Lecture 1a: Basic Concepts and Recaps

Cédric Archambeau

Centre for Computational Statistics and Machine Learning
Department of Computer Science
University College London

c.archambeau@cs.ucl.ac.uk

Advanced Topics in Machine Learning (MSc in Intelligent Systems)
January 2008

Outline

Lecture 1:

- Course info
- Notations, definitions, recaps ...
- Linear Models for regression
- Gaussian processes for regression

Lecture 2:

- Hidden Markov models and linear state space models
- Nonlinear state space models
- Applications of particle filters
- Guest speaker: **Frank Wood** (Gatsby unit)

Outline

Lecture 3:

- Dirichlet distribution and its representations
- Dirichlet processes and infinite mixtures
- Dirichlet process mixtures of regressors
- Guest speaker: **Yee Whye Teh** (Gatsby unit)

Lecture 4: (on Tuesday!)

- Unscented Kalman filters and extensions
- Dirichlet process mixtures of linear dynamical systems
- Guest speakers: **Simon Julier** (CS) and **David Barber** (CSML).

Lecture 5:

- Introduction to Ito calculus and stochastic differential equations
- Continuous-time stochastic processes
- Wiener process, Diffusion processes, Markov jump processes, ...

Practical info

Where, when?

- Week 1 to 5.
- Tuesdays 14:00-17:00: Malet Place room 1.04.
- Fridays 10:00-13:00: Rockefeller Building room 339.

What?

- Lectures
- Guest speakers
- Individual report

Exam:

- Written Examination (2.5 hours, 50%)
- Coursework (50%)

To pass you must obtain an average of at least 50% when the coursework (1 out of 2) and exam components (2 out of 4) are weighted together.

Practical info (continued)

Individual report:

- Project starts on Tuesday 05/02.
- Report is due **before 9:00 on Monday 25/02**: send an electronic copy to me via email **and** hand in a hard copy to the CS reception[1].
- Literature review, implementation and comparison.
- No longer than 10 pages (including figures), minimal font size 11pt, no less than 25 mm margins.
- Instructions will follow.

Reports that are handed in late will be penalised as follows: 25% penalty per day late. NO CREDIT will be given afterwards.

Questions: only by email (or after class).

[1] Malet Place Engineering building, 5th floor.

Some notations, definitions, etc.

Bold symbols denote column **vectors**:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} = (x_1, \ldots, x_D)^\top.$$

Capitalised bold symbols denote **matrices**:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1Q} \\ \vdots & \ddots & \vdots \\ x_{P1} & \ldots & x_{PQ} \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{P1} \\ \vdots & \ddots & \vdots \\ x_{1Q} & \ldots & x_{PQ} \end{pmatrix}^\top.$$

Some notations, definitions, etc.

If the function $p(\mathbf{x})$ is the **probability density function** of a continuous random variable X, then

$$\forall \mathbf{x} \in \mathbb{R}^D : p(\mathbf{x}) \geqslant 0, \quad \text{and} \quad \int p(\mathbf{x}) d\mathbf{x} = 1.$$

The **expectation** of $f(\mathbf{x})$ is defined as

$$\langle f(\mathbf{x}) \rangle = \int f(\mathbf{x})\ p(\mathbf{x}) d\mathbf{x}.$$

Examples:

- The **mean**: $\boldsymbol{\mu} = \langle \mathbf{x} \rangle$.
- The **covariance** matrix: $\boldsymbol{\Sigma} = \left\langle (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \right\rangle$.

The covariance is symmetric and positive semi-definite, i.e. all its eigenvalues are non-negative.

Sum rule of probability (marginalisation): $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y}$.

Product rule of probability: $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y}) p(\mathbf{y})$.

Some notations, definitions, etc. (continued)

Bayes' rule allows us to update a prior belief on some $\mathbf{y}$ into a posterior belief, based on the observations $\mathbf{x}$:

$$\underbrace{p(\mathbf{y}|\mathbf{x})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{x}|\mathbf{y})}^{\text{likelihood}}\overbrace{p(\mathbf{y})}^{\text{prior}}}{\underbrace{p(\mathbf{x})}_{\text{evidence}}}.$$

The normalising constant is known as the *evidence*, the *marginal likelihood* or the *partition function*.

Proof

Bayes' rule follows from the product rule:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}).$$

Multivariate Gaussian distribution

Let X be a D-dimensional Gaussian random vector. Its density is defined as

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-D/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\},$$

where $\boldsymbol{\mu} \in \mathbb{R}^{D \times 1}$ is the mean and $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ is the covariance matrix.

