6 Probability Density Functions (PDFs)

In many cases, we wish to handle data that can be represented as a real-valued random variable, or a real-valued vector $\mathbf{x} = [x_1, x_2, ..., x_n]^T$. Most of the intuitions from discrete variables transfer directly to the continuous case, although there are some subtleties.

We describe the probabilities of a real-valued scalar variable x with a Probability Density Function (PDF), written $p(x)$. Any real-valued function $p(x)$ that satisfies:

$$p(x) \geq 0 \qquad \text{for all } x \tag{1}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{2}$$

is a valid PDF. I will use the convention of upper-case P for discrete probabilities, and lower-case p for PDFs.

With the PDF we can specify the probability that the random variable x falls within a given range:

$$P(x_0 \leq x \leq x_1) = \int_{x_0}^{x_1} p(x)dx \tag{3}$$

This can be visualized by plotting the curve $p(x)$. Then, to determine the probability that x falls within a range, we compute the area under the curve for that range.

The PDF can be thought of as the infinite limit of a discrete distribution, i.e., a discrete distribution with an infinite number of possible outcomes. Specifically, suppose we create a discrete distribution with N possible outcomes, each corresponding to a range on the real number line. Then, suppose we increase N towards infinity, so that each outcome shrinks to a single real number; a PDF is defined as the limiting case of this discrete distribution.

There is an important subtlety here: a probability density is *not* a probability per se. For one thing, there is no requirement that $p(x) \leq 1$. Moreover, the probability that x attains any one specific value out of the infinite set of possible values is always zero, e.g. $P(x = 5) = \int_5^5 p(x)dx = 0$ for any PDF $p(x)$. People (myself included) are sometimes sloppy in referring to $p(x)$ as a probability, but it is not a probability — rather, it is a function that can be used in computing probabilities.

Joint distributions are defined in a natural way. For two variables x and y, the joint PDF $p(x, y)$ defines the probability that (x, y) lies in a given domain $\mathcal{D}$:

$$P((x, y) \in \mathcal{D}) = \int_{(x,y)\in\mathcal{D}} p(x, y)dxdy \tag{4}$$

For example, the probability that a 2D coordinate (x, y) lies in the domain $(0 \leq x \leq 1, 0 \leq y \leq 1)$ is $\int_{0\leq x\leq 1} \int_{0\leq y\leq 1} p(x, y)dxdy$. The PDF over a vector may also be written as a joint PDF of its variables. For example, for a 2D-vector $\mathbf{a} = [x, y]^T$, the PDF $p(\mathbf{a})$ is equivalent to the PDF $p(x, y)$.

Conditional distributions are defined as well: $p(x|\mathbf{A})$ is the PDF over x, if the statement $\mathbf{A}$ is true. This statement may be an expression on a continuous value, e.g. "$y = 5$." As a short-hand,

we can write $p(x|y)$, which provides a PDF for x for every value of y. (It must be the case that $\int p(x|y)dx = 1$, since $p(x|y)$ is a PDF over values of x.)

In general, for all of the rules for manipulating discrete distributions there are analogous rules for continuous distributions:

Probability rules for PDFs:
- $p(x) \geq 0$, for all x
- $\int_{-\infty}^{\infty} p(x)dx = 1$
- $P(x_0 \leq x \leq x_1) = \int_{x_0}^{x_1} p(x)dx$
- **Sum rule:** $\int_{-\infty}^{\infty} p(x)dx = 1$
- **Product rule:** $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$.
- **Marginalization:** $p(y) = \int_{-\infty}^{\infty} p(x, y)dx$
- We can also add conditional information, e.g. $p(y|z) = \int_{-\infty}^{\infty} p(x, y|z)dx$
- **Independence:** Variables x and y are independent if: $p(x, y) = p(x)p(y)$.

6.1 Mathematical expectation, mean, and variance

Some very brief definitions of ways to describe a PDF:

Given a function $f(\mathbf{x})$ of an unknown variable $\mathbf{x}$, the **expected value** of the function with repect to a PDF $p(\mathbf{x})$ is defined as:

$$E_{p(\mathbf{x})}[f(\mathbf{x})] \equiv \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{5}$$

Intuitively, this is the value that we roughly "expect" $\mathbf{x}$ to have.

