CS5487 Lecture Notes (2020)
Dr. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Lecture 1

Probability Review

Random variable

The r.v. X takes a value in $\mathcal{X}$ (set of possible values) according to some outcome of an event.

E.g. Event: flip a coin

$\mathcal{X} = \{H, T\}$

X has value H when the coin lands as heads.
" " T " " " " as tails.

Associated with r.v. X is a distribution $p(X=x)$ that describes the frequency of events occurring.

Examples

Discrete r.v.

indicator variable

$\mathcal{X} = \{0, 1\}$

# of people in class

$\mathcal{X} = \mathbb{Z}_+$ ← non-negative integers

probability mass function (pmf)

$P(X=x)$ = probability of taking value x

$$\sum_{x \in \mathcal{X}} p(X=x) = 1$$

$$0 \le P(X=x) \le 1, \quad \forall x \in \mathcal{X}$$

Continuous R.V.

sensor reading

$\mathcal{X} = \mathbb{R}$ ← real numbers

probability density function (pdf)

$p(x)$ = likelihood of x

$$P(a \le x \le b) = \int_a^b p(x)\,dx$$

$$\int p(x)\,dx = 1$$

$$0 \le p(x) \quad \forall x \in \mathcal{X}$$

Example Distributions

• **Bernoulli (coin)**

$\mathcal{X} = \{0, 1\}$, π = probability of 1

$$\begin{cases} p(x=1) = \pi \\ p(x=0) = 1-\pi \end{cases}$$

$$p(x) = \pi^x (1-\pi)^{1-x}$$

$x=0 \Rightarrow \pi^0 (1-\pi)^1 = 1-\pi$

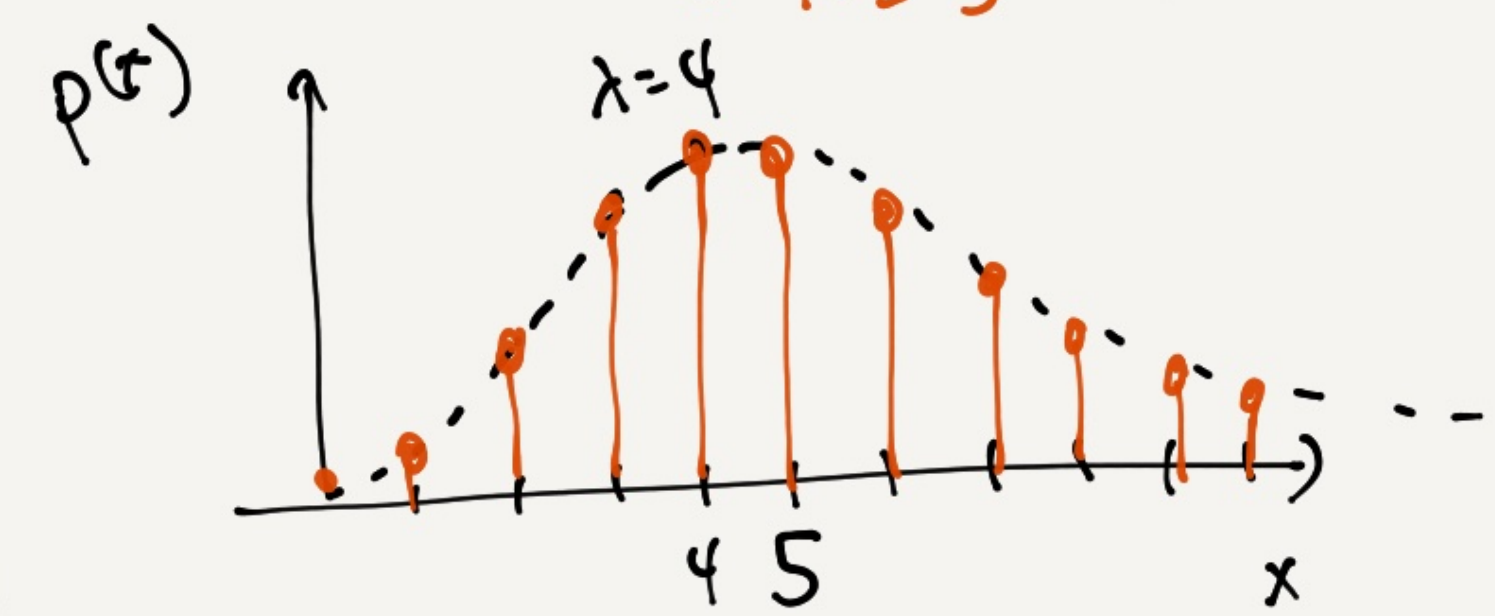
$x=1 \Rightarrow \pi^1 \underline{(1-\pi)^{1-1}} = \pi$

• **Poisson distribution**: # of arrivals over a fixed time period.

$\mathcal{X} = \mathbb{Z}_+$, λ = average arrival rate ≥ 0

$$p(x) = \frac{1}{x!} e^{-\lambda} \lambda^x$$

$x!$ = x factorial = $x(x-1)(x-2)\cdots 1$
$= 1 \cdot 2 \cdot 3 \cdot \ldots \cdot x$



