# Sketching for Linear Algebra

Michael Kapralov
EPFL

# Linear regression

**Input:**

- a sequence of $d$-dimensional data points $a_1, \ldots, a_n \in \mathbb{R}^d$
- values $b_j = f(a_j), j = 1, \ldots, n$

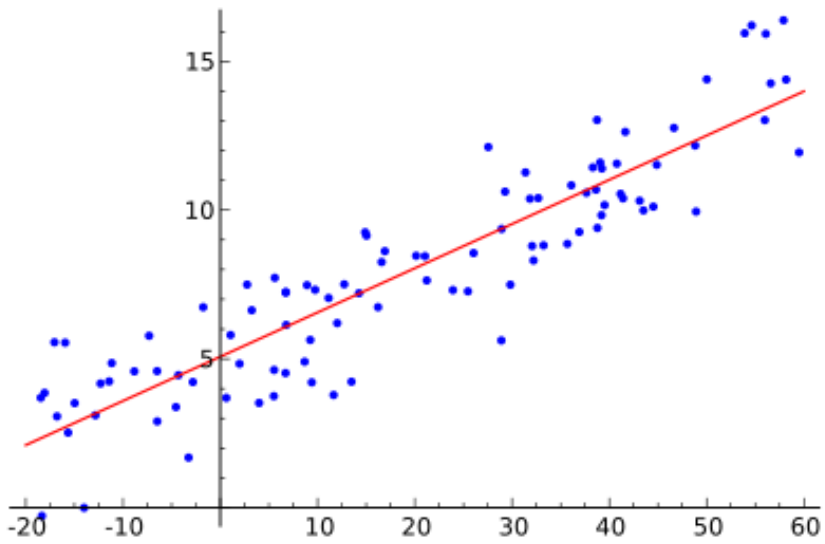
**Output:** linear approximation to $f$

# Linear regression

**Input:**

- a sequence of $d$-dimensional data points $a_1, \ldots, a_n \in \mathbb{R}^d$
- values $b_j = f(a_j), j = 1, \ldots, n$

**Output:** linear approximation to $f$

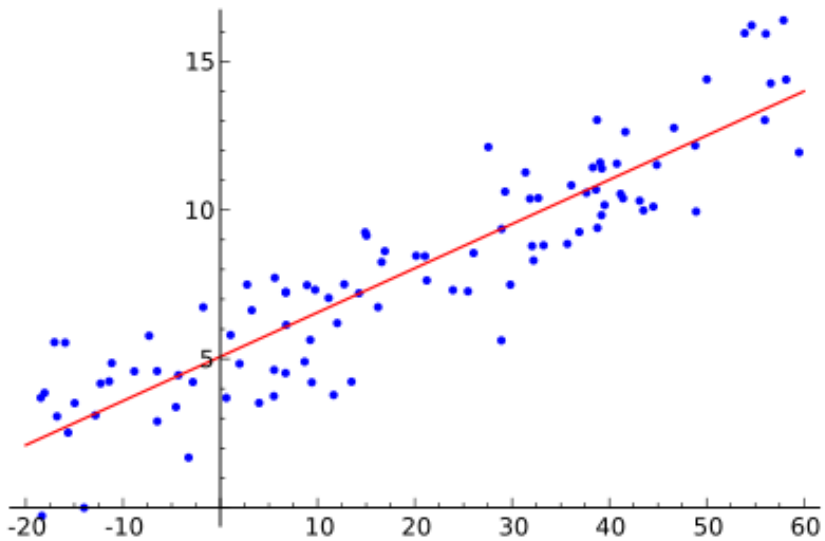


# Linear regression

**Input:**

- a sequence of $d$-dimensional data points $a_1, \ldots, a_n \in \mathbb{R}^d$
- values $b_j = f(a_j), j = 1, \ldots, n$

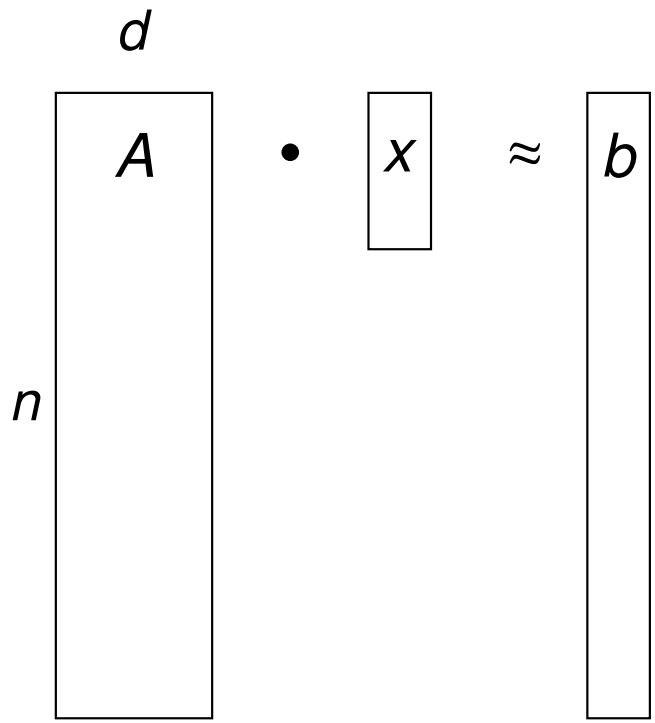
**Output:** linear approximation to $f$




Solve least squares problem:

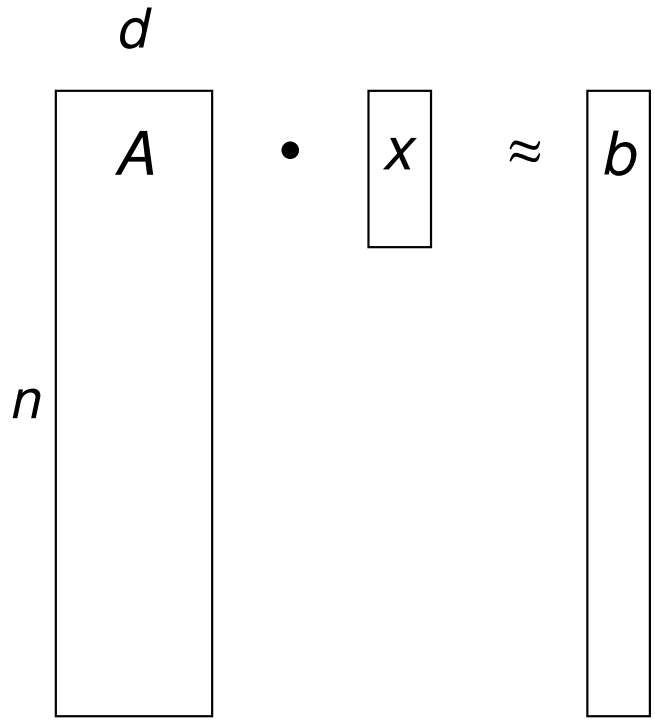
$$\min_{x \in \mathbb{R}^d} \sum_{j=1}^{n} (a_j x - b_j)^2 + \lambda ||x||_2^2$$

Think $n \gg d$, i.e. big data: lots of noisy samples



# Linear regression

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda ||x||_2^2$$



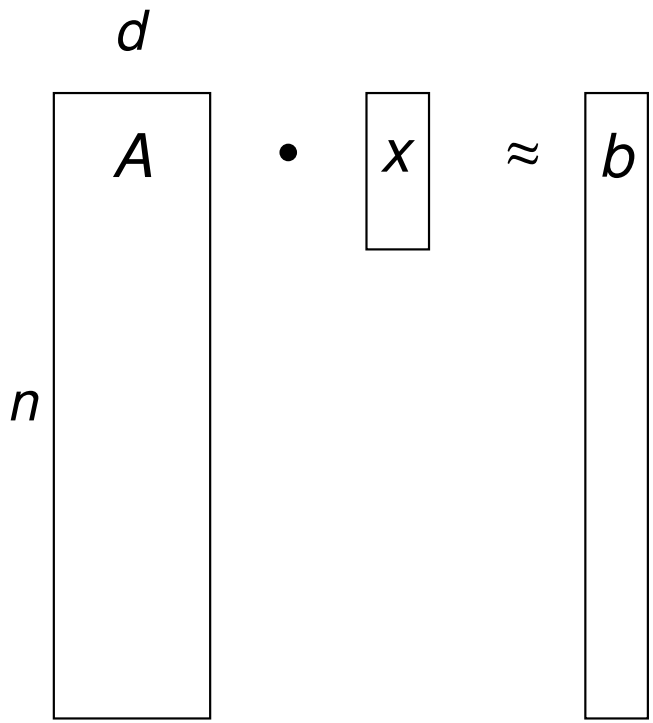

Exact solution:

$$x^* = \left(A^T A + \lambda I\right)^{-1} A^T b$$

Takes $nd^2$ time to solve naively, a bit faster with fast matrix multiplication

# Linear regression

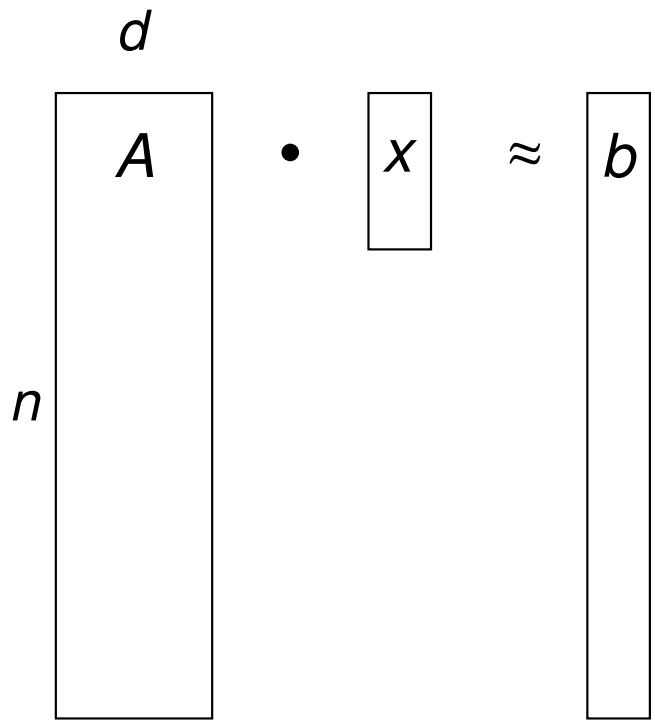
$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda ||x||_2^2$$




Nearly linear time in size of $A$ when $n \gg d$?

# Linear regression

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda ||x||_2^2$$

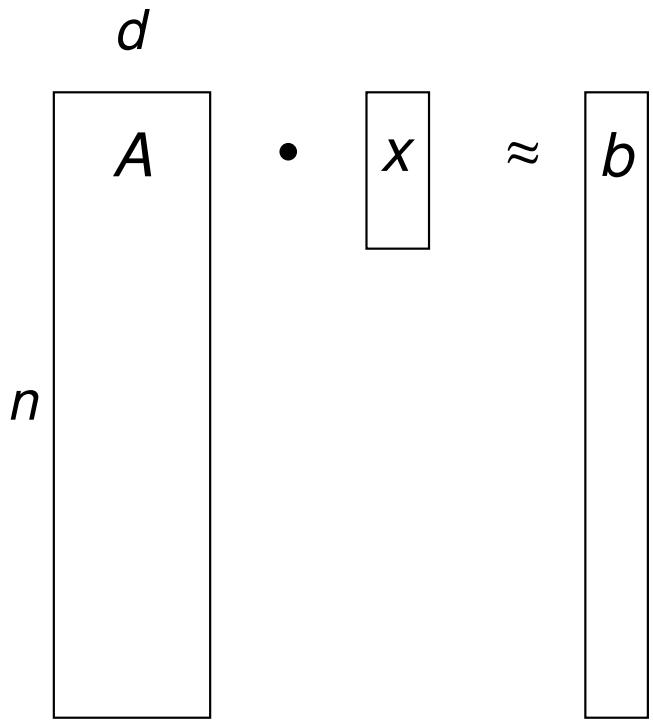



Nearly linear time in size of $A$ when $n \gg d$?

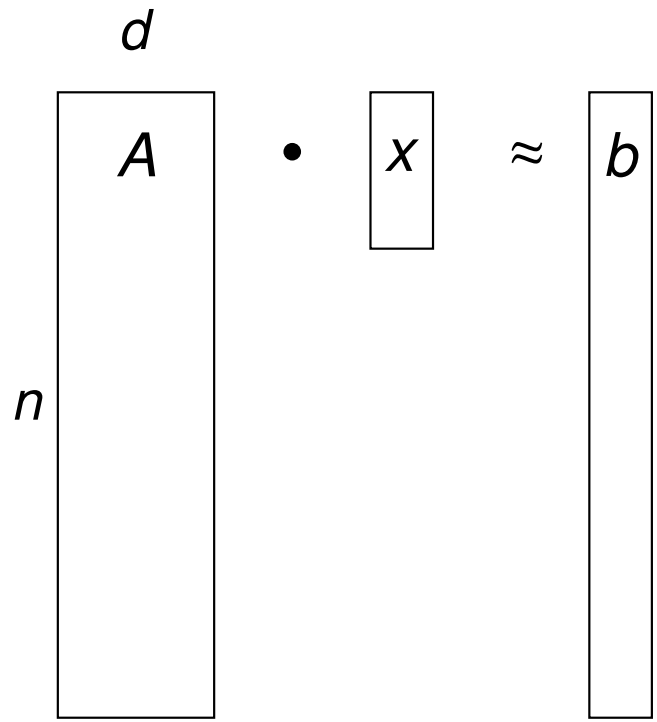
Approximately optimal quality of fit: find $x'$ such that

$$\|Ax' - b\|^2 + \lambda \|x'\|^2 \leq (1+\varepsilon)(\|Ax^* - b\|^2 + \lambda \|x^*\|^2)$$

# Linear regression

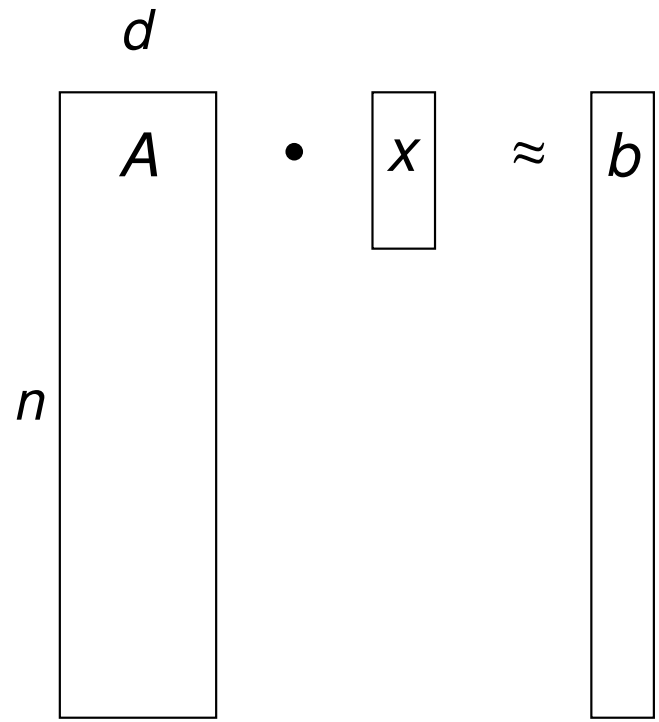


# Linear regression




Reduce # of rows in $A$ down to poly$(d/\varepsilon)$ while preserving solution cost, in time linear in size of $A$

