Information Theory and Statistics
Lecture 5: Exponential families

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Ph. D. Programme 2013/2014

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

PhD STUDIES

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Parametric family

Definition (parametric family)

- A *parametric family of distributions* is a family of probability distributions indexed by parameter $\theta \in \Theta$, which specify probabilities of a stochastic process $(X_i)_{i=-\infty}^{\infty}$.
- For discrete variables we write these distributions as

$$P(X_1^n = x_1^n|\theta).$$

- For real variables, we assume that there exists a probability density function $\rho(x_1^n|\theta)$ which satisfies

$$P(X_1^n \in A|\theta) = \int_A \rho(x_1^n|\theta) dx_1^n,$$

where $\int dx_1^n$ is the integral with respect to the **n**-dimensional Lebesgue measure.

Project co-financed by the European Union
within the framework of the European Social Fund

Random samples

1. Usually, parameter θ is a single real number or a vector.
2. It is also usually assumed that variables $\mathbf{X_i}$ are probabilistically independent (given the parameter θ). In that case, we call $\mathbf{X_1^n}$ a *random sample* of length $\mathbf{n}$ drawn from distribution $\mathbf{P(X_i = x_i|\theta)}$ or $\rho(\mathbf{x_i}|\theta)$, respectively. The first case will be called a *discrete random sample*, whereas the second will be called a *real random sample*.

Project co-financed by the European Union
within the framework of the European Social Fund

Examples of parametric families

A random sample of length $\mathbf{n}$ drawn from Bernoulli distributions with success probability $\boldsymbol{\theta}$ has probability distribution

$$\mathbf{P}(\mathbf{X}_1^n = \mathbf{x}_1^n | \boldsymbol{\theta}) = \prod_{i=1}^{n} \boldsymbol{\theta}^{x_i} (1 - \boldsymbol{\theta})^{1 - x_i},$$

where $\mathbf{x}_i \in \{0, 1\}$ and $\boldsymbol{\theta} \in (0, 1)$.

A random sample of length $\mathbf{n}$ drawn from normal (or Gauss) distributions with expectation $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$ has density

$$\rho(\mathbf{x}_1^n | \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right],$$

where $\mathbf{x}_i \in (-\infty, \infty)$, $\boldsymbol{\mu} \in (-\infty, \infty)$, and $\boldsymbol{\sigma} \in (0, \infty)$.

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY

Discrete exponential families

Definition (exponential family (discrete))

Let function $\mathbf{p} : \mathbb{X} \to (0, \infty)$ satisfy $\sum_{x \in \mathbb{X}} p(x) < \infty$. Having functions $T_l : \mathbb{X} \to \mathbb{R}$, we denote the canonical sum

$$Z(\theta) = \sum_{x \in \mathbb{X}} p(x) \exp\left(\sum_{l=1}^{s} \theta_l T_l(x) \right)$$

and define **s**-*parameter exponential family*

$$P(X_1^n = x_1^n | \theta) = \prod_{i=1}^{n} p(x_i) \exp\left(\sum_{l=1}^{s} \theta_l T_l(x_i) - \ln Z(\theta) \right)$$

for $\theta = (\theta_1, \theta_2, ..., \theta_s) \in \Theta := \{\omega \in \mathbb{R}^s : Z(\omega) < \infty\}$.

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
PhD STUDIES
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Bernoulli distributions as exponential family

Example

Bernoulli distributions form an exponential family because

$$\begin{aligned} P(X_1^n = x_1^n|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \prod_{i=1}^{n} \exp\left(x_i \ln\frac{\theta}{1-\theta} + \ln(1-\theta)\right) \\ &= \prod_{i=1}^{n} \exp\left(\eta x_i - \ln Z(\eta)\right), \end{aligned}$$

where $\eta = \ln\frac{\theta}{1-\theta}$ and $Z(\eta) = 1-\theta$. Function $\eta = \eta(\theta)$ is called the logit function.