The mean $\boldsymbol{\mu}$ of a distribution $p(\mathbf{x})$ is the expected value of $\mathbf{x}$:

$$\boldsymbol{\mu} = E_{p(\mathbf{x})}[\mathbf{x}] = \int \mathbf{x}p(\mathbf{x})d\mathbf{x} \tag{6}$$

The variance of a scalar variable x is the expected squared deviation from the mean:

$$E_{p(x)}[(x - \mu)^2] = \int (x - \mu)^2 p(x)dx \tag{7}$$

The variance of a distribution tells us how uncertain, or "spread-out" the distribution is. For a very narrow distribution $E_{p(x)}[(x - \mu)^2]$ will be small.

The **covariance** of a vector $\mathbf{x}$ is a matrix:

$$\boldsymbol{\Sigma} = \text{cov}(\mathbf{x}) = E_{p(\mathbf{x})}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(x)d\mathbf{x} \tag{8}$$

By inspection, we can see that the diagonal entries of the covariance matrix are the variances of the individual entries of the vector:

$$\boldsymbol{\Sigma}_{ii} = \text{var}(x_{ii}) = E_{p(\mathbf{x})}[(x_i - \mu_i)^2] \tag{9}$$

The off-diagonal terms are covariances:

$$\mathbf{\Sigma}_{ij} = \text{cov}(x_i, x_j) = E_{p(x)}[(x_i - \mu_i)(x_j - \mu_j)] \tag{10}$$

between variables x_i and x_j. If the covariance is a large positive number, then we expect x_i to be larger than μ_i when x_j is larger than μ_j. If the covariance is zero and we know no other information, then knowing $x_i > \mu_i$ does not tell us whether or not it is likely that $x_j > \mu_j$.

One goal of statistics is to infer properties of distributions. In the simplest case, the **sample mean** of a collection of N data points $\mathbf{x}_{1:N}$ is just their average: $\bar{\mathbf{x}} = \frac{1}{N}\sum_i \mathbf{x}_i$. The **sample covariance** of a set of data points is: $\frac{1}{N}\sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$. The covariance of the data points tells us how "spread-out" the data points are.

6.2 Uniform distributions

The simplest PDF is the **uniform distribution**. Intuitively, this distribution states that all values within a given range $[x_0, x_1]$ are equally likely. Formally, the uniform distribution on the interval $[x_0, x_1]$ is:

$$p(x) = \begin{cases} \frac{1}{x_1 - x_0} & \text{if } x_0 \leq x \leq x_1 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

It is easy to see that this is a valid PDF (because $p(x) > 0$ and $\int p(x)dx = 1$).

We can also write this distribution with this alternative notation:

$$x|x_0, x_1 \quad \sim \quad \mathcal{U}(x_0, x_1) \tag{12}$$

Equations 11 and 12 are equivalent. The latter simply says: x is distributed uniformly in the range x_0 and x_1, and it is impossible that x lies outside of that range.

The mean of a uniform distribution $\mathcal{U}(x_0, x_1)$ is $(x_1 + x_0)/2$. The variance is $(x_1 - x_0)^2/12$.

6.3 Gaussian distributions

Arguably the single most important PDF is the **Normal** (a.k.a., **Gaussian**) probability distribution function (PDF). Among the reasons for its popularity are that it is theoretically elegant, and arises naturally in a number of situations. It is the distribution that maximizes entropy, and it is also tied to the Central Limit Theorem: the distribution of a random variable which is the sum of a number of random variables approaches the Gaussian distribution as that number tends to infinity (Figure 1).

Perhaps most importantly, it is the analytical properties of the Gaussian that make it so ubiquitous. Gaussians are easy to manipulate, and their form so well understood, that we often assume quantities are Gaussian distributed, even though they are not, in order to turn an intractable model, or problem, into something that is easier to work with.

