Exponential Families of Dsns

Def If $(T_1, T_2, \ldots, T_k)$ are linearly independent real-valued functions on $\mathbb{R}^q$ or some discrete space, $h(x) \geq 0$ and the parameters $\eta \in \mathbb{R}^k$ pdfs or pmfs are functions of x

$$f(x|\eta) = c(\eta) h(x) \exp\left(\sum_{j=1}^{k} \eta_j T_j(x)\right)$$

we'll call the family of dsns an exponential family

Fact:

$$\mathcal{E} = \left\{ \eta \in \mathbb{R}^k \;\middle|\; \int h(x) \exp \sum_{j=1}^{k} \eta_j T_j(x)\, dx < \infty \right\}$$

or

$$\sum h(x) \exp \sum_{j=1}^{k} \eta_j T_j(x) < \infty$$

is a convex subset of $\mathbb{R}^k$ ($\eta, \eta' \in \mathcal{E}$ implies $\alpha\eta + (1-\alpha)\eta'$ for $\alpha \in (0,1)$ is also in $\mathcal{E}$)

This is the largest possible parameter space called the natural parameter space

($c(\eta) = \left(\int h(x) \exp\left(\sum_{j=1}^{k} \eta_j T_j(x)\right) dx \right)^{-1}$ when the integral or sum exists ... when it doesn't there is no such dsn)

Note: If $X_1, X_2, \ldots, X_n$ are iid from an exponential family the joint pdf/pmf is

$$f(x|\eta) = c(\eta)^n \left(\prod_{i=1}^{n} h(x_i)\right) \exp \sum_{j=1}^{k} \eta_j \left(\sum_{i=1}^{n} T_j(x_i)\right)$$

and the factorization Theorem says that

$$T(X) = \left(\sum T_1(X_i), \sum T_2(X_i), \ldots, \sum T_k(X_i) \right)$$

is sufficient for any $\mathcal{E}^* \subset \mathcal{E}$ — this is called the natural sufficient statistic for the parameter η (or family $\mathcal{P} = \{P_\eta\}_{\eta \in \mathcal{E}^*}$)

not only is the natural sufficient statistic sufficient, but provided $\mathcal{E}^*$ is big enough, it is minimal sufficient — e.g. there is

Thm If $\mathcal{E}^* \subset \mathcal{E}$ contains an open rectangle, then $T(X)$ above is minimal sufficient

The argument for this is on a handout... the standard one requires

i) showing $T(X)$ has a property called "completeness"
ii) appealing to "Bahadur's Thm" (that says CCS's are minimal)

The argument for i) requires establishing uniqueness of Laplace transforms

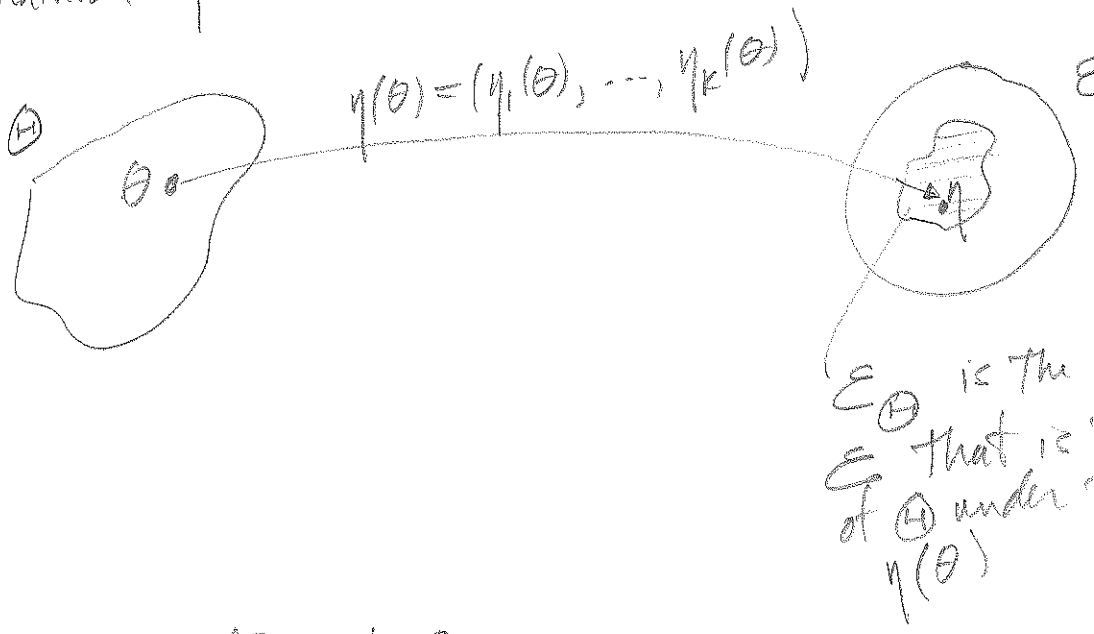
This is not the form in which this stuff is usually applied — rather, the following is more typical

