exponential families

Definition (Exponential family for measures in different spaces):

Let

- $(\Omega, \mathcal{F})$ be a measurable space
- $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \mu)$ be a measure space with $\mathcal{B}_{\mathbb{R}^n}$ being the Borel σ-algebra of $\mathbb{R}^n$ and μ being a σ-finite measure (usually, μ is the Lebesgue or counting measures)
- $\Theta \subseteq \mathbb{R}^s$
- $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$ be a family probability measures in Ω indexed by Θ, ie, the maps $\Theta \longrightarrow \mathcal{P}_\Theta$, $\theta \longmapsto P_\theta$ is bijective.

Suppose, for each $\theta \in \Theta$, $\exists$ a random vector $\vec{X} : \Omega \longrightarrow \mathbb{R}^n$ such that $P_\theta = \vec{X}^* \mu$ and $\vec{X}_* P_\theta \ll \mu$.

We call $\mathcal{P}_\Theta$ an Exponential family if exist

- $\eta : \Theta \longrightarrow \mathbb{R}^s$
- $\psi : \Theta \longrightarrow \mathbb{R}$

(of class $C^1(\Theta)$ when Θ is continuous)

- $T : \mathbb{R}^n \longrightarrow \mathbb{R}^s$ measurable
- $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ measurable such that $h(x) > 0 \ \forall x \in \mathbb{R}^n$.

such that the density function may be written as

$$\rho_{\vec{X}}(x;\theta) = e^{\langle \eta(\theta), T(x) \rangle - \psi(\theta)} h(x).$$

Remarks:
- When Θ is closed and continuous, $\eta, \psi \in C^1(\mathrm{Int}(\Theta))$ and this condition concerns the applicability of the ML method
- η is called natural function
- T is called natural sufficient statistics
- ψ is the log-cumulant function, responsable by the normalization of P_θ over Ω

- h is called base function, which induces a new measure $h\mu$ (usually denoted by $h\,d\mu$) and which connects P_Θ and μ by the exponential term in the calculation of the probability of an event in $\mathcal{F}$, as represented below,

$$P_\Theta(\vec{X}(A)) = \int_{\vec{X}(A)} e^{\langle \eta(\theta), T(x)\rangle - \psi(\theta)} h(x)\, d\mu(x), \quad \forall A \in \mathcal{F}$$

where the integral is reduced to a sum in the discrete case

- This definition specifies the probability measures in Ω such that the Radon-Nikodym derivative of them and μ are given by the exponential expression

- Usually, Exponential families are defined in $\mathbb{R}^n$. Then, as both μ and P_Θ are in $\mathbb{R}^n$, the comparison $P_\Theta \ll \mu$ makes sense. In our case, Ω isn't necessarily equal to $\mathbb{R}^n$ and, consequently, we need a measurable function (X, $\vec{X}$ or $[X]$) connecting them.

Definition (Exponential families): Let

- $(\Omega, \mathcal{F}, \mu)$ measure space such that μ is σ-finite and $\Omega \subset \mathbb{R}^n$
- $h: \Omega \longrightarrow \mathbb{R}$ be a measurable function, $h(x) > 0$ almost everywhere
- $T: \Omega \longrightarrow \mathbb{R}^d$ be a measurable function
- $\eta: \Theta \to \mathbb{R}^d$ be a measurable function ($\in C^1(\Theta)$ when Θ is continuous), where
  - $\Theta = \left\{ \theta \in \mathbb{R}^d : \psi(\theta) := \log \int_\Omega e^{\langle T(x), \eta(\theta)\rangle} h(x)\, d\mu(x) < \infty \right\}$

The family of probability densities $\{p_\eta := p(\cdot\,;\theta) : \theta \in \Theta\}$ which elements are defined as

$$\begin{aligned} p(\cdot\,;\theta) : \Omega &\longrightarrow \mathbb{R} \\ x &\longmapsto p(x;\theta) = \int_{\mathbb{R}^n} e^{\langle T(x), \eta(\theta)\rangle - \psi(\theta)} h(x)\, d\mu(x) \end{aligned}$$

is called a d-parameter exponential family. Θ is called a parameter space, ψ is called log-cumulant function, h is called base function (sometimes, it presented as $e^{K(x)}$, where K is called carrier measure) and T is a sufficient statistics for η.
When $\eta = id_{\Theta}$, $\{p_\theta : \theta \in \Theta\}$ is called a d-parameter exponential family in the canonical form and Θ is called the natural parameter space for that family

When the support of the density depends on the parameter, we call that family "irregular". They aren't exponential families

- Why is T sufficient statistics?
  - By the Fisher-Neyman factorization criteria, it's easy to see it

- A family of probability measures $\wp = \{P\}$ in $\Omega \subseteq \mathbb{R}^n$ is called a d-parametric exponential family of probability measures if
  - $P \ll \mu \quad \forall \; P \in \wp$
  - $\exists \; h : \Omega \longrightarrow \mathbb{R}$ a measurable function, $h(x) > 0$ almost everywhere
  - $\exists \; T : \Omega \longrightarrow \mathbb{R}^d$ be a measurable function for some $d \in \mathbb{N}$
  - $\exists \; \eta : \Theta \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ $(\in C^1(\Theta))$ measurable function such that
    - $\Theta \cdot \left\{ \theta \in \mathbb{R}^d : \psi(\theta) := \log \int_\Omega e^{\langle T(x), \eta(\theta)\rangle} h(x)\, d\mu(x) < \infty \right\}$
  - $\exists$ bijection $\Theta \longrightarrow \wp$, $\theta \longmapsto P_\theta \in \wp$

Such that the Radon-Nykodim derivatives of μ and P_η, for $\theta \in \Theta$, assume the exponential form

$$\rho(x;\theta) = e^{\langle T(x), \eta(\theta)\rangle - \psi(\theta)} h(x) \text{ almost everywhere}$$

- Examples:
  - Binomial distribution: $\theta = p$, $p \in (0,1)$, $\Omega = \mathbb{N}$, for $n \in \mathbb{N}$

$$\rho(x;p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} e^{x \ln\left(\frac{p}{1-p}\right) + n \ln(1-p)}$$

