Basics of Probability and Probability Distributions

Piyush Rai
Some Basic Concepts You Should Know About

- Random variables (discrete and continuous)
- Probability distributions over discrete/continuous r.v.’s
- Notions of joint, marginal, and conditional probability distributions
- Properties of random variables (and of functions of random variables)
  - Expectation and variance/covariance of random variables
- Examples of probability distributions and their properties
- Multivariate Gaussian distribution and its properties (very important)

**Note:** These slides provide only a (very!) quick review of these things. Please refer to a text such as PRML (Bishop) Chapter 2 + Appendix B, or MLAPP (Murphy) Chapter 2 for more details.

**Note:** Some other pre-requisites (e.g., concepts from information theory, linear algebra, optimization, etc.) will be introduced as and when they are required.
Informally, a random variable (r.v.) $X$ denotes possible outcomes of an event. It can be **discrete** (i.e., finite many possible outcomes) or **continuous**.

Some examples of **discrete r.v.**
- A random variable $X \in \{0, 1\}$ denoting outcomes of a coin-toss.
- A random variable $X \in \{1, 2, \ldots, 6\}$ denoting outcome of a dice roll.

Some examples of **continuous r.v.**
- A random variable $X \in (0, 1)$ denoting the bias of a coin.
- A random variable $X$ denoting heights of students in this class.
- A random variable $X$ denoting time to get to your hall from the department.
Discrete Random Variables

- For a discrete r.v. $X$, $p(x)$ denotes the probability that $p(X = x)$
- $p(x)$ is called the probability mass function (PMF)

\[
p(x) \geq 0
\]
\[
p(x) \leq 1
\]
\[
\sum_x p(x) = 1
\]
Continuous Random Variables

- For a continuous r.v. $X$, a probability $p(X = x)$ is meaningless.
- Instead we use $p(X = x)$ or $p(x)$ to denote the probability density at $X = x$.
- For a continuous r.v. $X$, we can only talk about probability within an interval $X \in (x, x + \delta x)$.
  - $p(x)\delta x$ is the probability that $X \in (x, x + \delta x)$ as $\delta x \to 0$.

The probability density $p(x)$ satisfies the following:

$$p(x) \geq 0 \quad \text{and} \quad \int_{x} p(x) \, dx = 1 \quad (\text{note: for continuous r.v., } p(x) \text{ can be } > 1)$$
A word about notation..

- $p(\cdot)$ can mean different things depending on the context
  - $p(X)$ denotes the distribution (PMF/PDF) of an r.v. $X$
  - $p(X = x)$ or $p(x)$ denotes the **probability** or **probability density** at point $x$

- Actual meaning should be clear from the context (but be careful)
- Exercise the same care when $p(\cdot)$ is a specific distribution (Bernoulli, Beta, Gaussian, etc.)
- The following means **drawing a random sample** from the distribution $p(X)$

$$x \sim p(X)$$
Joint Probability Distribution

Joint probability distribution $p(X, Y)$ models probability of co-occurrence of two r.v. $X$, $Y$. For discrete r.v., the joint PMF $p(X, Y)$ is like a table (that sums to 1)

$$
\sum_x \sum_y p(X = x, Y = y) = 1
$$

For continuous r.v., we have joint PDF $p(X, Y)$

$$
\int_x \int_y p(X = x, Y = y)\,dx\,dy = 1
$$
Marginal Probability Distribution

- Intuitively, the probability distribution of one r.v. regardless of the value the other r.v. takes.
- For discrete r.v.’s: \[ p(X) = \sum_y p(X, Y = y), \quad p(Y) = \sum_x p(X = x, Y) \]
- For discrete r.v. it is the sum of the PMF table along the rows/columns.

For continuous r.v.: \[ p(X) = \int_y p(X, Y = y)dy, \quad p(Y) = \int_x p(X = x, Y)dx \]

- Note: Marginalization is also called “integrating out”
Conditional Probability Distribution

- Probability distribution of one r.v. given the value of the other r.v.
- Conditional probability \( p(X|Y = y) \) or \( p(Y|X = x) \): like taking a slice of \( p(X, Y) \)
- For a discrete distribution:

\[
\begin{align*}
\text{For a continuous distribution}^1: \\
\text{picture courtesy: Computer vision: models, learning and inference (Simon Price)}
\end{align*}
\]
Some Basic Rules

- **Sum rule:** Gives the marginal probability distribution from joint probability distribution
  - For discrete r.v.: \( p(X) = \sum_{Y} p(X, Y) \)
  - For continuous r.v.: \( p(X) = \int_{Y} p(X, Y) dY \)