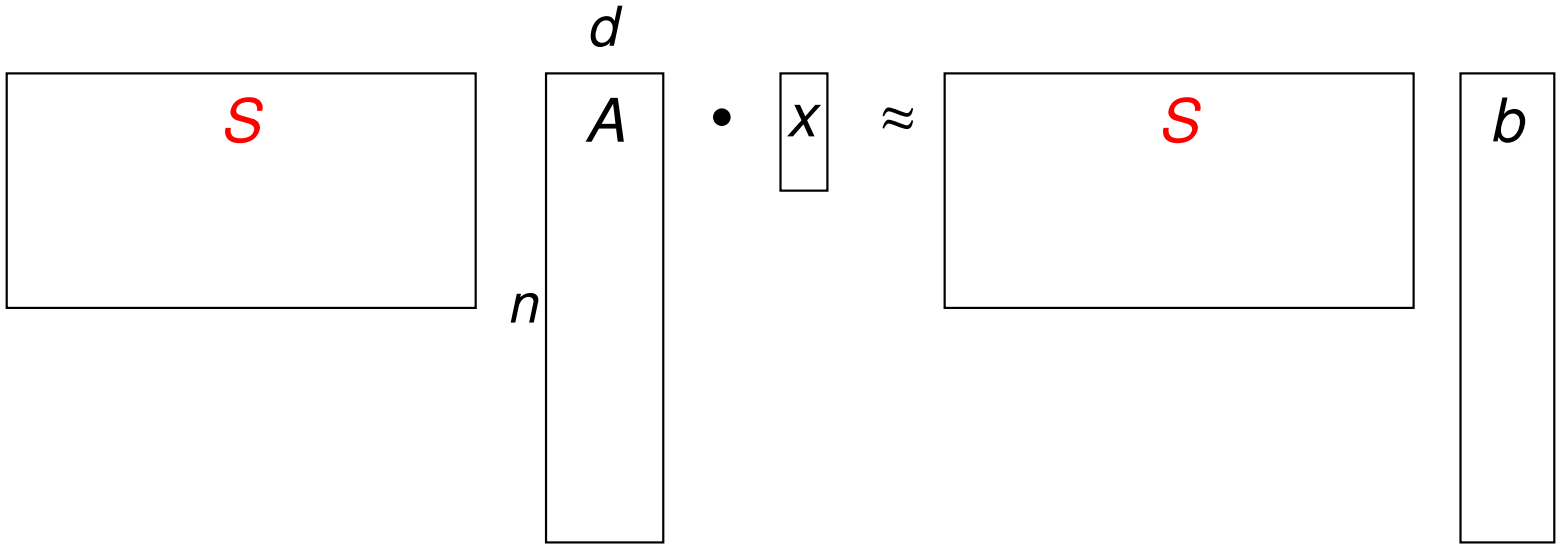
# Linear regression




Reduce # of rows in $A$ down to poly$(d/\varepsilon)$ while preserving solution cost, in time linear in size of $A$

$$\min_{x \in \mathbb{R}^d} \|S \cdot Ax - S \cdot b\|^2 + \lambda ||x||_2^2$$

# Linear regression

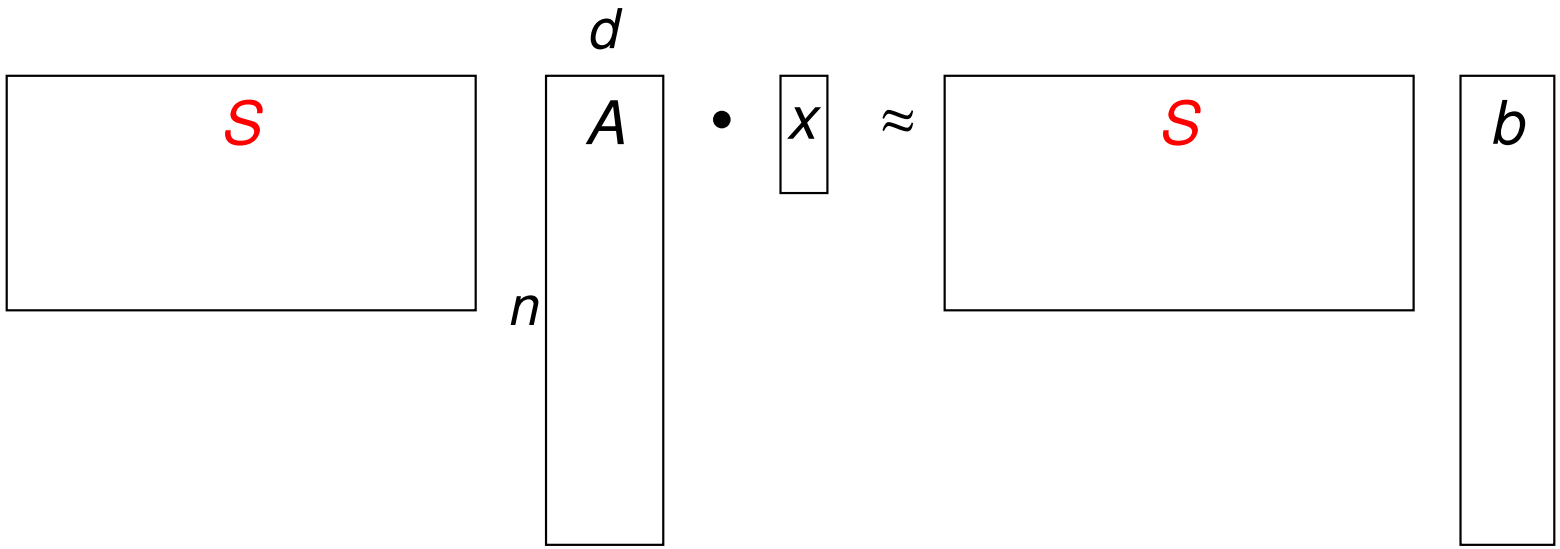



Reduce # of rows in $A$ down to poly$(d/\varepsilon)$ while preserving solution cost.

$$\min_{x \in \mathbb{R}^d} \|S \cdot Ax - S \cdot b\|^2 + \lambda ||x||_2^2$$

Assume $\lambda = 0$ for simplicity now

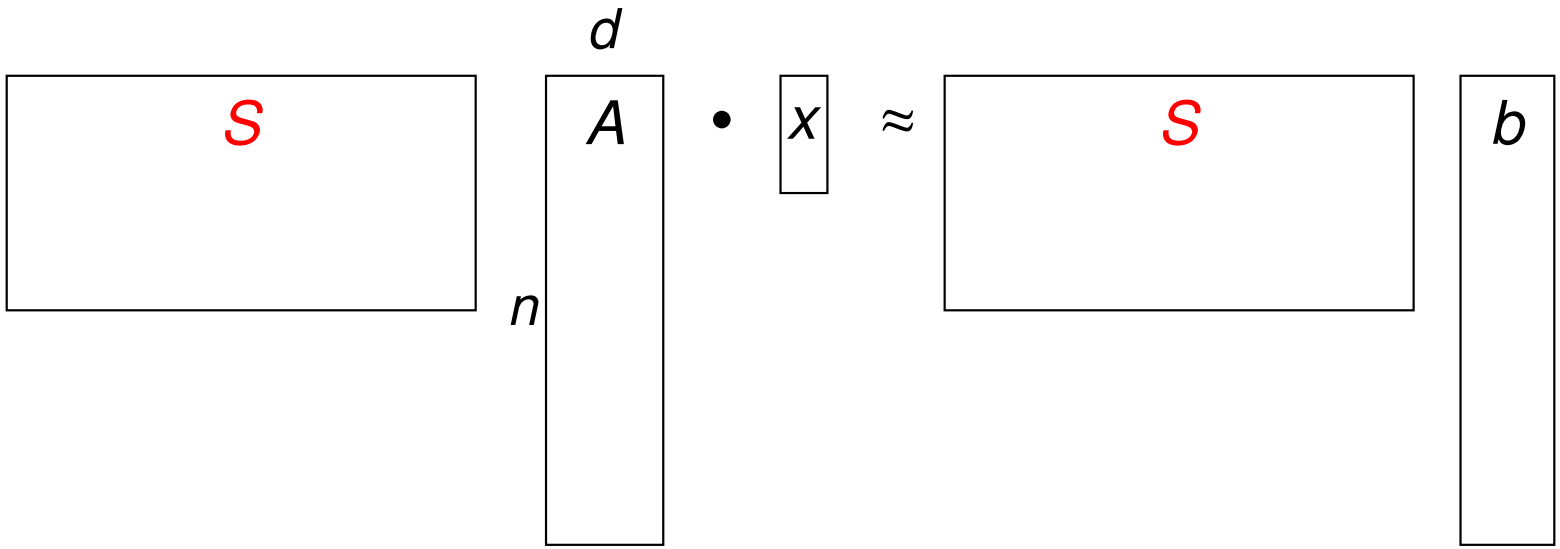
# Linear regression




Reduce # of rows in $A$ down to poly$(d/\varepsilon)$ while preserving solution cost.

$$\min_{x \in \mathbb{R}^d} \|S \cdot Ax - S \cdot b\|^2$$

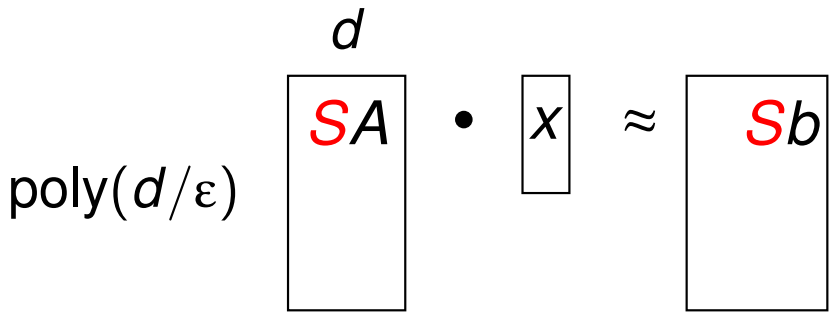
# Linear regression




Reduce # of rows in $A$ down to poly$(d/\varepsilon)$ while preserving solution cost.

$$\min_{x \in \mathbb{R}^d} \|S \cdot Ax - S \cdot b\|^2$$

# Linear regression




Reduce # of rows in $A$ down to $\text{poly}(d/\varepsilon)$ while preserving solution cost.

$$\min_{x \in \mathbb{R}^d} \|S \cdot Ax - S \cdot b\|^2$$

# Subspace embeddings

A random matrix $S$ is a $(d, \varepsilon, \delta)$-subspace embedding if for every subspace $P \subset \mathbb{R}^n$ of dimension $d$ one has

$$\mathbb{P}_S \left[ \left| \|Sy\|^2 - \|y\|^2 \right| \leq \varepsilon \|y\|^2 \text{ for all } y \in P \right] \geq 1 - \delta$$

# Subspace embeddings

A random matrix $S$ is a $(d, \varepsilon, \delta)$-subspace embedding if for every subspace $P \subset \mathbb{R}^n$ of dimension $d$ one has

$$\mathbb{P}_S \left[ \left| \|Sy\|^2 - \|y\|^2 \right| \leq \varepsilon \|y\|^2 \text{ for all } y \in P \right] \geq 1 - \delta$$

Ex.1 Identity map $S = I_n$

# Subspace embeddings

A random matrix $S$ is a $(d, \varepsilon, \delta)$-subspace embedding if for every subspace $P \subset \mathbb{R}^n$ of dimension $d$ one has

$$\mathbb{P}_S \left[ \left| \|Sy\|^2 - \|y\|^2 \right| \leq \varepsilon \|y\|^2 \text{ for all } y \in P \right] \geq 1 - \delta$$

Ex.1 Identity map $S = I_n$

Ex.2 Matrix with $\approx d/\varepsilon^2$ rows independent Gaussians

# Subspace embeddings

A random matrix $S$ is a $(d, \varepsilon, \delta)$-subspace embedding if for every subspace $P \subset \mathbb{R}^n$ of dimension $d$ one has

$$\mathbb{P}_S \left[ \left| \|Sy\|^2 - \|y\|^2 \right| \leq \varepsilon \|y\|^2 \text{ for all } y \in P \right] \geq 1 - \delta$$

Ex.1 Identity map $S = I_n$

Ex.2 Matrix with $\approx d/\varepsilon^2$ rows independent Gaussians

Ex.3 Subsampled randomized Hadamard transform

# Subspace embeddings

A random matrix $S$ is a $(d, \varepsilon, \delta)$-subspace embedding if for every subspace $P \subset \mathbb{R}^n$ of dimension $d$ one has

$$\mathbb{P}_S \left[ \left| \|Sy\|^2 - \|y\|^2 \right| \leq \varepsilon \|y\|^2 \text{ for all } y \in P \right] \geq 1 - \delta$$

Ex.1 Identity map $S = I_n$

Ex.2 Matrix with $\approx d/\varepsilon^2$ rows independent Gaussians

Ex.3 Subsampled randomized Hadamard transform

Ex.4 CountSketch

# Subspace embeddings

A random matrix $S$ is a $(d, \varepsilon, \delta)$-subspace embedding if for every subspace $P \subset \mathbb{R}^n$ of dimension $d$ one has

$$\mathbb{P}_S \left[ \left| \|Sy\|^2 - \|y\|^2 \right| \leq \varepsilon \|y\|^2 \text{ for all } y \in P \right] \geq 1 - \delta$$

Ex.1 Identity map $S = I_n$

Ex.2 Matrix with $\approx d/\varepsilon^2$ rows independent Gaussians

Ex.3 Subsampled randomized Hadamard transform

Ex.4 CountSketch

Ex.5 ...

# Subspace embeddings

A random matrix $S$ is a $(d, \varepsilon, \delta)$-subspace embedding if for every subspace $P \subset \mathbb{R}^n$ of dimension $d$ one has

$$\mathbb{P}_S \left[ \left| \|Sy\|^2 - \|y\|^2 \right| \leq \varepsilon \|y\|^2 \text{ for all } y \in P \right] \geq 1 - \delta$$

Ex.1 Identity map $S = I_n$

Ex.2 Matrix with $\approx d/\varepsilon^2$ rows independent Gaussians

Ex.3 Subsampled randomized Hadamard transform

Ex.4 CountSketch

Ex.5 ...

Why are subspace embeddings useful? Solve the (smaller) sketched problem!

# Subspace embeddings

A random matrix $S$ is a $(d, \varepsilon, \delta)$-subspace embedding if for every subspace $P \subset \mathbb{R}^n$ of dimension $d$ one has

$$\mathbb{P}_S \left[ \left| \|Sy\|^2 - \|y\|^2 \right| \leq \varepsilon \|y\|^2 \text{ for all } y \in P \right] \geq 1 - \delta$$

Ex.1 Identity map $S = I_n$

Ex.2 Matrix with $\approx d/\varepsilon^2$ rows independent Gaussians

Ex.3 Subsampled randomized Hadamard transform

Ex.4 CountSketch

Ex.5 ...

Why are subspace embeddings useful? Solve the (smaller) sketched problem!

$$\min_{x \in \mathbb{R}^d} \|S \cdot Ax - S \cdot b\|^2$$

Prove that $\|S \cdot Ax - S \cdot b\|^2$ is close to $\|Ax - b\|^2$?

Prove that $\|S \cdot Ax - S \cdot b\|^2$ is close to $\|Ax - b\|^2$?

Let $S$ be $(d+1, \varepsilon, 1/100)$-subspace embedding

Prove that $\|S \cdot Ax - S \cdot b\|^2$ is close to $\|Ax - b\|^2$?

Let $S$ be $(d+1, \varepsilon, 1/100)$-subspace embedding

Subspace $P$ is the column span of $A$ and $b$

Prove that $\|S \cdot Ax - S \cdot b\|^2$ is close to $\|Ax - b\|^2$?

Let $S$ be $(d+1, \varepsilon, 1/100)$-subspace embedding

Subspace $P$ is the column span of $A$ and $b$

By subspace embedding property whp over $S$

$$\|SAx - Sb\| \approx_\varepsilon \|Ax - b\| \quad \text{for all } x \in \mathbb{R}^d$$

Prove that $\|S \cdot Ax - S \cdot b\|^2$ is close to $\|Ax - b\|^2$?