$\eta(p) = \ln\left(\frac{p}{1-p}\right)$	$T(x) = x$	$\psi(p) = -n \log(1-p)$	$h(x) = \binom{n}{x}$	$\mu = \mu_c$

Normal distribution : $\theta = (\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$, $\Omega = \mathbb{R}$

$$p(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x - \left(\frac{\mu^2}{2\sigma^2} + \ln \sigma \right) \right]$$

$\eta(\theta) = \left(\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\right)$	$T(x) = (x, x^2)$	$\psi(\theta) = \frac{\mu^2}{2\sigma^2} + \ln\sigma$	$h(x) = 1/\sqrt{2\pi}$	$\mu = \mu_{Leb}$

Gamma : $\theta = (\alpha, \beta)$, $\overset{\text{shape}}{\alpha}, \overset{\text{scale}}{\beta} > 0$, $\Omega = (0, +\infty)$

$$p(x;\theta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} = e^{-x/\beta} \exp\left[\ln\left(\frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \right) \right]$$

$$= \exp\left[(\alpha-1)\ln(x) - \ln(\Gamma(\alpha)) - \alpha \ln\beta - x/\beta \right]$$

$$= \exp\left[(\alpha-1)\ln(x) - x/\beta - (\ln(\Gamma(\alpha)) + \alpha\ln\beta) \right]$$

$\eta(\theta) = (\alpha-1, 1/\beta)$	$T(x) = (\ln(x), -x)$	$\psi(\theta) = \ln(\Gamma(\alpha)) + \alpha\ln\beta$	$h(x) = 1$	$\mu = \mu_{Leb}$

↳ Distribution maximizes the entropy with the constrains $\mathbb{E}[X] = \alpha\beta$, $\mathbb{E}[\ln X] = \psi(\alpha) + \ln(\beta)$, ψ is called digamma function

Erlang : $\theta = (\overset{\text{scale}}{K}, \overset{1/\lambda \text{ shape}}{\lambda})$, $K \in \mathbb{N}$, $\lambda \in (0, +\infty)$, $\Omega = [0, +\infty)$

$$p(x;\theta) = \frac{\lambda^K x^{K-1} e^{-\lambda x}}{(K-1)!} = \exp\left[K\ln\lambda + (K-1)\ln x - \lambda x - \ln((K-1)!) \right]$$

$\eta(\theta) = (K-1, \lambda)$	$T = (\ln x, -x)$	$\psi(\theta) = \ln((K-1)!) - K\ln\lambda$	$h(x) = 1$	$\mu = \mu_{Leb}$

↳ It's the gamma for $K = \alpha \in \mathbb{N}$, $\lambda = 1/\beta$

- **Chi-Squared:** $\theta = K$, $K \in \mathbb{N}$ (degrees of freedom), $\Omega = (0, +\infty)$

$$p(x;\theta) = \frac{x^{\frac{K}{2}-1} e^{-\frac{x}{2}}}{2^{K/2}\Gamma(K/2)} = \exp\left[\frac{K}{2}\ln(x) - \frac{K}{2}\ln(2) + \ln\Gamma(K/2)\right] e^{-\ln(x) - x/2}$$

$\eta = K/2$	$T = \ln x$	$\psi(K) = \frac{K}{2}\ln(2) - \ln\Gamma(\frac{K}{2})$	$h(x) = e^{-\ln(x) - x/2}$	$\mu = \mu_{Leb}$

It's a gamma distribution for $\alpha = K/2$, $\beta = 2$

- **Beta:** $\theta = (\alpha, \beta)$, $\alpha > 0$, $\beta > 0$, $\Omega = (0,1)$

$$p(x;\theta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} = \exp\left[(\alpha-1)\ln x + (\beta-1)\ln(1-x) - \ln B(\alpha,\beta)\right]$$

$$= \exp\left[\alpha \ln x - \ln x + \beta \ln(1-x) - \ln(1-x) - \ln B(\alpha,\beta)\right]$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$\eta(\theta) = (\alpha, \beta)$	$T(x) = (\ln x, \ln(1-x))$	$\psi(\theta) = \ln B(\alpha,\beta)$	$h(x) = x(1-x)$	$\mu = \mu_{Leb}$

- **Bernoulli:** $\theta = p$, $p \in [0,1]$, $\Omega = \{0,1\}$

$$p(x;\theta) = p^x (1-p)^{1-x} = \exp(x\ln(p) + \ln(1-p) - x\ln(1-p))$$

$\eta(\theta) = \ln(p/1-p)$	$T(x) = x$	$\psi(p) = -\ln(1-p)$	$h(x) = 1$	$\mu = \mu_c$

Binomial for $n=1$

- **Poisson**: $\theta = \lambda$ (rate), $\lambda \in (0, +\infty)$, $\Omega = \mathbb{Z}_{\geq 0}$

$$p(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \exp\left[x \ln\lambda - \lambda - \ln(x!)\right]$$

$\eta(\theta) = \ln(\lambda)$	$T(x) = x$	$\psi(\theta) = \lambda$	$h(x) = 1/x!$	$\mu = \mu_c$

- **Exponential**: $\Theta = \lambda$, $\lambda > 0$, $\Omega = [0, +\infty)$

$$p(x;\theta) = \lambda e^{-\lambda x} = e^{\ln\lambda - \lambda x}$$

$\eta(\theta) = -\lambda$	$T(x) = x$	$\psi(\theta) = -\ln\lambda$	$h(x) = 1$	$\mu = \mu_{Leb}$

↳ Given a mean $\frac{1}{\lambda}$, it's the continuous distribution of maximum entropy

- **Dirichlet**: $\Theta = \vec{\alpha}$, given $K \in \mathbb{N} \setminus \{1\}$, $\vec{\alpha} = (\alpha_1, \dots, \alpha_K) \in \mathbb{R}^K$, $\alpha_i > 0 \ \forall\ 1 \leq i \leq K$

