Informal Notes for Mathematics for Machine Learning
Imperial College London

Lecture 6 & 7: Multivariate Probability

Lecturer: Matthew Wicker

1 Learning Objectives

In our last lecture we examined the basic learning choices (e.g., choice of loss function) of our linear models from a probabilistic perspective. However, this required working in detail with probability distributions. While simply plugging in a Gaussian density is all we required for that lecture, if we are to go further with establishing good probabilistic machine learning fundamentals we will need to introduce probability in a more general way. In this lecture (and in its accompanying slides) we start by introducing some concepts from measure theory and then provide some central tools in manipulating probability distributions that will become indispensable for the rest of the course. At the end of these notes we will also discuss a few probability distributions that we will continue to use throughout.

2 A Primer on Measure Theory

The field of measure theory lays the formal mathematical foundation of probability theory. While there is little that we will do throughout this course will call for an understanding specific measure theoretic details it is nonetheless an important topic to discuss to build up an intuition for probability theory and random variables. Lets start by building up to the definition of a measure with a sequence of definitions:

Definition 2.1. **σ-Algebra** A σ-Algebra on a non-empty set Ω is a collection $A \subseteq \mathcal{P}(\Omega)$ such that the collection is closed under compliments and countable unions. Formally:

1. (Closed under compliment): If a set $B \in A$, then that implies $\bar{B} \in A$, that its compliment is in A
2. (Closed under countable union): If a series of collections $B_1, B_2, \ldots, B_n \in A$, then that implies $\bigcup_{i=1}^{n} B_i \in A$

Perhaps the simplest example of a σ-Algebra over a set Ω would be $A = \{\emptyset, \Omega\}$. In the slides we derive the fact that Ω itself must be contained in any σ-Algebra. It may be worthwhile here to prove that fact for yourself without looking at the slides. We have introduced σ-Algebras in order to build the concept of a probability measure:

Definition 2.2. Probability Measure A probability measure, P over a σ-Algebra (Ω, A) is a function $P : A \mapsto [0, \infty]$ such that:

1. $P(\emptyset) = 0$ and $P(\Omega) = 1$
2. $P(\bigcup_{i=1}^{n} B_i) = \sum_{i=1}^{n} P(B_i)$ as long as each B_i is pairwise disjoint
3. $P(\Omega) = 1$

A probability measure over a given σ-Algebra is a function that satisfies three axioms which are also known as Kolmogorov's axioms, named after the mathematician that introduced them in 1933 (pretty recent compared to the rest of mathematics). Together, we have now established the basis for a random variable. Throughout this course will will exclusively work with random variables as *functions* which is all they really are in truth. It is convenient to introduce this notion of random variables as functions with the cumulative distribution function:

Definition 2.3. Cumulative Distribution Function A cumulative distribution function (c.d.f. or cdf) or a probability measure is a function $F : \mathbb{R} \mapsto \mathbb{R}$ such that:

1. $x \leq y \implies F(x) < F(y)$
2. $\lim_{\boldsymbol{x} \downarrow y} F(x) = F(y)$ (i.e., right continuous)
3. $\lim_{x \to \infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$

Though we will not prove it here, it is true that for each cumulative distribution function there is a unique probability measure and for each probability measure there is a unique cumulative distribution function. If the cumulative distribution function is continuous, then it has a corresponding probability density function (pdf). This is th we define the function we will deal with most often in this course and is one you have likely seen before even if it was not introduced this formally. We denote the probability at x, $p(x)$ which is related to the CDF F by $p(x) = \dfrac{dF(x)}{dx}$. In your previous courses on probability you have hopefully derived many rules and theorems that follow from these basic definitions. In these notes we will first recall a few of the key probability distributions that will appear throughout the course and will then develop some important rules of which you should keep a strong working knowledge.