Let $S$ be $(d+1, \varepsilon, 1/100)$-subspace embedding

Subspace $P$ is the column span of $A$ and $b$

By subspace embedding property whp over $S$

$$\|SAx - Sb\| \approx_\varepsilon \|Ax - b\| \quad \text{for all } x \in \mathbb{R}^d$$

So

$$\|Ax' - b\| \leq_\varepsilon \|SAx' - Sb\| \leq \|SAx^* - Sb\| \leq_\varepsilon \|Ax^* - b\|$$

## Theorem (Clarkson-Woodruff'13)

*For every* $d \geq 1$, $\varepsilon \in (0,1)$, CountSketch *with* $B$ *buckets is a* $(d, \varepsilon, O(d^2/(\varepsilon^2 B)))$*-subspace embedding.*

## Theorem (Clarkson-Woodruff'13)

*For every $d \geq 1$, $\varepsilon \in (0,1)$,* CountSketch *with $B$ buckets is a $(d, \varepsilon, O(d^2/(\varepsilon^2 B)))$-subspace embedding.*

Sketching matrix $S$ with $B$ rows (buckets) and $n$ columns:

## Theorem (Clarkson-Woodruff'13)

*For every $d \geq 1$, $\varepsilon \in (0,1)$,* CountSketch *with $B$ buckets is a $(d, \varepsilon, O(d^2/(\varepsilon^2 B)))$-subspace embedding.*

Sketching matrix $S$ with $B$ rows (buckets) and $n$ columns:

$$\begin{pmatrix} +1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & +1 & 0 & 0 & -1 \\ 0 & 0 & 0 & +1 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

## Theorem (Clarkson-Woodruff'13)

*For every $d \geq 1$, $\varepsilon \in (0,1)$,* CountSketch *with $B$ buckets is a $(d, \varepsilon, O(d^2/(\varepsilon^2 B)))$-subspace embedding.*

Sketching matrix $S$ with $B$ rows (buckets) and $n$ columns:

$$\begin{pmatrix} +1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & +1 & 0 & 0 & -1 \\ 0 & 0 & 0 & +1 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Reduce # of rows in $A$ down to about $d^2/\varepsilon^2$

## Theorem (Clarkson-Woodruff'13)

*For every $d \geq 1$, $\varepsilon \in (0,1)$,* CountSketch *with $B$ buckets is a $(d, \varepsilon, O(d^2/(\varepsilon^2 B)))$-subspace embedding.*

Sketching matrix $S$ with $B$ rows (buckets) and $n$ columns:

$$\begin{matrix}
+1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 \\
0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & +1 & 0 & 0 & -1 \\
0 & 0 & 0 & +1 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & 0
\end{matrix}$$

Reduce # of rows in $A$ down to about $d^2/\varepsilon^2$

Find $(1+\varepsilon)$-approximate fit in $\text{nnz}(A) + \text{poly}(d/\varepsilon)$ time

## Theorem (Clarkson-Woodruff'13)

*For every $d \geq 1$, $\varepsilon \in (0,1)$,* CountSketch *with $B$ buckets is a $(d, \varepsilon, O(d^2/(\varepsilon^2 B)))$-subspace embedding.*

Sketching matrix $S$ with $B$ rows (buckets) and $n$ columns:

$$\begin{pmatrix} +1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & +1 & 0 & 0 & -1 \\ 0 & 0 & 0 & +1 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Reduce # of rows in $A$ down to about $d^2/\varepsilon^2$

Find $(1+\varepsilon)$-approximate fit in $\text{nnz}(A) + \text{poly}(d/\varepsilon)$ time

Will show proof from Avron, Nguyen, Woodruff'19

Subspace embedding property: for every subspace $P$ of dimension $d$

$$\mathbb{P}_S \left[ \left| \|Sx\|^2 - \|x\|^2 \right| \leq \varepsilon \|x\|^2 \text{ for all } x \in P \right] \geq 1 - \delta$$

Subspace embedding property: for every subspace $P$ of dimension $d$

$$\mathbb{P}_S \left[ \left| \|Sx\|^2 - \|x\|^2 \right| \leq \varepsilon \|x\|^2 \text{ for all } x \in P \right] \geq 1 - \delta$$

$U = A(A^T A)^{-1/2}$ (an orthonormal basis for $P$)

Subspace embedding property: for every subspace $P$ of dimension $d$

$$\mathbb{P}_S\left[\left|\|Sx\|^2 - \|x\|^2\right| \le \varepsilon\|x\|^2 \text{ for all } x \in P\right] \ge 1-\delta$$

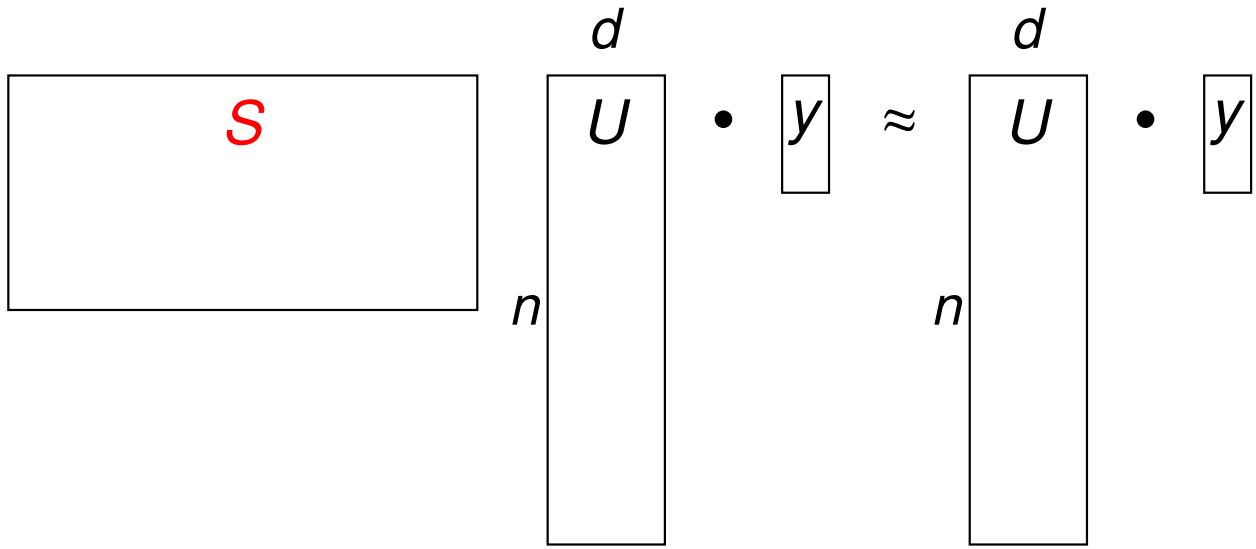
$$U = A(A^TA)^{-1/2} \text{ (an orthonormal basis for } P)$$

Subspace embedding property:

$$\mathbb{P}_S\left[\left|\|SUy\|^2 - \|y\|^2\right| \le \varepsilon\|y\|^2 \text{ for all } y \in \mathbb{R}^d\right] \ge 1-\delta$$

Subspace embedding property:

$$\mathbb{P}_S\left[\left|\|SUy\|^2 - \|y\|^2\right| \le \varepsilon\|y\|^2 \text{ for all } y \in \mathbb{R}^d\right] \ge 1 - \delta$$



Subspace embedding property:

$$\mathbb{P}_S\left[\left|\|SUy\|^2 - \|y\|^2\right| \le \varepsilon\|y\|^2 \text{ for all } y \in \mathbb{R}^d\right] \ge 1 - \delta$$

Subspace embedding property:

$$\mathbb{P}_S\left[\left|\|SUy\|^2 - \|y\|^2\right| \leq \varepsilon\|y\|^2 \text{ for all } y \in \mathbb{R}^d\right] \geq 1 - \delta$$

For all $y \in \mathbb{R}^d$

$$\|SUy\|^2 - \|y\|^2 = y^T U^T S^T SUy - y^T y$$

Subspace embedding property:

$$\mathbb{P}_S\left[\left|\|SUy\|^2 - \|y\|^2\right| \le \varepsilon\|y\|^2 \text{ for all } y \in \mathbb{R}^d\right] \ge 1-\delta$$

For all $y \in \mathbb{R}^d$

$$\begin{aligned}\|SUy\|^2 - \|y\|^2 &= y^T U^T S^T SUy - y^T y \\ &= y^T(U^T S^T SU - I)y\end{aligned}$$

Subspace embedding property:

$$\mathbb{P}_S\left[\left|\|SUy\|^2 - \|y\|^2\right| \le \varepsilon\|y\|^2 \text{ for all } y \in \mathbb{R}^d\right] \ge 1-\delta$$

For all $y \in \mathbb{R}^d$

$$\begin{aligned}\|SUy\|^2 - \|y\|^2 &= y^T U^T S^T SUy - y^T y \\ &= y^T(U^T S^T SU - I)y\end{aligned}$$

Subspace embedding property

$$\left|y^T(U^T S^T SU - I)y\right| \le \varepsilon\|y\|^2 \text{ for all } y \in \mathbb{R}^d$$

Subspace embedding property:

$$\mathbb{P}_S\left[\left|\|SUy\|^2 - \|y\|^2\right| \le \varepsilon\|y\|^2 \text{ for all } y \in \mathbb{R}^d\right] \ge 1-\delta$$

For all $y \in \mathbb{R}^d$

$$\begin{aligned}\|SUy\|^2 - \|y\|^2 &= y^T U^T S^T SUy - y^T y \\ &= y^T(U^T S^T SU - I)y\end{aligned}$$

Subspace embedding property

$$\left|y^T(U^T S^T SU - I)y\right| \le \varepsilon\|y\|^2 \quad \text{for all } y \in \mathbb{R}^d$$

equivalent to

$$\left\|U^T S^T SU - I_d\right\|_2 \le \varepsilon$$

For every $n \times d$ matrix $U$ with orthonormal rows

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

whp as long as $B \gg d^2/\varepsilon^2$?

For every $n \times d$ matrix $U$ with orthonormal rows

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

whp as long as $B \gg d^2/\varepsilon^2$?

Will show the stronger bound:

$$\left\| U^T S^T S U - I_d \right\|_F \leq \varepsilon$$

whp.

For every $n \times d$ matrix $U$ with orthonormal rows

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

whp as long as $B \gg d^2/\varepsilon^2$?

Will show the stronger bound:

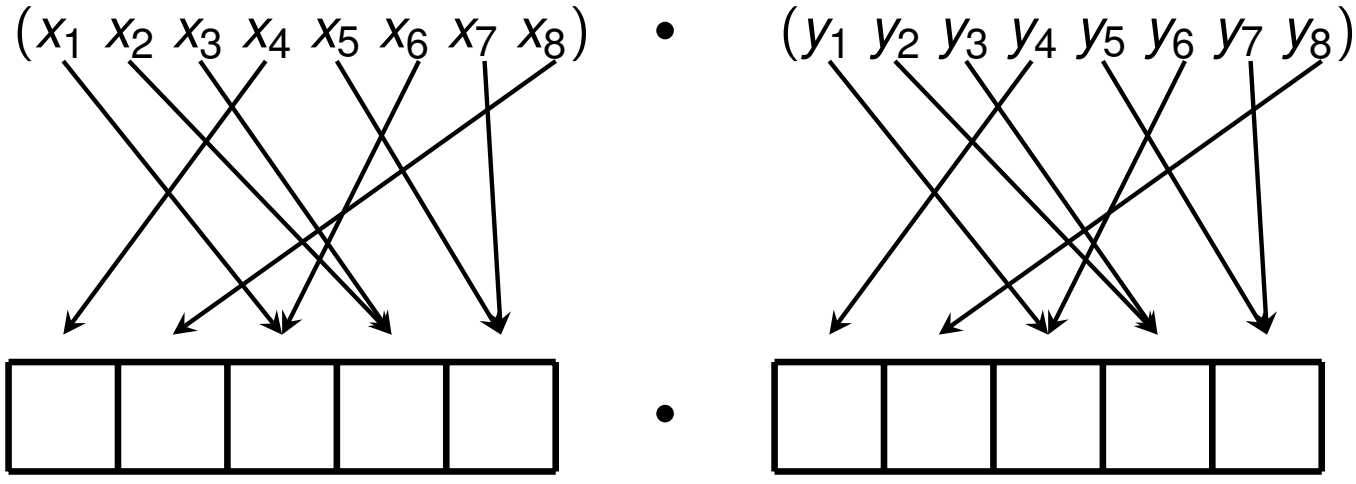
$$\left\| U^T S^T S U - I_d \right\|_F^2 \leq \varepsilon^2$$

whp.

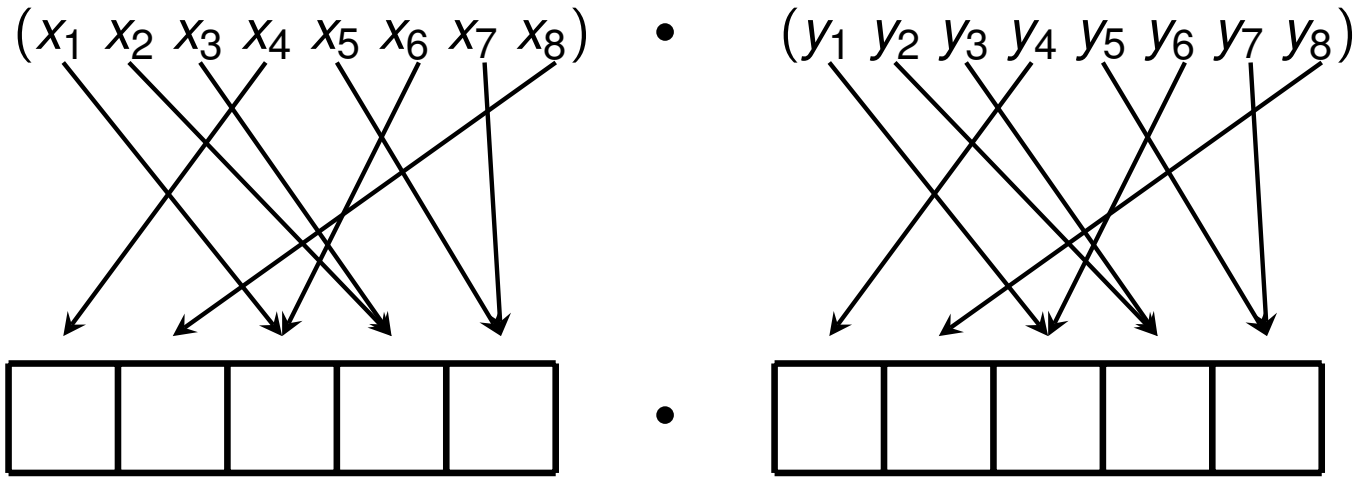
# Hashing almost preserves dot products?

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

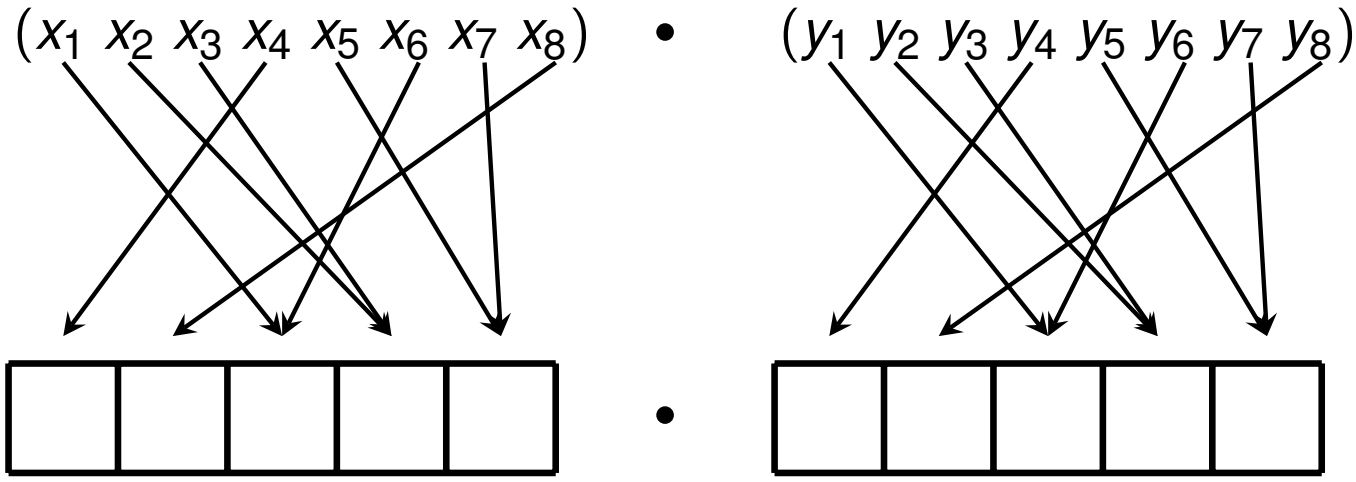
$$(Sx)^T Sy \approx x^T y?$$



# Dot products vs dot products over buckets



# Dot products vs dot products over buckets




For $x, y \in \mathbb{R}^8$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

Dot product:

$$x^T y = x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4 + x_5 y_5 + x_6 y_6 + x_7 y_7 + x_8 y_8$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

Dot product:

$$x^T y = x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4 + x_5 y_5 + x_6 y_6 + x_7 y_7 + x_8 y_8$$

Dot product of hashes:

$$\begin{aligned} (Sx)^T (Sy) = {} & {s_4}^2 x_4 y_4 \\ & + {s_8}^2 x_8 y_8 \\ & + (s_1 x_1 + s_6 x_6)(s_1 y_1 + s_6 y_6) \\ & + (s_2 x_2 + s_3 x_3)(s_2 y_2 + s_3 y_3) \\ & + (s_5 x_5 + s_7 x_7)(s_5 y_5 + s_7 y_7) \end{aligned}$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

Dot product of hashes:

$$\begin{aligned} (Sx)^T(Sy) = {} & {s_4}^2 x_4 y_4 \\ & + {s_8}^2 x_8 y_8 \\ & + (s_1 x_1 + s_6 x_6)(s_1 y_1 + s_6 y_6) \\ & + (s_2 x_2 + s_3 x_3)(s_2 y_2 + s_3 y_3) \\ & + (s_5 x_5 + s_7 x_7)(s_5 y_5 + s_7 y_7) \end{aligned}$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

Dot product of hashes:

$$\begin{aligned} (Sx)^T(Sy) = {} & {s_4}^2 x_4 y_4 \\ & + {s_8}^2 x_8 y_8 \\ & + (s_1 x_1 + s_6 x_6)(s_1 y_1 + s_6 y_6) \\ & + (s_2 x_2 + s_3 x_3)(s_2 y_2 + s_3 y_3) \\ & + (s_5 x_5 + s_7 x_7)(s_5 y_5 + s_7 y_7) \end{aligned}$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