Example Poisson(λ) pmf on $\{0, 1, 2, \ldots\}$ is

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\left(\frac{1}{x!}\right) e^{x(\log\lambda)}$$

$h(x)$ — $T_1(x)$ — η_1

That is, what one often has is a "standard" / "interpretable" parameterization and a (mathematically) "natural" parameterization



$T(X)$ sufficient for $\eta \in \mathcal{E}$

$\Downarrow$

$T(X)$ sufficient for $\eta \in \mathcal{E}_\Theta$

$\Updownarrow$

$T(X)$ sufficient for $\theta \in \Theta$

Back to Poisson Example

For $\lambda \in (0,\infty)$ (Θ)

$\eta(\lambda) = \log\lambda$ (so that $\lambda = e^\eta$)

$\mathcal{E} = \left\{\eta \mid \sum \frac{e^{\eta x}}{x!} < \infty\right\} = \mathbb{R}^1$

For n iid Poisson λ observations $T(X) = \sum X_i$ is minimal sufficient for $\eta \in E$

Since the image of $(0,\infty)$ under the log transform is $\mathbb{R}$, it contains an open interval, so $T(X)$ is minimal sufficient for $\lambda \in (0,\infty)$

Example (Normal)

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

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\eta_1 T_1(x) + \eta_2 T_2(x)\right) \exp\left(-\frac{\mu^2}{2\sigma^2}\right)$$

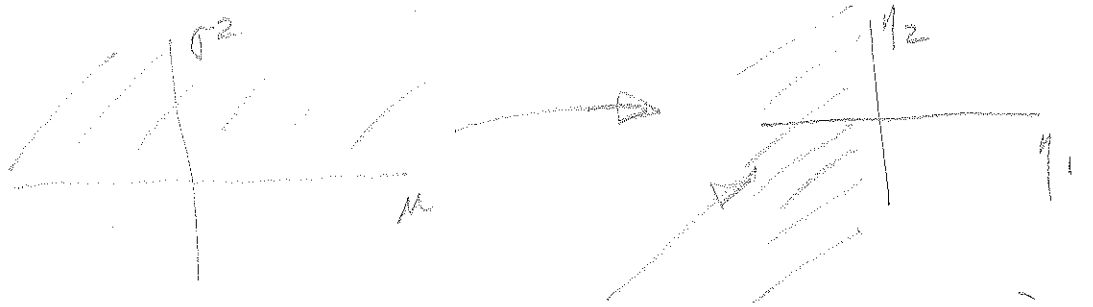
for $\eta_1 = \frac{1}{2\sigma^2}$ $\eta_2 = \frac{\mu}{\sigma^2}$

$T_1(x) = x^2$ $T_2(x) = x$

This is an exponential family of dens and for $X_1, X_2, \ldots, X_n$ iid $N(\mu,\sigma^2)$ $T(X) = (\sum X_i^2, \sum X_i)$ is is sufficient

$\eta(\mu,\sigma^2) = (\eta_1(\mu,\sigma^2), \eta_2(\mu,\sigma^2))$ maps

$\mathbb{R}^1 \times (0,\infty)$ to $(-\infty,0) \times \mathbb{R}^1$



This whole thing is the image of Θ under η - since it contains an open rectangle the natural sufficient statistic is minimal sufficient

day 10

What does this "open rectangle" business disallow?