3 Joint Probability Distributions

Now that we have covered the definition of a random variable and the prevailing view that we will take on them (as functions); we now turn to building multivariate probability distributions. To start, we express two joint random variables as $P(x, y) = P(X = x, Y = y)$ where X and Y are two random variables. In one sense, this gives us a two dimensional random variable, and expanding this idea further we get to a form that is familiar to us from our first lecture:

$$P(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

Now one critical difference to point out for the purposes of this discussion is that we do *not* assume that these joint random variables are independent and identically distributed. By making this assumption, we were able to try to unify the each random variable X_i under a single distribution through density estimation. Here we do not make this assumption and are interested in using joint distributions as a more flexible probabilistic model. Consider using a joint probability distribution to model the number of times you repeat yourself in conversation and your cellular service. Your friend makes a claim: *On a call to my friend I had like 3 or 4 bars but was repeating myself ten to twelve times!* This seems incredibly unfortunate, and so we may be curious about how we would go about computing just *how* unfortunate this is. We would do so with the expression:

$$P(3 \leq X \leq 4, 10 \leq Y \leq 12)$$

Looking on either side of the comma, we have something that is easily dealt with by a cumulative distribution function, but here, we need the joint cumulative distribution function: $F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$ which is a straight-forward extension of the CDF we just defined using our measure theory primer.

3.1 Marginal Distributions

Given that we have modelled the joint distribution $P(x, y)$ we might be interested in just $P(x)$ or $P(y)$. For example, after quantifying the likelihood of our friends claim from the above section, we might want to look at $P(X = x)$ or the distribution of cellular service bars in our area. To get the distribution $P(x)$ from $P(x, y)$ we carry out a process called *marginalization*. This is so named because the process of marginalization with tabular data requires you to look at your variable of interested and sum across every other dimension of values and you would write each sum in the margin next to your data table (we see an example of this in the lecture slides). The general formula to remember here is:

$$P(X = x) = \int_{\Omega_Y} P(X = x, Y = y)dy$$

$$P(X = x) = \int_{\Omega_Y} P(X = x|Y = y)p(Y = y)dy$$

3.2 Conditional Probability

Conditional probability asks us to think about one of the random variables in our joint model, $P(X = x, Y = y)$, given that we know some information about one of the random variables. We denote this as $P(X = x|Y = y)$ this represents the probability distribution of X *given* that the random variable Y takes the value y.

3.3 Independence

We say that two random variables X and Y are independent if and only if we have the case that: $P(X = x, Y = y) = P(X = x)P(Y = y)$. While this is indeed a critical property to be aware of, we do cover it in Lecture 1 and so omit a repeat discussion.

3.4 Conditional Independence

While independence is a critical property for simplifying our calculations in density estimation, sometimes we have a slightly weaker property that can still greatly simplify our analyses. We denote conditionally independent variables:

$$X_1 \perp\!\!\!\perp X_2 | X_3 \iff P(X_1, X_2|X_3) = P(X_1|X_3)P(X_2|X_3)$$

This can be an unintuitive property at first, however, in the lecture slides we give an example using biased coins that demonstrates the principle well. If necessary, we will add a further text description here.

4 Manipulation of Probability Distributions

We have now covered many important properties of random variables and their probability distributions. Here, we introduce several tools that we will use throughout the course.

4.1 Product Rule

We will often see the product of distributions when analyzing our machine learning algorithms. That is, expressions of the form $P(X)P(Y)$ while we have seen such expressions in our exposition of independence and conditional independence, one rule to remember is the product rule of probability. The product rule tells us that the joint distribution of two of our random variables can be expanded as the marginal distribution times the conditional distribution. Formally:

$$P(X = x, Y = y) = P(X = x|Y = y)P(Y = y)$$

Notice that on the right we have a conditional distribution (conditional on $Y = y$) and we are multiplying by $P(Y = y)$, hence the description of the rule. We further exemplify just a taste of the usefulness of this fact in our proof of the law of expectation below.

4.2 Conditional Expectation

We have seen the definition of the expectation of a random variable (or its mean), and it is reasonable to consider the expectation of a conditional distribution which we write: $\mathbb{E}[X|Y]$. We can write out the form of this as:

$$\mathbb{E}[X|Y] = \int_{\Omega_X} P(X = x|Y = y)dx$$

Notice that a the integration is just over the random variable X as Y has been fixed to a given value and so even though it is not excitedly clear from the notation above it is true that $Y = y$ is for some predetermined, fixed y.