- **Product rule:** \( p(X, Y) = p(Y|X)p(X) = p(X|Y)p(Y) \)

- **Bayes rule:** Gives conditional probability
  \[
  p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}
  \]
  - For discrete r.v.: \( p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_{Y} p(X|Y)p(Y)} \)
  - For continuous r.v.: \( p(Y|X) = \frac{p(X|Y)p(Y)}{\int_{Y} p(X|Y)p(Y) dY} \)

- Also remember the **chain rule**
  \[
  p(X_1, X_2, \ldots, X_N) = p(X_1)p(X_2|X_1) \ldots p(X_N|X_1, \ldots, X_{N-1})
  \]
Independence

- \( X \) and \( Y \) are independent (\( X \perp \perp Y \)) when knowing one tells nothing about the other

\[
\begin{align*}
p(X|Y=y) &= p(X) \\
p(Y|X=x) &= p(Y) \\
p(X,Y) &= p(X)p(Y)
\end{align*}
\]

- \( X \perp \perp Y \) is also called **marginal independence**

- **Conditional independence** (\( X \perp \perp Y|Z \)): independence given the value of another r.v. \( Z \)

\[
p(X, Y|Z = z) = p(X|Z = z)p(Y|Z = z)
\]
Expectation

- **Expectation** or **mean** $\mu$ of an r.v. with PMF/PDF $p(X)$

\[
\mathbb{E}[X] = \sum x p(x) \quad \text{(for discrete distributions)}
\]

\[
\mathbb{E}[X] = \int x p(x) \, dx \quad \text{(for continuous distributions)}
\]

- **Note:** The definition applies to **functions of r.v.** too (e.g., $\mathbb{E}[f(X)]$)

- **Linearity of expectation**

\[
\mathbb{E}[\alpha f(X) + \beta g(Y)] = \alpha \mathbb{E}[f(X)] + \beta \mathbb{E}[g(Y)]
\]

(a very useful property, true even if $X$ and $Y$ are not independent)

- **Note:** Expectations are always w.r.t. the underlying probability distribution of the random variable involved, so sometimes we'll write this explicitly as $\mathbb{E}_p[.]$, unless it is clear from the context
Variance and Covariance

- **Variance** $\sigma^2$ (or “spread” around mean $\mu$) of an r.v. with PMF/PDF $p(X)$

\[
\text{var}[X] = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mu^2
\]

- **Standard deviation:** $\text{std}[X] = \sqrt{\text{var}[X]} = \sigma$

- For two scalar r.v.’s $x$ and $y$, the **covariance** is defined by

\[
\text{cov}[x, y] = \mathbb{E} \left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]
\]

- For vector r.v. $x$ and $y$, the **covariance matrix** is defined as

\[
\text{cov}[x, y] = \mathbb{E} \left[\{x - \mathbb{E}[x]\}\{y^T - \mathbb{E}[y^T]\}\right] = \mathbb{E}[xy^T] - \mathbb{E}[x]\mathbb{E}[y^T]
\]

- Cov. of components of a vector r.v. $x$: $\text{cov}[x] = \text{cov}[x, x]$

- **Note:** The definitions apply to functions of r.v. too (e.g., $\text{var}[f(X)]$)

- **Note:** Variance of sum of independent r.v.’s: $\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]$
Suppose \( y = f(x) = Ax + b \) be a linear function of an r.v. \( x \)

Suppose \( \mathbb{E}[x] = \mu \) and \( \text{cov}[x] = \Sigma \)

- **Expectation of** \( y \)
  \[ \mathbb{E}[y] = \mathbb{E}[Ax + b] = A\mu + b \]

- **Covariance of** \( y \)
  \[ \text{cov}[y] = \text{cov}[Ax + b] = A\Sigma A^T \]

Likewise if \( y = f(x) = a^T x + b \) is a scalar-valued linear function of an r.v. \( x \):

- \( \mathbb{E}[y] = \mathbb{E}[a^T x + b] = a^T \mu + b \)
- \( \text{var}[y] = \text{var}[a^T x + b] = a^T \Sigma a \)

Another very useful property worth remembering
Common Probability Distributions

Important: We will use these extensively to model data as well as parameters

Some discrete distributions and what they can model:

- **Bernoulli**: Binary numbers, e.g., outcome (head/tail, 0/1) of a coin toss
- **Binomial**: Bounded non-negative integers, e.g., # of heads in $n$ coin tosses
- **Multinomial**: One of $K$ (>2) possibilities, e.g., outcome of a dice roll
- **Poisson**: Non-negative integers, e.g., # of words in a document
- .. and many others