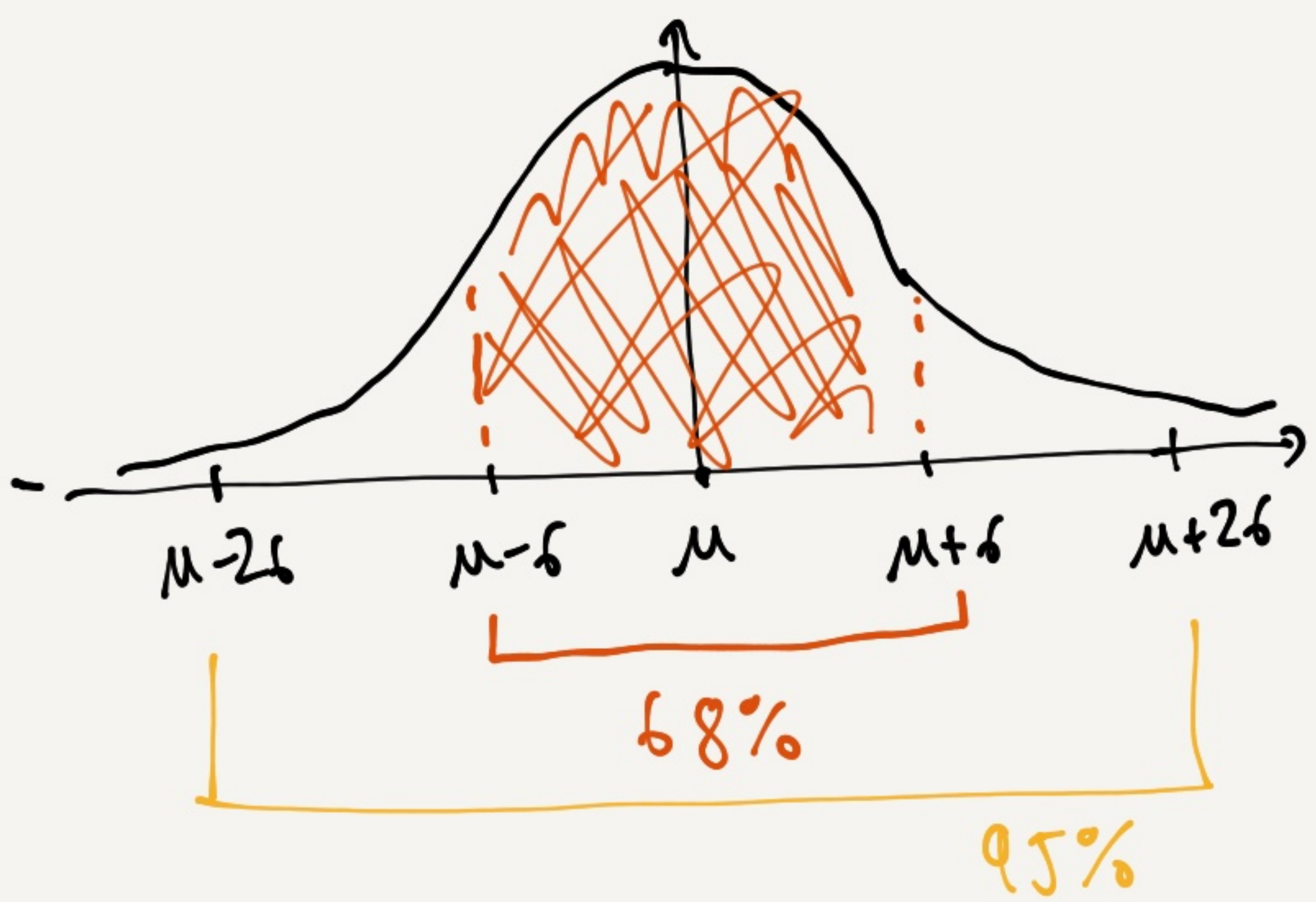
Gaussian (Normal)

$\mathcal{X} = \mathbb{R}$

μ = mean

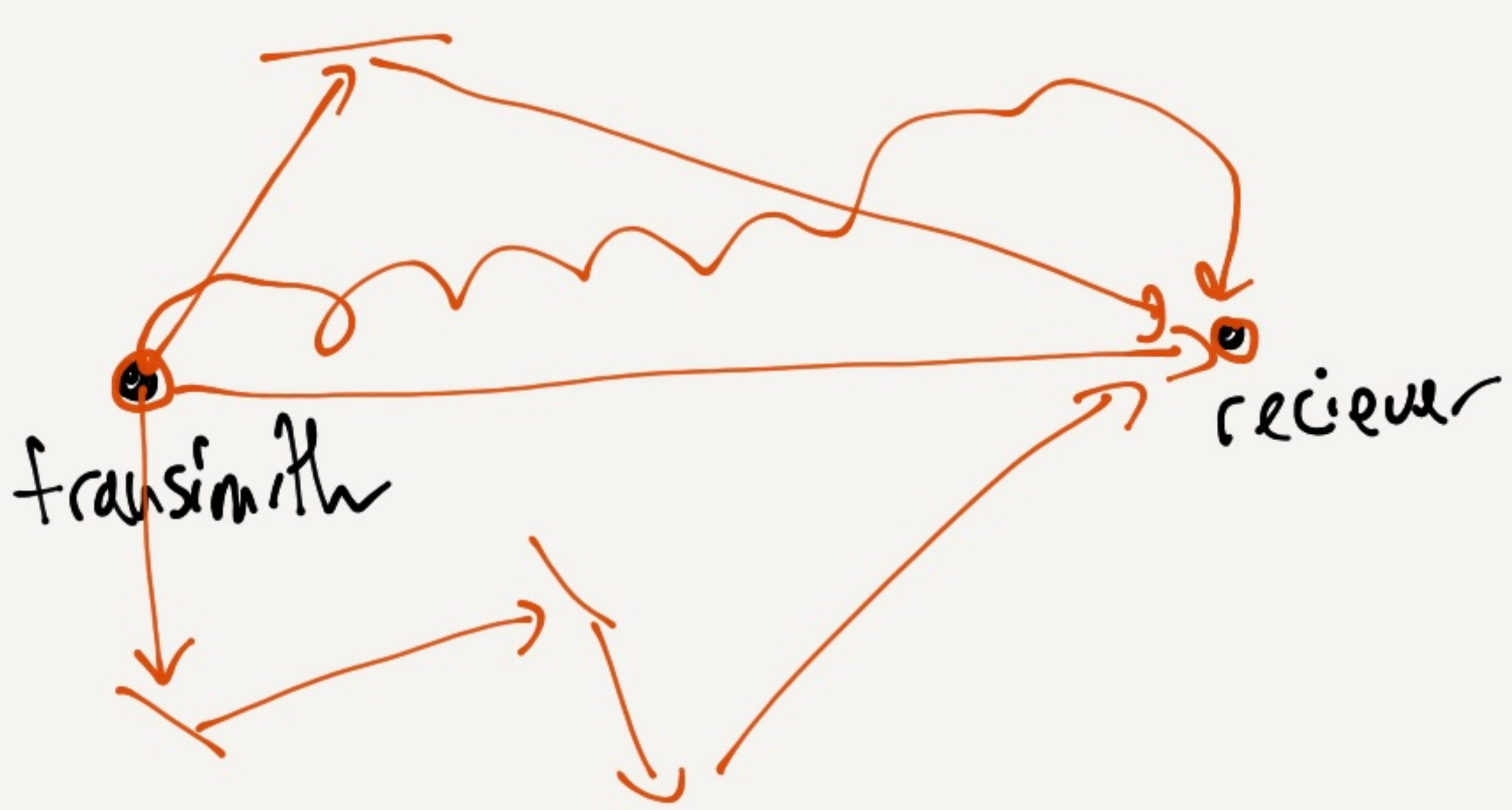
σ^2 = variance > 0 $\rightsquigarrow$ σ = standard deviation

