**Learning in graphical models & MCMC methods** **Fall 2014**

Cours 8 — November 19

*Enseignant: Guillaume Obozinski* *Scribe: Khalife Sammy, Maryan Morel*

The web page of the course: `http://www.di.ens.fr/~fbach/courses/fall2014/`

## 8.1 HMM (end)

As a reminder, the message propagation algorithm for Hidden Markov Models writes itself under 2 recursions :

One of them is the $\alpha - recursion$ : $\alpha_{t+1} = p\left(y_{t+1} \mid z_{t+1}\right) \sum\limits_{z_t} p\left(z_{t+1} \mid z_t\right) \alpha_t\left(z_t\right)$.

Since the probabilities are usually small, one will achieve the computation with logarithms. For instance we write matricially :

$\alpha_{t+1} = M\alpha_t \Leftrightarrow exp(\tilde{\alpha}_{t+1}) = Mexp(\tilde{\alpha}_t)$ avec $\tilde{\alpha}_t = \log(\alpha_t)$

For a more robust computation, one will use more precisally :

$y = \sum_{k=1}^{n} x_k$,
with $\tilde{y} = \log(y)$, $\tilde{x} = \log(x)$
and $y = \sum_{k=1}^{n} x_k \Leftrightarrow exp\left(\tilde{y}\right) = \sum_{k=1}^{n} exp\left(\widetilde{x_k}\right)$
then $\tilde{y} = \log[\sum_{k=1}^{n} exp(\widetilde{x_k})]$

To avoid the loss of soft maximisation, a normalization of the $\widetilde{x_k}$ is welcome :
Let $M = \max\limits_k(\widetilde{x_k})$ then we can write $\tilde{y} = \log \sum_{k=1}^{n} exp(\widetilde{x_k} - M) + M$, which yields a robust computation.

**Remark 8.1.1** *For hidden markov models, the max-product algorithm will give the most probable sequence for hidden states.*

## 8.2 Multiclass classification and link with logistic regression

In this part we will consider the following case :

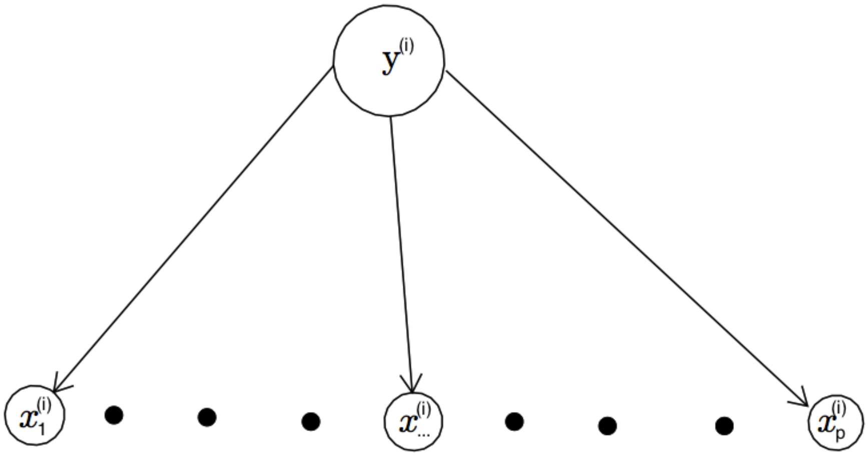
$X^i : \Omega \mapsto \{0,1\}^p$, with $X^i_j = 0$ if the word j is present in i
$Y^i : \Omega \mapsto \{0,1\}^K$

This model is used for example to classify documents by assuming the bag of words.
Let M classes of documents, and we want to assign a document to a class. For this we look at the presence of K words that will allow us to determine to which class belongs the document. X is called the binary vector of K bits giving the presence or absence of a word in the document.

We could try to directly estimate $p(y_i|x_i)$. This corresponds to a $2^k$ vector possibilities, and very limitative from a computational point of view. We therefore use another model, said Naïve Bayes.

**Exemple of generative model for K classes : Naïve Bayes**

It is here assumed that $X_i|Y_i$ are independant. This is clearly a strong hypothesis (because the presence of words can be highly correlated given the type of the document), but in practice this reveals to be quite relevant. We want to derive the model of prediction by comparaison with the logistic regression.