$\Omega = \mathrm{Simp}(K-1) := \{x \in (0,1]^K : \sum_{i=1}^K x_i = 1\}$

$$p(x;\theta) = \underbrace{\frac{\Gamma(\overbrace{\sum_{i=1}^K \alpha_i}^{\alpha_0})}{\prod_{j=1}^K \Gamma(\alpha_j)}}_{\frac{1}{B(\vec{\alpha})}} \prod_{i=1}^K x_i^{\alpha_i - 1} = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^K x_i^{\alpha_i - 1}$$

$$= \exp\left(-\ln B(\vec{\alpha}) + \sum_{i=1}^K (\alpha_i - 1)\ln x_i\right)$$

$$= \exp\left(-\ln B(\vec{\alpha}) + \sum_{i=1}^K \alpha_i \ln x_i - \sum_{j=1}^K \ln x_j\right)$$

$\eta(\theta) = \vec{\alpha}$	$T(x) = (\ln x_1, \dots, \ln x_K)$	$\psi(\theta) = \ln B(\vec{\alpha})$	$h(x) = \prod_{i=1}^K \frac{1}{x_i}$	$\mu = \mu_{Leb}$

↳ multivariated beta distribution

- **Categorical**: $\Theta = \vec{p}$, given $K \in \mathbb{N}$, $\vec{p} = (p_1, \dots, p_K) \in \mathbb{R}^K$, $p_i \geq 0 \ \forall\, 1 \leq i \leq K$, $\sum_{i=1}^{K} p_i = 1$

$\Omega = \{1, \dots, K\}$

$$p(x;\theta) = \prod_{i=1}^{K} p_i^{[x=i]} = \exp\left[\sum_{i=1}^{K} [x=i] \ln(p_i)\right]$$

Iverson braket $[x=\ell] = \begin{cases} 1, & \text{if it is true} \\ 0, & \text{otherwise} \end{cases}$

$\eta(\theta) = (\ln(p_1), \dots, \ln(p_K))$	$T(x) = ([x=1], \dots, [x=K])$	$\psi(\theta) = 0$	$h(x) = 1$	$\mu = \mu_c$

- **Geometrical**: $\Theta = p$, $p \in (0,1)$, $\Omega = \mathbb{N}$

success probability of each trial $K \in \mathbb{N}$

usually, it's defined for $p \in (0,1]$. In that case, the geometric distributions are not an exponential family. Then, we remove the degenerated case of sure success of all the trials

$$p(x;\theta) = (1-p)^{x-1} p = \exp[x \ln(1-p) - \ln(1-p) + \ln p]$$

$\eta(\theta) = \ln(1-p)$	$T(x) = x$	$\psi(\theta) = \ln\left(\frac{1-p}{p}\right)$	$h(x) = 1$	$\mu = \mu_c$

Given a mean $1/p$, it's the continuous distribution of maximum entropy

- **Von Mises**: $\Theta = (\mu, \kappa)$, $\mu \in \mathbb{R}$, $\kappa > 0$, $\Omega = [t, t+2\pi)$ for some $t \in \mathbb{R}$ ($\Omega \equiv \mathbb{S}^1$)

$$p(x;\theta) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)} = \frac{e^{\kappa(\cos x \cdot \cos\mu - \sin x \cdot \sin\mu)}}{2\pi I_0(\kappa)}$$

$\eta(\theta) = (\kappa \cos\mu, -\kappa \sin\mu)$	$T(x) = (\cos x, \sin x)$	$\psi(\theta) = \ln(2\pi I_0(\kappa))$	$h(x) = 1$	$\mu = \mu_{Leb}$

- **Inverse gaussian**: $\Theta = (\underset{\text{mean}}{\nu}, \underset{\text{shape}}{\lambda})$, $\nu, \lambda > 0$, $\Omega = (0, +\infty)$

$$p(x,\theta) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left[-\frac{\lambda(x-\nu)^2}{2\nu^2 x}\right]$$

$$= \exp\left[\frac{1}{2}\ln\lambda - \frac{1}{2}\ln(2\pi x^3) - \frac{\lambda x}{2\nu^2} - \frac{\lambda}{2x} + \frac{\lambda}{\nu}\right]$$

$\eta(\theta)=\left(\frac{\lambda}{2\mu^2},\frac{\lambda}{2}\right)$	$T(x)=\left(-x,-\frac{1}{x}\right)$	$\psi(\theta)=-\frac{1}{2}\ln\lambda-\frac{\lambda}{\mu}$	$h(x)=\frac{1}{2}\ln(2\pi x^3)$	$\mu=\mu_{Leb}$

↳ it name arises from the inverse role of these distributions in the Brownian movement

- **Wishart**: $\theta=([V],p)$, given $n\in\mathbb{N}$, $[V]\in S_n^{++}(\mathbb{R})\subset\mathbb{R}^{n\times n}$, $p\geq n$, $\Omega=\mathcal{C}_n$ the cone of symmetric positive semidefinite matrices in $\mathbb{R}^{n\times n}$

$$\rho([x];\theta)=\frac{\det([x])^{\frac{p-n-1}{2}}}{2^{np/2}\det([V])^{p/2}\Gamma_n\left(\frac{p}{2}\right)}e^{-\frac{1}{2}tr([V]^{-1}[x])}$$

$$=\exp\left[\left(\frac{p-n-1}{2}\right)\ln(\det[x])-\frac{np}{2}\ln 2-\frac{p}{2}\ln(\det[V])-\ln\left[\Gamma_n\left(\frac{p}{2}\right)\right]-\frac{1}{2}tr([V]^{-1}[x])\right]$$

$$=\exp\left[\frac{p\ln(\det[x])}{2}-\frac{n\ln(\det[x])}{2}-\frac{\ln(\det[x])}{2}-\frac{p}{2}\left(n\ln 2+\ln(\det[V])\right)-\ln\left[\Gamma_n\left(\frac{p}{2}\right)\right]-\frac{1}{2}tr([V]^{-1}[x])\right]$$

$\eta(\theta)=\left(p,-\frac{1}{2}[V]\right)$	$T([x])=\left(\ln(\det[x]),[x]\right)$	$\psi(\theta)=\frac{p}{2}\left(n\ln 2+\ln(\det[V])\right)+\ln\left[\Gamma_n\left(\frac{p}{2}\right)\right]$
$h([x])=\det[x]^{-\frac{(n+1)}{2}}$	$\mu=\mu_{Leb\ induced}$	

