**Exponential Families of Dsn**

**Def:** If \((T_1, T_2, \ldots, T_k)\) are linearly independent real-valued functions on \(\mathbb{R}^k\) or some discrete space, \(h(x) \geq 0\) and for parameters \(\eta \in \mathbb{R}^k\) \(f(x | \eta)\) 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 dsn's an exponential family.

**Fact:**  
\[
\mathcal{E} = \left\{ \eta \in \mathbb{R}^k \mid \int h(x) \exp \left( \sum_{j=1}^{k} \eta_j T_j(x) \right) \, dx < \infty \right\}
\]

or  
\[
\sum_{j=1}^{k} \eta_j T_j(x) < \infty
\]

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.

\[
(\mathcal{C}(\eta) = \left( \int h(x) \exp \left( \sum_{j=1}^{k} \eta_j T_j(x) \right) \, dx \right)^{-1}
\]

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

\[
f(x | \eta) = c(\eta) \prod_{i=1}^{n} h(x_i) \exp \left( \sum_{j=1}^{k} \eta_j \sum_{i=1}^{n} T_j(x_i) \right)
\]
and the factorization theorem says that

\[ T(X) = (\Sigma T_1(X_i), \Sigma T_2(X_i), \ldots, \Sigma T_k(X_i)) \]

is sufficient for any \( \mathcal{E} \subset \mathcal{E} \) - this is called the natural sufficient statistic for the parameter \( \theta \) (or family \( \mathcal{P} = \{ P_{\theta} \}_{\theta \in \mathcal{E}} \)).

Not only is the natural sufficient statistic sufficient, but provided \( \mathcal{E}^* \) is big enough, it is minimal sufficient - e.g. then is

Thus, 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

1) showing \( T(X) \) has a property called "completeness",
2) appealing to "Bahadur's Thm" (That says \( \mathcal{E}^* \)'s are minimal).

The argument for 1) 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 (\( \lambda \)) pmf on \( \{0, 1, 2, \ldots\} \) is
\[ f(x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \]
\[ = e^{-\lambda} \left( \frac{\lambda^x}{x!} \right) \]
\[ = e^{-\lambda} \left( \frac{\lambda^x}{x!} \right) \]
\[ = e^{-\lambda} \left( \frac{\lambda^x}{x!} \right) \]

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

\[ \eta(\theta) = (\eta_1(\theta), \ldots, \eta_k(\theta)) \]

\[ E \] is the part of \( \Theta \) that is the image of \( \Theta \) under the map \( \eta(\theta) \).

\[ T(X) \] sufficient for \( \eta \in E \)

\[ T(X) \] sufficient for \( \eta \in E_\Theta \)

\[ T(X) \] sufficient for \( \theta \in \Theta \)

**Back to Poisson Example**

For \( \lambda \in (0, \infty) \)

\[ \eta(\lambda) = \log \lambda \quad \text{(so that } \lambda = e^\eta) \]

\[ E = \{ \eta | \sum \frac{\eta^x}{x!} < \infty \} \subset \mathbb{R} \]
For \( n \) iid Poisson \( \lambda \) observations \( T(x) = \sum x_i \) is minimal sufficient for \( \lambda \in \mathbb{E} \).

Since the image of \((0,\infty)\) under the log transform is \( \mathbb{R} \), it contains an open interval \( 0 < \alpha < \beta < \infty \).

\[ T(x) \] is minimal sufficient for \( \lambda \in (0,\infty) \).

**Example. (Normal)**

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

for \( \eta_1 = \frac{1}{2\sigma^2}, \eta_2 = \frac{\mu}{\sqrt{2}} \)

\[ T_1(x) = x^2 \quad T_2(x) = x \]

This is an exponential family of denss and for \( X_1, X_2, \ldots, X_n \) iid \( N(\mu, \sigma^2) \) \( T(x) = (2x_1^2, 2x_2^2) \) is sufficient.

\[ \eta(\mu, \sigma^2) = \left( \eta_1(\mu, \sigma^2), \eta_2(\mu, \sigma^2) \right) \] maps \( \mathbb{R}^2 \times (0,\infty) \) to \( (-\infty,0) \times \mathbb{R}^1 \).

This whole thing is the image of \( \Theta \) under \( \eta \), since it contains an open rectangle. The natural sufficient statistic is minimal sufficient.
What does this "open rectangle" business allow?

