1 Mathematical Basics

1.1 Information Theory

The quantification of information content is contingent upon the probability distribution. We seek a measure that expresses the information content in a manner that is inversely related to its probability. Furthermore, the information gain from observing two independent events should equal the sum of the information gains obtained from observing each event individually. That is:

$$h(x, y) = h(x) + h(y)$$

Meanwhile, their joint probability is given by the product of their individual probabilities:

$$p(x, y) = p(x)p(y)$$

From these premises, we can deduce that the relationship between information content and probability is expressed as:

$$h(x) = -\log_2 p(x)$$

The negative sign ensures the non-negativity of information.

Definition 1.1.1. *Therefore, the average information content of a transmission process is obtained by taking the expectation with respect to $p(x)$:*

$$H[x] = -\sum_x p(x) \log_2 p(x)$$

This quantity is referred to as the entropy of the random variable x.

For continuous variable x, we define:

$$H[x] = -\int p(x) \ln p(x)\, dx$$

as the differential entropy of continuous variables.

Theorem 1.1.1. *Noiseless coding theorem: Entropy serves as a lower bound on the number of bits required to transmit the state of a random variable.*

Indeed, this calculates the same fundamental concept as Huffman's minimum path encoding. Both entropy and Huffman coding address the theoretical limit and practical achievement of efficient information representation.

For discrete distributions, the maximum entropy configuration occurs when the variable's probabilities are uniformly distributed among all possible states. For continuous variables, we proceed with the following analysis:

$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$
$$\int_{-\infty}^{\infty} xp(x)\, dx = \mu$$
$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx = \sigma^2$$

We can use the method of Lagrange multipliers to compute the constrained maximum.

$$\mathcal{L}(x) = -\int_{-\infty}^{\infty} p(x) \ln p(x)\, dx + \lambda_1 \left(\int_{-\infty}^{\infty} p(x)\, dx - 1\right) + \lambda_2 \left(\int_{-\infty}^{\infty} xp(x)\, dx - \mu\right) + \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx - \sigma^2\right)$$

$$\begin{aligned} \frac{\delta \mathcal{L}}{\delta p} &= \frac{\delta}{\delta p}[-p \ln p + \lambda_1 p + \lambda_2 xp + \lambda_3 (x-\mu)^2 p] \\ &= -\ln p - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0 \end{aligned}$$

Key Rule: For a functional $\mathcal{L} = \int F(p,x)dx$, the variational derivative $\frac{\delta \mathcal{L}}{\delta p}$ equals the partial derivative of the integrand F with respect to p (i.e., $\frac{\partial F}{\partial p}$).

$$\text{Let } \frac{\partial}{\partial p(x)} F(p) = -\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0$$

$$\begin{aligned} p(x) &= e^{\left(-1+\lambda_1+\lambda_2 x+\lambda_3(x-\mu)^2\right)} \\ &= e^{-1+\lambda_1} \cdot e^{\lambda_2 x+\lambda_3(x-\mu)^2} = Ce^{\lambda_2 x+\lambda_3(x-\mu)^2} \\ &= Ce^{\lambda_3\left(x^2-2\left(\mu-\frac{\lambda_2}{2\lambda_3}\right)x+\mu^2\right)} = Ce^{\lambda_3\left(x-\mu+\frac{\lambda_2}{2\lambda_3}\right)^2} \end{aligned}$$

Since $p(x) > 0$, so $C > 0$.$p(x)$ is symmetric about $\mu - \frac{\lambda_2}{2\lambda_3}$, so $\mathbb{E}[p(x)] = \mu - \frac{\lambda_2}{2\lambda_3} = \mu$. It follows that:

$$\lambda_2 = 0$$

The form thus becomes:

$$p(x) = Ce^{\lambda_3(x-\mu)^2}$$

Finally, substituting in the constraint conditions gives the answer:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Here we supplement some necessary knowledge of functional analysis. A **function** represents a mapping from numbers to numbers, while a **functional** represents a mapping from functions to numbers.

For instance, consider the functional $J(y) = \int_{x_1}^{x_2} \sqrt{1+(y_x)^2}\, dx$. The goal is to find a suitable function $y(x)$ such that $J(y)$ attains its minimum.

Generally, we define a functional concerned with y (which is twice differentiable on the interval $[a,b]$):

$$J(y) = \int_a^b F(x, y, y_x)\, dx$$

Given that $y(a)$ and $y(b)$ are known, and that F is twice differentiable with respect to all its arguments, what condition must the function $y(x)$ satisfy in this general case for the functional to attain a minimum value?

For the general case described above, when the functional $J(y)$ attains an extremum, the function $y(x)$ must satisfy the Euler-Lagrange equation:

Theorem 1.1.2. ***Euler-Lagrange Equation:*** *Let $J(y)$ be a functional defined by*

$$J(y) = \int_a^b F(x, y, y_x)dx$$

where F is twice continuously differentiable in all its arguments. If $y = u(x)$ yields an extremum of $J(y)$ among all functions satisfying the boundary conditions $y(a) = y_0$ and $y(b) = y_1$, then $u(x)$ must satisfy the following necessary condition:

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_x}\right) = 0$$

Proof. We assume that the function yielding an extremum for the functional $J(y)$ is $y = u(x)$. On the interval $[x_1 = a, x_2 = b]$, we define a family of functions:

$$y(x) = u(x) + \epsilon\eta(x)$$

where ϵ is a real parameter and $\eta(x)$ is defined as the difference between any other function $y(x)$ connecting points P_1 and P_2 and the function $u(x)$. Consequently, $\eta(x)$ must satisfy the boundary conditions $\eta(a) = \eta(b) = 0$, ensuring $y(x)$ also satisfies these conditions and represents an admissible path. For any fixed $\eta(x)$, the functional $J(y) = J(u + \epsilon\eta)$ becomes a function of the parameter ϵ only, denoted $J(\epsilon)$. Since $y(x) = u(x)$ when $\epsilon = 0$ (where $J(y)$ attains its extremum), it follows that:

$$\left.\frac{dJ(\epsilon)}{d\epsilon}\right|_{\epsilon=0} = 0$$

To satisfy this condition, we first compute $\frac{dJ(\epsilon)}{d\epsilon}$ for an arbitrary but fixed $\eta(x)$:

$$\begin{aligned}
\frac{dJ(\epsilon)}{d\epsilon} &= \frac{d}{d\epsilon}\int_a^b F(x, y(x), y_x(x))dx \\
&= \int_a^b \frac{\partial F(x, y(x), y_x(x))}{\partial \epsilon}dx \\
&= \int_a^b \left(\frac{\partial F}{\partial y}\frac{\partial y}{\partial \epsilon} + \frac{\partial F}{\partial y_x}\frac{\partial y_x}{\partial \epsilon}\right)dx \\
&= \int_a^b \frac{\partial F}{\partial y}\eta(x)dx + \int_a^b \frac{\partial F}{\partial y_x}\eta'(x)dx \\
&= \int_a^b \frac{\partial F}{\partial y}\eta(x)dx + \int_a^b \frac{\partial F}{\partial y_x}d\eta(x) \\
&= \int_a^b \frac{\partial F}{\partial y}\eta(x)dx + \left[\eta(x)\frac{\partial F}{\partial y_x}\right]_a^b - \int_a^b \frac{d}{dx}\left(\frac{\partial F}{\partial y_x}\right)\eta(x)dx \quad \text{(Integration by Parts)} \\
&= \int_a^b \left[\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_x}\right)\right]\eta(x)dx \quad \text{(since } \eta(a) = \eta(b) = 0\text{)}
\end{aligned}$$

Evaluating this derivative at $\epsilon = 0$ (where $y = u$ and $y_x = u_x$) yields:

$$\left.\frac{dJ(\epsilon)}{d\epsilon}\right|_{\epsilon=0} = \int_a^b \left[\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right)\right]\eta(x)dx = 0$$

Given that $\eta(x)$ is arbitrary and the expression $\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right)$ is a fixed function for given F, the integral can only be zero for all $\eta(x)$ if:

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right) = 0$$

This completes the proof. The function $u(x)$ that extremizes the functional $J(y)$ must satisfy this Euler-Lagrange equation.

□

Therefore, the distribution that maximizes the differential entropy is the **Gaussian distribution**. If we compute the differential entropy of the Gaussian distribution, we obtain:

$$H[x] = \frac{1}{2}(1 + \ln(2\pi\sigma^2))$$

It is noteworthy that, unlike discrete entropy, differential entropy can be negative. $H(x) < 0$, when $\sigma^2 < \frac{1}{2\pi e}$.

Definition 1.1.2. *To describe the average additional information required to encode the values of x, we define the **relative entropy**, also called Kullback and Leibler divergence:*

$$KL(p \parallel q) = -\int p(x)\ln q(x)d\,x - (-\int p(x)\ln p(x)d\,x) = -\int p(x)\ln\frac{q(x)}{p(x)}d\,x = -\mathbb{E}_p[\ln\frac{q(x)}{p(x)}]$$

Theorem 1.1.3. ***Jensen's inequality***
For any set of points $\{x_i\}$, the convex function $f(x)$ must satisfy:

$$f(\sum_{i=1}^{M} \lambda_i x_i) \leqslant \sum_{i=1}^{M} \lambda_i f(x_i)$$

where $\lambda_i \geqslant 0$ and $\sum_{i=1}^{M} \lambda_i = 1$.

$$\mathrm{KL}(p \parallel q) = -\int p(x) \ln \frac{q(x)}{p(x)} d\,x \geqslant -\ln \int q(x) d\,x = 0$$

Theorem 1.1.4. *Consider two variables x and y with joint distribution $p(x, y)$. The differential entropy of this pair of variables satisfies:*

$$H[x, y] \leqslant H[x] + H[y]$$

Proof. We prove this inequality using the non-negativity property of the Kullback-Leibler (KL) divergence. Consider the KL divergence between the joint distribution $p(x, y)$ and the product of marginal distributions $p(x)p(y)$:

$$\begin{aligned}
0 &\leqslant \mathrm{KL}(p(x,y) \| p(x)p(y)) \\
&= \iint p(x,y) \ln \left(\frac{p(x,y)}{p(x)p(y)} \right) dxdy \\
&= \iint p(x,y) \ln p(x,y) dxdy - \iint p(x,y) \ln[p(x)p(y)] dxdy \\
&= \iint p(x,y) \ln p(x,y) dxdy - \iint p(x,y) [\ln p(x) + \ln p(y)] dxdy \\
&= \iint p(x,y) \ln p(x,y) dxdy - \iint p(x,y) \ln p(x) dxdy - \iint p(x,y) \ln p(y) dxdy
\end{aligned}$$

Now we recognize that these integrals correspond to differential entropy terms:

$$\begin{aligned}
\iint p(x,y) \ln p(x,y) dxdy &= -H[x,y] \\
\iint p(x,y) \ln p(x) dxdy &= \int \ln p(x) \left[\int p(x,y) dy \right] dx = \int p(x) \ln p(x) dx = -H[x] \\
\iint p(x,y) \ln p(y) dxdy &= \int \ln p(y) \left[\int p(x,y) dx \right] dy = \int p(y) \ln p(y) dy = -H[y]
\end{aligned}$$

Substituting these expressions back into the inequality:

$$\begin{aligned}
0 &\leqslant -H[x,y] - (-H[x]) - (-H[y]) \\
0 &\leqslant -H[x,y] + H[x] + H[y]
\end{aligned}$$