↳ $\langle(a,[A]),(b,[B])\rangle=ab+tr([A]^T[B])$

- **Inverse Wishart**: $\Theta=([\Psi],p)$, given $n\in\mathbb{N}$, $[\Psi]\in S_n^{++}(\mathbb{R})\subset\mathbb{R}^{n\times n}$, $p\geq n$, $\Omega=\mathcal{C}_n$ cone of symmetric positive semidefinite matrices in $\mathbb{R}^{n\times n}$

$$\rho([x];\theta)=\frac{\det([\Psi])^{p/2}\det([x])^{-\frac{(p+n+1)}{2}}}{2^{\frac{pn}{2}}\Gamma_n\left(\frac{p}{2}\right)}e^{-\frac{1}{2}tr([\Psi][x]^{-1})}$$

$$=\exp\left[\frac{p}{2}\ln[\det([\Psi])]-\frac{p}{2}\ln(\det[x])-\frac{(n+1)}{2}\ln(\det[x])-\frac{pn}{2}\ln 2-\ln\left[\Gamma_n\left(\frac{p}{2}\right)\right]-\frac{1}{2}tr([\Psi][x]^{-1})\right]$$

$\eta(\Theta)=\left(-\frac{p}{2}, -\frac{1}{2}[\Psi]\right)$	$T([x])=\left(\ln(\det[x]), [x]\right)$	$h([x]) = \det[x]^{-\frac{(n+1)}{2}}$
$\psi(\Theta)=\frac{p}{2}\left(n\ln 2 - \ln[\det([\Psi])]\right)+\ln\left[\Gamma_n\left(\frac{p}{2}\right)\right]$		$\mu = \mu_{\text{Leb induced}}$

- **Isn't exponential:** Let's see an example of a family of probability distributions which have an "exponential format", but isn't an exponential family
  - <u>Wrapped normal</u>: $\Theta=(\mu,\sigma)$, $\mu\in\mathbb{R}$, $\sigma>0$, $\Omega=[t,t+2\pi)\equiv\mathbb{S}^1$ for $t\in\mathbb{R}$

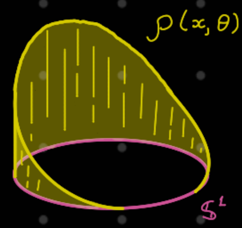
$$\rho(x;\Theta)=\frac{1}{2\pi}\,\vartheta\left(\frac{\theta-\mu}{2\pi},\frac{i\sigma^2}{2\pi}\right)=\frac{1}{\sigma\sqrt{2\pi}}\sum_{k=-\infty}^{\infty}\exp\left[-\frac{(x-\mu+2\pi k)^2}{2\sigma^2}\right]$$

Jacobi theta function

$$\vartheta(\alpha,\beta)=\sum_{k=-\infty}^{+\infty}\left[\left(e^{i\pi\alpha}\right)^2\right]^k\left(e^{i\pi\beta}\right)^{n^2}$$

normal distribution wrapping $\mathbb{S}^1$

To see that it doesn't admit the format of the exponential family, just observe the infinite sum over k. As the product $\langle\cdot,\cdot\rangle$ is always finite, that family of density are not exponential



Differential identities

- **Theorem:** Let $f:\mathbb{R}^n\longrightarrow\mathbb{R}$ be a measurable function and let Ξ_f be the set of values for $\Theta\in\mathbb{R}^d$ where

$$\int |f(x)|\, e^{\langle\Theta, T(x)\rangle}\, h(x)\, d\mu(x) < \infty$$

for μ measure in $\mathbb{R}^n$ and h and T following the suppositions in the definition of exponential families. Then, the function

$$g(\theta) = \int f(x)\, \exp^{\langle \theta, T(x)\rangle} h(x)\, d\mu(x) \in C^{\infty}(\mathrm{Int}\, \Psi_f)$$

and the derivatives can be computed by differentiation under the integral sign

Interpretation: Notice that we are not working necessarily with exponential families, because we don't have any condition about normalization. Otherwise, we can connect the log-cumulant term of normalization with the natural sufficient when the exponential family admits canonical form

For that, consider $f(x) = 1 \ \forall x \in \Omega \subset \mathbb{R}^n$ and, by definition, $\Psi_f \equiv \Psi$. Then, consider the function

$$g(\theta) = e^{\psi(\theta)} = \int e^{\langle \theta, T(x)\rangle} h(x)\, d\mu(x)$$

Differentiating under the integral sign if $\theta \in \mathrm{Int}(\Psi)$, we obtain

$$e^{\psi(\theta)} \frac{\partial \psi(\theta)}{\partial \theta_i} = \int T_i(x)\, e^{\langle \theta, T(x)\rangle} h(x)\, d\mu(x)$$

dividing by $e^{\psi(\theta)} \neq 0 \ \forall \theta$

$$\Rightarrow \frac{\partial \psi}{\partial \theta_i}(\theta) = \int T_i(x)\, e^{\langle \theta, T(x)\rangle - \psi(\theta)} h(x)\, d\mu(x)$$

We see the exponential family associated to ψ, h, T and μ appearing.

$$\frac{\partial \psi}{\partial \theta_i}(\theta) = \int T_i(x)\, \rho(x;\theta)\, d\mu(x)$$

Consequently, if a random variable $X: \mathbb{R}^n \to \mathbb{R}$ has density ρ_θ with respect to μ, then

$$\boxed{\therefore \frac{\partial \psi}{\partial \theta_i}(\theta) = \mathbb{E}_\theta[T_i(X)] \quad \forall \ \theta \in \mathrm{Int}\, \Psi}$$

Sometimes, we use the Dominated Convergence Theorem. For that, there are some useful bounds

- $|e^t - 1| \leq |t| e^{|t|} \quad \forall t \in \mathbb{R}$
- $|t| < e^{|t|} \quad \forall t \in \mathbb{R}$

Moments and Cumulants

Consider the random vector $\vec{T} = (T_1, \dots, T_d) : \mathbb{R}^n \longrightarrow \mathbb{R}^d$. The moment-generating function, in this case, is

