COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Andrew McGregor

Lecture 10

(ϵ, k)-FREQUENT ITEMS PROBLEM

Given stream of n items $x_1, \ldots, x_n$ where each $x_i \in U$. Return a set F, such that for every $x \in U$:

1. If $f(x) \geq n/k$ then $x \in F$
2. If $f(x) < (1 - \epsilon)n/k$ then $x \notin F$

where $f(x)$ is the number of times x appears in the stream.

(ϵ, k)-FREQUENT ITEMS PROBLEM

Given stream of n items $x_1, \ldots, x_n$ where each $x_i \in U$. Return a set F, such that for every $x \in U$:

1. If $f(x) \geq n/k$ then $x \in F$
2. If $f(x) < (1-\epsilon)n/k$ then $x \notin F$

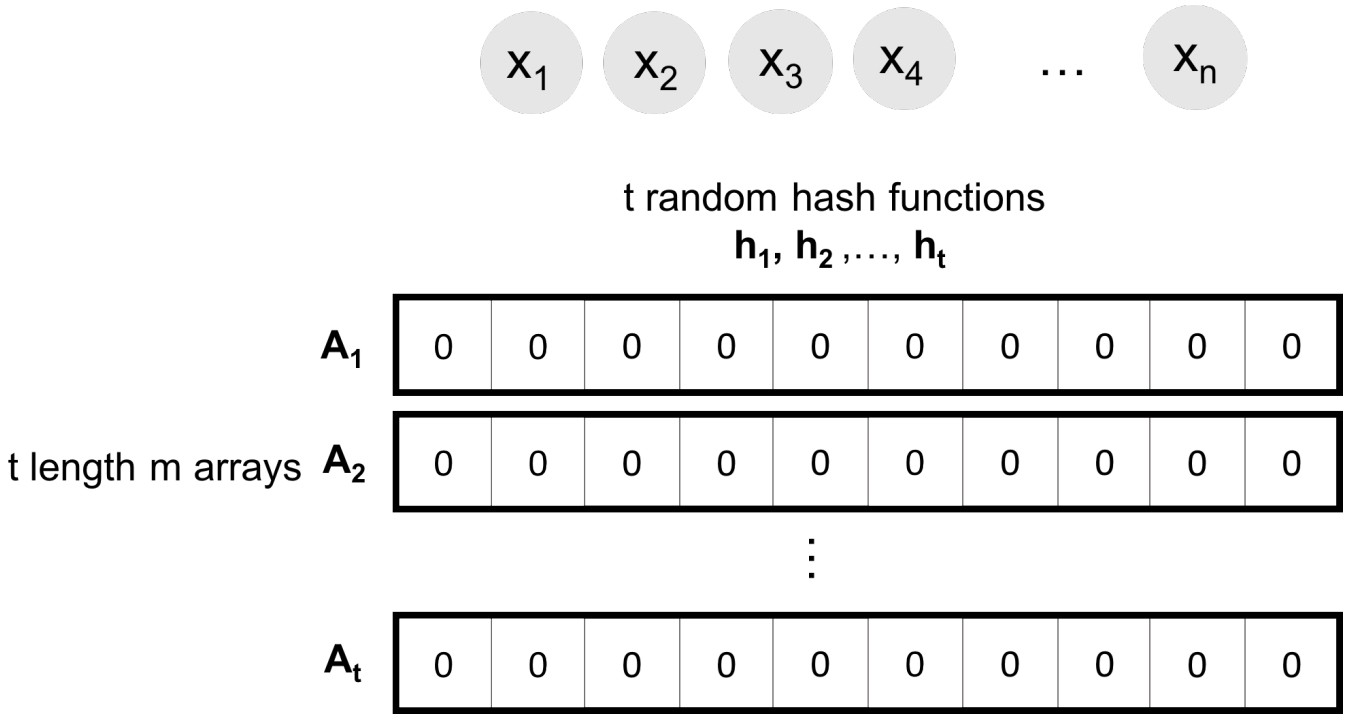
where $f(x)$ is the number of times x appears in the stream.

Relationship to Frequency Estimation. Note that if you have an estimate $\tilde{f}(x)$ for each each $f(x)$ such that

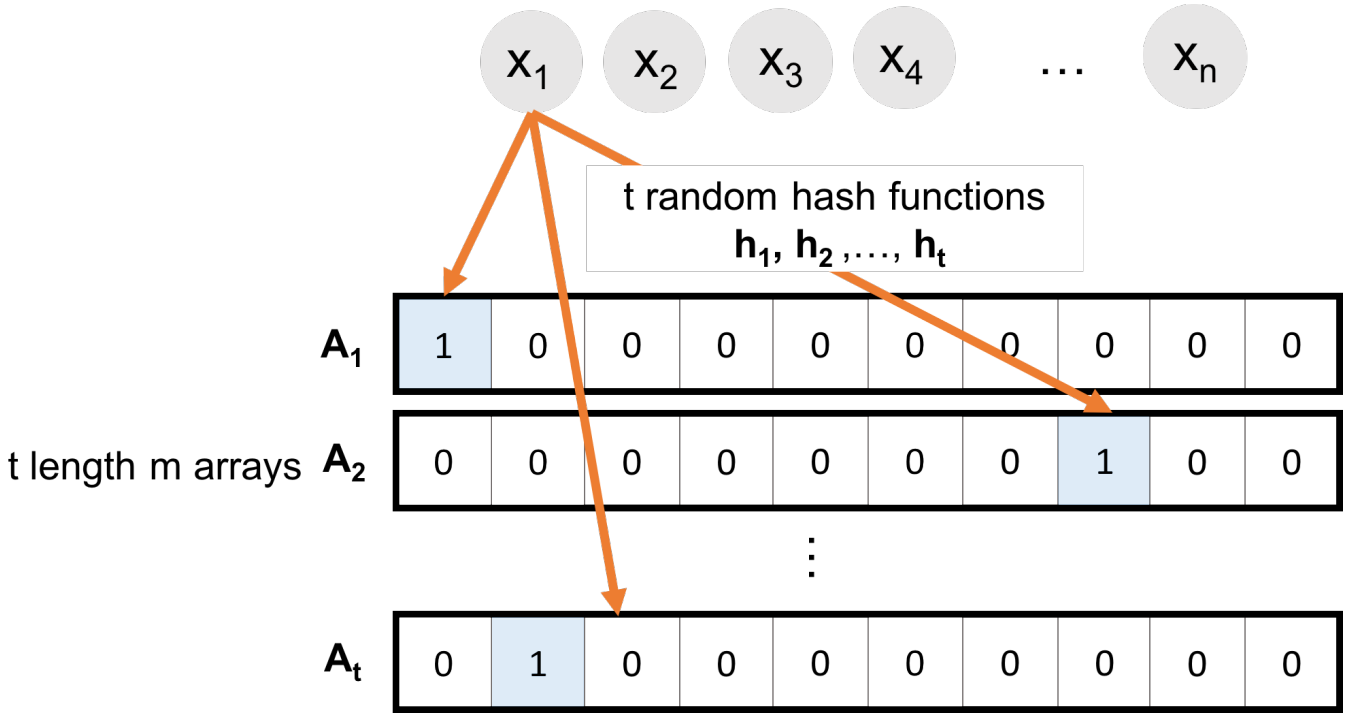
$$f(x) \leq \tilde{f}(x) \leq f(x) + \epsilon n/k$$

then you can solve the above problem.