Dot product of hashes:

$$\begin{aligned} (Sx)^T(Sy) = {} & \mathbf{{s_4}^2 x_4 y_4} \\ & + {s_8}^2 x_8 y_8 \\ & + (s_1 x_1 + s_6 x_6)(s_1 y_1 + s_6 y_6) \\ & + (s_2 x_2 + s_3 x_3)(s_2 y_2 + s_3 y_3) \\ & + (s_5 x_5 + s_7 x_7)(s_5 y_5 + s_7 y_7) \end{aligned}$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

Dot product after hashing:

$$\begin{aligned} (Sx)^T (Sy) = {} & x_4 y_4 \\ & + {s_8}^2 x_8 y_8 \\ & + (s_1 x_1 + s_6 x_6)(s_1 y_1 + s_6 y_6) \\ & + (s_2 x_2 + s_3 x_3)(s_2 y_2 + s_3 y_3) \\ & + (s_5 x_5 + s_7 x_7)(s_5 y_5 + s_7 y_7) \end{aligned}$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

Dot product after hashing:

$$\begin{aligned} (Sx)^T(Sy) = {} & x_4 y_4 \\ & + \mathbf{s_8^2 x_8 y_8} \\ & + (s_1 x_1 + s_6 x_6)(s_1 y_1 + s_6 y_6) \\ & + (s_2 x_2 + s_3 x_3)(s_2 y_2 + s_3 y_3) \\ & + (s_5 x_5 + s_7 x_7)(s_5 y_5 + s_7 y_7) \end{aligned}$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

Dot product after hashing:

$$\begin{aligned} (Sx)^T (Sy) = {} & x_4 y_4 \\ & + x_8 y_8 \\ & + (s_1 x_1 + s_6 x_6)(s_1 y_1 + s_6 y_6) \\ & + (s_2 x_2 + s_3 x_3)(s_2 y_2 + s_3 y_3) \\ & + (s_5 x_5 + s_7 x_7)(s_5 y_5 + s_7 y_7) \end{aligned}$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

Dot product after hashing:

$$\begin{aligned} (Sx)^T (Sy) = {} & x_4 y_4 \\ & + x_8 y_8 \\ & + (\mathbf{s_1 x_1} + \mathbf{s_6 x_6})(\mathbf{s_1 y_1} + \mathbf{s_6 y_6}) \\ & + (s_2 x_2 + s_3 x_3)(s_2 y_2 + s_3 y_3) \\ & + (s_5 x_5 + s_7 x_7)(s_5 y_5 + s_7 y_7) \end{aligned}$$

$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

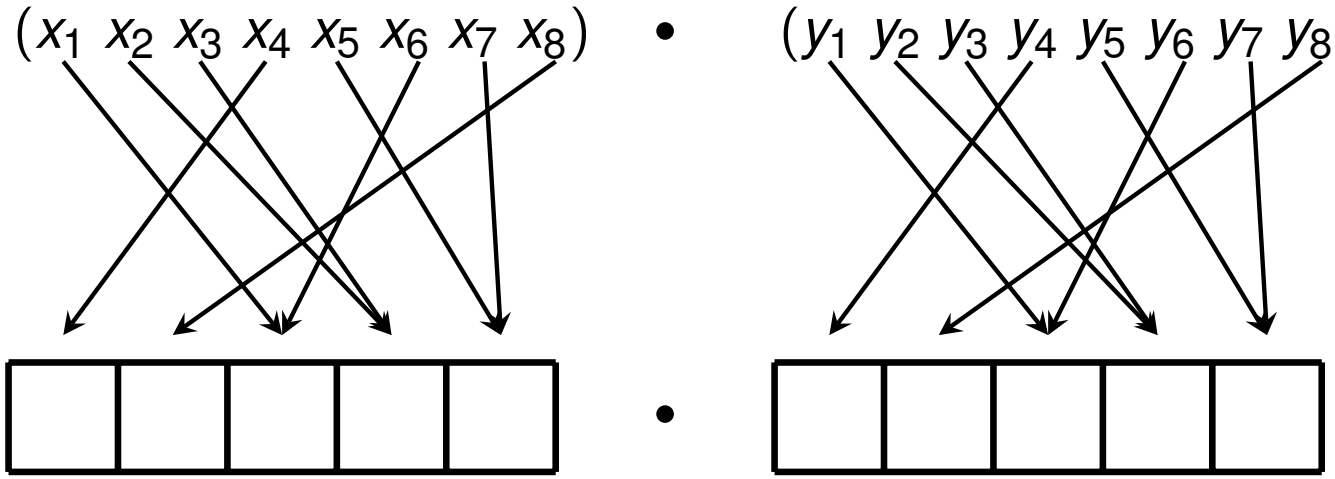
Dot product after hashing:

$$\begin{aligned} (Sx)^T (Sy) = {} & x_4 y_4 \\ & + x_8 y_8 \\ & + x_1 y_1 + x_6 y_6 + s_1 s_6 \cdot (x_1 y_6 + y_1 x_6) \\ & + (s_2 x_2 + s_3 x_3)(s_2 y_2 + s_3 y_3) \\ & + (s_5 x_5 + s_7 x_7)(s_5 y_5 + s_7 y_7) \end{aligned}$$

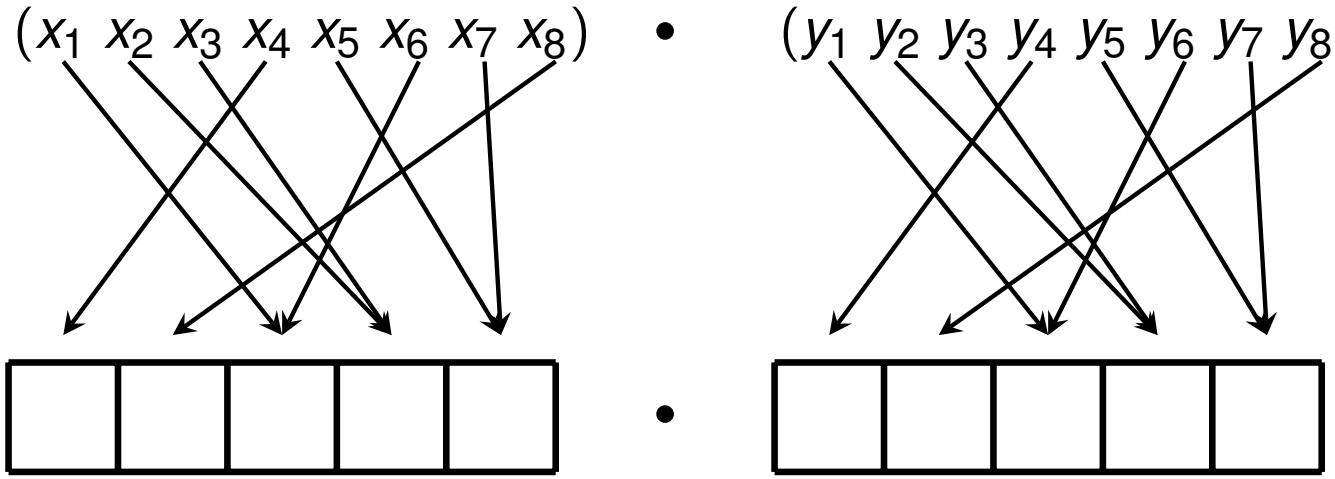
$$Sx = \begin{pmatrix} s_4 x_4 \\ s_8 x_8 \\ s_1 x_1 + s_6 x_6 \\ s_2 x_2 + s_3 x_3 \\ s_5 x_5 + s_7 x_7 \end{pmatrix} \quad \text{and} \quad Sy = \begin{pmatrix} s_4 y_4 \\ s_8 y_8 \\ s_1 y_1 + s_6 y_6 \\ s_2 y_2 + s_3 y_3 \\ s_5 y_5 + s_7 y_7 \end{pmatrix}$$

Dot product after hashing:

$$\begin{aligned} (Sx)^T (Sy) = {} & x_4 y_4 \\ & + x_8 y_8 \\ & + x_1 y_1 + x_6 y_6 + s_1 s_6 \cdot (x_1 y_6 + y_1 x_6) \\ & + x_2 y_2 + x_3 y_3 + s_2 s_3 \cdot (x_2 y_3 + y_3 x_2) \\ & + x_5 y_5 + x_7 y_7 + s_5 s_7 \cdot (x_7 y_5 + x_5 y_7) \end{aligned}$$

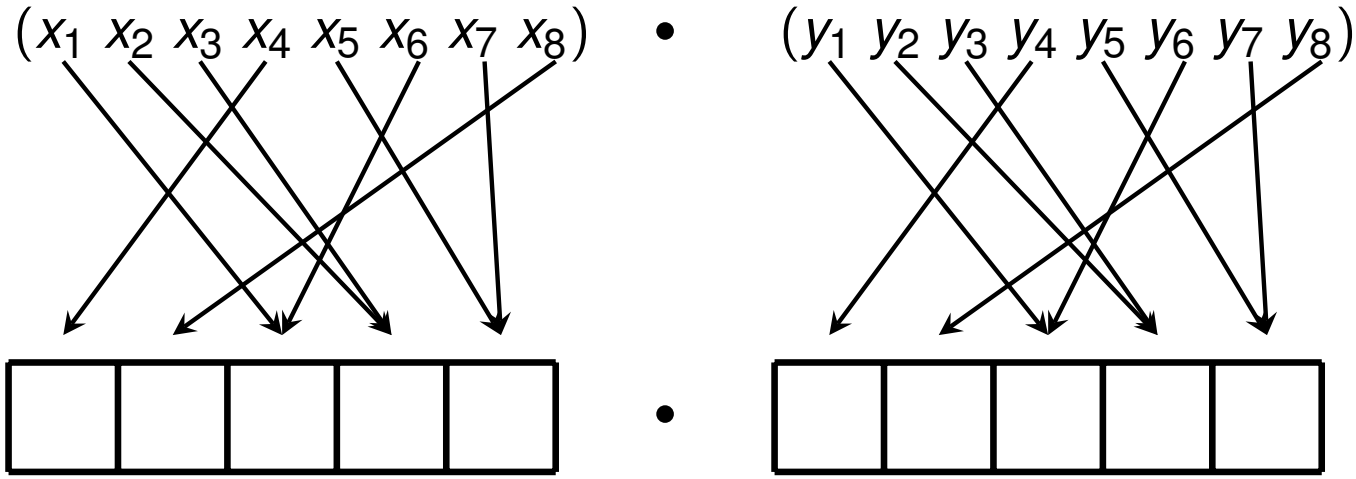



$$(Sx)^T(Sy) = x^T y + \sum_{b\in[B]} \sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j \cdot x_i y_j$$




$$(Sx)^T(Sy) = x^T y + \sum_{b\in[B]} \sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j \cdot x_i y_j$$

An unbiased estimator!

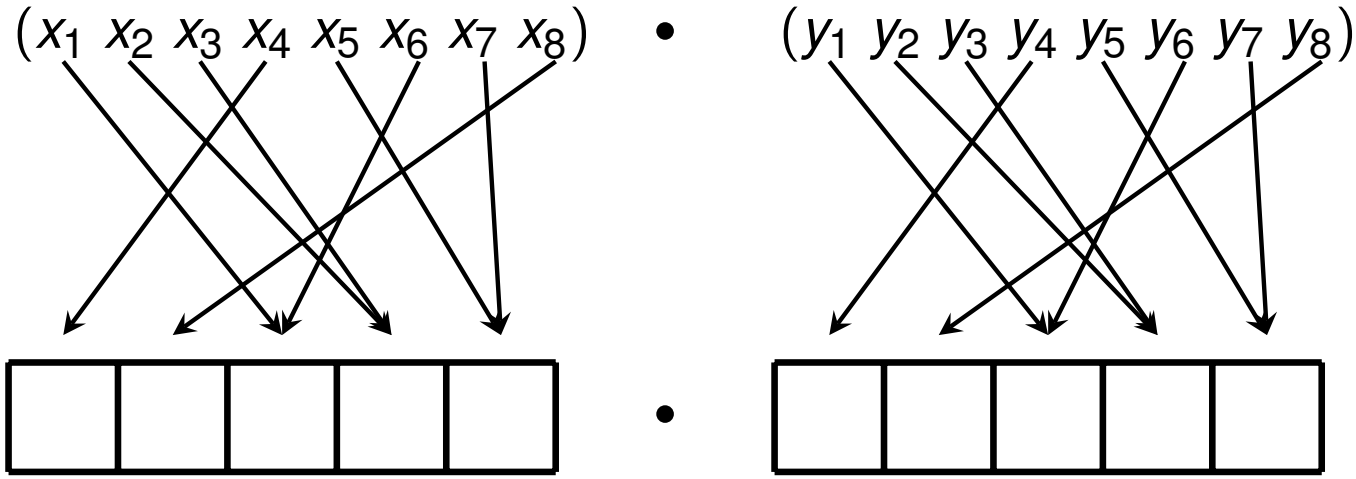



$$(Sx)^T(Sy) = x^T y + \sum_{b\in[B]} \sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j \cdot x_i y_j$$

An unbiased estimator!

$$\mathbb{E}\left[(Sx)^T(Sy)\right] = x^T y + \sum_{b\in[B]} \sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} \mathbb{E}[s_i s_j] \cdot x_i y_j$$

$$= x^T y$$



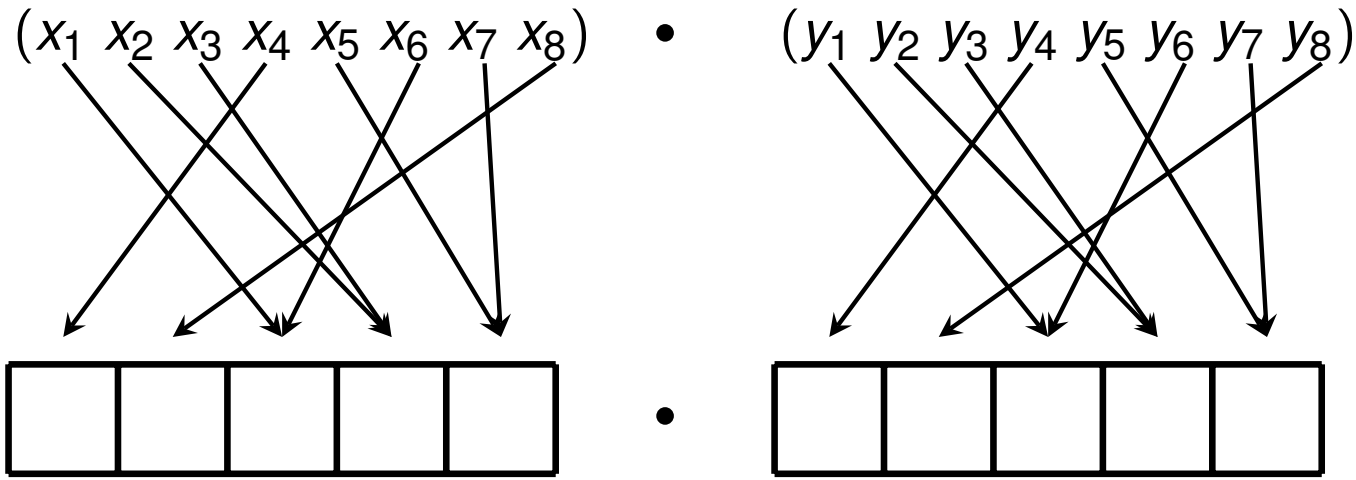

$$(Sx)^T(Sy) = x^T y + \sum_{b\in[B]} \sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j \cdot x_i y_j$$

An unbiased estimator!

$$\mathbb{E}\left[(Sx)^T(Sy)\right] = x^T y + \sum_{b\in[B]} \sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} \mathbb{E}[s_i s_j] \cdot x_i y_j$$

$$= x^T y$$

Variance?



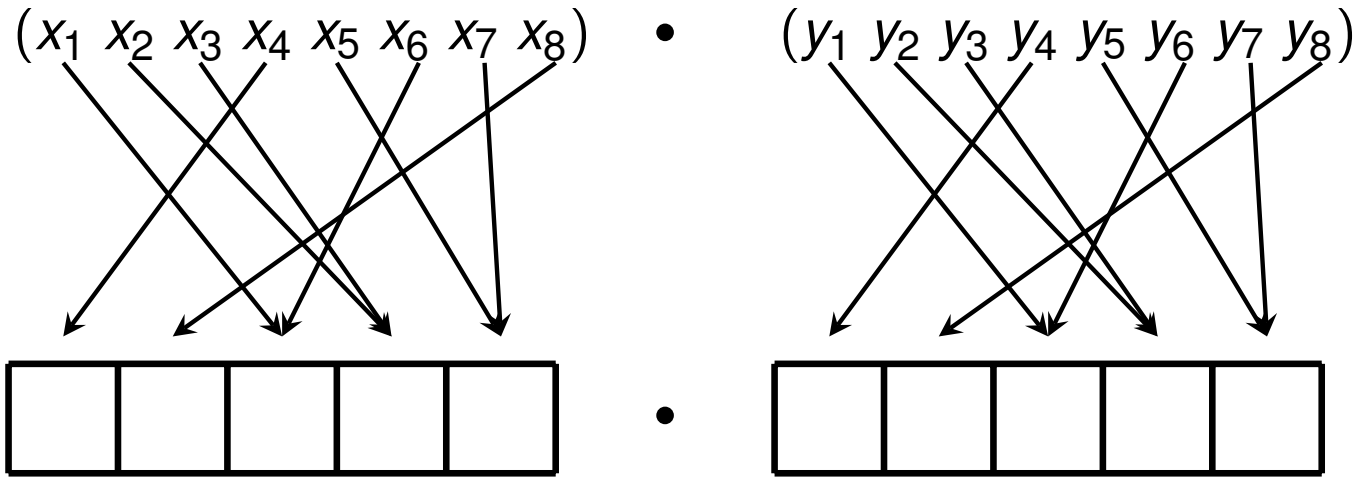

$$(Sx)^T(Sy) = x^T y + \sum_{b\in[B]} \sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j \cdot x_i y_j$$

An unbiased estimator!