Rearranging terms gives the desired result:

$$H[x, y] \leqslant H[x] + H[y]$$

Equality holds if and only if $p(x, y) = p(x)p(y)$ almost everywhere, which occurs when x and y are statistically independent. □

Suppose the data is generated from an unknown distribution $p(x)$ that we wish to model. We can attempt to approximate this distribution using a parametric distribution $q(x \mid \boldsymbol{\theta})$ controlled by a set of adjustable parameters $\boldsymbol{\theta}$. One method to determine $\boldsymbol{\theta}$ is to minimize the Kullback-Leibler divergence between $p(x)$ and $q(x \mid \boldsymbol{\theta})$. However, since we do not

know $p(x)$, this cannot be done directly. Nevertheless, assuming we have observed a finite set of training points—x_n generated from $p(x)$ for $n = 1, \ldots, N$—we can approximate the expectation with respect to $p(x)$ via a finite sum over these training points:

$$\mathrm{KL}(p\|q) \approx \frac{1}{N} \sum_{n=1}^{N} \left(-\ln q\left(x_n \mid \boldsymbol{\theta}\right) + \ln p\left(x_n\right)\right)$$

The second term on the right-hand side of the equation is independent of $\boldsymbol{\theta}$. The first term is the **negative log-likelihood function** of the distribution $q(x \mid \boldsymbol{\theta})$ evaluated using the training set. Therefore, it can be seen that minimizing this Kullback-Leibler divergence is equivalent to maximizing the log-likelihood function.

Definition 1.1.3. *Suppose x is given, the average additional information to know y is:*

$$H[y|x] = -\iint p(y,x) \ln p(y|x) d\, y d\, x$$

as the conditional entropy of y given x.

When two variables x and y are independent, their joint distribution can be factorized into the product of their marginal distributions, i.e., $p(x, y) = p(x)p(y)$. If these two variables are not independent, one can assess how "close" they are to independence by considering the Kullback-Leibler divergence between the joint distribution and the product of the marginal distributions, which is given by:

Definition 1.1.4.

$$I[x,y] \equiv \mathrm{KL}(p(x,y)\|p(x)p(y)) = -\iint p(x,y) \ln \left(\frac{p(x)p(y)}{p(x,y)}\right) dx\, dy$$

*$I[x, y]$ is called the **mutual information** between variables x and y. From the properties of the Kullback-Leibler divergence, we see that $I[x, y] \geqslant 0$, with equality holding if and only if x and y are independent. Using the sum and product rules of probability, we can see that the mutual information is related to the conditional entropy.*

$$I[x,y] = H[x] - H[x|y] = H[y] - H[y|x]$$

Therefore, the mutual information represents the reduction in uncertainty about x due to being informed of the value of y (or vice versa). From a Bayesian perspective, $p(x)$ can be viewed as the prior distribution of x, and $p(x|y)$ as the posterior distribution after observing new data y. In summary, the mutual information represents the reduction in uncertainty about x resulting from the new observation y.

The Bayesian approach to parameter estimation incorporates prior knowledge through the **maximum a posteriori** (MAP) framework. We express our belief about the parameters $\mathbf{w}$ before observing the data by specifying a **prior distribution** $p(\mathbf{w})$. A common and convenient choice is the Gaussian prior, which encodes a preference for smaller parameter values:

$$w_j \sim \mathcal{N}(0, \lambda^{-1}), \quad \text{for all } j.$$

This distributional assumption reflects the belief that the model weights should not be too large, with the hyperparameter λ controlling the strength of this belief.

The core objective is to find the parameter values $\mathbf{w}$ that maximize the **posterior probability** $p(\mathbf{w} \mid X, Y)$. According to Bayes' theorem, the posterior is proportional to the product of the likelihood and the prior:

$$p(\mathbf{w} \mid X, Y) \propto p(Y \mid X, \mathbf{w})\, p(\mathbf{w}).$$

Maximizing the posterior probability is equivalent to minimizing its negative logarithm:

$$-\ln p(\mathbf{w} \mid X, Y) = -\ln p(Y \mid X, \mathbf{w}) - \ln p(\mathbf{w}) + \text{constant}.$$

Substituting the Gaussian likelihood for the regression model, $p(Y \mid X, \mathbf{w}) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \mathbf{w}^T \mathbf{x}_i, \sigma^2)$, and the Gaussian prior, $p(\mathbf{w}) = \prod_{j=1}^{D} \mathcal{N}(w_j \mid 0, \lambda^{-1})$, yields the following expression:

$$-\ln p(\mathbf{w} \mid X, Y) = \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2 + \frac{\lambda}{2} \sum_{j=1}^{D} w_j^2 + \text{constant}.$$

The right-hand side of this equation is precisely the standard **regularized loss function** used in machine learning, composed of a mean squared error loss and an L2 regularization term:

$$J(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2 + \frac{\lambda}{2} \sum_{j=1}^{D} w_j^2.$$

From this Bayesian perspective, the regularization term $\frac{\lambda}{2} \sum_j w_j^2$ is not an ad-hoc penalty but arises naturally from the negative log of the Gaussian prior. It penalizes large weights, thereby encouraging a simpler, smoother model that is less prone to overfitting.

1.2 Probability Theory

The **convolution formula** is a fundamental result in probability theory that provides the probability density function (PDF) for the sum of two independent continuous random variables.

