# Lecture Notes

Lecture notes of the course Introduction to Machine Learning and Spoken Language Processing.

## Contents

# 1 Introduction

Every step in conversational translation consists of obtaining data, using models and algorithms and then evaluation by analysis of metrics.

| Automatic speech recognition (ASR) | |
|---|---|
| Data: | Transcribed speech, signal processing |
| Models or algorithms: | Hidden Markov models (HMMs), neural networks, weighted automata |
| Metrics: | Word error rate, discriminative modelling |

| Speech correction | |
|---|---|
| Data: | Paired ARS transcripts and corrected text |
| Models or algorithms: | Weighted finite state transducers, conditional random fields |
| Metrics: | Word error rate |

| Text-to-text translation | |
|---|---|
| Data: | Parallel text |
| Models or algorithms: | Stochastic grammars and neural networks, weighted automata |
| Metrics: | Translations not unique, metrics reflect relative quality |

| Text-to-speech | |
|---|---|
| Data: | Transcribed speech |
| Models or algorithms: | Hidden Markov models (HMMs), neural networks |
| Metrics: | No metric, can we train with humans? |

Discriminative modelling means modelling of the conditional probability distribution instead of the joint probability distribution. It does not allow one to generate samples, but this is not required for tasks such as classification.

The final step is integration, which requires

(1) **low latency**,

(2) **robustness**,

(3) uncertainty, which means that hard decisions should be avoided troughout the system,

(4) user interfaces and

(5) usage of obtained data to improve performance.

Also, the data should be captured since this kind of data is hard to come by.

## 1.1 Speech and the vocal tract

Speech is

(1) non-stationary, which means that its joint probability distribution changes when the signal is shifted in time and roughly means that its properties are time-dependent, and

(2) contains a mix of pseudo-periodic and random components.

### 1.1.1 Anatomy

The main components of human speech production mechanisms are

(1) a variably-shaped acoustic tube and

(2) an excitation source.

Rough differences in speech are due to difference in the excitation of the source and detailed differences are due to changes in the shape of the acoustic tube.

The acoustic tube consists of pharynx combined with either the oral or nasal cavity. The velum determines which cavity is used. The lip, teeth and tongue are called the articulators and can shape the oral cavity to change the sound of speech. The nasal cavity cannot be changed and is related to person-specific sounds.

### 1.1.2 Excitation

The articulators continually move to reach a next target vocal tract state. Often this state is never reached and the current sound is then further modified by anticipation of the following sounds. Therefore the exact realisation of each individual sound is heavily dependent on previous and succeeding sounds. We call these effects **co-articulation** effects.

At the articulatory level, a voiced sound is one in which the vocal folds vibrate, and an unvoiced sound is one in which they do not. The vocal folds are also called vocal cords. The folds can either relax or vibrate to produce pufs of sound. The vibration consists of contraction to build up pressure which is then released. The pressure is determined by the contraction and the pressure of the lungs.

The acoustic tube has three sources of excitation:

(1) vibration of the cords which generates **voiced** sounds;

(2) turbulence caused by forcing air through narrow constrictions of the acoustic tube which generates **frictive sounds** and

(3) turbulence caused by the release of air following a complete closure of the acoustic tube which generates **plosive** sounds.

Sounds may also have a mixed excitation as for example in “zoo” (mix of narrow constriction of the acoustic tube and vibration of the cords).

### 1.1.3 Sounds of English

A vowel is a sound pronounced with an open vocal tract so that there is no build-up of air pressure at any point above the glottis, which is the part of the larynx consisting of the vocal cords. This contrasts with consonants which have a constriction or closure at some point along the vocal tract.

We model the pronunciation of a word as a string of symbols which represent phones or segments. A phone is a speech sound; phones are represented with phonetic symbols that bear some resemblance to a letter in an alphabetic language like English.

A phoneme is generally regarded as an abstraction of a set of phones which are perceived as equivalent to each other in a given language; i.e. if the exchange of one phone in a word for another gives a new word with a different meaning, then the two phones belong to different phonemes. Different speech sounds that are realizations of the same phoneme are known as allophones. The arpabet is a phonetic transcription code.

Consonant sounds may be divided into five broad classes depending on the type of vocal tract constriction: plosives (stops), fricatives, affricates, liquids (semi-vowels) and nasals.

A diphthong is a sound made by combining two vowels, specifically when it starts as one vowel sound and goes to another.

# 2 Intro machine learning

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Active learning is choosing which points are most useful to label in a classification task where few points are labelled.

Clustering is determining clusters of points where two points belonging to the same cluster are generally more similar to each other than two points belonging to different clusters. Clustering can also be seen as mixture modelling.

## 2.1 Gamma function

**Theorem 1** (Bohr-Mollerup). *The Gamma function, defined for $x > 0$ by*

$$\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt$$

*is the only function $f$ on $x > 0$ that satisfies*

*(1)* $f(1) = 1$,

*(2)* $f(x+1) = xf(x)$ *for* $x > 0$ *and*

*(3)* $\log \circ f$ *is convex.*

The Bohr-Mollerup theorem shows that the gamma function is the natural extension of the factorial. That is, (1) and (2) characterise the factorial and (3) characterises the growth of the factorials. Note that $\Gamma(n) = (n-1)!$.

## 2.2 Multinomial distribution

For $N$ independent trials each of which leads to a success for exactly one of $K$ categories, with each category having a given fixed success probability $p_i$, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories. The multinomial distribution function is given by

$$p(x_1, \ldots, x_K) = \frac{N!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K} = \frac{\Gamma(\sum_{i=1}^K x_i + 1)}{\prod_{i=1}^K \Gamma(x_i + 1)} \prod_{i=1}^K p_i^{x_i}$$

where $\sum_{i=1}^K x_i = N$. The expected value is given by $E[X_i] = Np_i$ and the covariance matrix by

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} np_i(1-p_i) & \text{if } i = j, \\ -np_i p_j & \text{if } i \neq j. \end{cases}$$

The case $N = 2$ reduces to the binomial distribution.

## 2.3 Poisson distribution

Consider a process in which events occur in time. Then assume the following:

(1) The number of events in disjoint time intervals are independent.

