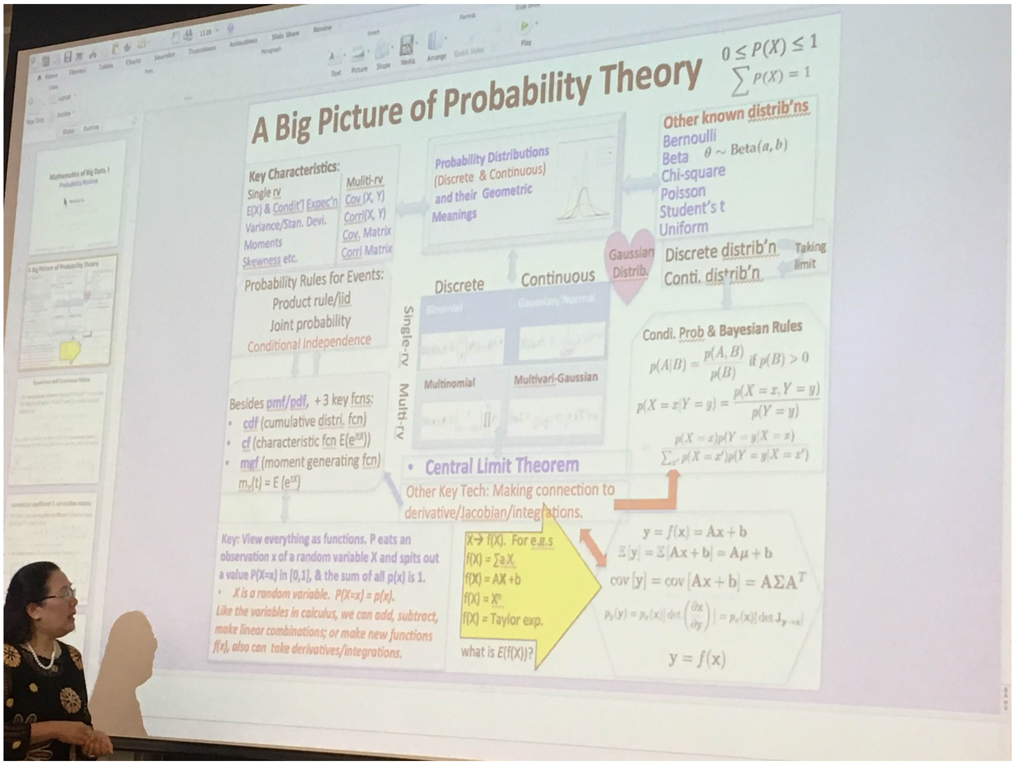
# Mathematics of Big Data I

Review 1: Overview of Probability Theory

September 8th, 2016

## 1 A Big Picture of Probability Theory



A Big Picture of Probability Theory

$0 \le P(X) \le 1$

$\sum P(X) = 1$

Key Characteristics:

Single rv
E(X) & Condit'l Expec'n
Variance/Stan. Devi.
Moments
Skewness etc.

Muliti-rv
Cov (X, Y)
Corr(X, Y)
Cov. Matrix
Corrl Matrix

Probability Distributions (Discrete & Continuous) and their Geometric Meanings

Other known distrib'ns
Bernoulli
Beta $\theta \sim \text{Beta}(a, b)$
Chi-square
Poisson
Student's t
Uniform

Gaussian Distrib.

Discrete distrib'n
Conti. distrib'n
Taking limit

Probability Rules for Events:
Product rule/iid
Joint probability
Conditional Independence

Discrete | Continuous

Single-rv: Binomial | Gaussian/Normal

Multi-rv: Multinomial | Multivari-Gaussian

Condi. Prob & Bayesian Rules

$$p(A|B) = \frac{p(A, B)}{p(B)} \text{ if } p(B) > 0$$

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

$$\frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')}$$

Besides pmf/pdf, +3 key fcns:
- cdf (cumulative distri. fcn)
- cf (characteristic fcn $E(e^{itX})$)
- mgf (moment generating fcn) $m_X(t) = E(e^{tX})$

- Central Limit Theorem

Other Key Tech: Making connection to derivative/Jacobian/integrations.