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = N(x \mid \mu, \sigma^2)$$



Central Limit Theorem (CLT)

sum N r.v. $\to$ Gaussian distribution for large N.



Joint Distribution

Distribution of more than one r.v.

$$p_{XY}(X=x, Y=y) = p(x,y)$$

probability/likelihood of $X=x$ AND $Y=y$.

Notation:

$p_X(x) \Rightarrow p(x)$

$P(X=x) \Rightarrow$

Example: $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{0,1\}$

joint $p(x,y)$	$Y=0$	$Y=1$	$p(x)$
$X=0$	0.08	0.12	0.20
$X=1$	0.32	0.48	0.80
$p(y)$	0.4	0.6	

- why is it called marginalization?

"margin of table"

Marginal Distribution

distribution over one r.v. in the joint.

$$p(X=x) = \sum_{y \in \mathcal{Y}} p(X=x, Y=y)$$

$$p(x) = \int_{y \in \mathcal{Y}} p(x,y)\,dy$$

"sum out the other variables" $\Rightarrow$ "marginalization"

Conditional Distribution

distribution of one r.v. when the value of another r.v. is known (given).

$$p(X=x \mid Y=y) = \frac{p(X=x, Y=y)}{p(Y=y)}$$

"given"

e.g.

$$p(X=0 \mid Y=0) = \frac{p(X=0, Y=0)}{p(Y=0)} = \frac{0.08}{0.4} = 0.2$$

$$p(X=1 \mid Y=0) = \frac{p(X=1, Y=0)}{p(Y=0)} = \frac{0.32}{0.4} = 0.8$$

distribution over X (with $Y=0$)

$$\to \quad p(x,y) = p(x|y)\,p(y) \quad ☆$$

Statistical Independence

distribution of a r.v. does not change when knowing the value of another r.v.

$X \perp Y$ iff $p(x|y) = p(x)$

$X \perp Y$ iff $p(x,y) = p(x)\,p(y)$

"joint is a product of the marginals"

☆ in our example $p(x|y) = p(x) \Rightarrow X \perp Y$

Bayes Rule

use defn of cond. prob.

$$\begin{cases} p(x,y) = p(x|y)\,p(y) \\ p(x,y) = p(y|x)\,p(x) \end{cases}$$

$$\Rightarrow \quad p(y|x)\,p(x) = p(x|y)\,p(y)$$

$$\Rightarrow \quad p(y|x) = \frac{p(x|y)\,p(y)}{p(x)}$$

we have $p(x) = \int p(x,y)\,dy = \int p(x|y)\,p(y)\,dy$

Bayes Rule: $$\boxed{p(y|x) = \frac{p(x|y)\,p(y)}{\int p(x|y)\,p(y)\,dy}}$$

"given only $p(x|y)$ & $p(y)$, we can "invert" the conditioning to obtain $p(y|x)$".

ρ "rho" p "p"

Expectations

Suppose we have a function $f(x)$ & a r.v. X.