Theorem 1.2.1. *Let X and Y be two independent continuous random variables with probability density functions $f_X(x)$ and $f_Y(y)$, respectively. Then the probability density function of $Z = X + Y$ is given by the convolution of f_X and f_Y:*

$$f_Z(z) = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy$$

An equivalent form, obtained by the substitution $x = z - y$, is:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx$$

Proof.

$$\begin{aligned} F_Z(z) &= P(Z \leqslant z) = P(X + Y \leqslant z) \\ &= \iint_{x+y \leqslant z} f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z-y} f_X(x) dx \right] f_Y(y) dy \\ &= \int_{-\infty}^{\infty} F_X(z - y) f_Y(y) dy \end{aligned}$$

The PDF of Z is the derivative of its CDF. Assuming the functions are sufficiently smooth (a condition validated by the Leibniz integral rule), we can differentiate under the integral sign:

$$\begin{aligned} f_Z(z) = \frac{d}{dz} F_Z(z) &= \frac{d}{dz} \left[\int_{-\infty}^{\infty} F_X(z - y) f_Y(y) dy \right] \\ &= \int_{-\infty}^{\infty} \frac{d}{dz} \left[F_X(z - y) \right] f_Y(y) dy \end{aligned}$$

By the chain rule, $\frac{d}{dz} F_X(z - y) = F_X'(z - y) \cdot \frac{d}{dz}(z - y) = f_X(z - y) \cdot 1$. Substituting this yields the final result:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy$$

□

Theorem 1.2.2. ***Leibniz Integral Rule***

*Let $I(z) = \int_{a(z)}^{b(z)} F(x,z)dx$, where the functions $a(z)$ and $b(z)$ are **differentiable** with respect to z, and the integrand $F(x,z)$ together with its **partial derivative** $\frac{\partial F}{\partial z}(x,z)$ are **continuous** functions over the region of integration. Under these conditions, the derivative of the integral is given by:*

$$I(z) = \int_{a(z)}^{b(z)} F(x,z)dx$$

where the parameter z appears in both the integration limits and the integrand. The derivative of this integral with respect to z is given by:

$$\frac{d}{dz}I(z) = \int_{a(z)}^{b(z)} \frac{\partial F}{\partial z}(x,z)dx + F(b(z),z)\cdot\frac{db}{dz} - F(a(z),z)\cdot\frac{da}{dz}$$

Theorem 1.2.3. *Let X and Y be random variables with joint distribution $p(x,y)$.*

1. *The law of total expectation states:*
$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$$
where $\mathbb{E}_X[X|Y]$ is the expectation of X under the conditional distribution $p(x|y)$.
2. *The law of total variance states:*
$$\text{var}[X] = \mathbb{E}_Y[\text{var}_X[X|Y]] + \text{var}_Y[\mathbb{E}_X[X|Y]]$$
where $\text{var}_X[X|Y]$ is the conditional variance of X given Y.

Proof. We begin with the definition of expectation and express it in terms of the joint distribution:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} xp(x)dx \\
&= \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} p(x,y)dy\right]dx \quad \text{(by the law of total probability)} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xp(x,y)dxdy
\end{aligned}$$

The conditional expectation of X given $Y = y$ is defined as:

$$\mathbb{E}_X[X|Y=y] = \int_{-\infty}^{\infty} xp(x|y)dx = \frac{\int_{-\infty}^{\infty} xp(x,y)dx}{p(y)}$$

where $p(y) = \int_{-\infty}^{\infty} p(x,y)dx$ is the marginal distribution of Y.

Now we take the expectation of this conditional expectation over Y:

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}_X[X|Y]] &= \int_{-\infty}^{\infty} \mathbb{E}_X[X|Y=y]p(y)dy \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} xp(x|y)dx\right]p(y)dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\frac{p(x,y)}{p(y)}p(y)dxdy \quad (\text{since } p(x|y) = \frac{p(x,y)}{p(y)}) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xp(x,y)dxdy \\
&= \mathbb{E}[X]
\end{aligned}$$

This completes the proof of the law of total expectation.

We begin with the definition of variance: $\text{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

We will prove the result by showing that the right-hand side equals this expression.

First, consider the term $\mathbb{E}_Y[\text{var}_X[X|Y]]$:

$$\begin{aligned}
\mathbb{E}_Y[\text{var}_X[X|Y]] &= \int_{-\infty}^{\infty} \text{var}_X[X|Y=y]p(y)dy \\
&= \int_{-\infty}^{\infty} \left[\mathbb{E}_X[X^2|Y=y] - (\mathbb{E}_X[X|Y=y])^2\right] p(y)dy \\
&= \int_{-\infty}^{\infty} \mathbb{E}_X[X^2|Y=y]p(y)dy - \int_{-\infty}^{\infty} (\mathbb{E}_X[X|Y=y])^2 p(y)dy \\
&= \mathbb{E}[X^2] - \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2]
\end{aligned}$$

Next, consider the term $\text{var}_Y[\mathbb{E}_X[X|Y]]$:

$$\begin{aligned}
\text{var}_Y[\mathbb{E}_X[X|Y]] &= \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2] - (\mathbb{E}_Y[\mathbb{E}_X[X|Y]])^2 \\
&= \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2] - (\mathbb{E}[X])^2 \quad \text{(by the law of total expectation)}
\end{aligned}$$

Now we add the two terms:

$$\begin{aligned}
&\mathbb{E}_Y[\text{var}_X[X|Y]] + \text{var}_Y[\mathbb{E}_X[X|Y]] \\
&= \left[\mathbb{E}[X^2] - \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2]\right] + \left[\mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2] - (\mathbb{E}[X])^2\right] \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&= \text{var}[X]
\end{aligned}$$

□

Theorem 1.2.4. *Let* $\mathbf{x}$ *be a continuous random vector with probability density function* $p(\mathbf{x})$ *and differential entropy* $H[\mathbf{x}]$*. Suppose we apply a nonsingular linear transformation to obtain a new random vector* $\mathbf{y} = \mathbf{A}\mathbf{x}$*, where* $\mathbf{A}$ *is an invertible matrix with nonzero determinant. Then the differential entropy of* $\mathbf{y}$ *is given by:*

$$H[\mathbf{y}] = H[\mathbf{x}] + \ln|\det(\mathbf{A})|$$

where $\det(\mathbf{A})$ *denotes the determinant of matrix* $\mathbf{A}$*.*

Proof. Let $p_{\mathbf{x}}(\mathbf{x})$ be the probability density function of $\mathbf{x}$. For the linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$, the inverse transformation is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$.