4.3 Law of Total Expectation

The law of total expectation states that:

$$\mathbb{E}_Y[\mathbb{E}_X[P(X = x|Y = y)]] = \mathbb{E}_X[P(X = x)]$$

. To prove this, we will use all of the rules we have proved up to this point. This is also one of our exercises, so I recommend that in your future revisions of these notes that you attempt to prove this fact on your own.

Lets start by expanding out the expectations:

$$\int_{\Omega_Y} \int_{\Omega_X} xP(X = x|Y = y)dxP(Y = y)dy$$

We can then see that, as we have seen this is just the conditional time marginal that we have seen and is sometimes called the product rule. Thus we can write the above equation in terms of the joint distribution:

$$P(X = x, Y = y) = P(X = x|Y = y)P(Y = y)$$

$$\int_{\Omega_X} \int_{\Omega_Y} xP(X = x, Y = y)dxdy$$

Grouping together the inner terms, we have:

$$\int_{\Omega_X} \left(\int_{\Omega_Y} xP(X = x, Y = y)dy \right) dx$$

Now, we should see that inside of the integral we have the process we introduced earlier, marginalization. So we know that we are just looking at the distribution $P(X = x)$ after marginalization, which means we can write out the above integral as:

$$\int_{\Omega_X} xP(X = x)dx$$

Which is exactly the definition of the expectation of X.

4.4 Law of Total Variance

We do not prove this here as it is a bit longer than the proof of the above law of total expectation, but a similar rule holds for variance which we call the law of total variance. This states that:

$$\mathbb{V}[Y] = \mathbb{E}_Y[\mathbb{V}[Y|X]] + \mathbb{V}[\mathbb{E}[Y|X]]$$

4.5 Change of Variables

In probability theory, the concept of *change of variables* is a fundamental technique that allows us to transform random variables from one probability distribution to another. This transformation is particularly useful when it simplifies the analysis or modeling of random phenomena. One

crucial property of the change of variables is that it *preserves event probabilities*, which in practice allows us to compute probabilities under Y given only knowledge about X and the functional relationship between the two. Consider two random variables, X and Y, where X follows a probability distribution with a probability density function (PDF) or probability mass function (PMF) denoted as $f_X(x)$, and Y is a transformed version of X, related by a one-to-one function g, such that $Y = g(X)$. The goal of the change of variables is to find the probability distribution of Y, which is denoted as $f_Y(y)$. The relationship between f_X and f_Y is given by the formula:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

Here, $g^{-1}(y)$ is the inverse function of g, and $\left|\frac{d}{dy} g^{-1}(y)\right|$ represents the absolute value of the derivative of g^{-1}. This formula is valid for both continuous and discrete random variables. Notice, that this formula is only valid when g is monotonically increasing or decreasing.

4.6 Law of the unconscious statistician

Recall that the definition of the mean of a random variable is:

$$\mathbb{E}_{p(x)}[x] = \int_{\Omega} xp(x)dx \qquad\qquad \mathbb{E}_{p(x)}[x] = \sum_{x \in \Omega} xp(x)$$

where the left is for continuous random variables and the right is for discrete random variables. The law of the unconscious statistician tells us how these expectations change when we apply a change of variables from above. That is, we are interested in the difference between $\mathbb{E}_{p(z)}[f(g(z))]$ and $\mathbb{E}_{p(x)}[f(x)]$ given that $X = g(Z)$.