Some continuous distributions and what they can model:

- **Uniform**: numbers defined over a fixed range
- **Beta**: numbers between 0 and 1, e.g., probability of head for a biased coin
- **Gamma**: Positive unbounded real numbers
- **Dirichlet**: vectors that sum of 1 (fraction of data points in different clusters)
- **Gaussian**: real-valued numbers or real-valued vectors
- .. and many others
Discrete Distributions
Bernoulli Distribution

- Distribution over a binary r.v. \( x \in \{0, 1\} \), like a coin-toss outcome
- Defined by a probability parameter \( p \in (0, 1) \)

\[
P(x = 1) = p
\]

- Distribution defined as: \( \text{Bernoulli}(x; p) = p^x(1 - p)^{1-x} \)

- Mean: \( \mathbb{E}[x] = p \)
- Variance: \( \text{var}[x] = p(1 - p) \)
Binomial Distribution

- Distribution over number of successes $m$ (an r.v.) in a number of trials
- Defined by two parameters: total number of trials ($N$) and probability of each success $p \in (0, 1)$
- Can think of Binomial as multiple independent Bernoulli trials
- Distribution defined as
  \[ \text{Binomial}(m; N, p) = \binom{N}{m} p^m (1 - p)^{N-m} \]

- Mean: $\mathbb{E}[m] = Np$
- Variance: $\text{var}[m] = Np(1 - p)$
Multinoulli Distribution

- Also known as the **categorical distribution** (models categorical variables)
- Think of a random assignment of an item to one of \( K \) bins - a \( K \) dim. binary r.v. \( \mathbf{x} \) with single 1 (i.e., \( \sum_{k=1}^{K} x_k = 1 \)): **Modeled by a multinoulli**

\[
\begin{bmatrix}
0 & 0 & 0 & \ldots & 0 & 1 & 0 & 0
\end{bmatrix}
\]

length = \( K \)

- Let vector \( \mathbf{p} = [p_1, p_2, \ldots, p_K] \) define the probability of going to each bin
  - \( p_k \in (0, 1) \) is the probability that \( x_k = 1 \) (assigned to bin \( k \))
  - \( \sum_{k=1}^{K} p_k = 1 \)

- The multinoulli is defined as: \( \text{Multinoulli}(\mathbf{x}; \mathbf{p}) = \prod_{k=1}^{K} p_k^{x_k} \)
- Mean: \( \mathbb{E}[x_k] = p_k \)
- Variance: \( \text{var}[x_k] = p_k(1 - p_k) \)
Multinomial Distribution

- Think of repeating the Multinoulli $N$ times
- Like distributing $N$ items to $K$ bins. Suppose $x_k$ is count in bin $k$
  \[ 0 \leq x_k \leq N \quad \forall \ k = 1, \ldots, K, \quad \sum_{k=1}^{K} x_k = N \]
- Assume probability of going to each bin: $p = [p_1, p_2, \ldots, p_K]$
- Multinomial models the bin allocations via a discrete vector $x$ of size $K$
  \[ [x_1 \ x_2 \ \ldots \ x_{k-1} \ x_k \ x_{k-1} \ldots \ x_K] \]
- Distribution defined as
  \[
  \text{Multinomial}(x; N, p) = \binom{N}{x_1x_2\ldots x_K} \prod_{k=1}^{K} p_k^{x_k}
  \]
- Mean: $\mathbb{E}[x_k] = Np_k$
- Variance: $\text{var}[x_k] = Np_k(1 - p_k)$
- Note: For $N = 1$, multinomial is the same as multinoulli
Poisson Distribution

- Used to model a non-negative integer (count) r.v. \( k \)
- Examples: number of words in a document, number of events in a fixed interval of time, etc.
- Defined by a positive rate parameter \( \lambda \)
- Distribution defined as
  \[
  \text{Poisson}(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad k = 0, 1, 2, \ldots
  \]

- Mean: \( \mathbb{E}[k] = \lambda \)
- Variance: \( \text{var}[k] = \lambda \)
Continuous Distributions
Uniform Distribution

- Models a continuous r.v. $x$ distributed uniformly over a finite interval $[a, b]$

\[
\text{Uniform}(x; a, b) = \frac{1}{b - a}
\]

- Mean: $\mathbb{E}[x] = \frac{(b+a)}{2}$
- Variance: $\text{var}[x] = \frac{(b-a)^2}{12}$
Beta Distribution