Example $X \sim N(\mu, \mu^2)$

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}|\mu|} \exp\left(-\frac{1}{2\mu^2}x^2 + \frac{x}{\mu} - \frac{1}{2}\right)$$

with $\eta_1(\mu) = -\frac{1}{2\mu^2}$ $\eta_2(\mu) = \frac{1}{\mu}$

This is some kind of an exponential family...

so, e.g., for $X_1, X_2, \ldots, X_n$ iid $N(\mu, \mu^2)$

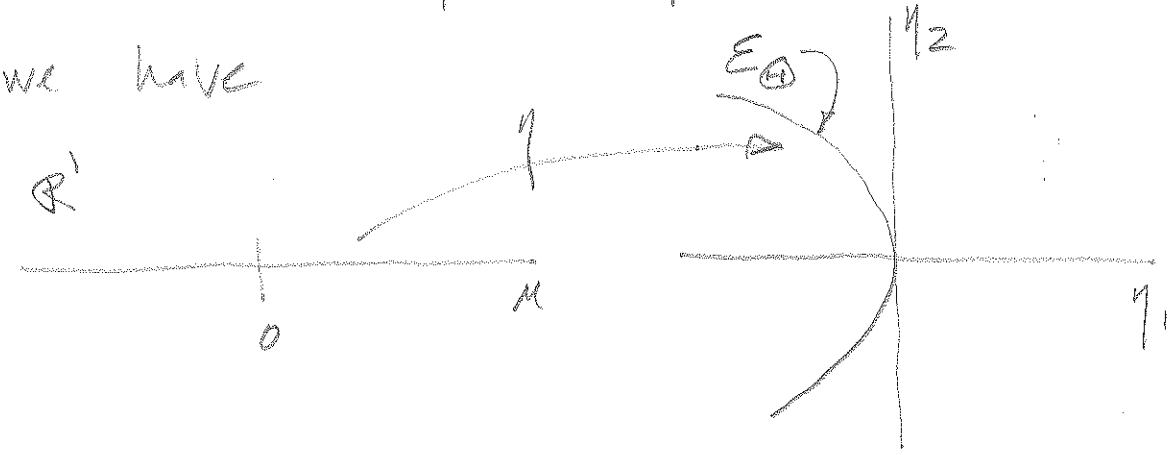
$T(\underline{x}) = \left(\sum x_i^2, \sum x\right)$ is sufficient for μ

Here (H) is really only "1-dimensional" and so also is the set of $\underline{\eta} = (\eta_1, \eta_2)$ under discussion here

i.e. since

$$\eta_1 = -\frac{1}{2}\eta_2^2$$

we have



so we can't immediately apply the theorem to get minimality of $T(\underline{x})$

BTW notice that

① $\mathcal{P} = \{N(\mu, \sigma^2)$ models $| (\mu, \sigma^2) \in \mathbb{R}^1 \times (0, \infty)\}$