Discrete Case: Let's begin with the discrete case, where X is a discrete random variable defined over a sample space Ω_X, and Z is another discrete random variable defined over a sample space Ω_Z. Furthermore, let $p_Z(z)$ be the probability mass function (PMF) of Z, and $p_X(x)$ be the PMF of X, related by $X = g(Z)$. We want to compute $\mathbb{E}_{p(z)}[f(g(z))]$ in terms of $\mathbb{E}_{p(x)}[f(x)]$. Using the definition of the expectation for discrete random variables, we have:

$$\mathbb{E}_{p(z)}[f(g(z))] = \sum_{z \in \Omega_Z} f(g(z))\, p_Z(z)$$

Now, since $X = g(Z)$, we can write:

$$\begin{aligned} \mathbb{E}_{p(x)}[f(x)] &= \sum_{x \in \Omega_X} f(x)\, p_X(x) \\ &= \sum_{x \in \Omega_X} f(g(z))\, p(X = x) \end{aligned}$$

To relating the first and second equations, we need to express the probability $p(X = x)$ in terms of $p_Z(z)$. Since $X = g(Z)$, we can write:

$$\begin{aligned} p(X = x) &= p(g(Z) = x) \\ &= \sum_{z \in \Omega_Z : g(z) = x} p(Z = z) \\ &= \sum_{z \in \Omega_Z : g(z) = x} p_Z(z) \end{aligned}$$

Substituting this into our second equation:

$$\mathbb{E}_{p(x)}[f(x)] = \sum_{x \in \Omega_X} f(g(z)) \left(\sum_{z \in \Omega_Z : g(z) = x} p_Z(z) \right)$$

Now, you can see that this is equivalent to the first equation. Therefore, in the discrete case, the Law of the Unconscious Statistician holds:

$$\mathbb{E}_{p(z)}[f(g(z))] = \mathbb{E}_{p(x)}[f(x)]$$

5 Refresher on Univariate Probability Distributions

We have recalled some of these distributions in previous notes, but to keep things self contained we repeat them here.

5.1 Bernoulli Distribution

The Bernoulli distribution is a fundamental probability distribution commonly used in machine learning. It models the probability of a binary outcome, where an event can have one of two possible outcomes: success (usually denoted as 1) or failure (usually denoted as 0). The distribution is characterized by a single parameter, which we denote θ, which represents the probability of success. The probability mass function (PMF) of the Bernoulli distribution is defined as:

$$P(X = x; \theta) = \begin{cases} \theta, & \text{if } x = 1 \\ 1 - \theta, & \text{if } x = 0 \end{cases}$$

Here, X is a random variable representing the outcome of a single trial. The Bernoulli distribution is a building block for more complex probability models and is often used to model binary events such as coin flips, where p represents the probability of getting heads.

5.2 Gaussian Distribution

The Gaussian distribution, also known as the normal distribution, is one of the most widely used probability distributions in statistics and machine learning. It is characterized by two parameters: the mean (μ) and the variance (σ^2), which determine the center and the spread of the distribution, respectively.

The probability density function (PDF) of the Gaussian distribution is given by:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The Gaussian distribution is often used to model continuous data that tends to cluster around a central value, making it a natural choice for many real-world applications. In machine learning, it plays a crucial role in algorithms such as linear regression and Gaussian Naive Bayes.

6 Multivariate Probability Distributions

6.1 Multinoulli Distribution

The Multinoulli distribution, also known as the categorical distribution, is an extension of the Bernoulli distribution to handle discrete outcomes with more than two possible categories. It is commonly used for modeling data with multiple discrete categories or classes. In the Multinoulli distribution, instead of a single parameter as in the Bernoulli distribution, we have a vector of probabilities ($\theta = [\theta_1, \theta_2, \ldots, \theta_k]$) representing the probabilities of each category. These probabilities sum to 1. The probability mass function (PMF) of the Multinoulli distribution is given by:

$$P(X = i) = \theta_i, \text{ for } i = 1, 2, \ldots, k$$

The Multinoulli distribution is frequently used in machine learning for tasks such as classification, where it models the probability distribution over multiple classes or categories.

6.2 Multivariate Gaussian Distribution

The Multivariate Gaussian distribution is an extension of the univariate Gaussian distribution to handle multiple dimensions. It is used to model continuous data with dependencies between multiple variables. In machine learning, it is a fundamental distribution for problems involving multivariate data. The Multivariate Gaussian distribution is characterized by a vector of means ($\boldsymbol{\mu}$) and a covariance matrix ($\boldsymbol{\Sigma}$), where $\boldsymbol{\mu}$ represents the means of each dimension, and $\boldsymbol{\Sigma}$ describes the relationships and variances between dimensions.