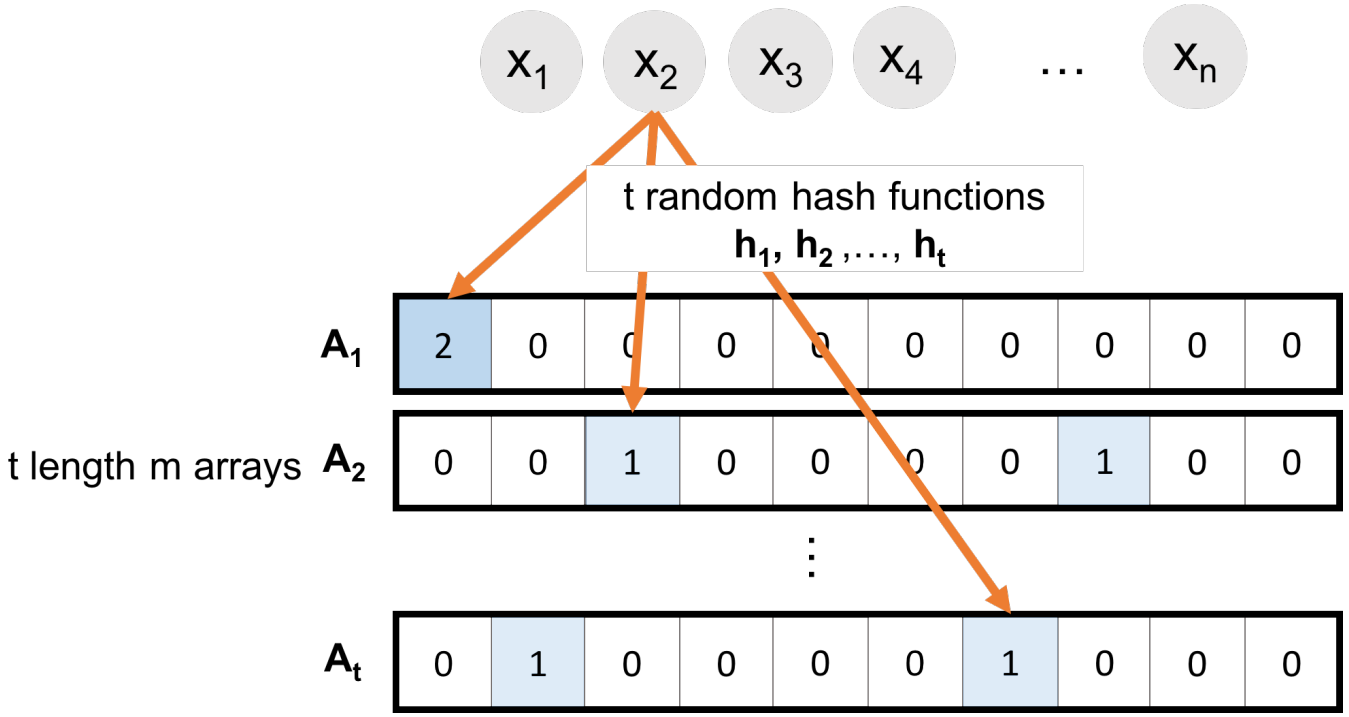
COUNT-MIN SKETCH ACCURACY



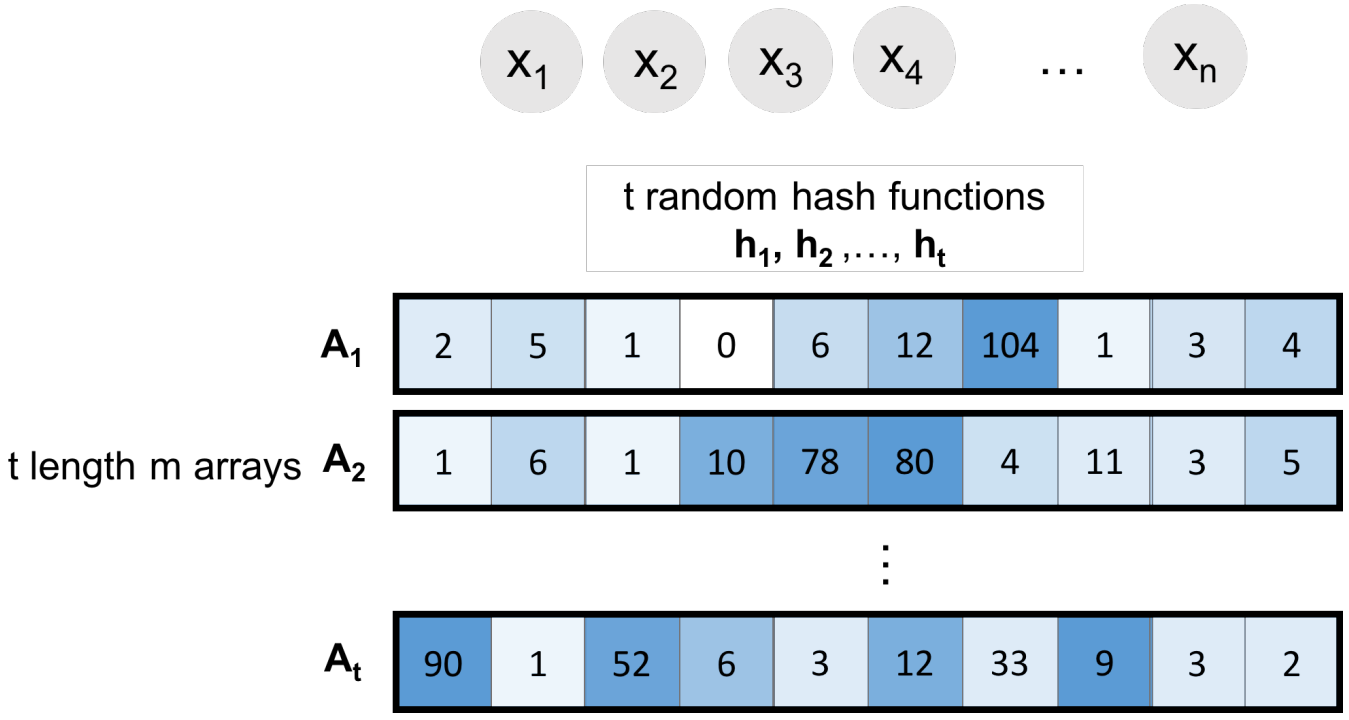
COUNT-MIN SKETCH ACCURACY



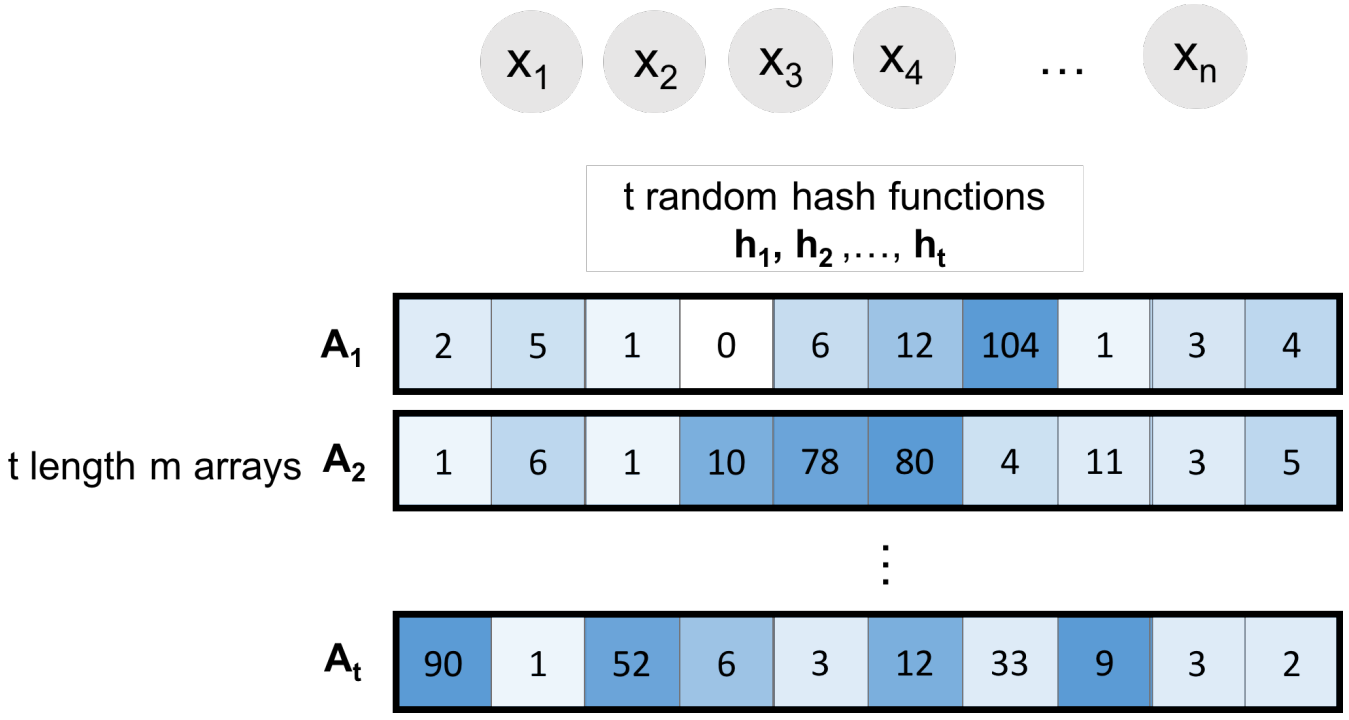
COUNT-MIN SKETCH ACCURACY



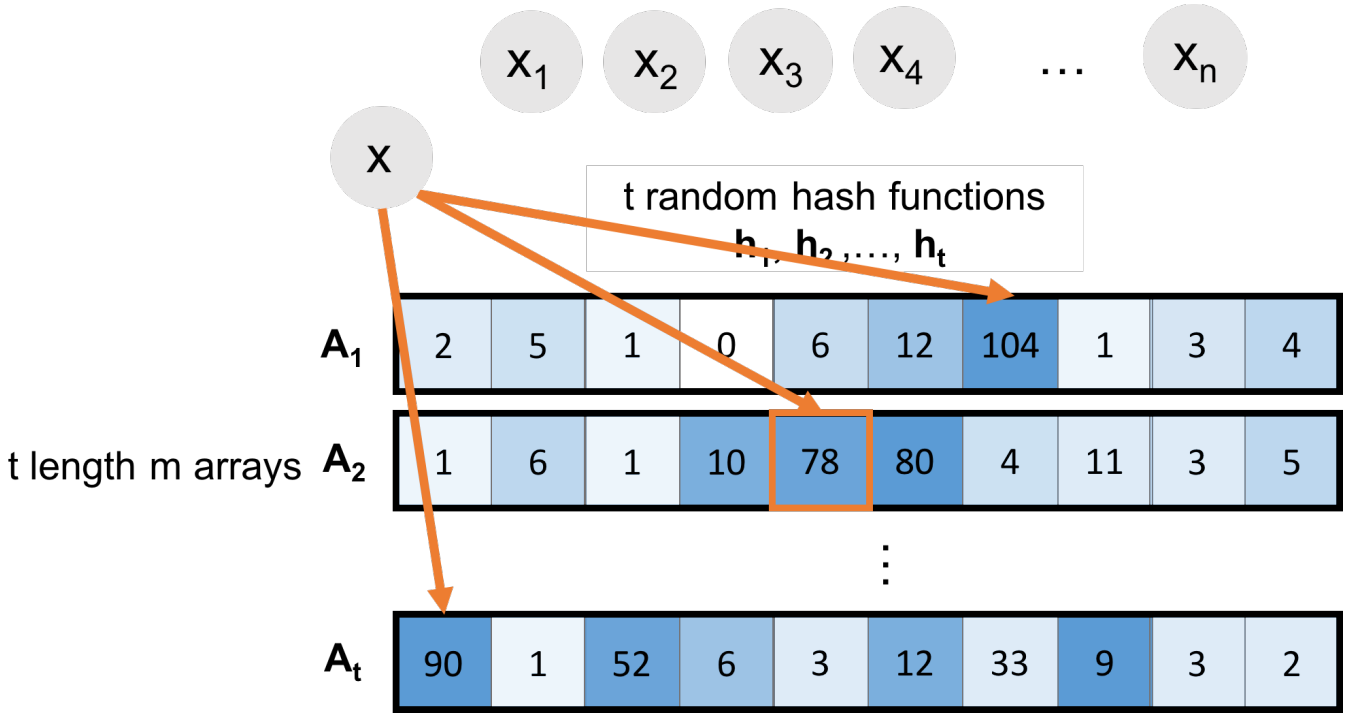
COUNT-MIN SKETCH ACCURACY