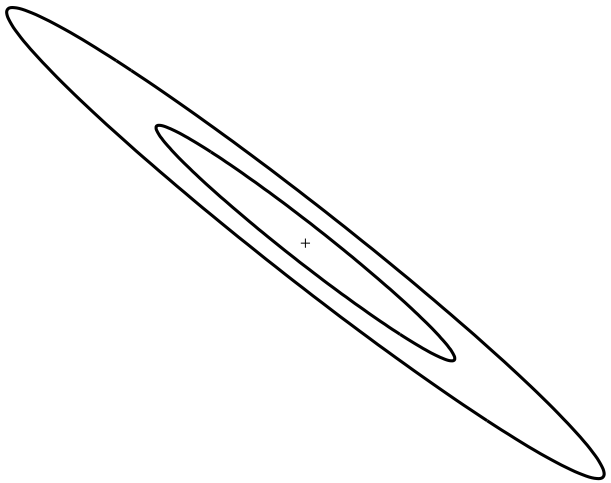


Figure: 2-dimensional Gaussian.

Gaussian identities

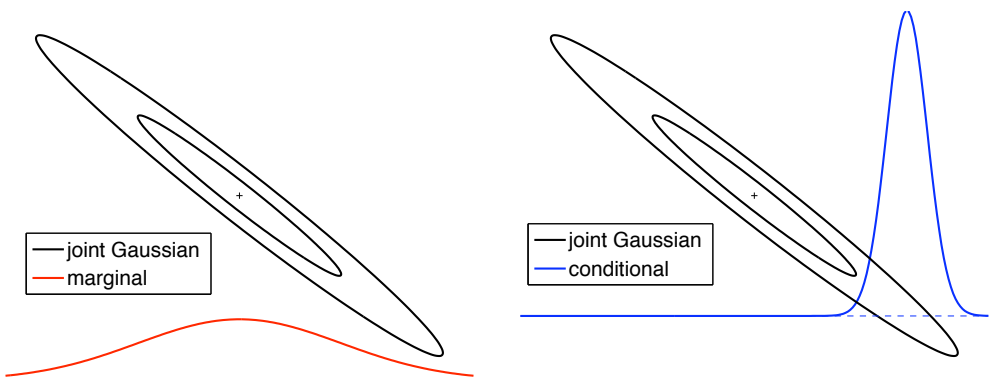
Let X and Y be **jointly Gaussian**:

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{y}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \boldsymbol{\Sigma}_{\mathbf{xy}}^{\top} & \boldsymbol{\Sigma}_{\mathbf{yy}} \end{bmatrix} \right).$$

The **marginal** $p(\mathbf{x})$ is Gaussian with mean $\boldsymbol{\mu}_{\mathbf{x}}$ and covariance $\boldsymbol{\Sigma}_{\mathbf{xx}}$.

The **conditional** $p(\mathbf{x}|\mathbf{y})$ is Gaussian with mean and covariance equal to

$$\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{xy}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}),$$
$$\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{xy}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \boldsymbol{\Sigma}_{\mathbf{xy}}^{\top}.$$



(a) Marginal. (b) Conditional.

Gaussian identities (continued)

Consider the following two Gaussian distributions:

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_{\mathbf{xx}}),$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Lambda}).$$

The **marginal** $p(\mathbf{y})$ is Gaussian with mean and covariance given by

$$\boldsymbol{\mu}_\mathbf{y} = \mathbf{A}\boldsymbol{\mu}_\mathbf{x} + \mathbf{b},$$
$$\boldsymbol{\Sigma}_{\mathbf{yy}} = \boldsymbol{\Lambda} + \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{A}^\top.$$

The **posterior** $p(\mathbf{x}|\mathbf{y})$ is Gaussian with mean and covariance equal to

$$\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\{\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_\mathbf{x} + \mathbf{A}^\top\boldsymbol{\Lambda}^{-1}(\mathbf{y} - \mathbf{b})\},$$
$$\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} = (\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} + \mathbf{A}^\top\boldsymbol{\Lambda}^{-1}\mathbf{A})^{-1}.$$

Gamma distribution

For $x \in \mathbb{R}^+$, the **Gamma** density is defined as follows:

$$\mathcal{G}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\}, \quad \alpha, \beta > 0,$$

where $\Gamma(u) \equiv \int_0^\infty v^{u-1} e^{-v} dv$ is the *gamma function*. We have

$$\langle x \rangle = a/b \quad \text{and} \quad \langle \ln x \rangle = \psi(a) - \ln b.$$

The function $\psi(\cdot) \equiv (\ln \Gamma)'(\cdot)$ is the *digamma* function.