Key: View everything as functions. P eats an observation x of a random variable X and spits out a value P(X=x) in [0,1], & the sum of all p(x) is 1.
- X is a random variable. P(X=x) = p(x).

Like the variables in calculus, we can add, subtract, make linear combinations; or make new functions f(x), also can take derivatives/integrations.

X→ f(X). For e.g.s
$f(X) = \sum a_i X_i$
$f(X) = AX + b$
$f(X) = X^n$
f(X) = Taylor exp.
what is E(f(X))?

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\mu + \mathbf{b}$$

$$\text{cov}[\mathbf{y}] = \text{cov}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\Sigma\mathbf{A}^T$$

$$p_y(\mathbf{y}) = p_x(\mathbf{x})\left|\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)\right| = p_x(\mathbf{x})|\det \mathbf{J}_{\mathbf{y}\to\mathbf{x}}|$$

$$\mathbf{y} = f(\mathbf{x})$$

The way we understand probability theory is by thinking of everything in terms of a function. We can view $P$, the probability measure, as a function that takes a random variable $X$ as input and outputs a number $P(X = x)$ which lies in the interval $[0, 1]$. The sum of all probabilities $P(x)$ should be one. There are four types of random variables we may come across:

1. **Single-Variable and Discrete**: for example, the binomial distribution.
2. **Single-Variable and Continuous**: for example, the Gaussian/Normal random variable.
3. **Multivariable and Discrete**: for example, the multinomial distribution.
4. **Multivariable and Continuous**: for example, the multivariable Gaussian.

It's important to know how to take transformations of random variables. For instance, we can take linear combinations of them, but we have to know how to modify the original probabilities so that our resulting probability will still have values $P(x) \in [0, 1]$. Let's consider an example of a matrix acting on a vector of random variables. In this case we can let $x_1, \ldots, x_n$ be $n$ random variables, as well as $y_1, \ldots, y_n$. We can write a matrix equation

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

and we can summarize this equation as $A\vec{x} = \vec{y}$. An example in which we might relate random variables in this way is by looking at stock trading.

One of the most important concepts in probability is that of the **expectation value**. If we let $X$ be a random variable, and important quantity to compute is the expectation of the $n$th power of $X$, or rather $E[X^n]$. This is called the **nth moment** of the random variable $X$.

It's also important to know the **conditional probability** of a random variable $X$. For instance, if $A$ and $B$ represent events of a random variable, we would like to know what the probability of $A$ occuring with prior knowledge that $B$ has already occured. We denote this quantity as $P(A|B)$ and this is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where $A \cap B$ is the event that both $A$ and $B$ occur. This gives rise to **Bayes' rules** which gives us a method of computing $P(B|A)$ if we already know $P(A|B)$.

Associated to any random variable we are always guaranteed a probability mass/density function for discrete/continuous random variables. However, there are three additional functions that we can always associate to random variables:

1. **Cumulative Distribution Function**: Given a random variable $X$, the cdf is given by $F_X(t) = P(X \leq t)$, which is valid for both discrete and continuous random variables.
2. **Characteristic Function**: This is defined as $E[e^{itX}]$.
3. **Moment-Generating Function**: This is given by $E[e^{tX}]$.

## 1.1 Covariance of Random Variables

Given two random variables $X$ and $Y$, we define the **covariance** of $X$ and $Y$, denoted by $Cov(X,Y)$ by

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

where $\mu_X = E[X]$ and $\mu_Y = E[Y]$. It's easy to show that if our random variables are independent, then their covariance is zero. and this results from the fact that $E[XY] = E[X]E[Y]$ in the case of independent random variables.

The important case of covariance is when we have vectors of random variables. Instead of dealing individual random variables we can take a vector of random variables

