Algorithms for Big Data (Fall 25)

Lecture 9

Applications of Sketching and Dimensionality Reduction

Ali Vakilian (vakilian@vt.edu)

Virginia Tech

Algorithms for Big Data
Fall 2025

Sparse Recovery

Sparsity is an important theme in optimization/algorithms/modeling

- Data is often explicitly sparse.

  Examples: graphs, matrices, vectors, documents (as word vectors)

- Data is often *implicitly* sparse: in a different representation the data is explicitly sparse.

  Examples: signals/images, topics, …

Algorithmic advantage

- To improve performance (speed, quality, memory, …)
- Find sparse representation to reveal information about data

  Examples: topics in documents, frequencies in Fourier analysis

Sparse Recovery

Problem. Given a vector/signal $x \in \mathbb{R}^n$, find a sparse vector z approximating x.

More formally, given $x \in \mathbb{R}^n$ and integer $k \geq 1$, find z s.t. z has at most k non-zeros ($\|z\|_0 \leq k$) s.t. $\|z - x\|_p$ is minimized for some $p \geq 1$.

What is the optimal offline solution?

How to solve in strict turnstile streaming for $p = 2$ using $\tilde{O}(k)$ space?

Sparse Recovery under ℓ_2 norm

Problem. Minimize $\mathrm{err}_2^k(x) = \min_{z:\|z\|_0 \le k} \|z - x\|_2$.

Interesting when $\mathrm{err}_2^k(x) \ll \|x\|_2$

- $\mathrm{err}_2^k(x) = 0$ iff $\|x\|_0 \le k$; so, related to distinct element problem.

Sparse Recovery under ℓ_2 norm

Problem. Minimize $\mathrm{err}_2^k(x) = \min_{z:\|z\|_0 \le k} \|z - x\|_2$.

Theorem. There is a linear sketch of size $O(\frac{k}{\varepsilon^2}\ \mathrm{polylog}(n))$ that returns z such that $\|z\|_0 \le k$, and with high probability,

$$\|x - z\|_2 \le (1 + \varepsilon) \cdot \mathrm{err}_2^k(x)$$

- Space is proportional to desired output sparsity which is typically $\ll n$.
- If x is k-sparse vector, it will be exactly reconstructed.
- The solution is based on CountSketch

Sparse Recovery under ℓ_2 norm

Problem. Minimize $\mathrm{err}_2^k(x) = \min_{z:\|z\|_0 \le k} \|z - x\|_2$.

Sparse Recovery (via CountSkecth):

let CS be a CountSketch with $w = \frac{3k}{\varepsilon^2}$ and $d = \Omega(\log n)$

% during stream

process the stream and update CS

% after stream

compute all $\tilde{x}_i$
output k coordinates with largest estimates

Sparse Recovery under ℓ_2 norm

Problem. Minimize $\mathrm{err}_2^k(x) = \min_{z:\|z\|_0 \le k} \|z - x\|_2$.

Theorem. There is a linear sketch of size $\frac{k}{\varepsilon^2}$ polylog(n) that returns z such that $\|z\|_0 \le k$, and with high probability, $\|x - z\|_2 \le (1 + \varepsilon) \cdot \mathrm{err}_2^k(x)$

Lemma I. CountSketch w/ $w = \frac{3k}{\varepsilon^2}$ and $d = O(\log n)$ w.h.p. guarantees that

$$\forall i \in [n], \qquad |\tilde{x}_i - x_i| \le \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)$$

Lemma II. Let $x, y \in \mathbb{R}^n$ s.t. $\|x - y\|_\infty \le \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)$. Then, $\|x - z\|_2 \le (1 + \varepsilon) \cdot \mathrm{err}_2^k(x)$, where z is as follows: $z_i = y_i$ for k largest absolute indices of y, and $z_i = 0$ for the rest.

Sparse Recovery under ℓ_2 norm

Problem. Minimize $\mathrm{err}_2^k(x) = \min_{z:\|z\|_0 \le k} \|z - x\|_2$.

Lemma I. CountSketch w/ $w = \frac{3k}{\varepsilon^2}$ and $d = O(\log n)$ w.h.p. guarantees that

$$\forall i \in [n], \quad |\tilde{x}_i - x_i| \le \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)$$

Sparse Recovery under ℓ_2 norm

Problem. Minimize $\mathrm{err}_2^k(x) = \min_{z:\|z\|_0 \le k} \|z - x\|_2$.

