Information Theory, Exponential families **2013/2014**

Lecture 5 — October 30th

Lecturer: Guillaume Obozinski *Scribe: Thomas Belhalfaoui, Lénaïc Chizat*

5.1 Information Theory

5.1.1 Entropy

We will use the following properties (Jensen Inequality):

1. if $f : \mathbb{R} \to \mathbb{R}$ is convex and if X is an integrable random variable :
$$\mathbb{E}_X(f(X)) \geq f(\mathbb{E}_X(X))$$
2. if $f : \mathbb{R} \to \mathbb{R}$ is strictly convex, we have equality if and only if X is constant a.s.

Definition 5.1 (Entropy) *Let X be a random variable taking values in the finite set $\mathcal{X}$. We denote $p(x) = P(X = x)$.*
In information theory, the quantity
$$I(x) = \log \frac{1}{p(x)}$$
can be interpreted as a quantity of information carried by the occurrence of x. (This is sometimes called self-information). Entropy is defined as the expected amount of information of the random variable.
$$H(X) = E_{p(x)}\left[I(X)\right] = -\sum_{x\in\mathcal{X}} p(x)\, \log\, p(x)$$
The base of the logarithm is the natural base or 2, the latter being more consistent with bit coding interpretations of entropy. In this course we will use the natural logarithm.

5.1.2 Kullback-Leibler divergence

Definition 5.2 (Kullback Leibler Divergence) *Let p and q be two finite distributions on $\mathcal{X}$. The Kullback Leibler Divergence between p and q is defined by*
$$\begin{aligned} D(p \parallel q) &= \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x\in\mathcal{X}} \frac{p(x)}{q(x)} \left(\log \frac{p(x)}{q(x)}\right) q(x) \\ &= E_{X\sim q}\left[\frac{p(X)}{q(X)} \log \frac{p(X)}{q(X)}\right] \end{aligned}$$

KL Divergence is *not* a distance as it is not symmetric.

Proposition 5.3 *$D(p \parallel q) \geq 0$ and equality holds if and only if $p = q$.*

Proof If there exists $x \in \mathcal{X}$ such that $q(x) = 0$ and $p(x) \neq 0$ then $D(p \parallel q) = +\infty$. Otherwise, we can without loss of generality assume that $q(x) > 0$ everywhere. We make this assumption in the rest of the proof. By convexity of the function $y \mapsto y \log y$, and by Jensen's inequality, we have

$$D(p \parallel q) = E_q\left[\frac{p(X)}{q(X)} \log\left(\frac{p(X)}{q(X)}\right)\right] \geq E_q\left[\frac{p(X)}{q(X)}\right] \log E_q\left[\frac{p(X)}{q(X)}\right] = 0$$

since

$$E_q\left[\frac{p(X)}{q(X)}\right] = \sum_{x \in \mathcal{X}} \frac{p(x)}{q(x)} q(x) = \sum_{x \in \mathcal{X}} p(x) = 1.$$

Furthermore, $D(p \parallel q) = 0$ iff there is an equality in Jensen's inequality above which implies that $p(x) = cq(x)$ q-a.s., but summing this last equality over x implies that $c = 1$, which in turn implies that $p = q$. ■

Proposition 5.4 *We have the following inequalities:*

1. *$H(X) \geq 0$ with equality if X is constant a.s*
2. *$H(X) \leq \log(Card(\mathcal{X}))$*

Proof Since $p(x) = \mathbb{P}_p(X = x) \leq 1$ then $-p(x) \log p(x) \geq 0$ which implies that $H(X) \geq 0$ with equality iff $-p(x) \log p(x) = 0$ for all $x \in \mathcal{X}$, which proves the first point. Then

$$\begin{aligned} D(p \parallel q) &= -\sum_{x \in \mathcal{X}} p(x) \log q(x) - (-\sum_{x \in \mathcal{X}} p(x) \log p(x)) \\ &= -\sum_{x \in \mathcal{X}} p(x) \log q(x) - H(X) \end{aligned}$$