The Jacobian matrix of this transformation is:

$$\mathbf{J} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$$

and the absolute value of its determinant is $|\det(\mathbf{J})| = |\det(\mathbf{A})|$.

By the change of variables formula, the probability density function of $\mathbf{y}$ is:

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{A}^{-1}\mathbf{y}) \cdot |\det(\mathbf{A})|^{-1} \tag{1}$$

Now we compute the differential entropy of $\mathbf{y}$:

$$\begin{aligned}
H[\mathbf{y}] &= -\int p_{\mathbf{y}}(\mathbf{y}) \ln p_{\mathbf{y}}(\mathbf{y}) d\mathbf{y} \\
&= -\int \left[p_{\mathbf{x}}(\mathbf{A}^{-1}\mathbf{y}) \cdot |\det(\mathbf{A})|^{-1}\right] \ln\left[p_{\mathbf{x}}(\mathbf{A}^{-1}\mathbf{y}) \cdot |\det(\mathbf{A})|^{-1}\right] d\mathbf{y}
\end{aligned}$$

Make the substitution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$, which implies $d\mathbf{y} = |\det(\mathbf{A})| d\mathbf{x}$:

$$\begin{aligned}
H[\mathbf{y}] &= -\int p_{\mathbf{x}}(\mathbf{x}) \cdot |\det(\mathbf{A})|^{-1} \ln\left[p_{\mathbf{x}}(\mathbf{x}) \cdot |\det(\mathbf{A})|^{-1}\right] |\det(\mathbf{A})| d\mathbf{x} \\
&= -\int p_{\mathbf{x}}(\mathbf{x}) \ln\left[p_{\mathbf{x}}(\mathbf{x}) \cdot |\det(\mathbf{A})|^{-1}\right] d\mathbf{x}
\end{aligned}$$

Using the logarithmic identity $\ln(ab) = \ln a + \ln b$:

$$\begin{aligned} H[\mathbf{y}] &= -\int p_{\mathbf{x}}(\mathbf{x}) \left[\ln p_{\mathbf{x}}(\mathbf{x}) + \ln|\det(\mathbf{A})|^{-1}\right] d\mathbf{x} \\ &= -\int p_{\mathbf{x}}(\mathbf{x}) \ln p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} - \int p_{\mathbf{x}}(\mathbf{x}) \ln|\det(\mathbf{A})|^{-1} d\mathbf{x} \\ &= H[\mathbf{x}] + \ln|\det(\mathbf{A})| \int p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \quad (\text{since } \ln|\det(\mathbf{A})|^{-1} = -\ln|\det(\mathbf{A})|) \\ &= H[\mathbf{x}] + \ln|\det(\mathbf{A})| \quad (\text{as } \int p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} = 1) \end{aligned}$$

□

1.3 STANDARD DISTRIBUTION

1.3.1 Multivariate Gaussian distribution

Definition 1.3.1.

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

where μ is a D-dimensional mean vector, Σ is a $D \times D$ covariance matrix, and $\det(\Sigma)$ *denotes the determinant of* Σ.

Definition 1.3.2.

$$\Delta^2 = (x-\mu)^T \Sigma^{-1}(x-\mu)$$

The quantity Δ is called the ***Mahalanobis distance*** *from $\boldsymbol{\mu}$ to $\boldsymbol{x}$. It reduces to the* ***Euclidean distance*** *when $\boldsymbol{\Sigma}$ is the identity matrix. The Gaussian distribution is constant on surfaces of constant value of this quadratic form in $\boldsymbol{x}$-space.*

Theorem 1.3.1 (Symmetry of the Precision and Covariance Matrices in the Multivariate Gaussian Distribution)**.** *Consider the multivariate Gaussian distribution:*

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

The precision matrix (inverse covariance matrix), $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$, can be decomposed into the sum of a symmetric matrix and an antisymmetric matrix. The antisymmetric component does not contribute to the exponent of the distribution. Consequently, the precision matrix can be assumed, without loss of generality, to be symmetric. Furthermore, since the inverse of a symmetric matrix is symmetric, the covariance matrix $\boldsymbol{\Sigma}$ can likewise be assumed, without loss of generality, to be symmetric.

Proof. Let the precision matrix be denoted by $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Any square matrix can be uniquely decomposed into a symmetric part and an antisymmetric part:

$$\begin{aligned} \boldsymbol{\Lambda} &= \boldsymbol{\Lambda}_S + \boldsymbol{\Lambda}_A, \quad \text{where} \\ \boldsymbol{\Lambda}_S &= \frac{\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top}{2} \quad (\text{symmetric: } \boldsymbol{\Lambda}_S^\top = \boldsymbol{\Lambda}_S), \\ \boldsymbol{\Lambda}_A &= \frac{\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top}{2} \quad (\text{antisymmetric: } \boldsymbol{\Lambda}_A^\top = -\boldsymbol{\Lambda}_A). \end{aligned}$$

The key term in the exponent of the Gaussian distribution is the quadratic form:

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$$

Substituting the decomposition into this form yields:

$$\begin{aligned}\Delta^2 &= (\mathbf{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Lambda}_S+\boldsymbol{\Lambda}_A)(\mathbf{x}-\boldsymbol{\mu}) \\ &= (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}_S(\mathbf{x}-\boldsymbol{\mu}) + (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}_A(\mathbf{x}-\boldsymbol{\mu})\end{aligned}$$