On average, what value of $f(x)$ would I observe?

$$E_X[f(x)] = \sum_{x \in \mathcal{X}} f(x)p(x)$$ ⇐ "weighted average of function values"

E ← expectation, X ← r.v.

$$E_X[f(x)] = \int_{x \in \mathcal{X}} f(x)p(x)\,dx$$

σ = standard deviation

- **mean:** $E[x] = \int x\,p(x)\,dx = \mu_x$
- **variance:** $var(x) = E[(x - E_x)^2] = \sigma_x^2$ ← variance
  $= E[x^2] - (\underbrace{Ex}_{\mu_x})^2$
- **covariance:** $cov(x,y) = E_{XY}[(x - EX)(y - EY)]$
  $= \iint (x - EX)(y - EY)\,p(x,y)\,dx\,dy$
  $= E_{XY}[xy] - E_X[x]E_Y[y]$
  $= \sigma_{xy}^2$

covariance — between x & y

Review of Linear Algebra

column vector: $X \in \mathbb{R}^d$ ← d-dim vector of real values

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

column-centric

matrix: $A \in \mathbb{R}^{m \times n}$ (m rows, n columns)

$$A = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

a_i is the i^{th} column of A

inner product: $x^T y = \sum_i x_i y_i$ (similarity btwn x & y)

length (norm): $\|x\| = \sqrt{x^T x} = \sqrt{\sum_i x_i^2}$

distance metric: $\|x - y\| = \sqrt{(x-y)^T(x-y)} = \sqrt{\sum_i (x_i - y_i)^2}$

outer product: $$xy^T = \begin{bmatrix} x_1y_1 & \cdots & x_1y_n \\ \vdots & \ddots & \vdots \\ x_ny_1 & \cdots & x_ny_n \end{bmatrix}$$ "pairwise products of entries in x & y"

matrix-vector multiplication

① $A \in \mathbb{R}^{m\times d}$, $x \in \mathbb{R}^d$

$$y = Ax = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_d \\ | & & | \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = \sum_i x_i a_i \in \mathbb{R}^m$$

(linear combo of the columns of A)

② $A \in \mathbb{R}^{d\times m}$, $x \in \mathbb{R}^d$

$$y = A^T x = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_m \\ | & & | \end{bmatrix}^T x = \begin{bmatrix} — a_1^T — \\ \vdots \\ — a_m^T — \end{bmatrix}\begin{bmatrix} | \\ x \\ | \end{bmatrix}$$

$$= \begin{bmatrix} a_1^T x \\ \vdots \\ a_m^T x \end{bmatrix} \in \mathbb{R}^m$$

(vector of innerproducts btwn columns of A and x)

matrix-matrix multiplication

① $\underset{\mathbb{R}^{m\times d}}{A}\ \underset{\mathbb{R}^{d\times n}}{B} = A\begin{bmatrix} | & & | \\ b_1 & \cdots & b_n \\ | & & | \end{bmatrix} = \begin{bmatrix} Ab_1 & \cdots & Ab_n \end{bmatrix}$ (A multiplied each column of B)

② $\underset{\mathbb{R}^{d\times m}}{A^T}\ \underset{\mathbb{R}^{d\times n}}{B} = \begin{bmatrix} — a_1^T — \\ \vdots \\ — a_m^T — \end{bmatrix}\begin{bmatrix} | & & | \\ b_1 & \cdots & b_n \\ | & & | \end{bmatrix} = \begin{bmatrix} a_i^T b_j \end{bmatrix}_{ij} \in \mathbb{R}^{m\times n}$

(matrix of innerproducts btwn columns of A & columns of B).

③ $\underset{\mathbb{R}^{m\times d}}{A}\ \underset{\mathbb{R}^{n\times d}}{B^T} = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_m \\ | & & | \end{bmatrix}\begin{bmatrix} — b_1^T — \\ \vdots \\ — b_n^T — \end{bmatrix} = \sum_i a_i b_i^T$

(sum of outerproducts of columns of A & B)

Vector r.v.

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \qquad \mathcal{X} = \mathbb{R}^d$$

Notation

$$p(x_1, \dots, x_d) = p(\underset{\text{vector}}{x}), \quad \int p(x)\,dx = 1$$

mean vector: $$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} = E[x] = \int_{\mathbb{R}^d} x\,p(x)\,dx$$

$$\iiiint x\,p(x)\,dx_1\,dx_2\,dx_3 \cdots dx_d$$

covariance matrix: $$\mathrm{cov}(x) = E[(x - Ex)(x - Ex)^T]$$

2d example $$= E\begin{bmatrix} (x_1-\mu_1)^2 & (x_1-\mu_1)(x_2-\mu_2) \\ (x_1-\mu_1)(x_2-\mu_2) & (x_2-\mu_2)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix}$$

covariance btwn x_1 & x_2.

variance.

$$\mathrm{cov}(x) = E[xx^T] - E[x]E[x]^T$$

Multivariate Gaussian

$\mathcal{X} = \mathbb{R}^d$ mean: $\mu \in \mathbb{R}^d$, cov matrix $\Sigma \in \mathbb{S}^d_{++}$ ← positive definite

$$p(x) = \frac{1}{\underbrace{(2\pi)^{d/2}|\Sigma|^{1/2}}_{\text{determinant of } \Sigma}} e^{\underbrace{-\frac{1}{2}\|x-\mu\|^2_\Sigma}_{\text{Mahalanobis distance}}} = N(x|\mu, \Sigma)$$

