# Probability and Statistics Refresher for Probabilisitic Machine Learning

Piyush Rai

CSE Department, IIT Kanpur

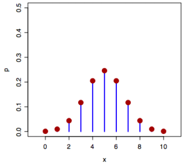
# Some Basic Concepts You Should Know About

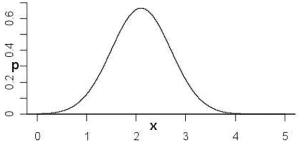
- Random variables (discrete/continuous), probability distributions over discrete/continuous r.v.'s
- Notions of joint, conditional, and marginal distributions
- Properties of random variables (and of functions of random variables)
  - Expectation and variance/covariance
- Examples of various probability distributions (and when is each appropriate) and their properties
  - Mean/mode/variance etc of a probability distribution
- Multivariate Gaussian distribution and its properties (very important)
- Functions of distributions, e.g., KL divergence, Entropy, etc.

**Note:** This is only a (very!) quick review of these things. Please refer to a text such as PRML (Bishop) Chapter 2 + Appendix B, or PML-1 (Murphy) Chapter 2 and 3 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
- Can be **discrete** (i.e., finite many possible outcomes) or **continuous**



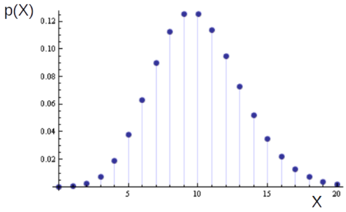




- Some examples of **discrete r.v.**
  - A random variable $X \in \{0, 1\}$ denoting outcomes of a coin-toss
  - A random variable $X \in \{1, 2, \ldots, 6\}$ denoteing outcome of a dice roll
- Some examples of **continuous r.v.**
  - A random variable $X \in (0, 1)$ denoting the bias of a coin
  - A random variable $X$ denoting heights of students in CS698S
  - A random variable $X$ denoting time to get to your hall from the department

# Discrete Random Variables

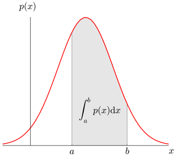
- For a discrete r.v. $X$, $p(x)$ denotes the probability that $p(X = x)$
- $p(x)$ is called the probability mass function (PMF)

$$\begin{aligned} p(x) &\geq 0 \\ p(x) &\leq 1 \\ \sum_x p(x) &= 1 \end{aligned}$$



# 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

$$p(x) \geq 0 \quad \text{and} \quad \int_x p(x)dx = 1 \qquad \text{(note: for continuous r.v., } p(x) \text{ can be } > 1)$$

# A word about notation..

- $p(.)$ 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 $p(.)$ is a specific distribution (Bernoulli, Beta, Gaussian, etc.)
- The following means **drawing a random sample** from the distribution $p(X)$

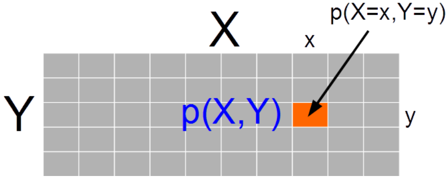
$$x \sim p(X)$$

# 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)

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

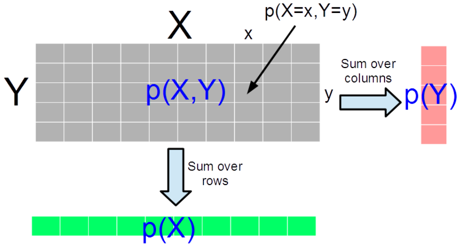



For continuous r.v., we have joint PDF $p(X, Y)$

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

# Marginal Probability Distribution

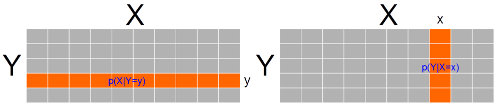
- Intuitively, the probability distribution of one r.v. regardless of the value the other r.v. takes
- For discrete r.v.'s: $p(X) = \sum_y p(X, Y = y), \quad p(Y) = \sum_x p(X = x, Y)$
- For discrete r.v. it is the sum of the PMF table along the rows/columns



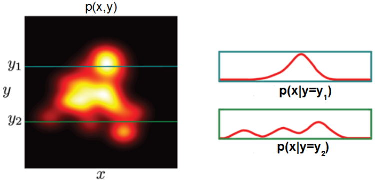

- For continuous r.v.: $p(X) = \int_y p(X, Y = y)dy, \quad p(Y) = \int_x p(X = x, Y)dx$
- Note: Marginalization is also called "integrating out" (especially in Bayesian learning)

# 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:




- For a continuous distribution[1]:




[1] Picture courtesy: Computer vision: models, learning and inference (Simon Price)

# 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

$$\boxed{p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}}$$

  - 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**

$$p(X_1, X_2, \ldots, X_N) = p(X_1)p(X_2|X_1)\ldots p(X_N|X_1, \ldots, X_{N-1})$$

# CDF and Quantiles

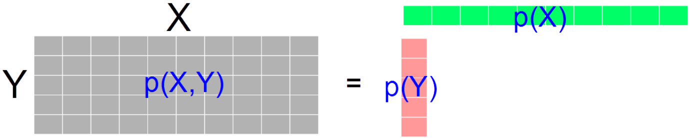
- Cumulative distribution function (CDF): $F(x) = p(X \leq x)$
- $\alpha \leq 1$ quantile is defined as the $x_\alpha$ s.t.

$$p(X \leq x_\alpha) = \alpha$$

# Independence

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

$$\begin{aligned} p(X|Y=y) &= p(X) \\ p(Y|X=x) &= p(Y) \\ p(X,Y) &= p(X)p(Y) \end{aligned}$$




- $X \perp\!\!\!\perp Y$ 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 $p(X)$

$$
\begin{aligned}
\mathbb{E}[X] &= \sum_x xp(x) \qquad \text{(for discrete distributions)} \\
\mathbb{E}[X] &= \int_x xp(x)dx \qquad \text{(for continuous distributions)}
\end{aligned}
$$

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

- **Rule of iterated/total expectation**

$$\mathbb{E}_{p(X)}[X] = \mathbb{E}_{p(Y)}[\mathbb{E}_{p(X|Y)}[X|Y]]$$

# 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. $\boldsymbol{x}$ and $\boldsymbol{y}$, the **covariance matrix** is defined as

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

- Cov. of components of a vector r.v. $\boldsymbol{x}$: $\text{cov}[\boldsymbol{x}] = \text{cov}[\boldsymbol{x},\boldsymbol{x}]$
- **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: $\text{var}[X+Y] = \text{var}[X] + \text{var}[Y]$

# KL Divergence

- Kullback–Leibler divergence between two probability distributions $p(X)$ and $q(X)$

$$
\begin{aligned}
KL(p||q) &= \int p(X)\log\frac{p(X)}{q(X)}dX = -\int p(X)\log\frac{q(X)}{p(X)}dX \qquad \text{(for continuous distributions)} \\
KL(p||q) &= \sum_{k=1}^{K} p(X=k)\log\frac{p(X=k)}{q(X=k)} \qquad \text{(for discrete distributions)}
\end{aligned}
$$

- It is non-negative, i.e., $KL(p||q) \geq 0$, and zero if and only if $p(X)$ and $q(X)$ are the same
- For some distributions, e.g., Gaussians, KL divergence has a closed form expression
- KL divergence is not symmetric, i.e., $KL(p||q) \neq KL(q||p)$

# Entropy

- Entropy of a continuous/discrete distribution $p(X)$

$$
\begin{aligned}
H(p) &= -\int p(X)\log p(X)dX \\
H(p) &= -\sum_{k=1}^{K} p(X=k)\log p(X=k)
\end{aligned}
$$

- In general, a peaky distribution would have a smaller entropy than a flat distribution
- Note that the KL divergence can be written in terms of expetation and entropy terms

$$
KL(p||q) = \mathbb{E}_{p(X)}[-\log q(X)] - H(p)
$$

- Some other definition to keep in mind: conditional entropy, joint entropy, mutual information, etc.

# Transformation of Random Variables

Suppose $\boldsymbol{y} = f(\boldsymbol{x}) = \mathrm{A}\boldsymbol{x} + \mathrm{b}$ be a linear function of an r.v. $\boldsymbol{x}$

Suppose $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$ and $\mathrm{cov}[\boldsymbol{x}] = \Sigma$

- Expectation of $\boldsymbol{y}$

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\mathrm{A}\boldsymbol{x} + \mathrm{b}] = \mathrm{A}\boldsymbol{\mu} + \mathrm{b}$$

- Covariance of $\boldsymbol{y}$

$$\mathrm{cov}[\boldsymbol{y}] = \mathrm{cov}[\mathrm{A}\boldsymbol{x} + \mathrm{b}] = \mathrm{A}\Sigma\mathrm{A}^T$$

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

- $\mathbb{E}[y] = \mathbb{E}[\boldsymbol{a}^T\boldsymbol{x} + b] = \boldsymbol{a}^T\boldsymbol{\mu} + b$
- $\mathrm{var}[y] = \mathrm{var}[\boldsymbol{a}^T\boldsymbol{x} + b] = \boldsymbol{a}^T\Sigma\boldsymbol{a}$

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

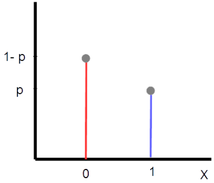
# Discrete Distributions

# 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: $\text{Bernoulli}(x; p) = p^x(1 - p)^{1-x}$

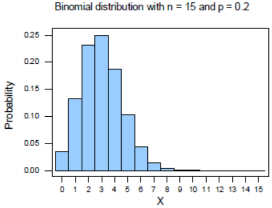



- Mean: $\mathbb{E}[x] = p$
- Variance: $\text{var}[x] = p(1 - p)$

# 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. $\boldsymbol{x}$ with single 1 (i.e., $\sum_{k=1}^{K} x_k = 1$): **Modeled by a multinoulli**

$$\underbrace{[0 \quad 0 \quad 0 \quad \ldots 0 \quad 1 \quad 0 \quad 0]}_{\text{length} = K}$$

- Let vector $\boldsymbol{p} = [p_1, p_2, \ldots, p_K]$ 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: $\text{Multinoulli}(\boldsymbol{x}; \boldsymbol{p}) = \prod_{k=1}^{K} p_k^{x_k}$
- 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 $k$

$$0 \leq x_k \leq N \quad \forall\ k = 1, \ldots, K, \qquad \sum_{k=1}^{K} x_k = N$$

- Assume probability of going to each bin: $\boldsymbol{p} = [p_1, p_2, \ldots, p_K]$
- Multonomial models the bin allocations via a discrete vector $\boldsymbol{x}$ of size $K$

$$[x_1 \quad x_2 \quad \ldots x_{k-1} \quad x_k \quad x_{k-1} \ldots \quad x_K]$$

- Distribution defined as

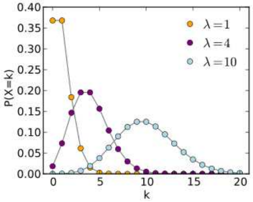
$$\text{Multinomial}(\boldsymbol{x}; N, \boldsymbol{p}) = \binom{N}{x_1 x_2 \ldots x_K} \prod_{k=1}^{K} p_k^{x_k}$$

- 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

$$\text{Poisson}(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad k = 0, 1, 2, \ldots$$




- Mean: $\mathbb{E}[k] = \lambda$
- Variance: $\text{var}[k] = \lambda$

# The Empirical Distribution

- Given a set of points $\phi_1, \ldots, \phi_K$, the empirical distribution is a discrete distribution defined as

$$p_{emp}(A) = \frac{1}{K}\sum_{k=1}^{K}\delta_{\phi_k}(A)$$

where $\delta_\phi(.)$ is the **dirac function** located at $\phi$, s.t.

$$\delta_\phi(A) = \begin{cases} 1 & \text{if } \phi \in A \\ 0 & \text{if } \phi \in A \end{cases}$$

- The "weighted" version of the empirical distribution is

$$p_{emp}(A) = \sum_{k=1}^{K} w_k \delta_{\phi_k}(A) \qquad \left(\text{where } \sum_{k=1}^{K} w_k = 1\right)$$

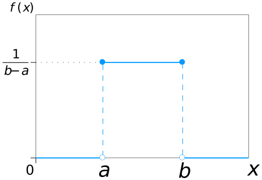
and the weights and points $(w_k, \phi_k)_{k=1}^{K}$ together define this discrete distribution

# Continuous Distributions

# Uniform Distribution

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

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

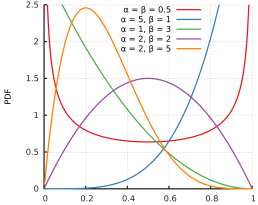



- Mean: $\mathbb{E}[x] = \frac{(b+a)}{2}$
- Variance: $\text{var}[x] = \frac{(b-a)^2}{12}$

# Beta Distribution

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

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

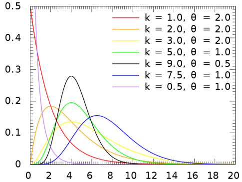



- Mean: $\mathbb{E}[p] = \frac{\alpha}{\alpha+\beta}$
- Variance: $\text{var}[p] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Often used to model the probability parameter of a Bernoulli or Binomial (also **conjugate** to these distributions)

# 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 (conjuate to Gaussian if mean known)

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

# Dirichlet Distribution

- Used to model non-negative r.v. vectors $\boldsymbol{p} = [p_1, \ldots, p_K]$ 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 $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ of positive reals
- Distribution defined as

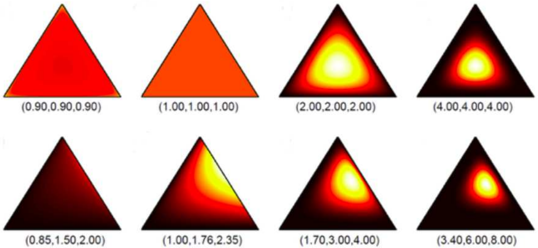
$$\text{Dirichlet}(\boldsymbol{p}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}$$

- 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 $\boldsymbol{p} = [p_1, p_2, \ldots, p_K]$ 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: $\text{var}[p_k] = \frac{\alpha_k(\alpha_0 - \alpha_k}{\alpha_0^2(\alpha_0+1)}$ where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$
- Note: $\boldsymbol{p}$ 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 $\boldsymbol{\alpha}$:

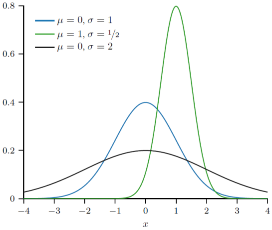


Now comes the
Gaussian (Normal) distribution..

# 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

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

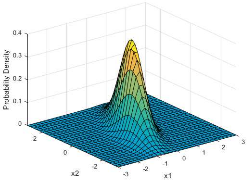



- Mean: $\mathbb{E}[x] = \mu$
- Variance: $\text{var}[x] = \sigma^2$
- Precision (inverse variance) $\beta = 1/\sigma^2$

# Multivariate Gaussian Distribution

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

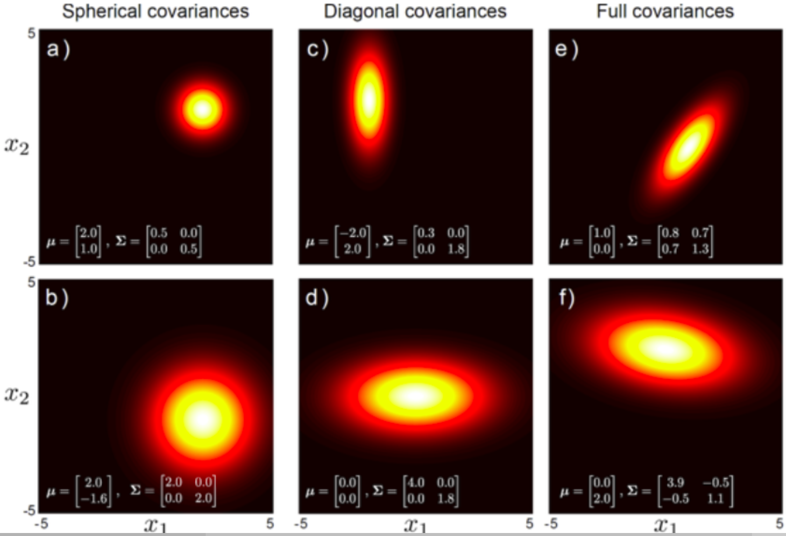
$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$




- The covariance matrix $\Sigma$ must be symmetric and positive definite
  - All eigenvalues are positive
  - $\boldsymbol{z}^\top \Sigma \boldsymbol{z} > 0$ for any real vector $\boldsymbol{z}$
- 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



# Some nice properties of the Gaussian distribution..

# Multivariate Gaussian: Marginals and Conditionals

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

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

- The marginal distribution is simply

$$p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \Sigma_{aa})$$

- The conditional distribution is given by

$$\begin{aligned} p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

**Thus marginals and conditionals of Gaussians are Gaussians**

# Multivariate Gaussian: Marginals and Conditionals

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

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right) \\ p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \end{aligned}$$

- Marginal of $\boldsymbol{y}$ and "reverse" conditional are given by

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \\ p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \end{aligned}$$

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

- Note that the "reverse conditional" $p(\boldsymbol{x}|\boldsymbol{y})$ is basically the posterior of $\boldsymbol{x}$ is the prior is $p(\boldsymbol{x})$
- Also note that the marginal $p(\boldsymbol{y})$ is the predictive distribution of $\boldsymbol{y}$ after integrating out $\boldsymbol{x}$
- Very useful property for probabilistic models with Gaussian likelihoods and/or priors. Also very handly for computing **marginal likelihoods**.

# Gaussians: Product of Gaussians

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

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

where

$$\begin{aligned} \mathbf{T} &= (\boldsymbol{\Sigma}^{-1} + \mathbf{P}^{-1})^{-1} \\ \boldsymbol{\omega} &= \mathbf{T}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{P}^{-1}\boldsymbol{\nu}) \\ Z^{-1} &= \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\nu}, \boldsymbol{\Sigma} + \mathbf{P}) = \mathcal{N}(\boldsymbol{\nu}; \boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{P}) \end{aligned}$$

# Multivariate Gaussian: Linear Transformations

- Given a $\boldsymbol{x} \in \mathbb{R}^d$ with a multivariate Gaussian distribution

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$$

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

$$\boldsymbol{y} = \mathrm{A}\boldsymbol{x} + \mathrm{b}$$

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

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

$$\mathcal{N}(\boldsymbol{y}; \mathrm{A}\boldsymbol{\mu} + \mathrm{b}, \mathrm{A}\Sigma\mathrm{A}^\top)$$

# 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)