Slide 5

What is the relationship between the Euclidean distance \( \|x - \mu\| \) and the argument of the exponent? Let us start with writing out the expression for the distance between two vectors \( x, y \in \mathbb{R}^d \):

\[
\|x - y\| \triangleq \sqrt{\sum_{j=1}^{d} (x_j - y_j)^2}
\]  \quad (1)

This is the familiar metric we normally use in Euclidean spaces; it measures a length of the shortest path between \( x \) and \( y \). A closely related quantity is the Euclidean norm of a vector \( x \),

\[
\|x\|^2 \triangleq \sum_{j=1}^{d} x_j^2.
\]

This is of course equal to \( x^T x \).

So, the norm of \( x - \mu \) is equal to the squared distance of between \( x \) and \( \mu \), and also to

\[
(x - \mu)^T (x - \mu).
\]

What happens if we put a matrix \( A \) “in the middle” of the dot product, \((x - \mu)^T A (x - \mu)\)? The simplest case is the identity \( A = \mathbf{I} \) - it of course does not change anything. Now suppose that \( A = a \mathbf{I} \), i.e. a diagonal matrix where all elements are zero except for the main diagonal which contains \( a \)’s.

Then,

\[
(x - \mu)^T a \mathbf{I} (x - \mu) = [x_1 - \mu_1, \ldots, x_d - \mu_d] \begin{bmatrix} a & 0 & \ldots & 0 \\ 0 & a & 0 & \ldots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & a \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ \vdots \\ x_d - \mu_d \end{bmatrix} = a \sum_{j=1}^{d} (x_j - \mu_j)^2.
\]

The first product yields \( a(x_1 - \mu_1), \ldots, a(x_d - \mu_d) \) and then the second yields

\[
a \sum_{j=1}^{d} (x_j - \mu_j)^2,
\]
that is, $A$ simply scaled the distance by a factor of $a$.

A slightly more general case is $A = \text{diag}(a_1, \ldots, a_d)$, i.e. a diagonal matrix with different elements on the main diagonal. It is easy to see that

$$(x - \mu)^T A (x - \mu) = \sum_{j=1}^{d} a_j (x_j - \mu_j)^2$$

This is a more complex scaling operation. In terms of distances, one interpretation is that $a_j$ specifies the “cost” of traveling in the direction parallel to the axis $x_j$.

We will defer the discussion of more general cases for later.

## Slide 9

Here is the derivation for the mean and covariance of $\hat{w}$, given the training data $X$. We will start dropping the part of the notation that specifies the distribution with respect to which the expectations are taken whenever that’s obvious.

$$E[\hat{w}|X] = E[w^* + (X^TX)^{-1}X^T \nu|X]$$

$$= w^* + E[(X^TX)^{-1}X^T \nu|X]$$

$$= w^* + X^TX)^{-1}X^T E[\nu|X]$$

$$= w^* + X^TX)^{-1}X^T E[\nu]$$

$$= w^* + X^TX)^{-1}X^T 0$$

$$= w^*$$

Justification for the steps above:

(3) $w^*$ is independent of a particular $X$; it’s the optimal linear regressor for the model, not the data. Therefore, it’s a constant with respect to $X$ and we can take it out of the expectation.

(4) $(X^TX)^{-1}X$ is of course also a constant w.r.t. $X$.

(5) since in our model the noise is independent of the inputs, $p(\nu|X) = p(\nu)$ and we can remove conditioning on $X$. 2
(6) assumption that the noise is white (zero-mean).

Recall that the covariance of two random variables $a$ and $b$ is defined as

$$\text{Cov}_{a,b} \triangleq E_{p(a,b)} [(a - \mu_a)(b - \mu_b)],$$

where $\mu_a = E_{p(a)} [a]$ and $\mu_b = E_{p(b)} [b]$ are the marginal means of the corresponding random variables. Intuitively, this means how $a$ and $b$ “co-vary” (i.e., if $\text{Cov}_{a,b}$ is positive it means that they tend, in probability, to deviate to the same direction from their means.) A related quantity is correlation,

$$\text{cor } a, b \triangleq \frac{\text{Cov}_{a,b}}{\sigma_a \sigma_b},$$

where $\sigma_a$ denotes the standard deviation (square root of variance) of $a$. Correlation measures the amount of linear relationship between the two variables. Note that both covariance and correlation are symmetric, in the sense that $\text{Cov}_{a,b} = \text{Cov}_{b,a}$.

The covariance matrix $\text{Cov}_Z$ of a random variable $z \in \mathbb{R}^d$ is a generalization of this concept. The $(i, j)$ entry in this matrix is the covariance of $z_i$ and $z_j$; the diagonal elements are therefore just the variances of $z_i$, $i = 1, \ldots, d$ (sometimes this matrix is called the variance-covariance matrix):

$$\text{Cov}_Z \triangleq \begin{bmatrix}
\sigma_{z_1}^2 & \text{Cov}_{z_1,z_2} & \cdots & \text{Cov}_{z_1,z_d} \\
\text{Cov}_{z_2,z_1} & \sigma_{z_2}^2 & \cdots & \text{Cov}_{z_2,z_d} \\
\vdots & \ddots & \ddots & \vdots \\
\text{Cov}_{z_d,z_1} & \text{Cov}_{z_d,z_2} & \cdots & \sigma_{z_d}^2 
\end{bmatrix}.$$ 

From this definition a few properties follow immediately:

1. $\text{Cov}_Z$ is square and symmetric.

2. The elements on the main diagonal are always non-negative (and can be zero only if the corresponding $z_i$ is constant).

It is easy now to show that under the above definition,

$$\text{Cov}_Z = E \left[ (z - \mu_z)(z - \mu_z)^T \right]$$

A quick “dimension sanity check”: this is an outer product, of a $d \times 1$ column vector by a $1 \times d$ row vector, yielding a $d \times d$ matrix.
Here is an often useful particular case. Suppose we have a random vector $z$ the components $z_j$ of which are known to be statistically independent. Then, by definition of covariance, the elements off the diagonal are zero, and the covariance matrix is $\text{Cov}_z = \text{diag}(\sigma^2_1, \ldots, \sigma^2_d)$. If the components of $z$ are \textit{identically} distributed, so that they all have the same variance $\sigma^2$, the covariance matrix is just $\sigma^2 I$. Note that under the statistical model we have assumed, this is exactly the case for our random noise vector $\nu$.

Now let’s get back to deriving the covariance of $\hat{w}$. Recall that under our current assumptions,

$$\hat{w} = w^* + (X^T X)^{-1} X \nu$$

We can use the result we have just derived that $E[\hat{w}] = w^*$

$$\text{Cov}_w E[\hat{w} | X] = E[(\hat{w} - w^*)(\hat{w} - w^*)^T | X]$$

$$= E \left[ ((X^T X)^{-1} X \nu) ((X^T X)^{-1} X \nu)^T | X \right]$$

$$= E \left[ (X^T X)^{-1} X \nu \nu^T X^T (X^T X)^{-1} | X \right]$$

$$= (X^T X)^{-1} X E \left[ \nu \nu^T X^T (X^T X)^{-1} | X \right]$$

$$= (X^T X)^{-1} X E \left[ \nu^T \right] X^T (X^T X)^{-1}$$

$$= (X^T X)^{-1} X \sigma^2 I (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1} X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1}.$$