is an exponential family with a genuinely 2-d η

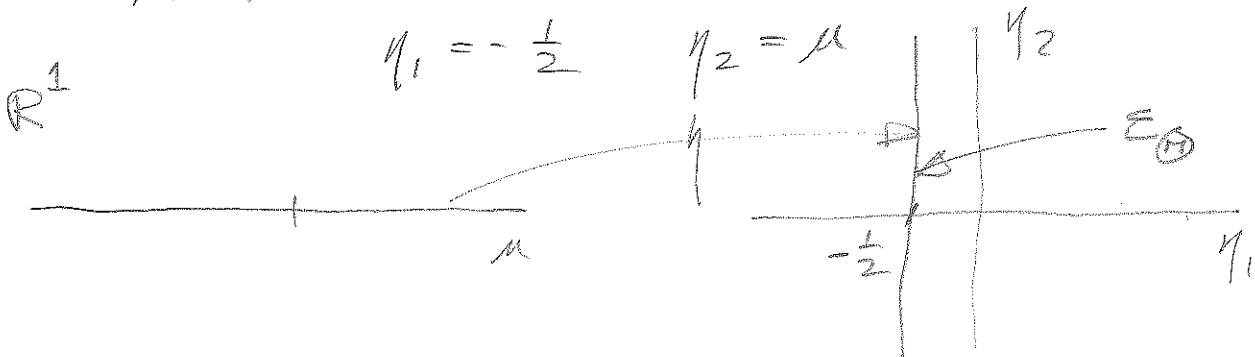
(2) $\mathcal{P} = \{ N(\mu, 1) \text{ models} \mid \mu \in R^1 \}$ has

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2 + x\mu - \frac{1}{2}\mu^2\right)$$

$\eta_1(\mu) = \mu$; $T_1(x) = x$

and with $\eta_1 = \mu$ we have an exponential family with a 1-dimensional η and single $T(x) = x$

we think of this as a sub-model of the full $N(\mu, \sigma^2)$ model we have



(3) The $N(\mu, \mu^2)$ example is something else - $E_{(1)}$ is "1-dimensional" but it isn't a case where a single $T_1(x)$ and 1-dimensional η will work

people would call this a "curved exponential family" involving k linearly independent T_j's, $E_{(1)}$ some $k' < k$ - dimensional subset of E but not writeable in terms of k' T_j's and k' - dimensional η

Measures of Information (about a parameter) in a random vector X (Fisher Information + Kullback-Leibler Information)

For classroom purposes suppose that $\Theta \subset \mathbb{R}^1$ (There is a "handout" on the web page covering / laying out the $\mathbb{R}^k$ version of this) - consider the (random) functions of θ

$L(\theta)$ the likelihood

$l(\theta) = \log L(\theta)$ the loglikelihood

$l'(\theta) = \frac{d}{d\theta} l(\theta)$ the "score function"

Note that "usually"

$$E_{\theta_0} l'(\theta) = E_{\theta_0} \frac{d}{d\theta} \log f(X|\theta)\Big|_{\theta=\theta_0}$$

$$\overset{\text{cont's case}}{=} \int \frac{\frac{d}{d\theta} f(x|\theta)\Big|_{\theta=\theta_0}}{f(x|\theta_0)} f(x|\theta_0)\,dx$$

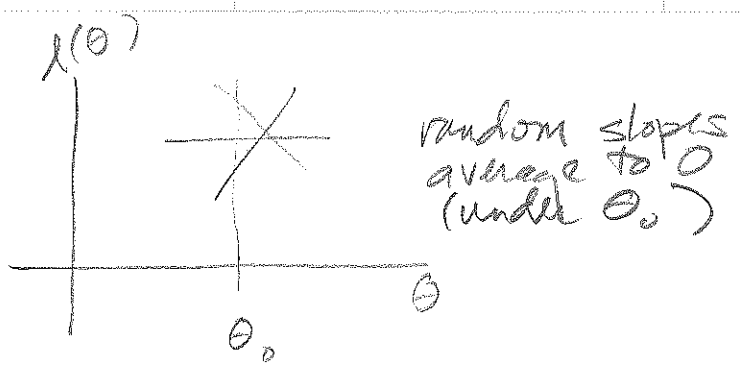
$$= \int \frac{d}{d\theta} f(x|\theta)\Big|_{\theta=\theta_0} dx$$

$$\overset{\text{AIP}}{=} \frac{d}{d\theta}\left(\int f(x|\theta)\,dx\right)\Big|_{\theta=\theta_0}$$

$$= \frac{d}{d\theta} 1\Big|_{\theta=\theta_0}$$

$$= 0$$

That is, the θ_0 mean of the score function is 0 - on (θ_0) average the derivative of the loglikelihood at θ_0 is 0 ...



