Probabilistic graphical models: Introduction

Guillaume Obozinski, Swiss Data Science Center, EPFL/ETHZ
Loïc Landrieu, IGN
Timothée Lacroix, FAIR & Ecole des Ponts

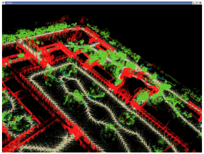
École des Pontes ParisTech

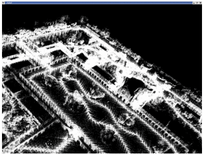
African Masters of Machine Intelligence 2018-2019, AIMS, Kigali

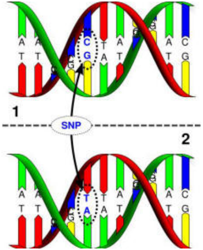
In this lecture...

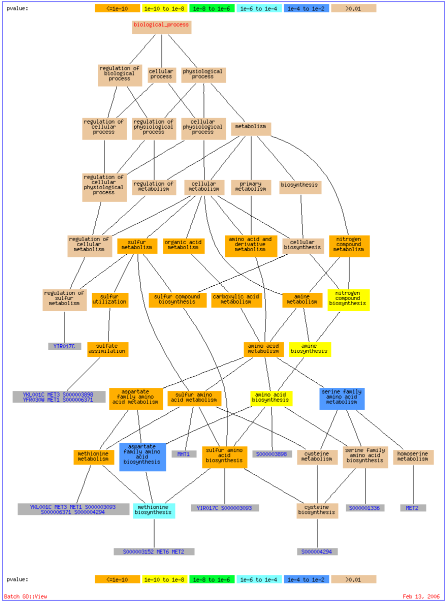
1. What are directed graphical models ?
2. What are the directed graphical models that you already know ?
3. What are the main questions that need to be answered in a theory of graphical models ?
4. Small review of formulations and computation for multinomial statistical models...

Structured problems in HD





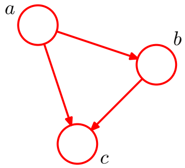




Directed graphical model or Bayesian Network

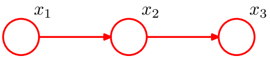
Let G be a *directed acyclic graph (DAG)*. We say that a distribution factorizes according to the graph if it can be written as a product of conditional distributions involving exactly each variable and its parent variables in the graph.

$$p(a, b, c) = p(a)\, p(b|a)\, p(c|b, a)$$



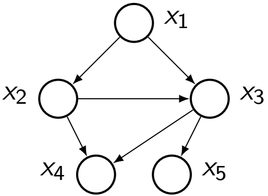
$$p(x_1, x_2) = p(x_1)p(x_2)$$

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$$



Conditional Probability Tables (CPT) :

The parameterization of DGM when variables are discrete.



Assume

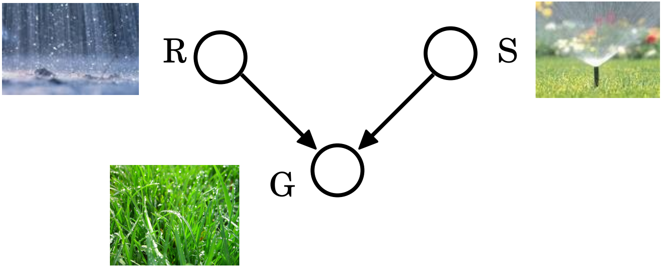
- $x_1 \in \{0, 1\}$
- $x_2 \in \{0, 1, 2\}$
- $x_3 \in \{0, 1, 2\}$

CPT for x_3 : θ_3

		$p(x_3 = k \mid x_1, x_2)$		
x_1	x_2	0	1	2
0	0	1	0	0
0	1	1	0	0
0	2	0.1	0	0.9
1	0	1	0	0
1	1	0.5	0.5	0
1	2	0.2	0.3	0.5

$$p(\mathbf{x}; \boldsymbol{\theta}) = p(x_1; \theta_1)\, p(x_2|x_1; \theta_2)\, \underline{p(x_3|x_2, x_1; \theta_3)}\, p(x_4|x_3, x_2; \theta_4)\, p(x_5|x_3; \theta_5)$$

The Sprinkler



$$P(S = 1) = 0.5$$

$$P(R = 1) = 0.2$$

- $R = 1$: it has rained
- $S = 1$: the sprinkler was on
- $G = 1$: le grass is wet

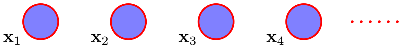
$P(G = 1 \mid S, R)$	R=0	R=1
S=0	0.01	0.8
S=1	0.8	0.95

Sequence modelling

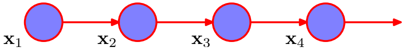
How to model the distribution of DNA sequences of length k ?