Let $Q = (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}_A(\mathbf{x}-\boldsymbol{\mu})$. Since Q is a scalar, $Q^\top = Q$. We can also express the transpose of Q using the property of matrix transposition:

$$\begin{aligned}Q^\top &= \left((\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}_A(\mathbf{x}-\boldsymbol{\mu})\right)^\top \\ &= (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}_A^\top(\mathbf{x}-\boldsymbol{\mu}) \\ &= (\mathbf{x}-\boldsymbol{\mu})^\top(-\boldsymbol{\Lambda}_A)(\mathbf{x}-\boldsymbol{\mu}) \quad \text{(due to the antisymmetry of } \boldsymbol{\Lambda}_A\text{)} \\ &= -(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}_A(\mathbf{x}-\boldsymbol{\mu}) \\ &= -Q\end{aligned}$$

Therefore, we have $Q = -Q$, which implies $Q = 0$. Thus,

$$(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}_A(\mathbf{x}-\boldsymbol{\mu}) = 0$$

and the quadratic form depends solely on the symmetric component:

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}_S(\mathbf{x}-\boldsymbol{\mu})$$

This result shows that the probability density $p(\mathbf{x})$ depends only on $\boldsymbol{\Lambda}_S$. Therefore, for the purpose of defining the distribution, one can always use the symmetric precision matrix $\boldsymbol{\Lambda}_S$ without loss of generality.

Finally, since $\boldsymbol{\Lambda}_S$ is symmetric and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}_S^{-1}$, it follows from a standard result in linear algebra that the covariance matrix $\boldsymbol{\Sigma}$ is also symmetric. Thus, $\boldsymbol{\Sigma}$ can likewise be assumed to be symmetric without loss of generality.

□

Theorem 1.3.2 (Properties of Eigenvalues and Eigenvectors of a Real Symmetric Matrix)**.** *Let $\boldsymbol{\Sigma}$ be a $D \times D$ real symmetric matrix ($\boldsymbol{\Sigma}^\top = \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$), with the characteristic equation:*

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i \tag{2}$$

Then the following properties hold:

1. *Its eigenvalues λ_i are all real numbers.*
2. *Eigenvectors corresponding to distinct eigenvalues are mutually orthogonal.*
3. *Its set of eigenvectors can be chosen to form an orthonormal basis, satisfying:*

$$\mathbf{u}_i^\top\mathbf{u}_j = \delta_{ij} \tag{3}$$

(where δ_{ij} is the Kronecker delta function), even if some eigenvalues $\lambda_i = 0$.

Proof. We prove the three properties in order.
Part 1: Proving the eigenvalues λ_i are real.

1. Assume the eigenvalue and eigenvector can be complex: Since $\boldsymbol{\Sigma}$ is a real matrix, its characteristic equation may yield complex eigenvalues and eigenvectors. Let $(\lambda_i, \mathbf{u}_i)$ be an eigenpair, where λ_i could be complex and $\mathbf{u}_i$ a complex vector:
$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i$$
2. Take the complex conjugate: Taking the conjugate of both sides, and noting that $\overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$ because $\boldsymbol{\Sigma}$ is real, yields:
$$\overline{\boldsymbol{\Sigma}\mathbf{u}_i} = \overline{\lambda_i\mathbf{u}_i} \quad \Rightarrow \quad \boldsymbol{\Sigma}\overline{\mathbf{u}_i} = \overline{\lambda_i}\,\overline{\mathbf{u}_i}$$
3. Left-multiply by $\overline{\mathbf{u}_i}^\top$ and equate:

- First, left-multiply the first equation by $\overline{\mathbf{u}_i}^\top$:

$$\overline{\mathbf{u}_i}^\top \mathbf{\Sigma} \mathbf{u}_i = \overline{\mathbf{u}_i}^\top \lambda_i \mathbf{u}_i = \lambda_i (\overline{\mathbf{u}_i}^\top \mathbf{u}_i)$$

- Second, left-multiply the second equation by $\mathbf{u}_i^\top$. Noting that $\mathbf{u}_i^\top \mathbf{\Sigma} \overline{\mathbf{u}_i} = (\mathbf{\Sigma}^\top \mathbf{u}_i)^\top \overline{\mathbf{u}_i} = (\mathbf{\Sigma} \mathbf{u}_i)^\top \overline{\mathbf{u}_i}$ (because $\mathbf{\Sigma}$ is symmetric), we get:

$$\mathbf{u}_i^\top \mathbf{\Sigma} \overline{\mathbf{u}_i} = \mathbf{u}_i^\top \overline{\lambda_i} \, \overline{\mathbf{u}_i} = \overline{\lambda_i} (\mathbf{u}_i^\top \overline{\mathbf{u}_i})$$

- Now, observe the left-hand sides of the two previous equations. Since $\overline{\mathbf{u}_i}^\top \mathbf{\Sigma} \mathbf{u}_i$ is a scalar, it is equal to its own transpose:

$$\begin{aligned} (\overline{\mathbf{u}_i}^\top \mathbf{\Sigma} \mathbf{u}_i)^\top &= \mathbf{u}_i^\top \mathbf{\Sigma}^\top \overline{\mathbf{u}_i} \\ &= \mathbf{u}_i^\top \mathbf{\Sigma} \overline{\mathbf{u}_i} \quad \text{(because } \mathbf{\Sigma} \text{ is symmetric)} \end{aligned}$$

  This shows the left-hand sides are equal. Therefore, their right-hand sides must also be equal:

$$\lambda_i (\overline{\mathbf{u}_i}^\top \mathbf{u}_i) = \overline{\lambda_i} (\mathbf{u}_i^\top \overline{\mathbf{u}_i})$$

  Note that $\overline{\mathbf{u}_i}^\top \mathbf{u}_i = \mathbf{u}_i^\top \overline{\mathbf{u}_i}$ (both represent the same real, positive scalar inner product of vector $\mathbf{u}_i$ with itself). Let $c = \overline{\mathbf{u}_i}^\top \mathbf{u}_i > 0$ (since $\mathbf{u}_i \neq \mathbf{0}$). We have:

$$\lambda_i c = \overline{\lambda_i} c \quad \Rightarrow \quad \lambda_i = \overline{\lambda_i}$$

  Conclusion: λ_i is equal to its own complex conjugate, therefore λ_i must be a real number. Q.E.D. for Part 1.