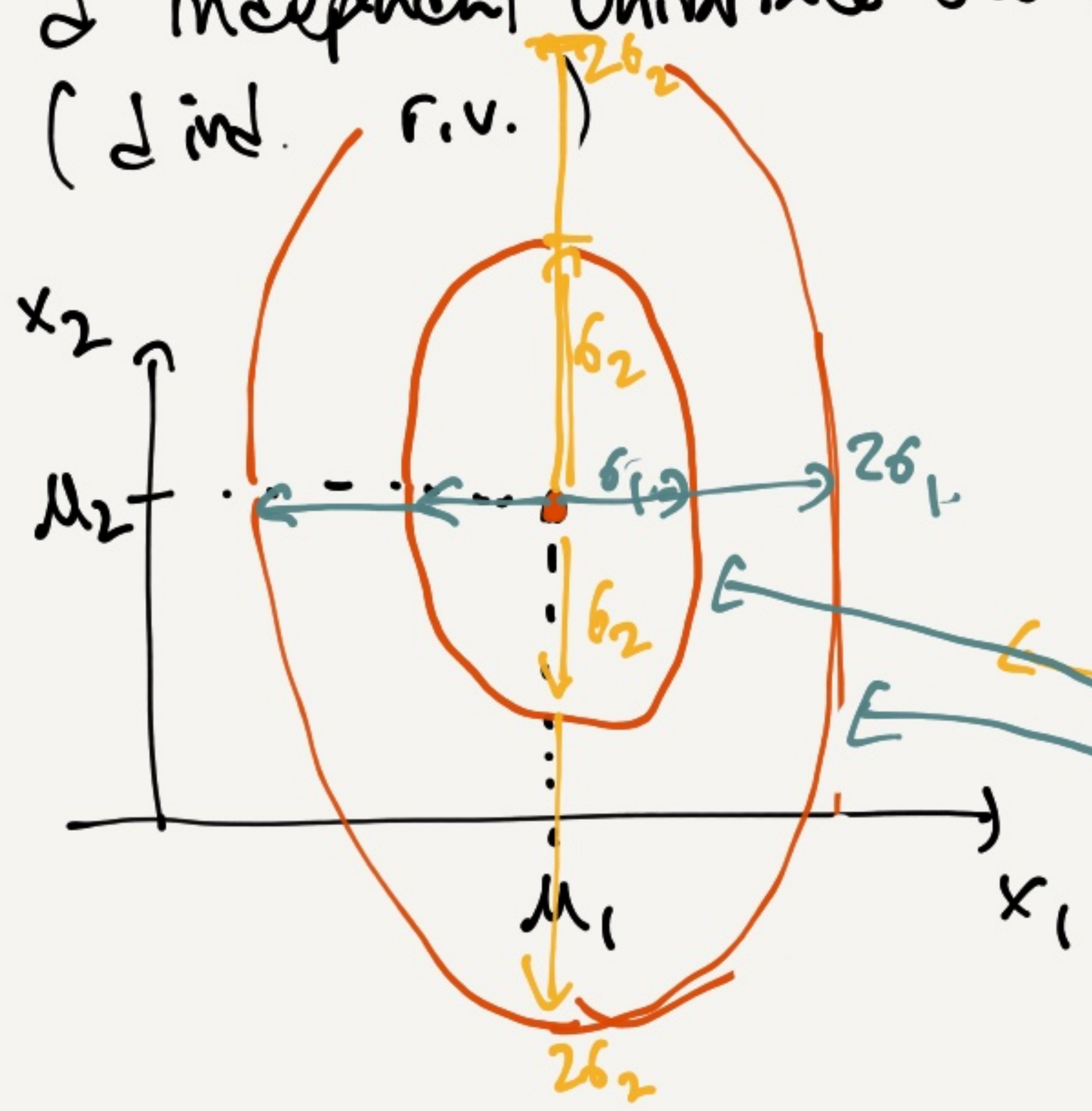
determinant of Σ
"volume of Gaussian

Mahalanobis distance
$\|x-\mu\|^2_\Sigma = (x-\mu)^T \Sigma^{-1}(x-\mu)$

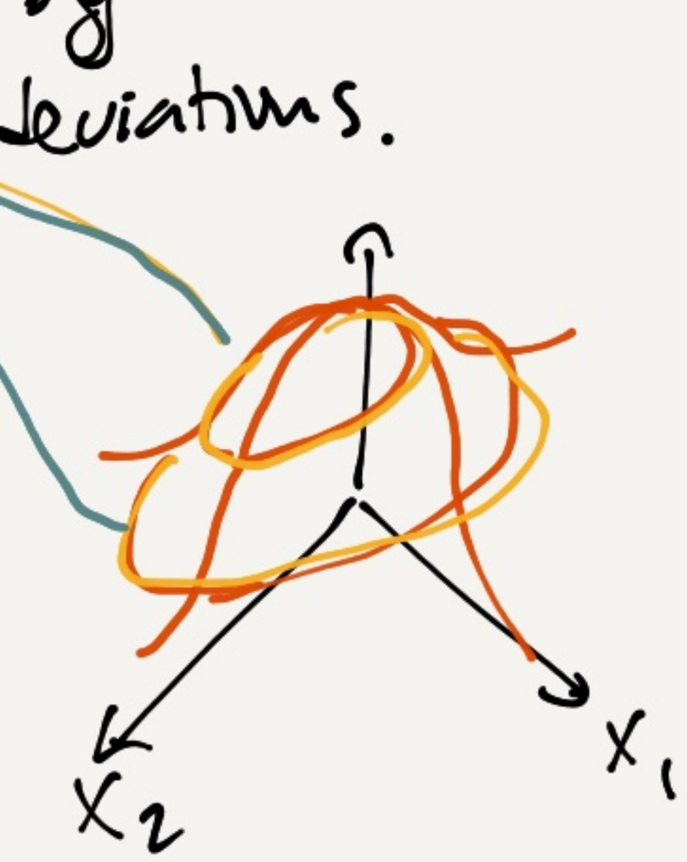
Special cases: Diagonal Σ

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix} \Rightarrow \underbrace{N(x|\mu,\Sigma)}_{\text{joint likelihood}} = \underbrace{\prod_{i=1}^d N(x_i|\mu_i, \sigma_i^2)}_{\text{product of univariate Gaussians}}$$

i.e. d independent univariate Gaussians.
(d ind. r.v.)



isocontours are ellipses w/ width determined by the standard deviations.