The most informative loglikelihoods, if they are not maximum (and thus have 0 slope at θ_0) are climbing steeply or dropping steeply at θ_0, i.e. have big $|l'(\theta_0)|$ — so a measure of "information" about θ at θ_0 might be

$$\mathrm{Var}_{\theta_0}\, l'(\theta_0)$$

the θ_0 variance of the score function at θ_0 — if this is big $L(\theta)$ tends to be informative about θ at θ_0

Def (See Web page for $\theta \in \mathbb{R}^k$ version of this where the $\|$ concept is the covariance matrix of $\nabla \log L(\theta)$ at θ_0) In the case where $\Theta \subset \mathbb{R}^1$ and regularity conditions hold

$$I(\theta_0) = \mathrm{Var}_{\theta_0}\left(l'(\theta)\right)^2$$
$$= E_{\theta_0}\left(l'(\theta)\right)^2$$

is called the Fisher Information in X about θ evaluated at θ_0

Example $X \sim \mathrm{Binomial}((n,p))$

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\ell(p) = \log\binom{n}{x} + x \log p + (n-x)\log(1-p)$$

$$\ell'(p) = \frac{x}{p} - \frac{n-x}{(1-p)}$$

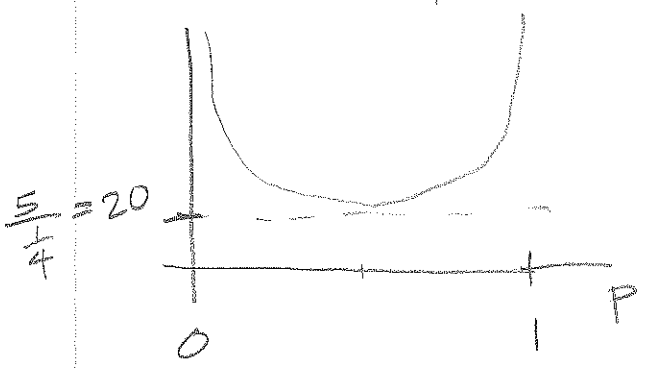
$$E_{p_0}\ell'(p_0) = \frac{np_0}{p_0} - \frac{n-np_0}{1-p_0} = n(1-1) = 0$$

$$E_{p_0}(\ell'(p_0))^2 = \mathrm{Var}_{p_0}\left(X\left(\frac{1}{p_0} + \frac{1}{1-p_0}\right) - \frac{n}{1-p_0}\right)$$

$$= \left(\mathrm{Var}_{p_0} X\right)\left(\frac{1}{p_0(1-p_0)}\right)^2$$

$$= \frac{n}{p_0(1-p_0)} = I(p_0)$$

To our collection of $Bi(5,p)$ plots we could add one of $I(p)$



There are a number of useful simple results about FI

Result (See handout for a careful and $\mathbb{R}^k$ version of this) Under appropriate regularity conditions

$$I(\theta_0) = -E_{\theta_0}(\ell''(\theta_0))$$

(The Fisher Information in X about θ at θ_0 is not only the variance of the slope of the log-likelihood at θ_0, but it is also the negative expected curvature of the log-likelihood at θ — The more curved the log-likelihood tends to be at θ_0, the more discriminating power one has to distinguish between θ_0 and θ's near θ_0, the better one's information about θ_0

"Pf" (Outline for the cont^s case)

$$l''(\theta) = \frac{d}{d\theta} l'(\theta) = \frac{d}{d\theta}\left(\frac{\frac{d}{d\theta} f(x|\theta)}{f(x|\theta)}\right)$$

$$= \frac{f(x|\theta)\frac{d^2}{d\theta^2} f(x|\theta) - \left(\frac{d}{d\theta} f(x|\theta)\right)^2}{(f(x|\theta))^2}$$

$$\text{So } E_{\theta_0} l''(\theta_0) = \int \frac{d^2}{d\theta^2} f(x|\theta)\Big|_{\theta=\theta_0} dx$$