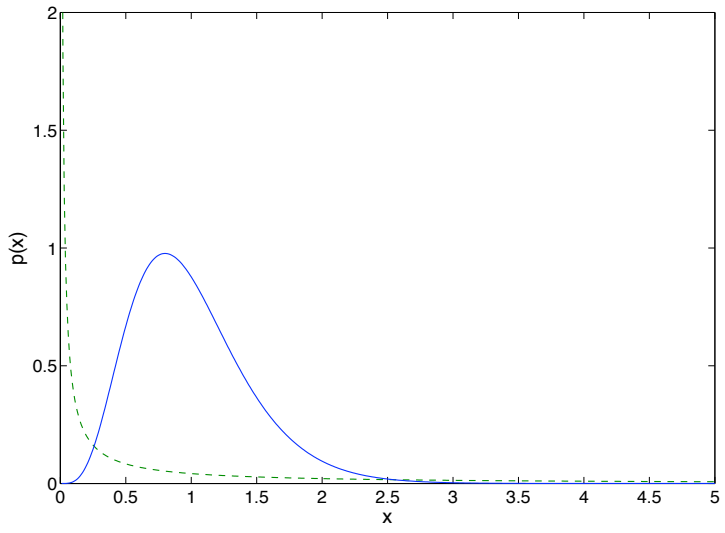


Figure: Gamma distribution for two values of a and b.

Multivariate Student-t distribution

The **Student-t** density[2] is defined as follows:

$$\mathcal{S}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \left(1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-\frac{\nu+D}{2}}.$$

Parameter $\nu > 0$ is the **shape parameter**:

- The Cauchy density is recovered for $\nu = 1$.
- The Gaussian density is recovered when $\nu \to \infty$.

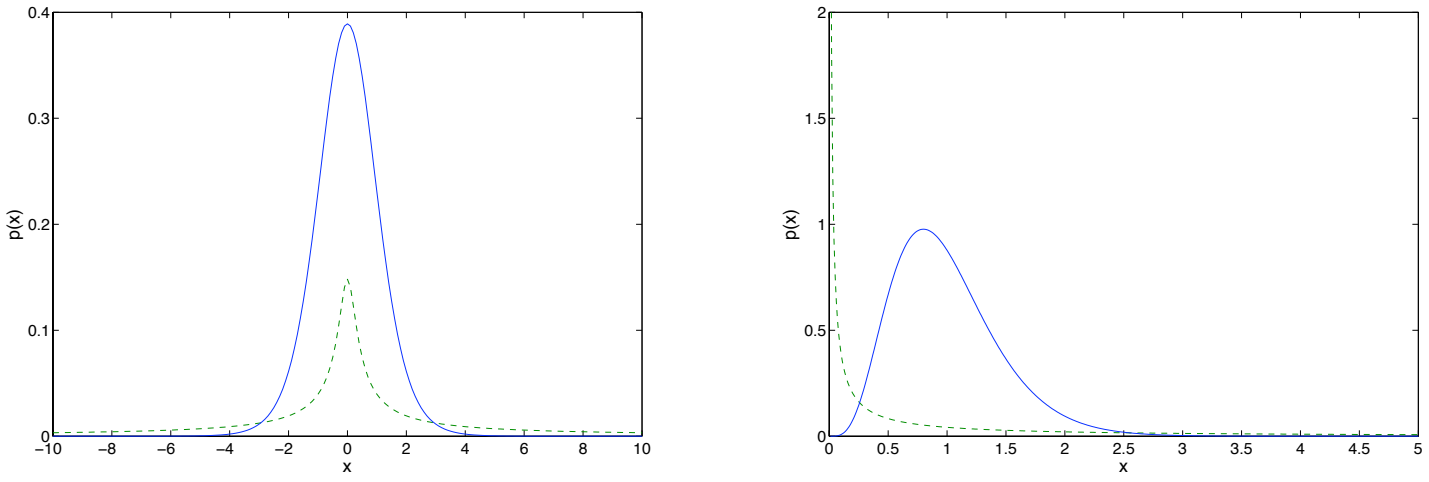
The Student-t density can be reformulated as an **infinite mixture of scaled Gaussians**:

$$\mathcal{S}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \int_0^\infty \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/u)\ \mathcal{G}(\tfrac{\nu}{2}, \tfrac{\nu}{2})\ du,$$

where u is a (latent) scale parameter.