Problem 1-6 M.v. Gaussian

$x \in \mathbb{R}^d$, mean $\mu \in \mathbb{R}^d$, cov matrix $\Sigma \in S_{++}^d$

$$p(x) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{\frac{1}{2}}}\, e^{-\frac{1}{2}\underbrace{\|x-\mu\|^2_\Sigma}}$$

Mahalanobis distance
$(x-\mu)^T \Sigma^{-1}(x-\mu)$

a) Σ is diagonal

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix}$$

determinant for diagonal matrix: $|\Sigma| = \prod_{i=1}^{d} \sigma_i^2$

distance: $(x-\mu)^T\Sigma^{-1}(x-\mu) = (x-\mu)^T \begin{bmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_d^2} \end{bmatrix}(x-\mu)$

$$\begin{bmatrix} a & & 0 \\ & b & \\ 0 & & c \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a} & & 0 \\ & \frac{1}{b} & \\ 0 & & \frac{1}{c} \end{bmatrix}$$

$$= \sum_{i=1}^{d} \frac{1}{\sigma_i^2}(x_i-\mu_i)^2$$

substitute into $p(x)$:

$$p(x) = \frac{1}{(2\pi)^{d/2}\left(\prod_{i=1}^{d}\sigma_i^2\right)^{1/2}}\, e^{-\frac{1}{2}\sum_{i=1}^{d}\frac{1}{\sigma_i^2}(x_i-\mu_i)^2}$$

$e^{a+b} = e^a e^b$

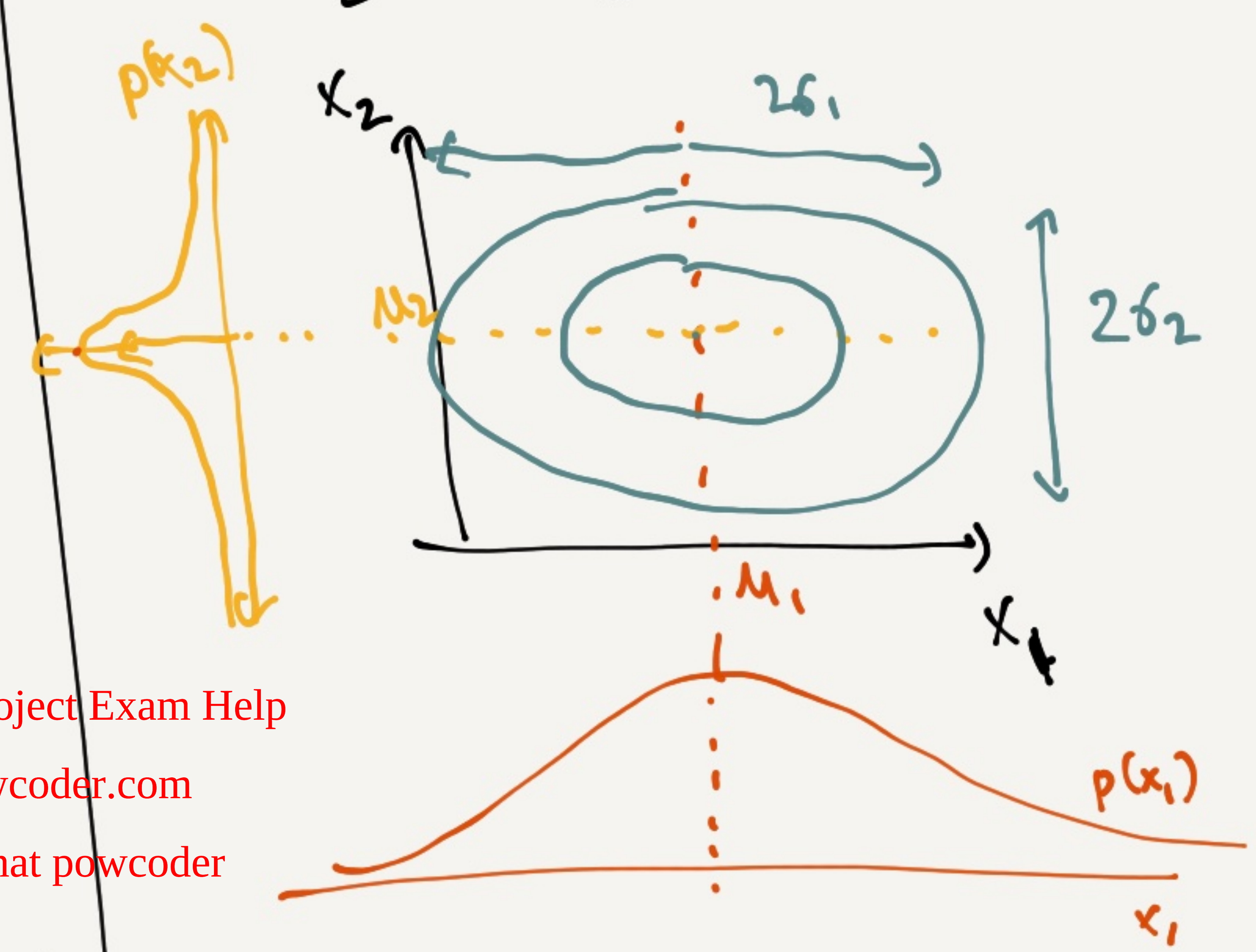
$$p(x) = \prod_{i=1}^{d}\underbrace{\left[\frac{1}{(2\pi)^{1/2}(\sigma_i^2)^{\frac{1}{2}}}\, e^{-\frac{1}{2}\frac{1}{\sigma_i^2}(x_i-\mu_i)^2}\right]}_{N(x_i\,|\,\mu_i,\,\sigma_i^2)}$$