Naive model $\to 4^n - 1$ parameters

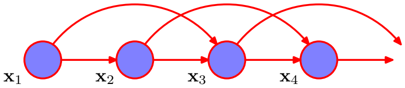
Independent model : $\to 3n$ parameters



First order Markov chain :



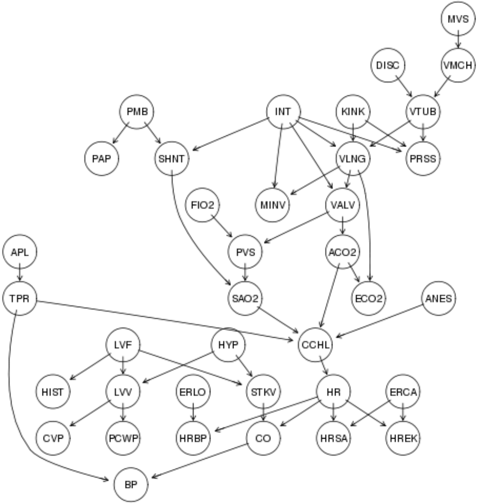
Second order Markov chain :



Number of parameters $\mathcal{O}(n)$ for chains of length n.

Anaesthesia alarm (Beinlich et al., 1989)

"The ALARM Monitoring system"



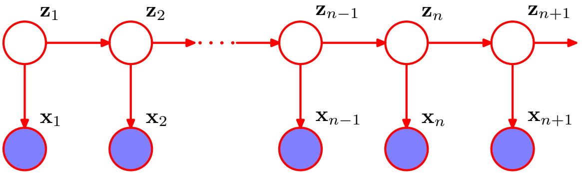
http://www.bnlearn.com/documentation/networks/

CVP	central venous pressure
PCWP	pulmonary capillary wedge pressure
HIST	history
TPR	total peripheral resistance
BP	blood pressure
CO	cardiac output
HRBP	heart rate / blood pressure.
HREK	heart rate measured by an EKG monitor
HRSA	heart rate / oxygen saturation.
PAP	pulmonary artery pressure.
SAO2	arterial oxygen saturation.
FIO2	fraction of inspired oxygen.
PRSS	breathing pressure.
ECO2	expelled CO2.
MINV	minimum volume.
MVS	minimum volume set
HYP	hypovolemia
LVF	left ventricular failure
APL	anaphylaxis
ANES	insufficient anesthesia/analgesia.
PMB	pulmonary embolus
INT	intubation
KINK	kinked tube.
DISC	disconnection
LVV	left ventricular end-diastolic volume
STKV	stroke volume
CCHL	catecholamine
ERLO	error low output
HR	heart rate.
ERCA	electrocauter
SHNT	shunt
PVS	pulmonary venous oxygen saturation
ACO2	arterial CO2
VALV	pulmonary alveoli ventilation
VLNG	lung ventilation
VTUB	ventilation tube
VMCH	ventilation machine

Models for speech processing

- Speech modeled by a sequence of unobserved phonemes
- For each phoneme a random sound is produced following a distribution which characterizes the phoneme

Hidden Markov Model : HMM (in fact Hidden Markov Chain)



$\rightarrow$ **Latent** variable models

The simplest graphical model

 $x \sim p_\theta$

- If there is no parameterization, then this is any distribution on x
- If there is a parameterization then this correspond to a

  statistical model $\mathcal{P}_\Theta = \big\{ p_\theta \mid \theta \in \Theta \big\}$

Examples of statistical models :

Bernoulli model

$$\mathcal{P}_{\text{Ber}} = \big\{ p_\pi \mid \pi \in [0,1] \big\} \qquad \text{for} \qquad p_\pi(y) = \pi^y (1-\pi)^{1-y}.$$