Part 2: Proving orthogonality of eigenvectors for distinct eigenvalues. Let $(\lambda_i, \mathbf{u}_i)$ and $(\lambda_j, \mathbf{u}_j)$ be two eigenpairs of $\mathbf{\Sigma}$ with $\lambda_i \neq \lambda_j$.

$$\mathbf{\Sigma} \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad \mathbf{\Sigma} \mathbf{u}_j = \lambda_j \mathbf{u}_j$$

Left-multiplying the first equation by $\mathbf{u}_j^\top$ and the second by $\mathbf{u}_i^\top$ gives:

$$\mathbf{u}_j^\top \mathbf{\Sigma} \mathbf{u}_i = \lambda_i (\mathbf{u}_j^\top \mathbf{u}_i), \quad \mathbf{u}_i^\top \mathbf{\Sigma} \mathbf{u}_j = \lambda_j (\mathbf{u}_i^\top \mathbf{u}_j)$$

Take the transpose of the second result. The left-hand side is:

$$\begin{aligned} (\mathbf{u}_i^\top \mathbf{\Sigma} \mathbf{u}_j)^\top &= \mathbf{u}_j^\top \mathbf{\Sigma}^\top \mathbf{u}_i \\ &= \mathbf{u}_j^\top \mathbf{\Sigma} \mathbf{u}_i \quad \text{(because } \mathbf{\Sigma} \text{ is symmetric)} \end{aligned}$$

The right-hand side of the transposed equation is $\lambda_j (\mathbf{u}_i^\top \mathbf{u}_j)^\top = \lambda_j (\mathbf{u}_j^\top \mathbf{u}_i)$ (since λ_j is real). Therefore, we have:

$$\mathbf{u}_j^\top \mathbf{\Sigma} \mathbf{u}_i = \lambda_j (\mathbf{u}_j^\top \mathbf{u}_i)$$

Comparing this with the first result, whose left-hand sides are identical, yields:

$$\lambda_i (\mathbf{u}_j^\top \mathbf{u}_i) = \lambda_j (\mathbf{u}_j^\top \mathbf{u}_i) \quad \Rightarrow \quad (\lambda_i - \lambda_j)(\mathbf{u}_j^\top \mathbf{u}_i) = 0$$

Since $\lambda_i \neq \lambda_j$, it follows that $(\lambda_i - \lambda_j) \neq 0$. Therefore, we must have:

$$\mathbf{u}_j^\top \mathbf{u}_i = 0$$

Conclusion: Eigenvectors $\mathbf{u}_i$ and $\mathbf{u}_j$ corresponding to distinct eigenvalues λ_i and λ_j are orthogonal. Q.E.D. for Part 2.

Part 3: Proving the eigenvectors can be chosen to form an orthonormal basis.

1. **Eigenvalue multiplicity:** For a real symmetric matrix, the algebraic multiplicity of each eigenvalue λ_i equals its geometric multiplicity. If an eigenvalue λ has multiplicity k, the dimension of its corresponding eigenspace $\ker(\mathbf{\Sigma} - \lambda \mathbf{I})$ is also k.

2. **Handling multiple eigenvalues:** For a single eigenvalue λ_i (whether simple or of multiplicity k), we can apply the Gram-Schmidt orthogonalization process within its k-dimensional eigenspace. From any basis of this subspace, we can construct an orthonormal basis (a set of k mutually orthogonal unit vectors) to serve as the eigenvectors for this eigenvalue.

3. **Handling all eigenspaces:**
   - For eigenvectors corresponding to *different* eigenvalues, they are automatically orthogonal (proven in Part 2).
   - For eigenvectors corresponding to the *same* eigenvalue (multiple roots), we make them orthogonal via the Gram-Schmidt process.
   - Combining the orthonormal bases from all eigenspaces yields an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_D\}$ for the entire space $\mathbb{R}^D$. This basis consists of the eigenvectors of $\mathbf{\Sigma}$.
4. **Including zero eigenvalues:** Even if some eigenvalues $\lambda_i = 0$, the above process still applies. Eigenvectors corresponding to zero eigenvalues form an orthonormal basis for the null space $\ker(\mathbf{\Sigma})$. These vectors remain orthogonal to eigenvectors of other eigenvalues because $0 \neq \lambda_j$.

Conclusion: Therefore, the set of eigenvectors of a real symmetric matrix $\mathbf{\Sigma}$ can always be chosen, without loss of generality, as an orthonormal basis satisfying $\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$. Q.E.D. for Part 3.

This property is the core content of the spectral theorem for real symmetric matrices. □

Theorem 1.3.3 (Spectral Decomposition of a Real Symmetric Matrix and its Inverse)**.** *Let $\mathbf{\Sigma}$ be a $D \times D$ real symmetric matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_D$ and corresponding orthonormal eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_D$ satisfying:*

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad \mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$$

Then the matrix $\mathbf{\Sigma}$ and its inverse $\mathbf{\Sigma}^{-1}$ (assuming all $\lambda_i \neq 0$) can be expressed as:

$$\mathbf{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top, \quad \mathbf{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\top$$

Proof. We prove the two parts of the theorem in order.

Part 1: Proof that $\mathbf{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$.