(2) The rate $\lambda$ at which events occur is contant over time: in a subinterval of length $u$ the expected number of telephone calls is $\lambda u$.

(3) The probability of two or more events in an interval of length $t$ tends to zero faster than $1/t$.

Then the distribution function of the number of events in an interval of length $t$ is given by

$$p(x) = \frac{(\lambda t)^x}{x!} \exp(-\lambda t).$$

The expected value is given by $E[X] = \lambda t$ and the variance by $\mathrm{Var}(X) = \lambda t$.

### 2.4 Gamma distribution

Since

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx$$

we have that

$$1 = \int_0^\infty \frac{x^{\alpha-1}}{\Gamma(\alpha)} \exp(-x) dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) dx$$

where $x > 0$ and $\beta > 0$. Now the gamma distribution is for $x > 0$ given by

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$

where $\alpha, \beta > 0$. The expected value is given by $E[X] = \alpha/\beta$ and the variance by $\mathrm{Var}(X) = \alpha/\beta^2$. You can prove that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Consider the situation process by the Poisson distribution and suppose that we want to wait until the $k$'th event occurs. Let $X$ be this waiting time. Then

$$F_X(x) = P(X \leq x) = 1 - P(X > x).$$

Here $P(X > x)$ can be interpreted as the statement that fewer than $k$ events occur in the interval $[0, x]$. So we calculate that

$$F_X(x) = 1 - \sum_{i=0}^{k-1} \frac{(\lambda x)^i}{i!} \exp(-\lambda x)$$

and therefore

$$\begin{aligned} f_X(x) &= \sum_{i=0}^{k-1} \frac{\lambda^{i+1} x^i}{i!} \exp(-\lambda x) - \sum_{i=1}^{k-1} \frac{\lambda^i x^{i-1}}{(i-1)!} \exp(-\lambda x) \\ &= \frac{\lambda^k}{\Gamma(k)} x^{k-1} \exp(-\lambda x) \end{aligned}$$

which is Gamma distributed with parameters $\alpha = k$ and $\beta = \lambda$. The case $\alpha = 1$ reduces to the exponential distribution. The Gamma distribution is thus the natural extension of the $\alpha$'th waiting time in a Poisson process with rate $\beta$.

### 2.5 Dirichlet distribution

Again consider the Poisson process. Let $W_i$ be $K$ independent Gamma-distributed variables with parameters $\alpha_i$ and $\beta$. Then the variables

$$X_i = \frac{W_i}{\sum_{i=1}^K W_i}$$

are Dirichlet-distributed with parameters $\alpha_i$. Thus, in a process consisting of subsequent waiting times with equal rates, the Dirichlet distribution describes the natural extension of the distribution of the fractions of those waiting times relative to the total time. It can be shown that the distribution function is given by

$$p(x_1, \ldots, x_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i - 1}$$

where $\sum_{i=1}^K x_i = 1$. If $\alpha_0 = \sum_{i=1}^K \alpha_i$, then the expected value is given by $E[X_i] = \alpha_i/\alpha_0$ and the covariance matrix by

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0+1)} & \text{if } i = j, \\ \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)} & \text{if } i \neq j. \end{cases}$$

The resemblance of this distribution function in comparison to the multinomial distribution function implies that it can be used to define a distribution over the parameters of a multinomial distribution. Indeed, in the case of $\alpha_1 - 1, \ldots, \alpha_K - 1$ successes in each category we have that

$$\begin{aligned} p(p_1, \ldots, p_K | \alpha_1 - 1, \ldots, \alpha_K - 1) &= p(p_1, \ldots, p_K) \frac{P(\alpha_1 - 1, \ldots, \alpha_K - 1 | p_1, \ldots, p_K)}{P(\alpha_1 - 1, \ldots, \alpha_K - 1)} \\ &= \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K p_i^{\alpha_i - 1} \end{aligned}$$

where we used the uniform prior $p(p_1, \ldots, p_K) = 1$. We also see that all $\alpha_i = 1$ yield the uniform distribution, which makes sense since all $\alpha_i = 1$ correspond to zero successes in each category.

Note that the distribution is also defined for $\alpha < 1$ and $\beta < 1$. This is a consequence of the natural extension of the Gamma distribution. Another useful interpretation of the parameters of the Dirichlet distribution is that of concentration parameters. If $\alpha_i \to 0$, then probability mass will be centered around $x_i = 0$ in a sparse manner.

The case $K = 2$ reduces to the Beta distribution which can be used to define a distribution over the parameter of the binomial distribution with $\alpha_1 - 1$ successes and $\alpha_2 - 1$ failures. Here we

denote $\alpha = \alpha_1$ and $\beta = \alpha_2$ such that

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

## 2.6 Conjugate priors

If the posterior distributions are in the same family as the prior probability distribution, the prior and posterior are then called conjugate distributions and the prior is called a conjugate prior for the likelihood function.

Let $\boldsymbol{P} \sim \mathrm{Dir}(\boldsymbol{\alpha})$. If the likelihood is distributed $\mathrm{Mult}(n, \boldsymbol{p})$, then

$$\begin{aligned} p(\boldsymbol{p}|\boldsymbol{x}) &\propto p(\boldsymbol{p})P(\boldsymbol{x}|\boldsymbol{p}) \\ &= \prod_i p_i^{\alpha_i - 1} \prod_i p_i^{x_i} \\ &= \prod_i p_i^{(\alpha_i + x_i) - 1} \end{aligned}$$

which shows that the posterior is distributed $\mathrm{Dir}(\boldsymbol{\alpha} + \boldsymbol{x})$. Therefore the Dirichlet distribution is the conjugate prior for the multinomial likelihood function.

## 2.7 Exponential family

The exponential familiy of distribution is given by

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x})g(\boldsymbol{\theta}) \exp\{\boldsymbol{\phi}(\boldsymbol{\theta})^T \boldsymbol{u}(\boldsymbol{x})\}.$$

First of all, the factorisation theorem tells us that $\boldsymbol{u}(\boldsymbol{x})$ is a sufficient statistic, meaning that the interaction between and $\boldsymbol{\theta}$ and the density is captured by $\boldsymbol{u}(\boldsymbol{x})$. Second, $\boldsymbol{\phi}(\boldsymbol{\phi})$ is called the vector of natural parameters whose space can be shown to be convex. Third, since $g(\boldsymbol{\theta})$ simply scales the density, it can be seen as a normalisation factor. It turns out that $-\log g(\boldsymbol{\theta})$ can be used to say something about the moments of the sufficient statistic.