We suppose that $Y_i$ follows a multnomial distribution of parameters $(\pi_1, ..., \pi_K)$, and we write $\mu_{j,k} = P(X_j^{(i)} = 1|Y_k^{(i)} = 1)$

$$p(X^i = x_i, Y^i = y_i) = p(x_i, y_i) = p(x_i|y_i)p(y_i)$$

We suppose here $X_1^i|Y_1^i, ..., X_p^i|Y_K^i$ are independant, and that $(Y_1^i, ..., Y_K^i)$ are also independant. Then :

$$p(x_i|y_i)p(y_i) = \prod_{j=1}^{p}\prod_{k=1}^{K}\mu_{jk}^{\delta(x_i^{(j)}=1,y_i^k=1)}(1-\mu_{jk})^{\delta(x_i^{(j)}=0,y_i^k=1)}\prod_{k=1}^{K}\pi_k^{y_i^{(k)}}$$

Then,

$$log[\,p(x_i|y_i)p(y_i)\,] = \sum_{j=1}^{p}\sum_{k=1}^{K}[\,log(\mu_{jk})x_i^{(j)}y_i^k + log(1-\mu_{jk})(1-x_i^{(j)})(1-y_i^k)\,] + \sum_{k=1}^{K}log(\pi_k)y_i^{(k)}\,]$$

Since $p(y_i|x_i) \sim p(x_i, y_i)$ with regards to the model,

$$log[\,p(y_i|x_i)\,] = \omega_i^t\phi(x^i, y^i)$$

which is a generalization of the logistic regression for binary classification.

# 8.3 Learning on graphical models

## 8.3.1 ML principle for general Graphical Models

**Directed graphical model**

**Proposition** : Let G be a directed graph with p nodes. Assume that (X(1), ... X(n)) are i.i.d, with p features : i.e ∀ i ∈ $\{1,..n\}$ $X_i \in \mathbb{R}^p$ , and that are fully observed, i.e there is no latent or hidden variable among them. Then the ML principle decouples in p optimisation problems.

**Proof** : Let us assume we have a decoupled model, i.e :

$p \in p_u = \{p_\theta(x) = \prod_j p(x_j|x_{\pi_j}, \theta_j), \theta j = (\theta_1, ..., \theta_p) \in H = H_1 \times ... \times H_p\}$

$L(\theta) = \prod\limits_{i=1k} p(x^i|\theta) = \prod\limits_{i=1,n}\prod\limits_{j=1,..p} p(x_j|x^i_{\pi_j}, \theta_j)$
$l(\theta) = \sum_{j=1..p}\sum_{i=1,..n} logp(x^i_j|x^i_{\pi j}), \theta_j)$

Then the ML principle reduces to solving p optimization problems :

$\max\limits_{\theta_j} l_j(\theta_j)$
s.t $\theta_j \in H$

With $l_j(\theta_j) = \sum_{i=1,..n} logp(x_j^i|x_{\pi j}^i), \theta_j)$ ■

**Undirected graphical model**

→ The ML problem is convex if : the data is fully observed (no latent or hidden variable), and the parameters are decoupled.
→ In general, if the data is not fully observed, the EM scheme or similar scheme is used.
If the parameters are coupled, the problem remains convex in some cases (e.g linear coupling), but not in general.
→ If the modelisation is a tree, reformulating it as a directed tree to get back to the directed case.

## 8.4 Approximate inference

### 8.4.1 Sampling methods

We often need to comptue the expectancy of a function $f$ under some distribution $p$ that cannot be computed. Let $X$ be a random variable following the distribution $p$, we want to compute $\mu = \mathbb{E}[f(X)]$.

**Example 8.4.1** $X = (X_1, ..., X_n)$,

$$f(X) = \delta(X = x_A)$$

$$\mathbb{E}[f(X)] = \mathbb{P}(X = x_A)$$

If we know how to sample from $p$, we can use the following method :

**Algorithm 1** Monte Carlo Estimation

```
1: Draw $X_1, ..., X_n \overset{i.i.d.}{\sim} p$
2: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$
```

This method relies on the two following propositions :

**Proposition 8.1 (Law of Large Numbers (LLN))**

$$\hat{\mu} \xrightarrow{a.s.} \mu \ if \ ||\mu|| < \infty$$

**Proposition 8.2 (Central Limit Theorem (CLT))** *For $X$ a scalar random variable, if $\mathbb{V}ar(f(X)) = \sigma^2 < \infty$, then*