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

\[
f(x; \mu) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (x - \mu)^2 - \frac{x^2}{\mu} - \frac{1}{2}\right)
\]

with \( y_1(\mu) = -\frac{1}{2\sigma^2} \) \( y_2(\mu) = \frac{1}{\mu} \)

This is some kind of an exponential family...

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

\( T(X) = (\sum X_i, \sum X) \) is sufficient for \( \mu \)

Here \( \mathbb{R} \) is really only "1-dimensional" and so also is the set of \( \eta = (y_1, y_2) \) under discussion here

i.e., since \( y_1 = -\frac{1}{2}y_2^2 \)

we have

so we can't immediately apply the theorem to get

minimality of \( T(X) \)

BTW notice that

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

is an exponential family with a genuinely 2-d \( \eta \)
2. \( \Omega = \{ \text{N}(\mu, \sigma^2) \text{ models } | \mu \in \mathbb{R}^k, \sigma^2 \in \mathbb{R}_+ \} \) has
\[
T(x | \mu) = \exp \left(-\frac{1}{2} x \mu + x \mu + \mu^2 \right)
\]
\[
T(x) = x
\]
and with \( \eta = \mu \) we have an exponential family with a 1-dimensional \( \eta \) and single \( T(x) \) - i.e.,

we think of this as a sub-model of the full \( \text{N}(\mu, \sigma^2) \) model. We have

\[
\begin{align*}
\eta_1 &= -\frac{1}{2} \\
\eta_2 &= \mu \\
\eta_3 &= \mu^2
\end{align*}
\]

3. The \( \text{N}(\mu, \sigma^2) \) example is something else – \( \eta \) is "1-dimensional," but it isn't a case where a single \( T(x) \) and 1-dimensional \( \eta \) will work.

people would call this a "curved exponential family" involving \( k \) linearly independent \( T_i(x) \) 's, \( \eta \) some \( k' < k \) - dimensional subset of \( \eta \) but not writable in terms of \( k' \) \( T_j(x) \) 's and \( k' \) - dimensional \( \eta \).
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 the $\mathbb{R}^k$ version of this) - consider

$$L(\theta) \text{ the likelihood}$$

$$l(\theta) = \log L(\theta) \text{ the loglikelihood}$$

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

Note that "usually"

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

$$= \left. \frac{d}{d\theta} f(x|\theta) \right|_{\theta=\theta_0} f(x|\theta_0) dx$$

$$= \left. \frac{d}{d\theta} f(x|\theta) \right|_{\theta=\theta_0} dx$$

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

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

$$= 0$$

That is, the $\theta_0$ mean of the score function is $0$ - on $(\theta_0)$ average. The derivative of the loglikelihood at $\theta_0$ is $0$ -

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

\[
\text{Var}_{\theta_0} l'(\theta_0)
\]

the \( \theta_0 \) variance of the score function at \( \theta_0 \) — if this is big \( l(\theta) \) tends to be informative about \( \theta \) at \( \theta_0 \).

\[\text{Det} \quad \text{(See Web page for \( \theta \in \mathbb{R}^k \) version of this)}
\]

where the \( \text{II} \) concept is the covariance matrix of \( \nabla \log l(\theta) \) at \( \theta_0 \). In the case where \( \Theta \subset \mathbb{R}^1 \) and regularity conditions hold

\[I(\theta_0) = \text{Var}_{\theta_0} (l'(\theta))^2 = E_{\theta_0} (l'(\theta))^2\]

is called the Frechet Information in \( X \) about \( \theta \) evaluated at \( \theta_0 \).

\[\text{Example} \quad X \sim \text{Bernoulli}(n, p)\]
\[ L(p) = \binom{n}{X} p^X (1-p)^{n-X} \]

\[ \ell(p) = \log(n) + X \log p + (n-X) \log(1-p) \]

\[ \ell'(p) = \frac{X}{p} - \frac{n-X}{1-p} \]

\[ \mathbb{E}_{p_0} \ell'(p_0) = \frac{n p_0}{p_0} - \frac{n-n p_0}{1-p_0} = n (1-1) = 0 \]

\[ \mathbb{E}_{p_0} (\ell'(p_0))^2 = \text{Var}_{p_0} \left( X \left( \frac{1}{p_0} + \frac{1}{1-p_0} \right) - \frac{n}{1-p_0} \right) \]

\[ = \left( \text{Var}_{p_0} X \right) \left( \frac{1}{p_0(1-p_0)} \right)^2 \]

\[ = \frac{n}{p_0(1-p_0)} = \mathbb{I}(p_0) \]