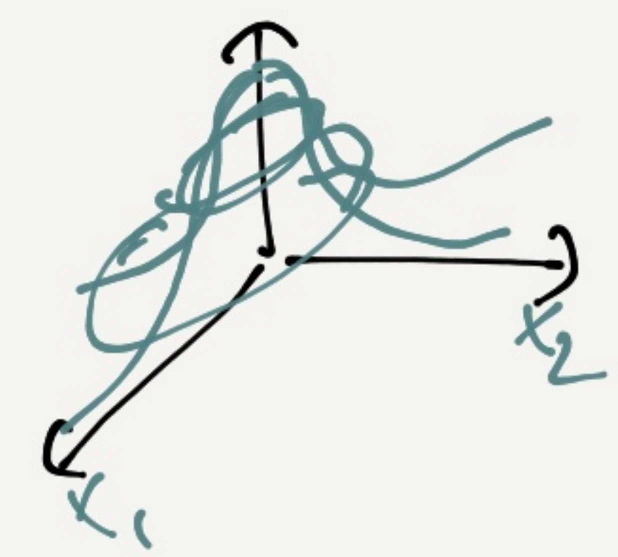
$$p(x) = \prod_{i=1}^{d} N(x_i\,|\,\mu_i, \sigma_i^2) = N(x\,|\,\mu, \Sigma)$$

joint = product of marginals $\Rightarrow$ x_i's are independent.

diagonal covariance $\Rightarrow$ d independent univariate Gaussians.

$$\Sigma = \begin{bmatrix} \ddots & 0 \\ 0 & \ddots \end{bmatrix} \quad \mathrm{cov}(x_i, x_j) = 0$$



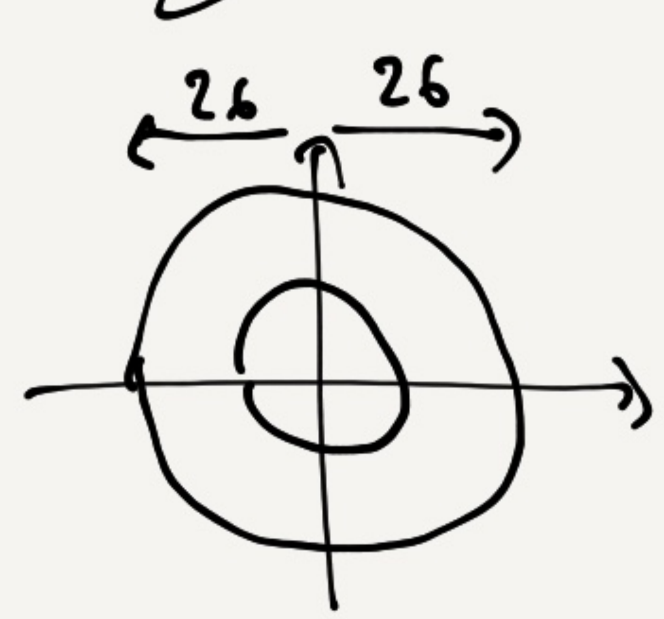


ellipses of the isocontours are aligned w/ the axes.

- All the values are the same in the diagonal

Isotropic Gaussian

$$\Sigma = \sigma^2 I = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}$$



isocontours are circles.

3) general case of Σ

• eigenvalues & eigenvectors.

$$(\underline{\Sigma} - \underset{\uparrow\,\text{eigenvalue}}{\lambda_i} I)\,\underset{\nwarrow\,\text{eigenvector}}{v_i} = 0$$

$\Sigma v_i = \lambda_i v_i$ ⇐ multiplying by vector gives a similar (scaled) vector.

• There are d eigenpairs for Σ:

define $V = [v_1 \cdots v_d]$, $\Lambda = \begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_d\end{bmatrix}$

$$\Sigma V = V\Lambda$$

Note: $\underline{V^T V = I} \Rightarrow V^T = V^{-1}$ (for symmetric matrices)

$$\underline{\Sigma} = \underline{V\Lambda V^T} \quad \text{(eigendecomposition of } \Sigma\text{)}$$

Need the inverse for Mahal distance:

$$\Sigma^{-1} = \underline{(V\Lambda V^T)}^{-1} \qquad (AB)^{-1} = B^{-1}A^{-1}$$

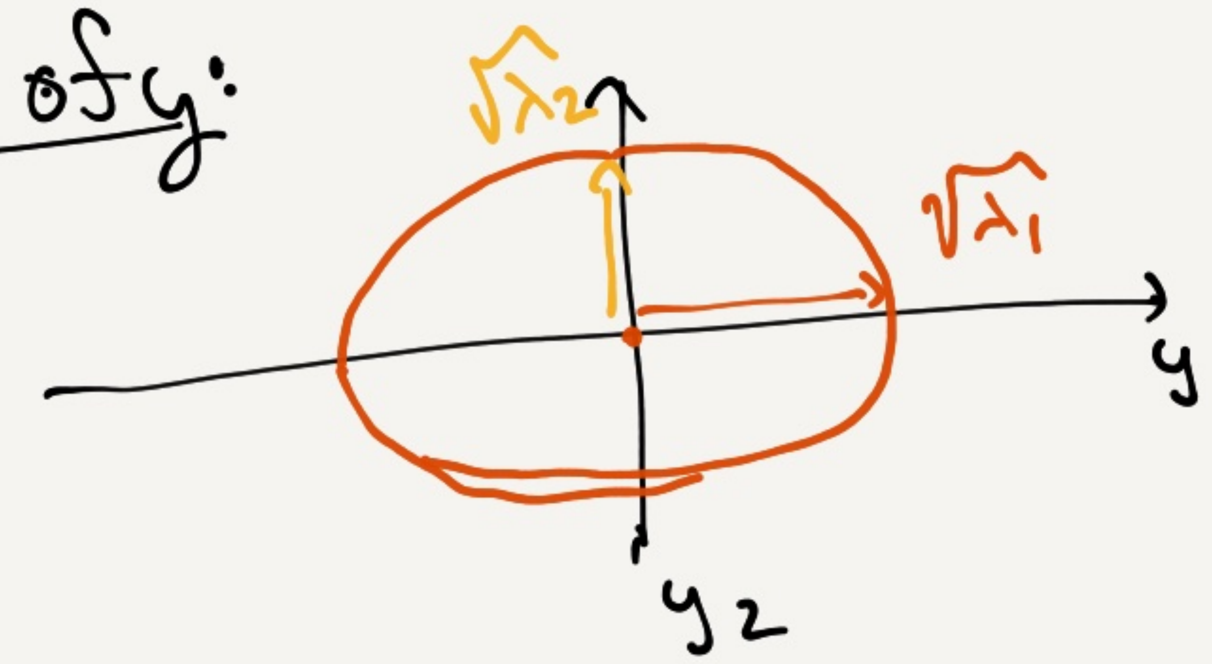
$$= \underbrace{V^{T^{-1}}}_{(V^{-1})^{-1} = V} \Lambda^{-1} \underbrace{V^{-1}}_{V^T}$$