The probability density function (PDF) of the Multivariate Gaussian distribution is given by:

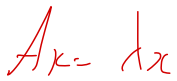
$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

The Multivariate Gaussian distribution is widely used in applications like multivariate statistics, clustering, and dimensionality reduction techniques like Principal Component Analysis (PCA).

6.2.1 Multivariate Gaussian Distribution (Geometric Interpretation)

The Multivariate Gaussian distribution not only provides a powerful statistical model but also offers a geometric interpretation that is particularly insightful. When visualizing a Multivariate Gaussian distribution in two dimensions (2D), it can be seen as an ellipse in the plane. This geometric interpretation reveals important information about the distribution's spread and orientation.

Let's consider a 2D Multivariate Gaussian distribution with means $\boldsymbol{\mu} = [\mu_1, \mu_2]$ and covariance matrix $\boldsymbol{\Sigma}$. The eigenvalues (λ_1 and λ_2) and eigenvectors ($\boldsymbol{v}_1$ and $\boldsymbol{v}_2$) of $\boldsymbol{\Sigma}$ are crucial in understanding the shape of the ellipse. The eigenvalues λ_1 and λ_2 represent the variances of the distribution along the directions defined by the corresponding eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. In other words, λ_1 and λ_2 indicate how spread out the distribution is along these axes. The eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ provide the directions of the major and minor axes of the ellipse, respectively. The lengths of these axes are determined by the square roots of the corresponding eigenvalues, $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$. In higher dimensions, the same principles apply, with each eigenvalue-eigenvector pair describing the properties of the distribution along a particular axis in the multidimensional space. This geometric view is fundamental for techniques like Principal Component Analysis (PCA) that aim to transform data to a new coordinate system aligned with the major axes of the Gaussian distribution.



Question 1 (Linear transform of a Gaussian random variable)**.** If X is a d-dimensional multivariate Gaussian random variable with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then what is the distribution of the random variable $Y = \mathbf{A}X$ with an invertible matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$?

Question 2 (Sum of independent Gaussian random variables)**.** If X, Y are two independent univariate Gaussian random variables (i.e., $X \perp\!\!\!\perp Y$), show that $Z = X + Y$ is also a univariate Gaussian random variable.

Question 3 (KL divergence and change-of-variables rule)**.** Show that KL divergence is invariant to change-of-variables, i.e., $\mathrm{KL}[p_X(x)||q_X(x)] = \mathrm{KL}[p_Y(y)||q_Y(y)]$ for $Y = T(X)$ with an invertible transformation T.

Question 4 (Independence of Gaussian variables)**.** Consider $X = (X_1, ..., X_N)$ as a multivariate random variable, which is distributed as a multivariate Gaussian with covariance matrix Σ. Show that $X_i \perp\!\!\!\perp X_j | X_{-ij}$ where X_{-ij} collect all the other X_n variables, if for the precision matrix $\Lambda := \Sigma^{-1}$ we have $\Lambda_{ij} = \Lambda_{ji} = 0$.

Question 5 (Independent vs uncorrelated variables)**.** Show that for the following definitions of X, Y, these two variables are uncorrelated (i.e., $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$), but not independent to each other: X is a univariate Gaussian variable with mean 0, and $Y = X^2$.

Question 6 (Expectation identities)**.** Prove the following expectation identities:

a. $\mathbb{V}_X[X] = \mathbb{E}_X[XX^\intercal] - \mathbb{E}_X[X]\mathbb{E}_X[X]^\intercal$, for $X \in \mathbb{R}^D$.

b. $\mathbb{C}_{X,Y}[X, Y] = \mathbb{E}_{X,Y}[XY^\intercal] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]^\intercal$, for $X \in \mathbb{R}^D, Y \in \mathbb{R}^E$.

c. $\mathbb{E}_{X,Y}[X + Y] = \mathbb{E}_X[X] + \mathbb{E}_Y[Y]$.

d. $\mathbb{V}_{X,Y}[X + Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y]$, for $X \perp\!\!\!\perp Y$.