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

*thus* $\mathbb{E}(||\hat{\mu} - \mu||_2^2) = \frac{\sigma^2}{n}$

**How to sample from a specific distribution ?**

1. Uniform distribution on [0, 1] : use `rand`
2. Bernoulli distribution of parameter $p$ : $X = \mathbf{1}_{\{U<p\}}$ with $U \sim \mathcal{U}([0,1])$
3. Using inverse transform sampling :

$$\forall x \in \mathbb{R} \qquad F(x) = \int_{-\infty}^{x} p(t)dt = \mathbb{P}(X \in [-\infty, x])$$

$$X = F^{-1}(U) \text{ avec } U \sim \mathcal{U}([0,1])$$

**Proof** $\mathbb{P}(X \leq y) = \mathbb{P}(F^{-1}(U) \leq y) = \mathbb{P}(U \leq F(y)) = F(y)$ ■

**Example 8.4.2** *Exponential distribution (one of the rare cases admitting an explicit inverse CDF[1])*

$$p(x) = \lambda e^{-\lambda x} \mathbf{1}_{\mathbb{R}_+}(x)$$

$$X = -\frac{1}{\lambda} \ln(U)$$

### 8.4.2 Rejection sampling

Assume that $p(x)$ is known up to a constant

$$p(x) = \frac{\tilde{p}(x)}{Z_p}$$

Assume that we can construct and compute $q_k$ such that

$$\tilde{p}(x) < kq_k(x)$$

with $q_k$ a probability distribution. Assume we can sample from $q$ We define the rejection sampling (R.S.) algorithm as :

**Algorithm 2** Rejection Sampling Algorithm

```
1: Draw $X$ from $q$
2: Accept $X$ with probability $\frac{\tilde{p}(x)}{kq_k(x)} \in [0,1]$, otherwise, reject the sample
```

[1] Cumulative Distribution Function

**Proof**

$$\begin{aligned}
\mathbb{P}(X = x, X \text{ is accepted}) &= \mathbb{P}(X = x, X \text{ is accepted}) \\
&= \mathbb{P}(X \text{ is accepted}|X = x)\mathbb{P}(X = x) \\
&= \frac{\tilde{p}(x)}{kq(x)}q(x) \\
&= \frac{\tilde{p}(x)}{k}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}(X \text{ is accepted}) &= \int \frac{\tilde{p}(x)}{k}\mathrm{d}x \\
&= \frac{Z_p}{k}
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{P}(X = x|X \text{ is accepted}) &= \frac{\tilde{p}(x)}{k}\frac{k}{Z_p} \\
&= p(x)
\end{aligned}$$

■

**Remark 8.4.1** *In practice, finding $q$ and $k$ such that acceptance has a reasonably large probability is hard.*

### 8.4.3 Importance Sampling

Assume $X \sim p$. We aim to compute the expectancy of a function $f$ :

$$\begin{aligned}
\mathbb{E}_p(f(X)) &= \int f(x)p(x)\mathrm{d}x \\
&= \int \frac{f(x)p(x)}{q(x)}q(x)\mathrm{d}x \\
&= \mathbb{E}_q\left(f(Y)\frac{p(Y)}{q(Y)}\right) \qquad \text{with } Y \sim q \\
&= \mathbb{E}_q(g(Y)) \\
&\approx \frac{1}{n}\sum_{j=1}^{n} g(Y_j) \qquad \text{with } Y_j \overset{iid}{\sim} q \\
&= \frac{1}{n}\sum_{j=1}^{n} f(Y_j)\frac{p(Y_j)}{q(Y_j)}
\end{aligned}$$

$w(Y_i) = \frac{p(Y_j)}{q(Y_j)}$ are called *importance weights.* Remind that

$$\mu = \mathbb{E}_p(f(X)) \approx \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

Thus we get :

$$\begin{aligned} \mathbb{E}(\hat{\mu}) &= \frac{1}{n}\sum \int f(x)\frac{p(x)}{q(x)}q(x)dx = \int f(x)p(x)dx \\ Var(\hat{\mu}) &= \frac{1}{n}Var_{q(x)}\left(\frac{f(x)p(x)}{q(x)}\right) \end{aligned}$$

**Lemme 8.3** *If* $\forall x,\ |f(x)| \leq M$,

$$Var(\hat{\mu}) \leq \frac{M^2}{n}\int \frac{p(x)^2}{q(x)}dx.$$

**Proof**

$$\begin{aligned} Var(\hat{\mu}) = &\ \tfrac{1}{n}Var_{q(x)}\left(\tfrac{f(x)p(x)}{q(x)}\right) \\ \leq &\ \tfrac{1}{n}\int \tfrac{f(x)^2 p(x)^2}{q(x)^2}q(x)dx \\ \leq &\ \tfrac{M^2}{n}\int \tfrac{p(x)^2}{q(x)}dx. \end{aligned}$$

■

**Remark 8.4.2**

$$\begin{aligned} \int \frac{p(x)^2}{q(x)}dx &= \int \frac{p^2(x) - 2p(x)q(x) + q^2(x)}{q(x)}dx + \int \frac{2p(x)q(x) - q^2(x)}{q(x)}dx \\ &= \underbrace{\int \frac{(p(x)-q(x))^2}{q(x)}dx}_{\chi^2 \text{ divergence between } p \text{ and } q.} +1 \end{aligned}$$

*Hence, importance sampling will give good results if q has mass where p has. Indeed, if for some y, $q(y) << p(y)$, importance weights $Var(\hat{\mu})$ may be very large.*

**Extension of Importance Sampling** Assume we only know $p$ and $q$ up to a constant : $p(x) = \frac{\tilde{p}(x)}{Z_p}$ and $q(x) = \frac{\tilde{q}(x)}{Z_p}$, and only $\tilde{p}(x)$ and $\tilde{q}(x)$ are known.

$$\begin{aligned}
\mathbb{E}\left(f(Y)\frac{\tilde{p}(Y)}{\tilde{q}(Y)}\right) &= \mathbb{E}\left(f(Y)\frac{p(Y)}{q(Y)}\frac{Z_p}{Z_q}\right) = \mu\frac{Z_p}{Z_q} \\
\hat{\hat{\mu}} &= \frac{1}{n}\sum_{i=1}^{n} f(Y_i)\frac{\tilde{p}(Y_i)}{\tilde{q}(Y_i)} \xrightarrow{a.s.} \mu\frac{Z_p}{Z_q}
\end{aligned}$$

Take $f$ to be a constant, we get

$$\begin{aligned}
\hat{Z}_{p/q} &= \frac{1}{n}\sum_{i=1}^{n}\frac{p(Y_i)}{q(Y_i)} \xrightarrow{a.s.} \frac{Z_p}{Z_q} \\
\hat{\mu} &= \frac{\hat{\hat{\mu}}}{\hat{Z}_{p/q}} \xrightarrow{a.s.} \mu
\end{aligned}$$

**Remark 8.4.3** *Even if $Z_p = Z_q = 1$, renormalizing by $\hat{Z}_{p/q}$ often improves the estimation.*

## 8.5 Markov Chain Monte Carlo (MCMC)

**Context** $x \in \mathcal{X}$, $\mathcal{X}$ finite. We aim to build a Markov chain $X_0, X_1, \ldots$ such that its density $q_t(x) = p(X_t = x)$ converges to a target distribution $p(x)$.

### 8.5.1 Reminder on Markov chains

Consider order 1 homogenous Markov chains, i.e.

$$\mathbb{P}(X_t = y | X_{t-1} = x) = \mathbb{P}(X_{t-1} = y | X_{t-2} = x)$$

**Definition 8.4 (Time Homogenous Markov chain)**

$$\begin{aligned}
\forall t \geq 0\ \forall (x,y) \in \mathcal{X} \qquad & p(X_{t+1} = y \mid X_t = x, X_{t-1}, \ldots, X_0) \\
& = p(X_{t+1} = y \mid X_t = x) \\
& = p(X_1 = y \mid X_0 = x) \\
& = S(x,y)
\end{aligned}$$

**Definition 8.5 (Transition matrix)** *Let $k = card(\mathcal{X}) < \infty$. We define the matrix $S \in \mathbb{R}^{k\times k}$ such that $\forall x, y \in \mathcal{X}, S(x,y) = \mathbb{P}(X_t = y | X_{t-1} = x)$. $S$ is called* transition matrix *of the Markov chain $(X_k)_k$.*