Since the set of eigenvectors $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_D\}$ forms an orthonormal basis for $\mathbb{R}^D$, any vector $\mathbf{x} \in \mathbb{R}^D$ can be uniquely expressed as a linear combination of these basis vectors:

$$\mathbf{x} = \sum_{j=1}^{D} c_j \mathbf{u}_j$$

where the coefficients c_j are given by the projection $c_j = \mathbf{u}_j^\top \mathbf{x}$. Substituting this expression for c_j yields:

$$\mathbf{x} = \sum_{j=1}^{D} (\mathbf{u}_j^\top \mathbf{x}) \mathbf{u}_j$$

Now, we compute $\mathbf{\Sigma}\mathbf{x}$ in two different ways.

Method 1 (Direct calculation using linearity and the eigenvalue equation):

$$\mathbf{\Sigma}\mathbf{x} = \mathbf{\Sigma} \left(\sum_{j=1}^{D} (\mathbf{u}_j^\top \mathbf{x}) \mathbf{u}_j \right) = \sum_{j=1}^{D} (\mathbf{u}_j^\top \mathbf{x}) \mathbf{\Sigma}\mathbf{u}_j = \sum_{j=1}^{D} (\mathbf{u}_j^\top \mathbf{x}) \lambda_j \mathbf{u}_j = \sum_{j=1}^{D} \lambda_j (\mathbf{u}_j^\top \mathbf{x}) \mathbf{u}_j$$

Method 2 (Calculation assuming the spectral decomposition form):

$$\mathbf{\Sigma}\mathbf{x} = \left(\sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{x} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i (\mathbf{u}_i^\top \mathbf{x})$$

Since the summation indices i and j are dummy variables, the expressions from Method 1 and Method 2 are identical:

$$\sum_{j=1}^{D} \lambda_j (\mathbf{u}_j^\top \mathbf{x}) \mathbf{u}_j = \sum_{i=1}^{D} \lambda_i (\mathbf{u}_i^\top \mathbf{x}) \mathbf{u}_i$$

Therefore, for any vector $\mathbf{x}$, we have:

$$\boldsymbol{\Sigma}\mathbf{x} = \left(\sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{x}$$

This implies the matrices themselves must be equal:

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$$

Part 2: Proof that $\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\top$.

To prove this is the inverse, we show that its product with $\boldsymbol{\Sigma}$ is the identity matrix $\mathbf{I}$.

$$\left(\sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\top \right) \boldsymbol{\Sigma}$$

Substituting the expression for $\boldsymbol{\Sigma}$ from Part 1 gives:

$$= \left(\sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\top \right) \left(\sum_{j=1}^{D} \lambda_j \mathbf{u}_j \mathbf{u}_j^\top \right) = \sum_{i=1}^{D} \sum_{j=1}^{D} \frac{1}{\lambda_i} \lambda_j \mathbf{u}_i \mathbf{u}_i^\top \mathbf{u}_j \mathbf{u}_j^\top$$

We now simplify this double sum using the orthonormality condition $\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$.

- If $i \neq j$, then $\mathbf{u}_i^\top \mathbf{u}_j = 0$, so the entire term is zero.
- If $i = j$, then $\mathbf{u}_i^\top \mathbf{u}_i = 1$ and $\frac{1}{\lambda_i} \lambda_i = 1$.

Thus, the double sum reduces to a single sum over the case $i = j$:

$$\sum_{i=1}^{D} \mathbf{u}_i \mathbf{u}_i^\top$$

We now prove that $\sum_{i=1}^{D} \mathbf{u}_i \mathbf{u}_i^\top = \mathbf{I}$. For any vector $\mathbf{x}$:

$$\left(\sum_{i=1}^{D} \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{x} = \sum_{i=1}^{D} \mathbf{u}_i (\mathbf{u}_i^\top \mathbf{x})$$

This is the expansion of the vector $\mathbf{x}$ in the orthonormal basis $\{\mathbf{u}_i\}$, which results in $\mathbf{x}$ itself. Therefore:

$$\sum_{i=1}^{D} \mathbf{u}_i \mathbf{u}_i^\top = \mathbf{I}$$

Connecting these results:

$$\left(\sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\top \right) \boldsymbol{\Sigma} = \mathbf{I}$$

By the definition of the inverse matrix, this proves that:

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\top$$

□

We substitute the spectral decomposition of $\boldsymbol{\Sigma}^{-1}$ into the formula for the Mahalanobis distance:

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

$$= (\mathbf{x} - \boldsymbol{\mu})^\top \left(\sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\top \right) (\mathbf{x} - \boldsymbol{\mu})$$

$$= \sum_{i=1}^{D} \frac{1}{\lambda_i} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{u}_i \mathbf{u}_i^\top (\mathbf{x} - \boldsymbol{\mu})$$

Now, we note that $(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{u}_i$ is a scalar (which can be viewed as the projection length of the vector $(\mathbf{x} - \boldsymbol{\mu})$ onto the basis vector $\mathbf{u}_i$). We denote it as y_i:

$$y_i = \mathbf{u}_i^\top (\mathbf{x} - \boldsymbol{\mu})$$

(Because the transpose of a scalar is equal to itself, $(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{u}_i = y_i$.)

Similarly, $\mathbf{u}_i^\top (\mathbf{x} - \boldsymbol{\mu})$ is also a scalar and equals y_i.

Therefore, each term in the previous expression can be rewritten as:

$$\frac{1}{\lambda_i} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{u}_i \mathbf{u}_i^\top (\mathbf{x} - \boldsymbol{\mu}) = \frac{1}{\lambda_i} \cdot y_i \cdot y_i = \frac{y_i^2}{\lambda_i}$$

Substituting each term back into the summation, we obtain the final scalar summation form:

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_D)^T$, then:

$$\boldsymbol{y} = \boldsymbol{U}(\boldsymbol{x} - \boldsymbol{\mu})$$

1.4 MATRIX-VECTOR DIFFERENTIATION