Question 7 (MML 6.11: Iterated Expectations)**.** Consider random variables X, Y with joint distribution $p(x, y)$. Show that

$$\mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]] \tag{1}$$

where $\mathbb{E}_X[X|Y]$ denotes the expectation under the conditional distribution $p(x|y)$.

Question 8 (MML 6.13: Probability Integral Transformation)**.** Given a continous r.v. X, with CDF $F_X(x)$, show that the r.v. $Y := F_X(X)$ is uniformly distributed.

Question 1 (Linear transform of a Gaussian random variable)**.** If X is a d-dimensional multivariate Gaussian random variable with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then what is the distribution of the random variable $Y = \mathbf{A}X$ with an invertible matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$?

$X \in \mathbb{R}^d$ $\int [\cdot]$ $X \sim N(\mu, \Sigma)$ $Y = AX$, $A \in \mathbb{R}^{d \times d}$

Change of variable formula $f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$

$y = g(x) \rightarrow Y = AX$, $g(x) = Ax$

$g^{-1}(y) = A^{-1}y$ $\frac{d}{dy} g^{-1}(y) = \frac{d}{dy} A^{-1}y = A^{-1}$

$J_x(A^{-1}Y) = |A^{-1}|$

mean of Y: $E(Y) = E(AX) = AE(X) = A\mu$

Var: $Var(Y) = E[(Y - E(Y))(Y - E(Y))^T]$
$= E[Ax - AE(x))(Ax - AE(x))^T)$ $= E(\cdot\cdot)(x^T A^T \cdot E(x)^T A^T)$
$= AE(x - E(x))E(x^T - E(x)^T)A^T$
$= A \Sigma A^T$

$|A^{-1}| \frac{1}{(2\pi)^{p/2} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(A^{-1}y - \mu)^T \Sigma^{-1}(A^{-1}y - \mu)}$

where $\mu = A\mu$, $\Sigma = A\Sigma A^T$

Question 2 (Sum of independent Gaussian random variables)**.** If X, Y are two independent univariate Gaussian random variables (i.e., $X \perp\!\!\!\perp Y$), show that $Z = X + Y$ is also a univariate Gaussian random variable.

$X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$, X, Y indep $\rightarrow p(x,y) = p(x)p(y)$

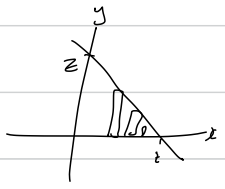
PDF of Z: $\int_{z=x+y} p(x,y)\,dx\,dy = \int_{z=x+y} p(x)p(y)\,dx\,dy$

$p(Z \le z^*) = p(X+Y \le z^* = x+y)$

$\int_{z \le z^*} p(z)\,dz = \int_{x+y \le z^*} p(x,y)\,dx\,dy$ ↓ X, Y indep

$= \int p(x)p(y)\,dx\,dy$

$p_Z(z) = \int_{x+y=z} p(x,y)\,dx\,dy = \int_{x+y=z} p(x)p(y)\,dx\,dy = \int_{x=-\infty}^{x=\infty} \int_{y=-\infty}^{y=z-x}$



$y = z - x$

$= \int_x p_X(x)\, p_Y(z-x)\,dx$

$= \int_x \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}} \times \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{1}{2}\frac{(z-x-\mu_y)^2}{\sigma_y^2}} dx$

$= \int_x \frac{1}{\sqrt{2\pi\sigma_x^2}\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(z-x-\mu_y)^2}{2\sigma_y^2}} dx$

$= \int_x e^{-\frac{1}{2}\left(\frac{x^2 - 2x\mu_x + \mu_x^2}{\sigma_x^2} + \frac{(z-x)^2 - 2(z-x)\mu_y + \mu_y^2}{\sigma_y^2}\right)} \times \text{(stuff)} = \int_x e^{-\frac{1}{2}\left(\cdots + \frac{z^2 - 2zx + x^2 - 2z\mu_y + 2x\mu_y + \mu_y^2}{\sigma_y^2}\right)} \times \text{(stuff)} = \int_x e^{-\frac{1}{2}\left(\cdots \right)}$