The exponential family is used for the following reasons:

(1) A number of important and useful calculations in statistics can be done all at one stroke within the framework of the exponential family.

(2) Exponential families have sufficient statistics that can summarize arbitrary amounts of independent identically distributed data using a fixed number of values.

(3) Exponential families have conjugate priors.

(4) The posterior predictive distribution of an exponential-family random variable with a conjugate prior can always be written in closed form.

## 2.8 Basic limit theorems

Let $X_1, \dots, X_n$ be an independent trials process with finite expected value $\mu$ and finite variance $\sigma^2$. Let $S_n = \sum_{i=1}^n X_i$.

**Theorem 2** (Law of Large Numbers). *For any $\varepsilon > 0$,*

$$\lim_{n\to\infty} P(|S_n/n - \mu| < \varepsilon) = 1.$$

*Proof.* Let $\varepsilon > 0$. Chebychev's inequality states that

$$P(|S_n/n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(S_n/n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Therefore

$$P(|S_n/n - \mu| < \varepsilon) = 1 - P(|S_n/n - \mu| \geq \varepsilon) \leq 1 - \frac{\sigma^2}{n\varepsilon^2} \to 1$$

as $n \to \infty$. □

So for any nonzero margin specified, with a large sample size there will be a very high probability that the average of the observations will be close to the expected value.

**Theorem 3** (Central Limit Theorem).

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$$

*as $n \to \infty$.*

*Proof.* Without loss of generality assume that $\mu = 0$ and $\sigma = 1$. Then

$$M_{S_n^*}(t) = E[\exp(t/\sqrt{n} \cdot S_n)] = M_{S_n}(t/\sqrt{n}) = [M_X(t/\sqrt{n})]^n.$$

Taylor expansion of $M_X(t)$ around $t = 0$ yields

$$\begin{aligned} M_X(t) &= M_X(0) + \frac{M_X^{(1)}(0)}{1!}t + \frac{M_X^{(2)}(0)}{2!}t^2 + \frac{M_X^{(3)}(c)}{3!}t^3 \\ &= 1 + \frac{t^2}{2} + Ct^3 \end{aligned}$$

for some $c \in (0, t)$. Therefore

$$[M_X(t/\sqrt{n})]^n = \left(1 + \frac{t^2/2}{n} + \frac{Ct^3}{n\sqrt{n}}\right)^n.$$

We now calculate directly

$$\begin{aligned}\lim_{n\to\infty}[M_X(t/\sqrt{n})]^n &= \exp\left\{\lim_{n\to\infty}\frac{\log\left(1+\frac{t^2/2}{n}+\frac{Ct^3}{n\sqrt{n}}\right)}{1/n}\right\}\\ &= \exp\left\{\lim_{n\to\infty}\frac{\frac{t^2}{2}+\frac{C't^3}{\sqrt{n}}}{1+\frac{t^2/2}{n}+\frac{Ct^3}{n\sqrt{n}}}\right\}\\ &= \exp(t^2/2)\end{aligned}$$

which is the moment generating function of the $N(0,1)$ distribution. □

# 3 Regression

It is important to always question the assumptions of your model.

## 3.1 Solving for maximum likelihood parameters

Since both are scalars,

$$\begin{aligned}\sum_n\left(\boldsymbol{\beta}^T\tilde{\boldsymbol{x}}^{(n)}\right)^2 &= \sum_n\left(\boldsymbol{\beta}^T\tilde{\boldsymbol{x}}^{(n)}\right)\left(\tilde{\boldsymbol{x}}^{(n)T}\boldsymbol{\beta}\right)\\ &= \sum_n\boldsymbol{\beta}^T\left(\tilde{\boldsymbol{x}}^{(n)}\tilde{\boldsymbol{x}}^{(n)T}\right)\boldsymbol{\beta}\\ &= \boldsymbol{\beta}^T\left(\sum_n\tilde{\boldsymbol{x}}^{(n)}\tilde{\boldsymbol{x}}^{(n)T}\right)\boldsymbol{\beta}.\end{aligned}$$

## 3.2 Vector and matrix calculus

### 3.2.1 Vector derivatives

**Definition 1.** *The derivative of $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ is given by the column vector*

$$\left[\frac{\partial f(\boldsymbol{x})}{\partial\boldsymbol{x}}\right]_i = \frac{\partial f(\boldsymbol{x})}{\partial x_i}.$$

**Theorem 4.**

$$\frac{\partial\boldsymbol{a}^T\boldsymbol{x}}{\partial\boldsymbol{x}} = \boldsymbol{a}.$$

*Proof.*

$$\left[\frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial \boldsymbol{x}}\right]_j = \frac{\partial \sum_i a_i x_i}{\partial x_j} = a_j.$$

□

**Theorem 5.**

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x}.$$

*Proof.*

$$\left[\frac{\partial \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\partial \boldsymbol{x}}\right]_k = \frac{\partial \sum_{i,j} a_{ij} x_i x_j}{\partial x_k} = \sum_i a_{ki} x_i + \sum_j a_{jk} x_j.$$

□

**Theorem 6.**

$$\frac{\partial \boldsymbol{f}(\boldsymbol{x})^T \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{f}'(\boldsymbol{x})^T \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{f}(\boldsymbol{x})^T \boldsymbol{g}'(\boldsymbol{x}).$$

*Proof.*

$$\left[\frac{\partial \boldsymbol{f}(\boldsymbol{x})^T \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right]_j = \frac{\partial \sum_i f_i(\boldsymbol{x}) g_i(\boldsymbol{x})}{\partial x_j} = \sum_i f_i'(\boldsymbol{x}) g_i(\boldsymbol{x}) + \sum_i f_i(\boldsymbol{x}) g_i'(\boldsymbol{x}).$$