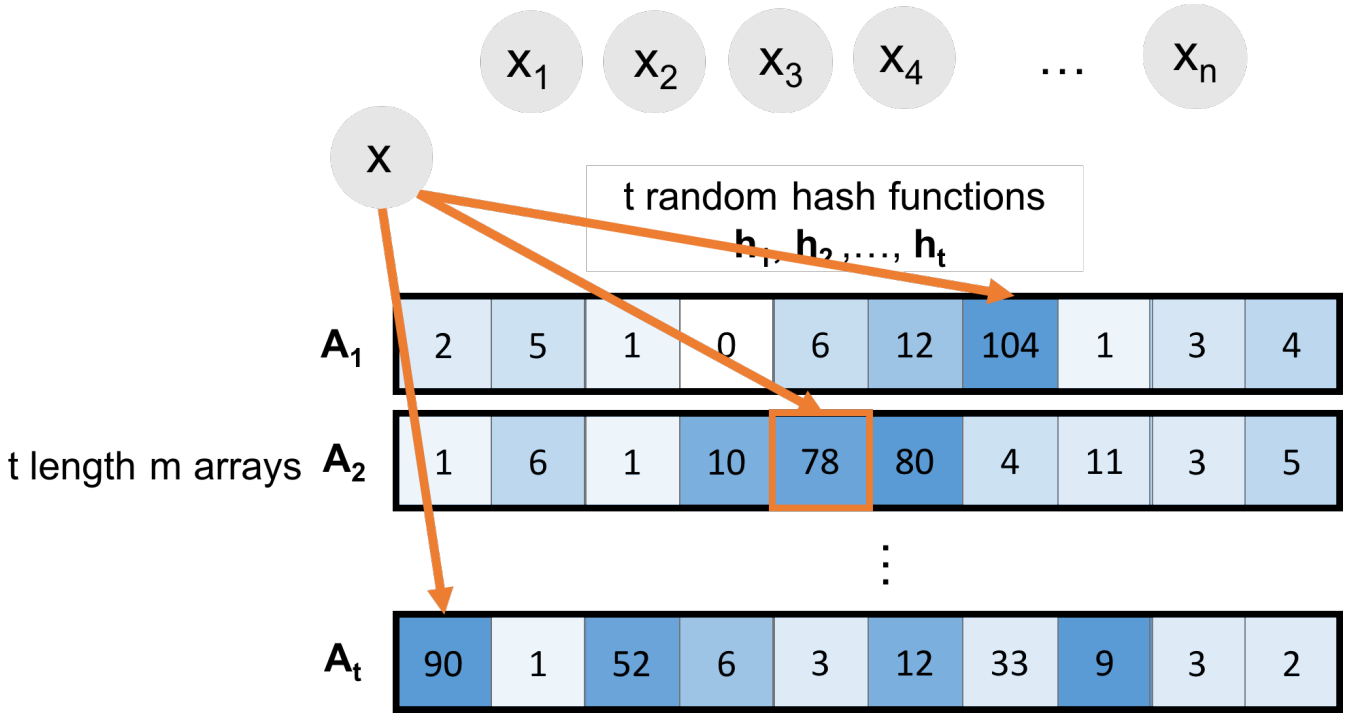
COUNT-MIN SKETCH ACCURACY



COUNT-MIN SKETCH ACCURACY

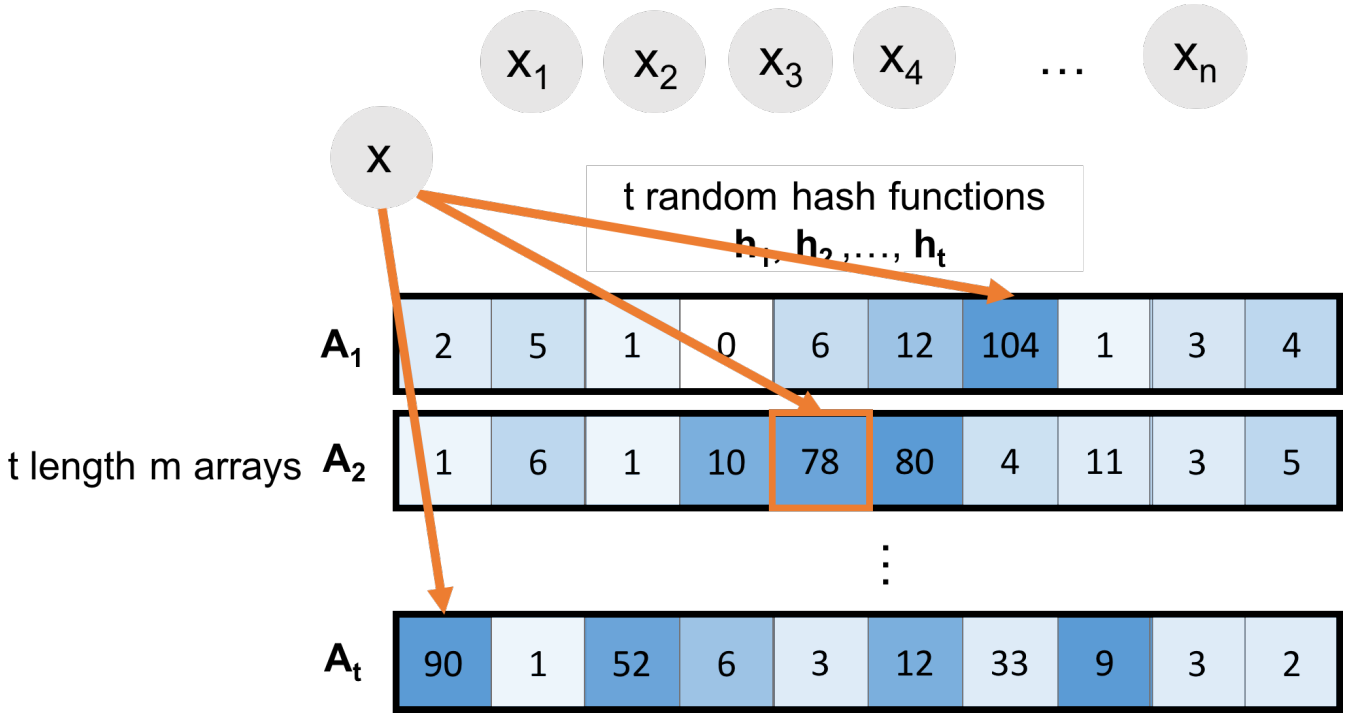


COUNT-MIN SKETCH ACCURACY



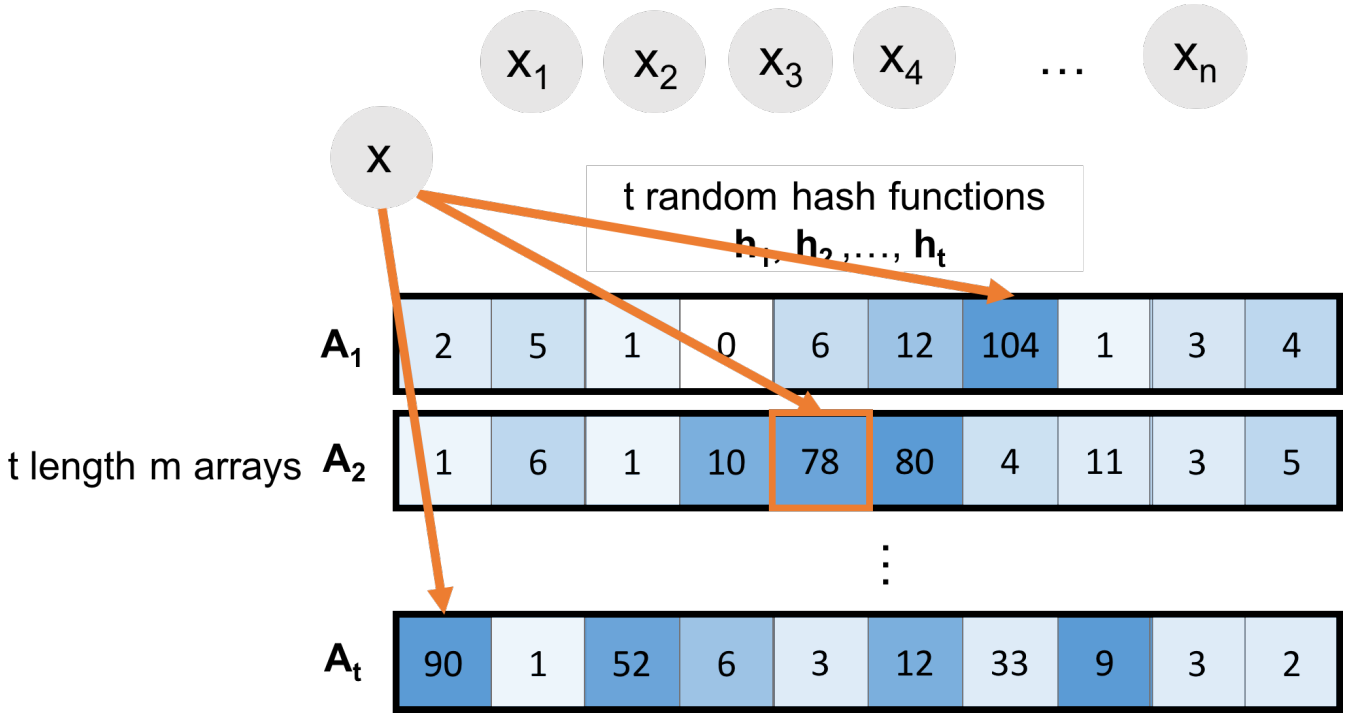
- Estimate $f(x)$ with $\tilde{f}(x) = \min_{i \in [t]} A_i[\mathbf{h}_i(x)]$.