Gaussian model

$$\mathcal{P}_{\text{Gauss}} = \big\{ p_{\mu,\sigma^2} \mid (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^* \big\} \quad \text{for}$$

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2\sigma^2} (x-\mu)^2 \right)$$

etc

Maximum likelihood principle

- Let a model $\mathcal{P}_\Theta = \{p(x;\theta) \mid \theta \in \Theta\}$
- Let an observation x

Sir Ronald Fisher
(1890-1962)

Likelihood :

$$\begin{aligned} \mathcal{L} : \Theta &\rightarrow \mathbb{R}_+ \\ \theta &\mapsto p(x;\theta) \end{aligned}$$

Maximum likelihood estimator :

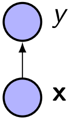
$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(x;\theta)$$

Case of i.i.d. data

For $(x_i)_{1 \le i \le n}$ a *sample* of i.i.d. data of size n :

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i;\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log\ p(x_i;\theta)$$

Simple graphical models with two nodes you already know



$p(\mathbf{x})\, p(y \mid \mathbf{x})$

Logistic regression for binary classification

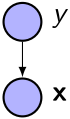
- $\mathbf{x} \in \mathbb{R}^p, \;\; y \in \{0, 1\}$
- $p_{\boldsymbol{\theta}}(y = 1|\mathbf{x}) = \dfrac{1}{1 + e^{-\langle \boldsymbol{\theta}, \mathbf{x} \rangle}}$
- $p(\mathbf{x})$ unspecified

Probabilistic model for Linear regression

- $\mathbf{x} \in \mathbb{R}^p, \;\; y \in \mathbb{R}$
- $\boldsymbol{\theta} = (\boldsymbol{w}, \sigma^2)$
- $p_{\boldsymbol{\theta}}(y \mid \mathbf{x}) = \mathcal{N}\big(y; \mu = \langle \boldsymbol{w}, \mathbf{x} \rangle, \sigma^2\big)$
- $p(\mathbf{x})$ unspecified

Those are examples of conditional models : we only model $p(y \mid \mathbf{x})$.

Simple graphical models with two nodes you already know



$p(y)\, p(\mathbf{x} \mid y)$

Gaussian mixture model ($K = 2$)

- Parameterization : $p(\mathbf{x}, y) = p_\pi(y)\, p_{\boldsymbol{\theta}}(\mathbf{x} \mid y)$

$$p_\pi(y) = \pi^y (1-\pi)^{1-y}$$
$$p_{\boldsymbol{\theta}}(\mathbf{x} \mid y = k) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- $\boldsymbol{\theta} = (\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1)$ with $\boldsymbol{\mu}_k \in \mathbb{R}^p$, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p \times p}$.

- From this model we can derive an equivalent $p(\mathbf{x})p(y \mid \mathbf{x})$ factorization using Bayes' rule : $p(y \mid \mathbf{x}) = \frac{p_\pi(y)\, p_{\boldsymbol{\theta}}(\mathbf{x}|y)}{p(y)}$.
- In particular, if $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$ this leads to Fisher's Linear Discriminant Analysis model [1] :

$$p(y \mid \mathbf{x}) = \frac{1}{1 + e^{-\langle \boldsymbol{w}, \mathbf{x} \rangle + b}} \quad \text{with} \quad \begin{cases} \boldsymbol{w} & = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ b & = \frac{1}{2} \boldsymbol{w} \boldsymbol{\Sigma} \boldsymbol{w} + \log \frac{\pi}{1-\pi}. \end{cases}$$

This is an example of a generative model : $p(x)$ is modeled as well.

1. See slide 7 in the lecture of Marc Deisenroth on logistic regression

GM associated with an *i.i.d. sample* (and plate notation)

Important to distinguish between :

- constituents of a structured random variable $X = (X_1, \dots, X_d)$ and
- an i.i.d. sample $X^{(1)}, \dots, X^{(n)}$ with $X \sim X^{(i)} = (X_1^{(i)}, \dots, X_d^{(i)})$.

When we sample data i.i.d., we have

$$p_\theta(x^{(1)}, \dots, x^{(n)}) = \prod_{i=1}^{n} p(x^{(i)}\,;\,\theta).$$

I.i.d. sampling itself corresponds to a graphical model :

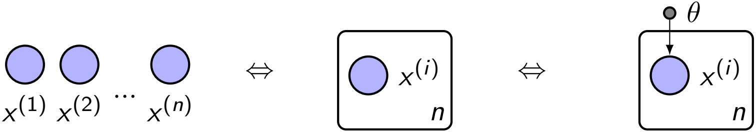


Plate Notation Plate + parameters explicit.

We use the plate notation to represent variables which :

- have the same conditional distribution
- and are independent from one another *when all other variables are fixed*.

Bayesian estimation

Bayesians treat the parameter θ as a random variable.

A priori

The Bayesian has to specify an *a priori* distribution $p(\theta)$ for the model parameters θ, which models his prior belief of the relative plausibility of different values of the parameter.

A posteriori

The observation contribute through the likelihood : $p(x|\theta)$.
The *a posteriori* distribution on the parameters is then

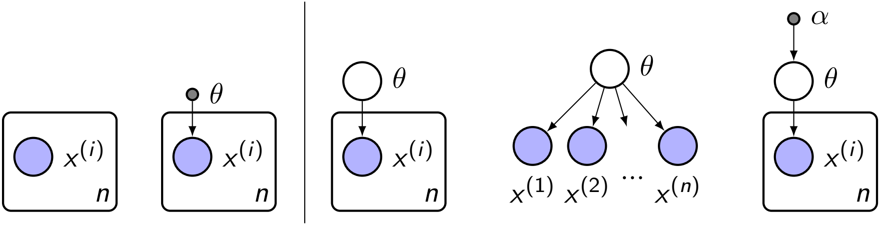
$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{p(x)} \propto p(x|\theta)\,p(\theta)\,.$$

$\rightarrow$ The Bayesian estimator is therefore a probability distribution on the parameters.

This estimation procedure is called Bayesian inference.

Bayesian GM for models with a single variable

- Frequentist model : $p_\theta(x^{(1)}, \ldots, x^{(n)}) = \prod_{i=1}^{n} p_\theta(x^{(i)})$
- Bayesian model : $p(x^{(1)}, \ldots, x^{(n)}, \theta) = \prod_{i=1}^{n} p(x^{(i)} \mid \theta)\, p_\alpha(\theta)$



Frequentist model

Bayesian formulation

- The $\left(x^{(i)}\right)_{i=1..n}$ are independent *when θ is fixed*.
- α is the parameter of the prior distribution : it is a hyperparameter.

Some exercises

1. Write the graphical model for an i.i.d. sample in which the parameters are made explicit :
   - For the binary logistic regression model
   - For the probabilistic linear regression model
   - For the mixture of Gaussian distributions
2. Do the same thing for the corresponding Bayesian models.

Three main operations on graphical models

1 - Probabilistic inference

2 - Decoding (MAP inference)

3 - Learning the parameters

Operations on GMs : 1 - Probabilistic inference

Computing probabilities in the model

- Given that the grass is wet, what is the probability that it rained ?
- Given the blood pressure, the ECG, and the measure of expelled CO_2, what is the probability that the patient suffers from a pulmonary embolus, that she did not received a sufficient dose of analgesic, that she is not well ventilated ?
- What is the probability that the 2nd word of the sentence was "cat" ?
- Computing a marginal distribution :

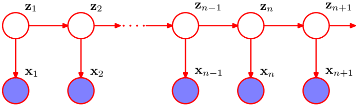
$$p(x_i) \quad \text{or} \quad p(x_i | x_1 = 3, x_7 = 0)$$

Operations on GMs : 2 - Decoding (MAP inference)

Computing most probable configurations of variable values

What is the most probable sequence of words pronounced given the sequence of phonemes heard ?

$$\text{argmax}_z \, p(z|x)$$



Operations on GMs : 3 - Learning

Given a parameterized graphical model

$$p(\mathbf{x};\boldsymbol{\theta}) = p(x_1;\theta_1)\, p(x_2|x_1;\theta_2)\, p(x_3|x_2,x_1;\theta_3)\, p(x_4|x_3,x_2;\theta_4)\, p(x_5|x_3;\theta_5),$$

in which $(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)$ are unknown.

- Given i.i.d. observations/measurement of **all** the variables,

$$\begin{aligned}
\mathbf{x}^{(1)} &= (x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, x_4^{(1)}, x_5^{(1)}) \\
\mathbf{x}^{(2)} &= (x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, x_4^{(2)}, x_5^{(2)}) \\
&\vdots \\
\mathbf{x}^{(n)} &= (x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, x_4^{(n)}, x_5^{(n)})
\end{aligned}$$

can we learn the CPTs or more generally the parameters θ_j ?

3 - Learning with **partially observed** data

Given a parameterized graphical model

$$p(\mathbf{x};\boldsymbol{\theta}) = p(x_1;\theta_1)\, p(x_2|x_1;\theta_2)\, p(x_3|x_2,x_1;\theta_3)\, p(x_4|x_3,x_2;\theta_4)\, p(x_5|x_3;\theta_5),$$

in which $(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)$ are unknown.

- Given i.i.d. observations of **a subset** the variables,

$$\begin{aligned}
\mathbf{x}^{(1)} &= (x_1^{(1)},\ ?\ ,\ ?\ , x_4^{(1)}, x_5^{(1)}) \\
\mathbf{x}^{(2)} &= (x_1^{(2)},\ ?\ ,\ ?\ , x_4^{(2)}, x_5^{(2)}) \\
&\vdots \\
\mathbf{x}^{(n)} &= (x_1^{(n)},\ ?\ ,\ ?\ , x_4^{(n)}, x_5^{(n)})
\end{aligned}$$

  can we learn the CPTs or more generally the parameters θ_j) ?

Operations on GMs : example of logistic regression

1 - Probabilistic inference

Given $\mathbf{x}$, compute $p(y \mid \mathbf{x})$ for $y \in \{0, 1\}$ $\rightarrow$ Just apply the formula.

2 - Decoding

Given $\mathbf{x}$, compute $\hat{y} = \arg\max_y \; p(y \mid \mathbf{x})$.

$$\text{Easy} \rightarrow \quad \begin{cases} \hat{y} = 1, & \text{if} \quad \mathbb{P}_{\boldsymbol{\theta}}(Y = 1 \mid \mathbf{x}) > \frac{1}{2}, \\ \hat{y} = 0, & \text{else.} \end{cases}$$

3 - Learning

Apply the *Maximum Likelihood Principle* :

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \mathrm{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(y^{(i)} \mid \mathbf{x}^{(i)}; \boldsymbol{\theta})$$

Since probabilistic inference and decoding are prediction tasks they are usually done after learning, and so with $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$.

Ops on GMs : example of Bayesian logistic regression

1 - Probabilistic inference on Y

Given $\mathbf{x}$, compute $\forall y,\ p(y \mid \mathbf{x}) = \int p(y \mid \mathbf{x}, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ or $\mathbb{E}[Y \mid x]$.

3 - Learning = Posterior (probabilistic) inference on $\boldsymbol{\theta}$ given D_n.

With **training set** $D_n = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1..n}$, compute $p(\boldsymbol{\theta}|D_n)$ or $\mathbb{E}[\boldsymbol{\theta}|D_n]$

3+1 - Posterior (probabilistic) inference on y^{new}

Given D_n, and $\mathbf{x}^{(\mathrm{new})}$, compute $\begin{cases} \mathbb{P}(Y^{(\mathrm{new})} = y \mid \mathbf{x}^{(\mathrm{new})}, D_n) & \text{or} \\ \mathbb{E}[Y^{(\mathrm{new})} \mid \mathbf{x}^{(\mathrm{new})}, D_n]. \end{cases}$

3+2 - Learning+Decoding = maximum a posteriori on Y

Typically $\hat{y} = \mathrm{argmax}_y \mathbb{P}(Y^{(\mathrm{new})} = y \mid \mathbf{x}^{(\mathrm{new})}, D_n)$.

(Since Y is binary, $\mathbb{P}(Y^{(\mathrm{new})} = 1 \mid \mathbf{x}^{(\mathrm{new})}, D_n) = \mathbb{E}[Y^{(\mathrm{new})} \mid \mathbf{x}^{(\mathrm{new})}, D_n]$)

Operations on graphical models : **Summary**

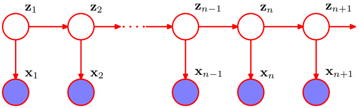
1 - Probabilistic inference

Computing a marginal distr. $p(x_i)$ or $p(x_i|x_1 = 3, x_7 = 0)$.

2 - Decoding (aka MAP inference)

Computing the most probable configuration for unobserved variables ?

$$\text{argmax}_z p(z|x)$$



3 - Learning

Schematically :

MLE	Bayesian
$\widehat{\boldsymbol{\theta}}_{MLE} = \text{argmax}_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(D_n)$	$p(\boldsymbol{\theta}\|D_n)$ or $p(\boldsymbol{y}^{(\text{new})}\|D_n, \mathbf{x}^{(\text{new})})$, etc

Other learning schemes/principles are possible (max-margin, moment methods, etc). We will focus mainly on (regularized) MLE in this course.

This week

Mo 18 am	Introduction to (directed) graphical model
Mo 18 pm	The mixture of unigram model
Tu 19 am	The mixture of unigram model and the EM algorithm
We 20 am	Practical session on EM
We 20 pm	Undirected models and graphical model theory
Th 21 am	Undirected models and graphical model theory
Fr 22 am	Quizz + Exact probabilistic inference
Fr 22 am	Practical session on exact inference

Tentative next week

Mo 25 am	HMMs
Mo 25 pm	Exponential families
Tu 26 am	Practical session on HMM
We 27 am	Exponential families / Gaussian graphical models
We 27 pm	Approximate inference
Th 28 am	Practical session on approximate inference
Fr 1 am	Quizz + TBD

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \ldots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \ldots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

For example if $K = 5$ and $c = 4$ then $\boldsymbol{y} = (0, 0, 0, 1, 0)^\top$.
So $\boldsymbol{y} \in \{0, 1\}^K$ with $\sum_{k=1}^K y_k = 1$.

$$\mathbb{P}(C = k) = \mathbb{P}(Y_k = 1) \quad \text{and} \quad \mathbb{P}(Y = \boldsymbol{y}) = \prod_{k=1}^{K} \pi_k^{y_k}.$$

Bernoulli, Binomial, Multinomial

$Y \sim \text{Ber}(\pi)$	$(Y_1, \ldots, Y_K) \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K)$
$p(y) = \pi^y (1-\pi)^{1-y}$	$p(\mathbf{y}) = \pi_1^{y_1} \ldots \pi_K^{y_K}$
$N_1 \sim \text{Bin}(n, \pi)$	$(N_1, \ldots, N_K) \sim \mathcal{M}(n, \pi_1, \ldots, \pi_K)$
$p(n_1) = \binom{n}{n_1} \pi^{n_1} (1-\pi)^{n-n_1}$	$p(\mathbf{n}) = \begin{pmatrix} & n & \\ n_1 & \ldots & n_K \end{pmatrix} \pi_1^{n_1} \ldots \pi_K^{n_K}$

with

$$\binom{n}{i} = \frac{n!}{(n-i)!i!} \qquad \text{and} \qquad \begin{pmatrix} & n & \\ n_1 & \ldots & n_K \end{pmatrix} = \frac{n!}{n_1! \ldots n_K!}$$

MLE for the Bernoulli model

Let $X_1, X_2, \ldots, X_n$ an i.i.d. sample $\sim \mathrm{Ber}(\theta)$. The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i;\, \theta) = \sum_{i=1}^{n} \log \left[\theta^{x_i}(1-\theta)^{1-x_i}\right]$$

$$= \sum_{i=1}^{n} \big(x_i \log\theta + (1-x_i)\log(1-\theta)\big) = N\log(\theta) + (n-N)\log(1-\theta)$$

with $N := \sum_{i=1}^{n} x_i$.

- $\theta \mapsto \ell(\theta)$ is strongly concave $\Rightarrow$ the MLE exists and is unique.
- since ℓ differentiable + strongly concave its maximizer is the unique stationary point

$$\nabla\ell(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \frac{N}{\theta} - \frac{n-N}{1-\theta}.$$

Thus

$$\hat{\theta}_{\mathrm{ML}} = \frac{N}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$