□

### 3.2.2 Matrix derivatives

**Definition 2.** *The derivative of $f(\boldsymbol{X})$ with respect to $\boldsymbol{X}$ is given by the matrix*

$$\left[\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}}\right]_{ij} = \frac{\partial f(\boldsymbol{x})}{\partial x_{ij}}.$$

**Theorem 7.** *If $\boldsymbol{X}$ is invertible, then*

$$\frac{\partial |\boldsymbol{X}|}{\partial \boldsymbol{X}} = |\boldsymbol{X}| \boldsymbol{X}^{-T}.$$

*Proof.*

$$\left[\frac{\partial |\boldsymbol{X}|}{\partial \boldsymbol{X}}\right]_{ij} = \frac{\partial \sum_k x_{ik} C_{ik}}{\partial x_{ij}}$$

where $C_{ik}$ denote the cofactors of $\boldsymbol{X}$. Note that all cofactors are independent of $x_{ik}$. So

$$\left[\frac{\partial|\boldsymbol{X}|}{\partial \boldsymbol{X}}\right]_{ij} = C_{ij}.$$

which means that by Cramer's rule

$$\frac{\partial|\boldsymbol{X}|}{\partial \boldsymbol{X}} = \operatorname{cof} \boldsymbol{X} = (\operatorname{adj} \boldsymbol{X})^T = (|\boldsymbol{X}|\boldsymbol{X}^{-1})^T = |\boldsymbol{X}|\boldsymbol{X}^{-T}.$$

□

**Theorem 8.** *If $\boldsymbol{X}$ is invertible, then*

$$\frac{\partial \ln|\boldsymbol{X}|}{\partial \boldsymbol{X}} = \boldsymbol{X}^{-T}.$$

*Proof.* By the trivial extension of the chain rule,

$$\frac{\partial \ln|\boldsymbol{X}|}{\partial \boldsymbol{X}} = \frac{1}{|\boldsymbol{X}|}\frac{\partial|\boldsymbol{X}|}{\partial \boldsymbol{X}} = \boldsymbol{X}^{-T}.$$

□

**Theorem 9.**

$$\frac{\partial \operatorname{tr}(\boldsymbol{AXB})}{\partial \boldsymbol{X}} = \boldsymbol{A}^T\boldsymbol{B}^T.$$

*Proof.* First note that

$$[\boldsymbol{AXB}]_{ij} = \sum_k [\boldsymbol{AX}]_{ik} b_{kj} = \sum_{k,l} a_{il}x_{lk}b_{kj}.$$

Therefore

$$\left[\frac{\partial \operatorname{tr}(\boldsymbol{AXB})}{\partial \boldsymbol{X}}\right]_{ij} = \frac{\partial \sum_{k,l,m} a_{ml}x_{lk}b_{km}}{\partial x_{ij}} = \sum_m a_{mi}b_{jm} = \sum_m [\boldsymbol{A}^T]_{im}[\boldsymbol{B}^T]_{mj}.$$

□

Note that Theorem 9 shows that

$$\operatorname{tr}(\boldsymbol{ABC}) = \sum_{k,l,m} a_{ml}b_{lk}c_{km}$$

which means that

$$\operatorname{tr}(\boldsymbol{ABC}) = \operatorname{tr}(\boldsymbol{CAB}) = \operatorname{tr}(\boldsymbol{BCA}).$$

**Theorem 10.**

$$\frac{\partial \operatorname{tr}(\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X})}{\partial \boldsymbol{X}} = \boldsymbol{A}^T \boldsymbol{X} + \boldsymbol{A} \boldsymbol{X}.$$

*Proof.* First note that

$$\left[\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}\right]_{ij} = \sum_k \left[\boldsymbol{X}^T \boldsymbol{A}\right]_{ik} x_{kj} = \sum_{k,l} [\boldsymbol{X}^T]_{il} a_{lk} x_{kj} = \sum_{k,l} x_{li} a_{lk} x_{kj}.$$

Therefore

$$\begin{aligned}\left[\frac{\partial \operatorname{tr}(\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X})}{\partial \boldsymbol{X}}\right]_{ij} &= \frac{\partial \sum_{k,l,m} x_{lm} a_{lk} x_{km}}{\partial x_{ij}} \\ &= \sum_{k,l,m} \frac{\partial x_{lm}}{\partial x_{ij}} a_{lk} x_{km} + \sum_{k,l,m} x_{lm} a_{lk} \frac{\partial x_{lm}}{\partial x_{km}} \\ &= \sum_k a_{ik} x_{kj} + \sum_k a_{ki} x_{kj}.\end{aligned}$$

□

### 3.2.3 Derivative of a vector with respect to vector

Remember that the hessian of $f$ is defined by

$$[\boldsymbol{H}]_{ij} = \frac{\partial^2 f}{\partial x_i x_j} = \frac{\partial}{\partial x_i} \frac{\partial f}{\partial x_j}.$$

This motivates the following definition.

**Definition 3.** *The second order derivative of $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ is given by the matrix*

$$\frac{\partial^2 f}{\partial \boldsymbol{x}^2} = \frac{\partial^2 f}{\partial x_i x_j}.$$

**Definition 4.** *The derivative of $\boldsymbol{f}(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ is given by the matrix*

$$\left[\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right]_{ij} = \frac{\partial f_j(\boldsymbol{x})}{\partial x_i}.$$

Thus we see that

$$\frac{\partial}{\partial \boldsymbol{x}} \frac{\partial f}{\partial \boldsymbol{x}} = \frac{\partial^2 f}{\partial \boldsymbol{x}^2}.$$

**Theorem 11.**

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^T.$$

*Proof.*

$$\left[\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}}\right]_{ij} = \frac{\partial \sum_k A_{jk}x_k}{\partial x_i} = A_{ji} = [\boldsymbol{A}^T]_{ij}.$$

□

### 3.3 Basis function regression

We can choose any basis function as long as the matrix inversion goes well. The Weierstrass approximation theorem can be use to approximate any continuous function by a polynomial function, which consists of basis functions.