**Properties 8.5.1** *If $k = card(\mathcal{X}) < \infty$, then:*

- $S \succeq 0$
- $S\mathbf{1} = \mathbf{1}$ *(i.e. column sum is equal to* 1*)*

*S is a stochastic matrix*

**Definition 8.6 (Stationary Distribution)** *The distribution $\pi$ on $\mathcal{X}$ is stationary if $S^T\Pi = \Pi$ where*

$$\Pi = \pi(x)_{x\in\mathcal{X}}$$

*Equivalently,*

$$i.e.\ \forall x, y \ \ \pi(y) = \sum_x \pi(x)S(x,y)$$

*If $\mathbb{P}(X_n = x) = \pi(x)$ with $\pi$ a stationary distribution of S, then we have $\mathbb{P}(X_{n+1} = y) = \sum_x \mathbb{P}(X_{n+1} = y|X_n = x)\mathbb{P}(X_n = x) = \sum_x S(x,y)\pi(x) = \pi(y)$*

**Theorem 8.7 (Perron-Frobenius)** *Every stochastic matrix S has* at least *one stationary distribution $\pi$*

**Definition 8.8 (Regular Markov Chain)** *A markov chain is regular (or equivalently aperiodic irreductible) if $\forall x, y \in \mathcal{X}, S(x,y) > 0$*

**Proposition 8.9** *If a Markov chain is regular, then its transition matrix has a unique stationary distribution $\pi$ and for any initial distribution $q_0$ on $X_0$, if $q_t(\cdot) = \mathbb{P}(X_t = \cdot)$, then $q_t \underset{t\to+\infty}{\longrightarrow} \pi$ Let $q_n$ be the distribution of $X_n$, then for all distribution $q_0$ we get*

$$q_n \to \pi$$

**Goal** We want to find

$$\pi(x) = \frac{1}{Z}\prod_{c\in\mathcal{C}} \psi_c(x_c)$$

We try to reverse engineer this distribution by finding a Markov chain converging to $\pi$

**Definition 8.10 (Detailed Balance)** *A Markov chain is* reversible *if for the transition matrix S,*

$$\exists\pi, \forall x, y \in \mathcal{X}, \pi(x)S(x,y) = \pi(y)S(y,x)$$

*This equation is called* detailed balance equation. *It can be reformulated*

$$\mathbb{P}(X_{t+1} = y, X_t = x) = \mathbb{P}(X_{t+1} = x, X_t = y)$$

**Proposition 8.11** *If $\pi$ satisfies detailed balance, then $\pi$ is a stationary distribution and $\sum_x S(x,y)p(x) = \sum_x p(y)S(y,x) = p(y)\sum_x S(y,x) = p(y)$*

### 8.5.2 Metropolis-Hastings Algorithm

**Proposal transition** $T(x,z) = \mathbb{P}(Z = z|X = x)$

**Acceptance probability** $\alpha(x,t) = \mathbb{P}(\text{Accept z } |X = x, Z = z)$

$\alpha$ is not a transition matrix.

**Algorithm 3** Metropolis Hastings

```
Initialize $x_0$ from $X_0 \sim q$
for $t = 1, \dots, T$ do
    Draw $z_t$ from $\mathbb{P}(Z = \cdot|X_{t-1} = x_{t-1}) = T(x_{t-1}, \cdot)$
    With probability $\alpha(Z_t, x_{t-1})$, set $x_t = z_t$, otherwise, set $x_t = x_{t-1}$
end for
```

**Proposition 8.12** *With that choice of $\alpha(x,z)$, if $T(\cdot,\cdot)$ is regular, then the Metropolis-Hastings algorithm defines a Markov chain that converges to $\pi$.*

**Explanation** $\mathbb{P}(X_t = x_t|X_{t-1} = x_{t-1}) = S(x_{t-1}, x_t)$

$$\begin{aligned} \forall z \neq x, S(x,z) &= T(x,z)\alpha(x,z) \\ S(x,x) &= T(x,x) + \sum_{z\neq x} T(x,z)(1-\alpha(x,z)) \end{aligned}$$

Let $\pi$ be given : we want to choose $S$ such that we have *detailed balance* :

$$\begin{aligned} \pi(x)S(x,z) &= \pi(z)S(z,x) \\ \pi(x)T(x,z)\alpha(x,z) &= \pi(z)T(z,x)\alpha(z,x) \end{aligned}$$

Then

$$\frac{\alpha(x,z)}{\alpha(z,x)} = \frac{\pi(z)T(z,x)}{\pi(x)T(x,z)} \ (*)$$

If

$$\alpha(x,z) = \min\left(1, \frac{\pi(z)T(z,x)}{\pi(x)T(x,z)}\right)$$

then

$$\begin{cases} \alpha(x,z) \in [0,1] \\ (*) \text{ is satisfied } \implies \text{ detailed balance} \end{cases}$$