Some Essentials of Probability for Probabilistic Machine Learning

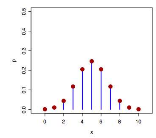
Piyush Rai
IIT Kanpur

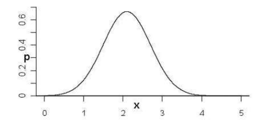
Probabilistic Machine Learning (CS772A)

Jan 4, 2016

Random Variables

- A random variable (r.v.) X denotes possible outcomes of an event
- Can be **discrete** (i.e., finite many possible outcomes) or **continuous**



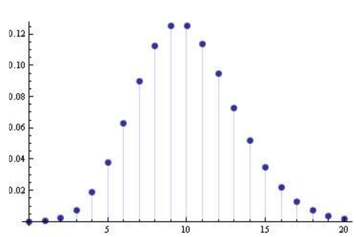


- Some examples of **discrete r.v.**
  - A random variable $X \in \{0, 1\}$ denoting outcomes of a coin-toss
  - A random variable $X \in \{1, 2, \ldots, 6\}$ denoteing outcome of a dice roll
- Some examples of **continuous r.v.**
  - A random variable $X \in (0, 1)$ denoting the bias of a coin
  - A random variable X denoting heights of students in CS772
  - A random variable X denoting time to get to your hall from the department
- An r.v. is associated with a probability mass function or prob. distribution

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)

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



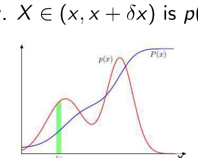
Picture courtesy: johndcook.com

Continuous Random Variables

- For a continuous r.v. X, a probability $p(X = x)$ is meaningless
- Instead we use $p(x)$ to denote the probability density function (PDF)

$$p(x) \geq 0 \quad \text{and} \quad \int_x p(x)dx = 1$$

- Probability that a cont. r.v. $X \in (x, x + \delta x)$ is $p(x)\delta x$ as $\delta x \to 0$



- Probability that X lies between $(-\infty, z)$ is given by the **cumulative distribution function** (CDF) $P(z)$ where

$$P(z) = p(X \leq z) = \int_{-\infty}^{z} p(x)dx \qquad \text{and} \quad p(x) = |P'(z)|_{z=x}$$

Picture courtesy: PRML (Bishop, 2006)

A word about notation..

- $p(.)$ can mean different things depending on the context
  - $p(X)$ denotes the 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 sample** from the distribution $p(X)$

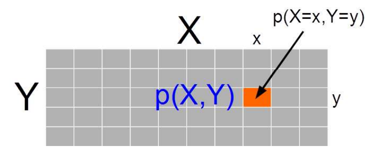
$$x \sim p(X)$$

Joint Probability

Joint probability $p(X, Y)$ 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 $p(X, Y)$

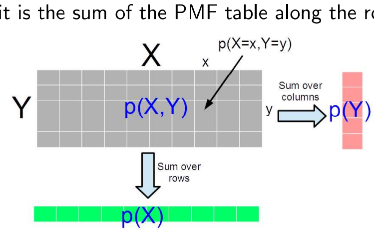
$$\int_x \int_y p(X = x, Y = y) dx dy = 1$$

Marginal Probability

- For discrete r.v.

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

- For discrete r.v. it is the sum of the PMF table along the rows/columns

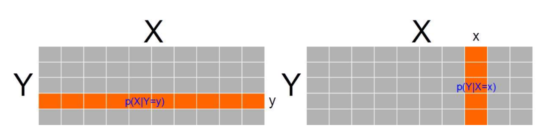


- For continuous r.v.

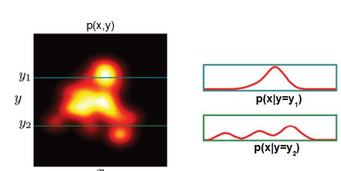
$$p(X) = \int_y p(X, Y = y) dy, \quad p(Y) = \int_x p(X = x, Y) dx$$

Conditional Probability

- Meaning: Probability of one event when we know the outcome of the other
- Conditional probability $p(X|Y)$ or $p(Y|X)$: 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
  - 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

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

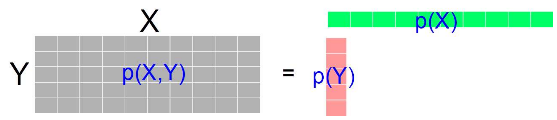
  - For discrete r.v.: $p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$
  - For continuous r.v.: $p(Y|X) = \frac{p(X|Y)p(Y)}{\int_Y p(X|Y)p(Y)dY}$
- Bayes rule is also central to parameter estimation (more on this later)
- Also remember the **chain rule**