### 3.4 Lagrange multipliers

**Theorem 12.** *Let $U \subseteq \mathbb{R}^n$ be open, $f : U \to \mathbb{R}$ in $C^1$ and $\boldsymbol{G} = (G_1, \dots, G_m) : U \to \mathbb{R}^m$ a differentiable transformation where $m < n$. Let $S = \{x \in U : \boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{0}\}$. If $f$ has a local extremum on $S$ in $\boldsymbol{a} \in S$ and $\boldsymbol{G}'(\boldsymbol{a})$ has full rank, then there exist $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ such that*

$$\nabla f(\boldsymbol{a}) = \lambda_1 \nabla G_1(\boldsymbol{a}) + \cdots + \lambda_m \nabla G_m(\boldsymbol{a}).$$

*Proof.* We provide a sketch of the proof. Let $\boldsymbol{r}(t)$ be any parametrised curve in $S$ such that $\boldsymbol{r}(0) = \boldsymbol{a}$ and let $g(t) = f(\boldsymbol{r}(t))$. Then

$$0 = g'(t)\big|_{t=0} = \nabla f(\boldsymbol{r}(0)) \cdot \boldsymbol{r}'(0) = \nabla f(\boldsymbol{a}) \cdot \boldsymbol{r}'(0)$$

which means that $\nabla f(\boldsymbol{a})$ is perpendicular to $\boldsymbol{r}'(0)$. So every curve in $S$ through $\boldsymbol{a}$ is perpendicular to $\nabla f(\boldsymbol{a})$, which implies that $\nabla f(\boldsymbol{a})$ is perpendicular to $S$. Now row $\boldsymbol{G}'(\boldsymbol{a})$ spans all vectors perpendicular to $S$, and the invertibility of $\boldsymbol{G}'(\boldsymbol{a})$ implies that we can find $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ such that $\boldsymbol{G}'(\boldsymbol{a})^T \boldsymbol{\lambda} = \nabla f(\boldsymbol{a})$. □

## 4 Classification

The hard work consists of finding the likelihood function. The rest then follows: choose a prior and one can do Bayesian inference or perform optimisation to find the maximum likelihood estimate.

### 4.1 Non-Gaussian noise

Laplacian noise is used to model outliers with heavier tails, or equivalently with higher kurtosis. To interpret kurtosis, let

$$Z = \frac{X - \mu}{\sigma}.$$

Then

$$\kappa = \frac{E[(Z-\mu)^4]}{E[(Z-\mu)^2]^2} = E[Z^4] = \mathrm{Var}(Z^2) + [E(Z^2)]^2 = \mathrm{Var}(Z^2) + 1.$$

The kurtosis can now be seen to be a measure of the dispersion of $Z^2$ around its expectation. This is Moors' interpretation.

The problem with Laplacian noise is that is cannot easily be solved for analytically. However the problem is convex, so any iterative algorithm can be used. The problem is related to the lasso problem, which minimises a weighted average of a sum of squares and a sum of absolute values.

### 4.2 Maximum a posteriori

The point of the Bayesian methods is that the posterior is calculated which then captures the uncertainty of the parameters. Maximum a posteriori is a point estimate and therefore just an optimisation instead of a Bayesian method.

Maximum a posteriori can be seen as a form of regularisation, since the prior corresponds with the regulariser. That is, a Gaussian prior corresponds with quadratic loss and a Laplacian prior is absolute loss. Ridge regression has quadratic loss and thus corresponds with a Gaussian prior. However, certain loss functions do not correspond to a prior probability distribution. This is because all distributions must integrate to one. An important difference between maximum a posteriori and regularisation is that in the case of a change of variables, the Jacobian comes into play which can change the maximum a posteriori estimate, but the estimate of regularisation always stays the same.

### 4.3 Softmax

Suppose that we want to test $\mathcal{H}_i$ upon observing $\boldsymbol{x}$ where $\{\mathcal{H}_1, \ldots, \mathcal{H}_n\}$ is a set of mutually exclusive and exhaustive hypotheses. Then

$$P(\mathcal{H}_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\mathcal{H}_i)P(\mathcal{H}_i)}{\sum_{j=1}^n P(\boldsymbol{x}|\mathcal{H}_j)P(\mathcal{H}_j)} = \frac{\exp(\ln \mathcal{L}_i + C_i)}{\sum_{j=1}^n \exp(\ln \mathcal{L}_j + C_j)}.$$

In the case that all distributions are from the same class belonging to the exponential family $\ln \mathcal{L}_i$ is linear in the natural parameters so that can write $\ln \mathcal{L}_i + C_i = \boldsymbol{\beta}_i^T \boldsymbol{\phi}(\boldsymbol{x})$ for some vector $\boldsymbol{\beta}_i$ and function $\boldsymbol{\phi}$. This motivates the softmax function used for classification:

$$P(\mathcal{H}_i|\boldsymbol{x}) = \frac{\exp[\boldsymbol{\beta}_i^T \boldsymbol{\phi}(\boldsymbol{x})]}{\sum_{j=1}^n \exp[\boldsymbol{\beta}_j^T \boldsymbol{\phi}(\boldsymbol{x})]}.$$

The two-dimensional case yields that

$$P(\mathcal{H}_1|\boldsymbol{x}) = \frac{\exp[\boldsymbol{\beta}_1^T \boldsymbol{\phi}(\boldsymbol{x})]}{\exp[\boldsymbol{\beta}_1^T \boldsymbol{\phi}(\boldsymbol{x})] + \exp[\boldsymbol{\beta}_2^T \boldsymbol{\phi}(\boldsymbol{x})]} = \frac{1}{1 + \exp[-\boldsymbol{\beta}^T \boldsymbol{\phi}(\boldsymbol{x})]}$$

where $\boldsymbol{\beta} = \boldsymbol{\beta_1} - \boldsymbol{\beta_2}$. This shows that the softmax function is overparametrised and that there are actually $n-1$ free parameters. Also, the contours of $P(\mathcal{H}_1|\boldsymbol{x})$ are given by all $\boldsymbol{x}$ such that $\boldsymbol{\beta}^T \boldsymbol{\phi}(\boldsymbol{x})$ is constant. This shows how the basis functions relate to the shape of the contours.