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Then we define the covariance of $\mathbf{X}$ by taking the matrix of cross-covariances of each random variable in the vector. We see that

$$Cov(x_1, x_1) = E[(x_1 - \mu_1)^2] = Var(x_1).$$

and this remains true for any diagonal entry. For cross terms we see that

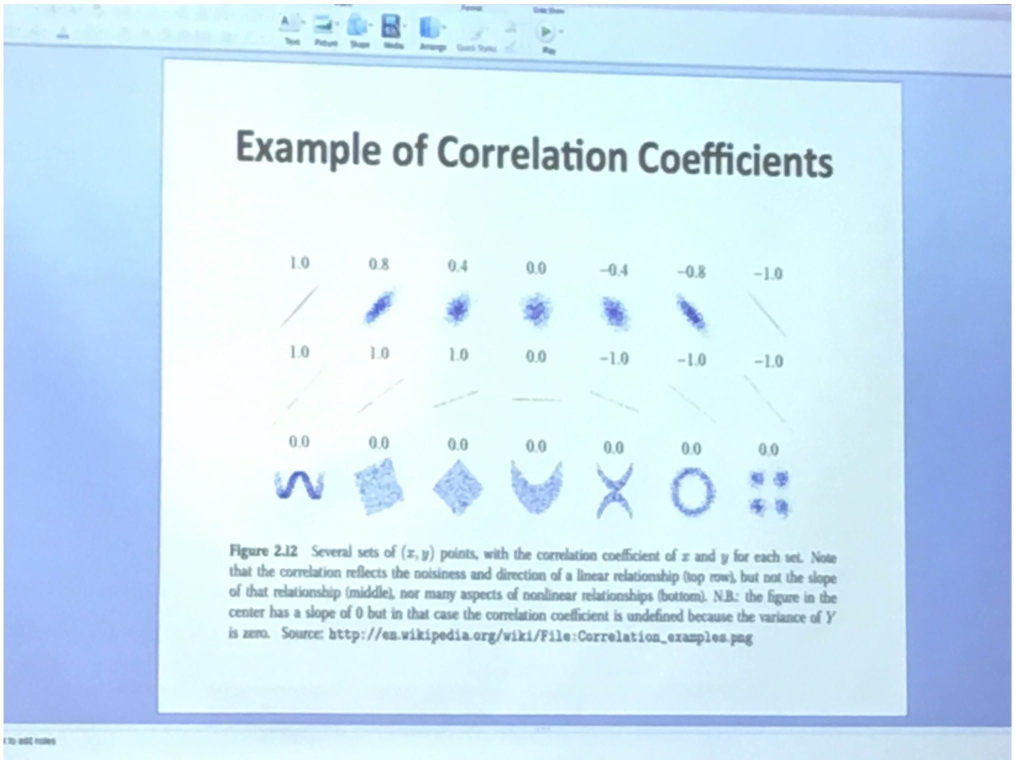
$$
\begin{aligned}
Cov(x_1, x_2) &= E[(x_1 - \mu_1)(x_2 - \mu_2)] \\
&= E[x_1 x_2] - \mu_1 \mu_2 \\
&= Cov(x_2, x_1)
\end{aligned}
$$

and thus we find that this matrix is symmetric. We see that the full covariance matrix is given by

$$
Cov(\mathbf{X}) = \begin{bmatrix} Var(x_1) & Cov(x_1, x_2) & \dots & Cov(x_1, x_n) \\ Cov(x_2, x_1) & Var(x_2) & \dots & Cov(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_n, x_1) & Cov(x_n, x_2) & \dots & Var(x_n) \end{bmatrix}.
$$

Related to covariance, is the **correlation** between random variables, which is defined by

$$
Corr[x, y] \triangleq \frac{Cov(x, y)}{\sqrt{Var(x) \cdot Var(y)}}.
$$



Example of Correlation Coefficients

**Figure 2.12** Several sets of $(x, y)$ points, with the correlation coefficient of $x$ and $y$ for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of $Y$ is zero. Source: `http://en.wikipedia.org/wiki/File:Correlation_examples.png`

An intuitive idea we can get about correlation is how related two variables are: if the correlation is zero then the variables are uncorrelated. If on the other hand the correlation is 1, then the relationship between random variables is linear.

## 1.2 The Multivariate Gaussian

Here we introduce the multivariate Gaussian distribution:

$$
\mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]
$$

where $\boldsymbol{\mu} = E[\mathbf{x}] \in \mathbb{R}^D$ and $\Sigma = \sigma^2 I_D \operatorname{cov}[x]$. Here we note that the covariance in this case is a matrix where $\Sigma^{-1}$ is the matrix inverse, and $|\Sigma|$ is the determinant of $\Sigma$. When the covariance is $\Sigma = \sigma^2 I_D$, we have only one parameter along a diagonal matrix. We call this kind of covariance matrix an **isotropic** or **spherical** covariance.