COUNT-MIN SKETCH ACCURACY



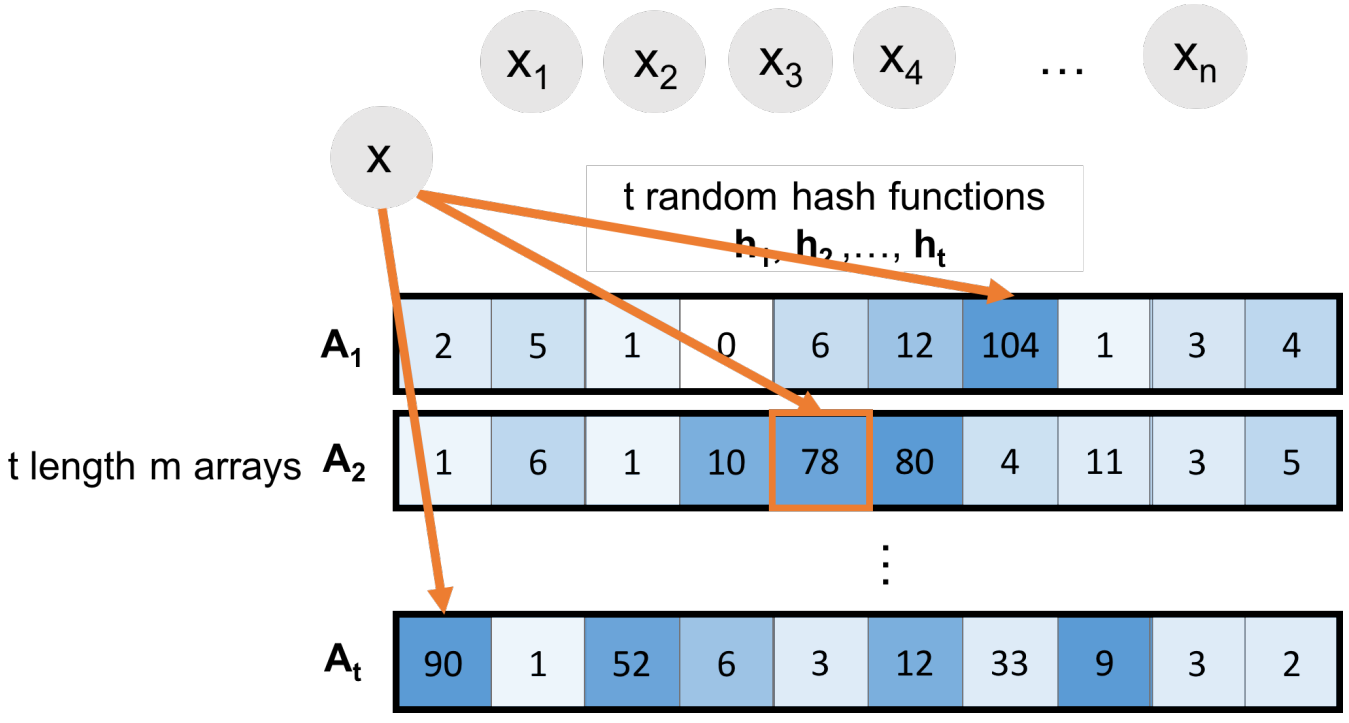
- Estimate $f(x)$ with $\tilde{f}(x) = \min_{i \in [t]} A_i[\mathbf{h}_i(x)]$.
- What is $\Pr[f(x) \leq \tilde{f}(x) \leq f(x) + 2n/m]$?

COUNT-MIN SKETCH ACCURACY



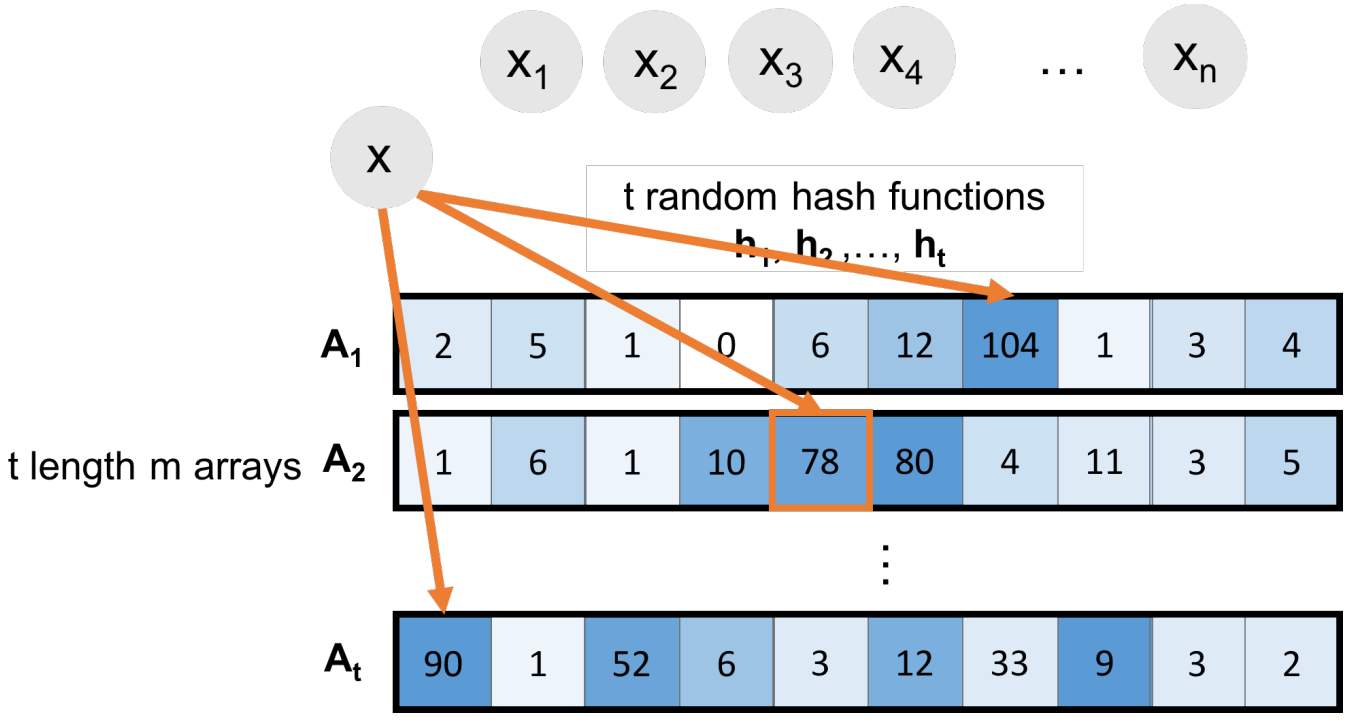
- Estimate $f(x)$ with $\tilde{f}(x) = \min_{i \in [t]} A_i[\mathbf{h}_i(x)]$.
- What is $\Pr[f(x) \leq \tilde{f}(x) \leq f(x) + 2n/m]$? Answer: $\geq 1 - 1/2^t$.

COUNT-MIN SKETCH ACCURACY



- Estimate $f(x)$ with $\tilde{f}(x) = \min_{i \in [t]} A_i[\mathbf{h}_i(x)]$.
- What is $\Pr[f(x) \leq \tilde{f}(x) \leq f(x) + 2n/m]$? Answer: $\geq 1 - 1/2^t$.
- Setting $t = \log(1/\delta)$ ensures probability is at least $1 - \delta$.

COUNT-MIN SKETCH ACCURACY



- Estimate $f(x)$ with $\tilde{f}(x) = \min_{i \in [t]} A_i[\mathbf{h}_i(x)]$.
- What is $\Pr[f(x) \leq \tilde{f}(x) \leq f(x) + 2n/m]$? Answer: $\geq 1 - 1/2^t$.
- Setting $t = \log(1/\delta)$ ensures probability is at least $1 - \delta$.
- Setting $m = 2k/\epsilon$ ensures the error $2n/m$ is $\epsilon n/k$ and this is enough to determine whether we need to output the element.

IDENTIFYING FREQUENT ELEMENTS

Count-min sketch gives an accurate frequency estimate for every item in the stream. But how do we identify the frequent items without having to look up the estimated frequency for $x \in U$?

IDENTIFYING FREQUENT ELEMENTS

Count-min sketch gives an accurate frequency estimate for every item in the stream. But how do we identify the frequent items without having to look up the estimated frequency for $x \in U$?

One approach:

- Maintain a set F while processing the stream:
- At step i:
  - Add ith stream element to F if it's estimated frequency is $\geq i/k$ and it isn't already in F.
  - Remove any element from F whose estimated frequency is $< i/k$.
- Store $O(k)$ items at any time and have all items with frequency $\geq n/k$ stored at the end of the stream.

Questions on Frequent Elements?

HIGH DIMENSIONAL DATA

'Big Data' means not just many data points, but many measurements per data point. I.e., very high dimensional data.

HIGH DIMENSIONAL DATA

'Big Data' means not just many data points, but many measurements per data point. I.e., very high dimensional data.

- Twitter has 321 million active monthly users. Records (tens of) thousands of measurements per user: who they follow, who follows them, when they last visited the site, timestamps for specific interactions, how many tweets they have sent, the text of those tweets, etc.

HIGH DIMENSIONAL DATA

'Big Data' means not just many data points, but many measurements per data point. I.e., very high dimensional data.

- Twitter has 321 million active monthly users. Records (tens of) thousands of measurements per user: who they follow, who follows them, when they last visited the site, timestamps for specific interactions, how many tweets they have sent, the text of those tweets, etc.
- A 3 minute Youtube clip with a resolution of 500×500 pixels at 15 frames/second with 3 color channels is a recording of $\geq$ 2 billion pixel values. Even a 500×500 pixel color image has $750,000$ pixel values.

HIGH DIMENSIONAL DATA

'Big Data' means not just many data points, but many measurements per data point. I.e., very high dimensional data.

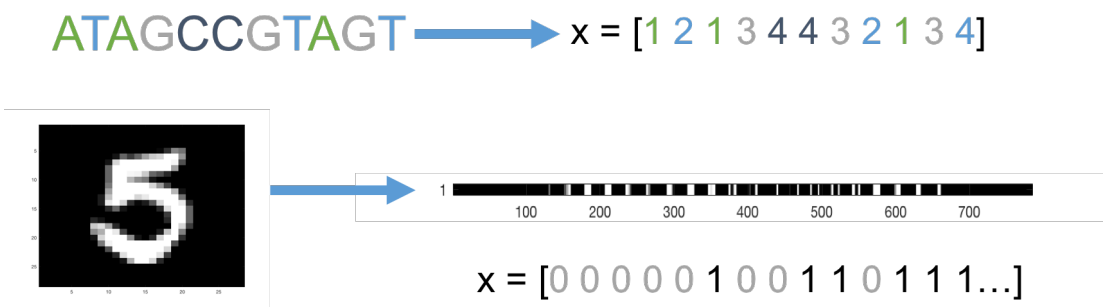
- Twitter has 321 million active monthly users. Records (tens of) thousands of measurements per user: who they follow, who follows them, when they last visited the site, timestamps for specific interactions, how many tweets they have sent, the text of those tweets, etc.
- A 3 minute Youtube clip with a resolution of 500×500 pixels at 15 frames/second with 3 color channels is a recording of $\geq$ 2 billion pixel values. Even a 500×500 pixel color image has $750,000$ pixel values.
- The human genome contains 3 billion+ base pairs. Genetic datasets often contain information on 100s of thousands+ mutations and genetic markers.

DATA AS VECTORS AND MATRICES

In data analysis and machine learning, data points with many attributes are often stored, processed, and interpreted as high dimensional vectors, with real valued entries.

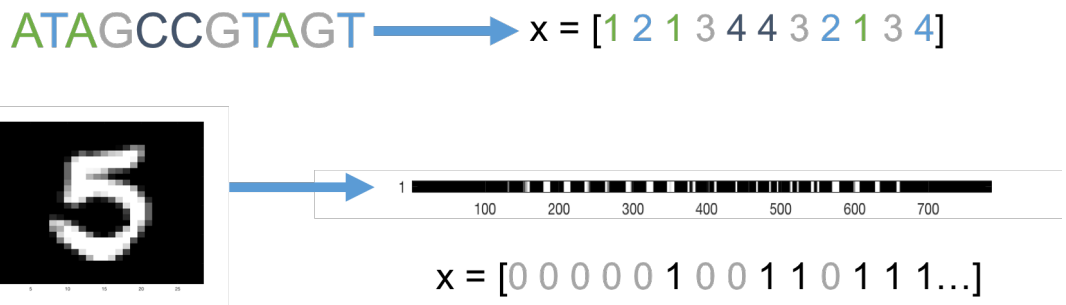
DATA AS VECTORS AND MATRICES

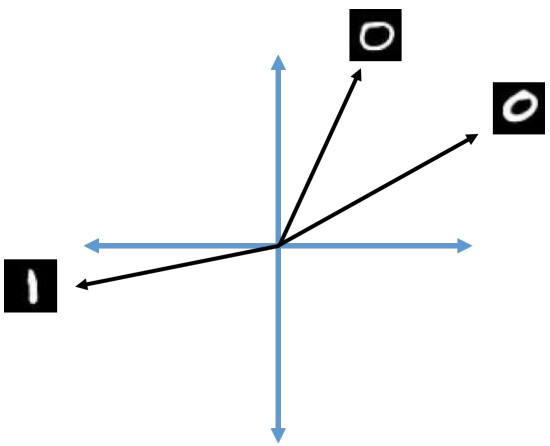
In data analysis and machine learning, data points with many attributes are often stored, processed, and interpreted as high dimensional vectors, with real valued entries.



DATA AS VECTORS AND MATRICES

In data analysis and machine learning, data points with many attributes are often stored, processed, and interpreted as high dimensional vectors, with real valued entries.





Similarities/distances between vectors (e.g., $\langle x, y \rangle$, $\|x - y\|_2$) have meaning for underlying data points.

DATASETS AS VECTORS AND MATRICES

Data points are interpreted as high dimensional vectors, with real valued entries. Data set is interpreted as a matrix.

Data Points: $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d$.

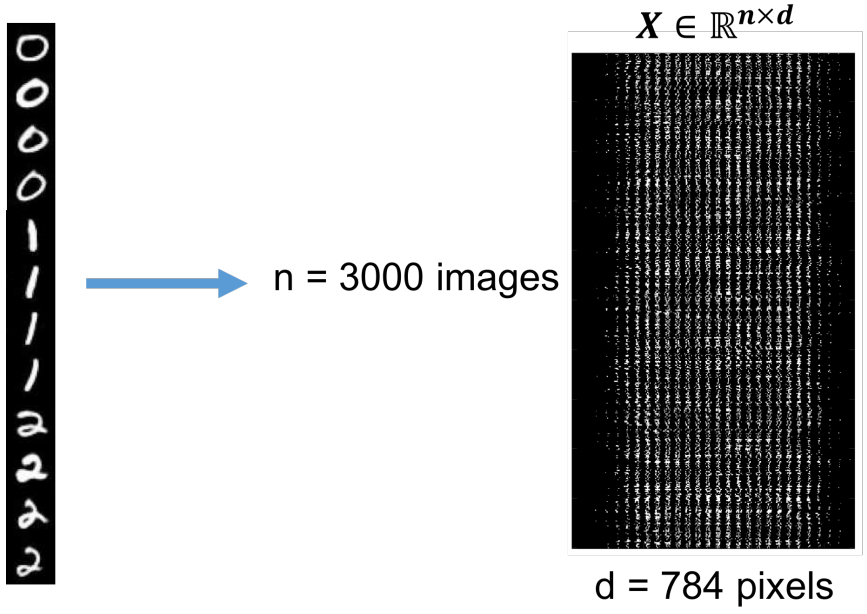
Data Set: $X \in \mathbb{R}^{n \times d}$ with i^{th} rows equal to $\vec{x}_i$.

DATASETS AS VECTORS AND MATRICES

Data points are interpreted as high dimensional vectors, with real valued entries. Data set is interpreted as a matrix.

Data Points: $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d$.

Data Set: $X \in \mathbb{R}^{n \times d}$ with i^{th} rows equal to $\vec{x}_i$.

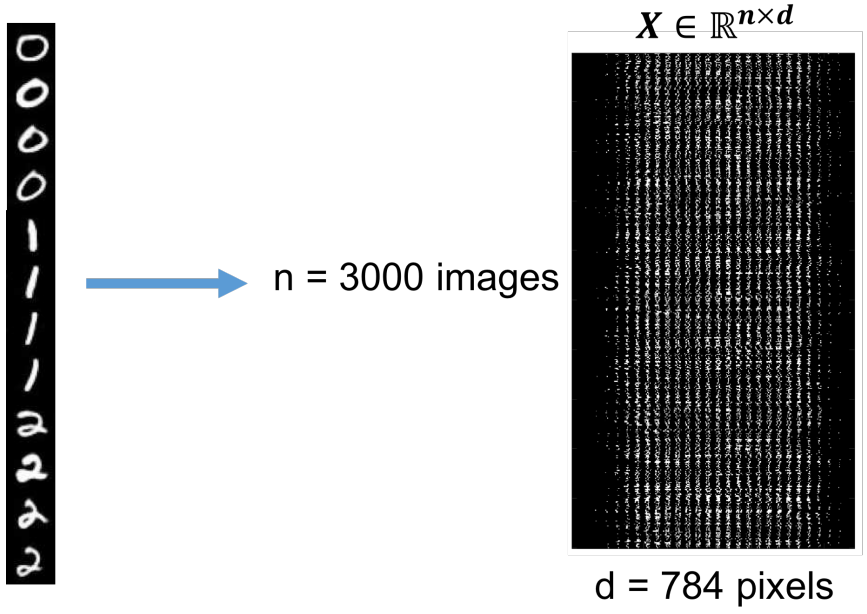


DATASETS AS VECTORS AND MATRICES

Data points are interpreted as high dimensional vectors, with real valued entries. Data set is interpreted as a matrix.

Data Points: $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d$.

Data Set: $X \in \mathbb{R}^{n \times d}$ with i^{th} rows equal to $\vec{x}_i$.



Many data points $n \implies$ tall. Many dimensions $d \implies$ wide.

DIMENSIONALITY REDUCTION

Dimensionality Reduction: Compress data points so that they lie in many fewer dimensions.

DIMENSIONALITY REDUCTION

Dimensionality Reduction: Compress data points so that they lie in many fewer dimensions.

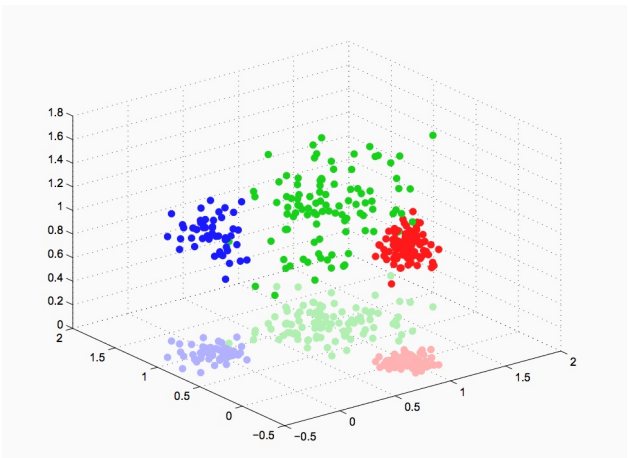
$$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d \rightarrow \tilde{x}_1, \ldots, \tilde{x}_n \in \mathbb{R}^m \text{ for } m \ll d.$$

DIMENSIONALITY REDUCTION

Dimensionality Reduction: Compress data points so that they lie in many fewer dimensions.

$$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d \rightarrow \tilde{x}_1, \ldots, \tilde{x}_n \in \mathbb{R}^m \text{ for } m \ll d.$$

'Lossy compression' that still preserves important information about the relationships between $\vec{x}_1, \ldots, \vec{x}_n$.

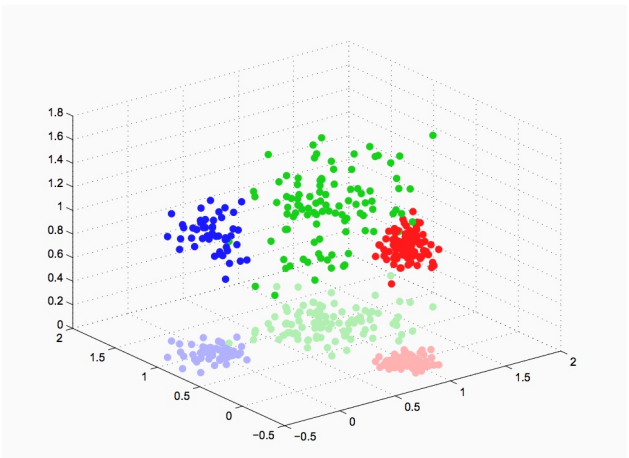


DIMENSIONALITY REDUCTION

Dimensionality Reduction: Compress data points so that they lie in many fewer dimensions.

$$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d \rightarrow \tilde{x}_1, \ldots, \tilde{x}_n \in \mathbb{R}^m \text{ for } m \ll d.$$

'Lossy compression' that still preserves important information about the relationships between $\vec{x}_1, \ldots, \vec{x}_n$.



Generally will not consider directly how well $\tilde{x}_i$ approximates $\vec{x}_i$.