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

Independence

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

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



- $X \perp\!\!\!\perp Y$ is also called **marginal independence**
- **Conditional independence** ($X \perp\!\!\!\perp Y|Z$): independence when another event Z is observed

$$p(X, Y|Z) = p(X|Z)p(Y|Z)$$

Expectation

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

$$\mathbb{E}[X] = \sum_x xp(x) \quad \text{(for discrete distributions)}$$

$$\mathbb{E}[X] = \int_x xp(x)dx \quad \text{(for continuous distributions)}$$

- **Note:** The definition applies to **functions of r.v.** too (e.g., $\mathbb{E}[f(x)]$)
- **Linearity of expectation** (very important/useful property)

$$\mathbb{E}[\alpha f(x) + \beta g(x)] = \alpha\mathbb{E}[f(x)] + \beta\mathbb{E}[g(x)]$$

Variance and Covariance

- **Variance** σ^2 (or "spread" around mean) 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:** $\text{std}[X] = \sqrt{\text{var}[X]} = \sigma$
- **Note:** The definition applies to functions of r.v. too (e.g., $\text{var}[f(X)]$)
- For r.v. x and y, the **covariance** is defined by

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

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

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

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

Transformation of Random Variables

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

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

- Expectation of $\boldsymbol{y}$

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

- Covariance of $\boldsymbol{y}$

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

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

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

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

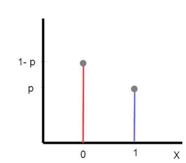
Discrete Distributions

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

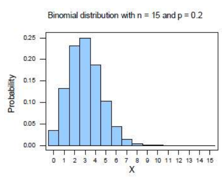


- Mean: $\mathbb{E}[x] = p$
- Variance: $\text{var}[x] = p(1-p)$

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**

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

- Let vector $\boldsymbol{p} = [p_1, p_2, \dots, p_K]$ 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

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

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

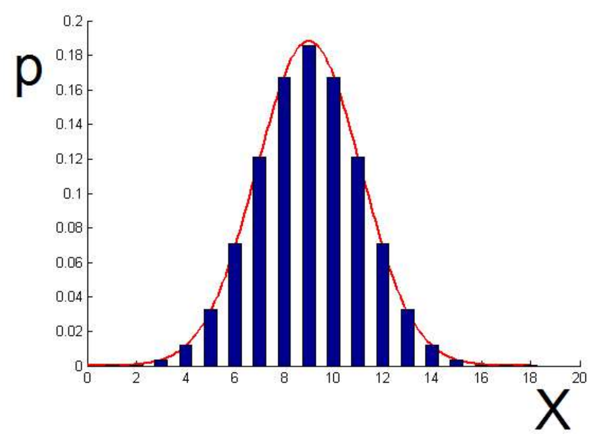
$$[x_1 \quad x_2 \quad \dots x_{k-1} \quad x_k \quad x_{k-1} \dots \quad x_K]$$

- Distribution defined as

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

- 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

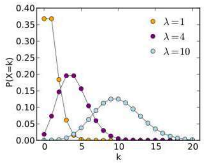
Multinoulli/Multinomial: Pictorially



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 λ
- Distribution defined as

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



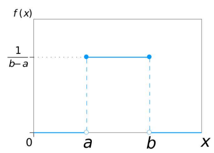
- Mean: $\mathbb{E}[k] = \lambda$
- Variance: $\text{var}[k] = \lambda$

Continuous Distributions

Uniform Distribution

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

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

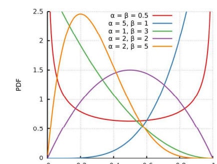


- Mean: $\mathbb{E}[x] = \frac{(b+a)}{2}$
- Variance: $\text{var}[x] = \frac{(b-a)^2}{12}$

Beta Distribution

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

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

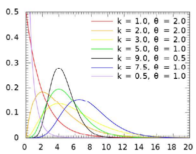


- Mean: $\mathbb{E}[p] = \frac{\alpha}{\alpha+\beta}$
- Variance: $\text{var}[p] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Often used to model the probability parameter of a Bernoulli or Binomial (also **conjugate** to these distributions)

Gamma Distribution

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

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



- 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, or to model the inverse variance of a Gaussian

Note: There is another equivalent parameterization of gamma in terms of **shape** and **rate** parameters

Dirichlet Distribution

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

$$0 \le p_k \le 1, \quad \forall k = 1, \dots, K \quad \text{and} \quad \sum_{k=1}^{K} p_k = 1$$

- Equivalent to a distribution over the $K - 1$ dimensional simplex
- Defined by a K size vector $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_K]$ of positive reals
- Distribution defined as

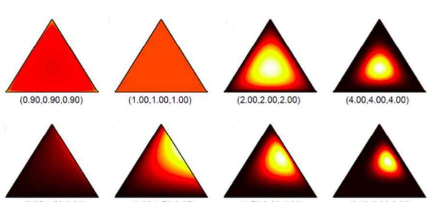
$$\text{Dirichlet}(\boldsymbol{p}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}$$

- Often used to model parameters of Multinoulli/Multinomial
- 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 $\boldsymbol{p} = [p_1, p_2, \dots, p_K]$ drawn from $\text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_K)$
  - 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: α_k's control where the peak(s) occur

Plot of a 3 dim. Dirichlet (2 dim. simplex) for various values of $\boldsymbol{\alpha}$:



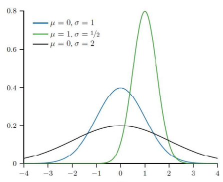
Picture courtesy: Computer vision: models, learning and inference (Simon Price)

Now comes the
Gaussian (Normal) distribution..