We choose $q_0(x) = \frac{1}{Card(\mathcal{X})}$. Then $H(X) = \log(Card(\mathcal{X})) - D$. Hence $H(X) \leq log(Card(\mathcal{X}))$. ■

Definition 5.5 (Mutual information) *Let X, Y be two random variables of joint distribution $p_{X,Y}(x, y) = P(X = x, Y = y)$ and with marginal distributions $p_X(x) = \sum_y p_{X,Y}(x, y)$ and $p_Y(y) = \sum_x p_{X,Y}(x, y)$. The mutual information of X and Y is defined by*

$$\begin{aligned} I(X, Y) &= \sum_{x,y} p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x)\, p_Y(y)} \\ &= D(p_{X,Y} \parallel p_X p_Y) \end{aligned}$$

Proposition 5.6 $I(X,Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$

Proof It directly follows from the fact that $D(p_{X,Y} \parallel p_X\, p_Y) = 0$ implies that $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ which is the definition of the independence of X and Y. ■

Independent $\Rightarrow$ not correlated **but** not correlated $\not\Rightarrow$ independence
The first implication comes from the fact that if $X \perp\!\!\!\perp Y$ then $E(X,Y) = E(X)E(Y)$ and then $Cov(X,Y) = 0$.
Counter-example for the reverse implication: if Θ is a r.v. following the uniform distribution on $[0,1]$ and we define the random variables X and Y by $X = \sin(2\pi\Theta)$ and $Y = \cos(2\pi\Theta)$ then X and Y are not correlated but dependent.

Remark 5.1.1 *The reverse is only true for Gaussian random variables.*

5.1.3 Relation between minimum Kullback-Leibler divergence and maximum likelihood principle

Definition 5.7 (Empirical distribution) *Let $x_1, ..., x_N \in \mathcal{X}$ be N i.i.d. observations of a random variable X.*
The empirical distribution of X derived from this sample is

$$\hat{p}(x) = \frac{1}{N}\sum_{n=1}^{N} \delta(x - x_n)$$

Where δ is the Dirac function, null everywhere except in 0 where it takes the value 1.

Proposition 5.8 *Let p_θ be a parameterized distribution on $\mathcal{X}$.*
Maximizing the likelihood $p_\theta(x)$ is equivalent to minimizing the KL Divergence $D(\hat{p}||p_\theta)$

Proof

$$\begin{aligned} D(\hat{p}||p_\theta) &= \sum_{x\in\mathcal{X}} \hat{p}(x) \log \frac{\hat{p}(x)}{p_\theta(x)} \\ &= -H(\hat{p}) - \sum_{x\in\mathcal{X}} \hat{p}(x)\ \log\ p_\theta(x) \\ &= -H(\hat{p}) - \frac{1}{N}\sum_{x\in\mathcal{X}}\sum_{n=1}^{N} \delta(x - x_n)\ \log\ p_\theta(x) \\ &= -H(\hat{p}) - \frac{1}{N}\sum_{n=1}^{N} \log\ p_\theta(x_n) \end{aligned}$$

The second term is equal to the opposite of the log-likelihood $p_\theta(x)$. Hence the conclusion. ■

Remark 5.1.2 *$p_\theta(x) = 0 \Rightarrow \hat{p}(x) = 0$, but $\hat{p}(x) = 0 \not\Rightarrow p_\theta(x) = 0$. So we should not try to compute $D(p_\theta||\hat{p})$, because this would rule out all the values of x that we have not encountered yet (i.e. such that $\hat{p}(x) = 0$).*

5.1.4 Maximum entropy principle

The maximum entropy principle is a different principle than the maximum likelihood principle and solves a different kind of problem. It assumes that we use the data to specify a constraint on the possible distribution we choose. The idea is to maximize the entropy $H(p)$ under the constraint that $p \in \mathcal{P}(\mathcal{X})$ where $\mathcal{P}(\mathcal{X})$ is a set of possible distribution typically specified from the data.

Let 's consider the following examples

1. A study on kangaroos estimated that $p = 3/4$ of the kangaroos are left-handed and $q = 2/3$ drink Foster beer. What is a reasonable estimate of the fraction of kangaroos that are both left-handed and drink Foster beer? The maximum entropy principle can be invoked to choose among all distributions of pairs of binary random variables. In particular, one way to formalize that we want to choose the least specific distribution that satisfies these constraints is to find the distribution with maximal entropy that satisfies the constraints on the marginals. If X is the variable "is left-handed" and Y "drinks Foster beer", then the problem is formalized as

$$\max_{p_{X,Y}} H(p_{X,Y}) \quad \text{s.t.} \quad p_{X,Y}(1,0) + p_{X,Y}(1,1) = p, \quad p_{X,Y}(0,1) + p_{X,Y}(1,1) = q.$$

   What is the solution to this problem? (Exercise)

2. Among all distributions on $\{1, \ldots, 10\}$ what is the distribution with expected value equal to 2 which has the largest entropy? (Exercise)

3. It is possible to show that the distribution on $\mathbb{R}$ with fixed mean μ and fixed variance σ^2 that has maximal differential entropy is the Gaussian distribution.

4. The principle of maximum entropy is also the principle invoked to construct distribution on angles with fixed mean and variance. It leads to the so-called *wrapped normal distribution*. A related distribution on angle which is also a maximum entropy distribution is the von Mises distribution.

The maximum entropy principle is used often when working with *contingency tables*.

5.1.5 Entropy and KL divergence for continuous random variables

Let X be a continuous random variable taking its values in the continuous space $\mathcal{X}$ and let p be its probability density function. We have the following adapted expressions of entropy and KL Divergence:

- Differential entropy:

$$H_{\text{diff}}(p) = -\int_{\mathcal{X}} p(x)\log(p(x))d\mu(x)$$

- Differential Kullback Leibler Divergence:

$$\begin{aligned} D_{\text{diff}}(p \parallel q) &= \int_{\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}d\mu(x) \\ &= E_{X\sim p}\left[\log\frac{p(X)}{q(X)}\right] \end{aligned}$$

In the continuous case, the entropy is not necessarily non-negative.

Remark 5.1.3 *The definition of $H_{diff}(p)$ depends on the reference measure μ. This means that $H_{diff}(p)$ does not capture any intrinsic properties of p any more, and loses its "physical interpretation" in terms of quantity of information, at least in an absolute sense. By contrast $D_{diff}(p \parallel q)$ does not depend on the choice of the reference measure and has therefore a stronger interpretation.*

5.2 Exponential families

Let $x_1, ..., x_N \in \mathcal{X}$ be N i.i.d. observations of a random variable X.

Definition 5.9 *A statistic Φ is just a function of the data: $x \mapsto \Phi(x) = \Phi(x_1, ..., x_N)$*

Definition 5.10 (Sufficient statistic (statistique exhaustive in French)) *A function $T : x \mapsto T(x)$ is a sufficient statistic for a model $\mathcal{P}_\Theta$ if and only if*

$$\forall\theta \in \Theta, \quad p_\theta(x) = h(x)\, g(T(x);\theta)$$

Note that in order to estimate θ from data x using the maximum likelihood principle the information of the statistics $T(x)$ carries all the information that is relevant.

Another way of interpreting what a sufficient statistic is is to take the Bayesian point of view. In Bayesian statistics, the parameter θ is modelled as a random variable and we then have:

$$p(x,\theta) = p(x|\theta)\, p(\theta) = h(x)\, g(T(x);\theta)\, p(\theta),$$

which means that $\theta \perp\!\!\!\perp X \mid T(X)$.