Project co-financed by the European Union
within the framework of the European Social Fund

Real exponential families

Definition (exponential family (real))

Let function $\mathbf{p} : \mathbb{R} \to (0, \infty)$ satisfy $\int \mathbf{p}(\mathbf{x})\mathbf{dx} < \infty$. Having functions $\mathbf{T}_l : \mathbb{X} \to \mathbb{R}$, we denote the canonical sum

$$\mathbf{Z}(\theta) = \int \mathbf{p}(\mathbf{x}) \exp\left(\sum_{l=1}^{s} \theta_l \mathbf{T}_l(\mathbf{x})\right) \mathbf{dx}$$

and define **s**-*parameter exponential family*

$$\rho(\mathbf{x}_1^n|\theta) = \prod_{i=1}^{n} \mathbf{p}(\mathbf{x}_i) \exp\left(\sum_{l=1}^{s} \theta_l \mathbf{T}_l(\mathbf{x}_i) - \ln \mathbf{Z}(\theta)\right)$$

for $\theta = (\theta_1, \theta_2, \ldots, \theta_s) \in \Theta := \{\theta' \in \mathbb{R}^s : \mathbf{Z}(\theta') < \infty\}$. The **s**-parameter exponential family is called of *full rank* if the interior of Θ is not empty and $\mathbf{T}_l$ do not satisfy a linear constraint of the form $\sum_{l=1}^{s} \mathbf{a}_l \mathbf{T}_l(\mathbf{x}_i) = \mathbf{c}$ for a constant $\mathbf{c}$.

Project co-financed by the European Union
within the framework of the European Social Fund

Normal distributions as exponential family

Example

Normal distributions form an exponential family because

$$\begin{aligned}\rho(x_1^n|\mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x_i^2}{2\sigma^2} + \frac{\mu x_i}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right] \\ &= \prod_{i=1}^{n} \exp\left(\alpha x_i^2 + \beta x_i - \ln Z(\alpha,\beta)\right),\end{aligned}$$

where $\alpha = -\frac{1}{2\sigma^2}$, $\beta = \frac{\mu}{\sigma^2}$, and $Z(\alpha,\beta) = \sigma\sqrt{2\pi}\exp\left[\frac{\mu^2}{2\sigma^2}\right]$.

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

PhD STUDIES

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The problem of maximum entropy

The problem of maximum entropy modeling is:

Problem (maximum entropy)

Find the probability density ρ that maximizes entropy

$$\mathbf{H}(\rho) = -\int \rho(\mathbf{x}) \ln \rho(\mathbf{x}) d\mathbf{x} \tag{1}$$

given constraints:

$$\int \rho(\mathbf{x}) d\mathbf{x} = 1, \tag{2}$$

$$\int \rho(\mathbf{x}) \mathbf{T}_i(\mathbf{x}) d\mathbf{x} = \alpha_i, \quad 1 \leq 1 \leq m. \tag{3}$$

Similar problems of maximizing entropy given some constraints appear in many applications, in machine learning in particular.

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
PhD STUDIES
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The solution of maximum entropy

Theorem

If there exists density

$$\rho^*(x) = \exp\left[\lambda_0^* + \sum_{i=1}^{m} \lambda_i^* T_i(x)\right], \tag{4}$$

where λ_i^* are chosen so that ρ^* satisfies conditions (2)–(3), then ρ^* maximizes entropy (1) on the space of probability densities that satisfy (2)–(3).

Project co-financed by the European Union
within the framework of the European Social Fund

Some remarks

1. The solution of the maximum entropy problem for discrete distributions is analogous, with probabilities replacing probability densities.
2. In certain maximization problems, there exists no ρ^* that satisfies (2)–(3). In such cases there is no distribution having the maximal entropy. This happens for example for constraints $\int x^k \rho(x) dx = \alpha_k$, where $k = 0, 1, 2, 3$. In that case we would obtain

$$\rho(x) = \exp\left[\lambda_0 + \lambda_1 x + \lambda_2 x^2 + \lambda_3 x^3\right],$$

which cannot be normalized for any $\lambda_3 \neq 0$ because $\rho(x)$ tends to infinity for either for $x \to \infty$ or $x \to -\infty$.