[2] Student's t density was published in 1908 by *William S. Gosset*, while he worked at Guinness Brewery in Dublin and was not allowed to publish under his own name.

Multivariate Student-t distribution (continued)



(a) (b)

Figure: (a) Student-t distribution for two values of the shape parameter and the corresponding (b) Gamma distribution.

Some notations, definitions, etc. (continued)

The differential **entropy** is defined as

$$\mathrm{H}[p(\mathbf{x})] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \, d\mathbf{x}.$$

The entropy of a Gaussian is given by $\frac{D}{2} \ln 2\pi e + \frac{1}{2} \ln |\mathbf{\Sigma}|$.

If the continuous random variable Y has the same mean and covariance as the Gaussian random variable X, then $\mathrm{H}[p(\mathbf{y})] \leqslant \mathrm{H}[p(\mathbf{x})]$.

The **Kullback-Leibler divergence** measures the difference between two densities:

$$\mathrm{KL}[q \| p] = \int q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} \, d\mathbf{x} \geqslant 0.$$

The KL is asymmetric (thus not a distance) and only zero if $q(\mathbf{x}) = p(\mathbf{x})$ for all $\mathbf{x}$.

Some notations, definitions, etc. (continued)

Jensens's inequality
For a *convex* function $f(\cdot)$, we have $\langle f(\mathbf{x})\rangle \geqslant f(\langle \mathbf{x}\rangle)$.

Proof
We proof Jensen's inequality for $x \in \mathbb{R}$. Consider the Taylor expansion of $f(x)$ around $\bar{x} = \langle x\rangle$:

$$f(x) = f(\bar{x}) + (x - \bar{x})f'(\bar{x}) + \frac{1}{2}(x - \bar{x})^2 f''(\bar{x}) + \ldots$$

A function f is convex if $f'' \geq 0$ for all x. Hence,

$$f(x) \geqslant f(\bar{x}) + (x - \bar{x})f'(\bar{x})$$

in a small neighbourhood around $\bar{x}$, which is fixed. Taking the expectation leads to

$$\langle f(x)\rangle \geqslant f(\bar{x}) + \underbrace{(\langle x\rangle - \bar{x})}_{=0} f'(\bar{x}).$$

Matrix identities

Woodbury identity:

$$(\boldsymbol{\Psi} + \mathbf{V}\boldsymbol{\Phi}\mathbf{W})^{-1} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\mathbf{V}(\boldsymbol{\Phi}^{-1} + \mathbf{W}\boldsymbol{\Psi}^{-1}\mathbf{V})^{-1}\mathbf{W}\boldsymbol{\Psi}^{-1},$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{N\times N}$, $\boldsymbol{\Phi} \in \mathbb{R}^{M\times M}$, $\mathbf{V} \in \mathbb{R}^{N\times M}$ and $\mathbf{W} \in \mathbb{R}^{M\times N}$.

- When $\boldsymbol{\Psi}^{-1}$ is known and $N \gg M$, this speeds up the **matrix inversion**.
- For determinants we have $|\boldsymbol{\Psi} + \mathbf{V}\boldsymbol{\Phi}\mathbf{W}| = |\boldsymbol{\Psi}|\ |\boldsymbol{\Phi}|\ |\boldsymbol{\Phi}^{-1} + \mathbf{W}\boldsymbol{\Psi}^{-1}\mathbf{V}|$.

Cholesky decomposition:
When $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times D}$ is symmetric, positive definite, it can decomposed as follows:

$$\boldsymbol{\Lambda} = \mathbf{Q}^{\top}\mathbf{Q},$$

where the cholesky factor $\mathbf{Q} \in \mathbb{R}^{D\times D}$ is upper triangular.

- Solving the linear system $\mathbf{Q}\mathbf{x} = \mathbf{b}$ by backward substitution is $\mathcal{O}(N)$.
- Computing the inverse of $\boldsymbol{\Lambda}$ by backward-forward substitution is $\mathcal{O}(N^2)$.
- The determinant of $\boldsymbol{\Lambda}$ is given by $\prod_d Q_{dd}^2$.

References

- Christopher M. Bishop: Pattern Recognition and Machine Learning. Springer, 2006.
- Tutorial on Gaussian processes at NIPS 2006 by Carl E. Rasmussen.
- The Matrix Cookbook by Kaare B. Petersen and Michael S. Pedersen.