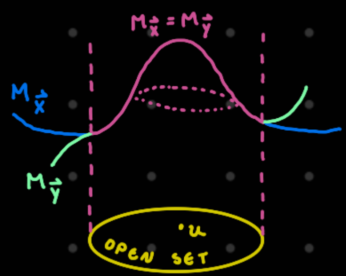
$$M_{\vec{T}}(u) = \mathbb{E}[e^{\langle u, \vec{T} \rangle}], \quad u \in \mathbb{R}^d$$

If $M_{\vec{T}}(u)$ exists, the cumulant generating function is

$$K_T(u) = \ln M_T$$

Lemma: If the moment generating functions $M_{\vec{X}}$ and $M_{\vec{Y}}$ for $\vec{X}, \vec{Y}$ random vectors are finite and coincide in some nonempty open set, then $\rho_X = \rho_Y$

↳ The moments, locally, determine the probability density globally



Now, let $\vec{X} : \mathbb{R}^n \to \mathbb{R}^d$ be a random vector with exponential probability density in the canonical form. Then, considering the random vector $T(\vec{X})$, we have

$$\begin{aligned}
\mathbb{E}_\theta[e^{\langle u, T(x) \rangle}] &= \int_\Omega e^{\langle u, T(x) \rangle} \rho_\theta(x)\, d\mu(x) \\
&= \int_\Omega e^{\langle u, T(x) \rangle} e^{\langle \theta, T(x) \rangle - \psi(\theta)} h(x)\, d\mu(x) \\
&= \int_\Omega e^{\langle u+\theta, T(x) \rangle - \psi(\theta)} h(x)\, d\mu(x)
\end{aligned}$$

Suppose that $u+\theta \in \Xi$. Then,

$$\mathbb{E}_\theta\left[e^{\langle u, T(x)\rangle}\right] = e^{\psi(u+\theta)-\psi(\theta)} \int_\Omega e^{\langle u+\theta, T(x)\rangle - \psi(u+\theta)} h(x)\, d\mu(x)$$

$$= e^{\psi(u+\theta)-\psi(\theta)} \overbrace{\int_\Omega \rho_{u+\theta}(x)\, d\mu(x)}^{1}$$

$$= e^{\psi(u+\theta)-\psi(\theta)}$$

Consequently, the cumulant-generating function is

$$K_T(u) = \psi(u+\theta) - \psi(\theta)$$

Note that, by definition, K_T should be computed using the probability density of $T(X)$, not of X. But it doesn't matter, because the density of $\vec{X}$ induces the density of T:

$$\rho_T(t) = \int_{\{x : T(x) = t\}} \rho_\theta(x)\, d\mu(x) = \int_\Omega \delta_{T(x)}(t)\, \rho_\theta(x)\, d\mu(x) \qquad \begin{cases} X : \mathbb{R}^n \longrightarrow \mathbb{R}^n \\ T : \mathbb{R}^n \longrightarrow \mathbb{R}^d \end{cases}$$

Then, on one hand,

$$\mathbb{E}_T\left[e^{\langle u, T(\vec{X})\rangle}\right] = \int_{\{t \in T(\Omega)\}} e^{\langle u, t\rangle} \rho_T(t)\, d\mu(t)$$

$$= \int_{\{t \in T(\Omega)\}} e^{\langle u, t\rangle} \left(\int_\Omega \delta_{T(x)}(t)\, \rho_\theta(x)\, d\mu(x)\right) d\mu(t)$$

$$= \int_\Omega \left(\int_{\{t \in T(\Omega)\}} e^{\langle u, t\rangle}\, \delta_{T(x)}(t)\, d\mu(t)\right) \rho_\theta(x)\, d\mu(x)$$

$$= \int_{\Omega} e^{\langle u, T(x) \rangle} \rho_\theta(x)\, d\mu(x)$$

$$= \mathbb{E}_\theta\left[e^{\langle u, T(\vec{x}) \rangle}\right]$$

It express the fact of the distribution of $T(\vec{X})$ be induced by the distribution of $\vec{X}$.

Then, in fact, the cumulant-generating function of a random vector which follows a exponential density (assuming that it admits canonical form) is a function of the log-cumulant function

$$K_T(u) = \psi(u+\theta) - \psi(\theta), \quad u+\theta \in \mathrm{Int}(\Xi)$$

Then, a K-th cumulant for T is

$$\mathcal{K}_K(T(x)) = \frac{\partial^{K_1}}{\partial u^{K_1}} \cdots \frac{\partial^{K_\ell}}{\partial u^{K_\ell}} \left(\psi(u+\theta) - \psi(\theta)\right)$$

$$= \frac{\partial^{K_1}}{\partial u_1^{K_1}} \cdots \frac{\partial^{K_\ell}}{\partial u_d^{K_\ell}} \psi(u+\theta)$$

changing the variables

$$= \frac{\partial^{K_1}}{\partial \theta_1^{K_1}} \cdots \frac{\partial^{K_\ell}}{\partial \theta_d^{K_\ell}} \psi(\theta), \quad K = \sum_{i=1}^{\ell} K_i$$

Then,

$$\boxed{\mathcal{K}_K(T(\vec{x})) = \frac{\partial^{K_1}}{\partial \theta_1^{K_1}} \cdots \frac{\partial^{K_\ell}}{\partial \theta_d^{K_\ell}} \psi(\theta), \quad K = \sum_{i=1}^{\ell} K_i}$$

K-th cumulant for random vectors are not unique, so sometimes people write $\mathcal{K}_{K_1,\dots,K_\ell}$ to make explicit the which term we are taking in the Maclaurin series for multiple variables

Curved exponential families

Definition (Full rank exponential family): An exponential family of densities $p_\theta(x) = e^{\langle \eta(\theta), T(x)\rangle - \psi(\theta)} h(x)$, $\theta \in \Xi$, is said to be of full rank if

- $\mathrm{Int}\, \eta(\Xi) \neq \emptyset$
- $\nexists\, v \in \mathbb{R}^d \setminus \{0\}$, $\nexists\, c \in \mathbb{R}$ such that $\langle v, T(x)\rangle = c$ a.e. with respect to μ

↳ The coordenates of the statistic T are "linear independents", ie, no-one is a linear combination of the other plus a cte

Each component of $T(x)$ contributes with aditional information about θ

If $\mathrm{Int}\,\eta(\Xi) = \emptyset$, η can not vary freely in one open set
↳ problems of inference (can not apply ML method, for example)