Definition 5.11 (Exponential family) *Let X be a random variable on $\mathcal{X}$. An exponential family is a family of distribution of the form*

$$p(x;\theta)\, d\mu(x) = h(x)\exp\left\{b(\theta)^T\phi(x) - \tilde{A}(\theta)\right\} d\mu(x),$$

where

- $h(x)$ *the ancillary statistic,*
- $h(x)d\mu(x)$ *the reference measure (or base measure),*
- $\phi(x)$ *the sufficient statistic (also called feature vector),*
- θ *the parameter,*
- $\eta = b(\theta)$ *the canonical parameter,*
- $\tilde{A}(\theta) = A(\eta)$ *the log-partition function.*

Proposition 5.12

$$A(\eta) = \log \int_{\mathcal{X}} h(x) \exp\left\{\eta^T \phi(x)\right\} d\mu(x)$$

Proof

$$1 = \int_{\mathcal{X}} p(x|\eta) d\mu(x) \;=\; e^{-A(\eta)} \int_{\mathcal{X}} h(x) \exp\left\{\eta^T \phi(x)\right\} d\mu(x)$$

■

Definition 5.13 (Canonical exponential family) *A canonical exponential family is an exponential family which such that* $b(\theta) = \theta = \eta$, *i.e.:*

$$p(x;\eta) = h(x) \exp(\eta^T \phi(x) - A(\eta))$$

Definition 5.14 (Domain) *The domain of an exponential family is defined by:*

$$\Omega \;= \{\eta \in \mathbb{R}^p \mid A(\eta) < \infty\}$$

Example 5.2.1 (Multinomial model) *Let* X *be a random variable on* $\mathcal{X} = \{0,1\}^K$. X *follows a multinomial distribution of parameter* $\pi \in [0,1]^K$.

$$\begin{aligned} p(x;\pi) &= \prod_{k=1}^{K} \pi_k^{x_k} \\ &= \exp\Big(\sum_{k=1}^{K} x_k \log \pi_k\Big) \\ &= \exp\Big(\sum_{k=1}^{K} x_k \eta_k\Big) \\ &= \exp(\langle x, \eta \rangle) \end{aligned}$$

In this expression we easily recognize:

- $\eta = (\log \pi_1, \log \pi_2, \cdots, \log \pi_K)^T$;
- $\phi(x) = x$;
- $d\mu(x)$ *the counting measure*
- $h(x) = 1$ *the constant function equal to one;*

But we don't recognize $A(\eta)$. Let us find it using Proposition 5.12:

$$\begin{aligned} A(\eta) &= \log\Big(\sum_{x\in\mathcal{X}} \exp(\eta^T x)\Big) \\ &= \log\Big(\sum_{k=1}^{K} \exp(\eta_k)\Big) \end{aligned}$$

$$\begin{aligned} p(x;\eta) &= \exp(\eta^T x - A(\eta)) \\ &= \exp\left(\sum_{k=1}^{K} \eta_k x_k - A(\eta)\right) \\ &= \exp\left(\sum_{k=1}^{K} (\eta_k - A(\eta)) x_k\right) \\ &= \exp\left(\sum_{k=1}^{K} \log\left(\frac{\exp \eta_k}{\sum_{k'=1}^{K} \exp \eta_k}\right) x_k\right) \end{aligned}$$

We see that in the first expression of the likelihood in its exponential form, we did not take into account the fact that $\sum_k \pi_k = 1$. There was a hidden constraint on η. Now we have a new expression for π_k and no more constraint over the values that η can take:

$$\tilde{\pi}_k = \frac{\exp(\eta_k)}{\sum_{k'} \exp(\eta_k)}.$$

Example 5.2.2 (Gaussian distribution (μ, σ) over $\mathbb{R}$)

$$\begin{aligned} p(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \exp\left\{x^2\left(\frac{-1}{2\sigma^2}\right) + x\frac{\mu}{\sigma^2} - \left[\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right]\right\} \end{aligned}$$

We recognize an exponential family with:

- $\phi(x) = (x,\ x^2)^T$
- $\eta = (\frac{\mu}{\sigma^2},\ -\frac{1}{2\sigma^2})^T = (\eta_1,\ \eta_2)^T$

- $A(\eta) = \frac{1}{2}\log\left(-\frac{2\pi}{2\eta_2}\right) - \frac{\eta_1^2}{4\eta_2}$

$$p(x) = \exp\left\{\phi(x)^T\eta - A(\eta)\right\}$$

on the domain: $\{\eta \in \mathbb{R}^2, \eta_2 < 0\}$.

Example 5.2.3 *Many other common distributions are exponential families: Binomial law, Poisson law ($\mathcal{X} = \mathbb{N}$), Dirichlet law, Gamma law, exponential law.*

5.2.1 Link with the graphical models

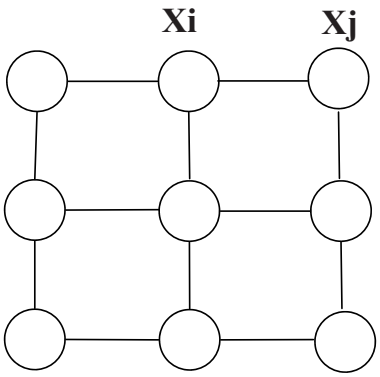

Figure 5.1. Ising model

Example 5.2.4 (Ising model)

$$p_\eta(x) = \frac{1}{Z(\eta)}\exp\sum_{(i,j)\in E}\psi_{ij}(x_i, x_j, \eta)$$

$$\psi_{ij}(x_i, x_j) = V_{ij}^{11}x_ix_j + V_{ij}^{10}x_i(1-x_j) + V_{ij}^{01}(1-x_i)x_j + V_{ij}^{00}(1-x_i)(1-x_j)$$

$$\eta = (V_{ij}^{kk'})_{\substack{(i,j)\in E \\ k,k'\in\{0,1\}}}$$

$$\phi(x) = \begin{pmatrix} x_ix_j \\ (1-x_i)x_j \\ \vdots \end{pmatrix}_{(i,j)\in E}$$

This first expression is overparametrized. We can rewrite the expression with just one parameter per pair (x_i, x_j):

$$p_\eta(x) = \frac{1}{Z}\prod_{(i,j)\in E}\exp\left(\tilde{\eta}_{ij}x_ix_j\right)\prod_{i\in V}\exp\left(\tilde{\eta}_i x_i\right).$$

Example 5.2.5 (General discrete graphical model) *In the general case of a discrete graphical model* ***such that*** $p(x) > 0$ *for all* $x \in \mathcal{X}$*, we have:*

$$\begin{aligned} p(x) &= \frac{1}{Z} \prod_{c \in \mathcal{C}} \Psi_c(x_c) \\ &= \frac{1}{Z} \exp\left\{ \sum_{c \in \mathcal{C}} \log \Psi_c(x_c) \right\} \\ &= \frac{1}{Z} \exp\left\{ \sum_{c \in \mathcal{C}} \sum_{y_c \in \mathcal{X}_c} \delta_{\{y_c = x_c\}} \log(\Psi_c(y_c)) \right\} \end{aligned}$$

Where $\mathcal{X}_c = \{$ *set of all possible values of the r.v. on the clique* $c\}$
We recognize:

$$\Phi(x) = \left(\delta_{(x_c = y_c)}\right)_{\substack{y_c \in \mathcal{X}_c \\ c \in \mathcal{C}}}$$

and

$$\eta = \left(\log(\Psi_c(y_c))\right)_{\substack{y_c \in \mathcal{X}_c \\ c \in \mathcal{C}}}$$

5.2.2 Minimal representation

Remark 5.2.1 *Let* $p_\eta(x) = \exp\left(\eta^\top \phi(x) - A(\eta)\right) h(x) d\mu(x)$.
The set $\mathcal{N}_\eta := \{x : p_\eta(x) = 0\}$ *actually does not depend on* η *but only on* $h(x)$.

Definition 5.15 (Common set of probability zero)

$$\mathcal{N} := \{x : h(x) = 0\}$$

Definition 5.16 (Affinely dependent statistics) *We denote* $\phi(x) = (\phi_1(x), \ldots, \phi_K(x))^\top$.
The sufficient statistics are said to be affinely dependent if:

$$\exists (c_0, \ldots, c_K) \neq 0, \quad \forall x \notin \mathcal{N},\ c_0 + c_1\phi_1(x) + \ldots + c_K\phi_K(x) = 0.$$

Definition 5.17 (Minimal representation of an exponential family) *A vector of sufficient statistics provides a minimal representation of the exponential family these statistics are affinely independent.*

Theorem 5.18 *Every exponential family admits at least one minimal representation (not necessarily unique) of unique minimal dimension K.*

Remark 5.2.2 *We will quite often use redundant (i.e. not minimal) representations.*

5.2.3 Exponential family of an i.i.d. sample

We consider an i.i.d. sample $X_1, \dots, X_n$ distributed according to p_η, which belongs to an exponential family. Then

$$p_\eta(x_1, \dots, x_n) = \prod_{i=1}^n p_\eta(x_i) = \prod_{i=1}^n \left[\exp\left(\eta^\top \phi(x_i) - A(\eta)\right) h(x_i)\right]$$
$$= \exp\left(\eta^\top \Big(\sum_{i=1}^n \phi(x_i)\Big) - nA(\eta)\right) \prod_i h(x_i)$$

1. The sufficient statistics is $n\bar{\phi}$, where $\bar{\phi} = \frac{1}{n}\sum_{i=1}^n \phi(x_i)$,
2. The canonical parameter η and the domain $\Omega = \{\eta \mid A(\eta) < \infty\}$ remain the same as for a single observation,
3. The log-partition function is $nA(\eta)$.

5.2.4 General exponential family

In general, in an exponential family, we can parametrize η with a function b such that $\eta = b(\theta)$ and θ in an open connected subset Θ of $\mathbb{R}^d$.

Definition 5.19 (Curved exponential family) *An exponential family is said to be curved if its Jacobian* $J = \left\{ \frac{\partial b_j(\theta)}{\partial \theta_i} \right\}_{i,j}$ *is not full rank.*

Example 5.2.6 $p_\mu(x) = \mathcal{N}(x; \mu, \mu^2)$

5.2.5 Convexity and differentiability in exponential families

Lemme 5.20 (Hölder's inequality)

$$\forall x, y \in \mathbb{R}^d,\ p, q \geqslant 1 \text{ such that } \frac{1}{p} + \frac{1}{q} = 1$$

$$|x^\top y| \leqslant ||x||_p ||y||_q \quad \text{where } ||x||_p = \left(\sum_{k=1}^n x_k^p \right)^{\frac{1}{p}}.$$

$$\forall f, g : \mathbb{R}^n \to \mathbb{R}, \quad \int |f(x)g(x)|dx \leqslant \left(\int |f(x)|^p dx \right)^{\frac{1}{p}} \left(\int |g(x)|^q dx \right)^{\frac{1}{q}}.$$

Theorem 5.21 (Convexity) *In a canonical exponential family, we have the following properties:*

1. Ω *is a convex subset of* $\mathbb{R}^p$
2. $Z : \eta \mapsto \int \exp\left(\eta^\top \phi(x)\right) h(x)dx$ *is a convex function*
3. $A : \eta \mapsto \log\left(Z(\eta)\right)$ *is a convex function*

Proof If $\Omega = \varnothing$ or Ω is a singleton, the result is trivial.