Lemma II. Let $x, y \in \mathbb{R}^n$ s.t. $\|x - y\|_\infty \le \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)$. Then, $\|x - z\|_2 \le (1 + \varepsilon) \cdot \mathrm{err}_2^k(x)$, where z is as follows: $z_i = y_i$ for k largest absolute indices of y, and $z_i = 0$ for the rest.

"Stronger" Guarantee for CountSketch

Lemma I. CountSketch w/ $w = \frac{3k}{\varepsilon^2}$ and $d = O(\log n)$ w.h.p. guarantees that

$$\forall i \in [n], \qquad |\tilde{x}_i - x_i| \le \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)$$

Analysis has two parts:

- First, similarly to the earlier analysis of CS, is to bound the variance and apply Chernoff but this time for all items other than k largest coordinates.
- Second, we show that there is no collision with k largest coordinates.

CountSketch Analysis

- Consider an item i and fix a row ℓ.
- Define $Z_\ell = g_\ell(i)C[\ell, h_\ell(i)]$ the value of counter in row ℓ that i is hashed to.

For $j \in [n]$ let Y_j be the indicator r.v. that is 1 if $h_\ell(i) = h_\ell(j)$; i.e., i and j collide in h_ℓ

$\mathbb{E}[Y_j] = \mathbb{E}\left[Y_j^2\right] = 1/w$ from pairwise independence of h_ℓ

$Z_\ell = g_\ell(i)C[\ell, h_\ell(i)] = g_\ell(i)f_i + \sum_{j \neq i} g_\ell(i)f_jY_j$

$\mathbb{E}[Z_\ell] = f_i + \sum_{j \neq i} \mathbb{E}[g_\ell(i)g_\ell(j)Y_j] \cdot f_j$

$= f_i$ // pairwise independence of g_ℓ

Since $\mathbb{E}\left[g_\ell(i)g_\ell(j)Y_j\right] = \mathbb{E}[g_\ell(i)g_\ell(j)]\,\mathbb{E}\left[Y_j\right] = 0$

From Lecture 8

CountSketch Analysis: Variance

- Define $Z_\ell = g_\ell(i)C[\ell, h_\ell(i)]$ the value of counter in row ℓ that i is hashed to.

For $j \in [n]$ let Y_j be the indicator r.v. that is 1 if $h_\ell(i) = h_\ell(j)$; i.e., i and j collide in h_ℓ

$\mathbb{E}[Y_j] = \mathbb{E}\left[Y_j^2\right] = 1/w$ from pairwise independence of h_ℓ

$$
\begin{aligned}
&= \mathbb{E}\left[(Z_\ell - f_i)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{j\neq i} g_\ell(i) g_\ell(j) Y_j f_j\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j\neq i} g_\ell(i)^2 g_\ell(j)^2 Y_j^2 f_j^2 + \sum_{j,j'\neq i} g_\ell(i)^2 g_\ell(j) g_\ell(j') Y_j Y_{j'} f_j f_{j'}\right] \\
&= \sum_{j\neq i} f_j^2\, \mathbb{E}[Y_j^2] \\
&\leq \|f\|_2^2/w
\end{aligned}
$$

Using Chebyshev, $\Pr[|Z_\ell - f_i| \geq \varepsilon\|f\|_2] \leq \frac{\mathrm{Var}(Z_\ell)}{\varepsilon^2\|f\|_2^2} \leq \frac{1}{\varepsilon^2 w} \leq 1/3$

From Lecture 8

Refining Analysis

$T_{\text{big}} = \{j \mid j \text{ is one of the } k \text{ largest coordinates (in absolute value)}\}$

$T_{\text{small}} = [n] \setminus T_{\text{big}}$

In particular, $\sum_{j \in T_{\text{small}}} x_j^2 = \left(\text{err}_2^k(x)\right)^2$

Lemma. $\Pr\left[|Z_\ell - x_i| \geq \frac{\varepsilon}{\sqrt{k}} \cdot \text{err}_2^k(x)\right] \leq 2/5.$

Refining Analysis (contd.)

$Z_\ell = g_\ell(i)C[\ell, h_\ell(i)] = x_i + \sum_{j \in T_{\text{big}}} g_\ell(i)g_\ell(j)Y_jx_j + \sum_{j \in T_{\text{small}}} g_\ell(i)g_\ell(j)Y_jx_j$

Let A_{big} be the event that $h_\ell(j) = h_\ell(i)$ for some $j \in T_{\text{big}}$ and $j \neq i$.

Lemma. W.p. at least $1 - \varepsilon^2/3$, no big coordinate collide with i under h_ℓ.

- For every $j \neq i$, Y_j is the indicator variable whether j is colliding with i under h_ℓ
- $\Pr[Y_j] = \frac{1}{w} \leq \frac{\varepsilon^2}{3k}$ (by pairwise independence of h_ℓ)
- Let $Y = \sum_{j \in T_{\text{big}}} Y_j$. By linearity of expectation, $\mathbb{E}[Y] \leq \varepsilon^2/3$.
- By Markov, $\Pr[A_{\text{big}}] = \Pr[Y \geq 1] \leq \varepsilon^2/3$

Refining Analysis (contd.)

Z'_ℓ

$$Z_\ell = g_\ell(i) C[\ell, h_\ell(i)] = x_i + \sum_{j \in T_{\mathrm{big}}} g_\ell(i) g_\ell(j) Y_j x_j + \sum_{j \in T_{\mathrm{small}}} g_\ell(i) g_\ell(j) Y_j x_j$$

Similar to earlier analysis for CountSketch,

Lemma. $\Pr\left[\left|Z'_\ell - x_i\right| \geq \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)\right] \leq 1/3.$

Lemma. W.p. at least $1 - \varepsilon^2/3$, no big coordinate collide with i under h_ℓ.

So, by union bound, for sufficiently small values of ε,

Lemma. $\Pr\left[\left|Z_\ell - x_i\right| \geq \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)\right] \leq \frac{1}{3} + \frac{\varepsilon^2}{3} \leq \frac{2}{5}.$

High probability estimates

Lemma. $\Pr\left[|Z_\ell - x_i| \geq \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)\right] \leq \frac{1}{3} + \frac{\varepsilon^2}{3} \leq \frac{2}{5}$.

Recall $\tilde{x}_i$= median$\{Z_1, \cdots, Z_d\}$,

- With $d = O(\log n)$, applying Chernoff bound,

$$\Pr\left[|\tilde{x}_i - x_i| \geq \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)\right] \leq 1/n^2$$

- By union bound, w.p. at least $1 - 1/n$, $|\tilde{x}_i - x_i| \leq \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)$ for all $i \in [n]$

Lemma. CountSketch with $w = \frac{3k}{\varepsilon^2}$ and $d = O(\log n)$ w.h.p. guarantees that

$$\forall i \in [n], \qquad |\tilde{x}_i - x_i| \leq \frac{\varepsilon}{\sqrt{k}} \cdot \mathrm{err}_2^k(x)$$

Dimensionality Reduction

JL Lemma and Subspace Embedding

Linear Sketching view of AMS-Sketch

- the sketch is mergeable.
- $z_{S \cup T} = z_S + z_T$

How to get the final estimate from the sketch z?

$$\hat{F}_2 = \text{median}_{g=1\ldots k}\left(\frac{1}{t}\sum_{j \in G_g} {z_j}^2\right)$$

where $G_1, \ldots, G_k$ are partition of the m rows ($m = tk$).

AMS-F_2-Sketch:

let $m = k \times t$

let Π be a $m \times n$ matrix with $\{-1, +1\}$ entries

(i) rows are independent and

(ii) in each row, entries are 4-wise indep.

$z \leftarrow 0$ is a $m \times 1$ vector initialized to $\mathbf{0}$

foreach item $\boldsymbol{i_j}$ in the stream do:

$z \leftarrow z + Me_{i_j}$

return z as sketch

From Lecture 6

Linear Sketching view of AMS-Sketch (contd.)

- **Geometric Interpretation**

Given a vector $x \in \mathbb{R}^n$, let M be the random map such that $z = Mx$ has the following properties:

- $\mathbb{E}[z_i] = 0$, $\mathbb{E}[z_i^2] = \|x\|_2^2$ for each $i \in [k]$ where k is the number of rows.
- Each z_i^2 is an estimate of length of x in Euclidean norm.
- With $k = \Theta(\varepsilon^{-2} \log(1/\delta))$, a $(1 \pm \varepsilon)$-estimate of $\|x\|_2$ can be driven via averaging and median technique.

In other words, x is compressed as a k-dimensional vector z that contains information to estimate $\|x\|_2$.

Do we need median trick? Is averaging enough?

Distributional Johnson-Lindenstrauss Lemma

Distributional JL Lemma. Fix $x \in \mathbb{R}^d$, and let $\Pi \in \mathbb{R}^{k \times d}$ be a matrix whose entries are chosen independently according to standard normal distribution $\mathcal{N}(0, 1)$. If $k = \Omega(\varepsilon^{-2} \log(1/\delta))$, then with probability at least $1 - \delta$,

$$\left\| \frac{1}{\sqrt{k}} \Pi x \right\|_2 = (1 \pm \varepsilon) \|x\|_2$$

i. We can instead choose entries from $\{-1, +1\}$ as well.

ii. Unlike AMS sketch, entries of Π are independent.

Basically, we've projected x from $\mathbb{R}^d$ into $\mathbb{R}^k$ while preserving length to a $(1 \pm \varepsilon)$-factor.

Distributional Johnson-Lindenstrauss Lemma

Distributional JL Lemma. Fix $x \in \mathbb{R}^d$, and let $\Pi \in \mathbb{R}^{k \times d}$ be a matrix whose entries are chosen independently according to standard normal distribution $\mathcal{N}(0,1)$. If $k = \Omega(\varepsilon^{-2} \log(1/\delta))$, then with probability at least $1 - \delta$,

$$\left\| \frac{1}{\sqrt{k}} \Pi x \right\|_2 = (1 \pm \varepsilon) \|x\|_2$$

i. We can instead choose entries from $\{-1, +1\}$ as well.
ii. Unlike AMS sketch, entries of Π are independent.

Basically, we've projected x from $\mathbb{R}^d$ into $\mathbb{R}^k$ while preserving length to a $(1 \pm \varepsilon)$-factor.

Distributional Johnson-Lindenstrauss Lemma

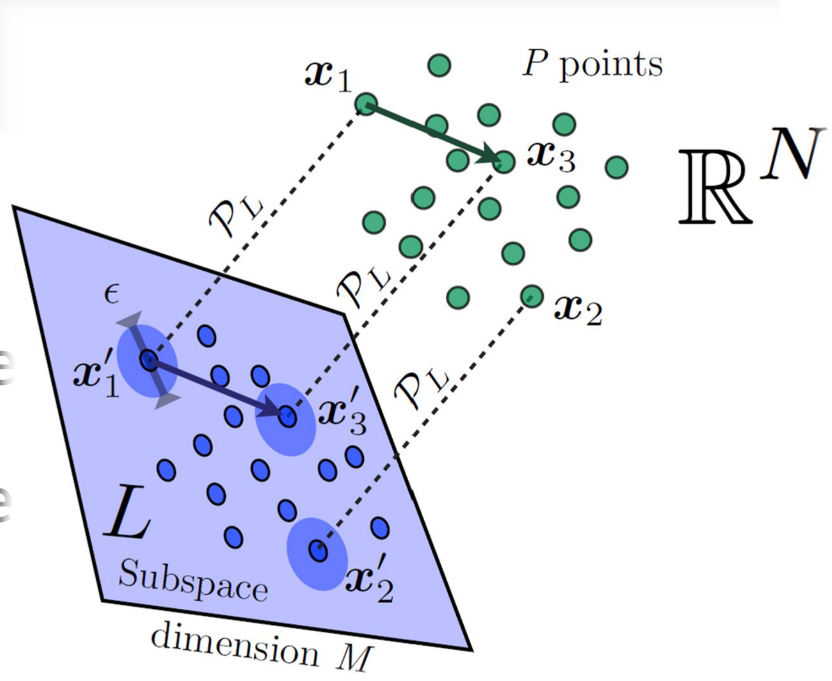
Distributional JL Lemma. Fix $x \in \mathbb{R}^d$, and let $\Pi \in \mathbb{R}^{k \times d}$ be a matrix whose entries are chosen independently according to standard normal distribution $\mathcal{N}(0, 1)$. If $k = \Omega(\varepsilon^{-2} \log(1/\delta))$, then with probability at least $1 - \delta$,

$$\left\| \frac{1}{\sqrt{k}} \Pi x \right\|_2 = (1 \pm \varepsilon) \|x\|_2$$

i. We can instead choose entries from $\{-1, +1\}$

ii. Unlike AMS sketch, entries of Π are independe

Basically, we've projected x from $\mathbb{R}^d$ into $\mathbb{R}^k$ while a $(1 \pm \varepsilon)$-factor.



Dimensionality Reduction

Metric JL Lemma. Let $v_1, \ldots, v_n$ be n points in $\mathbb{R}^d$. For any $\varepsilon \in (0, \frac{1}{2})$, there is a linear map $f: \mathbb{R}^d \to \mathbb{R}^k$ where $k \leq 8\varepsilon^{-2} \ln n$, such that for all $i \neq j \in [n]$,

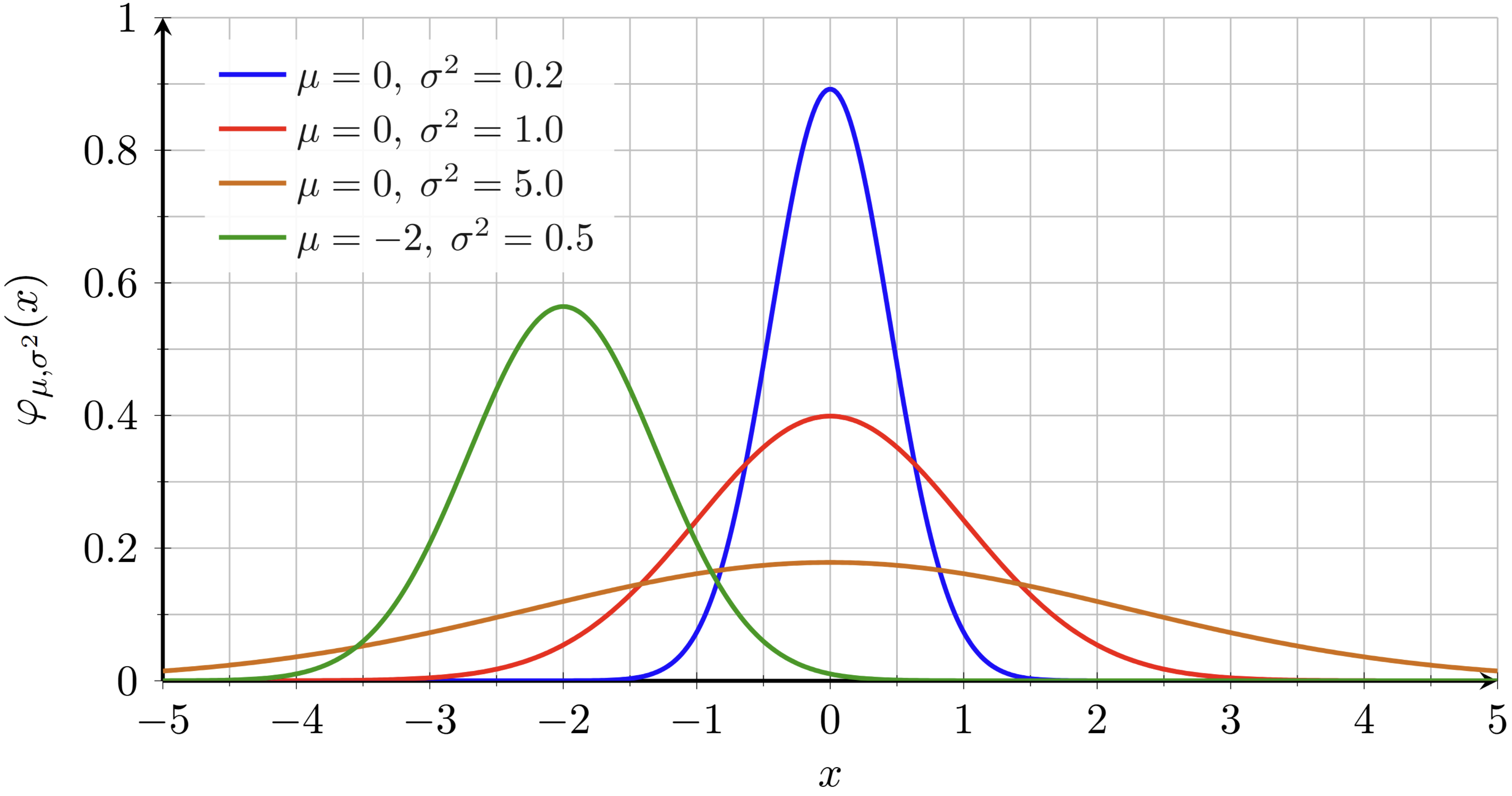
$$(\mathbf{1} - \boldsymbol{\varepsilon}) \left\| \boldsymbol{v_i} - \boldsymbol{v_j} \right\|_{\mathbf{2}} \leq \left\| \boldsymbol{f}(\boldsymbol{v_i}) - \boldsymbol{f}(\boldsymbol{v_j}) \right\|_{\mathbf{2}} \leq (\mathbf{1} + \boldsymbol{\varepsilon}) \left\| \boldsymbol{v_i} - \boldsymbol{v_j} \right\|_{\mathbf{2}}$$

- The linear map is simply given the random matrix Π; i.e., $\boldsymbol{f}(\boldsymbol{v}) = \boldsymbol{\Pi v}$
- The mapping is oblivious (to data)

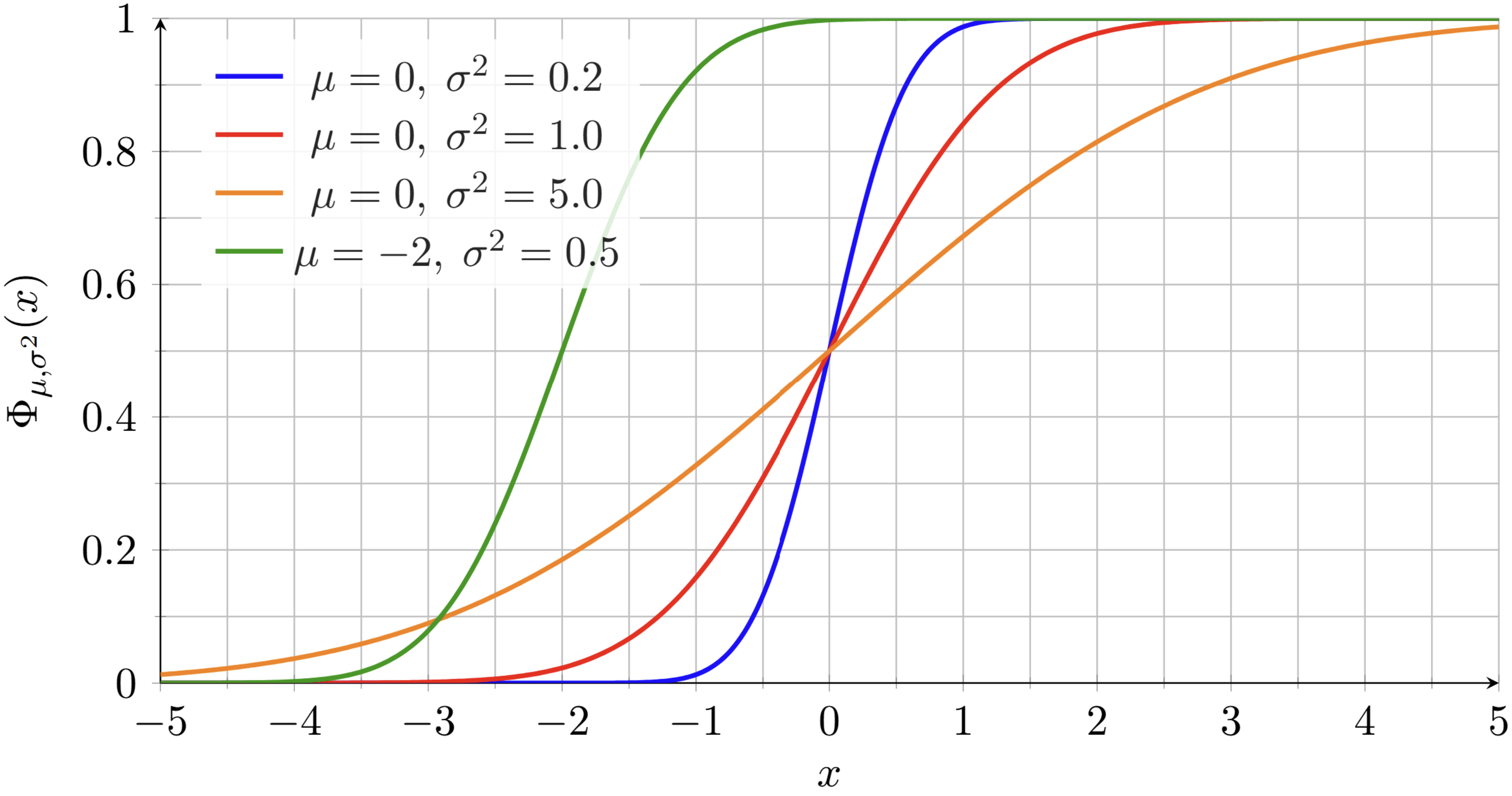
Proof. Apply DJL with $\delta = n^{-2}$, and union bound over the $\binom{n}{2}$ vectors $\boldsymbol{v_i} - \boldsymbol{v_j}$, for all pairs $\boldsymbol{i} \neq \boldsymbol{j} \in [\boldsymbol{n}]$.

Proof of DJL and Metric JL

Normal Distribution (PDF)



Normal Distribution (CDF)



Sum of Independent Normal Distribution

Lemma. Let X and Y be independent random variables.
Suppose $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Let $Z = X + Y$. Then,

$$\boldsymbol{Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)}$$

Corollary. Let X and Y be independent random variables. Suppose $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$. Let $Z = aX + bY$ where a, b are arbitrary real numbers. Then, $\boldsymbol{Z \sim \mathcal{N}(0, a^2 + b^2)}$

Normal distribution is a *stable distribution:* adding two indep. r.v. within the same class gives a distribution inside the class. Other exist and useful in $\boldsymbol{F_p}$ estimation for $\boldsymbol{p \in (0, 2)}$.

Random Gaussian Vector

One can consider higher dimensional normal distributions, also called multivariate Gaussian (or Normal) distributions.

Random Gaussian vector: $Z = (Z_1, \dots, Z_k)$ if $Z_i \sim \mathcal{N}(0,1)$ for each i, and $Z_1, \dots, Z_k$ are independent.

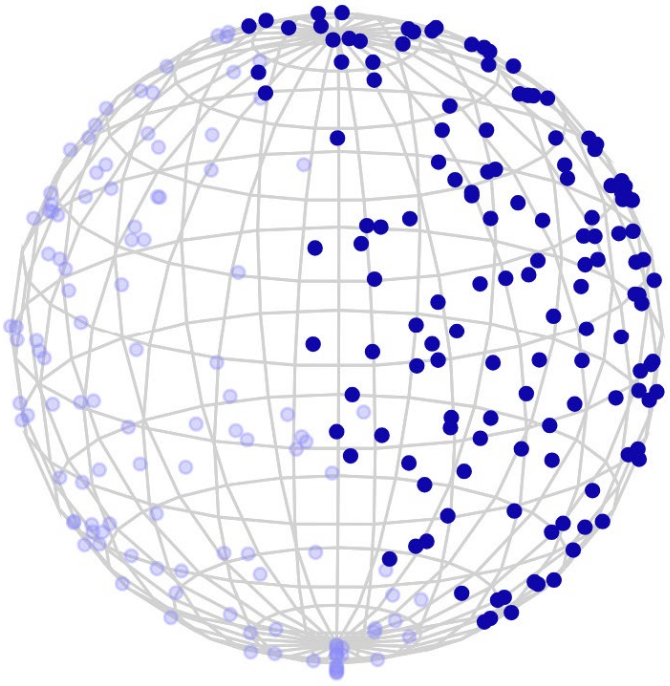
- Density function is $f(y_1, \dots, y_k) = \left(\frac{1}{\sqrt{2\pi}}\right)^k \exp\left(-\frac{y_1^2 + \cdots + y_k^2}{2}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^k e^{-\|y\|_2/2}$
- Only depends on $\|y\|_2$
- The distribution is **centrally symmetric**. (can be used to generate a random unit vector in $\mathbb{R}^k$). $U = \frac{Z}{\|Z\|}$ is uniform on the unit sphere.
- $\mathbb{E}\left[\|Z\|_2^2\right] = \sum_i \mathbb{E}\left[Z_i^2\right] = k$. Length is concentrated around k.