$$- \int \frac{\left(\frac{d}{d\theta} f(x|\theta)\Big|_{\theta=\theta_0}\right)^2}{(f(x|\theta_0))^2} f(x|\theta_0)\, dx$$

$$\overset{\text{AIP}}{=} \frac{d^2}{d\theta^2} \int f(x|\theta)\, dx\Big|_{\theta=\theta_0} - E_{\theta_0}(l'(\theta_0))^2$$

$$= 0 - I(\theta_0)$$

Example $X \sim Bi(n, p)$

$$l''(p) = \frac{X}{p^2} - \frac{n-X}{(1-p)^2}$$

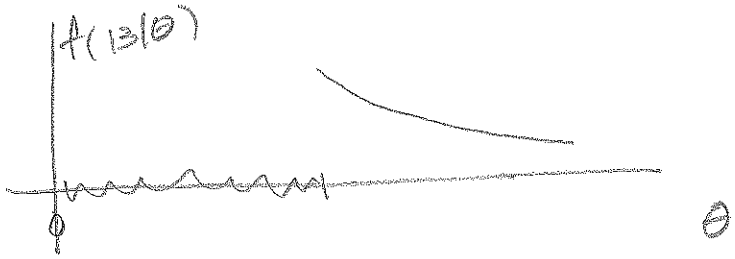
$$E_{p_0} l''(p_0) = \frac{np_0}{p_0^2} - \frac{n(1-p_0)}{(1-p_0)^2}$$

$$= -n\left(\frac{1}{p_0} + \frac{1}{1-p_0}\right)$$

$$= -\frac{n}{p_0(1-p_0)} = -I(p_0)$$

BTW, regularity conditions in a careful development of this stuff are meant to ensure that for all x the density $f(x|\theta)$ changes smoothly in θ at θ_0

e.g., they outlaw models like $U(0,\theta)$ dsn – for $x=13$ one has



$f(x|\theta)$ is not positive $\forall\theta$
There is a discontinuity in $f(13|\theta)$ at $\theta=13$
(so, obviously, $f(13|\theta)$ isn't differentiable at $\theta=13$)

A second useful property of the FI is that for models of independence, it is additive, i.e. if

$$X = (X_1, X_2, \ldots, X_n)$$

has dsn depending upon θ and $(\forall\theta)$ $X_1, X_2, \ldots, X_n$ are independent

$$I_X(\theta) = \sum_{i=1}^{n} I_{X_i}(\theta)$$

why? Take, for example, the iid case with marginal $f(x|\theta)$

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$\ell(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$$\ell'(\theta) = \sum_{i=1}^{n} \frac{\frac{d}{d\theta} f(X_i|\theta)}{f(X_i|\theta)}$$

for any θ a sum of iid r.v.'s under any θ_0

$$I_X(\theta_0) = \mathrm{Var}_{\theta_0}\, \ell'(\theta_0) = \sum_{i=1}^{n} \mathrm{Var}_{\theta_0} \left. \frac{\frac{d}{d\theta} f(X_i|\theta)}{f(X_i|\theta_0)} \right|_{\theta=\theta_0}$$

$$= \sum_{i=1}^{n} I_{X_i}(\theta_0)$$

$$= n\, I_{X_i}(\theta_0)$$

$$= n I_{X_1}(\theta_0)$$

Example $X_1, X_2, \ldots, X_n$ iid $\mathrm{Ber}(p)$

$X = (X_1, \ldots, X_n)$

Clearly $I_X(p) = \frac{n}{p(1-p)}$ and $I_{X_1}(p) = \frac{1}{p(1-p)}$

$$= n\, I_{X_1}(p)$$

Another (sensible) fact is that $I_{T(X)}(\theta) \le I_X(\theta)$

i.e. you can't increase FI by taking a function of X - "see the redundant" on the web page

Further $I_{T(X)}(\theta) = I_X(\theta) \quad \forall \theta$

$\iff T(X)$ is sufficient for θ

will be a HW problem

(provided all is well-defined, etc.)