LOW DISTORTION EMBEDDING

Low Distortion Embedding: Given $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$, distance function D, and error parameter $\epsilon \geq 0$, find $\tilde{x}_1, \ldots, \tilde{x}_n \in \mathbb{R}^m$ (where $m \ll d$) and distance function $\tilde{D}$ such that for all $i, j \in [n]$:

$$(1 - \epsilon)D(\vec{x}_i, \vec{x}_j) \leq \tilde{D}(\tilde{x}_i, \tilde{x}_j) \leq (1 + \epsilon)D(\vec{x}_i, \vec{x}_j).$$

LOW DISTORTION EMBEDDING

Low Distortion Embedding: Given $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$, distance function D, and error parameter $\epsilon \geq 0$, find $\tilde{x}_1, \ldots, \tilde{x}_n \in \mathbb{R}^m$ (where $m \ll d$) and distance function $\tilde{D}$ such that for all $i, j \in [n]$:

$$(1-\epsilon)D(\vec{x}_i, \vec{x}_j) \leq \tilde{D}(\tilde{x}_i, \tilde{x}_j) \leq (1+\epsilon)D(\vec{x}_i, \vec{x}_j).$$

We'll focus on the case where D and $\tilde{D}$ are Euclidean distances. I.e., the distance between two vectors x and y is defined as

$$\|\vec{x} - \vec{y}\|_2 = \sqrt{\sum_i (\vec{x}(i) - \vec{y}(i))^2}$$

This is related to the Euclidean norm, $\|\vec{z}\|_2 = \sqrt{\sum_{i=1}^n \vec{z}(i)^2}$.

THE JOHNSON-LINDENSTRAUSS LEMMA

Johnson-Lindenstrauss Lemma: For any set of points $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ and $\epsilon > 0$ there exists a linear map $\boldsymbol{M} : \mathbb{R}^d \to \mathbb{R}^m$ such that $m = O\left(\frac{\log n}{\epsilon^2}\right)$ and letting $\tilde{x}_i = \boldsymbol{M}\vec{x}_i$:

$$\text{For all } i, j: \ (1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2 \leq \|\tilde{x}_i - \tilde{x}_j\|_2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2.$$

Further, if $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ has each entry chosen independently from $\mathcal{N}(0, 1/m)$, it satisfies the guarantee with high probability.

THE JOHNSON-LINDENSTRAUSS LEMMA

Johnson-Lindenstrauss Lemma: For any set of points $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ and $\epsilon > 0$ there exists a linear map $\boldsymbol{M} : \mathbb{R}^d \to \mathbb{R}^m$ such that $m = O\left(\frac{\log n}{\epsilon^2}\right)$ and letting $\tilde{x}_i = \boldsymbol{M}\vec{x}_i$:

$$\text{For all } i, j: \ (1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2 \leq \|\tilde{x}_i - \tilde{x}_j\|_2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2.$$

Further, if $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ has each entry chosen independently from $\mathcal{N}(0, 1/m)$, it satisfies the guarantee with high probability.

For $d = 1$ trillion, $\epsilon = .05$, and $n = 100,000$, $m \approx 6600$.

THE JOHNSON-LINDENSTRAUSS LEMMA

Johnson-Lindenstrauss Lemma: For any set of points $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ and $\epsilon > 0$ there exists a linear map $\boldsymbol{M} : \mathbb{R}^d \to \mathbb{R}^m$ such that $m = O\left(\frac{\log n}{\epsilon^2}\right)$ and letting $\tilde{x}_i = \boldsymbol{M}\vec{x}_i$:

$$\text{For all } i, j: \ (1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2 \leq \|\tilde{x}_i - \tilde{x}_j\|_2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2.$$

Further, if $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ has each entry chosen independently from $\mathcal{N}(0, 1/m)$, it satisfies the guarantee with high probability.

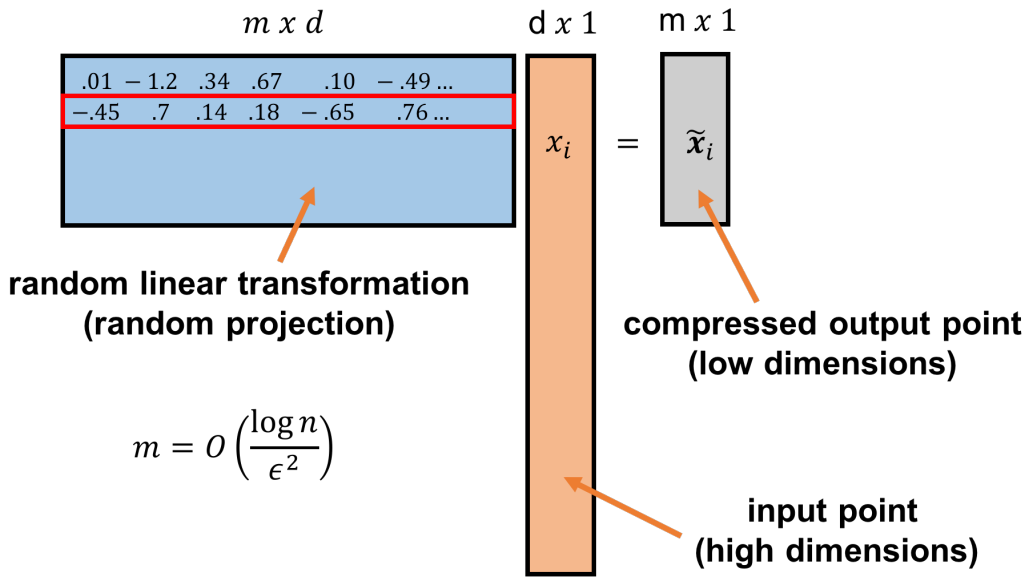
For $d = 1$ trillion, $\epsilon = .05$, and $n = 100,000$, $m \approx 6600$.

Very surprising! Powerful result with a simple construction: applying a random linear transformation to a set of points preserves distances between all those points with high probability.

RANDOM PROJECTION

For any $\vec{x}_1, \ldots, \vec{x}_n$ and $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ with each entry chosen independently from $\mathcal{N}(0, 1/m)$, with high probability, letting $\tilde{\mathbf{x}}_i = \boldsymbol{M}\vec{x}_i$:

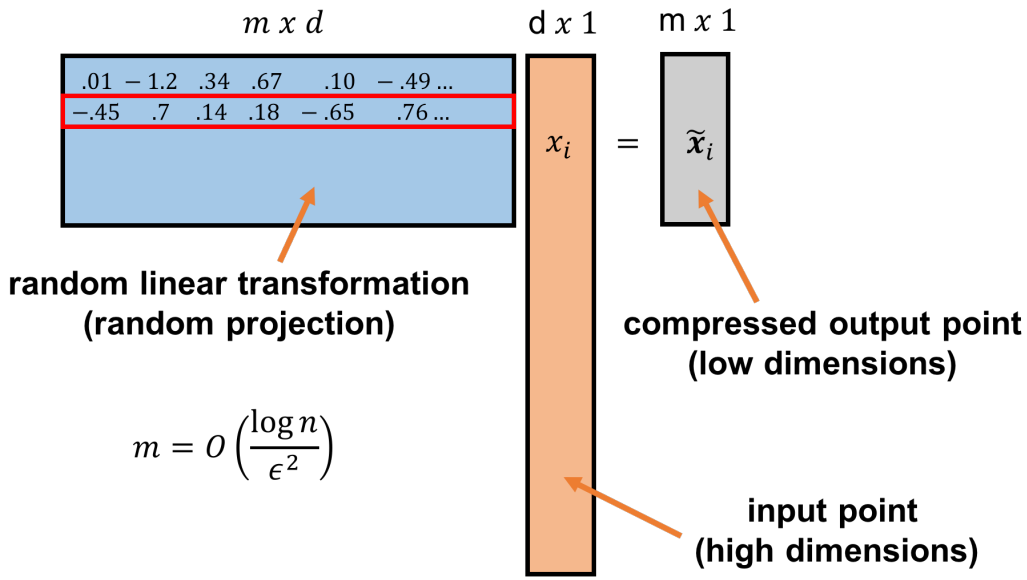
$$\text{For all } i, j: \ (1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2 \leq \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|_2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2.$$



RANDOM PROJECTION

For any $\vec{x}_1, \ldots, \vec{x}_n$ and $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ with each entry chosen independently from $\mathcal{N}(0, 1/m)$, with high probability, letting $\tilde{\mathbf{x}}_i = \boldsymbol{M}\vec{x}_i$:

$$\text{For all } i, j: \ (1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2 \leq \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|_2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2.$$

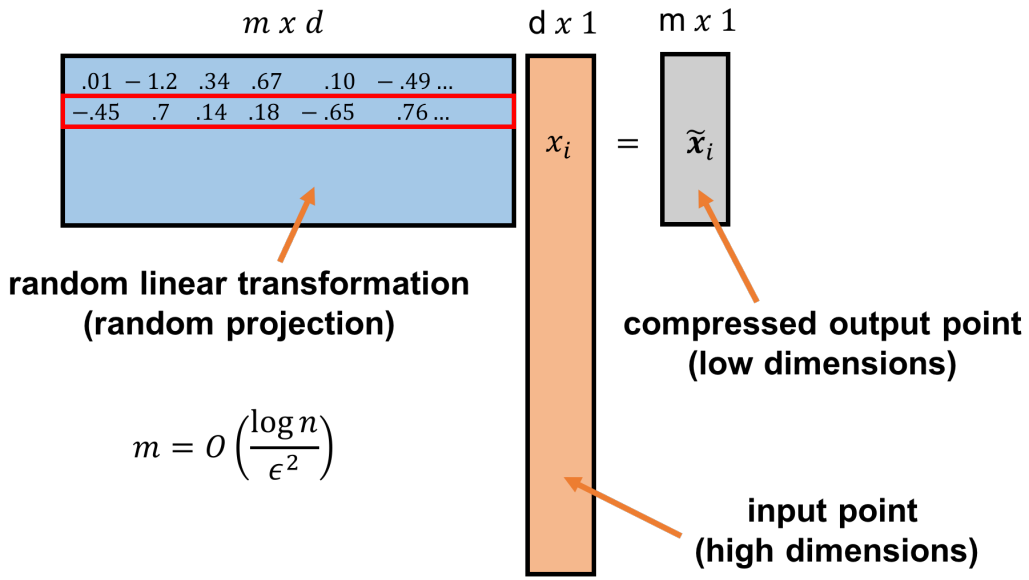


- $\boldsymbol{M}$ is known as a random projection. It is a random linear function, mapping length d vectors to length m vectors.