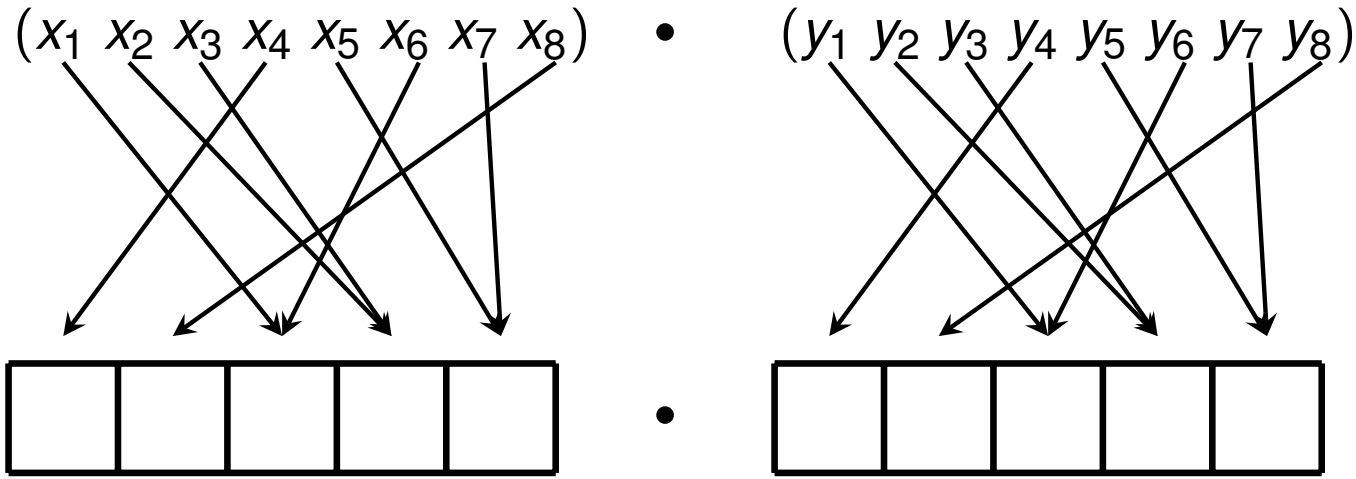
$$\mathbb{E}\left[(Sx)^T(Sy)\right] = x^T y + \sum_{b\in[B]} \sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} \mathbb{E}[s_i s_j] \cdot x_i y_j$$

$$= x^T y$$

Variance?



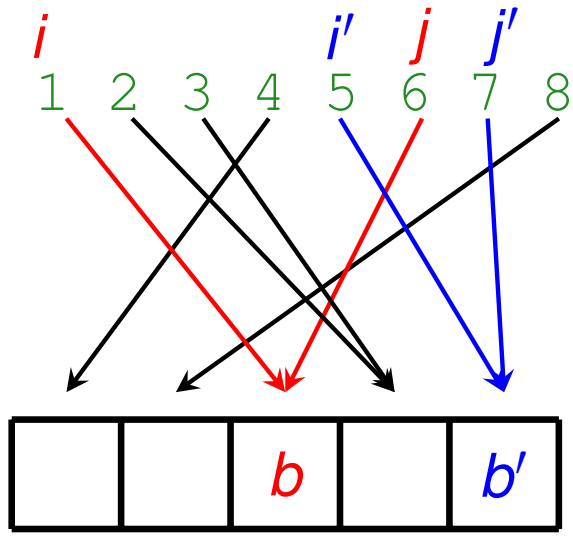

$$\mathbb{E}\left[\left((Sx)^T(Sy)-x^Ty\right)^2\right]=\mathbb{E}\left[\left(\sum_{b\in[B]}\sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_is_j\cdot x_iy_j\right)^2\right]$$



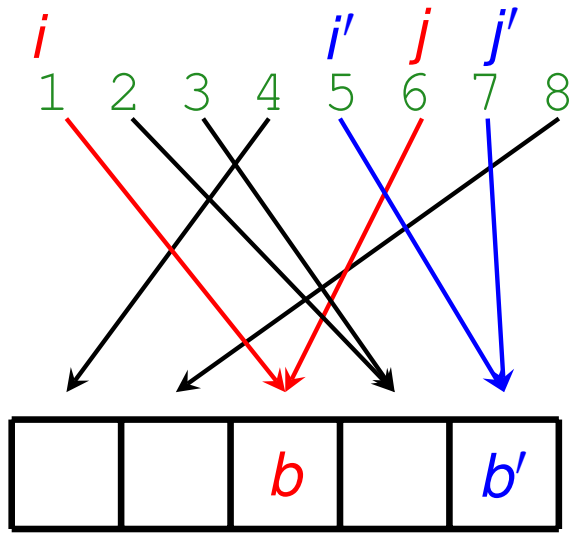

$$\left(\sum_{b\in[B]}\sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j x_i y_j\right)^2 = \sum_{b,b'\in[B]}\sum_{\substack{i,i',j,j'\in[n],i\neq j,i'\neq j',\\ h(i)=h(j)=b\\ h(i')=h(j')=b'}} s_i s_j s_{i'} s_{j'} \cdot x_i y_j x_{i'} y_{j'}$$

$$\left( \sum_{b \in [B]} \sum_{\substack{i,j \in [n], i \neq j, \\ h(i)=h(j)=b}} s_i s_j x_i y_j \right)^2 = \sum_{b,b' \in [B]} \sum_{\substack{i,i',j,j' \in [n], i \neq j, i' \neq j', \\ h(i)=h(j)=b \\ h(i')=h(j')=b'}} s_i s_j s_{i'} s_{j'} \cdot x_i y_j x_{i'} y_{j'}$$

$$\left( \sum_{b \in [B]} \sum_{\substack{i,j \in [n], i \neq j, \\ h(i)=h(j)=b}} s_i s_j x_i y_j \right)^2 = \sum_{b,b' \in [B]} \sum_{\substack{i,i',j,j' \in [n], i \neq j, i' \neq j', \\ h(i)=h(j)=b \\ h(i')=h(j')=b'}} s_i s_j s_{i'} s_{j'} \cdot x_i y_j x_{i'} y_{j'}$$



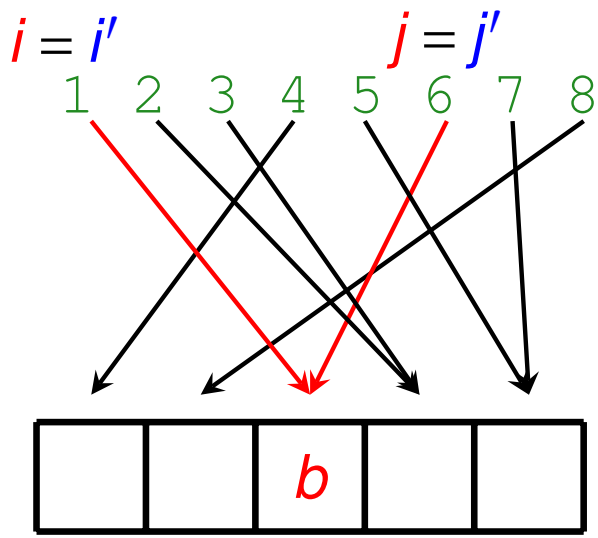
$$\left(\sum_{b\in[B]}\sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j x_i y_j\right)^2 = \sum_{b,b'\in[B]}\sum_{\substack{i,i',j,j'\in[n],i\neq j,i'\neq j',\\ h(i)=h(j)=b\\ h(i')=h(j')=b'}} s_i s_j s_{i'} s_{j'} \cdot x_i y_j x_{i'} y_{j'}$$



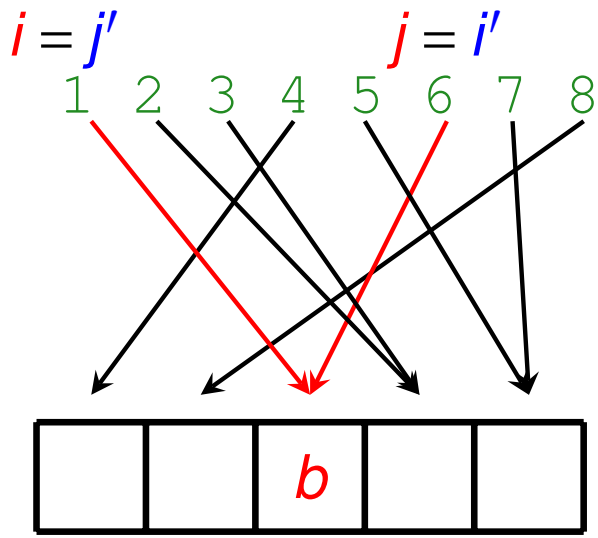

If $b \neq b'$, the $|\{i,j,i',j'\}| = 4$, so $\mathbb{E}\left[s_i s_j s_{i'} s_{j'}\right] = 0$

$$\mathbb{E}\left[\left(\sum_{b\in[B]}\sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j x_i y_j\right)^2\right] = \mathbb{E}\left[\sum_{b\in[B]}\sum_{\substack{i,i',j,j'\in[n],i\neq j,i'\neq j',\\ h(i)=h(j)=b,\\ h(i')=h(j')=b}} s_i s_j s_{i'} s_{j'} x_i y_j x_{i'} y_{j'}\right]$$

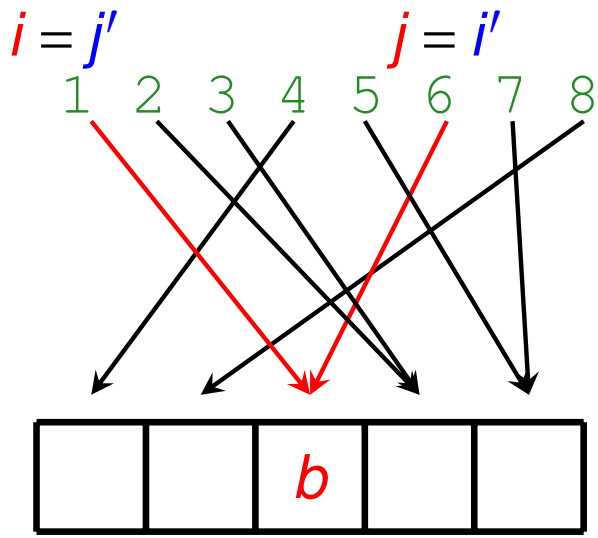
$$\mathbb{E}\left[\left(\sum_{b\in[B]}\sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j x_i y_j\right)^2\right] = \mathbb{E}\left[\sum_{b\in[B]}\sum_{\substack{i,i',j,j'\in[n],i\neq j,i'\neq j',\\ h(i)=h(j)=b,\\ h(i')=h(j')=b}} s_i s_j s_{i'} s_{j'} x_i y_j x_{i'} y_{j'}\right]$$



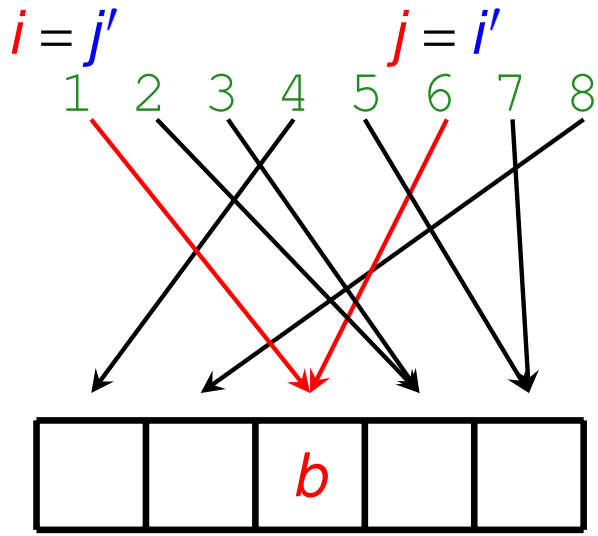
$$\mathbb{E}\left[\left(\sum_{b\in[B]}\sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j x_i y_j\right)^2\right] = \mathbb{E}\left[\sum_{b\in[B]}\sum_{\substack{i,i',j,j'\in[n],i\neq j,i'\neq j',\\ h(i)=h(j)=b,\\ h(i')=h(j')=b}} s_i s_j s_{i'} s_{j'} x_i y_j x_{i'} y_{j'}\right]$$



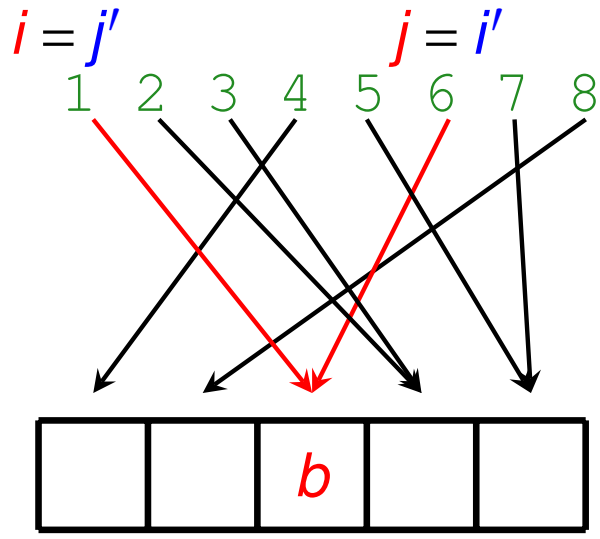
$$\mathbb{E}\left[\left(\sum_{b\in[B]}\sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j x_i y_j\right)^2\right] = \mathbb{E}\left[\sum_{b\in[B]}\sum_{\substack{i,i',j,j'\in[n],i\neq j,i'\neq j',\\ h(i)=h(j)=b,\\ h(i')=h(j')=b}} s_i s_j s_{i'} s_{j'} x_i y_j x_{i'} y_{j'}\right]$$



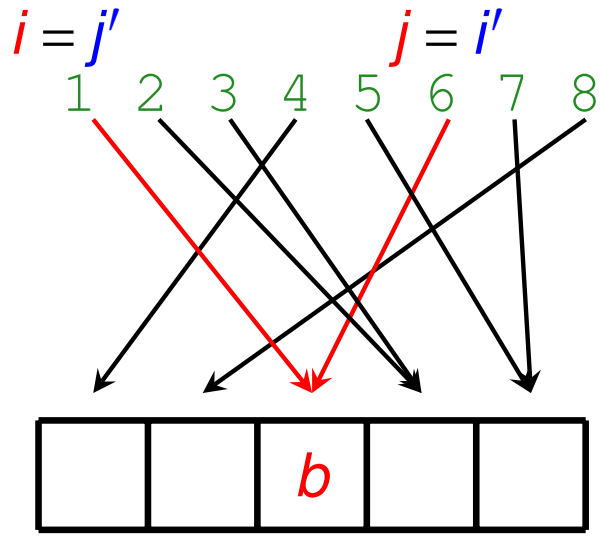

Summand zero in expectation over $s$ unless $i = i'$ and $j = j'$ or $i = j'$ and $j = i'$.




$$\mathbb{E}\left[\left(\sum_{b\in[B]}\sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j x_i y_j\right)^2\right] = \sum_{b\in[B]}\frac{1}{B^2}\sum_{i,j\in[n],i\neq j}(x_i^2 y_j^2 + x_i x_j y_i y_j)$$




$$\mathbb{E}\left[\left(\sum_{b\in[B]}\sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j x_i y_j\right)^2\right] = \frac{1}{B}\sum_{i,j\in[n],i\neq j}(x_i^2 y_j^2 + x_i x_j y_i y_j)$$




$$\mathbb{E}\left[\left(\sum_{b\in[B]}\sum_{\substack{i,j\in[n],i\neq j,\\ h(i)=h(j)=b}} s_i s_j x_i y_j\right)^2\right] = \frac{1}{B}\sum_{i,j\in[n],i\neq j}(x_i^2 y_j^2 + x_i x_j y_i y_j)$$

$$\begin{aligned}\sum_{i,j\in[n],i\neq j}(x_i^2 y_j^2 + x_i x_j y_i y_j) &\leq \sum_{i,j\in[n]}(x_i^2 y_j^2 + x_i x_j y_i y_j)\\ &= \|x\|^2\|y\|^2 + (x^T y)^2\\ &= 2\|x\|^2\|y\|^2\end{aligned}$$