The sigmoid function is related to the two-dimensional case of the softmax function and is given by

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

Note that $\sigma(z) + \sigma(-z) = 1$ and that $\sigma'(z) = \sigma(-z)$.

### 4.4 Logistic classification

The learning rule in logistic classification can be interpreted in a nice way. The rule is given by

$$\beta^{(t+1)} = \beta^{(t)} + \eta \sum_n (y^{(n)} - \sigma(\boldsymbol{\beta}^T \tilde{\boldsymbol{x}}^{(n)}))\tilde{\boldsymbol{x}}^{(n)}.$$

First note that $\sigma(\boldsymbol{\beta}^T \tilde{\boldsymbol{x}}^{(n)}) = P(y^{(n)} = 1|\tilde{\boldsymbol{x}}, \boldsymbol{\beta})$. So if $y^{(n)} = 1$, then $P(.)$ should be high. Thus $y^{(n)} - \sigma(.)$ is a measure of error. If the error is non-zero, then the plane should turn towards or away from $\boldsymbol{x}^{(n)}$ to improve the likelihood.

The learning rate has a different effect for batch and online learning and should therefore be scaled. Also, to improve convergence, the learning rate should grow like $1/t$. Online learning should be done randomly.

In the case that the data is perfectly linearly separable, $||\boldsymbol{\beta}|| \to \infty$ should yield infinite likelihood. We are then absolutely certain in our classification, which does not make sense. By assigning a prior on $\boldsymbol{\beta}$ we can penalise the length of $\boldsymbol{\beta}$ to avoid this problem.

# 5 Spectral analysis

A formant is a concentration of acoustic energy around a particular frequency in the speech wave. Formants are due to resonances in the vocal tract and are key in distinguishing phones from each other.

A stationary signal is one whose properties do not vary with time. For example, a sine wave of constant amplitude and frequency is stationair. More specifically, a stochastic process is stationary if the joint distribution is invariant in time, which has as a consequence that the mean and autocorrelation do not depend on time.

Speech is assumed to be stationary over an interval of 10-20 ms. Typical sampling frequencies are given by the following table.

| Quality | Sampling frequency | Bandwidth | Resolution |
|---|---|---|---|
| High | 16 kHz | 8 kHz | 16 bits uniformly/non-uniformly quantised |
| Low | 8 kHz | 4 kHz | 8 bits non-uniformly quantised |

Effects of quantisation are accounted for as noise. If a signal is non-periodic, then $f_0 \to 0$ which implies that the spectrum becomes continuous.

Vowels are mostly below 2 kHz. Friccatives are bursts, mainly consisting of high frequencies.

## 5.1 Fourier series

Every periodic signal of frequency $f_0$ can be represented as a Fourier series, i.e.

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{1}{T} F\left(j\frac{2\pi n}{T}\right) \exp\left(j\frac{2\pi n}{T}t\right)$$

where

$$F(j\omega) = \int_{-\infty}^{\infty} f(t) \exp(-j\omega t) dt.$$

# 6 Spectral analysis (cont)

A periodic signal is obtained by convolution with an impulse train, which means that it has a line spectrum where the lines are spaced by the fundamental frequency. The lowest frequency is called the fundamental frequency and the others harmonics.

The complex Fourier series is obtained by simply substituting Euler's formula for $\cos(.)$ and $\sin(.)$.

The terminology for short and large windows is given in the following table.

| Window length | Time resolution | Frequency resolution |
|---|---|---|
| Short | High | Low, can only observe *wide-band* patterns |
| Long | Low | High, can observe *narrow-band* patterns |

### 6.1 Phones

Vowel sounds are characterised by the first three formants $F_1$, $F_2$ and $F_3$. There is a simple relationship between the tongue and jaw positions and $F_1$ and $F_2$:

| | Tongue front | Tongue back |
|---|---|---|
| High jaw | Low $F_1$, high $F_2$ | Low $F_1$, low $F_2$ |
| Low jaw | High $F_1$, high $F_2$ | High $F_1$, low $F_2$ |

The first two formants are mainly responsible for the vowel quality.

Liquids are also characterised by formant positions, but the dynamics are more important and the overall energy is lower.

Nasals have a strong $F_1 \approx 250$ Hz and weak higher formants. There is ofter energy at $2.5$ kHz.

Fricatives have most energy in higher frequencies and show weak formant structure.

Stops are characterised by silence optionally followed by a burst of high energy.

## 7 Clustering

$K$-means tends to converge to a local minimum. We can prove its convergence by making use of the monotonic convergence theorem. Obviously, $C \geq 0$, so it is bounded from below. Furthermore, we show that every step decreases $C$. Step one is trivial. We can show step two as

follows:

$$\begin{aligned}
\frac{\partial C}{\partial \boldsymbol{m}_i} &= \frac{\partial}{\partial \boldsymbol{m}_i} \sum_{n=1}^{N} \sum_{k=1}^{K} s_{nk} ||\boldsymbol{x}_n - \boldsymbol{m}_k||^2 \\
&= \sum_{n=1}^{N} s_{ni} \frac{\partial}{\partial \boldsymbol{m}_i} ||\boldsymbol{x}_n - \boldsymbol{m}_i||^2 \\
&= \sum_{n=1}^{N} s_{ni} \frac{\partial}{\partial \boldsymbol{m}_i} (\boldsymbol{x}_n^T \boldsymbol{x}_n - 2\boldsymbol{x}_n^T \boldsymbol{m}_i + \boldsymbol{m}_i^T \boldsymbol{m}_i) \\
&= \sum_{n=1}^{N} s_{ni} (2\boldsymbol{m}_i - 2\boldsymbol{x}_n) \\
&= 0 \\
\implies \boldsymbol{m}_i &= \left( \sum_{n=1}^{N} s_{ni} \right)^{-1} \sum_{n=1}^{N} s_{ni} \boldsymbol{x}_n.
\end{aligned}$$

Taking the second derivative show that we are indeed dealing with a minimum.

$K$-means does not tell us how many clusters to choose or how to initialise. We can even choose another distance metric or mixing weights. The two main assumptions are the identity covariance matrix and equal mixing weights. Equal weights can cause $K$-means to place multiple centers in an area if that area contains a lot more points than other areas.

