Law of large Numbers :-

$$P\left(\left|\frac{\sum x_i}{n} - E(x)\right| \geq \varepsilon\right) \leq \frac{Var(x)}{n\varepsilon^2}$$

i.e. larger the variance ⇒ more is the probability that $\left|\frac{\sum x_i}{n} - E(x)\right|$ will Exceed

ε → error.

* Markov's inequality :-

x → non negative random variable then.

for $a > 0$, $\quad P(x \geq a) \leq \frac{E(x)}{a}$

* chebysheu's inequality :-

$$P(|x - E(x)| \geq c) \leq \frac{Var(x)}{c^2}$$

* In law of large numbers size of universe does not enters in to the bound.

* Geometry of High Dimensions:-

(i) most of the volume is near surface.

(ii) A, object in R^d, shrunk by ε, then

$$\text{Volume}((1-\varepsilon)A) = (1-\varepsilon)^d \text{Volum}(A)$$

(iii) partition A into infinitesimal cubes the $(1-\varepsilon)A$ is union of a set of cubes obtained by shrinking the cubes in A by a factor of $(1-\varepsilon)$.

When we shrink each of the 2d sides of a d-dimensional cube by a factor f its volume shrinks by a factor of f^d

$\because 1-x \leq e^{-x}$

$$\therefore \frac{\text{Volume}((1-\varepsilon)A)}{\text{Volume}(A)} = (1-\varepsilon)^d \leq e^{-\varepsilon d}$$

as $d \to \infty$, $\uparrow \longrightarrow 0$,

$\therefore$ all volume of A must be in portion of A that does not belong to $(1-\varepsilon)A$.

* Most of the volume of d-dimensional unit ball is contained in an annulus of width $O(1/d)$ near the boundry. If the ball is of radius r, then annulus width is $O(r/d)$

* Area & volume of unit ball

$$A(d) = \frac{2\pi^{d/2}}{\lceil d/2}$$ & $$V(d) = \frac{2\pi^{d/2}}{d\lceil d/2}$$

where, $d \rightarrow$ dimension of unit ball.

* Volume near the equator :-

Theorem 2.7 :- for $c \geq 1$ & $d \geq 3$, at least $a\left(1 - \frac{2}{c}e^{c^2/2}\right)$ fraction of volume of the d-dimensional unit ball has $|x_1| \leq \frac{c}{\sqrt{d-1}}$

* Most of the volume of upper hemisphere of the d-dimensional ball is below the plane $x_1 = \frac{c}{\sqrt{d-1}}$

$$\frac{\text{Upper bound above plane}}{\text{lower bound total hemisphere}} = \frac{\frac{V(d-1)}{c\sqrt{d-1}} e^{-\frac{c^2}{2}}}{\frac{V(d-1)}{2\sqrt{d-1}}}$$

$$\therefore \frac{\text{Upper bound above plane}}{\text{lower bound total plane}} = \frac{2}{c} e^{-\frac{c^2}{2}}$$

* Near orthogonality :-

One immediate implication of the above analysis is that if we draw to points at random from unit ball, with high probability their vectors will be nearly orthogonal to each other.

* Thm 2.8 :- Consider drawing n-points $x_1, x_2, x_3 \ldots, x_n$ at random from unit ball with high probability $1 - O(1/n)$

① $|x_i| \geq 1 - \frac{2\ln n}{d}$ for all i, and

② $|x_i \cdot x_j| \leq \frac{\sqrt{6\ln n}}{\sqrt{d-1}}$ for all $i \neq j$

* for $\mu = 0$ & $\text{var}(x) = 1$

then Gaussian,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

$$p(x) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{x_1^2 + x_2^2 + x_3^2 + \ldots + x_d^2}{2}}$$

& is spherically symmetric.

* ones a vector is normalized its co-ordinates are no longer statically independent.

* intuition of gaussian annulus theorem :-

in higher dimension the p.d. of gaussian is

$$p(x) = \frac{1}{(2\pi)^{d/2} \sigma^d} e^{-\frac{|x|^2}{2\sigma^2}}$$

& it essentially means that nearly all the probability is concentrated in a thin annulus of width $O(1)$ at radius $\sqrt{d}$.

• also value of pdf is max. at origin, but volume at origin is pretty low. While integrating pdf over unit ball we donot get significant integration till the radius of ball $\simeq \sqrt{d}$.

* Gaussian Annulus Theorem :-

→ for d-dimensional spherical gaussian with unit variance in each direction for any $\beta \leq \sqrt{d}$, all but most $3e^{-c\beta^2}$ of the probability mass lies within the annulus of $\sqrt{d} - \beta \leq |x| \leq \sqrt{d} + \beta$.

i.e. nearly all probability lies in an annulus of width $O(1+d)$ & radius $\sqrt{d}$.