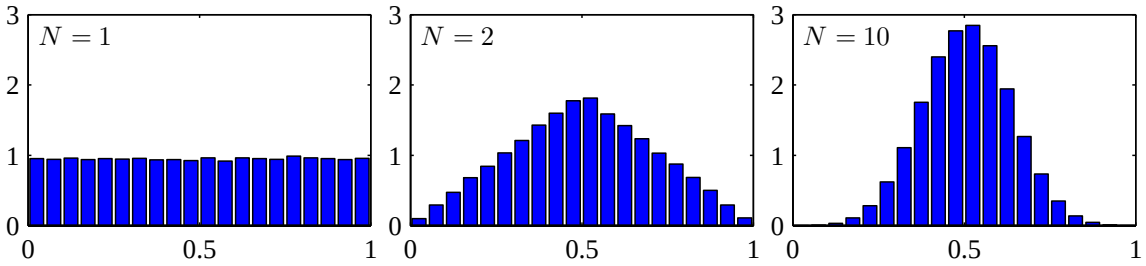

Figure 1: Histogram plots of the mean of N uniformly distributed numbers for various values of N. The effect of the Central Limit Theorem is seen: as N increases, the distribution becomes more Gaussian. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

The simplest case is a Gaussian PDF over a scalar value x, in which case the PDF is:

$$p(x|\mu, \sigma^2) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{13}$$

(The notation $\exp(a)$ is the same as e^a). The Gaussian has two parameters, the mean μ, and the variance σ^2. The mean specifies the center of the distribution, and the variance tells us how "spread-out" the PDF is.

The PDF for D-dimensional vector $\mathbf{x}$, the elements of which are jointly distributed with a the Gaussian denity function, is given by

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad = \quad \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left(-(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2\right) \tag{14}$$

where $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Sigma}$ is the $D \times D$ covariance matrix, and $|A|$ denotes the determinant of matrix A. An important special case is when the Gaussian is isotropic (rotationally invariant). In this case the covariance matrix can be written as $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. This is called a spherical or isotropic covariance matrix. In this case, the PDF reduces to:

$$p(\mathbf{x}|\boldsymbol{\mu}, \sigma^2) \quad = \quad \frac{1}{\sqrt{(2\pi)^D\sigma^{2D}}} \exp\left(-\frac{1}{2\sigma^2}||\mathbf{x}-\boldsymbol{\mu}||^2\right) . \tag{15}$$

The Gaussian distribution is used frequently enough that it is useful to denote its PDF in a simple way. We will define a function G to be the Gaussian density function, i.e.,

$$G(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left(-(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2\right) \tag{16}$$

When formulating problems and manipulating PDFs this functional notation will be useful. When we want to specify that a random vector has a Gaussian PDF, it is common to use the notation:

$$\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{17}$$

Equations 14 and 17 essentially say the same thing. Equation 17 says that $\mathbf{x}$ is Gaussian, and Equation 14 specifies (evaluates) the density for an input $\mathbf{x}$. The covariance matrix $\mathbf{\Sigma}$ of a Gaussian must be symmetric and positive definite

6.3.1 Diagonalization

A useful way to understand a Gaussian is to diagonalize the exponent. The exponent of the Gaussian is quadratic, and so its shape is essentially elliptical. Through diagonalization we find the major axes of the ellipse, and the variance of the distribution along those axes. Seeing the Gaussian this way often makes it easier to interpret the distribution.

As a reminder, the eigendecomposition of a real-valued symmetric matrix $\mathbf{\Sigma}$ yields a set of orthonormal vectors $\mathbf{v}_i$ and scalars λ_i such that

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{18}$$

Equivalently, if we combine the eigenvalues and eigenvectors into matrices $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_N]$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, ...\lambda_N)$, then we have

$$\mathbf{\Sigma}\mathbf{U} = \mathbf{U}\mathbf{\Lambda} \tag{19}$$

Since $\mathbf{U}$ is orthonormal:

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{20}$$

The inverse of $\mathbf{\Sigma}$ is straightforward, since $\mathbf{U}$ is orthonormal, and hence $\mathbf{U}^{-1} = \mathbf{U}^T$:

$$\mathbf{\Sigma}^{-1} = \left(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\right)^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T \tag{21}$$

(If any of these steps are not familiar to you, you should refresh your memory of them.)