$$\Sigma^{-1} = V\Lambda^{-1}V^T$$

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \underbrace{(x-\mu)^T V}_{y^T}\, \underline{\underline{\Lambda^{-1}}}\, \underbrace{V^T(x-\mu)}_{y}$$

$$= y^T \underset{\text{diagonal}}{\underline{\underline{\Lambda^{-1}}}} y \quad , \quad y = V^T(x-\mu)$$

in the space of y:

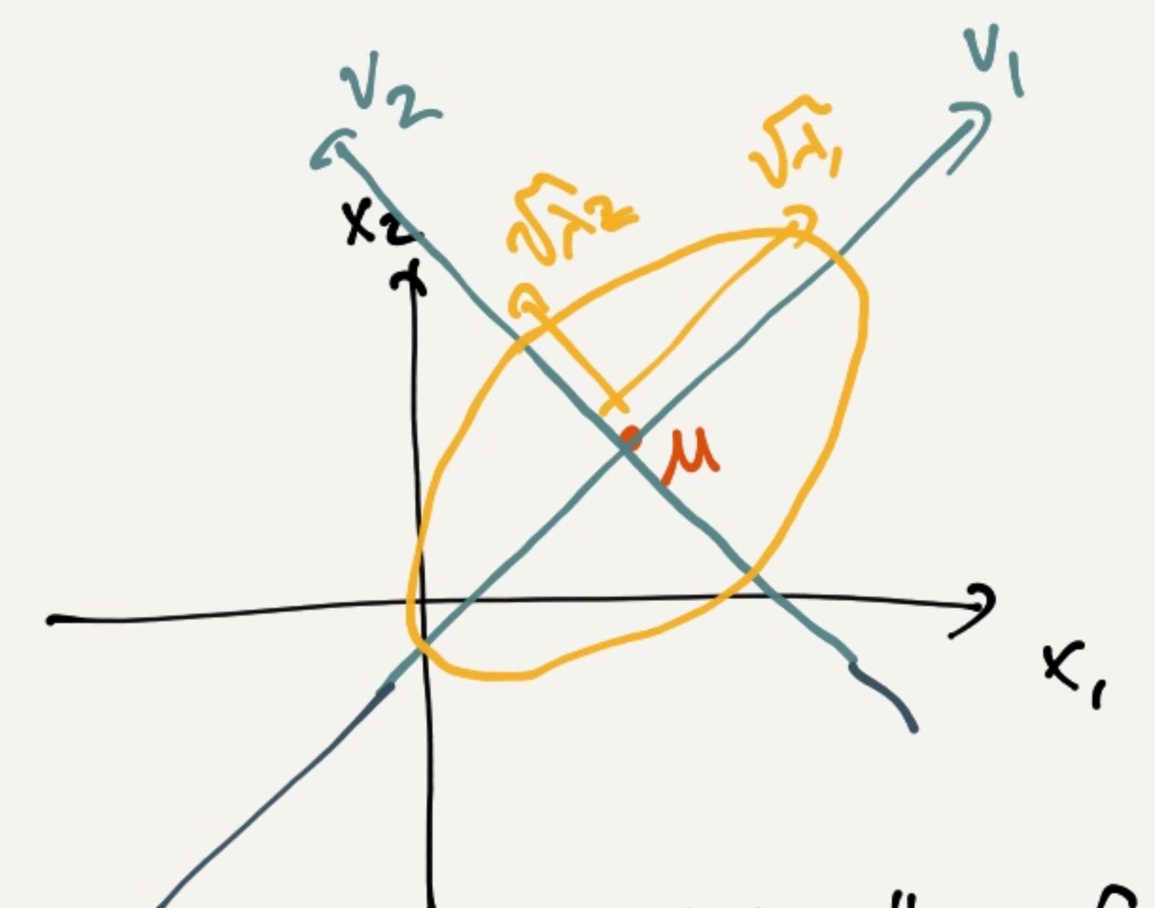


in the space of x:

$$y = V^T(x-\mu)$$

$$Vy = x-\mu$$

$$\Rightarrow x = \underset{\text{rotation matrix}}{\underline{V}}y + \underset{\text{translation}}{\underline{\mu}}$$



rotated ellipse for the iso contours.

(Note: $|\Sigma| = \prod_{i=1}^{d} \lambda_i$)