To our collection of Bi(5, p) plots we could add one of \( \mathbb{I}(p) \)

\[ \frac{n}{p_0(1-p_0)} = 20 \]

Thus are a number of useful simple results about PI

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

\[ \mathbb{I}(\theta_0) = -E_{\theta_0}(\ell''(\theta_0)) \]
The Fisher Information in $X$ about $\theta$ at $\theta_0$ is not only the variance of the slope of the log-likelihood at $\theta_0$, but it is also the negative expected curvature of the log-likelihood at $\theta$. The more curved the log-likelihood tends to be at $\theta$, the more discriminating power one has to distinguish between $\theta_0$ and $\theta$'s near $\theta_0$. The better one's information about $\theta_0$.

\[ P^\prime \] (Outline for the continuous case)

\[ l''(\theta) = -\frac{d^2}{d\theta^2} \log f(x|\theta) = \frac{d}{d\theta} \left( \frac{\frac{d}{d\theta} f(x|\theta)}{f(x|\theta)} \right) \]
\[ = \frac{f(x|\theta) \frac{d^2}{d\theta^2} f(x|\theta) - \left( \frac{\frac{d}{d\theta} f(x|\theta)}{f(x|\theta)} \right)^2}{(f(x|\theta))^2} \]

So \( E_{\theta_0} l''(\theta_0) = \int \frac{d^2}{d\theta^2} f(x|\theta) \bigg|_{\theta=\theta_0} \, dx \)
\[ = \int \frac{\left( \frac{d}{d\theta} f(x|\theta) \bigg|_{\theta=\theta_0} \right)^2}{(f(x|\theta_0))^2} f(x|\theta_0) \, dx \]
\[ \Theta \frac{d^2}{d\theta^2} \int f(x|\theta_0) \, dx \bigg|_{\theta=\theta_0} = E_{\theta_0} (l''(\theta_0))^2 \]

\[ = 0 - I(\theta_0) \]

Example \( X \sim \text{Bi}(n, p) \)

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

\[ E_{\theta_0} l''(p_0) = \frac{n p_0^2}{\theta_0^2} - \frac{n(1-p_0)}{(1-p_0)^2} \]
\[-n \left( \frac{1}{p_0} + \frac{1}{1-p_0} \right) \]
\[= -\frac{n}{p_0(1-p_0)} = -I(p_0)\]

BTW, regularity conditions in a careful development of this stuff are meant to ensure that for all \( \theta \) the density \( f(x|\theta) \) changes smoothly in \( \theta \) at \( \theta_0 \), e.g., they outlaw models like \( U(0, \theta) \) for \( \theta = 13 \) one has

\[f(13|\theta)\]

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 EI is that for models of independence, it is additive, i.e., if

\[X = (X_1, X_2, \ldots, X_n)\]

has \( \theta \) determined upon \( \theta \) 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

\[f(x|\theta)\]
\[ L(\Theta) = \prod_{i=1}^{n} f(X_i; \Theta) \]
\[ l(\Theta) = \sum_{i=1}^{n} \log f(X_i; \Theta) \]
\[ l'(\Theta) = \sum_{i=1}^{n} \frac{d}{d \Theta} f(X_i; \Theta) f(X_i; \Theta) \] for any \( \Theta \) a sum of iid r.v.'s under any \( \Theta_0 \)

\[ I_X(\Theta_0) = \text{Var}_{\Theta_0} l'(\Theta_0) = \sum_{i=1}^{n} \text{Var}_{\Theta_0} \frac{d}{d \Theta} f(X_i; \Theta_0) f(X_i; \Theta_0) \]
\[ = \sum_{i=1}^{n} I_{X_i}(\Theta_0) \]
\[ = n I_{X_i}(\Theta_0) \]
\[ = n I_X(\Theta_0) \]

Example: \( X_1, X_2, \ldots, X_n \) iid \( \text{Ber}(p) \)
\[ X = (X_1, \ldots, X_n) \]

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

\[ n I_{X_i}(p) \]

Another (sensible) fact is that \( I_{T(X)}(\Theta) \leq I_X(\Theta) \)

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

Further \( I_{T(X)}(\Theta) = I_X(\Theta) \ \forall \ \Theta \) will be a HW problem.

\( \iff T(X) \text{ is sufficient for } \Theta \)
(provided all is well-defined, etc.)