RANDOM PROJECTION

For any $\vec{x}_1, \ldots, \vec{x}_n$ and $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ with each entry chosen independently from $\mathcal{N}(0, 1/m)$, with high probability, letting $\tilde{\mathbf{x}}_i = \boldsymbol{M}\vec{x}_i$:

$$\text{For all } i, j: \ (1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2 \leq \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|_2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2.$$



- $\boldsymbol{M}$ is known as a random projection. It is a random linear function, mapping length d vectors to length m vectors.
- $\boldsymbol{M}$ is data oblivious. Stark contrast to methods like PCA.

ALGORITHMIC CONSIDERATIONS

- Alternative constructions: ± 1 entries, sparse (most entries 0), Fourier structured, etc. $\implies$ efficient computation of $\tilde{\mathbf{x}}_i = \boldsymbol{M}\vec{x}_i$.

ALGORITHMIC CONSIDERATIONS

- Alternative constructions: ± 1 entries, sparse (most entries 0), Fourier structured, etc. $\implies$ efficient computation of $\tilde{\mathbf{x}}_i = \boldsymbol{M}\vec{x}_i$.
- Data oblivious property means that once $\boldsymbol{M}$ is chosen, $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$ can be computed in a stream with little memory.
- Storage is just $O(nm)$ rather than $O(nd)$.

ALGORITHMIC CONSIDERATIONS

- Alternative constructions: ± 1 entries, sparse (most entries 0), Fourier structured, etc. $\implies$ efficient computation of $\tilde{\mathbf{x}}_i = \boldsymbol{M}\vec{x}_i$.
- Data oblivious property means that once $\boldsymbol{M}$ is chosen, $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$ can be computed in a stream with little memory.
- Storage is just $O(nm)$ rather than $O(nd)$.
- Compression can be performed in parallel on different servers.

ALGORITHMIC CONSIDERATIONS

- Alternative constructions: ± 1 entries, sparse (most entries 0), Fourier structured, etc. $\implies$ efficient computation of $\tilde{\mathbf{x}}_i = \boldsymbol{M}\vec{x}_i$.
- Data oblivious property means that once $\boldsymbol{M}$ is chosen, $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$ can be computed in a stream with little memory.
- Storage is just $O(nm)$ rather than $O(nd)$.
- Compression can be performed in parallel on different servers.
- When new data points are added, can be easily compressed, without updating existing points.

THE JOHNSON-LINDENSTRAUSS LEMMA

Johnson-Lindenstrauss Lemma: For any set of points $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ and $\epsilon > 0$ there exists a linear map $\boldsymbol{M} : \mathbb{R}^d \to \mathbb{R}^m$ such that $m = O\left(\frac{\log n}{\epsilon^2}\right)$ and letting $\tilde{x}_i = \boldsymbol{M}\vec{x}_i$:

$$\text{For all } i, j: \ (1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2 \leq \|\tilde{x}_i - \tilde{x}_j\|_2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2.$$

Further, if $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ has each entry chosen independently from $\mathcal{N}(0, 1/m)$, it satisfies the guarantee with high probability.

DISTRIBUTIONAL JL

The Johnson-Lindenstrauss Lemma is a direct consequence of:

Distributional JL Lemma: Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ have each entry chosen i.i.d. as $\mathcal{N}(0, 1/m)$. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $\vec{y} \in \mathbb{R}^d$, with probability $\geq 1 - \delta$

$$(1 - \epsilon)\|\vec{y}\|_2 \leq \|\boldsymbol{M}\vec{y}\|_2 \leq (1 + \epsilon)\|\vec{y}\|_2$$

$\boldsymbol{M} \in \mathbb{R}^{m \times d}$: random projection matrix. d: original dimension. m: compressed dimension, ϵ: embedding error, δ: embedding failure prob.

DISTRIBUTIONAL JL

The Johnson-Lindenstrauss Lemma is a direct consequence of:

Distributional JL Lemma: Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ have each entry chosen i.i.d. as $\mathcal{N}(0, 1/m)$. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $\vec{y} \in \mathbb{R}^d$, with probability $\geq 1 - \delta$

$$(1-\epsilon)\|\vec{y}\|_2 \leq \|\boldsymbol{M}\vec{y}\|_2 \leq (1+\epsilon)\|\vec{y}\|_2$$

I.e., applying a random matrix $\boldsymbol{M}$ to any vector $\vec{y}$ preserves the norm with high probability. Like a low-distortion embedding, but for the length of a compressed vector rather than distances between vectors.

$\boldsymbol{M} \in \mathbb{R}^{m \times d}$: random projection matrix. d: original dimension. m: compressed dimension, ϵ: embedding error, δ: embedding failure prob.

DISTRIBUTIONAL JL $\Longrightarrow$ JL

Distributional JL Lemma $\Longrightarrow$ JL Lemma: Distributional JL show that a random projection $\boldsymbol{M}$ preserves the norm of any y. The main JL Lemma says that $\boldsymbol{M}$ preserves distances between vectors.

DISTRIBUTIONAL JL $\Longrightarrow$ JL

Distributional JL Lemma $\Longrightarrow$ JL Lemma: Distributional JL show that a random projection $\boldsymbol{M}$ preserves the norm of any y. The main JL Lemma says that $\boldsymbol{M}$ preserves distances between vectors. Since $\boldsymbol{M}$ is linear these are the same thing!

DISTRIBUTIONAL JL $\implies$ JL

Distributional JL Lemma $\implies$ JL Lemma: Distributional JL show that a random projection $\boldsymbol{M}$ preserves the norm of any y. The main JL Lemma says that $\boldsymbol{M}$ preserves distances between vectors. Since $\boldsymbol{M}$ is linear these are the same thing!

Proof: Given $x_1, \ldots, x_n$, define $\binom{n}{2}$ vectors y_{ij} where $y_{ij} = x_i - x_j$.

DISTRIBUTIONAL JL $\implies$ JL

Distributional JL Lemma $\implies$ JL Lemma: Distributional JL show that a random projection $\boldsymbol{M}$ preserves the norm of any y. The main JL Lemma says that $\boldsymbol{M}$ preserves distances between vectors. Since $\boldsymbol{M}$ is linear these are the same thing!

Proof: Given $x_1, \ldots, x_n$, define $\binom{n}{2}$ vectors y_{ij} where $y_{ij} = x_i - x_j$.

- If we choose $\boldsymbol{M}$ with $m = O\left(\epsilon^{-2} \log 1/\delta'\right)$, for each y_{ij} with probability at least $1 - \delta'$ we have:

$$(1 - \epsilon)\|y_{ij}\|_2 \leq \|\boldsymbol{M} y_{ij}\|_2 \leq (1 + \epsilon)\|y_{ij}\|_2$$

DISTRIBUTIONAL JL $\implies$ JL

Distributional JL Lemma $\implies$ JL Lemma: Distributional JL show that a random projection $\boldsymbol{M}$ preserves the norm of any y. The main JL Lemma says that $\boldsymbol{M}$ preserves distances between vectors. Since $\boldsymbol{M}$ is linear these are the same thing!

Proof: Given $x_1, \ldots, x_n$, define $\binom{n}{2}$ vectors y_{ij} where $y_{ij} = x_i - x_j$.

- If we choose $\boldsymbol{M}$ with $m = O\left(\epsilon^{-2} \log 1/\delta'\right)$, for each y_{ij} with probability at least $1 - \delta'$ we have:

$$(1 - \epsilon)\|y_{ij}\|_2 \leq \|\boldsymbol{M} y_{ij}\|_2 \leq (1 + \epsilon)\|y_{ij}\|_2$$

DISTRIBUTIONAL JL $\implies$ JL

Distributional JL Lemma $\implies$ JL Lemma: Distributional JL show that a random projection $\boldsymbol{M}$ preserves the norm of any y. The main JL Lemma says that $\boldsymbol{M}$ preserves distances between vectors. Since $\boldsymbol{M}$ is linear these are the same thing!

Proof: Given $x_1, \ldots, x_n$, define $\binom{n}{2}$ vectors y_{ij} where $y_{ij} = x_i - x_j$.

- If we choose $\boldsymbol{M}$ with $m = O\left(\epsilon^{-2} \log 1/\delta'\right)$, for each y_{ij} with probability at least $1 - \delta'$ we have:

$$(1-\epsilon)\|x_i - x_j\|_2 \leq \|\boldsymbol{M}(x_i - x_j)\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2$$

DISTRIBUTIONAL JL $\implies$ JL

Distributional JL Lemma $\implies$ JL Lemma: Distributional JL show that a random projection $\boldsymbol{M}$ preserves the norm of any y. The main JL Lemma says that $\boldsymbol{M}$ preserves distances between vectors. Since $\boldsymbol{M}$ is linear these are the same thing!

Proof: Given $x_1, \ldots, x_n$, define $\binom{n}{2}$ vectors y_{ij} where $y_{ij} = x_i - x_j$.

- If we choose $\boldsymbol{M}$ with $m = O\left(\epsilon^{-2} \log 1/\delta'\right)$, for each y_{ij} with probability at least $1 - \delta'$ we have:

$$(1-\epsilon)\|x_i - x_j\|_2 \leq \|\boldsymbol{M}x_i - \boldsymbol{M}x_j\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2$$

DISTRIBUTIONAL JL $\implies$ JL

Distributional JL Lemma $\implies$ JL Lemma: Distributional JL show that a random projection $\boldsymbol{M}$ preserves the norm of any y. The main JL Lemma says that $\boldsymbol{M}$ preserves distances between vectors. Since $\boldsymbol{M}$ is linear these are the same thing!

Proof: Given $x_1, \ldots, x_n$, define $\binom{n}{2}$ vectors y_{ij} where $y_{ij} = x_i - x_j$.

- If we choose $\boldsymbol{M}$ with $m = O\left(\epsilon^{-2} \log 1/\delta'\right)$, for each y_{ij} with probability at least $1 - \delta'$ we have:

$$(1-\epsilon)\|x_i - x_j\|_2 \leq \|\boldsymbol{M}x_i - \boldsymbol{M}x_j\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2$$

DISTRIBUTIONAL JL $\implies$ JL

Distributional JL Lemma $\implies$ JL Lemma: Distributional JL show that a random projection $\boldsymbol{M}$ preserves the norm of any y. The main JL Lemma says that $\boldsymbol{M}$ preserves distances between vectors. Since $\boldsymbol{M}$ is linear these are the same thing!