2.7 Random Projection and Johnson-Lindenstrauss Lemma

One of the most frequently used subroutines in tasks involving high dimensional data is nearest neighbor search. In nearest neighbor search we are given a database of n points in $\mathbf{R}^d$ where n and d are usually large. The database can be preprocessed and stored in an efficient data structure. Thereafter, we are presented "query" points in $\mathbf{R}^d$ and are asked to find the nearest or approximately nearest database point to the query point. Since the number of queries is often large, the time to answer each query should be very small, ideally a small function of $\log n$ and $\log d$, whereas preprocessing time could be larger, namely a polynomial function of n and d. For this and other problems, dimension reduction, where one projects the database points to a k-dimensional space with $k \ll d$ (usually dependent on $\log d$) can be very useful so long as the relative distances between points are approximately preserved. We will see using the Gaussian Annulus Theorem that such a projection indeed exists and is simple.

The projection $f : \mathbf{R}^d \to \mathbf{R}^k$ that we will examine (many related projections are known to work as well) is the following. Pick k Gaussian vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k}$ in $\mathbf{R}^d$ with unit-variance coordinates. For any vector $\mathbf{v}$, define the projection $f(\mathbf{v})$ by:

$$f(\mathbf{v}) = (\mathbf{u_1} \cdot \mathbf{v}, \mathbf{u_2} \cdot \mathbf{v}, \ldots, \mathbf{u_k} \cdot \mathbf{v}).$$

The projection $f(\mathbf{v})$ is the vector of dot products of $\mathbf{v}$ with the $\mathbf{u_i}$. We will show that with high probability, $|f(\mathbf{v})| \approx \sqrt{k}|\mathbf{v}|$. For any two vectors $\mathbf{v_1}$ and $\mathbf{v_2}$, $f(\mathbf{v_1} - \mathbf{v_2}) = f(\mathbf{v_1}) - f(\mathbf{v_2})$. Thus, to estimate the distance $|\mathbf{v_1} - \mathbf{v_2}|$ between two vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ in $\mathbf{R}^d$, it suffices to compute $|f(\mathbf{v_1}) - f(\mathbf{v_2})| = |f(\mathbf{v_1} - \mathbf{v_2})|$ in the k-dimensional space since the factor of $\sqrt{k}$ is known and one can divide by it. The reason distances increase when we project to a lower dimensional space is that the vectors $\mathbf{u_i}$ are not unit length. Also notice that the vectors $\mathbf{u_i}$ are not orthogonal. If we had required them to be orthogonal, we would have lost statistical independence.

Theorem 2.10 (The Random Projection Theorem) *Let $\mathbf{v}$ be a fixed vector in $\mathbf{R}^d$ and let f be defined as above. There exists constant $c > 0$ such that for $\varepsilon \in (0, 1)$,*

$$Prob\left(\left||f(\mathbf{v})| - \sqrt{k}|\mathbf{v}|\right| \geq \varepsilon\sqrt{k}|\mathbf{v}|\right) \leq 3e^{-ck\varepsilon^2},$$

where the probability is taken over the random draws of vectors $\mathbf{u_i}$ used to construct f.

essentially mean, that probability of length of projection of vector differing from its expected value is exponentially small in k, dimension of target subspace.

Theorem 2.11 (Johnson-Lindenstrauss Lemma) *For any $0 < \varepsilon < 1$ and any integer n, let $k \geq \frac{3}{c\varepsilon^2} \ln n$ with c as in Theorem 2.9. For any set of n points in R^d, the random projection $f : R^d \to R^k$ defined above has the property that for all pairs of points $\mathbf{v_i}$ and $\mathbf{v_j}$, with probability at least $1 - 3/2n$,*

$$(1-\varepsilon)\sqrt{k}\,|\mathbf{v_i} - \mathbf{v_j}| \leq |f(\mathbf{v_i}) - f(\mathbf{v_j})| \leq (1+\varepsilon)\sqrt{k}\,|\mathbf{v_i} - \mathbf{v_j}|\,.$$

Proof: Applying the Random Projection Theorem (Theorem 2.10), for any fixed $\mathbf{v_i}$ and $\mathbf{v_j}$, the probability that $|f(\mathbf{v_i} - \mathbf{v_j})|$ is outside the range

$$\left[(1-\varepsilon)\sqrt{k}|\mathbf{v_i} - \mathbf{v_j}|, (1+\varepsilon)\sqrt{k}|\mathbf{v_i} - \mathbf{v_j}|\right]$$

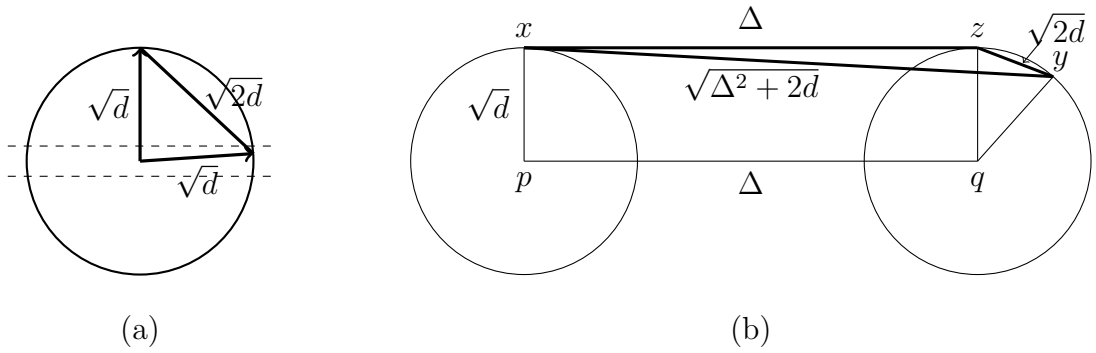
is at most $3e^{-ck\varepsilon^2} \leq 3/n^3$ for $k \geq \frac{3\ln n}{c\varepsilon^2}$. Since there are $\binom{n}{2} < n^2/2$ pairs of points, by the union bound, the probability that any pair has a large distortion is less than $\frac{3}{2n}$. ∎

Remark: It is important to note that the conclusion of Theorem 2.11 asserts for all $\mathbf{v_i}$ and $\mathbf{v_j}$, not just for most of them. The weaker assertion for most $\mathbf{v_i}$ and $\mathbf{v_j}$ is typically less useful, since our algorithm for a problem such as nearest-neighbor search might return one of the bad pairs of points. A remarkable aspect of the theorem is that the number of dimensions in the projection is only dependent logarithmically on n. Since k is often much less than d, this is called a dimension reduction technique. In applications, the dominant term is typically the $1/\varepsilon^2$ term.

for nearest neighbour problem, if database has n_1 points & n_2-queries are expected during lifetime of the algorithm take $n = n_1 + n_2$ & project database to random k-dimensional space.

* separating Gaussians :-

First, consider just one spherical unit-variance Gaussian centered at the origin. From Theorem 2.9, most of its probability mass lies on an annulus of width $O(1)$ at radius $\sqrt{d}$. Also $e^{-|\mathbf{x}|^2/2} = \prod_i e^{-x_i^2/2}$ and almost all of the mass is within the slab $\{\, \mathbf{x} \mid -c \leq x_1 \leq c \,\}$, for $c \in O(1)$. Pick a point $\mathbf{x}$ from this Gaussian. After picking $\mathbf{x}$, rotate the coordinate system to make the first axis align with $\mathbf{x}$. Independently pick a second point $\mathbf{y}$ from



(a) (b)

Figure 2.5: (a) indicates that two randomly chosen points in high dimension are surely almost nearly orthogonal. (b) indicates the distance between a pair of random points from two different unit balls approximating the annuli of two Gaussians.

this Gaussian. The fact that almost all of the probability mass of the Gaussian is within the slab $\{\mathbf{x} \mid -c \leq x_1 \leq c,\ c \in O(1)\}$ at the equator implies that $\mathbf{y}$'s component along $\mathbf{x}$'s direction is $O(1)$ with high probability. Thus, $\mathbf{y}$ is nearly perpendicular to $\mathbf{x}$. So, $|\mathbf{x} - \mathbf{y}| \approx \sqrt{|\mathbf{x}|^2 + |\mathbf{y}|^2}$. See Figure 2.5(a). More precisely, since the coordinate system has been rotated so that $\mathbf{x}$ is at the North Pole, $\mathbf{x} = (\sqrt{d} \pm O(1), 0, \ldots, 0)$. Since $\mathbf{y}$ is almost on the equator, further rotate the coordinate system so that the component of $\mathbf{y}$ that is perpendicular to the axis of the North Pole is in the second coordinate. Then $\mathbf{y} = (O(1), \sqrt{d} \pm O(1), 0, \ldots, 0)$. Thus,

$$(\mathbf{x} - \mathbf{y})^2 = d \pm O(\sqrt{d}) + d \pm O(\sqrt{d}) = 2d \pm O(\sqrt{d})$$

and $|\mathbf{x} - \mathbf{y}| = \sqrt{2d} \pm O(1)$ with high probability.