If not, there exist $\eta_1, \eta_2 \in \Omega$ such that $\eta_1 \neq \eta_2$. Let $\eta = \alpha\,\eta_1 + (1-\alpha)\,\eta_2, \quad \alpha \in]0,1[$.

$$\begin{array}{rcll} \exp(\eta^\top\phi(x)) & \leqslant & \alpha\exp(\eta_1^\top\phi(x)) & + \quad (1-\alpha)\exp(\eta_2^\top\phi(x)) \\ \displaystyle\int \ldots h(x)d\mu(x) & \leqslant & \displaystyle\alpha\int \ldots h(x)d\mu(x) & + \quad (1-\alpha)\int \ldots h(x)d\mu(x) \\ Z(\eta) & \leqslant & \alpha\, Z(\eta_1) & + \quad (1-\alpha)\, Z(\eta_2). \end{array}$$

Thus Z is a convex function. Moreover:

$$\eta_1, \eta_2 \in \Omega \quad \Rightarrow \quad Z(\eta) \leqslant \alpha Z(\eta_1) + (1-\alpha)Z(\eta_2) < \infty \quad \Rightarrow \quad \eta \in \Omega$$

which proves that Ω is a convex set.

$$Z(\eta) = \int \exp\left(\eta^\top\phi(x)\right) h(x)d\mu(x) = \int \underbrace{(\exp\eta_1^\top\phi(x))^\alpha h(x)^\alpha}_{f(x)^\alpha} \underbrace{(\exp\eta_2^\top\phi(x))^{1-\alpha} h(x)^{1-\alpha}}_{g(x)^{(1-\alpha)}} d\mu(x)$$

By taking $p = \frac{1}{\alpha}$, we obtain:

$$\begin{array}{rcll} \displaystyle\int f(x)^\alpha g(x)^{1-\alpha} d\mu(x) \leqslant & \displaystyle\left(\int f(x)^{\alpha p} d\mu(x)\right)^{\frac{1}{p}} & \displaystyle\left(\int g(x)^{(1-\alpha)q} d\mu(x)\right)^{\frac{1}{q}} & \\ Z(\eta) \leqslant & Z(\eta_1)^\alpha & Z(\eta_2)^{1-\alpha} & \\ A(\eta) = \log(Z(\eta)) \leqslant & \alpha A(\eta_1) & +(1-\alpha)A(\eta_2). & \end{array}$$

Hence A is a convex function. ■

Corollary 5.22 *In a canonical exponential family, the maximum likelihood estimator is the solution of a convex optimization problem.*

Proof The log-likelihood is concave:

$$\ell(\eta) = \log p_\eta(x) = \eta^\top \bar{\phi}(x) - A(\eta) + \log h(x).$$

■

Remark 5.2.3 *The theorem does not hold in any of those two cases:*

1. *The family is curved,*
2. *ϕ is not fully observed and we consider the marginal likelihood of the observations.*

Theorem 5.23 *If $\eta \in \overset{\circ}{\Omega}$, then Z is $\mathcal{C}^\infty$ (and so is A) and:*

$$\frac{\partial Z}{\partial \eta_k} = \mathbb{E}_\eta[\phi_k(x)]Z(\eta)$$

$$\frac{\partial^m}{\partial \eta_1^{m_1} \dots \partial \eta_K^{m_K}} Z(\eta) = \mathbb{E}_\eta[\phi_1(x)^{m_1} \dots \phi_K(x)^{m_K}]Z(\eta)$$

Proof It is a bit technical but standard to show using the dominated convergence theorem that one can exchange differentiation and expectation in the computations of the differentials of Z. One then has

$$\begin{aligned}
\frac{\partial Z}{\partial \eta_k} &= \int \phi_k(x) \exp\left\{\eta^\top \phi(x)\right\} h(x) d\mu(x) \\
&= \int \phi_k(x) \exp\left\{\eta^\top \phi(x) - A(\eta)\right\} h(x) d\mu(x) \underbrace{exp(A(\eta))}_{Z(\eta)} \\
&= \mathbb{E}_\eta[\phi_k(x)]Z(\eta),
\end{aligned}$$

which proves the first formula (the general one can be deduced by induction). ■