Univariate Gaussian Distribution

- Distribution over real-valued scalar r.v. x
- Defined by a scalar **mean** μ and a scalar **variance** σ^2
- Distribution defined as

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

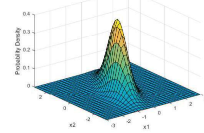


- Mean: $\mathbb{E}[x] = \mu$
- Variance: $\text{var}[x] = \sigma^2$
- Precision (inverse variance) $\beta = 1/\sigma^2$

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** $\boldsymbol{\Sigma}$

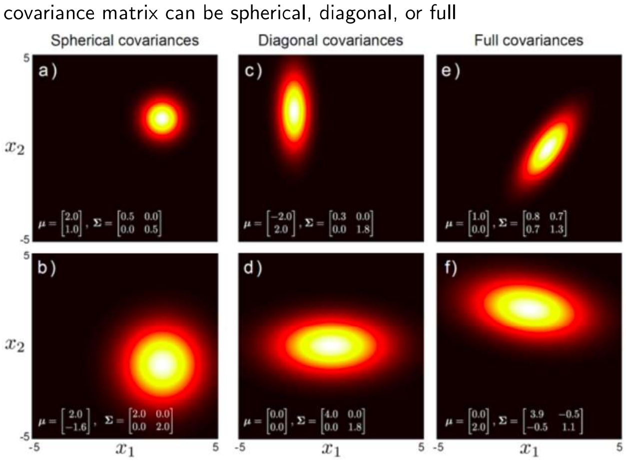
$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$



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

Multivariate Gaussian: The Covariance Matrix

The covariance matrix can be spherical, diagonal, or full



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

Some nice properties of the Gaussian distribution..

Multivariate Gaussian: Marginals and Conditionals

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

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

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

- The marginal distribution is simply

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

- The conditional distribution is given by

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

Thus marginals and conditionals of Gaussians are Gaussians

Multivariate Gaussian: Marginals and Conditionals

- Given the conditional and marginal of r.v. being conditioned on

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

- Marginal and "reverse" conditional are given by

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

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

- 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

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

Multivariate Gaussian: Affine Transforms

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

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

- Consider an affiine transform of $\boldsymbol{x}$ into $\mathbb{R}^D$

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

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

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

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

Exponential Family

- An exponential family distribution is defined as

$$p(x;\theta) = h(x)e^{\eta(\theta)T(x)-A(\theta)}$$

- θ is called the parameter of the family
- $h(x)$, $\eta(\theta)$, $T(x)$, and $A(\theta)$ are known functions
- $p(.)$ depends on x only through $T(x)$
- $T(x)$ is called the **sufficient statistics**: summarizes the entire $p(x;\theta)$
- Exponential family is the only family for which **conjugate priors** exist (often also in the exponential family)
- Many other nice properties (especially useful in Bayesian inference)

Many well-known distribution (Bernoulli, Binomial, categorical, beta, gamma, Gaussian, etc.) are exponential family distributions

`https://en.wikipedia.org/wiki/Exponential_family`

Binomial as Exponential Family

- Recall the exponential family distribution

$$p(x;\theta) = h(x)e^{\eta(\theta)T(x)-A(\theta)}$$

- Binomial in the usual form:

$$\text{Binomial}(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

- Can re-express it as

$$\binom{n}{x} e^{\left(x \log\left(\frac{p}{1-p}\right) + n\log(1-p)\right)}$$

  - $h(x) = \binom{n}{x}$
  - $\eta(\theta) = \log\left(\frac{p}{1-p}\right)$
  - $T(x) = x$
  - $A(\theta) = -n\log(1-p)$

Gaussian as Exponential Family

- Recall the exponential family distribution

$$p(x;\theta) = h(x)e^{\eta(\theta)T(x)-A(\theta)}$$

- Gaussian in the usual form:

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

- Can re-express it as $p(x;\theta) = h(x)e^{\eta(\theta)T(x)-A(\theta)}$ where
  - $h(x) = \frac{1}{\sqrt{2\pi}}$
  - $\eta(\theta) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T$
  - $T(x) = (x, x^2)^T$
  - $A(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$

Conjugate Priors

- Given a distribution $p(x|\theta)$
- We say $p(\theta)$ is conjugate to $p(x|\theta)$ if

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

  has the same form as $p(\theta)$
- Many pairs of distributions are conjugate to each other, e.g.,
  - Gaussian-Gaussian
  - Bernoulli-Beta
  - Poisson-Gamma
  - .. and many others
- More on this in the next class..

Next class: Parameter estimation in probabilistic models