## Theorem

*For every $x, y \in \mathbb{R}^n$, every $B \geq 1$*

$$\mathbb{E}\left[\left((Sx)^T Sy - x^T y\right)^2\right] \leq \frac{2}{B}\|x\|^2\|y\|^2.$$

## Theorem

*For every $x, y \in \mathbb{R}^n$, every $B \geq 1$*

$$\mathbb{E}\left[\left((Sx)^T Sy - x^T y\right)^2\right] \leq \frac{2}{B}\|x\|^2\|y\|^2.$$

So with constant probability

$$(Sx)^T Sy = x^T y \pm \frac{O(1)}{\sqrt{B}}\|x\|\|y\|$$

Need

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

Need

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

Enough to prove

$$\left\| U^T S^T S U - I_d \right\|_F^2 \leq \varepsilon^2$$

with high(?) probability

Need

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

Enough to prove

$$\left\| U^T S^T S U - I_d \right\|_F^2 \leq \varepsilon^2$$

with high(?) probability

$$\mathbb{E}\left[ \left\| U^T S^T S U - I_d \right\|_F^2 \right] = \sum_{i \in [d], j \in [d]} \left( (SU_i)^T SU_j - U_i^T U_j \right)^2$$

Need

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

Enough to prove

$$\left\| U^T S^T S U - I_d \right\|_F^2 \leq \varepsilon^2$$

with high(?) probability

$$\mathbb{E}\left[ \left\| U^T S^T S U - I_d \right\|_F^2 \right] = \sum_{i \in [d], j \in [d]} \left( (SU_i)^T S U_j - U_i^T U_j \right)^2$$

$$\leq \frac{2}{B} \sum_{i \in [d], j \in [d]} \|U_i\|^2 \|U_j\|^2$$

Need

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

Enough to prove

$$\left\| U^T S^T S U - I_d \right\|_F^2 \leq \varepsilon^2$$

with high(?) probability

$$\begin{aligned}
\mathbb{E}\left[ \left\| U^T S^T S U - I_d \right\|_F^2 \right] &= \sum_{i \in [d], j \in [d]} \left( (SU_i)^T S U_j - U_i^T U_j \right)^2 \\
&\leq \frac{2}{B} \sum_{i \in [d], j \in [d]} \| U_i \|^2 \| U_j \|^2 \\
&\leq \frac{2}{B} \left( \sum_{i \in [d]} \| U_i \|^2 \right)^2
\end{aligned}$$

Need

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

Enough to prove

$$\left\| U^T S^T S U - I_d \right\|_F^2 \leq \varepsilon^2$$

with high(?) probability

$$\begin{aligned}
\mathbb{E}\left[ \left\| U^T S^T S U - I_d \right\|_F^2 \right] &= \sum_{i \in [d], j \in [d]} \left( (SU_i)^T S U_j - U_i^T U_j \right)^2 \\
&\leq \frac{2}{B} \sum_{i \in [d], j \in [d]} \| U_i \|^2 \| U_j \|^2 \\
&\leq \frac{2}{B} \left( \sum_{i \in [d]} \| U_i \|^2 \right)^2
\end{aligned}$$

Need

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

Enough to prove

$$\left\| U^T S^T S U - I_d \right\|_F^2 \leq \varepsilon^2$$

with high(?) probability

$$\begin{aligned}
\mathbb{E}\left[ \left\| U^T S^T S U - I_d \right\|_F^2 \right] &= \sum_{i \in [d], j \in [d]} \left( (SU_i)^T S U_j - U_i^T U_j \right)^2 \\
&\leq \frac{2}{B} \sum_{i \in [d], j \in [d]} \| U_i \|^2 \| U_j \|^2 \\
&\leq \frac{2}{B} \left( \sum_{i \in [d]} \| U_i \|^2 \right)^2 \\
&\leq \frac{2}{B} \| U \|_F^4
\end{aligned}$$

Need

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

Enough to prove

$$\left\| U^T S^T S U - I_d \right\|_F^2 \leq \varepsilon^2$$

with high(?) probability

$$\begin{aligned}
\mathbb{E}\left[ \left\| U^T S^T S U - I_d \right\|_F^2 \right] &= \sum_{i \in [d], j \in [d]} \left( (SU_i)^T S U_j - U_i^T U_j \right)^2 \\
&\leq \frac{2}{B} \sum_{i \in [d], j \in [d]} \| U_i \|^2 \| U_j \|^2 \\
&\leq \frac{2}{B} \left( \sum_{i \in [d]} \| U_i \|^2 \right)^2 \\
&\leq \frac{2}{B} \| U \|_F^4
\end{aligned}$$

Need

$$\left\| U^T S^T S U - I_d \right\|_2 \leq \varepsilon$$

Enough to prove

$$\left\| U^T S^T S U - I_d \right\|_F^2 \leq \varepsilon^2$$

with high(?) probability

$$\begin{aligned}
\mathbb{E}\left[ \left\| U^T S^T S U - I_d \right\|_F^2 \right] &= \sum_{i \in [d], j \in [d]} \left( (SU_i)^T S U_j - U_i^T U_j \right)^2 \\
&\leq \frac{2}{B} \sum_{i \in [d], j \in [d]} \| U_i \|^2 \| U_j \|^2 \\
&\leq \frac{2}{B} \left( \sum_{i \in [d]} \| U_i \|^2 \right)^2 \\
&\leq \frac{2}{B} \| U \|_F^4 \\
&= \frac{2d^2}{B}
\end{aligned}$$

# Approximate least squares regression

Compute $SA, Sb$, let $x'$ be the minimizer of

$$\|SAx - Sb\|^2.$$

Time to compute $SA$ is $O(\text{nnz}(A))$, time to solve the resulting system $\text{poly}(d/\varepsilon)$.

# Approximate least squares regression

Compute $SA, Sb$, let $x'$ be the minimizer of

$$\|SAx - Sb\|^2.$$

Time to compute $SA$ is $O(\text{nnz}(A))$, time to solve the resulting system $\text{poly}(d/\varepsilon)$.

## Theorem

*Can find a* $(1+\varepsilon)$*-approximate solution to*

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2$$

*in time* $O(nnz(A)) + poly(d/\varepsilon)$.

# Ridge regression?

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda \|x\|^2$$

# Ridge regression?

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda \|x\|^2$$

Sketching dimension smaller than $d$ if $A$ is approximately low rank?

# Ridge regression?

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda \|x\|^2$$

Sketching dimension smaller than $d$ if $A$ is approximately low rank?

**Theorem**

*Can find a* $(1+\varepsilon)$*-approximate solution to*

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda \|x\|^2$$

*in time* $O(nnz(A)) + poly(s_\lambda/\varepsilon)d$*, where* $s_\lambda$ *is the statistical dimension of* $A$*.*

# Ridge regression?

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda \|x\|^2$$

Sketching dimension smaller than $d$ if $A$ is approximately low rank?

## Theorem

*Can find a* $(1+\varepsilon)$*-approximate solution to*

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda \|x\|^2$$

*in time* $O(nnz(A)) + poly(s_\lambda/\varepsilon)d$*, where* $s_\lambda$ *is the statistical dimension of* $A$*.*

$$s_\lambda(A) := \mathrm{tr}\left((A^T A + \lambda I)^+ A^T A\right) = \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda}$$

# Statistical dimension

$$s_\lambda(A) := \operatorname{tr}\left((A^T A + \lambda I)^+ A^T A\right) = \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda}$$

# Statistical dimension

$$s_\lambda(A) := \operatorname{tr}\left((A^T A + \lambda I)^+ A^T A\right) = \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda}$$

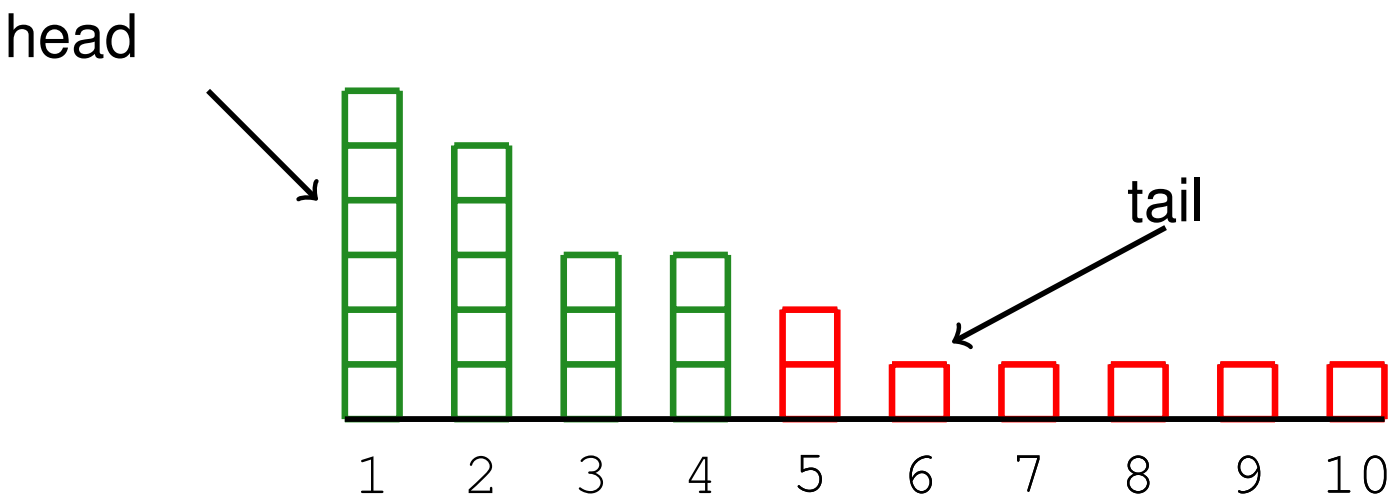
Statistical dimension$\approx$ # eigenvalues above $\lambda$
$+$(sum of eigenvalues below $\lambda$)$/\lambda$

# Statistical dimension

$$s_\lambda(A) := \mathrm{tr}\left((A^T A + \lambda I)^+ A^T A\right) = \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda}$$

Statistical dimension$\approx$ # eigenvalues above $\lambda$
$+$(sum of eigenvalues below $\lambda$)$/\lambda$

Recall: in CountSketch need $B \geq |\mathrm{HEAD}|$ and get estimation error $\approx \|x_{TAIL}\|/\sqrt{B}$.



# Ridge regression?

## Theorem

*Can find a* $(1+\varepsilon)$*-approximate solution to*

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda \|x\|^2$$

*in time* $O(nnz(A)) + poly(s_\lambda/\varepsilon)d$*, where* $s_\lambda$ *is the statistical dimension of* $A$*.*

# Ridge regression?

## Theorem

*Can find a* $(1+\varepsilon)$*-approximate solution to*

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda \|x\|^2$$

*in time* $O(nnz(A)) + poly(s_\lambda/\varepsilon)d$*, where* $s_\lambda$ *is the statistical dimension of* $A$*.*

(Roughly) $U = A(A^TA)^{-1/2}$ becomes $U = A(A^TA + \lambda I)^{-1/2}$

# Ridge regression?

## Theorem

*Can find a* $(1+\varepsilon)$*-approximate solution to*

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda \|x\|^2$$

*in time* $O(nnz(A)) + poly(s_\lambda/\varepsilon)d$*, where* $s_\lambda$ *is the statistical dimension of A.*

(Roughly) $U = A(A^TA)^{-1/2}$ becomes $U = A(A^TA + \lambda I)^{-1/2}$

$$\begin{aligned} \|U\|_F^2 &= \mathrm{tr}((A^TA + \lambda I)^{-1/2} A^TA (A^TA + \lambda I)^{-1/2}) \\ &= \mathrm{tr}((A^TA + \lambda I)^{-1} A^TA) \\ &= s_\lambda \end{aligned}$$

# Kernel ridge regression

**Input:**

- a sequence of $d$-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^d$
- values $y_j = f(x_j), j = 1, \ldots, n$

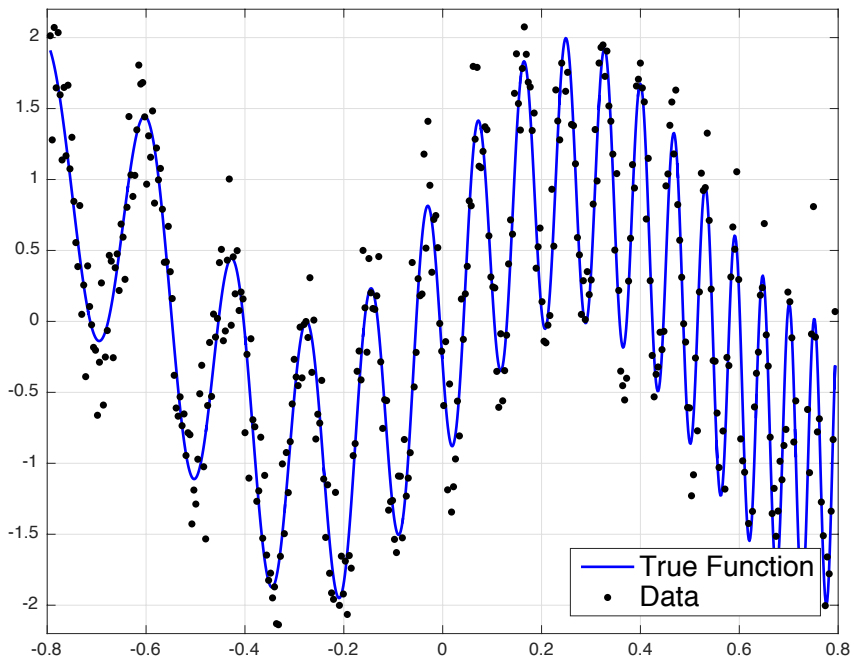
**Output:** approximation from class of 'smooth' functions on $\mathbb{R}^d$

# Kernel ridge regression

**Input:**

- a sequence of $d$-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^d$
- values $y_j = f(x_j), j = 1, \ldots, n$

**Output:** approximation from class of 'smooth' functions on $\mathbb{R}^d$



# Kernel ridge regression

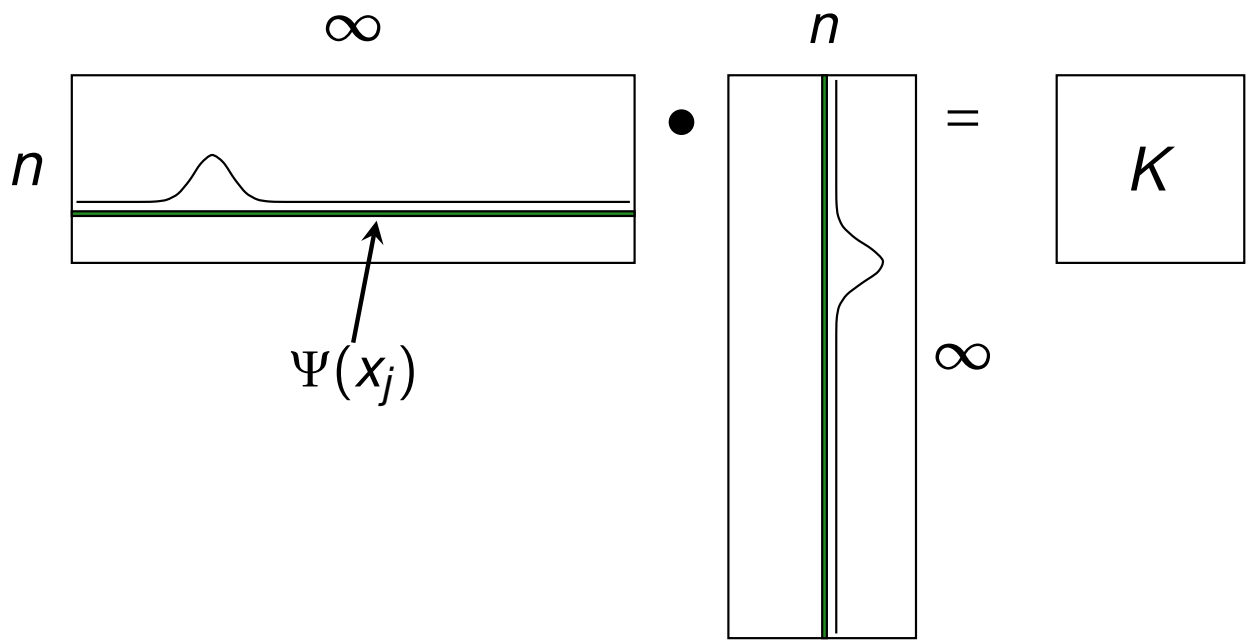
Main computational effort:

$$\boxed{(K + \lambda I)^{-1} y}$$

# Kernel ridge regression

Main computational effort:

$$(K + \lambda I)^{-1} y$$




The $(i,j)$-th entry of Gaussian kernel matrix $K$ is

$$K_{ij} = e^{-(x_i - x_j)^2/2}$$

# CountSketch for matrices?

Input: $r$ parties hold vectors $x_1, \ldots, x_r \in \mathbb{R}^n$

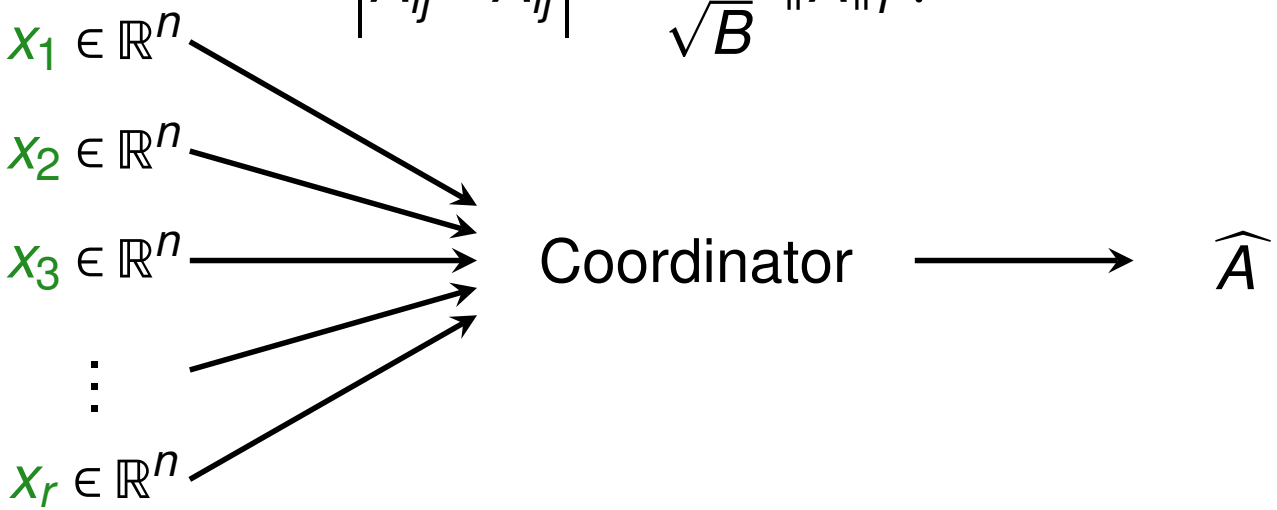
each party sends $O(B \log n)$ bits to coordinator