Theorem: In an exponential family $\{p_\theta = e^{\langle \eta(\theta), T(x)\rangle - \psi(\theta)} h(x) : \theta \in \Xi\}$ of full rank, T is always a complete statistic

Definition (Curved exponential families): Let $\mathcal{P} = \{P_\theta : \theta \in \Xi\}$ be a full rank d-parameter canonical exponential family with complete statistic T. Consider the subfamily $\mathcal{P}_0 \subsetneq \mathcal{P}$ of $\mathcal{P}$ not necessarily in the canonical form, but parametrized by $\zeta \in \tilde{\Xi} \subset \mathbb{R}^s$, with $\eta : \tilde{\Xi} \to \Xi$ being the parameter change function and $s < d$. Then, we can write

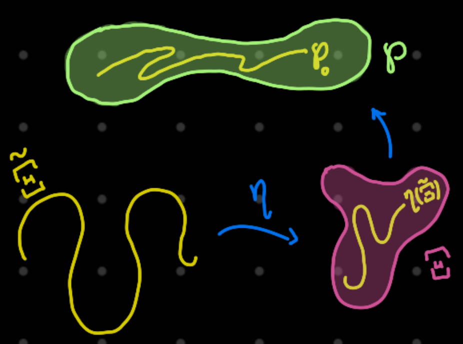
$$\mathcal{P}_0 = \{P_{\eta(\zeta)} : P_{\eta(\zeta)} \in \mathcal{P}, \zeta \in \tilde{\Xi}\}.$$

Usually, that is the common definition for curved exponential family, but we will put more one constraint which justifies the name.

If there is no $(d-s) \times s$ matrix M and no vector $\alpha \in \mathbb{R}^{d-s} \setminus \{0\}$ such that

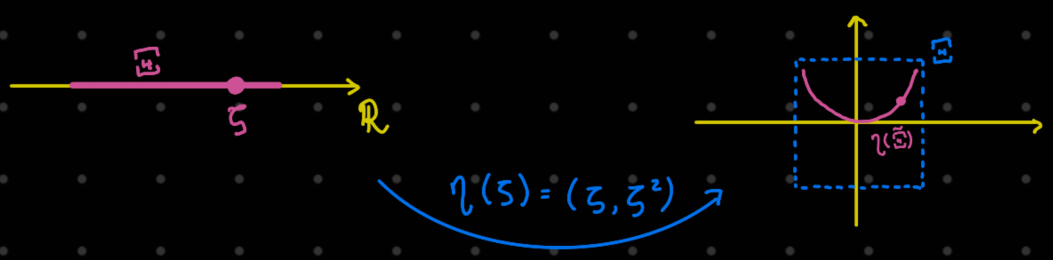
$$\eta(\tilde{\Xi}) = \{\eta(\theta) : M\eta(\theta) - \alpha = 0 \;\; \forall \theta \in \Xi\},$$

$\mathcal{P}_0$ is called a s-curved exponential family.



Interpretation: $\eta(\tilde{\Xi})$ can't be described as a set of solutions of a linear system of equations. Consequently, $\eta(\tilde{\Xi})$ is not a afin subset of Ξ. The consequence of the projection of T over $\eta(\Theta)$ is that T may not be complete anymore, but it's still minimal sufficient. There is no guarantee of any function of T with cte expectation is constant.

Example of non-linearity: $\tilde{\Xi} = (-a, a),\ a > 1$ $\qquad \Xi = (a^2+\varepsilon, a^2+\varepsilon) \times (a^2+\varepsilon, a^2+\varepsilon),\ \varepsilon > 0$



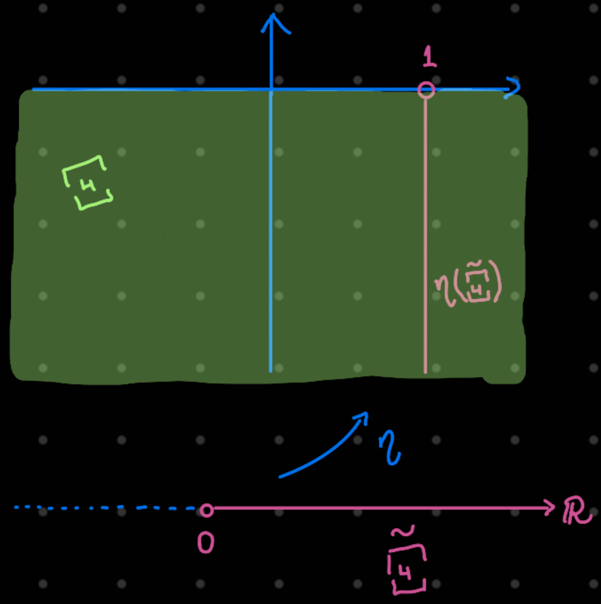
Example of linear exponential family: Consider $\{\mathcal{N}(\mu, \sigma^2)\}$, which the canonical parameters are $\eta = \left(\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\right)$ and it admits the complete sufficient statistic $T(X) = (X, X^2)$.

Consider the subfamily $\{\mathcal{N}(\theta, \theta)\},\ \theta > 0$. Then,

- if $\mu = \theta$ and $\sigma^2 = \theta \Rightarrow \begin{cases} \frac{\mu}{\sigma^2} = \frac{\theta}{\theta} = 1 \\ \frac{-1}{2\sigma^2} = \frac{-1}{2\theta} \end{cases}$

Consequently,

$$\eta : \tilde{\Xi} = \{1\} \times (0, +\infty) \longrightarrow \Xi = \mathbb{R} \times (-\infty, 0)$$
$$(1, \theta) \longmapsto \left(1, -\frac{1}{2\theta^2}\right)$$

