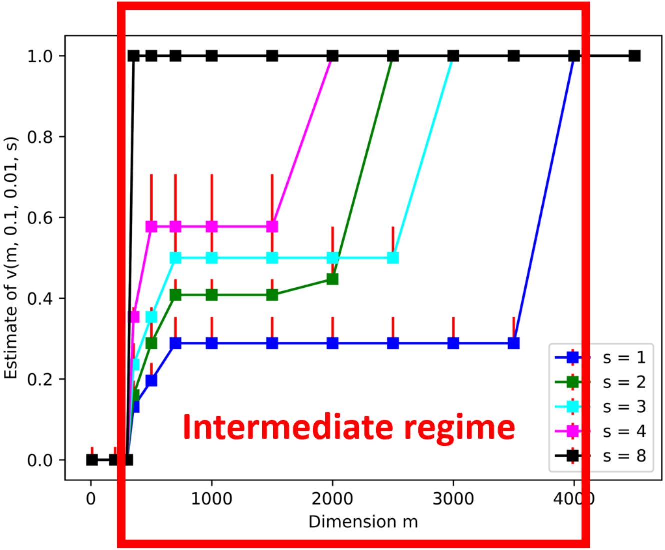
$v(m, \epsilon, \delta, s)$ on more synthetic data



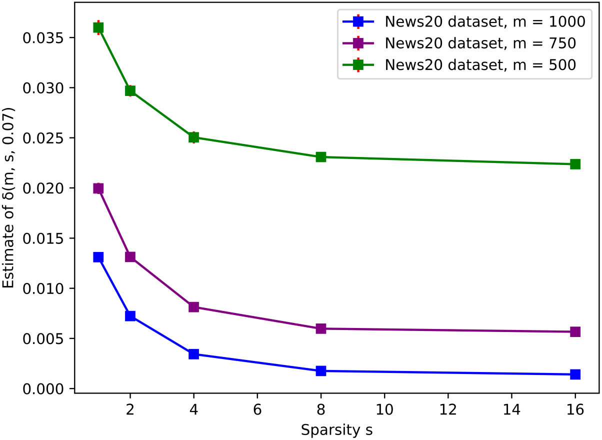
$v(m, \epsilon, \delta, s)$ on more synthetic data



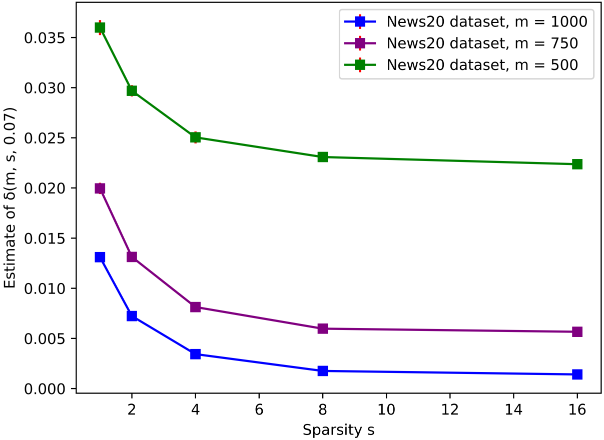
$v(m, \epsilon, \delta, s)$ on more synthetic data



Sparse JL on News20 dataset



Sparse JL on News20 dataset



Sparse JL with ≥ 4 hash functions can perform much better than feature hashing in practice.

Comparison to previous work

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

$v(m, \epsilon, \delta, s) :=$ sup over $v \in [0,1]$ s.t. sparse JL meets ℓ_2 goal on $x \in S_v$.

Comparison to previous work

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

$v(m, \epsilon, \delta, s) :=$ sup over $v \in [0, 1]$ s.t. sparse JL meets ℓ_2 goal on $x \in S_v$.

Bounds on v (Weinberger et al '09,..., Freksen et al. '18):

- $v(m, \epsilon, \delta, \mathbf{1})$ understood
- $v(m, \epsilon, \delta, s)$ bound for *multiple hashing* (a suboptimal construction)

Comparison to previous work

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

$v(m, \epsilon, \delta, s) :=$ sup over $v \in [0,1]$ s.t. sparse JL meets ℓ_2 goal on $x \in S_v$.

Bounds on v (Weinberger et al '09,..., Freksen et al. '18):

- $v(m, \epsilon, \delta, \mathbf{1})$ understood
- $v(m, \epsilon, \delta, s)$ bound for *multiple hashing* (a suboptimal construction)

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

$v(m,\epsilon,\delta,s) :=$ sup over $v \in [0,1]$ s.t. sparse JL meets ℓ_2 goal on $x \in S_v$.

Bounds on v (Weinberger et al '09,..., Freksen et al. '18):

- $v(m,\epsilon,\delta,\mathbf{1})$ understood
- $v(m,\epsilon,\delta,s)$ bound for *multiple hashing* (a suboptimal construction)

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

$v(m, \epsilon, \delta, s) :=$ sup over $v \in [0,1]$ s.t. sparse JL meets ℓ_2 goal on $x \in S_v$.

Bounds on v (Weinberger et al '09,..., Freksen et al. '18):

- $v(m, \epsilon, \delta, \mathbf{1})$ understood
- $v(m, \epsilon, \delta, s)$ bound for *multiple hashing* (a suboptimal construction)

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.*

$\implies$ *Characterization of sparse JL performance in terms of* ϵ, δ, *and* ℓ_∞*-to-*ℓ_2 *norm ratio for a general # of hash functions* s

Conclusion

Tight analysis of $v(m, \epsilon, \delta, s)$ for uniform sparse JL for a general s. Could inform how to optimally set s and m in practice.

Characterization of sparse JL performance in terms of ϵ, δ, and ℓ_∞-to-ℓ_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).

Proof technique involves a new perspective on analyzing JL distributions.

Thank you!

EXTRA MATERIAL: MAIN RESULT AND PROOF

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:

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

Sparse JL as a sparse random projection (KN '12)

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

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

Example (Block Sparse JL)

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

Sparse JL 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 $h_i : \{1, \ldots, n\} \to \{(m/s)(i-1)+1, \ldots, (m/s)(i)\}$

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 $\|x\|_2 = 1$:

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

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 $v = 1$ (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

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

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,

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 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:

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

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

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

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 ± 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 ± 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

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

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 ± 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 <u>tractable</u> 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)