- Used to model an r.v. $p$ between 0 and 1 (e.g., a probability)
- Defined by two shape parameters $\alpha$ and $\beta$

\[
\text{Beta}(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1 - p)^{\beta-1}
\]

- Mean: $\mathbb{E}[p] = \frac{\alpha}{\alpha + \beta}$
- Variance: $\text{var}[p] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$
- Often used to model the probability parameter of a Bernoulli or Binomial (also conjugate to these distributions)
Gamma Distribution

- Used to model positive real-valued r.v. $x$
- Defined by a **shape parameters** $k$ and a **scale parameter** $\theta$

\[
\text{Gamma}(x; k, \theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)}
\]

- Mean: $\mathbb{E}[x] = k\theta$
- Variance: $\text{var}[x] = k\theta^2$
- Often used to model the rate parameter of Poisson or exponential distribution (conjugate to both), or to model the inverse variance (precision) of a Gaussian (conjugate to Gaussian if mean known)

Note: There is another equivalent parameterization of gamma in terms of **shape** and **rate** parameters ($\text{rate} = 1/\text{scale}$). Another related distribution: Inverse gamma.
Dirichlet Distribution

- Used to model non-negative r.v. vectors $\mathbf{p} = [p_1, \ldots, p_K]$ that sum to 1

\[
0 \leq p_k \leq 1, \quad \forall k = 1, \ldots, K \quad \text{and} \quad \sum_{k=1}^{K} p_k = 1
\]

- Equivalent to a distribution over the $K - 1$ dimensional simplex

- Defined by a $K$ size vector $\mathbf{\alpha} = [\alpha_1, \ldots, \alpha_K]$ of positive reals

- Distribution defined as

\[
\text{Dirichlet}(\mathbf{p}; \mathbf{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}
\]

- Often used to model the probability vector parameters of Multinoulli/Multinomial distribution

- Dirichlet is conjugate to Multinoulli/Multinomial

**Note:** Dirichlet can be seen as a generalization of the Beta distribution. Normalizing a bunch of Gamma r.v.'s gives an r.v. that is Dirichlet distributed.
Dirichlet Distribution

- For \( p = [p_1, p_2, \ldots, p_K] \) drawn from Dirichlet(\( \alpha_1, \alpha_2, \ldots, \alpha_K \))
  
  - Mean: \( \mathbb{E}[p_k] = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k} \)
  
  - Variance: \( \text{var}[p_k] = \frac{\alpha_k}{\alpha_0^2(\alpha_0+1)} \)
    where \( \alpha_0 = \sum_{k=1}^{K} \alpha_k \)

- Note: \( p \) is a point on \((K-1)\)-simplex

- Note: \( \alpha_0 = \sum_{k=1}^{K} \alpha_k \) controls how peaked the distribution is

- Note: \( \alpha_k \)'s control where the peak(s) occur

Plot of a 3 dim. Dirichlet (2 dim. simplex) for various values of \( \alpha \):
Now comes the Gaussian (Normal) distribution..
Univariate Gaussian Distribution

- Distribution over real-valued scalar r.v. $x$
- Defined by a scalar mean $\mu$ and a scalar variance $\sigma^2$
- Distribution defined as
  \[ N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

  ![Graph of Gaussian distributions](image)

- Mean: $\mathbb{E}[x] = \mu$
- Variance: $\text{var}[x] = \sigma^2$
- Precision (inverse variance) $\beta = 1/\sigma^2$
Multivariate Gaussian Distribution

- Distribution over a multivariate r.v. vector $x \in \mathbb{R}^D$ of real numbers
- Defined by a mean vector $\mu \in \mathbb{R}^D$ and a $D \times D$ covariance matrix $\Sigma$

\[
\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}
\]

- The covariance matrix $\Sigma$ must be symmetric and positive definite
  - All eigenvalues are positive
  - $z^\top \Sigma z > 0$ for any real vector $z$
- Often we parameterize a multivariate Gaussian using the inverse of the covariance matrix, i.e., the precision matrix $\Lambda = \Sigma^{-1}$
Multivariate Gaussian: The Covariance Matrix

The covariance matrix can be spherical, diagonal, or full

- **Spherical covariances**
  - $\mu = \begin{bmatrix} 2.0 \\ 1.0 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}$

- **Diagonal covariances**
  - $\mu = \begin{bmatrix} -2.0 \\ 2.0 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 0.3 & 0.0 \\ 0.0 & 1.8 \end{bmatrix}$

- **Full covariances**
  - $\mu = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 0.8 & 0.7 \\ 0.7 & 1.3 \end{bmatrix}$