$\frac{x^2\sigma_y^2 + 2x\mu_y\sigma_y^2 - 2z\sigma_x^2 x + z^2\sigma_x^2 + \mu_x^2\sigma_y^2 - 2z\mu_y\sigma_x^2 + 2x\mu_x\sigma_y^2 - 2x\mu_x\sigma_y^2 + \mu_y^2\sigma_x^2}{\sigma_x^2\sigma_y^2}$

6b. $\underbrace{C_{X,Y}(X,Y)}_{\in \mathbb{R}^{D\times E}} = E(XY^T) - E(X)E(Y)^T$, $X\in\mathbb{R}^D$, $Y\in\mathbb{R}^E$

Want to show: $C_{X,Y}(X,Y)_{ij} = E(X_i Y_j) - E(X_i)E(Y_j)$

Definition of covar: $C(X,Y)_{ij} = E_{X,Y}\left[(X_i - E(X_i))(Y_j - E(Y_j))\right]$

$= E_{X,Y}\left(\underbrace{X_i Y_j - X_i E(Y_j) - Y_j E(X_i) + E(X_i)E(Y_j)}_{= f(X_i, Y_j)}\right)$

$E_{X,Y}(f(X_i,Y_j)) = \int_{x_1}\cdots\int_{x_d}\int_{y_1}\cdots\int_{y_E} f(X_i,Y_j)\, p(x_1\cdots x_d, y_1\cdots y_E)\, dx_1\cdots y_E$

$= \int_{X_i}\int_{Y_j} f(X_i,Y_j) \underbrace{\int_{x_1}\cdots\int_{x_d}\int_{y_1}\cdots\int_{y_E} p(x_1\cdots y_E)\, dx_1\cdots y_E}_{\text{all except } x_i, y_j}$

Marginalisation: $\iint_{x_2 x_3} p(x_1,x_2,x_3)\, dx_2 x_3 = p(x_1)$

$= \int_{X_i}\int_{Y_j} f(X_i,Y_j)\, p(X_i,Y_j)\, dX_i Y_j$ - integrated over all other vars

$= \int_{X_i}\int_{Y_j} \left(X_i Y_j - X_i E(Y_j) - Y_j E(X_i) + E(X_i)E(Y_j)\right) p(X_i,Y_j)\, dX_i Y_j$

$= \int_{X_i}\int_{Y_j} X_i Y_j\, p(X_i,Y_j)\, dX_i Y_j - E(Y_j)\underbrace{\iint_{X_i Y_j} X_i\, p(X_i,Y_j)\, dX_i Y_j}_{E(X_i)} - E(X_i)\underbrace{\iint_{X_i Y_j} Y_j\, p(X_i,Y_j)\, dX_i Y_j}_{= \int_{Y_j}\int_{X_i} p(X_i,Y_j)\, dX_i Y_j} + E(X_i)E(Y_j)\underbrace{\iint p(X_i,Y_j)\, dX_i Y_j}_{=1}$

$= \int_{Y_j} p(Y_j)\, dY_j$

$E(Y_j)$

$= E(X_i Y_j) - E(Y_j)E(X_i) - E(X_i)E(Y_j) + E(X_i)E(Y_j)$

$= E(X_i Y_j) - E(X_i)E(Y_j)$

7 show $E_x(x) = E_y[\underbrace{E_x[x|y]}_{\text{f of } y}]$

$$= \int_y E(x|y)\, p(y)\, dy$$

↓ definition

$$= \int_y \int_x x\, p(x|y)\, p(y)\, dx\, dy$$

$$= \int_x x \int_y p(x|y) p(y)\, dy\, dx$$

$p(x|y)$: $\frac{p(x,y)}{p(y)}$ → $p(x|y)p(y) = p(x,y)$

$$= \int_x x \int_y p(x,y)\, dy\, dx$$

↓ marginalization

$$= \int_x x\, p(x)\, dx$$

$$= E[x]$$