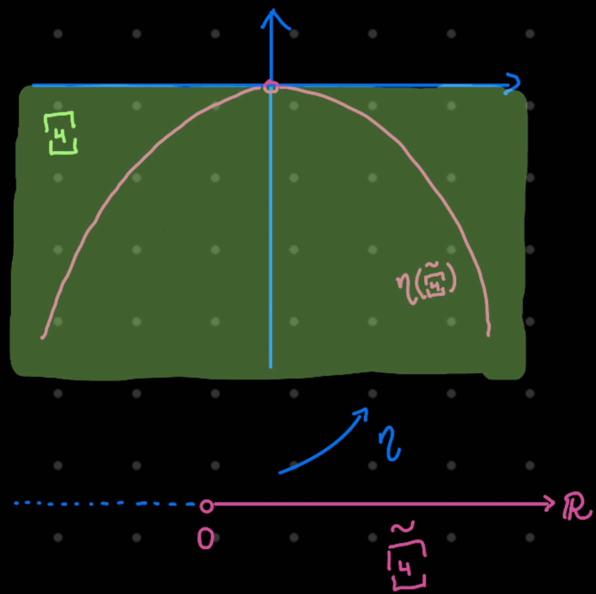


Note it's linear. For all $\theta \in \tilde{\Xi}$, $\eta(\tilde{\Xi})$ is subset of

$$\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \tilde{\eta}_1 \\ \tilde{\eta}_2 \end{pmatrix} = 1$$

Example of 1-curved exponential family: Consider now the subfamily $\{\mathcal{N}(\theta,\theta^2)\}$, $\theta>0$. Then, in this case,

- if $\mu=\theta$ and $\sigma^2=\theta \Rightarrow \begin{cases} \frac{\mu}{\sigma^2}=\frac{\theta}{\theta^2}=\frac{1}{\theta} \\ \frac{-1}{2\sigma^2}=\frac{-1}{2\theta^2} \end{cases}$



Consequently,

$$\eta : \tilde{\Xi}=(0,+\infty) \longrightarrow \Xi = \mathbb{R}\times(-\infty,0)$$
$$\theta \longmapsto \left(\frac{1}{\theta}, \frac{-1}{2\theta^2}\right)$$

Convexity

Theorem: Given an exponential family $\{p_\theta(x)=e^{\langle\theta,T(x)\rangle-\psi(\theta)}h(x)\}$ in the canonical form. The natural parameter space Ξ of this family is convex and ψ is a convex function on Ξ, ie,

$$\psi(\alpha\theta_1+(1-\alpha)\theta_2) \leq \alpha\psi(\theta_1)+(1-\alpha)\psi(\theta_2), \quad \forall\ \alpha\in(0,1),\ \forall\ \theta_1,\theta_2\in\Xi,\ \theta_1\neq\theta_2$$

Proposition: Let $\{p_\theta(x)=e^{\langle\theta,T(x)\rangle-\psi(\theta)}h(x)\}$ be an exponention family such that $\mathrm{Var}_\theta[T(X)]>0\ \forall\ \theta\in \mathrm{Int}(\Xi)$, where Ξ is the natural space of parameters. Then, $\mathbb{E}_\theta[T(X)]$ is a monotonic function on Ξ, ie, $\mathrm{grad}_\theta\, \mathbb{E}_\theta[T(X)]>0$.

Maximum entropy under linear constrains

Lemma: Let $(\Omega, \mathfrak{F})$ be a measurable space and $(V, \|\cdot\|_V)$ be a normed vector space. Consider $S(\Omega)$ the vector space of all finite measures of Ω and endowed with the total variation norm $\|\cdot\|_S$ (it will be clear in the proof). Let $\varphi: S(\Omega) \longrightarrow V$ be a continuous linear operator. Then, exists a $\mathfrak{F}$-measurable function $T: \Omega \to V$ such that, for all $\rho \in S(\Omega)$,

$$\varphi(\rho) = \int_\Omega T(\omega)\, d\rho(\omega)$$

Proof: Define the function

$$m: \mathfrak{F} \longrightarrow V$$
$$A \longmapsto \varphi(\delta_A),$$

where δ_A is the characteristic function of $A \in \mathfrak{F}$. By linearity of φ,

$m(A \cup B) = \varphi(\delta_{A\cup B}) = \varphi(\delta_A + \delta_B) = \varphi(\delta_A) + \varphi(\delta_B) = m(A) + m(B) \quad \forall\, A, B \in \mathfrak{F}$ such that $A \cap B = \emptyset$

Generalizing it, we have

$$m\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} m(A_i) \quad \forall \{A_i\}_{i\in\mathbb{N}} \subset \mathfrak{F} \text{ such that } A_j \cap A_k = \emptyset \ \forall j, k \in \mathbb{N},$$

and we know the right side converges since $\mathfrak{F}$ is closed by union and m is finite. Then, in other words, m is countable additive.

Now, for each $A \in \mathfrak{F}$, consider all the countable decomposition $\{A_i : A = \bigcup_i A_i, A_i \in \mathfrak{F}\}$ of A by measurable sets. For each one, define

$$S(A; \{A_i\}) = \sum_i \|m(A_i)\|_V = \sum_i \|\varphi(\delta_{A_i})\|_V$$

It induces the measure in Ω above, called total variation of m:

$$|m|(A) = \sup_{\{A_i\}} S(A; \{A_i\}) \quad \forall A \in \mathfrak{F}$$

By the triangular inequality, we can write

$$\|m(A)\|_V = \|\sum_i m(A_i)\|_V \leq \sum_i \|m(A_i)\|_V \qquad (1)$$

As φ is continuous, we have that $\exists\, C>0$ such that

$$\|m(A)\|_V = \|\varphi(\delta_A)\|_V \leq C\|\delta_A\|_S = C \qquad (2)$$

Joing (1) and (2), we obtain

$$\|m(A)\|_V \leq \sum_i \|m(A_i)\|_V$$
$$\|m(A)\|_V \leq C$$

These two expressions are true for all $\{A_i\}$, including $\{A_i\} = \{A\}$, and C doesn't depend on the partition. Then, $|m|(A) \leq C \;\; \forall A \in \mathfrak{F}$ and, consequently, $|m|$ is a finite measure. From that, we obtain that m and $|m|$ are σ-finite measures and $m \ll |m|$. By the Radon-NiKodym Theorem, we have $\exists$ $\mathfrak{F}$-measurable function $T: \Omega \to V$ such that

$$m(A) = \int_A T(\omega)\, d|m|(\omega)$$

To generalize this result, consider $\rho \in S(\Omega)$. By the Hahn Decomposition Theorem, $\exists$ $\rho^+, \rho^- \in S(\Omega)$ finite positive measures such that $\rho = \rho^+ - \rho^-$. Consider now a partition $\{A_i\}_{i=1}^n$ of Ω in disjoint measurable sets and define

$$\rho_n^+ := \sum_{i=1}^n \rho^+(A_i)\,\delta_{A_i} \;,\; \rho_n^- := \sum_{i=1}^n \rho^-(A_i)\,\delta_{A_i} \;,\; \rho_n := \rho_n^+ - \rho_n^-$$