Proof: Given $x_1, \ldots, x_n$, define $\binom{n}{2}$ vectors y_{ij} where $y_{ij} = x_i - x_j$.

- If we choose $\boldsymbol{M}$ with $m = O\left(\epsilon^{-2} \log 1/\delta'\right)$, for each y_{ij} with probability at least $1 - \delta'$ we have:

$$(1-\epsilon)\|x_i - x_j\|_2 \leq \|\boldsymbol{M}x_i - \boldsymbol{M}x_j\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2$$

- Union Bound: Every distance preserved with probability $1 - \binom{n}{2} \cdot \delta'$.

DISTRIBUTIONAL JL $\implies$ JL

Distributional JL Lemma $\implies$ JL Lemma: Distributional JL show that a random projection $\boldsymbol{M}$ preserves the norm of any y. The main JL Lemma says that $\boldsymbol{M}$ preserves distances between vectors. Since $\boldsymbol{M}$ is linear these are the same thing!

Proof: Given $x_1, \ldots, x_n$, define $\binom{n}{2}$ vectors y_{ij} where $y_{ij} = x_i - x_j$.

- If we choose $\boldsymbol{M}$ with $m = O\left(\epsilon^{-2} \log 1/\delta'\right)$, for each y_{ij} with probability at least $1 - \delta'$ we have:

$$(1-\epsilon)\|x_i - x_j\|_2 \leq \|\boldsymbol{M}x_i - \boldsymbol{M}x_j\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2$$

- Union Bound: Every distance preserved with probability $1 - \binom{n}{2} \cdot \delta'$.
- Setting $\delta' = \delta/\binom{n}{2}$ ensures all distances preserved with probability $1 - \delta$ and

$$m = O\left(\frac{\log(1/\delta')}{\epsilon^2}\right) = O\left(\frac{\log(\binom{n}{2}/\delta)}{\epsilon^2}\right) = O\left(\frac{\log(n/\delta)}{\epsilon^2}\right)$$

Distributional JL proof (part 1 of 3)

Distributional JL Lemma: Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ have independent $\mathcal{N}(0, 1/m)$ entries. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $y \in \mathbb{R}^d$, with probability at least $1 - \delta$

$$(1 - \epsilon)\|y\|_2 \leq \|\boldsymbol{M}y\|_2 \leq (1 + \epsilon)\|y\|_2.$$

Distributional JL proof (part 1 of 3)

Distributional JL Lemma: Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ have independent $\mathcal{N}(0, 1/m)$ entries. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $y \in \mathbb{R}^d$, with probability at least $1 - \delta$

$$(1 - \epsilon)\|y\|_2 \leq \|\boldsymbol{M}y\|_2 \leq (1 + \epsilon)\|y\|_2.$$

- Let $\tilde{y} = \boldsymbol{M}y$ and $\boldsymbol{M}_j$ be the j^{th} row of $\boldsymbol{M}$

Distributional JL proof (part 1 of 3)

Distributional JL Lemma: Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ have independent $\mathcal{N}(0, 1/m)$ entries. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $y \in \mathbb{R}^d$, with probability at least $1 - \delta$

$$(1 - \epsilon)\|y\|_2 \leq \|\boldsymbol{M}y\|_2 \leq (1 + \epsilon)\|y\|_2.$$

- Let $\tilde{y} = \boldsymbol{M}y$ and $\boldsymbol{M}_j$ be the j^{th} row of $\boldsymbol{M}$
- For any j, $\tilde{y}_j = \langle \boldsymbol{M}_j, y \rangle$

Distributional JL proof (part 1 of 3)

Distributional JL Lemma: Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ have independent $\mathcal{N}(0, 1/m)$ entries. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $y \in \mathbb{R}^d$, with probability at least $1 - \delta$

$$(1 - \epsilon)\|y\|_2 \leq \|\boldsymbol{M}y\|_2 \leq (1 + \epsilon)\|y\|_2.$$

- Let $\tilde{y} = \boldsymbol{M}y$ and $\boldsymbol{M}_j$ be the j^{th} row of $\boldsymbol{M}$
- For any j, $\tilde{y}_j = \langle \boldsymbol{M}_j, y \rangle = \sum_{i=1}^{d} \mathbf{g}_i \cdot y_i$ where $\mathbf{g}_i \sim \mathcal{N}(0, 1/m)$.

Distributional JL proof (part 1 of 3)

Distributional JL Lemma: Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ have independent $\mathcal{N}(0, 1/m)$ entries. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $y \in \mathbb{R}^d$, with probability at least $1 - \delta$

$$(1 - \epsilon)\|y\|_2 \leq \|\boldsymbol{M}y\|_2 \leq (1 + \epsilon)\|y\|_2.$$

- Let $\tilde{y} = \boldsymbol{M}y$ and $\boldsymbol{M}_j$ be the j^{th} row of $\boldsymbol{M}$
- For any j, $\tilde{y}_j = \langle \boldsymbol{M}_j, y \rangle = \sum_{i=1}^{d} \mathbf{g}_i \cdot y_i$ where $\mathbf{g}_i \sim \mathcal{N}(0, 1/m)$.
- By linearity of expectation:

$$\mathbb{E}[\tilde{y}_j] = \sum_{i=1}^{d} \mathbb{E}[\mathbf{g}_i] \cdot y_i = 0 \ .$$

Distributional JL proof (part 1 of 3)

Distributional JL Lemma: Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ have independent $\mathcal{N}(0, 1/m)$ entries. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $y \in \mathbb{R}^d$, with probability at least $1 - \delta$

$$(1 - \epsilon)\|y\|_2 \leq \|\boldsymbol{M}y\|_2 \leq (1 + \epsilon)\|y\|_2.$$

- Let $\tilde{y} = \boldsymbol{M}y$ and $\boldsymbol{M}_j$ be the j^{th} row of $\boldsymbol{M}$
- For any j, $\tilde{y}_j = \langle \boldsymbol{M}_j, y \rangle = \sum_{i=1}^{d} \mathbf{g}_i \cdot y_i$ where $\mathbf{g}_i \sim \mathcal{N}(0, 1/m)$.
- By linearity of expectation:

$$\mathbb{E}[\tilde{y}_j] = \sum_{i=1}^{d} \mathbb{E}[\mathbf{g}_i] \cdot y_i = 0 \ .$$

- Since $\mathbb{E}[\tilde{y}_j] = 0$ we have $\mathbb{E}[\tilde{y}_j^2] = \mathrm{Var}[\tilde{y}_j]$. Then, by linearity of variance:

$$\mathbb{E}[\tilde{y}_j^2] = \mathrm{Var}[\tilde{y}_j] = \sum_{i=1}^{d} \mathrm{Var}[\mathbf{g}_i \cdot y_i] = \sum_{i} y_i^2/m = \|y\|_2^2/m \ .$$

Distributional JL proof (part 1 of 3)

Distributional JL Lemma: Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ have independent $\mathcal{N}(0, 1/m)$ entries. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $y \in \mathbb{R}^d$, with probability at least $1 - \delta$

$$(1-\epsilon)\|y\|_2 \leq \|\boldsymbol{M}y\|_2 \leq (1+\epsilon)\|y\|_2.$$

- Let $\tilde{y} = \boldsymbol{M}y$ and $\boldsymbol{M}_j$ be the j^{th} row of $\boldsymbol{M}$
- For any j, $\tilde{y}_j = \langle \boldsymbol{M}_j, y \rangle = \sum_{i=1}^{d} \mathbf{g}_i \cdot y_i$ where $\mathbf{g}_i \sim \mathcal{N}(0, 1/m)$.
- By linearity of expectation:

$$\mathbb{E}[\tilde{y}_j] = \sum_{i=1}^{d} \mathbb{E}[\mathbf{g}_i] \cdot y_i = 0 \ .$$

- Since $\mathbb{E}[\tilde{y}_j] = 0$ we have $\mathbb{E}[\tilde{y}_j^2] = \mathrm{Var}[\tilde{y}_j]$. Then, by linearity of variance:

$$\mathbb{E}[\tilde{y}_j^2] = \mathrm{Var}[\tilde{y}_j] = \sum_{i=1}^{d} \mathrm{Var}[\mathbf{g}_i \cdot y_i] = \sum_{i} y_i^2/m = \|y\|_2^2/m \ .$$

- Hence $\mathbb{E}[\|\tilde{y}\|_2^2] = \mathbb{E}[\sum_j \tilde{y}_j^2] = \|y\|_2^2$.

Distributional JL proof (part 1 of 3)

Distributional JL Lemma: Let $\boldsymbol{M} \in \mathbb{R}^{m \times d}$ have independent $\mathcal{N}(0, 1/m)$ entries. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $y \in \mathbb{R}^d$, with probability at least $1 - \delta$

$$(1 - \epsilon)\|y\|_2 \leq \|\boldsymbol{M}y\|_2 \leq (1 + \epsilon)\|y\|_2.$$

- Let $\tilde{y} = \boldsymbol{M}y$ and $\boldsymbol{M}_j$ be the j^{th} row of $\boldsymbol{M}$
- For any j, $\tilde{y}_j = \langle \boldsymbol{M}_j, y \rangle = \sum_{i=1}^{d} \mathbf{g}_i \cdot y_i$ where $\mathbf{g}_i \sim \mathcal{N}(0, 1/m)$.
- By linearity of expectation:

$$\mathbb{E}[\tilde{y}_j] = \sum_{i=1}^{d} \mathbb{E}[\mathbf{g}_i] \cdot y_i = 0 \ .$$

- Since $\mathbb{E}[\tilde{y}_j] = 0$ we have $\mathbb{E}[\tilde{y}_j^2] = \mathrm{Var}[\tilde{y}_j]$. Then, by linearity of variance:

$$\mathbb{E}[\tilde{y}_j^2] = \mathrm{Var}[\tilde{y}_j] = \sum_{i=1}^{d} \mathrm{Var}[\mathbf{g}_i \cdot y_i] = \sum_{i} y_i^2/m = \|y\|_2^2/m \ .$$

- Hence $\mathbb{E}[\|\tilde{y}\|_2^2] = \mathbb{E}[\sum_j \tilde{y}_j^2] = \|y\|_2^2$. Remains to show $\|\tilde{y}\|_2^2$ is concentrated.