(assume shared randomness)

Output: find largest entries in $A = \sum_{i=1}^{r} x_i x_i^T$

more precisely, output approximation $\widehat{A}$

$$\left| \widehat{A_{ij}} - A_{ij} \right| = \frac{O(1)}{\sqrt{B}} \|A\|_F.$$




Every party $i$ sends CountSketch($x_i x_i^T$) into $O(B)$ buckets

(slow)

# CountSketch for matrices?

Input: $r$ parties hold vectors $x_1, \ldots, x_r \in \mathbb{R}^n$

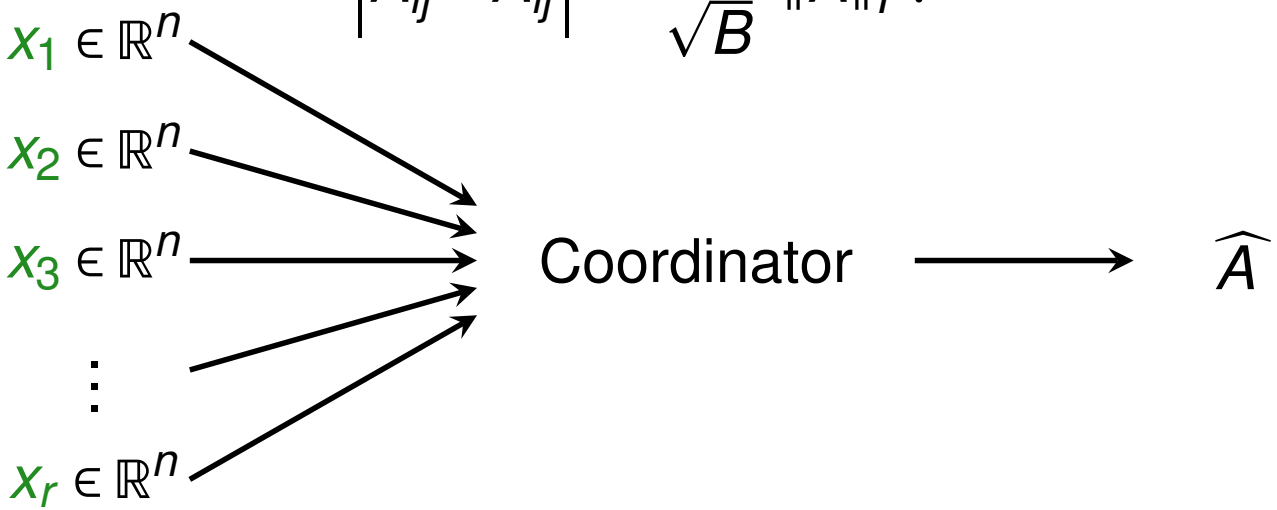
each party sends $O(B \log n)$ bits to coordinator

(assume shared randomness)

Output: find largest entries in $A = \sum_{i=1}^{r} x_i x_i^T$

more precisely, output approximation $\widehat{A}$

$$\left|\widehat{A_{ij}} - A_{ij}\right| = \frac{O(1)}{\sqrt{B}} \|A\|_F.$$




Every party $i$ sends SOMESKETCH($x_i$) into $B$ buckets?

(fast)

Define

$$A = \sum_{i=1}^{r} x_i x_i^T \in \mathbb{R}^{n \times n}.$$

Hash function

$$h : [n] \times [n] \to [B]$$

and random signs

$$s : [n] \times [n] \to \{-1, +1\}.$$

$$(Sx)_b = \sum_{i,j \in [n] : h(i,j) = b} s(i,j) \cdot x_i x_j.$$

CountSketch$(x_i x_i^T)$ takes $n^2$ time to compute...

Define

$$A = \sum_{i=1}^{r} x_i x_i^T \in \mathbb{R}^{n \times n}.$$

Hash function

$$h : [n] \times [n] \to [B]$$

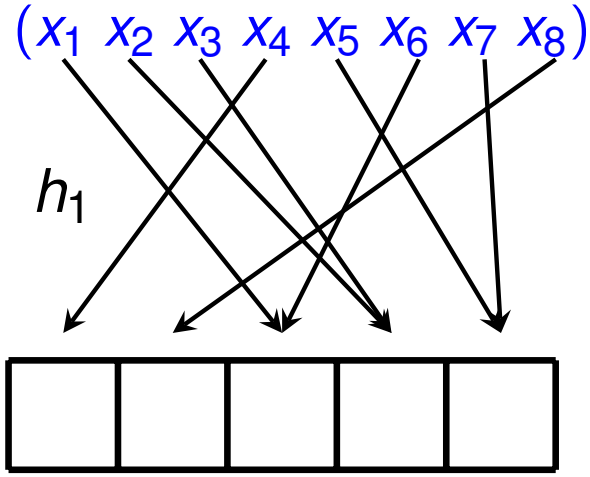
and random signs

$$s : [n] \times [n] \to \{-1, +1\}.$$

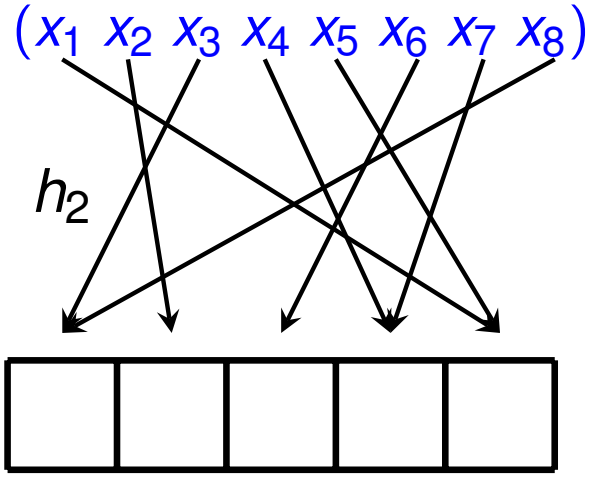
$$(Sx)_b = \sum_{i,j \in [n] : h(i,j) = b} s(i,j) \cdot x_i x_j.$$

CountSketch$(x_i x_i^T)$ takes $n^2$ time to compute...

Make hash functions 'separable'?




$S_1 x$




$S_2 x$

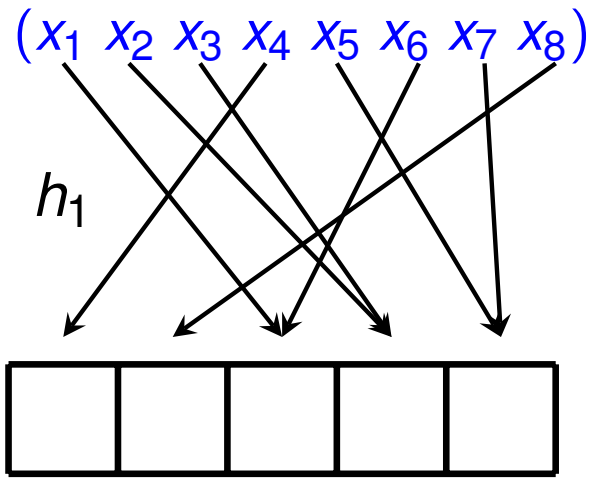
Take two independent instances of CountSketch: hash functions

$$h_1, h_2 : [n] \to [B],$$

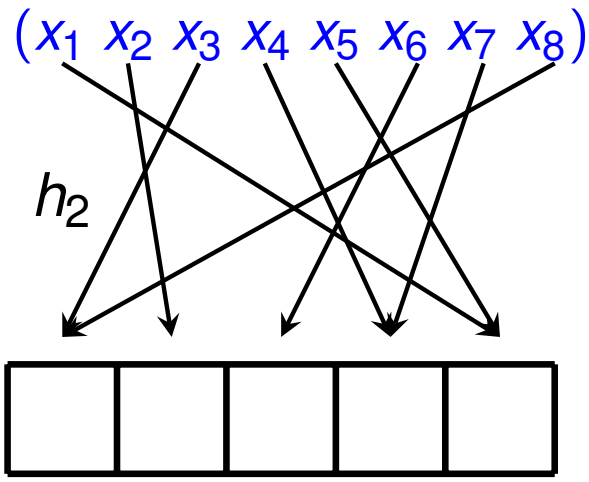
random signs

$$s_1, s_2 : [n] \to \{-1, +1\}$$

Tensor $\text{CountSketch}_1$ and $\text{CountSketch}_2$!



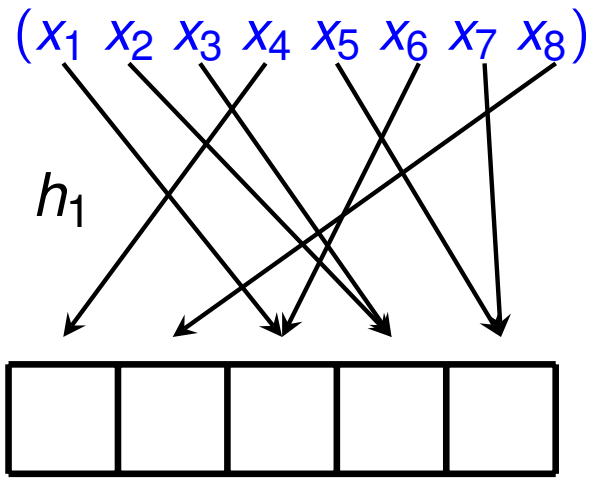

$S_1 x$



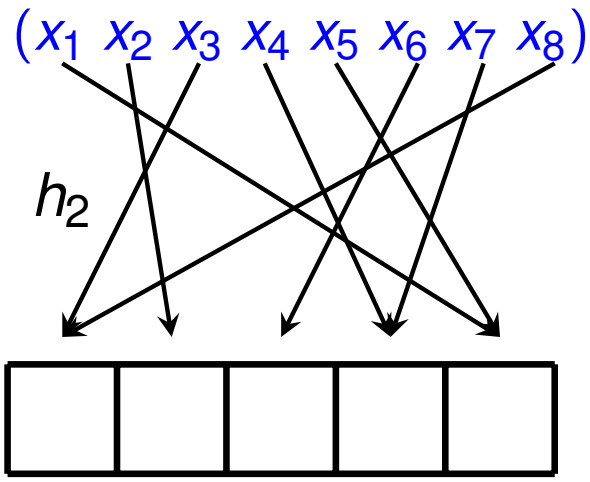

$S_2 x$

Define tensoring of CountSketch$_1$ and CountSketch$_2$:

$$h(i,j) = h_1(i) + h_2(j) \pmod{B}.$$




$S_1 x$

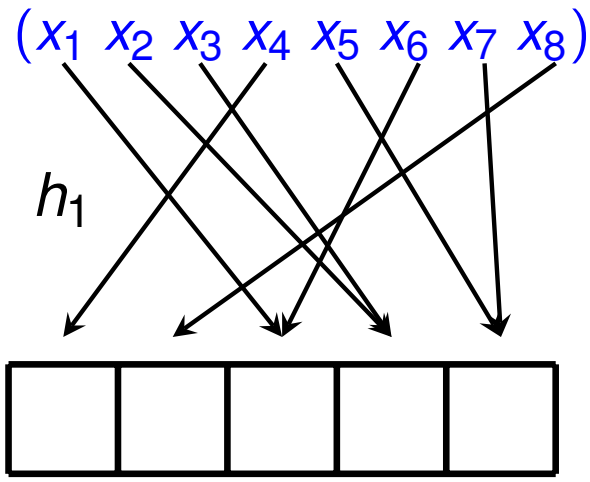



$S_2 x$

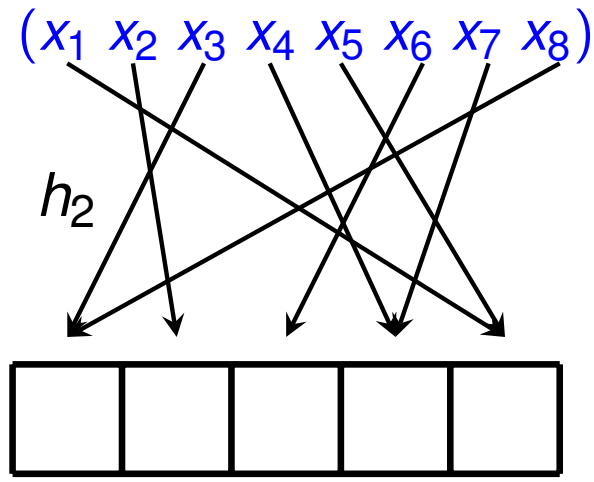
Define tensoring of CountSketch$_1$ and CountSketch$_2$:

$$h(i,j) = h_1(i) + h_2(j) \pmod{B}.$$

and $s(i,j) = s_1(i) \cdot s_2(j)$.