Recall that the covariance is symmetric and guaranteed to be positive semi-definite. In the case of the multivariate Gaussian we require the covariance to be positive definite (otherwise matrix inversion would be impossible, as well as having zero determinant in the denominator of the normalizing coefficient). There is an important way to interpret covariances geometrically since they are related to quadratic surfaces. We will illustrate this with an example. Let's consider the quadratic form $x^2 + 4xy + y^2$. We can write this in matrix form as

$$x^2 + 4xy + y^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \vec{x}^T A \vec{x}.$$

Another example of this is given by the form $3x^2 + 7xy + 6xz + y^2 - z^2 = 9$. In this case we get a $3 \times 3$ example

$$\begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} 3 & 7/2 & 3 \\ 7/2 & 1 & 0 \\ 3 & 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = 9$$

which we can write as $\vec{y}^T B \vec{y} = 9$. In both of these cases we find that $A$ and $B$ are both symmetric, in which we are guaranteed an orthonormal basis of eigenvectors. Therefore we can diagonalize the matrix $A$ by an orthogonal matrix $P$ (recall that an orthogonal matrix is one satisfying $P^{-1} = P^T$), meaning that $P^T A P = D$ where $D$ is diagonal.

**Definition 1** *An $n \times n$ symmetric matrix $A$ is called **positive-definite** if for all $\vec{x} \in \mathbb{R}^n$ we have $\vec{x}^T A \vec{x} \geq 0$ with equality holding if and only if $\vec{x} = 0$.*

It is a fact of linear algebra (which we encourage you to prove on your own) that $A$ is positive definite, which we write as $A \geq 0$ if and only if all of the eigenvalues of $A$ are positive. Note that if $A$ is positive-definite, we can take the square-root of the matrix $D$ given by $\sqrt{D}$, the matrix where each diagonal entry is the square-root of the corresponding entry in $D$. We can rewrite the matrix $A$ as

$$A = PDP^T = P\sqrt{D}\sqrt{D}P^T = P\sqrt{D}P^T P\sqrt{D}P^T = GG$$

where we define $G = P\sqrt{D}P^T$. Note that $G$ is symmetric so we can write this decomposition as $A = G^T G$. Now in the inner product we have that

$$\vec{x}^T A \vec{x} = \vec{x}^T G^T G \vec{x} = (G\vec{x})^T (G\vec{x}).$$

Now , if we let $\vec{y} = G\vec{x}$ we can reduce our matrix inner product as

$$\vec{x}^T A \vec{x} = \vec{y}^T \vec{y} = ||\vec{y}||^2 \geq 0$$

where $||\vec{y}|| = 0$ if and only if $\vec{y}$ is the zero vector. We can see then that $\vec{x}^T A \vec{x}$ is zero if and only if $\vec{y} = G\vec{x} = 0$ if and only if $\vec{x} = 0$ since we have that $G$ is invertible (recall that all of the entries in $D$ and hence $\sqrt{D}$ are positive). Another way of seeing these relationships is by simply writing $\vec{y} = P^T \vec{x}$, thus we see

$$\vec{x}^T A \vec{x} = \vec{x}^T PDP^T \vec{x} = \vec{y}^T D \vec{y} = \sum_{i=1}^{n} \lambda_i y_i^2$$