This process is called approximation of ρ^+, ρ^- by simple measures ρ_n. We know that, $\forall\, \varepsilon > 0$, $\|\tilde{\rho} - \tilde{\rho}_n\|_S < \varepsilon$ for all positive $\tilde{\rho} \in S(\Omega)$. Then, by the linearity of φ, we have

$$\varphi(\rho_n) = \varphi(\rho_n^+) - \varphi(\rho_n^-)$$
$$= \sum_{i=1}^{n}\left[\rho_n^+(A_i)\,\varphi(\delta_{A_i}) - \rho_n^-(A_i)\,\varphi(\delta_{A_i})\right]$$
$$= \sum_{i=1}^{n}\left[\left(\rho_n^+(A_i) - \rho_n^-(A_i)\right)\int_{A_i} T(\omega)\, d|m|(\omega)\right]$$
$$= \sum_{i=1}^{n} \rho_n(A_i)\int_{A_i} T(\omega)\, d|m|(\omega)$$
$$= \int_{\Omega} T(\omega)\left(\sum_{i=1}^{n} \rho_n(\omega)\,\delta_{A_i}(\omega)\right) d|m|(\omega)$$

Taking the limit of $n \to \infty$, by the convergence in $\|\cdot\|_S$ of ρ_n to ρ, we obtain

$$\varphi(\rho) = \int_{\Omega} T(\omega)\, d\rho(\omega)$$

$\blacksquare$

Problem: Let $S(\Omega)$ be the space of all finite measure on the measurable space $(\Omega, \mathcal{F})$ and consider $\mathcal{P}_+(\Omega) \subset S(\Omega)$ the subspace of probability measures. Let V be a vector space and $\varphi : S(\Omega) \longrightarrow V$ a linear map. Given $p \in V$, consider $S_v = S(\Omega) \cap \varphi^{-1}(p)$.
Show that $\rho_p = \underset{\rho \in S_v}{\arg\max}\, H(\rho)$ form an exponential family for each $p \in V$ (almost everywhere in Ω)

Proof: Let $p \in V$ and note we have two constrains in this case
i) $\varphi(\rho) = p$
ii) $\rho_p(\Omega) = 1$

Based on that, we can write the Lagrangian
$\mathcal{L}(\rho, \lambda, \alpha) = H(\rho) + \langle \lambda, p - \varphi(\rho) \rangle + \alpha(1 - \rho(\Omega))$,
where $\lambda \in V^*$ and $\alpha \in \mathbb{R}$ are called Lagrange multipliers. By the lemma above, we know $\exists$ $T : S(\Omega) \longrightarrow V$ measurable function such that

$$\varphi(\rho) = \int_\Omega T(\omega)\, d\rho(\omega)$$

Based on that, to find the optimal points, the first derivative of $\mathcal{L}$ with respect to ρ needs be zero for all $\delta\rho$. So, let's start computing the variation in each term of $\mathcal{L}$:

- $$\left.\frac{d}{dt}\right|_{t=0} H(\rho + t\,\delta\rho) = -\int_\Omega (\ln \rho(\omega) + 1)\, \delta\rho(\omega)\, d\omega$$

- $$\left.\frac{d}{dt}\right|_{t=0} \langle \lambda, p - \varphi(\rho + t\,\delta\rho) \rangle = -\left[\langle \lambda, \varphi(\delta\rho) \rangle + \int_\Omega \langle \lambda, T(\omega) \rangle\, \delta\rho(\omega)\, d\omega \right]$$

- $$\left.\frac{d}{dt}\right|_{t=0} \alpha(1 - (\rho + t\,\delta\rho)(\Omega)) = -\alpha \int_\Omega \delta\rho(\omega)\, d\omega$$

Then, we have

$$\left.\frac{d}{dt}\right|_{t=0} \mathcal{L}(\rho + t\,\delta\rho, \lambda, \alpha) = 0$$

$$\Leftrightarrow \int_\Omega (\ln \rho(\omega) + 1)\, \delta\rho(\omega)\, d\omega + \langle \lambda, \varphi(\delta\rho) \rangle + \int_\Omega \langle \lambda, T(\omega) \rangle\, \delta\rho(\omega)\, d\omega + \alpha \int_\Omega \delta\rho(\omega)\, d\omega = 0$$

Note that it holds for all δt. Then,

$$\ln \rho(\omega) + 1 + \langle \lambda, T(\omega) \rangle + \alpha = 0 \quad \text{a.e.}$$

$$\Rightarrow \ln \rho(\omega) = -1 - \langle \lambda, T(\omega) \rangle - \alpha \quad \text{a.e}$$

$$\Rightarrow \rho(\omega) = e^{-[\langle \lambda, T(\omega) \rangle + (1+\alpha)]} \quad \text{a.e}$$

Applying the normalization $\rho(\Omega)$, we obtain

$$e^{-(1+\alpha)} \int_\Omega e^{-\langle \lambda, T(\omega) \rangle} d\omega = 1$$

$$\Rightarrow \quad e^{-(1+\alpha)} = \left(\int_\Omega e^{-\langle \lambda, T(\omega) \rangle} d\omega \right)^{-1}$$

Replacing it in the expression for $\rho(\omega)$, we obtain

$$\rho(\omega) = e^{\langle \lambda, \tilde{T}(\omega) \rangle - \psi(\lambda)},$$

where

$$\psi(\lambda) := \int_\Omega e^{-\langle \lambda, T(\omega) \rangle} d\omega \quad \text{and} \quad \tilde{T}(\lambda) := -T(\lambda)$$

Clearly, $\{\rho_\lambda\}$ is an exponential family in the canonical form. The role of the linear map is to introduce the sufficient statistics connecting $S(\Omega)$ and V. ∎