Project co-financed by the European Union
within the framework of the European Social Fund

Proof

Let ρ satisfy constraints (2)–(3). We obtain

$$\begin{aligned} H(\rho) &= -\int \rho(x)\ln\rho(x)dx \\ &= -D(\rho||\rho^*) - \int \rho(x)\ln\rho^*(x)dx \\ &\leq -\int \rho(x)\ln\rho^*(x)dx \\ &= -\int \rho(x)\left[\lambda_0^* + \sum_{i=1}^{m}\lambda_i^* T_i(x)\right] dx \\ &= -\int \rho^*(x)\left[\lambda_0^* + \sum_{i=1}^{m}\lambda_i^* T_i(x)\right] dx \\ &= -\int \rho^*(x)\ln\rho^*(x)dx = H(\rho^*), \end{aligned}$$

with the equality if and only if ρ and ρ^* are equal.

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

PhD STUDIES

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Langrangian function

The remaining problem is to find the suitable λ_i^*.

Theorem

Consider the density ρ_λ *and the Lagrangian function* $\mathbf{L}(\lambda)$ *defined*

$$\rho_\lambda(x) = \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x) - \ln Z(\lambda)\right], \tag{5}$$

$$L(\lambda) = \ln Z(\lambda) - \sum_{i=1}^{k} \lambda_i \alpha_i, \tag{6}$$

where the canonical sum $Z(\lambda) = \int \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x)\right] dx$.
Function $L(\lambda)$ *has a single minimum and* $\lambda^* = (\lambda_1^*, \dots, \lambda_m^*)$ *for which (5) satisfies conditions (3) is the solution of*

$$\lambda^* = \arg\min_{\lambda} L(\lambda). \tag{7}$$

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI PhD STUDIES UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY

Proof

We have

$$\frac{\partial L(\lambda)}{\partial \lambda_j} = \frac{1}{Z(\lambda)} \int T_j(x) \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x)\right] dx - \alpha_j$$
$$= \int \rho_\lambda(x) T_j(x) dx - \alpha_j.$$

Hence the Lagrangian has an extremum if and only if ρ_λ satisfies conditions (3). Further analysis shows that there is only one extremum and it is a minimum because the Lagrangian is convex.

Project co-financed by the European Union
within the framework of the European Social Fund

Proof (continued)

Indeed we obtain

$$\frac{\partial^2 L(\lambda)}{\partial\lambda_j \partial\lambda_k} = \frac{\partial}{\partial\lambda_k}\left[\frac{1}{Z(\lambda)}\int T_j(x)\exp\left[\sum_{i=1}^m \lambda_i T_i(x)\right]dx\right]$$

$$= -\frac{1}{[Z(\lambda)]^2}\left[\int T_k(x)\exp\left[\sum_{i=1}^m \lambda_i T_i(x)\right]dx\right]$$

$$\times\left[\int T_j(x)\exp\left[\sum_{i=1}^m \lambda_i T_i(x)\right]dx\right]$$

$$+\frac{1}{Z(\lambda)}\int T_k(x)T_j(x)\exp\left[\sum_{i=1}^m \lambda_i T_i(x)\right]dx.$$

Writing $\mathbf{E}\,T = \int \rho_\lambda(x)T(x)dx$, we have

$$\frac{\partial^2 L(\lambda)}{\partial\lambda_j \partial\lambda_k} = \mathbf{E}\,(T_j T_k) - \mathbf{E}\,T_j \mathbf{E}\,T_k = \mathbf{E}\,[T_j - \mathbf{E}\,T_j]\,[T_k - \mathbf{E}\,T_k]\,.$$

Project co-financed by the European Union
within the framework of the European Social Fund

Proof (finished)

We observe that the second derivative of the Lagrangian is a covariance matrix, which is nonnegative definite, i.e.,

$$\sum_{j,k=1}^{m} a_j \frac{\partial^2 L(\lambda)}{\partial \lambda_j \partial \lambda_k} a_k = E\left[\sum_{j=1}^{m} a_j \left[T_j - E\, T_j\right]\right]^2 \geq 0.$$

Hence the Lagrangian is convex and the there is only one extremum.

Project co-financed by the European Union
within the framework of the European Social Fund

Recapitulation

- Coefficients $\boldsymbol{\lambda}_i^*$ can be found by minimizing Lagrangian $\mathbf{L}(\boldsymbol{\lambda})$.
- In many problems of machine learning this can be only done numerically.
- The suitable minimization can be performed using generic minimization algorithms, e.g. minimization by conjugate gradients, or algorithms dedicated for the Lagrangian, e.g. the iterative scaling.