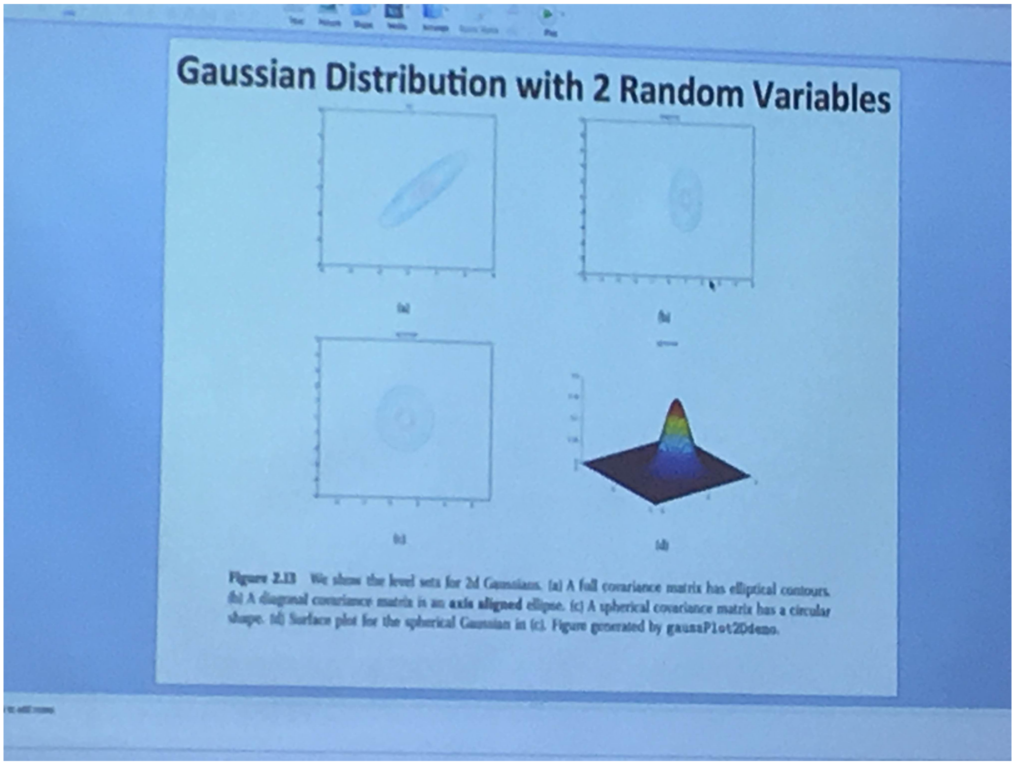
where the $\lambda_i$ are the eigenvalues of $A$, contained as the diagonal entries in $D$, and the $y_i$ are the entries in the vector $\vec{y}$.

If we come back to our example of the quadratic surface in three dimensions, this hints at the idea that we can transform our coordinates to an orthonormal frame in which our quadratic form is simplified. We can for instance take our $3 \times 3$ example above and write $\lambda_1 y_1^2 + \lambda_2 y_2^2 + \lambda_3 y_3^2 = 9$. We can rewrite this equation as

$$\left(\frac{y_1}{3/\sqrt{\lambda_1}}\right)^2 + \left(\frac{y_2}{3/\sqrt{\lambda_2}}\right)^2 + \left(\frac{y_3}{3/\sqrt{\lambda_3}}\right)^2 = 1$$

which is the equation for an ellipsoid. What we conclude from this is the following generalization: **symmetric positive-definite matrices can be interpreted, via quadratic forms, as ellipsoids of various sizes, shapes, and dimensions. The eigenvectors of the matrix correspond to the principle directions of the ellipsoid with the eigenvalues indicating how much the ellipsoid extends or "stretches" in that direction.** Note that while this method generalizes to any number of dimensions, we are only able to visualize the 2 and 3-dimensional cases.

How can we bring this knowledge of positive-definite covariance matrices back to multivariate Gaussian distributions?



Gaussian Distribution with 2 Random Variables

**Figure 2.13** We show the level sets for 2d Gaussians. (a) A full covariance matrix has elliptical contours. (b) A diagonal covariance matrix is an **axis aligned** ellipse. (c) A spherical covariance matrix has a circular shape. (d) Surface plot for the spherical Gaussian in (c). Figure generated by gaussPlot2Ddemo.

In general we can think of positive-definite matrices $A$ as inducing a **quadratic surface** $q(\vec{x}) = \vec{x}^T A\vec{x}$. We can see then that the Gaussian distribution as taking the exponential of this surface $\mathcal{N}(\mu, \Sigma) \sim \exp\left(-\frac{1}{2}q(\vec{x} - \boldsymbol{\mu})\right)$. In the two-dimensional case we can see this rather easily: since the matrix $\Sigma$ is assumed symmetric positive-definite, we automatically have that it's inverse $\Sigma^{-1}$ is as well (the inverse of the covariance is known as the **precision matrix**). We see that the graph of the Gaussian takes an ellipsoid and maps it as a bump in $\mathbb{R}^3$ with it's peak occuring when $\vec{x} = \boldsymbol{\mu}$.

## 1.3 Relationship With Inner Products and Tensors

If we look at the function $e^{-||x-y||^2}$ as a special case of the Gaussian distribution, where we see that the function peaks when $x$ is very close to $y$. By noting that the exponent is actually the Euclidean product $||x - y||^2 = \langle x - y, x - y\rangle$. We're however not constrained to the Euclidean dot product, but can in fact change how we take our inner product.

**Definition 2** *A* ***tensor*** *is a multi-linear map $T : \mathbb{R}^n \times \ldots \times \mathbb{R}^n \to \mathbb{R}$ where there are $k$ copies of $\mathbb{R}^n$ where the function $T$ is linear in each of it's entries.*

We see that the Euclidean dot product $\langle\cdot,\cdot\rangle$ is an example of a 2-tensor since it is a multi-linear map (note in the case of the complex Euclidean dot product we don't exactly have linearity because of the conjugate-symmetry on the second term). The inner product is in fact a symmetric bilinear form.

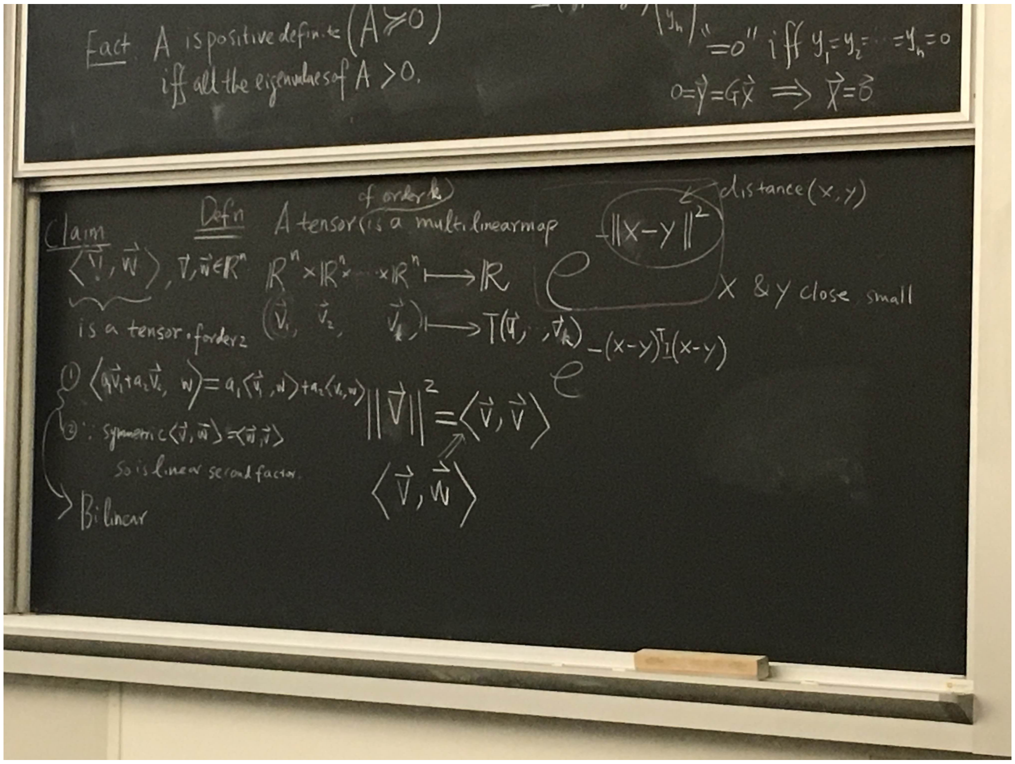
Let's explore why tensors are powerful tools. Let's take two vectors $\vec{v},\vec{w}\in\mathbb{R}^2$. Let $\{\vec{u}_1,\vec{u}_2\}$ be any basis of $\mathbb{R}^2$. We can then write our vectors as

$$\begin{aligned}\vec{v} &= a_{11}\vec{u}_1 + a_{12}\vec{u}_2\\ \vec{w} &= a_{21}\vec{u}_1 + a_{22}\vec{u}_2.\end{aligned}$$