- $\mu = \begin{bmatrix} 2.0 \\ -1.6 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 2.0 & 0.0 \\ 0.0 & 2.0 \end{bmatrix}$

- $\mu = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 4.0 & 0.0 \\ 0.0 & 1.8 \end{bmatrix}$

- $\mu = \begin{bmatrix} 0.0 \\ 2.0 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 3.9 & -0.5 \\ -0.5 & 1.1 \end{bmatrix}$
Some nice properties of the Gaussian distribution..
Multivariate Gaussian: Marginals and Conditionals

- Given \( x \) having multivariate Gaussian distribution \( N(x | \mu, \Sigma) \) with \( \Lambda = \Sigma^{-1} \). Suppose

\[
\begin{align*}
  x &= \begin{pmatrix} x_a \\ x_b \end{pmatrix}, \quad \mu &= \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \\
  \Sigma &= \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \quad \Lambda &= \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}
\end{align*}
\]

- The marginal distribution is simply

\[
p(x_a) = N(x_a | \mu_a, \Sigma_{aa})
\]

- The conditional distribution is given by

\[
\begin{align*}
  p(x_a | x_b) &= N(x | \mu_a | b, \Lambda_{aa}^{-1}) \\
  \mu_{a|b} &= \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b)
\end{align*}
\]

Thus marginals and conditionals of Gaussians are Gaussians.
Multivariate Gaussian: Marginals and Conditionals

- Given the conditional of an r.v. $y$ and marginal of r.v. $x$, $y$ is conditioned on $x$

$$
p(y|x) = \mathcal{N}(y|Ax + b, L^{-1})
$$

$$
p(x) = \mathcal{N}(x|\mu, \Lambda^{-1})
$$

- Marginal of $y$ and “reverse” conditional are given by

$$
p(x|y) = \mathcal{N}(x|\Sigma\{A^T L(y - b) + \Lambda \mu\}, \Sigma)
$$

$$
p(y) = \mathcal{N}(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^T)
$$

where $\Sigma = (\Lambda + A^T L A)^{-1}$

- Note that the “reverse conditional” $p(x|y)$ is basically the posterior of $x$ is the prior is $p(x)$

- Also note that the marginal $p(y)$ is the predictive distribution of $y$ after integrating out $x$

- Very useful property for probabilistic models with Gaussian likelihoods and/or priors. Also very handy for computing marginal likelihoods.
Gaussians: Product of Gaussians

- Pointwise multiplication of two Gaussians is another (unnormalized) Gaussian

\[ \mathcal{N}(x; \mu, \Sigma) \mathcal{N}(x; \nu, P) = \frac{1}{Z} \mathcal{N}(x; \omega, T), \]

where

\[ T = (\Sigma^{-1} + P^{-1})^{-1} \]
\[ \omega = T(\Sigma^{-1} \mu + P^{-1} \nu) \]
\[ Z^{-1} = \mathcal{N}(\mu, \nu, \Sigma + P) = \mathcal{N}(\nu, \mu, \Sigma + P) \]
Given a $x \in \mathbb{R}^d$ with a multivariate Gaussian distribution

$$N(x; \mu, \Sigma)$$

Consider a linear transform of $x$ into $y \in \mathbb{R}^D$

$$y = Ax + b$$

where $A$ is $D \times d$ and $b \in \mathbb{R}^D$

$y \in \mathbb{R}^D$ will have a multivariate Gaussian distribution

$$N(y; A\mu + b, A\Sigma A^\top)$$
Some Other Important Distributions

- **Wishart** Distribution and **Inverse Wishart** (IW) Distribution: Used to model $D \times D$ p.s.d. matrices
  - Wishart often used as a conjugate prior for modeling precision matrices, IW for covariance matrices
  - For $D = 1$, Wishart is the same as gamma dist., IW is the same as inverse gamma (IG) dist.

- **Normal-Wishart** Distribution: Used to model mean and precision matrix of a multivar. Gaussian
  - **Normal-Inverse Wishart (NIW):** Used to model mean and cov. matrix of a multivar. Gaussian
  - For $D = 1$, the corresponding distr. are Normal-Gamma and Normal-Inverse Gamma (NIG)

- **Student-t** Distribution (a more robust version of Normal distribution)
  - Can be thought of as a mixture of infinite many Gaussians with different precisions (or a single Gaussian with its precision/precision matrix given a gamma/Wishart prior and integrated out)

Please refer to PRML (Bishop) Chapter 2 + Appendix B, or MLAPP (Murphy) Chapter 2 for more details