### 7.1 Analogies

Least squares is modelling with additive Gaussian noise. $K$-means is assuming a mixture of Gaussians.

The parameters are distribution-related, e.g. the mean of a Gaussian, and hidden variables are data-related, e.g. which point originates from which cluster. The following table shows how different algorithms relate.

| | | Parameters | |
|---|---|---|---|
| | | Maximisation | Expectation |
| Hidden variables | Maximisation | $K$-means | ME algorithm |
| | Expectation | EM algorithm | Bayesian methods |

## 8 Dimensionality reduction

If data is projected onto a higher-dimensional space, then calculation will be done with needlessly many variables, since the data is also fully explained in the lower-dimensional space. Dimen-

sionality reduction is used to project data onto a lower-dimensional space such that the relevant information is preserved. Equivalently, it finds the linear projection which

(1) maximises the variance,

(2) minimises the reconstruction error,

(3) has the highest mutual information under a Gaussian model and

(4) is equivalent to the maximum likelihood estimate under a linear Gaussian factor model.

## 8.1 Maximum variance

Let $\boldsymbol{x}$ be nonzero data. We can summarise the variance by

$$V = E\left[||\boldsymbol{X} - E[\boldsymbol{X}]||^2\right] = E\left[||\boldsymbol{X}||^2\right] = \sum_i E[X_i^2].$$

We associate the variance in direction $\boldsymbol{e}_i$ with $E[X_i^2]$. We thus see that $V$ is the sum of the variances in all directions. We now want to rotate $\boldsymbol{x}$ such that the biggest variance is contained in direction $\boldsymbol{e}_1$ and the then-biggest variances in the following directions. If this rotation is given by $\boldsymbol{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)^T$, then the variance in direction $\boldsymbol{e}_1$ is given by

$$\operatorname{Var}(\boldsymbol{e}_i \cdot \boldsymbol{V}\boldsymbol{X}) = \operatorname{Var}(\boldsymbol{v}_1^T \boldsymbol{X}).$$

Thus we find that

$$\boldsymbol{v}_1 = \operatorname*{argmax}_{\boldsymbol{w}} \operatorname{Var}(\boldsymbol{w}^T \boldsymbol{X}).$$

A simple exercise shows that $\boldsymbol{v}_1, \dots, \boldsymbol{n}_n$ are the eigenvectors of $\operatorname{Cov}(\boldsymbol{X})$ with descending eigenvalues. Note that the rotated data has a diagonal covariance matrix and that the variances are given by the eigenvalues:

$$\operatorname{Cov}(\boldsymbol{V}\boldsymbol{X}) = \boldsymbol{V} \operatorname{Cov}(\boldsymbol{X}) \boldsymbol{V}^T = \boldsymbol{\Lambda}.$$

## 8.2 Reconstruction error

Suppose that the result of PCA is projection onto a plane $P$. Now the tangents of $P$ are the highest directions of variance. So movement of the data into those directions would amplify the error the most. Since projection onto $P$ yields zero movement along the tangents of $P$, the result of PCA is equivalent to minimal (reconstruction) error.

### 8.3 Mutual information

Let $X$ be a random variable. Its entropy $H(X)$ is a measure of information. Now suppose that we have another random variable $Y$. Since $X, Y$ is also just another random variable, its entropy is given by $H(X, Y)$. We call $H(X, Y)$ the joint entropy of $X$ and $Y$.

We now investigate how much learning $Y$ tells us about $X$. Each possible event $X|Y = y_i$ occurs with probability $P(Y = y_i)$. Therefore, by the consistence of entropy, we must have

$$H(X|Y) = \sum_i P(Y = y_i) H(X|Y = y_i) = E[H(X|Y = y)],$$

which is by the axioms of entropy the one and only valid measure of this quantity. Conditional entropy is related to the other quantities as follows:

$$\begin{aligned} H(X|Y) &= -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \\ &= -\sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(y)} \\ &= -\sum_{x,y} p(x, y) \log p(x, y) + \sum_{x,y} p(x, y) \log p(y) \\ &= H(X, Y) - H(Y). \end{aligned}$$

We can now state Bayes' rule for entropy:

$$H(X|Y) = H(X) + H(Y|X) - H(Y).$$

Mutual information is a quantity that measures a relationship between two random variables that are sampled simultaneously. In particular, it measures how much information is communicated, on average, in one random variable about another. Equivalently, the mutual information $I(X; Y)$ measures how much the realization of random variable $Y$ tells us about the realization of $X$, i.e. how by how much the entropy of $X$ is reduced if we know the realization of $Y$:

$$I(X; Y) = H(X) - H(X|Y).$$

Bayes' rule for entropy shows that the mutual information is symmetric; i.e. $I(X; Y) = I(Y; X)$.

### 8.4 Kernel trick

Many methods are based on planes in $\mathbb{R}^N$. These methods can be generalised to non-linear methods by making use of a non-linear transformation $\boldsymbol{\phi} : \mathbb{R}^N \to \mathbb{R}^M$. Planes in $\mathbb{R}^M$ correspond then to non-linear surfaces in $\mathbb{R}^N$. These surfaces are given by $\{\boldsymbol{x} : \boldsymbol{n} \cdot \boldsymbol{\phi}(\boldsymbol{x}) = c\}$. This shows what effect the kernel trick has.

# 9 Linear prediction analysis

Modelling the vocal tract as an lumped tube model where we assume the glottis to have infinite impedance and have unity reflection coefficient yields that the transfer function is an all-pole model. This model is only valid for vowel and approximant sounds. The all-pole model implies that

$$y[n] = \sum_{k=1}^{P} a_k y[n-k] + x[n] \iff x[n] = y[n] - \sum_{k=1}^{P} a_k y[n-k].$$

Since the glottis releases pressure spikes periodically, we can assume that during the quasi-stationary period we have that $x[n] = \sqrt{E}\delta[n_0]$, or equivalently that the input is spectrally flat with power $E$. Therefore the coefficients can be approximated by solving

$$0 = y[n] - \sum_{k=1}^{P} y[n-k] = y[n] - \hat{y}[n].$$

which means that $\hat{y}[n]$ can be seen as an estimator of $y[n]$, hence the name linear predictive. When we have estimated the coefficients, we can recontruct the input signal by

$$x[n] = y[n] - \hat{y}[n].$$

In this context, the input signal is also called the error. Note that the vocal tract gives rise to formants, and we wish to see its transfer function.