Now, consider the negative log of the Gaussian (i.e., the exponent); i.e., let

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \; . \tag{22}$$

Substituting in the diagonalization gives:

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T (\mathbf{x} - \boldsymbol{\mu}) && (23) \\ &= \frac{1}{2}\mathbf{z}^T\mathbf{z} && (24) \end{aligned}$$

where

$$\mathbf{z} = \mathrm{diag}(\lambda_1^{-\frac{1}{2}}, ..., \lambda_N^{-\frac{1}{2}})\mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu}) \tag{25}$$

This new function $f(\mathbf{z}) = \mathbf{z}^T\mathbf{z}/2 = \sum_i z_i^2/2$ is a quadratic, with new variables z_i. Given variables $\mathbf{x}$, we can convert them to the $\mathbf{z}$ representation by applying Eq. 25, and, if all eigenvalues are

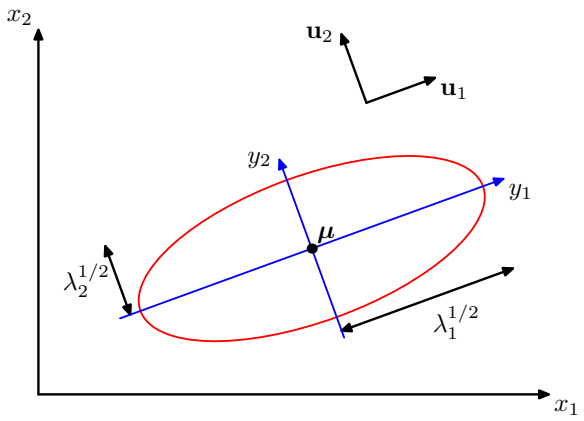

Figure 2: The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)(Note y_1 and y_2 in the figure should read z_1 and z_2.)

nonzero, we can convert back by inverting Eq. 25. Hence, we can write our Gaussian in this new coordinate system as[1]:

$$\frac{1}{\sqrt{(2\pi)^N}} \exp\left(-\frac{1}{2}||\mathbf{z}||^2\right) = \prod_i \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_i^2\right) \tag{26}$$

It is easy to see that for the quadratic form of $f(\mathbf{z})$, its level sets (i.e., the surfaces $f(\mathbf{z}) = c$ for constant c) are hyperspheres. Equivalently, it is clear from 26 that $\mathbf{z}$ is a Gaussian random vector with an isotropic covariance, so the different elements of $\mathbf{z}$ are uncorrelated. In other words, the value of this transformation is that we have decomposed the original N-D quadratic with many interactions between the variables into a much simpler Gaussian, composed of d independent variables. This convenient geometrical form can be seen in Figure 2. For example, if we consider an individual z_i variable in isolation (i.e., consider a slice of the function $f(\mathbf{z})$), that slice will look like a 1D bowl.

We can also understand the local curvature of f with a slightly different diagonalization. Specifically, let $\mathbf{v} = \mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu})$. Then,

$$f(\mathbf{u}) = \frac{1}{2}\mathbf{v}^T\boldsymbol{\Lambda}^{-1}\mathbf{v} = \frac{1}{2}\sum_i \frac{v_i^2}{\lambda_i} \tag{27}$$

If we plot a cross-section of this function, then we have a 1D bowl shape with variance given by λ_i. In other words, the eigenvalues tell us variance of the Gaussian in different dimensions.

[1]The normalizing $|\boldsymbol{\Sigma}|$ disappears due to the nature of change-of-variables in PDFs, which we won't discuss here.