$S_1 x$



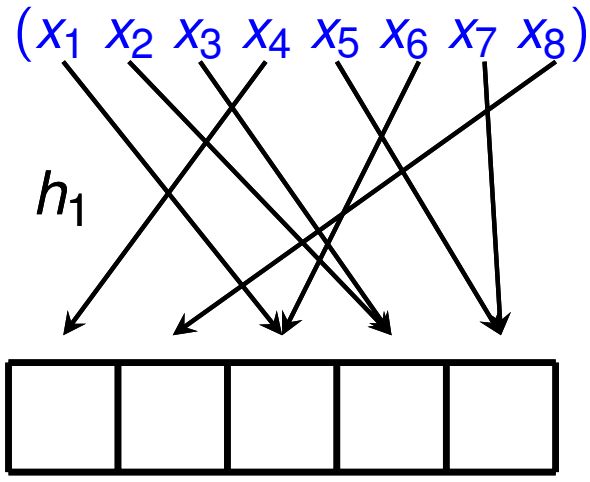

$S_2 x$

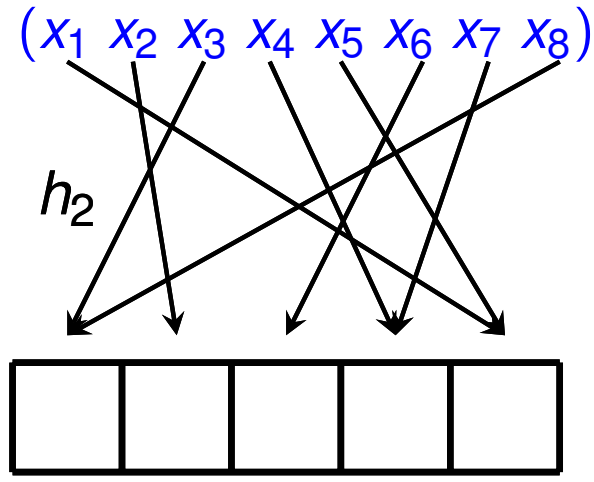
Define tensoring of CountSketch$_1$ and CountSketch$_2$:

$$h(i,j) = h_1(i) + h_2(j) \pmod{B}.$$

and $s(i,j) = s_1(i) \cdot s_2(j)$.

$$\begin{aligned}(Sx)_b &= \sum_{i,j\in[n]:h(i,j)=b} s(i,j)\cdot x_i x_j \\ &= \sum_{i,j\in[n]:h_1(i)+h_2(j)=b} s_1(i)\cdot s_2(j)\cdot x_i x_j\end{aligned}$$



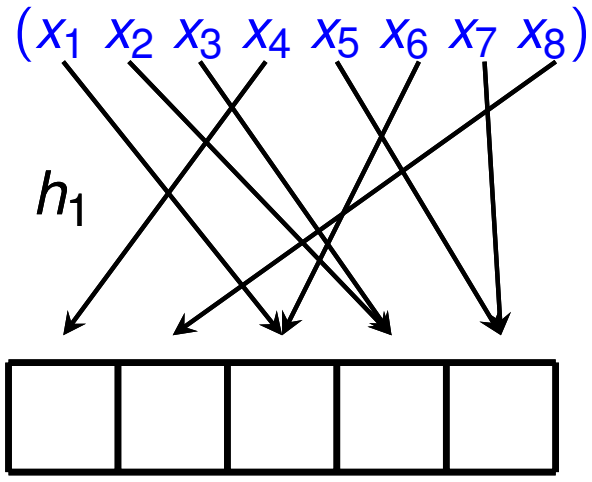

$S_1 x \quad * \quad S_2 x \quad = Sx$

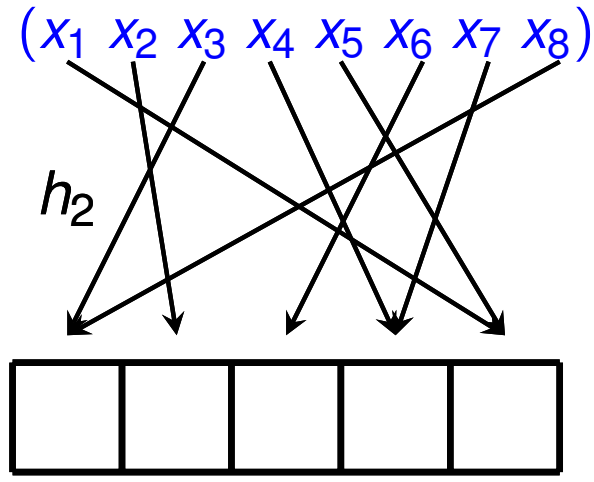
Convolution of $S_1 x$ and $S_2 x$: for $b \in [B]$

$$(Sx)_b = \sum_{i,j \in [d]: h_1(i)+h_2(j)=b} s_1(i) \cdot s_2(j) \cdot x_i x_j$$

$$= \sum_{a,a' \in [B]: a+a'=b \pmod B} (S_1 x)_a (s_2 x)_{a'}$$

$$= ((S_1 x) * (S_2 x))(b)$$




$S_1 x \quad * \quad S_2 x \quad = Sx$

Convolution of $S_1 x$ and $S_2 x$: for $b \in [B]$

$$(Sx)_b = \sum_{i,j \in [d]: h_1(i)+h_2(j)=b} s_1(i) \cdot s_2(j) \cdot x_i x_j$$

$$= \sum_{a,a' \in [B]: a+a'=b \pmod B} (S_1 x)_a (s_2 x)_{a'}$$

$$= ((S_1 x) * (S_2 x))(b)$$

This is the TENSORSKETCH of Pagh and Pham'13

$$(Sx)_b = \sum_{i,j \in [n] : h_1(i)+h_2(j)=b} s_1(i) \cdot s_2(j) \cdot x_i x_j.$$

Variance of $(Sx)^T(Sy)$? (exercise)

Hash function $h(i,j) = h_1(i) + h_2(j)$ is pairwise independent

Hash function $h(i,j) = h_1(i) + h_2(j)$ is pairwise independent

Sign function $s(i,j) = s_1(i) \cdot s_2(j)$ is not 4-wise independent

Hash function $h(i,j) = h_1(i) + h_2(j)$ is pairwise independent

Sign function $s(i,j) = s_1(i) \cdot s_2(j)$ is not 4-wise independent

$$A = \sum_i x_i x_i^T = \begin{pmatrix} +1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & \mathbf{+1} & -1 & -1 & -1 & \mathbf{-1} & -1 \\ -1 & -1 & +10 & +5 & +2 & -1 & -1 \\ -1 & -1 & +5 & 5 & +1 & -1 & -1 \\ -1 & -1 & +2 & +1 & +2 & -1 & -1 \\ -1 & \mathbf{-1} & -1 & -1 & -1 & \mathbf{+1} & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & +1 \end{pmatrix}$$

Hash function $h(i,j) = h_1(i) + h_2(j)$ is pairwise independent

Sign function $s(i,j) = s_1(i) \cdot s_2(j)$ is not 4-wise independent

$$A = \sum_i x_i x_i^T = \begin{pmatrix} +1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & \mathbf{+1} & -1 & -1 & -1 & \mathbf{-1} & -1 \\ -1 & -1 & +10 & +5 & +2 & -1 & -1 \\ -1 & -1 & +5 & 5 & +1 & -1 & -1 \\ -1 & -1 & +2 & +1 & +2 & -1 & -1 \\ -1 & \mathbf{-1} & -1 & -1 & -1 & \mathbf{+1} & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & +1 \end{pmatrix}$$

Note: for every $a,b,c,d \in [n]$ one has

$$s_{a,b} \cdot s_{b,c} \cdot s_{c,d} \cdot s_{d,a} = 1$$

Higher tensor powers?

Define

$$A = \sum_{j=1}^{r} x_j^{\otimes q} \in \mathbb{R}^{n^q}.$$

Define

$$A = \sum_{j=1}^{r} x_j^{\otimes q} \in \mathbb{R}^{n^q}.$$

Compute $\widehat{A}$ such that for every $\mathbf{i} = (i_1, i_2, \ldots, i_q) \in [d]$

$$\left| \widehat{A}_{\mathbf{i}} - A_{\mathbf{i}} \right| \leq \frac{O(1)}{\sqrt{B}} \cdot \|A\|_F$$

from sketches of $x_1, x_2, \ldots, x_r$?

Tensorized version of CountSketch: hash functions

$$h_1, h_2, \ldots, h_q : [n] \to [B],$$

and

$$h(\mathbf{i}) = h_1(i_1) + h_2(i_2) + \ldots + h_q(i_q) \pmod{B}.$$

Random signs

$$s_1, s_2, \ldots, s_q : [n] \to \{-1, +1\}$$

and

$$s(\mathbf{i}) = s_1(i_1) \cdot s_2(i_2) \cdots s_q(i_q)$$

Tensorized version of CountSketch: hash functions

$$h_1, h_2, \ldots, h_q : [n] \to [B],$$

and

$$h(\mathbf{i}) = h_1(i_1) + h_2(i_2) + \ldots + h_q(i_q) \pmod{B}.$$

Random signs

$$s_1, s_2, \ldots, s_q : [n] \to \{-1, +1\}$$

and

$$s(\mathbf{i}) = s_1(i_1) \cdot s_2(i_2) \cdots s_q(i_q)$$

Can apply to $x$ using $q$ FFTs of length $B$!

Tensorized version of CountSketch: hash functions

$$h_1, h_2, \ldots, h_q : [n] \to [B],$$

and

$$h(\mathbf{i}) = h_1(i_1) + h_2(i_2) + \ldots + h_q(i_q) \pmod{B}.$$

Random signs

$$s_1, s_2, \ldots, s_q : [n] \to \{-1, +1\}$$

and

$$s(\mathbf{i}) = s_1(i_1) \cdot s_2(i_2) \cdots s_q(i_q)$$

Can apply to $x$ using $q$ FFTs of length $B$!

**Theorem (Pagh and Pham'13)**

*With high probability for every* $\mathbf{i} \in [n]^q$

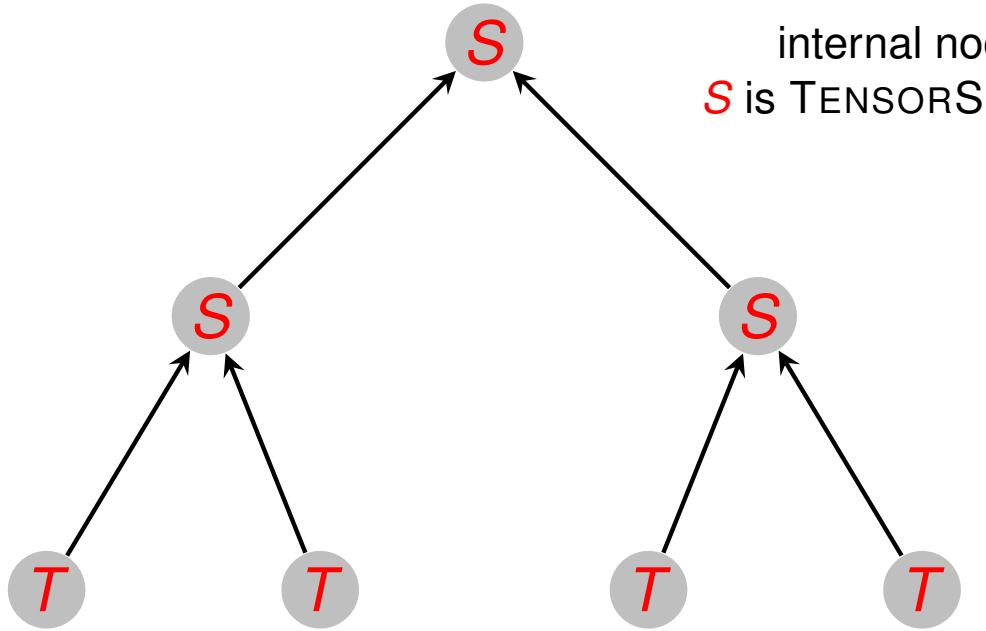
$$\left| \widehat{A_{\mathbf{i}}} - A_{\mathbf{i}} \right| = \frac{O(1)}{\sqrt{B}} \|A\|_F.$$

# TensorSketch as a subspace embedding?

Yes, but with an exponential dependence on degree $q$..

**Theorem (Avron, Nguyen, Woodruff'19)**
*Hashing into* $\approx 3^q s_\lambda^2/\varepsilon^2$ *buckets suffices.*

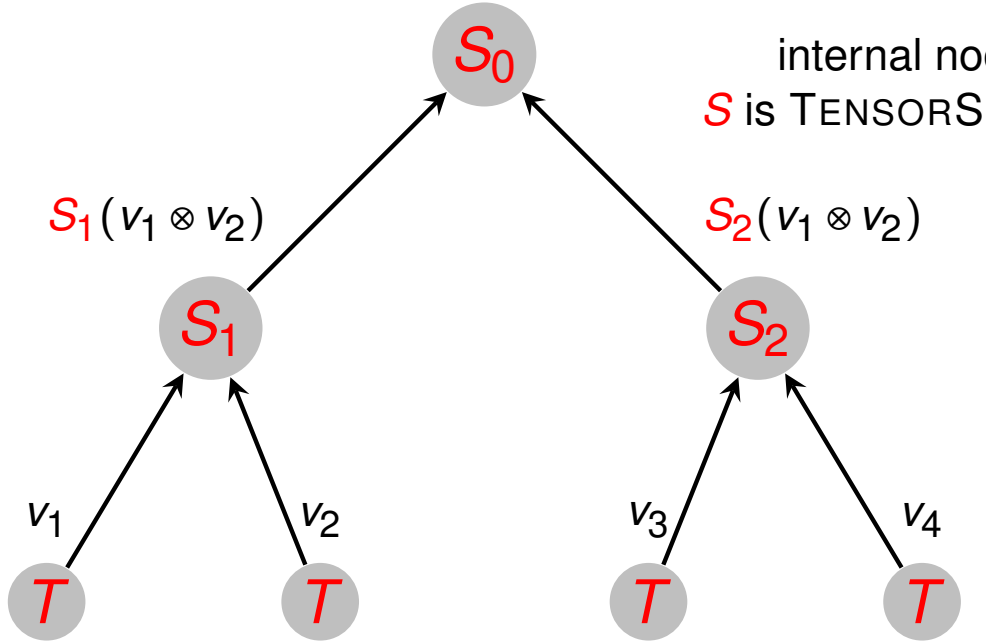
Polynomial dependence on $q$?




internal nodes:
$S$ is TENSORSKETCH or TENSORSRHT

leaves:
$T$ is COUNTSKETCH or OSNAP

Result of Ahle, K., Knudsen, Pagh, Velingker, Woodruff and Zandieh'20




internal nodes:
$S$ is TensorSketch or TensorSRHT

leaves:
$T$ is CountSketch or OSNAP

Only tensor twice, so at a single node

$$\mathbb{E}\left[\left\|Sx^{\otimes q}\right\|^4\right] = O\left(\frac{3^2}{B}\right) \cdot \left\|x^{\otimes q}\right\|^4$$

Set $B \approx {s_\lambda}^2 q$

Errors from $\leq 2q$ independent repetitions do not accumulate exponentially

# Gaussian kernel

$$K_{ij} = e^{-\|x_i - x_j\|^2/2} = e^{-\|x_i\|^2/2 - \|x_j\|^2/2 + x_i^T x_j}$$

Assume $\|x_i\| = 1$ for all $i$.

$$K_{ij} = e^{-1} \cdot e^{x_i^T x_j}$$

# Gaussian kernel

$$e^{x_i^T x_j} = \sum_{k \geq 0} \frac{(x_i^T x_j)^k}{k!} = \sum_{k \geq 0} \frac{((x_i^{\otimes k})^T (x_j^{\otimes k})}{k!}$$

# Gaussian kernel

$$e^{x_i^T x_j} = \sum_{k \geq 0} \frac{(x_i^T x_j)^k}{k!} = \sum_{k \geq 0} \frac{((x_i^{\otimes k})^T (x_j^{\otimes k})}{k!}$$

So

$$K = \sum_{k \geq 0} \frac{1}{k!} \cdot (X^{\otimes k})^T X^{\otimes k}$$

# Gaussian kernel

$$e^{x_i^T x_j} = \sum_{k \geq 0} \frac{(x_i^T x_j)^k}{k!} = \sum_{k \geq 0} \frac{((x_i^{\otimes k})^T (x_j^{\otimes k})}{k!}$$

So

$$K = \sum_{k \geq 0} \frac{1}{k!} \cdot (X^{\otimes k})^T X^{\otimes k}$$

# Gaussian kernel

$$e^{x_i^T x_j} = \sum_{k \geq 0} \frac{(x_i^T x_j)^k}{k!} = \sum_{k \geq 0} \frac{((x_i^{\otimes k})^T (x_j^{\otimes k})}{k!}$$

So

$$K = \sum_{k \geq 0} \frac{1}{k!} \cdot (X^{\otimes k})^T X^{\otimes k}$$

Assume bounded dataset: $\|x_i\|_2 \leq R$ for all $i$.

# Gaussian kernel

$$e^{x_i^T x_j} = \sum_{k \geq 0} \frac{(x_i^T x_j)^k}{k!} = \sum_{k \geq 0} \frac{((x_i^{\otimes k})^T (x_j^{\otimes k})}{k!}$$

So

$$K = \sum_{k \geq 0} \frac{1}{k!} \cdot (X^{\otimes k})^T X^{\otimes k}$$

Assume bounded dataset: $\|x_i\|_2 \leq R$ for all $i$.

Truncate series after $O(R \log(1/\varepsilon))$ terms, additive precision $\varepsilon$.

# Gaussian kernel

$$e^{x_i^T x_j} = \sum_{k \geq 0} \frac{(x_i^T x_j)^k}{k!} = \sum_{k \geq 0} \frac{((x_i^{\otimes k})^T (x_j^{\otimes k})}{k!}$$

So

$$K = \sum_{k \geq 0} \frac{1}{k!} \cdot (X^{\otimes k})^T X^{\otimes k}$$

Assume bounded dataset: $\|x_i\|_2 \leq R$ for all $i$.

Truncate series after $O(R \log(1/\varepsilon))$ terms, additive precision $\varepsilon$.

Sketch all terms in truncated Taylor expansion

Theorem (Ahle, K., Knudsen, Pagh, Velingker, Woodruff and Zandieh'20)

*Sketch with target dimension* $\mathrm{poly}(R, \log n, 1/\varepsilon) s_\lambda^2$.

Remove dependence on $R$?

A high dimensional version of Fast Multipole Methods of Greengard and Rokhlin'83?

Some recent progress in Charikar, K., Waingarten'??

# Statistical dimension?

Q: is $s_\lambda$ a good parameter for kernel matrices? Better bounds using geometric information?