## 9.1 DFT method

Remember that the coefficients are such that

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}.$$

Hence we see that

$$\mathcal{Z}^{-1}\{H^{-1}(z)\} = \mathcal{Z}^{-1}\left\{1 - \sum_{i=1}^{p} a_i z^{-i}\right\} = [1, -a_1, \ldots, -a_p, 0, \ldots].$$

Thus we can calculate $H^{-1}(j\Omega)$ by calculating the DFT of $[1, -a_1, \ldots, -a_p, 0, \ldots]$ and then invert to obtain $H(j\Omega)$.

# 10 Sequence modelling

Arrows in graphics models can be interpreted as "influences" such that the probability of the states are independent given the states which point to them. Greyed out states are observable, white states unobservable, square states discrete and round states continuous. These last four properties are often neglected.

$n$-th order Markov models corresponds to $n+1$-gram models, since unigram models do not depends on their past at al.

If each state is reachable from any other state, then the stationary distribution exists and is given by the eigenvector with eigenvalue $\lambda = 1$. Given an observed sequence, a maximum-likelihood estimate of $T$ can be made by simply counting. This however, is not realistic since zero probability can occur and thus a prior should be used.

The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term. A Gaussian-Markov model is often called an AR-1 model. Note that this is inaccurate true since an AR-1 model can have different emission densities.

In the case of Newtonian mechanics, the Markov assumption makes sense; if we know the speed, position and acceleration of each element at a time $t$, we can predict their states at time $t+1$.

The marginal of a multivariate Gaussian is obtained by simply picking the elements from the mean and covariance. This can be proved by the observation that the marginal is a linear transformation which yields exactly the wanted result.

Noise is assumed to be independent of the state. The expected value is assumed to be over the appropriate joint distribution. Since this distribution is always Gaussian, this reduces to the marginal (joint) distribution of the relevant variables.

## 10.1 Reversibility

We note that

$$\begin{aligned} p(y_{1:T}) &= p(y_1)p(y_2|y_1)\cdots p(y_T|y_{T-1}) \\ &= p(y_1)\frac{p(y_2,y_1)}{p(y_1)}\cdots\frac{p(y_T,y_{T-1})}{p(y_{T-1})} \\ &= \frac{p(y_2,y_1)}{p(y_2)}\cdots\frac{p(y_T,y_{T-1})}{p(y_T)}p(y_T) \\ &= p(y_1|y_2)\cdots p(y_{T-1}|y_T)p(y_T). \end{aligned}$$

In general, the conditionals are different. If they do equal, then the chain is called a reversable Markov chain. For AR-1 this is generally the case, for $N$-gram models not.

### 10.2 First order Markov

These models are characterised by

$$p(y_t|y_{t-1}) = \mathcal{G}(y_t; \Lambda y_{t-1}, \Sigma) \iff y_t = \Lambda y_{t-1} + \Sigma^{1/2}\varepsilon,\ \varepsilon \sim \mathcal{G}(0, I).$$

Assume that $\Lambda = \lambda$. We can calculate the stationary distribution by the following observations:

(1) since everything is linear and the initial state is Gaussian, the stationary distribution is Gaussian,

(2) $\lambda < 1$ can be interpreted as the stability condition such that the stationary distribution exists.

We then calculate

$$E[y_t] = \lambda E[y_t] + \sigma E[\varepsilon] \iff (1-\lambda)E[y_t] = 0 \implies E[y_t] = \mu_\infty = 0.$$

Similarly,

$$E[y_t^2] = \lambda^2 E[y_t^2] + \sigma^2 E[\varepsilon^2] \implies \sigma_\infty^2 = \frac{\sigma^2}{1-\lambda^2}.$$

### 10.3 Continuous observed state

In the case of discrete state, these models are characterised by

$$p(y_t|x_t = k) = \mathcal{G}(y_t; \mu_k, \Sigma_k).$$

We see that at any time

$$p(y_t) = \sum_k p(y_t|x_t = k)p(x_t = k) = \sum_k \pi_k^t \mathcal{G}(y_t; \mu_k, \Sigma_k).$$

Therefore this model is a mixture of Gaussian models with dynamic cluster assignments. The stationary distribution is determined by the stationary distribution of the Markov chain; i.e. $\pi_k^t \to \pi_k^\infty$. In the case of continuous state, which is called a linear Gaussian state space model, the argument is analogous where the output is still Gaussian by the closure of products of Gaussians.

| | Internal state | Observed state |
|---|---|---|
| Linear Gaussian state space | Continuous | Continuous |
| Discrete state HMM | Discrete | Continuous |

### 10.4 Inference

The "marignal" of $x_t, x_{t-1}$ is called the pairwise marginal. Now only in the case of LGSSMs the sequence of most probable states is equal to the most probable sequence of states. This can be proved as follows. If

$$p(x_{1:T}|y_{1:T}) = \mathcal{G}(x_{1:T}; \mu_{1:T}, \Sigma_{1:T}) \implies x'_{1:T} = \mu_{1:T},$$

then

$$p(x_t|y_{1:T}) = \mathcal{G}(x_t; \mu_t, \Sigma_{t,t}) \implies x^*_t = \mu_t$$

which implies that $x'_{1:T} = x^*_{1:T}$. Now suppose a binary two-state discrete model where

| $x_1$ | $x_2$ | $p(x_1, x_2\|y_1, y_2)$ |
|---|---|---|
| 0 | 0 | 0.3 |
| 0 | 1 | 0.4 |
| 1 | 0 | 0.3 |
| 1 | 1 | 0 |

then $x'_{1:2} = [0, 1]$ while $p(x_2 = 0|y_1, y_2) = 0.6$, meaning that $x^*_2 = 0$.

### 10.5 Kalman filter