Random Gaussian Vector

One can consider higher dimensional normal distributions, also called multivariate Gaussian (or Normal) distribution

Random Gaussian vector: $Z = (Z_1, \dots, Z_k)$ if Z
and $Z_1, \dots, Z_k$ are independent.

- Density function is $f(y_1, \dots, y_k) = \left(\frac{1}{\sqrt{2\pi}}\right)^k \exp\left(-\frac{y_1^2}{}\right.$
- Only depends on $\|y\|_2$
- The distribution is **centrally symmetric**. (can be use vector in $\mathbb{R}^k$). $U = \frac{Z}{\|Z\|}$ is uniform on the unit sphere.
- $\mathbb{E}\left[\|Z\|_2^2\right] = \sum_i \mathbb{E}\left[Z_i^2\right] = k$. Length is concentrated ar



Concentration of sum of squares of normally distributed variables

$\chi^2(k)$ **distribution:** distribution of sum of squares of k independent standard normally distributed random variables,

$$Y = \sum_{1 \le i \le k} Z_i^2 \text{ where each } Z_i \sim \mathcal{N}(0,1)$$

Lemma. Let $Z_1, \dots, Z_k$ be independent $\mathcal{N}(0,1)$ r.v.s. and let $Y = \sum_i Z_i^2$. Then, for $\varepsilon \in (0,1/2)$, there is a constant c such that,

$$\mathbf{Pr}\left[(1-\varepsilon)^2 k \le Y \le (1+\varepsilon)^2 k\right] \ge 1 - 2e^{c\varepsilon^2 k}$$

- Recall Chernoff for bounded independent non-negative rv. Z_i^2 are not bounded, however, Chernoff bounds extend to sums of random variables with exponentially decaying tails.

Proof of DJL Lemma

Distributional JL Lemma. Fix $x \in \mathbb{R}^d$, and let $\Pi \in \mathbb{R}^{k \times d}$ be a matrix whose entries are chosen independently according to standard normal distribution $\mathcal{N}(0,1)$. If $k = \Omega(\varepsilon^{-2} \log(1/\delta))$, then with probability at least $1 - \delta$,

$$\left\| \frac{1}{\sqrt{k}} \Pi x \right\|_2 = (1 \pm \varepsilon) \|x\|_2$$

Proof of DJL Lemma

Without loss of generality, assume $\|x\|_2 = 1$. $\boldsymbol{Z_i = \sum_{j=1} \Pi_{ij} x_i}$

- $Z_i \sim \mathcal{N}(0,1)$
- Z is a random Gaussian vector in k dimensions.
- $Y = \sum_i Z_i^2$. Y's distribution is $\chi^2(k)$ since each coordinate is i.i.d. Gaussians.
- Hence, $\mathbf{Pr}\left[(1-\boldsymbol{\varepsilon})^2 \boldsymbol{k} \leq \boldsymbol{Y} \leq (1+\boldsymbol{\varepsilon})^2 \boldsymbol{k}\right] \geq 1 - 2e^{c\varepsilon^2 k}$
- Since $k = \Omega(\varepsilon^{-2}\log(1/\delta))$, $\mathbf{Pr}\left[(1-\boldsymbol{\varepsilon})^2 \boldsymbol{k} \leq \boldsymbol{Y} \leq (1+\boldsymbol{\varepsilon})^2 \boldsymbol{k}\right] \geq 1 - \boldsymbol{\delta}$
- Therefore, $\|z\|_2 = \sqrt{Y/k}$ has the property that with probability $1-\delta$,

$$\|z\|_2 = (1 \pm \varepsilon)\,\|x\|_2$$