Consider two spherical unit variance Gaussians with centers $\mathbf{p}$ and $\mathbf{q}$ separated by a distance Δ. The distance between a randomly chosen point $\mathbf{x}$ from the first Gaussian and a randomly chosen point $\mathbf{y}$ from the second is close to $\sqrt{\Delta^2 + 2d}$, since $\mathbf{x} - \mathbf{p}, \mathbf{p} - \mathbf{q}$, and $\mathbf{q} - \mathbf{y}$ are nearly mutually perpendicular. Pick $\mathbf{x}$ and rotate the coordinate system so that $\mathbf{x}$ is at the North Pole. Let $\mathbf{z}$ be the North Pole of the ball approximating the second Gaussian. Now pick $\mathbf{y}$. Most of the mass of the second Gaussian is within $O(1)$ of the equator perpendicular to $\mathbf{z} - \mathbf{q}$. Also, most of the mass of each Gaussian is within distance $O(1)$ of the respective equators perpendicular to the line $\mathbf{q} - \mathbf{p}$. See Figure 2.5 (b). Thus,

$$\begin{aligned} |\mathbf{x} - \mathbf{y}|^2 &\approx \Delta^2 + |\mathbf{z} - \mathbf{q}|^2 + |\mathbf{q} - \mathbf{y}|^2 \\ &= \Delta^2 + 2d \pm O(\sqrt{d})). \end{aligned}$$

To ensure that the distance between two points picked from the same Gaussian are closer to each other than two points picked from different Gaussians requires that the upper limit of the distance between a pair of points from the same Gaussian is at most

the lower limit of distance between points from different Gaussians. This requires that $\sqrt{2d}+O(1) \leq \sqrt{2d+\Delta^2}-O(1)$ or $2d+O(\sqrt{d}) \leq 2d+\Delta^2$, which holds when $\Delta \in \omega(d^{1/4})$. Thus, mixtures of spherical Gaussians can be separated in this way, provided their centers are separated by $\omega(d^{1/4})$. If we have n points and want to correctly separate all of them with high probability, we need our individual high-probability statements to hold with probability $1-1/poly(n)$,[3] which means our $O(1)$ terms from Theorem 2.9 become $O(\sqrt{\log n})$. So we need to include an extra $O(\sqrt{\log n})$ term in the separation distance.

Algorithm for separating points from two Gaussians: Calculate all pairwise distances between points. The cluster of smallest pairwise distances must come from a single Gaussian. Remove these points. The remaining points come from the second Gaussian.

IMP

One can actually separate Gaussians where the centers are much closer. In the next chapter we will use singular value decomposition to separate points from a mixture of two Gaussians when their centers are separated by a distance $O(1)$.

IMP

2.9 Fitting a Spherical Gaussian to Data

Given a set of sample points, $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, in a d-dimensional space, we wish to find the spherical Gaussian that best fits the points. Let f be the unknown Gaussian with mean $\boldsymbol{\mu}$ and variance σ^2 in each direction. The probability density for picking these points when sampling according to f is given by

$$c \exp\left(-\frac{(\mathbf{x_1}-\boldsymbol{\mu})^2+(\mathbf{x_2}-\boldsymbol{\mu})^2+\cdots+(\mathbf{x_n}-\boldsymbol{\mu})^2}{2\sigma^2}\right)$$

where the normalizing constant c is the reciprocal of $\left[\int e^{-\frac{|\mathbf{x}-\boldsymbol{\mu}|^2}{2\sigma^2}} dx\right]^n$. In integrating from $-\infty$ to ∞, one can shift the origin to $\boldsymbol{\mu}$ and thus c is $\left[\int e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}} dx\right]^{-n} = \frac{1}{(2\pi)^{\frac{n}{2}}}$ and is independent of $\boldsymbol{\mu}$.

$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}$

The *Maximum Likelihood Estimator* (MLE) of f, given the samples $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, is the f that maximizes the above probability density.

Lemma 2.12 *Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a set of n d-dimensional points. Then $(\mathbf{x}_1-\boldsymbol{\mu})^2+(\mathbf{x}_2-\boldsymbol{\mu})^2+\cdots+(\mathbf{x}_n-\boldsymbol{\mu})^2$ is minimized when $\boldsymbol{\mu}$ is the centroid of the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, namely $\boldsymbol{\mu} = \frac{1}{n}(\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_n)$.*

Proof: Setting the gradient of $(\mathbf{x}_1-\boldsymbol{\mu})^2+(\mathbf{x}_2-\boldsymbol{\mu})^2+\cdots+(\mathbf{x}_n-\boldsymbol{\mu})^2$ with respect to $\boldsymbol{\mu}$ to zero yields

$$-2(\mathbf{x}_1-\boldsymbol{\mu})-2(\mathbf{x}_2-\boldsymbol{\mu})-\cdots-2(\mathbf{x}_n-\boldsymbol{\mu})=0.$$

Solving for $\boldsymbol{\mu}$ gives $\boldsymbol{\mu} = \frac{1}{n}(\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_n)$. ∎

[3] poly(n) means bounded by a polynomial in n.