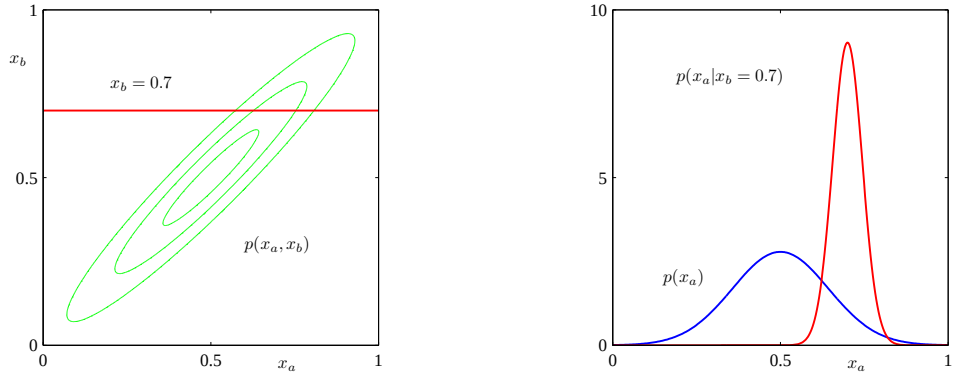

Figure 3: Left: The contours of a Gaussian distribution $p(x_a, x_b)$ over two variables. Right: The marginal distribution $p(x_a)$ (blue curve) and the conditional distribution $p(x_a|x_b)$ for $x_b = 0.7$ (red curve). (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)

6.3.2 Conditional Gaussian distribution

In the case of the multivariate Gaussian where the random variables have been partitioned into two sets $\mathbf{x}_a$ and $\mathbf{x}_b$, the conditional distribution of one set conditioned on the other is Gaussian. The marginal distribution of either set is also Gaussian. When manipulating these expressions, it is easier to express the covariance matrix in inverse form, as a "precision" matrix, $\Lambda \equiv \Sigma^{-1}$. Given that $\mathbf{x}$ is a Gaussian random vector, with mean $\boldsymbol{\mu}$ and covariance Σ, we can express $\mathbf{x}$, $\boldsymbol{\mu}$, Σ and Λ all in block matrix form:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_a} \\ \boldsymbol{\mu_b} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}, \quad (28)$$

Then one can show straightforwardly that the marginal PDFs for the components $\mathbf{x}_a$ and $\mathbf{x}_b$ are also Gaussian, i.e.,

$$\mathbf{x}_a \sim \mathcal{N}(\boldsymbol{\mu_a}, \Sigma_{aa}) , \qquad \mathbf{x}_b \sim \mathcal{N}(\boldsymbol{\mu_b}, \Sigma_{bb}). \quad (29)$$

With a little more work one can also show that the conditional distributions are Gaussian. For example, the conditional distribution of $\mathbf{x}_a$ given $\mathbf{x}_b$ satisfies

$$\mathbf{x}_a|\mathbf{x}_b \sim \mathcal{N}(\boldsymbol{\mu}_{a|b}, \Lambda_{aa}^{-1}) \quad (30)$$

where $\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu_a} - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$. Note that Λ_{aa}^{-1} is not simply Σ_{aa}. Figure 3 shows the marginal and conditional distributions applied to a two-dimensional Gaussian.

Finally, another important property of Gaussian functions is that the product of two Gaussian functions is another Gaussian function (although no longer normalized to be a proper density function):

$$G(\mathbf{x};\, \boldsymbol{\mu}_1, \Sigma_2)\, G(\mathbf{x};\, \boldsymbol{\mu}_2, \Sigma_2) \;\propto\; G(\mathbf{x};\, \boldsymbol{\mu}, \Sigma), \quad (31)$$

where

$$\boldsymbol{\mu} = \Sigma \left(\Sigma_1^{-1} \boldsymbol{\mu_1} + \Sigma_2^{-1} \boldsymbol{\mu_2} \right), \tag{32}$$

$$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}. \tag{33}$$

Note that the linear transformation of a Gaussian random variable is also Gaussian. For example, if we apply a transformation such that $\mathbf{y} = A\mathbf{x}$ where $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, we have $\mathbf{y} \sim \mathcal{N}(\mathbf{y}|A\boldsymbol{\mu}, A\Sigma A^T)$.