An **inner product** is any symmetric, bilinear, positive-definite form $\langle\cdot,\cdot\rangle:\mathbb{R}^2\times\mathbb{R}^2\to\mathbb{R}$, where the positive-definiteness means that $\langle\vec{x},\vec{x}\rangle\geq 0$ and equality holds if and only if $\vec{x}=0$. The power behind the inner product is that in using our example vectors above, we have that

$$\begin{aligned}\langle\vec{v},\vec{w}\rangle &= \langle a_{11}\vec{u}_1 + a_{12}\vec{u}_2,\ a_{21}\vec{u}_1 + a_{22}\vec{u}_2\rangle\\ &= a_{11}\langle\vec{u}_1, a_{21}\vec{u}_1 + a_{22}\vec{u}_2\rangle + a_{12}\langle\vec{u}_2, a_{21}\vec{u}_1 + a_{22}\vec{u}_2\rangle\\ &= a_{11}a_{21}\langle\vec{u}_1,\vec{u}_1\rangle + a_{11}a_{22}\langle\vec{u}_1,\vec{u}_2\rangle + a_{12}a_{21}\langle\vec{u}_2,\vec{u}_1\rangle + a_{12}a_{22}\langle\vec{u}_2,\vec{u}_2\rangle\\ &= \begin{bmatrix} a_{11} & a_{12}\end{bmatrix}\begin{bmatrix}\langle\vec{u}_1,\vec{u}_1\rangle & \langle\vec{u}_1,\vec{u}_2\rangle\\ \langle\vec{u}_2,\vec{u}_1\rangle & \langle\vec{u}_2,\vec{u}_2\rangle\end{bmatrix}\begin{bmatrix} a_{21}\\ a_{22}\end{bmatrix}.\end{aligned}$$

What we discern from this is that we can understand the entire inner product on any two vectors by simply specifying a basis for the space on which the tensor acts, and compute the cross-inner products on the basis elements. We can therefore compute the inner product on any two vectors in the space by representing them in the given basis and performing the indicated matrix multiplication given above. We see in the $2\times 2$ case, we can determine the rule for the inner product simply by computing 3 inner products. In the case of an $n\times n$ matrix, we need only compute $\frac{n(n+1)}{2}$ of these inner products. It is quite easy to tell immediately if the form is symmetric from it's matrix representation, but how do we determine whether it is positive-definite? The most direct method is to compute it's eigenvalues, which is a standard task for any computer given that the size $n$ of the matrix is of reasonable size.



Fact A is positive definite $(A \succcurlyeq 0)$ iff all the eigenvalues of $A > 0$.

$\begin{pmatrix} \vdots \\ y_n \end{pmatrix}$ "$=0$" iff $y_1 = y_2 = \cdots = y_n = 0$

$0 = \vec{y} = G\vec{x} \implies \vec{x} = \vec{0}$

Claim $\underbrace{\langle \vec{v}, \vec{w} \rangle}$, $\vec{v}, \vec{w} \in \mathbb{R}^n$ is a tensor of order 2

① $\langle a_1\vec{v}_1 + a_2\vec{v}_2, w \rangle = a_1\langle \vec{v}_1, w \rangle + a_2\langle v_2, w \rangle$

② ∴ symmetric $\langle \vec{v}, \vec{w} \rangle = \langle \vec{w}, \vec{v} \rangle$

So is linear second factor.

→ Bilinear

Defn A tensor (of order k) is a multilinear map $\mathbb{R}^n \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n \longmapsto \mathbb{R}$

$(\vec{v}_1, \vec{v}_2, \ \vec{v}_k) \longmapsto T(\vec{v}_1, \cdots, \vec{v}_k)$

$\|\vec{v}\|^2 = \langle \vec{v}, \vec{v} \rangle \Leftarrow \langle \vec{v}, \vec{w} \rangle$

$e^{-\|x-y\|^2}$ distance$(x,y)$

x & y close small

$e^{-(x-y)^T I (x-y)}$

Coming back to the Gaussian distribution we have that the argument of the exponent is given by $(\vec{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\vec{x} - \boldsymbol{\mu})$, and so this indicates a symmetric positive definite bilinear form, where the rule for for the form is given by the inverse of the covariance matrix. This metric given by the inverse of the covariance is known as the **Mahalanobis metric**.