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
PhD STUDIES
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Improved iterative scaling

Let us assume that

$$\mathbf{T_i(x) \geq 0.}$$

Then we can minimize Lagrangian $\mathbf{L(\lambda)}$ via the improved iterative scaling.

Let $\boldsymbol{\delta = (\delta_1, \dots, \delta_m)}$. By inequality $\mathbf{\log y \leq y - 1}$ we have

$$\mathbf{L(\lambda + \delta) - L(\lambda) \leq A(\lambda, \delta) := \frac{Z(\lambda + \delta)}{Z(\lambda)} - 1 - \sum_{i=1}^{m} \delta_i \alpha_i}$$

$$\mathbf{= \int \rho_\lambda(x) \exp\left[\sum_{i=1}^{m} \delta_i T_i(x)\right] - 1 - \sum_{i=1}^{m} \delta_i \alpha_i dx.}$$

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI

PhD STUDIES

UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY

Improved iterative scaling (continued)

Function $y \mapsto \exp y$ is convex, hence $\exp\left[\sum_{i=1}^m n_i g_i\right] \leq \sum_{i=1}^m n_i \exp g_i$ for $n_i \geq 0$, $\sum_{i=1}^m n_i = 1$ by the Jensen inequality. Then putting $T_+(x) = \sum_{i=1}^m T_i(x)$ and setting $n_i = T_i(x)/T_+(x)$ and $g_i = \delta_i T_+(x)$ we obtain

$$A(\lambda, \delta) \leq B(\lambda, \delta) := \int \rho_\lambda(x) \sum_{i=1}^m \frac{T_i(x)}{T_+(x)} \exp\left[\delta_i T_+(x)\right] dx - 1 - \sum_{i=1}^m \delta_i \alpha_i.$$

The derivatives of $B(\lambda, \delta)$ w.r.t δ are

$$\frac{\partial B(\lambda, \delta)}{\partial \delta_i} = B'(\lambda, \delta_i) := \int \rho_\lambda(x) T_i(x) \exp\left[\delta_i T_+(x)\right] dx - \alpha_i,$$

$$\frac{\partial^2 B(\lambda, \delta)}{\partial \delta_i^2} = B''(\lambda, \delta_i) := \int \rho_\lambda(x) T_i(x) T_+(x) \exp\left[\delta_i T_+(x)\right] dx.$$

In the improved iterative scaling algorithm, we approximate finding the minimum of $L(\lambda)$ via stepwise finding of the minima of $B(\lambda, \delta)$ using the Newton's method.

Project co-financed by the European Union
within the framework of the European Social Fund

Improved iterative scaling (continued)

The minimum of $\mathbf{B}(\lambda, \delta)$ corresponds to condition

$$\mathbf{B}'(\lambda, \delta_i) = 0$$

for all $\mathbf{i}$. In the Newton's method, the zero of the derivative of $\mathbf{B}'(\lambda, \delta_i)$ can be found by iteration

$$\delta_i \leftarrow \delta_i - \frac{\mathbf{B}'(\lambda, \delta_i)}{\mathbf{B}''(\lambda, \delta_i)}$$

until sufficient convergence is observed.

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI

PhD STUDIES

UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY

Improved iterative scaling (finished)

procedure IMPROVED ITERATIVE SCALING
    for $i \in \{1, \ldots, k\}$ **do**
        $\lambda_i \leftarrow 0$
    end for
    repeat
        for $i \in \{1, \ldots, k\}$ **do**
            $\delta_i \leftarrow 1$
            while $\left| \frac{B'(\lambda, \delta_i)}{B''(\lambda, \delta_i)} \right| > \epsilon$ **do**
                $\delta_i \leftarrow \delta_i - \frac{B'(\lambda, \delta_i)}{B''(\lambda, \delta_i)}$
            end while
            $\lambda_i \leftarrow \lambda_i + \delta_i$
        end for
    until $\max_{i \in \{1,\ldots,k\}} |\delta_i| > \epsilon$
    for $i \in \{1, \ldots, k\}$ **do**
        return λ_i
    end for
end procedure

Project co-financed by the European Union
within the framework of the European Social Fund

KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
PhD STUDIES
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY